D2.3.ca.en
Language Report for Catalan
(English version)

Version 1.0

2011-07-27

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	27-07-2011	Asunción Moreno, Núria Bel, Eva Revilla, Emília García, Sisco Vallverdú	UPF and UPC	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.ca.en Language Report for Catalan (English version)

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

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

Responsible: Asunción Moreno (WP2 coordinator)

Contributing Partners: Universitat Politècnica de Catalunya; Universitat Pompeu Fabra

Authors: Asunción Moreno, Núria Bel, Eva Revilla, Emília García, Sisco Vallverdú

Collaborative authors: Lluís Padró, José Adrián R. Fonollosa, Joan Soler, Ignasi Esquerra, Mireia Farrus

Reviewer: Paul Thompson

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 Catalan language demonstrates that a lively language technology industry and research environment exist in Catalonia. Although a number of technologies and resources for Catalan exist, there are fewer technologies and resources for the Catalan language than for the Spanish or English languages.

According to the assessment detailed in this report, immediate action must occur before any breakthroughs for the Catalan 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 whitepapers focus on the readiness of core technologies in 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 Europeans 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.

Catalan in the European Information Society

General Facts

Catalan is part of the Romance family of languages. The Catalan language has around 8 million native speakers, and there are almost 12 million people who can speak it. It is the co-official language in three regions of Spain, i.e., Catalonia, the Balearic Islands and the Region of Valencia, and also spoken in some border villages of Aragón and Murcia. It is the only official language of Andorra and is also spoken in the French department of Pyrénées Orientales (known as North Catalonia), and in the Italian city of Alghero in Sardinia.

The status of Catalan is different according to the areas where it is spoken. In Catalonia, the majority of people are bilingual. Studies made during 2010 by the Catalan Studies Institute confirm that 95.3% of the population understand Catalan, 60.6% can write it, and 77.5% can speak it, but this latter number increases to 96.4% when restricted to people born in Catalonia. These figures are also confirmed by other studies, such as the Programme for International Student Assessment (PISA) 2010, which found that almost 80% of the population of Catalonia can read Catalan.

As will be explained later in this chapter, the Catalan language can be studied in several countries all over the world, especially in European and North American Universities.

Particularities of the Catalan Language

The Catalan language has five clearly distinguished dialects (i.e., Northern, Central, North-Western, Balearic and Valencian) with special normalisation rules. The dialects differ mainly in the pronunciation of certain vowels, the set of function words used (e.g., articles, possessives pronouns and other pronouns) and also in certain words of the lexicon.

Catalan uses eight different vowel sounds and thirty one consonant sounds. The alphabet uses 26 letters plus 2 additional ones, the ç('ce trencada') and the *l·l* ('ela geminada').

Concerning the word order of sentences or utterances in Catalan, the main pattern used is Subject, Verb, Object. Nevertheless, the word order of Catalan is relatively free, and it is not uncommon to find the use of clitic elements that change the basic structure. For example, the sentence: 'La Maria ens portà els regals a nosaltres' (Mary brought us the presents) can be also phrased as: 'A nosaltres ens portà els regals la Maria' or 'Els regals la Maria ens els portà'.

Catalan is a pro-drop language. This means that it is possible to use a conjugated verb without using the personal pronoun that plays the subject role.

One of the ways in which Catalan differs from languages such as French or English is that it is not possible to separate verb constructions involving an auxiliary verb. For example, in English it is possible to say: 'I had always done this'; or in French: 'j'avais toujours fait ça'. In Catalan, it is possible to say: 'jo sempre he fet això', or 'jo he fet sempre això', but 'jo he sempre fet això' is incorrect.

The lexical roots of Catalan mainly evolve from the Latin language. Among the most commonly spoken Romance languages, Italian and French are the closest to Catalan, from both a lexical and phonetic point of view. This means that Catalan speakers are able to understand Italian or French easily.

The orthography of Catalan is more transparent than that of English, but less so than that of Spanish or Italian. For example, the vowels *a*, *e* and *o*, are pronounced differently in some dialects, depending on whether or not they are on a stressed syllable. Additionally, *b* and *v* have identical pronunciations in many dialects. Stress marks and dieresis are used in Catalan to help to mark the stress and pronunciation of some words.

Recent developments

After the Spanish Civil War (1936-1939) and during Franco's dictatorship (1939-1975) the Catalan language and culture were intensely persecuted and discriminated against. The use of Catalan was prohibited in education, in the administration and in any media and dissemination system (books, newspapers, radio, television and cinema).

In spite of that, civil society managed to keep a significant cultural activity, often clandestine, which lead to a relaxation of the prohibition in the 1970s. From 1974 onwards several newspapers were published in Catalan, and from 1978 Catalan was allowed as education language.

"La Nova Cançó" is an artistic and cultural movement that claimed, in the late 1950s, the right to use normally the Catalan language. The group of singers "Els setze jutges" was created in 1959 within this cultural movement, and the first recordings in Catalan appeared in 1962.

Òmnium Cultural[1] was born in 1961 with the mission of protecting and promoting Catalan culture. It was founded at a moment in history when Catalan culture was censured and oppressed by Franco's dictatorship: there was thus a national need to ensure its recovery and continued survival. Given this, the organisation became a social tool and a key means of national resistance, taking the place of Catalan institutions, which were non-existent during the dictatorship. Nowadays, Òmnium creates debate, becoming involved and taking positions with regard to the key issues of today affecting Catalan society. It also promotes and normalises the Catalan language, culture and identity.

In 1976, Ràdio 4, a Spanish regional public radio station, began broadcasting exclusively in Catalan.

In 1983, TV3 started its first trial broadcast. TV3 is a television channel operated by the public Catalan Corporation of Media (CCMA)[2], which broadcasts all the programs only in Catalan.

The production of films in Catalan was reactivated from the second decade of the 1970s, and it was consolidated with the creation of TV3, both in terms of own production and inclusion of film subtitling and dubbing.

[1] http://www.omnium.cat
[2] http://www.ccma.cat

In 1985, the Catalan Government and the Institute of Catalan Studies established the TERMCAT[3], the centre for terminology in the Catalan language. Its mission is to ensure the development and integration of Catalan terminology into both specialist sectors and society in general.

Language cultivation in Catalan

According to Article 6 of the Statute of Autonomy of Catalonia, the own language of Catalonia is Catalan. Catalan is the language normally used in public administration, public media of Catalonia and teaching. Both Catalan and Spanish are official languages in Catalonia. The citizens of Catalonia have the right and duty to know both languages.

The Institute of Catalan Studies[4], founded in 1907 by Enric Prat de la Riba, has as main objective to promote high scientific research, mainly on all the elements related to Catalan culture. After returning to normal in the late 70s, the IEC was divided into five sections. The Philological Section plays the role of academy of the Catalan language. This function involves the scientific study of this language, the establishment of linguistic rules and monitoring the process of implementing this regulation in the geographical area where Catalan is used. The IEC also publishes the Dictionary of the Catalan language, whose second edition was released in 2007. This dictionary is also a useful instrument to nourish the Catalan language.

The Catalan Encyclopaedia[5] is a private non-profit project, born in 1965, that has become a reference in the publication and consultation on various topics in Catalan language, especially encyclopaedias and dictionaries.

The Institut Ramon Llull[6] was created in 2002 by the Catalan Government and the Government of the Balearic Islands. Its mission is to promote Catalan language and culture internationally, in all of its variations and methods of expression, as well as teaching outside of its linguistic area. It has two headquarters, Barcelona and Palma, and office buildings in Berlin, London, New York and Paris.

The Consorci per a la Normalització Lingüística[7] is an entity created from the common will of the Catalan Government and many local councils in order to facilitate understanding, use and dissemination of the own language of Catalonia in all areas. One of its main functions is to provide non-Catalan speakers with teaching and support.

Language in Education

In Catalonia, from the late 1960s onwards, around 80 schools, created as cooperatives by parents or teachers, were the pioneers in restoring the use of Catalan in education. These schools were in-

[3] http://www.termcat.cat/
[4] http://www.iec.cat/
[5] http://www.enciclopedia.cat/
[6] http://www.llull.cat
[7] http://www.cpnl.cat/

spired by the pedagogical tradition existing before the Spanish Civil War (1936) and followed Maria Montessori's method.

With the restoration of democracy following the dictator's death in 1975, the 1978 Spanish Constitution recognized the linguistic plurality of the state. The 1979 Statute of Catalonia and the 1983 Statute of the Balearic Islands recognized Catalan as their own and official language, as well as Spanish. The 1982 Statute of the Comunitat Valenciana recognized its status as official language with the legal name of Valencian.

At that time, after years of exclusion from education, Catalan was in a clearly disadvantaged state in comparison to Spanish. In order to address this situation, the various autonomous governments adopted different strategies.

In Catalonia, the strategy adopted in 1983 was the so-called *language immersion*, which was inspired by a programme carried out in Quebec (Canada) to deal with issues of language contact similar to those faced in the Catalan-speaking regions of Spain. The model was based on the idea that children should not be segregated according to their native language, because this would create two different school models: one for Catalan-speaking children and another one for Spanish-speaking children. Using the language immersion strategy, children are schooled totally in Catalan, independently of the language they speak at home, and learn to read and write in this language. When children begin to master this language, Spanish is gradually introduced into the curriculum. In this way, when compulsory education finishes, students have an equivalent mastery of both Catalan and Spanish, and are bilingual and biliterate, as several research studies indicate[v].

In the Balearic Islands, the autonomous authorities adopted a language immersion programme similar to that of Catalonia, whereas in the Comunitat Valenciana, the model adopted established different types of centres and programmes according to the students' native language.

In France, in the Department of Pyrénées Orientales (in Languedoc-Roussillon region) the situation regarding Catalan is far worse than in the Catalan-speaking regions of Spain. Although Catalan is the native language of some proportion of the population of this department, the instructional language used at school is French. In 1976, La Bressola[vi] was created in Perpignan, as a network of 8 schools that adopted the language immersion programme in Catalan, as a means to recover the use of this language in the everyday lives of people in the region. Besides this initiative, there are also some bilingual schools in the region, in which the schooling is carried out in both French and Catalan.

The last PISA study, conducted in 2009, revealed that students in Catalonia, with a mean score of 498, performed above the OECD mean score (494) and the mean score of Spain (481) in relation to reading literacy[vii]. This means that the fact that children are schooled following the language immersion programme, in which Catalan is the main medium of instruction, does not affect their reading literacy performance in Spanish.

However, the PISA study also shows that there is a striking difference between the scores obtained by native students (Catalan- or Spanish-speakers) and those obtained by students with a migration background. These results have reinforced public awareness re-

garding the importance of language learning, focussing especially on social integration.

At the beginning of the 21st century, there is a new challenge for Catalan schools, i.e., the schooling of a large number of students with a migration background. Unlike in the 1960s, when the newcomers to Catalan-speaking regions were mainly Spanish-speakers, it is now the case that children come from many countries all over the world, and speak many different languages. To address this situation, the government has created "reception classrooms" (*aules d'acollida*). Inspired by the language immersion programme, these "reception classrooms" are conceived as a temporary support for the newcomers, while they are in the process of acquiring the minimal skills for communication with their classmates.

International aspects

Catalan is one of the so-called minority languages and it has been recognized as such by the Council of Europe in the European Chapter for Regional or Minority Languages, which "aims to protect and promote the historical regional or minority languages of Europe". The importance of these languages is attested by the fact that they are spoken in total by more than forty million citizens in the EU.

As a minority language, Catalan was represented in the European Bureau for Lesser Used Languages, which was set up in 1982 on the initiative of the European Parliament. The aim of this pan-European non-governmental organisation has been to encourage respect towards lesser protected languages within the EU, and to promote linguistic diversity. Catalan is also one of the languages dealt with in the Mercator Network[viii], a network of three research and documentation centres, whose main objective is to become a specialized resource centre and an information service dealing with European minority languages. Mercator has three branches, each one focussing on a thematic programme, i.e, education, legislation and media.

Taking into account all the languages spoken in Spain, only Spanish has the status of an official language in the EU. However, in November of 2004, the Spanish government delivered to the EU the translation of the European Constitution into the languages of the state which are also official in their respective territories: Catalan (with the name Catalan when used in Catalonia and the Balearic Islands, and the name Valencian when used in the Comunitat Valenciana), Galician and Basque.

In 2005, the Council of Ministers recognised the possibility of using official languages other than Spanish in European institutions. After signing administrative agreements with some EU institutions, recognising a limited use of Catalan, the status of Catalan is currently that of a semi-official language, i.e. a language of communication among citizens. This status means that citizens can write in Catalan to these institutions (European Commission, European Parliament, Council, European Ombudsman and Committee of the Regions), and, in turn, they have the right to be answered in the same language. Some publications and official documentation are also translated into Catalan. Moreover, in the European Commission's Representation in Barcelona, Catalan is used as the usual language of communication with citizens (information campaigns, publishing, press releases and websites).

The international projection of Catalan is rather limited. In the business world at international level, the use of Catalan is non-existent. In fact, English has become the common language of communication on both written and oral levels. A small number of large international companies are now using Catalan to deal with their Catalan customers, in order to add value to their products and to improve their customer services. These companies include Microsoft, IKEA and Toshiba.

In terms of opportunities for learning Catalan as a foreign language, the situation is a little better. The European Commission is developing an active policy on multilingualism, which aims at preserving and promoting linguistic diversity in Europe, fostering language learning (including regional and minority languages) and using multilingualism as a stimulus for competitiveness. In this context, the Lifelong Learning Programme 2007-13 contains a selection of projects promoting language learning. Among them, the Lingu@net Europa Plus multilingual online languages resource centre[ix] provides support and resources in 20 European languages, including Catalan. In addition, an important decision made by representatives of the EU Member States has been to include Catalan, as well as Basque and Galician, in the list of languages offered in the Erasmus Intensive Language Courses from the academic year 2010-2011[x]. These EU-funded language courses aim to prepare prospective Erasmus students for their study period in Catalan universities, where this language is used as a communication and academic language.

Despite being a minority language, the interest in learning Catalan in foreign universities is attested by the fact that, at present (academic year 2010-11), more than 160 universities all over the world offer Catalan studies[xi]. Some of the countries offering the widest choice of courses are Germany (26), the USA (23), the UK (21), France (20) and Italy (17). Catalan can be studied in 11 Spanish universities. The Ramon Llull Institute (IRL)[xii], a consortium consisting of the Governments of Catalonia and the Balearic Islands, has the aim of promoting the Catalan language and its culture internationally. The IRL is part of the Fundació Ramon Llull, set up by the Government of Andorra, the IRL, the General Council of the Eas-tern Pyrenees, the city of Alghero and the Network of Valencian Cities. This foundation involves governments and institutions from the Catalan linguistic domain, in an attempt to unify efforts for a common purpose.

The interest of the Catalan government and institutions in the international projection of the language and culture is also mirrored in the organisation of different international events where Catalan has been the main guest or guest of honour, such as the Frankfurt Book Fair 2007[xiii], the Catalan Days 2009 (New York)[xiv] or Expolangues 2010 (Paris)[xv]. Also, in the literary world, the PEN Català[xvi], which was founded in 1922 (only one year after the foundation of the PEN International by C.A. Dawson Scott), has been a platform for the international projection of Catalan literature and the writers of the Catalan linguistic domain.

Catalan on the Internet

Contrary to what might be expected of a minority language (after all, Catalan occupies the position 75 in the Ethnologue[xvii] classification of languages by language size), when it comes to its presence on the Internet, the situation is radically different. According to Luís Collado, the person responsible for Google Books and Google News in Spain and Portugal, Google places the Catalan language

among the 10 to 15 most active languages of the world on the web. Google considers Catalan a language with an activity that goes beyond the borders of its linguistic domain.

The WICCAC[xviii] association, which gathers together independent webmasters from the Catalan linguistic domain who have created t websites in this language, publishes a monthly barometer on the use of Catalan on the Internet. The barometer is updated by surveying the websites of companies and organisations located in Catalonia or other places within the linguistic domain. The survey also includes the websites of companies and organisations which are located outside of the domain, but which offer their products and services within the domain. The latest update (April 2011) shows that the global percentage of the use of Catalan is 59.88% (medium). This percentage has been slowly but steadily growing since the first barometer in August 2002 (40.71%).

This presence of Catalan on the web is due, on one hand, to the attitude of public institutions that foster the normalisation of the use of this language on the Internet, and, on the other hand, to private initiatives of organisations and people who are very committed to their language and culture.

It is worth mentioning the significance and strength of these private initiatives, which have placed Catalan among the most active languages on the web. Currently, for instance, Viquipèdia (the Catalan Wikipedia), with 337.514 articles, is the 13th largest Wikipedia in terms of the number of articles. Another example is the puntCAT Foundation[xix], which has launched an Internet campaign to promote navigation in Catalan by helping users to configure their navigators and make Catalan their default navigation language. At the time of its creation in 2004, the main goal of the puntCAT Foundation was the promotion of all kinds of activities with regard to the creation, management and control of the registry of the .cat domain name. Accordingly, in 2005, the ICANN approved the creation of the .cat sponsored top-level domain, which was intended to be used to promote Catalan language and culture. Nowadays, the number of websites using this domain is about 50.000[xx].

The fact that Google, YouTube or Facebook, among others, offer a Catalan version of their navigation interfaces is explained, at least partly, by the growing importance of the Catalan-speaking community on the web.

Another very significant initiative, as regards the use of Catalan in ICT, is Softcatalà[xxi], an association whose main aim is to foster the use of this language in computer science, on the Internet and in ICT. Softcatalà is based on the volunteer work of students, professionals and users (computer scientists, philologists, designers, translators), who develop, translate and distribute software in Catalan, such as navigators, Internet tools, office automation software, multimedia software, games, etc. The monthly mean of unique visitors has grown considerably between 2006 (285.186 visitors) and 2011 (625.296 visitors), as well as the monthly mean of total visits (689.142 in 2006 vs. 1.366.644 in 2011). In 2010, the total software downloads numbered approximately 816.000, among them the Catalan versions of OpenOffice (224.796) and Mozilla Firefox (74.212).

The web also offers a growing number of digital newspapers in Catalan, as well as some online courses to learn the language.

All in all, this significant Internet presence suggests that there is a vast amount of Catalan language data available on the web.

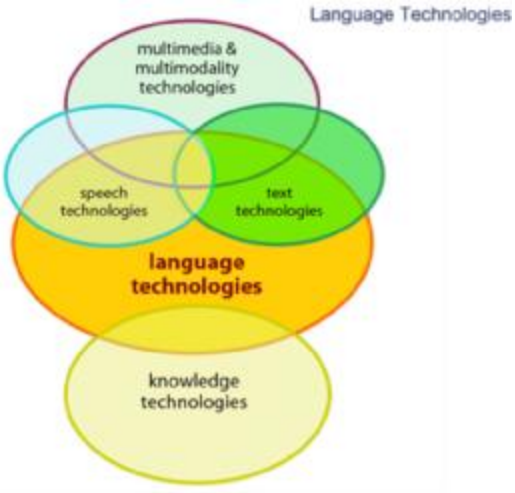
According to these data, we can conclude that, despite being rather small, the Catalan-speaking community is very active and committed to its language and culture, so that it is not cut off from digital communication. Thus, people want to exercise their right to use their own language on the web at all levels, either when searching for content or when creating content.

Selected Further Reading

Language Technology Support for Catalan

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, replacing "5è" by "cinquè" for Catalan, etc.
- Grammatical analysis: finding the verb and its objects, modifiers, etc.; detecting the sentence structure.
- Semantic analysis: disambiguation (Which meaning of *apple* is the right one in a 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 lookups 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 Catalonia.

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 Catalan.

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.'

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

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, *llibre anglès* is a much more probable word sequence than *llibre anglesa*. 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 Catalan with its flexible word order and richer inflection.

The use of Language Checking is not limited to word processing tools. 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.

Only a few companies and Language Service Providers offer products in this area for Catalan. Enciclopèdia Catalana, Maxigrammar and Inèdit have created and commercialised products that include spell and grammar checking for Catalan, as well as specific checking facilities adapted to different domains and styles. Softcatalà and Barcelona Media have also developed language tools that are offered to the community as web applications. A new version of "El Corrector" has recently been developed and commercialised as iPod, iPhone and iPad apps.

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[xxii].

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 facility for misspelled words, including Catalan, and, in 2009, they incorporated basic semantic search capabilities into their algorithmic mix[xxiii], 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, or even more loosely related terms. Again, these developments require language specific resources. A Catalan WordNet has been developed by the research centre TALP at Universitat Politècnica de Catalunya. The Catalan WordNet is freely available[xxiv].

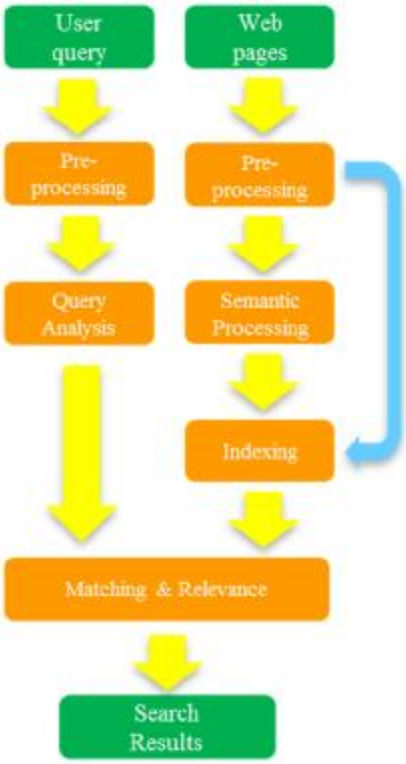


Figure 4: Web Search Architecture

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 recognizers.

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.

Small and Medium Enterprises such as Inbenta (www.inbenta.com) or RightNow (formerly q-go, www.q-go.es) offer specific semantic search engines, including those that are developed for Catalan.

There are also some interesting initiatives to group specific web search services for Catalan, such as http://www.cercat.cat/ or http://som-hi.com/, as one of the firs initiatives. An historical overview can be found at: http://www.gencat.cat/diue/doc_un/cis02_partal_uk.pdf

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 smart phones.

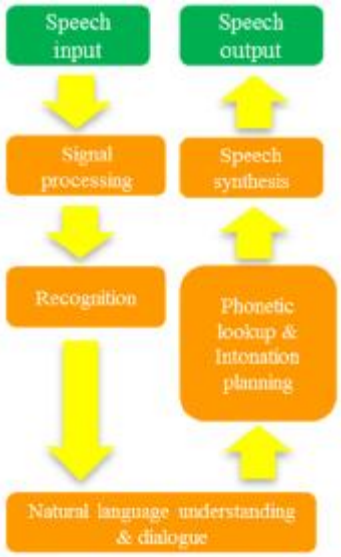


Figure 5: Simple Speech-based Dialogue Architecture

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 given system's purpose.
- 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 system's functionality.
- 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 has been characterised by 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, also for Catalan, although some smaller local companies are starting to compete, such as Verbio[xxv], which is a spin-off of Universitat Politècnica de Catalunya and has its own speech technology.

Regarding dialogue management technology and know-how, markets are strongly dominated by national players, which are usually SMEs. Most of the companies on the Catalan TTS market are essentially application developers. Key players in the Spanish market are: Indsys[xxvi] (Intelligent Dialogue Systems), Fonetic[xxvii], Ydilo[xxviii] and NaturalVoz[xxix].

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 or for closely related languages. 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

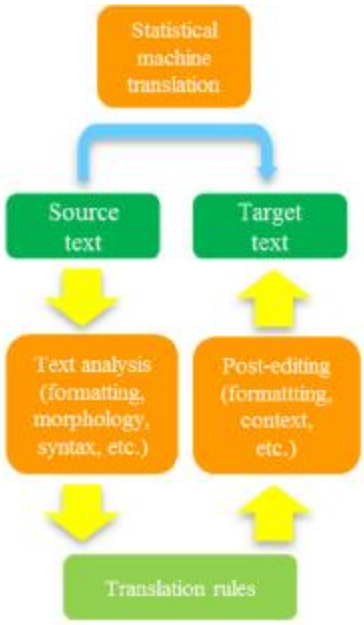


Figure 6: Machine translation (top: statistical; bottom: rule-based)

multiple levels, e.g., word sense disambiguation on the lexical level ('Jaguar' can mean a car or an animal) or the attachment of prepositional phrases on the syntactic level as in:

Passejava amb el nen cantant.

T1: I was walking with the singer child.

T2: I was walking and singing with the child

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 the 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, 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.

Due to the particular official situation of bilingualism in particular regions of Spain, and to the linguistic relatedness of Catalan with Spanish, the development of MT systems for this pair of languages has been quite successful. Initially supported by the autonomous Catalan leading MT systems like METAL (Siemens), ATLAS (Fujitsu) were located in Barcelona during the 1990's, and when the big companies ended their engagement, the systems were further developed by offspring and spin-off companies: METAL was further developed by a local SME, INCYTA, as well as by GMS and later by Lucy Software, which is currently the main (but not the only) vendor for MT systems involving Catalan. In fact, the Catalan-Spanish MT system bought by the Generalitat of Catalonia was offered as a public web service as early as 2005, while Google started offering it in 2007.

Other companies also developed MT systems with in-house technologies: T6 Estàndar Linguistic and AutomaticTrans. The system

developed by AutomaticTrans has its origin in the production of a bilingual newspaper, in Catalan and Spanish, El Periódico. Currently, there are three newspapers which are made available in the two languages through the use of Machine Translation technology. The other two are El Segre and La Vanguardia.

The Generalitat Valenciana supported the creation of a MT system specific for Valencian, SALT. More recently, the Universidad Politécnica de Valencia has also released a version of SiShiTra, a hybrid system, and the open-source OpenTrad also offers a particular version for Valencian.

The pair Spanish-Catalan was the origin of the open source system developed by the Transducens group of the Universitat d'Alacant, Apertium, which is the first open-source MT technology in the world. Its commercial exploitation is mainly carried out by Prompsit Language Engineering and OpenTrad Consortium.

Most of the systems introduced above are rule-based. While there is significant research on this technology in national and international contexts, data-driven and hybrid systems have been less successful in business than in research so far.

Provided with good adaptation in terms of user-specific terminology and workflow integration, the use of MT can increase productivity significantly. Thus, the main user of MT for Catalan is the public administration, including the Justice Department and other Ministries, since the beginning of the 21st century.

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, most of the existing parallel corpora are between Spanish and Catalan, so for the majority of translation solutions, Spanish is the pivot language.

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.

The situation in all these research areas for most of the languages 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 Catalan has never been prominent. Accordingly, there are hardly any annotated corpora or other resources for Catalan that relate to 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. Although it has not yet acquired a fixed place in any of the faculty

systems of the universities in the Catalan linguistic domain, Barcelona has a higher concentration of studies that consider the field.

Several subjects related to language technology are offered in bachelor degrees by different departments, such as the faculty of computer science (e.g., in the Universitat d'Alacant) or the faculty of linguistics (e.g., in the Universitat de Barcelona, the Universitat PompeuFabra and the Universitat Oberta de Catalunya).

As regards masters courses and postgraduate degrees, Universitat Autònoma de Barcelona offers the International Master in Natural Language Processing and Human Language Technology, in collaboration with foreign universities, and Universitat Politècnica de Catalunya runs the European Master's Degree in Language and Speech. Besides this, some masters are offered in which one of the research areas focusses on natural language processing (e.g., the Universitat de Barcelona, the Universitat Pompeu Fabra, the Universitat de Girona and the Universitat Rovira i Virgili). Modules in Language Technology are also offered to students of other master courses (e.g. the Universitat d'Alacant, the Universitat de Castelló and the Universitat Politècnica de València).

PhD programmes are also offered, in which one of the research areas is human language technologies (e.g. the Universitat d'Alacant, the Universitat de Barcelona and the Universitat Rovira i Virgili).

Language Technology Programs

Technology programs for the Catalan language have been supported by the Spanish and the Catalan Governments.

The existence of comparably lively LT activity, specifically in the Barcelona area, can be traced back to major LT programs and large research projects carried out in the last decades. One of the first programs was EUROTRA, the ambitious Machine Translation (MT) project established and funded by the European Commission from the late 1970s until 1994. Even though the EUROTRA project did not work with Catalan, the project, which had a long-term impact on language industries in Europe, was also crucial to highlighting the importance of languages in the technological world, and how it could affect small to medium languages. The Catalan government quickly began to understand the challenges and initiated different programs that basically addressed the localization of differrent software tools, including LT applications.

Machine translation, speech recognition, spelling and grammar checking research and industrial developments have been supported by different departments of the Generalitat de Catalunya for more than 20 years. The Secretaria de Política Lingüística, the Comissionat per a la Societat de la Informació and the Secretaria de Telecomunicacions i Societat de la Informació have been the main engines of the support policies. Besides, MT into or from Catalan has also benefited from Spanish funding programs. Projects such as Apertium (open source MT system) and OpenTrad, as well as a number of other small programs, have received funding from the Ministerio de Ciencia y Tecnología.

The CREL – Centre de Referència en Enginyeria Lingüística, 1996-2000, managed by the Institut d'Estudis Catalans and with participants from the major Catalan Universities, was created with the specific aim of promoting the creation of resources and tools for

the automatic processing of Catalan texts in a variety of applications.

As regards the presence of the Catalan language in European infrastructures, in 2008 the Catalan Government signed an agreement with Universitat Pompeu Fabra, the national representative of the European project CLARIN in Spain, with the aim of building a Catalan demonstrator. The main goal of this demonstrator (CLARIN-CAT-LAB), which is already available for research purposes[xxx], is to integrate language resources and technology for the Catalan language, thus guaranteeing the presence of this language in the European CLARIN infrastructure. In addition, the Biblioteca de Catalunya, Catalonia's national library, is one of the partners in the EUROPEANA project. Other Catalan institutions are also contributing content to the project.

From 2000 up until the present day, the Spanish Government supported supported several projects in the area of multilingual speech technologies within the National Plan of Research and Technology, i.e., TEHAM, AVIVAVOZ, and BUCEADOR. The main purpose of these projects was to improve the quality of speech recognition, speech translation and text to speech synthesis in all the official languages spoken in Spain, i.e., Basque, Galician, Catalan and Spanish.

In 2005, the Catalan Government launched a project to produce Language Resources for Speech Recognition and Speech Synthesis. As a result, Language Resources for telephone applications, in-car applications and microphone applications were produced. Later, the project TECNOPARLA (2007-2010) had as its objective the translation of speech between Spanish and Catalan. The speech signals were collected directly from TV programmes. The project resulted on advances in all the component speech technologies, i.e. diarization, speech recognition, speech translation and text to speech synthesis.

Availability of Tools and Resources for Catalan

The following table provides an overview of the current situation of Language Technology support for Catalan. The rating of existing tools and resources is based on educated estimations by several leading experts using the following criteria (each ranging from 0 to 6).

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	4	5	5	5	4	4
Parsing (shallow or deep syntactic analysis)	2	3	3	3	3	3	2
Sentence Semantics (WSD, argument structure, semantic roles)	1	3	2	2	2	2	2
Text Semantics(co-reference resolution, context, pragmatics, inference)	1	1	2	1	1	1	1
Advanced Discourse Processing (text structure, coherence, rhetorical structure/RST, argumentative zoning, argumentation, text patterns, text types etc.)	1	1	2	2	1	1	1
Information Retrieval (text indexing, multimedia IR, crosslingual IR)	3	1	3	1	3	2	2
Information Extraction (named entity recognition, event/relation extraction, opinion/sentiment recognition, text mining/analytics)	2	2	2	1	2	2	2
Language Generation (sentence generation, report generation, text generation)	1	2	3	1	3	3	1
Summarization, Question Answering, advanced Information Access Technologies	0	0	0	0	0	0	0
Machine Translation	3	3	4	3	4	3	2
Speech Recognition	3	3	3	3	3	3	2
Speech Synthesis	4	2	4	4	5	4	2
Dialogue Management (dialogue capabilities and user modelling)	1	6	4	2	3	3	3
Language Resources (Resources, Data, Knowledge Bases)							
Reference Corpora	3	3	4	3	3	3	3
Syntax-Corpora(tree banks, dependency banks)	3	2	3	3	3	2	2
Semantics-Corpora	2	1	1	1	1	1	1
Discourse-Corpora	1	6	2	2	3	3	3
Parallel Corpora, Translation Memories	2	1	3	2	3	1	1

	Quantity	Availability	Quality	Coverage	Maturity	Sustainability	Adaptability
Speech-Corpora (raw speech data, labelled/annotated speech data, speech dialogue data)	3	5	4	3	5	4	4
Multimedia and multimodal data (text data combined with audio/video)	1	4	2	2	3	3	3
Language Models	1	1	3	2	4	4	3
Lexicons, Terminologies	3	2	4	3	4	4	3
Grammars	2	3	2	2	2	2	2
Thesauri, Word Nets	2	2	4	2	2	2	2
Ontological Resources for World Knowledge (e.g. upper models, Linked Data)	2	2	3	2	2	2	2

Conclusions

The table can be summarized in the form of a number of key messages, which highlight crucial issues for the further development of automatic language processing of Catalan, on the basis of the present situation:

- While some specific corpora of high quality exist, a very large syntactically annotated corpus is not available.
- For Catalan, a large corpus exists, but it is not easily/cheaply accessible.
- Many of the resources lack standardisation, i.e., even if they exist, sustainability is not given; concerted programs and initiatives are needed to standardise data and interchange formats.
- Semantics is more difficult to process than syntax; text semantics is more difficult to process than word and sentence semantics.
- The more semantics a tool takes into account, the more difficult it is to find the right data; more efforts for supporting deep processing are needed.
- Standards do exist for semantics in the sense of world knowledge (RDF, OWL, etc.); they are, however, not easily applicable to NLP tasks.
- Speech processing is currently more mature than NLP for written text.
- There are many groups working in machine translation, particularly focused in Catalan-Spanish.
- Research has been successful in designing particular high quality software, but it is nearly impossible to come up with sustainable and standardised solutions, given the current funding situations.

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)

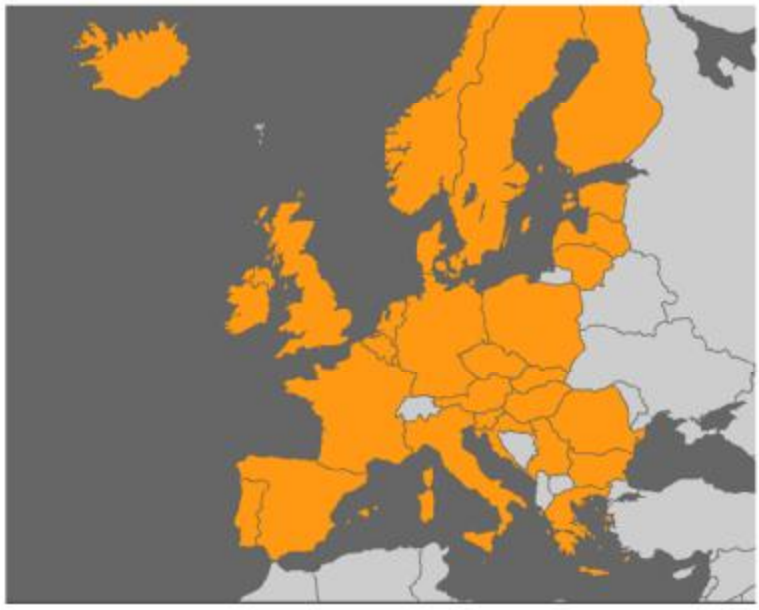


Figure 1: Countries Represented in META-NET

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 deve-lopers, 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	ConsiglioNazionaleRicerche, 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, *Userlanguage 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 theEU*, Kingston Upon Thames, 2009 (http://ec.europa.eu/dgs/translation/publications/studies).

[v] The educational authorities in Catalonia belonged to The Network of European Language Planning Boards (http://www.languageplanning.eu/home/Pages/index.aspx)

[vi] http://www.bressola.cat/index.php

[vii] http://stats.oecd.org/PISA2009Profiles/#

[viii] http://www.mercator-central.org/

[ix] http://www.linguanet-europa.org/plus/ca/home.jsp

[x] http://ec.europa.eu/education/news/news1518_en.htm

[xi] http://www.llull.cat/_eng/_cultura/cultura_catalana_mapa.shtml?seccio=cultura&subseccio=mapa

[xii] http://www.llull.cat/_eng/_home/index.cfm?seccio=inici&ampsubseccio=1

[xiii] http://www.frankfurt2007.cat/

[xiv] http://www.llull.cat/monografics/catalandays/

[xv] http://www.llull.cat/monografics/EXPOLANGUEs/index.cfm

[xvi] http://www.pencatala.cat/

[xvii] http://www.ethnologue.com/ethno_docs/distribution.asp?by=size

[xviii] http://wiccac.cat/index.php

[xix] http://www.navegaencatala.cat/

[xx] http://www.domini.cat/

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

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

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

[xxiv] http://nlp.lsi.upc.edu/web/index.php?option=com_docman&Itemid=135

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

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

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

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

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

[xxx] http://clarin-cat-lab.org/