|  |
| --- |
|  |
|  |
| META-NET White Paper Series |
| Languages in the European Information Society |
| – Spanish – |
|  |
|  |

|  |  |  |
| --- | --- | --- |
| flag_2colors.jpg | FP7-gen-CMYK.jpg | logo_ict_psp.2.gif |

The development of this white paper has been funded by the Seventh Framework Programme and the ICT Policy Support Programme of the European Commission under contracts T4ME (Grant Agreement 249119), CESAR (Grant Agreement 271022), METANET4U (Grant Agreement 270893) and META-NORD (Grant Agreement 270899).

# Preface

This white paper is part of a series that promotes knowledge about language technology and its potential. It addresses educators, journalists, politicians, language communities and others.

The availability and use of language technology in Europe varies between languages. Consequently, the actions that are required to further support research and development of language technologies also differ for each language. The required actions depend on many factors, such as the complexity of a given language and the size of its community.

META-NET, a Network of Excellence funded by the European Commission, has conducted an analysis of current language resources and technologies. This analysis focused on the 23 official European languages as well as other important national and regional languages in Europe. The results of this analysis suggest that there are many significant research gaps for each language. A more detailed expert analysis and assessment of the current situation will help maximise the impact of additional research and minimize any risks.

META-NET consists of 47 research centres from 31 countries that are working with stakeholders from commercial businesses, government agencies, industry, research organisations, software companies, technology providers and European universities. Together, they are creating a common technology vision while developing a strategic research agenda that shows how language technology applications can address any research gaps by 2020.

Contact

META-NET

DFKI Projektbüro Berlin

Alt-Moabit 91c

10559 Berlin

Germany

office@meta-net.eu

http://www.meta-net.eu

Author

Dr. Maite Melero, Barcelona Media Centre d‘Innovació

Table of Contents

Preface 3

Executive Summary 6

Risk for Our Languages and a Challenge for Language Technology 9

Language Borders Hinder the European Information Society 10

Our Languages at Risk 10

Language Technology is a Key Enabling Technology 11

Opportunities for Language Technology 11

Challenges Facing Language Technology 12

Language Acquisition in Humans and Machines 12

Spanish in the European Information Society 15

General Facts 15

Particularities of the Spanish Language 15

Recent Developments 16

Language Cultivation in Spain 17

Language in Education 19

International Aspects 19

Spanish on the Internet 21

Language Technology Support for Spanish 23

Application Architectures 23

Core Application Areas 24

Language Checking 24

Web Search 25

Speech Interaction 26

Machine Translation 28

Other Application Areas 30

Educational Programmes 32

National Projects and Efforts 33

Availability of Tools and Resources 33

Cross-language comparison 35

Conclusions 37

About META-NET 39

Lines of Action 39

Member Organisations 41

References 44

# Executive Summary

**Language technology builds bridges for Europe’s future**

During the last 60 years, Europe has become a distinct political and economic structure, yet culturally and linguistically it is still very diverse. This means that from Portuguese to Polish and Italian to Icelandic, everyday communication between Europe’s citizens as well as communication in the spheres of business and politics is inevitably confronted by language barriers. The EU’s institutions spend about a billion euros a year on maintaining their policy of multilingualism, i.e., translating texts and interpreting spoken communication. Yet does this have to be such a burden? Modern language technology and linguistic research can make a significant contribution to pulling down these linguistic borders. When combined with intelligent devices and applications, language technology will in the future be able to help Europeans talk easily to each other and do business with each other even if they do not speak a common language.

Language barriers can bring business to a halt, especially for SMEs who do not have the financial means to reverse the situation. The only (unthinkable) alternative to this kind of multilingual Europe would be to allow a single language to take a dominant position and end up replacing all other languages.

One classic way of overcoming the language barrier is to learn foreign languages. Yet without technological support, mastering the 23 official languages of the member states of the European Union and some 60 other European languages is an insurmountable obstacle for the citizens of Europe and its economy, political debate, and scientific progress.

The solution is to build key enabling technologies. These will offer European players tremendous advantages, not only within the common European market but also in trade relations with third countries, especially emerging economies. To achieve this goal and preserve Europe’s cultural and linguistic diversity, it is necessary to first carry out a systematic analysis of the linguistic particularities of all European languages, and the current state of language technology support for them. Language technology solutions will eventually serve as a unique bridge between Europe’s languages.

**Language technology as a key for the future**

The automated translation and speech processing tools currently available on the market still fall short of this ambitious goal. The dominant actors in the field are primarily privately-owned for-profit enterprises based in Northern America. Already in the late 1970s, the EU realised the profound relevance of language technology as a driver of European unity, and began funding its first research projects, such as EUROTRA. At the same time, national projects were set up that generated valuable results but never led to concerted European action. In contrast to this highly selective funding effort, other multilingual societies such as India (22 official languages) and South Africa (11 official languages) have recently set up long-term national programmes for language research and technology development.

The predominant actors in LT today rely on imprecise statistical approaches that do not make use of deeper linguistic methods and knowledge. For example, sentences are automatically translated by comparing a new sentence against thousands of sentences previously translated by humans. The quality of the output largely depends on the amount and quality of the available sample corpus. While the automatic translation of simple sentences in languages with sufficient amounts of available text material can achieve useful results, such shallow statistical methods are doomed to fail in the case of languages with a much smaller body of sample material or in the case of sentences with complex structures. The European Union has therefore decided to fund projects such as EuroMatrix and EuroMatrixPlus (since 2006) and iTranslate4 (since 2010) that carry out basic and applied research and generate resources for establishing high quality language technology solutions for all European languages. Analysing the deeper structural properties of languages is the only way forward if we want to build applications that perform well across the entire range of Europe’s languages.

European research in this area has already achieved a number of successes. For example, the translation services of the European Union now use MOSES open-source machine translation software that has been mainly developed through European research projects. Many of the research and development labs (e.g. IBM and Philips) have been closed down or moved elsewhere. Rather than building on the outcomes of its research projects, Europe has tended to pursue isolated research activities with a less pervasive impact on the market. The economic value of even the earliest efforts can be seen in the number of spin-offs. A company such as Trados, which was founded back in 1984, was sold to the UK-based SDL in 2005.

**Language Technology helps unify Europe**

Drawing on the insights gained so far, it appears that today’s 'hybrid' language technology mixing deep processing with statistical methods will be able to bridge the gap between all European languages and beyond. As this series of white papers shows, there is a dramatic difference in the state of readiness with respect to language solutions and the state of research between Europe’s member states. Yet even though Spanish is one of the ‘bigger’ languages, it too still needs further research before truly effective language technology solutions are ready for everyday use.

META-NET’s long-term goal is to introduce high-quality language technology for all languages in order to achieve political and economic unity through cultural diversity. The technology will help tear down existing barriers and build bridges between Europe’s languages. This requires all stakeholders - in politics, research, business, and society - to unite their efforts for the future.

This whitepaper series complements other strategic actions taken by META-NET (see the appendix for an overview). Up-to-date information such as the current version of the META-NET vision paper[[1]](#footnote-1) or the Strategic Research Agenda (SRA) can be found on the META-NET web site: http://www.meta-net.eu.

# Risk for Our Languages and a Challenge for Language Technology

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

We are witnesses to a digital revolution that is dramatically impacting communication and society. Recent developments in digital information and 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?

After Gutenberg’s invention, real breakthroughs in communication and knowledge exchange were accomplished by efforts such as Luther’s translation of the Bible into vernacular 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 exchanges across languages;
* the creation of editorial 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 has helped to automate and facilitate many of the processes:

* desktop publishing software has replaced typewriting and typesetting;
* Microsoft PowerPoint has replaced overhead projector transparencies;
* e-mail send and receive documents faster than a fax machine;
* Skype offers cheap 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, approximate translations;
* social media platforms such as Facebook, Twitter, and Google+ facilitate communication, collaboration, and information sharing.

Although such tools and applications are helpful, they are not yet capable of supporting a sustainable, multilingual European society for all where information and goods can flow freely.

## Language Borders Hinder the European Information Society

We cannot predict exactly what the future information society will look like. But there is a strong likelihood that the revolution in communication technology is bringing people speaking different languages together in new ways. This is putting pressure on individuals to learn new languages and especially on developers to create new technology applications to ensure mutual understanding and access to shareable knowledge. In a global economic and information space, more languages, speakers and content interact more quickly with new types of media. The current popularity of social media (Wikipedia, Facebook, Twitter, YouTube, and, recently, Google+) 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 from 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.[[2]](#endnote-1) 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 European (as well as Asian and Middle Eastern) languages has exploded.

Surprisingly, this ubiquitous digital divide due to language borders has not gained much public attention; yet, it raises a very pressing question: Which European languages will thrive in the networked information and knowledge society, and which are doomed to disappear?

## Our Languages at Risk

While the printing press helped step up the exchange of information in Europe, it also led to the extinction of many European languages. Regional and minority languages were rarely printed and languages such as Cornish and Dalmatian were limited to oral forms of transmission, which in turn restricted their scope of use. Will the Internet have the same impact on our languages?

Europe’s approximately 60 languages are one of its richest and most important cultural assets, and a vital part of its unique social model.[[3]](#endnote-2) While languages such as English and Spanish are likely to survive in the emerging digital marketplace, many European languages could become irrelevant in a networked society. This would weaken Europe’s global standing, and run counter to the strategic goal of ensuring 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.[[4]](#endnote-3)

The wide variety of languages in Europe is one of its richest and most important cultural assets.

## Language Technology is a Key Enabling Technology

In the past, investment efforts in language preservation focused on language education and translation. According to one estimate, the European market for translation, interpretation, software localisation and website globalisation was €8.4 billion in 2008 and is expected to grow by 10% per annum.[[5]](#endnote-4) Yet this figure covers just a small proportion of current and future needs in communicating between languages. The most compelling solution for ensuring the breadth and depth of language usage in Europe tomorrow is to use appropriate technology, just as we use technology to solve our transport, energy and disability needs among others.

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

Digital language technology (targeting all forms of written text and spoken discourse) helps people collaborate, conduct business, share knowledge and participate in social and political debate regardless of language barriers and computer skills. It often operates invisibly inside complex software systems to help us:

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

Language technology consists of a number of core applications that enable processes within a larger application framework. The purpose of the META-NET language white papers is to focus on how ready these core technologies are for each European language.

Europe needs robust and affordable language technology for all European languages.

To maintain our position in the frontline of global innovation, Europe will need language technology adapted to all European languages that is robust, affordable and tightly integrated within key software environments. Without language technology, we will not be able to achieve a really effective interactive, multimedia and multilingual user experience in the near future.

## Opportunities for Language Technology

In the world of print, the technology breakthrough was the rapid duplication of an image of a text (a page) using a suitably powered printing press. Human beings had to do the hard work of looking up, reading, translating, and summarizing knowledge. We had to wait until Edison to record spoken language – and again his technology simply made analogue copies.

Digital language technology can now automate the very processes of translation, content production, and knowledge management for all European languages. It can also empower intuitive language/speech-based interfaces for household electronics, machinery, vehicles, computers and robots. Real-world commercial and industrial applications are still in the early stages of development, yet R&D achievements are creating a genuine window of opportunity. For example, machine translation is already reasonably accurate in specific domains, and experimental applications provide multilingual information and knowledge management as well as content production in many European languages.

As with most technologies, the first language applications such as voice-based user interfaces and dialogue systems were developed for highly specialised domains, and often exhibit limited performance. But there are huge market opportunities in the education and entertainment industries for integrating language technologies into games, cultural heritage sites, edutainment packages, libraries, simulation environments and training programmes. Mobile information services, computer-assisted language learning software, eLearning environments, self-assessment tools and plagiarism detection software are just some of the application areas where language technology can play an important role. The popularity of social media applications like Twitter and Facebook suggest 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. It can help address the complex issue of multilingualism in Europe – the fact that different languages coexist naturally in European businesses, organisations and schools. But citizens need to communicate across these language borders criss-crossing the European Common Market, and language technology can help overcome this final barrier while supporting the free and open use of individual languages. Looking even further forward, innovative European multilingual language technology will provide a benchmark for our global partners when they begin to enable their own multilingual communities. Language technology can be seen as a form of ‘assistive’ technology that helps overcome the ‘disability’ of linguistic diversity and make language communities more accessible to each other.

Language technology helps overcome the “disability” of linguistic diversity.

Finally, one active field of research is the use of language technology for rescue operations in disaster areas, where performance can be a matter of life and death: Future intelligent robots with cross-lingual language capabilities have the potential to save lives.

## Challenges Facing Language Technology

The current pace of technological progress is too slow.

Although language technology has made considerable progress in the last few years, the current pace of technological progress and product innovation is too slow. Widely-used technologies such as the spelling and grammar correctors in word processors are typically monolingual, and are only available for a handful of languages. Online machine translation services, although useful for quickly generating a reasonable approximation of a document’s contents, are fraught with difficulties when highly accurate and complete translations are required. Due to the complexity of human language, modelling our tongues in software and testing them in the real world is a long, costly business that requires sustained funding commitments. Europe must therefore maintain its pioneering role in facing the technology challenges of a multiple-language community by inventing new methods to accelerate development right across the map. These could include both computational advances and techniques such as crowdsourcing.

## Language Acquisition in Humans and Machines

To illustrate how computers handle language and why it is difficult to program them to use it, let’s look briefly at the way humans acquire first and second languages, and then see how language technology systems work.

Humans acquire language skills in two different ways. Babies acquire a language by listening to the real interactions between its parents, siblings and other family members. From the age of about two, children produce their first words and short phrases. This is only possible because humans have a genetic disposition to imitate and then rationalize what they hear.

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

Learning a second language at an older age requires more effort, largely because the child is not immersed in a language community of native speakers. At school, foreign languages are usually acquired by learning grammatical structure, vocabulary and spelling using drills that describe linguistic knowledge in terms of abstract rules, tables and examples. Learning a foreign language gets harder with age.

The two main types of language technology systems ‘acquire’ language capabilities in a similar manner. Statistical (or ‘data-driven’) approaches obtain linguistic knowledge from vast collections of concrete example texts. While it is sufficient to use text in a single language for training, e.g., a spell checker, parallel texts in two (or more) languages have to be available for training a machine translation system. The machine learning algorithm then “learns” patterns of how words, short phrases and complete sentences are translated.

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

This statistical approach can require millions of sentences and performance quality increases with the amount of text analysed. This is one reason why search engine providers are eager to collect as much written material as possible. Spelling correction in word processors, and services such as Google Search and Google Translate all rely on statistical approaches. The great advantage of statistics is that the machine learns fast in continuous series of training cycles, even though quality can vary arbitrarily.

The second approach to language technology and machine translation in particular is to build rule-based systems. Experts in the fields of linguistics, computational linguistics and computer science first have to encode grammatical analyses (translation rules) and compile vocabulary lists (lexicons). This is very time consuming and labour intensive. Some of the leading rule-based machine translation systems have been under constant development for more than twenty years. The great advantage of rule-based systems is that the experts have 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. But due to the high cost of this work, rule-based language technology has so far only been developed for major languages.

As the strengths and weaknesses of statistical and rule-based systems tend to be complementary, current research focuses on hybrid approaches that combine the two methodologies. However, these approaches have so far been less successful in industrial applications than in the research lab.

As we have seen in this chapter, many applications widely used in today’s information society rely heavily on language technology. Due to its multilingual community, this is particularly true of Europe’s economic and information space. Although language technology has made considerable progress in the last few years, there is still huge potential in improving the quality of language technology systems. In the following, we will describe the role of Spanish in European information society and assess the current state of language technology for the Spanish language.

# Spanish in the European Information Society

## General Facts

Spanish is one of the languages with more demographic weight in the world today.

Spanish is one of the languages with more demographic weight in the world today. According to Ethnologue[[6]](#endnote-5) it is the second most widely spoken language in terms of native speakers after Chinese. Around 6% of the world population, or over 400 million speakers, have Spanish as first language.

Spanish is the official language of Spain, where it originated as an evolution of Vulgar Latin, but most Spanish speakers are in Latin America; of all countries with a majority of Spanish speakers, only Spain and Equatorial Guinea are outside the Americas. Mexico has the most native speakers of any country. Nationally, Spanish is the official language of Argentina, Bolivia, Chile, Colombia, Costa Rica, Cuba, Dominican Republic, Ecuador, El Salvador, Guatemala, Honduras, Mexico , Nicaragua, Panama, Paraguay, Peru, Puerto Rico, Uruguay, and Venezuela[[7]](#endnote-6).

In the Americas, Spanish coexists with a multitude of indigenous languages, some of them also co-official with Spanish in their respective countries, such as Quechuan and Aymara in Peru and Bolivia, and Guarani in Paraguay. In general, the situation can be characterised as being diglossic, with Spanish being the socially prestigious language, and the other language being used as communication vehicle within the family domain. The Spanish language functions in effect as a koiné and it is characterized by its linguistic homogeneity, especially among the educated, not forgetting that it comprises diverse sociolinguistic and geolinguistic varieties.

In Spain and in some parts of the Spanish speaking world, but not all, Spanish is called castellano (Castilian) as well as español (Spanish), that is, the language of the Castile region, contrasting it with other languages spoken in Spain, which are co-official in their territories, such as Catalan, Galician and Basque, spoken respectively by around six million, two million and half a million people.

In spite of its broad homogeneity, there exist variations between the language spoken in most of Spain, and the one spoken in much of southern Spain, the Canary Islands and Latin America. Apart from lexical local variations, it is well-known the different pronunciation of the letter c (as in the word Barcelona), being a voiceless dental fricative /θ/ (as in English thing) in the former and a sibilant /s/ in the latter. Other variations affect the use of second person pronouns: tú or vos, depending on the zones, for the singular; and ustedes or vosotros, for the plural.

## Particularities of the Spanish Language[[8]](#endnote-7)

Certain linguistic characteristics of Spanish are challenges for computational processing.

Spanish phonology exhibits a five-vowel system. Spanish is an inflected language, with a two-gender system and about fifty conjugated forms per verb, but limited inflection of nouns, adjectives, and determiners. It uses prepositions, and usually, though not always, places adjectives after nouns, as do most other Romance languages. Its syntax is generally subject-verb-object, though variations are common.

Spanish syntax provides overt marking for some direct objects: the so-called "personal a", that is, human direct objects are preceded by the preposition “a”:

*Vi a María (I saw Mary)* vs. *Vi la película (I saw the film).*

A "redundant" indirect object pronoun (*le, les*) may be used even in the presence of an explicit indirect object noun phrase:

*Le di las llaves a María (I gave the keys to Mary).*

With regard to subject pronouns, Spanish is a *pro-drop* language, meaning that the inflected verb phrase can often stand alone without the use of a subject pronoun (or a subject noun phrase). Compared to other Romance languages, Spanish has a somewhat freer syntax with relatively fewer restrictions on subject-verb-object word order.

Due to prolonged language contact with other languages, the Spanish lexicon contains loanwords from Old Basque/Iberian, Old Germanic, Arabic, and indigenous languages of the Americas, although around a 94% of its vocabulary comes from Latin.

Spanish orthography is far more transparent than, for example, English orthography. The pronunciation of a word can be easily figured out from its written form. Nonetheless a number of the writing system's rules lead to potential homophony. These include the silent ‹h› (*a/ha*), the lack of distinction between ‹b› and ‹v› (*baca/vaca*), or ‹g› and ‹j› before ‹e i› (*deje/general*), as well as some dialectal mergers such as that between ‹y› and ‹ll› (*pollo/poyo*), and between ‹c z› and ‹s› (*tensión / atención*). In this way, a number of spellings could represent the same pronunciation. This is a source of potential errors that spellers and grammar checkers need to be able to detect and correct.

Homophony is a source of potential errors that spellers and grammar checkers need to be able to detect and correct.

Written Spanish unequivocally marks stress through a series of orthographic rules. The default stress is on the next-to-last syllable on words that end in a vowel, ‹n› or ‹s› and on the final syllable when the word ends in any consonant other than ‹n› or ‹s›. Words that do not follow the default stress have an acute accent over the stressed vowel.

Capitalization in Spanish is sparse compared to other languages. In general, only personal and place names, some abbreviations (e.g. *Sr. López*, but *señor López*), the first word (only) in the title of a book, movie, song, etc., and the first word in a sentence are capitalized, as are names of companies, government bodies, etc. Names of nationalities or languages are not capitalized, nor (in standard style) are days of the week and months of the year.

## Recent Developments

Spanish is growing in economic importance in both Latin America and the USA.

Due to demographic reasons, particularly outside Spain (in the Americas), Spanish is escalating positions in the ranks of most spoken and used languages of the world.

The language is growing in economic importance in both Latin America and the USA, where it is challenging English in a number of towns that have predominantly Spanish speaking populations.

Spain is active in promoting itself as the global centre of authority for Spanish.

Spain is active in promoting itself as the global centre of authority for the language, with publicly supported institutions such as the Real Academia Española and the Instituto Cervantes and private fundations, such as Fundéu, that keeps watch over the new uses of the language and the influence of other languages (see next section).

Spain has also made exemplary progress in the development of Language Technologies competence to meet the needs of its linguistically complex population. With nearly 30% of Spain’s citizens being native speakers of a regional language other than Castilian, the research focus needs to take this diversity into account. Catalan, Basque, Galician and Gascon/Aranese are officially bilingual with Spanish in the regions where they are spoken, while both Aragonese and Asturian are protected within their Autonomous Communities (though they are not recognised as “official” languages).

Political devolution has enabled the Spanish regions to foster language technology development, with particular success in Catalan and Basque. The intrinsically broad linguistic focus of the Spanish research, far from diluting the focus on Spanish LT may provide a good model for Europe’s multilingual policies and LT community.

Spain’s commitment to multilanguage research is a model for the European Union.

On the one hand there is a depth of scientific experience with less dense languages, such as Catalan and Basque, that is virtually unique in Europe. Attention to developing cross-lingual products and services between regional languages is equally rare in other Member States. Overall, Spain is among the Member States with “strong potential” in LT. Its commitment to multilanguage research, especially through regional RTD programmes, is a model for the Union.

On the other hand, like Ireland (for English), the research community in Spain competes globally in the development of products for a world language. Spanish is well served by established first-generation tools (such as machine translation) that were developed in the US, or elsewhere in Europe, to take advantage of the market opportunity represented by a global Spanish-speaking population of four to five hundred million people.

## Language Cultivation in Spain

The Real Academia Española (RAE) exercises a standardizing influence through its publication of dictionaries and widely respected grammar and style guides.

The Real Academia Española (RAE or Royal Spanish Academy), founded in 1713, together with the 21 other national ones (one for each Spanish speaking country), exercises a standardizing influence through its publication of dictionaries and widely respected grammar and style guides. Because of influence and for other socio historical reasons, a standardized form of the language, mostly based on Castilian Spanish, is widely acknowledged for use in literature, academic contexts and the media.

The RAE, through its non-profit Foundation, has created two large Spanish textual corpora: the Corpus of Reference of Current Spanish (CREA) and the Diachronic Corpus of Spanish (CORDE). They have been used by the Academy as a source of information to write dictionaries and grammars, such as the Student's Dictionary, the Panhispanic Dictionary of Doubts, the Essential Dictionary and the New Grammar of the Spanish language, which incorporate many examples of the actual use of the language, taken mainly from CREA.

The Instituto Cervantes (Cervantes Institute) is a worldwide non-profit organization created by the Spanish government in 1991. This organization has branched out in over 20 different countries with 54 centres devoted to the Spanish and Hispanic American culture and Spanish language. The ultimate goals of the Institute are to promote the education, the study and the use of Spanish universally as a second language, to support the methods and activities that would help the process of Spanish language education, and to contribute to the advancement of the Spanish and Hispanic American cultures throughout non-Spanish-speaking countries.

From 1997 until today, the Cervantes Institute together with the Royal Spanish Academy and the Association of Spanish Language Academies have organised every three years an International Congress of the Spanish Language. This conference is an international forum for discussing the situation, problems and challenges of the Spanish language. It aims at raising awareness of individuals, governments and institutions in promoting the language, its unity and foster dialogue across the Hispanic cultural community.

The Fundéu BBVA (formerly The Foundation of Urgent Spanish)[[9]](#endnote-8) is a non-profit organization created in February 2005 in Madrid. The Foundation was created in collaboration with the Real Academia Española, and is founded under the Department of Urgent Spanish of the news agency Agencia EFE. It took its new name Fundéu BBVA in 2008. The main goal of the Foundation is to promote good usage of Spanish in all media, including, but not limited to, print, radio, television, and Internet.

The main goal of the Fundéu BBVA is to promote good usage of Spanish in all media, including print, radio, television, and Internet.

The Foundation has created and maintains a web page in which the materials elaborated by the foundation can be found. The Manual of Urgent Spanish is a reference guide for all those who want to write correct Spanish, free from the influence of other languages. An example from the Manual reads:

*Jugar un papel (play a role) – the phrase “jugar un papel” appears quite frequently in the news and is increasingly used in both written and spoken Spanish, ... “jugar un papel” is an Anglo-Gallicism which has its roots in the English "to play a role" and the French "jouer un rôle" ... in the sense of "fulfill a mission or assignment" , "agree with the position” ... it is an unacceptable expression, alien to our language as it would be “jugar una comedia” (to play a comedy) o “jugar al piano” (to play the piano).*

The manual also addresses other linguistic aspects, such as pronunciation, orthography, and punctuation. For example, with respect to “seseo” and “ceceo” (the varying local ways of pronouncing ‘s’, ‘c’ and ‘z’), the Manual of Urgent Spanish says:

*The seseo, that is, the pronunciation of the letters c and z as s, is characteristic of some southern areas of the Peninsula, the Canary Islands and America. This pronunciation is supported by the educated norm. Not so the ceceo (or lisp) (pronunciation of the s as z).*

The Foundation of Urgent Spanish in 2008 launched Wikilengua[[10]](#endnote-9), a wiki which serves as a resource tool of the accepted use of Spanish. With the contribution of the Spanish community, frequent doubts can be consulted and corrected. Being open and accessible to everyone, Wikilengua may also serve as a medium to reflect the diversity and richness of Spanish in its multiple varieties. It is constructed and grows around a community of authors, translators, editors, linguists, professors, students, journalists, and in general all people who are interested in Spanish, individually or as part of some entity, that would like to share their knowledge with hundreds of millions of other Spanish-speakers.

## Language in Education

The Article 3 of the Spanish Constitution states that “Spanish is the official language of the State. All Spaniards have a duty to understand and the right to use it. Other Spanish languages may also be official languages within autonomous communities depending on their individual statutes. The rich diversity of the languages of Spain constitutes a cultural heritage which deserves our particular respect and protection.” Each ‘historical’ autonomous community (Catalonia, the Basque Country and Galicia) possesses laws on the use of Spanish and their own language. Other communities (the Balearic Islands, the Valencian Community, Navarre, Aragon and Asturias) have progressively implemented various legal provisions to promote the use of their language within the education system and civil society.

Each bilingual community has a Linguistic Policy Department to promote the use of its own language.

Spanish is the official language of the State. All Spaniards have a duty to understand and the right to use it.

Each bilingual community has a Linguistic Policy Department to promote the use of its own language. This department proposes, implements and controls the application of diverse legal measures which have been taken to promote the languages and provide guidelines for their use. It works closely with the community’s Education Department. Legal provisions have been made in the context of the Education Law of 3 May 2006 (LOE) to guarantee the teaching of and equal status as an official language of the second language in bilingual communities.

As far as legal provisions about foreigners are concerned, it is stipulated that all foreigners under the age of 18 have the right (and duty) to be educated under the same conditions as Spaniards. This means access to an elementary education from 6 to 16 years of age, to include primary teaching (6-12 years) and compulsory secondary schooling (12-16 years). Different provisions are available to enable young people (in the school system by classes and induction groups) and adults to learn the official languages of Spain. Teaching is also available for vocational training and recap classes with the aim of obtaining specific qualifications.

Different provisions are available to enable young people and adults to learn the official languages of Spain.

The relevant induction programmes are the responsibility of the autonomous communities and vary from one community to the other. However, induction and linguistic integration are systematically offered.

In the context of decentralising the education system, the Law of Education (LOE) reaffirms the sharing of responsibility between the Spanish Ministry for Education and the Education Departments in the autonomous communities. For monolingual communities, the State is responsible for 65% of teaching and the community for 35%. For bilingual communities, the State’s responsibility is reduced to 55% and the community’s increased to 45%. The teaching of a second foreign language is also subject to provisions made by the autonomous communities, only the first modern language is compulsory from primary school and its study remains compulsory until the end of secondary schooling (16 years)[[11]](#endnote-10) .

## International Aspects

Spanish is the second most studied language in the world, after English, with nearly two million students in secondary education in the European Union.

Spanish is the second most studied language in the world, after English[[12]](#endnote-11), with nearly two million students in secondary education in the European Union. It has also become the most popular foreign language studied in the United States (chosen by 70% of the students). In Brazil its study in high school is compulsory. Spanish is also a language of culture, endowed with a rich literature.

However, despite the demographic weight of Spanish, its position as an international language of communication and the current demand for Spanish as second language, the competitiveness of Spanish as a scientific language is seriously challenged by English. In the technological domain, English is clearly the dominant language. The capacity of the English language to generate technical terms and consolidate their use is very difficult to compete with. In the field of natural and environmental sciences, Spanish is also virtually nonexistent. To address this situation, a number of possible solutions have been proposed. These include the strengthening of the teaching of Spanish in science and technology, training of foreign students at Spanish universities, the organization of regular workshops on technical writing, and the establishment of a science news agency in Spanish.

In the technological domain, English is clearly the dominant lan-guage.

On the other hand, Spanish as a business language is in a position to challenge English dominance, being the key to a huge Latin-American market. The economic value of the Spanish language has even been quantified[[13]](#endnote-12). It is considered to contribute over 15% to the GDP of Spain (therefore, in 2009, more than 37,000 million)[[14]](#endnote-13). According to a research by the Telefonica Foundation, using the same language multiplies by two or three the market share of exports from Spain to Spanish-speaking countries.

The economic value of the Spanish language is considered to contribute over 15% to the GDP of Spain.

Spanish is not only a booming language for business but it is a business itself. There is a new tourism for foreigners coming to Spain to study Spanish, which since 1995 has increased at a rate of about 10% annually. In 2006, about 150,000 tourists travelled to Spain with this goal, and left 255 million. The Cervantes Institute has specific Spanish courses for Business, which seem to be very successful in the countries of Eastern Europe, because of their relationship with Spain, especially in construction and public works, and because of emerging Latin-American markets.

The President of the Governing Council of INSEAD Business School (Universidad Autónoma de Madrid) foresees that in 2030 Spanish will become the second language for economic exchange in the world, only behind Chinese. Its widespread use demands that more business masters and training are offered in this language. Also the traditional teaching methods are starting to give way to new technologies and e-learning. So far the greatest number of interested students has come from Latin American countries. But the range has recently opened to students from other origins, especially Portuguese-speaking professionals, and German and French as well. Those seeking this type of courses are experienced professionals who wish to advance in their careers. Also companies that, in order to achieve greater competitiveness, promote their staff training.

Spanish is recognised as one of the official languages of the United Nations, the European Union, the Organisation of American States, the Organization of Ibero-American States, the African Union, the Union of South American Nations, the Latin Union, and the Caricom and has legal status in the North American Free Trade Agreement.

Spanish is an official language of the EU but not a working language. The Spanish government has also made efforts in favour of including Catalan, Galician and Basque among the official languages of the European institutions. They are currently considered semi-official, together with Scottish, Gaelic and Welsh. Not being a working language, there are a lot of documents that are not translated into Spanish and an even greater lot not translated into the rest of the languages of Spain. Most of these documents are available only in English, some of them are also in French or/and German. The same happens with some of the EU web sites.

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

Language technology can help diminish personal and economic disadvantages naturally faced by non-native speakers of English.

## Spanish on the Internet[[15]](#endnote-14)

It is often claimed that English dominates computers and the Internet, and that those wishing to use either must first learn English. That may have been true in the early days of the technology, but lack of English is no longer the barrier it once was. What began as an Anglophone phenomenon has rapidly become a multilingual affair. Software has been made capable of displaying many different kinds of script. Many corporate websites now employ multilingual strategies making choice of language a ‘user preference’. Machine translation of web content is only a mouse-click away. Furthermore, the Internet is proving to be a very useful resource to those interested in learning lesser-used languages. An analysis published in November 2005 by Byte Level Research con-cluded that “the next Internet revolution will not be in English“. While English isn’t becoming any less important on the Internet, other languages, such as Chinese, Russian, Spanish, and Portuguese, are becoming comparatively more important[[16]](#endnote-15).

The next Internet revolution will not be in English.

Global internet usage statistics for 2007 show Spanish as the third most commonly used language on the Internet, after English and Chinese[[17]](#endnote-16). The total number of Internet users from the different Spanish speaking countries was over 153 million in 2010, with a 20.8% from Spain. Penetration of Internet use in Spain reached 62.6% in 2010 according to Nielsen Online[[18]](#endnote-17).

The total number of Internet users from the different Spanish speaking countries was over 153 million in 2010, with a 20.8% from Spain.

The study entitled “Socio-Demographic Profile of Internet Users, Analysis of INE (Spanish National Statistics Institute) 2010 Data”, produced by the Spanish Observatory for Telecommunications and the Information Society (ONTSI)[[19]](#endnote-18), is based on data produced by the INE in its “Survey of Equipment and Use of Information and Communication Technologies in Households 2010” (ICT-H 2010) and highlights the increasingly intensive use of the Internet:

* 44.4% of the Spanish population aged between 16 and 74, that is to say, 15.4 million people, use the Internet on a daily basis.
* Almost 27 million individuals aged 10 or over have accessed the Internet at some time.
* The difference between men and women in Internet use has been reduced over the last years, mainly in the group of users who access the Internet every week.
* The age variable is a differentiating factor: the higher percentage of Internet users is found in the younger age groups. Almost all students have accessed the Internet at some time.
* The level of completed education is also a determining factor when it comes to Internet use: the higher the level of education, the higher the percentage of Internet users.

Another report prepared by the ONTSI is based on data extracted from the survey on ICT use in companies, also carried out by the INE, and asserts that 100% of large companies have computers, Internet access and e-mail. It also points out that mobile phones and computers were available in 2 out of 3 microenterprises in the first quarter of 2010. Over 90% of large and medium-sized companies and nearly 40% of microenterprises interact with the public administration via the Internet. The two main reasons for interacting are obtaining information and downloading forms. The most notable growth in SMEs and large companies has been observed in complete electronic management, which has risen by 5.2 points over the previous year.

All sectors have Internet penetration rates of over 90%, and in the financial sector it is of 100%. 97.2% of all SMEs and large enterprises have Internet access. Searching for information and using financial and banking services are the most common activities for companies with Internet, with percentages of 96.4% and 90.2%, respectively. The Web is used as a communication platform on 86.8% of occasions.

All sectors have Internet penetration rates of over 90%. Searching for information and using financial and banking services are the most common activities for companies with Internet

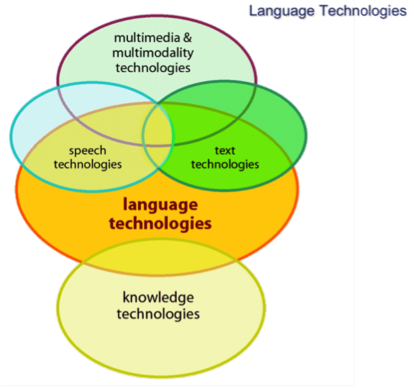
In 2010, 58.1% of microenterprises have Internet access, 2 points more than the previous year, accumulating a growth of 5 points in two years. 75.5% of microenterprises with Internet access use it as a communication platform, and 69.3% use it to access banking and financial services. Most notable are microenterprises in the financial sector, the IT, telecommunications and audiovisual sector, and the real estate and administrative activities sector, with over 80% of them using the Internet to access banking and financial services.

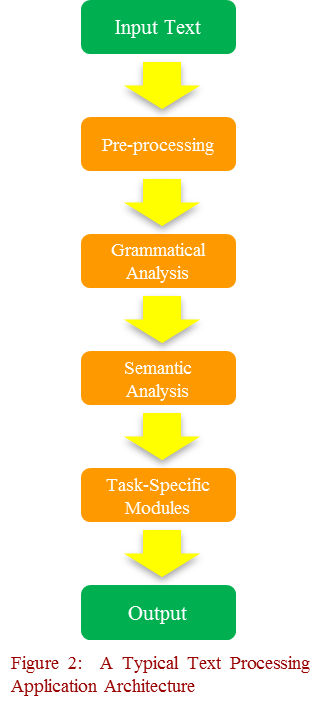
For language technology, the growing importance of the internet is important in two ways. On the one hand, the large amount of digitally available language data represents a rich source for analysing the usage of natural language, in particular by collecting statistical information. On the other hand, the internet offers a wide range of application areas involving language technology[[20]](#endnote-19).

The most commonly used web application is certainly web search, which involves the automatic processing of language on multiple levels, as we will see in more detail in the next section. It involves sophisticated language technology, differing for each language. For an inflected language such as Spanish, a lemmatiser, capable of generalising over surface inflected forms, is a very important tool. But Internet users and providers of web content can also profit from language technology in less obvious ways, for example if it is used to automatically translate web contents from one language into another. Considering the high costs associated with manually translating these contents, it may be surprising how little usable language technology is built compared to the anticipated need. However, it becomes less surprising if we consider the complexity of the Spanish language and the number of technologies involved in typical LT applications. In the next chapter, we will present an introduction to language technology and its core application areas as well as an evaluation of the current situation of LT support for Spanish.

# Language Technology Support for Spanish

Language technologies are software systems designed to handle human language and are therefore often called “human language technology”. Human language comes in spoken and written forms. While speech is the oldest and in terms of human evolution the most natural form of language communication, complex information and most human knowledge is stored and transmitted in written texts. Speech and text technologies process or produce these different forms of language, though they both use dictionaries and rules of grammar and semantics. This means that language technology (LT) links language to various forms of knowledge, independently of the media (speech or text) it is expressed in. The figure on the right illustrates the LT landscape. When we communicate, we combine language with other modes of communication and information media – for example speaking can involve gestures and facial expressions. Digital texts link to pictures and sounds. Movies may contain language in spoken and written form. In other words, speech and text technologies overlap and interact with other technologies that facilitate processing of multimodal communication and multimedia documents.

In the following, we will discuss the main application areas of language technology, i.e., language checking, web search, speech interaction, and machine translation. This includes applications and basic technologies such as

* spelling correction
* authoring support
* computer-assisted language learning
* information retrieval
* information extraction
* text summarization
* question answering
* speech recognition
* speech synthesis

Before discussing the above application areas, we will shortly describe the architecture of a typical LT system.

## Application Architectures

Software applications for language processing typically consist of several components that mirror different aspects of language. The figure on the right shows a highly simplified architecture that can be found in a typical text processing system. The first three modules handle the structure and meaning of the text input:

1. Pre-processing: cleans the data, analyses or removes formatting, detects the input language, normalizes text, and so on.
2. Grammatical analysis: finds the verb, its objects, modifiers and other parts of speech as well as detects the sentence structure.
3. Semantic analysis: performs disambiguation (i.e., computes the appropriate meaning of words in a given context); resolves anaphora (i.e., which pronouns refer to which nouns in the sentence) and substitute expressions; and represents the meaning of the sentence in a machine-readable way.

After analysing the text, task-specific modules can perform other operations, such as automatic summarization and database look-ups. This is a simplified and idealised description of the application architecture and illustrates the complexity of LT applications.

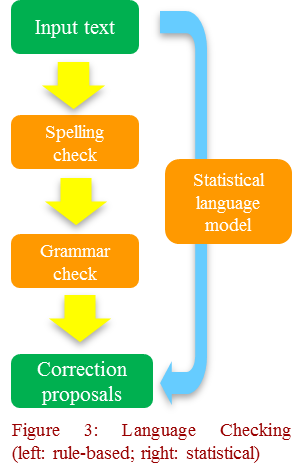
After introducing the core application areas for language technology, we shall provide a brief overview of the state of LT research and education today, and end with an overview of past and present research programmes. We shall then present an expert estimate of core LT tools and resources in terms of various dimensions such as availability, maturity and quality. The general situation of LT for the Spanish language is summarized in a table.

## Core Application Areas

Tools and resources that are underlined in the text can also be found in the table at the end of this chapter.

In this section, we focus on the most important LT tools and resources, and give an overview of LT activities in Spain. Tools and resources that are underlined in the text can also be found in the table at the end of this chapter.

### Language Checking

Anyone who has used a word processor such as Microsoft Word knows that it has a spelling checker that highlights spelling mistakes and proposes corrections. The first spelling correction programs compared a list of extracted words against a dictionary of correctly spelled words. Today these programs are far more sophisticated. Using language-dependent algorithms for text analysis, they detect errors related to morphology (e.g., plural formation) as well as syntax–related errors, such as a missing verb or a conflict of verb-subject agreement (e.g., *she \*write a letter*). But most spell checkers will not find any errors in the following text:

Eye have a spelling chequer,

It came with my Pea Sea.

It plane lee marks four my revue

Miss Steaks I can knot sea.[[21]](#endnote-20)

Handling these kinds of errors usually requires an analysis of the context. For Spanish, even spell checking requires analyzing the context in many cases. A typical case is when the orthographic error transforms one word into another, which also exits. In the following example, the first sentence contains a couple of frequent errors (problems with orthographic accents and omission of silent /h/). The second sentence is the corrected version of the first:

Mí calculo es que hoy a venido mas publico que ayer.

[Me I-estimate that today to come but I-publish than yesterday]

Mi cálculo es que hoy ha venido más público que ayer.

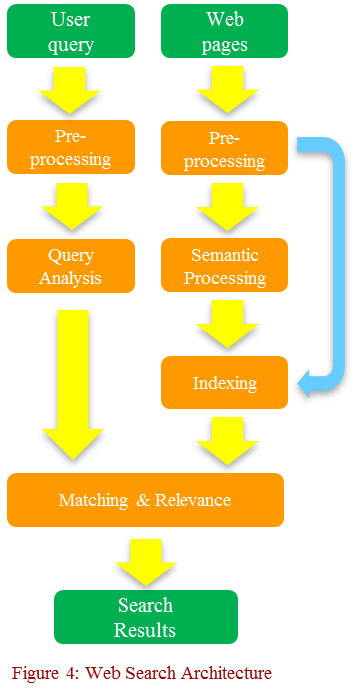
[My estimation is that today has come more people (audience) than yesterday]

This type of analysis either needs to draw on language-specific grammars laboriously coded into the software by experts, or on a statistical language model. In this case, a model calculates the probability of a particular word as it occurs in a specific position (e.g., between the words that precede and follow it). For example, *Mi cálculo* or *ha venido* are far much probable word sequences than *Mí calculo* or *a venido* respectively. A statistical language model can be automatically created by using a large amount of (correct) language data (called a text corpus). Most of these two approaches have been developed around data from English. However, they do not necessarily transfer well to other languages, e.g. highly inflectional ones or languages with a flexible word order like Spanish. For these more complex languages, an advanced high-precision language checker may require the development of more sophisticated methods, involving a deeper linguistic analysis.

Language checking is not limited to word processors; it is also used in “authoring support systems”, i.e., software environments in which manuals and other documentation are written to special standards for complex IT, healthcare, engineering and other products. Fearing customer complaints about incorrect use and damage claims resulting from poorly understood instructions, companies are increasingly focusing on the quality of technical documentation while targeting the international market (via translation or localization) at the same time. Advances in natural language processing have led to the development of authoring support software, which helps the writer of technical documentation use vocabulary and sentence structures that are consistent with industry rules and (corporate) terminology restrictions.

The use of language checking is not limited to word processors; it also applies to authoring support systems.

Only a few Spanish companies offer products in this area. For example Daedalus, born as a spin-off of two research groups from the Universidad Politécnica de Madrid (UPM) and the Universidad Autónoma de Madrid (UAM), has developed STILUS, a spelling, grammar and stylistic proof-reader for texts in Spanish. The technology is freely available through their web portal, but it can also be integrated in any content management tool. WinCorrect is a speller and grammar checker developed by Maxigramar (also with a free online version) and Signum is a Ecuador based SME that developed the spellchecker for Spanish licensed by Microsoft for their Office suite.

Besides spell checkers and authoring support, language checking is also important in the field of computer-assisted language learning. And language checking applications also automatically correct search engine queries, as found in Google's *Did you mean…* suggestions.

### Web Search

Searching the Web, intranets or digital libraries is probably the most widely used yet largely underdeveloped language technology application today. The Google search engine, which started in 1998, now handles about 80% of all search queries. The Google search interface and results page display has not significantly changed since the first version. Yet in the current version, Google offers spelling correction for misspelled words and has recently incorporated basic semantic search capabilities that can improve search accuracy by analysing the meaning of terms in a search query context.[[22]](#endnote-21) The Google success story shows that a large volume of available data and efficient indexing techniques can deliver satisfactory results for a statistically-based approach.

For more sophisticated information requests, it is essential to integrate deeper linguistic knowledge for text interpretation. Experiments using lexical resources such as machine-readable thesauri or ontological language resources (e.g. WordNet) have demonstrated improvements in finding pages using synonyms of the original search terms, such as e.g. *“Energía atómica” and “Energía nuclear”*, or even more loosely related terms.

The next generation of search engines will have to include much more sophisticated language technology, in particular in order to deal with search queries consisting of a question or other sentence type rather than a list of keywords. For the query, “Give me a list of all companies that were taken over by other companies in the last five years,” the LT system needs to analyse the sentence syntactically and semantically as well as provide an index to quickly retrieve relevant documents. A satisfactory answer will require syntactic parsing to analyse the grammatical structure of the sentence and determine that the user wants companies that have been acquired, not companies that acquired other companies. For the expression *last five years*, the system needs to determine the relevant years. And, the query needs to be matched against a huge amount of unstructured data to find the piece or pieces of relevant information the user wants. This is called “information retrieval”, and involves searching and ranking relevant documents. To generate a list of companies, the system also needs to recognize a particular string of words in a document as a company name, a process called “named entity recognition”.

The next generation of search engines will have to include much more sophisticated language technology.

A more demanding challenge is matching a query in one language with documents in another language. Cross-lingual information retrieval involves automatically translating the query into all possible target languages and then translating the results back into the source language.

Now that data is increasingly found in non-textual formats, there is a need for services that deliver multimedia information retrieval by searching images, audio files and video data. In the case of audio and video files, a speech recognition module must convert the speech content into text (or into a phonetic representation) that can then be matched against a user query.

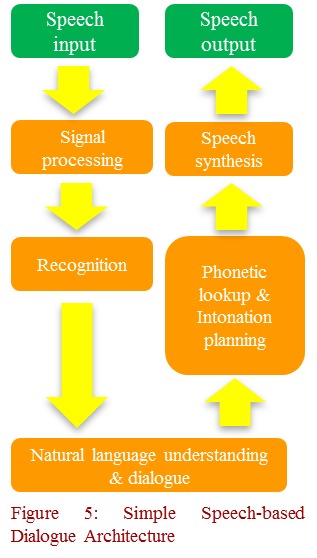
In Spain, a few SMEs develop linguistic technology aimed at multilingual search and information retrieval, both from the Internet and from internal information systems. Among the main ones, we find the following: iSOCO, Daedalus, Inbenta, Bitext and Thera, the latter also as spin-off, in this case from the Universitat de Barcelona. Their technology incorporates tools of automatic translation, as well as components of named entity recognition, fuzzy search and semantic tagging.

These companies focus their development on providing add-ons and advanced search engines for special interest portals by using topic-relevant semantics. Due to the constant high demand for processing power, such search engines are only cost-effective when handling relatively small text corpora. The processing time is several thousand times higher than that needed by a standard statistical search engine like Google. These search engines are in high demand for topic-specific domain modelling, but they cannot be used on the Web with its billions and billions of documents.

### Speech Interaction

Speech interaction technology is the basis for creating interfaces that allow a user to interact with spoken language instead of a graphical display, keyboard and mouse.

Speech interaction technology is used to create interfaces that enable users to interact in spoken language instead of a graphical display, keyboard and mouse. Today, voice user interfaces (VUI) are usually used for partially or fully automated telephone services provided by companies to customers, employees or partners. Business domains that rely heavily on VUIs include banking, supply chain, public transportation, and telecommunications. Other uses of speech technology include interfaces to car navigation systems and the use of spoken language as an alternative to the graphical or touch-screen interfaces in smartphones.

Speech interaction comprises four technologies:

1. Automatic speech recognition (ASR) determines which words are actually spoken in a given sequence of sounds uttered by a user.
2. Natural language understanding analyses the syntactic structure of a user’s utterance and interprets it according to the system in question.
3. Dialogue management determines which action to take given the user input and system functionality.
4. Speech synthesis (text-to-speech or TTS) transforms the system’s reply into sounds for the user.

One of the major challenges of ASR systems is to accurately recognise the words a user utters. This means restricting the range of possible user utterances to a limited set of keywords, or manually creating language models that cover a large range of natural language utterances. Using machine learning techniques, language models can also be generated automatically from speech corpora, i.e., large collections of speech audio files and text transcriptions. Restricting utterances usually forces people to use the voice user interface in a rigid way and can damage user acceptance; but the creation, tuning and maintenance of rich language models will significantly increase costs. VUIs that employ language models and initially allow a user to express their intent more flexibly — prompted by a *How may I help you?* greeting — tend to be automated and are better accepted by users.

Companies tend to use pre-recorded utterances by professional speakers for generating the output of the voice user interface. For static utterances where the wording does not depend on particular contexts of use or personal user data, this can deliver a rich user experience. But more dynamic content in an utterance may suffer from unnatural intonation because bits of audio files have simply been strung together. Today’s TTS systems are getting better (though they can still be optimized) at producing natural-sounding dynamic utterances.

Interfaces in the market for speech interaction technology have been considerably standardised during the last decade in terms of their various technology components. There has also been strong market consolidation in speech recognition and speech synthesis. The national markets in the G20 countries (economically resilient countries with high populations) have been dominated by just five global players, with Nuance (USA) and Loquendo (Italy) being the most prominent players in Europe[[23]](#endnote-22), also for Spanish, although some smaller local companies are starting to compete, such as Verbio , 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 Spanish TTS market are essentially application developers. Key players in the Spanish market are: Indsys (Intelligent Dialogue Systems), Fonetic , Ydilo and NaturalVoz . Rather than relying on a software license-driven product business, these companies are mainly positioned as full-service providers that create voice user interfaces as part of a system integration service. In the area of speech interaction, there is as yet no real market for syntactic and semantic analysis-based core technologies.

The demand for voice user interfaces in Spain has grown fast in the last five years, driven by increasing demand for customer self-service, cost optimisation for automated telephone services, and the increasing acceptance of spoken language as a media for human-machine interaction.

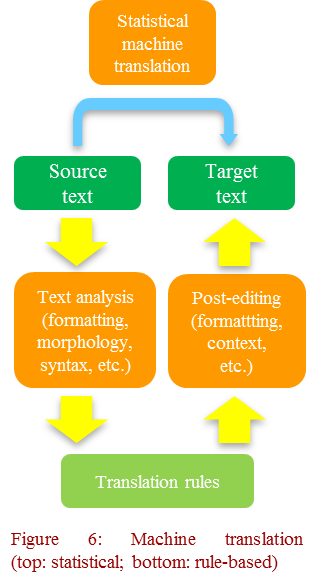
Looking forward, there will be significant changes due to the spread of smartphones as a new platform for managing customer relationships in addition to fixed telephones, the Internet and e-mail. This will also affect how speech interaction technology is used. In the long run, there will be fewer telephone-based VUIs and spoken language will play a far more central role as a user-friendly input for smartphones. This will be largely driven by stepped improvements in the accuracy of speaker-independent speech recognition via speech dictation services already offered as centralised services to smartphone users.

### Machine Translation

At its basic level, Machine Translation simply substitutes words in one natural language with words in another language.

The idea of using digital computers to translate natural languages goes back to 1946 and was followed by substantial funding for research during the 1950s and again in the 1980s. Yet machine translation (MT) still cannot meet its initial promise of across-the-board automated translation.

The most basic approach to machine translation is to automatically replace the words in a text in one natural language by words in another language. This can be useful in subject domains that have a very restricted, formulaic language such as weather reports. But to produce a good translation of less standardized 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 is that human language is ambiguous. Ambiguity creates challenges on multiple levels, such as word sense disambiguation on the lexical level (a *jaguar* is a brand of car or an animal) or the assignment of case on the syntactic level, for example:

The woman saw the car and her husband, too.

[La mujer vio el coche y su marido también.]

[La mujer vio el coche y **a** su marido también..]

One way to build an MT system is to use linguistic rules. For translations between closely related languages, a direct substitution translation may be feasible in cases like the above example. But, rule-based (or linguistic knowledge-driven) systems often analyse the input text and create an intermediary symbolic representation from which the text can be generated into the target language. 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 skilled linguists. This is a very long and therefore costly process.

In the late 1980s when computational power increased and became cheaper, there was more interest in statistical models for machine translation. Statistical models are derived from analysing bilingual text corpora, such as the Europarl parallel corpus, which contains the proceedings of the European Parliamentin 11 European languages. Given enough data, statistical MT works well enough to derive an approximate meaning of a foreign language text by processing parallel versions and finding plausible patterns of words. But unlike knowledge-driven systems, statistical (or data-driven) MT often generates ungrammatical output. Data-driven MT is advantageous because less human effort is required, and it can also cover special particularities of the language (e.g., idiomatic expressions) that can get ignored in knowledge-driven systems.

The strengths and weaknesses of knowledge-driven and data-driven machine translation tend to be complementary, so that nowadays researchers focus on hybrid approaches that combine both methodologies. One approach uses both knowledge-driven and data-driven systems together with a selection module that decides on the best output for each sentence. However, results for sentences longer than say 12 words will often be far from perfect. A better solution is to combine the best parts of each sentence from multiple outputs; this can be fairly complex, as corresponding parts of multiple alternatives are not always obvious and need to be aligned.

Leading international MT developer Lucy Software has an important subsidiary in Spain, Lucy Iberica , former Translendium. Lucy Iberica is responsible for the development of language pairs that include Spanish and all language pairs involving any other Iberian language (Catalan, Portuguese, Galician and Basque). Word Magic is a popular Spanish-English (and viceversa) system developed by a US-based company. Both Lucy and Word Magic systems are grammar rule-based. While there is significant research in data-driven and hybrid systems in national and international contexts, this technology has been less successful in business than in research so far, with just a few companies offering statistical MT customization, such as Pangeanic.

Apertium is a free open-source machine translation platform that provides a language-independent machine translation engine initially designed by the Transducens group at the Universitat d'Alacant and subsequently developed in the framework of the nationally funded Opentrad project. Among current MT systems using Apertium technology, we find interNOSTRUM (Spanish-Catalan), Traductor Universia (Spanish-Portuguese) and Matxin (Basque-Spanish), the former developed by Transducens and the latter by the IXA group at Euskal Herriko Unibertsitatea. It is possible to use Apertium to build machine translation systems for a variety of language pairs (there are over 20 to date); to that end, Apertium uses simple XML-based standard formats to encode the linguistic data needed (either by hand or by converting existing data), which are compiled using the provided tools into the high-speed formats used by the engine.

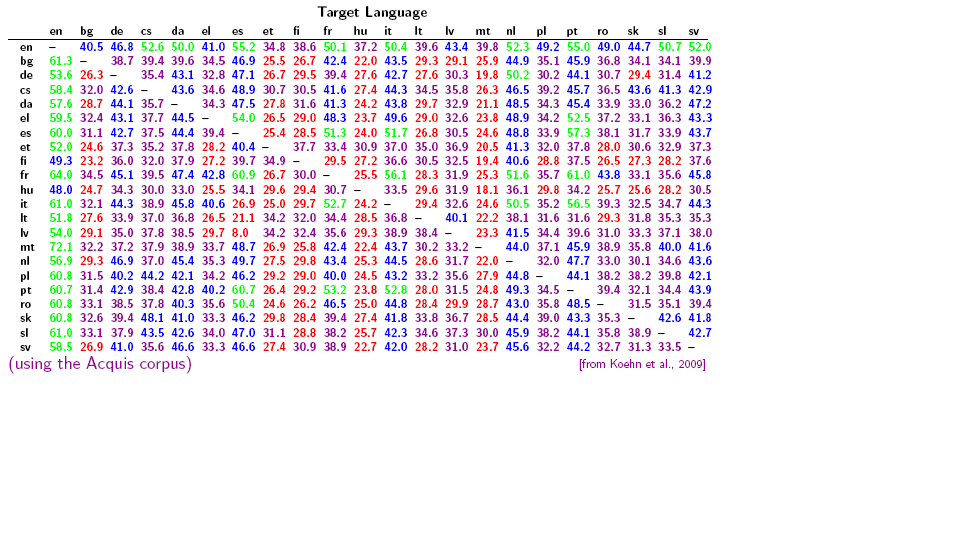
The use of machine translation can significantly increase productivity provided the system is adapted to user-specific terminology and integrated into a workflow.

The use of machine translation can significantly increase productivity provided the system is intelligently adapted to user-specific terminology and integrated into a workflow.

There is still a huge potential for improving the quality of MT systems. The challenges involve adapting language resources to a given subject domain or user area, and integrating the technology into workflows that already have term bases and translation memories. Another problem is that most of the current systems are English-centred and only support a few languages from and into Spanish. This leads to friction in the translation workflow and forces MT users to learn different lexicon coding tools for different systems.

Evaluation campaigns help compare the quality of MT systems, the different approaches and the status of the systems for different language pairs. The table below, which was prepared during the EC Euromatrix+ project, shows the pair-wise performances obtained for 22 of the 23 official EU languages. (Irish Gaelic was not compared.) The results are ranked according to a BLEU score, which indicates higher scores for better translations.[[24]](#endnote-23) (A human translator would achieve a score of around 80 points.)

The best results (in green and blue) were achieved by languages that benefit from a considerable research effort in coordinated programs and from the existence of many parallel corpora (e.g., English, French, Dutch, Spanish and German). The languages with poorer results are shown in red. These languages either lack such development efforts or are structurally very different from other languages (e.g., Hungarian, Maltese and Finnish).



Performance of Machine Translation for Language Pairs in the Euromatrix+ Project

## Other Application Areas

Language technology applications often provide significant service functionalities “under the hood” of larger software systems.

Building language technology applications involves a range of subtasks that do not always surface at the level of interaction with the user, but they provide significant service functionalities “under the hood” of the system in question. They all form important research issues that have now evolved into individual sub-disciplines of computational linguistics.

Question answering, for example, is an active area of research for which annotated corpora have been built and scientific competitions have been initiated. The concept of question answering goes beyond keyword-based searches (in which the search engine responds by delivering a collection of potentially relevant documents) and enables users to ask a concrete question to which the system provides a single answer. For example:

Question: How old was Neil Armstrong when he stepped on the moon?

Answer: 38.

While question answering is obviously related to the core area of web search, it is nowadays an umbrella term for such research issues as what different types of questions there are, and how they should be handled; how a set of documents that potentially contain the answer can be analysed and compared (do they provide conflicting answers?); and how specific information (the answer) can be reliably extracted from a document without ignoring the context.

This is in turn related to information extraction (IE), an area that was extremely popular and influential when computational linguistics took a statistical turn in the early 1990s. IE aims to identify specific pieces of information in specific classes of documents, such as detecting the key players in company takeovers as reported in newspaper stories. Another common scenario that has been studied is reports on terrorist incidents. The problem here is to map the text to a template that specifies the perpetrator, target, time, location and results of the incident. Domain-specific template-filling is the central characteristic of IE, which makes it another example of a “behind the scenes” technology that forms a well-demarcated research area that in practice needs to be embedded into a suitable application environment.

Text summarization and text generation are two borderline areas that can act either as standalone applications or play a supporting role “under the hood”. Summarization attempts to give the essentials of a long text in a short form, and is one of the features available in Microsoft Word. It mostly uses a statistical approach to identify the “important” words in a text (i.e., words that occur very frequently in the text in question but less frequently in general language use) and determine which sentences contain the most of these “important” words. These sentences are then extracted and put together to create the summary. In this very common commercial scenario, summarization is simply a form of sentence extraction, and the text is reduced to a subset of its sentences. An alternative approach, for which some research has been carried out, is to generate brand new sentences that do not exist in the source text. This requires a deeper understanding of the text, which means that so far this approach is far less robust. On the whole, a text generator is rarely used as a stand-alone application but is embedded into a larger software environment, such as a clinical information system that collects, stores and processes patient data. Creating reports is just one of many applications for text summarization.

For the Spanish language, research in most text technologies is much less developed than for the English language.

For the Spanish language, research in these text technologies is much less developed than for the English language. Question answering, information extraction, and summarization have been the focus of numerous open competitions in the USA since the 1990s, primarily organised by the government-sponsored organisations DARPA and NIST. These competitions have significantly improved the start-of-the-art, but their focus has mostly been on the English language. As a result, there are hardly any annotated corpora or other special resources needed to perform these tasks in Spanish. When summarization systems use purely statistical methods, they are largely language-independent and a number of research prototypes are available. For text generation, reusable components have traditionally been limited to surface realization modules (generation grammars) and most of the available software is for the English language.

Apart from the experimental systems being developed by the research groups, there are a few SMEs offering this kind of services. Among them Daedalus and Inbenta, and some international companies with a significant presence in the Spanish market, such as Q-go and Artificial Solutions .

## Educational Programmes

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

Graduate courses offer a more targeted professional training. There are several doctoral programs which offer masters or subjects related to language and speech processing. Certain universities such as the Universitat Politècnica de Catalunya also participate in the European Masters in Language and Speech sponsored by ELSNET (European Network of Excellence in Human Language Technologies). Masters are often offered by a group of universities, either at state or at European level. For example, the Universitat Autònoma de Barcelona offers the International Master in Natural Language Processing and Human Language Technology, in collaboration with foreign universities. Modules in Language Technology are also offered to students of other master or PhD courses, particularly in Translation (e.g. Autònoma de Barcelona, Alacant, Castelló, Politècnica de València, Granada).

There are over 30 research groups in Spain spread across the universities, working on speech recognition, natural language processing, text-to-text translation and speech synthesis. The Sociedad Española para el Procesamiento del Lenguaje Natural (SEPLN, Spanish Society for Natural Language Processing), is a non-profit organization with over 300 members, both from academia and industry, which was created in 1984 with the purpose to promote and spread activities related to teaching, research and development of NLP, on both national and international level. SEPLN organizes seminaries, symposiums and conferences and promotes collaboration with national and international institutions.

There are over 30 research groups in Spain spread across the universities, working on speech recognition, natural language processing, text-to-text translation and speech synthesis.

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

## National Projects and Efforts

The Ministry of Education, through the Interministerial Commission of Science and Technology (CICYT), which is the body that coordinates and monitors national strategy plans for Science & Technology in Spain, has supported research in the field of information technologies through national research programs. These programs have impelled numerous research projects and collaboration with international research centers and companies. The basis of technology development and commercial applications for automated processing of the Spanish language has been partly created as a result of these projects.

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

The CENIT (National Strategic Consortiums for Technological Re-search) programme seeks to stimulate cooperation in R&D between the private sector, universities, public research organisations and centres, science and technology parks and technological centres, boosting public and private-sector cooperation in R&D. CENIT projects last at least four years and have a minimum budget of €5 mill. a year during which they will receive minimum funding of 50% from the private sector. At least 50% of public funding will be allocated to public research centres or technological centres. Information Technology and Communication is one of the programme’s priority areas. Projects in this area sometimes include research in Language Technologies.

The aim of the AVANZ@ and INNPACTO Plans are to bring the Information Society to ordinary citizens, and to private and public sectors. Promoting the use of ICT technologies will have a knock-on effect on the whole sector in Spain, therefore on its innovation status. The Plan’s objectives include increasing the percentage of businesses using e-commerce; promoting the use of electronic billing; extending the electronic public sector by implementing an electronic identity card and electronic registration; attaining a rate of one Internet-connected computer for every two students in schools; and doubling the number of homes with Internet access. Among their priorities is to facilitate the use of new technologies to old people and people with disabilities, as an ideal means to achieve social integration, avoid exclusion and improve their quality of life. User-friendly language technology tools offer the principal solution to satisfy this goal, for example by offering speech synthesis for the blind.

Previous projects and programmes have led to the development of a number of LT tools and resources for the Spanish language. In the following section, the current state of LT support for Spanish is summarized.

## Availability of Tools and Resources

The following table summarizes the current state of language technology support for the Spanish language. The rating for existing tools and resources is based on educated estimates of leading experts.

|  | **Quantity** | **Availability** | **Quality** | **Coverage** | **Maturity** | **Sustainability** | **Adaptability** |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Language Technology (Tools, Technologies and Applications)** | | | | | | | |
| Speech Recognition | 4 | 2 | 3’6 | 4’8 | 4 | 4 | 3 |
| Speech Synthesis | 4 | 2 | 4’8 | 4’8 | 4 | 4 | 3 |
| Text analysis | 3’5 | 3 | 5’4 | 5’4 | 4’5 | 3’5 | 3’5 |
| Text interpretation | 1’5 | 2 | 2’4 | 2’4 | 2 | 2 | 1 |
| Language generation | 1 | 2 | 2’4 | 2’4 | 2 | 1 | 2 |
| Machine translation | 5 | 4 | 6 | 4’8 | 5 | 3 | 3 |
| **Language Resources (Resources, Data and Knowledge Bases)** | | | | | | | |
| Text corpora | 3 | 2’5 | 3’6 | 3’6 | 3 | 3’5 | 2,5 |
| Speech corpora | 3 | 1 | 3’6 | 2’4 | 3 | 4 | 3 |
| Parallel corpora | 4 | 3 | 4’8 | 3’6 | 3 | 2 | 2 |
| Lexical resources | 3 | 3 | 3’6 | 3’6 | 3,5 | 3,5 | 2,5 |
| Grammars | 2 | 3 | 3’6 | 3’6 | 4 | 2 | 3 |

The key results for the Spanish language can be summed up as follows:

* Speech processing currently seems to be slightly more mature than the processing of written text. In fact, speech technology has already been successfully integrated into many everyday applications, from spoken dialogue systems and voice-based interfaces to mobile phones and car navigation systems.
* Research has successfully led to the design of medium to high quality software for basic text analysis, such as tools for morphological analysis and syntactic parsing. But advanced technologies that require deep linguistic processing and semantic knowledge are still in their infancy.
* As to resources, there is a large reference text corpus with a balanced mix of genres for the Spanish language, but it is not easily available for research. There are a number of corpora annotated with syntactic, semantic and discourse structure mark-up, but again, there are not nearly enough language corpora containing the right sort of content to meet the growing need for more deep linguistic and semantic information.
* In particular, there is a lack of the sort of parallel corpora that form the basis for statistical and hybrid approaches to machine translation. Parallel corpora exist, between Spanish and English, and between Spanish and other languages from Spain. However, parallel corpora between Spanish and other languages are mostly missing.
* Many of these tools, resources and data formats do not meet industry standards and cannot be sustained effectively. A concerted programme is required to standardise data formats and APIs.
* An unclear legal situation restricts making use of digital texts, such as those published online by newspapers, for empirical linguistic and language technology research, for example, to train statistical language models. Together with politicians and policy makers, researchers should try to establish laws or regulations that enable them to use publicly available texts for language-related R&D activities.
* The cooperation between the Language Technology community and those involved with the Semantic Web and the closely related Linked Open Data movement should be intensified with the goal of establishing a collaboratively maintained, machine-readable knowledge base that can be used both in web-based information systems and as semantic knowledge bases in LT applications – ideally, this endeavour should be addressed in a multilingual way on the European scale.

To conclude, in a number of specific areas of Spanish language research, we have software with limited functionality available today. Obviously, further research efforts are required to meet the current deficit in processing texts on a deeper semantic level and to address the lack of resources such as parallel corpora for machine translation.

## Cross-language comparison

The current state of LT support varies considerably from one language community to another. In order to compare the situation between languages, this section will present an evaluation based on two sample application areas (machine translation and speech processing) and one underlying technology (text analysis), as well as basic resources needed for building LT applications.

Cluster 4

Cluster 3

Cluster 1

Cluster 2

English, French, German, Spanish, Italian, Dutch, Czech,   
Danish, Portuguese, Finnish

Icelandic, Irish

Latvian, Lithuanian, Maltese, Norwegian

Romanian, Slovak

Basque, Bulgarian, Catalan, Croatian, Estonian, Galician

Greek, Hungarian

Polish, Serbian

Slovene, Swedish

Figure 1: Language clusters for Speech Processing

Cluster 4

Cluster 3

Cluster 2

Cluster 1

Basque, Bulgarian, Croatian, Danish, Dutch, Estonian, Finnish, Galician, Greek, Hungarian, Icelandic, Irish, Latvian, Lithuanian, Maltese, Norwegian, Polish, Portuguese, Romanian, Serbian, Slovak, Slovene, Swedish

Spanish, Catalan, German, Italian, French, Czech

English

Figure 2: Language clusters for Machine Translation

*Cluster description (for Speech Processing and Machine Translation)*

* Cluster 1 (excellent LT support): Technologies with excellent quality and performance exist that can be used in practically all relevant applications.
* Cluster 2 (good support): Technologies with reasonable quality and performance exist whose usability is limited to some applications or domains.
* Cluster 3 (medium support): Research prototypes, first commercial applications or free services exist with varying quality and performance.
* Cluster 4 (low to almost no support): From the drawing board to rudimentary prototypes with very limited quality and performance.

Cluster 4

Cluster 5

Cluster 3

Cluster 2

Cluster 1

English

Czech, Dutch, German

Croatian, Estonian, Greek, Icelandic, Latvian, Lithuanian, Maltese, Serbian, Slovak, Slovene

Bulgarian, French, Norwegian, Polish, Portuguese, Spanish, Swedish, Basque, Catalan, Danish, Finnish, Galician, Hungarian, Irish, Italian, Romanian

Figure 3: Language clusters for Text Analysis

Cluster 3

Cluster 2

Cluster 1

Cluster 5

Cluster 4

Italian, Finnish, Croatian, Slovene, Icelandic, Irish, Greek, Latvian, Slovak, Serbian, Lithuanian, Maltese

Spanish, Norwegian, Portuguese, Danish, Polish, Romanian, Bulgarian,   
Catalan, Basque, Galician, Estonian

German, Hungarian, Swedish, French, Dutch, Czech

English

Figure 4: Language clusters for Resources

*Cluster description (for Text Analysis and Resources)*

* Cluster 1 (excellent LT support): Technologies/resources exist that are in widespread use and cover practically all linguistic phenomena (vocabulary, compounds, grammar, metaphors etc.) of a language.
* Cluster 2 (very good support): Technologies/resources exist that are used in a variety of applications and cover the most important linguistic phenomena.
* Cluster 3 (good support): Technologies/resources exist that cover a reasonable amount of linguistic phenomena and are used in applications that are usually domain-specific.
* Cluster 4 (medium support): Research prototypes/resources exist, but quality and coverage varies.
* Cluster 5 (low to almost no support): From the drawing board to rudimentary prototypes (very limited quality and coverage, toy systems)

The above tables show that, thanks to large-scale LT funding in recent decades, the Spanish language is better equipped than most other languages. It compares well with most large languages, such as French and German. But LT resources and tools for Spanish clearly do not yet reach the quality and coverage of comparable resources and tools for the English language, which is in the lead in almost all LT areas. And there are still plenty of gaps in English language resources with regard to high quality applications.

For speech processing, current technologies perform well enough to be successfully integrated into a number of industrial applications such as spoken dialogue and dictation systems. Today’s text analysis components and language resources already cover most surface linguistic phenomena of Spanish and form part of many applications involving mostly shallow natural language processing, e.g. spelling correction and authoring support.

However, for building more sophisticated applications, such as machine translation, there is a clear need for resources and technologies that cover a wider range of linguistic aspects and allow a deep semantic analysis of the input text. By improving the quality and coverage of these basic resources and technologies, we shall be able to open up new opportunities for tackling a vast range of advanced application areas, including high-quality machine translation.

## Conclusions

**In this series of white papers, we have made an important initial effort to assess language technology support for 30 European languages, and provide a high-level comparison across these languages. By identifying the gaps, needs and deficits, the European language technology community and related stakeholders are now in a position to design a large scale research and development programme aimed at building a truly multilingual, technology-enabled Europe.**

We have seen that there are huge differences between Europe’s languages. While there are good quality software and resources available for some languages and application areas, others (usually ‘smaller’ languages) have substantial gaps. Many languages lack basic technologies for text analysis and the essential resources for developing these technologies. Others have basic tools and resources but are as yet unable to invest in semantic processing. We therefore still need to make a large-scale effort to attain the ambitious goal of providing high-quality machine translation between all European languages.

We can be cautiously optimistic about technology support for the Spanish language. There is an LT industry and research scene in Spain, which was previously supported by large research programs, many of them in large industrial corporations. A number of large-scale resources and state-of-the-art technologies have been produced and distributed for Spanish. However, the size of the resources and the number of tools is still very limited when compared to the resources and tools for the English language, and they are simply not extensive enough to develop the technologies that are required to support a truly multilingual knowledge society.

Unfortunately, there is only a relatively small language technology industry at work on the Spanish language. Most large companies have stopped or severely decreased their LT work, leaving the field to a small population of specialized SMEs that are unable to address an international market in which the language barrier is a key factor holding back cross-border e-commerce in the EU.[[25]](#endnote-24)

It is clear that there must be a greater effort to create LT resources for Spanish, and drive research, innovation and development in general. The need for large amounts data and the extreme complexity of language technology systems makes it vital to develop a new infrastructure to spur greater sharing and cooperation.

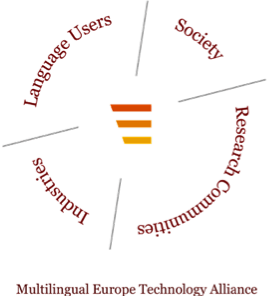
There is also a lack of continuity in research and development funding. Short-term coordinated programmes tend to alternate with periods of low or sparse funding. In addition, there is an overall lack of coordination with programmes in other EU countries and at the European Commission level.

A large coordinated effort focused on language technologies would help the Spanish language, together with other languages, and establish a genuine multilingual agenda for Europe and the world as a whole.[[26]](#endnote-25)

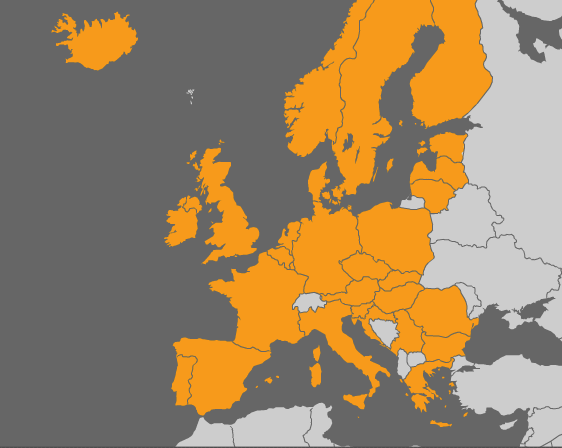
META-NET’s long-term goal is to introduce high-quality language technology for all languages in order to achieve political and economic unity through cultural diversity. The technology will help tear down existing barriers and build bridges between Europe’s languages. This requires all stakeholders - in politics, research, business, and society - to unite their efforts for the future.

# 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)



Countries Represented in META-NET

META-NET fosters the technological foundations for a truly multilingual European information society that:

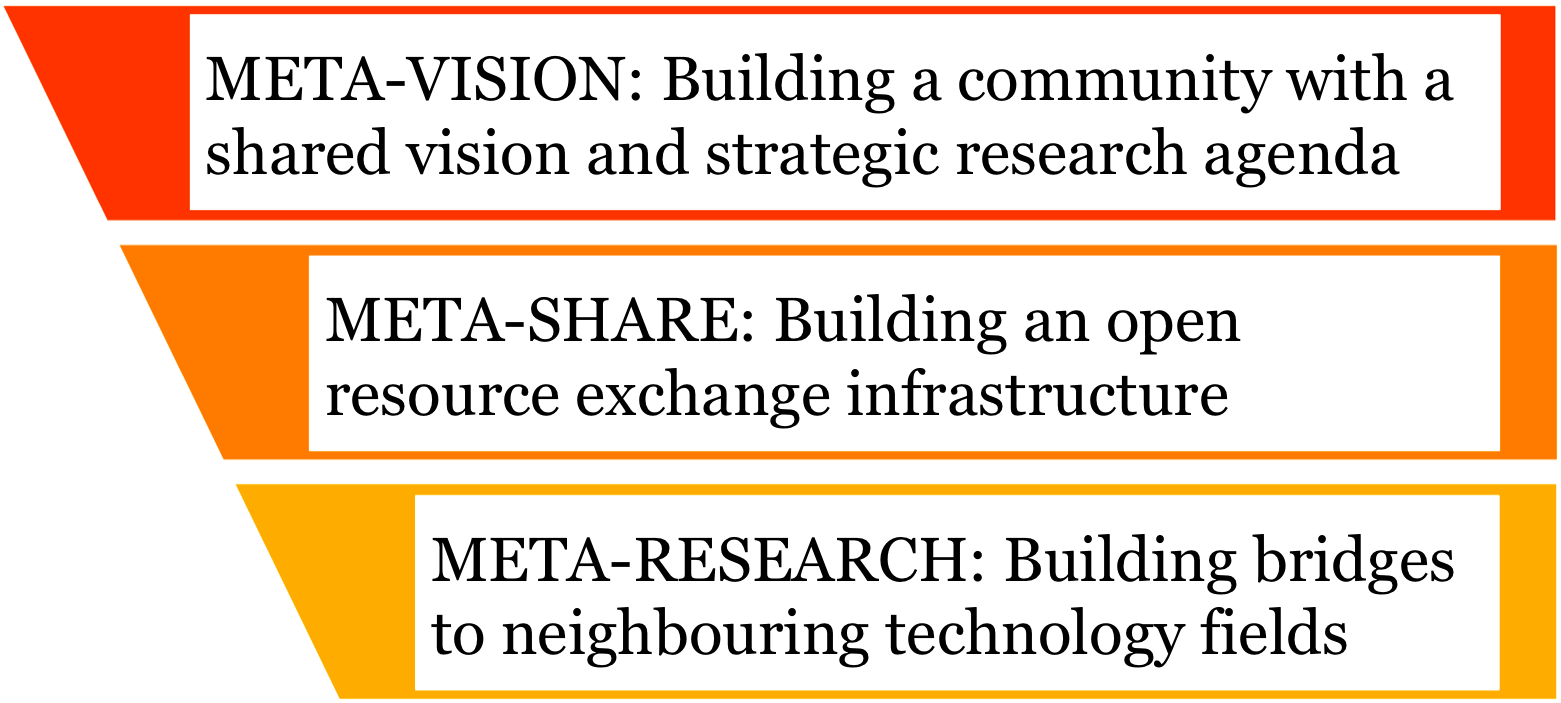
* makes communication and cooperation possible across languages;
* provides equal access to information and knowledge in any language;
* offers advanced and affordable networked information technology to European citizens.

The network supports a Europe that unites as a single digital market and information space. It stimulates and promotes multilingual technologies for all European languages. These technologies support automatic translation, content production, information processing and knowledge management for a wide variety of subject domains and applications. They also enable intuitive language-based interfaces to technology ranging from household electronics, machinery and vehicles to computers and robots.

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 launched on 1 February 2010 with the goal of advancing research in language technology (LT). 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.



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. The present White Paper was prepared together with volumes for 29 other languages. The shared technology vision was developed in three sectorial Vision Groups. The META Technology Council was established in order to discuss and to prepare the SRA based on the vision in close interaction with the entire LT community.

**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-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, the action line focuses on conducting leading-edge research in machine translation, collecting 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.

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

## Acknowledgments

Dr. Toni Badia, Universitat Pompeu Fabra and Barcelona Media Centre d‘Innovació

Dra. Núria Bel, Universitat Pompeu Fabra

Dr. Aljoscha Burchardt, DFKI

Kathrin Eichler, DFKI

Dra. Mª Antònia Martí, Universitat de Barcelona

Dra. Asunción Moreno, Universitat Politècnica de Catalunya.

Dr. Lluís Padró, Universitat Politècnica de Catalunya

Dr. GeorgRehm, DFKI

# References

1. http://www.meta-net.eu/vision/reports/meta-net-vision-paper.pdf [↑](#footnote-ref-1)
2. European Commission Directorate-General Information Society and Media, *User language preferences online*, Flash Eurobarometer #313, 2011 (http://ec.europa.eu/public\_opinion/flash/fl\_313\_en.pdf). [↑](#endnote-ref-1)
3. 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). [↑](#endnote-ref-2)
4. UNESCO Director-General, *Intersectoral mid-term strategy on languages and multilingualism*, Paris, 2007 (http://unesdoc.unesco.org/images/0015/001503/150335e.pdf). [↑](#endnote-ref-3)
5. European Commission Directorate-General for Translation, *Size of the language industry in the EU*, Kingston Upon Thames, 2009 (http://ec.europa.eu/dgs/translation/publications/studies). [↑](#endnote-ref-4)
6. <http://www.ethnologue.org> [↑](#endnote-ref-5)
7. Moreno Fernández, Francisco y Otero Roth, Jaime (2007) Atlas de la lengua española en el mundo. Editorial Planeta. [↑](#endnote-ref-6)
8. -Nueva gramática de la lengua española (2009), Real Academia Española, Colección Nebrija y Bello, Espasa. [↑](#endnote-ref-7)
9. <http://www.fundeu.es/> [↑](#endnote-ref-8)
10. <http://www.wikilengua.org/index.php/Portada> [↑](#endnote-ref-9)
11. <http://www.efnil.org/documents/language-legislation-version-2007/spain/spain> [↑](#endnote-ref-10)
12. -Vivanco Torres, Hiram. Enseñanza del español como lengua extranjera. Ponencia Congreso Internacional de la Lengua Española. Cartagena. 2007. (<http://congresosdelalengua.es/cartagena/ponencias/seccion_3/36/vivanco_hiram.htm>) [↑](#endnote-ref-11)
13. Garcia Delgado, Jose Luis (2007) Economia del español, una introducción. Editorial Planeta. [↑](#endnote-ref-12)
14. <http://www.fundacionlengua.com/es/valor-economico-espanol-genera-35-millones-empleos/art/1599/> [↑](#endnote-ref-13)
15. Rojo, Guillermo y Sánchez, Mercedes (2010) El español en la red. Editorial Ariel. (<http://www.fundacion.telefonica.com/es/debateyconocimiento/media/publicaciones/espanol_red.pdf>) [↑](#endnote-ref-14)
16. [http://www.britishcouncil.org/learning-research-english-next.pdf](%20http:/www.britishcouncil.org/learning-research-english-next.pdf) [↑](#endnote-ref-15)
17. [http://www.internetworldstats.com/stats7.htm](%20http:/www.internetworldstats.com/stats7.htm) [↑](#endnote-ref-16)
18. [http://www.nielsen-online.com/intlpage.html](%20http:/www.nielsen-online.com/intlpage.html) [↑](#endnote-ref-17)
19. <http://www.ontsi.red.es/> [↑](#endnote-ref-18)
20. Quirós Romero, Cipriano (2010) Lengua y tecnologías de la información y las comunicaciones. Editorial Planeta. [↑](#endnote-ref-19)
21. ## Jerrold H. Zar: “Candidate for a Pullet Surprise”, [*Journal of Irreproducible Results*](http://www.jir.com/pullet.html), Jan/Feb. 1994, page 13 (first verse).

    [↑](#endnote-ref-20)
22. http://www.pcworld.com/businesscenter/article/161869/google\_rolls\_out\_semantic\_search\_capabilities.html [↑](#endnote-ref-21)
23. In 2011, Nuance announced the acquisition of Loquendo, which represents a further step in market consolidation. [↑](#endnote-ref-22)
24. K. Papineni, S. Roukos, T. Ward, W.-J. Zhu. BLEU: A Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting of ACL*, Philadelphia, PA. [↑](#endnote-ref-23)
25. <http://ec.europa.eu/consumers/strategy/docs/com_staff_wp2009_en.pdf> [↑](#endnote-ref-24)
26. V. Reding, J.Figel’, Preface, in *Human Language Technologies for Europe*, TC-Star project, <http://www.tcstar.org/pubblicazioni/D17_HLT_ENG.pdf> [↑](#endnote-ref-25)