

**D2.3.pt.en**  
**Language Report for**  
**Portuguese**  
**(English version)**

Version 1.0

2011-07-29



# METANET4U

[www.metanet4u.eu](http://www.metanet4u.eu)

The central objective of the Metanet4u project is to contribute to the establishment of a pan-European digital platform that makes available language resources and services, encompassing both datasets and software tools, for speech and language processing, and supports a new generation of exchange facilities for them.

This central objective is articulated in terms of the following main goals:

**Assessment:** to collect, organize and disseminate information that permits an updated insight into the current status and the potential of language related activities, for each of the national and/or language communities represented in the project. This includes organizing and providing a description of: language usage and its economic dimensions; language technologies and resources, products and services; main actors in different areas, including research, industry, government and society in general; public policies and programs; prevailing standards and practices; current level of development, main drivers and roadblocks; etc.

**Collection:** to assemble and prepare language resources for distribution. This includes collecting languages resources; documenting these language resources; upgrading them to agreed standards and guidelines; linking and cross-lingual aligning them where appropriate.

**Distribution:** to distribute the assembled language resources through exchange facilities that can be used by language researchers, developers and professionals. This includes collaboration with other projects and, where useful, with other relevant multi-national forums or activities. It also includes helping to build and operate broad inter-connected repositories and exchange facilities.

**Dissemination:** to mobilize national and regional actors, public bodies and funding agencies by raising awareness with respect to the activities and results of the project, in particular, and of the whole area of language resources and technology, in general.

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



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



and by the participating institutions:



Faculty of Sciences, University of Lisbon



Instituto Superior Técnico



University of Manchester



University *Alexandru Ioan Cuza*



Research Institute for Artificial Intelligence,  
Romanian Academy



University of Malta



Technical University of Catalonia



Universitat Pompeu Fabra

## Revision History

Version	Date	Author	Organisation	Description
0.1	18-07-2011	Amália Mendes, António Branco, Paulo Henriques, Sílvia Pereira, Isabel Trancoso, Thomas Pellegrini, Hugo Meinedo, Paulo Quaresma	ULX, IST	Draft version
0.2	28-07-2011	Núria Bel	UPF	Review notes
1.0	29-07-2011	Amália Mendes, António Branco, Paulo Henriques, Sílvia Pereira, Isabel Trancoso, Thomas Pellegrini, Hugo Meinedo, Paulo Quaresma	ULX, IST	Final version

### Statement of originality:

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



# METANET4U

## **D2.3.pt.en Language Report for Portuguese (English version)**

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

**Deliverable D2.3.pt.en**

Completion: Final

Status: Submitted

Dissemination level: Public

Responsible: Asunción Moreno (WP2 coordinator)

Contributing Partners: ULX, IST

Authors: Amália Mendes, António Branco, Paulo Henriques, Sílvia Pereira,  
Isabel Trancoso, Thomas Pellegrini, Hugo Meinedo, Paulo Quaresma

Reviewer: Núria Bel

© all rights reserved by FCUL on behalf of METANET4U

**META-NET White Paper Series**

# **Languages in the European Information Society**

**– Portuguese –**



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).

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 European Commission Network of Excellence, 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.

META-NET  
DFKI Projektbüro Berlin  
Alt-Moabit 91c  
10559 Berlin  
Germany

office@meta-net.eu  
<http://www.meta-net.eu>

## Authors

Dr. Amália Mendes, CLUL-University of Lisbon  
Prof. António Branco, University of Lisbon  
Paulo Henriques, CLUL-University of Lisbon  
Sílvia Pereira, University of Lisbon  
Prof. Isabel Trancoso, INESC-ID / IST  
Dr. Thomas Pellegrini, INESC-ID  
Dr. Hugo Meinedo, INESC-ID  
Prof. Paulo Quaresma, University of Évora

## Acknowledgements

The publisher is grateful to the authors of the German white paper for permission to reproduce materials from their paper.

# Table of Contents

<b>Executive Summary .....</b>	<b>7</b>
<b>A Risk for Our Languages and a Challenge for Language Technology .....</b>	<b>8</b>
Language Borders Hinder the European Information Society.....	8
Our Languages at Risk.....	9
Language Technology is a Key Enabling Technology .....	10
Opportunities for Language Technology .....	10
Challenges Facing Language Technology .....	11
Language Acquisition .....	11
<b>Portuguese in the European Information Society .....</b>	<b>13</b>
General Facts .....	13
Particularities of the Portuguese Language .....	14
Recent developments .....	15
Language cultivation in Portugal and abroad .....	15
Language in Education .....	16
International aspects .....	17
Portuguese on the Internet .....	17
Selected Further Reading.....	18
<b>Language Technology Support for Portuguese.....</b>	<b>19</b>
Language Technologies .....	19
Language Technology Application Architectures .....	19
Core application areas.....	20
<i>Language checking</i> .....	20
<i>Web search</i> .....	21
<i>Speech interaction</i> .....	22
<i>Machine translation</i> .....	24
Language Technology ‘behind the scenes’ .....	26
Language Technology in Education .....	28
Language Technology Programs .....	29
Availability of tools and resources for Portuguese .....	31
Table of Tools and Resources .....	32
Conclusions .....	33
<b>About META-NET .....</b>	<b>35</b>
Lines of Action .....	35
Member Organisations .....	37
<b>References .....</b>	<b>40</b>

## Executive Summary

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

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

Currently, language technology services are primarily offered by commercial providers from the United States. Google Translate, a free service, is just one example. Another example, illustrating the immense potential of language technology, is the recent success of Watson, from IBM. This is a computer system that recently won an episode of the Jeopardy game show against human candidates.

As Europeans, we have to ask ourselves urgent questions:

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

This whitepaper for the Portuguese language demonstrates that a dynamic research environment exists in Portugal, which needs to be further enhanced to offer support for an emerging language technology industry. Although a number of linguistic resources and processing tools have been developed for Portuguese, there are fewer solutions for this language than for several other, better resourced languages in the European Union.

According to the assessment detailed in this report, immediate action must occur so that relevant progress for the Portuguese language can be achieved.

## A Risk for Our Languages and a Challenge for Language Technology

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

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

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

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

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

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

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

## Language Borders Hinder the European Information Society

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

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

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

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

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

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

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

## Our Languages at Risk

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

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

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

## Language Technology is a Key Enabling Technology

In the past, investment efforts have focused on language education and translation. For example, according to some estimates, the European market for translation, interpretation, software localisation and website globalisation was € 8.4 billion in 2008 and was expected to grow by 10% per annum.<sup>iv</sup> Yet, this existing capacity is not enough to satisfy current and future needs.

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

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

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

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

## Opportunities for Language Technology

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

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

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

There are huge market opportunities in the education and entertainment industries for the integration of language technologies in games, edutainment offerings, simulation environments or training programmes. Mobile information services, computer-assisted language learning software, eLearning environments, self-assessment tools and plagiarism detection software are just a few more examples where language technology can play an important role. The popularity of social media applications like Twitter and Facebook 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 that makes both economic and cultural sense. Multilingualism in Europe has become the rule. European businesses, organisations and schools are also multinational and diverse. Citizens want to communicate across the language borders that still exist in the European Common Market. Language technology can help overcome such remaining barriers while supporting the free and open use of language. Furthermore, innovative, multilingual language technology for European can also help us communicate with our global partners and their multilingual communities. Language technologies support a wealth of international economic opportunities.

*Multilingualism is the rule, not an exception.*

## Challenges Facing Language Technology

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

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

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

## Language Acquisition

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

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

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

Learning a second language usually requires much more effort when a child is not immersed in a language community of native speakers. At school age, foreign languages are usually acquired by learning their grammatical structure, vocabulary and orthography from books and educational materials that describe linguistic knowledge in terms of abstract rules, tables and example texts. Learning a foreign language takes a lot of time and effort, and it gets more difficult with age.

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

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

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

## Portuguese in the European Information Society

### General Facts

Portuguese is the third European language in the world, with around 200 million native speakers, and a total of 220 million speakers (native and second language) in 4 continents: Europe, America, Africa and Asia<sup>v</sup>.

It is the official language of Portugal, Brazil, Angola, Cape Verde, Guinea-Bissau, Mozambique, S. Tome and Principe, Timor-Lorosae, Macau, Goa, and, since 2010, of Equatorial Guinea.

Due to the emigration flow<sup>vi</sup>, Portuguese is also spoken by Portuguese communities in many countries, occupying in some of them an important ranking position among the foreign population. That is the case, in Europe, of Luxembourg (around 25% of the population), Andorra (around 11% of the population), France, Germany, United Kingdom, Switzerland, Spain and Belgium<sup>vii</sup>.

Portuguese is an official language of the European Union, the Mercosul (economic market including Brazil and other South American countries), and the African Union. With the development of the alphabetization in the former colonies in Africa and in East Timor, Portuguese has a high possibility of growing as a second language.

The expeditions and coastal trade that Portugal maintained during several centuries have linguistic counterparts: Portuguese incorporated words from African, Amerindian and Asian languages, but also gave its lexical contribution to many languages in the world, including the Lingua Franca of the Mediterranean Sea, and several Pidgins and Creoles of the Atlantic Ocean, the Pacific Ocean and the Indian Ocean<sup>viii</sup>.

The geographical division of dialects in Portugal<sup>ix</sup> distinguishes between Southern-Central dialects and Northern dialects. Differences are easily identifiable at phonetic/phonological and lexical level, although dialects are mutually understandable (possibly with the exception of some dialects of the Azores islands). The Northern dialects can be distinguished by the lack of the phonological distinction between /b/ and /v/, with prevalence of /b/; the preservation of ancient diphthongs; and the existence of apico-alveolar fricatives.

As Brazil is a very large country, it is not feasible to present here its linguistic varieties. For geographical, political and social reasons, neither is it possible to talk about the standard variety of Brazilian Portuguese. Instead, experts tend to talk about ‘cultivated urban varieties’, which present some differences with standard European Portuguese.

The African varieties of Portuguese also differ from the European, but to a lesser extent, and share some features with Brazilian Portuguese. The situation among the African varieties differ greatly: while in Angola and Mozambique the number of speakers of Portuguese has been increasing since the independence from Portugal, in other cases, like S. Tome and Principe and Cape Verde, Creole languages are widely used and Portuguese is a second language.

## Particularities of the Portuguese Language

Portuguese is a Romance language which implies that most of its lexicon is derived from Latin. It adopted also many words from a large variety of other languages, at different times, which, in many cases, remain among the most frequent words (e.g. pre-Latin: *bar-ranco* ‘ravine’, *seara* ‘corn-field’, *bruxa* ‘witch’; Germanic: *luvas* ‘gloves’, *bando* ‘band’, *guerra* ‘war’; and mostly Arabic: *aldeia* ‘village’, *açúcar* ‘sugar’, *laranja* ‘orange’).

The Portuguese language may often sound like a sequence of consonants to a foreign listener. This is due to the fact that, differently from the other Romance languages, the Portuguese unstressed vowels are often weakened or even not pronounced. This vowel weakening is a late change in European Portuguese and it did not affect the variety spoken in Brazil.

The basic word order in Portuguese is SVO - Subject Verb Object (*Ele leu o livro* ‘he read the book’). In certain pragmatic contexts (e.g. emphatic reading), the VSO order is also encountered (*lês tu o livro* ‘read you the book’) and the OSV or OVS order are possible in marked constructions termed as topicalized sentences (*O livro, ele não leu* ‘the book, he not read’).

Portuguese is a null subject language, that is the subject of the sentence may not be realized by any phonetically overt expression (*li o livro* ‘[I] read the book’). When the subject is a first person pronoun, its non-realization is in fact the default option and there is usually no expletive pronoun in impersonal constructions (*Ø há um livro sobre esse tema*, ‘is a book on that subject’). This characteristic of Portuguese represents a specific challenge for the automatic syntactic analysis of Portuguese texts and speech.

The inflection paradigm in Portuguese is much richer than the English one, especially in the case of verbs: for instance, a verb following a regular paradigm will have different markers for aspect/tense/mood, person and number, reaching more than 70 different inflected forms.

Also, there are two verb inflectional paradigms which do not exist in the other official Romance languages and are very frequent in Portuguese: the inflected infinitive and the future subjunctive. The former shares the theme with the non inflected infinitive (e.g. *cantar* ‘to sing’) to which the aspect/tense/mood constituent, and person/number markers are adjoined (*para eu cantar* ‘for I to sing’, *para tu cantares* ‘for you to sing’, *para eles cantarem* ‘for them to sing’). The inflected forms of the subjunctive future are homonyms to the ones of the non inflected infinitive, except with irregular verbs, and this increases the number of ambiguous forms in the verb paradigm.

The position of clitic pronouns in the sentence is another characteristic that raises specific challenges to the automatic processing of Portuguese language. Clitic pronouns can occur before and after the verb, but, in the future and conditional tenses, they can also be realized in the middle of the verb form (*cantar-lhe-ei uma canção* ‘I will sing him/her a song’). Furthermore, the presence of a third person clitic in the middle position (and also in the final position) can affect the verb: for example, in the final sequence -ar, the -r falls and the vowel is stressed (*cantá-la-ei* ‘I will sing it’).

## Recent developments

English being the most widespread language in the world, its influence on other languages, including Portuguese, is increasingly noticeable. Movies and television, especially American series, music and the Internet open a window to the regular presence of English in daily life and many words are eventually integrated into the Portuguese language.

It is mainly in the business language and on the Internet that English words are more visible, like *CEO*, *stock*, *manager*, *briefing*, *Casual Day* or *download*, *USB key*, *delete*, *upload*, *refresh*, *online*, *site* and also *lifting*, *e-learning*, *shopping*. The English influence is felt in European Portuguese and also in other Portuguese varieties in the world, especially the Brazilian one, which tends to adapt these loan words, like *deletar*, *googlar* and *twitar*.

In what concerns music, although there are many musical projects with English lyrics targeted to a younger audience, the projects sung in Portuguese like *Fado* and other traditional types of Portuguese music, which were considered less trendy for some time by younger people, are now regaining a large audience of all ages, and this reflects strongly on the Portuguese language.

In the last decade there has been a growth in the economic relevance of Portuguese in an international context, particularly due to the economic development of Brazil. Within the United Nations, Portuguese has played an increasingly important role, with ongoing initiatives for Portuguese to become one of its working languages, as it is already the case in the European Union and the Mercosul.

The growing importance of Portuguese at the international level is reflected in the increasing number of people taking Portuguese courses worldwide.

## Language cultivation in Portugal and abroad

There is no institution with the role of establishing the norm for the Portuguese language, unlike French (Académie Française) and Spanish (Real Academia Española), for example. The Academy of Sciences of Lisbon and the Brazilian Academy of Letters offer contributions in this direction, in particular with the publication of reference dictionaries: the Dictionary of Contemporary Portuguese, in Portugal, and the Dictionary of the Brazilian Academy of Letters in Brazil.

The Instituto Camões is an institution under the Portuguese Foreign Affairs Ministry and its main objective is the promotion of Portuguese in the world by giving support to cultural activities related to language, by awarding scholarships to nationals and foreigners in order to promote Portuguese, and by supporting Portuguese as a communication language on international levels, particularly in international institutions like the United Nations. This institution also coordinates Portuguese teaching abroad by establishing and supporting Portuguese language courses in foreign universities and centers for Portuguese language and culture.

The Community of Countries with Portuguese as Official Language (CPLP) is an inter-government organization for cooperation that has been active in the dissemination and promotion of the Portuguese language. The International Institute for the Portuguese Language (IILP) has been created in the scope of CPLP but is waiting for a stronger commitment by policy makers. It was also in the framework of CPLP that efforts were undertaken to prepare a new

agreement for the Portuguese orthography to support the expansion of the language and its consolidation in the international economic and political arena. After some initial resistance, this new Orthographic Agreement (started in 1990) includes all countries that have Portuguese as an official language: Portugal, Brazil, Angola, Mozambique, Guinea-Bissau, Cape Verde, Sao Tome and Principe and East Timor.

The Portuguese public radio and television have been engaged in the promotion of the Portuguese language by means of broadcast programs that seek to teach good practices regarding the use of standard Portuguese. For example, the weekly program "Watch your language" is both educational and entertaining and publicizes the New Orthographic Agreement. Also, public television and public radio issue daily a short program to clarify some frequent doubts regarding the Portuguese norm, and there are regular talks in public radio regarding good practices in Portuguese speaking and writing. There are also many publications concerned with the "safeguard" of the Portuguese language, seeking to attract more audiences to the appropriate use of Portuguese. All these programs and publications address a vivid interest by the Portuguese population regarding language issues.

The use of Portuguese is supported in the music sector by means of a quota system in the Portuguese radios, introduced by law, where there is a mandatory proportion of Portuguese music in broadcasted programs. This law first stipulated a quota ranging from 25% to 40% of Portuguese music and was eventually fixed in 25%.

The Portuguese language is also promoted through the increasing international projection of the cultural work of Portuguese speaking authors, like the philosophers José Gil and Eduardo Lourenço, as well as fictions writers like Antonio Lobo Antunes, Gonçalo M. Tavares, José Luis Peixoto, and the recently deceased Nobel prize José Saramago, whose works have been translated worldwide.

## Language in Education

In recent years, there has been a large investment in the development of a network of school libraries. This has been done under the scope of the National Plan for Reading whose key goal is to foster the literacy level of Portuguese students from various learning levels, but with special focus on the early years of school.

Another recent initiative has been the widespread integration of new Information Technologies in schools. Younger students have been granted the possibility to acquire at very low cost laptops especially designed for their different levels of education. In tandem with this access to individual laptops, educational software programs have been designed, where Portuguese is the language used, and in many cases where the learning of Portuguese grammar is specifically fostered. The results achieved by the students in the years to come will allow an in-depth assessment of this major investment on new technologies.

Recent results from the 2009 Programme for International Student Assessment (PISA) reveal a notorious comparative progress of the Portuguese students in terms of their reading, science and mathematics skills, with special highlights to the reading component.

It will be important to follow, in the near future, the impact of this investment in a National Plan for Reading and new technologies as well as of the recent measure to increase the compulsory school

attendance to 12 years and see its implications in the forthcoming PISA assessments.

## International aspects

Portuguese is a global language, with around 220 million speakers. In Portugal there are around 10 million speakers<sup>x</sup> and in many African countries Portuguese is an official language but co-exists with many other national languages (mostly Bantu languages in Angola and in Mozambique, Tetum in East Timor, and Portuguese-based Creoles in Cape Verde, Guinea-Bissau and Sao Tome and Principe). It is in fact Brazil which hosts the largest share of the global Portuguese-speaking community with its 190 million native speakers of Portuguese. On a par to the size of its population, Brazil is contributing to the increasingly larger international projection of the Portuguese language as a consequence of its economical development and of its position in the international arena as one of the emergent powers of the 21st century.

Therefore, a recent increase of interest in the Portuguese language is observed, especially in Latin-American countries, but also in Macau and in Spain, for instance. The Portuguese language is taught in many countries around the world<sup>xi</sup>. Several Chambers of Commerce have been interested in offering Portuguese lessons for potential investors in Portugal, as it was recently the case of the Italian Chamber, just to cite one case among many others. The Portuguese emigrant communities also promote the teaching of Portuguese in several European countries.

As a consequence of the historical undertaking of the Portuguese maritime explorations, geographical discoveries and settling of new global trade routes, which started in the 12th Century, the Portuguese language has been projected for centuries all over the world as one of the most prominent languages for business and trade. It is, nowadays, one of the 23 official languages of the European Union and has been included in many research projects funded by the European Commission targeting the development of language resources and technology. The Portuguese language is also an official, administrative or working language of 27 international organizations, including, for example, CPLP (The Community of Portuguese-speaking countries), the Latin Union, Mercosul and FIFA (Fédération Internationale de Football Association). Moreover, in recent years, some efforts have been undertaken to include the Portuguese language as an official language of the United Nations.

On a par with its progressive projection, the Portuguese language faces challenges in some contexts when it comes to its standing as an international language of communication. In South America, with around 190 million of native speakers, Portuguese co-exists with large Spanish speaking nations. In Europe, Portuguese has little more than 10 million speakers. In Asia, Portuguese is an official language only in East Timor and Macau. And in Africa, besides the fact that many native languages co-exist with Portuguese, English and French are languages with a vigorous projection in that continent.

## Portuguese on the Internet

An overview on statistical data about Portuguese language reveals that it is one of the most used languages in the internet. According to the last estimates, Portuguese is the fifth most common language on the web, being surpassed only by English, Chinese, Spanish and Japanese<sup>xii</sup>. This survey shows that about 82.5 million users are

surfing the web in Portuguese, and that in one decade, from 2000 to 2010, it registered an astonishing expansion of 990%.

Portuguese is particularly well positioned when it comes to its presence in social networks. A semantic and quantitative study of 2.8 million tweets, performed by Semiocast, reveals that Portuguese is the third language most used on Twitter, coming right after English and Japanese.<sup>xiii</sup>

This is in line with the boom of Internet access in Brazil, especially among the young people. This country has one of the largest numbers of Internet users worldwide (76 million)<sup>xiv</sup>, and a census questionnaire revealed that the number of people aged 10 or older using the Internet jumped by 12 million since 2008<sup>xv</sup>. Portugal in turn has around 5 million Internet users<sup>xvi</sup> and has also registered a notorious growth in terms of Internet access. Statistics reveal that the number of Internet subscribers has steadily increased: in 2001 there were 466.813 subscribers, and the last counting indicates 1.898.026<sup>xvii</sup>. They reveal also that in 2010 54% of Portuguese households had an Internet connection, that in 2008 more than 90% of individuals aged between 10 and 15 years used a computer (96.6%) and the Internet (92.7%), and that in 2006 95% of companies with ten or more employees used computers, while 84% used e-mail and 83% had access to the Internet<sup>xviii</sup>.

On a par to the effort of assuring the presence of public institutes, agencies and services on the internet, a National Plan for the Promotion of Accessibility has been implemented in Portugal, in 2007, together with specific legislation<sup>xix</sup> targeted to foster social inclusion through the Information Society and to allow e-content access to citizens with disabilities.

An increasing usage of the Portuguese language in the internet is thus clear. Along with the data shown above, it is worth pointing out that Portuguese is present in several websites of political and economical institutions, as in the sites of the European Union or the Mercosul, just to give two examples, though efforts should be continued so that it will be present in a number of others where it is not yet an option.

### Selected Further Reading

Cardeira, Esperança, 2006, *O Essencial sobre a História do Português* Lisboa: Editorial Caminho.

Lewis, M. Paul (ed.), 2009. *Ethnologue: Languages of the World*, Sixteenth edition. Dallas, Tex.: SIL International. Online version: <http://www.ethnologue.com>.

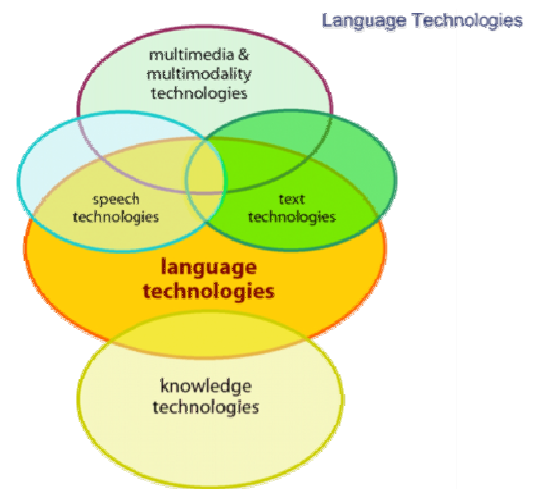
Pinto, Paulo Feytor, 2010, *Novo Acordo Ortográfico da Língua Portuguesa*. Lisboa: INCM.

Centro Virtual Camões(<http://cvc.instituto-camoes.pt/conhecer/bases-tematicas.html>)

# Language Technology Support for Portuguese

## Language Technologies

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



## Language Technology Application Architectures

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

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

Task-specific modules then perform many different operations such as automatic summarization of an input text, database look-ups and many others. Below, we will illustrate **core application areas** and highlight certain of the modules of the different architectures in each section. Again, the architectures are highly simplified and idealised, serving for illustrating the complexity of language technology applications in a generally understandable way.

After the introduction of the core application areas, we will shortly give an overview of the situation in LT research and education, concluding with an overview of (past) funding programs. In the end of this section, we will present an expert estimation on the situation regarding core LT tools and resources in a number of

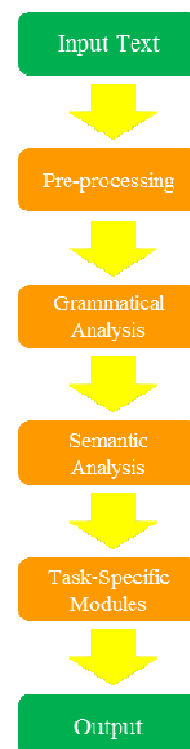


Figure 2: A Typical Text Processing Application Architecture

dimensions such as availability, maturity, or quality. This table gives a good overview on the situation of LT for Portugal.

The most important tools and resources involved are underlined in the text and can also be found in the table at the end of the chapter. The sections discussing the core application areas also contain an overview of the industries active in the respective field in Portugal and Brazil.

## Core application areas

### Language checking

Anyone using a word processing tool such as Microsoft Word has come across a spell checking component that indicates spelling mistakes and proposes corrections. 40 years after the first spelling correction program by Ralph Gorin, language checkers nowadays do not simply compare the list of extracted words against a dictionary of correctly spelled words, but have become increasingly sophisticated. In addition to language-dependent algorithms for handling morphology (e.g. plural formation), some are now capable of recognizing syntax-related errors, such as a missing verb or a verb that does not agree with its subject in person and number, e.g. in 'She *\*write* a letter.' However, most available spell checkers (including Microsoft Word) will find no errors in the following first verse of a poem by Jerrold H. Zar (1992):

*Eye have a spelling chequer,*

*It came with my Pea Sea.*

*It plane lee marks four my revue*

*Miss Steaks I can knot sea.*

For handling this type of errors, analysis of the context is needed in many cases, as in the following Portuguese examples:

*Fizemos jogos tradicionais, incluindo o jogo do pião.*

*[We played traditional games, including the whipping top game.]*

*Fizemos jogos tradicionais, incluindo o jogo do peão.*

*[We played traditional games, including the game of the pedestrian.]*

This either requires the formulation of language-specific grammar rules, i.e. a high degree of expertise and manual labour, or the use of a so-called statistical language model. Such models calculate the probability of a particular word occurring in a specific environment (i.e., the preceding and following words). For example, *o jogo do pião* is a much more probable word sequence than *o jogo do peão*. A statistical language model can be automatically derived using a large amount of (correct) language data (i.e. a corpus). Up to now, these approaches have mostly been developed and evaluated on English language data. However, they do not necessarily transfer straightforwardly to Portuguese with its richer inflection.

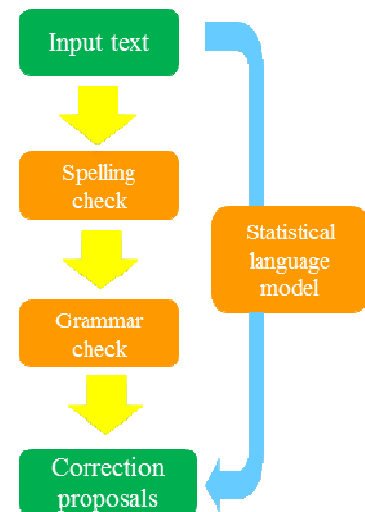


Figure 3: Language Checking (left: rule-based; right: statistical)

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

Additionally to the one provided by Microsoft Word, there are some other Language Checking tools for Portuguese. In Portugal, the Priberam company created FLIP, a language checker available for Portuguese (both European Portuguese and Brazilian Portuguese) and Spanish, which suggests syntactic and orthographic corrections. CoGrOO is a Brazilian Portuguese grammar checker for Open Office. Also for this variety, NILC, an Interinstitutional Center for Research and Development in Computational Linguistics, developed ReGra, which is available as an integral part of the Microsoft Word and the word processor REDATOR.

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

## Web search

Search on the web, in intranets, or in digital libraries is probably the most widely used and yet underdeveloped Language Technology today.

The search engine Google, which started in 1998, is nowadays used for about 80% of all search queries world-wide<sup>xx</sup>. The verb *googlar* even has an entry in the Porto Editora online dictionary<sup>xxi</sup>. Neither the search interface nor the presentation of the retrieved results has significantly changed since the first version. In the current version, Google offers a spelling correction for misspelled words and also, in 2009, incorporated basic semantic search capabilities into their algorithmic mix<sup>xxii</sup>, which can improve search accuracy by analysing the meaning of the query terms in context. The success story of Google shows that with a lot of data at hand and efficient techniques for indexing these data, a mainly statistically-based approach can lead to satisfactory results.

However, for a more sophisticated request for information, integrating deeper linguistic knowledge is essential. In the research labs, experiments using machine-readable thesauri and ontological language resources like WordNet have shown improvements by allowing to find a page on the basis of synonyms of the search terms (e.g. atomic energy, atomic power, and nuclear energy) or even more loosely related terms. In this connection, the WordNets for Portuguese (e.g., MWN.PT and WordNet.PT) will be useful towards this end.

The next generation of search engines will have to include much more sophisticated Language Technology. If a search query consists of a question or another type of sentence rather than a list of keywords, retrieving relevant answers to this query requires an

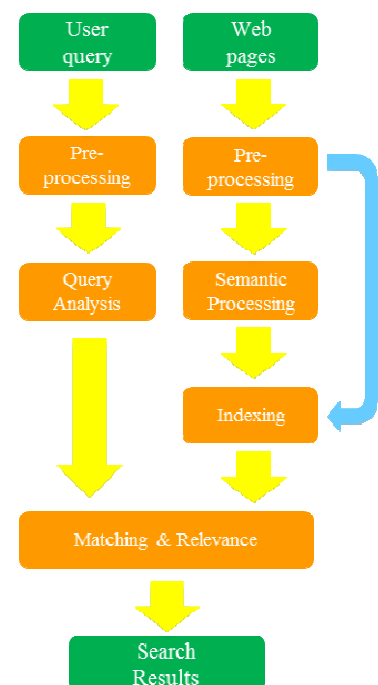


Figure 4: Web Search Architecture

analysis of this sentence on a syntactic and semantic level as well as the availability of an index that allows for a fast retrieval of the relevant documents. For example, imagine a user inputs the query ‘Give me a list of all companies that were taken over by other companies in the last five years’. For a satisfactory answer, syntactic parsing needs to be applied to analyse the grammatical structure of the sentence and determine that the user is looking for companies that have been taken over and not companies that took over others. Also, the expression *last five years* needs to be processed in order to find out which years it refers to.

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

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

In the late 90's, several search engines started being developed in Portugal. AEIOU, which came up in 1996, was later bought by Impresa and developed further to a content portal<sup>xxiii</sup>. Sapo was launched in 1997 as a search engine as well, becoming then into a portal and being now an internet service provider owned by PT Multimédia<sup>xxiv</sup>. In the meanwhile, Sapo created search engine versions for Angola, Cape Verde, Mozambique and East Timor. As of today, although many other Portuguese search engines have been created (Clix, Tumba, Busca Online, Guianet, Netindex, among others)<sup>xxv</sup>, only few Portuguese companies keep providing self-developed search engine services, and the search engine Google.pt is clearly the most popular.

The Brazilian situation is somewhat different. There are examples of Web Search engines that are directed to Brazilian sites only, such as Achei<sup>xxvi</sup> or Giga Busca<sup>xxvii</sup>, but they are fewer than in Portugal, and their coverage and outreach is fairly limited. Therefore, Google is largely the dominant search engine also in Brazil.

## Speech interaction

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

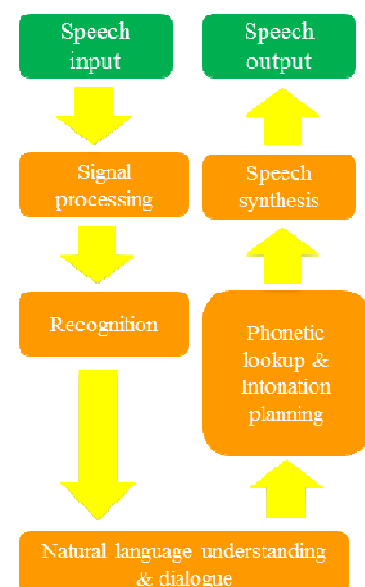


Figure 5: Simple Speech-based Dialogue Architecture

spoken language as an alternative to the input/output modalities of graphical user interfaces, e.g. in smartphones.

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

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

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

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

Regarding the market for speech interaction technology, the last decade has been characterised by a strong standardisation of the interfaces between the different technology components, as well as by standards for creating particular software artefacts for a given application. There also has been strong market consolidation within the last ten years, particularly in the field of ASR and TTS.

Here, the national markets in the G20 countries – i.e. economically strong countries with a considerable population - are dominated by less than 5 players worldwide, with *Nuance* and *Loquendo* being the most prominent ones in Europe.

On the Portuguese TTS market, there further exist some smaller companies like *SVOX* and *Voice Interaction*, and the latter has a differentiating focus by providing voices not only for European and Brazilian Portuguese but also for the African varieties of Portuguese.

Regarding dialogue management technology and know-how, DigA is the only complete framework, especially built for European Portuguese: it is open-domain but is not available as open-source. The open-source Olympus SDS was adapted to Portuguese with success, but not extensively tested so far. From the various modules required by Spoken Dialogue Systems, the dialog manager is the only module that is language-independent. These other modules exist, although usually not available for free and not as open-source frameworks, but the language adaptation task is time- and effort-consuming.

Finally, within the domain of *speech* interaction, a genuine market for the linguistic core technologies for syntactic and semantic analysis does not exist yet.

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

As for dialogue management technology, it will be crucial that it opens its scope to support multimodal interaction scenarios, as those created by smartphone usages, as well as multiple user interfaces channels, given some common model of domain specific interaction behaviour. Apart from ongoing research on the optimisation of ASR and TTS, it is the latter fields of domain-specific customisation of linguistic core technologies and of generic dialogue management that appear most relevant as fields of transformation between applied research and industrial productization.

## Machine translation

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

At its basic level, MT simply substitutes words in one natural language by words in another. This can be useful in subject domains with a very restricted, formulaic language, e.g., weather reports. However, for a good translation of less 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 here lies in the fact that human language is ambiguous, which yields challenges on multiple levels. For example, word sense disambiguation is a challenge on the lexical level: ‘Jaguar’ can mean a car or an animal and ‘banco’ in Portuguese has at least two meanings, ‘bank’ or ‘bench’:

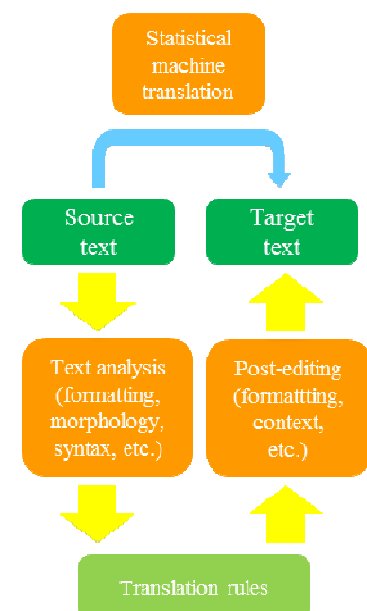


Figure 6: Machine translation (top: statistical; bottom: rule-based)

*O rapaz viu a rapariga no banco.*

*[The boy saw the girl at the bank / on the bench.]*

Syntactic ambiguity is also a challenge, e.g., the attachment of prepositional phrases in these two sentences:

*O polícia viu o homem com o telescópio.*

*[The policeman observed the man with the telescope.]*

*O polícia viu o homem com o revólver.*

*[The policeman observed the man with the revolver.]*

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

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

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

For Portuguese, the lack of effective Word Sense Disambiguation (WSD) mechanisms is one of the main reasons why the results of the existent MT systems are often insufficient.

Besides, while languages like German, for instance, form compounds as one word, the tendency in Portuguese is to write compounds as phrases, i.e., separate words which form a lexical unit. This can be a specific challenge for MT involving languages like Portuguese in this respect.


Leading MT rule-based systems, like LOGOS, Apertium and SYSTRAN, are available for Portuguese. While there is significant research in this technology in national and international contexts,

data-driven and hybrid systems have been less successful in business than in research so far.

Provided good adaptation in terms of user-specific terminology and workflow integration, the use of MT can increase productivity significantly. Special systems for interactive translation support were developed, e.g., at Siemens. Language portals, such as the one of Volkswagen, provide access to dictionaries and company-specific terminology, translation memory and MT support.

The quality of MT systems is still considered to have huge improvement potential. Challenges include the adaptability of the language resources to a given subject domain or user area and the integration into existing workflows with term bases and translation memories. In addition, most of the current systems are English-centred and support only few languages being translated from and into Portuguese, which leads to shortcomings in the total translation workflow, and e.g. forces MT users to learn different lexicon coding tools for different systems.

Evaluation campaigns allow for comparing the quality of MT systems, the various approaches and the status of MT systems for the different languages. The following table<sup>xxviii</sup>, presented within the EC Euromatrix+ project, shows the pairwise performances obtained for 22 official EU languages (Irish Gaelic is missing) in terms of BLEU score<sup>xxix</sup>.



# Translating between all EU-27 languages

	Target Language																					
	en	bg	de	cs	da	el	es	et	fi	fr	hu	it	lt	lv	mt	nl	pl	pt	ro	sk	sl	sv
en	—	40.5	46.8	52.6	50.0	41.0	55.2	34.8	38.6	50.1	37.2	50.4	39.6	43.4	39.8	52.3	49.2	55.0	49.0	44.7	50.7	52.0
bg	61.3	—	38.7	39.4	39.6	34.5	46.9	25.5	26.7	42.4	22.0	43.5	29.3	29.1	25.9	44.9	35.1	45.9	36.8	34.1	34.1	39.9
de	53.6	26.3	—	35.4	43.1	32.8	47.1	26.7	29.5	39.4	27.6	42.7	27.6	30.3	19.8	50.2	30.2	44.1	30.7	29.4	31.4	41.2
cs	58.4	32.0	42.6	—	43.6	34.6	48.9	30.7	30.5	41.6	27.4	44.3	34.5	35.8	26.3	46.5	39.2	45.7	36.5	43.6	41.3	42.9
da	57.6	28.7	44.1	35.7	—	34.3	47.5	27.8	31.6	41.3	24.2	43.8	29.7	32.9	21.1	48.5	34.3	45.4	33.9	33.0	36.2	47.2
el	59.5	32.4	43.1	37.7	44.5	—	54.0	26.5	29.0	48.3	23.7	49.6	29.0	32.6	23.8	48.9	34.2	52.5	37.2	33.1	36.3	43.3
es	60.0	31.1	42.7	37.5	44.4	39.4	—	25.4	28.5	51.3	24.0	51.7	26.8	30.5	24.6	48.8	33.9	57.3	38.1	31.7	33.9	43.7
et	52.0	24.6	37.3	35.2	37.8	28.2	40.4	—	37.7	33.4	30.9	37.0	35.0	36.9	20.5	41.3	32.0	37.8	28.0	30.6	32.9	37.3
fi	49.3	23.2	36.0	32.0	37.9	27.2	39.7	34.9	—	29.5	27.2	36.6	30.5	32.5	19.4	40.6	28.8	37.5	26.5	27.3	28.2	37.6
fr	64.0	34.5	45.1	39.5	47.4	42.8	60.9	26.7	30.0	—	25.5	56.1	28.3	31.9	25.3	51.6	35.7	61.0	43.8	33.1	35.6	45.8
hu	48.0	24.7	34.3	30.0	33.0	25.5	34.1	29.6	29.4	30.7	—	33.5	29.6	31.9	18.1	36.1	29.8	34.2	25.7	25.6	28.2	30.5
it	61.0	32.1	44.3	38.9	45.8	40.6	26.9	25.0	29.7	52.7	24.2	—	29.4	32.6	24.6	50.5	35.2	56.5	39.3	32.5	34.7	44.3
lt	51.8	27.6	33.9	37.0	36.8	26.5	21.1	34.2	32.0	34.4	28.5	36.8	—	40.1	22.2	38.1	31.6	31.6	29.3	31.8	35.3	35.3
lv	54.0	29.1	35.0	37.8	38.5	29.7	8.0	34.2	32.4	35.6	29.3	38.9	38.4	—	23.3	41.5	34.4	39.6	31.0	33.3	37.1	38.0
mt	72.1	32.2	37.2	37.9	38.9	33.7	48.7	26.9	25.8	42.4	22.4	43.7	30.2	33.2	—	44.0	37.1	45.9	38.9	35.8	40.0	41.6
nl	56.9	29.3	46.9	37.0	45.4	35.3	49.7	27.5	29.8	43.4	25.3	44.5	28.6	31.7	22.0	—	32.0	47.7	33.0	30.1	34.6	43.6
pl	60.8	31.5	40.2	44.2	42.1	34.2	46.2	29.2	29.0	40.0	24.5	43.2	33.2	35.6	27.9	44.8	—	44.1	38.2	38.2	39.8	42.1
pt	60.7	31.4	42.9	38.4	42.8	40.2	60.7	26.4	29.2	53.2	23.8	52.8	28.0	31.5	24.8	49.3	34.5	—	39.4	32.1	34.4	43.9
ro	60.8	33.1	38.5	37.8	40.3	35.6	50.4	24.6	26.2	46.5	25.0	44.8	28.4	29.9	28.7	43.0	35.8	48.5	—	31.5	35.1	39.4
sk	60.8	32.6	39.4	48.1	41.0	33.3	46.2	29.8	28.4	39.4	27.4	41.8	33.8	36.7	28.5	44.4	39.0	43.3	35.3	—	42.6	41.8
sl	61.0	33.1	37.9	43.5	42.6	34.0	47.0	31.1	28.8	38.2	25.7	42.3	34.6	37.3	30.0	45.9	38.2	44.1	35.8	38.9	—	42.7
sv	58.5	26.9	41.0	35.6	46.6	33.3	46.6	27.4	30.9	38.9	22.7	42.0	28.2	31.0	23.7	45.6	32.2	44.2	32.7	31.3	33.5	—

(using the Acquis corpus)

[from Koehn et al., 2009]

Philipp Koehn, U Edinburgh

EuroMatrixPlus

23 March 2010

The best results (shown in green and blue) were achieved by languages which benefit from considerable research efforts within coordinated programs, from the existence of many parallel corpora or from being similar to other languages (e.g. English, French, Dutch, Spanish, Portuguese, German), the worst (in red) by languages that did not benefit from similar efforts, or that are very different from other languages (e.g. Hungarian, Maltese, Finnish).

## Language Technology ‘behind the scenes’

Building Language Technology applications involves a range of subtasks that do not always surface at the level of interaction with the user, but provide significant service functionalities ‘under the

hood' of the system. Therefore, they constitute important research issues that have become individual sub-disciplines of Computational Linguistics in academia.

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

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

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

In these areas, the Portuguese language has been less researched than English, where question answering, information extraction, and summarization have since the 1990s been the subject of nu-

merous funded competitions, primarily those organized by DARPA/NIST in the United States. These have significantly improved the state of the art, but the focus has always been on English; some competitions have added multilingual tracks, but Portuguese, like other languages, has not received sufficient support. Accordingly, there are hardly any annotated corpora or other resources for these tasks. Summarization systems, when using purely statistical methods, to a considerable extent are often language independent, and thus some research prototypes are available. However, a summarization tool using statistical methods but based on the gist of the text already exists specifically for Portuguese. For text generation, reusable components have traditionally been limited to the surface realization modules (the "generation grammars"); again, most available software is for English, and in this area there are no available tools for Portuguese. Similarly, we can find only a very limited number of question answering systems for Portuguese.

## Language Technology in Education

The area of Language Technology (LT) stands out for being inherently interdisciplinary, involving a wide range of scientific fields such as linguistics and computer science, but also statistics, engineering and psycholinguistics, among many others. Portugal has a reasonable offer in this area with respect to higher education, where the relevant courses are usually integrated in departments offering undergraduate studies in Translation, Language Science or Computer Science.

The area of LT has been fostered in several universities, both in education (majors, Masters and PhD degrees) and in research centres. At the University of Lisbon, on a par to several courses at different levels of education, including a minor in Natural Language Processing and an MA and PhD programs in Cognitive Science, there are major research centers focusing on LT. The Natural Language and Speech Group (NLX), from the Department of Informatics, is the national leading team in the computational processing of Portuguese and has an online center providing a comprehensive set of linguistic processing services (Lx-Center). The Center of Linguistics (CLUL), from the Faculty of Letters, has a long tradition in producing standard, dialectal and historical language resources, including a large-scale corpus and smaller and specific data sources available online.

The Instituto Superior Técnico (IST), located in Lisbon also offers courses in LT and has a doctoral program in Computer Science in collaboration with other Portuguese universities and with the Carnegie Mellon University. INESC is a research institution associated to IST and its Laboratory of Spoken Language Systems (L2f) is the national leader in speech recognition and synthesis.

The New University of Lisbon also has courses and research units working in the LT field, namely its Centre for Research in Computing and Information Technology (CITI) and its Center of Linguistics (CLUNL).

In Lisbon, there is also ILTEC, an institute devoted to theoretical and computational linguistics. Other universities in the country also provide courses in the area of LT and host other research units: the Centre for Research in Information Technology in the University of Evora; the Center for General and Applied Linguistic Studies (CELGA) in Coimbra; the Centre for Human Language Technology and Bioinformatics (HULTIG), in the University of

Beira Interior; the Center of Linguistics (CLUP) and the Laboratory for Artificial Intelligence and Computer Science (LIACC), in the University of Porto; the Center for Humanities Studies (CEHUM), in the University of Minho, Braga. And the University of Algarve is cooperating in an European Erasmus MA in Natural Language Processing and Human Language Technology.

In Brazil there has been also considerable activity in LT both in terms of education and research, that concentrates mostly around the axis São Paulo – Rio de Janeiro, and around Porto Alegre, in the South. Courses in this area have been offered more at the post-graduation level, in MA and PhD programs, rather than at the undergraduate level.

In the other Portuguese-speaking countries, the LT area shows little or no development, the data collection and the development of resources and tools targeted to Portuguese varieties in Africa are being undertaken mostly by Portuguese research centers.

## Language Technology Programs

The activity in LT in Portugal can be traced back to projects, programs or initiatives carried out in the last decades. One of the first important programs in this area was EUROTRA, an ambitious Machine Translation project established and funded by the European Commission from the late 1970s until 1994. The participation of Portugal in this project since 1986, was undertaken by the Institute of Theoretical and Computational Linguistics (ILTEC), specifically created for this purpose. This project had a long-lasting impact on the language industries in Europe, with Portugal being no exception. The EUROTRA project promoted a significant starting step for consistently pursued LT activities in Portugal and for the setting up of a Portuguese community of researchers in this area.

Another European key project in LT involving Portuguese was LE-PAROLE, developed in the late 90's, with the participation of CLUL and INESC. Its main achievement was the building of corpora and lexicons according to integrated models of composition and materials description. For each language, a 20 million word corpus was built with harmonized design, composition and codification, including a 250.000 word tagged subcorpus. Each language lexicon is composed of 20.000 entries with syntactic and morphosyntactic information.

This corpus has been enriched and enlarged under the national project TagShare, in Portugal, conducted at the University of Lisbon, in the Department of Informatics (NLX) and in the Center of Linguistics (CLUL), in 2005. This project enabled the development of a set of linguistic resources and software component tools to support the computational processing of Portuguese. The outcome was a 1 Million word corpus linguistically annotated and fully verified by experts – the CINTIL corpus<sup>xxx</sup> –, and a whole range of processing tools for tokenization, morphosyntactic category (POS) tagging, inflection analysis, lemmatization, multi-word lexeme recognition, named entity recognition, etc. and. The annotation schemes developed in the project became de facto standards for Portuguese in the field of LT and have been further used in the Reference Corpus of Contemporary Portuguese (CRPC).

The Corpus de Extractos de Textos Electrónicos MCT/Público (CETEMPúblico) is a corpus of about 180 million words from texts of a Portuguese daily newspaper, released in 2000. It is intended primarily to support the development of processing tools for the

Portuguese language which need raw texts for their construction and testing. This corpus was created by the project Computational Processing of Portuguese, under a protocol between the Ministry of Science and Technology (MCT) and that newspaper. This project subsequently evolved into Linateca, a long term project for Portuguese LT<sup>xxxi</sup>.

On the industry side, it is worth mentioning the important contribution for the emerging of an LT industry in Portugal of the establishment of the international Microsoft Language Development Center, near Lisbon, since 2005.

More recently, Portuguese and Brazilian institutions have been participating in the ongoing CLARIN project, aiming at establishing an integrated and interoperable European **research infrastructure** of language resources and technology.

In Brazil, relevant efforts in LT support to Portuguese have been also undertaken. To mention just a couple of illustrative examples, in the early 90's, under the DIRECT project the Bank of Portuguese was created at the Pontifical Catholic University of São Paulo. Since its inception, the Bank of Portuguese has been a source of data for corpus-based studies for several projects. Also worth mentioning is the Summ-it corpus, a corpus built to support the study of summarization along with the phenomena of anaphoric and rhetorical relations in Portuguese. This resource was developed under the PLN-BR project, by the Núcleo Interinstitucional de Lingüística Computacional (NILC), driven by the University of S. Paulo and gathering researchers from other institutions.

The above notes cover only a few illustrative examples of projects, programmes and initiatives in LT addressing the Portuguese language. Although these are part of positive developments for the Portuguese language in recent years, the fact is that there is a large gap with respect to the LT activity on other more researched languages, for which the development of language resources and technology is far more advanced.

Compared to the level of funding for LT in the U.S., the support for this area in Portugal and in other European countries is still very low. In Portugal, funding for this area comes mainly from the Ministry of Science, Technology and Higher Education, through the Foundation for Science and Technology (FCT). However, obtaining support for LT projects is particularly difficult because project proposals in this area are accepted and evaluated under the Electrical Engineering tracks in calls for project proposals, where they have to compete with hundreds of proposals on totally unrelated issues. On a par with FCT, the Fundação Calouste Gulbenkian occasionally funds some LT projects.

In Brazil, funding for research in general, and for LT activities in particular, comes mainly from governmental agencies. The National Council for Scientific and Technological Development (CNPq), the São Paulo Research Foundation (FAPESP), the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), and Research and Projects Financing (FINEP) are the four institutions that significantly support research in this country. Some of them have provided also special joint university-industry funding programs. For instance, FAPESP and Microsoft Research recently formed a partnership to fund socially relevant projects in the state of São Paulo, which included, for instance, the PorSimples text simplification project in the area of LT.

## Availability of tools and resources for Portuguese

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

- 1 **Quantity:** Does a tool/resource exist for the language at hand? The more tools/resources exist, the higher the rating.
  - 0: no tools/resources whatsoever
  - 6: many tools/resources, large variety
- 2 **Availability:** Are tools/resources accessible, i.e., are they Open Source, freely usable on any platform or only available for a high price or under very restricted conditions?
  - 0: practically all tools/resources are only available for a high price
  - 6: a large amount of tools/resources is freely, openly available under sensible Open Source or Creative Commons licenses that allow re-use and re-purposing
- 3 **Quality:** How well are the respective performance criteria of tools and quality indicators of resources met by the best available tools, applications or resources? Are these tools/resources current and also actively maintained?
  - 0: toy resource/tool
  - 6: high-quality tool, human-quality annotations in a resource
- 4 **Coverage:** To which degree do the best tools meet the respective coverage criteria (styles, genres, text sorts, linguistic phenomena, types of input/output, number languages supported by an MT system etc.)? To which degree are resources representative of the targeted language or sublanguages?
  - 0: special-purpose resource or tool, specific case, very small coverage, only to be used for very specific, non-general use cases
  - 6: very broad coverage resource, very robust tool, widely applicable, many languages supported
- 5 **Maturity:** Can the tool/resource be considered mature, stable, ready for the market? Can the best available tools/resources be used out-of-the-box or do they have to be adapted? Is the performance of such a technology adequate and ready for production use or is it only a prototype that cannot be used for production systems? An indicator may be whether resources/tools are accepted by the community and successfully used in LT systems.
  - 0: preliminary prototype, toy system, proof-of-concept, example resource exercise
  - 6: immediately integratable/applicable component
- 6 **Sustainability:** How well can the tool/resource be maintained/integrated into current IT systems? Does the tool/resource fulfil a certain level of sustainability concerning documentation/manuals, explanation of use cases, front-ends, GUIs etc.? Does it use/employ standard/best-practice programming environments (such as Java EE)? Do indus-

try/research standards/quasi-standards exist and if so, is the tool/resource compliant (data formats etc.)?

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

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

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

## Table of Tools and Resources

	Quantity	Availability	Quality	Coverage	Maturity	Sustainability	Adaptability
Language Technology (Tools, Technologies, Applications)							
Tokenization, Morphology (tokenization, POS tagging, morphological analysis/generation)	4	2	4	5	5	2	5
Parsing (shallow or deep syntactic analysis)	2	4	4	3	4	3	4
Sentence Semantics (WSD, argument structure, semantic roles)	1	3	4	3	3	3	4
Text Semantics (coreference resolution, context, pragmatics, inference)	2	1	3	1	2	2	1
Advanced Discourse Processing (text structure, coherence, rhetorical structure/RST, argumentative zoning, argumentation, text patterns, text types etc.)	1	2	2	2	2	2	2
Information Retrieval (text indexing, multimedia IR, crosslingual IR)	0	0	0	0	0	0	0
Information Extraction (named entity recognition, event/relation extraction, opinion/sentiment recognition, text mining/analytics)	2	2	4	3	2	1	3
Language Generation (sentence generation, report generation, text generation)	0	0	0	0	0	0	0
Summarization, Question Answering, advanced Information Access Technologies	2	3	4	2	3	1	2
Machine Translation	3	2	2	3	4	2	2
Speech Recognition	2	3	4	2	2	2	4

	Quantity	Availability	Quality	Coverage	Maturity	Sustainability	Adaptability
Speech Synthesis	3	3	4	4	4	3	4
Dialogue Management (dialogue capabilities and user modelling)	1	1	3	3	4	2	4
Language Resources (Resources, Data, Knowledge Bases)							
Reference Corpora	4	3	4	5	4	5	5
Syntax-Corpora (treebanks, dependency banks)	2	3	4	4	4	4	4
Semantics-Corpora	1	1	4	3	3	4	4
Discourse-Corpora	1	1	2	2	1	1	1
Parallel Corpora, Translation Memories	2	4	3	2	2	3	3
Speech-Corpora (raw speech data, labelled/annotated speech data, speech dialogue data)	4	2	4	4	4	3	3
Multimedia and multimodal data (text data combined with audio/video)	0	0	0	0	0	0	0
Language Models	0	0	0	0	0	0	0
Lexicons, Terminologies	5	4	5	4	4	3	3
Grammars	1	4	5	2	2	2	2
Thesauri, WordNets	2	2	4	2	4	3	3
Ontological Resources for World Knowledge (e.g. upper models, Linked Data)	2	2	2	2	3	2	1

## Conclusions

The situation of Portuguese concerning language technology support have been steadily improving but still requires continued effort to reach a sustained ground of development. Immediate action must occur so that important progress for the Portuguese language can be attained.

For Portuguese, a number of resources and processing tools exist, but far less than for English. Still, this comparison has to be taken with care. Even for English, language technology support today is by far not in a state that is required for supporting a true multilingual knowledge society. Noteworthy is the fact that a network of research centers, both from Portugal and Brazil, has been set up and should promote the advancement of language technology for Portuguese in the near future if funding will be properly secured.

In this Whitepaper Series, the first effort has been made to assess the overall situation of many European languages with respect to

language technology support in a way that allows for high level comparison and identification of gaps and needs.

For Portuguese, key results regarding technologies and resources include the following:

- Although certain specific subareas in the field have been very active, Portuguese is a less resourced language specially if compared to languages from countries with much larger expenditure in R&D, like English, German or Dutch;
- Two large corpora were compiled for Portuguese, but one lacks representativeness, as it covers only one text type (newspaper), and the other is not fully available due to copyright restrictions;
- For less studied varieties of Portuguese, corpora are being compiled during the last years but they still need to receive more attention;
- A de facto standard 1M word tagged corpus is available together with the respective POS tagger, though it needs to be upgraded to international agreed formats;
- Concerning speech technologies, a variety of commercial systems exist for both European and Brazilian varieties, for speech recognition, speech synthesis and statistical dialog management but, although Portuguese and Brazilian teams are very active in the field, tools and annotated corpora are usually reserved for internal use and not freely available;
- While many corpora have POS annotation and other types of morphological information, syntactically annotated corpora are more rare;
- Some parsers were developed but most are still very limited, as well as summarization and question answering systems;
- Annotated corpora with semantic information are missing, leading to the worrisome situation that no processing tools or research exists yet for word sense disambiguation in Portuguese;
- Parallel corpora for machine translation which include Portuguese are essentially the ones made available by EU initiatives and are consequently very limited in terms of text type;
- More work needs to be dedicated to lexical resources and word-nets;
- Tools addressing text and discourse annotation are few and partial;
- The more linguistic and semantic knowledge a tool takes into account, the more gaps exist (see, e.g., information retrieval vs. text semantics)
- More efforts for supporting deep linguistic processing are thus needed.

From this, it is clear that more efforts need to be directed into the creation of resources for Portuguese and into research, innovation, and development of processing tools. The need for large amounts of data and the high complexity of language technology systems make it also mandatory to develop new infrastructures for sharing and cooperation.

## 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.

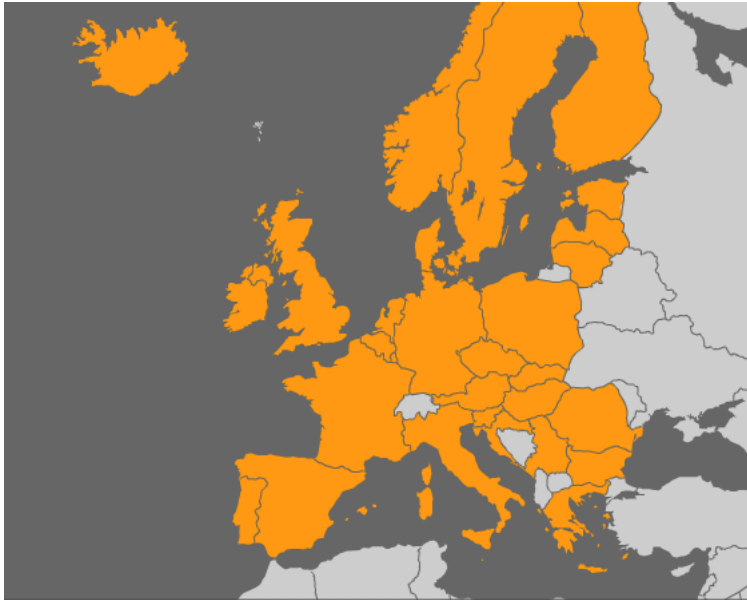


Figure 1: Countries Represented in META-NET

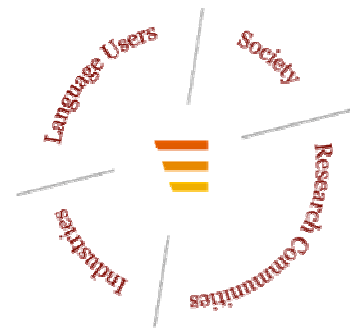
META-NET cooperates with other initiatives like the Common Language Resources and Technology Infrastructure (CLARIN), which is helping establish digital humanities research in Europe. META-NET fosters the technological foundations for the establishment and maintenance of a truly multilingual European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

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

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

## Lines of Action

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



*The Multilingual Europe Technology Alliance (META)*

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



Figure 2: Three Lines of Action in META-NET

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

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

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

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

## Member Organisations

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

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

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

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

## References

---

- <sup>i</sup> 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](http://ec.europa.eu/public_opinion/flash/fl_313_en.pdf)).
- <sup>ii</sup> 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](http://ec.europa.eu/education/languages/pdf/com/2008_0566_en.pdf)).
- <sup>iii</sup> UNESCO Director-General, *Intersectoral mid-term strategy on languages and multilingualism*, Paris, 2007 (<http://unesdoc.unesco.org/images/0015/001503/150335e.pdf>).
- <sup>iv</sup> 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>).
- <sup>v</sup> Cf. <http://observatorio-lp.sapo.pt/pt/dados-estatisticos/falantes-de-portugues>) and [www.ethnologue.com](http://www.ethnologue.com).
- <sup>vi</sup> <http://stats.oecd.org/Index.aspx?datasetcode=MIG>;  
<http://www.observatorioemigracao.secomunidades.pt/np4/home.html>)
- <sup>vii</sup> Cf. <http://www.portugal-linha.pt/>
- <sup>viii</sup> Cf. <http://cvc.instituto-camoes.pt/index.php>; d'Andrade et. al. (1999) orgs. *Crioulos de Base Portuguesa*. Lisboa: APL.
- <sup>ix</sup> Cf. Lindley Cintra, L. F. (1971) "Nova proposta de classificação dos dialectos galego-portugueses", *Boletim de Filologia*. Lisboa: Centro de Estudos Filológicos, 22, pp. 81-116.
- <sup>x</sup>According to the census of 2001.
- <sup>xi</sup> <http://www.instituto-camoes.pt/missao-do-instituto-camoes/instituto-camoess-mission.html>
- <sup>xii</sup> Cf. <http://www.internetworldstats.com/stats7.htm>.
- <sup>xiii</sup> Cf. <http://twtrcon.com/2010/02/25/top-5-language-on-twitter-are-english-japanese-portuguese-malay-and-spanish/>.
- <sup>xiv</sup> Cf. <http://www.internetworldstats.com/top20.htm>
- <sup>xv</sup> Cf. <http://mybroadband.co.za/news/internet/15031-Internet-access-Brazil-booms.html>.
- <sup>xvi</sup> Cf. <http://www.internetworldstats.com/stats4.htm>,  
<http://www.internetworldstats.com/stats15.htm>.
- <sup>xvii</sup> Cf. <http://www.pordata.pt>.
- <sup>xviii</sup> Cf. <http://www.pordata.pt>.
- <sup>xix</sup> [http://www.unic.pt/index.php?option=com\\_content&task=view&id=2777&Itemid=40](http://www.unic.pt/index.php?option=com_content&task=view&id=2777&Itemid=40)
- <sup>xx</sup> Cf. <http://www.spiegel.de/netzwelt/web/0,1518,619398,00.html>
- <sup>xxi</sup> <http://www.infopedia.pt/verbos-portugueses/googlar>
- <sup>xxii</sup>  
Cf. [http://www.pcworld.com/businesscenter/article/161869/google\\_rolls\\_out\\_semantic\\_search\\_capabilities.html](http://www.pcworld.com/businesscenter/article/161869/google_rolls_out_semantic_search_capabilities.html)
- <sup>xxiii</sup> <http://www.aeiou.pt>
- <sup>xxiv</sup> <http://www.sapo.pt>
- <sup>xxv</sup> <http://www.searchenginecolossus.com/Portugal.html>

<sup>xxvi</sup> <http://www.achei.com.br>

<sup>xxvii</sup> <http://www.gigabusca.com.br>

<sup>xxviii</sup> Ph. Koehn, A. Birch and R. Steinberger. 462 Machine Translation Systems for Europe, Machine Translation Summit XII, p. 65-72, 2009.

<sup>xxix</sup> The higher the score, the better the translation, a human translator would get around 80. 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.

<sup>xxx</sup> <http://cintil.ul.pt>

<sup>xxxi</sup> <http://www.linguateca.pt>