D2.3.gl.en
Language Report for Galician
(English version)

Version 1.0

2011-09-07

METANET4U

www.metanet4u.eu

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

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

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

and by the participating institutions:

Faculty of Sciences, University of Lisbon

Instituto Superior Técnico

University of Manchester

University *Alexandru Ioan Cuza*

Research Institute for Artificial Intelligence, Romanian Academy

University of Malta

Technical University of Catalonia

Universitat Pompeu Fabra

Revision History

Version	Date	Author	Organisation	Description
1.0	07-09-2011	Carmen García-Mateo, Montserrat Arza	UVIGO	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.gl.en Language Report for Galician (English version)

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

Deliverable Number: D2.3.gl.en

Completion: Final
Status: Submitted
Dissemination level: Public

Responsible: Asunción Moreno (WP2 coordinator)

Contributing Partners: University of Vigo; Universitat Politècnica de Catalunya

Authors: Carmen García-Mateo, Montserrat Arza

Reviewer: Asunción Moreno

Table of Contents

Executive Summary

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

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

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

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

This whitepaper for the Galician language demonstrates that a rather limited language technology industry and research environment exist for Galician. Although a number of technologies and resources for Galician exist, there are fewer technologies and resources for the Galician language than for the English language. The technologies and resources also have a poorer quality.

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

META-NET contributes to building a strong, multilingual European digital information space. By realising this goal, a multicultural union of nations can prosper and become a role model for peaceful and egalitarian international cooperation. If this goal cannot be achieved, Europe will have to choose between sacrificing its cultural identities or suffering economic defeat.

A Risk for Our Languages and a Challenge for Language Technology

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

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

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

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

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

- desktop publishing software replaces typewriting and typesetting;
- Microsoft PowerPoint replaces overhead projector transparencies;
- e-mail sends and receives documents often faster than with 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, can they sufficiently implement a sustainable, multilingual European information society, a modern and inclusive society where information and goods can flow freely?

Language Borders Hinder the European Information Society

We cannot precisely know what the future information society will look like. When it comes to discussing a common European energy strategy or foreign policy, we might want to listen to European foreign ministers speak in their native language. We might want a platform where people, who speak many different languages and who have varying language proficiency, can discuss a particular subject while technology automatically gathers their opinions and generates brief summaries. We also might want to speak with a health insurance help desk that is located in a foreign country.

It is clear that communicative 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.[1] A few years ago, English might have been the lingua franca of the web—the vast majority of content on the web was in English. The situation has now changed drastically. 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 lead 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.[2] While popular languages like English

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

[1] 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).

[2] 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).

or Chinese [FIXME: Spanish] will certainly maintain their presence in the emerging digital society and market, many European languages could be cut off by digital communications and become irrelevant for the Internet society. Such developments would certainly be unwelcome. On one hand, a strategic opportunity would be lost, which would weaken Europe's global standing. On the other hand, such developments would conflict with the goal of equal participation for every European citizen regardless of language. According to a UNESCO report on multilingualism, languages are an essential medium for the enjoyment of fundamental rights, such as political expression, education and participation in society.[3]

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.[4] 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, searching for information online or booking a flight. We benefit from language technology when we:
search for and translate web pages,

- use the spelling and grammar checking features in a word processor;
- view product recommendations at an online shop;
- hear the verbal instructions of a synthetic voice in a navigation system;
- translate web pages with an online service.

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

The language technologies detailed in this paper are an essential part of innovative future applications. Language technology is typically an enabling technology within a larger application framework like a navigation system or a search engine. These white papers focus on the readiness of core technologies in each language.
In the near future, we need language technology for all European languages that is available, affordable and tightly integrated within larger software environments. An interactive, multimedia and multilingual user experience is not possible without language technology.

One can think of language technology as the operating system for the content and user interaction.

[3] UNESCO Director-General, *Intersectoral mid-term strategy on languages and multilingualism*, Paris, 2007 (http://unesdoc.unesco.org/images/0015/001503/150335e.pdf).
[4] 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).

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. The current rate of progress creates a genuine window of opportunity with research steadily progressing during the last few years. For example, machine translation (MT) already delivers a reasonable amount of accuracy within specific domains, and experimental applications provide multilingual information and knowledge management as well as content production in many European languages.

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

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

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

cant improvements to be made that can further communication and productivity in our multilingual environment.

Language technologies with broad use, such as the grammar and spell checking 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 its native language via examples. Exposure to concrete, linguistic specimens by language users, such as parents, siblings and other family members, helps babies from the age of about two produce their first words and short phrases. This is only possible because of a special genetic disposition humans have for learning their first language.

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. At school age, foreign languages are usually acquired by learning their grammatical structure, vocabulary and orthography from books and educational materials that describe linguistic knowledge in terms of abstract rules, tables and example texts. Learning a foreign language takes a lot of time and effort, and it gets more difficult with age.

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

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

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

Galician in the European Information Society

General Facts

Galician is part of the Romance family of languages. It's the co-official language in the Spanish region of Galicia. Galicia has over 2,800,000 inhabitants. Approximately, two million persons are speakers of Galician and the rest speaks it as a second language [5,6].

The Galician-speaking territory is delimited by the Autonomous Community of Galicia and the farthest western area of Asturias, León and Zamora, as well as three small areas in Extremadura. Apart from this, and due to the historical circumstances of the Galician emigration around the world, there are some regions with a large concentration of people of Galician origin. This population preserved its language as communication vehicle -not only in the private field but also in the public field- through periodicals, literary publications or even in the radio in host countries. There are still large Galician-speaking communities in other regions of Spain (Madrid, Barcelona, the Basque Country and the Canary Islands), in Europe (Portugal, France, Switzerland, Germany, the United Kingdom and the Netherlands) and in America (Argentina, Uruguay, Brazil, Venezuela, Cuba, Mexico and the United States).

Galician is the historical language of Galicia. The oldest document written in Galician and preserved in Galicia dates back to 1228, from the reign of Alfonso IX, and it is presently in the Archives of the House of Alba in Madrid.

Galicia is -by constitutional recognition- an autonomous community with its own institutions: its Parliament, its own government, security corps, its own public media, flag, etc. The Statute of Autonomy of Galicia -passed in 1981- recognized Galician as the "own" language of Galicia and the co-official language of the Community, which "everyone has the right to know and use", and at the same time it made the authorities responsible for the normalisation of Galician in all fields. The Linguistic Normalisation Act -passed unanimously on June 15th, 1983 by the Galician Parliament- guarantees and regulates citizens' linguistic rights, particularly those regarding the fields of administration, education and the media.
In accordance with the Linguistic Normalisation Act, the local and autonomic administrations are obliged to write all of their official documents in Galician; the use of Galician is established in the whole educational system and the promotion of the language is guaranteed in those countries with emigrant Galician communities as well as in Galician-speaking areas bordering the Community.

In the 1981 Galician Statute of Autonomy, Galician is declared co-official and Galicia's "own" language, and the autonomous institutions are given full competence in the normalisation process

Since the death of Franco, the situation of Galician, especially regarding its legal status and promotion, has remarkably advanced[7]. Nevertheless, all these improvements didn't come with what really matters; a clear growth in the spoken use of the language, and full legal equality with Spanish has not been reached yet.

[5] Information extracted from the "Xunta de Galicia" web page http://www.xunta.es/linguagalega/datos_basicos_da_lingua_galega
[6] Information extracted from the Council of Galician Culture web page http://www.consellodacultura.org/
[7] LOIA project of the Council of Galician Culture: //www.consellodacultura.org/arquivos/cdsg/loia/historia.php?idioma=2&id=76

Language proficiency in Galician[8]

	Comprehension	Speaking	Reading	Writing
2001 Census	99.16%	91.04%	68.65%	57.64%
1991Census	96.96%	91.39%	49.30%	34.85%

Particularities of the Galician Language

Galician is closely related to Portuguese. It is also related to other Romance languages like Spanish or French. Galician uses seven different vowel sounds and nineteen consonant sounds[9]. The Galician alphabet has 23 letters (*a, b, c, d, e, f, g, h, i, l, m, n, ñ, o, p, q, r, s, t, u, v, x, z*) and six digraphs (*ch, gu, ll, nh, qu, rr*). The letters *ç, j, k, w* and *y* are only used in foreign words. The accent mark (´) is used to mark the accented syllable in polysyllabic words and also as a diacritical mark to distinguish between pairs of words that are differentiated in the spoken language because one is stressed where the other is unstressed (dá, verb *dar* / *da*, preposition *de* + article *a*), or because one of them has (a half open vowel) an open-mid vowel while the other has the corresponding close vowel (*vés*, verb v*ir* / *ves*, verb *ver*). In writing, *é* and ó can represent both the open-mid vowels as well as the close vowel.

Concerning the word order of the sentences or utterances in Galician, the main patron used is Subject, Verb, Object. Nevertheless, Galician is almost free and it is not rare to find the use of clitic elements changing the basic structure. The passive voice, which is formed using the auxiliary verb *ser* (to be) and the past participle of the main verb, is not often used in Galician, except in legal, journalistic and scientific documents. Other constructions are used instead to express the idea of passivity: the usual word order is inverted (*Ese libro lino eu cando era pequeno, Esa película rodárona na Coruña*), active verb forms are used with the third person reflexive pronoun (*Esa película rodouse na Coruña*), and there also exists an impersonal construction in which the active verb is formed in the third person singular without an explicit subject, but preceded by the pronoun se (*Véndese viño*).

Yes/no questions are normally formed by reversing the order of the subject and verb (*Veu Antón?* – Has Antón arrived?). If we want to add emphasis, we can add a final interrogative particle (*Veu Antón ou non?).* Negations are usually expressed by placing the adverb non before the verb: *Carme non dixo nada interesante* (Carme didn't say anything interesting). As can be seen, "double negatives" do exist in Galician.

[8] Chart taken from the General Plan for the Normalisation of the Galician Language. Comparative data from the 1991 Census and provisional data from the 2001 Census. Source: Instituto Galego de Estatística (Galician Institute of Statistics)

[9] http://www.consellodacultura.org/arquivos/cdsg/loia/gramatica.php?idioma=2&seccion=6

Galician is a pro-drop language; it is possible to use the conjugation of the verb without the personal pronoun involved that plays the subject role.

The orthography in Galician is more transparent than in English, but less than in Spanish or Italian. For example, vowels *e* and *o*, can be pronounced different in some dialects.

The three main dialectal areas are: (1) eastern Galician, which includes the dialects spoken outside the Galician administrative area, the most important of which is the Galician spoken in Asturias; (2) central Galician, among which the Mondoñedo and Lugo-Ourense varieties stand out; and (3) western Galician, where the dialects of the Fisterra region in the north and of Tui and Baixa Limia in the south stand out.

The main dialectal features are:

(1) phonetic features: *gheada* (there exists a fricative phoneme or approximant, either voiceless or voiced, in place of the voiced velar occlusive /g/, in words such as *gato* [cat] and *pagar* [to pay]), is characteristic of western Galician and a large part of central Galician; *seseo* (absence of /θ/ and the presence of /s/ in the positions where /θ/ occurs in common Galician, in words such as *cen* [hundred] and *cazar* [hunt]), is characteristic of western Galician;

(2) morphological features: in nouns, the ending -án (<Latin -ANU & -ANA: *irmán* <Latín GERMANU, GERMANA) in western dialects, as against the ending -ao (<Latín -ANU) and -á (<Latin -ANA) (*irmao/irmá* [brother/sister]) in the dialects of the central and eastern areas; the formation of the plural of nouns ending in -n, the ending -óns (<Latín -ONES) in the western areas, as against the ending -ós in the central area and -ois in the eastern areas; in verbs, the personal suffix -is for the second person plural (*andais*) in the eastern dialects, as against the suffix -des in common Galician (*andades*). The eastern dialects (especially Galician spoken in Asturias) also have many other peculiarities.

Recent developments

The Media and Cultural Industries

There are not any newspapers available in Galician. In some newspapers, Galician is not completely absent, although it is relegated to cultural information and the opinions columns. In the non-daily press, both a weekly newspaper offering general information (*A Nosa Terra*), which has been coming out regularly for more than twenty years, and a monthly magazine offering information and debate (*Tempos Novos*), stand out. Although they are of a more restricted diffusion, we highlight the tri-monthly magazine *Grial* (of a long standing tradition), *Encrucillada*, *A Trabe de Ouro* and *Agália*.

With regard to television, in 1985 the publicly owned *Compañía de Radio-Televisión de Galicia* was created, and from then on Galician television began to broadcast, basically in Galician, with a noticeable audience and some outstanding successes.

Regarding radio stations, it is undoubtedly the *Radio Galega* that shows the greatest commitment with the use and promotion of Galician.

The publication of books in Galician increased dramatically during this period, rising from 187 titles in 1980, to 1,826 in 2005. However, some problems should be pointed out, such as the overwhelming presence of institutional publications, an excessive amount of small Galician publishing houses and the dangerous dependency on the school market.

As regards musical production, the renewed fashion for "roots" music must be highlighted; In our case, this means music inspired (more or less vaguely) in popular-traditional music, or simply "Celtic"; and also the adaptation, from the Galician perspective, of popular contemporary international music, that is to say, pop-rock.

A number of products and services has been developed in the last years aiming at incorporating the Galician to the ICT society. Operating systems, grammar checkers and phone applications are some examples[10].

Language cultivation

The Royal Galician Academy[11] (Galician*: Real Academia Galega*, RAG) is an institution, dedicated to the study of Galician culture and especially the Galician language; it promulgates norms of grammar, spelling, and vocabulary and works to promote the language.

The "Consello da Cultura Galega"[12] is a legal institution dedicated to the promotion and preservation of the Galician culture. Galician language promotion is one of its aims. Its LOIA project[13], carried out through its Language Section and the Sociolinguistics Documentation Centre of Galicia, aims at spreading the basic elements needed to know the Galician language, its history, its cultural production and its social situation and prospects for the future, in a wide and concise manner. Most of the material reflected in this chapter has been extracted from the LOIA web site.

Language in Education

The 1981 Statute of Galician recognised Galician as their own and official language, as well as Spanish. The introduction of Galician language in the education took place in 1979. The development of the Linguistic

[10] http://www.xunta.es/linguagalega/o_galego_nas_novas_tecnoloxias
[11] http://www.realacademiagalega.org/
[12] http://consellodacultura.org/
[13] http://www.consellodacultura.org/arquivos/cdsg/loia/index.php?idioma=2

Standardisation Act aims that students shall have the same writing and oral skills in Galician and Spanish.

In Galicia, children have the right to receive primary education in their mother tongue, and the educational authorities are obliged to provide the "means necessary to promote the progressive use of Galician in education", establishing as the minimum aim the "on finishing the two cycles in which Galician is obligatory, students should know this language, on both an oral and written level, to the same extent as Castilian".

Since the early eighties, an intense task of the linguistic readjustment of the Galician primary and secondary school teachers was undertaken; this was achieved through intense courses of Galician literature and language, and they were attended by a large part of the practising teachers over the course of the decade. From the early nineties, provision for the creation of teams of linguistic normalisation were adopted, and plans of normalisation in educative centres were developed; aids to encourage activities that promoted Galician were also established.

In general, it can be said that at present, the different initiatives have been centred around what are considered the two main goals in the area: to convert Galician into the instrument (vehicle language) of the education system; and to ensure that students obtain full linguistic competence in both official languages (Galician and Castilian) by the end of obligatory education. Nevertheless, despite the unquestionable achievements –unequal depending on the educational level- there is still a long way to go before these goals are fully met.

International aspects

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

As a minority language, Galician was represented in the European Bureau for Lesser Used Languages, which was set up in 1982 on the initiative of the European Parliament. The aim of this pan-European non-governmental organisation has been to encourage respect towards lesser protected languages within the EU and to promote linguistic diversity.

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

In 2005 the Council of Ministers recognised the possibility of using official languages other than Spanish in the European institutions. After signing administrative agreements with some EU institutions, recognising a restricted limited use of Galician, the status of Galician is currently that of a semi-official language, a language of communication with the citizens. This status means that citizens can write in Galician to these institutions (European Commission, European Parliament, Council, European Ombudsman and Committee of the Regions), and, in turn, they have the right to be answered in the same language. Some publications and official documentation is translated into Galician, as well.

The international projection of Galician is quite limited. In the business world at international level, the use of Galician is non-existent. In fact, English has become the common language of communication on written and oral level. Currently, from the customers' point of view, a few big international companies are using Galician to deal with their Galician customers, as an added value to their products and as an improvement of their customer services. Some of these companies are Microsoft, or Telefónica.

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

As regards learning Galician as a foreign language, the situation is a little better. The European Commission is developing an active policy on multilingualism, which aims at preserving and promoting linguistic diversity in Europe, fostering language learning (including regional and minority languages) and using multilingualism as a stimulus for competitiveness. In this context, the Lifelong Learning Programme 2007-13 contains a selection of projects promoting language learning. Among them, the Lingu@net Europa Plus multilingual online languages resource centre[14] provides support and resources in 20 European languages, including Galician. In addition, an important decision made by representatives of the EU Member States has been to include Galician, as well as Basque and Catalan in the list of languages offered in the Erasmus Intensive Language Courses from the academic year 2010-2011[15]. These EU-funded language courses aim to prepare prospective Erasmus students for their study period in Galician universities, where this language is used as a communication and academic language.

The services of linguistic normalization of the three Galician universities, as well as those of some town councils, organise on a regular basis Galician language courses. During summer, there is also the possibility of attending the *Cursos de Verán de Lingua e Cultura Galegas para Estranxeiros e para Españois de Fóra de Galicia* (Summer Courses in

[14] http://www.linguanet-worldwide.org/lnetww/gl/home.jsp
[15] http://ec.europa.eu/education/news/news1518_en.htm

Galician Language and Culture for Foreigners and Spaniards from Outside of Galicia).

The General Secretariat for Language Policy ("Secretaria Xeral de Política Lingüística") has various agreements of collaboration with different universities outside Galicia, with the aim of creating chairs and posts for language assistants that promote and spread Galician in the international sphere. Currently, there are forty-seven centres of Galician studies located at several universities in Europe, America and the Australian Continent.

Thanks to the development of new technologies, it's possible to come closer to the learning of the Galician language using new tools available on the web, as interactive on-line courses: *é-galego, A Palabra Herdada, Galingua.*

Galician on the Internet

The presence of Galician on the Internet is rather limited (after all, Galician occupies the position 160 in the Ethnologue[16] classification of languages by language size). Nevertheless, there are some initiatives that try to increase the presence of Galician on the web. Galipedia[17] (the Galician Wikipedia) with around 75.000 articles is in the same group as some EU official languages like Greek or Latvian. Another example is the PuntoGal[18] initiative that is trying to obtain a domain on the Internet for the Galician language and culture. Through this domain, Galician society would have more visibility on the net and throughout the world. Google or Facebook, among others, offer a Galician version for their navigation interfaces.

The Regional Government has launched some initiatives to support the creation of webs in Galician. Additionally, the *Mancomun*[19] web offers a number of open-software tools in Galician developed with the Regional Government support. For example *Galinux*[20] is a GNU/Linux distribution in Galician designed for educational purposes.

The web also offers a growing number of digital newspapers in Galician (or Spanish newspapers with a plug-in tool for translation into Galician), as well as some online courses to learn the language.

[16] http://www.ethnologue.com/ethno_docs/distribution.asp?by=size
[17] http://gl.wikipedia.org/wiki/Portada
[18] http://www.puntogal.org/
[19] http://www.mancomun.org
[20] http://www.galinux.org/

Language Technology Support for Galician

Language Technologies

Language technologies are information technologies that are specialised for dealing with human language. Therefore these technologies are also often subsumed under the term Human language technology. Human language occurs in spoken and written form. While speech is the oldest and most natural mode of language communication, complex information and the bulk of human knowledge is recorded and transmitted in written texts. Speech and text technologies process or produce language in these two forms. But language also has aspects common to both forms such as dictionaries, most of the grammar, and the meaning of sentences. Thus, large parts of language technology cannot be subsumed under either speech or text technologies. Knowledge technologies include technologies that link language to knowledge. Figure 1 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. Texts can be combined with pictures and sounds. Movies may contain language in spoken and written form. Thus, speech and text technologies overlap and interact with many other technologies that facilitate the processing of multimodal communication and multimedia documents.

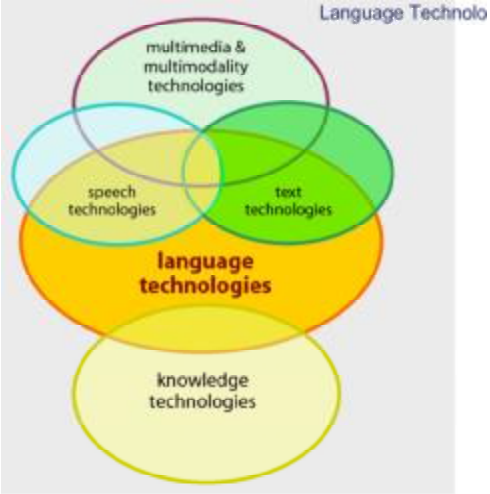

Figure 1: The language technology Landscape

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. Figure 2 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 a given context?), resolving anaphora and referring expressions like *she*, *the car*, etc.; representing the meaning of the sentence in a machine-readable way

Task-specific modules then perform many different operations such as automatic summarisation of an input text, database look-ups and many others. Below, we will illustrate core application areas and highlight their core modules. Again, the architectures of the applications are highly simplified and idealised, to illustrate the complexity of language technology (LT) applications in a generally understandable way.

After introducing the core application areas, we will give a short overview of the situation in LT research and education, concluding with an overview of past and ongoing research programs. At the end of this section, we will present an expert estimation on the situation regarding core LT tools and resources on a number of dimensions such as avai-

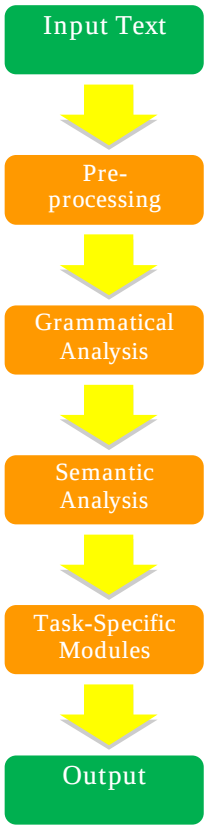

Figure 2: A Typical Text Processing Application Architecture

lability, maturity, or quality. This table gives a good overview on the situation of LT for Galician.

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 for Galician.

Core application areas

Language Checking

Anyone using a word processing tool such as Microsoft Word has come across a **spell checking** component that indicates spelling mistakes and proposes corrections. Forty years after the first spelling correction program by Ralph Gorin, language checkers nowadays do not simply compare the list of extracted words against a dictionary of correctly spelled words, but have become increasingly sophisticated. In addition to language-dependent algorithms for handling morphology (e.g. plural formation), some are now capable of recognising syntax–related errors, such as a missing verb or a verb that does not agree with its subject in person and number, e.g. in "She **write* a letter." However, for other common error types, the above described methods are not sufficient. For example, take a look at 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.*

Most available spell checkers (including Microsoft Word) will find no errors in this poem because they mostly look at words in isolation. However, for detecting so-called homophone errors (e.g. "Eye" instead of "I"), the language checker needs to consider the context in which a word occurs. For Galician, even spell checking requires analysing the context in many cases. A typical case is when the orthographic error transforms one word into another, which also exits. In the following example, the first sentence contains a frequent error (problems with orthographic accents). The second sentence is the corrected version of the first one.

*A casa do meu tío **e** a casa da miña avoa.*
[The house of my uncle and the house of my grandmother]

*A casa do meu tío **é** a casa da miña avoa.*
[The house of my uncle is the house of my grandmother]

To automatically correct these errors, it is not enough to check each word in a dictionary, since all words in the first sentence are correct in isolation. 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, "é a" is a much more

probable word sequence than “e a”. A statistical language model can be derived automatically 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 well to other languages, e.g. highly inflectional ones or languages with a flexible word order like Galician. For these more complex languages, an advanced high-precision language checker may require the development of more sophisticated methods, involving a deeper linguistic analysis.

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

Only few companies and Language Service Providers offer products in this area for Galician. *Imaxin software*[21] is one example with some online free-to-use services for translation and grammar checking. *OrtoGal* software from Computational Linguistics Group (SLI)[22] of the University of Vigo offers spell and grammar checking. There is also plug-in software for OpenOffice like *Golfiño*[23] developed by *Imaxin Software* and supported by the Galician Regional Government.

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. 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 also for Galician and, in 2009, they incorporated basic semantic search capabilities into their algorithmic mix, 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 in-

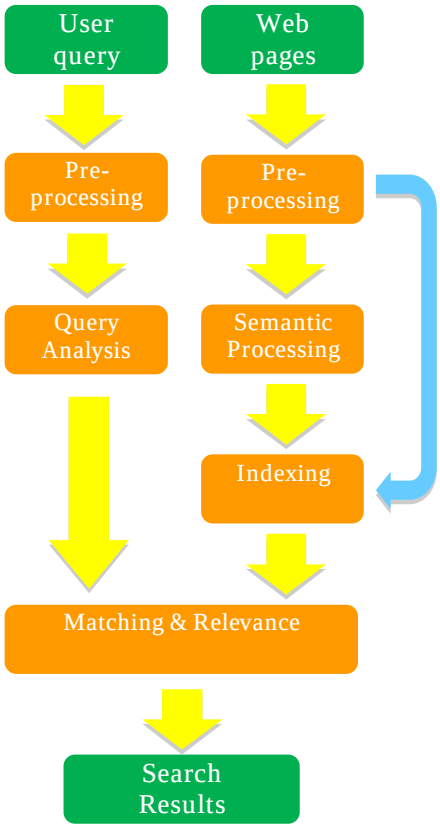

Figure 3: Web Search Architecture

[21] http://www.imaxin.com/
[22] http://webs.uvigo.es/sli/index_en.html
[23] http://www.mancomun.org/descargarprogramas/detalledeproduto/nova/golfino/

dexing these data, a mainly statistically-based approach can lead to satisfactory results.

However, for a more sophisticated request for information, integrating deeper linguistic knowledge is essential. In the research labs, experiments using machine-readable thesauri and ontological language resources like WordNet have shown improvements by allowing the possibility of finding a page on the basis of synonyms of the search terms, or even more loosely related terms. Again, these developments require of language specific resources. A Galician WordNet has been developed by the research centre "Centro Ramón Piñeiro para la Investigación en Humanidades"[24]. The Galician WordNet is called GALWORDNET.

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

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

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

To the best of our knowledge there is no linguistic technology at companies aimed at multilingual search and information retrieval, both from the Internet and from internal information systems on Galician.

[24] http://www.cirp.es

Speech Interaction

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

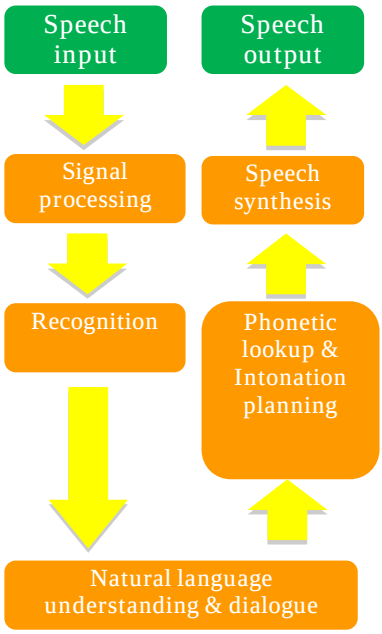

Figure 4: A Simple Speech-based Dialogue Architecture

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

- Automatic speech recognition (ASR) is responsible for determining which words were actually spoken given a sequence of sounds uttered by a user.
- Syntactic analysis and semantic interpretation deal with analysing the syntactic structure of a user's utterance and 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 increase the costs significantly. However, VUIs that employ language models and initially allow a user to flexibly express their intent – evoked, e.g., by a 'How may I help you' greeting – show both a higher automation rate and a higher user acceptance and may therefore be considered as advantageous over a less flexible directed dialogue approach.

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

trast, today's TTS systems prove superior, though optimisable, regarding the prosodic naturalness of dynamic utterances.

Regarding the market for speech interaction technology, the last decade has been characterised by a strong standardisation of the interfaces between the different technology components, as well as by standards for creating particular software artefacts for a given application. There also has been strong market consolidation within the last ten years, particularly in the field of ASR and TTS. Here, the national markets in the G20 countries – i.e. economically strong countries with a considerable population - are dominated by less than 5 players worldwide, with *Nuance* and *Loquendo* being the most prominent ones in Europe, also for Galician (Loquendo), although some smaller local companies are starting to compete, such as Verbio[25], which is a spin-off of Universitat Politècnica de Catalunya and has its own speech technology, or the Galician *2Mares*[26].

Regarding dialogue management technology and know-how, markets are strongly dominated by national players, which are usually SMEs. Most of the companies on the Spanish TTS market (some offer Galician) are essentially application developers. Key players in the Spanish market are: Indsys[27] (Intelligent Dialogue Systems), Fonetic[28], Ydilo[29], NaturalVoz[30], and 2Mares.

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

Machine Translation

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

[25] http://www.verbio.com/
[26] http://www.2mares.com/
[27] http://www.indisys.es/default.aspx
[28] http://www.fonetic.es/
[29] http://www.ydilo.com/esp/index.php
[30] http://www.naturalvox.com/

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

> O policía observou ao home co telescopio.
> [The policeman observed the man with the telescope.]
>
> O policía observou ao home co 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.

Leading international MT developer Lucy Software has an important subsidiary in Spain, Lucy Iberica[31], former Translendium. Lucy Iberica is responsible for the development of language pairs that include Spanish and all language pairs involving any other Iberian language (Catalan, Portuguese, Galician and Basque). Lucy system is grammar rule-based. The Regional Government ("Xunta de Galicia")[32] offers a translation service on the Internet that uses the technology of the Lucy Iberica. While there is significant research in data-driven and hybrid systems in national and international contexts, this technology has been less successful in business than in research so far.

Apertium is a free open-source machine translation platform that provides a language-independent machine translation engine initially designed by the Transducens group at the Universitat d'Alacant and subsequently developed in the framework of the nationally funded Opentrad project. Among current MT systems using Apertium technology, we find interNOSTRUM (Spanish-Catalan), Traductor Universia (Spanish-Portuguese) and Matxin (Basque-Spanish), the former developed by Transducens and the latter by the IXA group[33] at Euskal Herriko Unibertsitatea, Imaxin Software (Galician-Spanish). It is possible to use Apertium to build machine translation systems for a variety of language pairs (there are over 20 to date); to that end, Apertium uses simple XML-based standard formats to encode the linguistic data needed (either by hand or by converting existing data), which are compiled using the provided tools into the high-speed formats used by the engine.

Provided good adaptation in terms of user-specific terminology and workflow integration, there is a wide consensus that the use of MT can increase productivity significantly. The quality of MT systems is still considered to have huge improvement potential. Challenges include the adaptability of the language resources to a given subject domain or user area and the integration into existing workflows with term bases and translation memories. In addition, many language pairs are still missing.

Language Technology

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

[31] http://www.lucysoftware.com/
[32] http://www.xunta.es/tradutor/
[33] http://ixa.si.ehu.es/Ixa

the system providing a single answer: 'At what age did Neil Armstrong step on the moon?' - '38'. While this is obviously related to the aforementioned core area Web Search, question answering nowadays is primarily an umbrella term for research questions such as what *types* of questions should be distinguished and how should they be handled, how can a set of documents that potentially contain the answer be analysed and compared (do they give conflicting answers?), and how can specific information - the answer - be reliably extracted from a document, without unduly ignoring the context.

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

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

For Galician, the situation in these research areas is much less developed than it is for English, where, since the 1990s, question answerring, information extraction, and summarization have been the subject of numerous open 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 Galician was never a

targeted language. Accordingly, there are hardly any annotated corpora or other resources for these tasks. Summarization systems, when using purely statistical methods, are often to a good extent language-independent, and thus some research prototypes are available. For text generation, reusable components have traditionally been limited to the surface realization modules (the "generation grammars"); again, most available software is for English.

Apart from the experimental systems being developed by the research groups, there are no SMEs offering this kind of services. Since 2000 up till today, the Spanish Government supported within the National Plan of Research and Technology several projects in the area of Multilingual Speech Technologies: TEHAM, AVIVAVOZ, and BUCEADOR. Their main purpose was to improve the quality of Speech Recognition, Speech Translation and Text to Speech Synthesis in all the official languages spoken in Spain: Basque, Galician, Catalan and Spanish.

Language Technology in Education

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

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

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

el Procesamiento del Lenguaje Natural (SEPLN, Spanish Society for Natural Language Processing), is a non-profit organisation with over 300 members, both from academia and industry, which was created in 1984 with the purpose to promote and spread activities related to teaching, research and development of NLP, on both national and international level. SEPLN organizes seminaries, symposiums and conferences and promotes collaboration with national and international institutions.

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

The Spanish Network on Speech Technology (RTTH)[34] is a common forum where researchers (presently more than 25o researchers) in Speech Technology combine efforts and share experiences in order to:

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

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

Language Technology Programs

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

The Centre for the Development of Industrial Technology (CDTI) is a Spanish public organisation, under the Ministry of Science and Innovation, whose objective is to help Spanish companies increase their tech-

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

nological profile. CDTI evaluates and finances R&D projects through programmes such as CENIT and AVANZA.

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

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

The Galician Regional Government supports research through the "Plan Galego de Investigación, Desenvolvemento e Innovación Tecnolóxica (PGIDIT)". Language Technology is not a priority line, but along the years research groups from the universities and some companies have gotten grants for doing research and developments in LT.

Availability of Tools and Resources for Galician

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

1. **Quantity**: Does a tool/resource exist for the language at hand? The more tools/resources exist, the higher the rating.
    - 0: no tools/resources whatsoever
    - 6: many tools/resources, large variety
2. **Availability**: Are tools/resources accessible, i.e. are they Open Source, freely usable on any platform or only available for a high price or under very restricted conditions?

- 0: practically all tools/resources are only available for a high price
- 6: a large amount of tools/resources is freely, openly available under sensible Open Source or Creative Commons licenses that allow re-use and re-purposing

3. **Quality**: How well are the respective performance criteria of tools and quality indicators of resources met by the best available tools, applications or resources? Are these tools/resources current and also actively maintained?
   - 0: toy resource/tool
   - 6: high-quality tool, human-quality annotations in a resource
4. **Coverage**: To which degree do the best tools meet the respective coverage criteria (styles, genres, text sorts, linguistic phenomena, types of input/output, number languages supported by an MT system etc.)? To which degree are resources representative of the targeted language or sublanguages?
   - 0: special-purpose resource or tool, specific case, very small coverage, only to be used for very specific, non-general use cases
   - 6: very broad coverage resource, very robust tool, widely applicable, many languages supported
5. **Maturity**: Can the tool/resource be considered mature, stable, ready for the market? Can the best available tools/resources be used out-of-the-box or do they have to be adapted? Is the performance of such a technology adequate and ready for production use or is it only a prototype that cannot be used for production systems? An indicator may be whether resources/tools are accepted by the community and successfully used in LT systems.
   - 0: preliminary prototype, toy system, proof-of-concept, example resource exercise
   - 6: immediately integratable/applicable component
6. **Sustainability**: How well can the tool/resource be maintained/integrated into current IT systems? Does the tool/resource fulfil a certain level of sustainability concerning documentation/manuals, explanation of use cases, front-ends, GUIs etc.? Does it use/employ standard/best-practice programming environments (such as Java EE)? Do industry/research standards/quasi-standards exist and if so, is the tool/resource compliant (data formats etc.)?
   - 0: completely proprietary, ad hoc data formats and APIs
   - 6: full standard-compliance, fully documented
7. **Adaptability**: How well can the best tools or resources be adapted/extended to new tasks/domains/genres/text types/use cases etc.?
   - 0: practically impossible to adapt a tool/resource to another task, impossible even with large amounts of resources or person months at hand
   - 6: very high level of adaptability; adaptation also very easy and efficiently possible

Table of Tools and Resources

	Quantity	Availability	Quality	Coverage	Maturity	Sustainability	Adaptability
Language Technology (Tools, Technologies, Applications)							
Tokenization, Morphology (tokenization, POS tagging, morphological analysis/generation)	4	5	4	5	4	4	4
Parsing (shallow or deep syntactic analysis)	4	5	5	4	3	4	4
Sentence Semantics (WSD, argument structure, semantic roles)	2	1	3	2	2	1	2
Text Semantics (co-reference resolution, context, pragmatics, inference)	1	1	3	2	2	2	1
Advanced Discourse Processing (text structure, coherence, rhetorical structure/RST, argumentative zoning, argumentation, text patterns, text types etc.)							
Information Retrieval(text indexing, multimedia IR, crosslingual IR)	2	1	2	2	1	2	1
Information Extraction (named entity recognition, event/relation extraction, opinion/sentiment recognition, text mining/analytics)	3	1	3	1	2	1	1
Language Generation (sentence generation, report generation, text generation)							
Summarization, Question Answering, advanced Information Access Technologies	2	1	1	2	1	1	1
Machine Translation	3	5	4	5	5	4	4
Speech Recognition	3	2	5	5	5	5	5
Speech Synthesis	4	3	5	5	5	5	4
Dialogue Management (dialogue capabilities and user modelling)	1	0	1	1	0	0	0
Language Resources (Resources, Data, Knowledge Bases)							
Reference Corpora	5	4	5	5	5	5	4
Syntax-Corpora (treebanks, dependency banks)	1	1	2	2	2	2	1
Semantics-Corpora	1	1	1	1	1	1	1
Discourse-Corpora							
Parallel Corpora, Translation Memories	3	5	5	5	5	5	5
Speech-Corpora (raw speech data, labelled/annotated speech data, speech dialogue data)	3	4	4	5	4	4	5
Multimedia and multimodal data (text data combined with audio/video)	2	1	4	2	3	3	3
Language Models	2	1	5	4	4	4	4
Lexicons, Terminologies	5	4	5	4	5	4	4
Grammars	2	2	2	2	2	2	2
Thesauri, WordNets	4	5	4	4	3	3	3
Ontological Resources for World Knowledge (e.g. upper models, Linked Data)	2	1	1	1	1	1	1

Conclusions

The situation of Galician concerning language technology support gives rise to cautious optimism. Supported by some research projects in the past, an emerging language technology industry and research scene exists in Spain that develops products and services for Galician. The industry consists of SMEs, most of which originally were spin-offs of a project or a research group.

For Galician, a number of technologies and resources exist, but far less than for English. Still, even for English and major languages, language technology support today is by far not in a state that is needed for offering the support a true multilingual knowledge society needs.

In this Whitepaper Series, a 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 Galician, key results regarding technologies and resources include the following:

- Speech processing currently seems to be more mature than processing of written text. Advanced information access technologies are in their infancies and for Galician in particular, almost non-existent.
- 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 needed.
- Research was successful in designing particular high quality software, but many of the resources lack standardization, i.e., even if they exist, sustainability is not always given; concerted programs and initiatives are needed to standardize data and interchange formats.
- For Galician, a large reference text corpus (with a balanced mixture of various genres) exists, as well as other specialised corpora, but they are not easily/cheaply accessible.
- While some specific corpora of high quality exist, a very large syntactically annotated corpus is not available.
- There are very few annotated corpora with syntactic, semantic, or discourse information; again, the situation is worse the more deep linguistic and semantic information is needed.
- Speech processing is currently more mature than NLP for written text.
- Parallel corpora exist between Galician and Spanish and they have been used to develop machine translation systems. However, parallel corpora between Galician and other languages are missing.

- Multimedia data is a huge gap.

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

META-NET

META-NET is a Network of Excellence funded by the European Union. It currently consists of 44 members, representing 31 European countries, which are listed below. META-NET is fostering the Multilingual Europe Technology Alliance (META), a growing community of language technology professionals and organisations in Europe.

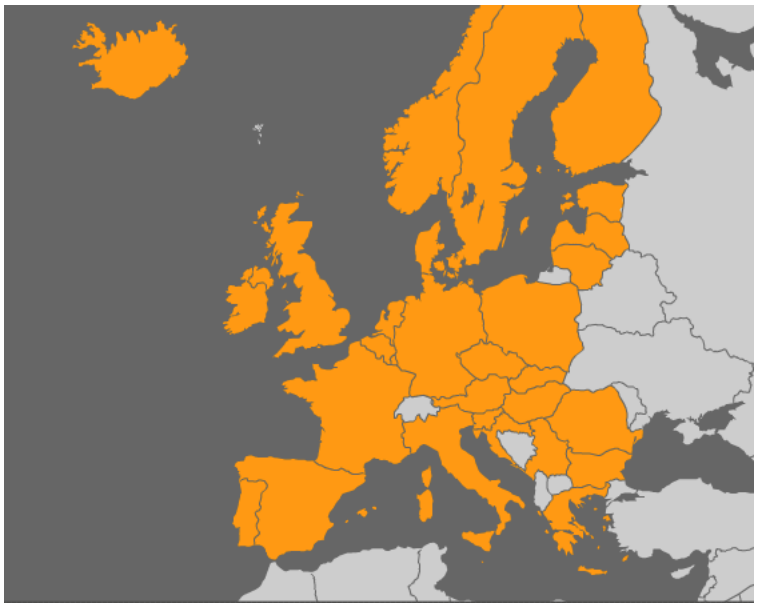

Figure 5: Countries Represented in META-NET

META – The Multilingual Europe Technology Alliance

META-NET cooperates with a dozen other large initiatives like CLARIN, which is helping social sciences establish the Digital Humanities field in Europe. META-NET is dedicated to fostering the technological foundations for establishing and maintaining a truly multilingual European information society that

- makes possible communication and cooperation across languages,
- safeguards equal access to information and knowledge for users of any language,
- offers advanced functionalities of networked information technology to all citizens at affordable costs.

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.

META-NET's Three Lines of Action

META-NET was launched on 1 February 2010 with the goal of advancing research in language technology. The initiative supports a Europe that unites as a single, digital market and information space. META-NET has conducted several activities that further its goals. META-VISION, META-SHARE and META-RESEARCH are the network's three lines of action.

Figure 6: 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 META-NET's first year, presentations at the FLaReNet Forum (Spain), language technology Days (Luxembourg), JIAMCATT 2010 (Luxembourg), LREC 2010 (Malta), EAMT 2010 (France) and ICT 2010 (Belgium) centred on public outreach. According to initial estimates, META-NET has already contacted more than 2,500 LT professionals to share its goals and visions with them. At the META-FORUM 2010 event in Brussels, META-NET shared 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 collecting data, preparing data sets and organising language resources for evaluation purposes; compiling inventories of tools and methods; and organising workshops and training events for members of the community. 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.

Composition of the META-NET Network of Excellence

Country	Member (Affiliation)	Contacts
Austria	Universität Wien	Gerhard Budin
Belgium	University of Antwerp	Walter Daelemans
	University of Leuven	Dirk van Compernolle
Bulgaria	Bulgarian Academy of Sciences	Svetla Koeva
Croatia	Zagreb University	Marko Tadic
Cyprus	University of Cyprus	Jack Burston
Czech Rep.	Charles University in Prague*	Jan Hajic
Denmark	University of Copenhagen	Bente Maegaard, Bolette Sandford Pedersen
Estonia	University of Tartu	Tiit Roosmaa
Finland	Aalto University*	Timo Honkela
	University of Helsinki	Kimmo Koskenniemi, Krister Linden
France	CNRS, LIMSI*	Joseph Mariani
	ELDA*	Khalid Choukri
Germany	DFKI*	Hans Uszkoreit, Georg Rehm
	RWTH Aachen*	Hermann Ney
Greece	ILSP, R.C. “Athena”*	Stelios Piperidis
Hungary	Hungarian Academy of Sciences	Tamás Váradi
	Budapest Technical University	Géza Németh, Gábor Olaszy
Iceland	University of Iceland	Eirikur Rögnvaldsson
Ireland	Dublin City University*	Josef van Genabith
Italy	Consiglio Nazionale Ricerche*	Nicoletta Calzolari
	Fondazione Bruno Kessler*	Bernardo Magnini
Latvia	Tilde	Andrejs Vasiljevs
	University of Latvia	Inguna Skadina
Lithuania	Institute of the Lithuanian Language	Jolanta Zabarskaitë
Luxembourg	Arax Ltd.	Vartkes Goetcherian
Malta	University of Malta	Mike Rosner
Netherlands	Universiteit Utrecht*	Jan Odijk
Norway	University of Bergen	Koenraad De Smedt
Poland	Polish Academy of Sciences	Adam Przepiórkowski
	University of Łódź	Piotr Pezik
Portugal	University of Lisbon	Antonio Branco

	Inst. for Systems Engineering and Computers	Isabel Trancoso
Romania	Romanian Academy of Sciences	Dan Tufis
	University Alexandru Ioan Cuza	Dan Cristea
Serbia	Belgrade University	Dusko Vitas, Cvetana Krstev, Ivan Obradovic
Slovakia	Slovak Academy of Sciences	Radovan Garabik
Slovenia	Jozef Stefan Institute*	Marko Grobelnik
Spain	Barcelona Media*	Toni Badia
	Technical University of Catalonia	Asunción Moreno
	University Pompeu Fabra	Núria Bel
Sweden	University of Gothenburg	Lars Borin
UK	University of Manchester	Sophia Ananiadou

An * represents the founding members.

How to Participate?

META-NET and META offer many opportunities for participation. Please check out www.meta-net.eu for information on upcoming events and activities.

	Pupin Institute	SanjaVranes