META-NET White Paper Series

Languages in the European Information Society

– Maltese –

Early Release Edition

META-FORUM 2011

27-28 June 2011

Budapest, Hungary

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

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

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

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

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

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

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

Authors

Michael Rosner, University of Malta
Jan Joachimsen, University of Malta

Acknowledgements

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

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?

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

A Risk for Our Languages and a Challenge for Language Technology

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

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

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

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

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

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

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

Language Borders Hinder the European Information Society

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

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

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

Today, we can transmit gigabytes of text around the world in a few seconds before we recognize that it is in a language we do not understand. According to a recent report requested by the European Commission, 57% of Internet users in Europe purchase goods and services in languages that are not their native language. (English is the most common foreign language followed by French, German and Spanish.) 55% of users read content in a foreign language while only 35% use another language to write e-mails or post comments on the web.[1] A few years ago, English might have been the lingua franca of the web—the vast majority of content on the web was in English—but the situation has now drastically changed. The amount of online content in other languages (particularly Asian and Arabic languages) has exploded.

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

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

Our Languages at Risk

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

The approximately 60 languages of Europe are one of its richest and most important cultural assets. Europe's multitude of languages is also a vital part of its social success.[2] While popular languages like English or Spanish will certainly maintain their presence in the emerging digital society and market, many European languages could be cut off from digital communications and become irrelevant for the Internet society. Such developments would certainly be unwelcome. On the one hand, a strategic opportunity would be lost that would weaken Europe's global standing. On the other hand, such developments would conflict with the goal of equal participation for every European citizen regardless of language. According to a UNESCO report on multilingualism, languages are an essential medium for the enjoyment of fundamental rights, such as political expression, education and participation in society.[3]

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

Language Technology is a Key Enabling Technology

In the past, investment efforts have focused on language education and translation. For example, according to some estimates, the European market for translation, interpretation, software localisation and website globalisation was € 8.4 billion in 2008 and was expected to grow by 10% per annum.[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, conducting an online search or booking a flight. We benefit from language technology when we:

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

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

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

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

Opportunities for Language Technology

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

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

There are huge market opportunities in the education and entertainment industries for the integration of language technologies in games, edutainment offerings, simulation environments or training programmes. Mobile information services, computer-assisted language learning software, eLearning environments, self-assessment tools and plagiarism detection software are just a few more examples where language technology can play an important role. The popularity of social media applications like Twitter and Facebook suggest a further need for sophisticated language technologies that can monitor posts, summarise discussions, suggest opinion trends, detect emotional responses, identify copyright infringements or track misuse.

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

Multilingualism is the rule, not an exception.

Challenges Facing Language Technology

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

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

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

Language Acquisition

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

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

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

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

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

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

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

Maltese in the European Information Society[5]

General Facts

Maltese is the national language of the Maltese archipelago, which consists of the islands Malta, Gozo (*Għawdex*) and Comino (*Kemmuna*).

Together with English, Maltese is also the official language of Malta. According to the *Demographic Review 2009* by the National Statistics Office of Malta[6], the estimated Maltese population (excluding foreigners) in Malta for the end of the year 2009 was 396,278. It is estimated that today, due to emigration phases from Malta mostly in the 1950s and 1960s, roughly the same number of expatriate native speakers lives abroad (mostly in the United Kingdom, Australia, USA and Canada).

Although Maltese belongs to the South Arabic branch of the Semitic language family, it differs considerably from the other neo-Arabic languages. Its structure is the result of different language contact situations that emerged under different rulers of the islands in the course of a millenium. While the core of Maltese is Semitic, it also contains a Romance superstrate and English adstrate. Also, Maltese is the only Semitic language written in a (modified) Latin alphabet.

The Semitic core of the Maltese language stems from the Arab conquest in 870 AD and its subsequent repopulation with Arabic speaking settlers. In 1090, Malta was conquered by the Normans, who installed Sicilian as the official language, while the population still used their Arabic vernacular in everyday life. Malta was more and more cut off politically, culturally and linguistically from the Arabic world. In the following centuries, under the influence of the official Romance languages of the rulers, more and more Romance loan words entered the Arabic dialect. When Malta was under British rule in 1800, the official language changed from Italian to English, which brought an increasing number of English loan words into Maltese. The following sentence taken from a newspaper article[7] can illustrate the different influences of the languages in contact (Romance loan words are in boldface, English loans underlined):

> *Il-ħold-up sar minn żażugħ li kien liebes* ***nuċċali skur*** *tax-xemx.*
>
> *[The robbery happened by a young man who was wearing dark sunglasses]*

One remarkable fact about Maltese is that despite its relatively small number of speakers and the small area in which it is spoken, there is a comparatively rich number of variants or dialects. In general, a main distinction can be made between the Standard variety spoken in the urban areas like Valletta and Sliema and non-standard varieties spoken in the rural areas. Outside of Malta, the Maltese spoken in Australia has developed into an ethnolect of its own called *Maltraljan*. It differs from Standard Maltese mainly in terms of its lexicon (i.e., the vocabulary) that are the result of extensively borrowing words from (Australian) English and subsequent change in meaning.

With English being the second official language in Malta, many Maltese are bilingual. Between the poles of monolingualism and full bilingualism, there is a continuum of language-mixing and codeswitching. Most Maltese speak only Maltese at home and among each other. English, on the other hand, is the language used in the written context of higher education and in communication with foreigners.

Particularities of the Maltese Language

Maltese is the only Semitic language in the European Union and the only Semitic language written in a Latin alphabet. The Maltese alphabet makes use of some special graphemes that differ from other Latin alphabets (the sound values are given in the International Phonetic Alphabet): ċ [tʃ], ġ [dʒ], għ (mostly silent), ħ [h], ż [z]

Some particular characteristics of Maltese are

- free word order
- semitic morphology (i.e. “word design”)
- aspect-based temporal system
- finite verbs in complex predicates

Even though there are no case endings, Maltese has a very free word order. The sentence *Il-kelb gidem il-qattus ilbieraħ* 'The dog bit the cat yesterday.' has the word order S(ubject) V(erb) O(bject) but could also be expressed as:

Ilbieraħ il-kelb gidem il-qattus. *(SVO)*

Gidem il-qattus il-kelb ilbieraħ. *(VOS)*

Il-qattus gidem il-kelb ilbieraħ. *(OVS)*

In the last example, since there is no case marking, *il-qattus* could be mistaken as being the subject of the written sentence. However, in spoken discourse, this OVS sentence would have a different intonation from an SVO sentence so that this ambiguity would not occur.

As a Semitic language, Maltese shows a non-concatenative morphology, i.e. inflected and derived word forms change internally:

In languages like English, word forms are made up of stems and affixes, i.e. concatenatively. The verb *shoot* can be inflected for third person by attaching the affix *-s* to the stem as in *(he) shoot-s*. Also, from the verbal stem a noun can be derived by adding the affix *-er* as in *shoot-er*. Hence both inflection and derivation take place without internal changes to the structure, i.e. concatenatively.

In Maltese, the basic “unit” within a word is not a stem but a root made up of three (sometimes four) consonants in a fixed order that carry a general meaning. Word stems with their specific meaning are formed by arranging the consonants according to a certain pattern. For example, the root *k-t-b* carries the meaning of everything connected with “writing”. In the following, patterns are represented as numbers **1,2,3** for the root consonants and **v** for the vowels between them, for example **1v2v3**. By applying the pattern **1v2v3** and filling the vowel positions between the root consonants 1,2 and 3 with the vowel sequence i-e, one gets the perfective verb *kiteb* 'he wrote'. Inflection of this verb for plural takes place by affixation of the plural affix *-u*, giving the form *kitbu* 'they wrote'.

Applying the pattern **1v22v:3** to the root renders the agent noun *kittieb* 'writer'. Inflection of the noun by adding the affix *-a* gives the plural *kittieba* 'writers'. Note that the plural suffix *-a* looks similar to the feminine marker *-a* so that *kittieba* could also refer to a female writer. The other Semitic Maltese plural suffixes are *-in* as in *mħallef* 'judge', *mħallfin* 'judges';*- at/-iet* as in *kittieba* '(female) writer', *kittiebat* '(female) writers'; *-ijiet* as in *żmien* 'time', *żminijiet* 'times'.

Plural nouns in Maltese can also be formed non-concatenatively (the so-called broken plural forms), i.e. no affixation takes place, but the noun is changed internally, e.g. *ktieb* 'book' vs. *kotba* 'books'.

Loan verbs today are mostly imported using a special verb class thatcan accommodate undigested stems. For example, the English stem *park-* became the basis of the Maltese verb forms *pparkjajt, pparkjat, pparkja* 'I/ she/ he parked'. Today, this formerly marginal Semitic special verb class has increased in size due to the influx of English loan verbs. It is highly productive, often giving way to ad-hoc loans of English verbs which already have a Semitic counterpart in Maltese. For example 'to download (a file)' can be expressed using the Semitic verb *niżżel* (originally meaning 'he caused to come down'). Taking the English stem *download* and importing it via the special verb class instead gives forms like *ddawnlowdjajt, ddawnlowdjat, ddawnlowdja* 'I/ she/ he downloaded'. This strategy is often criticised as corrupting the language (Fabri, forthcoming: p. 17).

Verbs in Maltese are marked for *aspect*, i.e. as to whether an action is completed (perfective) or not completed (imperfective). In the absence of any other grammatical markers, verbs in the perfective are interpreted as 'past tense' and verbs in the imperfective as 'present tense': *Andrew kiteb* 'Andrew wrote'; *Andrew jikteb* 'Andrew writes'. Combination of the imperfective verb with *kien,* the perfective form of the verb for 'to be', expresses habitual past: *Andrew kien jikteb* 'Andrew used to write'. Adding word *qed* 'progressive' (like the English *-ing* form) gives *Andrew kien qed kikteb* 'Andrew was writing' etc.

Maltese verbs do not have infinitive forms. Thus, in complex predicates like in the English sentence 'Andrew wants to write', both verbs are morphologically finite: *Andrew irid jikteb* (literally: 'Andrew he wants he writes').

Recent developments

With the rise of English to an international language and language of technology after the Second World War, the amount of English loan words in Maltese has grown to a great extent. Many of them have become “nativized”, i.e. they are adopted in regular use so much that even derived Semitic words cannot replace them. For example, instead of the commonly used word *ajruport* (from English *airport*), the Semitic word *mitjar* once was proposed (derived from *tar* 'he flew'). However, it became never accepted by the language community. On the other hand, loan words enter the language very rapidly, being imported spontaneously, even though there are already proper Maltese words for them (for example *ddownlowdja* vs *niżżel* 'he downloaded'). This fuels fears among some that the language might become “corrupted” (Fabri, forthcoming: p. 17).

Another recent development for Maltese is its status as an official language of the European Union. This has both advantages and disadvantages (Fabri forthcoming: p. 20). On the one hand, Maltese has finally become an internationally recognised language, a status that it did not have for a long time, being marginalised as a "kitchen language" centuries before. On the other hand, Maltese EU translators are confronted with certain challenges: many technical and legal terms have yet to be "invented" for Maltese. This results eventually in lexical expansion of the language (definitely a positive aspect), which, however, has to be coordinated by a central body so that individual translators do not come up with different terms for the same concepts independently from each other (which is a serious problem). The central body to deal with this challenge is the National Council for the Maltese Language (*Il-Kunsill Nazzjonali tal-Ilsien Malti*).

Other developments in recent years concern the Maltese orthography. Maltese (together with English) became the official language of Malta on January 1, 1934 – in the orthography released by the Union of Maltese Writers (*Għaqda tal-Kittieba tal-Malti*) in 1924. Since then, the orthography has undergone three revisions (1984, 1992 and 2008).

The last reform was released in 2008. Its aim was to reduce writers' insecurities that resulting from a considerable numbers of spelling variants for certain words. A great amount of variants could be reduced by finding a consistent balance between grammatical and phonetic spelling. Thus the four variants *zobtu, zoptu, sobtu* and *soptu* ('suddenly, unexpectedly') could be reduced to the two variants *zoptu* for [ˈzɔp.tʊ] and *soptu* for [ˈsɔp.tʊ]. For a similar reason, the word *skond* [skɔnt] 'according to', was changed to *skont* since its other grammatical forms do not justify spelling with *d* (derived from Italian *secondo*), as e.g. *skontok* [ˈskɔn.tɔk] 'according to you'.

For the third area (loan words), the principle remains to write loan words according to the Maltese orthography if they are regarded as "nativised" and if it does not result in conflicts with the pronunciation or with other Maltese writing rules. However, many Maltese prefer to write English loan words with their original spelling, since they have become used to them. In fact, during a public seminar on the treatment of English loan words in April 2008, there were emotional discussions among the audience when it came to words like *email* and their proposed new spellings as *imajl*. Factors like the habits of a language community make the standardisation of spellings even more difficult than finding the balance between grammatical and phonetic principles.

These examples only give a slight idea of the hard work that the *National Council for the Maltese Language* is undertaking as part of language cultivation in Malta. The next section will give an insight into the history of language cultivation in Malta.

Language cultivation in Malta

Compared to other languages of Europe, the status of Maltese as an official language (since 1934) itself is a recent development. Thus language cultivation, too, had a late start.

For centuries, Maltese was only the spoken medium of the Maltese population and marginalised in comparison to the respective official language of Malta's rulers. This started to change with the language movement of the mid-/ late 18th century when first systematic studies of the language were conducted by Agius de Soldanis (1750) and Mikiel Anton Vassalli (1797). Especially Vassalli promoted the Maltese language by promoting its use in every domain of everyday life. Fortunato Panzavecchia's bible translations of the mid 19th century contributed to further standardisation of the language. And with the move towards a standardised orthography in the early 20th century, an important step was made by the foundation of the Union of Maltese Writers (*Għaqda tal-Kittieba tal-Malti*) in 1920. The orthographic system, which was developed by this organisation, became Malta's official orthography in 1934 and, with some changes and additions, has been in use since.

In 1964, after gaining independence from Great Britain, the status of Maltese as national language and as official language together with English was written into the constitution. When Malta joined the EU in 2004, Maltese became an official language of the EU. As noted in the section above, this results in certain challenges, which can only be solved by a body that coordinates standardisation and common practice in translation work.

The body in Malta to do this work is the National Council for the Maltese Language (*Il-Kunsill Nazzjonali tal-Ilsien Malti*). It was founded in 2005 as the first government organisation to officially deal with language matters and language planning for the Maltese language. The Council's tasks are, as formulated in the Maltese Language Act (ACT No. V of 2004): promoting the Maltese language, to “adopt a suitable linguistic policy backed by a strategic plan” and put it into practice. Moreover, the Council's task is to update the Maltese orthography and decide on correct spellings (taking over the task from the Academy of Maltese and being mainly responsible for the Maltese orthography reform of 2008). On its website, the Council also offers training courses for proofreaders and Maltese language courses for foreigners[8].

Before the Council was founded, standardisation of orthography was the task of the Academy of Maltese *(Akkademja tal-Malti)*. It emerged 1964 from the Union of Maltese Writers *(Għaqda tal-Kittieba tal-Malti)*, which had been the founding body for the first official orthography in 1924/1932. The Academy's main aim today is to promote academic studies in the Maltese language and literature, promote the use of Maltese in every domain of everyday life and to build up contacts to people who are friends of the language and who use it outside of Malta[9]. The Academy works closely together with the National Council for the Maltese Language.

The motivation behind the Maltese Language Act was the idea that one national language which is shared by all individuals within that nation forms the basis for cultural and national identity. This of course calls for standardisation of the language. Indeed, from the language cultivation movement of the 19th century until today, Maltese has risen from a formerly marginalised vernacular to a national language of high prestige. This is also reflected in the ever-growing amount of literary works in Maltese during the same timespan and in the high number of influential organisations and bodies for the Maltese language and literature (see Fabri forthcoming, p. 22).

Language in Education

Particularly in a bilingual society like in Malta, several aspects play a role when it comes to language in education.

One aspect is the language of instruction, i.e. the language that is used officially by the teachers during the lessons in school or in the seminars at the university.

Another factor is the language used in certain school books. With English being the language of technology and natural sciences, most of the school books on these topics are in English. In fact, efforts to translate technical and scientific terms into Maltese have encountered several problems, one of them being the acceptance by the language community. Hence the school subjects, too, possibly determine the language of instruction for certain lessons, although it can also be that English school books (and the English terminology contained therein) are used while the language of instruction is Maltese.

Yet another aspect is the language used by individuals. Bilingual speakers not only use different languages in different social settings (“domains”) , e.g. Maltese with the family at home, English with foreigners, Maltese or English during school lessons etc. They also tend to mix both languages, either by language mixing (e.g. English words are mixed into a conversation conducted in Maltese) or by code-switching (e.g. a conversation in Maltese switches to English and back again, with the English parts being larger than just single words, often consisting of several sentences). Thus even during school lessons that are taught in one language, conversations between teachers and students can switch between the languages.

Keeping these three factors in mind, it becomes clear that the actual exposure of students to the respective language in schools or at the university is something different from the chosen language of instruction.

Regarding the official language of instruction in education, both Maltese and English can be found in schools and at the university, since Maltese and English share the status as official languages in Malta. In schools, both are taught as subjects from early on. Which language is used as language of instruction depends on the type of school. Private schools tend to use English more than Maltese (sometimes to a greater extent), while in state schools Maltese is slightly preferred to English. Church schools have their individual preferences in that some traditionally prefer one language over the other.

As was mentioned before, most science books that are used in school are in English. Thus, with the introduction of more and more scientific subjects later in school and even more so at the university, students are exposed to the two languages at the same time, using them for different situations: they might have their lessons taught in Maltese, but read their books and write their essays in English. Especially for students at university, conversations between them, friends and lecturers often take place in Maltese, sometimes code-switching/mixing between Maltese and English, or they are even in English only (the latter for example with international students or lecturers).

At home with their family and friends, however, most Maltese speak Maltese, some mix languages and only a few families speak English only.

As can be seen from the examples above, despite the fact that both Maltese and English are used as languages in education, there is a clear distribution when it comes to their use in society. Sciriha and

Vassallo (2001, p. 29, cited in Fabri, forthcoming: p. 18) point out that "70% of the respondents claimed to use Maltese at work, while 90% said they communicate with their family members at home in Maltese. ... the percentages for spoken Maltese are extremely high but go down for other skills like reading and writing."

This distribution of Maltese being used mainly as the spoken medium and English mainly as the written medium bears a certain risk, as it can have an impact on different skills of its native speakers when it comes to speaking, reading or writing. In order to give reasons to this statement, one has to look at the basic characteristics of spoken and written language.

In general, written texts differ from spoken discourse in a number of ways. What they have in common is that both are ways of transferring information between two parties, i.e. speaker and hearer, and writer and reader, respectively. However, they differ in the way how information is passed on between them. Putting it in a simple way, a written text unlike spoken discourse is set outside a concrete interactive communicative situation. Spoken discourse, on the one hand, depends on the interaction between speaker and hearer. The speaker has to structure the information in a certain way. This is important because of the limited human short-term memory: a hearer in a conversation can only take in a certain amount of information before he has to interrupt and ask the speaker to make sure that he understood.

A written text, on the other hand, is non-interactive in so far as the reader cannot ask for more specific information. He can however, browse back and forth in the text (something that a hearer cannot due in discourse). In that way, a written text itself serves as the long-term memory for the reader. Thus, a written text structures information differently than would be done in a spoken conversation. For example, a text has to provide more background information in order to provide a common ground with the reader before the actual information flow starts. This is not a problem, given that a text can serve as a long-term memory for the reader. In fact, it allows for a more elaborated structure than spoken discourse, i.e. it usually contains longer sentences and a higher amount of subordinate clauses.

This register (i.e. "language style") distinction is what in the literature has been dubbed *orate* versus *literate* text structures. Of course, a text can also be written in an orate register that resembles spoken conversations (e.g. in forum chats or informal emails). But it is not the register normally used in e.g. essays. Ideally, native speakers acquire the literate register already from an early age on, e.g. by their parents reading stories to them. Later in school, this knowledge is deepened by active exercise in writing essays, for example.

A literate register develops over time in a language with a literary tradition. Maltese, compared to its short history as an officially written language (since 1934) has a long and rich literary history. Even though the oldest literature discovered is very sparse (*Il Cantilena* by Pietro Caxaro, dating back to about 1450), a literary tradition started to form around the 1740s. In the 19th century, the amount of literature in Maltese was growing (Fabri, forthcoming: p. 25), and with it, Maltese was expanding. Today it is a language with a fully fledged literate register.

This register, however, needs to be exercised in order to keep up the status of the language as a both conversational and literary language. The trend in higher education to write more essays in English than in Maltese, at least theoretically, bears the risk of reducing Maltese to the orate register. A higher amount of Maltese websites of all genres is desirable to cover both registers and their subtypes in order to ensure a stable status of the language in all its richness.

International aspects

Bearing the previous sections in mind, it should be clear now that the international aspects of Maltese differ to a great extent from other languages. With under a million native speakers worldwide, Maltese is considered a “lesser-spoken” language. In its history, it was the language of occupiers but rather one of the occupied. As a result of this, Maltese has never became what is traditionally considered an international language or lingua franca as was the case for e.g. Latin, Spanish, Portuguese or English, all of which being the languages of conquerers. It did spread to other countries, where it is still spoken today (Australia, Canada, USA and UK), but only as a community language. It took nearly 200 years from the first interest of Maltese grammarians in their own language until it eventually gained the status of an official language. And even then, the other official language, English, served as the language for international relations.

A change for Maltese to become an internationally visible language came with Malta's joining of the EU in 2004. Since then, it has been an official language inside the European Union, with all the benefits and challenges which are connected to this status.

Academically, Maltese was discovered as a subject language in the field of Linguistics as early as 1936 with Sutcliffe's *Grammar of the Maltese Language*. It was not until the 1960s, however, that it gained wider international academic awareness through the publications by Joseph Aquilina (e.g. *Papers in Maltese Linguistics* (1961)). Since then, more and more scholars outside Malta have taken an interest in Maltese. 2007 saw the foundation of the International Association of Maltese Linguistics (*Għaqda Internazzjonali tal-Lingwistika Maltija*)[10], an association of linguists who are interested in the Maltese language. The main aim of GĦILM, as stated on its website, is to provide “a connection between interested scholars from all subdisciplines of Linguistics”, thus facilitating research on Maltese. It also wants to bring together people from different backgrounds who work with the Maltese language (linguists, translators, students and others).

Maltese on the Internet

A survey of the National Statistics Office of Malta in the second quarter of 2009[11] shows that among a population of roughly 400,000 persons, 67 per cent had access to a computer and 64 per cent had access to the internet. A recent Eurobarometer survey (published in May 2011)[12] among European internet users' browsing habits showed that only 6.5 per cent of Maltese internet users use exclusively Maltese on the internet when reading, consuming content or communicating. Instead, 90.6 per cent choose to browse websites in English and 20.1 per cent Italian, respectively. These figures formed the basis of an article in the Maltese daily newspaper “The Times of Malta”, which provoked a lively discussion mostly among Maltese readers of the online edition[13].

The exact findings in the survey, however, point to the conclusion that this habit is not a deliberate choice:

When asked which language Maltese considered their mother tongue, 89.5 per cent of the respondents claimed that Maltese was their mother tongue (opposed to only 7.6 per cent for English and 0.2 per cent for Italian).

Languages other than respondents' own used to read or watch content on the internet were English (90.6 per cent) and Italian (20.1 per cent). Only 6.5 per cent responded that they only use their own language, which is not surprising, given that most Maltese are bilingual in Maltese and English and a considerable number speaks Italian as well.

When writing on the internet, numbers in favour of Maltese are higher than when reading or watching content: 87 per cent claimed they used Maltese, 85 per cent English and 8 per cent Italian.

The reason for the majority to use English as the language for consuming online content may be just the limited number of websites in Maltese rather than the favour for English per se. Remember that most respondents did not regard English as their own language and that the usage of Maltese increased when producing content on the web, even though this use of Maltese in most cases takes place in chat forums and social platforms, hence in a colloquial style, i.e. in the orate register.

A peculiarity about the Maltese used by the younger generation in social platforms and chat forums is its phonetic spelling, without the silent characters like *għ* and *h*. Thus *għax* 'because' is written as *ax*, *tiegħi* 'my' as *tiei* etc. The reason for this may be the late introduction of Maltese special characters into the PC world. Although Maltese has been implemented in the Unicode framework since its start, computers and operation systems followed much later. The *Maltese Standards Authority* released a standardised Maltese keyboard layout in 2002, and Microsoft's operation system Windows has been available in a Maltese language version since as late as 2006 (with Windows XP). In the case of mobile phones, the special Maltese letters are still not implemented. Hence it will be seen whether the ad-hoc orthography of the chat forums will give way to a spelling with special characters once they are available on mobile phones or whether this phonetic orthography will survive as a “sociolect” of the younger generation.

As for the amount of Maltese on the internet in general, it is hard to come up with exact numbers, not least because the number of websites is changing constantly. But there are other factors which give an idea about the amount of Maltese online in comparison to other languages.s

A first look at the amount of Wikipedia entries (on June 1st, 2011) showed that there were about 2,820 entries in Maltese in contrast to more than 3,640,000 entries in English and more than 1,238,000 entries in German.

Comparing the number of top level domains (TLD), the TLD .mt occupies rank 213 (out of 358) with an unspecified number of registered .mt domains (a member of the Network Information Centre Malta gave an estimate of about 5,000), opposed to 21,336,063 registered domains for .com (commercial, rank 1) and 5,459,604 domains for .de (Germany, rank 2). Of course, the number of registered domains does not tell anything about the language in which the pages under a certain domain are written.

Some rough numbers of the amount of Maltese language on the internet can be calculated using a procedure proposed by Kilgarriff and Grefenstette (2003). The basic idea is that function words (e.g. *but, for, this* etc) are more frequent than content words (e.g. nouns, verbs, adjectives). Moreover, they are a finite set in a language, i.e. new function words emerge less often than content words. The percentage of the function words in a language are stable in a text sample as the size of the sample increases (Zipf's Law). Thus, one can calculate the amount of words for any language on the internet as follows:

In the first step, one calculates the amount of selected function words of Maltese in a corpus (i.e. a text collection) whose size is known. In the second step, one uses a search engine (e.g. Google) to find out the frequency for the same function words on the web. In the third step, the frequency from the corpus count is extrapolated to the Google search and then an average is calculated for the frequency of function words in the search results.

A calculation for Maltese in 2010 (28/03/2010 by Albert Gatt, who did a Google search restricted on Maltese webpages) resulted in an average of 80,000,000 words. Compared with other languages, Maltese is more represented than Albanian, Breton, Welsh, Lithuanian and Latvian but less represented than Catalan, Malay, Turkish and Croatian. Gatt states that the estimate of 80 million words is very low, even for a conservative estimate. The reasons for this are that many Maltese webpages are written in English and that many webpages in Maltese are written without using Maltese special letters available in Unicode. Also, the search could only be performed on visible webpages, i.e. pages that had a visible URL. Nevertheless this calculation shows that Maltese is one of the rarer languages on the internet.

Apart from private home pages and weblogs, there are a number of official websites in Maltese. First of all, there is the home page of the Maltese government[14], which is available in both Maltese and English. Also, there are the internet editions of the Maltese language daily and weekly newspapers: *In-Nazzjon, L-Orizzont* (daily), *Illum, Il-ĠENSillum, KullĦadd, Leħen is-Sewwa, It-Torċa* (weekly).

The websites of the Maltese TV and radio stations show a mixture of both English and Maltese to different degrees. For example, the website of the stations NET TV[15] and Super 1[16] show a framework in English, with some articles in Maltese, even though their programme contains both Maltese and English titles. The church-owned radio station RTK[17] (Maltese and English) lets the user choose between the two languages. The website of the Public Broadcasting Services (PBS)[18] contains sections in English and sections in Maltese as has the website of Radio 101[19]. This mixture between English and Maltese reflects the language use in everyday life. Within the programmes, however, the situation is a clearer, since the *Maltese Broadcasting Authority* has issued strict guidelines for the use of Maltese on TV and the radio. Following those, presenters should speak in either Maltese or English and not switch between the two languages (Fabri, forthcoming: p. 28). Hence the programmes of the stations contain broadcasts in Maltese only and others in English only. Those are often available online as well, either as live stream or as podcasts.

Outside Malta, a big collection for Maltese texts is within the EUR-Lex[20] that hosts all official law and other documents of the European Union since 1951 in its 23 official languages.

Many if not all of these openly available web documents are used in corpus projects, e.g. the *JRC-Acquis Multilingual Parallel Corpus*[21], which is a parallel corpus containing the complete text of the European Union Law in 22 languages. Another corpus that contains a growing number of visible web documents in Maltese is the corpus on the MLRS (Maltese Language Resource Server)[22].

Selected Further Reading

Ambros, Arne (1998): Bonġornu, kif int?: Einfu hrung in die maltesische Sprache. Wiesbaden: Reichert.

Borg, Albert J. & Azzopardi-Alexander, Marie (1997): Maltese. Routledge.

Brincat, Joseph M.: Malta 870-1054 (1995): Al-Himyarī's Account and its Linguistic Implications. Malta: Said International.

Bovingdon, Roderick (2001): The Maltese language of Australia: Maltraljan: a lexical compilation with linguistic notations & a social, political and historical background. Munich: LINCOM Europa.

Fabri, Ray (to appear): "Maltese" In: C. Delcourt and P. van Sterkenburg (eds.) The Languages of the 25. Revue belge de Philologie et d'Histoire: RBPH, 85 (2007) 3 and Amsterdam-Philadelphia: John Benjamins.

Kilgarriff, A. and Grefenstette, G. (2003): 'Introduction to the special issue on the web as corpus', Computational Linguistics, 29: 333-347.

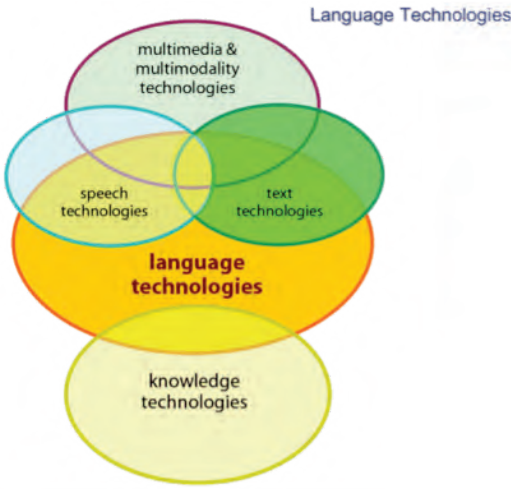
Kontzi, Reinhold (2005). Sprachkontakt im Mittelmeer: Gesammelte Aufsa tze zum Maltesi-schen. Narr, Tu bingen.

Mifsud, M. (1995). Loan verbs in Maltese: a descriptive and comparative study. Leiden etc: Brill.

Language Technology Support for Maltese

Language Technologies

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



Language Technology Application Architectures

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

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

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

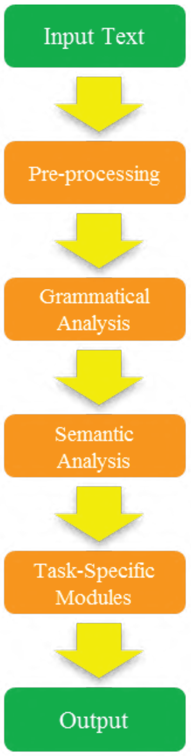


Figure 2: A Typical Text Processing Application Architecture

After the introduction of the core application areas, we will briefly give an overview of the situation in LT research and education, concluding with an overview of (past) funding programs. In the end of this section, we will present an expert estimation on the situation regarding core LT tools and resources in a number of dimensions such as availability, maturity, or quality. This table gives a good overview on the situation of LT for Maltese.

Core application areas

Language checking

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

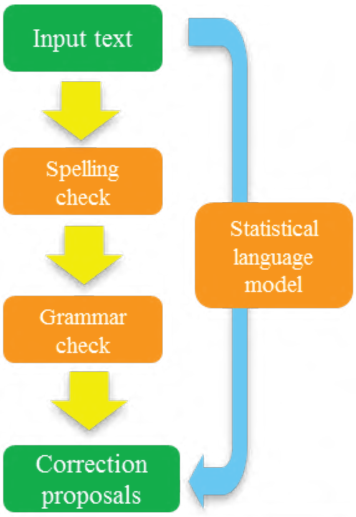


Figure 3: Language Checking (left: rule-based; right: statistical)

Eye have a spelling chequer,
It came with my Pea Sea.
It plane lee marks four my revue
Miss Steaks I can knot sea.

For handling this type of errors, analysis of the context is needed in many cases, e.g., for deciding in which position in a Maltese verb the silent *għ* has to be written, as in:

a) *...in-negozjati li kien għamel il-Gvern ...*

[...the negotiations that the government had made...]

b) *Pawlu, agħmel l-eżamijiet!*

[Paul, do the exams!]

c) **...in-negozjati li kien agħmel il-Gvern ...*

Both *għamel* 'he made' and *agħmel* 'make!' are pronounced [ˈɐː.mɛl].

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, *kien għamel* is much more probable word sequence than *kien agħmel*. A statistical language model can be automatically derived using a large amount of (correct) language data (i.e. a corpus). Up to now, these approaches have mostly been developed and evaluated on English language data. However, they do not necessarily transfer well to highly inflectional languages like Maltese, where a given word type, such as a vcrb, can yield a large number of orthographic forms.

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

As with other languages, a means to determine whether a given string is a valid word is not a sufficient condition for spelling-error detection, but it is a necessary condition. As yet, no such means exists for Maltese, though various attempts have been made.

One of the earliest was by Mangion (1999), which tried to solve this problem using a rudimentary form of rule-driven morphological analysis. Essentially a word was considered valid if it could be derived by rule from a stem found in a dictionary. The problem with this approach is that it requires a complete list of all stems, and course, the rules have to be very accurate, so results were somewhat limited by the absence of a lexicon of roots and the imperfect nature of the rules.

A second approach looked to statistics for a solution. The intuitive idea is that for a given language, certain sequences of characters are highly unlikely. In English, for example, we never find the sequence "kk", so if that occurs as a substring in a written word, we can confidently assert that the word contains a spelling error. We can calculate the probability of an entire string as a function of the probabilities of all its substrings, so that, more generally, we can adopt the principle that the probability of a string of characters must exceed a certain threshold to count as a valid word.

A statistical spell checker making use of such a principle was developed (Mizzi 2001). It did not require a lexicon, being based instead on the distribution of character n-grams found in a newspaper corpus. It became clear that for this approach to succeed (i) a more accurate language model is needed for which more language data was required than was then available, and (ii) that string probability alone is insufficient to accurately classify an orthographic word as an error. As suggested above, other information is necessary, such as part of speech information from the surrounding context.

Other attempts to develop a spell-checker for Maltese include an online checker that has been developed by Mr. Ramon Casha of the Linux User Group[23]. This is based on a wordlist of around 1 million word types, some of which are generated by rule. Its accuracy has not been officially established. Microsoft has also been working on a spell checker for inclusion with their Maltese language interface pack.

The use of language checking is not limited to word processing tools. Other application areas are **authoring support**, for example to assist the writer of technical documentation to use technical vocabulary consistently, and the field of **computer-assisted language learning**. Language checking is also applied to automatically correct queries sent to search engines, e.g. Google's "Did you mean..." suggestions.

Apart from an interactive CD picture dictionary (Sciriha 1997), no such applications have been developed for Maltese to date.

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-wide24. Since 2004, the verb *google* even has an entry in the Cambridge Advanced Learner's dictionary. Neither the search interface nor the presentation of the retrieved results has significantly changed since the first version. In the current version, Google offers a spelling correction for misspelled words and also, in 2009, incorporated basic semantic search capabilities into their algorithmic mix25, which can improve search accuracy by analysing the meaning of the query terms in context. The success story of Google shows that with a lot of data at hand and efficient techniques for indexing these data, a mainly statistically-based approach can lead to satisfactory results.

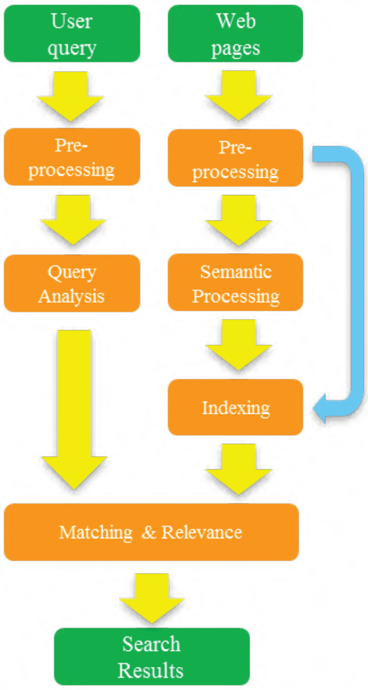


Figure 4: Web Search Architecture

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 (or the equivalent German GermaNet), have shown improvements by allowing to find a page on the basis of synonyms of the search terms, e.g. *Atomkraft*, *Kernenergie* and *Nuklearenergie* (atomic energy, atomic power, and nuclear energy) or even more loosely related terms.

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.

In Malta, there are a number of search websites that are specifically oriented towards Malta[26]. In addition there are a small number of Malta based SMEs that incorporate relatively sophisticated language processing techniques within search applications. Charonite[27], for example, is a local SME dealing with search engine optimisation. However, at the time of writing there are no commercially available search engines that are specifically oriented towards the Maltese language, apart from a prototype for cross lingual information retrieval developed within the scope of a European FP6 research project called LT4EL[28] which used multilingual language technology tools and semantic web techniques for improving the retrieval of learning material.

Speech interaction

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

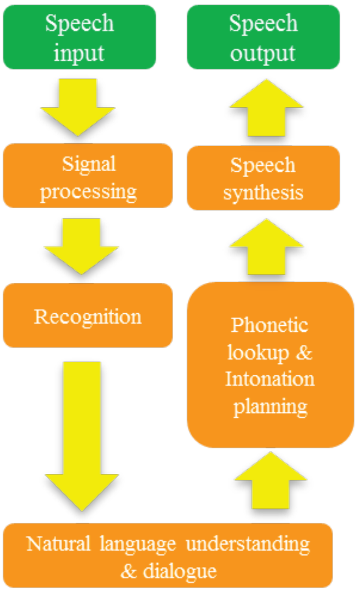

Speech input
Speech output
Signal processing
Speech synthesis
Recognition
Phonetic lookup & Intonation planning
Natural language understanding & dialogue

Figure 5: Simple Speech-based Dialogue Architecture

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

- Automatic speech recognition (ASR) is responsible for determining which words were actually spoken given a sequence of sounds uttered by a user.
- Syntactic analysis and semantic interpretation deal with analysing the syntactic structure of a user's utterance and 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 contrast, today's TTS systems prove superior, though optimisable, regarding the prosodic naturalness of dynamic utterances.

Regarding the market for Speech Interaction technology, the last decade underwent a strong standardisation of the interfaces between the different technology components, as well as by standards for creating particular software artefacts for a given application. There also has been strong market consolidation within the last ten years, particularly in the field of ASR and TTS. Here, the national markets in the G20 countries – i.e. economically strong countries with a considerable population - are dominated by less than 5 players worldwide, with Nuance and Loquendo being the most prominent ones in Europe.

Most speech technology in Malta has concentrated on text-to-speech (TTS). Some pioneering work was initially carried out by P. Micallef (1997) and this was followed by a number of Master's dissertations (Calleja 2002). Some preliminary work a web-based TTS system (Buhagiar and Micallef 2008).

A significant development in Maltese speech synthesis was the winning of a government tender the development of a speech synthesiser by the local company Crimson Wing Malta Ltd. This work is partly financed by the EU Regional Development fund and commissioned by the Maltese Foundation for Information Access (FITA). The prototype will be SAPI compliant and will include three voices (male, female, and child). According to a recent presentation (Borg et al. 2011) the work is advancing well and a prototype, expected in 2012, will be freely available for download.

Work on speech recognition is less advanced. A prototype for recognising numerals was created by (Calleja 2004) in simple domains. With respect to speech, the fundamental problem remains a lack of suitably annotated data since this requires significant manual effort. The creation of a corpus and descriptive framework for the study of Maltese intonation was initiated by the Institute of Linguistics carried out by Vella and Farrugia (Vella and Farrugia 2006). It is expected that the work by Crimson Wing will also yield some corpora which will be made available for research.

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

Machine translation

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

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

Il-Kuntistabbli osserva lir-ragel bit-teleskopju.

[The policeman observed the man with the telescope.]

Il-Kuntistabbli osserva lir-ragel r-ragel bir-rivolver.

[The policeman observed the man with the revolver.]

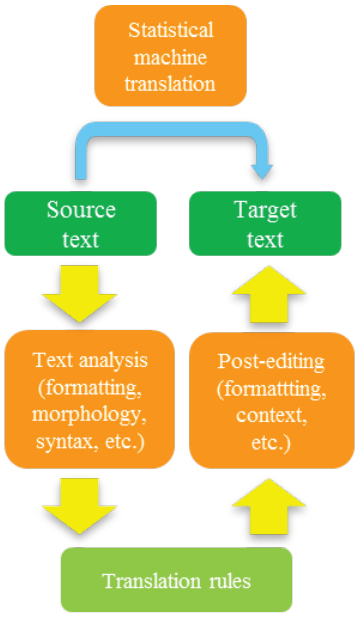

Statistical machine translation
Source text
Target text
Text analysis (formatting, morphology, syntax, etc.)
Post-editing (formattting, context, etc.)
Translation rules

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

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.

In Malta work carried out in Machine Translation has been restricted to just a few Bachelors and Masters dissertations. A transfer system based on LFG was developed for English/Maltese by Farrugia (2000) and successfully translated weather forecasts.

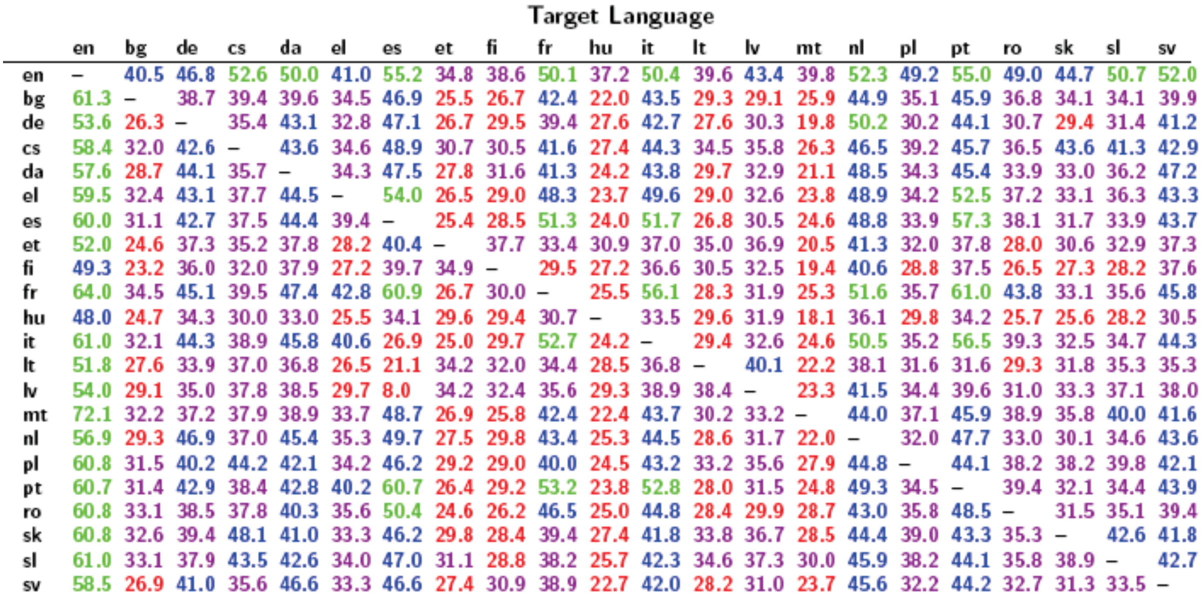
Later J. Bajada (Bajada 2004, 2009) worked on statistical MT (SMT) with the emphasis on techniques for producing language and translation models. The earlier work concerned word-based models, whilst the latter developed techniques for gathering bilingual phrase data from a limited corpus.

Like in so many other areas, the underlying problem is a lack of large quantities of suitably annotated bilingual data. For this reason, perhaps, the benchmark system against which to judge advances remains Google Translate.

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

Evaluation campaigns allow for comparing the quality of MT systems, the various approaches and the status of MT systems for the different languages. Table 1, presented within the EC Euromatrix+ project, shows the pairwise performances obtained for 22 official EU languages (Irish Gaelic is missing) in terms of BLEU score29.

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



Target Language

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

Table 1: Pairwise performances obtained for 22 official EU languages in Machine Translation (source: Euromatrix

Language Technology 'behind the scenes'

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

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

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

Two 'borderline' areas, which sometimes play the role of 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.

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.

In Malta the vast majority of research and education in LT has taken place at the University of Malta. However, it was established rather late. One reason for this was the late appearance of Computer Science as a curriculum subject at the University. The turbulent political leadership of the country during the 1970s and 1980s had not foreseen the information revolution to come and it was not until the early 1990s that an undergraduate option in Computing with Mathematics was offered through the Faculty of Science.

The roots of change same in 1994, when a national strategic initiative was undertaken to recognise and strengthen the role of IT in commercial, political, and above all, educational sectors. One immediate consequence of this was the introduction of a substantial four-year Bachelors programme - the BSc. IT (Hons) - at University as well as the founding of a new Department of Computer Science and Artificial Intelligence (CSAI[30]). A course in NLP was included as an advanced option, and this led, four years later, to a series of undergraduate final year projects tackling language processing issues including computational approaches to Maltese[31]. The Department of Computer Communications Engineering also participated in the programme, and this led to another set of undergraduate projects in speech technology

Another important influence on research is the University's Institute of Linguistics (IOL), founded in 1988 with the aim of teaching as well as promoting and coordinating research in both General and Applied Linguistics, furthering research involving the description of particular languages, not least Maltese, fostering the study of the various sub-fields of linguistics, and promoting interdisciplinary research involving academics in practical cooperation that cuts across departmental and faculty boundaries. abroad. The Institute of Linguistics runs two undergraduate programmes: a B.A. in General Linguistics and a new B.Sc. in Human Language Technology which will be on offer in October 2011. It is also possible to do a Masters Degree and a Ph.D. in Linguistics with the Institute.

In 1997, an interdisciplinary group of computer scientists and linguists[32] embarked on Maltilex, a project to create a computational lexicon, which was sustained by a small grant from the University supported by the Mid-Med Bank. A simple web-based interface was developed to enable the creation and maintenance of entries, as reported in Rosner-et-al (1998) at the first ACL Workshop on Computational Approaches to Semitic Languages (Rosner 1998). Several thousand such entries were created by hand, but the project ran into legal problems, the compilation of entries having been largely inspired by Joseph Aquilina's existing paper dictionary (Aquilina 1987).

Effort then shifted from paper dictionaries to extraction of lexical entries from other sources. Two dissertations (Dalli 2001, Attard 2006) used techniques based on alignment derived from bioinformatics - to cluster lexical entries and this was used as a means of structuring the lexicon automatically.

Despite lack of funding, the Maltilex effort continued in a somewhat piecemeal fashion, supported by staff at the IOL and CSAI Department. It was not until 2005 that Malta's Council for Science and Technology (MCST) launched the country's first Research and Technology Development Initiative and a joint proposal for a Maltese Language Resource Server (MLRS) was accepted, providing sufficient financial support to employ a researcher full time between 2006 and 2008. The project had the twin goals of creating both a lexicon and a corpus (Rosner 2008), and it laid the foundations for the present MLRS server.

The research mentioned above mainly deals with the written language. Two branches of speech-related work are also ongoing.

The first, initiated from the signal-processing tradition within the Engineering Faculty, yielded a prototype speech synthesizer (Micallef 1997). His work has influenced several other projects aimed at improving speech synthesis from a low-resource perspective including Calleja (2002), Farrugia (2004), Camilleri (2010) , Borg-et-al (2011).

The second, tackles the issue of intonation (Vella 2007) from a linguistic perspective. Some pioneering work to create a corpus and descriptive framework for the study of Maltese intonation was carried out by Vella and Farrugia (Vella and Farrugia 2006).

Outside Malta, two research groups that are in active collaboration with local LT-oriented efforts deserve a special mention.

At the University of Arizona a group led by led by linguist Adam Ussishkin is particularly interested in the psycholinguistic issues pertaining to Semitic languages including Maltese. To study these issues a online corpus has been made available (Ussishkin et-al 2009).

At the University of Bremen, Prof. Thomas Stolz has been actively involved with the academic study of Maltese but is particularly known for having hosted the first conference on Maltese Linguistics in Bremen (Comrie et al 2009), founded a periodical[33] and the International Association of Maltese Linguistics, also based in Bremen, that exists alongside the Malta-based Council for the Maltese Language

As mentioned, the LT-sensitive communities existing at the University of Malta mainly inhabit the Faculty of ICT, the Institute of Linguistics. There is also a potential interest in Faculty of Arts (Department of Maltese) and other Humanities subjects though up until now computational linguistics tends to be regarded as an exotic topic located in the more scientific computer science faculties or in the humanities and, therefore, the research topics dealt with only overlap only partially.

Curiously, Malta does not lack for LT-related international events. LREC 2010 was held in Valletta, drawing 1200 participants. The annual EAMT conference was also held in Malta in 1994, and there have also been a number of smaller workshops held during the last 10 years.

Language Technology Programmes

Malta joined the EU in 2004 and this event immediately conferred to Maltese the status of being an official EU language. With this status came new obligations - in particular to translate large quantities of official documents, and in addition, a recognition, at European level, that as a national language, it should have "first-class" status from a technological as well as a social perspective, and be accorded all the rights and privileges enjoyed by "larger" European languages (i.e. having larger numbers of native speakers).

The government's National IT Strategy 2008-10 included a number of objectives related to Maltese Language including (i) the development of online government in Maltese, (ii) creation of Maltese language tools, in collaboration with the University, and (iii) support for Maltese online communities. At the time of writing in 2011, not all the objectives have been realised. However the longer term effects of this strategy are beginning to take shape.

Currently language technology scene in Malta is currently under the influence of four main initiatives:

1. First of all, a government-supported project partly funded by EU regional development funds is under way to bring speech technology within the reach of disabled persons. The project is currently focused on Maltese speech synthesis, and at this point the relevant language models are in the process of being developed. The consortium, which consists of an SME[34], a foundation[35], and the University, has pledged that these resources will be made available for research purposes. It remains to be seen whether components of the speech synthesiser will be made available to resource sharing networks inspired by CLARIN and META.
2. Second, as is evident from the current report, Malta participates in METANET4U and is thus in receipt of significant EU funding aimed at the enhancement and distrubution of resources and tools that are specifically for Maltese. The University of Malta is a member of META-NET and intends to fulfil its obligations towards the aims of META, particularly regarding the identification of stakeholders, actually and potential.
3. Third, the Maltese Language Resource Server (MLRS) has come to fruition and significant efforts are under way at University, through the Institute of Linguistics[36] and the Department of Intelligent Computer Systems[37], to maintain and develop it. Currently MLRS is online at http://mlrs.research.um.edu.mt. The corpus comprises some 80M words, and the system includes some basic services that include KWIC search and display, pattern-directed search, various kinds of statistical analysis etc. Further tools are currently planned including a part-of-speech tagger and a spell-checker.
4. Finally, a new undergraduate programme in Human Language Technology is destined to be launched by the Institute of Linguistics in October 2011. This will cover a full range of topics and will inevitably have a positive long-term effect on the study of Maltese from a computational perspective.

Besides these, a project to develop an electronic version of the Aquilina dictionary is currently in preparation. This is a collaborative effort between the University of Malta who are supplying the linguistic expertise, the University of Arizona, who have already digitised the dictionary into machine readable form, and the publishers Midsea Books of Valletta. The dual aims of the project are to update the content, and to confer upon researchers the flexibility to swiftly access the text. At the same time, an effort is in progress locally, to organise the right level of lexicographic expertise necessary to update the content of the original paper dictionary.

We should also mention Malta's relationship to CLARIN, a proposed EU research infrastucture addressing the provision of language resources for the Humanities and Social Sciences. During specification phase, the University was able to participate thanks to a small support grant from the local Council for Science and Technology. However, it has turned out to be more challenging to secure the longer term funding required for the construction phase of CLARIN. Identification of a suitable government entity to take responsibility for the programme has so far been without success. Consequently, Malta's future participation in the construction phase currently hangs in the balance.

Availability of Tools and Resources for Maltese

The following table provides an overview of the current situation of language technology support for Maltese. The rating of existing technologies 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 for Maltese

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

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

Conclusions

In this Whitepaper Series, the first effort has been made to assess the overall situation of many European languages with respect to language technology support in a way that allows for high level comparison and identification of gaps and needs.

For Maltese, the most evident characteristics revealed by the table are that

- most entries are blank, and
- the highest grade scored is 4.

The fact that most entries are blank reflects the immature state of LT-related research and development in Malta. Although there are signs that the situation is improving, investment in language technology remains at a low level, and as a result, despite modest local achievements, the effort is fragmentary, both in terms of coverage of different areas, and in terms of sustainability of research: there have been too many projects involving just one area, just one researcher, and just one or two years. The collective efforts don't add up as they should.

So what has been achieved? We can see by looking at the non-blank entries, whose average score yields the following ordering:

- Tools:
  1. Tokenisation, Speech Synthesis
  2. Speech Recognition
- Resources:
  - Reference Corpora
  - Parallel Corpora
  - Lexicons, Terminology (this should be understood to include wordlists)
  - Language Models

With respect to tools:

- Low level text extraction and processing tools are available, including a tokeniser. A POS-tagger is under development, but its performance is not state-of-the-art, pending further training with better annotated data.
- Higher level tools (syntactic or semantic analysis, classification tools, information extraction etc. are entirely lacking. A consequence is that, for example, there are no treebanks available for Maltese.
- Prototype speech recognition tools have been developed at University but are not readily available at the time of writing. However, the government-funded speech engine mentioned earlier should yield a working speech synthesizer by 2013. Whilst this is a very positive development, it is highly focused on the synthesis side of speech. Almost now work on speech recognition is planned at this stage.

With respect to resources, the situation is a little more structured, in so far as there already exists MLRS, an extensible computational infrastructure in the form of a server providing the basic functionality to enable access over the web to available corpora, some services, and a rudimentary system to facilitate the submission of contributions. MLRS currently provides some very basic services for the extraction, representation, search and analysis of text.

The existing MLRS corpus is currently around 100 million tokens in length. It is predominently textual and monolingual. It is also somewhat non-representative: there is no shortage of legalistic material, but there is currently a lack of academic text and works of fiction.

As things stand, these materials can only be searched and analysed through the server and cannot be accessed directly. The reasons are legalistic. With access restricted in this way, the complications of IPR and copyright have been neatly sidestepped. The price is that these complications will eventually have to be confronted in the future, and in fact META is in the process of formulating a set of licence agreements to suit the distribution of resources, like MLRS.

In this report, we have tried to convey the paradoxical state of Maltese Language Technology. The paradox arises because there are significant efforts made by a small number of well qualified people across a spectrum of LT-related activities to improve the state of the art, whether this be in terms of tools, or resources, or both. It is also clear that within the wider context of educational, commercial and cultural activities in the country, there is a place for LT to make an important contribution. The problem is that efforts that have been made are uncoordinated, short term, and fragmentary, so progress is slower than it has to be.

Sustained and directed coordination of effort is, in our opinion, the only way in which the benefits of LT for Maltese will be realised in a reasonable time. We believe that even in a country as small as Malta, the work needs to be shared out amongst different stakeholders. We must arrive at a workable roadmap via a localised version of the tripartite division of labour advocated by META: identification of a community with a shared vision; extension of an infrastructure to facilitate the sharing of resources, and reinforcement of connections between LT and neighbouring fields of research and development.

Bibliography

Acquilina Guze, Maltese-English Dictionary, Valletta: Midsea Publications, 1987.

Attard, D., A Lexicon Server Toolkit for Maltese, Dept CSAI, University of Malta, 2005

Bajada, J., Investigation of Translations Equivalences from Parallel Texts, Dept CSAI, University of Malta, 2004

Bajada, J., Phrase Extraction for Machine Translation, MSc, Dept CSAI, University of Malta, 2009.

Buhagiar, I, and Micallef, P., Web Based Maltese Language Text to Speech Synthesiser, Proceedings of WICT08, University of Malta, 2008

Bonnici, S., A Hypertext Implementation of a Dictionary for the Maltese Language, Dept. CSAI, University of Malta, 1997

Borg, M., K Bugeja, C Vella, G Mangion, C Gafa, Preparation of a free-running text corpus for Maltese concatenative speech synthesis, GĦILM 3rd Conference on Maltese Linguistics, Valletta, 2011.

Calleja, S., Speech Synthesis, MSc, Dept CSAI, University of Malta, 2002.

Camilleri, R., Speech Annotation System, Roberta Camilleri, B.Eng Project, Dept Computer Communications Engineering, M.Sc Disssertation, 2010.

Comrie, B., Fabri, R., Mifsud, M., Stolz, T., & Vanhove, M. (Eds.). Introducing Maltese Linguistics. Proceedings of the 1st International conference on Maltese Linguistics (Bremen/Germany, October, 2007). Studies in Language Companion Series. Amsterdam ; Philadelphia: John Benjamins, 2009

Dalli, A., Computational Lexicon for Maltese, MSc, Dept CSAI, University of Malta, 2001.

Galea, David, A System for the Analysis of Maltese Verbs, Dept CSAI, University of Malta, 1999.

Farrugia, Paulseph, An Automatic Translation System for Maltese/English, Dept CSAI, University of Malta, 1999.

Farrugia, A., MULTIMORPH: A Computational Analysis of the Maltese Broken Plural, Dept CSAI, University of Malta, 2008

Farrugia, C., A Portal for Acquisition, Representation and Presentation of Current Affairs in Malta, Dept CSAI, University of Malta, 2009

Farrugia, R, SAMILS – A Semi-Automatic Machine Indexing for Legal Systems, Dept CSAI, University of Malta, 2000

Felter, P. an Optical Character Recognition System for Maltese, Dept CSAI, University of Malta, 2001

Mangion, Gordon, Spelling Correction for Maltese, Dept CSAI, University of Malta, 1999.

Micallef, P., "A Text to Speech Synthesis System for Maltese", Ph.D. Thesis, University of Surrey, December 1997.

Mizzi, R., The Development of a Statistical Spell Checker for Maltese, Dept CSAI, University of Malta, 2000.

Psaila, A. Speech Annotation using Hidden Markov Models, Dept Computer Communications Engineering, M.Sc Disssertation, 2008

Rosner, M., J. Caruana, and R. Fabri. Maltilex: A computational lexicon for Maltese. In M. Rosner, editor, Computational Approaches to Semitic Languages: Proceedings of the Workshop held at COLING-ACL98, Université de Montréal, Canada, pp 97–105, 1998.

Rosner, M., R. Fabri, D. Attard, and Albert Gatt. Maltese language resource server. In Proceedings of CSAW06, University of Malta, pages 90–98, November 2006.

Rosner. M., Electronic language resources for Maltese. In Proceedings of Bremen Workshop on Maltese Linguistics, University of Bremen, 2008.

Sciriha, L., The Maltese Interactive Picture Dictionary, Protea Textware, ISBN 978-0958733007, 1997.

Ussishkin, A., Francom, J.,Woudstra, D., Creating a Web-based Lexical Corpus and Information-extraction Tools for the Semitic Language Maltese, Proc SALTMIL Workshop, Donostia, ISBN: 978-84-692-4940-6September 2009.

Vella A. and P. Farrugia, MalToBI – Building an Annotated Corpus of Spoken Maltese, 2006.

Vella, A., On Maltese Prosody or The Intonational Phonology of Maltese, Proc. 2nd International Conference of, Maltese Linguistics, Bremen, Oct 2007

Vella, G., Automatic Summarization of Legal Documents, Dept. CSAI, University of Malta, 2010

About META-NET

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

Language Users
Society
Research Communities
Industries

The Multilingual Europe Technology Alliance (META)

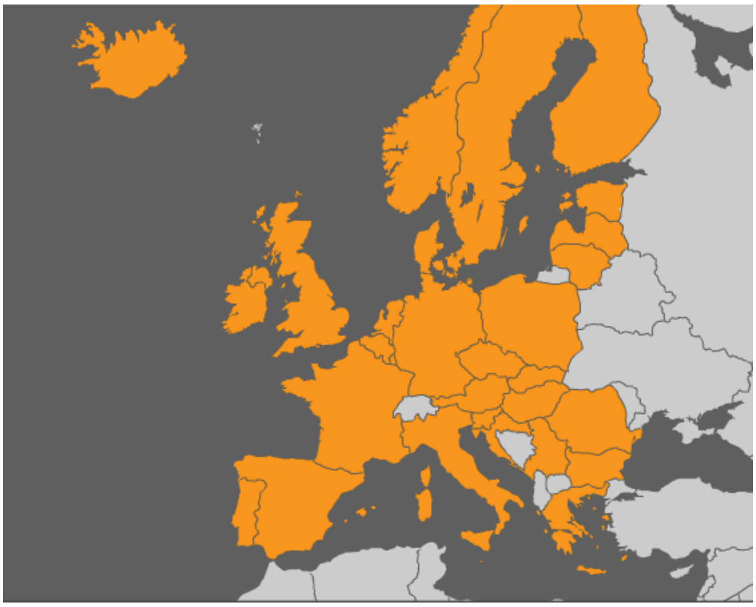


Figure 1: Countries Represented in META-NET

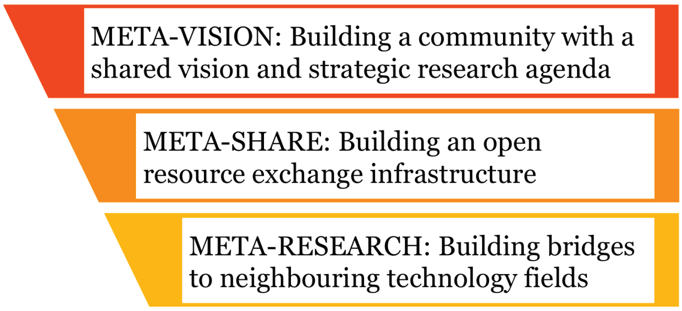
META-NET cooperates with other initiatives like the Common Language Resources and Technology Infrastructure (CLARIN), which is helping establish digital humanities research in Europe. META-NET fosters the technological foundations for the establishment and maintenance of a truly multilingual European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

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

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

Lines of Action

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



META-VISION: Building a community with a shared vision and strategic research agenda

META-SHARE: Building an open resource exchange infrastructure

META-RESEARCH: Building bridges to neighbouring technology fields

Figure 2: Three Lines of Action in META-NET

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

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

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

Member Organisations

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

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

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

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

References

[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).

[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).

[5] The authors are indebted to Prof. Ray Fabri, whose forthcoming article, as cited in the bibliography, was the inspiration for much of the content and many of the examples in this section.

[6] http://www.nso.gov.mt/statdoc/document_file.aspx?id=2840

[7] l-Orizzont from September 7th, 1995; reproduced in Ambros 1998: p. 135

[8] http://www.kunsilltalmalti.gov.mt/

[9] http://www.akkademjatalmalti.com/

[10] http://www.fb10.uni-bremen.de/ghilm/default.aspx

[11] http://www.nso.gov.mt/statdoc/document_file.aspx?id=2651

[12] http://europa.eu/rapid/pressReleasesAction.do?reference=IP/11/556&format=HTML&aged=0&language=EN&guiLanguage=en

[13] "Maltese language hardly used on the internet", The Times of Malta, 16/05/2011.

[14] http://www.gov.mt/

[15] http://www.nettv.com.mt/

[16] http://www.one.com.mt

[17] http://www.rtk.org.mt/

[18] http://www.pbs.com.mt/

[19] http://www.radio101.com.mt/

[20] http://eur-lex.europa.eu

[21] http://langtech.jrc.it/JRC-Acquis.html

[22] http://mlrs.research.um.edu.mt/

[23] see http://linux.org.mt/spellcheck

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

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

[26] see http://www.philb.com/cse/malta.htm

[27] see http://www.charonite .com

[28] http://www.let.uu.nl/lt4el/

[29] The higher the score, the better the translation, a human translator would get around 80. K. Papineni, S. Roukos, T. Ward, W.-J. Zhu. BLEU: A Method for Automatic Evaluation of Machine Translation. In Proceedings of the 40th Annual Meeting of ACL, Philadelphia, PA.

[30] renamed "Department of Intelligent Computer Systems (ICS)" in 2009

[31] (Galea (1999), Mangion (1999), Farrugia (1999), Farrugia (2000), Mizzi (2000), Bajada (2004), Attard(2005), Farrugia (2008), Farrugia (2009), Vella (2010),

[32] M. Rosner, R. Fabri, J. Caruana, M. Montebello and others

[33] ILSIENNA/Our Language: Journal of the International Association of Maltese Linguistics (GHILM), Universitätsverlag Brockmeyer, Bochum.

[34] Crimson Wing Ltd

[35] FITA (Foundation for IT Access)

[36] A. Gatt, C. Borg, R. Fabri

[37] M. Rosner