META-NET White Paper Series

Languages in the European Information Society

– Hungarian –

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

Author

Eszter Simon, Research Institute for Linguistics, Hungarian Academy of Sciences

Editor

Piroska Lendvai, Research Institue for Linguistics, Hungarian Academy of Sciences

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?

This whitepaper for the Hungarian language demonstrates that a lively language technology industry and research environment exists in Hungary. Although a number of technologies and resources for Hungarian exist, there are fewer technologies and resources for the Hungarian language than for the English language. The technologies and resources also have a poorer quality.

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

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

A Risk for Our Languages and a Challenge for Language Technology

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.

Hungarian in the European Information Society

General Facts

Hungarian is the most widely spoken non-Indo-European language in Europe. It is the official language of the Republic of Hungary, where ca. 97% of the population of 10 million claims Hungarian as their native language. It is also spoken by Hungarian communities in the seven neighbouring countries. The largest one is an approximately 1.5 million diaspora in Romania. With its 13 million speakers Hungarian is 12th on the list of the most populous European languages. Abroad, Hungarian is an official language in Vojvodina, as well as in three municipalities in Slovenia. Hungarian is officially recognized as a minority or regional language in Austria, Croatia, Romania, Ukraine, and Slovakia. Additionally, emigrant communities use it worldwide, primarily in the United States, Canada and Israel.

It is interesting that Hungarian barely has any major variety: its dialects differ very little from each other and the standard, and spelling is particularly uniform. This may be the result of a long-term neighbourly existence, which – by continuously clashing with other languages – may have launched speakers on a road to standardisation. According to the traditional categorization, there are seven dialects identified in the present area of Hungary. These dialects are, for the most part, mutually intelligible. Two additional Hungarian dialects exist in Romania: Székely and Csángó.

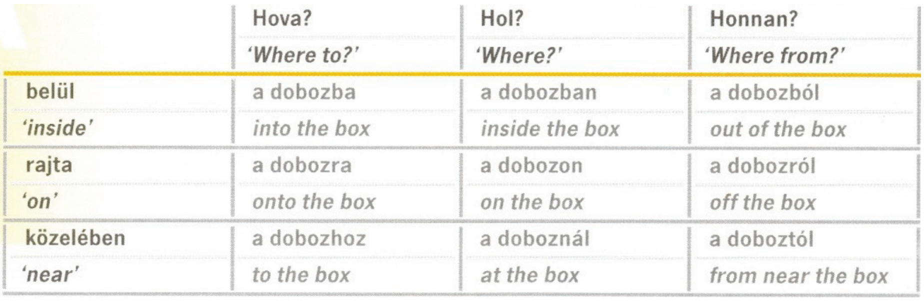
There is scant difference between the Hungarian used in the Republic of Hungary and that used in neighbouring countries. Of course, minor but characteristic differences are present. While the variety in Hungary developed under fundamentally German influence, Romanian Hungarian is more under the influence of Romanian. The Csángó minority group has been largely isolated from other Hungarian people, and they therefore preserved a dialect closely resembling medieval Hungarian.

Particularities of the Hungarian Language

Most European languages belong to the Indo-European family of languages, but not Hungarian! It is a Uralic language, part of the Ugric group, related to Finnish, Estonian and a number of minority languages spoken in the Baltic states and in present-day Russia.

Uralic languages share a few ancient characteristics, such as:

- There is no gender in Hungarian: the same word (ő) expresses the concepts of both ‘he’ and ‘she’.
- There are only two verb tenses: present and past. Their variations and the future tense may be circumscribed.
- The so-called 'direction triad’: there are 3x3 of each set of location cases, as shown by the following table using the word doboz (‘box’) (the article determiner a (‘the’) not being subject to declension):

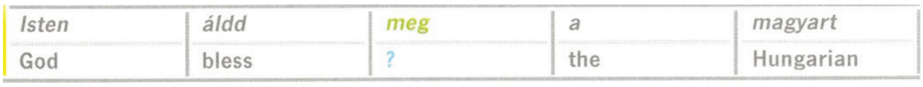


	Hova?	Hol?	Honnan?
	'Where to?'	*'Where?'*	*'Where from?'*
belül	a dobozba	a dobozban	a dobozból
'inside'	*into the box*	*inside the box*	*out of the box*
rajta	a dobozra	a dobozon	a dobozról
'on'	*onto the box*	*on the box*	*off the box*
közelében	a dobozhoz	a doboznál	a doboztól
'near'	*to the box*	*at the box*	*from near the box*

Hungarian is written in Roman letters, nonetheless, Hungarian texts do not resemble any other European language. Below are two lines from a classic poem, in simple literal translation (from Ferenc Kölcsey's 1817 poem *Hymnus*, forming the lyrics of the Hungarian national anthem):

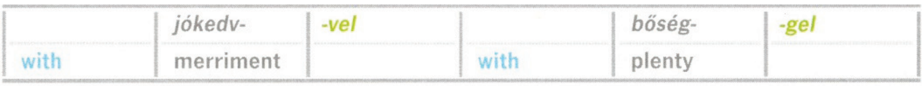
Isten, áldd meg a magyart	*"God bless the Hungarians*
Jókedvvel, bőséggel.	*With merriment and plenty."*

Not a single word is recognisable on the basis of the average European vocabulary; not only do Hungarians refer to God as *Isten*, they do not even call themselves "Hungarian"; they call themselves *Magyar*. But there is more to this than differences in individual words:



Isten	*áldd*	*meg*	*a*	*magyart*
God	bless	?	the	Hungarian

The word denoted with the question mark does not exist in most languages: its name is *igekötő* ('verb-binder', or technically 'verbal prefix'). It plays a multitude of roles here expressing the perfect tense, i.e. indicates a completed action. One of the beauties (and difficulties) of the Hungarian language lies precisely within the usage of verbal prefixes. Now let us examine the second line:



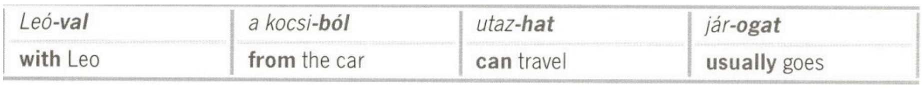
	jókedv-	*-vel*		*bőség-*	*-gel*
with	merriment		with	plenty	

Where English uses the preposition *with*, Hungarian uses suffixes. Hungarian does not have any preposition. In this example it uses the suffixes *–vel* and *–gel* to express what English expresses by means of *with*.

Another important feature of Hungarian is the possessive structure, the reverse of its counterparts in Indo-European languages. For example, in *Paul's radio*, Hungarian does not attach a suffix to the possessor (*Paul*), but rather to his possession, the radio: *Pál rádió-ja*, literally: '*Paul radio-his*'.

It is more of a cultural historical rather than a linguistic curiosity that in Hungarian the family name comes first, with the 'utónév' ('given, name, Christian name') behind, thus the regular order is Liszt Ferenc (=Franz Liszt), Bem József (=Józef Bem), Bartók Béla, Márai Sándor, etc.

Hungarian is called a "synthetic" language: for the most part, it expresses grammatical elements in a single word with affixes, as opposed to "analytical" languages, which prefer to use separate words – prepositions, pronouns, auxiliaries. For example, the Hungarian equivalent of the English auxiliary *can* is the suffix –*hat/-het*.



*Leó-**val***	*a kocsi-**ból***	*utaz-**hat***	*jár-**ogat***
with Leo	**from** the car	**can** travel	**usually** goes

Suffixes, often multiple ones, must be attached to the word stem in strict order, so words can grow to stunning lengths. This type of synthetic word formation is called agglutination (meaning 'gluing of words'). For example:

> *bolondozhattunk "we could fool [around]" (='fool-verb-can-past-we')*
>
> *ösztönözhettünk "we could stimulate" (='stimulus-verb-can-past-we')*

The structure of the two words is identical – the apparent difference is caused by the vowels, due to the so-called vowel harmony (otherwise known as assimilation). The vowels are relegated into one of two classes: "deep": *a o u*, and "high": *e i ö ü*. In the suffixes the vowel appears to fit the stem: *bolond* is deep, thus the other vowels are also deep: *o - o + o - a - u*, while *ösztön* is high, therefore the other vowels are high as well: *ö - ö + ö - e - ü*.[5]

Recent developments

In a way Hungarian has always been a minority language and continuously adopted words from other peoples. There have always been people of various other languages in the Carpathian basin: Slavs (primarily Slovaks, Serbs, Croats) and later Germans, Romanians, Jewish and Roma populations. Latin was used as the official language as late as the beginning of the 19th century, being the language of public administration and science. The Hungarian Parliament introduced legislative sessions in Hungarian only from 1844 onward.

Hungarian was always more of an importer than an exporter. The current vocabulary contains a number of words derived from Slavic, Latin, Romanian, and Italian. The German influence was the strongest, since Hungary was a part of the Habsburg Empire for 400 years. There are a vast number of words of German origin, including *tánc* 'dance' and *hering* 'herring'. Lexical borrowing continues to this day: from French *fritőz* 'friteuse', *bagett* 'baguette'; from Italian *maffiózó* 'mafia member', *paparazzi*; from English *fitnesz* 'fitness', *szerver* 'server' etc. Nowadays loan words are usually anglicisms – due to the strong influence of American films, popular music, and technology (including the Internet).

Language cultivation in Hungary

The Balassi Bálint Institute, founded on January 1, 2002, was launched to promote Hungarian language culture, similarly to the well-known British Council and Goethe Institute. The Balassi Institute contributes to the teaching of Hungarian language and Hungarian studies for non-Hungarians living in Hungary. It also performs cultivation of the Hungarian language and education of Hungarians living beyond the borders of Hungary, and participates in the linguistic and terminological follow-up training of teachers of Hungarian and other experts beyond the national borders. It organises courses on Hungarian studies and minority rights. In cooperation with the international network of institutions for Hungarian studies, it promotes the education and research of the Hungarian language and cultureabroad.[6]

The Research Institute for Linguistics is among the leading institutions in the field of research on the Hungarian language. It was founded in 1949, and placed under the direction of the Hungarian Academy of Sciences in 1951. Its primary tasks include research in Hungarian linguistics, general, theoretical and applied linguistics, Uralic linguistics, and phonetics. The Institute's tasks include the preparation of a comprehensive dictionary of the Hungarian language, and the maintenance of its archival materials. Its research projects investigate various aspects of Hungarian as well as minority languages in and outside Hungary, and deal with issues of language policy within the framework of the European integration. Further activities include the compilation of linguistic corpora and databases, and the laying of the linguistic groundwork for language technology softwares and applications. Besides, the Institute operates a public counselling service on language and linguistics, and runs the Theoretical Linguistics undergraduate and PhD programmes, jointly with Eötvös Loránd University.[7]

The orthography in Hungary is under strict academic control: the rules of Spelling Committee of Hungarian Academy of Sciences are intended to use. The regulations are not obligatory, but misspellings can certainly cause loss of prestige.

These days many enthusiastic traditionalists argue that the neologisms originating from the English language threaten the Hungarian rather than enrich it. Due to their "language protecting" work the so-called "language law" was ratified in 2002, which demands that every English advertisements and slogans must be replaced by Hungarian equivalents. Additional measures for protecting the status of the Hungarian language have also been taken. For example, a television and radio quota regulating the percentage of music sung in Hungarian was introduced at the beginning of 2011.

Language in Education

From 1844, when Hungarian became the official language of public administration, science and education, elementary school children have the possibility to learn Hungarian. Hungarian became the language of higher education after the Education Reform act of 1868. Diplomas in Hungarian may be earned at numerous institutions of higher education beyond the border through Hungarian universities and colleges: from Nyitra (Nitra, Slovakia) all the way to Újvidék (Novi Sad, Serbia) or Kolozsvár (Cluj-Napoca, Romania).

From the 19th century onward, Hungarian language and literature have played an important role in education. The study of Hungarian is compulsory from age 6 to 18. In elementary school, from age 6 to 10, the teaching requirements are divided into key areas of reading, writing and composition. After age 10 grammar and literature are taught separately.

According to the PISA 2009 study[8] that aims to measure literary reading skills of teenagers, Hungary became the member of countries whose results are not statistically significantly different from the OECD average. The overall reading score in Hungary is comparable with those of Germany, France and the UK.

International aspects

Hungary has a great number of world famous physicists (Ede Teller, Eugene Wigner and Leo Szilard, who contributed to the Manhattan Project), mathematicians (Alfréd Rényi, Paul Erdős, is the latter being the author of the Erdős number), and musicians (Franz Liszt, Béla Bartók). Hungarian scientists have won several Nobel prizes in physics, chemistry, and medicine.

As everywhere in the scientific world, Hungarian scholars likewise face a great deal of pressure to publish in international, English-language journals, leading to a self-perpetuating cycle that promotes the increasing importance of English. The situation is similar in the business world: in many large and internationally active companies, English has become the lingua franca, both in written and oral communication. According to a survey in 2005[9], the number of foreign language speakers in Hungary was below the European average: the percentage of Hungarian people who speak at least one foreign language is 35%.

Language technology can address this challenge from a different perspective by offering services like machine translation or cross-lingual information retrieval, and thus help diminish personal and economic disadvantages naturally faced by non-native speakers of English.

Hungarian on the Internet

In 2009, 61.6% of the people in Hungary were internet users[10]. Among young people, aged 14-17, the proportion is even higher. The internet penetration is below the European average, but it has been increasing permanently since the political change in 1990. According to a European study in 2010, the usage of community pages like Facebook is above the European average – probably due to the pre-existence of a quite popular Hungarian community site named iWiw. The existence of a quite active Hungarian-speaking web community is also mirrored by the fact that the Hungarian Wikipedia is the 19th largest, before more commonly used European languages like Turkish, Romanian or Danish and world languages like Arabic or Korean.

For Hungarian Language Technology, the growing importance of the Internet is important in two ways. On the one hand, the large amount of digitally available language data represents a rich source for analysing the usage of natural language, in particular by collecting statistical information. On the other hand, the Internet offers a wide range of application areas for Language Technology.

The most commonly used web application is certainly Web Search, which involves the automatic processing of language on multiple levels, as we will see in more detail the second part of this paper. It involves sophisticated Language Technology, differing for each language. For Hungarian, this comprises for instance taking into account the different inflectional endings of nouns, adjectives and verbs, and different stem forms like *ló* ('horse, single') and *lovak* ('horses, plural').

Internet users and providers of web content can also profit from Language Technology in less obvious ways, e.g., if it is used to automatically translate web contents from one language into another. Considering the high costs associated with manually translating these contents, comparatively little usable Language Technology is developed and applied, compared to the anticipated need. This may be due to the complexity of the Hungarian language and the number of technologies involved in typical Language Technology applications. In the next chapter, we will present an introduction to Language Technology and its core application areas as well as an evaluation of the current situation of Language Technology support for Hungarian.

Selected Further Reading

Did you know? Educational publication about the Hungarian language published by Balassi Institute, Secretariat of National Anniversaries.

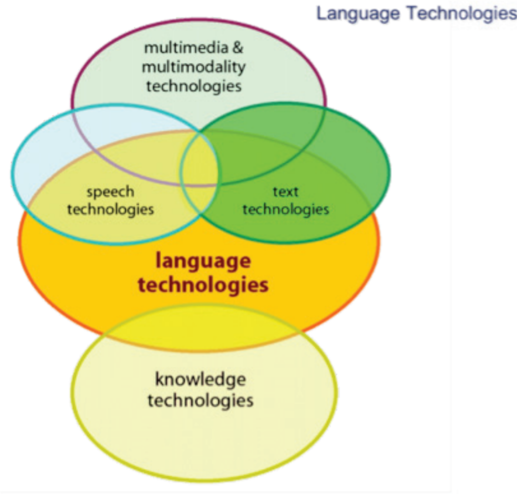
http://en.wikipedia.org/wiki/Hungarian_language

Péter Rebrus, Anna Babarczy: Hungarian descriptive grammar. In: Katalin S. Nagy, István Szakadát (eds.): Média és társadalom: válogatás a Szociológia és Kommunikáció Tanszék Média Oktató és Kutató Központ munkatársainak legújabb munkáiból. (Media and society: selection of new publications of MOKK) Budapest, 2006. pp. 331-381.

Language Technology Support for Hungarian

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, detecting the input language, handling the specific accented letters *(á,é,í,ö,ő,ú,ü,ű)* in Hungarian, etc.
- Grammatical analysis: finding the verb and its objects, modifiers, etc.; detecting the sentence structure.
- Semantic analysis: disambiguation (Which meaning of “apple” is the right one in the given context?), resolving anaphora and referring expressions like “she”, “the car”, etc.; representing the meaning of the sentence in a machine-readable way.

Task-specific modules then perform many different operations such as automatic summarization of an input text, database look-ups and many others.

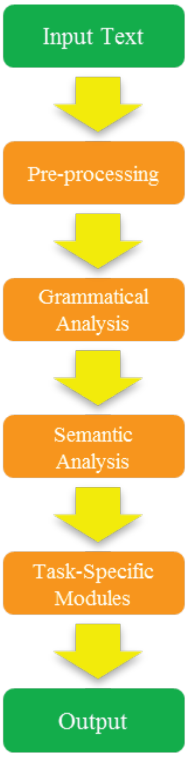


Figure 2: A Typical Text Processing Application Architecture

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 simplyfied and idealised, serving for illustrating the complexity of language technology applications in a generally understandable way.

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

After introducing the core application areas, we will give a short overview of the situation in LT research and education, concluding with an overview of past and ongoing research programs. At the end of this section, we will present an expert estimation on the situation regarding core LT tools and resources on a number of dimensions such as availability, maturity, or quality. This table gives a good overview on the situation of LT for Hungarian.

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):

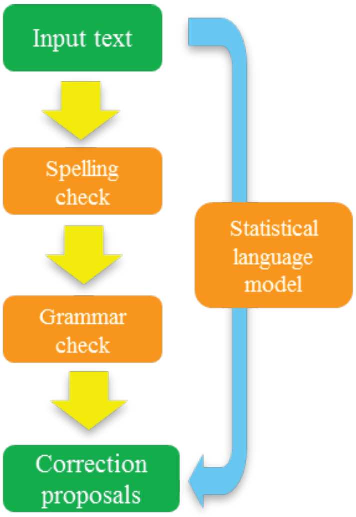


Figure 3: Language Checking (left: rule-based; right: statistical)

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. In Hungarian, there are inflected word forms that can hold more several meanings, e.g.:

várunk$_1$ ('we are waiting')

várunk$_2$ ('our castle')

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, *várunk* is probably not a verb if the sentence contains an other finite verb. A statistical language model can be automatically derived using a large amount of (correct) language data (i.e. a **corpus**). Up to now, these approaches have mostly been developed and evaluated on English language data. However, they do not necessarily transfer straightforwardly to Hungarian with its flexible word order and richer inflection.

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.

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

As Hungarian is a highly agglutinative language, a Hungarian spell checker must contain a morphological analyzer for handling the great number of affixes and complex words. The first spell checker for Hungarian has been developed by combining a spell checking system and a morphological model by a Hungarian SME named MorphoLogic[11] in the late 80s. Their program (*Helyes-e?*) is available for MS Office, QuarkXPress, Adobe InDesign and other desktop publisher packages. MorphoLogic developed grammar and style checkers as well, which recognizes spelling errors that word checkers can not find because they do not analyse the context. The program does not necessarily mark errors, it also gives warnings. Most of the marks indicating actual errors, or drawing the attention to a possible mistake, leaving it to the user to decide whether it is a real mistake.

An open source spell checker for Hungarian also exists as well. Hunspell[12] is based on MySpell, and it has been integrated into OpenOffice, Mozilla Firefox 3, Mozilla Thunderbird and Google Chrome.

Web search

Search on the web, in intranets, or in digital libraries is probably the most widely used and yet underdeveloped Language Technology today. The search engine Google, which started in 1998, is nowadays used for about 80% of all search queries world-wide.[13] The verb *guglizni* is commonly used in Hungarian, even though it has not made its way into printed dictionaries. Neither the search interface nor the presentation of the retrieved results has significantly changed since the first version. In the current version, Google offers a spelling correction for misspelled words and also, in 2009, incorporated basic semantic search capabilities into their algorithmic mix[14], 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.

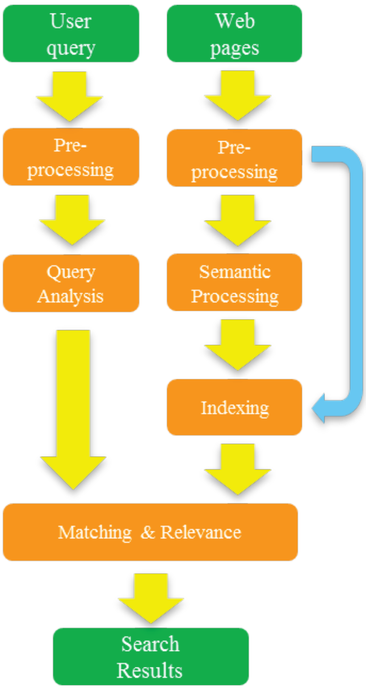


Figure 4: Web Search Architecture

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 similar resources for other languages), have shown improvements by allowing to find a page on the basis of synonyms of the search terms, e.g. *atomenergia*, *magenergia* and *nukleáris energia* (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 recognisers**.

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.

For inflectional languages like Hungarian, it is important to be able to search for all the inflected forms of a word simultaneously, instead of having to enter each different form separately. For this purpose, several morphological parsers exist for Hungarian. NP chunkers for identifying noun phrases provide higher level parsing: a statistical and a rule-based application have been developed for Hungarian.

Due to the variable word order characteristic of Hungarian, we cannot rely on exploiting particular linear configurations alone when syntactic parsers are developed. On the other hand, Hungarian is an agglutinative language with rich case marking, and morphological case markers and postpositions lend themselves to being used as cues for parsing. A database of Hungarian verbs and case markers of their arguments was developed at the Research Institute for Linguistics, which has been built in higher level parsing applications, e.g. for automatic acquisition of verb argument frames, or rule-based syntactic parsing. More syntactic parser for Hungarian exist – one of them was built in the Hungarian treebank (Szeged Treebank) and in a rule-based machine translator (MetaMorpho).

Focus on development for HLT companies and research institutes lies on providing trend- and text-analysis tools which integrate natural language processing tools to find the relevant information in unstructured text. For this purpose part-of-speech taggers, dependency parsers and named entity recognisers have been developed for Hungarian, which are mostly based on statistical learning algorithms.

A meta-search and clustering engine is PolyMeta[15]. It enables organizations and individuals to simultaneously search diverse information resources on the web with a common interface. It employs natural language processing and Information Retrieval algorithms in its query analysis and refinement, search strategy, relevancy ranking, focused drill-down and exploration of multi-dimensional information spaces.

But not only SMEs try to extract information by natural language processing tools. Several projects have been running at universities and research institutes with the aim of developing a model-based semantic search system, creating the framework of a unified Hungarian ontology, or creating a semantically structured, general purpose Hungarian concept set on the basis of the results and formalism of EuroWordNet language ontology (Hungarian WordNet).

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.

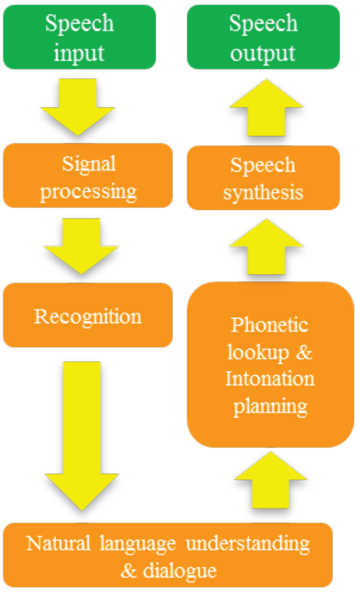

Speech input
Speech output
Signal processing
Speech synthesis
Recognition
Phonetic lookup & Intonation planning
Natural language understanding & dialogue

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.

Due to the specific characteristics of Hungarian, the widely used methods in Speech Interaction technology are difficult or impossible to adapt for Hungarian. However, the methods developed for Hungarian can be applied for similar languages, e.g. Finnish, Turkish, Arabic, in the field of TTS and ASR.

The Hungarian TTS market is dominated by research groups at Budapest University of Technologies[16]. The most widely used TTS system is Profivox, which has been built into SMS- and email-reader softwares, into in-car and mobilephone GPS systems, and into e-book- and screen-reader applications which can help the integration of blind people into information society.

On the Hungarian ASR market there are additional smaller companies, such as Applied Logic Laboratory, Aitia, Digital Natives, as well as research groups, e.g. at the University of Szeged. In spite of the linguistic difficulties mentioned above, more speech recogniser applications for Hungarian have been developed over the last few years. One of them is a prosodic recogniser that was prepared by a cross-lingual study for agglutinative, fixed stressed languages, like Hungarian and Finnish, about the segmentation of continuous speech on word level by examination of supra-segmental parameters. Another system helps the work of doctors: during examining the patient they dictate the diagnosis which will be automatically transcribed. Further application areas are call centers, dialogue systems, or indexing and searching media databases.

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

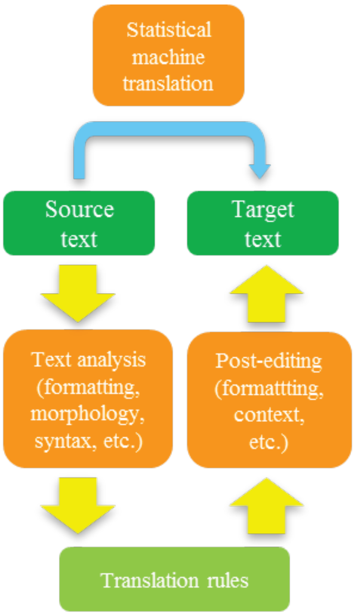

Statistical machine translation
Source text
Target text
Text analysis (formatting, morphology, syntax, etc.)
Post-editing (formattting, context, etc.)
Translation rules

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

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:

> *A rendőr látta az embert a távcsővel.*
>
> *[The policeman observed a man with a telescope.]*
>
> *A rendőr látta az embert a revolverrel.*
>
> *[The policeman observed a man with a revolver.]*

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

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

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

There are knowledge- and data-driven solutions on the Hungarian MT market, as well. MorphoLogic, a private R&D company offers both desktop machine translation programs and online services. The programs translate between English and Hungarian. Their MT system integrates rule-based and statistical methods, but its main component is a parser which creates an intermediary representation, from which it produces the text in the target language.

Availability of large amounts of bilingual texts is really the key in statistical MT. The Hunglish Corpus is a free sentence-aligned Hungarian-English parallel corpus of about 54.2 m words in 2.07 m sentences. At present this is the largest Hungarian-English parallel corpus. The sentence alignment was performed with hunalign, which is one of the most used sentence level aligner, developed by

Hungarian researchers at Budapest University of Technologies. The corpus may be searched through an online sentence search service[17], which can be used as a raw translator or a smart bilingual lexicon.

The iTranslate4.eu[18] project started off in March 2010, which intends to provide online translation solution for all European languages. It does not only offer full coverage of EU languages, but also provides for each language pair the best quality available at the time and mediates easy transfer to professional translators. The project is carried out by a consortium of European MT companies that have developed the best translation system for at least one language pair. The project has two Hungarian participants: the consortium leader is the Research Institute for Linguistics of Hungarian Academy of Sciences, and MorphoLogic provides the common API for the services.

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 summarisation** and **text generation**. Summarisation, 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, summarisation equals sentence extraction: the text is reduced to a subset of its sentences. All commercial summarisers make use of this idea. An alternative approach, to which some research is devoted, is to actually synthesise *new* sentences, i.e., to build a summary of sentences that need not show up in that form in the source text. This requires a certain amount of deeper understanding of the text and therefore is much less robust. All in all, a text generator is in most cases not a stand-alone application but embedded into a larger software environment, such as into the clinical information system where patient data is collected, stored and processed, and report generation is just one of many functionalities.

The identification and classification of named entities serves as a base for several IE applications. Manually annotated Hungarian NE corpora were constructed and more NER systems have been developed and successfully applied to Hungarian and English business news and English clinical texts. In medical documents (e.g. findings or case histories) there is a huge amount of information encoded in free text format. Automated processing of these texts would make these data easily accessible. A Hungarian research group's results in this field involve automatic coding of radiological findings and anonymization of medical documents.

Due to the exponentially growing number of publications, the necessity for automatic information extraction is strong in the biomedical domain as well. A Hungarian research group's activities in this field mainly focus on the disambiguation of biological terms and the detection of expressions containing uncertainty or negation.

For Hungarian, the situation in question answering and text generation is much less developed than it is for English, where these research areas have since the 1990s been the subject of numerous open competitions, primarily those organized by DARPA/NIST in the United States. These have significantly improved the state of the art, but the focus has always been on English; some competitions have added multilingual tracks, but Hungarian was never prominent. Summarization systems, when using purely statistical methods, are often to a good extent language-independent, and thus some research prototypes are available. For text generation, reusable components have traditionally been limited to the surface realization modules (the "generation grammars"); again, most available software is for English.

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. As such, it has not yet acquired a fixed place in the Hungarian faculty system. In Hungary there is no university with an established department of Computational Linguistics. However, programmes are offered by related departments, such as the faculty of computer science or the faculty of linguistics. Some universities offer Master or Bachelor courses only, or modules in Language Technology to students of other courses of study. Many of these programs and courses have only recently been introduced. Currently, six Hungarian universities offer at least courses in the field of Language Technology.

In spite of the efforts in recent years to find the way of regular teaching of CL into the Hungarian faculty system, the education of next generation computational linguists does not achieve the required level. The aim of the Hungarian CL community is to develop a high quality curriculum of BSc-MSc-PhD sequence, which fits into the European standards. The relatively low salaries and scholarships of young researchers pose further problems, which could partly be solved by strengthening the relationships between research and industry.

Language Technology Programs

Similar to other countries, the beginnings of natural language processing in Hungary are connected with Machine Translation. The first attempts were made in the 60s – in those years from Russian to Hungarian. In the 70s-80s the lexicographers' work gave the impetus that led to the development of the first computational morphosyntactic systems for Hungarian. In those years there were no regular nationally financed projects, moreover, Hungary was separated from European support.

After the political change, in the 90s new sections were formed at universities (e.g. the Natural Language Processing Group at Szeged University) and in research institutes (e.g. the Department for Corpus Linguistics at the Research Institute for Linguistics). Since 2000, there has been a significant increase in the number of projects supported by European funds and nationally financed projects, supported mainly by the Fund of the Ministry of Education, or the Agency for Research Fund Management and Research Exploitation.

As a consequence, over the past decade a number of important electronic language resources (dictionaries, corpora, lexical databases) as well as processing resources (spell checking, morphological analyser etc.) have been developed. Activities however have not been synchronized, and not uncommonly similar resources have been developed in different places (e.g. there are at least three morphological analysers for Hungarian). A range of different formalisms or standards have been used in these, which in the majority of cases are either incompatible or difficult to convert from; there is also a lack of documentation and in many cases copyright issues are unclear. Nevertheless, in recent years the international trends of standardisation and uniformization of existing resources have reached Hungary as well. Several projects started off with the objective of integration and interoperability, e.g. creating a unified Hungarian ontology, or harmonizing the different coding systems of separately developed morphological analysers.

In 2008, prominent Hungarian academic institutions and R&D companies formed the Hungarian Platform for Speech and Language Technology[19], which aims to help sharing and integration of high quality knowledge accumulated in centers that worked in isolation beforehand; to work out detailed strategic and implementation plans and to help their subsequent implementation; to disseminate its analyses and proposals among the members of the IT sector; to represent the Hungarian interests on the international level; and to disseminate the achievements of the Platform among the potential users of the technology. Hungarian institutions are also involved in the CLARIN project.

Status of Tools and Resources for Hungarian

The following table provides an overview of the current situation of language technology support for Hungarian. 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 in-

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

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

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

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

Table of Tools and Resources

	Quantity	Availability	Quality	Coverage	Maturity	Sustainability	Adaptability
Language Technology (Tools, Technologies, Applications)							
Tokenization, Morphology (tokenization, POS tagging, morphological analysis/generation)	6	2	4	5	5	1	5
Parsing (shallow or deep syntactic analysis)	3	2	4	4	3	5	4
Sentence Semantics (WSD, argument structure, semantic roles)	1	3	3	1	0	0	3
Text Semantics (co-reference resolution, context, pragmatics, inference)	1	3	2	0	0	0	3
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)	1	0	2	1	0	1	1
Information Extraction (named entity recognition, event/relation extraction, opinion/sentiment recognition, text mining/analytics)	6	6	6	6	5	5	5
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	6	1	5	5	6	5	6
Speech Recognition	3	0	4	2	4	3	3

	Quantity	Availability	Quality	Coverage	Maturity	Sustainability	Adaptability
Speech Synthesis	4	3	4	4	5	3	3
Dialogue Management (dialogue capabilities and user modelling)	0	0	0	0	0	0	0
Language Resources (Resources, Data, Knowledge Bases)							
Reference Corpora	6	6	6	6	6	6	4
Syntax-Corpora (treebanks, dependency banks)	1	6	6	5	6	6	4
Semantics-Corpora	3	6	6	1	3	5	5
Discourse-Corpora	0	0	0	0	0	0	0
Parallel Corpora, Translation Memories	6	4	6	6	6	6	6
Speech-Corpora (raw speech data, labelled/annotated speech data, speech dialogue data)	2	2	4	2	4	4	0
Multimedia and multimodal data (text data combined with audio/video)	1	0	1	1	1	0	0
Language Models	6	3	4	3	6	6	5
Lexicons, Terminologies	5	1	6	6	2	2	6
Grammars	3	3	6	5	6	4	3
Thesauri, WordNets	1	1	6	3	5	5	3
Ontological Resources for World Knowledge (e.g. upper models, Linked Data)	2	6	1	1	1	4	2

Conclusions

The situation of Hungarian concerning language technology support in the last few years gives rise to cautious optimism. Supported by mostly national funds, there exist a language technology research scene in Hungary. The Hungarian NLP market is dominated by research groups at universities and academic institutes, however there are additional smaller companies on the market.

For Hungarian, a number of technologies and resources exist, but far less than for English. The Hungarian human language technology is in a specific situation: it is rather a follower of the international, English-centered technologies, but due to its specific characteristics application of new methods is needed.

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 comaprison and identification of gaps and needs.

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

- While some specific corpora of high quality exist, a very large syntactically annotated corpus is not available.
- There is one syntactically highly elaborately annotated corpus for Hungarian. The corpus is available for free, which results in a wide range applications developed based on the corpus.
- Many of the resources lack standardization, i.e., even if they exist, sustainability is not addressed; concerted programs and initiatives are needed to standardize data and interchange formats.
- Semantics is more difficult to process than syntax; text semantics is more difficult to process than word and sentence semantics.
- The more semantics a tool takes into account, the more difficult it is to find the right data for developing it; more efforts for supporting deep processing are needed.
- Standards do exist for semantics in the sense of world knowledge (RDF, OWL, etc.); they are, however, not easily applicable to NLP tasks.
- The standard preprocessing steps (tokenization, morphology, shallow parsing etc.) are complete for Hungarian, but the more difficult semantics and discourse analysis need further research.
- Research was successful in designing particular high quality software, but it is nearly impossible to come up with sustainable and standardized solutions given the current funding situations.
- There is a large variation in the different areas of NLP in Hungarian: in some fields (e.g. morphology, IE, MT, parallel corpora) there are higher ratings, while in other fields (e.g. advanced discourse processing, dialogue management) ratings are close to zero.
- Speech Recognition and Machine Translation of Hungarian is studied at several universities and workplaces, but free tools and data are not available. It is quite typical in the Hungarian NLP market: the number of free databases and open source programs is quite low.

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

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)

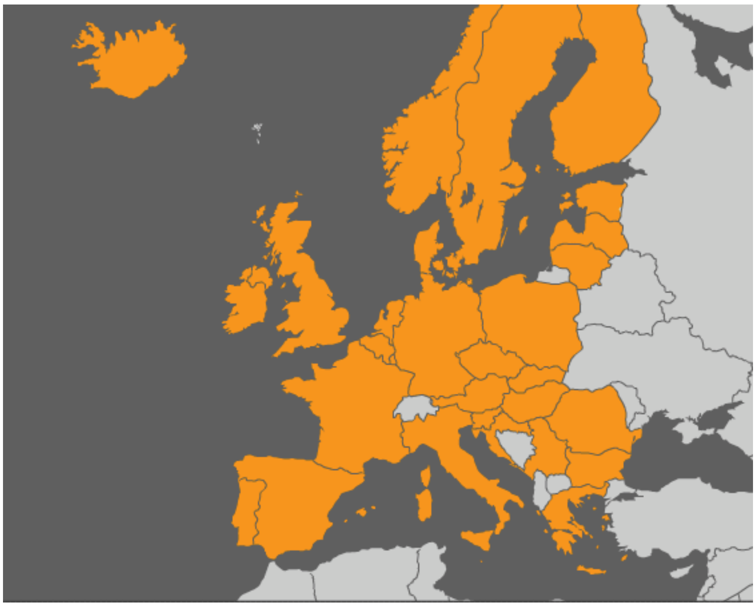


Figure 1: Countries Represented in META-NET

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 University of Edinburgh	Sophia Ananiadou 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] Source: Educational publication about the Hungarian language published by Balassi Institute, Secretariat of National Anniversaries. Text: Ádám Nádasdy

[6] http://www.bbi.hu/index.php?id=99&fid=110

[7] http://www.nytud.hu/eng/index.html

[8] http://www.oecd.org/document/61/0,3343,en_2649_35845621_46567613_1_1_1_1,00.html

[9] http://www.tarki.hu/tarkitekinto/20050412.html

[10]http://www.google.com/publicdata?ds=wb-wdi&met_y=it_net_user_p2&idim=country:HUN&dl=hu&hl=hu&q=internethaszn%C3%A1lat

[11] http://www.morphologic.hu/

[12] http://hunspell.sourceforge.net/

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

[14] http://www.pcworld.com/businesscenter/article/161869/google_rolls_out_semantic search_capabilities.html

[15] http://www.weblib.com/

[16] http://www.tmit.bme.hu/home

[17] http://szotar.mokk.bme.hu/hunglish/search/corpus

[18] http://itranslate4.eu/

[19] http://hlt-platform.hu/