META-NET White Paper Series

Languages in the European Information Society

– Bulgarian –

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

Assoc. Prof. Dr. Diana Blagoeva, Institute for Bulgarian
Prof. Dr. Svetla Koeva, Institute for Bulgarian
Prof. DScs. Vladko Murdarov, Institute for Bulgarian

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 Bulgarian language demonstrates that a lively language technology industry and research environment exists in Bulgaria. Although a number of technologies and resources for Standard Bulgarian exist, there are fewer technologies and resources for the Bulgarian language than for the English language. The technologies and resources are also of poorer quality.

According to the assessment detailed in this report, immediate action must occur before any breakthroughs for the Bulgarian 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 helps people collaborate, conduct business, share knowledge and participate in social and political debates across different languages.

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

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

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

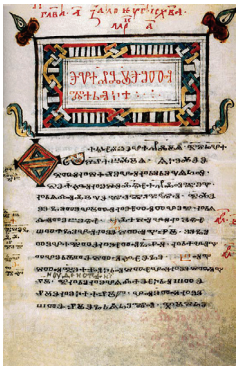
Bulgarian in the European Information Society

General Facts

Bulgarian is the official language of the Republic of Bulgaria.

Bulgarian is spoken by approximately 9 million native speakers, mainly in Bulgaria. but also in Greece, Macedonia, Romania, Serbia, Turkey (Europe), Ukraine, Australia, Canada, USA, Germany and Spain. It is also spoken in Croatia, Czech Republic, Hungary, Israel, Moldova, Romania, Russian Federation (Europe), Slovakia[5]. Bulgarian communities in the neighbouring countries and communities of immigrants around the world are supported by the Government agency for Bulgarians Abroad[6] at ministerial level.

Preliminary data provided by the National Statistical Institute[7] from the population census of Bulgaria indicates that as of the 1st of February, 2011, the population was 7 351 633. The number of Bulgarians living in countries of the European Union will become clear in 2012 when the data from EU populations censuses is available.



The official alphabet is Cyrillic. Bulgarian is the first Slavic language with its own writing system which dates from the 9th century. In 886 AD the Glagolitic alphabet was introduced in Bulgaria. Glagolitic was created by Sts. Cyril and Methodius but was gradually replaced with the Cyrillic alphabet, created at the literary schools of Ohrid and Preslav in the beginning of the 10th century. On the 1st of January 2007 when Bulgaria became a full member of the European Union, Cyrillic became the third official alphabet of the European Union, following the Latin and Greek alphabets.

Bulgarian dialects are those regional, colourful varieties of the Bulgarian language spoken both inside and outside the borders of the country. The main isogloss separating the Bulgarian dialects into Eastern and Western is the "Yat Border" which marks the different mutations of the Old Bulgarian "yat" form (**ѣ**), pronounced in certain conditions as either /'a/ or /e/ to the east (бял [b'al], but plural бели [beli], '*white*') and strictly as /e/ to the west of it (бел [bel], plural бели [beli]).

Particularities of the Bulgarian Language

Bulgarian belongs to the family of South Slavic languages and forms part of the Balkan linguistic union (Balkan Sprachbund). Consequently Bulgarian displays similarities to both language groups. As a Slavic language Bulgarian possesses a rich inflectional and derivational morphology, verb aspect pairs, etc. However, due to the mutual influence of Balkan languages Bulgarian has lost noun cases (except vocative) and completely has lost the infinitive form.

In addition to a common lexicon some of the most important grammatical features of Bulgarian, which distinguish it as a Slavic language, are (some of these features are also shared with other Balkan languages):

- Rich system of inflections. The rich inflectional system poses specific difficulties to LT systems: lemmatisation, for example, might face the notorious problem of recognition of certain inflectional types that can belong to different parts of speech.

Such a case of homography, for instance is the word бели which can be:

- plural of the noun беля '*trouble*';
- plural go the adjective бял '*white*';
- 3rd person singular presence of the verb беля '*peel*';
- 2nd or 3rd person singular aorist of the verb беля '*pee*l';
- 2nd person singular imperative of the verb беля '*peel*'.

- Rich derivational system: diminutive and augmentative nouns: стол - столче ('*chair*' - '*small seat*'); feminine suffixes for nouns relating to persons (professions): учител - учителка ('*teacher*' - '*female teacher*'); etc.
- Aspectual verb pairs: родя - раждам ('to give birth' – 'to be giving birth'.
- A pro-drop language, i.e. subjects are typically not overtly expressed whenever they are inferable from context: Аз чета книга - Чета книга ('*I am reading a book*').

During the historical development of Bulgarian and as a result of its contacts with neighbouring non-Slavic languages in the Balkans, significant changes have occurred when compared with the other Slavic languages. Typical features of the Balkan Sprachbund are:

- Simplification of the nominal case structure - only relics of the accusative and dative cases remain in the system of pronouns and the vocative form in certain masculine and feminine animate nouns or persons.
- Existence of post-fixed definite article in Bulgarian (жена - жената '*a woman*' - '*the woman*'), Albanian, Aroumanian and Romanian. This phenomenon arose most probably under the influence of Old Bulgarian since it does not exist in Modern Greek, where the definite article is prefixed.
- Loss of the infinitive and its replacement with finite verb forms, etc.

Bulgarian orthography is far more transparent than, for example, English orthography. The pronunciation of a word can be easily divined from its written form. Nonetheless, a number of the writing system's rules combined with pronunciation rules (assimilation between voiced and voiceless consonants, reduction of vowels, etc.) lead to potential homophony: кос [kos] '*blackbird*' and коз [kos] '*trump*'. In this way, a number of spellings could represent the same pronunciation. This is a source of potential errors that spell and grammar checkers are able to detect and correct.

Capitalization in Bulgarian is sparse compared with other languages. In general, only personal and place names, some abbreviations (e.g. Стара Загора '*Stara Zagora*'), the first word (only) in the title of a book, movie, song, etc., and the first word in a sentence are capitalized, as are names of companies, government bodies, etc. Names of nationalities or languages are not capitalized, nor are days of the week and months of the year.

The Bulgarian language possesses certain features which can pose significant difficulty for the automatic definition of its syntactic structure. For example, it has a relatively free word order – while the order of adjectives as pre-positive modifiers of nouns or of adverbs as pre-positive modifiers of adjectives and other adverbs is fixed, the order of the subject and object of the verb is relatively loose. Thus in a sentence of three constituents (subject, verb, ob-

ject), all six permutations are possible, in a sentence containing four constituents (subject, verb, direct object, indirect object) - twenty four permutations, and so on. For example, in the following sentence: Дивата котка гони злото куче цяла сутрин *'The wild cat chases the bad dog the whole morning'* it is not clear without a broader context which constituent is the subject - дивата котка '*wild-the cat*' or злото куче '*bad-the dog*'. Unlike other languages which show a relatively loose word order (the other Slavonic languages, for example) Bulgarian does not possess nominal case inflection to indicate the syntactic relations between the words.

Yet another characteristic feature of Bulgarian which poses difficulty for syntactic parsing is the free omission of the subject which, when combined with the possibility of shifting the positions of the subject and object, makes the task even harder. For example the following sentence: Извика Анна '**called Anna*' could mean that Anna called somebody, i.e. Anna is the subject, or that Anna had been called by someone else, i.e. that she is the object (there is no morphological marker to distinguish the subject from the object, except in the case of the graphical representation of the definite article in masculine nouns in the subject position: попита учителят '**asked teacher*' means that the teacher asked something, while попита учителя '**asked teacher*' means that the teacher was asked something.

The rather flexible word order which when combined with the lack of morphological distinction for nominal cases and subject omission is a real challenge for natural language processing of Bulgarian.

Recent developments

Since 1990 American movies and television series have begun to dominate Bulgarian broadcasting. Foreign films and series are either dubbed (mainly by the national television and other bigger television companies, nationally broadcast) or subtitled (mainly at smaller private television companies). However, whatever the way in which they are translated, the strong presence of the American way of life in the media has influenced Bulgarian culture and language. Due to the continuing triumph of English and American music since the 1960s (much more since 1990), Bulgarians were exposed to a lot of English during their adolescence. English quickly acquired the status of a 'cool/hip' language in certain areas, which it has kept up to the present day.

During the last 20 years there has been a noticeable trend towards the "internationalization" of the Bulgarian lexicon as a result of the influence of English. Bulgarian has accepted new words, meanings and collocations originating predominantly from English (so-called Anglicisms), although many others have been taken from other European and non-European languages.

In the Dictionary of New Words in Bulgarian (2010)[8] which records new words, phraseology and terminological expressions from the past 20 years, about 4300 new lexical units have been registered. Of these about a quarter (about 1020) are borrowings from English. There are a number of terminological areas in which the lexicon has developed almost entirely under the influence of English: computer technology and the internet (файл '*file*', сайт '*site*'), finances, economics and business (дилър '*dealer*', брокер '*broker*'), contemporary music (диджей 'dj', техно '*techno*', клип '*clip*'), sport (джогинг '*jogging*', бодибилдинг '*bodybuilding*'). The influx of English borrowings has also been seen in commonly used words - e.g.

тостер '*toaster*', стикер '*sticker*', бодигард '*bodyguard*', and teenage slang.

The dictionary also records over 700 new word meanings, the majority of which have arisen under the influence of English. These are semantic calques, many of which are in the area of computer terminology: мишка '*mouse*', папка '*folder*', гласова поща '*voice mail*' etc. The influence of English in these cases is not always obvious since it typically affects the word senses, rather then formal composition of words.

Many of the new borrowings from English cause difficulties for speakers of Bulgarian. Some are difficult to pronounce – блокбъстър '*blockbuster*', мърчандайзинг '*merchandizing*', while others are difficult to adapt morphologically - some words give rise to uncertainty when used in the plural for example, бодигарди or бодигардове '*bodyguards*', чипсети or чипсетове '*chipsets*'. Older generation Bulgarians who do not speak English find these borrowings hard to understand.

A new phenomenon unknown in Bulgarian before the 1990's is the graphic representation of many foreign words in the Latin alphabet, in particular English borrowings. A particular case in point are abbreviations (such as CV, CD-ROM, PR), as well as widely used words such as internet, or even prefixed component parts such *e-* for electronic. The use of Latin is particularly prevalent in certain specialized areas such as: computer technology, in the names of companies, commercial establishments etc., but also in commonly used language, such as advertising. Some people are worried that the use of the Latin alphabet for writing Bulgarian (called шльокавица [shl'okavitsa] and used mainly in unofficial modes - sms, e-mails, etc.)[9] will somehow affect the quality of spoken and written Bulgarian and eliminate the use of the distinctive Cyrillic alphabet.

The few examples given demonstrate the importance of raising awareness of developments which might lead to the risk of excluding large parts of society from participation in the information society, namely those who are not familiar with English.

Language cultivation in Bulgaria

Bulgarian is the official language in the Republic of Bulgaria, as stated in the Bulgarian constitution. Constituted by decree of the Council of Ministers, the Institute for Bulgarian Language of the Bulgarian Academic of Science is the official institution which observes changes in the Bulgarian language, determines literary norms and reflects these changes in both orthography and speech. Its primary tasks include research in Bulgarian linguistics, general, theoretical, applied and computational linguistics, as well as the preparation of a comprehensive dictionary of the Bulgarian language, and the maintenance of its archival materials. Other research projects investigate Bulgarian dialects in and outside Bulgaria, including issues of language policy within the framework of European integration. Further tasks include the assembly of linguistic corpora and databases, and laying the linguistic groundwork for computational software and applications. Over the past 60 years as a product of these functions the Institute has published the academic journal Български език '*Bulgarian Language*', in collaboration with Bulgarian National Television has produced Език мой '*My Language*' broadcast and has provided information and advice through its Language service. Regular broadcasts on Bulgarian National Radio are also provided.

In parallel with these activities many of the national and local daily newspapers carry articles concerning matters of language: for example the regular column in the Труд '*Trud*' national daily newspaper, written by professors of the University of Sofia. Since 2002 the University of Sofia has published Родна реч '*Native tongue*', a journal whose main aim is to respond to a range of current language issues and assist in raising linguistic culture amongst Bulgarian society.

The first academic spelling dictionary was published in 1983 and reflected the accepted standards of spelling and speech. In 2002 a new spelling dictionary was published which was augmented by new words and certain changes which mirrored the development of the language. As well as dictionaries which serve to define standard language, other spelling and pronunciation dictionaries are also published to reflect standard usage. Numerous periodicals connected with problems of spelling and pronunciation are also published on a regular basis.

Most of these activities are of purely academic interest and are not sufficiently popular among the younger population in particular. Media language seeks to attract and entertain and in many cases deviates from the proscriptions of standard usage. No official quotas regulate the percentage of Bulgarian language music song (even on the National Radio and Television). This should be compared to the regulations for example in France, Hungary, Slovenia. Nevertheless, the current Law on Radio and Television states that Bulgarian National Radio shall set aside for the creation and performance of Bulgarian music and radio drama not less than 5% of the subsidy of the state budget; while Bulgarian National Television shall set aside not less than 10% of the same subsidy for Bulgarian television film production.

The most recent orthographic reform was carried out in 1945. Its aim was to bring the spelling of Bulgarian up to date and consisted of the removal of certain letters related historical orthographic forms. On a number of occasions members of the Bulgarian National Assembly have submitted draft versions of a new Law on Bulgarian Language primarily in the aims of preserving the purity of the language. These have led to long and rather emotional debates. In 2007 the Law on Transliteration was passed in the aims of standardizing the diverse practices of rendering words (personal names) in Latin letters.

The situation of the Bulgarian language is disadvantageous when compared to languages such as French, which is strongly promoted by the global community of French-speaking peoples within the so-called Francophone Union. The wide usage of Language technology can make an important contribution in this area by offering media, internet and mobile communications sophisticated language services - spell and grammar checkers, style correction, dictionary checks for synonyms, and the correct pronunciation of words.

Language in Education

From the 19th century onwards Bulgarian language and literature has had a very important role in the education. According to the legislation in Bulgaria, all education and teaching provided as part of the current state curriculum, from pre-school through to university level, must be in Bulgarian. Special arrangements exist for children whose mother tongue is not Bulgarian. The study of Bulgarian is compulsory for the elementary and secondary school.

In PISA'2009[10] (the capacity to use scientific knowledge, to identify questions and to draw evidence) Bulgaria is in 46th place out of 65 countries. In PISA'2006 Bulgaria is in 43th place out of 55 countries. In PISA'2000 - in 33th place out of 41 countries.

In PIRLS'2006[11] (reading literacy as the ability to understand and use those written language forms required by society) Bulgaria was in 14th place with 547 points. Again the results were lower than in PIRLS'2001 - 4th place with 550 points.

The Bulgarian PISA and PIRLS results can be used as an indicator (or International Benchmark) to determine to what extent international educational standards are satisfied within the National School Curriculum. In 2009 the Bulgarian students demonstrated basic literacy and almost half of them found it difficult to interpret and analyze a text.

The insufficiencies of Bulgarian language teaching in high schools (for example) can be summarised in a number of points: insufficient allocated time – 36 hours (lessons) annually; non-communicative organisation of the teaching process; inadequate content.

According to the most recent State Educational Requirements Bulgarian language teaching is conducted within the framework of a cultural and education study sector - Bulgarian Language and Literature. This sector is traditional within Bulgarian schools and the universities train specialist - middle and high-school teachers in this subject. One of the ways of increasing the effectiveness of Bulgarian language teaching is for it to be focused upon as a specific and important scientific area. Although traditionally seen as a humanitarian discipline, linguistics is concerned with the formulation of rules according to which language units are inter-combined, and is thus close to the sciences.

At university level, there is a general shortage of courses in Bulgarian (at some Universities) that would enable future experts for successful professional communication and appropriate functional literacy.

Language skills are the key qualification needed in education as well as for personal and professional communication. Increasing the volume of Bulgarian language teaching in schools is one possible step towards providing students with the language skills required for active participation in society. Language technology can make an important contribution here by offering so-called computer-assisted language learning (CALL) systems. Such systems allow students to experience language through play; for example by linking special vocabulary in an electronic text to comprehensible definitions or to audio or video files supplying additional information, e.g., the pronunciation of a word.

International aspects

The usage and influence of the Bulgarian language beyond the borders of Bulgaria and its speakers is limited. For many years the Ministry of Education, Youth and Science has held competitions to send Bulgarian university lecturers to work as lectors in Bulgarian language, literature and culture in a number of foreign universities where Bulgarian is studied. The Ministry of Culture established the National Culture Fund which holds regular competitions for the translation of Bulgarian literature into foreign languages. Translations of the works of classical Bulgarian writers, such as Христо

Ботев '*Hristo Botev*' and Иван Вазов '*Ivan Vazor*', into almost all European languages have, of course, been in existence for many years.

Bulgaria has a great number of world famous singers (Николай Гяуров '*Nikolay Gyaurov*', Гена Димитрова '*Gena Dimitrova*', Борис Христов '*Boris Hristov*', Валя Балканска, '*Valya Balkanska*'), authors and artists (Христо Явашев '*Hristo Yavashev*', Доньо Донев '*Donyo Donev*', Юлия Кръстева '*Yuliya Krasteva*', Цветан Тодоров '*Tsvetan Todorov*') among many others. The poet, Пенчо Славейков '*Pencho Slaveikov*' (1866-1912), a central figure in Bulgarian literature, is the only Bulgarian to have been nominated for the Nobel Prize. Outside certain narrow specialised circles overseas where the Bulgarian languages is taught and studied, Bulgarian is an unknown and exotic language.

As everywhere in the scientific world, Bulgarian scientists face a great deal of pressure to publish in visible (usually international) journals, most of which are now in the English language. 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. At the same time 44% of mature Bulgarians do not speak a foreign languages according to research published by the European statistical service, Eurostat[12], in 2009.

Bulgarian has acquired the status as an official administrative language of the European Union on the same basis as English, German and French, since the European Union is based on solidarity and equality amongst its members. Since the 1st January 2007, Bulgarian has been used in the following situations in the context of relations between Bulgaria and the EU:

- The official bulletin containing the rights of citizens and the texts of EU law is published in Bulgarian.
- The Bulgarian authorities are entitled to speak in Bulgarian in Council of the European Union.
- Bulgarian citizens are entitled to use Bulgarian in their correspondence with the European institutions.

The fact of Bulgaria's membership of the European Union together with the idea of unity and diversity, globalisation while preserving national identity, provides a real opportunity for the egalitarian use of Bulgarian together with the major European languages.

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

Bulgarian on the Internet

Bulgarian internet usership in 2009 increased by 31% in comparison with 2007 and already 46% of the total population uses the internet. According to a study by gemiusAudience[13], published in the report "Do you CEE?"[14] Bulgaria is amongst the countries with the highest percentage of internet penetration.

According to data published by internetworldstats.co[15] there are about 3.5 million internet users in Bulgaria, and after the statistics published by Gemius the growth of sites observed by analysts is

almost 10.7% on an annual basis. In 2010 there was a further 5% increase in usership.

In addition to the ubiquitous international web sites, the most popular web sites on the Bulgarian part of the Internet are Bulgarian news portals (dir.bg, gbg.bg news.bg, etc.). Bulgarian Wikipedia as an important source for natural language processing contains app. 117.000 articles, a considerably smaller size than the biggest Wikipedias – English, German and French – but in the number of articles it is in the 34th position[16] among 270 Wikipedias in other languages.

It is often claimed that English dominates computers and the internet, and that those wishing to use either must first learn English. That may have been true in the early days of the technology but lack of English is no longer the barrier it once was. What began as an anglophone phenomenon has rapidly become a multilingual affair. Software has been made capable of displaying many different kinds of script. Many corporate websites now employ multilingual strategies making choice of language a 'user preference'. Machine translation of web content is only a mouse-click away.

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

The most commonly used web application is web search which involves the automatic processing of language on multiple levels, as will be seen in more details in the second part of this paper. It involves sophisticated language technology, differing for each language. For Bulgarian, this comprises, for instance, matching 52 different verb forms associated with a single transitive imperfective verb lemma. But internet users and providers of web content can also profit from language technology in less obvious ways, for example if it is used to automatically translate web contents from one language into another. Considering the high costs associated with manually translating these contents, it may be surprising how little usable language technology is built in compared to the anticipated need.

However, it becomes less surprising if we consider the complexity of the Bulgarian 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 Bulgarian.

Selected Further Reading

Съвременен български език Фонетика. Лексикология. Словообразуване. Йордан Пенчев, Иван Куцаров, Тодор Бояджиев, Издателска къща Петър Берон, София, 1999, 654 с.

(Contemporary Bulgarian language, Phonetics, Lexicology, Word formation. Iordan Penchev, Ivan Kutsarov, Todor Boyadzhiev. Petar Beron Publishing House, Sofia, 1999, 654p.)

Речник на новите думи в българския език. Емилия Пернишка, Диана Благоева, Сия Колковска, Наука и изкуство, София, 2010, 515 с.

(Dictionary of New Words in Bulgarian. Emiliya Pernishka, Diana Blagoeva, Siya Kolkovska. Nauka i izskustvo publishers, Sofia, 2010, 515p.)

Български диалектен атлас. Обобщаващ том. Ч. I-III. Фонетика. Акцентология. Лексика. (авторски колектив). Книгоиздателска къща Труд. София, 2001, 538 с.

(Bulgarian Dialect Atlas. Summarise volume 4. I-III. Phonetics. Accentology, Lexicon. (Authors' collective) Trud publishing house. Sofia, 2001, 538 p.)

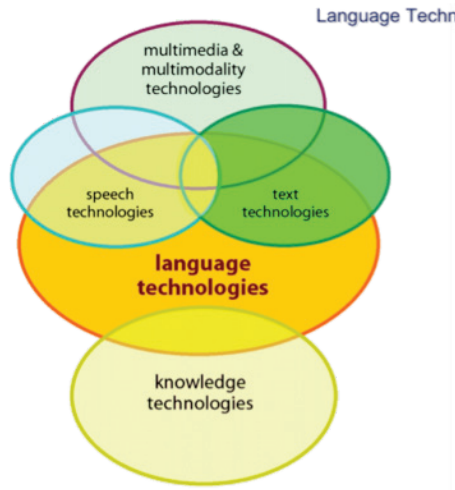
Българите, книжовността, езикът XIX – XX век. (авторски колектив) състав. В. Мурдаров. София, ЕМАС, 2008, 389 с.

(Bulgarians, Literacy, Language XIX-XX century. (Authors' collective), compiled V. Murdarov. Sofia, EMAC, 2008, 389 p.)

Language Technology Support for Bulgarian

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, etc.
- Grammatical analysis: finding the verb and its objects, modifiers, etc.; detecting the sentence structure.
- Semantic analysis: disambiguation (Which meaning of *bank* is the right one in a given context?), resolving anaphora and referring expressions like *she*, *the car*, etc.; representing the meaning of the sentence in a machine-readable way

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

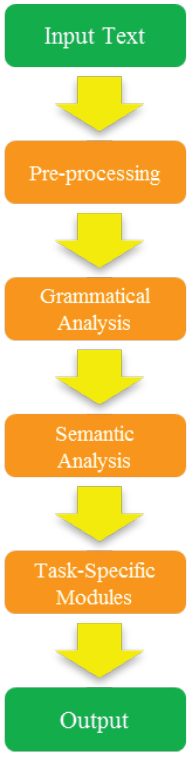


Figure 2: A Typical Text Processing Application Architecture

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 Bulgarian.

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 Bulgaria.

Core application areas

Language Checking

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

Eye have a spelling chequer,

It came with my Pea Sea.

It plane lee marks four my revue

Miss Steaks I can knot sea.

For handling this type of error, context analysis is needed in many cases, e.g., to decide if a word needs to be written in upper case, as in:

Тя живее в Стара Загора.

[She lives in Stara Zagora.]

Тя е стара жена.

[She is an old woman.]

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). 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 Bulgarian with its flexible word order, subject pro-drop 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

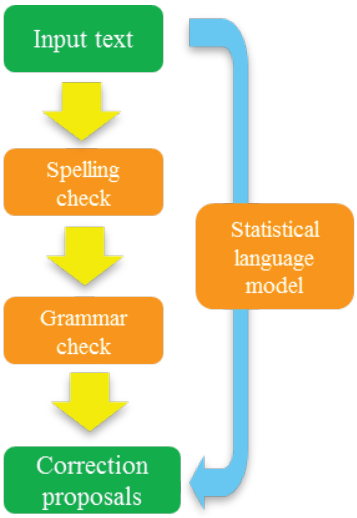


Figure 3: Language Checking (left: rule-based; right: statistical)

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.

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[17]. 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[18], 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.

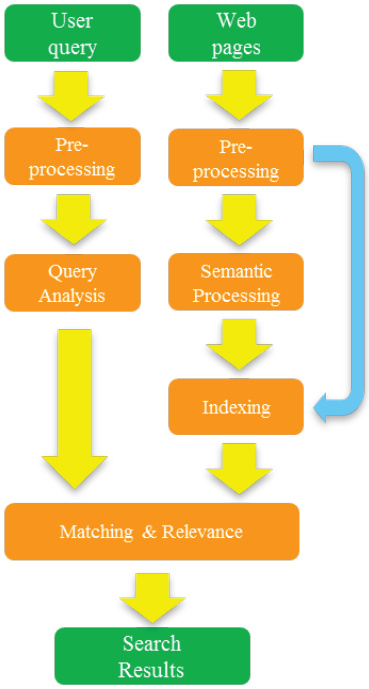


Figure 4: Web Search Architecture

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 Bulgarian BulNet), have shown improvements by allowing to find a page on the basis of synonyms of the search terms, e.g. атомна енергия and ядрена енергия (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 Bulgaria Jabse.com is a search engine (Jabse is an acronym meaning: Just Another Bulgarian Search Engine). It indexes sites from the Bulgarian internet space based on open-source codes. Jabse possesses its own evaluation system to define the importance of pages and terms contained therein on the basis of a range of criteria. Certain Bulgarian portals have crawler software similar to those used by global search engines designed to index sites included within its categories. For example, Dir.bg, one of the first and largest web portals in Bulgaria launched a standalone service – Diri.bg. Diri (in Bulgarian дири) is an old word for '*search*'. This new service is in direct competition with the existing Jabse and it is still to be seen whether Jabse or Diri.bg will develop sufficiently to become a significant factor in the Bulgarian internet space to rival Google. Open source based technologies like Lucene and SOLr are often used by search-focused companies to provide the basic search infrastructure. Other search-based companies rely on international search technologies like, e.g., FAST or Exalead.

Developmental focus for these companies lies in providing add-ons and advanced search engines for special-interest portals by exploiting topic-relevant semantics. Due to the still high demands in processing power, such search engines are only economically usable on relatively small text corpora. Processing time easily exceeds that of a common statistical search engine as, e.g., provided by Google by a magnitude of thousands. These search engines also have high demand in topic-specific domain modelling, making it not feasible to use these mechanisms on web scale.

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.

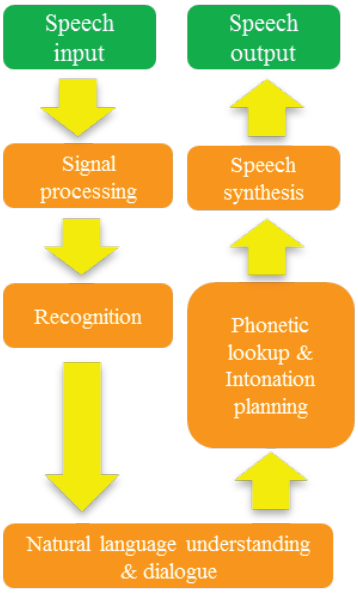


Figure 5: Simple Speech-based Dialogue Architecture

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

On the Bulgarian TTS market, there are a few Bulgarian text-to-speech systems. One of these is SpeechLab 2.0 provided free-of-charge to computer users with visual disabilities. Only a few companies such as Сиела '*Siela*' – a Bulgarian publisher of legal literature - have developed their own system for Bulgarian speech recognition. Finally, within the domain of Speech Interaction, a genuine market for the linguistic core technologies for syntactic and semantic analysis does not exist yet.

As for the actual employment of VUIs, demand in Bulgaria has strongly increased within the last 5 years. This tendency has been driven by end customers' increasing demand for customer self-service and the considerable cost optimisation aspect of automated telephone services, as well as by a significantly increased acceptance of spoken language as a modality for man-machine interaction.

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 in 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.

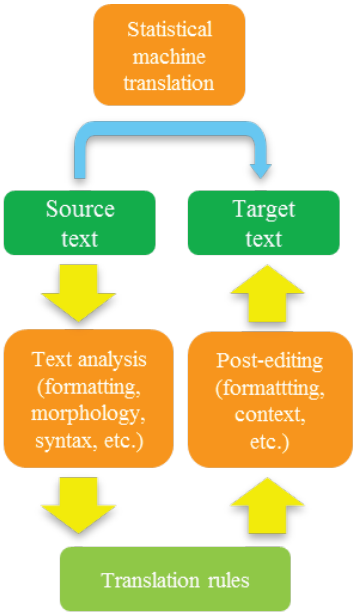


Figure 6: Machine translation (top: statistical; bottom: rule-based)

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:

Полицаят наблюдава престъпника с телескопа.

[The policeman observed the man with the telescope.]

Полицаят наблюдава престъпника с пистолета.

[The policeman observed the man with the revolver.]

One way of approaching the task is based on linguistic rules. For translations between closely related languages, a direct translation may be feasible in cases like the example above. However, 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 the methodologies of both. This can be done in several ways. One is to use both knowledge- and data-driven systems and have a selection module decide on the best output for each sentence. However, for longer sentences, no result will be perfect. A better solution is to combine the best parts of each sentence from multiple outputs, which can be fairly complex, as corresponding parts of multiple alternatives are not always obvious and need to be aligned.

For Bulgarian, MT is particularly challenging. The lack of noun case inflection; free word order and subject pro-drop pose problems for analysis. Extensive inflection in verb morphology is a challenge for generating words with proper markings.

One of the good examples is WebTrance by SkyCode – a machine translation system which automatically translates texts, help files, menus, windows and internet pages from English, German, French, Spanish, Italian and Turkish into and from Bulgarian. Meaning-based translation, rather than word-for-word translation, is a challenge for many people studying a foreign language. The aim of WebTrance is to provide meaningful translation of texts. Provided good adaptation in terms of user-specific terminology and workflow integration, the use of MT can increase productivity significantly.

The quality of MT systems is still considered to have huge improvement potential. Challenges include the adaptability of the language resources to a given subject domain or user area and the integration into existing workflows with term bases and translation memories. In addition, most of the current systems are English-centred and support only few languages from and into Bulgarian, 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 score[19].

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 sub-tasks 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.

For Bulgarian, the situation in all these research areas is much less developed than it is for English, where question answering, information extraction, and summarization 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 Bulgarian was never prominent. Accordingly, there are hardly any annotated corpora or other resources for these tasks. Summarization systems, when using purely statistical methods, are often to a good extent language-independent, and thus some research prototypes are available. For text generation, reusable components have traditionally been limited to the surface realization modules (the "generation grammars"); again, most available software is for English.

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 Bulgarian faculty system. In Bulgarian universities courses in computational linguistics are only partially available. They are usually designed either for humanities students or mathematicians but not both. Two years ago the University of Plovdiv began to offer a Bachelors programme in Linguistics with Information Technologies. 30 students enrolled in 2009-2010 and twice that number in 2010-2011. Students are offered a wide range of courses connected with the fundamentals of linguistics, mathematics and programming as well as language technology applications. The Bachelors degree in Informatics at the University of Plovdiv traditionally offers a lecture course in Computational Linguistics.

Since 2004 the Faculty of Mathematics and Informatics of the University of Sofia has been offering a Masters programme in Artificial Intelligence. The University of Sofia also offers a Masters programme in Computational Linguistics. The programme includes subjects from the sphere of mathematics, logic, programming and theoretical linguistics. Basic courses in computer linguistics are also offered such as statistical methods for text processing, machine translation, semantic networks and ontology, etc. Graduates possess a good grounding to begin academic research in the area of computational linguistics as well as broader computational and linguistic literacy allowing them ability to develop unconventional practical applications. The success of the programme is evident from the professional development of the students from the previous intakes. They are employed in leading IT companies, popular national and specialised media and particularly in academic research.

Language Technology Programs

The first international and national funding supporting language technologies for Bulgarian began at the very beginning of 1990s. Over a short period of time financing for a number of research projects from European institutions was got: LaTeSLav (1991-1994) – aimed at developing a prototype of a grammar checker, BILEDITA (1996-1998) – for the development of bi-lingual electronic dictionaries, GLOSSER (1996-1998) – aimed at supporting foreign language training and others. The Multext-East (1995-1997) extension of the previous Multext and EAGLES EU projects provided the Bulgarian language resources in a standardized format with standard mark-up and annotation, and these resources were later expanded and upgraded in the ELAN (European Language Activity Network) 1998-1999), TELRI I in II (Trans European Language Resources Infrastructure 1995-1998 / 1999-2001) and Concede (Consortium for Central European Dictionary Encoding 1998-2000) projects.

The Ministry of Education, Youth and Research through the National Scientific Fund (NSF), has supported research through national research programs. These programs have impelled numerous research projects and collaboration with international research centres and companies. The basis of technology development and commercial applications for automated processing of the Bulgarian language has been partly created as a result of these projects.

A number of years ago five Bulgarian academic institutions founded a consortium to create and develop an integrated national academic infrastructure for language resources. Bulgarian institutions are also involved in the CLARIN project. Other ongoing projects include those comprised by EUROPEANA aimed at developing the basic technologies and standards necessary to make knowledge on the Internet more widely available in the future. In addition to many other smaller-scale funded projects, the abovementioned projects have led to the development of competences in the field of Language Technology as well as a basic technological infrastructure of language tools and resources for Bulgarian. However, public funding for LT projects in Bulgaria is dramatically lower than that for comparable projects in Europe, as well as in comparison to investments into areas such as language translation and multilingual information access by the USA[20].

Availability of Tools and Resources for Bulgarian

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

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

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

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 Bulgarian, key results regarding technologies and resources include the following:

- For morphologically related tools such as tokenizers, part of speech taggers and morphological analyzers, the situation in Bulgaria is reasonably good. Even if the tools are not all freely available, the resources are of relatively high quality and the coverage is good.
- With regard to resources such as reference corpora, lexicons, and wordnets, the situation is also reasonably good for Bulgarian since substantial resources have been developed in

recent years. While some reference corpora of high quality and quantity exist, i.e. the Bulgarian National Corpus, large syntactically and semantically corpora annotated by experts are not available.

- Semantics is more difficult to process than syntax; text semantics is more difficult to process than word and sentence semantics. Semantic tools and resources are scored *very* low. Thus, programs and initiatives are needed to substantially boost this area both with regard to basic research and the development of annotated corpora.
- There also exist individual products with limited functionality in subfields such as speech synthesis, speech recognition and machine translation, and a few others.

There are insufficient parallel corpora for machine translation. Translation of Bulgarian into and from another language works best since most data exists.

- There is a huge gap in multimedia data.
- Several of the resources lack standardization, i.e., even if they exist, sustainability is not supported; concerted programs and initiatives are needed to standardize data and interchange formats.

To sum up, the results indicate that Bulgarian stands reasonably well with respect to the most basic language technology tools and resources, such as tokenizers, PoS taggers, morphological analyzers, reference corpora. Furthermore, some tools do exist for word sense disambiguation, machine translation, as well as resources like parallel corpora, and specialized corpora. However, these tools and resources are rather simple and have a limited functionality for some of the areas. For instance, parallel corpora only exist for very few language pairs and for limited text genres. When it comes to more advanced fields such as text semantics, language generation, and annotated multimodal data, Bulgarian clearly lacks the basic tools and resources even if some of these are currently under development.

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 National Scientific Fund of the Ministry of Education, Youth and Science.

As a consequence over the past decade a number of important electronic language resources (dictionaries, corpora, lexical data bases) as well as programmes for their processing (word sense disambiguation tool, spell checking, etc.) have been developed.

In general, it can be stated that over the last two decades language technology for Bulgarian has not been supported by a consistently devised national funding scheme. The process of development of HLT applications, tools and resources for Bulgarian has been, therefore, a mixture of international projects extending their scope from Western European languages to Central and Eastern Europe, also with a view to the EU enlargement process, national research funding, and the enthusiasm of researchers involved in LT.

From this it is clear that more efforts need to be directed towards the development of resources for Bulgarian as well as into research, innovation, and development. The need for large amounts data and the high complexity of language technology systems make it man-

datory that new infrastructures for sharing and cooperation are also developed.

It is also to be hoped that Bulgaria's participation in CESAR and META-NET will make it possible to develop, standardise and make available several important LT resources and thus contribute to the growth of Bulgarian language technology.

About META-NET

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

The Multilingual Europe Technology Alliance (META)

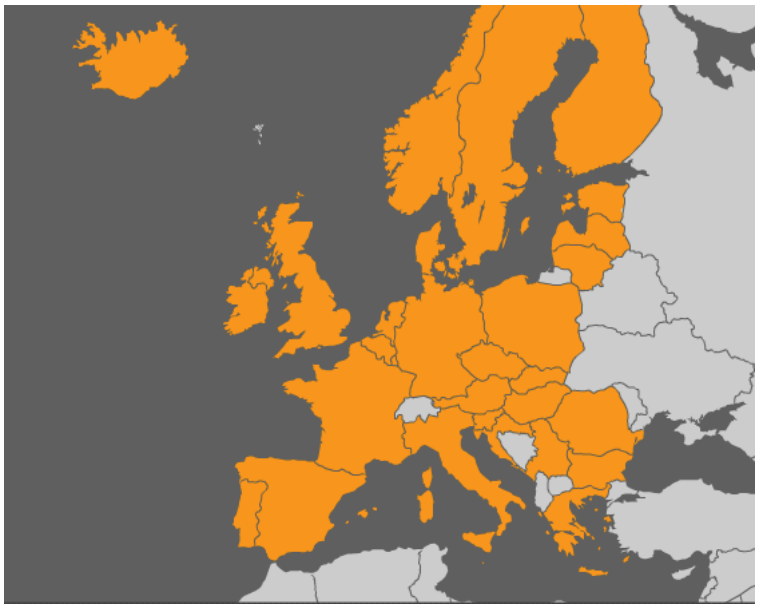


Figure 1: Countries Represented in META-NET

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.

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 collect-

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

Member Organisations

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

Country	Organisation	Participant(s)
Austria	University of Vienna	Gerhard Budin
Belgium	University of Antwerp	Walter Daelemans
	University of Leuven	Dirk van 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] http://www.ethnologue.com/show_language.asp?code=bul

[6] http://www.aba.government.bg/?show=english

[7] http://www.nsi.bg/census2011/pagebg2.php?p2=36&sp2=37

[8] Dictionary of New Words in Bulgarian (2010)

[9] http://bg.wikipedia.org/wiki/Шльокавица

[10] http://www.oecd.org/document/61/0,3746,en_32252351_46584327_46567613_1_1_1_1,00.html

[11] http://nces.ed.gov/surveys/pirls/

[12] http://epp.eurostat.ec.europa.eu/portal/page/portal/eurostat/home/

[13] http://www.gemius.com

[14] http://www.internetcee.com

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

[16] Wikipedia metadata: http://meta.wikimedia.org/wiki/List_of_Wikipedias.

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

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

[19] 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.

[20] Gianni Lazzari: „Sprachtechnologien für Europa“, 2006: http://tcstar.org/pubblicazioni/D17_HLT_DE.pdf