META-NET White Paper Series

Languages in the European Information Society

– Slovak –

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, Editors and Contributors

Mária Šimková, Ľ. Štúr Institute of Linguistics, Slovak Academy of Sciences
Radovan Garabík, Ľ. Štúr Institute of Linguistics, Slovak Academy of Sciences
Agáta Karčová, Ľ. Štúr Institute of Linguistics, Slovak Academy of Sciences
Katarína Gajdošová, Ľ. Štúr Institute of Linguistics, Slovak Academy of Sciences
Michal Laclavík, Institute of Informatics, Slovak Academy of Sciences
Jozef Juhár, Technical University of Košice
Karol Furdík, Technical University of Košice
Peter Ďurčo, University of St. Cyril and Methodius in Trnava
Helena Ivoríková, Studia Academica Slovaca, Comenius University
Jozef Ivanecký, European Media Laboratory
Július Zimmermann, Prešov University

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 **3**

A Risk for Our Languages and a Challenge for Language Technology **5**

Language Borders Hinder the European Information Society 5

Our Languages at Risk 6

Language Technology is a Key Enabling Technology 7

Opportunities for Language Technology 7

Challenges Facing Language Technology 8

Language Acquisition 8

Slovak in the European Information Society **10**

General Facts 10

Slovak Dialects 11

Western Slovak dialects *12*

Central Slovak dialects *12*

Eastern Slovak dialects *12*

Particularities of the Slovak Language 13

Slovak Language on the Internet 15

Slovak language as a foreign language 15

Slovak Online *15*

Studia Academica Slovaca – The Centre for Slovak as a Foreign Language *16*

Summer School of Slovak Language and Culture *16*

Courses of Slovak as a Foreign Language *17*

Distance Education *17*

Methodical seminary on Slovak as a foreign language *17*

Educational Programme Slovak as a Foreign Language *17*

Language Technology Support for Slovak **19**

Language Technologies 19

Language Technology Application Architectures 19

Core application areas 20

Language Checking *20*

Web Search *21*

Speech Interaction *23*

Machine Translation *25*

Language Technology 'behind the scenes' 27

Language Technology in Education 28

Availability of Tools and Resources for Slovak 29
Table of Tools and Resources 30
Conclusions 31
Bibliography 33
About META-NET 37
Lines of Action 37
Member Organisations 39
References 42

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 Slovak language demonstrates that a basic language research environment exists in Slovakia, although the technology industry is not really developed. Despite the fact that a small number of technologies and resources for standard Slovak exist, there are fewer technologies and resources for the Slovak language than for the close Czech and Polish languages, and far fewer than for the major EU languages, like English, German or French. The Slovak language technologies and resources also are of a poorer quality.

According to the assessment detailed in this report, sustained and focused action must occur before the Slovak language resources approach in their quality and quantity similar resources existing for other neighbouring languages.

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

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?

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.

Slovak in the European Information Society

General Facts

The Slovak Republic is a country in the Central Europe neighbouring with both Slavic (Czech Republic, Poland, Ukraine) and non-Slavic countries (Hungary, Austria). Its geographic location, mostly mountainous landscape, and historical development caused considerable multiethnic and multicultural character of the country, differentiation of Slovak dialects and subsequent codification of (modern) standard Slovak as a over-regional communication mean as lately as in 1843. Although part of the territory of Slovakia belonged to the historic Great Moravia, where Constantine and Methodius, invited from the Byzantine Empire in the 9th century were spreading the Christian religion and scholarship through Old Church Slavonic and Glagolitic alphabet. Later development of Slovakia and Slovak language was influenced by Latin alphabet and Roman culture. There subsequently occurred several influences that left traces on the Slovak language as well.

The Slovak language belongs – in the Indo-European family of languages, together with Polish, Czech, Lower and Upper Sorbian – to the West branch of Slavic languages. Linguistic, historic, and archaeological sources prove that Slovak developed directly from Proto-Slavic. The Proto-Slavic basis of Slovak was formed in the area between the Carpathians, the Danube, and the Upper Moravia. The Slavonians, predecessors of the Slovaks, came to this area in the 6th century from the south-east. The reconstructed language of the Great Moravian ethnic group, which was divided into dialects but formed a certain cultural form can be regarded as the basis of Slovak. The Slovak language went through fast development in the 10th to 12th centuries (jer vocalisation, disappearance of nasal vowels), and stabilised in the 13th to 15th centuries. In the 16th to 18th centuries, Czech was used as the cultural language in Slovakia, together with several types of cultural Slovak, such as cultural West Slovak, cultural Central Slovak and cultural East Slovak. From the end of the 18th century, attempts at the formation of literary Slovak started. At the end of the 18th century, Anton Bernolák based his codification on cultural West Slovak, but failed to get wide recognition due to changed social and economic conditions. Ľudovít Štúr used Central Slovak as the basis, his idea took hold very soon, and with certain modifications (Martin Hattala, Michal Miloslav Hodža) lasts to these days.

The Slovak language is the official language in Slovak Republic, since May 2004 it is also one of administrative languages of the European Union. Slovak is spoken by 4.5 million inhabitants of Slovakia, more than one million emigrants in the United States, and approx. 300 thousand persons in the Czech Republic. Smaller language groups of Slovaks are situated in Hungary, Romania, Serbia, Croatia, Bulgaria, Poland, United Kingdom, France, Germany, Belgium, Austria, Norway, Denmark, Finland, Sweden, Italy, Switzerland, Netherlands, Cyprus, Russia, Ukraine, Kyrgyzstan, Israel, Canada, South Africa, Argentina, Brazil, Uruguay, Australia, New Zealand, and other countries. Slovaks abroad pertain to different groups: they are descendants of indigenous inhabitants of Slovakia, who moved to other areas of the former Austro-Hungary; descendants of later migrants from Slovakia, living overseas (emigration wave from the late 19th to the mid 20th century); political and economic migrants after 1945, 1948, and 1968 and their de-

scendants; and, finally, mostly young people settled abroad after the year 1990. It is estimated that some 270 000 Slovaks went abroad in the last wave of emigration in the years 2007 – 2008. A special group consists of descendants of the Slovaks, who remained abroad due to political-geographical changes after the year 1918 or the year 1945. At the same time, there are ethnic minorities living in Slovakia (Hungarians, Gypsies, Czechs, Ruthenians, Ukrainians, Germans, Poles, Moravians, Croatians, Bulgarians, Jews), which together account for 14.2% of population of Slovakia.

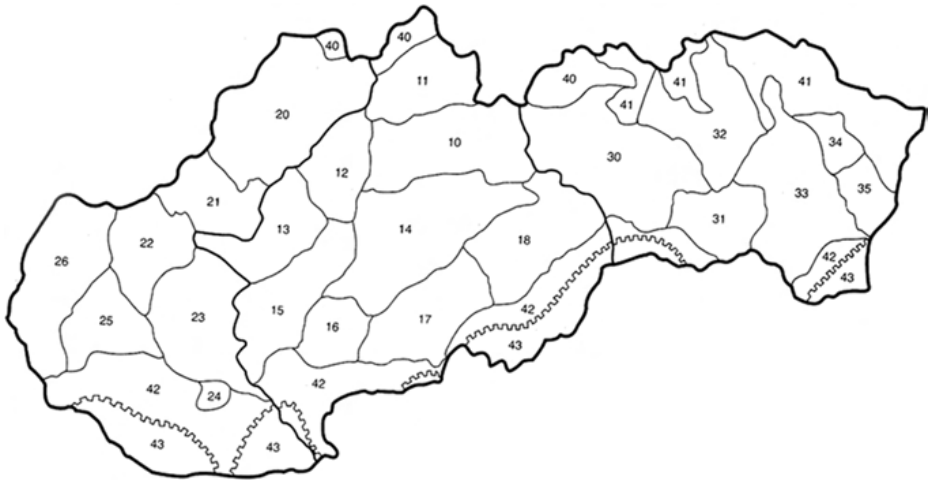
Slovak language has several forms: Standard Slovak is mainly used in written form and in the official communication, colloquial Slovak represents a standard mainly used in verbal communication. Each form has specific subgroups, which form the system of stratification of Slovak language: literary language / nationwide standard language / nationwide substandard language / regional variant / local variant, territorial variant (dialects), social variant (slang, jargon, argot, professional languages). Responsibility for control over language and language policy is borne by the Ministry of Culture (Act on State Language, Central Language Board). Its decisions should be based on knowledge and opinions of the scientific and professional community, led by the Ľudovít Štúr Institute of Linguistics of the Slovak Academy of Sciences. The Institute is a founder and coordinator of several commissions with nationwide coverage: spelling committee, orthoepic committee, onomastic committee, and the committee for codification. The committees prepare and recommend codification of orthoepic, spelling, grammatical and lexical rules. Spelling rules are subject to a broader discussion with involvement of general public, but due to interconnection of many factors and social impact of any changes they are not amended too often. The last amendments, especially in the rules of rhythmic alternation and capitalization, were made in 1991.

Territorial arrangement of Slovakia (the territory with size of almost 50 thousand km² is mainly situated lengthwise; the length between eastern and western borderline is almost 430 km) and specifics of individual dialects affect also forms of Slovak language in specific regions and locations, which represents a problem to be coped with mainly by the foreigners learning Slovak and moving in the territory of the Slovak Republic.

Slovak Dialects

Slovak dialects are a means of communication of the autochthonous population of the respective dialect areas in everyday social and working relations with the nearest environment. Slovak dialects are inherited from one generation to the next in verbal form, although the process of levelling can be observed in this area.

Vocabularies of individual dialects in Slovakia are described in more detail in the Dictionary of Slovak Dialects and several dialects are described in separate studies with extension to other linguistic levels.



Slovak dialects are divided into three basic groups:

Western Slovak dialects

The Western Slovak dialects are spread in the Trenčín, Nitra, Trnava, Myjava areas and other regions.

20. Upper Trenčín dialects

21. Lower Trenčín dialect

22. Váh river dialect

23. Central Nitra dialects

24. Lower Nitra dialects

25. Trnava area dialects

26. Záhorie dialect

Central Slovak dialects

Central Slovak dialects are spoken in the regions of Liptov, Orava, Turiec, Tekov, Hont, Novohrad, Gemer and in the Zvolen area.

10. Liptov dialects

11. Orava dialects

12. Turiec dialect

13. Upper Nitra dialects

14. Zvolen dialects

15. Tekov dialects

16. Hont dialect

17. Novohrad dialects

18. Gemer dialects

Eastern Slovak dialects

Eastern Slovak regions can be found in the regions of Spiš, Šariš, Zemplín and Abov.

30. Spiš dialects

31. Abov dialects

32. Šariš dialects

33. Zemplín dialect

34. Soták dialects

35. Už dialects

40. Goral dialects

41. Ukrainian dialects

42. Various dialects

43. Hungarian dialects

These groups are further divided into a variety of subdialects (each village has its own dialect); especially mountain regions have very varied dialects. In the past, the mountain character of the country caused certain (language) isolation of the population in the individual provinces. These specific characteristics were also caused by the reorganisation and migration of the population, colonization, mixing of different dialect types, influence of neighbouring Slavic and non Slavic languages, changes in the employment of the population, etc. According to the nature of dialects and occurrence of the individual characteristics, Slovak dialects in Hungary, Serbia, Croatia, Romania, Bulgaria and other countries, where large compact groups moved in the past, can be included to these groups. In view of the limited number of old written monuments, Slovak dialects are the basic source of Slovak historical grammar.

Particularities of the Slovak Language

Slovak language has started to develop independently directly from Old Church Slavic language since the 10th century. Main changes were ongoing in it and were stabilized before the 15th century; some of them equally (reduction of the nasal vowels), the other differentially (vocalization of hard jers in eastern and western parts of contemporary Slovakia was of western Slavic type and in the central part it was of non-western Slavic type). The part of these changes was also decomposition of Old Church Slavic syllable structure, which influenced the changes in declension and conjugation. Although Slovak and Czech languages were developing under different conditions for a long period (Slovakia became a part of the Kingdom of Hungary in the 11th century), they have remained close to each other. However, some specific features of Slovak language (the forms *l**a**keť*/elbow, *Če**s**i*/the Czech, the suffix *-m* in the first person singular, etc.) are parallel in South Slavic languages. With some less significant characteristics, Slovak resembles Polish (prefix *pre-* unlike the Czech *pro-*, preservation of consonant *dz*, and several expressions such as *teraz, pivnica*). By other characteristics it approaches East Slavic languages. Therefore we talk about the central position of Slovak among the Slavic languages and about good understandability of Slovak for the members of other Slavic nations.

Slovak uses modified Latin script with diacritical marks. The palatalization of consonants is marked with a háček (*ď, ť, ň, ľ*; also used for graphemes *ž, š, č, dž*) and the length of vowels by an acute accent (*á, é, í, ý, ó, ú, ý*). Vowels are not subject to reduction, they are pronounced in full form in each position. Besides for vowels and consonants, several diphthongs occur in Slovak language. A phonetic speciality of Slovak standard language (and of Central Slovak dialects) is so-called rhythmic rule, which is a tendency not to have

two long syllables adjacent (*pekný – krásny*, *prosím – smútim*). Slovak has a dynamic accent bound to the first syllable of the word that is not very strong (it is weaker than in Russian or Polish). In prepositional phrases with one-syllable prepositions, the accent is usually put on the prepositions: *v škole* [in the school].

Unlike Russian or Czech, Slovak has a simpler structure of declension and conjugation paradigms. However, the system of substantive and verbal forms is clearly structured, in spite of unification tendencies. Slovak language has six grammatical cases (nominative, genitive, dative, accusative, locative and instrumental). Unlike Czech, the vocative is not frequently used in Slovak; it is usually identical with the nominative. Slovak recognizes 4 genders: masculine animate and masculine inanimate, feminine, and neuter of substantives and related adjectives, pronouns and numerals. Masculine and feminine genders with animate concretums are determined according to natural gender, in other cases it is a matter of convention, which is not signalized by any article, only sometimes by ending (e.g.: *strom/tree* – masc. inanimate, *jabloň/apple tree* – fem., *jablko/aple* – neuter.). For each *g*ender there are given several patterns in school books and their paradigms differ especially – in G/A sing. and N/G plur. (e.g.: masculine animate *chlap* / *chlap**a*** / *chlap**i*** / chlap**ov**, *hrdina* / *hrdin**u*** / *hrdin**ovia*** / *hrdin**ov***; *žena* / žen**y** / *žen**u*** / *žen**y*** / *žien*, *dlaň* / *dlan**e*** / *dlaň* / *dlan**e*** / *dlan**í***). In some patterns and cases there is some significant homonymy: G and A sing. of animate masculine, N and A sing. of inanimate masculine, in feminine gender of G sing. and N plur. etc. There are possible transitions among the paradigms, e.g. the feminine paradigm *kosť* is nowadays more productive than the paradigm *dlaň*. Words formally assigned to a certain paradigm quite often do not follow the pattern, which is the reason for many exceptions; in NLP literature a much larger number of paradigms is mentioned (Páleš, 1994; Benko et al., 1998; Sokolová, 2007).

In the conjugation of verbs, three tenses are distinguished: past, present, and future. In addition to the three forms – indicative, imperative, and conditional, most of the verbs exist in two aspects – perfective (*volať*) and imperfective (*zavolať*). Slovak is a highly inflectional language with elements of analytical constructions (especially in verb forms such as *budem písať*, *bol by som prišiel*). The grammar function of words is clearly designated by inflection, therefore the word order in a sentence is relatively free. In the syntactic typology, Slovak is characterised by a basic construction scheme S(ubject) – V(erb) – O(bject), however, it is rather theoretical scheme, whose realization varies as a consequence of the free word order. Cases are helpful for unambiguous determination of S and O (S is in N case, O is usually in A or G, D cases, rarely in other cases), homonymy of the forms, however, can be a cause of an uncertainty in subject and object functions (especially in foreign proper names but also in several other cases). Special problems for foreigners and computer processing of Slovak language are caused by highly movable verbal morphemes *sa, si*, which can be situated in front of the verb or behind it and in distance of several words, or even in the different part of the sentence structure (*Netrvalo dlho, keď **sa** im ich hviezda, ktorú predtým videli v diaľke, zrazu **priblížila***). In Slovak language, two-member sentences with a subject (agents) are the most frequent but one-member constructions without agents are also frequently used (*Prší., Prišlo mu zle., Na stavbe sa tvrdo pracuje.*). Subject is known from the context and the form of the predicative verb is not expressed formally (*Našiel som ho.*); its presence in the sentence in the form of personal pronoun marks an emphasis. (*Ja som ho našiel!*).

Slovak Language on the Internet

At the end of 2010 the size of Slovak Internet population reached approximately 2394 000, which is more than 44 % of all Slovak inhabitants. In case of younger generation, this percentage has been much higher as young people spend much time on the Internet during the day. By the end of 2010 the number of Slovak domains exceeded the level of 231 thousand5. The fraction of *.sk* domains on the world wide web was about *1 ‰6 by the end of* 2010. The style of Internet communication and the texts to be found on the Internet are interesting for the natural language research but also for the text collecting purposes. Internet is also a place for the usage of various applications, which use language data as a source.

Shared with many other European languages, a specific feature of early Slovak language presence on the internet[7] was the habit of using the language without diacritics. Owing to the "character encoding mess" in the late 80's and 90's and to the lack of software support for different character encodings, the "proper" language on the internet started to dominate only in the late 1990's. Nowadays, with almost universal Unicode and UTF-8 encoding, there are no more outstanding problems and the diacritics are used universally (however, in informal contexts, such as in e-mails and discussion fora, and especially in SMS, Slovak language without diacritics is common).

A special category consists of bilingual dictionaries, which are freely accessible to Slovak users of three major Slovak portals (*azet.sk*, *centrum.sk, zoznam.sk).*

Google develops freely accessible, automatic text translator from various languages into Slovak and vice versa. The rate of correctness is, however, low in case of the majority of languages. There is an interesting result in case of mutual translation between closely related languages of Slovak↔Czech, where the percentage of correctness of the translation is good. Of course, even these translations are sometimes incorrect, however, they are much more successful than the translations between Slovak and English, German, French, and other major languages.

The use of Internet by Slovak Internet users is reflected in more than 60 000 registered Slovak users of Internet encyclopedia Wikipedia in Slovak language. Slovak Wikipedia *includes more than 285 000 articles.*

Slovak language as a foreign language

Slovak Online

Slovak Online is a project providing a web portal enabling free-of-charge studies of Slovak language by means of e-learning. Provided language courses of different levels (mini course for tourists, courses A1 and A2 according to Common European Framework of Reference for Languages) are divided into topical chapters and they are supplemented by audio and video recordings and exercises. The site includes an outline of Slovak grammar and orthography, dictionary and language games. It also provides some basic information and trivia about Slovakia and Slovak language, a library with the extracts of Slovak literary works and the possibility of instant messaging communication of the registered users.

Target group of the site consists of foreigners living in Slovakia, partners in the mixed marriages, inhabitants of border area, Slovak

people living abroad, slovakists and slavists, immigrants, students and tourists. Currently, the site has a German, English, Esperanto, Lithuanian, Polish and Slovak version.

The project, the first one of its kind, came into existence on the basis of experience gained by the operation of the *lernu!*8 site – the biggest portal for the Esperanto language studies. It was supported by European Committee in the frame of the KA2 programme – languages – lifelong learning. The project is coordinated by a civic association Edukácia@Internet (Slovakia), with partnership of Ľudovít Štúr Institute of Linguistics (Slovakia), Studio GAUS (Germany), Vilniaus universitetas (Lithuania), Wyższa Szkoła Informatyki, Zarządzania i Administracji w Warszawie (Poland) and Slovak Centre London (UK).

Studia Academica Slovaca – The Centre for Slovak as a Foreign Language

Studia Academica Slovaca – The Centre for Slovak as a Foreign Language (SAS) is a specialized centre at the Faculty of Arts[9], Comenius University (FF UK) in Bratislava. The pedagogical and research activities focus on the education of foreigners interested in Slovak language and culture, propagation of Slovak science, culture and art abroad, implementation and coordination of the research of Slovak as a foreign language, realization of international and domestic research projects and activities aimed at creating and publishing academic Slovakist material and textbooks of Slovak as a foreign language. Besides the SAS being an expert centre for Slovak as a foreign language, it also traditionally participates in scientific methodical preparation for lecturers of Slovak as a foreign language at universities abroad. The result of the cooperation with the lectorates and foreign Slavists builds a database of Slavonic studies abroad.

Another part of the Centre's activities is the annual organization and realization of a Summer School of Slovak Language and Culture Studia Academica Slovaca, which has been offered to foreign applicants since 1965. The Methodical Centre SAS reassumed its successful history in 1992, and in 2006 it was transformed into SAS – The Centre for Slovak as a Foreign Language. In its almost half-century of existence, SAS has not only shaped its contents and methodology but has also become a respected and coveted institution, with almost 6 000 foreign alumni interested in Slovak language, culture and realia from more than 50 countries all over the world. On the grounds of Studia Academica Slovaca the basis of scientific description and didactics of Slovak as a foreign language was laid, and thanks to its co-operators, the first textbooks and didactics of Slovak as a foreign language were written. In relation to its wide tradition and experience, SAS currently works as a coordination and information centre with slovakwide as well as an exterior sphere of activity.

A product of the implementation of the project by the Studia Academica Slovaca group "Educational programme Slovak as a Foreign Language", the Faculty of Arts of Comenius University has obtained the initiative award European label 2007 from the European Commission in the sphere of language education.

Summer School of Slovak Language and Culture

The Summer School of Slovak Language and Culture Studia Academica Slovaca is aimed at Slovakists and Slavists abroad, cultural workers, managers, lecturers, language teachers, translators and all

those interested in studying Slovak language and culture. The aim of the course is to enable students to acquire and improve their Slovak language competence on various levels, as well as to extend their knowledge in Slovak linguistics, literature, history and culture.

Established in 1965, Summer School SAS is the oldest summer university in Slovakia and has been under the name Studia Academica Slovaca since 1966. Since its establishment, SAS has continually maintained its profile of Slovakist academic studies. The Summer School SAS is usually attended by approximately 150 participants from more than 25 countries all over the world. Those creating and leading the seminars are professional teachers and lecturers, experts in teaching Slovak as a foreign language, often experienced in teaching in Slovakia as well as abroad.

Courses of Slovak as a Foreign Language

In the winter and summer semestre of the academic year, Studia Academica Slovaca – The Centre for Slovak as a Foreign Language provides Slovak as a foreign language courses for beginners (levels A1 and A2) for foreign students, scholarship holders, host students and others who are interested.

In 2006 the SAS Centre acquired accreditation from the Ministry of Education of the Slovak Republic for providing educational activities concerning Slovak as a Foreign Language – language courses in contact and distance form for all levels of language development including beginners (A1, A2), intermediate and upper-intermediate (B1, B2) and advanced (C1, C2). Their contents are published in printed version (Pekarovic ová et al., 2007) and published on the web10.

Distance Education

Studia Academica Slovaca – The Centre for Slovak as a Foreign Language offers those who are interested in Slovak language a Slovak e-learning course for level A1 (Basic User – Breakthrough) and level A2 (Basic User – Waystage). The course is a part of the project Educational Programme Slovak as a Foreign Language, whose implementation is based on a grant from the Ministry of Education of the Slovak Republic. The e-learning course is aimed especially at Slovakists and Slavists abroad, participants of the Summer School Studia Academica Slovaca as well as at all those interested in learning the Slovak language.

Methodical seminary on Slovak as a foreign language

Every year a Methodical seminary on Slovak as a foreign language for teachers of grammar and secondary schools abroad and for university lectors takes place to inform about new approaching linguistics, literature, culture and didactics of Slovak as a foreign language.

Educational Programme Slovak as a Foreign Language

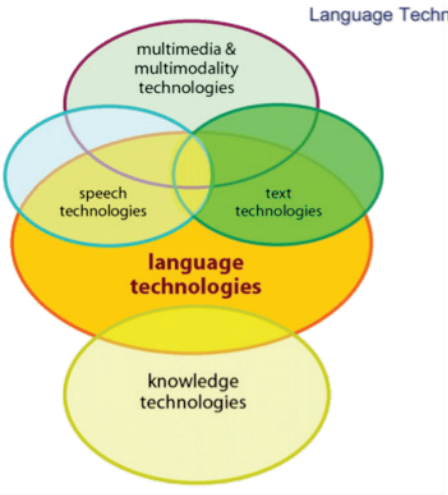
This educational project was implemented by Studia Academica Slovaca – The Centre for Slovak as a Foreign Language at the Faculty of Arts, Comenius University in Bratislava, based on the grant task from the Ministry of Education of the Slovak Republic. Those working on the project are internal and external employees who are authors of didactic and scientific materials, working also as lecturers.

The objective of the project is to create both content and forms of language development for foreigners on individual levels corresponding to The Common European Framework of Reference for Languages, as well as to specify individual criteria of the evaluation and certification of language competence. The main scope is the preparation of standard and specialized learning materials for students and methodical materials for teachers.

Language Technology Support for Slovak

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, detecting accents (“ città ” and “ citta’ ”) and apostrophes (“dell’UE” and “della UE”) for Italian, etc.
- Grammatical analysis: finding the verb and its objects, modifiers, etc.; detecting the sentence structure.
- Semantic analysis: disambiguation (Which meaning of “apple” is the right one in the given context?), resolving anaphora and referring expressions like “she”, “the car”, etc.; representing the meaning of the sentence in a machine-readable way.

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

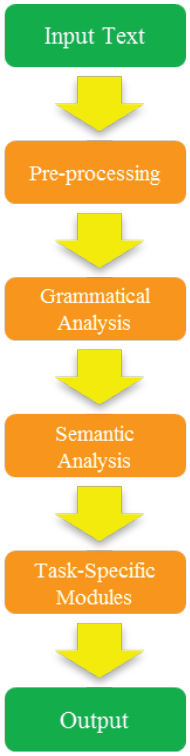


Figure 2: A Typical Text Processing Application Architecture

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 Slovak.

Core application areas

Language Checking

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

Eye have a spelling chequer,

It came with my Pea Sea.

It plane lee marks four my revue

Miss Steaks I can knot sea.

For handling this type of errors, analysis of the context is needed in many cases, e.g., for deciding if a word needs to be written with "y" or "i", as in:

Kto chce psa biť, palicu si nájde.

[He who wants to beat a dog will find a stick.]

Kto chce psom byť, pána si nájde.

[He who wants to be a dog will find his master.]

This either requires the formulation of language-specific grammar rules, i.e. a high degree of expertise and manual labour, or the use of a so-called statistical language model. Such models calculate the probability of a particular word occurring in a specific environment (i.e., the preceding and following words). For example, *chce psom byť* is a much more probable word sequence than *chce psom biť*, and *chce psa biť* is a much more probable sentence than *chce psa byť* (neverthless, we can contrive contexts where all four sequences are grammatical). 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 Slovak with its flexible word order and richer inflection.

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

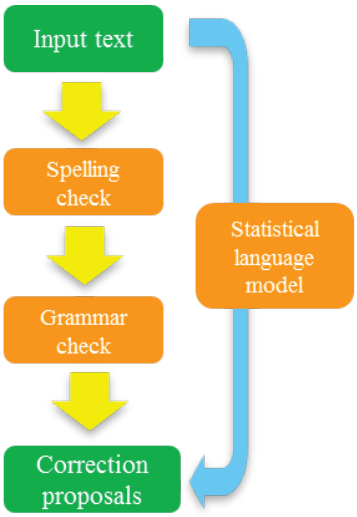


Figure 3: Language Checking (left: rule-based; right: statistical)

consistent with certain rules and (corporate) terminology restrictions.

The existing spelling checkers for Slovak are mostly based on a dictionary of basic word forms (lemmas) combined with a set of morphological rules enabling the analysis or generation of all (correct) word forms. Although this simple approach seems to be satisfactory, it has two substantial drawbacks. The first issue concerns the superficially correct word forms appearing in a wrong context. The second drawback is the inability to distinguish between real spelling errors and word forms which are correct, but which are not contained in the dictionary. Such words will always exist due to the natural enhancement of a lexicon by newly created words, by new scientific or technical terms etc.

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. In 2006, the verb *googlovať/googliť* very narrowly missed being included in the first volume of the new Dictionary of Contemporary Slovak Language (*Slovník súčasného slovenského jazyka*), a fact that is over and over being used to reproach the dictionary authors for. Neither the search interface nor the presentation of the retrieved results have 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 mix11, 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.

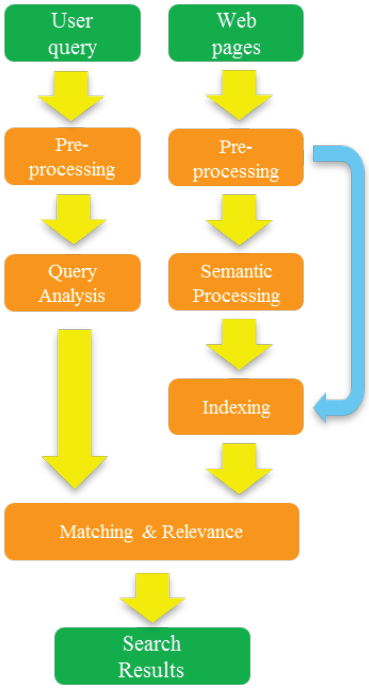


Figure 4: Web Search Architecture

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, have shown improvements by allowing to find a page on the basis of synonyms of the search terms, e.g. *jadrová*, *atómová* and *nukleárna energia* (nuclear, atomic and nuclear *again* 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 Slovakia, there were several different SMEs developing search technologies, or search technologies developed by Czech SMEs were used. The first Slovak search engine taking into account Slovak morphology[12] has been *morfeo.sk,* run by internet portal *centrum.sk,* which started to provide fulltext search of the *.sk* domain webpages in 2003. It used lemmatization and morphology annotation to look for inflected words in order to be able to provide the user with more relevant results than those including the basic forms of the words. It also included fuzzy search possibilities and search by synonyms. By 2009 the number of indexed pages was over 117 million; since at that time, Google already included Slovak morphology support and surpassed the number of the indexed pages, and *centrum.sk* switched to customized Google search.

One of the enterprises engaged in this field is Forma s.r.o., a company that developed three linguistic modules: proofreader, lemmatizer, and thesaurus, on the basis of data obtained from the Ľ. Štúr Institute of Linguistics of the Slovak Academy of Sciences. The company also developed separate programs for full-text Slovak search and still operates online versions of some older dictionaries.

Focus on development for search technologies lies on 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.

Research in this field is mainly performed by the Institute of Informatics of the Slovak Academy of Sciences, which has started to deal with processing of written natural language in 2006. At the same time, WIKT workshops, containing several articles or even entire section dedicated to processing of Slovak language in each year, have been initiated. Since 2006, the research in the Institute of Informatics has been mainly performed within NAZOU project aimed at development of the tools for obtaining, processing, organizing and presentation of information from Internet. Job offers represented a specific application with the tools having been tested on Slovak job offers as well. The Institute prepared an analysis of

processing of texts in Slovak (Laclavík, 2007a) and, at the same time, Ontea, a tool for extraction of information (Laclavík, 2007b, 2009) was developed. The tool was later integrated with the tools for language identification (Vojtek, 2006) and lemmatization (Krajči, 2007)

Ontea works on the basis of searching for patterns, which can either be linguistically dependent patterns, such as use of prepositions, sentence structure, but also simpler patterns, such as use of capitals and abbreviations, e.g. *s. r. o.* and *a. s.* for searching for businesses, *SK, SKK, EUR, EURO, €* for price searching, or abbreviations of Slovak first names for searching of persons in a text. A principle is applicable to various languages, but the patterns have to be made for a specific language, e.g. Slovak. At the present, the Ontea tool is being improved for use in processing of e-mail communication. The system was tested within AIIA project13 (Laclavík, 2010) on Slovak e-mails of Anasoft company and SANET association. Ontea uses not only the patterns, but also dictionaries (gazetteers) as well as their combinations in order of extraction and identification of entities in a text. Since use of dictionaries (but also some patterns) can cause problems with identification of an entity that is in other than basic form, use of lemmatizer seems to be appropriate. Since the entities are mostly of a nomenclatural nature, such as people, locations, product names, names of projects or services, they are difficult to be lemmatized. Although the problems have not yet been successfully resolved, they could be settled by a new method of combination of dictionary, character based tokenization, lemmatization, and verification of an entity in dictionary.

Extraction of entities using patterns was also used in an experiment with large group of data, when Slovak websites were processed with an aim of extraction of geographical data (Slovak addresses) and their subsequent finding (Dlugolinský et al., 2010).

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.

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.

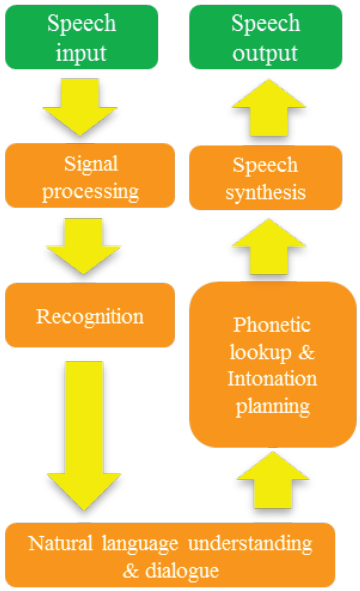


Figure 5: Simple Speech-based Dialogue Architecture

- 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. Fundamental requirement for good performance is also well trained acoustic model based on huge amount of recorded data covering different accent, age group, gender etc. 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 acoustic and language models may increase the costs significantly. However, VUIs that employ language models and initially allow a user to flexibly express their intent – evoked, e.g., by a 'How may I help you' greeting – show both a higher automation rate and a higher user acceptance and may therefore be considered as advantageous over a less flexible directed dialogue approach.

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

Regarding the market for Speech Interaction technology, the last decade underwent a strong standardisation of the interfaces between the different technology components, as well as by standards for creating particular software artifacts 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, Loquendo and SVOX being the most prominent ones in Europe.

Speech recognition in Slovakia has a long history but it has been done only at a universities or scientific institutions. Most of the places focus on basic research and solutions of specific problems of speech recognition. The Department of Speech Analysis and Synthesis of the Institute of Informatics of the Slovak Academy of Sciences as a participant of SpeechDat-E project focuses mainly on acoustic models for telephony systems. With growing number of speech data other than from that of the telephony domain the institute has been trying to create widely usable acoustic models for application such as dictation, talk transcription, etc. The main focus of the Department of Telecommunication of the Slovak Technical University is processing of speech signal in noisy conditions (speech/silence detection, features extraction, etc.). Among others the department created several small speech recognition systems to compare performance and usability of different free speech recognition systems for the Slovak language. At TU Košice there are several departments focusing on automatic speech recognition. The Department of Electronics and Multimedia Communications was originally focusing mainly on basic research for digital processing

of speech signal. Today the most noticeable output represent the activities in the field of language modelling for the Slovak language. The current language model created at the department contains 2109 tokens. The second important workplace at TU Košice is the Department of Cybernetics and Artificial Intelligence where the first voice retrieval information dialogue system and SAMPA for the Slovak language were created. Today the speech recognition activities at the department plays rather minor role. The Department of Applied Mathematics and Statistics of the Faculty of Mathematics, Physics and Informatics of the Comenius University in Bratislava is working mainly on speech recognition of isolated words for children voices. The results were applied in educational process for verification of a read text by children. From the audio data recorded for the acoustic model training two speech databases have been created *(Alica* and *Viktória)*. These databases are available also for other research institutions. The main institution for the speech recognition at University of Žilina is the Department of Telecommunications and Multimedia. Its team focuses mainly on digital signal processing for the speech recognition and recognition of isolated words using Hidden Markov Models.

A close cooperation between the Department of Electronics and Multimedia Communications of TU Košice and the Department of Speech Analysis and Synthesis of Institute of Informatics of the Slovak Academy of Sciences resulted in first visible success. The target of the cooperation should be a large vocabulary of the Slovak recognition system for continuous speech.

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

Machine Translation

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

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

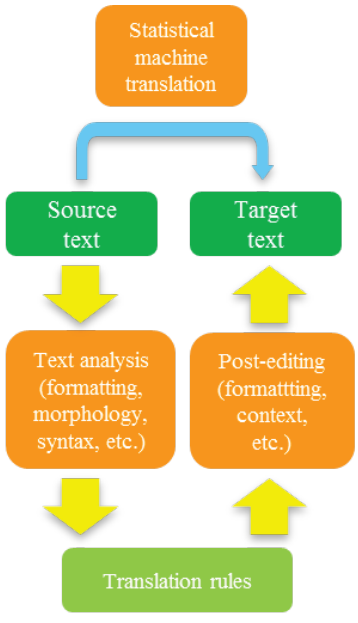


Figure 6: Machine translation (top: statistical; bottom: rule-based)

animal or an operating system) or the attachment of attributes on the syntactic level as in:

Otcovi priatelia neprišli, moji áno.

[Father's friends did not come, mine did.]

Otcovi priatelia neprišli, mne áno.

[The friends did not come to the father, [but] to me.]

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

Beginning in the late 1980s, as computational power increased and became less expensive, more interest was shown in statistical models for MT. The parameters of these statistical models are derived from the analysis of bilingual text corpora, such as the Europarl parallel corpus, which contains the proceedings of the European Parliament in 21 European languages. Given enough data, statistical MT works well enough to derive an approximate meaning of a foreign language text. However, unlike knowledge-driven systems, statistical (or data-driven) MT often generates ungrammatical output. On the other hand, besides the advantage that less human effort is required for grammar writing, data-driven MT can also cover particularities of the language that go missing in knowledge-driven systems, for example idiomatic expressions.

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

In the 1990s the prototype of MT between closely related languages was proposed for the pair Czech and Slovak at Charles University in Prague.

TEOS Trenčín markets the first practical multilingual MT software for the Slovak language, bundled with their PC dictionary software. However, since the system did not use any further linguistic analysis and simply substituted words from one language with words in the other language (mostly limited to lemmas), its usability was limited to languages that do not have much morphology – i.e. English. Later version allowed to translate webpages on the fly, a functionality that is particularly useful in the English→Slovak translation, which coincidentally was the only translation direction that "worked".

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 (not limited to the Slovak language) are English-centred. In particular, Google Translator offers the best translation quality for translations from/to English.

Availability of large amounts of bilingual texts is really the key in statistical MT. For Slovak, corpora of parallel texts with several other languages are currently being created. The largest data – in total several million pairs of sentences – is available in the Slovak-Czech and Slovak-English parallel corpora, compiled at the Ľ. Štúr Institute of Linguistics. The corpora contain mostly fiction and are automatically sentence aligned.

Language Technology 'behind the scenes'

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

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

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

The JBOWL (Java Bag-Of-Words Library) software library developed at the Centre for Information Technologies (FEI-CIT) in Košice for the support of NLP and Text Mining applications. JBOWL is a modular system enabling maintenance of textual documents. It provides functions and means supporting the processing of natural language texts (e.g., tokenization, morphological analysis, lemmatization, disambiguation, syntactic analysis based on ATN networks, clustering and phrase identification, term weighting and indexing) as well as the knowledge discovery and

mining from unstructured textual documents. In addition, the system provides implementations of several algorithms of controlled and uncontrolled machine learning with customizable input parameters and methods for evaluating the quality of Text Mining models.

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

Language Technology in Education

Language Technology is a highly interdisciplinary field, involving the expertise of linguists, computer scientists, mathematicians, philosophers, psycholinguists, and neuroscientists, among others. As such, it has not yet acquired a fixed place in the Slovak faculty system.

Since 2007 the researchers from the Institute of Informatics of the Slovak Academy of Sciences (Michal Laclavík and Martin Šeleng) have been teaching the Information retrieval course14 in the Faculty of Information Technologies of the Slovak Technical University. This course focuses on such themes as information retrieval, information extraction, graph algorithms for their support as well as processing of large amount of data. The students solve various practical projects in this domain, while many of them use Slovak text sources, and some of them solve directly the NLP problems of the Slovak language processing. As an example let us mention several projects aiming at creation of statistical, dictionary oriented or algorithmic stemmer, based on the "snowball" or "Egothor projects, and at determination of the efficiency and statistics for the simple stemmers which function on the principle of omitting the vowels, diacritic marks or, eventually, word endings etc. At the same time there are also projects of statistical translation or the automatic dictionary creation between the Slovak or other languages (English, Czech). Finally, let us mention the projects utilising dictionaries or frequency language dictionaries for applications such as T9, named entities extraction using computer learning methods and libraries such as OpenNLP, creation of POS tagging algorithms as well as extraction of events from e-mails or from Slovak webpages and the like.

There is no regular CL study programme otherwise.

Availability of Tools and Resources for Slovak

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

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

programming environments (such as Java EE)? Do industry/research standards/quasi-standards exist and if so, is the tool/resource compliant (data formats etc.)?

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

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

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

Table of Tools and Resources

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

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

Conclusions

The table can be summarized in the form of a number of key messages, which highlight crucial issues for the further development of automatic language processing of Slovak on the basis of the present situation:

- While some specific corpora of high quality exist, a very large syntactically annotated corpus is not available.
- For Slovak, the Slovak National Corpus is the reference language corpus, but only the query interface is generally available, due to licensing restrictions.
- On the other hand, Corpus of Spoken Slovak is not encumbered by copyright law and is therefore publicly available, but its size is miniscule compared with the corpus of written language.
- Many of the resources lack standardization, i.e., even if they exist, sustainability is not given; concerted programs and initiatives are needed to standardize data and interchange formats.

- Semantics is more difficult to process than syntax; text semantics is more difficult to process than word and sentence semantics.
- There is an ontological resource for Slovak (even mapped to English ontological resources) but its coverage is limited.
- Standards do exist for semantics in the sense of world knowledge (RDF, OWL, etc.); they are, however, not easily applicable to NLP tasks.
- Written text processing is more mature than speech processing (especially speech recognition)
- Many of the resources taken as standard in other languages are missing for Slovak; NLP language research in Slovakia is severely underfunded.
- Some of the research and development activities for the Slovak language is carried out in the Czech Republic by Czech universities and Czech SMEs.
- Speech Recognition of the Slovak language is studied at several universities and workplaces but the amount of free tools and data is limited.
- In contrast with speech recognition, speech synthesis is less covered by universities and other workplaces.
- In the field of speech synthesis, there are open source packages available together with several other simple synthesizers but the speech synthesis with more natural voices is not available.
- Slovak dialogue systems are very little extended due to poor accessibility of high quality speech recognition modules of the Slovak language.

Bibliography

Bednár, P., Butka, P., Paralič, J.: Java Library for Support of Text Mining and Retrieval. In: Proceedings of the 4th annual conference Znalosti. Eds. L. Popelínský, M. Krátký. Ostrava : Vysoká škola báňská – Technická univerzita, 2005. pp. 162 – 169. ISBN 80-7097-523-7

Bočák, M., Garberová, B., Gregová, R., Mochňacká, B., Oborník, P., Rusnák, J., Sabol, J. S., Smoláková, V.: Texty elektronických médií: Stručný výkladový slovník. Prešov : Prešovská univerzita v Prešove, 2010. 290 p. ISBN 978-80-555-0256-4

Čerešňa, M.: Výpočtový model na analýzu viet slovenského jazyka. Diplomová práca. Fakulta matematiky, fyziky a informatiky Univerzity Komenského v Bratislave, 2002. 95 p. http://www.dbai.tuwien.ac.at/staff/ceresna/ling/nl-parsing-model.pdf.

Dlugolinský, Š., Laclavík, M., Hluchý, L.: Towards a search system for the Web exploiting spatial data of a web document. In: DEXA 2010: Database and Expert Systems Applications: proceedings. Ed. R. R. Wagner. Los Alamitos : IEEE Computer Society, 2010, pp. 27 – 31. ISBN 978-0-7695-4174-7

EUROMAP Study. Benchmarking HLT progress in Europe – The EUROMAP Study By Rose Lockwood and Andrew Joscelyne. 2003. http://www.hltcentral.org/page-243.0.shtml

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

Furdík, J., Furdík, K.: Slovotvorný slovník slovenčiny - softvérové riešenie. In: Varia 9. Zborník materiálov z IX. kolokvia mladých jazykovedcov. Ed. M. Nábělková, M. Šimková. Bratislava : Slovenská jazykovedná spoločnosť pri SAV, 2002. pp. 305 – 316. ISBN 80-89037-04-6

Furdík, K.: Získavanie informácií v prirodzenom jazyku s použitím hypertextových štruktúr. Dizertačná práca. Fakulta elektrotechniky a informatiky Technickej univerzity v Košiciach, 2003. 150 p.

Galamboš, L.: Lemmatizer for Document Information Retrieval Systems in JAVA. In: SOFSEM 2001: Theory and Practice of Informatics, 28th Conference on Current Trends in Theory and Practice of Informatics. Eds. Leszek Pacholski, Peter Ružička. Springer, 2001. ISBN 3-540-42912-3

Galamboš, L.: Multilingual Stemmer in Web Environment. PhD Thesis. Faculty of Mathematics and Physics, Charles University in Prague, 2004.

Galamboš, L.: Semi-automatic stemmer evaluation. In: Intelligent Information Processing and Web Mining. Eds. Mieczyslaw A. Klopotek, Slawomir T. Wierzchon, Krzysztof Trojanowski. Springer, 2004. 641 p. ISBN 3-540-21331-7

Garabík, R.: Štruktúra dát v Slovenskom národnom korpuse a ich vonkajšia anotácia. In: Slovenčina na začiatku 21. storočia. Ed. Mária Imrichová. Prešov : Fakulta humanitných a prírodných vied Prešovskej univerzity, 2004. pp. 164 – 173. ISBN 80-8068-526-6

Garabík, R., Gianitsová, L., Horák, A., Šimková, M.: Tokenizácia, lematizácia a morfologická anotácia Slovenského národného korpusu. Interný materiál. 2004. http://korpus.juls.savba.sk/publications/block2/tokenizacia-lematizacia-a-morfologicka-anotacia-slovenskeho-narodneho-korpusu/Tagset-aktualny.pdf

Garabík, R.: Slovak morphology analyzer based on Levenshtein edit operations. In: 1st Workshop on Intelligent and Knowledge oriented Technologies - WIKT 2006 Proceedings. Eds. Michal Laclavík, Ivana Budinská, Ladislav Hluchý. Bratislava, 2007. ISBN 978-80-969202-5-9

Garabík, R.: Nástroje pri tvorbe a používaní projektov Slovenského národného korpusu. In: Proceedings of the 5th Workshop on Intelligent and Knowledge oriented Technologies - WIKT 2010. Bratislava, 2010. pp. 2 – 7. ISBN 978-80-970145-2-0

Genči, J: Contribution to Processing of Slovak Language at DCI FEEI TUKE. In: Slovanské a východoeurópske jazyky v počítačovom spracovaní. Bratislava : VEDA, 2005. pp. 67 – 72. ISBN 80-224-0895-6 http://korpus.juls.savba.sk/~slovko/2005/proc/slovko.pdf

Ivoríková, H. et al.: Krížom-krážom. Slovenčina A1+A2. Cvičebnica. Bratislava : Univerzita Komenského, 2010. ISBN: 978-80-223-2809-8

Kamenárová, R. et al.: Krížom-krážom. Slovenčina A1. Bratislava : Univerzita Komenského, 2007. 188 p. ISBN 978-80-223-2441-0

Kamenárová, R. et al.: Krížom-krážom. Slovenčina A2. Bratislava: Univerzita Komenského, 2009. 207 p. ISBN 978-80-223-2608

Kostolanský, E., Hašanová, J., Benko, V.: Model morfologickej databázy slovenčiny (počítačové spracovanie jazyka). Trnava : Fakulta prírodných vied Univerzity sv. Cyrila a Metoda v Trnave, 2004. 188 p. ISBN 80-89034-70-5

Krajči, S., Novotný, R.: Hľadanie základného tvaru slovenského slova na základe spoločného konca slov. In: 1st Workshop on Intelligent and Knowledge oriented Technologies - WIKT 2006 Proceedings. Eds. Michal Laclavík, Ivana Budinská, Ladislav Hluchý. Bratislava, 2007. ISBN 978-80-969202-5-9

Laclavík, M., Ciglan, M., Krajčí, S., Hluchý, L., Furdík, K.: Dostupné zdroje a výzvy pre počítačové spracovanie informačných zdrojov v slovenskom jazyku. In: 1st Workshop on Intelligent and Knowledge oriented Technologies - WIKT 2006 Proceedings. Eds. Michal Laclavík, Ivana Budinská, Ladislav Hluchý. Bratislava, 2007. pp. 92 – 97. ISBN 978-80-969202-5-9

Laclavík, M., Ciglan, M., Šeleng, M., Krajči, S., Vojtek, P., Hluchý, L: Semi-automatic semantic annotation of Slovak Texts. In: Computer treatment of Slavic and East European languages. Eds. Jana Levická, Radovan Garabík. Bratislava : Slovak National Corpus, Ľ. Štúr Institute of Linguistics Slovak Academy of Sciences, 2007. pp. 126 – 138. ISBN 978-80-87139-05-9

Laclavík, M., Šeleng, M., Ciglan, M., Hluchý, L.: Ontea: Platform for pattern based automated semantic annotation. In: Computing and informatics. 2009, vol. 28, no. 4, pp. 555 –579. ISSN 1335-9150

Laclavík, M., Šeleng, M., Gatial, E., Dlugolinský, Š., Balogh, Z., Hluchý, L., Jeckel, E., Horváth, P.: AIIA: adaptívna platforma na podporu interoperability v súkromnom a verejnom sektore. In: Znalosti 2010: Sborník příspěvků 9. ročníku konference. Ed. Pavel Smrž. Praha : Vysoká škola technická v Praze, 2010, pp. 227 – 230. ISBN 978-80-245-1636-3

Nižníková, J., Sokolová, M.: Valenčný slovník slovenských slovies. Prešov : Filozofická fakulta Prešovskej univerzity, 1998. ISBN 80-88885-53-1

Nižníková, J.: Valenčný slovník slovenských slovies (Na korpusovom základe). 2nd volume. Prešov : Filozofická fakulta Prešovskej univerzity v Prešove, 2006. ISBN 80-88885-53-1

Nižníková, J.: Vetné modely v slovenčine. Prešov : Filozofická fakulta Prešovskej univerzity, 2001. ISBN 80-8068-052-3

Páleš, E.: Sapfo – parafrázovač slovenčiny. Bratislava : VEDA, 1994. ISBN 80-224-0109-9.

Pekarovičová, J. et al.: Slovenčina pre cudzincov. Praktická fonetická príručka. Bratislava : Stimul, 2005. 244 p. ISBN 80-89236-04-9

Pekarovičová, J.: Slovenčina ako cudzí jazyk – predmet aplikovanej lingvistiky. Bratislava : Stimul, 2004. 208 p. ISBN 80-88982-87-1

Pekarovičová, J., Žigová, Ľ., Mošaťová, M.: Vzdelávací program Slovenčina ako cudzí jazyk. Jazykový kurz v kontaktnej a dištančnej forme. Bratislava : Stimul, 2007. 83 p. ISBN 978-80-89236-28-2. http://www.fphil.uniba.sk/fileadmin/user_upload/editors/sas/slavic/Vzdelavaci_program.pdf

Pekarovičová, J.: Slovakistika v zahraničí. Bratislava : Stimul, 2001. 165 p. ISBN 80-88982-41-3

Pekarovičová, J., Vojtech, M.: Slovacicum. Súčasné Slovensko. Bratislava : Stimul, 2006. 239 p. ISBN 80-89236-10-3

Sabol, J., Zimmermann, J.: Komunikačný štatút prízvuku v spisovnej slovenčine. In: Acta Facultatis Philosophicae Universitatis Šafarikanae. Spoločenskovedný zošit 10. Prešov : Filozofická fakulta Univerzity P. J. Šafárika, 1994.

Sabol, J., Zimmermann, J.: Základy akustickej fonetiky. Košice: Rektorát Univerzity P. J. Šafárika v Košiciach, 1986.

Sabol, J.: Metamorfózy kvantity v spisovnej slovenčine. In: Studia Academica Slovaca. 33. Prednášky XL. letnej školy slovenského jazyka a kultúry. Eds. Jozef Mlacek, Miloslav Vojtech. Bratislava : Stimul – Centrum informatiky a vzdelávania Filozofickej fakulty Univerzity Komenského, 2004. pp. 183 – 203. ISBN 80-88982-82-0

Sabol, J.: Vzťah znaku a slova a supraznaku a vety v mediálnom texte. In: Médiá a text. Ed. Juraj Rusnák, Michal Bočák. Prešov: Filozofická fakulta Prešovskej univerzity, 2005, pp. 20 – 26. ISBN 80-8068-408-1

Sabol, J.: Historicko-synchronické morfologické a derivačné signály kvantity v slovenčine. In: Kvantita v spisovnej slovenčine a v slovenských nárečiach. Ed. Matej Považaj. Bratislava : VEDA, 2005, pp. 9 – 32. ISBN 80-224-0858-1

Sokolová, M.: Nový deklinačný systém slovenčiny. Prešov: Filozofická fakulta Prešovskej univerzity v Prešove, 2007. ISBN 80-8068-550-9

Sokolová, M., Ološtiak, M., Ivanová, M. et al.: Slovník koreňových morfém slovenčiny. Prešov : Filozofická fakulta Prešovskej univerzity v Prešove, 2007. ISBN 80-8068-319-0.

Sokolová, M., Moško, G., Šimon, F., Benko, V.: Morfematický slovník slovenčiny. Prešov : Náuka, 1999. ISBN 80–968202–1-4

Sondy do morfosyntaktického výskumu slovenčiny na korpusovom materiáli. Eds. Miloslava Sokolová, Martina Ivanová. Prešov : Filozofická fakulta Prešovskej univerzity v Prešove, 2006. ISBN 80-8068-545-2

Soria, C., Mariani, J.: Report on Existing Projects and Initiatives META-NET study. 2011.

Šimková, M.: Korpusová lingvistika na Slovensku. In: Jazykovedný časopis, 2008, roč. 59, č. 1 – 2, s. 11 – 24.

Vojtek, P., Grlický, V.: Identification of Natural Language using n-grams and Markov processes. In: Tools for Acquisition, Organisation and Presenting of Information and Knowledge. Eds. Pavol Návrat et al. Bratislava : Vydavateľstvo Slovenskej technickej univerzity, 2006. pp. 154 – 161. ISBN 80-227-2468-8

Zimmermann, J.: Spektrografická a škálografická analýza akustického rečového signálu. Prešov : Náuka, 2002. ISBN 80-89038-22-0

Žigová, Ľ.: Slovenčina pre cudzincov. Gramatická a pravopisná cvičebnica. Bratislava : Univerzita Komenského, 2005. 164 p. ISBN 80-223-1926-0

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)

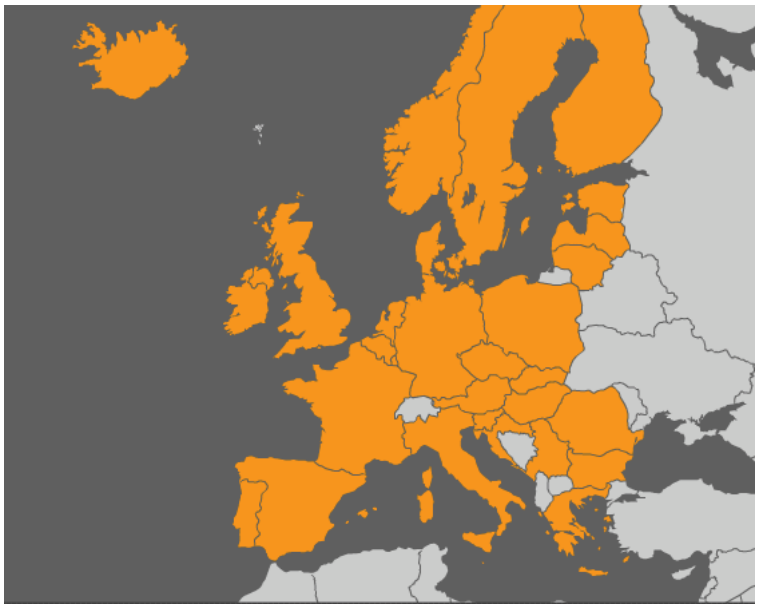


Figure 1: Countries Represented in META-NET

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] https://www.sk-nic.sk/documents/pdf/2010-12-31_SK-NIC_PS.pdf

[6] Number of all domains according to http://*www.verisigninc.com was reaching approximately 200 million at the end of year 2010*.

[7] And generally, in anything computer related

[8] http://www.lernu.net

[9] Officially also called Faculty of Philosophy

[10] http://www.fphil.uniba.sk/fileadmin/user_upload/editors/sas/slavic/Vzdelavaci_program.pdf

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

[12] Developed at the Faculty of Mathematics and Physics, Charles University, Prague

[13] http://aiia.ui.sav.sk/

[14] http://vi.ikt.ui.sav.sk/