META-NET White Paper Series

Languages in the European Information Society

– Serbian –

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

Prof. Duško Vitas, University of Belgrade
Prof. Ljubomir Popović, University of Belgrade
Prof. Cvetana Krstev, University of Belgrade
Dr. Mladen Stanojević, Pupin Institute
Prof. Ivan Obradović, University of Belgrade

Acknowledgements

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

Table of Contents

Executive Summary

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

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

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

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

This whitepaper for Serbian demonstrates that some of basic language technologies are developed (components depending on morphology), especially in the research environment, and some of them are applied in industry and business, e.g. speech based. However, the interest of the market for HLT products is still low.

According to the assessment presented in this report, immediate action must be taken before any breakthroughs for the Serbian language can be achieved.

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

A Risk for Our Languages and a Challenge for Language Technology

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

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

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

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

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

Opportunities for Language Technology

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

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

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

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

Multilingualism is the rule, not an exception.

Challenges Facing Language Technology

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

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

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

Language Acquisition

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

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

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

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

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

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

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

Serbian in the European Information Society

General Facts

Standard Serbian is the standard national language of Serbs and the official language in the Republic of Serbia. It was formed on the basis of ekavian and ijekavian Neo-Štokavian South Slavic dialects and its form was determined by the reformer of written language of the Serbs Vuk Karadžić (1787-1864), who at the same time reformed both the Cyrillic alphabet and orthogrpahy. In the 20th century, in the common state of Yugoslavia, this language was officially encompassed by *Serbo-Croatian*, a name that implied a linguistic unity with Croats (and later with other nations whose languages were based on Neo-Štokavian dialects). In the last decade of the 20th century in Serbia the name Serbo-Croatian has been replaced in general usage by the name Serbian.[5]

According to the census from 2002 the population of Serbia is 7.498.001,[6] and Serbian is the mother tongue for 88.3% of the population.[7] To this number one should add the population of Serbian nationality in other parts of former FR Yugoslavia (a number not easy to be determined). The Serbian Diaspora, mainly originating from people leaving the country in search of a job abroad and emigration for economical reasons, lives primarily in a number of countries of Central and Western Europe, in USA, Canada and Australia[8] (their knowledge of Serbian is mainly determined by the generation of émigrés they belong to).

Serbia is a multilingual community. The minority nationals,[9] according to the census from 2002, are the Hungarians (3.91%), Bosniaks (2.1%), Roma (1.44%), Croats (0.94%), Montenegrins (0.92%), Albanians (0.82%), Slovaks (0.79%), Yugoslavs (1.08%) and other ethnic minorities (Ashkali/Balkan Egyptians, Bulgarians, 'Bunjevci', Aromanians, Czechs, 'Gorani', Jews, Macedonians, Germans, Muslims, Romanians, Rusyns (Carpatho-Rusyns), Slovenes, Turks, Ukrainians and Vlachs, 2.45%). The structure of the minority nationals according to language is the following: 3.8% Hungarian, 1.8% Bosnian, 1.1% Roma, 0.8% Albanian, 0.8% Slovak, 0.7% Vlach, 0.5% Romanian, 0.4% Croatian, 0.2% Bulgarian and 0,2% Macedonian. The remaining languages are spoken by 0.5% of the population, whereas for 0.8% of the population these data are unknown. Elementary and middle school education in some of the minority languages exists in Serbia, namely in Albanian (55 elementary/4 middle schools), Hungarian (108/38), Bulgarian (26/-), Romanian (27/2), Rusyn (3/2), Slovak (15/2) and Croatian (7/1).[10] The instruction is supported by published textbooks and readers (for example, in 2005 a total of 526 textbooks for elementary and 283 for middle school were published).[11] Official use of minority languages is regulated by the Law on official use of language and alphabet[12] which provides for publishing laws and legal acts in languages of minority nationals, in accordance with a special law. This includes the right to address republic authorities in one's own language, as well as the right to be answered in that language (depending on the size of the minority community).

Translations to and from Serbian represent an important activity. During 2010 a total of 2549 titles were translated (1438 from English, 215 from French, 170 from German, 191 from Italian, 74 from Spanish, 149 from Hungarian). Part of the translations are from Slavonic languages (225 from Russian, 4 from Czech, 13 from

Polish, 21 from Slovak, 19 from Slovene, 18 from Macedonian, 12 from Bulgarian). Translations from Serbian to another language published in Serbia in 2010 comprise 591 titles.

Particularities of the Serbian Language

Serbian has its specific features which make its computational processing a complex task. These specific features will be outlined per linguistic areas.

Phonetics, phonology, morphophonology

The vowel system is simple (five vowels), but the consonant system is rather complex (twenty five consonants). The vibrant *r* in some positions is pronounced as a vowel and functions as a syllable nucleus, e.g. *p**r**st* '*finger*' or *v**r**sta* '*species*'. There is a large number of morphophonemic alternations in inflection and word formation, which are in some cases combined in such a way that two forms of a word can be very distant, e.g., the nominative singular of the noun *misao* is *misao* '*thought*' whereas its instrumental singular is *mišlju* **(**alternations *a/∅, o/l, l+j/lj/ s/š)*.

The accent system comprising four accents is based on two intersecting parameters: length opposition (long : short) and tone opposition (rising : falling). The distribution of rising and falling accents is regulated by special rules. Accentual alternations are common in inflection and word formation. As accents marks are not used, written texts contain homographs. For example *luk* with a short falling accent means *onion*, whereas with a long falling accent it means *arc* or *bow*.

For many words and grammatical forms the codified norm prescribes the pronunciation of post-accentual lengths, but they are increasingly disregarded in current usage, especially in certain positions.

Almost all words have an accent, but clitics also exist: proclitics (the majority of conjunctions and prepositions and the negative particle *ne* before verbs) and enclitics (non-accentuated forms of pronouns and verbs and the interrogative particle *li*).

As for borrowed words, their pronunciation is phonemically adapted to Serbian. However, combinations of consonants in borrowings often deviate from those typical of original Serbian words, e.g. *so**ftv**er* '*software*', *ha**rdv**er* '*hardware*', *interfe**js*** '*interface*'. In some cases, they are stressed on the final syllable, which is a deviation from the normative distribution of accents in Serbian.

A certain number of lexemes and word forms reflect the dialectally determined differences between the ekavian and ijekavian pronunciation of the old Slavic vowel called *jat*, as shown in the Table.

		ekavian	ijekavian
flower	**singular**	*cv**e**t* (long **e**)	*cv**ije**t*
	plural	*cv**e**tovi* (short **e**)	*cv**je**tovi*

Morphology (parts of speech, inflection, word formation)

There are ten parts of speech (word classes), with a large number of subclasses. The systems of pronouns and numerals are especially complex. The article does not exist.

Nouns are classified according to grammatical gender (masculine, feminine or neuter). However, classification according to semantic gender (male, female) is also relevant. For instance, the noun *gazda 'boss'* declines like a feminine gender noun but designates a male person.

Verbs are classified according to verbal aspect (perfective or imperfective). A certain number of verbs have both aspects. There are several types of so called reflexive verbs.

There are three types of inflection: (a) declension (nouns are inflected for number and case, while adjectives are inflected for gender, number, case and adjectival aspect);

4 types of nominal inflection	singular	paucal (2-4)	plural
window (masc.)	*prozor*	*prozor**a***	*prozor**i***
egg (neut.)	*jaj**e***	*jajet**a***	*jaj**a***
woman (fem.)	*žen**a***	*žen**e***	
news (fem.)	*vest*	*vest**i***	

(b) conjugation (which is highly complex); and (c) comparison (gradable adjectives and adverbs). Within all three types of inflection there are different paradigms, with a number of exceptions. Inflection is accompanied by numerous morphophonemic (and accentual) alternations. In all types of inflection, formal syncretism of certain grammatically different word forms is not uncommon. As a consequence of inflection, for a dictionary of 120000 lemmas, at least 4.5 million inflected grammatical forms exist (however, not as much formal words as some forms in certain paradigms are identical).

Personal pronouns (including the reflexive pronoun) and the auxiliary, copulative and existential verb *jesam*, as well as the auxiliary verbs *biti* and *hteti* have enclitic forms, which are used much more frequently than the corresponding stressed forms, e.g. *mu* is the enclitic form of *njemu* - the dative form of the pronoun *on 'he'*.

Word formation comprises suffixation, prefixation (especially important for verbs), certain types of composition, and, to a lesser degree, other word formation processes.

Calques and coinages, as well as so-called exocentric noun compounds, are frowned upon by language purists, as something that is not characteristic of authentic Štokavian word formation. This attitude does not facilitate lexical and terminological elaboration through word formation, and is one of the reasons for the very large number of borrowings.

Borrowings fit into existing inflectional paradigms, but there are also some exceptions, e.g. some foreign word do not inflect, such as the nouns *Meri 'Mary'* or *skvo 'squaw'* or adjectives *fer 'fair'* or *braon 'brown'*.

Well developed word formation (suffixation, prefixation, and to a lesser extent composition and various combined word formation processes) results in the fact that the majority of lexemes can be grouped into word families, and nested entries in dictionaries. It is very important that in a number of cases the meaning of the derived word is based on the systematic modifications of the meaning of the word it is derived from, which greatly facilitates the lexicographic processing of such cases. For example, for *prst 'finger'*, diminutive *prstić* and augmentative *prstetina,* adjective *prstni (*pertaining to a finger)*, prstast* (in the shape of a finger), *prstasto* (akin to finger), etc.

Lexis, phraseology, terminology, onomastics

Borrowings are, in general, phonologically and morphologically adapted, that is, adjusted to the pronunciation and morphology of Serbian. They also form word families according to Serbian word formation rules.

The composition of the vocabulary reflects, on the one hand, the fact that it is based on the Štokavian dialect, not only with regard to the original inventory but also with regard to new words formed according to Štokavian word formation processes. On the other hand, the vocabulary reflects the cultural and linguistic history of the Serbian nation, including borrowings from Church Slavonic, Turkish (*megdan 'battle'*), Russian (*zapeta 'comma'*), German (*štrudla 'strudel'*), French (*ruž 'lipstick'*), and, especially today, English (*parking 'parking'*). In addition, there are many internationalisms based on classical languages (Greek and Latin), especially in terminologies of specific fields.

Phraseology includes various kinds of idiomatic expressions, either of Serbian origin or created by the calquing of foreign expressions, today primarily English ones.

In the field of terminology and nomenclature, Serbian has always greatly relied on foreign languages; foreign terms have either been translated, with occasional deviations from word formation norms, or borrowed, especially in the case of terminological internationalisms. Endeavors aimed at finding original Serbian solutions or adapting existing terms to Serbian have yielded some results, but cannot keep pace with the growing needs in the fields of terminology and nomenclature.

Onomastics (anthroponymy, hydronymy, oronymy, etc.) represents an important segment of the vocabulary of Serbian, the more so as word families are also generated from these words.

Syntax, text linguistics

In terms of distribution of sentence constituents (subject, predicate, object, etc.), Serbian belongs to SVO languages with free word order (more precisely, with free distribution of mobile sentence constituents). This means that, in general, all permutations of mobile sentence constituents are permitted, but that the preferred order is: subject – predicate – object. However, free does not mean anarchic; on the contrary, the selection of a particular order is based on a very complex functional system, i.e. regulated by com-

binations of various syntactic, semantic, pragmatic and stylistic factors. Consider e.g. the sentence

Marija dade Jovanu jabuku.

[Mary gave John an apple.]

In Serbian, this idea can be expressed in 24 = 4! = 1*2*3*4 (number of permutations of four words) different ways

Marija dade Jovanu jabuku.

Marija dade jabuku Jovanu.

Marija Jovanu dade jabuku.

Marija jabuku dade Jovanu.

Jovanu dade Marija jabuku.

Jovanu Marija dade jubuku.

Jabuku Marija dade Jovanu.

Jabuku Jovanu dade Marija.

Dade Marija jabuku Jovanu.

Dade Jovanu jabuku Marija, etc.

Certain constituents are also expressed by enclitics, which are distributed in a very specific manner.

Subject pronouns need not be expressed; instead, they can be implied (the so-called zero subject). For example; ***Ja*** *se zovem Marko* vs. *Zovem se Marko* '*My name is Marko*'. A considerable number of sentence patterns are formed with various types of semantic subjects.

Besides the active and passive voice, there is another special way of formulating sentences with a non-specified human subject by using a reflexive form of the verb.

Negation is applied both to the verb and to the pronominal constituent (the so-called double negation), e.g. *Ovde* ***ne*** *poznajem* ***nikog*** 'I don't know anybody here'.

There are seven cases: nominative, genitive, dative, accusative, vocative, instrumental and locative.

an example of noun declension	singular	paucal	plural
nominative	*prozor*		*prozor**i***
genitive	*prozor**a***	*prozor**a***	*prozor**a***
dative	*prozor**u***		*prozor**ima***
accusative	*prozor*	*prozor**a***	*prozor**e***
vocative	*prozor**e***	*prozor**a***	*prozor**i***

an example of noun declension	singular	paucal	plural
instrumental	*prozor**om***		*prozor**ima***
locative	*prozor**u***		*prozor**ima***

Oblique cases can all be combined with prepositions (the locative always so). All these cases and prepositional phrases are polysemous. Conversely, the same meaning can often be expressed by different cases or prepositional phrases (case synonymy). There are also a number of expressions functioning as prepositions, e.g. *prilikom* (+ genitive) '*on the occasion of*'.

In Serbian there is a well-developed system of personal verb forms for expressing temporal and modal meanings (the aspect is the classification category); all these forms are polysemous. One of the features of the verb system is that the construct *da* + present tense tend to increasingly supplant the infinitive.

Agreement in gender, number, case and person is one of the characteristic aspects of Serbian syntax, and it is also important for establishing textual cohesion. Categorization of agreement controllers (especially certain types of nouns, constructions with numerals and coordinated noun phrases), as well as the ways this control is expressed in different agreement positions represents an extremely complex area.

The majority of subordinate clauses (especially relative, temporal, conditional and causal) have several formal and semantic subtypes.

In the case of coordinated clauses the inventory of conjunctions for copulative and for adversative relations is especially rich.

Relations between expressions in a text are established by various kinds of textual coordinators and textual connectors. The choice of the order of sentence constituents is important for topic-comment distribution and focus prominence. The so-called zero subject and enclitic pronoun forms are important tools for sentence contextualization.

Orthography (alphabet, orthography type, punctuation, orthographic adaptation of foreign words)

The traditional Serbian alphabet is Cyrillic, which consists of thirty graphemes. Today the Latin alphabet is also increasingly used. It also consists of thirty graphemes (three of them digraphs) which stand in a bijective (one-to-one) relation to Cyrillic graphemes. However, the official alphabet is only Cyrillic.

	Serbian letters														
Cyrillic	А а	Б б	В в	Г г	Д д	Ђ ђ	Е е	Ж ж	З з	И и	Ј ј	К к	Л л	Љ љ	М м
Latin	A	B	V	G	D	Đ	E	Ž	Z	I	J	K	L	Lj	M

	Serbian letters														
	a	b	v	g	d	đ	e	ž	z	i	j	k	l	lj	m
Cyrillic	Н н	Њ њ	О о	П п	Р р	С с	Т т	Ћ ћ	У у	Ф ф	Х х	Ц ц	Ч ч	Џ џ	Ш ш
Latin	N n	Nj nj	O o	P p	R r	S s	T t	Ć ć	U u	F f	H h	C c	Č č	Dž dž	Š š

As to the relation between the graphemic and the phonemic systems, graphemes and phonemes stand in principle in a bijective relation to each other.

At the level of coding schemes, Latin digraphs *lj*, *nj*, *dž* can be coded either as ligatures or as digraphs. In the first case, Unicode[13] provides special codes, for example, for ligatures *LJ*, *Lj* and *lj*, whereas in the second case, as digraphs, they represent a combination of two ASCII codes, for example for *L* and *J*. This can lead to problems in transliteration, which, in general, can nevertheless be performed automatically in the majority of cases. For example, in Serbian Wikipedia each article can be displayed both in Cyrillic and in Latin alphabet.

The Latin alphabet does not envisage the use of Latin characters *q*, *x*, *y*, *w*, nor the use of Latin characters for writing Roman numerals, which can lead to a distortion of the message when a text is transliterated from Latin to Cyrillic. Thus, for example *www* can become *њњњ*, and Latin *Petar II* may become *Петар ИИ* instead of *Петар II*.

Both alphabets are present in contemporary publishing production. According to the data of the National Library of Serbia, a total of 12574 monographs were published in 2010. Out of this number, 6459 were in Cyrillic, 6050 in Latin and 65 in other alphabets. As for daily newspapers with a wider audience, *Politika* and *Večernje novosti* are published in Cyrillic, whereas the majority of other daily newspapers (*Blic, Kurir, Danas*,...) are published in Latin alphabet.

The orthography is of a quasiphonemic type: with a few exceptions, the word is written the same way it is pronounced (the rule: "*Write as you speak*!"), more precisely, according to its phonemic composition.

The punctuation is of a logical, rather than grammatical type (akin to French and English).

According to the orthographic norm, foreign words are written both in Cyrillic and Latin alphabets the way they are pronounced, i.e. they are transcribed. Foreign names are also transcribed (for example, instead of *Shakespeare*, the proper way to write, and pronounce, is *Šekspir*).

Serbian and other languages of Štokavian provenance

The common Štokavian basis, mutual influences and coexistence within a common state and – conceptually – within the common Serbo-Croatian language, resulted in the fact that computational processing of other languages of Štokavian provenance (Croatian,

Bosnia, Montenegrin) has to solve similar problems. This opens great possibilities for synergy, or at least productive cooperation, as well as for a rational and economical approach to solving common problems. It is also supported by the existence of considerable resources for the former common Serbo-Croatian language (grammars and dictionaries), where, truth be told, due attention had not been paid to differences within the Štokavian standard language field. In general, the issue here is not translation from one foreign language to another, but rather *adaptation* of texts composed in standard languages with the same dialectical basis and strongly interconnected in their development. The main problems pertain, in fact, to the phenomena related to the elaboration of the Štokavian core, and especially, the terminology.

Recent developments

The developments at the end of the 20th and the beginning of the 21st century include the following:

Instead of common standard Serbo-Croatian there are now four national standard languages. More specifically, the official language in Serbia is now Serbian, not Serbo-Croatian any more. Due to recent population resettlement provoked by wartime events the dialect picture in Croatia and Bosnia and Herzegovina (in the parts affected by war) has changed.

Increasing changes in lexis and phraseology as well as in terminology can be observed, related to political, social and economic changes in Serbia, its opening towards the world, but also due to harmonization of legal acts, standards and terminology with those existent in the European Union. The influence of English can especially be observed, not only due to cultural and economic factors, which is true for other countries as well, but also due to the fact that in harmonization with the European Union the source texts used are texts in English.

The use of Latin alphabet is increasing (except in official texts).

Texts in Serbian are more and more realized in digital form (use of computers, electronic publishing, internet, SMS-messages).

Language cultivation in Serbia

Work on standardization and cultivation of the language in compliance with the new official language identity

In 1997 an inter-academy and inter-university body was formed as the *Board for Standardization of Serbian*,[14] composed of representatives from relevant institutions from Serbia, Montenegro and the Republic Srpska (in Bosnia and Herzegovina).

Instead of the former Serbo-Croatian standard, the standard of Serbian is now being specified.

There is no purism towards Croatisms (word borrowed from Croatian usage).

A new Serbian orthography was produced.

The use of Cyrillic alphabet is supported, as it is viewed as endangered by the Latin alphabet, especially with younger generations.

Curricula and textbooks in primary and secondary schools are harmonized with the new language situation.

Modernization of the standards

The Board for standardization of Serbian organized the production of a series of descriptive-normative monographs with the aim of presenting the actual state of the language and offering standardized solutions (to date the following topics were processed: word formation, syntax and phonology). A number of standardizing recommendations have been issued. The official orthography has twice been modernized so far.

Cultivation of language usage

The Board for standardization of Serbian (by way of its recommendations), The Society for Serbian Language and Literature (by way of its publications and by organizing Serbian language competitions for students of primary and secondary schools), Serbian Matica (Matica srpska) (by organizing work on the production of orthography, by way of its publications and by organizing round tables and conferences on Serbian language), Vuk's foundation (by way of its publication and by organizing round tables and conferences on Serbian language) and various other institutions, some publishing houses, editorial boards of daily newspapers and editorial boards of radio and TV broadcasts, as well as language experts and mother tongue enthusiasts are endeavoring to contribute to the preservation of the regularity and purity of Serbian in its written and oral usage.

Response to the rising influence of English

A need for substitution of English words by Serbian is emphasized, as well as of calqued translations from English (authentic) by Serbian words and expressions. (In a wider context, the resistance towards the increasing usage of Latin alphabet also belongs here).

Improvement of the situation in the field of lexicography

More and more attention is being given to lexicography, both monolingual and bilingual. A large one volume dictionary of modern Serbian has been published, which was greatly needed.The work on the development of a large academy dictionary of Serbian is being modernized.

European Union laws and regulations are being translated[15] as well as international standards,[16] including terminological standards.

Language in Education

The subject *Serbian Language and Literature* is one of the most important subjects in primary and secondary school. However, the instruction is focused on proper writing and speech, knowledge about the language (grammar and lexis), knowledge about the history of literary (written) languages of the Serbs and about the origin of standard Serbian. Mother tongue competitions (starting from the upper elementary school grades) are directed towards this type of instruction. So, insufficient attention is given to practical use of language and functional literacy.

The wish to bring the goals and standards of instruction closer to the instruction in the European Union, as well as the unsatisfactory results of students on PISA testing, serves as impulses for modernization of language instruction and for putting a greater emphasis on functional literacy and communicational skills. This is being reflected both in the current educational reform (goals of language instruction, standards to be reached, syllabi), as well as in

the improvement of the quality of textbooks. At the university level, there is a general shortage of courses in Serbian that would systematically enable future experts for successful professional communication and appropriate functional literacy.

The application of HLT methods could certainly contribute to the modernization of instruction, for example, by way of computer-assisted language learning (CALL) systems.

International aspects

The use and instruction of Serbian for parts of the Serbian nation living in neighboring countries is regulated by laws of these countries.

The disappearance of the common Serbo-Croatian language and the official existence of distinct languages of Štokavian provenience is reflected in the organization of instruction of the former Serbo-Croatian language, as well as in the names of departments where this instruction had been held: for these languages, hence for Serbian language (and literature) as well, distinct curricula and diplomas now exist, with a greater or lesser combination of subjects, whereas departments have collective names.

The practice of organizing summer schools for foreigners is continued in Serbia, but now for Serbian, not for Serbo-Croatian any more. Domestic experts are also being sent to work as lectors on departments abroad.

Supplementary mother tongue instruction is organized in some countries for children of Serbian origin.

The need for harmonization of legal systems and terminology with those in the European Union, the influence of Anglo-American culture in the field of entertainment and media, as well as the overall atmosphere of globalization, are gradually making Serbian more and more closely linked to other languages, especially English, thus giving a rising impulse and importance to the translation industry.

Serbian on the Internet

A survey[17] from 2010 showed that 50.8% of the population uses the computer and Internet on a regular basis, whereas 43.7% of the population never used a computer. According to another source,[18] as much as 55.9% of the population uses Internet with an increase rate of 926.8% in the period 2000-2010. According to the same source, there were 2,237,680 Facebook users in Serbia on August 31 2010 which represents 30.5% of the total population. Public services (e-government) are used by only 13.2% of the population, whereas 38.5% claimed they would never use such services. Trading via Internet has been used by only 13% of the population. According to the Statistical Office of the Republic of Serbia[19] the usage of ICT equipment shows the following growth:

Households in Serbia had	2006	2010
a computer	26.5%	50.4%
a laptop	1.5%	11.2%
access to the Internet	18.5%	39%

Households in Serbia had	2006	2010
cable TV	30.2%	42.6%
a mobile phone	71.2%	82%

According to the same source, the number of companies using Internet was 96.8% in 2010 (compared to 90.2% in 2006); the number of companies having their own web site was 67.5% in 2010 (compared to 52.9% in 2006). In 2010, 70.6% of them used e-government services.

The data of the Statistical Office of the Republic of Serbia (RZS) from a last year survey on a sample of 2,400 households and the same number of individuals aged from 16 to 74, show that 39% of respondents have an Internet connection, the highest percentage of 51% being in Belgrade.[20] Access to internet is income dependent, as 83% of households with a monthly income over 600 euro have Internet, while for households with monthly income less than 300 euro the percentage decreases to only 29%. The majority of population accesses the global web from desktop computers, one fifth from cell phones, and a little less from laptops.

As for connection type, almost one half of the households in Serbia have an ADSL connection, one quarter have cable internet, whereas 29% of the respondents use mobile devices for connection. In the majority of cases access is from home (84%), then from work, from some other person's home, from school or university, and as little as 3,8% from internet cafes. Students are the most largely represented category on the web, with as much as 95%. Other than for business purposes, internet is most commonly used for e-mail (78%), then for entertainment (games, movies, music – 55%), for reading electronic press (41%) and for learning (23%).

The most popular web sites on the Serbian part of the Internet are Serbian news portals (Blic,[21] B92,[22] Naslovi,[23] RTS[24]). The most visited domestic portal is *Krstarica*[25] which includes a search engine, up-to-date daily news from Serbia, a directory of local sites grouped by topics and a variety of other content. An experiment initiated in 2005 with the introduction of a local search engine *Pogodak*, where the search was adjusted to morphology of Serbian, ended in 2010 as unprofitable.

Serbian Wikipedia represents a source of various language data. It contains a little more than 142,000 articles, an it holds the 28th position[26] in the world regarding the number of articles. The alternative Wikipedia in Serbo-Croatian[27] is smaller and contains about 40,000 articles. Free content language data projects can also be found within the portals *Rastko*,[28] *Antologija srpske književnosti*[29] (Antology of Serbian Literature) and *Transpoetika*[30] where primarily literary texts are stored.

The visibility of a number of pages with content in Serbian has dramatically fallen during 2010, due to the change of the domain from .yu to .rs.

The most commonly used web application is web search, which involves automatic processing of language on multiple levels, as will be described in more detail in the second part of this paper. It involves sophisticated Language Technology, differing for each

language. For Serbian, as we have already mentioned, the problem arises from the relation between Cyrillic and Latin alphabet, ekavian and ijekavian dialects, graphemic variations in the form of the lemma, as well as morphological richness.

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

Selected Further Reading

Enciklopedija Jugoslavije, knj. 6, Zagreb: Jugoslavenski leksikografski zavod, 1990, str. 48-94.

Ivić, Pavle, *Srpski narod i njegov jezik*, Beograd: Srpska književna zadruga, 1971.

Piper, Predrag, *Srpski između velikih i malih jezika*, III izd., Beograd: Beogradska knjiga, 2010.

Popović Ljubomir, Od srpskohrvatskog do srpskog i hrvatskog standardnog jezika: srpska i hrvatska verzija, in: G. Neweklowsky (ed), *Bosanski-Hrvatski-Srpski / Bosnisch-Kroatisch-Serbisch, Aktuelna pitanja jezika Bosnjaka, Hrvata, Srba i Crnogoraca* (=*Wiener slawistischer Almanach*, 57), Wien, 2003. 201-224

Popović Ljubomir, From standard Serbian through standard Serbo-Croatian to standard Serbian, in: Ranko Bugarski and Celia Hawkesworth (eds), *Language in the former Yugoslav lands,* Bloomington, Indiana: Slavica, 2004, 25-40.

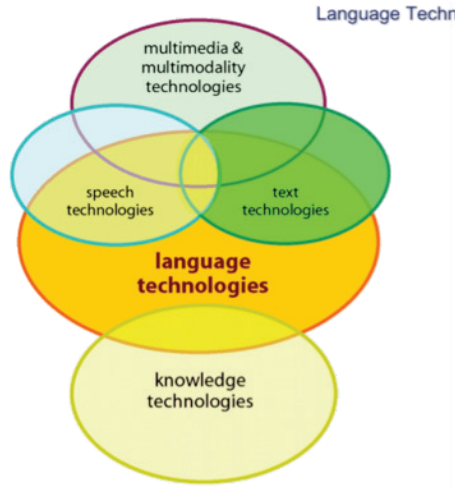
Radovanović, Milorad (ed.), *Srpski jezik na kraju veka*, Beograd: Institut za srpski jezik SANU – Službeni glasnik, 1996.

Cvetana Krstev, *Processing of Serbian – Automata, Texts and Electronic dictionaries* Faculty of Philology, University of Belgrade, Belgrade, 2008.

Language Technology Support for Serbian

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, modificators, 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.

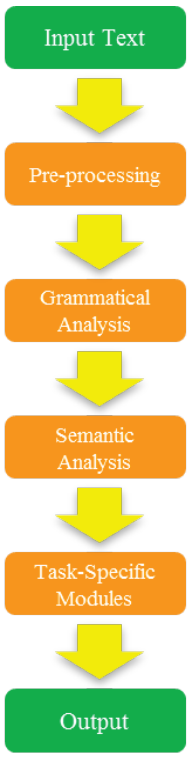


Figure 2: A Typical Text Processing Application Architecture

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 Serbian.

The most important tools and resources involved are underlined in the text and can also be found in the table at the end of the chapter.

Core application areas

Language Checking

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

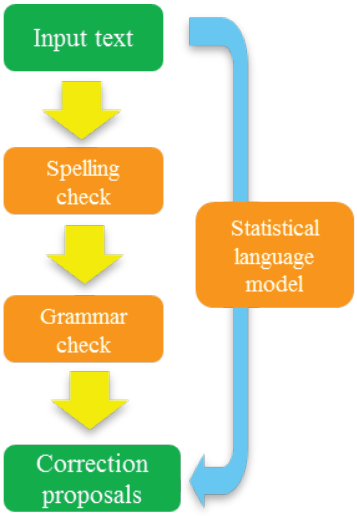


Figure 3: Language Checking (left: rule-based; right: statistical)

Eye have a spelling chequer,

It came with my Pea Sea.

It plane lee marks four my revue

Miss Steaks I can knot sea.

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

Divio se Ruži.

[He admired Rose.]

Divio se ruži.

[He admired the rose.]

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, *Crvena zvezda (name of a football club)* is a much more probable word sequence than *crvena zvezda (red star)*. 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 Serbian with its flexible word order and rich inflection.

The first attempts to develop spell checking software for Serbian dates back to the end of the 1970s[31] motivated by problems confronted by large publishing houses. To date, free spelling checking modules for Serbian are available for OpenOffice[32] on different operating systems, and there exists also a handicraft product, the RAS package,[33] developed by the Srbosof company (individualized installation).

The use of Language Checking is not limited to word processing tools, but it is also applied in authoring support systems. Accompanying the rising number of technical products, the amount of technical documentation has rapidly increased over the last decades. Fearing customer complaints about wrong usage and damage claims resulting from bad or badly understood instructions, companies have begun to focus increasingly on the quality of technical documentation, at the same time targeting the international market. Advances in natural language processing lead to the development of authoring support software, which assists the writer of technical documentation to use vocabulary and sentence structures consistent with certain rules and (corporate) terminology restrictions, but such systems are not available for Serbian. An unfinished experiment along these lines should be noted, which was devised to establish control over the language used in textbooks for elementary and middle schools, with the aim of restricting excessive use of expert terminology.

Besides spell checkers and authoring support, Language Checking is also important in the field of computer-assisted language learning[34] 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.[35] The verbs *guglati/izguglati* are in common use in Serbian. Neither the search interface nor the presentation of the retrieved results has significantly changed since the first version. In the current version, Google offers a spelling correction for misspelled words and also, in 2009, incorporated basic semantic search capabilities into their algorithmic mix,[36] which can improve search accuracy by analysing the meaning of the query terms in context. The success story of Google shows that with a lot of data at hand and efficient techniques for indexing these data, a mainly statistically-based approach can lead to satisfactory results.

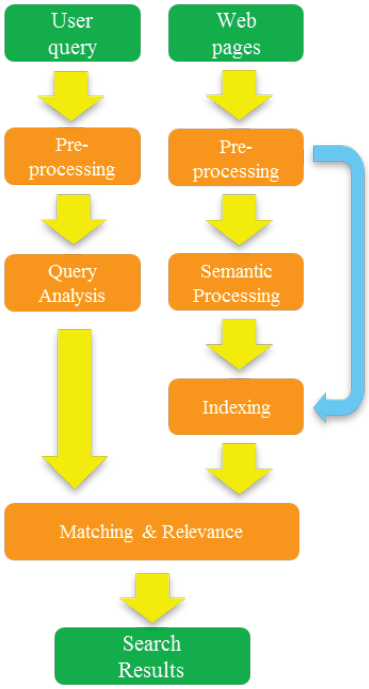


Figure 4: Web Search Architecture

However, for a more sophisticated request for information, integrating deeper linguistic knowledge is essential. In the research labs, experiments using machine-readable thesauri and ontological language resources like WordNet (or the Serbian equivalent SrbNet), have shown improvements by allowing to find a page on the basis of synonyms of the search terms, e.g. *nuklearna energija*, *atomska energija* (nuclear energy, atomic energy) or even more loosely related terms, e.g. *beli luk* and *češnjak* (synonyms for garlic).

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.

Popular sites in Serbia offering search capabilities, such as B92 and Krstarica, mostly rely on Google services.[37] An attempt to introduce a search engine which would perform exclusively a top-down search of the .rs domain, and which would partly be adjusted to the specific features of Serbian, was abandoned in 2010 as unprofitable. A certain number of SMEs is working on the enhancement of search services, albeit mainly for foreign partners and for English.

For research purposes, experiments have been performed with query expansion, by sending queries expanded on basis of morphological dictionaries and multilingual semantic networks to search engines. The experiments yielded interesting and useful results[38] in various domains.

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.

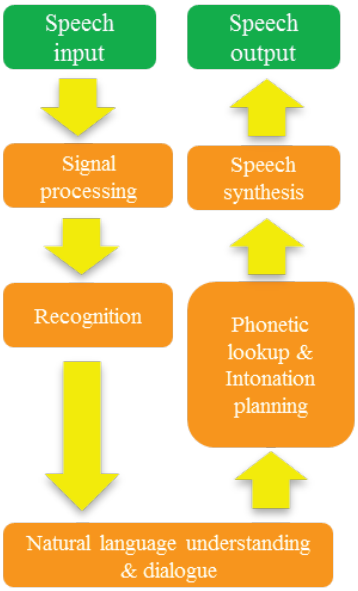


Figure 5: Simple Speech-based Dialogue Architecture

- Dialogue management is required for determining, on the part of the system the user interacts with, which action shall be taken given the user's input and the functionality of the system.
- Speech synthesis (Text-to-Speech, TTS) technology is employed for transforming the wording of that utterance into sounds that will be output to the user.

One of the major challenges is to have an ASR system recognise the words uttered by a user as precisely as possible. This requires either a restriction of the range of possible user utterances to a limited set of keywords, or the manual creation of language models that cover a large range of natural language user utterances. Whereas the former results in a rather rigid and inflexible usage of a VUI and possibly causes a poor user acceptance, the creation, tuning and maintenance of language models may increase the costs significantly. However, VUIs that employ language models and initially allow a user to flexibly express their intent – evoked, e.g., by a 'How may I help you' greeting – show both a higher automation rate and a higher user acceptance and may therefore be considered as advantageous over a less flexible directed dialogue approach.

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

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

The speech synthesis and recognition methods in Serbia (and ex-yu) were developed mainly in electrical engineering environments in cooperation with phonetics experts. These early endeavors were focused on recognition of isolated phonemes. A substantial breakthrough in this area was made by a group from the Faculty of Technical Sciences at the University of Novi Sad, when they developed, in addition to speech databases, a lexical database with more than 4,000,000 accentuated word forms for Serbian and more than 3,000,000 word forms for Croatian. Various applications in the fields of TTS and ASR have been developed based on these resources. Serbian speech recognition and generation has been commercialized by AlfaNum company, a spin-off of the University of Novi Sad. This company is successfully conducting business activities in other ex-yu states as well (Croatia, Macedonia, Bosnia and Montenegro). The AlfaNum company has a considerable number of users within Serbian companies. When translating to Serbian, Google translator also offers an elementary TTS for translation results (albeit without built-in accents).

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

Machine Translation

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

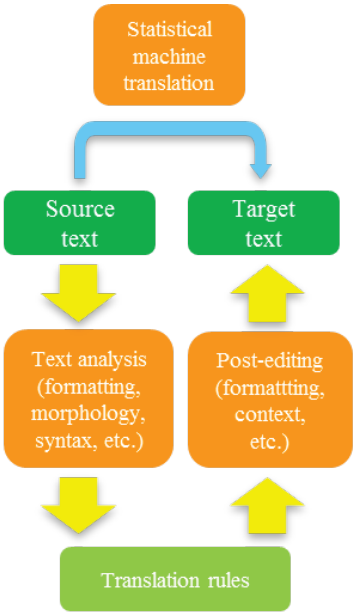


Figure 6: Machine translation (top: statistical; bottom: rule-based)

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

> *Policajac je uspeo da primeti čoveka bez teleskopa.*
>
> *[The policeman managed to notice the man without the telescope.]*
>
> *Policajac je uspeo da primeti čoveka bez revolvera.*
>
> *[The policeman managed to notice the man without the revolver.]*

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

Beginning in the late 1980s, as computational power increased and became less expensive, more interest was shown in statistical models for MT. The parameters of these statistical models are derived from the analysis of bilingual text corpora, such as the Europarl parallel corpus, which contains the proceedings of the European Parliament in 11 European languages. Given enough data, statisti-

cal 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.

When the relation between Serbian and other foreign languages is concerned, the problems depend on the nature of the specific language (whether its morphology is developed or not, whether it has a free or fixed distribution of sentence constituents, whether it possesses an article or not, whether it is written in Cyrillic or Latin alphabet, whether it uses logical or grammatical punctuation, etc.) However, there is not only an issue of problems here, but also of possibilities for cooperation in solving similar problems. In that sense, cooperation with projects related to computational processing of other Slavonic languages is especially useful. However, lexical-terminological relations are also important, namely to what extent a foreign language influenced the elaboration of Serbian. In this field, cooperation should be sought with projects aimed at computational processing of languages which served and are still serving as the backbone for elaboration of Serbian, namely, English, French, German and Russian in the first place.

It should also be added that contrastive research on the relation between Serbian and some foreign languages is also taking place. However, there is unfortunately insufficient cooperation between linguists dealing with Serbian as mother tongue and those, who being experts for foreign languages, engage in contrastive research. Another problem is the insufficient number of large bilingual dictionaries.

The greatest need for LT in Serbia is in the area of translation. There are some specialized associations (e.g. Association of Literary Translators of Serbia, Association of Scientific and Technical Translators of Serbia), some local SMEs (e.g. Elitence and Proverbum) and some foreign companies (e.g. WorldLingo) that offer professional translation services or free, phrase-based machine translation (e.g. Google Translate, WorldLingo). Some of them use proprietary electronic dictionaries in their work, while WorldLingo also offers enhanced machine translation services (web sites, texts, documents, emails, APIs, etc.).

Apart from the well-known freely available Google statistical translation systems which also include Serbian, no other MT systems were produced for Serbian, beside some preliminary work (e.g. done in the scope of the SEE-ERA project) and toy experimental systems.

However, generic statistical MT systems such as Google Translate support Serbian to a considerable degree, especially in translation

from and into English. Nevertheless, for other language pairs, the performance is low and the results far from understandable, sometimes even ridiculous. This is due to the scarcity of parallel corpora that are used to train statistical MT.

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

Evaluation campaigns allow for comparing the quality of MT systems, the various approaches and the status of MT systems for the different languages. Table 1, presented within the EC Euromatrix+ project, shows the pairwise performances obtained for 22 official EU languages (Irish Gaelic is missing) in terms of BLEU score.[39]

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

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

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

Language Technology 'behind the Scenes'

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

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

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

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

Within the aforementioned areas, highly successful experiments for Serbian are underway related to named entity extraction as a part of the information extraction problem. A speedy development of IE and QA is expected, given the extent of developed morphological dictionaries and local grammars.

There are other fields in which linguistic technology is being applied. One of them is plagiarism detection, which uses language-independent technologies, but may be enhanced with search for simple paraphrases of the text. A research along these lines for scientific articles in Serbian has been realized by the CEON company.[40]

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 Serbian higher education system and is largely limited to isolated courses within more general post-graduate study programmes. Paradoxically, despite this state of affairs, short research seminars on topics related to computational linguistics for high school students are organized within the Petnica science centre[41] each year.

At the level of university studies, topics from the field of computational linguistics are present within computer science, electronics, library science, linguistics and psychology studies at the Universities of Belgrade and Novi Sad. Courses offered to students cover the basic concepts of natural language processing, but they are in the function of educating students for other profiles. Within graduate studies at the Faculty of Mathematics in Belgrade, courses in lexical analysis and text mining are offered, in addition to courses providing basic mathematical knowledge necessary in the field of natural language processing (especially statistics, algebra, and logic), whereas a greater choice of courses in the HLT field exist at the level of doctoral studies. The most comprehensive education in the HLT field is offered to students at the Department of library science at the Faculty of Philology, whereas at other departments students take at most one introductory course. Within Serbian language studies education in the field of NLP is not envisaged. Faculties of Philosophy in Belgrade and Novi Sad provide courses in psycholinguistics, where students can get acquainted with methods of statistical text processing. Methods of interest for speech processing are studied at technical faculties. None of the faculties offer a curriculum giving expertise in the field of computational linguistics or language technologies.

Language Technology Programs

Due to various reasons the LT industry in Serbia is relatively undeveloped compared to the leading EU economies. The main driving force behind the development of LT in Serbia are mainly domestic SMEs but also some foreign companies, which sometimes provide support for Serbian language in various LT-related applications. As a national program to support the development of language technologies does not exist, their development and application are often realized in an uncoordinated manner. The introduction of language technologies in Serbia follows at least three different directions (a) through state supported scientific and technology development projects (b) through (mainly) foreign companies which, in addition to computer equipment, also offer some sort of language support and (c) through in-house development within domestic organizations such as publishing houses and translation agencies. Except in rare cases, these three lines of activities are realized independently from each other.

On the other hand, the computer literate population in Serbia is accustomed to using English GUIs even though some of them may not speak English. They often find the localized versions awkward and imprecise, so they are reluctant to use them. The only applications that massively use Serbian GUI are various business, financial and accountant applications including the SAP ERP system. However, there are also some examples of GUI localized by other renowned software vendors like Microsoft (e.g. MS Windows, MS Office), Google or Oracle (Open Office[42]).

Interdisciplinarity has been recognized only in the latest cycle of scientific projects (for the 2011-2014 period) funded by the Ministry of Education and Science. Until 2010 scientific projects (and hence criteria for their evaluation) have been strictly divided among the fields of mathematics (including computer science as its part), language, and technological disciplines. In such a setting it was hard to realize the natural combination of disciplines which form the basis of language technology development. In this context it was necessary to establish connections between research in the field of Serbian language and informatics.

The first project along these lines entitled "Interactions between text and dictionaries" was conceived in 2002 as a joint project of the departments for Serbian at the Faculty of Philology in Belgrade and the Faculty of Philosophy in Novi Sad, as well as the Faculty of Mathematics in Belgrade. In the scope of this project the first corpus of contemporary Serbian was developed,[43] accessible via the web, currently having more than 300 registered users from different domestic and foreign universities and institutes. Development of an electronic morphological dictionary of Serbian following the so called LADL format was also initiated within the scope of this project. The project was later continued as a joint project of the Department of Serbian at the Faculty of Philology and the Faculty of Mathematics in the period from 2006 to 2010 under the name "Theoretical and methodological framework for modernization of Serbian" and from 2011 to 2014 under the name "Serbian and its resources: theory, description and applications". Within the scope of these projects the development of the electronic dictionary of simple words was finalized, and development of the dictionary of compounds initiated, aligned French-Serbian and English-Serbian corpora of literary texts were developed, as well as local grammars for certain segments of Serbian (especially for named entities). Different software tools were also developed, among which special attention should be given to Leximir, a workstation which enables integration and transformation of heterogeneous lexical resources.

In parallel with this research in the field of language, a project was funded within the social sciences field under the name "Fundamental cognitive processes and functions", realized by the Department of Psychology at the Faculty of Philosophy in Belgrade. The aim of this project, amongst others, was to investigate the possibility of automatic annotation of texts based on an annotated corpus,[44] developed during the 1950s and converted to electronic form in the 1990s.

Speech synthesis and recognition is being realized at the Faculty of Technical Sciences of the University of Novi Sad in the scope of projects of technological development from 2005, namely "Development of speech technologies in Serbian and their application in Telekom Serbia" (2005-2007), "Man-machine speech communication" (2008-2010), "Development of dialogue systems for Serbian and other South-Slavonic languages" (2011-2014). They provide

support for different TTS and ASR applications and services including IVR systems, private branch exchanges, call centers, audio logging, track commercials, word spotter, etc.

Other single resources of interest for HLT have been developed within other scientific areas, albeit without any direct interaction with the aforementioned projects. Let us just mention a few examples such as the Serbian-English geological thesaurus[45] and the folkloristic database DABI of the Institute of Balkan studies SASA.[46]

In parallel with national projects, Serbian scientific institutions have also taken part in various international projects related to the HLT field. A certain level of activities was maintained during the UN sanctions due to the participation in projects TELRI I and II.[47] Although Serbian research groups could not participate at that time in the project MULTEXT-East[48] they nevertheless produced useful resources in formats defined by that project: morphosyntactic description of Serbian, the aligned version of Serbian translation of Orwell's *1984*, its lemmatized morphosyntactically tagged version and a comprehensive dictionary covering *1984* lexica.

After the sanctions were lifted, of particular importance was the BalkaNet[49] project which enabled the development of a WordNet type semantic network for Serbian. The Serbian part of the multilingual lexical database of proper names Prolex[50] was developed within the scope of bilateral cooperation with France, whereas a one million large aligned English-Serbian project, lemmatized and morphologically annotated, was developed within the scope of the Intera project. This corpus was used for tagger training, as well as for experiments in alignment at the word level and in automatic translation.

Serbian participants were also involved in two regional projects. One of them was the SEE-ERA.NET - Building Language Resources and Translation Models for Machine Translation focused on South Slavonic and Balkan Languages (ICT 10503 RP, 2007-2008). Its main contribution was the development of unidirectional translation models that rely on large-scale multilingual resources, namely *The Acquis Communautaire*. However, since documents that are the base of this resource were not yet translated to Serbian at that time no translation model was produced Serbian.[51] On its part, the Serbian team contributed by developing another multilingual aligned resource based on Verne's novel *Around the world in 80 days* (in 16 languages at that time). The other project was WISE - An Electronic Marketplace to Support Pairs of Less Widely Studied European Languages (BSEC 009 / 05.2007, 2007 - 2008) with the aim not only to produce cross-lingual lexical resources enriched with linguistic meta-data but also to develop and promote an electronic marketplace for the less widely studies Balkan languages, including Serbian.

Further activities encompass, in the first place, the development of procedures for syntactic analysis of Serbian, which, due to the free order of words and morphological richness, represents and extremely complex task. This means that new resources need to be developed, new types of dictionaries and corpora in the first place, as well as accompanying tools.

Availability of Tools and Resources for Serbian

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

tools and resources is based on educated estimations by several leading experts using the following criteria (each ranging from 0 to 6).

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

7 **Adaptability**: How well can the best tools or resources be adapted/extended to new tasks/domains/genres/text types/use cases etc.?
   - 0: practically impossible to adapt a tool/resource to another task, impossible even with large amounts of resources or person months at hand
   - 6: very high level of adaptability; adaptation also very easy and efficiently possible

Table of Tools and Resources

	Quantity	Availability	Quality	Coverage	Maturity	Sustainability	Adaptability
Language Technology (Tools, Technologies, Applications)							
Tokenization, Morphology (tokenization, POS tagging, morphological analysis/generation)	4	3	5	5	5	4	4
Parsing (shallow or deep syntactic analysis)	1	2	5	3	2	2	2
Sentence Semantics (WSD, argument structure, semantic roles)	0	0	0	0	0	0	0
Text Semantics (coreference resolution, context, pragmatics, inference)	0	0	0	0	0	0	0
Advanced Discourse Processing (text structure, coherence, rhetorical structure/RST, argumentative zoning, argumentation, text patterns, text types etc.)	0	0	0	0	0	0	0
Information Retrieval(text indexing, multimedia IR, crosslingual IR)	3	1	3	3	2	2	3
Information Extraction (named entity recognition, event/relation extraction, opinion/sentiment recognition, text mining/analytics)	1	2	2	2	3	2	3
Language Generation (sentence generation, report generation, text generation)	0	0	0	0	0	0	0
Summarization, Question Answering, advanced Information Access Technologies	1	1	0	1	0	1	1
Machine Translation	1	1	0	1	0	1	1
Speech Recognition	2	2	1	1	1	1	0
Speech Synthesis	2	2	4	4	5	5	1
Dialogue Management (dialogue capabilities and user modelling)	0	0	0	0	0	0	0

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

Conclusions

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

For Serbian, the state of resources and technologies could be described as follows:

- When morphological issues and issues related to them are concerned, it is safe to say that the level of development of technologies and resources is satisfactory, mainly due to the existence of large electronic dictionaries and local grammars. An immediate consequence of this fact is that necessary tools for information retrieval and information extraction are available. Some of the dictionaries are ready for wider use, whereas some need to be upgraded, as for example SrbNet.
- A reference corpus of contemporary Serbian in ekavian dialect is available, as well as several parallel aligned corpora, all of which are available to researchers of Serbian. Current research is focused on upgrading of the reference corpus and its expanding with the ijekavian dialect.

- Speech technologies are well developed, and they have found wide usage in business, but research needs to be further expanded, in order to expand the area of their usability.
- Software aimed at enhancing the productivity of lexicographical work has been developed, but the issue of accepting new technologies in traditionally oriented lexicographic environments is an impediment to speedier development of lexicography.
- Successful experiments have been performed in some areas, such as shallow parsing, summarization, machine translation, ontological resources, in a strictly research environment. However, the results obtained are still far from the level of development reached for developed European languages. The attention of researchers is also attracted by the multimedia and multimodal document, especially in the context of digitization of cultural heritage.
- Given the complexity of Serbian syntax, areas based on deep parsing simply do not exist: sentence semantics, text semantics, and language generation. This results in the absence of a formalized syntax of Serbian and restricts the development of syntactically and semantically annotated corpora. Formalization of Serbian syntax is thus the most urgent task for further expansion of HLT.

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)

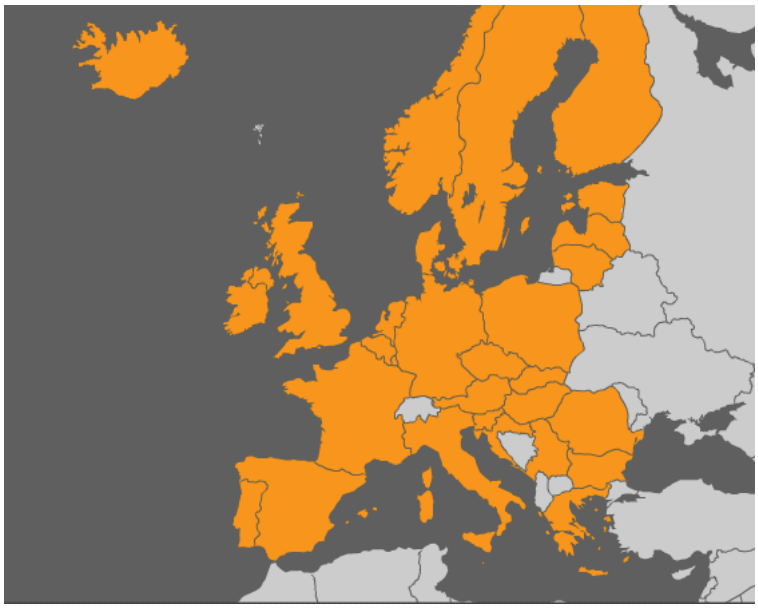


Figure 1: Countries Represented in META-NET

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] The Constitution of the Republic of Serbia from 2006 prescribes: "Serbian language and Cyrillic script shall be in official use in the Republic of Serbia. http://www.srbija.gov.rs/cinjenice_o_srbiji/ustav.php?change_lang=en

[6] The 2002 Census: http://webrzs.stat.gov.rs/axd/Zip/VJN3.pdf

[7] According to the 2005 UNDP Human Development Report for Serbia, ISBN: 86-7728-012-x, p. 33

[8] According to the 2002 census the majority of Serbs abroad live in Germany (102799), then in Austria (87844) and Switzerland (65751).

[9] http://www.ombudsman.rs/pravamanjina/index.php/sr_YU/podaci

[10] http://webrzs.stat.gov.rs/WebSite/repository/documents/00/00/18/48/god2010pog22.pdf

[11] According to the 2005 UNDP Human Development Report for Serbia, ISBN: 86-7728-012-x, p. 69

[12] "Sl. glasnik RS", br. 45/91, 53/93, 67/93, 48/94, 101/2005 - dr. zakon i 30/2010

[13] http://unicode.org/charts/PDF/U0180.pdf

[14] http://en.wikipedia.org/wiki/Board_for_Standardization_of_the_Serbian_Language

[15] http://www.seio.gov.rs/home.50.html

[16] http://www.iss.rs/

[17] http://webrzs.stat.gov.rs/WebSite/Public/PageView.aspx?pKey=204

[18] http://www.internetworldstats.com/europa2.htm#rs

[19] http://webrzs.stat.gov.rs/WebSite/

[20] http://webrzs.stat.gov.rs/WebSite/repository/documents/00/00/10/40/PressICT2010.pdf

[21] http://www.blic.rs/

[22] http://www.b92.net/

[23] http://www.naslovi.net/

[24] http://www.rts.rs/

[25] http://www.krstarica.com/

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

[27] http://sh.wikipedia.org/

[28] http://www.rastko.rs/

[29] http://www.ask.rs/

[30] http://transpoetika.org/

[31] Zoran Urošević: *Statistička metoda otkrivanja i korekcije slovnih grešaka supstitucionog tipa u tekstu na srpskohrvatskom jeziku*, BIGZ, Beograd, 1975.

[32] http://extensions.services.openoffice.org/en/node/1572/releases

[33] http://www.rasprog.com/html/3_0_korektor.html

[34] One of the rare such courses for Serbian was offered at http://www.azbukum.org.rs/index.php

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

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

[37] http://www.alexa.com/topsites/countries/CS

[38] http://www.ncd.matf.bg.ac.rs/casopis/12/NCD12065.pdf

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

[40] http://ceon.rs/index.php?option=com_content&task=view&id=224&Itemid=106

[41] http://www.petnica.rs/

[42] Localization of OpenOffice was funded in the period from 2008 to 2011 by the Ministry for Telecommunications and Information Society through a project at the Faculty of Mathematics: http://ooo.matf.bg.ac.rs/

[43] http://www.korpus.matf.bg.ac.rs/

[44] http://www.serbian-corpus.edu.rs/ns/eindex.htm

[45] http://www.rgf.bg.ac.rs/geolissterm/Index.aspx

[46] http://www.balkaninstitut.com/srp/projekti/sikimic/stratifikacija_balkana.html

[47] http://telri.nytud.hu/

[48] http://nl.ijs.si/ME/

[49] http://cordis.europa.eu/ictresults/index.cfm?section=news&tpl=article&ID=73737

[50] http://www.cnrtl.fr/lexiques/prolex/

[51] Translation of EU legislation is underway, and part of the translated material can be viewed at http://prevodjenje.seio.gov.rs/evroteka/index.php?jezik=srpc