META-NET White Paper Series

Languages in the European Information Society

Croatian

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 for educators, journalists, politicians, language communities and others, who want to establish a truly multilingual Europe.

This white paper is part of a series that promotes knowledge about language technology and its potential. 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 differs 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 regional languages in Europe. The results of this analysis suggests 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 44 research centres from 31 countries who 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. Dr. Marko Tadić, University of Zagreb, Faculty of Humanities and Social Sciences
Prof. Dr. Dunja Brozović-Rončević, Institute of Croatian Language and Linguistics
Prof. Dr. Amir Kapetanović, Institute of Croatian Language and Linguistics

Acknowledgements

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

Table of Contents

Table of Contents 2
Executive Summary 3
A Risk for our Languages and a Challenge for Language Technology 4
Language Borders Hinder the European Information Society 5
Our Languages at Risk 5
Language Technology is a Key Enabling Technology 6
Opportunities for Language Technology 6
Challenges Facing Language Technology 7
Language Acquisition 8
Croatian in the European Information Society 10
General Facts 10
Croatian dialects 12
Standardization of Croatian language 14
Characteristics of the Croatian language 14
Phonetics, phonology, morphonology 14
Morphology 15
Vocabulary, phraseology, terminology 17
Syntax 18
Orthography 18
Onomastics 19
The relationship between the Croatian standard language and other Štokavian-structured languages 19
Linguistic cultivation in Croatia 20
Language in education 21
International aspects 22
Croatian on the Internet 22
Language Technology Support for Croatian 24
Language Technologies 24
Language Technology Application Architectures 24
Core application areas 25
Language Checking 25
Web Search 26
Speech Interaction 28
Machine Translation 29
Language Technology 'behind the scenes' 32
Language Technology in Education 34
Language Technology Programs 34
Availability of tools and resources for Croatian 35
Table of Tools and Resources 37
Conclusions 38
Interpretation of the table of Croatian resources and tools 38
What needs to be done? 38
References 40
META-NET 42
What is the goal? 42
First META-NET Events in 2010 42
Current composition of the META Technology Council 44
Composition of the META-NET Network of Excellence (*: founding members) 44

Executive Summary

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

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

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

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

This whitepaper for the Croatian language demonstrates that a basic language research environment exists in Croatia, although the technology industry is not really developed. Despite the fact that a small number of technologies and resources for Croatian exist, there are fewer of them developed for the Croatian language than for other Slavic languages, e.g. Czech, and far fewer than for the major EU languages, like English, German or French. According to the assessment detailed in this report, focused action must be taken in order to bring the Croatian language resources and tools at the level of quality and quantity of language resources and tools that already exist for other European languages.

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

A Risk for our Languages and a Challenge for Language Technology

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

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

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

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

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

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

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

Language Borders Hinder the European Information Society

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

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

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

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]

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.

Multilingualism is the rule, not an exception.

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.

Challenges Facing Language Technology

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

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.

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: learning examples and learning the underlying language rules.

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.

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 in a similar manner as humans.

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.

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 rulebased system is very time consuming and labour intensive. Rulebased 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.

Croatian in the European Information Society

General Facts

The Croatian language belongs to the West-South Slavic subgroup of the Slavic branch of the Indo-European linguistic family. Currently over 5.5 million people speak Croatian as their native language. The Croatian language consists of the dialects and standard national language of the Croats, which is the official language of more then 4 million people in the Republic of Croatia and is, along with Bosnian and Serbian, one of three official languages in Bosnia and Herzegovina, where it is spoken by about 700,000 people. However, the Croatian language is also spoken by members of national minorities in Croatia as well as by autochthonous Croatian ethnic and linguistic minorities in Serbia, Montenegro, Slovenia, Hungary, Austria, Slovakia and Italy, who either reside upon territories of former Croatian lands or emigrated due to historically conditioned exoduses throughout the centuries. Due to intensive economically and politically conditioned emigration after the two World Wars in the 20th century, Croatian is also spoken within the Croatian linguistic community in a number of other European countries and overseas. The largest Croatian economic diaspora is located in Germany, followed by the USA, Canada and Australia, and they also occasionally use the Croatian language. Their active use of the Croatian language mainly depends on the generation of emigration they belong to. However, in many countries, especially in Europe, there are additional school programs in Croatian organized and financed by the Croatian government.

The official status of the Croatian language in Croatia is defined by the Constitution of the Republic of Croatia. According to the Article 12 of the Constitution: "The Croatian language and the Latin script shall be in official use in the Republic of Croatia. In individual local units, another language and Cyrillics or some other script may be introduced into official use together with the Croatian language and Latin script under conditions specified by law." Since Croatia is expected to join the European Union in 2013, the Croatian language will then become the 24th administrative language of the EU.

In Croatia there is still not a unified "language law" stipulating the usage of Croatian as an official language in public matters. Efforts to introduce a language act have been undertaken on a few occasions since Croatia gained independence, but so far none of them succeeded in gaining the support of the Croatian Government and did not enter parliamentary procedure. The last attempt was made in April 2010. However, certain articles regarding the usage of Croatian as an official state language in official matters are found within acts on education, court procedures etc. So far, legislation states no requirement for a compulsory test or examination as a prerequisite for naturalization. The Citizenship Act[5] presupposes that a foreign person applying for Croatian citizenship is familiar with the Croatian language and alphabet.

According to the 2001 census, Croatia has 4,437,460 residents, of whom 89.63% are Croats. Serbs are the most significant national

minority, comprising 4.54% of the population, while each remaining national minority makes up less than 0.5% of the population: the Bosniaks (0,47%), Albanians (0,34%), Slovenians (0,30%), Montenegrins (0,11%) and others in less significant numbers. The Croatian language is the native language of 96% of all residents. National minorities declared to speak these languages: Albanian, Bosnian, Bulgarian, Czech, Hebrew, Hungarian, German, Istro-Romanian, Italian, Macedonian, Montenegrin, Polish, Roma, Romanian, Russian, Rusyn, Slovak, Serbian, Turkish and Ukrainian. Four minority languages, Serbian, Hungarian, Italian and Czech, have earned the right to the official use of their minority language and script in certain districts according to their share in the population, which must amount to 1/3 of the general population in a local government district. As of 2009, there are 27 districts in Croatia where national minorities have the right to the official use of their language in local administration. That right is used to a high degree in Istarska County, where Italian is the native language of 20,521 residents, but bilingual street signs can be found even in areas where there is no Italian minority. The Republic of Croatia ratified the European Charter for Regional or Minority Languages in 1997.[6]

The recently conducted 2011 census, which was carried out according to international statistical standards, and thus enumerated all citizens of the Republic of Croatia, foreign citizens and stateless persons who reside in the Republic of Croatia, has not yet provided official figures on language usage.

Croatia has a numerous diaspora that often still speaks the Croatian language. Croatian ethnic and linguistic minorities live in many European countries due to historical migrations beginning from the 16th century, as well as recent, mostly economical and political emigration. The most numerous groups are the so-called Burgenland Croats in Austria (presumably about 50,000), and about the same number of Croats live in Hungary. In Austria, the Croats actively use Burgenland Croatian. This variant of Croatian, which has been standardized according to somewhat different rules then standard Croatian, is one of Austria's official minority languages. There are a number of kindergartens and schools in Burgenland that use Burgenland Croatian. On the other hand, the Croatian standard language is the official minority language in Hungary. In Italy at the moment live about 3,000 Croats, who use a variant of Croatian called Molise Croatian that is also taught in schools in three communities inhabited by Croats in Molise. The number of Croats in Serbia, specifically in the province of Vojvodina where Croats are a national minority, is difficult to establish, since a number of ethnic Croats are declared as so-called "Bunjevci", mostly for political reasons. Although many Croats were expelled from Serbia after Croatia gained independence from Yugoslavia, it is assumed that there are still more then 100,000 Croats in Serbia. In other European countries, a Croatian autochthonous minority lives in Montenegro (7,000 to 10,000), the Czech Republic (less then 1,000), Slovakia (4,000) and Romania (7,500). The number of Croats in Slovenia is about 50,000, but only a small number of them are an autochthonous minority, mostly in settlements along the border, and more of them are recent economic emigrants. So, as a minority language, Croatian has the status of an official minority language in Serbia (as one of seven official languages

in the province of Vojvodina), Montenegro, Austria and Hungary, and in Italy a variant of Croatian called Molise Croatian is recognized as a linguistic minority.

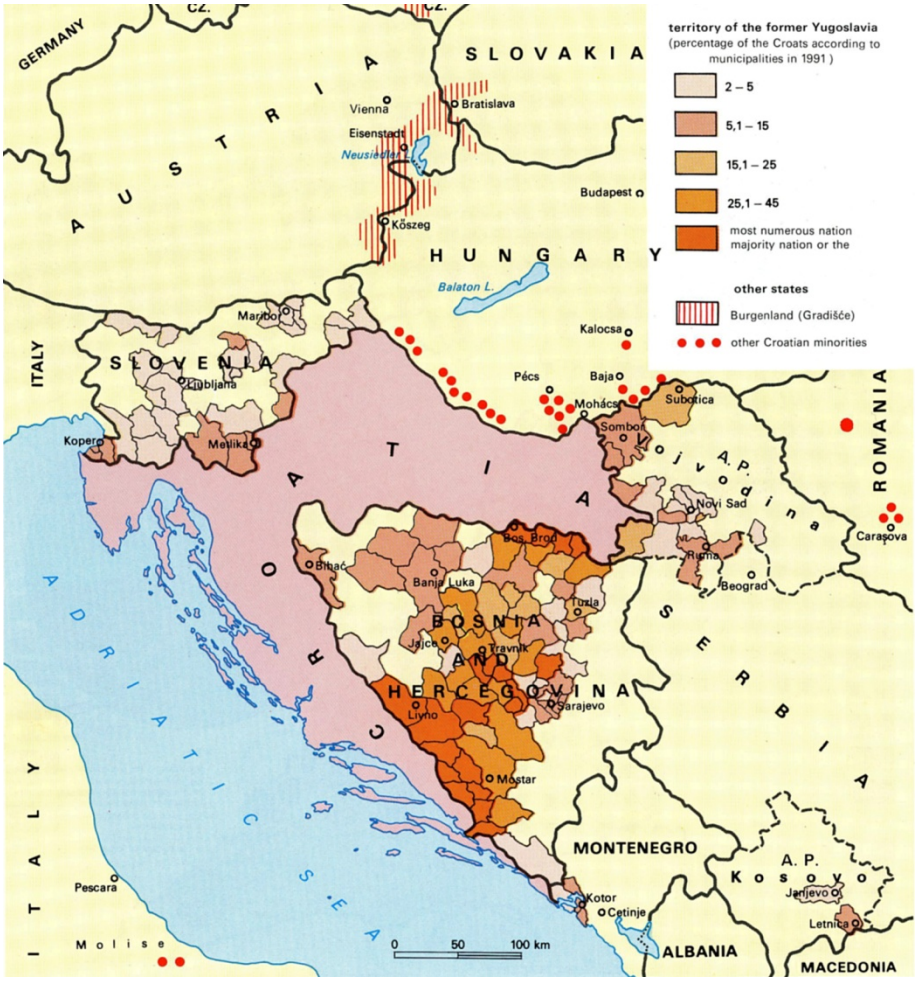


Figure 1: Croats in neighbouring states[7]

Croatian dialects

The dialectal picture of Croatia is composed of three dialectal groups: Čakavian, Kajkavian and Štokavian. Dialects belonging to all three dialectal groups are spoken throughout the Republic of Croatia. All Croatian dialects belong to the Central South Slavic diasystem of the Slavic linguistic branch, and on the South-Slavic territory it comprises part of the dialectal continuum between the Slovenian type in the North-West and the Macedonian-Bulgarian type in the South-East. The names of those dialectal groups are based upon the use of the interrogative pronouns *ča, kaj* and *što* 'what' (lat. *quid*). However, on the South Slavic territory, this classification is relevant only for Croatian dialects and it results from the needs of the Croatian linguistic community. The Slovenes use the pronoun *kaj* but the Slovenian language is not a *Kajkavian dialect*. The Bosniaks, Montenegrins, Serbs, as well as the Bulgarians, Macedonians and Eastern Slavs use *što*, but their languages are not Štokavian dialects in the same sense as the Croatian Štokavian dialect. The Serbs, the Montenegrins and the Bosniaks do not have this pronominal word as a criterion of dialectal

classification. As far as Štokavian dialects are concerned, the archaic šćakavian (the so-called Slavonian) is spoken only by Croats, Neo-Štokavian ikavian and ijekavian-šćakavian is spoken by Croats and Bosniaks, and Neo-Štokavian ijekavian by Croats in some areas in the wider Dubrovnik region, but also by other South Slavic peoples. Croats in Burgenland (Austria, Hungary, Slovakia) mostly speak Čakavian, and rarely the Štokavian or Kajkavian dialects; Croats in the Italian province of Molise speak an archaic Štokavian dialect, and Karaševo Croats in Romania speak a Torlak dialect.

Due to numerous, often forced migrations, the areal distribution of certain Croatian dialects has changed drastically since the Middle Ages. Both Čakavian and Kajkavian were historically distributed throughout a much wider area, but at present the Štokavian dialect prevails. Prior to migrations, the Čakavian dialects were spoken as far North as the rivers Kupa and Sava, and as far east as the Una-Dinara-Cetina line. After migrations, Čakavian dialects were ousted mostly to the coastal regions and islands, while the Čakavian dialects inland began to differ according to the degree of Štokavian influence. The Kajkavian dialects were also once spoken much further to the East, where the Štokavian prevails today.

The Čakavian, Kajkavian and Štokavian dialectal groups differ on all linguistic levels: phonological, morphological, syntactic and lexical, and each level includes a number of archaisms and innovations specific to a particular dialectal group.

Figure 2: Map of Croatian dialects in the Republic of Croatia

Standardization of Croatian language

The millennial history of the Croatian language is attested to by texts written as early as the end of the 10th or the beginning of the 11th century, the period in which the three Croatian dialects (Čakavian, Štokavian, Kajkavian) began to form. All three Croatian dialects played an important part in the formation of the Croatian literary language (various dialectal stylizations) and the moulding of the Croatian linguistic culture that led to the Croatian standard language with a Štokavian foundation.

The first clear trends towards the shaping of the Croatian standard language became apparent in the 17th century, when the majority of the Croatian ethnic community—especially after the grammar and other works of Bartol Kašić (1575-1650) and a flourishing of Renaissance and Baroque literature from Štokavian Dubrovnik—recognised the linguistic structure of the Štokavian dialect (firstly with the ikavian *jat* reflex, but later with the jekavian reflex) as the best starting point for the construction of a supra-regional Croatian literary language. Despite the choice of one linguistic structure in the construction of their standard language, the Croats did not dismiss the achievements of the centuries-old linguistic cultures of various dialectal stylisations within the Croatian literary language (Kajkavian, Štokavian, Čakavian, hybrid) that had marked its history within the Croatian ethnic community. Although the standardisation of the language of the Croats based upon the Štokavian dialect began very early, national linguistic unity was only achieved during the time of the Illyrian national revival (starting in 1835), when smaller groups of Croats who had until then expressed themselves in the Kajkavian idiom also accepted the Štokavian Croatian standard language. Throughout most of the 20th century, the Croatian standard language developed in various South Slavic state units under various names, and was presented as a variant of the so-called Croato-Serbian (Serbo-Croatian) language. This was abandoned during the socio-political changes of 1990.

Did you know that the etymology of the word „cravatte“ ('tie') comes from the „Croatian“ and from French in 17th century it spread to other languages?

Different stylisations of the Croatian language were shaped in diaspora long in the past (e.g. Burgenland Croatian, Molise Croatian).
Croatian written culture is marked by the use of three alphabets (Glagolitic, Cyrillic, Latin), and the Latin script has been the foremost of the three among the Croats since the 16th century. Its usage was neither normed nor systematised until 1835, when Ljudevit Gaj gave the Croatian Latin alphabet its modern-day form.

Characteristics of the Croatian language

Phonetics, phonology, morphonology

The phoneme inventory of the Croatian standard language consists of 5 vowels (*a, e, i, o, u*) and 25 consonants (*m, v, n, l, r, j, nj, lj, p, b, f, s, z, c, t, d, ć, đ, š, ž, č, dž, h, k, g*). The acoustic and articulatory characteristics of the vowels do not change depending on their placement (regardless whether in a short, long, accented or unaccented syllable). In addition to these 5 vowels, there also exist the syllabic *r* (*crn* 'black') and the diphthong *ie*, which is marked in writing as *je/ije* (*djelo, odijelo*).

The prosodic system consists of 4 accents (two long accents with a descending and ascending tone and two short accents with descending and ascending tone) and unaccented post-accentual lengths. The accentual system of the Croatian standard language is neo-Štokavian, although it exists today with many differentiations from the prosodic models codified in the second half of the 19th century. Accent location is not fixed to a specific syllable, but the distribution of accents does have some limitations (e.g. the last syllable of a multi-syllable word cannot in principle be accentuated, descending accents are realised only in the initial syllables of non-compound words). These rules are broken in everyday speech, especially in large urban centres that are not located in Neo-Štokavian regions (e.g. *kontinuitêt* / *kontinuìtēt*). Accent and length can have a differentiating role as they occasionally differentiate the meaning of lexemes or their wordforms, e.g. *gr̂ad* (= 'hail') : *grâd* (= 'town, city'), *žènē* (Gen. sing.) : *žène* (Nom.plur.).

In Croatian some words do not have their own accent (clitic), but in an accentual unit proclitics can carry an accent passed over from an accented word with a descending accent in the initial syllable (*grâd* : *u̍ grād*), while enclitics cannot do this. The transfer of an accent onto a proclitic is becoming ever more rare in everyday speech, especially in urban centres not located in neo-Štokavian regions.

The Croatian standard language is characterised by a number of phonologically (Nom. sing. *sla**d**ak* : Gen. sing. *sla**t**koga*, Nom. sing. *dio* : Gen. sing. *dijela*) and morphonologically conditioned alternations (Nom. sing. *maj**k**a* : Dat. sing. *maj**c**i*, Nom. sing. *juna**k*** : Voc. sing. *juna**č**e*).

Regional implementation of the Croatian standard language is often influenced in speech by dialects located in a given region, e.g. in the Čakavian Kvarner region the prevalence of the plosive *t'* in place of the voiceless africate *ć*, or in the northwestern (Kajkavian) region, the non-differentiation of *č* – *ć* and *đ* – *dž*.

Morphology

The Croatian standard language differentiates between ten parts of speech, of which five inflect (nouns, adjectives, numbers, pronouns, verbs) and four do not inflect (prepositions, conjunctions, particles, exclamations), while adverbs inflect only in comparation.

Grammatical categories that characterise the majority of declinable words are gender (three values: masculine, neuter, feminine), number (two values: singular, plural), case (seven values: nominative, genitive, dative, accusative, vocative, locative, instrumental). Some declinable words have special categories (e.g. definiteness is marked on adjectives with a full set of inflectional endings; animacy is marked by ending in masculine nouns and adjectives; nouns can be concrete, material, categorial or collective etc.). Words that are conjugated (verbs) are characterised by the categories of: manner (four values: indicative, imperative, conditional, optative), person (three values: 1st, 2nd, 3rd), number (two values: singular, plural), voice (two values: active, passive), tense (seven values: present, aorist, imperfect, perfect, pluperfect, future 1, future 2). The verbs *biti* ('to be') and *htjeti* ('to will') are auxiliary in Croatian. Verbs also have complicated aspectual system (imperfective and perfective with additional subvalues such as

inchoativity, iterativity etc.) and they also encode the feature of transitivity. Adjectives and adverbs can take comparative forms (three values: positive, comparative, superlative).

Declension has two main types: noun declension (nouns and indefinite form of adjectives) and pronoun-adjective declension (pronouns, definite form of adjectives, numbers). Each noun gender has its own declension (a-type for masculine and neuter gender, e-type for feminine gender), and there is a special i-type (feminine gender nouns).

Noun declension	N and G singular	N plural
a-type masculine	*opis, opis**a***	*opis**i***
a-type neuter	*sunce, sunc**a***	*sunc**a***
e-type feminine	*žen**a**, žen**e***	*žen**e***
i-type feminine	*noć, noć**i***	*noć**i***

Suffixes for adjective-pronoun declension are shown in this table:

Case	Masculine	Neuter	Feminine
Singular			
N	-i	-o -e	-a
G	-og(a) -eg(a)		-e
D	-om(u/e) -em(u/e)		-oj
A	= N / = G	= N	-u
V	= N	= N	= N
L	-om(u/e) -em(u/e)		= D
I	-im		-om
Plural			
N	-i	-a	-e
G	-ih		
D	-im(a)		
A	-e	= N	= N
V	= N	= N	= N
L	= D		
I	= D		

Words in Croatian are formed by derivation and compounding. There are a few different methods of formation: suffix formation (*star-ac*), prefix-suffix formation (*do-život-an*), compound non-suffix formation (*plačidrug*), compound suffix formation (*vanjskopolitički*), coalescence (*uz-brdo*), formation through compound abbreviations (*Varteks*) and conversion (*mlada*). Suffix formation is the most common.

Vocabulary, phraseology, terminology

The foundational lexical layer of the Croatian standard language, aside from proto-Slavonic lexical heritage, consists of Štokavian vocabulary with an admixture of vocabulary from other Croatian dialects or vocabulary inherited from the literary language of various dialectal stylisations from older periods (e.g. from Kajkavian, *kukac*, *hlače*, *rječnik*, or Čakavian, *spužva*). Aside from this, the Croatian language as a whole bears witness to direct and indirect contact with other cultures. The Croatian language stands out among the remaining South Slavic languages in significant lexical influence received from Romance languages (substrate traces of the Dalmatic language, e.g. *jarbol*, *tunj*). Italian significantly influenced the coastal regions of Croatia (especially the parts formerly under Venetian control), while German and, to an extent, Hungarian influenced the continental part.

The Church Slavonic literary language left traces in older historical periods of the Croatian language, and so it did not present a great influence during the time in which the standard language was being shaped. Russian did not leave as a deep mark on Croatian as it did on the neighbouring Serbian standard language. The influence of the vocabulary of classical languages (Latin and Greek) is omnipresent in Croatian culture, especially in intellectual vocabulary, and scientific terminology. During the middle-Croatian period (16th – 18th century), Turkish loan words intensively entered the Croatian language, especially words related to everyday life. It is interesting to note that Burgenland Croatian, due to early migrations, does not have any Turkish loanwords, not even those that are in standard Croatian no longer perceived as foreign words (e.g. *bubreg, čizma, jastuk,* etc.). In contrast to those loan-words, Burgenland Croatian uses older Croatian words of common Slavic origin and is therefore very important for the history of Croatian lexical inventory. German and French once had an influence on Croatian vocabulary, and in the second half of the 20th century, the influence of English has been ever stronger. The Czech language, although not in direct contact, has had a strong influence on Croatian vocabulary in several episodes, especially in the 19th century in professional terminology enriched by Bogoslav Šulek (e.g. *časopis*, *kisik, dušik, vodik*). During the period of Yugoslavia, Croatian was influenced by the Serbian language, especially because of common federal state administration. Purist tendencies in vocabulary came about occasionally from the 16th to the 20th century (e.g. Zoranić, Ritter Vitezović, Reljković, the period from 1941–1945).

Continuity from ancient times to the modern-day Croatian standard language and the participation of three dialects in the construction of the Croatian standard language can be seen in its well-developed and rich phraseology (e.g., in his 16th century stylised texts, Marulić uses the phraseme *zgubiti glas* = 'to be ashamed, to lose face', while Zoranić uses the phraseme *u magnutje oka* = 'immediately', which are nearly the same as the phrasemes *izgubiti glas* and *u trenu oka* in the Štokavian-structured standard language).
Terminology in specific professional fields began to develop as early as the 16th century, confirmed by the numerous Croatian (mostly multi-lingual) dictionaries compiled from the 16th to the 20th century. In the 19th century, German and Czech had especially strong influence on Croatian terminology, and English has today assumed this role.

Syntax

The Croatian language belongs to a group of languages characterised by an SVO syntactic structure (*Marija voli Ivana*) and relatively free word order (numerous permutations of components are possible with some limitations, such as clitic placement). As concerns the information structure of sentences, it is a basic rule for structuring stylistically unmarked discourse that the first place is taken by the *theme* (old information), which is followed by the *rheme* (comment, new information).

The subject of a sentence does not have to be explicitly stated, and its omission is desirable insofar as it is repeated a number of times within a narrow context. Double-negation is required (*Nitko ga nije volio*). The agreement of components in gender, number and case is typical of Croatian sentence structure.

There are seven cases in the Croatian standard language, and case forms are combined with prepositions (obligatory for the locative case). An important characteristic of Croatian verbs is their aspect while verb forms also express both tense and modal meaning. Sentence organisation can be both coordinated and subordinated (with the aid of conjunctions or without them). A relatively new occurrence in the modern language is the ever-less common use of the Slavonic genitive (*Nije volio vina*), genitive expressions of possession are avoided in favour of possesive adjectives (*majčina kuća* instead of *kuća majke*), and the use of preterite tenses is reduced (imperfect, aorist and pluperfect). In modern Croatian passive constructions are rarer than in the older Croatian language.

Orthography

Although the history of Croatian culture has been marked by the use of three scripts (Glagolitic, Cyrillic and Latin script), the Latin script has been the dominant script used by Croats since the 16th century. The Croatian Latin alphabet was not fully standardized until 1835, when Ljudevit Gaj gave it its current-day form. It is composed of 30 characters, of which three are double characters (*dž*, *lj*, *nj*), and the rest are single characters, of which five have diacritics (*č*, *ć*, *đ*, *š*, *ž*). In academic circles, especially in the printing of texts from Croatian written heritage, the dual-characters *dž*, *lj* and *nj*, are replaced by *ǵ*, *ļ* and *ń* respectively. The characters *q*, *x*, *y*, *w* do not exist in the Croatian alphabet originally, although they are being used for writing foreign names.

Capital letters														
A	B	C	Č	Ć	D	Dž	Đ	E	F	G	H	I	J	K
L	Lj	M	N	Nj	O	P	R	S	Š	T	U	V	Z	Ž
Lowercase letters														
a	b	c	č	ć	d	dž	đ	e	f	g	h	i	j	k
l	lj	m	n	nj	o	p	r	s	š	t	u	v	z	ž

Croatian orthography is phonological-morphonological, since it presents a confluence of two orthographic principles: dominant phonological (e.g. the marking of assimilation) and subordinate

morphonological (e.g. *podcrtati*). Interword separation is logical, and not grammatical (as it once was). It is typical of Croatian orthography that the writing of foreign names is not adjusted to their pronunciation or the graphic inventory of the Croatian alphabet (e.g. *John*, not *Džon* or *Washington*, not *Vašington*).

Onomastics

Croatian names represent important linguistic monuments of the linguistic, cultural and social heritage of the people who created them. Thus, both personal names (anthroponyms) and place names (toponyms) are an important segment of Croatian linguistic culture. The territory of present-day Croatia, roughly bound by the river Drava in the North, the river Danube in the East and the Adriatic Sea in the South, is very picturesquely reflected in its complex stratification of geographical names. The complex stratification of Croatian toponymy reflects centuries of coexistence of the various ethnic groups that have settled on the Eastern coast of the Adriatic and its hinterland throughout history. Centuries of linguistic interpenetration and the merging of various cultural traditions have left an indelible imprint on Croatian toponymy. Furthermore, place names attestations are frequently the oldest witnesses to the oldest changes in the Croatian language itself. Since Croatian developed across religious (pre-christian and christian), cultural and civilisational borders, traces of both East and West have been left on Croatian names. With regards to personal names, Croatians were the first Slavic nation to bear family names (since the 12th century) along the Adriatic coast due to direct Romance cultural influence. The oldest layer of Croatian names is founded upon proto-Slavic name forms that are following common Indo-European name formation patterns. The patronymics form the basis for the largest part of inventory of family names but, unlike in Russian, they are not productive any more and remain unchanged as frozen family names that are incorporated in inflectional system as nouns. In contrast to the Croatian toponomastic system, where we found almost no Turkish influence, many Croatian family names were formed upon Turkish loan-words with Croatian suffixes, since most family names in Croatia were created after the Council of Trent in the 16th century, at the time when a large portion of Croatian lands was under Turkish rule.

The relationship between the Croatian standard language and other Štokavian-structured languages

The four national languages, Croatian, Serbian, and recently Bosnian and Montenegrin, all share Štokavian structural basis, however the traditions and superstructures of these languages are fairly different. What is specific to Croatian linguistic history and culture among other the South Slavic languages is the relationship between its three dialects (Kajkavian, Čakavian, Štokavian), which continually enriches the Štokavian-structured Croatian standard language. Because of different starting points (the non-existence of a basic, common standard) and traditions in language cultivation and standardisation, the disunity of neo-Štokavian structure and differences in linguistic superstructure, one monolithic standard language was never formed during the existence of the Yugoslavian states, although there were serveral

attempts of political imposition of the common name (*Serbo-Croato-Slovenian* during the Kingdom of Yugoslavia; *Serbo-Croatian* or *Croato-Serbian, Croatian or Serbian* during the communist Yugoslavia). During the Second World War and few year later all official documents in Yugoslavia were published in four official languages (Croatian, Macedonian, Serbian, Slovenian), but soon a lot of political effort was put again into convergence of Croatian and Serbian. Despite all attempts to recognise the official existence of Croatian as a language on its own, the forcing of unified terminology, vocabulary, orthography and other linguistic norms in Yugoslavia, led to the official recognition of one standard language (*Serbo-Croatian*) with two variants (*eastern* or *Serbian* and *western* or *Croatian*). The reaction from Croatia came in the form of *Declaration on the Position of the Croatian Language* that openly advocated the recognition of the independent Croatian language and was unanimously signed in 1967 by leading scientific, cultural and educational institutions as well as leading intellectuals throughout Croatia who took a great risk with such an open political move in communist times.

In the past 20 years, the four Štokavian-structured standard languages have developed autonomously as national standard languages in naturally diverging way, and no agreement or coordination exists concerning their norming, which has increased differences between them.

Linguistic cultivation in Croatia

The Croatian Language Council was founded by a decision of the Ministry of Science, Education and Sport taken on 14th April 2005. Its basic task is the systematic and scientific care of the Croatian standard language. The specific tasks of the Council are:

- to tend to the Croatian standard language;
- to discuss current dilemmas and open issues in the Croatian standard language;
- to warn of cases of infractions of the constitutional decree on the position of Croatian as the official language of the Republic of Croatia;
- to promote the culture of the Croatian standard language in written and oral communication;
- to tend to the status and role of the Croatian standard language in light of Croatia's integration into the European Union;
- to make decisions on further standardization processes of the Croatian standard language;
- to take care of language issues and set principles for the orthographic standardization.

The Croatian Language Council meets regularly and draws conclusions after thorough debate. The Institute of Croatian Language and Linguistics hosts the Council, provides technical and administrative support as well as linguistic expertise when necessary.

The Institute of Croatian Language and Linguistics[8] is the central Croatian institution for the research of the Croatian language, and one department of the Institute (the Croatian Standard Language Department) is dedicated to the description of the Croatian standard

language, with special attention paid to linguistic culture (e.g. work on offering linguistic advice to the public and the writing of language handbooks). Advice on proper language usage and linguistic expertise are permanent duties of the Institute. Advices are given by phone, e-mail and in written form. Furthermore, the answers to the most frequently asked questions are available on the Language Advice Portal[9] on the Institute web site.

The Institute's STRUNA project[10], which develops the Croatian professional terminology, deserves a special mention. The goal of this project is to establish a system of coordinating terminological work in all professional fields in Croatia, and in doing so contribute to the improvement of the quality and effectiveness of higher education and scientific research work through the creation of unified and verified terminology that can be used by experts in all fields, as well as by interested participants from the general public. The establishment of a research terminology network and scientific cooperation between institutions that deal in various aspects of terminological work is also planned.

Besides this, other Croatian scientific institutions (several universities with their departments of Croatian language and literature) and cultural institutions (such as *Matica hrvatska*) also take part in the care of the Croatian language. Public media, such as state radio-television and some newspaper publishers, have well-developed proofreading services for the Croatian standard language and pay special attention to the quality of language they use in their public text production.

Language in education

Croatian language is official in all primary and secondary schools, except in regions with national minority residents. However, it is not defined as obligatory for the use at universities. There is a pronounced tendency in Croatia, especially in so-called "hard sciences" to teach in the English language. There were agreeable opinions that it could be functional and useful, but also harmful and unacceptable not to teach in the Croatian language at universities. It would have devastating effects for the development of the Croatian scientific terminology and occupational phraseology. Therefore "The Croatian Language Council" advised the Ministry to legally define the language usage at higher education.

In primary and secondary schools, Croatian language and literature is taught as a subject, and takes up considerable space in the curriculum. As part of this subject, Croatian grammar, vocabulary and literature is studied, and written and verbal expression in Croatian is developed. The PISA test, which tests the skills of pupils at the global level, has been executed in Croatia since 2006, and the first results of testing showed that Croatian 15-year-olds took 26th place of world countries, placed ahead of ten European Union member states and the United States of America.

Besides Croatian, in primary and secondary education it is obligatory to study at least one foreign language from the fourth grade. However, English language (only rarely French or German) is often taught

already in kindergartens. English language is usually the first foreign language in primary education. The most widespread second foreign language is German, then Italian and French. In secondary education Russian and Spanish are occasionally taught as second or third foreign languages. Latin and Old Greek are taught in all classics-program schools that start from the fifth grade of primary school. Furthermore, Latin language is still obligatory in all humanistic secondary schools. In a Jewish minority school (which is open to general public), it is also possible to study Hebrew. Education on minority languages, from the kindergarten level to secondary education, is available and financed by the Croatian government for the Serbian, Czech, Hungarian and Italian minority.

International aspects

The use of the Croatian standard language in countries in the region is regulated by the laws of these countries. The status of the Croatian standard language as one of the official languages of neighbouring Bosnia and Herzegovina is especially important, and so Croatian institutions pay special attention to cooperation with scientific and cultural institutions of the Croatian nation in Bosnia and Herzegovina. The Republic of Croatia's cultural institutions establish cooperation with many Croatian diasporic institutions throughout the world.

Lectures of Croatian language are organised in schools abroad for the children of Croatian citizens who reside either temporarily or permanently in other countries. The Croatian language is taught at many foreign institutions and Slavic studies centres (there are 36 official exchange instructorships for the Croatian language and literature as well as 2 centres for Croatian studies in Australia and Canada in the jurisdiction of and financed by the Ministry of Science, Education and Sport of the Republic of Croatia). A number of centres for the study of Croatian as a second or foreign language operate in Croatia, the best-known of which is *Croaticum*[11].

Croatian on the Internet

According to the statistical information of the Croatian Bureau of Statistics, the use of information and communications technology in enterprises and households looks as follows in terms of percentages:

Usage of information and communication technologies (ICT) in enterprises (%)			
	2008	2009	2010
Computer usage	98	98	97
Internet access	97	95	95
Web site	64	57	61
Usage of financial and banking services	84	84	85
E-government usage	56	61	63

Households equipped with information and communication technologies (ICT) (%)			
	2008	2009	2010
Personal computer	53	55	60
Internet access	45	50	57
Mobile phone	81	82	–

The most-visited Croatian websites are: *net.hr* (a news, sports, entertainment and events portal), *index.hr* (general web portal, info, services, news, sports, entertainment, automotive, gastronomy), *jutarnji.hr* (the website of the daily newspaper "Jutarnji list"), *24sata.hr* (website of the daily newspaper "24 sata"), *tportal.hr* (newsportal of HT, Croatian Telecomm), *njuskalo.hr* ("Njuškalo" advertisments portal), *vecernji.hr* (website of the daily newspaper "Večernji list"), *forum.hr* (the largest Croatian web forum, discussing society, culture, entertainment, etc.). Seven daily Croatian newspapers publish their articles on their own dedicated portals in addition to their paper versions.

The Institute of Croatian Language and Linguistics maintains the web page about the Croatian that features the comprehensive list of mono- and multilingual dictionaries, grammars and orthographies. At the Faculty of Humanities and Social Sciences a similar web page is maintained[12].

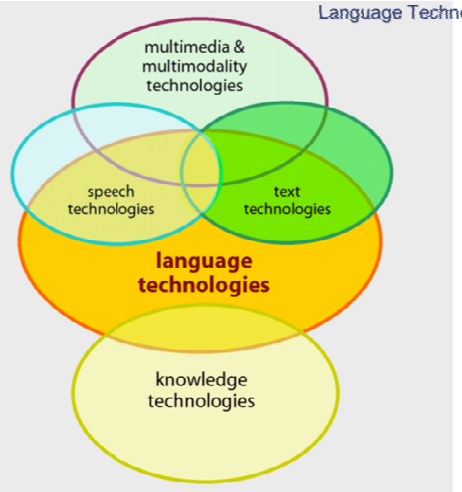
The Croatian-language Wikipedia was founded in 2003 and has 100,708 articles, being the 30rd Wikipedia by number of official articles.

Access to resources in Croatian has been made easier in recent times by Croatian institutions and organisations undergoing the digitalisation process (including significantly projects supported by Ministry of Science, Education and Sports and Ministy of Culture for digitising Croatian cultural heritage) which has increased the visibility of the Croatian language among internet sources.

Language Technology Support for Croatian

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.



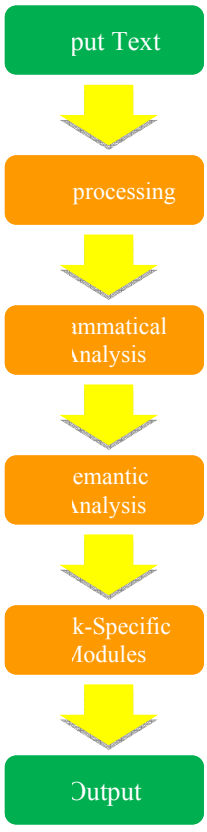
hnology

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

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



After introducing the core application areas, we will give a short overview of the situation in LT research and education, concluding

ssing

with an overview of past and ongoing research programs for Croatian[13]. 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 Croatian.

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):

> *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 an Croatian noun should be written with capital first letter (female personal name) or not (common noun), as in:

> *Slatka je ova višnja.*
> *[This cherry is sweet.]*
>
> *Slatka je ova Višnja.*
> *[This Cherry is sweet.]*

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, *jaz između* ('gap between') is much more probable word sequence than *jaz generacija* ('generation gap'). 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 Croatian because of its flexible word order rich inflection that contribute abundantly to the data sparsness problem in such systems.

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 yet available for Croatian.

Although the research on computational models of inflectional morphology existed in 1980s the first industry-strength spelling checker for Croatian *Hrvatski računalni pravopis* has been published in 1996[14]. Soon after it was bought by Microsoft and today it represents the integral part of Croatian MS Office proofing tools and it is widely used. Other spelling checkers have also been developed by several private companies, but none of them has been so successful. An open source spelling checker for Croatian also exist, it can be used with OpenOffice on different operating systems and is based on Ispell/Aspell. These programs are based on the very large lexicon of correct wordforms which have two drawbacks: 1) strings that represent correct wordforms appearing in a wrong co-text; 2) inability to distinguish betwee real spelling errors and wordforms which are correct, but which are unknown to the lexicon.

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

Web Search

Search on the web, in intranets, or in digital libraries is probably the most widely used and yet underdeveloped Language Technology today. The search engine Google, which started in 1998, is nowadays used for about 80% of all search queries world-wide.[15] Since 2004, the verb *guglati/googlati* and its derivatives *(iz-/na-/pre-/pro-/u-)guglati/(iz-/na-/pre-/pro-/u-)googlati* is used in Croatian, even though it has not made its way into printed dictionaries (even more complex derivatives such as *ugugljiv* 'googlable' are recorded). 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[16], which can improve search accuracy by analysing the meaning of the query terms in context. With the help of this algorithm it also started to cover some of the wordforms in which Croatian lexemes could appear in texts. Unlike in e.g. English nouns where only four wordforms are possible for a noun lexeme (*hand, hand's, hands, hands'*) in Croatian theoretically it can appear in 14 different wordforms, but they are represented on average with 10 different types (*ruka, ruke, ruci, ruku, rukom, rukama...*). Google search engine can retrieve forms like *ruka, ruke*, but *ruci* is still not connected to lemma *ruka*. There is a room for

improvement when Google has to deal with inflectionally rich languages where lexemes appear in many different wordforms. 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 but they also depend heavily on the language structure.

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 its equvalent for Croatian, CroWN), have shown improvements by allowing to find a page on the basis of synonyms of the search terms, e.g. *nuklearna energija* and *atomska energija* (nuclear energy and atomic energy) or even more loosely related terms.

The next generation of search engines will have to include much more sophisticated language technology. If a search query consists of a question or another type of sentence rather than a list of keywords, retrieving relevant answers to this query requires an analysis of this sentence on a syntactic and semantic level as well as the availability of an index that allows for a fast retrieval of the relevant documents. For example, imagine a user inputs the query 'Give me a list of all companies that were taken over by other companies in the last five years'. For a satisfactory answer, syntactic parsing needs to be applied to analyse the grammatical structure of the sentence and determine that the user is looking for companies that have been taken over and not companies that took over others. Also, the expression *last five years* needs to be processed in order to find out which years it refers to.

Finally, the processed query needs to be matched against a huge amount of unstructured data in order to find the piece or pieces of information the user is looking for. This is commonly referred to as information retrieval and involves the search for and ranking of relevant documents. In addition, generating a list of companies, we also need to extract the information that a particular string of words in a document refers to a company name. This kind of information is made available by so-called named-entity recognisers.

Even more demanding is the attempt to match a query to documents written in a different language. For cross-lingual information retrieval, we have to automatically translate the query to all possible source languages and transfer the retrieved information back to the target language. The increasing percentage of data available in non-textual formats drives the demand for services enabling multimedia information retrieval, i.e., information search on images, audio, and video data. For audio and video files, this involves a speech recognition module to convert speech content into text or a phonetic representation, to which user queries can be matched.

For inflectional languages like Croatian, it is important to be able to search for all the inflectional forms of a word at once, instead of having to enter each different form separately. This can be done with the aid of the Croatian Lemmatisation Server that has been developed at the

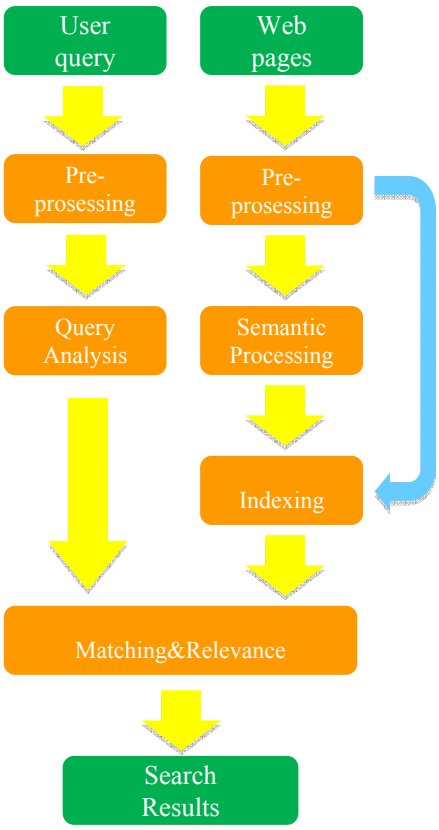


Figure 5: A Web Search Architecture

Department of Linguistics, Faculty of Humanities and Social Sciences at the University of Zagreb and is freely Internet accessible[17] providing an interface to the Croatian Morphological Lexicon, a comprehensive full wordforms database. It contains over 110,000 lexemes yielding over 4 million inflectional wordforms where each entry contains lemma, wordform and full MSD tag and it is MulText East[18] compliant.

In 2009 as a result of a joint Flemish-Croatian project CADIAL[19], the governmental agency HIDRA enabled the public web access to all Croatian legislative documents using the inflectionally sensitive search engine[20]. This engine also enables cross-lingual document retrieval since all documents are indexed with EUROVOC descriptors thus allowing the usage of English EUROVOC descriptors in queries.

Speech Interaction

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

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

- Automatic speech recognition (ASR) is responsible for determining which words were actually spoken given a sequence of sounds uttered by a user.
- Syntactic analysis and semantic interpretation deal with analysing the syntactic structure of a user's utterance and interpreting the latter according to the purpose of the respective system.
- Dialogue management is required for determining, on the part of the system the user interacts with, which action shall be taken given the user's input and the functionality of the system.
- 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 world-wide, with Nuance and Loquendo being the most prominent ones in Europe.

Although the Croatian diphone base was developed within the MBROLA[21] project in 1998 in which Department of Phonetics, Faculty of Humanities and Social Sciences, University of Zagreb participated, up to now, there has been no commercial application of Croatian TTS or ATS systems developed in Croatia. Research in this field has been conducted also at the Faculty of Electrical Engineering and Computing of the same university as well as at the University of Rijeka where a strong group works on the development of resources and tools for speech processing of Croatian.

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

Machine Translation

The idea of using digital computers for translation of natural languages came up in 1946 by A. D. Booth and was followed by substantial funding for research in this area in the 1950s and beginning again in

the 1980s. Nevertheless, Machine Translation (MT) still fails to fulfil the high expectations it gave rise to in its early years.

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

> *Policajac je uočio čovjeka bez teleskopa.*
> [The policeman spotted a man without a telescope.]
>
> *Policajac je uočio čovjeka bez pištolja.*
> [The policeman spotted a man without a pistol.]

One way of approaching the task is based on linguistic rules. For translations between closely related languages, a direct translation may be feasible. 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(s).

Beginning in the late 1980s, as computational power increased and became less expensive, more interest was shown in statistical models for MT. The parameters of these statistical models are derived from the analysis of bilingual text corpora, such as the Europarl parallel corpus, which contains the proceedings of the European Parliament in 21 European languages or JRC-Acquis multilingual parallel corpus[22], the total body of European Union (EU) law applicable in the EU Member States in 23 European languages. Given enough data, statistical MT works well enough to derive an approximate meaning of a foreign language text. However, unlike knowledge-driven systems, statistical (or data-driven) MT often generates ungrammatical output. Still, the current methods do not work equally well for all language pairs. Regarding the European languages, acceptable translations can be obtained for English and the Romance languages, but the quality is downgraded substantially for Germanic, Slavic, Finno-Ugric and Baltic languages[23].

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.

For Croatian, MT is particularly challenging. The free word order and extensive inflection is a challenge for generating words with proper endings that mark grammatical categories of gender, case, number, mood, tense, etc. Also the required agreement in all these categories between e.g. attributes and their nouns or only in number and gender for subject and predicate represent additional challenge.

Several EC co-funded collaborative projects were undertaken for advanced research and development of machine translation for under-resourced languages, including Croatian. The CIP ICT PSP project LetsMT![24] and FP7 project ACCURAT[25] are developing innovative methods for making it easier to gather data for MT and to create customized MT systems for different domains and usage scenarios. In both these projects the group from the Faculty of Humanities and Social Sciences, University of Zagreb is taking part.

The ACCURAT project[26] researches novel methods that exploit comparable corpora to compensate for the shortage of linguistic resources to improve MT quality for under-resourced languages and narrow domains[27]. The ACCURAT project's target is to achieve strong improvement in translation quality for a number of new EU official languages and languages of associated countries (Croatian, Estonian, Greek, Latvian, Lithuanian and Romanian), and propose novel approaches for adapting existing MT technologies to specific narrow domains, significantly increasing language and domain coverage of automated translation.

The LetsMT! project[28] builds an innovative online collaborative platform for data sharing and MT generation. This cloud-based platform provides all categories of users with an opportunity to upload their proprietary resources to the repository and receive a tailored statistical MT system trained on such resources. The latter can be shared with other users who can exploit them further on. The translation services of the LetsMT! project can be used in several ways: through the web portal, through a widget provided for free inclusion in a web-page, through browser plug-ins, and through integration in computer-assisted translation (CAT) tools and different online and offline applications.

Google Translate has offered translations to and from Croatian since 2008. The quality of the translations was rather poor in the beginning, but is getting better as more and more parallel Croatian-English data is available on-line.

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 machine aided translation workflows with term bases and translation memories.

Evaluation campaigns allow for comparing the quality of MT systems, the various approaches and the status of MT systems for the different languages. The 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[29].

	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)

The best results (shown in green and blue) were achieved by languages which 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 are very different from other languages (e.g., Hungarian, Maltese, Finnish).

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.

In 2009 the Croatian Newswire Agency (HINA)[30] started to develop the system for (pre)processing of their news streams that included lemmatisation, named entity recognition[31] and classification, classification of news to a predefined topic schema and keyword extraction. This system was developed jointly by the Faculty of Electrical Engineering and Computing[32] and the Faculty of Humanities and Social Sciences, both from the University of Zagreb.

Two 'borderline' areas, which sometimes play the role of a 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 sentences that contain many important words. Such 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.

None of the technologies discussed in two 'borderline' areas exist for Croatian apart from some experiments have been performed on text summarization[33].

Language Technology in Education

Language Technology is a highly interdisciplinary field, involving the expertise of linguists, computer scientists, mathematicians, philosophers, psycholinguists, cognitive scientists and neuroscientists, among others. Since at the Department of Linguistics, Faculty of Humanities and Social Sciences the Algebraic linguistic approaches have been studied continuously since 1950s, it was easy to introduce in 2005 the Language Technologies topics collected in the special study direction of Computational Linguistics at the two-year Master's programme in Linguistics at the same department. Similar programme was launched at the University of Zadar since 2010.

Language Technology Programs

There are only about 5.5 million people speaking Croatian, and this is not enough to sustain costly development of new products supported from the commercial sources. It costs just as much to build language resources and tools for Croatian as for languages with hundreds of millions of speakers. As a result, the number of commercial companies in the language technology industry in Croatian is close to zero. This role was partially taken by the state, but certainly not to the extent necessary to develop all the resources and tools needed.

In Croatia activities for collecting language resources, i.e. computer corpora, started as early as 1967 when the first computer corpus of Croatian text was collected by Željko Bujas and its concordance produced[34] at the Institute of Linguistics, Faculty of Humanities and Social Sciences of the University of Zagreb. Since then, this institution has become a central institution for corpus linguistics research in Croatia. In 1968 the first usage of computer parallel corpus in contrastive linguistics ever, was led by Rudolf Filipović[35]. The computer processing of old Croatian authors was going on in 1970s and 1980s while the collection of the *One-million corpus of Croatian literary language* started in 1976, lead by Milan Moguš. On the basis of this corpus the first Croatian frequency dictionary was produced[36]. The collection of the Croatian National Corpus[37] started in 1998[38] and it reached 101 million in 2004[39]. Today, the largest Croatian corpus is the hrWaC collected at the same Faculty in 2010 and it reached 1.3 billion tokens crawled from the .hr internet domain. In 2000 at the same Faculty, led by Damir Boras, a large campain of digitalisation of Croatian old mono- and multilingual dictionaries started[40].

Did you know that the oldest Croatian printed dictionary Dictionarium quinque nobilissimarum Europae linguarum Latinae, Italicae, Germanicae, Dalmaticae et Ungaricae *by Faust Vrančić (1595) is also the oldest Hungarian printed dictionary?*

At the Institute of Croatian Language and Linguistics the collection of a comprehensive language corpus *The Croatian Language On-line Repository* (*Riznica*)[41] that includes Croatian written texts from the 11th century onward started in 2004. This Repository is organized into three major corpora (Old Croatian, Middle Croatian, Modern Croatian) where for the first two a substantial problems characteristic for diachronic corpora have to be solved, e.g. transliteration of three different scripts (Glagolitic, Cyrillic and Latin), no standardized orthographies, individual variations in the the usage of certain characters etc.

After the research programmes in 1970s and 1980s, that were typically oriented to literary and linguistic computing, most of research activities in the fields of computational linguistics, corpus linguistics and language technology today is funded by the Ministry of Science, Education and Sports through LT related projects. The first one *Computational Processing of the Croatian Literary Language* started in 1991, and was followed in 1996 by *Computational Processing of the Croatian Language* and in 2002 by *Development of the Croatian Language Resources*. In 2007 three main research programmes oriented to the development of LT for Croatian, encompassing several research projects were funded from the same source:

- *Computational Lingustic Models and Language Technologies for Croatian*[42] where a number of resources and tools are being produced and maintained (e.g. Croatian National Corpus, Croatian-English Parallel Corpus, Croatian Morphological Lexicon, Croatian Dependency Treebank[43], Croatian Wordnet[44], hybrid tagger[45] and lemmatiser[46], dependency parser, NERC system etc.);
- *Sources for Croatian Heritage and Croatian European Identity*[47] with projects dealing with digitisation of old-Croatian dictionaries and building the Croatian valency dictionary[48];
- *Croatian Language Repository*[49] where a number of projects deal with different linguistic problems starting from Croatian dialects and etymological research up to development of semantic networks in building lexical resources. All this projects include digitisation of collected linguistic data thus enriching the pool of available language resources for Croatian.

This programmes opened the possibility to catch up with the level of LT development in other European languages and enabled the participation of Croatian research teams in current FP7 and ICT-PSP projects, since the last one that they participated in (TELRI II) finished in 2002.

From Croatia the Faculty of Humanites and Social Sciences, University of Zagreb is a partner in the CLARIN project—a pan-European effort to create language resource infrastructure for researchers in humanities and social sciences—and Croatia is one of the founding countries of the CLARIN ERIC. The same institution takes part in FP7 project ACCURAT and ICT-PSP projects LetsMT! and CESAR. The University of Zadar is a partner in the ICT-PSP project ATLAS.

Availability of tools and resources for Croatian

The following table provides an overview of the current situation of language technology support for Croatian. 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 sub-languages?
    - 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)	3	2	5	5	3	2	5
Parsing (shallow or deep syntactic analysis)	1	1	2	2	1	0	4
Sentence Semantics (WSD, argument structure, semantic roles)	1	0	1	3	0	0	3
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)	1	0	5	2	3	2	3
Information Extraction (named entity recognition, event/relation extraction, opinion/sentiment recognition, text mining/analytics)	3	1	4	3	2	1	3
Language Generation (sentence generation, report generation, text generation)	1	1	4	0	3	0	0
Summarization, Question Answering, advanced Information Access Technologies	1	0	1	0	0	0	0
Machine Translation	1	0	1	3	0	0	0
Speech Recognition	2	2	3	3	2	3	3
Speech Synthesis	3	3	3	4	4	4	4
Dialogue Management (dialogue capabilities and user modelling)	1	2	1	1	0	2	2
Language Resources (Resources, Data, Knowledge Bases)							
Reference Corpora	3	3	3	4	4	4	2
Syntax-Corpora (treebanks, dependency banks)	1	1	3	4	2	1	2
Semantics-Corpora	0	0	0	0	0	0	0
Discourse-Corpora	0	0	0	0	0	0	0
Parallel Corpora, Translation Memories	3	2	3	3	3	1	2
Speech-Corpora (raw speech data, labelled/annotated speech data, speech dialogue data)	3	1	3	3	4	3	4
Multimedia and multimodal data (text data combined with audio/video)	1	1	4	3	3	3	3
Language Models	0	0	0	0	0	0	0
Lexicons, Terminologies	3	3	4	3	4	3	3
Grammars	0	0	0	0	0	0	0
Thesauri, WordNets	2	3	3	4	3	2	2
Ontological Resources for World Knowledge (e.g. upper models, Linked Data)	0	0	0	0	0	0	0

Conclusions

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

Interpretation of the table of Croatian resources and tools

The table can be summarised in the form of a number of key messages, which highlight crucial issues for the further development of LT for Croatian on the basis of the present situation:

- Croatian stands reasonably well with respect to the most basic language technology tools and resources, such as reference corpora, smaller parallel corpora, large inflectional lexicons, tokenisers, MSD taggers, lemmatisers, NERC system etc.
- However, a large syntactically annotated corpus is missing as well as large parallel corpus (e.g. Croatian translations of Acquis Communautaire). Many of existing resources lack standardization so initiatives are needed to standardize the data and interchange formats.
- Experiments have been conducted in some areas, such as shallow parsing (chunking), summarization, application of ontological resources, but only in an academic research environment. However, the results obtained are far from the level of development that other European languages demonstrate. The multimedia and multimodal document processing, is gaining attraction, particularly the digitisation in the context of preserving the cultural heritage, but language technologies for Croatian are not involved in these processes as needed.
- There exist also individual products with limited functionality in subfields such as speech synthesis, speech recognition and information extraction, and a few others.
- Tools and resources for more advanced language technology such as deep parsing, machine translation, text semantics, discourse processing, language generation, dialogue management, etc., simply do not exist.

What needs to be done?

Public funding for LT in Europe is relatively low compared to the expenditures for language translation and multilingual information access by the USA[50]. In Croatia public funding is even lower than in many other European countries, including neighboring countries Slovenia and Hungary. Although there is a pressing need of recognising the importance of LT in ensuring sustainable development of Croatian in 21st century and in challenges that EU membership will bring with the role of Croatian as one of the EU official languages, no initiative has been launched, that would foster the creation of large-scale resources and tools/services for Croatian, as well as a partnership between government, academia and industry to develop an expertise cluster in Croatian language technology. We believe that this initiative should be institutionally supported by a special-purpose competence

centre that could be funded by the EU Structural Funds in order to stimulate business research and promote sectoral cooperation between companies and research institutions to develop innovative products and technologies to improve the competitiveness of enterprises at EU market that Croatia is about to join as a member state in 2013.

References

Agić, Ž.; Tadić, M.; Dovedan, Z. (2008) Improving Part-of-Speech Tagging Accuracy for Croatian by Morphological Analysis, Informatica 32, 4; pp 445 451.

Agić, Ž.; Tadić, M.; Dovedan, Z. (2009) Evaluating Full Lemmatization of Croatian Texts. In: Klopotek, M.; Przepiorkowski, A.; Wierzchon, S.; Trojanowski, K. (eds) Recent Advances in Intelligent Information Systems, Academic Publishing House EXIT, Warsaw, pp 175-184.

Batnožić, S.; Ranilović, B.; Silić, J. (1996) Hrvatski računalni pravopis (uz računalni program spelling-checker), Matica hrvatska–SYS, Zagreb.

Bekavac, B.; Tadić, M. (2007) Implementation of Croatian NERC system. Proceedings of the Workshop on Balto-Slavonic Natural Language Processing 2007, Special Theme: Information Extraction and Enabling Technologies, ACL, Prag, Czech Republic, pp 11-18.

Brozović Rončević D., Ćavar D. (2008) Hrvatska jezična riznica kao podloga jezičnim i jezičnopovijesnim istraživanjima hrvatskoga jezika. "Vidjeti Ohrid", Proceedings of the 14th international Slavistic Congress in Ohrid, Hrvatsko filološko društvo–Hrvatska sveučilišna naklada, Zagreb, 173-186.

Bujas, Ž. (1974) Osman, kompjutorska konkordancija, Sveučilišna naklada Liber.

Eisele, A.; Xu, J. (2010) Improving Machine Translation Performance Using Comparable Corpora. Proceedings of the 3rd Workshop on Building and Using Comparable Corpora, European Language Resources Association (ELRA), La Valletta, Malta, pp 35-41.

Koehn, P.; Birch, A.; Steinberger, R. (2009) 462 machine translation systems for Europe. Proceedings of the Twelfth Machine Translation Summit (MT Summit XII). International Association for Machine Translation.

Mikelić Preradović, N.; Lauc, T.; Boras, D. (2007) CROXMLSUM – the System for XML Document Summarization in Croatian. International Journal of Mathematics and Computers in Simulation, 1, 1, pp 81-89.

Mikelić Preradović, N. (2010) CROVALLEX lexicon improvements: Subcategorization and semantic constraints. WSEAS transactions on computers, 9

Moguš, M.; Bratanić, M.; Tadić, M. (1999) Hrvatski čestotni rječnik, Zavod za lingvistiku Filozofskoga fakulteta Sveučilišta u Zagrebu – Školska knjiga.

Papineni, K.; Roukos, S.; Ward, T.; Zhu, W.-J. (2002) BLEU: A Method for Automatic Evaluation of Machine Translation. In Proceedings of the 40th Annual Meeting of ACL, Philadelphia, PA, pp 311-318.

Skadiņa, I.; Vasiļjevs, A.; Skadiņš, R.; Gaizauskas, R.; Tufiş, D.; Gornostay, T. (2010) Analysis and Evaluation of Comparable Corpora for Under Resourced Areas of Machine Translation. Proceedings of the 3rd Workshop on Building and Using Comparable Corpora. European Language Resources Association (ELRA), La Valletta, Malta, pp 6-14.

Steinberger, R.; Pouliquen, B.; Widiger, A.; Ignat, C.; Erjavec, T.; Tufiş, D.; Varga, D. (2006) The JRC-Acquis: A multilingual aligned parallel corpus with 20+ languages. Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC2006). Genoa, Italy.

Tadić, M. (1997) Računalna obradba hrvatskih korpusa: povijest, stanje i perspektive. Suvremena lingvistika, 23, 43-44; pp 387-394.

Tadić, M. (2002) Building the Croatian National Corpus, Proceedings of the 3rd International Conference on Language Resources and Evaluation (LREC2002). Las Palmas, Spain, pp 441-446.

Tadić, M. (2003) Jezične tehnologije i hrvatski jezik, Exlibris, Zagreb.

Tadić, M. (2009) New version of the Croatian National Corpus. In: Hlaváčková, D.; Horák, A.; Osolsobě, K.; Rychlý, P. After Half a Century of Slavonic Natural Language Processing, Masaryk University, Brno, pp 199-205.

Vasiļjevs, A., Gornostay, T.; Skadiņš, R. (2010) LetsMT! – Online Platform for Sharing Training Data and Building User Tailored Machine Translation. Proceedings of the Fourth Baltic conference 'Human Language Technologies – the Baltic Perspective'.

META-NET

META-NET is a Network of Excellence dedicated to fostering the technological foundations of a multilingual European information society. For realising applications that enable automatic translation, multilingual information and knowledge management and content production across all European languages, a concerted, substantial, and continent-wide effort in language technology research and engineering is needed. To this end, META-NET is pursuing three lines of actions:

META-NET's three Lines of Action

What is the goal?

A key goal of META-NET is to build the Multilingual European Technology Alliance (META), bringing together researchers, commercial technology providers, private and corporate language technology users, language professionals and other information society stakeholders. META will prepare the necessary ambitious joint effort towards furthering language technologies as a means towards realising the vision of a Europe united as one single digital market and information space.

META-NET booth at LREC (Malta, May 2010)

First META-NET Events in 2010

In the short period of its existence since February 2010, META-NET has already established high visibility in a number of key stakeholder communities. Here is a selection of past events organised by META-NET or with META-NET participation:

- **Language Technology Days 2010** (March 2010, Luxembourg): Presenting META-NET to ca. 250 key representatives from the European Language Technology R&D landscape; networking with new and upcoming projects and initiatives.
- **LREC 2010** (Language Resources and Evaluation Conference, May 2010, Malta): biggest conference in Computational Linguistics and Language Technology with a focus on language resources (ca. 1500 participants). META-NET was present with multiple presentations in the main conference and workshops, with a booth in the EC Projects Village, and as a sponsor.
- **theMETAnk 2010** (June 2010, Berlin): brainstorming meeting with about 120 key Language Technology researchers mostly from academia.

Translingual Europe (Berlin, June 2010)

- **Translingual Europe 2010** (June 2010, Berlin): invitation only industry conference with 150 participants; organised by META-NET.

Current composition of the META Technology Council

Name	Affiliation	Role	Country
Nicoletta Calzolari	Consiglio Nazionale d. Ricerche	Director of Research	Italy
Bill Dolan	Microsoft Research	Head of NLP	USA
Josef van Genabith	Dublin City University, CNGL	Director	Ireland
Yota Georgakopolou	European Captioning Institute	Managing Director	UK, Greece
Gregory Grefenstette	Exalead	Chief Science Officer	France
Jan Hajic	Charles University	Professor	Czech Republic
Theo Hoffenberg	Softissimo	CTO	France
Thomas Hofmann	Google	Dir. Engineering	Switzerland
Keith Jeffrey	ERCIM	President	UK
Stefan Kreckwitz	Across	CTO	Germany
Claude de Loupy	Syllabs	CEO	France
Elisabeth Maier	CLS Communication	CTO	Switzerland
Daniel Marcu	Language Weaver	CTO	USA, Romania
Joseph Mariani	CNRS-LIMSI, IMMI	Director	France
Penny Marinou	EUATC	President	Greece
Jaap van der Meer	TAUS	Director	Netherlands
Roger Moore	University of Sheffield	Professor	UK
Stelios Piperidis	ILSP, Research Centre “Athena”	Head of Department	Greece
Gabor Proszeky	Morphologic	CEO	Hungary
Georg Rehm	DFKI	Senior Consultant	Germany
C.M. Sperberg-McQueen	World Wide Web Consortium	Technical Staff	USA
Daniel Tapias	Sigma Technologies	CEO	Spain
Alessandro Tescari	Pervoice	CEO	Italy
Hans Uszkoreit	DFKI	Scientific Director	Germany
AndrejsVasiljevs	Tilde	CEO	Latvia
Michel Vérel	Vecsys	CEO	France
Alex Waibel	CMU, University of Karlsruhe	Professor	USA/Germany

Composition of the META-NET Network of Excellence (*: founding members)

Country	Member (Affiliation)	Contacts
Austria	Universität Wien	Gerhard Budin
Belgium	University of Antwerp	Walter Daelemans
	University of Leuven	Dirk van Compernolle
Bulgaria	Bulgarian Academy of Sciences	Svetla Koeva
Croatia	Zagreb University, Department of Linguistics	Marko Tadić
Cyprus	University of Cyprus	Jack Burston
Czech Rep.	Charles University in Prague*	Jan Hajic
Denmark	University of Copenhagen	Bente Maegaard, Bolette Sandford Pedersen

Estonia	University of Tartu	Tiit Roosmaa
Finland	Aalto University*	Timo Honkela
	University of Helsinki	Kimmo Koskenniemi, Krister Linden
France	CNRS, LIMSI*	Joseph Mariani
	ELDA*	Khalid Choukri
Germany	DFKI*	Hans Uszkoreit, Georg Rehm
	RWTH Aachen*	Hermann Ney
Greece	ILSP, R.C. "Athena"*	Stelios Piperidis
Hungary	Hungarian Academy of Sciences	Tamás Váradi
	Budapest Technical University	Géza Németh, Gábor Olaszy
Iceland	University of Iceland	Eirikur Rögnvaldsson
Ireland	Dublin City University*	Josef van Genabith
Italy	Consiglio Nazionale Ricerche*	Nicoletta Calzolari
	Fondazione Bruno Kessler*	Bernardo Magnini
Latvia	Tilde	Andrejs Vasiljevs
	University of Latvia	Inguna Skadina
Lithuania	Institute of the Lithuanian Language	Jolanta Zabarskaitë
Luxembourg	Arax Ltd.	Vartkes Goetcherian
Malta	University of Malta	Mike Rosner
Netherlands	Universiteit Utrecht*	Jan Odijk
Norway	University of Bergen	Koenraad De Smedt
Poland	Polish Academy of Sciences	Adam Przepiórkowski
	University of Łódź	Piotr Pezik
Portugal	University of Lisbon	Antonio Branco
	Inst. for Systems Engineering and Computers	Isabel Trancoso
Romania	Romanian Academy of Sciences	Dan Tufis
	University Alexandruloan Cuza	Dan Cristea
Serbia	Belgrade University	Dusko Vitas, Cvetana Krstev, Ivan Obradovic
	Pupin Institute	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
	University Pompeu Fabra	Núria Bel
Sweden	University of Gothenburg	Lars Borin
UK	University of Manchester	Sophia Ananiadou

[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] Article 8 (NN 53/91, 28/91, 113/93, 4/94).
[6] Međunarodne novine 18/97.
[7] A Concise Atlas of the Republic of Croatia, The Miroslav Krleža Lexicographical Institute, Zagreb, 1993, pp 66.
[8] http://www.ihjj.hr.
[9] http://savjetnik.ihjj.hr.
[10] http://struna.ihjj.hr/o-programu.php.
[11] http://croaticum.ffzg.hr/.
[12] http://www.hrvatskijezik.eu.
[13] See also Tadić (2003).
[14] Batnožić et al. (1996).
[15] http://www.spiegel.de/netzwelt/web/0,1518,619398,00.html.
[16] http://www.pcworld.com/businesscenter/article/161869/google_rolls_out_semantic _search_capabilities.html
[17] http://hml.ffzg.hr.
[18] http://nl.ijs.si/MTE.
[19] http://www.cadial.org.
[20] http://cadial.hidra.hr.
[21] http://tcts.fpms.ac.be/synthesis/mbrola.html.
[22] Steinberger et al. (2006).
[23] Koehn et al. (2009).
[24] http://www.letsmt.eu.
[25] http://www.accurat-project.eu.
[26] Skadiņa et al. (2010).
[27] Eisele & Xu (2010).
[28] Vasiļjevs et al. (2010).
[29] The higher the score, the better the translation, a human translator would get around 80. Papineni et al. (2002).
[30] http://www.hina.hr.
[31] Bekavac & Tadić (2007).
[32] http://ktlab.fer.hr.
[33] Preradović Mikelić et al. (2007).
[34] Bujas, Ž. (1974).
[35] Tadić, M. (1997).
[36] Moguš et al. (1999).
[37] http://hnk.ffzg.hr.
[38] Tadić, M. (2002).
[39] Tadić, M. (2009).
[40] http://crodip.ffzg.hr
[41] http://riznica.ihjj.hr.
[42] http://rmjt.ffzg.hr.
[43] http://hobs.ffzg.hr.
[44] http://rmjt.ffzg.hr/p3_en.html.
[45] Agić et al. (2008).
[46] Agić et al. (2009).
[47] http://zprojekti.mzos.hr/page.aspx?pid=97&lid=1.
[48] Mikelić Preradović, N. (2010).
[49] http://zprojekti.mzos.hr/page.aspx?pid=97&lid=1&progID=382&projID=1486.
[50] Gianni Lazzari: „Sprachtechnologien für Europa“, 2006: http://tcstar.org/pubblicazioni/D17_HLT_DE.pdf.