

META-NET White Paper Series

Languages in the European Information Society

– Czech –

Early Release Edition

META-FORUM 2011

27-28 June 2011

Budapest, Hungary



The development of this white paper has been funded by the Seventh Framework Programme and the ICT Policy Support Programme of the European Commission under contracts T4ME (Grant Agreement 249119), CESAR (Grant Agreement 271022), METANET4U (Grant Agreement 270893) and META-NORD (Grant Agreement 270899).

This white paper is part of a series that promotes knowledge about language technology and its potential. It addresses educators, journalists, politicians, language communities and others.

The availability and use of language technology in Europe varies between languages. Consequently, the actions that are required to further support research and development of language technologies also differ for each language. The required actions depend on many factors, such as the complexity of a given language and the size of its community.

META-NET, a European Commission Network of Excellence, has conducted an analysis of current language resources and technologies. This analysis focused on the 23 official European languages as well as other important national and regional languages in Europe. The results of this analysis suggest that there are many significant research gaps for each language. A more detailed, expert analysis and assessment of the current situation will help maximise the impact of additional research and minimize any risks.

META-NET consists of 47 research centres from 31 countries that are working with stakeholders from commercial businesses, government agencies, industry, research organisations, software companies, technology providers and European universities. Together, they are creating a common technology vision while developing a strategic research agenda that shows how language technology applications can address any research gaps by 2020.

META-NET
 DFKI Projektbüro Berlin
 Alt-Moabit 91c
 10559 Berlin
 Germany

office@meta-net.eu
<http://www.meta-net.eu>

Authors

Jarmila Panevová
 Jiří Mírovský
 Barbora Vidová Hladká

Contributors

Ondřej Bojar, Silvie Cinková, Jan Cuřín, Vladislav Kuboň, Karel Oliva, Nino Peterek, Johanka Spoustová, Magda Ševčíková, Ivan Šmilauer, Daniel Zeman and Zdeněk Žabokrtský

Acknowledgements

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

Table of Contents

Executive Summary	3
A Risk for Our Languages and a Challenge for Language Technology.....	5
Language Borders Hinder the European Information Society.....	5
Our Languages at Risk.....	6
Language Technology is a Key Enabling Technology.....	7
Opportunities for Language Technology	7
Challenges Facing Language Technology	8
Language Acquisition.....	8
Czech in the European Information Society.....	10
General Facts	10
Particularities of the Czech Language	11
Recent developments.....	13
Language cultivation in the Czech Republic.....	14
Language in Education.....	15
International aspects	16
Czech on the Internet.....	17
Selected Further Reading	18
Language Technology Support for Czech	19
Language Technologies	19
Language Technology Application Architectures.....	19
Core application areas	20
<i>Web search</i>	<i>20</i>
<i>Language checking.....</i>	<i>21</i>
<i>Speech interaction.....</i>	<i>23</i>
<i>Machine translation.....</i>	<i>24</i>
Information management/“LT behind the scenes“	27
<i>Miscellaneous</i>	<i>30</i>
LT Industry and Programs	30
LT Research and Education	31
Status of Tools and Resources for Czech.....	32
Conclusions	35
Bibliography.....	37
About META-NET	38
Lines of Action.....	38
Member Organisations	40



References..... 43

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

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

A Risk for Our Languages and a Challenge for Language Technology

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

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

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

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

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

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

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

Language Borders Hinder the European Information Society

We cannot precisely know what the future information society will look like. When it comes to discussing a common European energy strategy or foreign policy, we might want to listen to European

foreign ministers speak in their native language. We might want a platform where people, who speak many different languages and who have varying language proficiency, can discuss a particular subject while technology automatically gathers their opinions and generates brief summaries. We also might want to speak with a health insurance help desk that is located in a foreign country.

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

A global economy and information space confronts us with more languages, speakers and content.

Today, we can transmit gigabytes of text around the world in a few seconds before we recognize that it is in a language we do not understand. According to a recent report requested by the European Commission, 57% of Internet users in Europe purchase goods and services in languages that are not their native language. (English is the most common foreign language followed by French, German and Spanish.) 55% of users read content in a foreign language while only 35% use another language to write e-mails or post comments on the web.¹ A few years ago, English might have been the lingua franca of the web—the vast majority of content on the web was in English—but the situation has now drastically changed. The amount of online content in other languages (particularly Asian and Arabic languages) has exploded.

An ubiquitous digital divide that is caused by language borders has surprisingly not gained much attention in the public discourse; yet, it raises a very pressing question, “Which European languages will thrive and persist in the networked information and knowledge society?”

Which European languages will thrive and persist in the networked information and knowledge society?

Our Languages at Risk

The printing press contributed to an invaluable exchange of information in Europe, but it also led to the extinction of many European languages. Regional and minority languages were rarely printed. As a result, many languages like Cornish or Dalmatian were often limited to oral forms of transmission, which limited their continued adoption, spread and use.

The approximately 60 languages of Europe are one of its richest and most important cultural assets. Europe’s multitude of languages is also a vital part of its social success.² While popular languages like English or Spanish will certainly maintain their presence in the emerging digital society and market, many European languages could be cut off from digital communications and become irrelevant for the Internet society. Such developments would certainly be unwelcome. On the one hand, a strategic opportunity would be lost that would weaken Europe’s global standing. On the other hand, such developments would conflict with the goal of equal participation for every European citizen regardless of language. According to a UNESCO report on multilingualism, languages are an essential medium for the enjoyment of fundamental rights, such as political expression, education and participation in society.³

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

Language Technology is a Key Enabling Technology

In the past, investment efforts have focused on language education and translation. For example, according to some estimates, the European market for translation, interpretation, software localisation and website globalisation was € 8.4 billion in 2008 and was expected to grow by 10% per annum.⁴ Yet, this existing capacity is not enough to satisfy current and future needs.

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

- 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.

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

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

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

Multilingualism is the rule, not an exception.

Challenges Facing Language Technology

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

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

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

Language Acquisition

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

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

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

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

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

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

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

Czech in the European Information Society

General Facts

Czech Republic (CR in sequel) consists of three historical parts: Bohemia (Czechia), Moravia and Silesia. The language used in all three „countries“ is Czech, one of west Slavonic languages. Czech language has about 10 million speakers.⁵ In the other parts of the world it has about 200 000 speakers, mostly emigrants and the children of emigrants who left the country in sizable migration waves around World War I and World War II, and the years 1948 and 1968. Many Czech speakers can be found esp. in Austria (mostly in Vienna), Poland, Germany, Ukraine, Croatia (mostly in Daruvar area), in Western Romania (in Banat), in Australia, and Canada. Several tens of thousands of Czechs have continued to live in the Slovak Republic after the split of Czechoslovakia in 1993. However, the largest group of Czech speakers outside lives in United States, in cities like New York, Chicago or Cleveland and in number of communities in Texas, Wisconsin, Minnesota, and Nebraska. According the US Census, more than 90,000 Czech speakers lived in United States in 1990.⁶

Czech language is an official language in CR, since May 2004 it is also one of administrative languages of EU. According the data from 2001 (when the last population census was finished) 5,4 % of citizens belong to the minorities. During administrative, judicial and other official proceedings the standard (literary) Czech is used. The manuals and description of imported goods must contain their Czech translation.

Czech language has several varieties (formations), especially in its spoken form. Standard (literary) Czech is a prestige variety used in school education, and strongly preferred in official negotiating and in mass media. However, the using of standard Czech is not prescribed by any law. For the regulation of language and for the language policy the Institute of Czech Language of Academy of Science of CR is responsible. The regulation is provided after a wide discussion among specialists in languages, and this part of public which is interested in language development (journalists, actors, professional speakers etc.). The Institute of Czech Language prepares the handbooks recommending the codified versions of the orthoepy, orthography, morphology and lexicon. The public is very sensitive as to language changes, esp. in the domain of orthography, therefore the last, very particular changes in orthography were provided in 1993.

In the common communication most people prefer rather other varieties of Czech than a literary Czech. The most spread variety is so-called common Czech (based on the Central Czech interdialect), in Moravia and Silesia the rest of dialects (Hanak, Lach, Czecho-Moravian) are used actively in the spoken form, in Bohemia the traces of Northeast and Southwest dialects can be heard.⁷ The common Czech and dialects differ from its literary variant esp. in morphology, less in the lexicon and pronunciation, the other differences are marginal. All variants of Czech are mutually intelligible. Anyway, so-called code-switching present during the communication by the particular native speakers and dependent on the official and private type of communication, on the speaker's education etc. are often confusing for the foreigners who study Czech.

Czech along with Slovak, Polish, and the Upper and Low Sorbian belongs to the western Slavonic group. However, Czech separated itself from the other Slavonic languages by a number of changes, most of which took place in the 10th through 16th centuries (sound changes such as a' > ě, g > h, r' > ř; in 15th century Czech lost the dual number and two of the Slavic past tenses – the aorist and imperfect); on the other hand the verbal aspect had grown more significant and the number of declensions had increased. For the written form the medieval Latin alphabet was used, later (at the beginning of 15th century) the diacritical markers were introduced by religious reformer Jan Hus („háček“ for the palatal/palatalized consonants – č, ď, ň, ř, š, ť, ž; „čárka“ for long vowels – á, é, í, ó, ú, ý), the only digraph surviving in modern Czech is *ch*, special mark for long u was applied – ů (coming from the chain of changes ó > uo > ů).

Particularities of the Czech Language

Czech language is a highly inflectional language with a very complicated morphology. The noun declension differs 7 cases (Nom, Gen, Dat, Accus, Voc, Loc, Instr), 2 numbers (sg, pl) and 4 genders (masc. anim, masculinanim, fem, neutr); every category has several types of declension (e.g. masc inanim has in school grammars two types of declension “hrad” [the castle] and “stroj” [the machine] with Gen sg „hradu“, „stroje“, respectively. However, some nouns classified as the “hrad” type have in Gen sg the ending –a (“lesa”), some have both endings (bez rybníku [without a pond], do rybníka [to a pond]).

The noun gender is only partially influenced by the natural gender, mostly it is determined by the ending of the lemma (word), though the ending itself is not unambiguous; for foreigners this means to learn the new words together with their gender similarly as in German, where the nouns are accompanied in the lexicon by the determiners “der/die/das” (e.g. “nůž” [the knife] – masc inanim, “mříž” [lattice] - fem, “tabule” [the desk] – fem, “pole” [the field] – neutr).

The morphological component of inflectional language is a necessary prerequisite of any automatic analysis as well as of a language generation and it is obvious that the classification given in school grammars is not sufficient for this purpose. Moreover, the flexion requires not only the connection of a stem with an ending, but many times the morphonemic changes of the stem are part of word-forms process, see e.g. “hoch” (Nom sg), “hoši” (Nom pl) – [the boy/boys], “řeka” (Nom sg) – „o řece“ (Loc sg) [the river], brána (Nom sg), branou (Instr sg) – [the gate], pásek (Nom sg), pásku (Gen sg) – [the belt].

In Czech a rich ambiguity of endings exists (e.g. the ending –a within noun paradigm means *Gen sg masc anim*, *Accus sg masc anim*, *Nom sg masc anim*, *Gen sg masc inanim*, *Nom sg fem*, *Gen sg neutr*, *Nom pl neutr*, *Accus pl neutr*).

The verbal morphology is complicated as well. The formal ambiguity is present e.g. with the form „prosí“ – 3. **sg.** ind. praes./ 3. **pl.** ind. praes. The analytical (complex) verbal forms bring another complication: With the forms such as “psal jsem” [I wrote/I was writing], “psal by” [he would write] the auxiliary verbs “jsem”, “by” behave as clitics moving along the sentence and usually separated from their main verbal forms.

From the other side, the agreement between noun and adjective is a good (and very often sufficient) indication for the solution of ambiguities (“velké stavení” – Nom sg, „velkého stavení“ – Gen sg, „velkému stavení“ – Dat sg [large building], while a single form „stavení“ means *Nom sg, Gen sg, Dat sg, Accus sg, Loc sg, Nom pl, Gen pl, Accus pl.*).

Czech has so-called free word order. It means that the scheme SVO is not obligatory pattern of Czech sentence. The case endings are again a good means helping with identification of subject, (direct) object, (indirect) object and other syntactic functions of the words in the sentence. Compare examples (1), (2), (3)

- (1) *Syn (Nom) poslal matce (Dat) dárek (Accus).*

[lit. The son - sent – mother - a gift]

[The son sent to his mother a gift]

- (2) *Dárek (Accus) poslal matce (Dat) syn (Nom)*

[lit. The gift - sent - mother – the son]

[The son sent to his mother a gift]

- (3) *Dárek (Accus) poslal syn (Nom) matce (Dat).*

[lit. The gift – sent – the son –mother]

[The son sent to his mother a gift]

The noun cases are in all three sentences the same and they allow assigning the syntactic functions to the nouns. These three variants differ as to their information structure in that sense, which information is known and which is introduced as the new one. In ex. (1) a “neutral” word order is used and the sentence fits at the beginning of the text or discourse. In ex. (2) the words “gift” and “mother” are known from the context and the agent of the action (son) is introduced as a piece of new information. In ex. (3) the addressee (“mother”) is focused as a new piece of information.

However, the possibility of word movements in the sentence along with word-form ambiguities is often a great obstacle for a correct parsing of the sentence. Recently, the OVS pattern is often used, esp. in the newspaper headlines and in spoken commentaries, see ex. (4), (5), and (6) where the two meanings are closed to each other, though different. In (5) the ambiguity is multiplied by the lexical ambiguity of the verb in one of the sentence readings:⁸

- (4) *Třicet nemocnic (Nom/Accus) chce zrušit Ministerstvo zdravotnictví (Nom/Accus)*

[lit. Thirty – hospitals – want – to liquidate – Ministry of Health]

[Thirty hospitals want to liquidate the Ministry of Health]

- (5) *Dítě (Nom/Accus) vyzvedne taxík (Nom/Accus)*

[lit. A child – take/lift – a taxi]

[A child will take/lift a taxi]

- (6) *Anna (Nom) představila přítelkyni (Dat/Accus) tchyni (Dat/Accus).*

[lit. Anne – introduce - the girl-friend – the mother-in-law] with

two translations:

[Anne introduced her girl-friend to her mother-in-law/Anne introduced to her girl-friend her mother-in-law]

These ambiguities can be solved only by the semantic and pragmatic knowledge.

The possibility of word-order shifts causes so-called distance dependencies (as in ex. (7)), which represent some troubles for NLP systems. For an automatic syntactic analysis of any type the separation of constituents is difficult to solve:

(7) *Tu knihu se Pavel rozhodl do knihovny vrátit až zítra.*

[lit. This – book – REFL – Paul – decide – to library – to return – only – tomorrow]

[Paul decided to return this book to library only tomorrow]

Recent developments

Though the Czech language preserves 98% of its vocabulary from the Old Slavonic language, it is not insensitive to the influence of other languages.⁹ Till the 19th century the German was the main language in contact (see e.g. words as “knedlík” [der Knödel; the dumpling], “šunka” [der Schinken; the ham], “taška” [die Tasche; the bag], “brýle” [die Brille; the glasses], “blok” [der Block; the block], “cihla” [der Ziegel; the brick], “muset” [müssen; must]).

In 20^{ies} century CR was under the political influence of Russia (Soviet Union), the new words connected with politics and socialistic ideology were adapted for Czech, however, recently they disappeared step-by-step or become unknown to the young generation together with disappearance of the notions and objects they referred to (“kulak” [the rich farmer], “pětiletka” [five years economic plan], “celiny” [great fields], “chozrasčot” [the state economic plan], “prověrka” [the personal checking]).

Recently, the English is the language which influences the lexicon and also phraseology of Czech more and more. The loans in the domain of the sports terminology (“football” [the football], “hokej” [the hockey]) are not quite new, while the IT terminology spreads following the fast development of information technologies and the user’s access to them (“harddisk”, “byte”, “software”, “resetovat”, “flesh disk”, “odlogovat se” etc.). Some of them have Czech equivalents which are used only rarely (“pevný disk” [hard-disk]), some loans are without a Czech counterpart (“software”, “reset”).

The older loaned words were fully adapted in Czech entering the word-formation and word-derivation system (e. g. “weekend” with orthographic variant “víkend” – „víkendu“ (Gen sg), „víkendový“ – adjective), some stand out of Czech grammatical system („prodávájízajezdy all inclusive“ [they sell the trips all inclusive], “nový PR manažér“ [a new PR manager] with pronunciation “pí-ár”).

Occasionally, the whole English phrases are translated word-by-word (so-called “calques”) and they are used as fashionable, e. g. “mějte hezký den” [have a nice day], “opatrujte se!” [take care!]. The names of companies, firms, shops, restaurants and other local proper names often consist of a combination of the Czech and foreign parts (Novodvorská Plaza, Langhans Galerie); while foreign part of these compounds remains without morphological changes, the Czech part is properly inflected, so that an untypical and non-Czech syntactic constructions appear (“navštívil Langhans Galerie”

[he visited Langhans Gallery] instead of normal syntactic construction with the postponed attribute in Nominative (in all cases of this noun phrase) “navštívil Galerii Langhans” [he visited the Gallery Langhans], “šel do Galerie Langans” [he went to the Gallery Langhans].

The young generation uses expressions and phrases to demonstrate that they are “cool” (“houmlesák” [homeless person], “chodí do fitka” [he attends a fitness center], “sorác” [sorry], “lúzn” [looser], “prezoška” [presentation]).

From the other side, the Czech language enriched the international lexicon not only by the word “robot” (used in 30-ies by the famous writer Karel Čapek and his brother Josef in the play R.U.R.), but recently by the words „tunel“, „tunelovat“ with new connotations concerning dishonest economical activities.

The neologisms in Czech language were published in 3 books edited by Olga Martinčová and others (see Martinčová et al., 1998, 2004, 2005).

The facts presented in this section and following the influence of languages in contact are as to the development of Czech vocabulary marginal and they do not represent any danger for the system of Czech scientific terminology.

Language cultivation in the Czech Republic

We have mentioned in the section General Facts that the Institute of Czech Language of Academy of Science of Czech Republic is responsible for the language policy in CR. Its Advisory Department is an important section of this Institute. The experienced linguists working here answer the questions asked by public in written form, by e-mail or by the direct telephone calls. The practical popular handbooks are published as a reaction on a deep interest of civic society in linguistic culture and problems of language planning. We have in mind e.g. handbooks “Na co se nás často ptáte” [The frequent questions for us] (Černá et al., 2002), „Jak používat čárku a další interpunkční znaménka“ (Janovec et al., 2006) [How to put commas and other graphic separators]. The Institute provides special web pages as a supplement service for the public.¹⁰

On the other hand, the language policy in CR is in the main stream far from prescriptive approach. The functional point of view introduced by the members of the classical Prague Linguistic Circle (founded in 1926) continues to describe the language development through the studies of the concrete results of communication acts. The members of the Prague Linguistic Circle demonstrated the impropriety of the purist approach to the language policy based on the principle “correct” vs. “incorrect”. They point out that the stratification of the standard Czech into several varieties (see section General Facts) is a source of a rich choice of a proper variety for a proper situation. The fact that the Czech native speakers does not use the same variety at school or at the official meetings addressing the wide public as in the common talks at home, at shops or during the chats with friends was respected and the research of the talks in different communicative situations is described and presented on the functional basis.

There are many of platforms for the linguistic discussion connected with language policy and language planning as well as of the results of a research (Jazykovědné sdružení ČR [Linguistic Society of Czech Republic], Pražský lingvistický kroužek [Prague Linguistic

Circle], Kruh přátel českého jazyka [The Circle of Friends of Czech Language]), the special interest to the questions of language cultivation is devoted in the journal Naše řeč [Our Speech].

The results of wide discussions are reflected in the normative grammars and in other normative handbooks. The language descriptions in them are formulated as a recommendation for the users who are interested in a cultural way of expression in their native language. Using of normative handbooks is required as obligatory in the elementary and secondary schools by the Ministry of Education, Youth and Sport of CR.

Language in Education

Czech language is an obligatory subject in all types of the elementary and secondary (high) schools; it belongs also to the obligatory subjects for school-leaving examination. However, the subject “Czech language” covers teaching of the language (its grammar and other types of language skills) as well as the literature (including some notions from the literary science). Because there is no subject in school curricula which involves the world literature, a brief survey of it is included under the roof of the subject “Czech language”. The discussion among the specialists in the didactics, psychology, Czech language and literary scientists and the Ministry of Education, Youth and Sport about the separation of Czech language on one side and the Czech and World literature on the other in the school curriculum took place 4-5 years ago. Unfortunately, the discussion failed, and in this respect the situation has not changed. The trend of decreasing of language skills of school children and teenagers is obvious.

In CR there are no serious problems connected with language education of migrants such as in France or Germany. However, the knowledge of Czech language attested by the certificates given by the accredited institutes (e. g. Institute for language and Preparatory Studies of Charles University in Prague, Czech Centers in Berlin, London, Moscow and at the Warsaw University) about applicant’s reaching of the corresponding knowledge of Czech is required for particular professions as well as for the university studies of applicants who apply to study according to the schedule valid for Czech students. The certificate confirms a particular level of the knowledge of Czech language. The levels A1, A2, B1, B2, C1, C2, are defined according to the “Common European Framework of Reference for Languages: Learning, Teaching, Assessment” which was put together by the Council of Europe.¹¹ Level A1 means that the applicant is able to understand Czech in common everyday situations, while the level C2 qualifies the applicant as a person understanding Czech very well and speaking Czech fluently in all situations.

The development of LT is a very good and useful basis for interactive tools for teaching and especially for exercises in language education. Several tools for checking language abilities in Czech were developed, some of them are closely connected with the existence of the annotated corpus of Czech language - the Prague Dependency Treebank (PDT 2.0 in sequel, in details, see Chapter 3, section Core application areas).

The STYX system was proposed and implemented for the exercising of the Czech morphology and syntax. It makes schoolchildren familiar with the Prague Dependency Treebank, the largest annotated corpus of Czech language and it is designed as an electronic corpus-based exercise book of Czech morphology and syntax with

sentences directly selected from the Prague Dependency Treebank. The exercise book offers complex sentence processing with respect to both morphological and syntactic phenomena, i.e. the exercises allow students of elementary and secondary schools to practice the identification of parts of speech, parsing the sentences and classifying syntactic functions of words. The STYX system includes almost 12,000 sentences to practice and tools for viewing these sentences, for composing exercises and practicing. Even more, the STYX system consists of the module 'Capek', a simple annotation editor tailored to schoolchildren to involve them in text annotation. The editor offers a possibility to annotate any sentence, not just those provided by the exercise book.¹²

Another type of tool for teaching and exercising of Czech was developed originally for French students learning Czech as second language. It is called CETLEF¹³ and it is a web-based application featuring fill-in-the-blank exercises on Czech declension where the task of the learner is to fill an inflected form of a given word in a specific syntactic context. The system represents an example of a Computer Assisted Language Learning tool (CALL) using some NLP techniques. First, NLP is used for analyzing learner's productions in order to provide a linguistically motivated feedback on errors. Secondly, it enriches the pedagogical environment with automatically generated linguistic annotation. The idea behind the error diagnosis is that most erroneous forms, existing or not in the language, can be reproduced artificially with a corresponding model of inflection (containing ending paradigms and contextual rules for morphological alternations). Hence they are interpretable in terms of violations of morphological categories. The diagnosis is carried out by matching the learner production with dynamically generated hypothetical word-forms. The most likely interpretations, chosen by a small number of heuristic rules, are used for error tagging and for the generation of the feedback. CETLEF is also used as an alternative source of learner data suitable for a research on second language acquisition. Beside learner's corpora containing mainly students' essays, linguistic productions resulting from grammatical exercises allows focusing on some more specific aspects of the target language, which can be an advantage for the study of acquisition of some complex systems like Czech inflection.

The collection of students' errors in special corpora (so-called students' corpora) is also a promising application of computers during the process of language teaching and learning. The errors are classified according to their sources and the feedback between teachers and students is reflected (see also the webpage of Technical University in Liberec¹⁴).

International aspects

Czech Republic is a small country covering 78,867 km² with a small language – Czech. After the defeat at the battle of White Mountain in 1620 the Literary Czech was under the danger of disappearance due to the German pressure. On the basis of concentrated efforts of Czech writers, poets, translators and teachers during the National Revival it survived. These efforts influenced the shape of the Literary Czech and caused the differences between the norm of literary Czech and its really spoken variants as mentioned above in the section General Facts. However, there were established conditions for the development of rich cultural life: fiction, poems as well as the scientific texts from the different areas were written and published since the end of the 18th century. Many books written in Czech were translated into foreign languages (esp. since the end of

the 19th century). Among many others let us mention here “The Good Soldier Švejk“ (written by Jaroslav Hašek in 1923 and translated into 54 languages), fictions of Karel Čapek and Bohumil Hrabal, to quote the most famous writers in 20ies century. The Czech poet Jaroslav Seifert received the Nobel prize in literature (1984). One of the most famous contemporary writers in the world Milan Kundera, born in CR, wrote his early books in Czech; after his emigration he published his fictions and essays in French.

In the 19th century the Czech botanical and chemical terminology was constituted by J. S. Presl (1791 – 1849). Nowadays, the communication in science in CR undergoes changes characteristic for the globalization of the world and it is influenced by the newly opened possibilities of Czech scholars to have regular contacts with the world of science after 1989. English became the main means of scientific communication. This concerns first of all the technical and natural science; the humanities, esp. the branches concerning Czech history, language and folklore, are not influenced by English so deeply; however, the evaluation criteria of the results of research proposed and applied by the administratively constituted commissions puts a strong pressure on the scientific community to use English.

The discussion about the danger for a small national language to be lost for the process of communication among scholars however has arrived at the conclusion that the Czech language will survive and it will serve as a means of inner communication in the science as well as in other communication areas such as mass-media, economy, law, industry etc.

Czech on the Internet

In the year 2010, almost 60% of Czechs were internet users. Most of them stated to be online every day. Among young people, the proportion of users is even higher. In January 2011, more than 750 thousand .cz domains were registered. These numbers give us a vague idea of the vast amount of Czech language data available on the web.

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

The most commonly used web application is certainly web search, which involves the automatic processing of language on multiple levels, as we will see in more detail in the next. It involves sophisticated language technology, different for each language. For Czech, every language processing system must deal with rich morphology, free word order and various encodings of the diacritical characters or even with lack of the diacritics (especially in blogs or web discussion platforms).

Internet users and providers of web content can also profit from language technologies in less obvious ways, for example if it is used for automatic translation of web contents from one language into another. Considering the high costs associated with manual translation of these contents, it may be surprising how little usable language technology is built in comparison with the anticipated need.

However, it becomes less surprising if we consider the complexity of the Czech language and the number of technologies involved in typical LT applications. In the next chapter, we will present an introduction to language technology and its core application areas as well as an evaluation of the current situation of LT support for Czech.

Selected Further Reading

ČERNÁ, A.. – SVOBODOVÁ, I.. – ŠIMANDL, J.. – UHLÍŘOVÁ, L. (2002): Na co se nás často ptáte. Praha: Scientia

JANOVEC, L. – BUŠOVÁ, L. – ŘÍHOVÁ, A. – ŠAMALOVÁ, M. (2006) : Jak používat čárku a další interpunkční znaménka. Praha: Klett.

KOMÁREK, M. (2006): Studie z diachronní lingvistiky. Olomouc : Univerzita Palackého.

MARTINCOVÁ, O. a kol. (2005) : Neologizmy v dnešní češtině. Praha: Academia.

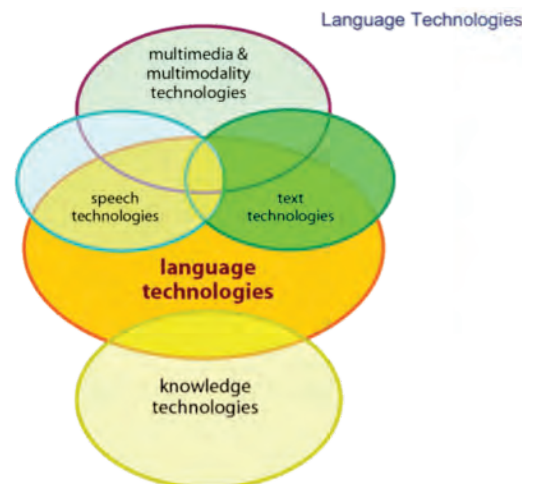
SGALL, P. – PANEVOVÁ, J. (2004) : Jak psát a jak nepsat česky. Praha: Karolinum.

ŠMILAUER, I. (2008). Acquisition du tchèque par les francophones: analyse automatique des erreurs de déclinaison. *The Prague Bulletin of Mathematical Linguistics* (90):33–56.

Language Technology Support for Czech

Language Technologies

Language technologies are information technologies that are specialized for dealing with human language. Therefore these technologies are also often subsumed under the term Human Language Technology. Human language occurs in spoken and written form. Whereas speech is the oldest and most natural mode of language communication, complex information and most of human knowledge is maintained and transmitted in written texts. Speech and text technologies process or produce language in these two modes of realization. But language also has aspects that are shared between speech and text such as dictionaries, most of grammar and the meaning of sentences. Thus large parts of language technology cannot be subsumed under either speech or text technologies. Among those are technologies that link language to knowledge. The figure on the right illustrates the Language Technology landscape. In our communication we mix language with other modes of communication and other information media. We combine speech with gesture and facial expressions. Digital texts are combined with pictures and sounds. Movies may contain language and spoken and written form. Thus speech and text technologies overlap and interact with many other technologies that facilitate processing of multimodal communication and multimedia documents.



Language Technology Application Architectures

Typical software applications for language processing consist of several components that mirror different aspects of language and of the task they implement. The figure on the right displays a highly simplified architecture that can be found in a text processing system. The first three modules deal with the structure and meaning of the text input:

- ❑ Pre-processing: cleaning up the data, removing formatting, detecting the input language, etc.
- ❑ Grammatical analysis: finding the verb and its objects, modifiers, etc.; detecting the sentence structure.
- ❑ Semantic analysis: disambiguation (Which meaning of “apple” is the right one in the given context?), resolving anaphora and referring expressions like “she”, “the car”, etc.; representing the meaning of the sentence in a machine-readable way

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

After the introduction of the core application areas, we will shortly give an overview of the situation in LT research and education. In the end of this section, we will present an expert estimation on the situation regarding core LT tools and resources in a number of dimensions such as availability, maturity, or quality. This table gives a good overview on the situation of LT for Czech.

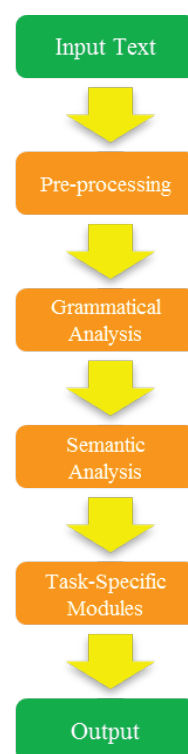


Figure 2: A Typical Text Processing Application Architecture

Core application areas

Web search

In the Czech Republic, there is a long tradition of using local web search engines. The most widely used web search engines are Seznam.cz, Google.com, Morfeo.cz and Jyxo.cz. Therefore, the situation is rather different from other countries, where Google.com has an 80% majority. In the local market, there is enough room both for improving existing search engines through academia-industry collaboration, or for introducing a new one (especially if it would be restricted on a specific domain or specific task, e.g. question answering).

To the best of our knowledge, Google's results are considered to be the most relevant. Google started in 1998 and neither the search interface nor the presentation of the retrieved results has significantly changed since the first version. 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.

However, for a more sophisticated information need, integrating deeper linguistic knowledge is essential. In particular, if a search query consists of a question or a complete sentence rather than a list of keywords, retrieving relevant answers to this query requires an analysis of this question or sentence on a syntactic and semantic level as well as the availability of an index that allows for a fast retrieval of 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“. A simple keyword-based approach will not take us very far here. Expanding the query terms by synonyms, for example using an ontological language resource like WordNet, may improve the results. However, for a satisfactory answer, a deeper query analysis is necessary. For example, applying a syntactic parser to analyze the grammatical structure of the sentence, we can determine that the user is looking for companies that have been taken over and not companies that took over others. We also need to process the expression “last five years” to find out which years it refers to.

For Czech, the sentence analysis task is rather complicated, because we must deal with rich morphology and free word order. The local search engines have already incorporated some kinds of morphological analyses into their systems, but their quality varies.

Finally, the processed query needs to be matched to a massive amount of unstructured data in order to find the piece or pieces of information the user is looking for. This involves the retrieval 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 tagged using a named-entity recognizer.

We face an additional challenge if we want to match a query to documents written in a different language. For multilingual search, we have to automatically translate the query to all possible source languages and map the retrieved information back to the target language. Again, this requires a linguistic analysis of all texts involved. For users with a very specialized information need, an expansion of the query may require additional knowledge resources

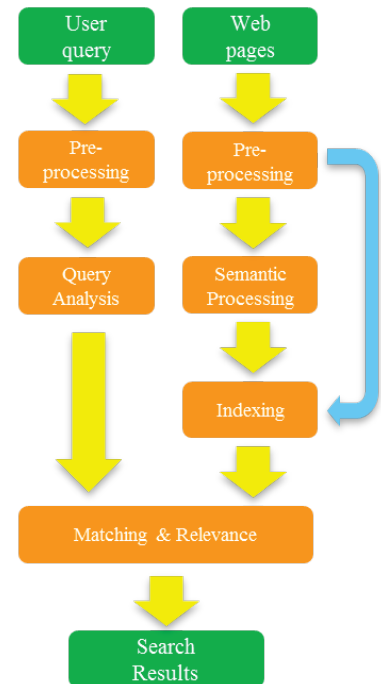


Figure 4: Web Search Architecture

like a domain-specific ontology, representing the concepts relevant within the domain and the relationships between those concepts.

The increasing share of data available in non-textual format also drives the demand for services enabling multimedia search, 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.

Language checking

The morphological and syntactic properties of Czech constitute a great challenge for both spelling and grammar checking. Although the corresponding tools already exist for both kinds of checking (first spelling checkers date back to early 1990's, the development of a first grammar checker for Microsoft Office took much longer time, it has been included as late as in 2005), there are still many issues waiting for an efficient solution.

The existing spelling checkers for Czech are based on a dictionary of basic word forms (lemmas) combined with a set of morphological rules enabling the analysis or generation of all correct word forms. Although this simple approach seems to be satisfactory, it has two substantial drawbacks. The first issue concerns the spelling errors which are actually correct word forms appearing in a wrong context. Due to the isolated handling of individual word forms it is virtually impossible to discover such errors; some more advanced error detection algorithms would definitely be useful. The second drawback is the inability to distinguish between real spelling errors and word forms which are correct, but which are not contained in the dictionary. Such words will always exist due to the natural enhancement of a lexicon by newly created words, by new scientific or technical terms etc. The ability to capture this distinction would bring the spelling checkers to a new level.

Some attempts to make the spelling checking more context sensitive have already been made in the past. For example, one of the most frequent errors in Czech is a wrong use of a personal pronoun *já* [I] in the genitive, dative, accusative and locative cases. The forms used in those cases, namely *mě* [gen., acc] and *mně* [dat., loc.] cannot be distinguished in a spoken language and thus many people use them incorrectly in a written text. To determine a proper form with regard to a given context automatically, theoretically requires a complete analysis of a given sentence, because the proper case usually cannot be determined without taking into account syntax and/or verbal valency. This makes it a challenge even for sophisticated grammar checkers.

The solution implemented in Microsoft Word exploits the fact that these pronouns almost exclusively directly follow a preposition. If this preposition requires using only a particular case (like, e.g., the preposition *k(to)*), then the correct pronominal form can be determined with almost 100% reliability just on the basis of this local context. The Autocorrect tool of MS Word thus contains a simple list of prepositions with this property followed by a proper pronominal form.

This example demonstrates that a further research of similar morphological or syntactic properties may improve the quality of a contextual spelling checker quite substantially.

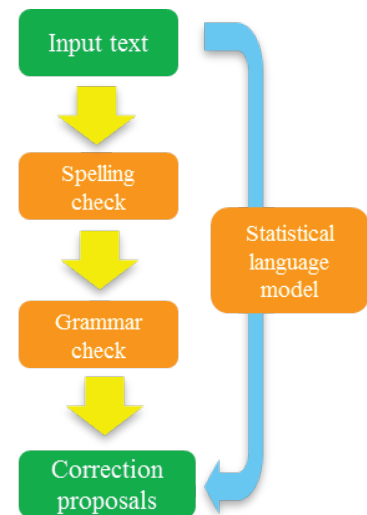


Figure 3: Language Checking (left: rule-based; right: statistical)

Czech proves to be even more difficult for grammar checking than it is for spelling checking. As a language with a great degree of word-order freedom it makes it difficult to apply error pattern checking, a standard method used for languages with stricter word-order like English. The order of words in a Czech sentence is not completely arbitrary (for example, Czech clitics always occupy a sentence-second position), but in some cases it is even possible to tear apart an adjective from a nominal group and put it almost anywhere in a given sentence, like, e.g., in a frequently used example *Vánoční nadešel čas* [*Christmas(adj.) came time / A Christmas time has arrived*]. Such constructions, called long-distance dependencies, or non-projective constructions, constitute a huge challenge for any grammar checker. The investigation of the Prague Dependency Treebank, a syntactically annotated corpus of Czech, shows that about 14% of sentences in the corpus contain at least one non-projective construction. This number clearly demonstrates that this phenomenon cannot be ignored.

These constructions make the grammar checking even more difficult for one more reason. The fact that a dependent word may be located very far from its governor also blurs the distinction between correct and incorrect sentences. Let us demonstrate this fact by means of the following example:

Které děvčata chtěla dostat šaty?

[*Which girls wanted to_get dresses?*]

This sentence may either be understood as a syntactically correct, but non-projective sentence which can be translated as *Which dresses the girls wanted to get?* or as syntactically incorrect, but projective sentence *Which girls wanted to get dresses?* (This reading is syntactically incorrect because of the wrong form of the interrogative pronoun *Které* [Which] – the proper form being *Která*.) To resolve this ambiguity is virtually impossible – non-projective constructions are an integral part of Czech language and their presence in a sentence does not indicate anything unusual.

The syntactic complexity of non-projective constructions in Czech is even higher than the previous example may suggest. A simple Czech sentence may contain more than one such construction, as, e.g. the sentence:

Tuto knihu jsem se mu rozhodl dát k narozeninám.

[*Lit.: This book I_am myself to him decided to_give to birthday.*]

[*I decided to give him this book to his birthday.*]

The pattern present in this example, namely the combination of a finite and infinite verb with intertwined dependent constituents, is very productive in Czech and it theoretically allows for an unlimited number of non-projective constructions in a single clause.

Although non-projective constructions constitute a great challenge for grammar checking because they make simple error-pattern based methods insufficient, they are not the only syntactic challenge in Czech. At least equally important seems to be another syntactic property of Czech – its ability to drop a subject of a subsequent sentence if it is clear from the context what the subject would be. Let us present yet another example:

Sportovci házely plyšáky.

[Sportsmen were throwing cuddle-bears.]

This sentence is syntactically incorrect if it is isolated due to a gender disagreement between the subject (*sportovci* – masc. anim.) and the predicate (*házely* – fem. or masc. inanimate). The situation changes, if we change the context in the following way:

Dívky křičely. Sportovci házely plyšáky a rozhodčím shnilá rajčata.

[Girls shouted. They were throwing cuddle-bears to the sportsmen and rotten tomatoes to referees.]

The fact that the subject may be omitted from a sentence makes it extremely difficult for existing grammar checkers of Czech to discover one of the most frequent types of grammatical errors, the errors in subject – predicate agreement. Further improvement may be achieved only if broader context than a single sentence is involved, this constitutes a great challenge for further research.

Although the first generation of spelling and grammar checkers for Czech already exist, other language checking tools do not. For example, in the field of **authoring support** we face a total lack of tools. This is caused to a certain extent by the fact that Czech is usually a target language for technical documentation of various products, not a source one, and thus the need for authoring tools is not as pressing as it is the case with more widely used languages. Nevertheless, the need for such tools will definitely grow in the future and thus the research in the natural language processing tools in this area will be more important.

Speech interaction

Several universities departments and some their spin-off affiliated companies in Brno, Liberec, Plzeň and Prague do speech processing research and application development.

General recognition of spoken Czech is still in its infants. Simple applications that work with a small vocabulary and grammar have a high reliability, because Czech does not have a complex sound system. The main problem of large applications with large vocabularies and more general language models are the large number of inflection forms of words, a relatively free word order and an informal common Czech, which prevent the use of statistical language modelling methods with similar results as in English. There are several commercial systems with large vocabularies (SpeechTechs.r.o., [HTTP://WWW.SPEECHTECH.CZ](http://www.speechtech.cz), OptimSys, s.r.o., [HTTP://WWW.OPTIMSYS.COM](http://www.optimsys.com), NewtonTechnologies, a.s. [HTTP://WWW.DIKTOVANL.CZ](http://www.diktovanl.cz)), but they only work in dictation systems with a high quality audio input or very limited language domains like sporting events or parliamentary speeches. It is possible to buy a separate recognition engine with an open Media Resource Control Protocol interface that allows the involvement of the recognition module into other applications. Companies offer applications generating off-line transcripts of multimedia archives allowing search. All these products have a relatively good configuration option, but they are not open source applications. For development of an open source recognizer of Czech we lack freely available acoustic training data at present, which would allow the preparation of free acoustic models for speaker independent recog-

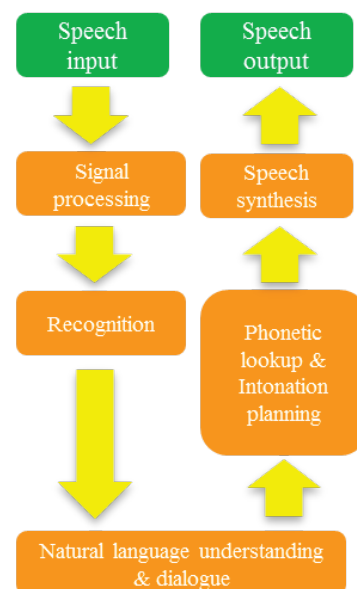


Figure 5: Simple Speech-based Dialogue Architecture

nizer. Universal open source tools and libraries are available for the speech recognition, but the reliable method is still missing for the recognition of spontaneous Czech speech with all its word forms and free word order.

Czech speech synthesis has several commercial voices on a good quality level (Eris, [HTTP://WWW.SPEECHTECH.CZ](http://www.speechtech.cz), Acapela Group, [HTTP://WWW.ACAPELA-GROUP.COM](http://www.acapela-group.com)), there are even open-source Czech synthetic voices, but with lower quality (Festival Czech, [HTTP://DEVEL.FREEBSOFT.ORG/FESTIVAL-CZECH](http://devel.freebsoft.org/festival-czech), Epos TTS System, [HTTP://EPOS.URE.CAS.CZ/](http://epos.ure.cas.cz/), MBROLA, [HTTP://TCTS.FPMS.AC.BE/SYNTHESIS/MBROLA.HTML](http://tcts.fpms.ac.be/synthesis/mbrola.html)). The speech synthesis modules are embeddable into Interactive Voice Response systems that support many open standards. To create more open source voices we are missing open source audio recordings again, which would allow the development of freely available high-quality voices. Recent research in speech synthesis orients on better naturalness and better emotional modulation of voices.

Dialog systems are also in their developmental infancy relying on the previous two technologies. Czech dialog systems without restrictions are the goal of cooperative research of more universities. Some of speech departments are working on many projects in the speech field, being able to offer simple dialog systems, covering most of voice technology.

Research on spoken Czech focuses on improving the language model. In addition to the proven method of increasing the amount of language model training data, which require time-consuming manual transcriptions, specific procedures are explored for the Czech language.

One line of research orients on conversion of spoken Czech to the formal written form, for which there already exist methods developed on text corpora.

Apart from simple cases of replacement of suffixes the method needs to address the replacement of whole word phrases by their correct form. Eliminated in a similar manner are the other phenomena of spontaneous speech such as filler words, repairs or listener responses.

The second line of research focuses on the development of syntactic and semantic analysers of Czech with the help of manually annotated tree corpora of written and spoken Czech. In conjunction with a morphological analysis, this method should help address problems of free word order and of the large number of word forms.

Current research on synthesized spoken Czech tries to develop more natural voices. Hopes are placed in advanced syntactic and semantic analysis of input texts, which should significantly improve the naturalness of utterances.

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 formulated in the early years.

The idea of the translation by the computers became attractive for linguists and mathematicians in the Czech Republic very soon after the first experiments with MT in the world 1954 in USA, 1955 in Soviet Union). In January 1960 the first experiment with English-Czech MT of several sentences by the computer of the 1st generation SAPO, made in earlier Czechoslovakia, was carried out due to the efforts of the small research group from Charles University and the Research Institute of Mathematical Machines. The development of the methods used in MT was continuously followed by this linguistic research group and some experimental rule-based systems of English-Czech and Czech-Russian MT systems were developed for the computers of 2nd generation (made in GDR and in USSR). They were domain-restricted and served mainly for a verification of formally expressed grammatical rules. In the 1990s the prototype of MT between closely related languages was proposed for the pair Czech and Slovak at Charles University; however, its practical application fails due to practical reasons (such as high costs connected with its maintenance etc.). The strategy of statistical methods or combination of statistical and rule-based methods was chosen as a more prospective one for the future.

At its basic level, MT simply substitutes words in one natural language by words in another. This can be useful in domains where a very restricted, formulaic language is used, e.g. weather reports. However, for a good translation of a less standardized text, 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 lays within the fact that human language is ambiguous, which presents challenges on multiple levels, for example *word sense disambiguation* on the lexical level or the attachment of prepositional phrases on the syntactic level.

One way of approaching the task is based on linguistic rules. For translations between closely related languages, a direct translation may be applied. But often, rule-based 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.

For Czech, there are several commercial and academic rule- and lexicon-based translation systems. One of them is based on a linguistic theory elaborated in Prague since 1960s. The system follows the above mentioned analysis-transfer-synthesis scenario. Despite the linguistic adequacy of such approach, the system still suffers from a number of practical difficulties, such as from relatively high error rate of current syntactic analysers and from computational issues related to high number of contextual features that should be taken into account when translating individual words.

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. (Czech has been added recently and the size of Czech data is still orders of magnitude smaller than for established languages.) Given enough data, statistical machine translation works well enough to get an approximate meaning of a foreign language text. However, unlike rule-based

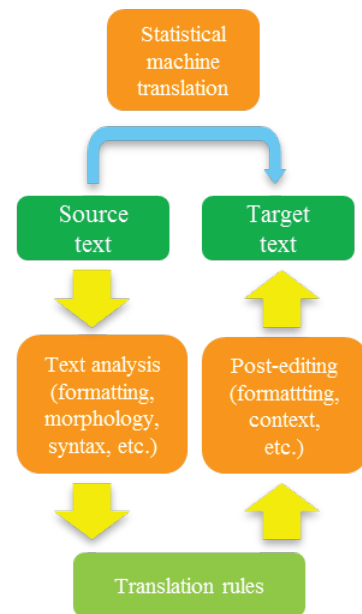


Figure 6: Machine translation (top: statistical; bottom: rule-based)

systems, statistical MT often generates ungrammatical output. On the other hand, besides the advantage that less human effort is required for grammar writing, statistical MT can also cover particularities of the language missing in the rule-based system, for example idiomatic expressions.

Availability of large amounts of bilingual texts is really the key in statistical MT. For Czech, corpora of parallel texts with several other languages are currently being created. The largest data – in total several million pairs of sentences – is available for the English-Czech language pair. The corpus contains for example EU law texts, newspaper texts, technical documentation, and electronic books. The most challenging problem related to the contemporary parallel corpora is the quality of alignment (pairing of corresponding parts of a text and its translation). Not only that exact word-to-word linkage is impossible due to differences in morphology and syntax of the two languages, but reliable sentence-to-sentence and sometimes even document-to-document alignment is difficult to achieve too. Needless to say that compilation of such corpora has to be fully automatic – human processing is completely out of question because of the data size.

Languages with richer morphology like Czech also pose specific challenges for state-of-the-art statistical systems: the system has to choose not only the correct word but also the appropriate form to satisfy grammatical context. Very few statistical systems to date can handle morphological richness explicitly and thus often fall short of vocabulary: all the necessary word forms are not available even in large parallel corpora.

As the strengths and weaknesses of rule-based and statistical MT are complementary, it is nowadays more or less consensus to target **hybrid approaches** combining methodologies of both. This can be done in several ways. One is to use both rule-based and statistical 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.

Another, more challenging approach is to design a new setup that combines the advantages of the two paradigms by integrating the good features of each. For example, making a rule-based system adaptive by adding a module for rule learning, or, making a statistical MT system syntax-aware by adding syntactical constraints.

Completely separate is the question of **evaluating MT output quality**, both manually and automatically. Experience shows that different systems score differently under various manual evaluations: rule-based systems tend to preserve the meaning better while statistical systems produce output more fluent locally. In e.g. question-answering evaluation, the meaning is more important. On the other hand, local fluency impacts the impression more when the user is directly comparing system outputs. Automatic evaluation (based on the comparison of MT output to one or more manually constructed reference translations) is vital in development of MT systems. It has been shown that such automatic evaluation is unreliable esp. for languages with richer morphology and it is also acknowledged that some automatic fine-grained reporting on MT quality and error types would be very useful.

Information management/“LT 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 on the academic side of Computational Linguistics.

For example, question answering has become an active area of research, for which annotated corpora have been built, competitions have been started, etc. 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?); how can specific information - the answer - be reliably extracted from a document, without unduly ignoring the context; etc.

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 for instance be the detection of the key players in company take over as they are reported in newspaper stories. Another scenario that has been worked on is reports on terrorist incidents, where the problem is to map the text to a template specifying the perpetrator, the target, time and location of the incident, and the results of the incident. Domain-specific template-filling is the central characteristic of IE, which for this reason is another example of a "behind the scenes" technology that constitutes a well-demarcated research area but for practical purposes then needs to be embedded into a suitable application environment.

Two "borderline" areas, which sometimes play the role of standalone application and sometimes that of supportive, "under the hood" component are *text summarization* and *text generation*. Summarization, obviously, refers to the task of making a long text short, and is offered for instance as a functionality within MS Word. It works largely on a statistical basis, by first identifying "important" words in a text (that is, for example, words that are highly frequent in this text but markedly less frequent in general language use) and then determining those sentences that contain many important words. These sentences are then marked in the document, or extracted from it, and are taken to constitute the summary. In this scenario, which is by far the most popular one, summarization equals sentence extraction: the text is reduced to a subset of its sentences. All commercial summarizers make use of this idea.

An alternative approach, to which some research is devoted, is to actually synthesize new sentences, i.e., to build a summary of sentences that need not show up in that form in the source text. This requires a certain amount of deeper understanding of the text and therefore is much less robust; furthermore, such an approach is to

a good extent geared towards a particular domain or text genre, since particular knowledge is needed to perform the step of abstracting from the source text to its "content". Synthesizing a summary now in turn is a case of text generation - the production of new text, either from other text (as in summarization), or from a set of non-textual data. This can be applied whenever reports are needed that describe how certain data streams develop over time. Such systems have been built for generating weather and air quality reports, or for summaries of medical diagnosis data. However, a text generator is in most cases not a stand-alone application but embedded into a larger software environment, such as into a clinical information system where patient data is collected, stored and processed, and report generation is just one of many functionalities.

There are many Czech research groups working on international (e.g. English) applications. Only a part of the HLT effort in the Czech Republic is dedicated particularly to Czech. There are many NLP-components for Czech, such as spell-checkers, corpora, morphological taggers and valency lexicons, along with a Czech collocation analyser (Word Sketch Engine for Czech, developed at the Masaryk University in Brno, (Horák, Rychlý, Kilgariff, 2009)) and a manifold research of speech recognition and generation, but not many more complex HLT applications ready to use in the industry.

To the best of our knowledge, there is just one working question-answering system reported for Czech, developed by researchers at the Masaryk University in Brno – UIO (standing for the Czech “Artificial Intelligence of a Monkey”), (Svoboda, 2003). UIO can ask databases and the Web. In its current version, UIO can be used for asking questions about train and coach timetables, cinema and theatre performances, about currency exchange rates, name-days and on the Diderot Encyclopedia. For all domains UIO has an accuracy rate about 80%. A competing (so far no-name) system is being developed at the University of West Bohemia.

A simple conversational dialog system was developed at the Institute of Formal and Applied Linguistics, Charles University in Prague in collaboration with the Faculty of Cybernetics at the University of West Bohemia in Pilsen and, to a lesser extent, with some other consortium partners in the FP-6 Companions project ([HTTP://WWW.COMPANIONS-PROJECT.ORG](http://www.companions-project.org)). A human-like avatar converses with seniors about their respective personal photograph collections and life stories (Ptáček et al., 2010; Romportl et al., 2010; Guber, Tihelka, 2010).

The Text-Mining Research Group at the University of West Bohemia is developing a User Profile Generation system (Grolmus, 2003). This system performs text-mining on the documents gathered and viewed by a user. It uses the (user-approved) information to recommend particular documents on further searches as well as to estimate the user's expertise in a given domain. This application can be used e.g. as a support of digital libraries.

WebGen, a similarly-named application developed at the Masaryk University in Brno (LSD lab, [HTTP://LSD.FI.MUNI.CZ/WEBGEN/INDEX.PHP?PAGE=UVOD&LANG=EN](http://lfd.fi.muni.cz/webgen/index.php?page=uvod&lang=en)) is a dialog-based system that helps visually impaired people generate web presentations in Czech. It is still in development (Bártel, Plhák, 2008).

The Department of Computer Graphics and Multimedia FIT BUT Faculty of Information Technology at the Brno University of Tech-

nology in Brno
[HTTP://WWW.FIT.VUTBR.CZ/RESEARCH/VIEW_PRODUCT.PHP.CS?ID=157&NOTITLE=1](http://www.fit.vutbr.cz/research/view_product.php?ID=157&NOTITLE=1) delivered speech-processing software that adds semantic labels to speech transcripts (Speech Tagging, 2010). The client side is an HTML user interface in a web browser accessing functionality provided by the server. The server enables upload and analysis of speech records. The user is able to define and manage so called "tags", which are groups of semantic related keywords. If some keyword is found in some record, the record is tagged correspondingly. This service would be useful in e.g. crisis management, when it is suitable to classify phone calls according to words spoken, but, to our knowledge, it has not been employed in real applications yet.

The Faculty of Cybernetics at the University of West Bohemia in Pilsen has developed several speech-based applications for Czech, such as a dialog system with train timetables or a dialog system for students registering for exams on the phone (University VoiceXML information system, [HTTP://VOICE.ZCU.CZ/](http://voice.zcu.cz/)). Their research groups run numerous projects aimed at assisting people with hearing impairments, e.g. by translating between Czech and (Czech) sign language.

Another useful application (developed in Pilsen) is a voice-controlled system for dentists (Nagy et al.). It works in two modes: in the first, it reads the record of a tooth in the mouth of the patient. In the second mode, it records the information that the dentist dictates and updates the status of the given tooth. Voice-control is essential there, since the dentist is not allowed to touch either a screen or controls on a dictation device while examining the patient.

The research groups from the University of West Bohemia and from the Institute of Formal and Applied Linguistics at the Charles University in Prague participated in the international MALACH project (MALACH stands for "Multilingual Access to Large Spoken Archives" and means "messenger angel" in Hebrew), (Psutka et al., 2005). They were in charge of speech recognition and semantic indexing of testimonies recorded in Czech and other Slavic languages. The Charles University now hosts one of the local access points to the archive of testimonies of holocaust survivors. (Other access points are located in the USA, Germany, Hungary, Izrael and Australia)

The nearly 52,000 videotaped testimonies of the Shoah Foundation Institute's Visual History Archive were recorded primarily between 1994 and 1999 in 56 countries and in 32 languages. While the majority of the interviews are with Jewish Holocaust survivors, the archive also includes the testimonies of political prisoners, Sinti and Roma (Gypsy) survivors, Jehovah's Witness survivors, survivors of eugenics policies, and homosexual survivors as well as rescuers and aid providers, liberators, and participants in war crimes trials.

The archive is accessible through an on-line interface, which enables the users browsing and viewing the testimonies, deploying an index of 55 thousand keywords and key phrases. The access point in Prague stores more than 500 testimonies in Czech, with average duration 2 hours. Other testimonies have to be ordered online from the other access points, which usually takes a few hours.

Miscellaneous

It would be misleading to judge the NLP-HLT research of the respective countries only on the basis of how many resources and applications for their national language they have produced. In fact, there is a vicious circle in the NLP-HLT research for small languages: the grant agencies as well as the government want to support only the best teams. The best teams are the ones that produce the most internationally recognized publications. These are significantly easier to achieve in research that has international impact. While almost any improvement in any issue is interesting to report on big or strategic languages such as English, Chinese or Arabic, a research with the same outcome has a grossly humbler impact when reported on languages that are interesting only for their native speakers. To produce a good publication on a small language, a real breakthrough is needed, whereas, obviously, breakthroughs cannot be counted on to happen regularly. Besides, even if a language-dependent result for a small language is considered a breakthrough by the local research community, it is still difficult to present to international reviewers who are not familiar with the language.

Also, language-independent solutions are generally preferred to the language-dependent ones, since their commercial application is cheaper. English is the natural first-choice language to experiment on in the European context, as there are comprehensive high-quality resources available for English. Also, the results are more easily compared within the international community.

As a consequence, national teams focus on English rather than relying on research on their national language. This is to be kept in mind when assessing the quality of national HLT/NLP research and development. A poor inventory of good HLT applications and resources for a small language does not necessarily imply poor research, but it can be a serious indicator of lacking governmental support policy. Targeted governmental support of national-language HLT is vital for language communities whose markets are too small for national-language HLT to be endorsed by the private sector.

LT Industry and Programs

- ❑ Industrial deployment of language technologies is not widespread in Czechia. Businesses specialized in LT are rare. The same holds for research & development departments of larger companies. One notable exception is the IBM research team in Prague working on voice technologies; however, their work is not so much focused on the Czech language.
- ❑ Web search engines and services (Seznam, Centrum, Google etc.) are nowadays generally capable of performing morphological analysis and lemmatization. Google offers phrase-based machine translation of websites and user-supplied text both to and from Czech. Seznam provides on-line dictionaries between Czech on one side and English, German, French, Italian, Spanish or Russian on the other side. However, they don't provide translation of running text.
- ❑ There are companies developing and publishing bilingual electronic dictionaries as Windows applications. These typically contain morphological analysis / lemmatization, some of them also a sort of ontology.
- ❑ Cell phone manufacturers can use a Czech version of T9.

- Office software packages (such as Microsoft Office 2010) provide Czech spellchecking, grammar checking, sometimes also machine translation and automatic speech recognition (voice input).
- Phone switchboards and help/information applications employing automatic speech recognition are virtually unheard of. There have been pilot projects with ASR by university teams specialized in ASR (most notably the University of West Bohemia in Plzeň and the Technical University in Liberec) but there is no wide industrial application of such technologies.
- Czech speech recognition was commercialized by Newton Technologies company—a spin-off of the Technical University in Liberec.
- Most of the government-originating funding programs are maintained by the Czech Science Foundation (GAČR) and focused on basic research. Recently (2009), a new Technological Agency of the Czech Republic (Technologická agentura České republiky, TAČR) has been established, which shall focus on applied research. However, there are probably no LT-related projects funded by TAČR yet.

LT Research and Education

From the historical point of view, the names *computational linguistics*, *natural language processing* and *speech recognition* have been used for a longer time than the term *language technologies*, at least in the context of research and education. No matter the name, the disciplines related to the natural language comprise a number of related objects of research and education: theoretical linguistics, corpus linguistics, computer science, mathematics, machine learning etc.

Czech Republic has a number of institutes that devote their research and teaching to computational linguistics and spoken language processing. Below we provide a detailed list of them including the information what topics they are interested in at most (CL – computational l., CoL – corpus l., TL – theoretical l., ASR – automatic speech recognition) and what study programs they offer, if ever.

Charles University in Prague, Faculty of Mathematics and Physics is offering the European Master Program in Language and Communication Technologies as a part of its MSc. study program as well as PhD program. Thanks to this activity, the Faculty can welcome the international students who give new impulses to their Czech colleagues.

The IBM Research Prague is the only natural language research and development lab outside the universities placed in the Czech Republic.

The study programs the institutes offer emphasis on both theory and practice. Unfortunately, a firm demand for such experts is very low in the Czech Republic.

1 Charles University in Prague

- Institute of Formal and Applied Linguistics
([HTTP://UFAL.MFF.CUNI.CZ](http://ufal.mff.cuni.cz)); CL, TL, ASR; BSc, MSc, PhD;

- Institute of Czech National Corpus
([HTTP://UCNK.FF.CUNI.CZ/ENGLISH/INDEX.PHP](http://ucnk.ff.cuni.cz/english/index.php)); CoL; PhD.
 - Institute of Theoretical and Computational Linguistics
([HTTP://UTKL.FF.CUNI.CZ](http://utkl.ff.cuni.cz)); CL; PhD;
- 2 University of Economics in Prague
 - Department of information and knowledge engineering
([HTTP://KIZ15.VSE.CZ/](http://kiz15.vse.cz/)); datamining, semantic web, ontologies; BSc, MSc, PhD;
- 3 Czech Technical University in Prague
 - Department of Cybernetics (<http://cyber.felk.cvut.cz/>); robotics, artificial intelligence; BSc, MSc, PhD;
 - Department of Circuit Theory
(<http://noel.feld.cvut.cz/speechlab/start.php?page=projects&lang=en#2>); ASR; BSc, MSc, PhD
- 4 Masaryk University, Brno
 - Natural Language Processing Centre
([HTTP://NLP.FI.MUNI.CZ/EN/NLPLAB](http://nlp.fi.muni.cz/en/nlplab)); CL, ASR;
 - Department of Czech Language
([HTTP://WWW.MUNI.CZ/PHIL/211700?LANG=EN](http://www.muni.cz/phil/211700?lang=en)); CL, TL; BSc, MSc, PhD;
- 5 Brno University of Technology
 - Speech Processing Group
([HTTP://SPEECH.FIT.VUTBR.CZ/](http://speech.fit.vutbr.cz/)); ASR;
 - Natural Language Processing Research Group
(<http://www.fit.vutbr.cz/research/groups/nlp/index.php?lang=en>); CL;
- 6 University of West Bohemia
 - Department of Cybernetics
([HTTP://WWW.KKY.ZCU.CZ/EN](http://www.kky.zcu.cz/en)); ASR; BSc, MSc, PhD;
- 7 Technical University of Liberec
 - Laboratory of Computer Speech Processing
([HTTPS://WWW.ITE.TUL.CZ/SPEECHLABE/](https://www.ite.tul.cz/speechlab/)); ASR;

Status of Tools and Resources for Czech

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

- 1 **Quantity:** Does a tool/resource exist for the language at hand? The more tools/resources exist, the higher the rating.
 - 0: no tools/resources whatsoever
 - 6: many tools/resources, large variety
- 2 **Availability:** Are tools/resources accessible, i.e., are they Open Source, freely usable on any platform or only available for a high price or under very restricted conditions?
 - 0: practically all tools/resources are only available for a high price

- 6: a large amount of tools/resources is freely, openly available under sensible Open Source or Creative Commons licenses that allow re-use and re-purposing
- 3 **Quality:** How well are the respective performance criteria of tools and quality indicators of resources met by the best available tools, applications or resources? Are these tools/resources current and also actively maintained?
 - 0: toy resource/tool
 - 6: high-quality tool, human-quality annotations in a resource
- 4 **Coverage:** To which degree do the best tools meet the respective coverage criteria (styles, genres, text sorts, linguistic phenomena, types of input/output, number languages supported by an MT system etc.)? To which degree are resources representative of the targeted language or sublanguages?
 - 0: special-purpose resource or tool, specific case, very small coverage, only to be used for very specific, non-general use cases
 - 6: very broad coverage resource, very robust tool, widely applicable, many languages supported
- 5 **Maturity:** Can the tool/resource be considered mature, stable, ready for the market? Can the best available tools/resources be used out-of-the-box or do they have to be adapted? Is the performance of such a technology adequate and ready for production use or is it only a prototype that cannot be used for production systems? An indicator may be whether resources/tools are accepted by the community and successfully used in LT systems.
 - 0: preliminary prototype, toy system, proof-of-concept, example resource exercise
 - 6: immediately integratable/applicable component
- 6 **Sustainability:** How well can the tool/resource be maintained/integrated into current IT systems? Does the tool/resource fulfil a certain level of sustainability concerning documentation/manuals, explanation of use cases, front-ends, GUIs etc.? Does it use/employ standard/best-practice programming environments (such as Java EE)? Do industry/research standards/quasi-standards exist and if so, is the tool/resource compliant (data formats etc.)?
 - 0: completely proprietary, ad hoc data formats and APIs
 - 6: full standard-compliance, fully documented
- 7 **Adaptability:** How well can the best tools or resources be adapted/extended to new tasks/domains/genres/text types/use cases etc.?
 - 0: practically impossible to adapt a tool/resource to another task, impossible even with large amounts of resources or person months at hand
 - 6: very high level of adaptability; adaptation also very easy and efficiently possible

	Quantity	Availability	Quality	Coverage	Maturity	Sustainability	Adaptability
Language Technology (Tools, Technologies, Applications)							
Tokenization, Morphology (tokenization, POS tagging, morphological analysis/generation)	4	1	5	5	3	2	4
Parsing (shallow or deep syntactic analysis)	4	4	4	3	4	3	4
Sentence Semantics (WSD, argument structure, semantic roles)	2	2	3	3	2	3	4
Text Semantics(coreference resolution, context, pragmatics, inference)	2	1	3	3	2	2	3
Advanced Discourse Processing (text structure, coherence, rhetorical structure/RST, argumentative zoning, argumentation, text patterns, text types etc.)	0	0	0	0	0	0	0
Information Retrieval(text indexing, multimedia IR, crosslingual IR)	4	1	4	5	4	1	1
Information Extraction (named entity recognition, event/relation extraction, opinion/sentiment recognition, text mining/analytics)	2	4	4	4	3	2	4
Language Generation (sentence generation, report generation, text generation)	2	1	3	3	3	2	4
Summarization, Question Answering, advanced Information Access Technologies	0	0	0	0	0	0	0
Machine Translation	5	4	2	3	4	2	3
Speech Recognition	4	5	4	4	3	4	4
Speech Synthesis	3	3	3	4	3	3	2
Dialogue Management (dialogue capabilities and user modelling)	4	5	5	5	4	4	4
Language Resources (Resources, Data, Knowledge Bases)							
Reference Corpora	4	3	5	5	5	4	1
Syntax-Corpora(treebanks, dependency banks)	4	3	6	3	6	5	4
Semantics-Corpora	4	3	4	3	4	5	4
Discourse-Corpora	2	1	3	2	2	3	3
Parallel Corpora, Translation Memories	2	4	3	4	2	2	3
Speech-Corpora (raw speech data, labelled/annotated speech data, speech dialogue data)	4	1	4	2	3	3	2

	Quantity	Availability	Quality	Coverage	Maturity	Sustainability	Adaptability
Multimedia and multimodal data (text data combined with audio/video)	4	2	4	2	3	3	2
Language Models	3	1	4	4	4	2	3
Lexicons, Terminologies	3	2	3	4	2	3	2
Grammars	1	1	3	2	2	1	1
Thesauri, WordNets	5	2	3	4	3	3	2
Ontological Resources for World Knowledge (e.g. upper models, Linked Data)	2	2	3	2	3	2	1

Conclusions

- ❑ While some specific corpora of high quality exist, a very large syntactically annotated corpus is not available.
- ❑ There is a highly elaborated syntactically annotated corpus for Czech. However, the corpus is not available for free (can be bought via LDC). Several extending annotations (coreference, discourse etc.) are being performed on top of the corpus, but they are not yet finished.
- ❑ For Czech, a large text corpus exists, but it is not available for automatic processing (only for on-line searching).
- ❑ Many of the resources lack standardization, i.e., even if they exist, sustainability is not given; concerted programs and initiatives are needed to standardize data and interchange formats.
- ❑ Semantics is more difficult than syntax; text semantics is more difficult than word and sentence semantics.
- ❑ Research was successful in designing particular high quality software, but it is nearly impossible to come up with sustainable and standardized solutions given the current funding situations.
- ❑ There is an ontological resource for Czech (even mapped to other European ontological resources) but its coverage is limited.
- ❑ Speech Recognition of Czech is studied at several universities and workplaces but free tools and data are not available.
- ❑ The main problems of large vocabulary recognizers are in specific Czech language modeling.
- ❑ In the field of speech synthesis, there are available open source packages, but the speech synthesis with more natural voices is available only in commercial applications.
- ❑ Czech dialogue systems are very little extended due to poor accessibility of high quality speech recognition modules of Czech.

- For the web search, there is enough room both for improving existing popular local search engines through the academia-industry collaboration, or for introducing a new one.

Bibliography

Bártek, Luděk, Plhák, Jaromír. Visually Impaired Users Create Web Pages. In: *11th International Conference on Computers Helping People with Special Needs*. Berlin : Springer - Verlag, 2008. 2008, Linz, Austria, p. 466 - 473.

Grolmus P, Hynek J., Ježek K. User Profile Identification Based on Text Mining, In: *Proceedings of 6th International Conference on Information Systems Implementation and Modelling – ISIM '03* Brno, Czech Republic : MARQ, 2003, p. 109-116.

Gruber, Martin; Tihelka, Daniel. Expressive Speech Synthesis for Czech Limited Domain Dialogue System - Basic Experiments. In *Proceedings of the 10th International Conference on Signal Processing, ICSP 2010*, vol. 1.

Horák, Aleš - Rychlý, Pavel - Kilgarriff, Adam. Czech Word Sketch Relations with Full Syntax Parser. In *After Half a Century of Slavonic Natural Language Processing*. Brno, Czech Republic : Masaryk University, 2009. p. 101-112.

Nagy M., Hanzlicek P., Zvarova J., Dostalova T., Seydlova M., Hippman R., Smidl L., Trmal J., Psutka J. Voice-controlled data entry in dental electronic health record.

Psutka Josef, Ircing Pavel, Psutka Josef V., Hajič Jan, Byrne William, Mírovský Jiří: Automatic transcription of Czech, Russian, and Slovak Spontaneous Speech in the MALACH Project. In: *Proceedings of Eurospeech 2005*, Lisboa, Portugal, p. 1349-1352, 2005

Ptáček, Jan; Ircing, Pavel; Spousta, Miroslav; Romportl, Jan; Loose, Zdeněk; Cinková, Silvie; Relano Gil, Jose and Raul Santos (2010). Integration of Speech and Text Processing Modules into a Real-Time Dialogue System. In *Text, Speech and Dialogue, Proceedings of the 13th International Conference TSD 2010*, Lecture Notes in Artificial Intelligence, vol. 6231, p. 552-559, Springer, Berlin-Heidelberg, Germany, 2010.

Romportl, Jan; Zovato, Enrico; Santos, Raul; Ircing, Pavel; Relano Gil, Jose; Danieli, Morena. Application of Expressive TTS synthesis in an Advanced ECA System. In *Proceedings of the ISCA Tutorial and Research Workshop on Speech Synthesis*, 2010, pp. 120-125.

Speech Tagging, software, 2010. Authors: Smrž Pavel, Schmidt Marek, Zuzanaček Jiří, Příbyl Bronislav, Navrátil Jan, Láník Aleš, Burget Lukáš, Cipr Tomáš, Fapšo Michal, Glembek Ondřej, Grézl František, Chalupníček Kamil, Karafiát Martin, Matějka Pavel, Schwarz Petr, Szöke Igor.

Svoboda, L.: UIO, a dialog system for question answering. In: *Proc. Znalosti 2003 Workshop* (V. Svátek, ed.), 2003.

About META-NET

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

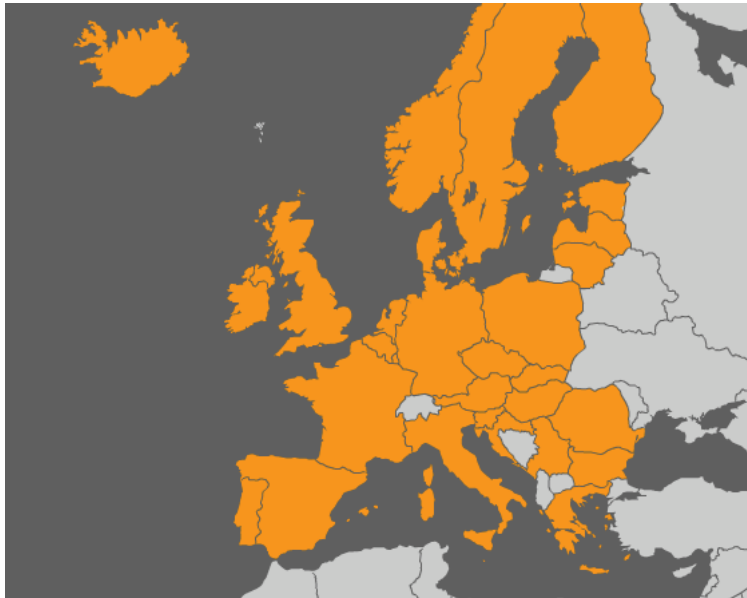


Figure 1: Countries Represented in META-NET

META-NET cooperates with other initiatives like the Common Language Resources and Technology Infrastructure (CLARIN), which is helping establish digital humanities research in Europe. META-NET fosters the technological foundations for the establishment and maintenance of a truly multilingual European information society that:

- ❑ makes communication and cooperation possible across languages;
- ❑ provides equal access to information and knowledge in any language;
- ❑ offers advanced and affordable networked information technology to European citizens.

META-NET stimulates and promotes multilingual technologies for all European languages. The technologies enable automatic translation, content production, information processing and knowledge management for a wide variety of applications and subject domains. The network wants to improve current approaches, so better communication and cooperation across languages can take place. Europeans have an equal right to information and knowledge regardless of language.

Lines of Action

META-NET launched on 1 February 2010 with the goal of advancing research in language technology (LT). The network supports a Europe that unites as a single, digital market and information space. META-NET has conducted several activities that further its



The Multilingual Europe Technology Alliance (META)

goals. META-VISION, META-SHARE and META-RESEARCH are the network's three lines of action.



Figure 2: Three Lines of Action in META-NET

META-VISION fosters a dynamic and influential stakeholder community that unites around a shared vision and a common strategic research agenda (SRA). The main focus of this activity is to build a coherent and cohesive LT community in Europe by bringing together representatives from highly fragmented and diverse groups of stakeholders. In the first year of META-NET, presentations at the FLReNet Forum (Spain), Language Technology Days (Luxembourg), JIAMCATT 2010 (Luxembourg), LREC 2010 (Malta), EAMT 2010 (France) and ICT 2010 (Belgium) centred on public outreach. According to initial estimates, META-NET has already contacted more than 2,500 LT professionals to develop its goals and visions with them. At the META-FORUM 2010 event in Brussels, META-NET communicated the initial results of its vision building process to more than 250 participants. In a series of interactive sessions, the participants provided feedback on the visions presented by the network.

META-SHARE creates an open, distributed facility for exchanging and sharing resources. The peer-to-peer network of repositories will contain language data, tools and web services that are documented with high-quality metadata and organised in standardised categories. The resources can be readily accessed and uniformly searched. The available resources include free, open source materials as well as restricted, commercially available, fee-based items. META-SHARE targets existing language data, tools and systems as well as new and emerging products that are required for building and evaluating new technologies, products and services. The reuse, combination, repurposing and re-engineering of language data and tools plays a crucial role. META-SHARE will eventually become a critical part of the LT marketplace for developers, localisation experts, researchers, translators and language professionals from small, mid-sized and large enterprises. META-SHARE addresses the full development cycle of LT—from research to innovative products and services. A key aspect of this activity is establishing META-SHARE as an important and valuable part of a European and global infrastructure for the LT community.

META-RESEARCH builds bridges to related technology fields. This activity seeks to leverage advances in other fields and to capitalise on innovative research that can benefit language technology. In particular, this activity wants to bring more semantics into machine translation (MT), optimise the division of labour in hybrid MT, exploit context when computing automatic translations and prepare an empirical base for MT. META-RESEARCH is working with other fields and disciplines, such as machine learning and the Semantic Web community. META-RESEARCH focuses on collect-

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

Member Organisations

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

Country	Organisation	Participant(s)
Austria	University of Vienna	Gerhard Budin
Belgium	University of Antwerp	Walter Daelemans
	University of Leuven	Dirk van Compernelle
Bulgaria	Bulgarian Academy of Sciences	Svetla Koeva
Croatia	University of Zagreb	Marko Tadić
Cyprus	University of Cyprus	Jack Burston
Czech Republic	Charles University in Prague	Jan Hajic
Denmark	University of Copenhagen	Bolette Sandford Pedersen and Bente Maegaard
Estonia	University of Tartu	Tiit Roosmaa
Finland	Aalto University	Timo Honkela
	University of Helsinki	Kimmo Koskenniemi and Krister Linden
France	CNRS/LIMSI	Joseph Mariani
	Evaluations and Language Resources Distribution Agency	Khalid Choukri
Germany	DFKI	Hans Uszkoreit and Georg Rehm
	RWTH Aachen University	Hermann Ney
	Saarland University	Manfred Pinkal
Greece	Institute for Language and Speech Processing, "Athena" R.C.	Stelios Piperidis
Hungary	Hungarian Academy of Sciences	Tamás Váradi

Country	Organisation	Participant(s)
	Budapest University of Technology and Economics	Géza Németh and Gábor Olaszy
Iceland	University of Iceland	Eiríkur Rögnvaldsson
Ireland	Dublin City University	Josef van Genabith
Italy	Consiglio Nazionale Ricerche, Istituto di Linguistica Computazionale "Antonio Zampolli"	Nicoletta Calzolari
	Fondazione Bruno Kessler	Bernardo Magnini
Latvia	Tilde	Andrejs Vasiljevs
	Institute of Mathematics and Computer Science, University of Latvia	Inguna Skadina
Lithuania	Institute of the Lithuanian Language	Jolanta Zabarskaitė
Luxembourg	Arax Ltd.	Vartkes Goetcherian
Malta	University of Malta	Mike Rosner
Netherlands	Utrecht University	Jan Odijk
	University of Groningen	Gertjan van Noord
Norway	University of Bergen	Koenraad De Smedt
Poland	Polish Academy of Sciences	Adam Przepiórkowski and Maciej Ogrodniczuk
	University of Lodz	Barbara Lewandowska-Tomaszczyk and Piotr Pezik
Portugal	University of Lisbon	Antonio Branco
	Institute for Systems Engineering and Computers	Isabel Trancoso
Romania	Romanian Academy of Sciences	Dan Tufis
	Alexandru Ioan Cuza University	Dan Cristea
Serbia	University of Belgrade	Dusko Vitas, Cvetana Krstev and Ivan Obradovic
	Institute Mihailo Pupin	Sanja Vranes
Slovakia	Slovak Academy of Sciences	Radovan Garabik
Slovenia	Jozef Stefan Institute	Marko Grobelnik
Spain	Barcelona Media	Toni Badia
	Technical University of Catalonia	Asunción Moreno
	Pompeu Fabra University	Núria Bel

Country	Organisation	Participant(s)
Sweden	University of Gothenburg	Lars Borin
UK	University of Manchester	Sophia Ananiadou
	University of Edinburgh	Steve Renals

References

- ¹ European Commission Directorate-General Information Society and Media, *User language preferences online*, Flash Eurobarometer #313, 2011 (http://ec.europa.eu/public_opinion/flash/fl_313_en.pdf).
- ² European Commission, *Multilingualism: an asset for Europe and a shared commitment*, Brussels, 2008 (http://ec.europa.eu/education/languages/pdf/com/2008_0566_en.pdf).
- ³ UNESCO Director-General, *Intersectoral mid-term strategy on languages and multilingualism*, Paris, 2007 (<http://unesdoc.unesco.org/images/0015/001503/150335e.pdf>).
- ⁴ European Commission Directorate-General for Translation, *Size of the language industry in the EU*, Kingston Upon Thames, 2009 (<http://ec.europa.eu/dgs/translation/publications/studies>).
- ⁵ The number of citizen in CR was in the middle of 2010 10,5 million (according the Czech Statistical Office).
- ⁶ See <http://www.czech-language/overview>
- ⁷ See <http://www.ethnologue.org>
- ⁸ The context for ex. (5) is introducing of the new service for parents: They can ask for a taxi with a nun, which takes care of a child.
- ⁹ See <http://cs.wikipedia.org>
- ¹⁰ [HTTP://PRIRUCKA.UJC.CAS.CZ/](http://PRIRUCKA.UJC.CAS.CZ/)
- ¹¹ See http://www.coe.int/t/dg4/linguistc/CADRE_EN.asp#TopOfPage
- ¹² More info on the STYX system is available at [HTTP://UFAL.MFF.CUNI.CZ/STYX](http://UFAL.MFF.CUNI.CZ/STYX).
- ¹³ CETLEF is a French acronym for *Understanding and Correction of Errors in Czech as a Foreign Language for French Learners*. It is available at [HTTP://WWW.CETLEF.FR](http://WWW.CETLEF.FR).
- ¹⁴ See [HTTP://WWW.C2J.CZ/ATTACHMENTS/105_2010_WKSH_LIBEREC_NIKI3.PDF](http://WWW.C2J.CZ/ATTACHMENTS/105_2010_WKSH_LIBEREC_NIKI3.PDF)