VEDA
vydavateľstvo
Slovenskej
akadémie
vied

SLOVENSKÁ AKADÉMIA VIED
Jazykovedný ústav Ľudovíta Štúra

VEDECKÝ REDAKTOR
prof. PhDr. Pavol Žigo, CSc.

RECENZENTI
Mgr. Michal Křen, PhD.
Mgr. Jana Levická, PhD.

EDITORKY
Mgr. Katarína Gajdošová
Mgr. Adriána Žáková

J A Z Y K O V E D N É Š T Ú D I E XXXI

Rozvoj jazykových technológií a zdrojov
na Slovensku a vo svete
(10 rokov Slovenského národného korpusu)

VEDA
vydavateľstvo
Slovenskej
akadémie
vied
Bratislava
2014

Zborník Jazykovedné štúdie XXXI je zostavený z príspevkov, ktoré odzneli na medzinárodnej konferencii Rozvoj jazykových technológií a zdrojov na Slovensku a vo svete (10 rokov Slovenského národného korpusu). Vedecko-informačné podujatie sa konalo v dňoch 7. a 8. júna 2012 v Bratislave pri príležitosti 10. výročia vzniku špecializovaného pracoviska Slovenského národného korpusu Jazykovedného ústavu Ľ. Štúra Slovenskej akadémie vied. Konferenciu organizačne pripravilo oddelenie Slovenského národného korpusu Jazykovedného ústavu Ľudovíta Štúra Slovenskej akadémie vied.

Jazykovedné štúdie XXXI sú financované z projektu Budovanie Slovenského národného korpusu a elektronizácia jazykovedného výskumu na Slovensku (tretia etapa) a projektu CEntral and South-east europeAn Resources, č. 271022.

ISBN 978-80-224-1391-6

Obsah

Marta Cimbáková
Slovo na úvod *7*

Georg Rehm
META-NET: A Research and Technology Strategy for Multilingual Europe *9*

Tamás Váradi
Supporting Multilingual Europe The CESAR initiative *27*

Mária Šimková – Radovan Garabík
Slovenský národný korpus (2002 – 2012): východiská, ciele a výsledky pre výskum a prax *35*

Katarína Gajdošová – Mária Šimková
Slovenský hovorený korpus (2008 – 2012) *65*

Róbert Sabo – Sakhia Darjaa – Milan Rusko
Praktické aplikácie automatického spracovania reči v Ústave informatiky SAV *85*

František Čermák
InterCorp: jeho povaha a možnosti *97*

Karel Pala – Pavel Rychlý
Building Large Corpora and Tools for Computer Lexicography *113*

Ludmila Dimitrova
Multilingual Resources with Bulgarian – Recent Developments (IMI-BAS Experience) *123*

Leonid Iomdin
Automatic Text Processing and Deeply Annotated Text Corpora of Russian:
Interaction and Mutual Impact *136*

Pavol Žigo
Počítačové kartografovanie nárečí v Slovanskom jazykovom atlase *147*

Július Kravjar
Národný korpus bakalárskych, diplomových, dizertačných, rigoróznych a habilitačných prác
slovenských vysokých škôl a boj proti plagiátorstvu *154*

Tibor Pintér
Maďarský národný korpus 2. Pokus o nový korpus maďarského jazyka *163*

Daniel Hládek – Ján Staš – Jozef Juhár
Building Organized Text Corpora for Speech Technologies in the Slovak Language *173*

Velislava Stoykova
Collaboratively Developed Lexical Resources for Bulgarian with Application to Dictionaries and
Reference Sources Compilation *182*

Slovo na úvod

Ctené dámy, vážení páni, milí hostia,
stretávame sa na podujatí, ktorého obsah charakterizujú prívlastky národný aj medzinárodný, lingvistický aj počítačový, teoretický aj aplikovaný, retrospektívny aj perspektívny. V rámci národného prezentačného dňa členov medzinárodného projektu CESAR sa prezentuje najmä interdisciplinárne pracovisko Slovenského národného korpusu Jazykovedného ústavu Ľudovíta Štúra Slovenskej akadémie vied, ktoré využilo doterajšie výsledky teoretického bádania v oblasti jazykového systému, ďalej ich rozvinulo a posunulo do roviny moderného počítačového spracovania prirodzeného jazyka. Veľmi dôležité je, že všetky vytvorené databázy, všeobecné či špecializované, sú prístupné širokej verejnosti, školám na Slovensku i v zahraničí, vedecko-výskumným pracoviskám, kultúrnym inštitúciám, vydavateľstvám a všetkým záujemcom. Toto sprístupnenie a využívanie bolo možné vďaka nekomerčnému charakteru projektu a jeho štátnej podpore. Ministerstvo školstva, vedy, výskumu a športu SR od začiatku stálo pri zrode tohto pracoviska, keď spolu s Ministerstvom kultúry SR a SAV podporilo vládny projekt Vybudovanie Národného korpusu slovenského jazyka a elektronizácia jazykovedného výskumu v rokoch 2002 – 2006 (uznesenie vlády SR č. 137 z 13. 2. 2002) a neskôr zaradilo novoutvorené pracovisko medzi projekty Štátneho programu výskumu a vývoja Aktuálne otázky rozvoja spoločnosti (2003 – 2006). Tieto projekty boli v tejto oblasti vôbec prvými projektmi so štátnou podporou, ktorá prišla prakticky v hodine dvanástej, keďže okolité krajiny už pred koncom tisícročia disponovali rozvinutými počítačovými a korpusovými pracoviskami i verejne prístupnými korpusmi svojich jazykov. Po ukončení prvej fázy projektu, ktorý verejnosť veľmi uvítala a ktorý bolo potrebné rozvíjať ďalej, Ministerstvo školstva, vedy, výskumu a športu SR prispelo k zachovaniu a ďalšiemu rozvoju pracoviska Slovenského národného korpusu v druhej etape trvajúcej do konca r. 2011 a v súčasnosti je signatárom zmluvy o spolupráci pri budovaní Slovenského národného korpusu a elektronizácii jazykovedného výskumu na Slovensku na obdobie do r. 2016. V súčasnej finančne náročnej situácii nie je jednoduché založiť nové vedecko-výskumné pracovisko a udržať jeho dlhoročné financovanie. No prínosy pre široké spektrum používateľov sa nedajú prehliadnuť ani sa nedá predpokladať, že by sa databázy slovenského jazyka mohli budovať niekde v zahraničí. O výsledky počítačového spracovania slovenčiny sa zaujímajú aj za našimi hranicami, o čom svedčí zapojenie Slovenského národného korpusu do medzinárodných projektov. Želám tomuto rokovaniu zdarný priebeh a všetkým jeho účastníkom dobre využitý čas pri výmene poznatkov a skúseností v oblasti jazykových a informačných technológií.

Bratislava 7. jún 2012

Marta Cimbáková
generálna riaditeľka sekcie vedy a techniky
MŠVVaŠ SR

META-NET: A Research and Technology Strategy for Multilingual Europe

Georg Rehm

German Research Center for Artificial Intelligence, Berlin, Germany

Abstract. Speaking one's mother tongue, be it Latvian, Hungarian, or Portuguese, must not become a social or economic disadvantage in the networked European information society of the 21st century. Language Technology has the potential to become the key solution to this crucial challenge if it is robust, cost-effective as well as available for all European languages and to all European citizens. However, in order to achieve these goals, the pace of research and development has to be accelerated by means of a major, dedicated push. A push with the magnitude needed can only be put into motion through a joint action of all stakeholder groups involved such as, among others, researchers, user and provider industries, technology integrators, language communities including the national institutions for language, politicians and society in general. To this end, META-NET – a European Network of Excellence that consists of 60 Language Technology research centres from 34 countries – is building META, the open and constantly growing Multilingual Europe Technology Alliance. This article introduces the META-NET White Paper Series in which we provide surveys on the state of language technology support for 30 European languages. Additionally, we present three priority research themes as strategic and unifying umbrella topics for the next 10 to 15 years of future European language technology research.

1 Introduction

Many European languages run the risk of becoming victims of the digital age as they are under-represented and under-resourced online. Huge regional market opportunities remain untapped because of language barriers (Directorate-General for Translation of the European Commission, 2009). If we do not take action now, speaking their mother tongue will become a severe social and economic disadvantage for many European citizens.

Innovative multilingual Language Technology (LT) is the ultimate intermediary that can help all European citizens to participate in an egalitarian, inclusive, and economically successful knowledge and information society. It can also help to establish and to further the single digital market. However, the degree to which LT is used and actually can be used in Europe varies enormously from language to language.

We are currently witnessing a revolution whose impact on language and society is comparable to that of Gutenberg's invention of the printing press. Digitisation and networked communication technology make possible an unlimited exchange of information and services – at any place, at any time. The downside is that certain groups (for example, people who live in rural areas or senior citizens) have difficulties participating in this new information-driven society. This problem is known as the digital divide.

Digital communication will have far-reaching and dramatic effects on Europe's languages, just like modern printing did five hundred years ago. Back then the new opportunities of large-scale communication triggered orthographic and grammatical standardisation for some languages and made the rapid dissemination of new scientific and intellectual ideas possible. At the same time, small languages and regional dialects were rarely put to print. This turned out to be a considerable disadvantage as it limited their sphere of use to oral conversation and sometimes even contributed to their eventual extinction.

Today's multitude of official and unofficial languages as well as regional dialects is one of Europe's richest and most important cultural assets and it is also a vital part of its social success story. While big languages such as English or Chinese will certainly be well represented in the emerging digital society, many European languages are in real danger if we do not act now.

The key for protecting and furthering the highly heterogeneous group of more than 60 European languages is Language Technology. Research in this area has made considerable progress in the last few years. Machine Translation (MT) delivers a reasonable amount of accuracy, albeit only in specific domains, and experimental applications provide multilingual information and knowledge management as well as content production across languages. Relevant related areas are the development of intuitive language-based interfaces to technology ranging from household appliances, to heavy-duty machinery, vehicles and robots. The entertainment sector including games and mobile information services also holds many opportunities, as does the educational sector with computer assisted language learning and self-assessment software. While prototypes for several of these technologies exist, they are, however, by no means perfect and not ready for production use, yet. Nevertheless, it is safe to say that current progress opens a genuine window of opportunity.

Unfortunately, the current pace of technological progress is much too slow to arrive at substantial software products and services that are able to move communication in a multilingual environment significantly forward within the next 10 to 20 years. Those basic technologies that are already widely used nowadays are usually monolingual and only available for a handful of languages. Well-known examples of the broad use of LT are the spelling and, recently, grammar correction features in modern text processing systems.

Applications for multilingual communication such as machine translation require a certain level of sophistication. Services that are available online, such as Google Translate or Bing Translator, are helpful when it comes to getting a rough idea of what a document in a foreign language is about. However, both products and also professional MT applications are fraught with multiple difficulties, especially if precise and also complete translations are needed.

The internet connects people, nationally and internationally, through a massively growing number of, in recent years especially mobile, devices. Multilingual LT enables instantaneous, cheap and effortless communication and interaction as well as business transactions across language borders. Once the necessary infrastructure is in place and

research breakthroughs have been achieved, LT will allow people to collaborate, do business, share knowledge, and to participate in social and political opinion forming. LT offers tremendous opportunities for the European Union and its highly multilingual environment, both from the viewpoint of the economy and also from that of the citizen. It also opens up transcontinental economic opportunities as experiences made developing, using and refining multilingual LT within the EU could be adapted to the specific needs of other multilingual communities, e.g., the citizens of India or Africa.

2 Three Lines of Action

META-NET is a European Network of Excellence forging the Multilingual Europe Technology Alliance (META) through a concerted effort to build a strong and powerful European community for and around LT (Rehm and Uszkoreit, 2011). Its goal is to prepare the grounds for multilingual applications that enable automatic translation, information and knowledge management, including localisation, as well as content production and applications in related areas across all European languages. The objective is to advance LT so that communication and cooperation across languages becomes possible and to secure users of any language equal access to information and knowledge. META-NET, which started work on February 1st, 2010, aims to advance research in LT as a means towards realising the vision of a Europe united in a single digital market and information space and is supporting these goals by pursuing three lines of action: META-VISION, META-SHARE and META-RESEARCH (see Figure 1).

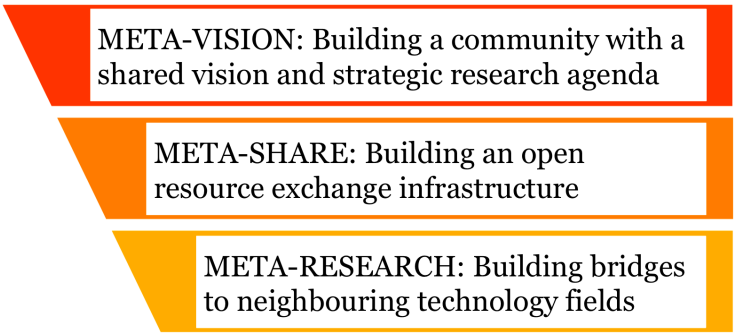


Fig. 1. META-NET's three lines of action

2.1 META-VISION: Fostering a Dynamic and Influential Community around a Shared Vision and Strategic Research Agenda

This first line of action is concerned with a goal that is not only important but strategically indispensable for the overall success of the initiative: building up a coherent and homogeneous European LT community by bringing together representatives from the highly fragmented and heterogeneous stakeholder groups. These are comprised of, amongst others, researchers, user industries as well as provider industries, administrators, politicians, technology integrators and representatives of the language communities and national institutions for language. Significant steps towards realising this goal have been taken through various means such as, for example, by mobilising ca. 70 participants for three think tanks and focus groups called Vision Groups; these are made up of external experts from industry who provide seed ideas for innovative LT application scenarios for the future knowledge and information society. The Vision Groups focus on the areas of "Translation and Localisation", "Media and Information Services" and "Interactive Systems". Furthermore, META-NET has been engaged in intense dissemination activities with presentations at multiple national as well as international conferences. In addition to the wide visibility provided by successful mobilisation activities, META-NET organised four conferences of its own: theMETAnk 2010 was a brainstorming workshop at which about 100 researchers presented and discussed their long-term visions for the field of Human Language Technologies (June 4/5, Berlin). At Translingual Europe 2010 (June 7, Berlin) researchers discussed current problems and visions with representatives from the provider industries (such as Microsoft, Asia Online and ProMT) and Language Technology as well as Machine Translation users (European Patent Office, Symantec, EC DGT). At META-FORUM 2010 (November 17/18, Brussels, Belgium) the initial results of the vision building process were showcased to more than 250 participants. In a series of interactive sessions the participants provided their feedback and views on the visions presented by the project. At the follow-up conference, META-FORUM 2011 (June 27/28, Budapest, Hungary), more than 300 participants from research and industry came together for, among many other agenda items, an update on the vision building process, drafts of the META-NET Language White Paper Series (see section The META-NET Language White Paper Series: Language Technology Support for Europe's Languages), a first outline of META-NET's Strategic Research Agenda (see section The Strategic Research Agenda for Multilingual Europe) and the META Prize and META Seal of Recognition award ceremonies; in the META Exhibition 40 exhibitors showcased LT products and recent research results. At META-FORUM 2012 – held in Brussels again on June 20/21, 2012 and collocated with the Digital Agenda Assembly 2012 – we focused upon the strategic aspects of our initiative and presented as well as discussed the Strategic Research Agenda and the final META-NET White Paper Series, among several other topics. Videos of the presentations and discussions as well as reports about the events are available on the META-NET website.

2.2 META-SHARE: Creating an Open Resource Exchange Infrastructure

In its second line of action META-NET is building META-SHARE, a sustainable peer-to-peer network of repositories of language data, tools and web services that are documented with high-quality metadata and aggregated in inventories allowing for uniform search and access. Data and tools can be both open and with restricted access rights, free and for-a-fee. META-SHARE targets existing but also new and emerging language data, tools and systems required for building and evaluating new technologies as well as innovative products and services. In this respect, reuse, combination, repurposing and re-engineering of language data and tools play a crucial role. META-SHARE will eventually become an important component of an LT marketplace for researchers and developers, language professionals (translators, localisation experts, etc.), as well as for industrial players including SMEs and big enterprises. In this role, META-SHARE will cater for the full development cycle of LT, from research through to innovative products and services. In this regard, designing, building and successfully establishing META-SHARE as an important and valuable piece of infrastructure within the European and also global LT community is one of META-NET's decisive goals (Piperidis, 2012). Among the important relevant components of the META-SHARE infrastructure is a universal metadata scheme for the description of Language Resources and Language Technologies that was developed by a working group that consists of experts from within the initiative and several other European specialists (Gavrilidou et al., 2012). We also explored thoroughly the landscape of language resources licensing and, with the help of legal experts, prepared a set of licensing templates. META-SHARE favours and aligns itself with the growing open data and open source movement, especially the Creative Commons Initiative. A first, fully functional prototype of META-SHARE was presented at META-FORUM 2010. Currently META-SHARE is in production use within the network of excellence. An improved version will be rolled out for use both by META-NET and the public at large in the autumn of 2012 (Federmann et al., 2012).

2.3 META-RESEARCH: Building Bridges to Neighbouring Technology Fields

The third line of action consists of innovative research work with regard to leveraging advances in other fields to help LT. Specifically the work focuses on bringing more semantics into Machine Translation (MT), optimising the division of labour in hybrid MT, preparing an empirical base for MT and exploiting the context when computing an automatic translation. To this end, META-NET is carrying out research by building bridges to other fields and disciplines such as Machine Learning and the Semantic Web community. META-RESEARCH is concerned with collecting data, preparing data sets and language resources for evaluation purposes, compiling inventories of tools and methods, and organising workshops and advanced training events for its staff members. Among its current major outcomes are the clear identification of issues in Machine Translation in

which semantics has shown potential to positively impact the state of the art, recommendations for approaching the problem of integrating semantic information in MT, and a list of tools and resources that could be employed for this purpose. A new language resource for MT, the Annotated Hybrid Sample MT Corpus, provides data for the language pairs English-German, English-Spanish and English-Czech. A third important outcome is software for the collection of multilingual hidden-web corpora. The tool clusters news articles in different languages discussing the same topic or event and clusters pages identified as being translations of each other. The research that is carried out in this line of action is meant to advance significantly the state of the art in MT.

2.4 Extension – Impact – Collaborations

META-NET has a founding consortium that consists of 13 partners in 10 countries. As the initiative operates on a European level we began to extend the network in November 2010. In the autumn of 2012 the enlarged network consists of 60 members in 34 countries (see Table 5). Most of the new members participate in three EU-funded projects that support the META-NET objectives by systematically collecting language resources and language technologies, curating and describing them with metadata records and making them available through META-SHARE, mobilising the communities in their respective countries and organising general awareness raising activities. These three projects – CESAR, METANET4U and META-NORD – commenced their work on February 1st, 2011 – exactly one year after the start of META-NET.

In addition to the network of excellence and the open technology alliance META (Multilingual Europe Technology Alliance), META-NET drafted and signed individual collaboration agreements with more than 40 projects and initiatives funded by the European Union. Among these are machine translation projects such as ACCURAT, Let's MT, EuroMatrix Plus and iTranslate4.eu, Language Technology projects such as PANACEA and ATLAS, web- and W3C (World Wide Web Consortium)-related projects such as Multilingual Web and Multilingual Web-LT as well as large European initiatives and networks such as CLARIN and FLaReNet.

We expect META-NET to have a significant, hitherto unprecedented, long-term impact on the European LT landscape. The three tightly integrated lines of action all aim at the same goal, albeit on different levels: to provide technological antidotes for the language barriers Europe has been facing for quite some time, in the form of robust, precise, high-performance, multilingual Language Technology; to stimulate the development of novel and innovative LT applications; to assemble and strengthen the European LT community; to raise awareness about the enormous potential Language Technology has for the European information society, for the single digital market and for society at large; to foster innovative LT research.

3 The META-NET Language White Paper Series: Language Technology Support for Europe's Languages

The META-NET Language White Paper Series describes the current state of language technology support for 30 different European languages and is a complement to the Strategic Research Agenda (see section The Strategic Research Agenda for Multilingual Europe). The individual volumes are meant to raise awareness for the topic of language technology. The target audience are mostly national but also international politicians, journalists, decision makers and the public at large.

The white papers include a language-specific assessment of existing technologies and resources, shortcomings and gaps, as well as a cross-language comparison. The documents were written for the following European languages (including all 23 EU member state languages): Basque, Bulgarian, Catalan, Croatian, Czech, Danish, Dutch, English, Estonian, Finnish, French, Galician, German, Greek, Hungarian, Icelandic, Irish, Italian, Latvian, Lithuanian, Maltese, Norwegian (bokmål and nynorsk), Polish, Portuguese, Romanian, Serbian, Slovak, Slovene, Spanish, and Swedish. Each Language White Paper is written in the language it reports upon and includes a complete English translation.

Draft versions of the Language White Papers were distributed at META-FORUM 2011 in Budapest, the first final versions were presented at META-FORUM 2012 in Brussels; the documents including additional information such as quotes and testimonials from politicians are available online at `http://www.meta-net.eu` and can also be purchased as print editions. In total, more than 200 authors and additional experts contributed to preparing the Language White Papers.

The current state of LT support varies considerably from one language community to another. In order to compare the situation between languages, the Language White Papers introduce an evaluation based on two sample application areas (machine translation and speech processing) and one underlying technology (text analysis) as well as basic resources needed for building LT applications. LT support for the languages was categorised using a five-point scale (1. excellent support; 2. good support; 3. moderate support; 4. fragmentary support; 5. weak or no support) and measured according to the following key criteria:

Speech Processing: quality of existing speech recognition and synthesis technologies, coverage of domains, number and size of existing speech corpora, amount and variety of available speech-based applications.

Machine Translation: quality of existing MT technologies, number of language pairs covered, coverage of linguistic phenomena and domains, quality and size of existing parallel corpora, amount and variety of available MT applications.

Text Analysis: quality and coverage of existing text analysis technologies (morphology, syntax, semantics), coverage of linguistic phenomena and domains, amount and variety of available applications, quality and size of existing (annotated) text corpora, quality and coverage of existing lexical resources (e.g. WordNet) and grammars.

Language Resources: quality and size of existing text corpora, speech corpora and parallel corpora, quality and coverage of existing lexical resources and grammars.

The more than 200 contributing authors to the Language White Papers prepared an initial language-specific assessment of LT support using an approach in which ca. 25 typical application areas and tools as well as resource types were assessed along seven different axes and criteria. Later on, the 30 individual and language-specific matrices were condensed in multiple iterations in order to arrive at a single score per language and area.

Tables 1 to 4 show that there are dramatic differences in language technology support between the various European languages and technology areas. For all LT areas, English is ahead of any other language but even support for English is far from being perfect. While there are good quality software and resources available for some languages and application areas, others, usually smaller languages, have substantial gaps. Many languages lack basic technologies for text analysis and essential resources. Others have basic tools and resources but the implementation of, for example, semantic methods is still far away. Therefore, a large-scale effort is needed to attain the ambitious goal of providing high-quality language technology support for all European languages, for example through high quality machine translation.

Excellent support	Good support	Moderate support	Fragmentary support	Weak/no support
	English	Czech Dutch Finnish French German Italian Portuguese Spanish	Basque Bulgarian Catalan Danish Estonian Galician Greek Hungarian Irish Norwegian Polish Serbian Slovak Slovene Swedish	Croatian Icelandic Latvian Lithuanian Maltese Romanian

Table 1. Speech processing – state of LT support for 30 European languages

Excellent support	Good support	Moderate support	Fragmentary support	Weak/no support
	English	French Spanish	Catalan Dutch German Hungarian Italian Polish Romanian	Basque Bulgarian Croatian Czech Danish Estonian Finnish Galician Greek Icelandic Irish Latvian Lithuanian Maltese Norwegian Portuguese Serbian Slovak Slovene Swedish

Table 2. Machine translation – state of LT support for 30 European languages

Excellent support	Good support	Moderate support	Fragmentary support	Weak/no support
	English	Dutch French German Italian Spanish	Basque Bulgarian Catalan Czech Danish Finnish Galician Greek Hungarian Norwegian Polish Portuguese Romanian Slovak Slovene Swedish	Croatian Estonian Icelandic Irish Latvian Lithuanian Maltese Serbian

Table 3. Text analysis – state of LT support for 30 European languages

Excellent support	Good support	Moderate support	Fragmentary support	Weak/no support
	English	Czech Dutch French German Hungarian Italian Polish Spanish Swedish	Basque Bulgarian Catalan Croatian Danish Estonian Finnish Galician Greek Norwegian Portuguese Romanian Serbian Slovak Slovene	Icelandic Irish Latvian Lithuanian Maltese

Table 4. Speech and text resources – state of LT support for 30 European languages

As the META-NET Language White Papers serve as important communication instruments, high-quality paper copies and also ebook versions are available both through a scientific publishing house and the META-NET website in order to ensure a wide distribution to non-specialists, journalists, politicians, administrators and other stakeholders (Rehm and Uszkoreit, 2012).

4 The Strategic Research Agenda for Multilingual Europe

In addition to building up a coherent, dynamic and influential European LT community and to preparing the META-NET Language White Paper Series, another important goal of the META-VISION line of action is collaboratively – within and by the community – to prepare, establish and also promote a Strategic Research Agenda (SRA) for the European LT landscape. The SRA is intended to be a long-term instrument that will serve as a unifying umbrella for both industrial and academic research and development in the period leading up to 2020. The SRA contains high-level recommendations and suggestions for joint actions to be presented to the European Commission and national as well as regional bodies and funding agencies. The process of preparing the SRA is complex and includes representatives of META-NET, the abovementioned three Vision Groups (see section META-VISION: Fostering a Dynamic and Influential Community around a Shared Vision and Strategic Research Agenda) and external experts.

After the three Vision Groups had collected literally hundreds of attractive and powerful technology visions in 2010 and early 2011 (see Mariani and Magnini, 2010; Koutsombogera and Piperidis, 2010; Burchardt and Rehm, 2010; Burchardt, Rehm, and Sasaki, 2011; for more details), the META Technology Council, a group that consists mostly of industry representatives and several researchers, took over in the complex process of preparing the SRA. The Technology Council discussed these visions in several meetings, reducing the number of potential visions to a shortlist of seven. The key criterion in these discussions was that the respective technology vision needed to be attractive and powerful enough to assemble behind it a very large proportion of the European LT research and innovation landscape. At the same time the vision needed to present a convincing solution for the issue of technology-enabled multilingualism in Europe. Towards the end of 2011 we agreed upon three main research strands that we call priority themes (Burchardt, Rehm, and Uszkoreit, 2012):

1. **Translation Cloud –** The goal of this priority theme is a multilingual European society, in which all citizens can use any service, access all knowledge, enjoy all media and control any technology in their mother tongues. Written and spoken communication is not to be hindered anymore by language boundaries. Costs for large volume and specialized high-quality translation will be truly affordable. The citizen, the professional, the organization, or the software application in need of cross-lingual communication will use a single access point for channelling text or speech through a gateway that will instantly return the translations into the requested languages in the required quality and desired format. Behind this access point will be a network of generic and special-purpose services combining automatic translation or interpretation, language checking, post-editing, as well as human creativity and quality assurance where needed for achieving the demanded quality. The service will be free for small volume use and for high-volume baseline quality but it will offer extensive business opportunities for a wide range of language service and language technology providers.
2. **Social Intelligence and e-Participation –** The goal is to improve decision-making in business and society. The quality, speed and acceptance of decisions are the main factor for the success of social systems such as enterprises, communities, states and supranational organisations. Business intelligence and analytics programmes search online data for relevant information, decision support systems evaluate and order the information and apply decision rules. Social intelligence builds upon improved text analytics methodologies for the analysis of large volumes of social media, comments, blogs, forum postings etc. of citizens, customers, patients, consumers and other members of arbitrary stakeholder communities. Part of the analysis is directed to the status, opinions and acceptance associated with individual information units. As the formation of collective opinions and attitudes is highly dynamic, new developments need to be detected and trends to be analysed. Emotions and sentiment play an important part in

actions such as voting, buying, supporting, donating and in collective opinion formation. Social intelligence does not just analyse but also support collective deliberation processes. Precise and robust multilingual technologies are needed to support discussion and deliberation processes on an international scale.

3. **Socially Aware Interactive Assistant –** Socially aware interactive assistants are conversational agents realized with or without a physical shell (from robots to different types of graphical or voice interfaces). Their behaviour leverages from the combination of analysis and synthesis of non-verbal, speech and semantic signals. It is the proper time to develop, implement and deploy socially aware and also multilingual assistants that can support and enhance the interaction of humans with their environment. This includes classical Human-Computer Interaction, Human-Artificial Agent (or robot) Interaction, and Computer-mediated Human-Human Interaction. Those assistants must be able to act in indoor environments (such as meeting rooms, offices, apartments), outdoor environments (streets, cities, transportation, roads) and virtual environments, and also be able to communicate, exchange information and understand the other agents' intentions. They must be suitable and/or able to adapt to the user's needs and environment. They must have the capacity to learn incrementally from all interactions and other sources of information. The ideal socially aware multilingual assistant can interact naturally with humans, in any language and in any communication modality, it can adapt and be personalized to individual communication abilities, it can recognize and generate speech incrementally and fluently.

From the short descriptions of the three priority themes one can easily see that the proposed research strands overlap in technologies and challenges. This intended overlap reflects the coherence and maturation of the field. At the same time, the division of labour and sharing of resources and results is a precondition for the realization of this highly ambitious research programme.

All three areas need to benefit from progress in core technologies of human language analysis and production such as morphological, syntactic and semantic parsing and generation. But each of the three areas will concentrate on one central area of LT: the Translation Cloud will focus on cross-lingual technologies such as translation and interpretation, the Social Intelligence strand will take care of knowledge discovery, text analytics and related technologies, and the research dedicated to the Interactive Assistant will take on interface technologies such as speech and multimodal interfaces (see Figure 2).

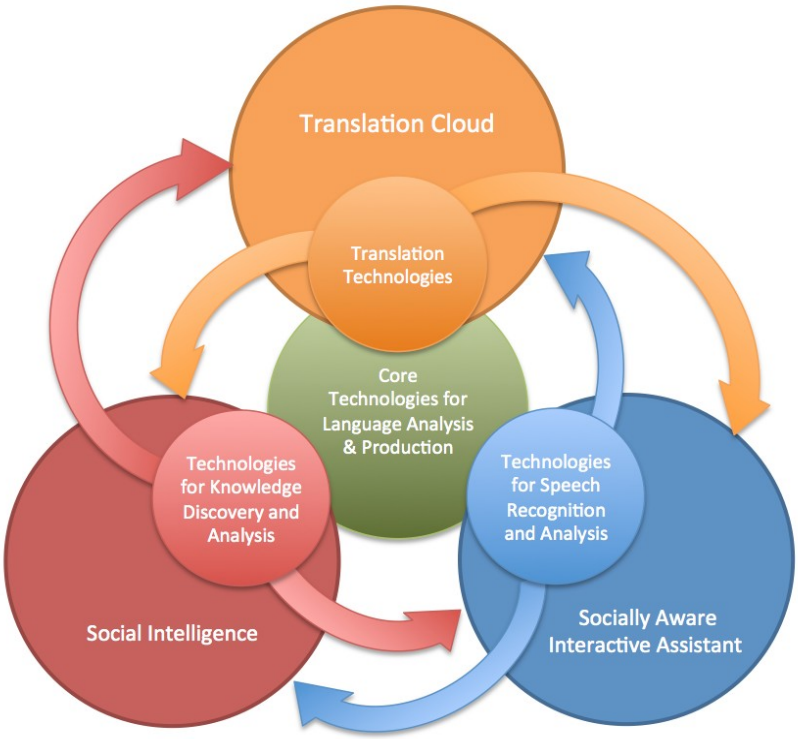


Fig. 2. The three priority themes proposed by the META-NET Strategic Research Agenda and topics for scientific cooperation among the themes

In addition to these three priority themes the Strategic Research Agenda contains lists of technology and application visions, plans, suggestions for the organisation of research and roadmaps for the path to a truly multilingual Europe, realised through high-performance, robust and precise Language Technology. A first draft version of the SRA was presented at META-FORUM 2012 in Brussels, it is also available online on the META-NET website. The open discussion process of the SRA will be completed in the autumn of 2012. Shortly after, the final version of the SRA will be presented to politicians and decision makers on the national and international levels.

5 Concluding Remarks: Get Involved and Participate

With a large and diverse community behind our goals, META-NET and META can achieve the critical mass needed to really make a difference as to how Language Technology can enable and secure multilingualism in Europe's future (European Commission, 2008; Directorate-General of the UNESCO, 2007). To researchers, technologists, professionals and administrators developing, providing or using language technologies and also to the European language communities the Multilingual Europe Technology Alliance (META) offers a unique opportunity to stay informed, contribute ideas or advocate on behalf of specific languages, while participating in expert discussions, working groups and planning activities that will shape Europe's linguistic future. META is an open and growing technology alliance that currently has more than 640 members including multiple research centres and universities, companies that develop and provide as well as companies that make use of language technologies and also many organisations that represent Europe's language communities.

Research and technology development projects are invited to join META-SHARE to access a pool of language resources and technologies while helping to validate and further to shape its services. Commercial enterprises are welcome to contribute their visions for products and services, to participate in our planning process and to use META to grow profitable partnerships. Schools and educators, journalists and the media, politicians, public institutions and organisations are encouraged to participate in open discussions on the vision of and way towards a truly multilingual information society.

Your voice is important, just like the language in which you express yourself. In joining our open technology alliance META and spreading the word, you'll be helping to shape the future of the European linguistic and also language technology landscape. Interested companies, research centres, institutions, organisations and individuals can join META without any financial obligations through a simple registration form: `http://www.meta-net.eu/join`.

6 Acknowledgements

This article is based on joint work with Aljoscha Burchardt, Kathrin Eichler, Tina Klüwer, Felix Sasaki and Hans Uszkoreit (all DFKI), the 60 member organisations of the META-NET network of excellence, the ca. 70 members of the Vision Groups, the ca. 30 members of the META Technology Council, the more than 200 authors of and contributors to the META-NET Language White Papers and the more than 130 representatives from industry and research who contributed to the META Strategic Research Agenda.

The META-NET Network of Excellence is co-funded by the 7th Framework Programme of the European Commission through the following grant agreements: T4ME

Net (no. 249119), CESAR (no. 271022), METANET4U (no. 270893) and META-NORD (no. 270899).

More information on META-NET and META is available at http://www.meta-net.eu and via office@meta-net.eu.

References

Burchardt, A. and Rehm, G. (eds.) (2010). *Vision Group Translation and Localisation – Results of first two Meetings. 2010. META-NET.* Available at: http://www.meta-net.eu/vision/reports/VisionGroup-TranslationLocalisation-draft.pdf

Burchardt, A., Rehm, G., and Sasaki, F. (eds.) (2011). *The Future European Multilingual Information Society – Vision Paper for a Strategic Research Agenda. META-NET.* Available at: http://www.meta-net.eu/vision/reports/meta-net-vision-paper.pdf

Burchardt, A., Rehm, G., and Uszkoreit H. (eds.) (2012). *LT 2020 – Vision and Priority Themes for Language Technology Research in Europe until the Year 2020. Towards a draft of the META-NET Strategic Research Agenda. META-NET. 2012.*

Directorate-General of the UNESCO (2007). Intersectoral Mid-term Strategy on Languages and Multilingualism. Available at: http://unesdoc.unesco.org/images/0015/001503/150335e.pdf

Directorate-General for Translation of the European Commission (2009). *Size of the Language Industry in the EU.* Available at: http://ec.europa.eu/dgs/translation/publications/studies

European Commission (2008). *Multilingualism: An Asset for Europe and a Shared Commitment.* Available at: http://ec.europa.eu/languages/pdf/comm2008_en.pdf

Federmann, Ch., Giannopoulou, I., Girardi, Ch., Hamon, O., Mavroeidis, D., Minutoli, S., and Schröder, M. (2012). META-SHARE Version 2: An Open Network of Repositories for Language Resources including Data and Tools. In Calzolari, N., Choukri, K., Declerck, T., Doğan, M. Uğur, Maegaard, B., Mariani, J., Odijk, J., and Piperidis, S., editors, *Proceedings of the Eighth International Conference on Language Resources and Evaluation – LREC 2012, 23-25 May, Istanbul, Turkey,* pages 3300–3303, Paris, France. ELRA.

Gavrilidou, M., Labropoulou, P., Desipri, E., Piperidis, S.; Monachini, M., Frontini, F., Declerck, T., Francopoulo, G., Arranz, V., and Mapelli, V. (2012). The META-SHARE Metadata Schema for the Description of Language Resources. In Calzolari, N., Choukri, K., Declerck, T., Doğan, M. Uğur, Maegaard, B., Mariani, J., Odijk, J., and Piperidis, S., editors, *Proceedings of the Eighth International Conference on Language Resources and Evaluation – LREC 2012, 23-25 May, Istanbul, Turkey,* pages 1090–1097, Paris, France. ELRA.

Koutsombogera, M. and Piperidis, S., editors, (2010). *Vision Group Media and Information Services – Results of first two Meetings.* META-NET. Available at: `http://www.meta-net.eu/vision/reports/VisionGroup-MediaInformationServices-draft.pdf`

Mariani, J. and Magnini, B., editors, (2010). *Vision Group Interactive Systems – Results of first two Meetings.* META-NET. Available at: `http://www.meta-net.eu/vision/reports/VisionGroup-InteractiveSystems-draft.pdf`

Piperidis, S. (2012). The META-SHARE Language Resources Sharing Infrastructure: Principles, Challenges, Solutions. In Calzolari, N., Choukri, K., Declerck, T., Doğan, M. Uğur, Maegaard, B., Mariani, J., Odijk, J., and Piperidis, S., editors, *Proceedings of the Eighth International Conference on Language Resources and Evaluation – LREC 2012, 23–25 May, Istanbul, Turkey,* pages 36–42 , Paris, France. ELRA.

Rehm, G. and Uszkoreit, H. (2011). Multilingual Europe: A challenge for language tech. In *MultiLingual*, 22(3): 51–52.

Rehm, G. and Uszkoreit, H. (eds.) (2012). *Europe's Languages in the Digital Age*. 31 volumes.

Country	Member (Affiliation)	Contacts
Austria	Universität Wien	Gerhard Budin
Belgium	University of Antwerp	Walter Daelemans
	University of Leuven	Dirk van Compernolle
Bulgaria	Bulgarian Academy of Sciences	Svetla Koeva
Croatia	Zagreb University	Marko Tadić
Cyprus	University of Cyprus	Jack Burston
Czech Rep.	Charles University in Prague	Jan Hajič
Denmark	University of Copenhagen	Bolette Sandford Pedersen, Bente Maegaard
Estonia	University of Tartu	Tiit Roosmaa
Finland	Aalto University	Timo Honkela
	University of Helsinki	Kimmo Koskenniemi, Krister Linden
France	CNRS, LIMSI	Joseph Mariani
	ELDA	Khalid Choukri
	University of Le Mans	Holger Schwenk
	University of Avignon	Georges Linares
Germany	DFKI	Hans Uszkoreit, Georg Rehm
	RWTH Aachen	Hermann Ney
	Saarland University	Manfred Pinkal
	University of Stuttgart	Jonas Kuhn, Hinrich Schütze
	Karlsruhe Institute of Technology	Alex Waibel
Greece	ILSP, R.C. “Athena”	Stelios Piperidis
Hungary	Hungarian Academy of Sciences	Tamás Váradi
	Budapest Technical University	Géza Németh, Gábor Olaszy
Iceland	University of Iceland	Eirikur Rögnvaldsson
Ireland	Dublin City University	Josef van Genabith
Israel	Bar-Ilan University	Ido Dagan
Italy	Consiglio Nazionale Ricerche	Nicoletta Calzolari
	Fondazione Bruno Kessler	Bernardo Magnini
Latvia	Tilde	Andrejs Vasiļjevs
	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	Universiteit Utrecht	Jan Odijk
	University of Groningen	Gertjan van Noord
Norway	University of Bergen	Koenraad De Smedt
	University of Oslo	Stephan Oepen
Poland	Polish Academy of Sciences	Adam Przepiórkowski
	University of Lódź	Barbara L.-Tomaszczyk
	Adam Mickiewicz University	Zygmunt Vetulani

Country	Member (Affiliation)	Contacts
Portugal	University of Lisbon	Antonio Branco
	Institute for Systems Engineering and Computers	Isabel Trancoso
Romania	Romanian Academy of Sciences	Dan Tufiş
	University Alexandru Ioan Cuza	Dan Cristea
Serbia	Belgrade University	Duško Vitas, Cvetana Krstev
	Pupin Institute	Sanja Vraneš
Slovakia	Ľ. Štúr Institute of Linguistics	Radovan Garabík
Slovenia	Jožef Stefan Institute	Marko Grobelnik
Spain	Barcelona Media	Toni Badia
	Technical University of Catalonia	Asunción Moreno
	University Pompeu Fabra	Núria Bel
	University of the Basque Country	Inma Hernaez Rioja
	University of Vigo	Carmen García Mateo
Sweden	University of Gothenburg	Lars Borin
Switzerland	Idiap Research Institute	Hervé Bourlard
Turkey	Tübitak Bilgem	Mehmet Uğur Doğan
UK	University of Manchester	Sophia Ananiandou
	University of Edinburgh	Steve Renals
	University of Wolverhampton	Ruslan Mitkov
	University of Sheffield	Rob Gaizauskas

Table 5. Current composition of the META-NET Network of Excellence (autumn of 2012)

Supporting Multilingual Europe
The CESAR initiative

Tamás Váradi

Research Institute for Linguistics, Hungarian Academy of Sciences,
Budapest, Hungary

Abstract. Multilingualism is an inherent asset of the European Union, yet it also presents great challenges. Without doubt, the key to meeting the challenges of Multilingual Europe in the age of digital cultures lies in language technology. The CESAR (Central and South-east European Resources) project was created to address the multilingual challenges within the geolinguistic area represented by the consortium (see Figure 1). A central concern of the project that will also be the focus of the present article is to contribute to a pan-European digital resource exchange facility by collecting resources and by documenting, linking and upgrading them to agreed standards and guidelines.

1 The CESAR consortium

The consortium consists of 9 partners covering 6 countries and 6 native languages[1]. The project is coordinated by the Research Institute for Linguistics of the Hungarian Academy of Sciences (Hungary), while the other partners are the following: Department of Telecommunications and Media Informatics, Budapest University of Technology and Economics (Hungary), University of Zagreb, Faculty of Humanities and Social Sciences (Croatia), Institute of Computer Science, Polish Academy of Sciences (Poland), University of Łodz (Poland), Faculty of Mathematics at the University of Belgrade (Serbia), Institut Mihajlo Pupin (Serbia), Institute for Bulgarian Language (Bulgaria) and Ľ. Štúr Institute of Linguistics, Slovak Academy of Sciences (Slovakia).

2 Project objectives

The main objective of the project is to make available a comprehensive set of language resources and tools covering Bulgarian, Croatian, Hungarian, Polish, Serbian and Slovak. The coverage of these languages brings about an added benefit of the project, anticipating and meeting foreseeable requirements with respect to resources from these languages. Building on a wide range of already existing resources and national or international activities, the project creates, populates and operates a comprehensive language-resource platform enabling and supporting large-scale multi- and cross-lingual products and services. The resources already involved (its number is continuously growing) in the project include interoperable mono- and multilingual speech databases, mono- and bilingual corpora, dictionaries, wordnets and relevant language technology processing tools such as tokenisers, lemmatisers, taggers and parsers (Váradi, 2011).

[1] Various data and materials (e.g. the public deliverables and presentations) can be found at `http://cesar-project.net` website.

Fig. 1. Geo-linguistic spread of CESAR

3 The main goals of the CESAR project are the following

CESAR is a large scale project with the aim to standardize the best language resources achievable in the regions covered. The scale of resources (and tools) is covering a wide variety of resource types and tools harvesting and processing information and data contained in them.

For that reason the main goals of the project could be summarized as the following:

- to provide a description of the national landscape in terms of language use; language-savvy products and services, language technologies and resources; main actors (research, industry, government and society); public policies and programmes; prevailing standards and practices; current level of development, main drivers and roadblocks;
- to contribute to a pan-European digital resources exchange facility by collecting resources and by documenting, linking and upgrading them to agreed standards and guidelines;
- to collaborate with other partner projects, in particular concurrent ICT-PSP 6.1 pilot projects and mainly the META-NET network of excellence – and where useful, with other relevant multi-national forums or activities, such as FlaReNET and CLARIN – to ensure consistent approaches, practices and standards facilitating a wider accessibility of, easier access to and reuse of quality language resources and tools;
- to help build and operate broad, non-commercial, community-driven, interconnected repositories that can be used by language researchers, developers and professionals (META-SHARE nodes and META-SHARE managing nodes);

- to mobilise national and regional stakeholders, public bodies and funding agencies by raising awareness, organizing meetings and other focused events;
- to reinvigorate cooperation between key technology partners in the region, building on previous collaboration in TELRI, MULTEXT-East and other projects;
- to bridge the technological gap between this region and the other parts of Europe by filling obvious and important blind spots in language resources and tools infrastructure.

4 CESAR operates as integral part of META-NET

The CESAR project is a part of the wider network of excellence called META-NET, a Network of Excellence funded by the European Union. It currently consists of 44 members, representing 31 EU countries. META-NET cooperates with a dozen other large initiatives like CLARIN, which is helping social sciences to establish the field Digital Humanities in Europe.

CESAR actively collaborates with other partner projects within META-NET, ensures consistent approaches, practices and standards aimed at ensuring a wider accessibility and easier access and reuse of quality language resources.

The CESAR project aims to stimulate ICT-based cross-lingual communication, collaboration and participation and thereby contribute to the creation of a pan-European digital single market by stimulating ICT-based cross-lingual communication, collaboration and participation.

5 Gaps and Challenges

One of the main goals of CESAR project is to bridge the technological gap between the Central and South-east European region and the other parts of Europe by filling obvious and important gaps in language resources and tools infrastructure. ICT research has started in this region with a lag behind Western European countries. Language technology has emerged in the respective participating countries autonomously, i.e. with national support both in the academic and in the private sector. As a result, the resources developed often reflect the isolated circumstances of their creation, and still often lack standardisation. The main actors of ICT research are however now ready to reinvigorate cooperation between key technology partners in the region, and to integrate national resources on a higher level in order to make them more accessible and interoperable, making them available to the wider language technology community to ease and speed up the provision of multilingual online services. To this end, existing resources are going to be assembled and upgraded so that they comply with widely used standards or community practices.

6 CESAR in META-SHARE

Key resources covered by the CESAR project are now linked and made interoperable using the facilities of the META-SHARE repository, aiming to build an open resource exchange infrastructure. The target user community of the resources practically embraces all stakeholders at

the modern digital market: everyday end-users, professional end-users (business, administration, media, education, libraries, etc.) as well as expertise holders (researchers, industrialists, policy makers, etc). Its concern is a careful investigation of the needs of various types of users – from individual users to large multinational organisations – from the perspective of the current status as well as from the near future prospects.

The CESAR project is contributing valuable resources to META-SHARE, which will eventually be an important component of a language technology marketplace for HLT researchers and developers, language professionals (translators, interpreters, content and software localisation experts, etc.), as well as for industrial players, especially SMEs, catering for the full development cycle of HLT, from research through to innovative products and services.

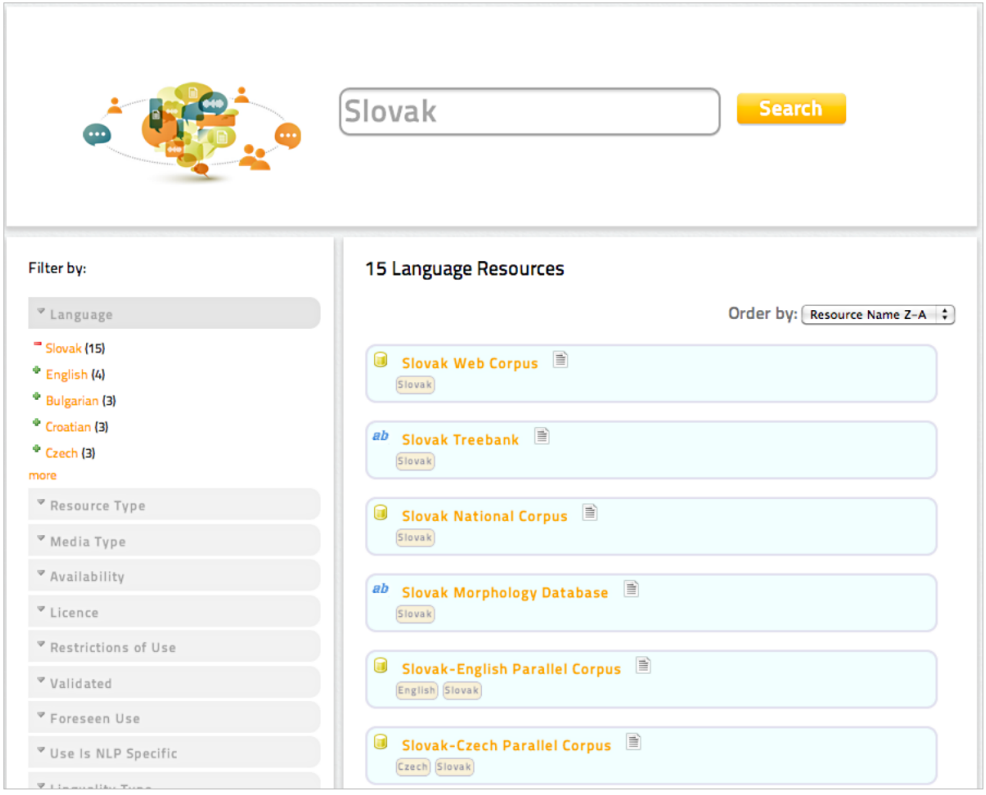


Fig. 2. Slovak resources in META-SHARE (`http://nlp.ipipan.waw.pl/metashare`)

7 Timeline

The project started on the 1st of February 2011 and has been running for two years. The resources are expected to be published in three batches. The result of the actions had been already achieved on the so called “first batch” of resources, which was published in early December 2011 and on

the "second batch" published in July 2012. The first batch covers 52 resources for six languages and contains 31 corpora, 11 lexical resources and 10 tools (technology, tools or services) while the second batch contains 32 corpora, 13 lexical resources and 20 tools[2].

The first batch was contributed by the project partners, but in the other two batches CESAR aims to involve resources from other, nationally relevant centres to increase the number of enhanced resources and spread the CESAR quality (which became well-known in involved countries).

8 Criteria for being involved

In the beginning of the project a methodology for selection was develop by which the identified language resources could be evaluated. A query was distributed among the partners to solicit suggestions on how to approach the evaluation procedure as it was confirmed that no single current methodology can be accepted as a standard. Instead, the consortium developed a list of four general indicators that were considered representative and indicative for the selection of language resources. The indicators determine the general requirements to which the selection should be subjected. Different sets of specific criteria have been defined for each indicator. The indicators are described in the following subsections (Ogrodniczuk et al., 2012).

In the process of enhancement the resources and tools the general evaluation is carried out in three flows: resource upgrade, extension, and cross-lingual alignment. Among these indicators the following criteria must be fulfilled:

For upgraded resources:

- All selected resources are state-of-the-art representatives of their type for a given language
- Equally valuable representatives are all included in the selection
- Current status of resources have superior quality at least on regional level without the need of excessive further development
- Licensing issues allow free processing and access to resources and resource-related materials or the consortium succeeds in reaching an agreement with respective copyright holders.

For extended/linked resources:

- The extension of resources provides considerable value to the community, at least on regional level
- The emphasis is on providing building blocks to the existing tools rather than major restructuring
- Additional resources are integrated with the existing ones only if they significantly improve the quality of resulting resources

[2] The third (and last) seems to cover at least 73 language resources which means a constantly increasing number of included resources. This amount shows the growing and lasting interest among the language resource creators (some of the resources are being upgraded during more than one batch. The third batch will be launched in the last month of the project, in January 2013.

- If more than one representative of a certain resource type for a language has been selected, they are very likely to be interlinked to benefit from strong sides of both solutions
- If less-developed, but still very popular resources can benefit from the enhancement due to their well-developed equivalent, their enhancement is also considered
- Experience of other consortium members/other consortia is extensively used in the process of extension of national resources to provide strong foundation for cross-lingual coverage
- Tools that are language-neutral or cross-lingual, are preferred.

For resources aligned across languages:

- No more than one tool of a certain type for each language is used
- Whenever applicable, the largest set of languages is selected
- Language Processing Tools in NooJ
- Language-independence is targeted to a great extent
- The quality of a result is of immense concern.

The soundness of specification cannot be judged without knowing the broader context of usage or adequacy of the language resource. To estimate the quality, quantity and importance, every case is thoroughly examined, taking into account regional determinants, popularity of the format outside its home institution. These indicators require a complex assessment of language resources in the context of the whole set of the established criteria. The partners not only appraise whether the selected resources fulfil the established criteria but also provide concrete examples and detailed explanations based on a thorough analysis.

9 Slovakia as part of CESAR

One of the pillars of the CESAR project is the Ľ. Štúr Institute of Linguistics (LSIL) with its valuable language resources – especially with a wide range of corpora connected with the Slovak language. Amongst the dedicated resources supported by the LSIL the most important are the monothematic (special) monolingual as well as the bilingual corpora covering several Slavic languages. These corpora compiled with the state-of-the-art technologies and techniques are representing the huge effort of the LSIL in representing Slovakia in the European digital market (Garabík, 2010; Šimková and Ľos, 2009).

As the list and short descriptions show the resources covered so far are representing the most essential resources for the language community and for the language technology business:

Batch 1.

- *Slovak National Corpus* – Slovak language written texts, contains about 770 million (lemmatised and MSD tagged) tokens
- *Corpus of Spoken Slovak* – Corpus of sound recordings of different types of (mostly spontaneous) speech. The recordings are transcribed orthographically and phonemically.
- *Morphology database* – Full paradigms of 77 000 lemmas, together with MSD tags, as used in the Slovak National Corpus.

- *Slovak-Czech parallel corpus* – Contains mostly fiction translated between Slovak and Czech (in both direction), with small amount of non-fiction texts and some translations from third language into both Czech and Slovak. The texts are automatically sentence-aligned, with some amount of texts aligned manually. 700 000 sentence pairs.
- *Slovak-English parallel corpus* – Contains original English fiction texts and their Slovak translations, with automatically aligned sentences.

Batch 2.

- *Balanced Slovak Corpus* – Corpus of ⅓ fiction, ⅓ informational text, ⅓ professional text (including popular science). The texts were selected from the Slovak National Corpus according to their style-genre annotation.
- *Dictionary of Slovak Collocations* – Contains selected (from Slovak National Corpus) multiword lexemes and phrasemes as well as typical collocations with restricted collocability.
- *Manually Annotated Slovak Corpus* – Full paradigms of 77 000 lemmas, together with MSD tags, as used in the Slovak National Corpus.
- *Slovak Web Corpus* – Corpus contains texts downloaded from the `.sk` domain. The texts are automatically lemmatized and morphologically tagged. The first version of the corpus contains 900 million tokens.
- *Slovak Legal Texts Corpus* – Contains entire body of law of the Slovak Republic, it has about 146 million tokens.
- *Slovak-French Parallel Corpus* – Corpus contains original French fiction texts and their Slovak translations, with automatically aligned sentences.
- *Slovak-Russian Parallel Corpus* – Corpus contains original Russian fiction texts and their Slovak translations, with automatically aligned sentences.
- *Slovak Terminology Database* – Monolingual database which at present contains 4 500 entries.
- *Slovak Treebank* – consists of 50 000 manually syntactically annotated sentences, using the Prague Dependency Treebank methodology.
- *Slovak WordNet* – A network of lexical-semantic relations, an electronic thesaurus with a structure modelled on that of the Princeton WordNet.

10 Concluding remarks

META-NET, and CESAR within it, is an excellent opportunity to promote language technology across Europe and to mobilize all stakeholders of the involved countries around a Strategic Research Agenda (which will be published in the last months of 2012). It serves as a most useful facility not only for the purpose of reaching all relevant business and government entities[3], and spreading information on language technology but it also to create an invaluable stock of

[3] The Language White Paper series (`http://www.meta-net.eu/whitepapers`) created through a huge META-NET wide collaborative effort, represents a pan-European horizontal perspective of the state of the art of the respective languages.

resources and tools (META-SHARE nodes and META-SHARE managing nodes). Slovak resources and tools are valuable components of this giant machinery (Garabík and Šimková, 2011; Šimková et al., 2012).

Although a huge part of the work of collecting and updating of language resources has been already done, the major work ahead to bridge the technological gap remains a task to carry out in the future – a task which indeed reaches beyond the timelines of the CESAR project.

References

Garabík, R. (2010). Slovak National Corpus tools and resources. In Laclavík, M. and Hluchý, L., editors, *Proceedings of the 5th Workshop on Intelligent and Knowledge oriented Technologies (WIKT 2010)*, pages 2–7, Bratislava, Slovakia.

Garabík, R. and Šimková, M. (2011). Slovak language in computer processing. *CLARIN Newsletter*, (11–12):19.

Ogrodniczuk, M., Garabík, R., Koeva, S., Krstev, C., Pęzik, P., Pintér, T., Przepiórkowski, A., Szaszák, G., Tadić, M., Váradi, T., and Vitas, D. (2012). Central and South-European language resources in META-SHARE. *Infotheca*, 12(1):3–26.

Váradi, T. (2011). Introducing the CESAR project. *Infotheca*, (12(1)):71–74.

Šimková, M., Garabík, R., Gajdošová, K., Laclavík, M., Ondrejovič, S., Juhár, J., Genči, J., Furdík, K., Ivoríková, H., and Ivanecký, J. (2012). *Slovenský jazyk v digitálnom veku – The Slovak Language in the Digital Age*. META-NET White Paper Series. Georg Rehm and Hans Uszkoreit (Series Editors). Springer, Berlin – New York. Available at: `http://www.meta-net.eu/whitepapers`.

Šimková, M. and Ľos, M. (2009). Frequency of Words and Forms in Contemporary Slovak (Based on the Slovak National Corpus). In Levická, J. and Garabík, R., editors, *Slovko 2009. Počítačové spracovanie prirodzeného jazyka, korpusová lingvistika a gramatický výskum*, pages 340–348, Brno, Czech Republic. Tribun.

11 References of the freely accessible corpora

Corpus of Spoken Slovak – `http://korpus.sk/shk.html`
Slovak-French Parallel Corpus – `http://korpus.juls.savba.sk/frask`
Slovak-Russian Parallel Corpus – `http://korpus.juls.savba.sk/parus`
Slovak-Czech parallel corpus – `http://korpus.juls.savba.sk/skcs.html`
Slovak-English parallel corpus – `http://korpus.juls.savba.sk/sken.html`
Slovak Terminology Database – `http://data.juls.savba.sk/std`

Slovenský národný korpus (2002 – 2012): východiská, ciele a výsledky pre výskum a prax

Mária Šimková – Radovan Garabík

Jazykovedný ústav Ľ. Štúra, Slovenská akadémia vied, Bratislava, Slovenská republika

Abstract. The Slovak National Corpus (SNK) project had been started since 2002 with the support of the Ministry of Education, the Ministry of Culture and the Slovak Academy of Sciences. SNK comprises several interconnected projects primarily for linguistic research and language teaching. First is the Slovak National Corpus itself – a huge corpus of modern written Slovak (since the 1953 orthography reform). Currently, the whole corpus prim-6.0 contains approximately $1.1 \cdot 10^9$ tokens and is constantly increasing in volume and improving in the quality of text conversion and annotation. Documents in the corpus keep rich metadata description, including detailed style and genre annotation, which is the base of dividing the corpus into specialized subcorpora (fiction, professional texts, journalistic texts, original Slovak fiction, balanced subcorpus). The corpus is automatically morphologically annotated and is publicly accessible for non commercial research purposes. The search interface (NoSketchEngine) provides CQL compatible query syntax with rich possibilities of statistical analysis and collocation extraction. Noticeable related corpora are: Manually morphologically annotated corpus (used for training of NLP tools), at 1.2 million words; Slovak WebCorpus, currently at about 10^9 tokens; Corpus of legal texts (a corpus of legal regulations of the Slovak Republic), prepared in collaboration with the Ministry of Justice of the Slovak Republic; Corpus of Spoken Slovak, aimed to provide a sample of spoken standard Slovak, including annotation of sound events and simple phonemic transcription, currently at 353 hours recordings = 2.61 million tokens; parallel corpora (Slovak-French, Slovak-Russian, Slovak-Czech, Slovak-English, Slovak-Latin). Another, separate project is the Slovak morphology database, Slovak Dependency Treebank, Slovak Terminology Database, Slovak WordNet, Corpus of Historical Slovak. Altogether the databases in the department of SNK contain nearly $3 \cdot 10^9$ tokens. The SNK team carried out research on the content of these resources in both national and international projects. Additional resources have been significantly expanded and their quality has been improved thanks to external financial support.

1 Východiská projektu Budovanie Slovenského národného korpusu a elektronizácia jazykovedného výskumu na Slovensku

V Slovenskom národnom korpuse Jazykovedného ústavu Ľ. Štúra Slovenskej akadémie vied v Bratislave (ďalej SNK JÚĽŠ SAV) sa za desať rokov existencie vybudovalo viacero textových, hovorených, všeobecných aj špecializovaných korpusov a databáz, v ktorých sa nachádzajú takmer 3 miliardy tokenov (slovných foriem a iných znakov alebo reťazcov znakov obsiahnutých v textoch) s pridanými externými a internými lingvistickými informáciami o pôvode, štýle, žánri a ďalších charakteristikách každého spracovaného textu (bibliografická a štýlovo-žánrová anotácia) a o lexikálno-gramatických kategóriách

každého slova v korpuse (lematizácia a morfologická, resp. morfosyntaktická anotácia). K elektronizácii jazykovedného výskumu prispievajú aj dve desiatky v SNK digitalizovaných lingvistických zdrojov, medzi ktorými sa nachádzajú viaceré staršie pravopisné príručky a lexikografické opisy slovenčiny vrátane piatich zväzkov slovníka A. Bernoláka z r. 1825, monografie, jednotlivé i periodicky vydávané zborníky, kompletné ročníky a čísla časopisov[1] vydávaných JÚĽŠ SAV a i. Korpusy, databázy a všetky digitalizované zdroje spolu s elektronickými verziami novších, resp. v súčasnosti vydaných slovníkov vrátane kodifikačných príručiek skoncipovaných v JÚĽŠ SAV[2] sú prostredníctvom slovníkového rozhrania SNK bezplatne k dispozícii širokej odbornej aj laickej verejnosti na Slovensku a v zahraničí na vyhľadávanie teoretických i praktických informácií o slovenčine prostredníctvom počítačovej siete Internet (`http://korpus.sk/dicts.html`, resp. `http://slovniky.juls.savba.sk/`).

Dosiahnutiu tohto pre aktuálny lingvistický výskum slovenčiny pomerne priaznivého stavu a aj medzinárodne porovnateľných a uznávaných výsledkov v oblasti budovania korpusov a korpusovej lingvistiky predchádzala fáza a) teoretických uvažovaní o štruktúre a rozsahu elektronicky spracovaného textového materiálu slovenského jazyka a jazykových informácií, b) praktického testovania internej databázy textov, c) hľadania nevyhnutných finančných prostriedkov a v neposlednom rade aj d) hľadania vhodného sídla korpusového pracoviska a riešiteľského kolektívu na jeho budovanie.

a) Situáciu v oblasti matematickej, počítačovej a kvantitatívnej lingvistiky na Slovensku v 60. – 80. rokoch 20. st. a neskoršiu koncepčnú, najmä makroštruktúrnu prípravu elektronickej bázy dát slovenského jazyka v JÚĽŠ SAV pod vedením J. Horeckého (1990a)[3] mapujú o. i. A. Jarošová (2001) a M. Šimková (2003, 2004, 2006, 2008) – tam aj ďalší prehľad literatúry. Súvisiace úvahy o mikroštruktúre jazykovej banky na báze kódovania kategoriálnych významov a o lexikálnych tezauroch prezentoval P. Žigo (1988a, 1988b, 1990), metodologické východisko počítačovej analýzy slovotvorného systému slovenčiny vypracoval J. Furdík (1990), automatizované generovanie slovenských slov a tvarov testoval E. Páleš (1994). Teoretické prípravy spočiatku prebiehali bez náležitého počítačového vybavenia, na základe poznania vývoja na zahraničných pracoviskách, možností počítačovej techniky a potrieb domáceho výskumu, od r. 1993 však narastal počet počítačových pracovných staníc a ich používateľov v rôznych oblastiach lingvistického výskumu v JÚĽŠ SAV i mimo neho (Jarošová, 1993; Šimková, 1996; Ďurčo, 1996, 1997;

[1] Celý časopis chápeme v súbore zdrojov ako jednu položku, pričom napr. Slovenská reč predstavuje za r. 1932 – 1996 približne 430 samostatných čísel (1 číslo = 1 dokument; viaceré čísla sú v rôznych formátoch, celkový počet dokumentov je teda vyšší), ktoré sa v SNK postupne digitalizovali a sprístupňovali. Od r. 1997 sa časopis Slovenská reč ukladal a spracovával v JÚĽŠ SAV v elektronickej podobe a jeho sprístupňovanie bolo postupne jednoduchšie.

[2] Elektronické formáty časti aktuálnej lexikografickej produkcie JÚĽŠ SAV sa technicky pripravovali v Oddelení spracúvania lingvistických dát JÚĽŠ SAV; podrobnejšie porov. ďalej v časti 2.7.

[3] Po zastavení prác v oddelení matematickej lingvistiky a fonetiky (1962 – 1970), ktoré J. Horecký založil a viedol v Ústave slovenského jazyka SAVU, od r. 1966 JÚĽŠ SAV, nadviazal na svoje predchádzajúce rozvíjanie moderných interdisciplinárnych metód v lingvistike opäť koncom 80. r. 20. st., keď sa príprava elektronickej bázy dát slovenčiny stala v JÚĽŠ SAV nevyhnutnou (Horecký, 1986a, 1986b, 1986/1987, 1990b).

Benko, 2001). Ďalšie smerovanie tvorby všeobecného textového korpusu slovenčiny, špecifických korpusov a súvisiacich databáz i nástrojov sa modifikovalo a konkretizovalo stále väčším zapájaním výpočtovej techniky do vedecko-výskumnej činnosti JÚĽŠ SAV (porov. napr. Benko, 1997; Šimková, 2004, 2012) a postupným zosúlaďovaním základných postupov pri tvorbe korpusov so štandardmi vypracovanými najmä v rámci TEI (Text Encoding Initiative) a ďalšími známymi štandardmi a postupmi osvedčenými na pracoviskách podobného typu (Garabík, 2004, 2005a).

b) V r. 1993 – 2001 sa v JÚĽŠ SAV budoval interný korpus textov, avšak bez dostatočného technického a personálneho zabezpečenia. Traja zainteresovaní pracovníci sa mu mohli venovať iba obmedzene, popri iných prácach (porov. aj Jarošová, 2001, s. 9), čomu zodpovedal jeho neveľký konečný rozsah (necelých 30 miliónov textových jednotiek; istú časť predstavovala lexikálna databáza slovníkových diel z produkcie JÚĽŠ SAV) a takmer výlučne textová podoba bez vnútornej lingvistickej anotácie na úrovni slova. No aj takýto korpus niekoľkonásobne presiahol rozsah papierovej kartotéky, ktorá sa v JÚĽŠ SAV budovala v predchádzajúcich desaťročiach (5 miliónov excerpčných lístkov), čím sa významne rozšírila a aktualizovala materiálová báza potrebná pri tvorbe koncepcie a v začiatkoch koncipovania výkladového Slovníka súčasného slovenského jazyka, ako aj pri príprave nových vydaní už existujúcich kodifikačných príručiek (Krátky slovník slovenského jazyka, 1997, 2003; Pravidlá slovenského pravopisu, 1998).

c) Opakované úsilie o finančné zabezpečenie budovania národného korpusu porovnateľného s korpusmi iných jazykov z vlastných zdrojov SAV – prostredníctvom viacerých projektov v grantovej agentúre VEGA – neprinieslo požadovaný výsledok, keďže z jej sústavne poddimenzovaného rozpočtu sa na tvorbu technicky a finančne náročného elektronického korpusu dostalo iba hraničné minimum z celkového objemu pôvodne plánovaných nevyhnutných prostriedkov. Získavanie textového materiálu, tvorba, archivovanie, zabezpečovanie fungovania verejne prístupných elektronických jazykových zdrojov veľkého rozsahu a vyvíjanie potrebných počítačových nástrojov však nebolo možné bez adekvátneho softvérového a hardvérového vybavenia a personálneho zabezpečenia (čo sa i počas existencie Slovenského národného korpusu ustavične potvrdzovalo a potvrdzuje tak v materiálno-technickej oblasti, ako aj, a to predovšetkým, v oblasti ľudských zdrojov). Na plynulú realizáciu všetkých súvisiacich prác bolo nevyhnutné hľadať potrebné finančné prostriedky aj z externého prostredia.

d) Po viacerých rokovaniach a konzultáciách v rámci SAV i s ďalšími zainteresovanými odborníkmi a inštitúciami sa na prelome tisícročí dospelo ku konečnému rozhodnutiu založiť v JÚĽŠ SAV nové oddelenie so zameraním na budovanie verejne prístupného korpusu slovenských textov. O inštitucionálnu a finančnú podporu sa JÚĽŠ SAV uchádzal s dvoma rámcovými projektmi – na tvorbu vlastného korpusu a na celkovú elektronizáciu lingvistického výskumu vrátane počítačového spracovania jestvujúcich jazykových zdrojov a tvorby jazykových technológií pre slovenčinu, na ktorých príprave participovali A. Jarošová, V. Benko a M. Šimková. Vo výslednej podobe boli obidva projekty fakticky spojené a pod záštitou Ministerstva kultúry SR, Ministerstva školstva SR a Predsedníctva SAV bolo 13. 2. 2002 prijaté uznesenie vlády SR č. 137 schvaľujúce *Projekt vybudovania Národného korpusu slovenského jazyka a elektronizácie jazykovedného výskumu v rokoch 2002 – 2006*, ktorým sa zabezpečilo financovanie zriadenia oddelenia Slovenského

národného korpusu JÚĽŠ SAV a jeho fungovania do konca r. 2006. To znamenalo fyzicky vybudovať a zariadiť nové priestory (rekonštruovala sa nevyužitá povala budovy JÚĽŠ SAV, kde vznikli podkrovné pracovné priestory) a najmä vytvoriť zodpovedajúci riešiteľský kolektív v podstate z externého prostredia. V JÚĽŠ SAV bolo síce vyššie spomínané know-how, no na riešenie takéhoto projektu neboli voľné vhodné riešiteľské kapacity. Pritom na žiadnej vysokej škole na Slovensku sa v tom čase nepripravovali špecialisti na počítačovú či korpusovú lingvistiku.

Vybudovaním nového pracoviska bola od 1. 4. 2002 poverená spolupracovníčka na príprave dovtedajšieho interného korpusu textov slovenského jazyka M. Šimková[4]. Postupne prijímaní noví pracovníci oddelenia SNK sa v procese tvorby koncepcie a zárodkov komplexu národného korpusu zároveň intenzívne vzdelávali v oblasti korpusovej lingvistiky a formalizácie prirodzeného jazyka. Na pôde SNK JÚĽŠ SAV sa v r. 2002 – 2004 konávali pravidelné semináre korpusovej a počítačovej lingvistiky, otvorené pre všetkých záujemcov o moderné metódy a nimi realizované lingvistické výskumy, na ktorých odznelo vyše 30 prednášok, z toho 20 zahraničných. Časť prednesených príspevkov je zhrnutá v publikácii *Insight into the Slovak and Czech Corpus Linguistics* (2006; zoznam všetkých vtedy uskutočnených prednášok a workshopov sa nachádza na s. 206 – 207). Poznatky a skúsenosti získavali jednotliví členovia kolektívu SNK aj na pracoviskách počítačovej, formálnej a korpusovej lingvistiky európskeho, ba i svetového významu v Brne (u K. Palu na Fakulte informatiky Masarykovej univerzity) a v Prahe (u J. Hajiča, E. Hajičovej a kol. v Ústave formálnej a aplikovanej lingvistiky Matematicko-fyzikálnej fakulty, u F. Čermáka v Ústave Českého národného korpusu a u V. Petkeviča v Ústave teoretickej a komputačnej lingvistiky Filozofickej fakulty Univerzity Karlovej). Viacerí pracovníci SNK sa zúčastnili aj na tzv. Mathesiovských seminároch v Prahe, kde prednášali poprední počítačoví a korpusoví lingvisti z celého sveta. Nové informácie a poznatky si nenechávali len pre seba, ale ich promptne sprostredkúvali slovenskej odbornej verejnosti v Jazykovednom časopise (Furdík – Šimková, 1998; Horák, 2003; Domin – Forróová – Garabík, 2003; Horák – Ološtiak – Ivanová – Gianitsová, 2004).

[4] Budovanie internej korpusovej a lexikálnej databázy v JÚĽŠ SAV koordinovali V. Benko a A. Jarošová. M. Šimková tu vykonávala viacero činností, ktoré mohla neskôr so znalosťou koordinovať v rámci oddelenia SNK: zisťovanie dostupnosti textov, ich získavanie od poskytovateľov (na tejto zložke sa nepravidelne podieľali aj ďalší pracovníci JÚĽŠ SAV), prvotné spracovanie textov vrátane tvorby konverzných tabuliek, vonkajšia anotácia textov, indexácia, testovanie dostupných softvérov, zaúčanie pracovníkov JÚĽŠ SAV do práce s počítačmi a s elektronickým korpusom, príprava dokladového materiálu podľa potrieb konkrétnych výskumov. V prvých fázach budovania pracovného kolektívu SNK a tvorby podrobnej celkovej koncepcie i čiastkových koncepcií jednotlivých podprojektov mali pre neskoršiu hlavnú riešiteľku projektu veľký význam poznatky nadobudnuté v novembri 2001 počas pobytu v Inštitúte nemeckého jazyka v Mannheime u Cyrila Belicu, ktorý jej nezištne poskytol svoje bohaté skúsenosti z vedenia korpusového projektu COSMAS (prehľadné zhrnutie je dostupné na WWW: `http://www1.ids-mannheim.de/kl/projekte/cosmas_I/gesamtkonzept.html`) a projektu kookurenčnej databázy (Belica, 2001). Korpusové technológie Ústavu nemeckého jazyka neskôr prezentoval C. Belica viacerým záujemcom na Slovensku v rámci seminárov SNK (27. 1. 2003). Významné sú i ďalšie jeho práce v oblasti korpusovej lingvistiky (napr. Belica – Steyer, 2005) a na projekte nového referenčného korpusu nemčiny (Kupietz – Belica – Keibel – Witt, 2010).

Od polovice r. 2003 sa oddelenie SNK pod vedením M. Šimkovej stalo aj riešiteľom úlohy *Komplexné spracovanie slovenského jazyka a jeho elektronizácia na účely jazykovedného výskumu* v rámci Štátneho programu výskumu a vývoja Aktuálne otázky rozvoja spoločnosti. Počas trvania tohto projektu sa pravidelne dvakrát ročne hodnotilo čiastkové plnenie stanovených úloh na základe veľmi podrobných priebežných správ, ktoré vypracúvala zodpovedná riešiteľka postupne aj s ďalšími členmi kolektívu, a následnej obhajoby dosiahnutých výsledkov pred komisiou Ministerstva školstva SR zloženou z domácich i zahraničných členov a troch zahraničných oponentov z odboru počítačovej a korpusovej lingvistiky. Pri predkladaní Záverečnej správy o riešení danej úlohy výskumu a vývoja (39 strán + 227 strán príloh) sa konštatovala vysoká úroveň a mimoriadna efektívnosť vrátane nízkych finančných nákladov pracoviska SNK, na ktorom sa za 4 roky vytvorili potrebné nástroje a vybudovali korpusy kvantitatívne aj kvalitatívne porovnateľné so zahraničím a od začiatku hojne využívané záujemcami o slovenský jazyk.[5] Kolektív oddelenia SNK v zložení M. Šimková, R. Garabík, H. Ivoríková, J. Levická, R. Hladík a T. Vančo získal v r. 2005 za dosiahnuté výsledky *Cenu Slovenskej akadémie vied za budovanie infraštruktúry pre vedu*.

Založenie oddelenia Slovenského národného korpusu JÚĽŠ SAV bolo výsledkom súhry, dozretia vonkajších aj vnútorných faktorov. K všeobecnejším a dlhšie trvajúcim vonkajším podmienkam v podobe technologizácie a informatizácie, rozvoja korpusov a korpusovej lingvistiky vo svete aktuálne pribudlo na konci predchádzajúceho tisícročia poznávanie možností jazykových technológií vrátane strojového prekladu a naliehavá potreba ich zabezpečenia pre slovenčinu v procese prístupových konaní Slovenskej republiky pred prijatím do Európskej únie. Vnútorné faktory sa v zásade týkali oblasti jazyka, jeho skúmania a opisu:

a) dlhoročná lexikografická tradícia JÚĽŠ SAV, pre napĺňanie ktorej bol nevyhnutný zodpovedajúci dokladový materiál na opis slovenského jazyka a v ktorej sa rozhodlo pokračovať aj začiatkom 90. rokov 20. storočia, keď sa pristúpilo k tvorbe koncepcie a následne ku koncipovaniu Slovníka súčasného slovenského jazyka;

b) dlhoročné zhromažďovanie jazykového materiálu v podobe manuálnej excerpcie a papierových kartoték v JÚĽŠ SAV na opis rôznych foriem slovenského jazyka, neskôr aj spomínaného interného korpusu textov súčasnej slovenčiny – bola tu teda značná skúsenosť s budovaním materiálovej bázy na výskum slovenčiny i aktuálna potreba jej vytvorenia zo súčasných textov, keďže excerpcia do všeobecnej papierovej kartotéky bola začiatkom 90.

[5] V oponentských posudkoch sa napr. uvádza: „Odborná úroveň řešení úkolu je na velmi dobré mezinárodní úrovni. SNK se kvalitou svého zpracování řadí k podobným projektům v Evropě a v některých ohledech je i předčí.“ (K. Pala); „Výslednou podobu SNK pokládám za srovnatelnou, co do velikosti i kvality, s korpusy českými (např. SYN2000, SYN2005, PUB2006) a anglickými ... Z užší oblasti slovanských jazyků nemá SNK kromě češtiny vůbec konkurenci...“ (K. Oliva); „Odborná způsobilost týmu řešitelky úlohy je mimo vší pochybnost. Přes těžkosti při získávání pracovníků s potřebnou a vhodnou kvalifikací, zejména pro technické zabezpečení úlohy, bylo řešení úlohy zcela v souladu se světovými postupy, standardy a časovými zvyklostmi. Lze říci, že pokud jde o časový plán řešení, tak tým SNK dokázal prakticky těchže výsledků dosáhnout v mnohem kratším čase, než obdobná pracoviště v zahraničí.“ (J. Hajič); „Celkově hodnotím dosavadní práci řešitelského kolektivu (stejně jako v letech 2003 a 2004) na předloženém projektu velmi vysoko, na skutečně špičkové mezinárodní úrovni...“ (J. Hajič)

rokov 20. st. z finančných dôvodov zastavená, dovtedy zhromaždený materiál už bol v podstate opísaný v predchádzajúcich slovníkoch a nedostatočne zabezpečený interný korpus nespĺňal požiadavky na rýchlejší rast objemu jazykových dát a kvalitu ich spracovania;

c) Koncepcia starostlivosti o štátny jazyk Slovenskej republiky schválená vládou SR v r. 2001, v rámci ktorej bolo stanovené za jednu z hlavných úloh v oblasti jazykovedy a jazykového výskumu zabezpečenie vyhovujúcich materiálových zdrojov, teda vybudovanie Národného korpusu slovenského jazyka.

Všetko vynaložené úsilie a všetky podniknuté kroky smerujúce k budovaniu Slovenského národného korpusu boli v súlade s tradíciou slovenskej lingvistiky, v ktorej sa výskum jazyka vo všetkých rovinách a formách opieral o jazykový materiál, hoci jeho zvyčajne individuálne zhromažďovanie (ručná excerpcia zameraná na konkrétny, práve riešený jednotlivý problém alebo jazykový jav) predstavovalo v predchádzajúcich desaťročiach časovo veľmi náročnú fázu vedecko-výskumnej práce. Relevantných dokladov nebol v zásade nikdy dostatok[6], skúmanie celkovej synchrónnej dynamiky jazyka v horizonte niekoľkých desaťročí a uplatnenie štatistických nástrojov na objektívne zhodnotenie fungovania jazykových prostriedkov z paradigmatického či syntagmatického hľadiska bolo bez existencie elektronického korpusu prakticky nemožné. Ustavičný dopyt po materiáli sa rozšíril aj na dopyt po všeobecne dostupných elektronických jazykových zdrojoch a zvyšoval sa najmä pod vplyvom tých lingvistov, ktorí pružne prechádzali na prácu s počítačom a zisťovali, aké možnosti na výskum jazyka sa pred nimi otvárajú. Rozsah a dynamiku tohto procesu sme naznačili vyššie (v bode a) na s. 38) a vystihuje ju o. i. aj záver recenzie zborníka príspevkov zo sympózia konaného v rámci 7. zasadnutia Lexikologicko-lexikografickej komisie pri Medzinárodnom komitéte slavistov v Nových Vozokanoch 24. – 26. apríla 1989: „Ak sa text príspevku J. Horeckého *Projekt bázy dát slovenského jazyka* vnímal v čase konania sa sympózia ako vzdialená a ešte iba v hrubých kontúrach formulovaná utópia, možno dnes konštatovať, že budovanie korpusu slovenského jazyka v Jazykovednom ústave Ľ. Štúra SAV (išlo o interný korpus; pozn. M. Š.) nadobudlo medzi časom už oveľa konkrétnejšiu podobu. Takisto sa už stalo realitou využívanie počítačov v rozličných štádiách prípravy slovníkov, na ktorých sa v súčasnosti v tejto jazykovednej inštitúcii pracuje. To je azda najvýrečnejšie svedectvo o aktuálnosti problematiky v recenzovanom zborníku“ (Buzássyová, 1992, s. 146).

Aj keď v zahraničí sa korpusové databázy často budovali a budujú v komerčných lexikograficky orientovaných spoločnostiach alebo sú súčasťou matematických či informatických pracovísk, ukotvenie Slovenského národného korpusu v Jazykovednom ústave Ľ. Štúra SAV nebolo náhodné, ale kontinuálne nadväzovalo na predchádzajúce práce v oblasti zhromažďovania materiálových zdrojov a bolo zároveň prirodzenou odpoveďou na aktuálne potreby slovenskej lingvistiky. Veľký záujem používateľov (vôbec nie bežný v porovnaní s okolitými krajinami, kde korpusové databázy spočiatku ostávali bez širšieho využitia lingvistickou obcou) sa potvrdzoval od samého začiatku: už interný korpus slovenských textov mal svoj stály okruh používateľov, ktorý sa akoby synergicky

[6] Kritici neraz vyčítali jazykovedcom, ako i ostatným spoločenským a humanitným vedám nedostatok exaktnosti a značnú mieru induktívnosti, introspekcie až subjektívnosti. Sprístupňovanie korpusov a nástup korpusovej lingvistiky rozprúdili v tomto smere nové diskusie aj medzi lingvistami navzájom (porov. ďalej).

rozširoval spolu s rozširovaním databáz SNK. Osobitne pozitívne a širokou verejnosťou mimoriadne vítané bolo postupné sprístupňovanie elektronických verzií kodifikačných príručiek a ďalších, i starších lexikografických opisov slovenčiny, ako aj inej knižnej a časopiseckej lingvistickej produkcie, s čím sa v SNK začalo už v r. 2003, keď takéto sprostredkúvanie výsledkov vedy a výskumu na Internete ešte nebolo štandardné ani na Slovensku, ani v iných krajinách (porov. časť 2.7).

Existencia elektronického korpusu ako špecifického materiálového zdroja na výskum a opis jazyka a korpusová lingvistika ako nový odbor sa zároveň aj na Slovensku museli vyrovnávať s viacerými otázkami týkajúcimi sa ich podstaty a metód, očakávaní a potrieb používateľov, ale aj istých obmedzení vyplývajúcich zo samotného jazyka, z možností počítačového spracovania jazykových javov i z možností počítačových technológií (rozsah a zloženie textov, typy a detailnosť anotácií, rýchlosť a kvalita vyhľadávacích softvérov a pod.).[7] S rozvojom a sprístupňovaním korpusových databáz pre slovenčinu a s rastúcim počtom používateľov sa i tu otázniky typu *načo korpus?*, *aký korpus?*, *komu korpus?* postupne nahrádzali stupňujúcimi sa požiadavkami na adekvátne, dostatočne rozmanité a lingvisticky predpripravené materiálové zabezpečenie aktuálneho jazykovedného výskumu, najmä lexikografického opisu súčasnej slovenčiny v novom výkladovom Slovníku súčasného slovenského jazyka, ako aj rozvíjajúcich sa potrieb počítačového spracovania slovenčiny v súlade s celosvetovým rozvojom počítačových a informačných technológií. Po ukončení prvej etapy projektu sa spolupráca Ministerstva školstva SR, Ministerstva kultúry SR a SAV na financovaní projektu *Budovanie Slovenského národného korpusu a elektronizácia jazykovedného výskumu na Slovensku* dohodla aj na druhú etapu (2007 – 2011). V súčasnosti pokračujú práce v SNK s finančnou podporou všetkých troch subjektov v tretej etape projektu (2012 – 2016).

2 Ciele Slovenského národného korpusu a ich napĺňanie

Príprava vyššie spomínaných prvých dvoch projektov na vybudovanie národného korpusu a na elektronizáciu jazykovedného výskumu na Slovensku prebiehala na prelome tisícročí a vychádzala z vtedajšieho stavu poznania v oblasti tvorby korpusov a korpusovej lingvistiky a z vtedajšieho stupňa rozvoja počítačových technológií. Jej súčasťou bol predpoklad, že počas piatich rokov je nevyhnutné vybudovať prakticky všetky elektronické zdroje pre slovenčinu aspoň v základnom rozsahu. Úlohy na r. 2002 – 2006 boli naplánované značne maximalisticky, avšak s minimalistickým personálnym obsadením

[7] Diskusie, kritiky, nedôvera, ba až istá nevraživosť časti klasických alebo inak orientovaných lingvistov voči korpusovým databázam boli vo všeobecnosti sprievodným znakom rozvoja korpusov a korpusovej lingvistiky vo svete s obsahom a priebehom charakteristickým pre danú dobu a dané jazykové, resp. vedecko-výskumné spoločenstvo (Searle, 1972; McEnery – Wilson, 2001), pričom názorové rozdiely a vyhranené postoje boli neraz podobné tým, ktorými prešla staršia generácia korpusových lingvistov v konfrontácii s postojmi N. Chomského (porov. Leech, 1991 – podľa prekladu 2000, s. 39). Niektoré očakávania a názory sa opakovane vracali, vytvorené mýty bolo treba opakovane vysvetľovať (Perkuhn – Belica, 2006; porov. aj Čermák, 2001, s. 125; Možnosti a meze..., 2006; Sokolová – Šimková – Ivanová, 2006; Cvrček – Kováříková, 2011).

(7 pracovných miest). Východiskovým modelom bol vtedy približne 10-členný Ústav Českého národného korpusu (ďalej ÚČNK; v súčasnosti ho tvorí vyše 40-členný pracovný kolektív vrátane pracovníkov v čiastkovom pomere a doktorandov), pričom už od jeho začiatkov sa na budovaní českých korpusových zdrojov podieľali viaceré bohemistické, počítačové a informatické pracoviská z celej ČR, ktorých pracovníci participovali na príprave rôznych typov korpusov a anotácií, resp. pre potreby českého korpusu vyvíjali jednotlivé softvérové nástroje: lematizátor, morfologický analyzátor, dezambiguátor, korpusový manažér a ďalšie. Pri tvorbe zdrojov a nástrojov SNK neboli žiadne iné pracoviská a všetky potrebné nástroje pre slovenčinu sa museli vyvinúť alebo implementovať v rámci SNK.

V projekte vybudovania Národného korpusu slovenského jazyka a elektronizácie jazykovedného výskumu v rokoch 2002 – 2006, ktorý prešiel náročným legislatívnym konaním, bolo stanovené ako hlavný cieľ zachytenie slovenského jazyka v čo najširšom rozsahu. Národný korpus mal predstavovať materiálové východisko na všestranný jazykovedný výskum (príprava akademickej gramatiky a lexikológie slovenského jazyka), no predovšetkým na lexikografický opis slovenčiny: na tvorbu viaczväzkového slovníka súčasnej slovenčiny, ortoepického slovníka, frekvenčného a retrográdneho slovníka, ale aj na prípravu nových vydaní už existujúcich praktických jazykových príručiek (Krátkeho slovníka slovenského jazyka, Pravidiel slovenského pravopisu, Synonymického slovníka slovenčiny). Uvažovalo sa aj o vytvorení súpisov mien a názvov, terminologickej banky a v spolupráci so zainteresovanými vládnymi orgánmi, najmä s Ministerstvom spravodlivosti SR a Ústavom pre aproximáciu práva, aj o vytvorení špecializovaného podkorpusu legislatívnych textov na podporu rozvoja terminologickej kultúry v oblasti právneho jazyka (porov. Jarošová, 2001b; materiál predložený na rokovanie vlády SR, dostupný na WWW: `http://www.rokovania.sk`). V schválenej etapizácii projektu sa predpokladalo v prvom rade vypracovanie podrobnejších koncepcií riešenia konkrétnych čiastkových úloh. Okrem všeobecného jednojazyčného korpusu písaných textov súčasnej slovenčiny sa v rámci neskôr schváleného Štátneho programu výskumu a vývoja mali v komplexe národného korpusu slovenského jazyka vytvoriť aj osobitné korpusy historických textov, prepisov nárečových a štandardných hovorených prejavov, korpusy paralelných textov, databáza lexikografických diel a terminologická databáza.

Rozsiahlosť stanovených úloh odrážala vyššie spomínané potreby slovenskej lingvistiky, počítačového spracovania prirodzeného jazyka (Natural Language Processing – NLP), ale aj ďalších záujemcov o slovenský jazyk. A hoci sa už bolo možné oprieť o viaceré poznatky a skúsenosti z počítačového spracovania jazykov (aj flektívnych, najmä češtiny a poľštiny), predsa každá čiastková úloha a každý typ korpusu či anotácie slovenských textov si vyžadovali vlastnú koncepčnú prípravu, testovanie prijatých postupov na vstupných vzorkách, dopĺňanie koncepcie, v niektorých prípadoch aj oponentské konania, priebežné kontroly a opravy.[8] V celom procese boli nezanedbateľné prínosy pracovných

[8] Napríklad manuál bibliografickej a štýlovo-žánrovej anotácie textov SNK (`http://korpus.sk/bibstyle.html`) mal vyše 20 verzií, kým sa ustálila jeho definitívna podoba, a aj v súčasnosti sa vyskytnú nové javy typu elektronické knihy, ktoré vedú k jeho nevyhnutným korekciám a doplneniam; podobne sa vytvárali a počas anotácie priebežne dolaďovali pravidlá morfologickej anotácie – slovenský tagset (porov. ďalej).

pobytov a konzultácií na vyššie uvedených pracoviskách v ČR, z ktorých (najmä z ÚČNK) pochádzali viaceré inšpirácie pri tvorbe základných manuálov na spracovanie a anotáciu textov v SNK nielen v podobe nadviazania na zistenia o správnosti zvoleného riešenia, ale aj upozornení na nevhodnosť niektorých postupov, resp. možnosť využitia novších a vhodnejších metód. Veľkým pomocníkom bolo aj ÚČNK vydané kompendium korpusovej lingvistiky obsahujúce zásadné štúdie zahraničných autorov, pre túto oblasť napr. S. Atkinsová – J. Clear – N. Ostler (2000), F. Čermák (2000). Prvá verzia lematizácie a morfologickej anotácie textov SNK bola realizovaná v r. 2004 na báze českého tagsetu a pomocou softvéru z Ústavu formálnej a aplikovanej lingvistiky MFF UK v Prahe. Postupne sa v SNK vypracúvali vlastné, značne odlišné pravidlá morfologickej anotácie v súlade so slovenskou gramatickou tradíciou a hoci sa v rámci spoločného projektu[9] uvažovalo o možnosti automatizovaného prekladu a plného využitia týchto nástrojov na automatizovanú anotáciu slovenského korpusu, čím by sa ušetrili značné náklady, táto cesta sa neukázala ako schodná (Hlaváčová, 2005). Bolo teda nevyhnutné uskutočniť v SNK vlastné ručné značkovanie trénovacích dát (s podporou automatického predznačkovania; porov. ďalej) a pristúpilo sa aj k tvorbe slovenskej morfologickej databázy a vlastného morfologického analyzátora (Garabík, 2005b). Na vyhľadávanie v korpusových dátach bol v prvej fáze zakúpený korpusový manažér Manatee s klientom Bonito z Fakulty informatiky v Brne (`http://www.textforge.cz/products`; autor P. Rychlý), zároveň sa však v SNK vyvíjalo vlastné webové rozhranie, ktoré sa naďalej používa na jednoduché vyhľadávania aj po uvoľnení klienta Bonito a jeho prechode na webovú aplikáciu medzi používateľmi známu ako Bonito 2.

V závere prvej etapy projektu budovania Národného korpusu slovenského jazyka a elektronizácie jazykovedného výskumu boli síce viaceré čiastkové ciele prekročené (napr. hlavný korpus bol o 100, resp. 150 mil. tokenov väčší oproti plánu), no ukázalo sa, že 7-členný pracovný kolektív naozaj nemôže splniť všetky úlohy, ktoré napr. v Českej republike riešilo vyše sto odborných a vedeckých pracovníkov na spomínaných štyroch špecializovaných pracoviskách. Niektoré úlohy sa počas prvej etapy čiastočne modifikovali, iné boli v plnom rozsahu presunuté do druhej etapy (napr. hovorený korpus) alebo v danom čase úplne vylúčené z úloh Slovenského národného korpusu – išlo o nárečový korpus a historický korpus, ktorých tvorba si vyžaduje okrem skúseností z korpusového spracovania dát aj účasť odborníkov z príslušných oblastí lingvistiky a ktoré boli opäť začlenené do plánov projektu SNK v 3. etape, keď dozreli potreby týchto špecializovaných korpusov i možnosti ich tvorby. Z pôvodne uvádzaných plánov na využitie Národného korpusu sa nemohlo naplniť ani to, aby slúžil ako virtuálna knižnica textov nespadajúcich pod autorský zákon, pretože legislatíva ostala v tomto smere nezmenená.

Náplňou projektu Budovanie Slovenského národného korpusu a elektronizácia jazykovedného výskumu na Slovensku v II. etape v r. 2007 – 2011 bolo predovšetkým ďalšie rozširovanie hlavného korpusu, vybudovanie základnej verzie Slovenského hovoreného korpusu a Slovenskej terminologickej databázy, príprava slovensko-českého

[9] Využitie spoločných vlastností češtiny a slovenčiny na budovanie anotovaných národných jazykových korpusov. Bilaterálna dohoda medzi MŠ SR a MŠMaT ČR č. 130/2003 (`http://korpus.sk/other_grants.html`).

paralelného korpusu, vývoj a skvalitnenie nástrojov na automatizované spracovanie slovenčiny. V III. etape (2012 – 2016) sa tieto úlohy rozšírili o ďalšie paralelné korpusy, nárečový korpus, historický korpus, webový korpus a korpusovolingvistické výstupy v podobe príručiek a špecializovaných slovníkov.

Od prvých predstáv o korpuse slovenských textov v rozsahu 20 miliónov slov, ktoré by boli presne percentuálne rozložené do 6 komponentov podľa komunikačných sfér (porov. Jarošová, 1993), cez základnú úlohu I. etapy projektu vytvoriť 200-miliónový korpus (porov. Šimková, 2003, 2004) sa obsah databáz SNK rozrástol v súčasnosti na takmer 3 miliardy tokenov, z ktorých sa jedna miliarda nachádza v aktuálne prístupnej verzii hlavného korpusu, druhá miliarda vo webovom korpuse, ďalšie stovky miliónov v paralelných korpusoch, databazách a v archíve či banke SNK (porov. ďalej). Projekt SNK sa stal komplexom, ktorý značne prekročil pôvodné zameranie takmer výlučne na lexikografické využitie, hoci príprava aktuálneho materiálu na lexikografické a gramatické opisy súčasnej slovenčiny stále patrí k jeho hlavným cieľom. V databázach Slovenského národného korpusu sa nachádza materiál z rôznych foriem slovenského jazyka a komunikačných sfér, v ktorých sa slovenčina uplatňuje, spolu s jazykovými informáciami na široké spektrum vedecko-výskumných (zďaleka nielen lingvistických), učebných i bežných laických využití.

2.1 Primárny korpus SNK

Na základe koncepcie Slovenského národného korpusu (Šimková, 2003, 2004) sa presnejšie vymedzilo budovanie všeobecného jednojazyčného korpusu písaných textov súčasného slovenského jazyka, ktorý bol neskôr označený ako vlastný, primárny korpus (verzie a podkorpusy série *prim*), a jednotlivých špecifických korpusov a databáz ako samostatných súčastí SNK (porov. ďalej). So zreteľom na potreby koncipovania nového výkladového slovníka mal primárny korpus pokrývať slovnú zásobu slovenčiny od polovice 20. st., ale v súvise s reformou slovenského pravopisu v r. 1953 a jej postupným zavádzaním do praxe bol za východiskový rok zaraďovania textov do nového korpusu určený rok 1955.[10] Takto rozsiahlo časovo stanovený záber nie je pre tvorcov korpusov v súčasnosti celkom štandardný, národné korpusy sú zvyčajne postavené na báze textov posledných dvoch až troch desaťročí, ktoré sú priamo dostupné v elektronickej podobe. V harmonograme prác SNK predstavovalo skenovanie, rozpoznávanie a rekonštrukcia textov, ktoré nikdy neexistovali alebo sa nezachovali v elektronickej podobe[11], množstvo

[10] Kombinácia textov v starom a novom pravopise by bola v tejto fáze spôsobovala značné problémy pri počítačovom spracovaní textov v korpuse.

[11] Na Slovensku sa s elektronickými verziami aktuálne vydávaných textov mohlo dať počítať približne od začiatku 90. rokov minulého storočia, no situácia bola v tejto oblasti značne komplikovaná a dynamická: menili sa editory a zalamovacie programy, takže bolo potrebné vytvárať množstvo konverzných skriptov a zohľadňovať rôzne špecifiká textov, vznikali a zanikali vydavateľstvá, archivovanie elektronických textov vôbec nebolo bežnou praxou a neuchovávali si

ručnej práce a nemalý časový, finančný aj manažérsky vklad, keď sa predovšetkým v druhej etape projektu dostávalo takýmto spôsobom do korpusu v priemere vyše 60 000 strán ročne. Na tejto zložke spracúvania textov do korpusu významne participovali mnohí externí spolupracovníci SNK, najmä študenti. Aj najnovšie texty, ktoré už bývajú zväčša priamo v elektronickej podobe, treba technicky čistiť (odstraňuje sa grafika, obrázky, tabuľky a pod.) do podoby čistého súvislého textu a konvertovať do jednotného formátu, čo neraz skomplikujú nové editovacie a zalamovacie programy. Text sa potom segmentuje na slová, textové jednotky, ku ktorým sa pridávajú lingvistické informácie.

Vnútornú štruktúru Slovenského národného korpusu tvoria úrovne A, B, C, D (porov. aj Garabík, 2004):

– v **A**RCHÍVE, ku ktorému majú prístup iba priamo zainteresovaní členovia oddelenia SNK, sa získané texty uchovávajú ako dokumenty v takom formáte, ako boli do SNK poskytnuté, so základnou informáciou o ich pôvode/zdroji, forme a obsahu; archív aktuálne zaberá približne 1,5 TB úložného priestoru;
– po odstránení znakov a symbolov editorov a programov, v ktorých texty vznikli, a po odstránení grafických súčastí sa texty prevedú do jednotného formátu, v ktorom sa zaznamenávajú štruktúrne vlastnosti textu; dokumenty tvoriace jeden text vo viacerých častiach sa spájajú do prirodzeného celku a naopak, ak dokument obsahuje texty s rôznymi charakteristikami (napr. kniha skladajúca sa z umeleckého textu a odborného predslovu alebo doslovu), v procese anotácie sú ručne rozdelené; ku každému textu sa následne doplní vonkajšia anotácia (bibliografické a štýlovo-žánrové údaje podľa príslušného manuálu SNK, porov. `http://korpus.sk/bibstyle.html`) a táto podoba korpusu tvorí **B**ANKU;
– v ďalšej fáze sa text rozsegmentovaný na základné jednotky (slová, interpunkcia, značky, symboly) lingvisticky značkuje: ku každému slovu – tokenu sa pridajú príslušné jazykové informácie (základný tvar slova – lema, slovný druh, morfologické kategórie tvaru slova v danom kontexte a pod. podľa pravidiel morfologickej anotácie SNK, porov. `http://korpus.sk/morpho.html`) a vytvorí sa **C**ORPUSOID;

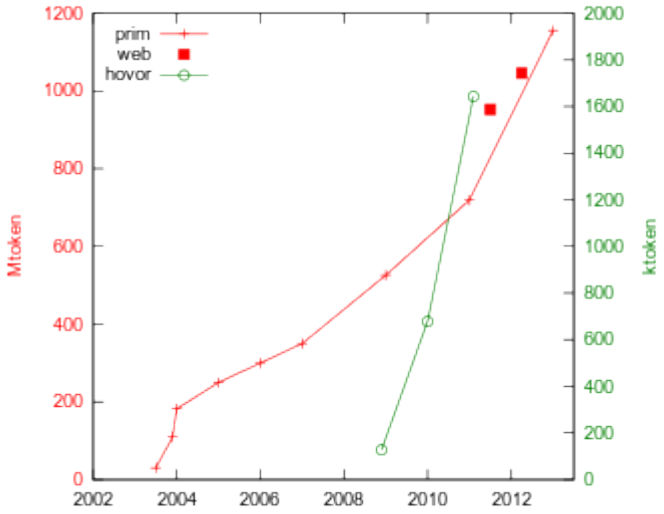
ich ani samotní autori. Napriek snahe o efektívnosť a čo najmenšie využívanie manuálnej práce bolo neraz potrebné naskenovať a v plnom rozsahu ručne spracovať aj novšie vydaný text, ak to bolo pre obsah a používateľov korpusu dôležité. S posunom hornej časovej hranice trvania projektu SNK v ďalších etapách sa posúvala aj hranica, od ktorej sa dalo očakávať získanie elektronických verzií textov: v druhej etape to už bol prelom tisícročí, v tretej je to zhruba posledných 5 rokov. Archív SNK sa neraz stal zdrojom elektronickej verzie konkrétneho textu aj pre jeho autora či majiteľa autorských práv, ak mal záujem svoj text opätovne vydať, no jeho elektronická podoba nebola uchovaná nikde inde. Na druhej strane sú ešte stále aj takí autori, ktorí majú svoj text uložený dlhé roky a prekvapia klasickými disketami so súbormi vo formáte T602, ktorý je pre spracovanie v korpuse často jednoduchší ako moderné, problematicky konvertovateľné formáty.

– takto spracované texty, ktoré majú od poskytovateľov textov licenciu na verejné využívanie, sa ako **DÁTA** sprístupňujú na vyhľadávanie jazykových prostriedkov a ich zobrazenie v presne vymedzenom krátkom kontexte na internete (`http://korpus.sk`) všetkým záujemcom o slovenský jazyk a bádateľom v oblasti korpusovej lingvistiky, ktorí súhlasia s podmienkami nekomerčného používania Slovenského národného korpusu.

Systematickému budovaniu korpusu v podobe opísaného spracovania textov predchádza systematický zber textových dát najrôznejších štýlov, žánrov, autorských, generačných či vydavateľských úzov zo všetkých regiónov Slovenska a sčasti aj od Slovákov žijúcich v zahraničí v tradičných i novovznikajúcich enklávach. Od začiatku existencie Slovenského národného korpusu bolo oslovených takmer 2 000 potenciálnych poskytovateľov textov, licenčné zmluvy na použitie ich diel v databázach SNK v súlade s autorským zákonom sa podarilo uzavrieť približne s jednou tretinou z nich. Niekoľko autorských dráh sa medzičasom, žiaľ, ukončilo (L. Ballek, J. Lenčo, M. Rúfus, L. Ťažký a i.), o to vzácnejšia je prítomnosť ich textov v SNK. Zoznam zazmluvnených poskytovateľov s poďakovaním sa pravidelne aktualizuje na stránke `http://korpus.sk/contributors.html`, úplný zoznam bibliografií textov tvoriacich príslušnú verziu korpusu sa nachádza vždy pri informáciách o danej verzii. Viaceré vydavateľstvá, autori a prekladatelia sú dlhoročnými prispievateľmi textov do SNK, kontakty s nimi sa priebežne obnovujú a na aktuálne vydané diela sa podpisujú nové licenčné zmluvy. Táto fáza budovania korpusu predstavuje časovo a komunikačne veľmi náročnú zložku, každý získaný text však prispieva k objektívnejšiemu poznaniu reálnej podoby a fungovania súčasnej slovenčiny.

Hlavný korpus v každej sprístupnenej verzii zahŕňa texty napísané a/alebo publikované v slovenskom jazyku od r. 1955 po rok, v ktorom bola zverejnená nová verzia. Princípom výstavby tohto i ostatných korpusov celku SNK je navrstvovanie textov na seba, teda nová verzia nepredstavuje separátnu, neprienikovú množinu textov, ale vždy ide o kvantitatívne rozšírenie predchádzajúcej verzie a v niektorých prípadoch aj o kvalitatívne zlepšenie spracovania (segmentácia, tokenizácia, lematizácia, anotácia). Staršie verzie sú uchované v archíve a používatelia ich na požiadanie môžu mať k dispozícii podľa potreby svojho výskumu ľubovoľne dlhý čas. Na otázku časového rozostupu medzi zhromažďovaním textov a sprístupňovaním nových korpusov nebola ťažká odpoveď: používatelia potrebovali čo najskôr relevantné množstvo jazykového materiálu a v každej chvíli uvítajú elektronicky spracovanú čo najaktuálnejšiu slovnú zásobu. Od začiatku existencie korpusu takto dostala verejnosť k dispozícii na online vyhľadávanie jazykových javov už deviatu verziu SNK, hoci jej číslo je v r. 2012 prim-6.0. Rozdiel je spôsobený tým, že prvé, testovacie verzie mali poradové čísla začínajúce sa 0 (prim0.1, prim0.2 – obidve boli sprístupnené v priebehu pol roka, a to leto – zima 2003) a vo verzii prim-2.0 došlo k viacerým opravám v segmentácii a tokenizácii textov, takže nasledujúca verzia dostala číslo 2.1. V r. 2004 – 2007 bol každý rok sprístupnený nový korpus, od r. 2009 sa prešlo na dvojročný interval.

Priebeh sprístupňovania jednotlivých verzií hlavného korpusu a ich rozsahy sú zobrazené v nasledujúcom grafe spolu s ďalšími korpusmi SNK.



Graf 1. Verzie hlavného korpusu SNK a slovenského webového korpusu v miliónoch tokenov a Slovenského hovoreného korpusu v tisícoch tokenov ku koncu r. 2012

Od verzie prim-2.1 (2006), ktorá už mala dostatočne veľký rozsah na to, aby sa z nej dali vytvoriť zmysluplné menšie celky, sa na špecifické analýzy a použitia vytvárali samostatné podkorpusy podľa hlavných štýlov: odborné texty (prf), publicistické texty (inf), umelecké texty (img). Tzv. vyvážený korpus (vyv) obsahoval texty z týchto troch hlavných štýlov v rovnakom tretinovom zastúpení. Pre potreby koncipovania Slovníka súčasného slovenského jazyka sa postupne tvorili samostatné podkorpusy textov, ktorých pôvodný jazyk bol slovenský: prim-*-public-sk, prim-*-public-img-sk. Všetky zverejňované zdroje (public) sú dostupné na vyhľadávanie jazykových javov na základe súhlasu autorov a/alebo majiteľov autorských práv vyjadreného v licenčnej zmluve. Texty, na ktoré nie je možné získať súhlas ani výnimku z autorského zákona (napr. zaniknutý časopis Kultúrny život, ktorý vychádzal v r. 1946 – 1968 a 1990 – 1993, resp. aj 2000 – 2002), sú zahrnuté iba do interne prístupného korpusu (juls, do verzie prim-5.0 aj snk). Najväčší, východiskový korpus (all) obsahuje všetky texty SNK spracované v čase jeho finalizácie na úrovni C (corpusoid; porov. vyššie). Z neho sa selektujú niektoré texty s osobitnými vlastnosťami (bez diakritiky, lingvistické texty, texty zahraničných Slovákov) a vzniká tzv. očistený korpus (sane), z ktorého sa tvoria všetky ostatné podkorpusy. Osobitnými súčasťami hlavného korpusu sú texty spred roka 1989 (r55az89) a ručne morfologicky značkovaný korpus (r-mak), na ktorom sa trénuje automatizovaná anotácia a lematizácia pomocou tagera Morče vyvinutého v Ústave formálnej a aplikovanej lingvistiky MFF UK v Prahe (`http://ufal.mff.cuni.cz/morce/index.php`) a natrénovaného na slovenskú sústavu značiek. Vyhľadávanie v štýlovo, žánrovo, jazykovo, časovo či inak vymedzených textoch sa dá realizovať aj pomocou nastavenia podmienok vo vyhľadávacom rozhraní, ale takto predpripravené podkorpusy môžu používateľom viaceré kroky zjednodušiť.

Štruktúru hlavného korpusu s rozsahmi vo verzii prim-6.0 znázorňuje nasledujúca tabuľka.

	prim-6.0-public-	*prim-6.0-juls-*
-all	1 156	1 257
-sane	1 121	1 196
-vyv	313	
-sk	905	
-inf	889	
-prf	106	
-img	114	
-img-sk	35	
r55az89-3.0	63	
r-mak-4.0	1.2	

Tabuľka 1. Podkorpusy hlavného korpusu prim-6.0 a ich rozsahy v miliónoch tokenov

V koncepcii SNK sa počítalo a už v r. 2003 sa aj začalo s prípravou vlastného morfologického tagsetu (súboru značiek a pravidiel anotácie) pre prvú, morfologickú rovinu internej lingvistickej anotácie (Forróová – Garabík – Gianitsová – Horák – Šimková, 2003; Garabík – Gianitsová – Horák – Šimková, 2004). Jej priebeh, postupné riešenia vznikajúcich problémov a dosahované výsledky sú opísané vo viacerých štúdiách (napr. Garabík – Gianitsová-Ološtiaková, 2005; Karčová – Šimková, 2006; Garabík – Šimková, 2012a; Garabík – Šimková, 2012b). Ručne morfologicky anotované podkorpusy r-mak boli postupne zverejňované spolu s novými verziami všeobecného, základného korpusu prim od r. 2006, pričom sa v ročných časových intervaloch približne zdvojnásobovali: od prvého rozsahu vyše 300 tisíc textových jednotiek v r-mak-1.0 cez 512 tisíc textových jednotiek v r-mak-2.0 až po vyše 1 200 000 textových jednotiek v r-mak-3.0, resp. necelých 1,2 mil. v opravenej verzii r-mak-4.0.

Od r. 2005 sa v SNK zároveň buduje morfologická databáza – slovník paradigiem slovenských slov, pomocou ktorej sa zlepšuje kvalita automatizovanej anotácie (Garabík, 2005b, 2007; Karčová, 2008). Východiskom jej tvorby bol model morfologickej databázy slovenčiny (Benko – Hašanová – Kostolanský, 2004) zakúpený od jej autorov. V súčasnosti obsahuje morfologický slovník SNK plné paradigmy všetkých ohybných slov z Krátkeho slovníka slovenského jazyka (2003), najfrekventovanejších slov z korpusu, frekventovaných skratiek, značiek a vlastných mien: spracovaných je približne 97 tisíc slov s 3,3 mil. tvarov v úplných paradigmách.

Vzhľadom na flektívnosť slovenského jazyka a bohatú tvarovú homonymiu nebude automatizovaná anotácia slov nikdy absolútne bezchybná, aj v iných jazykoch sa jej úspešnosť štandardne pohybuje na úrovni 94 – 96 %. Používatelia sa s tým učia počítať a pracovať. Úspešnosť automatizovaných procedúr anotácie a dezambiguácie sťažujú pribúdajúce nové slová, najmä vlastné mená a názvy, ale aj preklepy, rôzne značky, kombinácie písmen a číslic, slová z jazykov s inými znakovými sústavami a pod. Výhoda korpusu v podobe zhromaždenia veľkého množstva reálnych textov je v takýchto prípadoch komplikáciou, ktorá sa nedá vždy vyriešiť jednoduchou selekciou nežiaducich javov bez toho, aby sa neodstránili aj väčšie celky textov, ktoré spĺňajú požiadavky na bezchybný slovenský text.

Na vyhľadávanie textových jednotiek a jazykových informácií v korpusoch SNK sa od začiatku používal korpusový manažér Manatee s klientom Bonito, zakúpený v r. 2003 z Fakulty informatiky MU v Brne, a po jeho uvoľnení, doplnení ďalších funkcií a prechode na webové rozhranie sa používa aj v súčasnosti. Zároveň sa v prvej etape fungovania SNK vyvinulo a takisto dodnes používa vlastné webové rozhranie na jednoduché vyhľadávanie v dvoch základných korpusoch (prim-*-public-all, r-mak) pre neregistrovaných používateľov.

2.2 Slovenský webový korpus

Súčasťou komplexu SNK v súbore písaných textov je od r. 2011 tzv. webový korpus (`http://korpus.sk/web.html`), ktorý obsahuje korpusovo spracované slovenské texty z dostupných internetových zdrojov. Prvá verzia tohto korpusu bola pripravená v spolupráci s pracovníkmi Fakulty informatiky Masarykovej univerzity v Brne, kde sa budujú webové korpusy rôznych jazykov. Poskytnuté dáta boli v SNK ďalej spracované nástrojmi a postupmi štandardne používanými na písané texty (lematizácia, morfologická anotácia SNK) a do druhej verzie (2012; porov. Graf 1 v predchádzajúcej časti) boli doplnené texty získané z internetových zdrojov v rámci SNK. Napriek filtrácii sa v týchto dátach nachádzajú v istom rozsahu aj cudzojazyčné, najmä české texty.

2.3 Paralelné korpusy SNK

Paralelné korpusy vo všeobecnosti obsahujú rovnaké texty v dvoch alebo viacerých jazykoch, obvykle so zarovnaním (spárovaním) ekvivalentných častí textu, najčastejšie viet, prípadne jednotlivých slov alebo naopak, celých odsekov. Slúžia na porovnávacie lingvistické či translatologické výskumy, na tvorbu prekladových slovníkov, nástrojov automatizovaného prekladu i na výučbu niektorého z jazykov ako cudzieho.

V rámci SNK majú používatelia aktuálne k dispozícii niekoľko druhov paralelných korpusov, ktoré sa začali budovať od r. 2005. V tých, ktoré vznikali neskôr, sa uplatňuje novšia verzia webového rozhrania, systému anotácie a konverzie. Prvé paralelné korpusy SNK sa v procese rozširovania postupne konvertujú do nového systému. Slovenské

paralelné korpusy majú viacero podobných čŕt z hľadiska návrhu, spracovania, ako aj prístupov poskytnutých používateľom. Medzi spoločné vlastnosti patrí pomenovanie jednotlivých paralelných korpusov[12], ich anotácia a využiteľnosť, výber textov, štruktúra korpusu, spôsob zarovnania (párovania), použité počítačové nástroje.

Pri výbere textov do paralelných korpusov uprednostňujeme priame, vzájomné preklady medzi slovenčinou a druhým jazykom (v oboch smeroch), no nevylučujeme z týchto korpusov ani preklady z tretieho jazyka, a to ani v prípade, že korpus nemá nedostatok priamo preložených textov. Prítomnosť prekladov cez tretí jazyk prináša do korpusu žiadanú variabilitu jazyka – niektoré jazykové prostriedky sa inak používajú v originálnych textoch a inak v preložených.

Paralelné korpusy v SNK sú postavené na backende systému Manatee, ktorý umožňuje vyhľadávanie v jednotlivých častiach (slovenskej a druhého jazyka). Nad týmto systémom sme vybudovali aplikačnú vrstvu, ktorá sprístupňuje prepojenie medzi obidvomi časťami korpusu na základe odkazov medzi jednotlivými vetami. Nad touto vrstvou sa nachádza webové rozhranie pre používateľov napísané v programovacom jazyku Python (verzia 2.*) s využitím webového frameworku Karrigell (`http://karrigell.sourceforge.net/`).

Každý paralelný korpus v SNK predstavuje samostatnú položku a skladá sa z dvoch častí textov – nie sú to rozsahom rovnaké polovice, pretože zúčastnené jazyky predstavujú odlišné systémy s odlišnými vyjadrovacími sústavami. Jednu časť v každom paralelnom korpuse tvoria slovenské texty, druhú tie isté texty v príslušnom inom jazyku. Každá z častí je vnútorne rozdelená na dokumenty, kde jeden dokument zodpovedá jednej vstupnej jednotke textu, spravidla ide o knihu, niekedy aj menšie časti. Každý dokument má svoju bibliografickú a štýlovo-žánrovú anotáciu, ktorá je úspornejšia ako rovnaká anotácia textov v hlavnom korpuse, ale okrem štandardných položiek obligatórne obsahuje aj informáciu o originálnom jazyku, v ktorom bol text pôvodne napísaný, a meno prekladateľa/prekladateľov. Text sa ďalej delí na vety, z ktorých každá má svoj identifikátor vo forme prirodzeného čísla, monotónne sa zvyšujúceho od začiatku celého korpusu bez ohľadu na rozdelenie na dokumenty, a odkaz na zodpovedajúcu vetu alebo vety v druhej časti korpusu. V prípade, ak je veta spárovaná iba s jednou vetou, je tento odkaz tvorený identifikátorom párovej vety. Ak jednej vete v texte jedného jazyka zodpovedá v texte druhého jazyka viacero viet, odkaz je tvorený zoznamom identifikátorov spojených znakom U+002B PLUS SIGN. Ak veta nemá ekvivalent, tak je odkaz prázdny.

V slovenských textoch paralelných korpusov využívame štandardnú lematizáciu a anotáciu SNK, texty druhých jazykov sú v rámci možností lematizované a anotované pomocou dostupných nástrojov (bližšie porov. `http://korpus.sk/par.html`).

Jednotlivé paralelné korpusy sprístupnené v celku SNK sa vyznačujú aj istými špecifickými vlastnosťami, ktoré vyplývajú z dostupnosti textov, osobitostí druhého jazyka a pod.

[12]Poradie jazykov v názvoch paralelných korpusov SNK (napr. slovensko-latinský paralelný korpus), kde slovenčinu uvádzame na prvom mieste, je konvencia a neodráža vždy, niekedy takmer vôbec, smer prekladu konkrétneho textu.

	rok sprístupnenia	počet viet – slovenčina [tis.]	počet viet – druhý jazyk [tis.]
en (beletria)	2012	4 378	4 310
en (voľný)	2012	10 380	10 372
cs (beletria)	2011	741	742
cs (voľný)	2011	5 592	5 695
ru	2005	178	176
fr	2006	14	16
la	2012	33	28
bg	2012	21	20

Tabuľka 2. Paralelné korpusy SNK a ich rozsahy v tisíckach viet (stav ku koncu r. 2012)

Slovensko-anglický paralelný korpus

Najväčší z paralelných korpusov SNK je slovensko-anglický paralelný korpus, ktorý sa skladá z dvoch podkorpusov: z podkorpusu beletrie a z podkorpusu voľne dostupných textov – takých, ktorých licenčné podmienky umožňujú redistribúciu. Podkorpus beletrie obsahuje takmer 570 dokumentov (kníh) v rozsahu zhruba 4 milióny párov viet – 63 miliónov tokenov v anglickej časti a 54 miliónov tokenov v slovenskej časti. Paralelný slovensko-anglický korpus vznikol v SNK v rámci projektu 7. rámcového programu *EuroMatrixPlus – Bringing Machine Translation for European Languages to the User* (`http://www.euromatrixplus.net`; porov. aj v časti 3.3), ktorý mal za cieľ rozšírenie a skvalitnenie strojového prekladu medzi európskymi jazykmi. Tomu zodpovedalo aj hlavné (avšak nie jediné) zameranie tohto korpusu.

Slovensko-český paralelný korpus

Druhý najväčší paralelný korpus SNK je slovensko-český paralelný korpus, ktorý sa takisto skladá z podkorpusu beletrie a z podkorpusu voľne dostupných textov. Podkorpus beletrie obsahuje 147 dokumentov (kníh) v rozsahu cca 740 tisíc párov viet – okolo 10 miliónov tokenov pre každý z oboch jazykov. Korpus vznikol v spolupráci s Ústavom Českého národného korpusu Filozofickej fakulty Univerzity Karlovej v Prahe (vzájomná výmena textov) a neskôr bol rozšírený a upravený v rámci spomínaného projektu EuroMatrixPlus.

Slovensko-ruský paralelný korpus

Slovensko-ruský paralelný korpus je prvým z paralelných korpusov SNK. Vznikol v r. 2005 v spolupráci s Katedrou matematickej lingvistiky Filologickej fakulty Petrohradskej štátnej univerzity v Petrohrade. Obsahuje zhruba 178 tisíc dvojíc viet a je zložený prevažne z beletrie s menšou zložkou literatúry faktu. Na jeho rozšírení a aktualizácii spracovania i možností vyhľadávania sa priebežne pracuje.

Slovensko-francúzsky paralelný korpus

Slovensko-francúzsky paralelný korpus vznikol v SNK ako druhý paralelný korpus v poradí. Obsahuje niekoľko diel z beletrie francúzskych autorov a ako prvý z paralelných korpusov SNK bol rozšírený o voľne dostupné preklady Európskej únie. Vytvorenie tohto korpusu bolo ovplyvnené potrebami konkrétneho porovnávacieho výskumu, na jeho vývoji sa momentálne nepracuje.

Slovensko-latinský paralelný korpus

Slovensko-latinský paralelný korpus sa svojím zameraním, anotáciou a v menšej miere aj spracovaním odlišuje od ostatných doteraz spomínaných paralelných korpusov SNK: a) nie je určený pre používateľov ovládajúcich latinčinu a študujúcich slovenčinu ako cudzí jazyk – predpokladáme, že takíto záujemcovia sa k slovenčine dostanú skôr prostredníctvom iného jazyka ako latinčiny; b) obsahuje výlučne preklady z latinčiny do slovenčiny; c) nie je zameraný na moderný (alebo modernizovaný) jazyk, ale na texty v klasickej alebo stredovekej latinčine, o ktoré zrejme môžu mať používatelia tohto paralelného korpusu hlavný záujem; d) je tu značné množstvo textov napísaných v inom ako rodnom jazyku (toto obzvlášť platí v prípade stredovekej a modernej latinčiny).

Na pomenovanie anotačných kľúčov sme v prípade slovensko-latinského paralelného korpusu zvolili nie angličtinu, ale práve latinčinu vzhľadom na vznešenosť a vysoký status tohto jazyka. Podobne sú v latinčine aj hodnoty anotácie s výnimkou mien autorov a názvov slovenskojazyčných diel.

Slovensko-bulharský paralelný korpus

Slovensko-bulharský paralelný korpus vzniká v SNK v spolupráci s Inštitútom matematiky Bulharskej akadémie vied v rámci projektu medziakademickej dohody *Electronic Corpora – Contrastive Study with Focus on Design of Bulgarian-Slovak Digital Language Resources (continuation of the project)*. V súčasnosti obsahuje iba malé množstvo textov, aktuálne sa však pracuje na jeho rozširovaní a skvalitňovaní.

2.4 Slovenská terminologická databáza

Relatívne samostatnou zložkou Slovenského národného korpusu, nie celkom štandardnou v porovnaní s inými korpusovolingvistickými pracoviskami, je Slovenská terminologická databáza (`http://data.juls.savba.sk/std/`), ktorá v súčasnosti obsahuje približne 5 000 terminologických záznamov zo 16 oblastí (napr. bezpečnostnoprávna oblasť, história, sociálna práca, šach). Jej cieľom je sumarizácia a štandardizácia terminológie na Slovensku najmä v súvislosti s osobitnými potrebami harmonizácie právnej a ekonomickej terminológie s terminológiou Európskej únie (podrobnejšie Levická, 2007, 2008). Existencia STD na korpusovom pracovisku je výhodná vzhľadom na to, že korpus môže pomôcť a pomáha pri budovaní terminologickej databázy ako zdroj textov, ktoré môžu tvoriť samostatné podkorpusy z akejkoľvek oblasti. V celku SNK je to napr. korpus právnych textov vytvorený v r. 2011 v spolupráci s Ministerstvom spravodlivosti SR (`http://korpus.sk/legal.html`) a pripravujú sa ďalšie špecializované korpusy. Z nich

sa dajú termíny čiastočne automatizovane extrahovať, frekvenčne a distribučne zhodnotiť a po manuálnom overení a prípadnej modifikácii sa z dostupných textov v korpuse selektujú jednoznačné kontexty dokumentujúce definíciu, ktoré sa zapracujú do štruktúry terminologického záznamu obsahujúceho aj prípadné cudzojazyčné ekvivalenty.

Vedúca projektu Slovenskej terminologickej databázy J. Levická viedla v r. 2009 – 2011 v spolupráci s Ekonomickou univerzitou v Bratislave samostatný projekt VEGA *Spracovanie obchodnovednej terminológie pre potreby Slovenskej terminologickej databázy* s dôrazom na analýzu terminologických neologizmov, ktorého výsledkom bol aj súbor štúdií *Neologizmy v terminológii marketingu* (2010) a *Malý lexikón marketingu* (2012).

2.5 Slovenský hovorený korpus

Projekt Slovenského hovoreného korpusu sa začal realizovať ako súčasť Slovenského národného korpusu v r. 2008, všeobecné princípy jeho tvorby a jednotlivé zásady prepisu sa pripravovali od r. 2007 tak, aby Slovenský hovorený korpus vyhovoval rôznym výskumom, predovšetkým však bežným lingvistickým využitiam v JÚĽŠ SAV i na ďalších slovakistických pracoviskách.

Zvolený dvojúrovňový princíp prepisu získaných zvukových záznamov (základný textový, v zásade ortografický prepis a základný ortoepický prepis so zachytením vybraných výslovnostných a rečových javov) predstavuje pragmatické a počítačovo dobre spracovateľné riešenie, ktoré bolo v tom čase jedným z nadštandardných riešení a spolu s dostupnosťou zvukového záznamu prepojeného s prepisom poskytlo možnosť objektívneho výskumu súčasnej hovorenej slovenčiny.

Podrobnosti prepisu, sprístupňovania, zloženia a verzií Slovenského hovoreného korpusu, ako aj možnosti jeho používania sú opísané v nasledujúcom samostatnom príspevku v tejto publikácii (Gajdošová – Šimková, s. 65 – 84).

2.6 Historický korpus slovenčiny

Všetky vyššie opísané korpusy tvoriace celok SNK sú koncipované ako synchrónne – zachytávajúce súčasný jazyk konkrétneho, v danom prípade nedávneho časového obdobia. Historický korpus slovenčiny je koncipovaný ako diachrónny, spracúvajúci slovenský jazyk a jeho vývin z hľadiska širšieho časového obdobia. Cieľom tohto korpusu je obsiahnuť predspisovné obdobie slovenčiny, t. j. jazyk od 15. do 19. storočia. Vzhľadom na veľmi špecifické vlastnosti historických textov nie je možné v diachrónnom korpuse aplikovať bežné metódy lingvistického spracovania slovenčiny, ako je napríklad lematizácia a morfologická analýza. Prostredníctvom historického korpusu sa používateľom sprístupňuje jazyk v pôvodnej podobe, preto je jeho anotácia orientovaná najmä na normalizáciu tvarov slov a informatívne metadáta. Prvá verzia Historického korpusu slovenčiny, obsahujúca texty z publikácie *Pramene k dejinám slovenčiny*, bola sprístupnená

koncom roka 2012 v rozsahu 371 tisíc tokenov (bližšie porov. Garabík – Kajanová, 2013). Priebežne sa v SNK pracuje na prepise a anotácii ďalších historických dokumentov, ktoré budú postupne zaraďované do aktualizovaných verzií korpusu.

2.7 Lingvistické zdroje

Tvorba databázy lexikografických diel rozšírenej o ďalšie lingvistické zdroje nepatrila medzi hlavné úlohy riešiteľského kolektívu projektu SNK, no u záujemcov o jazyk a jazykové zdroje vzbudila najväčší záujem. Vytvorenie slovníkového rozhrania v r. 2003, prostredníctvom ktorého bol verejnosti ako prvý sprístupnený Krátky slovník slovenského jazyka, a jeho postupné dopĺňanie o ďalšie v SNK digitalizované i pôvodné elektronické lexikografické zdroje JÚĽŠ SAV prinášalo naozaj široké využívanie výsledkov elektronizácie jazykovedného výskumu a oceňovanie práce JÚĽŠ SAV napr. aj v podobe pochvalných vyjadrení mnohých používateľov na internete (blogy) i na iných miestach a zaraďovania slovenských slovníkov medzi obľúbené linky. Elektronické lingvistické zdroje sprístupňované na stránkach `http://slovniky.korpus.sk`, resp. `http://slovniky.juls.savba.sk/` obsahujú v súčasnosti pomerne rozsiahly súbor aktuálnej i staršej lingvistickej produkcie, o ktorú má slovenská i zahraničná verejnosť stále veľký záujem. Slovníky a databázy zverené SNK na ďalšie spracovanie a sprístupnenie v slovníkovom rozhraní už v elektronických formátoch (Krátky slovník slovenského jazyka, Pravidlá slovenského pravopisu, Synonymický slovník slovenčiny, Databáza priezvisk na Slovensku a i.), ktoré vznikali priamo v riešiteľských kolektívoch alebo sa pripravovali v oddelení spracúvania lingvistických dát JÚĽŠ SAV (porov. Benko, 2001), bolo potrebné konvertovať do požadovanej podoby a aj v súčasnosti prebiehajú ich priebežné opravy na základe zistených drobných textových chýb. Mnoho knižných a časopiseckých diel sa však postupne digitalizovalo, následne opravovalo a spracúvalo v rámci projektu SNK, pričom celý proces si vyžadoval premyslenú logistiku, administráciu a v neposlednom rade aj finančnú investíciu. V špecifických prípadoch bolo treba hľadať riešenia netriviálnych problémov nielen na úrovni počítačového spracovania, ale aj z lingvistického hľadiska (porov. Garabík – Kajanová, 2012). Do lingvistických zdrojov sa okrem spomínaných slovníkov postupne dopĺňali napríklad tieto publikácie a celé vydania časopisov:

- A. Bernolák: Slowár Slowenskí Češko-Laťinsko-Ňemecko-Uherskí, 1825
- Ľ. Štúr: Nauka reči Slovenskej, 1846
- Ľ. Štúr: Nárečja Slovenskuo alebo Potreba písaňja v tomto nárečí, 1846
- S. Czambel: Rukoväť spisovnej reči slovenskej, 1902
- Pravidlá slovenského pravopisu s abecedným pravopisným slovníkom, 1931
- Pravidlá slovenského pravopisu s pravopisným slovníkom, 1940
- Morfológia slovenského jazyka, 1966
- Dynamika slovnej zásoby súčasnej slovenčiny, 1989
- Pramene k dejinám slovenčiny 1, 1992

- Jazykovedný časopis, od r. 1954
- Kultúra slova, od r. 1967
- Československý terminologický časopis, 1962 – 1966
- Slovenská reč, od r. 1932
- Sociolinguistica Slovaca 1
- VARIA I – VIII
- Slovenskí jazykovedci, 1925 – 1975

Aj vďaka sprístupneniu uvedenej lingvistickej produkcie JÚĽŠ SAV a ďalších relevantných titulov na internete sú výsledky slovenskej lingvistiky dostupné každému používateľovi v akomkoľvek čase, čo oceňujú nielen slovakisti v zahraničí, ale aj lingvisti a iní záujemcovia na území Slovenska.

3 Zapojenie SNK do medzinárodných projektov

3.1 **Mondilex** (`http://www.mondilex.org/`)

Cieľom projektu 7. rámcového programu *Conceptual Modelling of Networking of Centres for High-Quality Research in Slavic Lexicography and Their Digital Resources* (2008 – 2010) bolo preskúmať možnosti vytvorenia konceptuálnej schémy výskumnej infraštruktúry podporujúcej centrá vysokokvalitného výskumu v slovanskej lexikografii, podporiť ich vedeckú kapacitu, integrovať ich digitálne zdroje a otvoriť ich pre spoluprácu s európskou akademickou komunitou. Projekt poskytol stratégie na koordináciu, unifikáciu a rozširovanie existujúcich a tvorbu nových digitálnych zdrojov v súlade s vývojom v oblasti korpusovej lingvistiky a počítačovej lexikografie a v súlade s medzinárodnými štandardmi. V rámci projektu sa nadviazali dôležité vzťahy medzi stredo- a východoeurópskymi pracoviskami riešiacimi ekvivalentné problémy počítačovej lingvistiky, počítačového spracovania prirodzeného jazyka a počítačovej lexikografie. Významným prínosom z hľadiska slovenského jazyka bolo vypracovanie morfologickej špecifikácie slovenčiny (Garabík, 2011) v súlade so špecifikáciami projektu Multext East (Erjavec, 2012).

3.2 **Slovak Online** (`http://slovake.eu/`)

Projekt *Slovak Online* (2009 – 2011) bol zameraný na vyučovanie slovenčiny ako cudzieho jazyka prostredníctvom internetu. Cieľom projektu bolo vypracovať online platformu na výučbu slovenčiny (úroveň A1, A2) s využitím moderných metód počítačového výskumu prirodzeného jazyka. V SNK sa pri riešení projektu využili inovatívne postupy kompilácie viacjazyčných elektronických slovníkov zameraných na sémantické vzťahy medzi slovami pri tvorbe slovensko-anglicko-nemecko-poľsko-litovského slovníka, resp. glosára. Jednotlivé položky-významy sú prepojené s databázou anglického WordNet-u v3.0. V rámci slovenskej časti slovníka vznikla a ďalej sa rozširuje (aj po ukončení projektu

Slovak Online) databáza zachytávajúca sémantické vzťahy medzi slovenskými slovami – slovenský WordNet. Z litovskej časti vznikol na pracovisku SNK litovský WordNet, ktorý je úzko previazaný so slovenským WordNetom.

3.3 **EuroMatrixPlus** (`http://www.euromatrixplus.eu/`)

Projekt *EuroMatrixPlus – Bringing Machine Translation for European Languages to the User* (EM+X; 2010 – 2012) mal za cieľ rozšírenie a vylepšenie systému automatického prekladu, osobitne použitím systému MOSES. Hlavný prínos SNK JÚĽŠ SAV spočíval v doplnení automatického prekladu slovenčiny, konkrétne v prekladoch z/do anglického a českého jazyka. Jedným z veľmi úspešných výstupov projektu bol vyššie spomínaný slovensko-anglický paralelný korpus, ktorý v rámci EM+X vznikol, a slovensko-český paralelný korpus, ktorý bol rámci projektu podstatne rozšírený.

Experimentálny systém automatického prekladu medzi slovenčinou a angličtinou/češtinou (obojsmerne) vytvorený v SNK je založený na štatistických metódach prekladu. Tabuľky automaticky zarovnaných fráz boli ďalej spracované do zjednodušeného slovníkového tvaru a prostredníctvom webového rozhrania poskytnuté verejnosti. Ich využitie je vhodné a s úspechom sa aj realizuje pri lexikografických dvojjazyčných projektoch (napr. v spoločnom projekte GAČR ***Konfrontační popis současného českého a slovenského lexika (systémové vztahy a komunikační koexistence)*** s Ústavom pro jazyk český AV ČR Praha).

Obidva paralelné korpusy sa priebežne dopĺňajú a rozširujú v súlade s potrebami používateľov aj po skončení projektu.

3.4 **CESAR** (`http://www.cesar-project.net/`)

Riešitelia projektu CEntral and South-east europeAn Resources (2011 – 2013) sa v úzkej spolupráci s celoeurópskym projektom META-NET (sieť excelentnosti venovaná podpore technologických základov multilingválnej európskej informačnej spoločnosti, `http://www.meta-net.eu/`) zamerali na rozširovanie, sprístupňovanie, skvalitňovanie, štandardizáciu a prepájanie širokého spektra jazykových zdrojov a nástrojov s cieľom prispieť k vybudovaniu otvorenej lingvistickej infraštruktúry. Do projektu sa zapojili jazykové zdroje bulharského, chorvátskeho, maďarského, poľského, slovenského a srbského jazyka v naozaj širokom rozsahu: databázy hovoreného jazyka, rôzne korpusy písaných textov, slovníky, wordnety a zodpovedajúce nástroje počítačového spracovania prirodzeného jazyka, ako sú tokenizátory, lematizátory, taggery a parsery.

Významným prínosom projektu bolo vypracovanie monografie *The Slovak Language in the Digital Age – Slovenský jazyk v digitálnom veku* (Šimková a kol., 2012), v ktorej sa podrobne opisuje aktuálna situácia v oblasti počítačového spracovania prirodzeného jazyka a existujúcich jazykových zdrojov na Slovensku. V rámci projektu sa sprevádzkoval slovenský uzol (`https://metashare.korpus.sk`) celoeurópskej siete META-SHARE zameranej na sprístupňovanie a propagáciu jazykových zdrojov.

Medzi zdroje slovenského jazyka, ktoré sa vďaka projektu CESAR podstatne rozšírili, patrí predovšetkým morfologická databáza slovenčiny. Ako nové boli v SNK vytvorené a voľne sprístupnené n-gramy zo Slovenského národného korpusu, zo zdrojov SNK bola automaticky vytvorená databáza slovenských kolokácií a na základe bulharského, chorvátskeho, maďarského, slovenského a srbského WordNetu vznikol viacjazyčný glosár synsetov.

4 Využívanie korpusov a databáz SNK

Jedným z leitmotívov každej koncepčnej a prípravnej práce hlavného projektu a všetkých súvisiacich projektov v celku SNK je navrhnutie vnútornej a vonkajšej štruktúry, ako aj anotácie konkrétneho korpusu či databázy tak, aby boli nielen užívateľsky ústretové, čo do istej miery závisí aj od pôvodných tvorcov vyhľadávacích nástrojov, a zrozumiteľné každému používateľovi, ale najmä aby obsahovali texty a jazykové informácie, ktoré budú čo najvšestrannejšie využiteľné pre akýkoľvek výskum, hoci jeho ciele a metódy pri tvorbe koncepcie ešte ani neboli známe. Značná časť doterajšieho času riešenia projektu Budovanie Slovenského národného korpusu a elektronizácia jazykovedného výskumu na Slovensku bola preto venovaná základnému výskumu v oblasti počítačového spracovania prirodzeného jazyka a jeho formalizácie, budovania korpusov a ich reálneho i potenciálneho využitia v lingvistike, najmä v lexikografii a terminológii, ale aj v iných vedných odboroch či v bežnom použití. Detailné prepracovanie vlastnej bibliografickej, štýlovo-žánrovej a morfologickej anotácie, vlastného spôsobu prepisovania zvukových záznamov do Slovenského hovoreného korpusu, koncepcia vnútornej štruktúry a budovania všetkých súčastí Slovenského národného korpusu metódou navrstvovania a časovo primeraného sprístupňovania aktuálnych textových dát, ako aj postupnej tvorby rôznych špecializovaných podkorpusov a databáz, to všetko predstavuje súbor teoreticko-analytických i praktických, testovacích prác, z ktorých sa viaceré nevyhnutne cyklicky opakujú. Cenným prínosom pri skvalitňovaní štruktúry textov, ich technického spracovania a lingvistickej anotácie boli v začiatkoch budovania SNK rady a pripomienky skúsenejších počítačových a korpusových lingvistov z partnerských pracovísk v ČR a ďalších okolitých krajinách, ako aj oponentov – účastníkov viacerých oponentských konaní pri schvaľovaní čiastkových koncepcií a kontrole plnenia jednotlivých fáz projektu. Osobitne dôležitú spätnú väzbu poskytovali a poskytujú používatelia, ktorých už od začiatku existencie SNK nebolo málo a ktorí pri svojich výskumoch objavovali zjavnejšie i skrytejšie nedostatky či rezervy alebo možnosti skvalitnenia sprístupnených korpusov a databáz. Každý dobre mienený relevantný a realizovateľný návrh na zlepšenie dát SNK a jednotlivých korpusových výstupov sa v rámci možností zapracoval do príslušných manuálov a krokov spracovania textov alebo ich anotácie a výsledok sa premietol v novej verzii korpusu alebo príslušnej databázy.

Okrem členov kolektívu koncipujúceho nový výkladový Slovník súčasného slovenského jazyka boli v prvých rokoch budovania SNK jeho najčastejšími používateľmi pracovníci, doktorandi a študenti filologických katedier Filozofickej fakulty Prešovskej univerzity v Prešove. V rámci spoločného projektu FF PU a SNK JÚĽŠ SAV *Morfosyntaktická analýza Slovenského národného korpusu*, riešeného pod vedením

M. Sokolovej v r. 2004 – 2006, sa uplatnila nová metodika korpusovolingvistických analýz rozsiahleho materiálu elektronickej databázy slovenských textov s využitím kvantitatívnych metód a otvorili sa možnosti tvorby nového gramatického opisu slovenčiny na korpusovom materiáli. Pri riešení konkrétnych otázok sa zároveň ukázali aj isté obmedzenia či nevýhody práce s korpusom[13] a poukázalo sa na potrebu skvalitňovania korpusových dát z hľadiska ich rozsahu a pestrosti zastúpených textov (pre okrajové jazykové javy) a z hľadiska dobrého výberu textov do vyváženého korpusu (pre frekventované jazykové javy). Na druhej strane sa potvrdil prínos stratégie tvorby a sprístupňovania SNK a jeho rôznych podkorpusov, ktoré sa osvedčili pri konkrétnych čiastkových výskumoch. Výsledky projektu sú zhrnuté v štúdiách v publikácii Sondy do morfosyntaktického výskumu slovenčiny na korpusovom materiáli (2006).

S rozširovaním korpusových a iných databáz SNK, skvalitňovaním anotácie textov a vyhľadávacích nástrojov pribúdali aj ďalší používatelia a nové, špecializované využitia korpusových zdrojov. V rámci spoločných projektov Filozofickej fakulty Univerzity Cyrila a Metoda v Trnave a SNK JÚĽŠ SAV vedených P. Ďurčom v r. 2008 – 2010 (*Konfrontačný výskum kolokácií v slovenčine a v nemčine*) a v r. 2011 – 2013 (*Sémantická a distribučná analýza adjektív v nemčine a slovenčine*) sa rozpracoval spôsob tvorby kolokačných profilov substantív a adjektív, ktoré sa koncipujú v osobitných slovníkoch slovných spojení (Majchráková – Ďurčo, 2010). Celok SNK i jeho jednotlivé súčasti sa postupne stávali zdrojom informácií o rôznych jazykových prostriedkoch a ich reálnom fungovaní v súčasnej slovenčine, ktoré využili mnohí študenti, doktorandi, ako aj etablovaní lingvisti v rozličných oblastiach svojho slovakistického i kontrastívneho výskumu.

Využívanie textového materiálu a jazykových informácií celku SNK v lingvistike je síce rozsiahle a ako primárne stálo pri formulovaní východiskových cieľov projektu Budovanie Slovenského národného korpusu a elektronizácia jazykovedného výskumu na Slovensku, ale medzi jeho používateľov patria aj vedecko-výskumní a iní pracovníci z oblasti počítačového spracovania prirodzeného jazyka, z rôznych matematických, počítačových, informatických a ďalších technických odborov, ako aj z oblasti neurológie, psychológie, logopédie, translatológie a pod. Počet registrovaných používateľov SNK sa z počiatočných 150 – 200 ročne[14] postupne zvýšil na približne 600 ročne, pričom nemalú

[13] Jazykové javy sa v reálnych textoch neraz nachádzajú v takých výskytoch, použitiach a kombináciách, s akými sa lingvista pri klasickej analýze jazyka bez elektronických korpusových zdrojov často ani nestretol, resp. ich nemusel brať do úvahy, a ktoré vyhľadávacie nástroje nedokážu spracovať v jednom kroku či na základe jedného pokynu. Pritom zadanie príkazu na vyhľadávanie musí byť jednak počítačovo presné (syntax príkazov obsahuje kombinácie operátorov tvorených interpunkčnými znamienkami a každá úvodzovka či zátvorka musí byť presne umiestnená), jednak si používateľ musí vedieť zadefinovať všetky parametre vyhľadávania podľa svojich potrieb a pri akejkoľvek zmene nezabudnúť na jej väzby a možnú nevyhnutnosť zmeniť aj súvisiace položky a nastavenia.

[14] Databázy SNK sú bezplatne prístupné na vyhľadávanie jazykových informácií aj bez registrácie, ale v rámci tohto prístupu sú k dispozícii iba dva hlavné korpusy so základnými možnosťami vyhľadávania a paralelné korpusy. Na systematickú a efektívnu prácu so všetkým dostupným materiálom pomocou špecializovaných korpusových nástrojov je potrebná registrácia (`http://korpus.sk/usage.html`), na základe ktorej dostane záujemca individuálne konto.

časť tvoria používatelia zo zahraničných výskumných pracovísk (slovakisti, slavisti aj nelingvisti z ČR, Poľska, Nemecka, Rakúska, Švajčiarska, Holandska, Francúzska, Nórska, USA, Číny a ďalších krajín). Osobitnú skupinu používateľov predstavujú (súčasní i budúci) učitelia základných a stredných škôl, ktorí sa o možnostiach a spôsoboch práce s elektronickými lingvistickými a jazykovými zdrojmi vo výučbe dozvedajú aj z didakticky koncipovaných informačných prehľadov (Očenáš, 2010; Sabol – Andričík – Hevier – Kesselová – Milčák, 2010). Záujemcovia o štúdium slovenského jazyka ako cudzieho a ich učitelia (slovenskí lektori v zahraničí) dostávajú informácie o SNK a jeho zdrojoch každoročne v rámci letnej školy Studia Academica Slovaca, ale neraz si vyžiadajú osobitnú prezentáciu a seminár o práci s korpusom aj počas svojich jazykovo-poznávacích pobytov na Slovensku.

Riešitelia projektu *Budovanie Slovenského národného korpusu a elektronizácia jazykovedného výskumu na Slovensku* si boli počas 10 rokov existencie vždy vedomí toho, že nič z plánovaného projektu by sa nebolo mohlo naplniť bez textov a ich poskytovateľov, že značne problematické by bolo plnenie viacerých úloh bez predchádzajúcich výskumov a spolupráce so zahraničnými pracoviskami, že veľa prác by sa nebolo mohlo realizovať bez množstva externých spolupracovníkov, ktorí skenovali, opravovali a anotovali texty, prepisovali zvukové nahrávky, vypracúvali terminologické záznamy a pod., a že celý projekt by vôbec nemal zmysel bez používateľov jednotlivých výsledkov a korpusových výstupov. Všetkým menovaným i nemenovaným účastníkom tohto spektra ďakujeme za spoluprácu a osobitné poďakovanie vyjadrujeme Ministerstvu kultúry SR, Ministerstvu školstva, vedy, výskumu a športu SR a Predsedníctvu SAV za finančnú i morálnu podporu projektu Budovanie Slovenského národného korpusu a elektronizácia jazykovedného výskumu na Slovensku počas doterajších etáp jeho riešenia.

Literatúra[15]

ATKINSOVÁ, Sue – CLEAR, Jeremy – OSTLER, Nicholas: Kritéria pro výstavbu korpusu. In: Studie z korpusové lingvistiky. Acta Universitatis Carolinae. Philologica 3 – 4. Praha: Univerzita Karlova – Nakladatelství Karolinum 2000, s. 75 – 115.

BELICA, Cyril: Das korpuslinguistische Gesamtkonzept der COSMAS-Plattform. Dostupný z WWW: `http://www1.ids-mannheim.de/kl/projekte/cosmas_I/gesamtkonzept.html`

Registrácia je platná len v rámci kalendárneho roka, vždy na začiatku nového roka sa neaktívnym používateľom konto zamrazí a tým, ktorí prejavia záujem o prácu s korpusom aj naďalej, sa pridelí nové heslo. V každom roku evidujeme iba počet aktívnych používateľov.

[15] Publikačné výstupy pracovníkov oddelenia SNK JÚĽŠ SAV sú dostupné v elektronickej podobe na WWW: `http://korpus.sk/publications.html`, v prípade publikovania v časopisoch vydávaných JÚĽŠ SAV aj priamo na ich stránkach.

BELICA, Cyril: Kookkurrenzdatenbank CCDB – Eine korpuslinguistische Denk- und Experimentierplattform für die Erforschung und theoretische Begründung von systemisch-strukturellen Eigenschaften von Kohäsionsrelationen zwischen den Konstituenten des Sprachgebrauchs, 2001. Dostupný z WWW: http://corpora.ids-mannheim.de/ccdb/

BELICA, Cyril – STEYER, Kathrin: Korpusanalytische Zugänge zu sprachlichem Usus. In: AUC (Acta Universitatis Carolinae) Germanistica Pragensia, Bd. XX. Praha: Karolinum Verlag 2005, s. 7 – 24. Dostupný z WWW: http://www1.ids-mannheim.de/fileadmin/lexik/uwv/dateien/Belica_Steyer_2005_01.pdf

BENKO, Vladimír: Korpus textov slovenského jazyka – súčasný stav a budúcnosť. In: Sociolinguistica Slovaca. 3. Ed. S. Ondrejovič. Bratislava: Veda 1997, s. 297 – 303.

BENKO, Vladimír: Počítačová podpora slovenských lexikografických projektov – retrospektívny pohľad. In: Slovenčina a čeština v počítačovom spracovaní. Ed. A. Jarošová. Bratislava: Veda 2001, s. 181 – 194.

BENKO, Vladimír – HAŠANOVÁ, Jana – KOSTOLANSKÝ, Eduard: Model morfologickej databázy slovenčiny. Počítačové spracovanie jazyka. Trnava: Univerzita sv. Cyrila a Metoda 2004. 188 s.

BUZÁSSYOVÁ, Klára: Metódy výskumu a opisu lexiky slovanských jazykov. Materiály zo sympózia konaného v rámci 7. zasadnutia Lexikologicko-lexikografickej komisie pri Medzinárodnom komitéte slavistov (Nové Vozokany 24. – 26. apríla 1990). Red. V. Blanár. Bratislava 1990. In: Jazykovedný časopis, 1992, roč. 43, č. 2, s. 144 – 146. (recenzia)

CVRČEK, Václav – KOVÁŘÍKOVÁ, Dominika: Možnosti a meze korpusové lingvistiky. In: Naše řeč, 2011, roč. 94, č. 3, s. 113 – 133.

ČERMÁK, František: Jazykový korpus: prostředek a zdroj poznání. In: Studie z korpusové lingvistiky. Acta Universitatis Carolinae. Philologica 3 – 4. Praha: Univerzita Karlova – Nakladatelství Karolinum 2000, s. 15 – 37.

ČERMÁK, František: Český národní korpus: stav v roce 2001. In: Slovenčina a čeština v počítačovom spracovaní. Ed. A. Jarošová. Bratislava: Veda 2001, s. 121 – 135.

DOMIN, Pavol – FORRÓOVÁ, Martina – GARABÍK, Radovan: Mathesiovské semináre. Euro Summer School – Vilém Mathesius Lecture Series 18. In: Jazykovedný časopis, 2003, roč. 54, č. 1 – 2, s. 137 – 141. (správa)

ĎURČO, Peter: Počítačový korpus vlastných mien a automatický výslovnostný transkriptor. In: Zápisník slovenského jazykovedca, 1996, roč. 15, č. 1 – 4, s. 31 – 34 (tézy prednášky konanej dňa 4. 6. 1996 v Slovenskej jazykovednej spoločnosti pri SAV v Bratislave).

ĎURČO, Peter: Počítačové spracovanie vlastných mien na Slovensku. In: Slovenčina na konci 20. storočia, jej normy a perspektívy. Sociolinguistica Slovaca. 3. Red. S. Ondrejovič. Bratislava: Veda 1997, s. 312 – 325.

ERJAVEC, Tomaž: MULTEXT-East: Morphosyntactic Resources for Central and Eastern European Languages. In: Language Resources and Evaluation, 2012, roč. 46, č. 1, s. 131 – 142.

FORRÓOVÁ, Martina – GARABÍK, Radovan – GIANITSOVÁ, Lucia – HORÁK, Alexander – ŠIMKOVÁ, Mária: Návrh morfologického tagsetu SNK. In: Slovanské jazyky v počítačovom spracovaní. Slovko 2003. Zborník z medzinárodnej konferencie nebol publikovaný. Rkp. 19 s. dostupný na WWW: http://korpus.sk/publications.html

FURDÍK, Juraj: Pokus o komplexný kvantitatívny výskum slovotvorného systému slovenčiny pomocou počítača. (Metodologické východisko.) In: Metódy výskumu a opisu lexiky slovanských jazykov. Red. V. Blanár et al. Bratislava: Jazykovedný ústav Ľudovíta Štúra SAV 1990, s. 254 – 264.

FURDÍK, Karol – ŠIMKOVÁ, Mária: Mathesiovské semináre. Vilem Mathesius Lecture Series 12. In: Jazykovedný časopis, 1998, roč. 49, č. 1 – 2, s. 152 – 154. (správa)

GAJDOŠOVÁ, Katarína – ŠIMKOVÁ, Mária: Slovenský hovorený korpus (2008 – 2012). In: Jazykovedné štúdie XXXI. Ed. K. Gajdošová – A. Žáková. Bratislava: Veda 2014, s. 67 – 86.

GARABÍK, Radovan: Štruktúra dát v Slovenskom národnom korpuse a ich vonkajšia anotácia. In: Slovenčina na začiatku 21. storočia. Ed. Mária Imrichová. Prešov: Prešovská univerzita, Fakulta humanitných a prírodných vied 2004, s. 164 – 173.

GARABÍK, Radovan: Corpus Construction Tools. In: Труды международной конференции MegaLing'2005. Прикладная лингвистика в поиске новых путей. Red. В. П. Захаров – С. С. Дикарева. С.-Петербург: Издательство «Осипов» 2005(a), s. 26 – 32.

GARABÍK, Radovan: Levenshtein Edit Operations as a Base for a Morphology Analyzer. In: Computer Treatment of Slavic and East European Languages. Zborník z medzinárodnej vedeckej konferencie Slovko 2005. Red. R. Garabík. Bratislava: Veda 2005(b), s. 50 – 58.

GARABÍK, Radovan: Slovak morphology analyzer based on Levenshtein edit operations. In: 1st Workshop on Intelligent and Knowledge oriented Technologies. Proceedings of the WIKT'06 conference. Ed.: M. Laclavík – I. Budinská – L. Hluchý. Bratislava: Institut of Informatics Slovac Academy of Sciences 2007, s. 2 – 5.

GARABÍK, Radovan: Slovak MULTEXT-East morphology tagset. In: Jazykovedný časopis, 2011, roč. 62, č. 1, s. 19 – 39.

GARABÍK, Radovan – GIANITSOVÁ, Lucia – HORÁK, Alexander – ŠIMKOVÁ, Mária: Tokenizácia, lematizácia a morfologická anotácia Slovenského národného korpusu. 2004. Interný materiál – východiskový manuál na ručnú anotáciu. Dostupný z WWW: http://korpus.sk/publications.html

GARABÍK, Radovan – GIANITSOVÁ-OLOŠTIAKOVÁ, Lucia: Manual Morphological Annotation of the Slovak Translation of Orwell's Novel 1984 – Methods and Findings. In: Computer Treatment of Slavic and East European Languages. In: Zborník z medzinárodnej vedeckej konferencie Slovko 2005. Ed. R. Garabík. Bratislava: Veda 2005, s. 59 – 66.

GARABÍK, Radovan – KAJANOVÁ, Michaela: Problémy a výsledky počítačového spracovania diela *Slowár Slowenski Češko-Laťinsko-Ňemecko-Uherski seu Lexicon Slavicum Bohemico-Latino-Germanico-Ungaricum*. In: Slovo v slovníku. Ed. K. Buzássyová – B. Chocholová – N. Janočková. Bratislava: Veda 2012, s. 294 – 300.

GARABÍK, Radovan – KAJANOVÁ, Michaela: Digitalizácia a anotácia Prameňov k dejinám slovenčiny. In: Prirodzený vývin jazyka a jazykové kontakty, 2013. V tlači.

GARABÍK, Radovan – ŠIMKOVÁ, Mária: Slovak Morphosyntactic Tagset. In: Journal of Language Modelling, 2012(a), roč. 0, č. 1, s. 41 – 63.

GARABÍK, Radovan – ŠIMKOVÁ, Mária: The Slovak National Corpus and its Corpus Linguistic Resources. In: Prace filologiczne, tom LXIII. Warszawa: Wydział polonistyki Uniwersytetu Warszawskiego 2012(b), s. 109 – 119.

HLAVÁČOVÁ, Jaroslava: Orwell's 1984 – Playing with Czech and Slovak Version. In: Computer Treatment of Slavic and East European Languages. Zborník z medzinárodnej vedeckej konferencie Slovko 2005. Red. R. Garabík. Bratislava: Veda 2005, s. 116 – 123.

HORÁK, Alexander: Mathesiovské semináre. Euro Summer School – Vilém Mathesius Lecture Series 17. In: Jazykovedný časopis, 2003, roč. 54, č. 1 – 2, s. 133 – 137. (správa)

HORÁK, Alexander – OLOŠTIAK, Martin – IVANOVÁ, Martina – GIANITSOVÁ, Lucia: Mathesiovské semináre. Euro Summer School – Vilém Mathesius Lecture Series 19. In: Jazykovedný časopis, 2004, roč. 55, č. 1, s. 73 – 78. (správa)

HORECKÝ, Ján: Matematizácia v jazykovede. In: J. Dubnička a kol.: Matematizácia a formalizácia vo vedeckom poznaní. Bratislava: Ústav filozofie a sociológie SAV 1986(a), s. 174 – 179.

HORECKÝ, Ján: Těšitelová, M. a kol.: Kvantitativní charakteristiky současné češtiny. Praha 1985. In: Jazykovedný časopis, 1986(b), roč. 37, č. 1, s. 92 – 93 (recenzia).

HORECKÝ, Ján: Možnosti využiť elektroniku v slovenskom jazyku. In: Slovenský jazyk a literatúra v škole, 1986/87, roč. 33, s. 33 – 35.

HORECKÝ, Ján: Projekt bázy dát slovenského jazyka. In: Metódy výskumu a opisu lexiky slovanských jazykov. Red. V. Blanár et al. Bratislava: Jazykovedný ústav Ľudovíta Štúra SAV 1990(a), s. 251 – 253.

HORECKÝ, Ján: Možnosti komputerizácie v lexikológii. In: Metódy výskumu a opisu lexiky slovanských jazykov. Red. V. Blanár et al. Bratislava: Jazykovedný ústav Ľudovíta Štúra SAV 1990(b), s. 287.

Insight into the Slovak and Czech Corpus Linguistics. Ed. M. Šimková. Bratislava: Veda 2006. 208 s.

JAROŠOVÁ, Alexandra: Korpus textov slovenského jazyka. In: Slovenská reč, 1993, roč. 58, č. 2, s. 89 – 95.

JAROŠOVÁ, Alexandra: Malá inventúra pred hľadaním spoločného jazyka. In: Slovenčina a čeština v počítačovom spracovaní. Ed. A. Jarošová. Bratislava: Veda 2001(a), s. 7 – 10.

JAROŠOVÁ, Alexandra: Národný korpus slovenského jazyka: lingvistické a počítačové aspekty. In: Kultúra slova, 2001(b), roč. 35, č. 4, s. 193 – 199.

KARČOVÁ, Agáta: Príprava a uskutočňovanie projektu morfologického analyzátora. In: Varia. 15. Zborník materiálov z XV. kolokvia mladých jazykovedcov (Banská Bystrica – Tajov 7. – 9. 12. 2005). Ed. Anna Gálisová — Alexandra Chomová. Bratislava: Slovenská jazykovedná spoločnosť pri SAV – Katedra slovenského jazyka a literatúry FHV UMB v Banskej Bystrici 2008, s. 286 – 292.

KARČOVÁ, Agáta – ŠIMKOVÁ, Mária: Морфологічна анотація текстів словацького національного корпусу. In: Лексикографічний бюлетень 13. Київ: Інститут української мови Національної академії наук України, 2006, s. 71 – 76.

KUPIETZ, Marc – BELICA, Cyril – KEIBEL, Holger – WITT, Andreas: The German Reference Corpus DeReKo: A primordial sample for linguistic research. In: Proceedings of the seventh conference on International Language Resources and Evaluation. Eds. N. Calzolari et al. LREC 2010, s. 1848 – 1854. Dostupný z WWW: http://www.lrec-conf.org/proceedings/lrec2010/pdf/414_Paper.pdf

LEVICKÁ, Jana: Terminology and Terminological Activities in the Present-Day Slovakia. In: Computer Treatment of Slavic and East European Languages. Zborník z medzinárodnej vedeckej konferencie Slovko 2007. Ed. J. Levická – R. Garabík. Brno: Tribun 2007, s. 139 – 151.

LEVICKÁ, Jana: Slovenská terminologická databáza. In: Kultúra slova, 2008, roč. 42, č. 3, s. 139 – 157.

MAJCHRÁKOVÁ, Daniela – ĎURČO, Peter: Compiling the First Electronic Dictionary of Slovak Collocations. In: LEXICOGRAPHICA. Feste Wortverbindungen und Lexikographie. Kolloquium zur Lexikographie und Wörterbuchforschung. Ed. P. Ďurčo. Berlin: De Gruyter 2010, s. 105 – 114.

McENERY, Tony – WILSON, Andrew: Corpus Linguistics: An Introduction. 2nd edition. Edinburgh: Edinburgh University Press 2001. 235 s.

Možnosti a meze české gramatiky. Ed. F. Štícha. Praha: Academia 2006. 304 s.

Neologizmy v terminológii marketingu. Ed. J. Levická – K. Viestová. Brno: Tribun 2010. 138 s.

OČENÁŠ, Ivan: Lingvistické zdroje v komunikačnom prostredí internetu. In: Odkazy a výzvy modernej jazykovej komunikácie. Ed. Jana Klincková. Banská Bystrica: Univerzita Mateja Bela 2010, s. 349 – 366.

PÁLEŠ, Emil: SAPFO. Parafrázovač slovenčiny. Počítačový nástroj na modelovanie v jazykovede. Bratislava: Veda 1994. 308 s.

PERKUHN, Rainer – BELICA, Cyril: Korpuslinguistik – das unbekannte Wesen oder Mythen über Korpora und Korpuslinguistik. In: Sprachreport, 2006, Jahrgang 22, Heft 1, s. 2 – 8.

Projekt budovania Národného korpusu slovenského jazyka a projekt elektronizácie jazykovedného výskumu v rokoch 2002 – 2006. Materiál predložený na rokovanie vlády SR. Dostupný z WWW: http://www.rokovania.sk

SABOL, Juraj – ANDRIČÍK, Marián – HEVIER, Daniel – KESSELOVÁ, Jana – MILČÁK, Marián: Využitie informačných a komunikačných technológií v predmete Slovenský jazyk a literatúra pre stredné školy. Košice: Elfa 2010. 195 s.

SEARLE, John R.: Chomsky's revolution in linguistics. In: New York Review of Books, 1972, č. 18, s. 12 – 29.

SOKOLOVÁ, Miloslava – ŠIMKOVÁ, Mária – IVANOVÁ, Martina: Možnosti a medze lingvistického výskumu v Slovenskom národnom korpuse. In: Sondy do morfosyntaktického výskumu slovenčiny na korpusovom materiáli. Ed. M. Sokolová – M. Ivanová. Prešov: Filozofická fakulta Prešovskej univerzity v Prešove 2006, s. 7 – 14.

Sondy do morfosyntaktického výskumu slovenčiny na korpusovom materiáli. Ed. M. Sokolová – M. Ivanová. Prešov: Filozofická fakulta Prešovskej univerzity v Prešove 2006, 196 s.

ŠIMKOVÁ, Mária: Počítačová lingvistika v JÚĽŠ SAV. In: Zápisník slovenského jazykovedca, 1996, roč. 15, č. 1 – 4, s. 50 (tézy prednášky konanej dňa 21. 11. 1996 v Slovenskej jazykovednej spoločnosti pri SAV v Bratislave).

ŠIMKOVÁ, Mária: Počítačové spracovanie prirodzeného jazyka a Slovenský národný korpus. In: Počítačová podpora prekladu. Ed. Marián Smolík – Jaroslav Šoltys – František Tomášik. Bratislava: Slovenská spoločnosť prekladateľov odbornej literatúry 2003, s. 15 – 19.

ŠIMKOVÁ, Mária: Slovenský národný korpus – východiská a plány. In: Slovenčina na začiatku 21. storočia. Ed. Mária Imrichová. Prešov: Prešovská univerzita, Fakulta humanitných a prírodných vied 2004, s. 150 – 158.

ŠIMKOVÁ, Mária: Slovak National Corpus – History and Current Situation. In: Insight into Slovak and Czech Corpus Linguistics. Ed. M. Šimková. Bratislava: Veda 2006, s. 152 – 159.

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

ŠIMKOVÁ, Mária: Desať rokov národného korpusu. In: Správy Slovenskej akadémie vied, 2012, roč. 48, č. 9, s. 12 – 14.

ŠIMKOVÁ, Mária: Язык – корпус – словарь. In: 70 години българска академична лексикография. София: Академично издателство „Проф. Марин Дринов“ 2013, s. 39 – 47.

ŠIMKOVÁ, Mária – GARABÍK, Radovan – GAJDOŠOVÁ, Katarína – LACLAVÍK, Michal – ONDREJOVIČ, Slavomír – JUHÁR, Jozef – GENČI, Ján – FURDÍK, Karol – IVORÍKOVÁ, Helena – IVANECKÝ, Jozef: The Slovak Language in the Digital Age – Slovenský jazyk v digitálnom veku. White Paper Series/Séria bielych kníh. Ed. G. Rehm – H. Uszkoreit. Berlin – New York: Springer 2012. 85 s.

VIESTOVÁ, Kristína – ŠTOFILOVÁ, Jana: Malý lexikón marketingu. Bratislava: Vysoká škola ekonómie a manažmentu verejnej správy 2012. 223 s.

ŽIGO, Pavol: Počítačové spracovanie jazyka. In: Jazykovedný časopis, 1988(a), roč. 39, č. 2, s. 153 – 164.

ŽIGO, Pavol: Využitie počítačov v slovenskom jazykovednom výskume. In: Studia Academica Slovaca. 17. Red. J. Mistrík. Bratislava: Veda 1988(b), s. 469 – 487.

ŽIGO, Pavol: Počítačové tezaury apelatívnej a propriálnej lexiky. In: Metódy výskumu a opisu lexiky slovanských jazykov. Red. V. Blanár et al. Bratislava: Jazykovedný ústav Ľudovíta Štúra SAV 1990, s. 265 – 272.

Slovenský hovorený korpus (2008 – 2012)

Katarína Gajdošová – Mária Šimková

Jazykovedný ústav Ľ. Štúra, Slovenská akadémia vied, Bratislava, Slovensko

Abstract. Corpus of Spoken Slovak is an essential part of the Slovak National Corpus which is carried out by Ľ. Štúr Institute of Linguistics of the Slovak Academy of Sciences. The creation of the corpus started in 2008. The database of speech audio files and text transcriptions exist in four versions. The current version s-hovor-4.0 contains 2.6 million tokens. The database is recorded from speakers having different age, gender, educational background and originating from different regions. Corpus of Spoken Slovak includes sociolinguistic information about speakers and audio files as well as rich annotation of text transcriptions (unintelligibly pronounced words, pauses, lapses, unfinished words, loud breathing, sneezing, yawning, etc.). Deviations from standard pronunciation are marked by the specific attribute *pron*. You can search in the Corpus of Spoken Slovak through the corpus query tool NoSketch Engine or via the WWW interface available at the Slovak National Corpus website.

1 Úvod

Ku koncu 20. storočia sa v oblasti korpusového spracovania textov začali dostávať do popredia hovorené korpusy, ktoré sú v porovnaní s rozsiahlymi písanými korpusmi zvyčajne mnohonásobne menšie. Dôvodom tohto nepomeru je oveľa vyššia náročnosť budovania hovoreného korpusu počnúc zberom materiálu cez prepis zvukových záznamov do textovej podoby, ručnú anotáciu a viacstupňové kontroly až po technické spracovanie takto získaných dát. Finančný rozpočet, ktorý je na celý proces potrebný, je takisto vyšší ako pri tvorbe písaného korpusu. Za veľké hovorené korpusy sa považujú už korpusy v rozsahu niekoľko miliónov tokenov (textových jednotiek), vo väčšine jazykov býva ich rozsah 1 – 4 milióny tokenov, najväčšie dosahujú cca 10 miliónov tokenov. Medzi prvé hovorené korpusy patrí hovorená časť Britského národného korpusu (British National Corpus, 100 mil. tokenov), ktorá predstavuje desatinu jeho celkovej veľkosti (teda približne 10 mil. tokenov). Rozsahom sa mu približuje napr. deväťmiliónový Korpus hovorenej holandčiny (The Spoken Dutch Corpus). Komplexnejší opis existujúcich hovorených korpusov podáva napr. F. Čermák (2006), ich ďalší rozvoj možno sledovať na stránkach korpusových pracovísk alebo príslušných projektov. Prehľad doterajšieho výskumu hovorenej slovenčiny a koncepčnú prípravu Slovenského hovoreného korpusu (ďalej SHK) obsahujú príspevky Hovorený korpus slovenčiny (Šimková – Garabík – Karčová – Gajdošová, 2008), Corpus of Spoken Slovak Language (Rusko – Garabík, 2007) a informačné texty na `http://korpus.sk/shk.html`.

Cieľom projektu SHK je zhromaždiť na jednom mieste rozsahom dostatočný a do istej miery predspracovaný materiál na výskum súčasnej štandardnej hovorenej slovenčiny. SHK sa preto buduje ako databáza textových prepisov zvukových záznamov rôznych prehovorov v štandardnej slovenčine z celého Slovenska so zachytením špecifických rečových

a iných nerečových prvkov a s pridanými sociolingvistickými informáciami o hovoriacich a ďalšími informáciami o zvukovom zázname. Textový prepis zvukových záznamov v databáze je obohatený aj o vnútornú lingvistickú anotáciu – každému slovu je priradená lema (základný tvar slova) a morfologická značka, prostredníctvom ktorých sa text spracúva a ktoré zároveň slúžia na vyhľadávanie konkrétnych jazykových a rečových javov.[1] Databáza SHK je bezplatne prístupná všetkým záujemcom v rámci Slovenského národného korpusu (ďalej SNK) v komplexe korpusov zaradených do vyhľadávača NoSketch Engine, ale i v osobitnom webovom rozhraní (porov. podrobnejšie v časti 6).

Databázy hovoreného korpusu označujeme podľa úzu SNK jednotne skratkou *s-hovor* a číslom verzie. Subkorpusy v rámci aktuálnej verzie sa začali tvoriť od verzie *s-hovor-4.0*. Subkorpus *s-hovor-4.0-upn* obsahuje len prepisy nahrávok z Ústavu pamäti národa, s ktorým oddelenie Slovenského národného korpusu Jazykovedného ústavu Ľ. Štúra SAV dlhodobo spolupracuje (podrobnejšie v nasledujúcej časti). Subkorpus *s-hovor-4.0-sane* tvoria všetky ostatné nahrávky, ktoré sa nachádzajú v korpuse príslušnej verzie *s-hovor*, teda okrem nahrávok z Ústavu pamäti národa. Prehľad verzií SHK, čas ich sprístupnenia a veľkosti jednotlivých korpusov podľa počtu zvukových záznamov, počtu prepísaných hodín a im zodpovedajúceho počtu tokenov uvádzame v tabuľke č. 1.

Názov korpusu/subkorpusu	Prístupný od	Počet nahrávok	Časový rozsah [hh:mm]	Počet tokenov
s-hovor-1.0	2008	71	12:45	127 714
s-hovor-2.0	2010	154	72:17	678 592
s-hovor-3.0	2011	246	180:23	1 643 118
s-hovor-4.0	2012	353	282:37	2 611 504
s-hovor-4.0-sane	2012	291	154:40	1 564 260
s-hovor-4.0-upn	2012	62	127:57	1 047 244

Tabuľka 1. Verzie a subkorpusy Slovenského hovoreného korpusu a ich rozsahy

2 Zdroje zvukových záznamov

Nahrávky do SHK sa získavajú rôznymi spôsobmi z rôznych zdrojov. Základom je vlastné nahrávanie prehovorov pracovníkmi SNK JÚĽŠ SAV v teréne tak, aby bolo pokryté celé územie Slovenska[2]. Nahrávajú sa predovšetkým spontánne rozhovory s vopred vybranými alebo náhodnými respondentmi na rozličné témy (práca, záľuby, šport, školstvo, zdravie a pod.). Istú vzorku prehovorov predstavujú spontánne prednesené (nie čítané) vysokoškolské alebo konferenčné prednášky a prejavy homiletického charakteru. Zvukový záznam

[1] Podrobnejšie o jednotlivých spôsoboch anotácie, celkového spracúvania písaných textov a využitia ich výsledkov pri analýze jazykových javov porov. v predchádzajúcej štúdii v tomto zborníku (Šimková – Garabík, s. 35 – 64).

[2] Podrobnejšiu lokalizáciu všetkých prehovorov, ktoré sa nachádzajú v *s-hovor-4.0*, podľa miesta narodenia hovoriaceho obsahuje mapa v prílohe.

sa do databázy SHK zaradí len so súhlasom respondenta. Nahrávajúci na začiatku alebo na konci rozhovoru osloví respondenta: *„Ak súhlasíte s tým, aby bola táto nahrávka prepísaná, zaradená do databázy Slovenského hovoreného korpusu a slúžila na vedecko-výskumné ciele, povedzte, prosím, áno, súhlasím.“* V prípade súhlasu by mal respondent odpovedať: *„Áno, súhlasím.“*

Okrem zvukových záznamov získaných v teréne sa do databázy SHK zaraďujú aj vzorky z médií. Podmienkou na zaradenie relácie do korpusu je istá miera spontánnosti, takže sa v SHK nachádzajú rôzne diskusné rozhlasové alebo televízne relácie, ale vylúčené sú napr. čítané správy. Pri získavaní nahrávok sa uplatňuje princíp rovnomerného zastúpenia viacerých zdrojov. Aktuálne sú súčasťou databázy relácie z TV Lux, TV Bratislava, Rádio Lumen, Rádio 7 a Hornet rádio, a to v rozsahu približne 10 hodín zvukových záznamov z jedného média. S veľkými (štátnymi ani komerčnými) médiami sa zatiaľ nepodarilo dosiahnuť podpísanie zmluvy na zaradenie výberu z ich relácií do SHK. V súčasnosti sa rokuje so zástupcami študentských rádií a internetových diskusných fór.

Dôležitými poskytovateľmi hovorených komunikátov do SHK sú vysokoškolské pracoviská. Viaceré slovakistické katedry realizovali vlastné výskumy hovorenej podoby slovenčiny v rámci špecificky zameraných projektov, napr. vývin reči dieťaťa[3], reč mládeže daného regiónu (Prešov), jazyk mesta (Banská Bystrica), mediálny prehovor (Nitra) a pod. Vzhľadom na čiastkový charakter týchto výskumov, po ktorých ukončení sa nahrávky ani ich prípadné prepisy na jednotlivých pracoviskách nearchivovali, má zaradenie zozbieraného materiálu do databázy SHK aj širší význam. Zvukové záznamy a ich jednotnou metódou realizované prepisy sú k dispozícii väčšiemu okruhu používateľov, môžu s nimi pracovať viacerí naraz, všetok materiál je archivovaný, dá sa k nemu opätovne vrátiť a výsledky výskumu kedykoľvek verifikovať. V súčasnosti sa v databáze SHK nachádza zvukový materiál z týchto partnerských pracovísk: Inštitút slovakistických, mediálnych a knižničných štúdií Filozofickej fakulty Prešovskej univerzity v Prešove, Katedra slovenského jazyka a literatúry Filozofickej fakulty Katolíckej univerzity v Ružomberku, Katedra slovenského jazyka a literatúry Fakulty humanitných vied Univerzity Mateja Bela v Banskej Bystrici. Systematická spolupráca na zbere, spracovaní a analýze hovorených prejavov sa aktuálne rozvíja s Pedagogickou fakultou Univerzity Komenského v Bratislave.

Osobitnou skupinou poskytovateľov zvukových záznamov sú individuálni poskytovatelia. Sú to zväčša študenti vysokých škôl, ktorí zbierali materiál na výskum v rámci postupových prác a na odporúčanie svojich pedagógov ho poskytli do databázy SHK, ale aj iní poskytovatelia, ktorí súhlasili so zaradením svojich vlastných nahrávok do väčšieho celku.

S poskytovateľmi zvukových záznamov z médií, inštitúcií aj z radov jednotlivcov sa podpisujú licenčné zmluvy podobne ako s poskytovateľmi písaných textov do databázy SNK. Všetci prispievatelia do SHK sú uvedení na stránke `http://korpus.sk/shk_Poskytovatelia_zvukovych_zaznamov.html`.

[3] Hovoriaci v SHK by mali mať vek vyšší ako 20 rokov. Detská reč prekračuje výskumný záber SHK a keďže sa jej dlhodobo a systematicky venujú v rámci viacerých projektov riešiteľské tímy z Prešovskej univerzity, nie je potrebné zahŕňať túto oblasť do SHK.

Samostatnú súčasť tvorby SHK predstavuje spolupráca s Ústavom pamäti národa (ďalej ÚPN) v rámci projektu Oral History – Svedkovia z obdobia neslobody. V archíve ÚPN sa nachádzajú videozáznamy výpovedí pamätníkov, ktorí boli v predchádzajúcom režime z politických dôvodov prenasledovaní, nespravodlivo súdení, väznení, často vo väzení týraní. Rozhovory majú zväčša rovnakú štruktúru. Na vyzvanie moderátora respondent krátko porozpráva o svojom detstve, o pomeroch, z ktorých pochádza, o rokoch dospievania, potom podrobnejšie o zatknutí, vypočúvaní, súdnom procese, väzení, prípadne aj o návrate do života po prepustení z väzby. V rámci spolupráce s ÚPN sa vybrané nahrávky z tohto projektu prepisujú v oddelení SNK, na základe prepisu sa v ÚPN ďalej spracovávajú a po doplnení o profily jednotlivých pamätníkov sprístupňujú verejnosti prostredníctvom portálu ÚPN[4] alebo portálu medzinárodného projektu Paměť národa[5], ktorého je Slovensko členom spolu s ďalšími európskymi krajinami. Prepísané výpovede pamätníkov sa po technických úpravách a doplneniach podľa koncepcie SHK stávajú spolu s ich zvukovým záznamom súčasťou hovorenej databázy v komplexe SNK.

3 Metadáta o hovoriacich a nahrávkach

3.1 Informácie o hovoriacom

Písané texty spracúvané a verejne sprístupňované vo veľkých korpusoch obsahujú štandardnú alebo podrobnú bibliografickú a štýlovo-žánrovú anotáciu. Na základe týchto (vonkajších) informácií o texte sa tvoria špecifické subkorpusy (napr. iba jedného štýlu, iba ženských autoriek) alebo sa dajú selektovať texty z určitého obdobia (napr. 70. roky 20. storočia), z konkrétnej vecnej oblasti (napr. technika, ekonómia, šport) a pod. Pri vonkajšej anotácii prepisov zvukových záznamov sa klasická bibliografická ani štýlovo-žánrová anotácia použiť nedá – na výskum hovorenej podoby sú potrebné údaje zväčša sociolingvistického charakteru (porov. Gajdošová, 2010). Keďže účastníci rozhovorov sú v SHK anonymizovaní, ich mená ani osobné údaje sa nezisťujú ani nikde neuvádzajú. Potrebné lingvistické informácie o hovoriacom/-ich a nahrávke sa zapisujú do osobitného formulára pri nahrávaní rozhovoru. Počas technického spracúvania sa viaceré informácie kódujú, niektoré z nich sa potom použijú v procese anotácie. Pri analýze konkrétnej časti prehovoru sa vždy dá zistiť, koľko rokov mal hovoriaci, odkiaľ pochádza, či ovláda aj iné jazyky a pod.

Formulár na zistenie relevantných údajov o hovoriacom obsahuje 12 položiek, pričom jedna z nich (aké nárečie používa) je viazaná na kladnú odpoveď pri predchádzajúcej otázke (používanie nárečia). Spôsob zápisu jednotlivých položiek v databázach SHK korešponduje so štandardnými zápismi v rámci korpusov (*spk* na začiatku zápisov označuje charakteristiky týkajúce sa hovoriaceho – *speakera*, uvádzame ich v tabuľke č. 2, *doc* na začiatku zápisov označuje metadáta o nahrávke – *dokumente*, uvedené v tabuľke č. 4).

[4] http://www.upn.gov.sk/filmy/projekt-oral-history-svedkovia-z-obdobia-neslobody

[5] http://www.pametnaroda.cz/

Kategória	Zápis v SHK	Hodnoty							
Pohlavie	spk.sex	žena				muž			
Vek	spk.age	< 20	20 – 29	30 – 39	40 – 49	50 – 59	60 – 69	70 – 79	80 ≤
Vzdelanie	spk.education	vysokoškolské		stredoškolské		odborné		nižšie	
Najdlhšie vykonávané povolanie	spk.profession	*napr. knihovník, programátor; zjednotené podľa všeobecného názvu, napr. učiteľ ZŠ, učiteľ SŠ, učiteľ matematiky, učiteľ hudobnej výchovy → učiteľ*							
Miesto narodenia	spk.birthplace	*uvádza sa presne alebo najbližšie väčšie (okresné) mesto, napr. Gelnica*							
Miesto najdlhšieho pobytu	spk.liveplace								
Miesto súčasného pobytu	spk.place								
Materinský jazyk	spk.L1	*zápis podľa európskej normy, napr. sk, hu, pl*							
Iné jazyky, ktoré hovoriaci aktívne ovláda	spk.languages	*zápis podľa európskej normy, napr. de, en, fr, ru*							
Dialekt	spk.dialect	používa				nepoužíva			
Ak používa dialekt, aký?	spk.dialects	*napr. abovský, gemerský, trnavský*							
Informácia o nahrávaní poskytnutá vopred	spk.informed	áno				nie			

Tabuľka 2. Informácie o hovoriacom zaznamenávané v SHK[6]

Na základe uvedených sociolingvistických informácií sa v databáze SHK každému účastníkovi zaznamenávanej komunikácie (aj nahrávajúcemu, ak je aktívnym účastníkom prehovoru) priradí kód. Kód je jedinečný len v rámci jednej nahrávky, nie celej databázy. Jednoznačné určenie hovoriaceho predstavuje kombinácia jeho kódu s kódom nahrávky. (Spôsob kódovania hovoriacich uvádzame v tabuľke č. 3.) Napríklad žene narodenej v Trnave s vysokoškolským vzdelaním vo veku 30 – 39 rokov je v databáze priradený kód *Trytria*. Ak by sa v jednej nahrávke vyskytovali viaceré ženy s rovnakými sociolingvistickými parametrami, odlíšili by sa poslednou samohláskou v kóde (*a, e, i, o, u*), v prípade viacerých mužov by sa na odlíšenie použila posledná spoluhláska (*m, n, p, r* atď.).

[6] Zloženie aktuálnej verzie *s-hovor-4.0* podľa pohlavia, veku, vzdelania a najdlhšie vykonávaného povolania hovoriacich sa nachádza v prílohe.

Okrem štandardných päť-, šesť- a sedemmiestnych kódov sú v databáze aj osobitné štvormiestne kódy. Označujú pracovníkov a spolupracovníkov SNK, ktorí nahrávali väčšinu z terénnych rozhovorov. Ide o kódy Masi, Heiv, Kaga, Agka, Jusl a pod.

Špeciálne kódy Jane, John a all sa prideľujú neznámym hovoriacim, ktorí sa v súvislosti s komunikačnou situáciou nakrátko vyskytli v nahrávke (John = napr. čašník v kaviarni, ktorý sa pýta, čo si debatujúci želajú; Jane = napr. žena, ktorá vstúpi do miestnosti s otázkou, či je v nej niekto, koho hľadá; all = napr. auditórium na prednáške, ktoré hromadne prejavuje súhlas so slovami prednášajúceho). Títo neplánovaní účastníci zaznamenávanej komunikácie nie sú v danej nahrávke cieľovými respondentmi, keďže sa v nej však vyskytujú, je potrebné – vzhľadom na koncepciu prepisu nahrávok v SHK – prepísať aj ich akýkoľvek krátky prehovor a priradiť ho konkrétnemu hovoriacemu. Informácie o ňom nie sú v celom kontexte natoľko relevantné, aby sa osobitne zisťovali a zapisovali, ich replika však zvyčajne vyžaduje reakciu a môže na kratšiu alebo dlhšiu chvíľu zmeniť smer či tému rozhovoru. Ak je takýchto „okoloidúcich“ viac, v jednej nahrávke sa môže vyskytnúť niekoľko rôznych hovoriacich, ktorí sú vždy označení tým istým kódom John alebo Jane. Z formy mena sa dá vyčítať ich pohlavie, z kontextu rozhovoru a komunikačnej situácie sa o nich v prípade potreby dajú dedukovať aj niektoré ďalšie charakteristiky. Výnimku predstavuje kategória *all*, kde sa v podstate nedá zistiť nič (ak nejde o zhromaždenie predstaviteľov výlučne jedného pohlavia, jednej vekovej skupiny, z jedného mesta a pod.).

Kategória	**Hodnoty**	**Zložky kódu**
Miesto narodenia	*uvádzajú sa prvé dve písmená mesta*	napr. *Levoča – Le*
Vzdelanie	nižšie	a
	odborné	e
	stredoškolské	o
	vysokoškolské	y
Vek	< 20	u
	20 ≤ vek < 30	du
	30 ≤ vek < 40	tri
	40 ≤ vek < 50	kva
	50 ≤ vek < 60	kvi
	60 ≤ vek < 70	si
	70 ≤ vek < 80	se
	80 ≤	o
Pohlavie	muž	m (n, p, r, s...)
	žena	a (e, i, o, u, y...)

Tabuľka 3. Kódovanie respondentov v SHK

V procese prvotného technického spracovania nahrávky sa časti nahrávok obsahujúce názvy a mená osôb, ktoré by prípadne boli nejakým spôsobom kompromitované alebo bezdôvodne stavané do negatívneho svetla, vystrihnú a vymažú. Mená a priezviská, ktoré by viedli k bližšej neželanej identifikácii osôb, sú prekryté technickým zvukom. Ide zväčša o nahrávky z projektu Oral History (ÚPN).

3.2 Informácie o nahrávke

Podobne ako sa v nahrávkach kódujú hovoriaci, sú kódované aj celé nahrávky. Kód nahrávky sa skladá z dátumu nahrávania, z prvých dvoch písmen mesta, v ktorom sa nahrávanie uskutočnilo, a z náhodne vygenerovaného štvormiestneho reťazca písmen, napr. *2012-08-01-Leusbo* (nahrávka nahratá 1. augusta 2012 v Leviciach). Jednoznačné určenie nahrávky aj v prípade viacerých miest s rovnakými dvoma začiatočnými písmenami (Levoča, Levice, Letanovce, Leopoldov, Lemešany...) zaručujú dátumy a najmä náhodne generované posledné štyri písmená. Prípadná, málo pravdepodobná, úplná zhoda by sa zachytila a eliminovala pri umiestňovaní nahrávky s rovnakým kódom do archívu SHK.

Kategória	**Zápis v SHK**	**Hodnoty**			
Téma/y rozhovoru	`doc.topic`	*napr. zdravie, rybárstvo, olympijské hry*			
Rozhovor	`doc.spontaneous`	spontánny	`y`	riadený	`n`
Rozhovor	`doc.formal`	formálny	`y`	neformálny	`n`
Účastníci rozhovoru sa	`doc.familiar`	poznajú	`y`	nepoznajú	`n`
Účastníci rozhovoru sú	`doc.equal`	rovnocenní komunikační partneri	`y`	nerovnocenní komunikační partneri	`n`
Komentáre	`doc.comment`	*napr. hovoriaci má rečovú chybu, je fajčiar*			
Dátum rozhovoru	`doc.date`	*v ISO 8601 formáte, napr. 2012-08-01*			
Miesto rozhovoru	`doc.place`	*názov obce (katastra) a konkrétne prostredie, kde sa rozhovor konal (napr. Brezno, trieda v budove školy počas prestávky)*			
Počet účastníkov rozhovoru	`doc.speakers`	*vrátane nahrávajúceho, ak je zároveň aktívnym účastníkom rozhovoru*			

Tabuľka 4. Informácie o nahrávke zaznamenávané v SHK a ich kódovanie

Najfrekventovanejšie témy v SHK súvisia s projektom Oral History: prenasledovanie v čase totality, komunizmus a november 1989. V nahrávkach v podkorpuse *s-hovor-4.0-sane* sú témy veľmi rôznorodé. Vyskytujú sa tu rozhovory o cudzích krajinách (India, Kazachstan, Chorvátsko, pobaltské krajiny), o rôznych športoch (futbal, hokej, hádzaná,

atletika, stolný tenis, speedminton, pétanque), o záľubách (ryby, psy, literatúra), o práci (oprava obuvi, knihovníctvo, učenie a riadenie školy), ale aj o zdravotných problémoch, jazykových otázkach, otázkach viery a pod.

4 Spôsob prepisu nahrávok v SHK

Prepis zvukových záznamov do databázy SHK uskutočňujú školení anotátori – externí spolupracovníci oddelenia SNK JÚĽŠ SAV. Anotátori prepisujú zvukové záznamy v anotačnom nástroji Transcriber[7], v ktorom je k dispozícii súčasne zvuková stopa aj textový prepis. Prepis zvukového záznamu prechádza po anotácii dvojstupňovou kontrolou korektorov.

V koncepcii prepisu zvukových záznamov v SHK sa vymedzili dve roviny prepisu. Na prvej úrovni je štandardný textový prepis podľa pravopisných pravidiel slovenčiny, za lomkou (/) je na druhej rovine zachytená viac-menej reálna výslovnosť. Ak hovoriaci vysloví slovo neštandardne, anotátor zaznačí jeho reálnu podobu za lomku.

Tento spôsob prepisu hovorenej reči do elektronickej databázy bol v čase prípravy koncepcie SHK ojedinelý (napr. korpusy ORAL v Českom národnom korpuse predstavovali v tom čase v podstate štandardizovaný textový prepis bez verného zachytenia priebehu komunikácie, bez prelinkovania so zvukom a bez možnosti analyzovať pôvodný prehovor). V SHK sa na prvej úrovni nachádza doslovný, ale v zásade ortografický prepis, čo umožňuje výsledný text spracovať nástrojmi natrénovanými v písaných textoch a tieto nástroje použiť aj pri vyhľadávaní v SHK. Na druhej úrovni používateľ nájde alebo overuje reálnu výslovnosť. Rovina ortoepického prepisu je limitovaná viacerými faktormi: hovorené korpusy všeobecného charakteru nie sú zamerané na podrobné foneticko-fonologické transkripcie, spolupracujúci anotátori nie sú natoľko špecializovaní, aby rozoznali napr. stupne mäkkosti u hovoriacich z rôznych regiónov Slovenska, pravidlá slovenskej výslovnosti nie sú pre takúto rozsiahlu databázu súčasnej slovenčiny dostatočne prepracované. Materiál spracovaný v SHK do textovej podoby prepojenej so zvukom by však mohol poslúžiť pri analýze a kodifikácii hovorenej podoby súčasnej slovenčiny.

Jav		**Spôsob zápisu**	**Príklad**
dĺžka	samohlások	dvojbodka (:)	*mama/ma:ma*
	spoluhlások	zdvojenie predĺženej spoluhlásky	*všetko/fšetko*
krátenie		krátka samohláska	*výber/vyber*
mäkkosť		mäkčeň	*televízor/ťelevízor*
tvrdosť		horné úvodzovky (")	*veľa/vel"a*

Tabuľka 5. Zápis reálnej výslovnosti v SHK

[7] Viac informácií o nástroji sa nachádza na `http://trans.sourceforge.net/en/presentation.php`.

Ak je slovo vyslovené štandardne, teda v súlade s kodifikáciou alebo akceptovaným úzom, pri anotácii zvukového záznamu nepíše anotátor za príslušným slovom nič. V prípade (výraznej) odchýlky od výslovnostnej normy sa za lomkou zapisuje reálna výslovnosť, napr. tvrdo vyslovená slabika *ľa* v slove *veľa – veľa/vel"a*. Pri finálnom spracovaní konkrétnej verzie hovoreného korpusu sa ku každému slovu, ktoré nemá za sebou lomku a paralelný zápis výslovnostnej podoby, automaticky doplní lomka a za ňu taký istý reťazec znakov, aký je zachytený na prvej rovine prepisu.

5 Označovanie neverbálnych súčastí prehovorov v SHK

Okrem uvedených dvoch rovín prepisu (štandardnej/reálnej) sa v SHK zaznamenávajú aj nerečové udalosti alebo isté ruptúry vo výpovedi. Značky používané v SHK na označenie neverbálnej zložky komunikácie rozdeľujeme do niekoľkých skupín (komplexný prehľad značiek platných od verzie *s-hovor-4.0*[8] sa nachádza v prílohe).

Prvou skupinou sú ruchy (*noise*). Ide najmä o fyziologické prejavy (napr. kýchanie, smrkanie, kašeľ) vyskytujúce sa v prehovore. V skupine ruchy sa osobitne zapisuje, či patria práve hovoriacemu účastníkovi komunikácie (na konci majú písmeno *h*) alebo sa vzťahujú na komunikačného partnera, ktorý počúva (na konci majú písmeno *p*). Do kategórie ruchov patrí aj krátkodobý zvuk v pozadí (*poz*), neartikulovaný zvuk hovoriaceho (*mm*) a neartikulovaný zvuk počúvajúceho (*hh*).

Ďalšia skupina značiek označuje spôsob výslovnostnej realizácie (*pronounce*) jedného slova alebo niekoľkých slov v kontexte. Sem zaraďujeme rečové udalosti, ktoré sa vzťahujú na ruptúrnosť výpovede (napr. nedokončené slovo, lapsus, skomolené slovo), ale aj na jej plynulosť (pauzy, skratky).

Osobitnú skupinu tvoria 2 značky – *pers* na označenie vlastného mena osoby a *loc* na označenie názvu miesta, ktoré sú prekryté technickým zvukom.

Súvislý prehovor v cudzom jazyku je označený dvojmiestnym kódom príslušného jazyka podľa normy ISO 639-2[9].

Okrem konkrétnych hodnôt sa týmto značkám prideľuje aj istý rozsah trvania (*extent*). Začiatok (*begin*) a koniec (*end*) sú párové značky a používajú sa pri tých situáciách, ktorých podstata môže byť spojená s istým časovým úsekom, napr. kým jeden účastník komunikácie hovorí, druhý – jeho komunikačný partner aktuálne v pozícii počúvajúceho – nahlas zívne. Jeho zívanie trvá istý časový úsek, počas ktorého hovoriaci vypovie niekoľko slov. Značka zívanie počúvajúceho (*zivp*) je v rozsahu trvania (*extent*) označená ako začiatok <*desc=zivp extent=begin*> a umiestnená za tým slovom hovoriaceho, kde počúvajúci začal zívať. Jej párová značka <*desc=zivp extent=end*> sa potom umiestni za tým slovom hovoriaceho, kde počúvajúci prestal zívať (napr. <*desc=zivp extent=begin*> *hovorím to v súvislosti* <*desc=zivp extent=end*>).

[8] Značky používané v prvých troch verziách SHK sú uvedené na stránke `http://korpus.sk/shk/Znacky_pouzivane_v_SHK_old.html`. Vzhľadom na ich nepriezračnosť pre používateľov sa vo verzii *s-hovor-4.0* uskutočnili niektoré zjednodušenia a zjednotenia.

[9] `http://www.loc.gov/standards/iso639-2/php/ code_list.php`

Rozsahom môže byť aj časový bod. Značka okamžitej udalosti (instantaneous) je spojená s aktuálnym miestom vo výpovedi hovoriaceho. Nachádza sa na konkrétnom mieste medzi dvoma vypovedanými slovami. Napr. cinknutie lyžičky v šálke počas pauzy medzi dvoma slovami *sused* a *hovoril* (*sused <desc=poz extent=instantaneous> hovoril*).

Udalosť vzťahujúca sa na predchádzajúce slovo má rozsah *previous* (*extent=previous*). Táto situácia môže nastať len pri niektorých značkách a mohli by sme ju považovať za totožnú s rozsahom značiek begin-end pri udalosti vzťahujúcej sa len na jedno slovo (napr. *<desc=zivp extent=begin> kdeže <desc=zivp extent=end>*). Vzhľadom na rozsiahlosť zložených alebo párových štruktúrnych značiek a snahu o prehľadnosť textového prepisu sme pri tvorbe štvrtej verzie hovoreného korpusu pristúpili k unifikácii zápisu takéhoto typu situácie a jednotne používame už len zápis s rozsahom *previous* (napr. *kdeže <desc=zivp extent=previous>*). Väčšina značiek z kategórie výslovnostná realizácia (*pronounce*) môže byť vo svojom rozsahu len konkrétna (*instantaneous*), čo vyplýva z povahy samotnej udalosti, a preto nie je potrebné uvádzať pri nej rozsah (napr. *skomolenina/skolomenina <desc=skom extent=instantaneous>*). V tabuľke č. 6 sú všetky značky používané v databáze SHK od štvrtej verzie rozčlenené podľa typu značky (*event type*) a podľa rozsahu (*extent*).

		extent					extent					extent					extent			
event type	**noise**	ins	beg	end	prev	**pron**	ins	beg	end	prev	**lang**	ins	beg	end	prev	**ent**	ins	beg	end	prev
desc	dychh	+	+	+	+	ps-	+	–	–	–	ar	+	+	+	–	pers	+	–	–	–
	smrkh	+	+	+	+	ps--	+	–	–	–	cs	+	+	+	–	loc	+	–	–	–
	kychh	+	+	+	+	skrcit	+	–	–	–	el	+	+	+	–					
	smiechh	+	+	+	+	skrhlas	+	–	–	–	en	+	+	+	–					
	sepkh	+	+	+	+	skrcudz	+	–	–	–	fr	+	+	+	–					
	zivh	+	+	+	+	laps1	+	–	–	–	he	+	+	+	–					
	stikh	+	+	+	+	laps2	+	–	–	–	hu	+	+	+	–					
	dychp	+	+	+	+	laps3	+	–	–	–	it	+	+	+	–					
	smrkp	+	+	+	+	laps4	+	–	–	–	la	+	+	+	–					
	kaslp	+	+	+	+	laps5	+	–	–	–	nl	+	+	+	–					
	kychp	+	+	+	+	laps6	+	–	–	–	pl	+	+	+	–					
	smiechp	+	+	+	+	laps7	+	–	–	–	ru	+	+	+	–					
	sepkp	+	+	+	+	laps8	+	–	–	–	cu	+	+	+	–					
	zivp	+	+	+	+	laps9	+	–	–	–	sr	+	+	+	–					
	stikp	+	+	+	+	laps10	+	–	–	–	zh	+	+	+	–					
	poz	+	+	+	+	skom	+	–	–	–	eo	+	+	+	–					
	mm	+	–	–	–	ned	+	–	–	–										
	hh	+	+	+	+	nezr	+	+	+	–										
						mum	+	+	+	–										
						dial	+	+	+	–										

		spev	+	+	+	+	
		***	+	-	-	-	

Tabuľka 6. Zoznam značiek na označenie rečových a nerečových udalostí a ich rozsahov v SHK

Vysvetlivky:
ins (instantaneous) = okamžitá udalosť; beg (begin) = začiatok udalosti; end (end) = koniec udalosti; prev (previous) = predchádzajúca udalosť; pron (pronounce) = výslovnostná realizácia; lang (languages) = jazyky; ent (entities) = zatreté mená; + = značka môže mať prislúchajúcu hodnotu; - = značka nemôže mať prislúchajúcu hodnotu; značky zo skupín ruchy, výslovnostná realizácia a zatreté mená sú podrobnejšie rozpísané v prílohe.

Značkou *laps1* sa označuje prerieknutie sa hovoriaceho v jednom slove, značka má vzťah iba k jednému predchádzajúcemu slovu, preto nie je potrebné bližšie určenie značkou <*entent=previous*>. *Laps* s vyššími číslami sa vzťahuje na prerieknutie sa v prislúchajúcom počte slov spätne, preto sa táto značka vždy umiestňuje za posledné slovo, ktorým sa hovoriaci preriekol, napr. *laps7* znamená, že prerieknutie sa vzťahuje na sedem predchádzajúcich výrazov.

Neartikulovaný zvuk hovoriaceho, ktorý sa označuje značkou *mm*, sa môže realizovať len v rámci výpovede v konkrétnom časovom bode podobne ako slovo. Preto musí byť značka umiestnená medzi dvoma konkrétnymi slovami s rozsahom <*extent= instantaneous*>. Pre neartikulovaný zvuk počúvajúceho, označený značkou *hh*, je okrem konkrétneho umiestnenia v čase (*instantaneous*) typické aj umiestnenie v rozsahu begin-end. Počas prehovoru hovoriaceho môže počúvajúci neartikulovaným zvukom vyjadrovať ne/súhlas s prehovorom hovoriaceho a jeho neartikulované zvuky môžu trvať aj počas niekoľkých slov. Preto je v tabuľke zaznačená možnosť jeho realizácie vo všetkých štyroch rozsahoch (*instantaneous, begin, end, previous*).

6 Možnosti práce s databázami SHK

Záujemcovia o skúmanie hovorenej slovenčiny majú v súčasnosti možnosť pracovať s databázami SHK dvoma spôsobmi.

6.1 Bonito / NoSketch Engine

V textovom prepise zvukových záznamov je možné vyhľadávať prostredníctvom korpusového manažéra Manatee s klientom Bonito[10]. Klient Bonito sa však už prestal vyvíjať. Registrovaní používatelia SNK aktuálne pracujú s databázami SNK prostredníctvom vyhľadávacieho nástroja NoSketch Engine cez webovú stránku `http://bonito.korpus.sk`. Pri práci s textovým prepisom databáz SHK má používateľ k dispozícii všetky možnosti, ktoré sú známe z písaných korpusov SNK. Štandardné zápisy slova sú od prvej verzie hovoreného korpusu automatizovane lematizované a morfologicky anotované, preto je možné v databáze SHK vyhľadávať aj na základe lemy a tagu. Samozrejmosťou je vyhľadávanie pomocou konkrétneho tvaru slova (*word*). Rovnako si používateľ môže nastaviť parameter výslovnostnej realizácie (atribút *pron*) a hľadať podľa reálnej výslovnosti, ktorá sa nachádza v prepise za lomkou. Okrem toho je možné používať

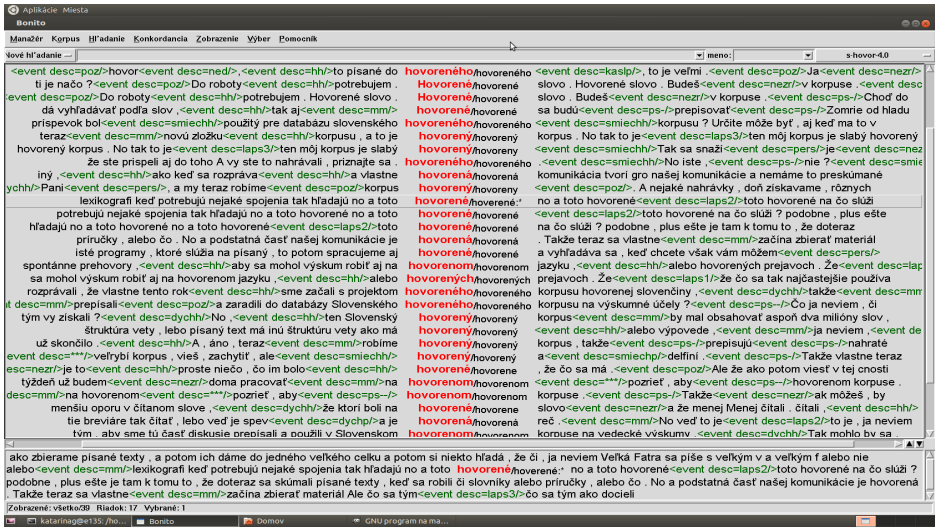
[10]Inštalačné balíčky pre OS Windows a OS Linux sú prístupné na stránke `http://korpus.sk/Bonito_old.html`.

aj regulárne výrazy a kombinácie všetkých uvedených možností (*word, lemma, tag, pron*) použitím CQL výrazu.

Od štvrtej verzie SHK sa v kliente Bonito zobrazujú aj sociolingvistické informácie o hovoriacich. Podľa týchto údajov sa dá automaticky zistiť distribúcia konkrétneho javu (napr. výslovnosť *n"e* miesto *nie*) podľa pohlavia, veku, pobytu hovoriacich a ďalších sociolingvistických charakteristík zaznamenaných vo formulároch SHK.

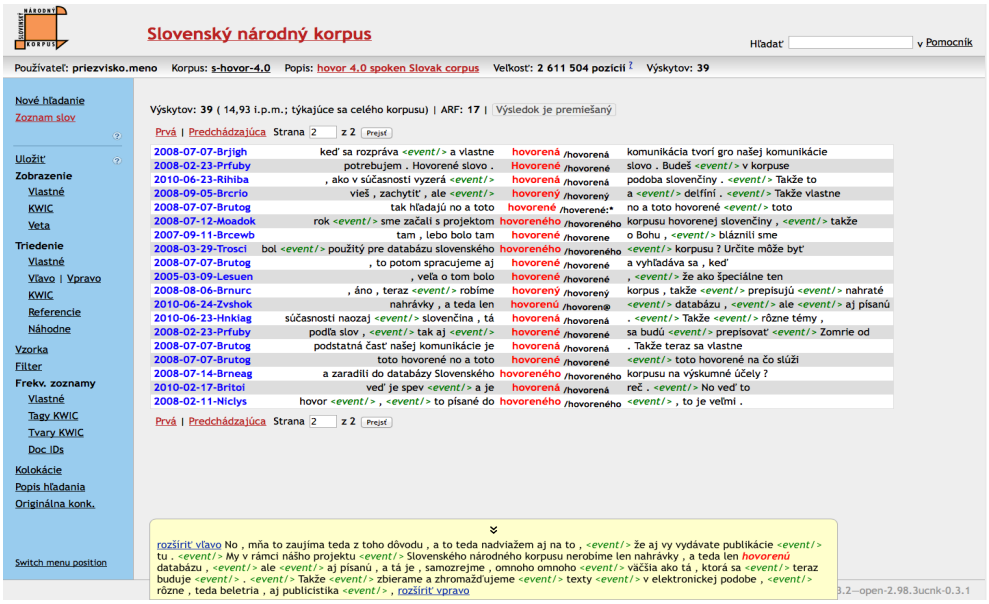
Pri vyhľadávaní sa značky použité v hovorenom korpuse zobrazujú ako štruktúrne značky zatiaľ v kliente Bonito a vo webovom rozhraní na stránke SNK. V prostredí klienta Bonito si používateľ môže zvoliť zobrazenie všetkých alebo len vybraných atribútov značky: typ značky (*event.type*), jej rozsah (*event.extent*) a opis (*event.desc*). Súčasne s hľadaným reťazcom je vhodné použiť aj regulárny výraz pre slovo, ktorým je pravá a ľavá hranatá zátvorka, čiže na zobrazenie štandardnej konkordancie s akýmkoľvek slovom, ktoré predstavuje prečítanú podobu cudzojazyčnej skratky, zapíšeme do vyhľadávacieho okienka: *[] <event desc="skrcudz"/>*.

So zapnutou voľbou zobrazenia všetkých troch ukazovateľov (*event.type, event.extent, event.desc*) môže pôsobiť štruktúrna značka na prvý pohľad komplikovane. Príkladom takej značky je napr. reťazec *<event type=pronounce desc=laps1 extent=instantaneous/>*, ktorý označuje príslušnosť ku kategórii spôsob výslovnostnej realizácie (*event type=pronounce*), prerieknutie vzťahujúce sa na 1 slovo späť (*desc=laps1*) a rozsah – podľa tabuľky č. 6. Používateľ si môže sám zvoliť, ktorú informáciu o značke chce vidieť. Zobrazenie typu značiek (*event type*) je vhodné v prípade, keď si používateľ nie je ešte celkom istý delením značiek na jednotlivé skupiny (*noise, pronounce, entities, language*). Najdôležitejšou informáciou je značka *desc*, ktorou sa priamo označuje nejaká udalosť. Rozsah značky (*extent*) je vhodné použiť najmä pri značkách, ktoré môžu mať rôzne trvanie (konkrétne v čase, počas trvania istej časti výpovede alebo slova).



Obr. 1. Konkordancia lemy *hovorený* v kliente Bonito so zobrazením štruktúrnych značiek *event desc* a výslovnostnej realizácie v atribúte *pron*.

V nástroji NoSketch Engine je možné zobraziť spolu so základnými atribútmi (*word, lemma, pron, tag, prec*) príslušné štruktúry, napr. *event, turn, who*.



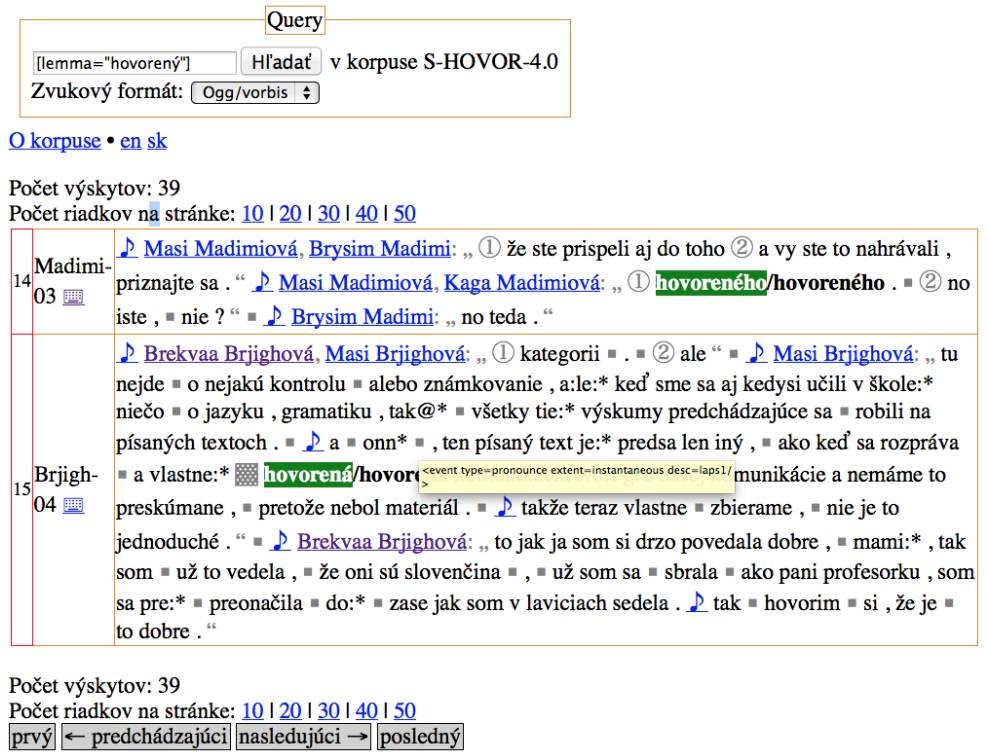
Obr. 2. Konkordancia lemy *hovorený* v nástroji NoSketch Engine so zobrazením štruktúrnej značky *event* a výslovnostnej realizácie v atribúte *pron*.

6.2 Webové rozhranie SNK

Druhý spôsob predstavuje práca s databázami SHK prostredníctvom webového rozhrania (`http://korpus.juls.savba.sk:8086/oral/`) na stránkach SNK. Osobitosťou databázy SHK je možnosť prístupu nielen k čistému textu, ale aj k jeho pôvodnému zvukovému záznamu, s ktorým je prelinkovaný. Mnohé svetové korpusy, ani hovorené korpusy v okolitých krajinách, touto možnosťou nedisponujú. V súčasnosti sa pracuje na prepojení textového prepisu a zvukovej realizácie prehovoru aj v rámci vyhľadávacieho nástroja NoSketch Engine.

Prostredníctvom webového rozhrania má používateľ k dispozícii vyhľadávanie na základe všetkých známych atribútov (word, lemma, tag, pron, CQL), pričom sa priamo zobrazuje prepis reálnej výslovnosti. Pod vyhľadávacím políčkom si používateľ vyberie jeden z troch zvukových formátov, v ktorom si môže vypočuť príslušnú časť zvukového záznamu. Po vyhľadaní zadaného výrazu sa okrem konkordancie zobrazí aj informácia o počte vyhľadaných výskytov a možnosti zmeniť počet riadkov zobrazených na strane. Pod grafickými obrazcami (štvorce) sú v texte skryté značky, ktoré sa zobrazia pri pohybe

myšou po značke. Po kliknutí na notu v časti textového prepisu si môže používateľ vypočuť príslušnú časť zvukového záznamu, ktorá je prepísaná v texte, vo zvolenom zvukovom formáte (Ogg Vorbis, Speex, Flac).



Obr. 3. Konkordancia lemy *hovorený* vo webovom rozhraní so zobrazenou štruktúrnou značkou

V ľavom stĺpci sa nachádzajú kódy nahrávok (napr. *Madimi-03*[11]), z ktorých sa zobrazuje krátky kontext hľadaného výrazu. Po kliknutí na ikonu klávesnice vedľa kódu nahrávky vidí používateľ metadáta týkajúce sa nahrávky. Repliky jednotlivých hovoriacich sú uvedené ich kódmi. Kód hovoriaceho vyzerá ako meno a priezvisko. Meno tvorí vlastný kód hovoriaceho, priezvisko kód príslušnej nahrávky. Hovoriace ženského pohlavia majú koncovku *-ová* na rýchlejšiu orientáciu pri určovaní pohlavia, napr. *Brekvaa Brjighová*. Po kliknutí na kód hovoriaceho sa zobrazia metadáta o respondentovi.

[11] Niektoré nahrávky sú rozdelené na viacero častí z dôvodu príliš veľkého rozsahu alebo vystrihnutia časti prehovoru neželaného obsahu. Jednotlivé časti nahrávok sú potom označené číslom (01, 02...), nadväznosť nahrávky na predchádzajúcu časť je zaznačená symbolom +, nenadväznosť nahrávky (nejaká časť nahrávky bola z istých dôvodov vystrihnutá) je označená symbolom – (napr. *Madimi-03* je tretia časť nahrávky s kódom *2008-06-26-Madimi*, pričom nahrávka nenadväzuje plynule na predchádzajúcu časť *Madimi-02*).

Použitie webového rozhrania na prístup k databázam SHK nie je podmienené registráciou a prihlasovaním používateľa. Všetkým záujemcom je bezplatne dostupné tak, ako aj vybrané písané korpusy.

Brekvaa Brjighová, žena, vo veku 40+ rokov, vzdelanie stredoškolské, miesto narodenia Bratislava, miesto najdlhšieho pobytu Bratislava, miesto súčasného pobytu Bratislava, materinský jazyk sk, ovláda jazyky: en, bola informovaná o nahrávaní.

```
sex: f
age: 40+
profession:
education: stredoškolské
birthplace: Bratislava
liveplace: Bratislava
place: Bratislava
L1: sk
languages: en
dialect: n
dialects:
informed: y
```

Obr. 4. Metainformácie o hovoriacej *Brekvaa Brjighová* vo webovom rozhraní

7 Záver

Počas piatich rokov budovania hovoreného korpusu štandardnej slovenčiny sa táto špecializovaná databáza komplexu Slovenského národného korpusu rozrástla na takmer 4 milióny textových jednotiek, z ktorých je v aktuálnej verzii *s-hovor-4.0* vyše 2,6 mil. jednotiek sprístupnených na výskum. Na finalizácii novej verzie sa začne pracovať začiatkom r. 2014 – doplnia sa prehovory a ich prepisy, ktoré pribudli od poslednej verzie, odstránia sa zistené chyby.

Rozsahom zvukových záznamov a ich prepisov patrí Slovenský hovorený korpus už teraz medzi hovorené korpusy štandardnej veľkosti, ktoré predstavujú istý referenčný zdroj na výskum hovorenej podoby jazyka. Možnosťami vyhľadávania a celkovým spracovaním (lematizácia, morfologická anotácia, dve úrovne prepisu, zaznamenávanie nerečových súčastí komunikácie a najmä prepojenie prepisu so zvukom online prístupné všetkým záujemcom) predstavuje Slovenský hovorený korpus nadštandardnú databázu záznamov a prepisov hovorenej reči. Pridanou hodnotou je aj mapovanie reálnej podoby hovorenej slovenčiny na celom území Slovenska, k čomu prispieva spolupráca s Ústavom pamäti národa a s viacerými univerzitnými slovakistickými pracoviskami. Okrem rozširovania materiálovej základne, skvalitňovania prepisov a celého spracovania hovorenej databázy patrí medzi najbližšie úlohy v tejto oblasti príprava koncepcie a pilotného nárečového korpusu slovenčiny, pri ktorom sa môžu využiť viaceré skúsenosti, metódy a nástroje uplatňované pri budovaní štandardného hovoreného korpusu, ako aj podrobnejšia transkripcia vybranej vzorky prepisov zvukových záznamov.

Slovenský hovorený korpus je už v tejto podobe predmetom rôznych analýz a štatistických prehľadov (porov. napr. Gajdošová, 2010; Hoffmannová, 2010; Kesselová, 2011; Šimková, 2011) a môže sa stať dobrým zdrojom na tvorbu aktuálnej ortoepickej príručky súčasnej slovenčiny.

Literatúra

ČERMÁK, František: Mluvené korpusy. In: Korpusová lingvistika. Stav a modelové přístupy. Ed. F. Čermák – R. Blatná. Praha: Nakladatelství Lidové noviny – Ústav Českého národního korpusu 2006, s. 53 – 67.

GAJDOŠOVÁ, Katarína: Cudzojazyčné výrazy v Slovenskom hovorenom korpuse. In: Slovo o slove. 16. Zborník Katedry komunikačnej a literárnej výchovy Pedagogickej fakulty Prešovskej univerzity. Ed. Ľ. Liptáková – M. Andričíková – M. Klimovič. Prešov: Katedra komunikačnej a literárnej výchovy, Pedagogická fakulta Prešovskej univerzity v Prešove 2010, s. 190 – 197.

GAJDOŠOVÁ, Katarína: Metadáta v Slovenskom hovorenom korpuse. In: Varia. 17. Zborník materiálov zo XVII. kolokvia mladých jazykovedcov (Liptovská Osada – Škutovky 7. – 9. 11. 2007). Zost. V. Kováčová. Ružomberok: Katolícka univerzita v Ružomberku – Slovenská jazykovedná spoločnosť pri JÚĽŠ SAV v Bratislave 2010, s. 115 – 120.

HOFFMANNOVÁ, Jana: České *jako* a slovenské *ako/akože* v mluvených projevech (malý konfrontační pokus). In: Slovo – Tvorba – Dynamickosť. Ed. M. Šimková. Bratislava: Veda 2010, s. 359 – 371.

KESSELOVÁ, Jana: Zlučovací vzťah v komunikačných a kognitívnych súvislostiach. In: Jazykovedný časopis, 2011, roč. 62, č. 2, s. 81 – 94.

RUSKO, Milan – GARABÍK, Radovan: Corpus of Spoken Slovak Language. In: Computer Treatment of Slavic and East European Languages. Zborník z medzinárodnej vedeckej konferencie Slovko 2007. Ed. J. Levická – R. Garabík. Brno: Tribun 2007, s. 222 – 236.

Slovenský hovorený korpus. Verzia s-hovor-4.0. Dostupný z WWW: `http://korpus.juls.savba.sk`

Slovenský národný korpus. Dostupný z WWW: `http://korpus.juls.savba.sk`.

ŠIMKOVÁ, Mária – GARABÍK, Radovan: Slovenský národný korpus (2002 – 2012): východiská, ciele a výsledky pre výskum a prax. In: Jazykovedné štúdie XXXI. Ed. K. Gajdošová – A. Záková. Bratislava: Veda 2014, s. 37 – 66.

ŠIMKOVÁ, Mária – GARABÍK, Radovan – KARČOVÁ, Agáta – GAJDOŠOVÁ, Katarína: Hovorený korpus slovenčiny. In: Čeština v mluveném korpusu. Ed. M. Kopřivová – M. Waclawičová. Praha: Nakladatelství Lidové noviny – Ústav Českého národního korpusu 2008, s. 227 – 233.

ŠIMKOVÁ, Mária: Frekvencia slov a tvarov v súčasnej slovenčine. In: Slovenská reč, 2011, roč. 76, č. 5 – 6, s. 322 – 333.

Prílohy

I. Zoznam značiek použitých v SHK na označenie ruchov, výslovnostných realizácií, technických zvukov a cudzích jazykov.

Ruchy (noise)

dychh – hlasné dýchanie (hovoriaci)
smrkh – smrkanie (hovoriaci)
kaslh – kašľanie (hovoriaci)
kychh – kýchanie (hovoriaci)
smiechh – smiech (hovoriaci)
sepkh – šepkanie (hovoriaci)
zivh – zívanie (hovoriaci)
stikh – štikútanie (hovoriaci)
mm – neartikulovaný zvuk hovoriaceho
hh – neartikulovaný zvuk počúvajúceho

dychp – hlasné dýchanie (počúvajúci)
smrkp – smrkanie (počúvajúci)
kaslp – kašľanie (počúvajúci)
kychp – kýchanie (počúvajúci)
smiechp – smiech (počúvajúci)
sepkp – šepkanie (počúvajúci)
zivp – zívanie (počúvajúci)
stikp – štikútanie (počúvajúci)
poz – krátkodobý zvuk v pozadí
spev – spievanie

Výslovnostná realizácia (pronounce)

ps- – krátka pauza
ps-- – stredne dlhá pauza
ps--- – dlhá pauza
skom – skomolené slovo
ned – nedokončené slovo
skrcit – skratka prečítaná, napr. SNK/snk
skrhlas – skratka hláskovaná, napr. SNK/esenká
skrcudz – skratka v cudzom jazyku, napr. IBM/ajbíem
laps1 – prerieknutie vzťahujúce sa na predchádzajúce slovo
laps2 – prerieknutie vzťahujúce sa na 2 predchádzajúce slová
laps3 – prerieknutie vzťahujúce sa na 3 predchádzajúce slová
laps4 – prerieknutie vzťahujúce sa na 4 predchádzajúce slová
laps5 – prerieknutie vzťahujúce sa na 5 predchádzajúcich slov
laps6 – prerieknutie vzťahujúce sa na 6 predchádzajúcich slov
laps7 – prerieknutie vzťahujúce sa na 7 predchádzajúcich slov
laps8 – prerieknutie vzťahujúce sa na 8 predchádzajúcich slov
laps9 – prerieknutie vzťahujúce sa na 9 predchádzajúcich slov
laps10 – prerieknutie vzťahujúce sa na 10 predchádzajúcich slov
nezr – nezrozumiteľne vyslovené slovo
mum – mumlavý, zachrípnutý spôsob prehovoru
dial – vypovedané v nárečí
******* – anotátor nevedel danú časť prehovoru identifikovať; jedna značka *** = jedno slovo

Prekrytie vlastných mien (entities)

pers – miesto technického prekrytia vlastného mena osoby
loc – miesto technického prekrytia vlastného mena lokality, obchodného mena alebo adresy

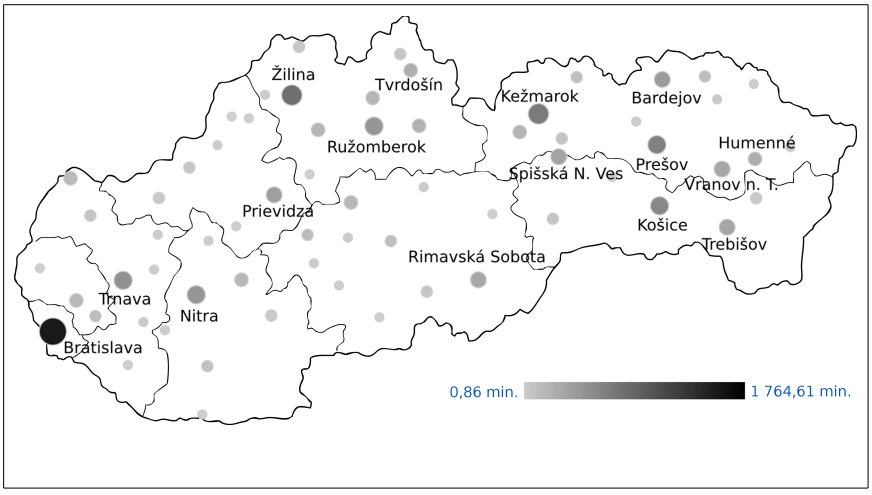
Jazyky (languages)

ar – arabčina
cs – čeština
de – nemčina
el – moderná gréčtina
en – angličtina
fr – francúzština
he – hebrejčina
hu – maďarčina
it – taliančina
la – latinčina
nl – holandčina
pl – poľština
ru – ruština
cu – cirkevná slovančina/staroslovienčina
sr – srbčina
zh – čínština
eo – esperanto

II. Zloženie aktuálnej verzie *s-hovor-4.0* podľa vybraných parametrov

Mapa Slovenska s vyznačenými okresnými mestami, z ktorých pochádzajú hovoriaci – účastníci komunikácií zaznamenaných a prepísaných v databázach SHK.

V metadátach o hovoriacom sa miesto jeho narodenia uvádza buď priamo, alebo ako najbližšie väčšie či okresné mesto. Okresné mestá zakreslené v mape zahŕňajú celý okres. Pre lepšiu prehľadnosť sme vyznačili iba názvy 16 najviac zastúpených okresov (od 1 764,61 min. do 314,54 min.), hovoriaci z ďalších naznačených lokalít sú zastúpení menej ako 300 minútami. V dôsledku rôznych migračných pohybov obyvateľov SR, ktoré sú v súčasnosti značne rozsiahle, rozloženie hovoriacich podľa miesta aktuálneho či najdlhšieho pobytu by predstavovalo odlišný obraz vrátane nezriedkavých presahov za hranice SR.



Obr 4. Mapa Slovenska s vyznačenými mestami, z ktorých pochádzajú hovoriaci

Zloženie verzie *s-hovor-4.0* podľa pohlavia hovoriacich. Neurčené pohlavie sa týka heterogénnej skupiny all – kolektívneho účastníka prednášok a iných verejných komunikačných situácií.

Pohlavie hovoriaceho	Rozsah prehovorov			
	v minútach		v tokenoch	
	abs.	%	abs.	%
m	9 941,01	58,63	1 486 090	56,91
f	6 724,76	39,66	1 125 378	43,09

Zloženie verzie *s-hovor-4.0* podľa veku hovoriacich v čase zaznamenania komunikácie.

Vek hovoriaceho	Rozsah prehovorov			
	v minútach		v tokenoch	
	abs.	%	abs.	%
00 – 09	3,29	0,02	617	0,02
10 – 19	99,90	0,59	16 905	0,65
20 – 29	1 602,26	9,45	296 774	11,36
30 – 39	1 937,30	11,43	369 528	14,15
40 – 49	1 392,54	8,21	252 597	9,67
50 – 59	1 568,48	9,25	249 598	9,56
60 – 69	1 424,70	8,40	207 151	7,93
70 – 79	3 295,50	19,44	463 776	17,76
80 – 89	3 529,81	20,82	460 939	17,65
90 – 99	96,50	0,57	12 940	0,50
neurčený	1 715,03	10,11	280 679	10,75

Zloženie verzie *s-hovor-4.0* podľa vzdelania hovoriacich.

Vzdelanie hovoriaceho	Rozsah prehovorov			
	v minútach		v tokenoch	
	abs.	%	abs.	%
vysokoškolské	9 263,23	54,63	1 489 451	57,03
stredoškolské	3 869,23	22,82	567 228	21,72
odborné	1 505,50	8,88	224 283	8,59
nižšie	57,26	0,34	8 392	0,32
žiadne	1,03	0,01	141	0,01
nezistené	1 969,80	11,62	322 009	12,33

Zloženie verzie *s-hovor-4.0* podľa najdlhšie vykonávaného povolania hovoriacich. Pre lepšiu prehľadnosť uvádzame číselné hodnoty iba v absolútnom počte tokenov.

Povolanie	**Tokeny**
nezistené	549 534
vedecký pracovník	323 960
učiteľ	154 105
kňaz	120 495
študent	101 199
stavebný technik	98 861
manažér	71 721
skladník	67 355
lekár	62 467
stavebný inžinier	57 992
úradník	47 774
vysokoškolský pedagóg	43 261
technický pracovník	42 066
redaktor	37 360
knihovník	35 041
športovec	30 564
elektrozámočník	25 784
obchodník	25 173
rehoľník	23 845
odborný pracovník	23 443
revízny technik	22 898
opravár	22 584
administratívny pracovník	22 312
novinár	21 458
múzejný pracovník	20 830
robotník	20 755
právnik	20 702
ekonóm	19 385
krčmár	18 107
akademický maliar	17 859
spisovateľ	17 672
dramaturg	17 033
zdravotnícky asistent	15 722
astronóm	15 511
pracovník v štátnej správe	15 140
historik	15 060
výskumný pracovník	15 005
podnikateľ	14 646
imunológ	14 438
maliar - natierač	14 162
žena v domácnosti	13 552
psychológ	13 481
bábkoherec	12 529
účtovník	11 797
rabín	11 676
stavbyvedúci	11 097

Povolanie	**Tokeny**
doktor	10 135
opatrovateľ	10 013
programátor	9 858
tréner	8 740
zootechnik	8 651
kultúrny pracovník	8 289
pracovník telekom. spol.	8 173
farmaceut	8 076
muzeológ	8 041
etnológ	7 932
informatik	7 910
IT pracovník	7 905
projektant	7 485
hovorca	7 312
realitný maklér	7 120
zdravotná sestra	7 025
zámočník	6 880
sociálny pracovník	6 651
zamestnanec vysokej školy	6 421
predavač	5 846
zbormajster	5 828
krajčír	5 789
šepkár	5 695
politológ	5 580
štátny zamestnanec	4 549
policajt	4 486
pracovník v knižnici	4 476
finančný analytik	4 333
optik	4 333
premietač	4 314
kunsthistorik	4 183
geológ	3 671
inšpicient bábkového divadla	2 762
obuvník	2 705
operátor	1 957
hudobný teoretik	1 889
dentálny hygienik	1 803
lekárnik	1 446
architekt	1 413
strihač filmov	1 103
upratovač	941
referent	660
stavbár	658
bibliograf	574
vojak z povolania	301
kameraman	181

Praktické aplikácie automatického spracovania reči v Ústave informatiky SAV

Róbert Sabo – Sakhia Darjaa – Milan Rusko

Ústav informatiky, Slovenská akadémia vied, Bratislava, Slovensko

Abstract. The Department of Speech Analysis and Synthesis of the Institute of Informatics of the Slovak Academy of Sciences has been working in the area of basic and applied research in speech communication and automatic speech processing since 1989. In the early '90s the department developed the first software synthesizers in Slovak with low memory consumption. The recognizers using the DTW (Dynamic Time Warping) algorithm were able to recognize words from a 1000 word vocabulary. In the late nineties the diphone concatenative synthesizers Kempelen 1.0 were developed for Slovak. They were purchased by all three telephone operators in Slovakia in the beginning of the 21st century. One of them is still in use by Telecom in its SMS TO VOICE service. The following research led to the development of the Kempelen 2.0 Unit Selection synthesizer with high intelligibility and naturalness of the synthesized speech. Currently, the department is developing statistical parametric synthesizers Kempelen 3.0 aimed at modeling personality features of the synthetic speaker and the emotional characteristics such as urgency, warning or reassuring tone of voice. After the speech database SpeechDat E was designed, the department had enough speech material to start the experiments with statistical speech recognition and know-how for other speech database building. The department was then involved in the development of the speech operated information system "IRKR", system for automatic transcription of the parliament speeches, TV and radio broadcast and a dictation system for the Ministry of Justice, which has been installed to the computers of 1800 judges and assistants so far.

1 Úvod

Oddelenie analýzy a syntézy reči Ústavu informatiky Slovenskej akadémie vied sa od roku 1989 venuje základnému a aplikovanému výskumu v oblasti rečovej komunikácie a automatického spracovania reči. Začiatkom deväťdesiatych rokov tu boli vyvinuté prvé funkčné softvérové syntetizátory reči v slovenčine s mimoriadne nízkymi nárokmi na pamäť. Rozpoznávače reči, využívajúce ešte algoritmus DTW (Dynamic Time Warping), boli schopné rozpoznať slová zo slovníka, ktorý obsahoval 1 000 slov. Koncom deväťdesiatych rokov sa začala používať difónová konkatenatívna metóda. Na tomto princípe pracujú syntetizátory Kempelen 1.0, ktoré si začiatkom 21. storočia zakúpili všetci traja vtedajší telefónni operátori na Slovensku. Dodnes sa používa takýto syntetizátor v službe doručovania SMS správ na pevný linku (SMS TO VOICE) v sieti Telecom. Nasledoval ďalší výskum v oblasti syntézy a bol vyvinutý syntetizátor Kempelen 2.0, ktorý používal výber rečových jednotiek z veľkej databázy (Unit Selection) a ktorý dosahuje okrem vysokej zrozumiteľnosti aj veľkú prirodzenosť reči. V súčasnosti vyvíja oddelenie syntetizátory Kempelen

3.0 založené na princípe štatistickej parametrickej syntézy s dôrazom na modelovanie personality syntetického rečníka a emočných charakteristík ako nástojčivý a varovný alebo naopak upokojujúci tón reči. Po vytvorení rečovej databázy SpeechDat E získalo oddelenie jednak rečový materiál na trénovanie rozpoznávačov a jednak znalosti potrebné na tvorbu ďalších databáz. Nasledovala účasť na vývoji informačného systému ovládaného rečou – IRKR, systému na automatický prepis parlamentných debát, TV a rádiových relácií a diktačného systému pre Ministerstvo spravodlivosti SR, ktorý má už dnes nainštalovaný asi 1 800 sudcov a asistentov. Okrem syntézy a rozpoznávania reči predstavíme v článku aj webový slovník gest DiGest a vyučovací systém na kontrolu výslovnosti EURONOUNCE.

2 Slovník gest DiGest

Hlavnou motiváciou na vytvorenie slovníka gest DiGest (z angl. DIctionary of GESTures) bolo iniciovať a umožniť výskum gest, identifikovať problémy spojené s výskumom prejavov neverbálnej reči a jeho aspektov v rôznych kultúrach a modalitách a umožniť testovanie navrhovaných riešení týchto problémov (pozri Rusko, 2012). V súčasnosti sa výskumníci v oblasti syntézy a rozpoznávania reči snažia čo najlepšie do detailov opísať nielen verbálny, ale aj neverbálny prejav. Pri skúmaní neverbálneho prejavu, ktorý niekedy nesie väčšie množstvo informácie ako verbálny, je potrebné disponovať (podobne ako pri skúmaní verbálneho prejavu) dostatočným množstvom materiálu. Takýto materiál sa snaží poskytnúť slovník gest DiGest. Slovník umožňuje rýchly prístup k mnohovrstvovým informáciám o gestách a posunkoch a ich vzájomné porovnávanie v rôznych kultúrach. Technicky je to databázový systém, ktorý na ukladanie obsahu používa databázu MySQL a webové rozhranie napísané v PHP. Skúšobná verzia časti tohto slovníka je k dispozícii na `http://ui.sav.sk/gestures/`. Databáza je len čiastočne naplnená údajmi a má slúžiť na demonštráciu platformy.

Prvý základný súbor gest a ich opis v prvej verzii slovníka DiGest sme prevzali z publikácie E. Ružičkovej (2001). V súčasnosti sú všetky zaznamenané modality gest uvedených v slovníku hrané. Informácie o širšom kontexte gesta či posunku je možné v slovníku uviesť prostredníctvom dlhšieho obrazového alebo zvukového záznamu, prípadne spolu s anotačným súborom. Gestá a posunky sa považujú za prototypy (predstavujú triedu gest či posunkov) a ich záznamy vo všetkých modalitách ilustrujú ich základné komunikačné charakteristiky.

Všetky jazyky a kultúry uvedené v tomto slovníku majú mať prísne rovnaké postavenie. Museli sme však označiť jazyk popisu, ako aj používateľského rozhrania. Zvolili sme angličtinu, lebo je to jazyk medzinárodnej vedeckej komunikácie. Angličtina sa používa pre všetky javy nezávislé od jazyka a na navigáciu. Súčasná verzia obsahuje aj obsah závislý od jazyka a kultúry pre americkú angličtinu, slovenčinu, taliančinu a mongolčinu; na implementácii pre obsah v japončine, čínštine a maďarčine sa pracuje. Na obrázku 1 je súčasné rozhranie slovníka DiGest. V hornom obdĺžniku sú kultúrne nezávislé informácie, v dolnom kultúrne závislé. Naviac sú na obrázku v samostatných oknách otvorené dve fotografie a jeden zvukový súbor.

Obr. 1. Súčasné grafické rozhranie Slovníka DiGest

Slovník má slúžiť ako viacúrovňový a multimodálny výskumný nástroj na štúdium vzťahu medzi modalitami v používaní gest a ich potenciálnou implementáciou v automatizovaných aplikáciách spracovania reči, pri medzijazykovom a medzikultúrnom výskume a na sledovanie vzťahu medzi formou a komunikatívnou funkciou gest.

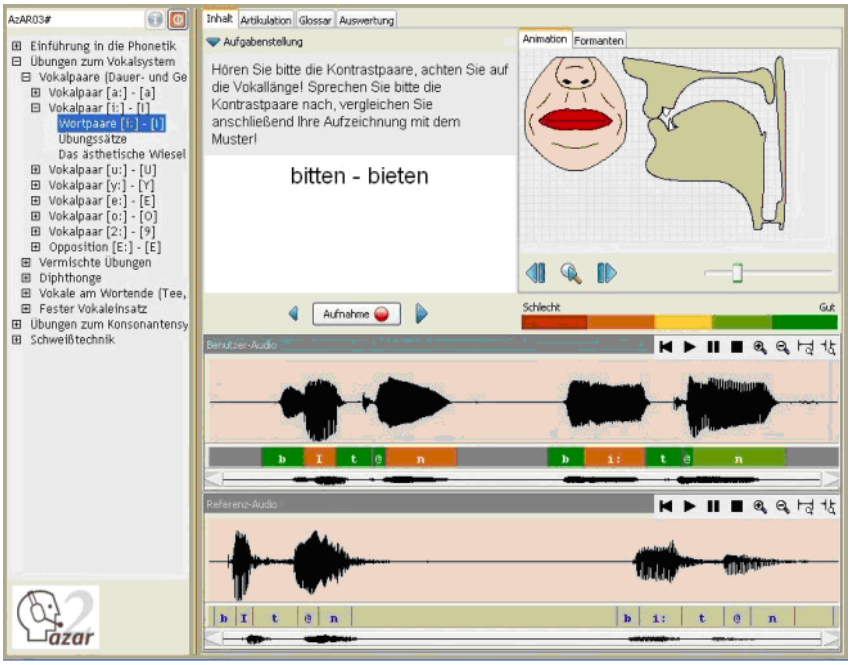
3 AZAR – vyučovací systém na kontrolu výslovnosti

Program AZAR (z nem. Automat zur Akzent Reduktion; Jokisch – Koloska – Hirschfeld – Hoffmann, 2005), ktorý si kladie za cieľ zlepšenie kvality výslovnosti a redukciu cudzieho prízvuku, bol vyvinutý v spolupráci so zahraničnými partnermi v rámci projektu EURONOUNCE. Ústav informatiky SAV sa podieľal na vytvorení jazykového páru nemčina – slovenčina, v jazykovej kombinácii L1SK – L2DE a L1DE – L2SK (pričom skratka L1 označuje materinský jazyk študenta). Každý výučbový softvér cudzieho jazyka zameriavajúci sa na trénovanie správnej výslovnosti pracuje na princípe porovnávania akustickej realizácie vzorovej (správne vyslovenej) nahrávky s akustickou realizáciou nahrávky študenta (používateľa softvéru). Sledované slovo alebo veta však môžu byť vyslovené hovoriacimi, ktorí majú rôznu farbu hlasu, rôzne tempo, rôznu hlasitosť a pod. Preto potrebujeme nielen jednu vzorovú realizáciu, ale aj model, ktorý bude obsahovať rôzne možné akustické realizácie. Rovnako je potrebné vybudovanie robustnej akustickej databázy. Pri výbere hovoriacich sme sa snažili o rovnomerné zastúpenie z hľadiska pohlavia, veku a regionálneho pôvodu (možný vplyv nárečia).

V rámci projektu EURONOUNCE (Jokisch et al., 2008) bola vybudovaná akustická databáza obsahujúca nahrávky od 16 hovoriacich v jazykovom páre L1SK – L2DE a 16 hovoriacich v jazykovom páre L1DE – L2SK. Databáza obsahovala prejavy študentov z každej znalostnej úrovne ovládania cudzieho jazyka (podľa Common European Framework of Reference for Languages úrovne A1 – C2). Nahrávky boli realizované v akusticky utlmenom priestore nahrávacieho štúdia. Pri budovaní programu a trénovaní akustických modelov sa nevyužívali iba nahrávky rodených hovoriacich, teda správna výslovnosť, ale analyzovali sa aj najčastejšie chyby študentov v jednotlivých úrovniach. Tomu bol prispôsobený napríklad aj výber testovacích viet do výučbového programu.

4 Výučbový program AZAR

Výučba správnej výslovnosti v cudzom jazyku prebieha prostredníctvom grafického používateľského rozhrania programu Azar (Jokisch – Koloska – Hirschfeld – Hoffmann, 2005). Ponúka úvod do fonetiky daného jazyka, ale predovšetkým cvičenia na nácvik správnej výslovnosti, ktorú si môže používateľ trénovať na kontrastívnych dvojiciach slov alebo na vetách. Pri každej vete alebo slove sa zobrazuje animácia pohybu artikulačných orgánov (obr. 2).



Obr. 2. Pracovné prostredie programu AZAR so zobrazením správnej polohy artikulačných orgánov

Vyslovenú realizáciu hlásky program vyhodnotí po porovnaní so správnym modelom realizácie hlásky získanej z rečovej databázy nahovorenej rodenými hovoriacimi a podľa miery akustickej podobnosti označí farebne mieru nesprávnej výslovnosti.

5 Syntéza reči v slovenčine

Výskum a vývoj prvej generácie softvérového syntetizátora reči (text to speech system – TTS) – Kempelen 0.1 – sa v Oddelení analýzy a syntézy reči Ústavu informatiky SAV začal v roku 1989. Vzhľadom na vtedajšie hardvérové obmedzenia bolo potrebné vyvinúť originálnu metódu kompresie. Neznelé spoluhlásky a hláskové prechody boli zachované celé, kým zo samohlások a znelých spoluhlások sa do pamäti uložili len počiatočné a záverečné dve periódy (mikrosegmenty) a jedna perióda zo strednej, ustálenej časti hlásky, ktorá sa pri syntéze opakovala v slučke. Opakovanie identického mikrosegmentu s rovnakou amplitúdou a samozrejme aj frekvenčným zložením viedlo k neprirodzene vysokej miere periodicity v týchto oblastiach a tým aj k „robotickej“ kvalite hlasu, avšak zrozumiteľnosť bola prijateľná.

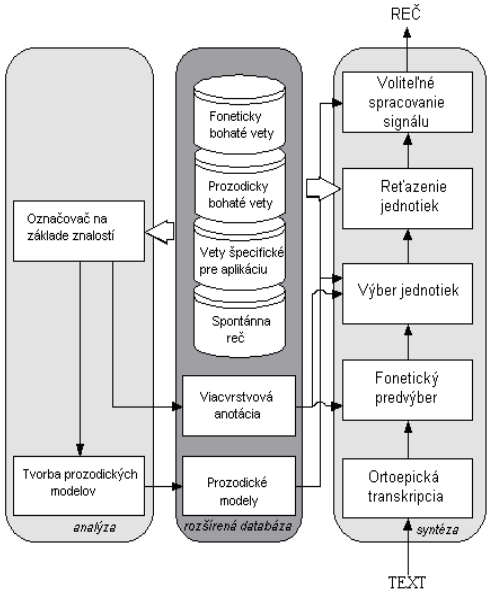
Výskum v oblasti difónovej syntézy a vývoj takéhoto konkatenatívneho syntetizátora začal na Slovensku približne v roku 1994. Táto generácia slovenského syntetizátora reči – Kempelen 1.x – bola založená na nadpájaní krátkych úsekov prednahratého rečového signálu, hlavne difón. Výslovnosť je v syntetizátoroch Kempelen 1.x riadená blokom ortograficko-ortoepickej transkripcie (blokom výslovnosti), založeným na prepracovanej sústave pravidiel, ktorá je doplnená slovníkom výslovnosti a zoznamom. Difónový syntetizátor Kempelen 1.0 priniesol syntetickú reč s lepšou zrozumiteľnosťou a vyššou prirodzenosťou spolu s prepracovaným používateľským rozhraním, ktoré otvorilo tomuto syntetizátoru cestu k širokému nasadeniu aj v profesionálnych telekomunikačných aplikáciách.

5.1 Unit-selection syntéza reči

Tretia generácia syntetizátorov Kempelen 2.x je založená na princípe výberu jednotiek rôznej dĺžky z rečovej databázy (tzv. Unit selection synthesis alebo aj Corpus-based speech synthesis). Hlavnou myšlienkou je mať k dispozícii čo najviac rečového materiálu, v ktorom sa môžu hľadať čo najdlhšie súvislé úseky, ktoré by bolo možné použiť pri syntéze reči podľa zadaného textu. Predpokladá sa, že na takýchto úsekoch sa prakticky nevyskytuje skreslenie a minimalizuje sa tak množstvo rušiacich zásahov do signálu. Výsledkom by mal byť signál s vysokou prirodzenosťou reči.

Vo všeobecnosti sa dá povedať, že za základnú jednotku syntézy bola v tomto systéme zvolená slabika. Ale keďže v databáze sú anotované aj hranice každej fonémy, je možné v prípade potreby (neexistencie vhodnej slabiky v databáze) použiť aj menšie jednotky.

Cieľom fonetického predvýberu elementov je vyhnúť sa nevhodným, respektíve problematickým bodom spojenia len na základe fonetických vedomostí o hláskach a ich spojeniach. Pozorne sa skúma pôvodný fonetický kontext spájaných hlások. Ak nie je v databáze k dispozícii hláska s identickým kontextom, hľadá sa v databáze element s kontextom patriacim do tej istej fonetickej kategórie ako požadovaný element. Úplne rozdielny fonetický kontext je prípustný len v najhoršom prípade, pretože spájanie elementov pochádzajúcich z nevhodného fonetického kontextu spôsobuje počuteľné nespojitosti timbru (farby zvuku), ako aj zhoršenie zrozumiteľnosti.



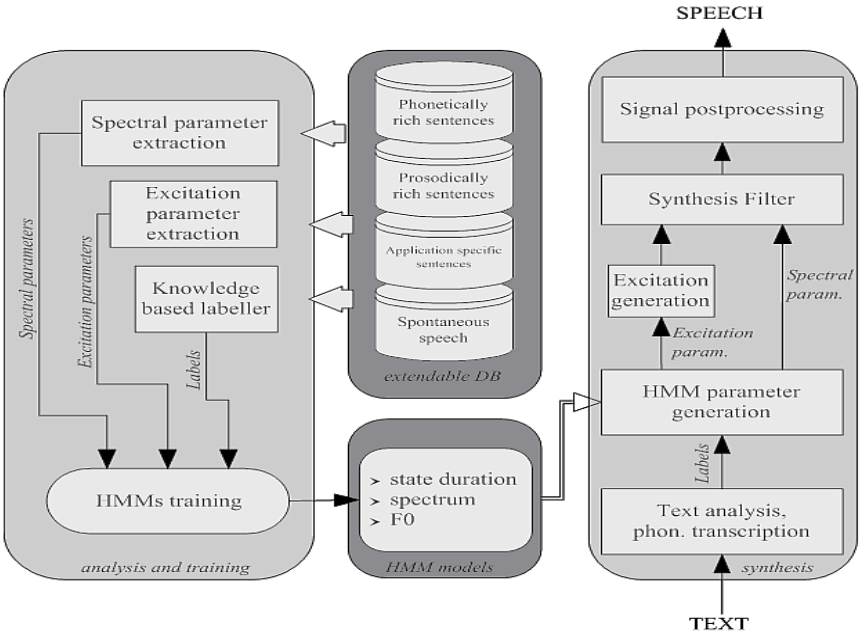
Obr. 3. Schéma syntetizátora Kempelen 2.1

Pri niektorých trifónach možno očakávať na ich centrálnej foneme takú silnú úroveň koartikulácie, že je prakticky nemožné správne určiť hranice takejto fonémy automatickým anotačným programom. Preto bol definovaný zoznam trifón, ktoré možno rozdeliť, len ak už neexistuje žiadna iná cesta syntézy. Typickými reprezentantmi sú kombinácie spoluhláska – samohláska – spoluhláska (VCV) so sonórami *l, r, j* alebo frikatívou *h* ako svojou centrálnou fonémou. Predvýber berie do úvahy aj polohu slabiky v slove (počiatočná, vnútorná a koncová) a polohu slova vo vete (koncové slovo vety).

5.2 HMM expresívna syntéza reči

Na generovanie expresívnej umelej reči sme zvolili prístup využívajúci skryté Markovove modely, tzv. HMM syntéza (Hidden Markov models). Vyvinuli sme HMM syntetizátor pre slovenčinu Kempelen 3.0 využívajúci trénovaciu procedúru (Yamagishi – Watts, 2010),

ktorá vytvorí tzv. priemerný hlas (Yamagishi, 2006). Trénovacia procedúra navrhnutá na využitie paralelného spracovania je založená na HTS (`http://hts.sp.nitech.ac.jp/`) a Sun Grid Engine (SGE) (`http://wikis.sun.com/display/GridEngine/Home`). Na obrázku 4 je znázornená schéma HMM syntetizátora Kempelen 3.0, ktorá využíva štatistické akustické modely a generuje parametre pre vokóder na generovanie reči.



Obr. 4. Schéma syntetizátora Kempelen 3.0

Na testovanie syntetizátora bolo vytvorené grafické rozhranie umožňujúce syntetizovať ľubovoľný zadaný text, pričom je možné meniť aj parametre syntetickej reči, napríklad výšku tónu alebo dĺžku hlások (tempo reči).

Existujú dva základné možné prístupy k tvorbe expresívneho syntetizátora s modelovaním pomocou skrytých Markovových modelov (Hidden Markov Models, HMM). Pri jednom sa na priamy tréning modelov používajú iba nahrávky s vysokým expresívnym nábojom (priama syntéza). Pri druhom prístupe sa neutrálny hlas prispôsobuje vyššiemu stupňu expresívnosti pomocou databázy expresívnej reči (adaptovaná syntéza).

Na pokusy v syntéze reči bol použitý systém HTS[1]. Na vývoj upraveného hlasu sa použila a posteriori lineárna regresia s obmedzeným štrukturálnym maximom (Constrained Structural Maximum A-Posteriori Linear Regression – CSMALPR; Nakano – Makoto – Yamagishi – Kobayashi, 2006) ako súčasná najnovšia technika úpravy HMM-TTS. Vyskúšali sme obidva prístupy a pri oboch sme získali veľmi sľubné výsledky. Napriek

1 `http://hts.sp.nitech.ac.jp/`

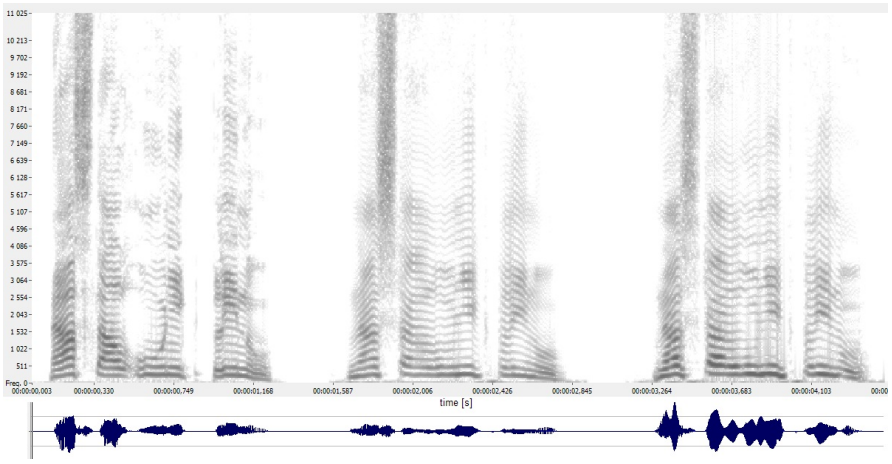
nízkemu počtu nahraných viet sme mohli natrénovať funkčný syntetizátor s vysokým stupňom expresívnosti. Podľa našich neformálnych percepčných testov si syntetizovaná reč veľmi dobre zachováva kvalitu hlasu, rytmus a výšku tónu zdrojových nahrávok.

5.3 HMM syntéza s nástojčivou emóciou

Syntéza s nástojčivou emóciou obsahuje tri stupne expresívnosti:

1. neutrálny,
2. vyššia naliehavosť ako vážny pokyn alebo príkaz,
3. extrémna naliehavosť, pokyn alebo vyhlásenie, aké sa používa v situácii priameho ohrozenia ľudského života (tretí stupeň expresívnosti).

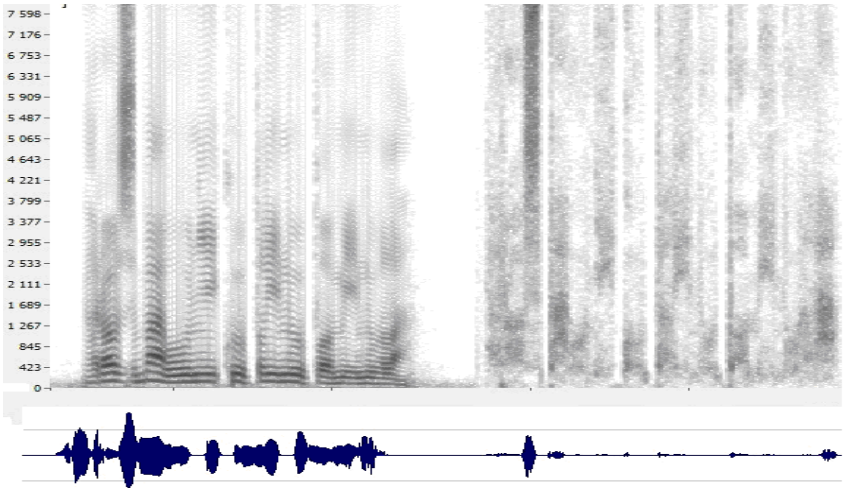
Na obrázku 5 vidíme výsledky syntézy expresívnej reči. Ide o spektrogram a oscilogram výstražného hlásenia *Hrozí únik plynu!* v treťom stupni expresívnosti, vysloveného človekom (vľavo), po upravenej syntéze (v strede) a po priamej syntéze (vpravo). Kým pri metóde priamej syntézy je zafarbenie a intonácia veľmi podobná expresívnej reči pôvodného hovoriaceho, zdá sa, že celková kvalita metódy upravenej syntézy je o niečo vyššia (signál obsahuje menej rušivých šumov). Pravdepodobným dôvodom je skutočnosť, že pre neutrálnu reč bolo k dispozícii oveľa viac trénovacích dát, ktoré sa pri predchádzajúcej metóde nepoužívali. Na druhej strane je pri upravenej syntéze o niečo nižší expresívny náboj.



Obr. 5. Spektrogram a oscilogram jednej prečítanej a dvoch syntetizovaných viet *Hrozí únik plynu!* s najvyšším stupňom expresívnosti

5.4 Syntéza šepotu

Prvá verzia HMM syntézy šepotu sa vyznačuje vysokou prirodzenosťou, keďže v trénovacom rečovom signáli sa nevyskytujú žiadne znelé časti, na ktorých by sa prejavoval hlavný nedostatok HMM syntézy – rušivé artefakty spôsobené vokóderom, ktoré sa prejavujú ako bzukot alebo brum (pozri obr. 6).

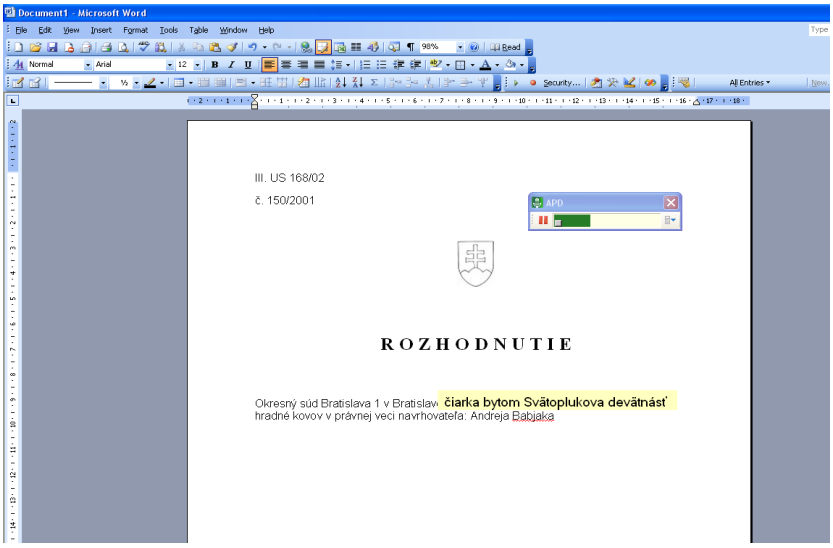


Obr. 6. Spektrogram a oscilogram jednej syntetizovanej vety Hrozba úniku plynu pominula! Zľava 1. neutrálna emócia, 2. šepot

Syntéza šepotu môže byť využitá pri poskytovaní informácií cez telefón v prípade, ak nie je žiadúce, aby bola syntetizovaná správa počutá ľuďmi nachádzajúcimi sa v okolí recipienta správy.

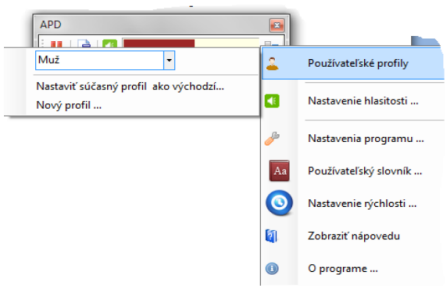
6 Automatické rozpoznávanie reči v slovenčine

V roku 2009 sa v oddelení analýzy a syntézy reči začal vývoj rozpoznávača reči pre Ministerstvo spravodlivosti Slovenskej republiky. V priebehu riešenia tejto úlohy bolo potrebné nazbierať stovky hodín rečových databáz a gigabajty textových databáz, z ktorých boli následne natrénované akustické a jazykové modely. Vytvorený program pod názvom APD, ktorý v súčasnosti využívajú slovenskí sudcovia, umožňuje hovoriacemu diktovať právne dokumenty priamo v prostredí programu MS Word (pozri obr. 7).



Obr. 7. Ukážka diktovania textu z oblasti judikatúry do textového editora MS Word

V menu používateľských profilov je možné zvoliť jeden z dvoch predinštalovaných profilov – muž, resp. žena, avšak natrénovaním vlastného používateľského profilu dosiahne používateľ rapídne zlepšenie funkčnosti systému (pozri obr. 8). Zároveň si program pamätá slová, ktoré boli pridané do slovníka, a všetky nastavenia, ktoré boli v rámci profilu vykonané.

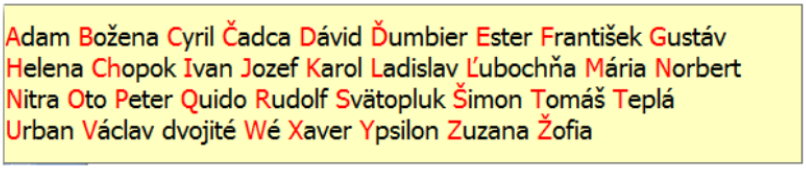


Obr. 8. Ukážka rolovacieho používateľského menu programu APD

Prvá záložka v menu nastavení ponúka nástroje na správu korekcií, resp. tzv. automatických zámen. Korekcie slúžia na nahradzovanie diktovaného textu počas písania napríklad skratkami a symbolmi – po vyslovení frázy „a tak ďalej“ sa do diktovaného dokumentu vpíše skratka atď. Na ilustráciu uvádzame niektoré ďalšie príkazy, ktoré dokumentujú aj vybrané funkcionality programu:

začiatok diktovania – spustí sa prepis,
koniec diktovania – pozastaví sa prepis,
alternatíva, alternatívy – zobrazí alternatívne hypotézy k poslednej fráze, vyberá sa vyslovením poradového čísla,
hláskovanie, hlásky, hláskovať – prepnutie na režim hláskovania,
naspäť – zoberie späť (zmaže) poslednú frázu,
opravujem, zmazať slovo – zmaže posledné slovo,
tučné písmo, tučné – zapne/vypne tučné písmo,
kurzíva, šikmé písmo – zapne/vypne písanie kurzívou,
podčiarkni – zapne/vypne podčiarkovanie,
zarovnaj na stred, do stredu – zarovnávanie textu na stred,
nadpis – zmení štýl na "Nadpis1",
dnešný dátum – vloží dnešný dátum,
pomlčka, bodka, výkričník, otáznik – vloží daný znak.

Režim hláskovanie slúži na diktovanie skratiek a hláskovanie. Zapína sa povelmi „hláskovanie“, „hlásky“, „hláskovať“. Hláskovanie prebieha formou diktovania krstných mien (pozri obr. 9), ktorých počiatočné písmena sú zapisované. Vypnúť režim hláskovania je možné povelom „koniec“.



Obr. 9. Ukážka zoznamu vlastných mien používaných v režime hláskovania. Tabuľka sa diktujúcemu automaticky zobrazí pri zapnutí módu diktovania.

Pri nastavení systému je možné ovplyvniť jeho rýchlosť a zároveň presnosť rozpoznávania. Pre pomalšie počítače je potrebné nastaviť vyššiu rýchlosť, pre rýchlejšie počítače je možné zvýšiť kvalitu rozpoznávania. Menu systému APD ponúka tiež položku *Používateľský slovník*, pomocou ktorého je možné pridávať nové slová, ktoré sa nenachádzajú v štandardnom slovníku, napríklad priezviská, názvy firiem a pod. Ďalšie funkcionality systému APD sú zhrnuté v dokumente opisujúcom používanie programu (`http://www.kempelen.sk/APD/NavodNaPouzitieAPD.pdf`).

Rozpoznávač reči vyvinutý v Ústave informatiky SAV úspešne využívajú aj tvorcovia portálu *Demagog.sk* pri prepise politických diskusií, pričom deklarujú značnú úsporu času pri prepise a vyhľadávaní kľúčových výrazov v jednotlivých diskusiách.

7 Záver

V článku sme uviedli praktické aplikácie automatického spracovania reči v Ústave informatiky SAV. Prezentovali sme nielen výsledky v oblasti rečovej syntézy a rozpoznávania reči, ale aj aplikácie z oblasti spracovania neverbálnych rečových prejavov – DiGest a odhaľovania cudzieho akcentu – program AZAR, na ktorého tvorbe sa oddelenie analýzy a syntézy reči podieľalo v rámci projektu EURONOUNCE (Jokisch et al., 2008). Medzi najvýznamnejšie aplikácie v oblasti rečovej syntézy patrí využitie difónového syntetizátora Kempelen v telekomunikáciách, ale aj v inteligentnom rečovom komunikačnom rozhraní IRKR. V oblasti rozpoznávania reči je najvýraznejším úspechom inštalácia programu APD (Automatický prepis diktátu) 1 800 sudcom na celom Slovensku, ktorí môžu program používať pri diktovaní súdnych rozhodnutí a zápisníc.

Oznámenie. Táto práca bola podporená z prostriedkov grantu VEGA číslo 2/0202/11.

Literatúra

JOKISCH, O. – JÄCKEL, R. – RUSKO, M. – DEMENKO, G. – CYLWIK, N. – RONZHIN, A. – HIRSCHFELD, D. – KOLOSKA, U. – HANISCH, L. – HOFFMANN, R.: The EURONOUNCE Project – an intelligent language tutoring system with multimodal feedback functions, roadmap and specification. In: Proc. ESSV Frankfurt/M. 2008, s. 116 – 123.

JOKISCH, O. – KOLOSKA, U. – HIRSCHFELD, D. – HOFFMANN, R.: Pronunciation Learning and Foreign Accent Reduction by an Audiovisual Feedback System. In: 1st International Conference on Affective Computing and Intelligent Interaction. Beijing: ACII 2005, s. 419 – 425.

NAKANO, Y. – MAKOTO, T. – YAMAGISHI, J. – KOBAYASHI, T.: Constrained Structural Maximum A Posteriori Linear Regression for Average-Voice-Based Speech Synthesis. In: Proc. of ICSLP'06, 2006.

RUSKO, Milan: Modelovanie prozodických javov v slovenčine. Dizertačná práca. Košice: 2012. (Rkp.)

RUŽIČKOVÁ, Eva: Picture dictionary of gestures (American, Slovak, Japanese, and Chinese). Bratislava: Comenius University Publishing House 2001. 195 s.

Slovník gest Digest. Dostupné z WWW: `http://ui.sav.sk/gestures`

YAMAGISHI, J. – WATTS, O.: The CSTR/EMIME HTS System for Blizzard Challenge 2010. In: Proc. Blizzard Challenge 2010.

YAMAGISHI, J.: Average-Voice-Based Speech Synthesis. Tokyo Institute of Technology 2006.

`http://hts.sp.nitech.ac.jp/` (downloaded: October 2011)

`http://rezo-computer.dyndns.org/euronounce/`

`http://wikis.sun.com/display/GridEngine/Home` (downloaded: October 2011)

`http://www.kempelen.sk/APD/NavodNaPouzitieAPD.pdf`

InterCorp: jeho povaha a možnosti

František Čermák

Filozofická fakulta, Univerzita Karlova v Praze, Česká republika

Abstract. The contribution is concerned with the nature, goals, shape, possibilities offered and contemporary state of the InterCorp Project, a parallel multilingual corpus built on over twenty languages which are linked through Czech texts in the Czech National Corpus. Over and over again, it becomes evident that the old need to compare languages in the context of modern times and needs is re-emerging in a new shape, bringing new possibilities. The major goal of the InterCorp Project is to link together a large number of languages, specifically on the basis of available modern translations, mostly from the field of fiction, and thus to enable their complex study, also utilizing the new software PARK which is being developed. After an up-to-date survey of the contemporary state and composition of the texts and their types that are already included (the Project goes on and constantly grows), a brief illustration is given of some possibilities of how to search this type of parallel corpus for the given word or collocation in rich contexts that is offered. This enables the user to set against each other more than two languages at the same time, if the corresponding translation texts are available, and thus to arrive at optimal equivalents. These examples are briefly commented on.

Further and by way of a mere illustration only, other options for quest of equivalents in the InterCorp are given. It is pointed out how problematic different views can be when regarding closely-related languages such as Slovak and Czech. On more languages (Czech, Polish, Russian, Finnish, German, Norwegian, English, Spanish, and French) possibilities of comparison of naming constructions on the typological background are shown, too.

The contribution winds up by deliberations on possibilities how, within this framework, to contribute both to applied and theoretical linguistics and how important it is to constitute comparative corpus linguistics, specifically in the areas of both closely-related and distant languages, though always firmly embedded in the background joint meaning.

1 Korpus: Jednojazyčný a vícejazyčný

Ten protiklad se může zdát normální, ale není. Záleží totiž na hledisku, které zvolíme. A necháme-li stranou aspekty historické a technické, je jasné, že o protiklad vlastně nejde. Ten první krok je totiž nutnou podmínkou pro vznik toho druhého. Bez více napřed jednojazyčných korpusů vybavených k sobě jinojazyčnou protiváhou, jsou paralelní korpusy pochopitelně nemyslitelné. Nicméně ještě před tím, než se pustíme do vlastního výkladu o InterCorpu, na místě snad bude pár obecných úvah, které je třeba mít na zřeteli.

Potřeba odstupu při úvaze o vlastním jazyku a projektu, který se ho týká, je zdravá, ba nutná. S určitým odstupem lze nejen získat určitou objektivnost, ale snad i dospět k jistým zobecněním, která jsou možná jen díky srovnávání s více jazyky. Lidská komunita zpravidla nikdy nežije v úplné izolaci a vždy jí obklopují jiné, které obvykle mluví jazyky jinými, jak blízkými jako je vztah komunity české a slovenské, či nikoliv, a jen v etymologicky dávné minulosti se mohlo stát, jako v případě polštiny, slovenštiny či češtiny, že se další soused metonymicky nazval v tomto vztahu *Němec, Niemec* či *Nemec* a jednoduše se mu tak upřel jazyk jakýkoliv, protože ho naši předkové měli za němého, jelikož vydával jim nesrozumitelné zvuky. Dneska víme už trochu víc: nesrozumitelné zvuky vydávají všichni mluvčí všech jiných jazyků, pokud se je nenaučíme, což vyžaduje určitou námahu. Avšak od respektování jazykové odlišnosti je jen krok ke (1) **srovnávání jazyků**, tj. k hledání podobností a odlišností, zvláště máme-li k takovému srovnání potřebu, ať už profesionální jako lingvisté, či čistě praktickou, dorozumívací. Je zřejmé, že první takové srovnání bývá vždy napřed bilingvní, jak nám napovídají vůbec první slovníky, které vznikaly mezi mluvčími jazyků v pravidelném styku, zvláště u jazyků sousedících (začátky je třeba hledat v Mezopotámii). Srovnávání jazyků však přineslo a přináší takové množství poznatků a vhledů, které mj. zdravě krotí přebujelý nacionalismus, že dalo vznik řadě oblastí zahrnujících jak na jedné straně praktickou lexikografii, tak na straně druhé typologii, univerzálie a obecnou lingvistiku vůbec. V praxi však toto poznávání mělo své jasné meze: bylo jen málo lidí, kteří kvalifikovaně znali více jazyků, jakkoliv znát pár sousedních jazyků se odjakživa chápalo za výhodu a dnes dokonce za slušnost, bylo ale především málo dat, z nichž se dalo vycházet, stejně tak jako jen málo opravdu multilingvních mluvčích, kteří by byli multilingvního výzkumu schopni. To ale poněkud předbíháme, protože obrovská většina srovnávání byla dvoujazyčná, a z něho bylo jak obtížné tak nebezpečné dělat širší závěry o čemkoliv.

Až příchod (2) **korpusů** dramaticky situaci změnil a dnešní velké korpusy, jako je Český národní korpus (přes 3 miliardy slov, ČNK), ještě větší německý v Mannheimu, ale i slovenský a další, prokázaly za relativně velmi krátkou dobu svou nezastupitelnou užitečnost pro veřejnost i jazykovědu. Dobrým příkladem je nezbytnost korpusu například pro probíhající tvorbu současného velkého slovenského slovníku. Je třeba nekompromisně trvat na tom, že každou informaci je možné a nutné hledat napřed v textech, a to díky kontextům, ve kterých se slovo vyskytuje. Jen z nich lze pak dospívat k potřebným a prokazatelným generalizacím. Korpusy se takto nabízenou informací maximálně podobají realnému světu komunikace kolem nás a jsou tak ve skutečnosti jeho nejlepším záznamem i aproximací, kterou máme. Ve srovnání s minulostí stará předkorpusová lingvistika nikdy neměla dost dat, resp. dost dat a jejich kontextů. Obecně lingvisticky pak je zřejmé, že v korpusovém pohledu kontexty, které jsou vytvářeny z kombinací slov, nás reorientují ze starého metodologického postoje prvek a jeho pozice (item-and-slot), resp. člen a jeho třída, tj. přístupu paradigmatického, k tolik potřebnému přístupu syntagmatickému, založenému na kombinacích a jejích typech.

V případě masívního nárůstu korpusů a možností, které dává, však neplatí, rozhlédneme-li se poněkud, že s jídlem roste chuť, jen se tak znovuobjevuje už i zmíněná potřeba podpořit starou tradici srovnávání jazyků pro studium nejrůznějších souvislostí z něj vyplývajících. Takové studium bez rostoucího (3) **paralelního korpusu**, v našem

případě specificky slovensko-českého a česko-slovenského korpusu, by vůbec nemohlo ani začít. Kombinuje jak možnost srovnávání tak výhody korpusu (1 a 2), není však jen jejich pouhou sumou. Je jasné, že zvýšený důraz, lingvistický i politický, na takovéto aktivity se dobře hodí do rámce toho, čemu se dnes říká globalizace, která by však neměla končit u pouhých proklamací, v lingvistice může mít velmi reálnou a dokonce potřebnou podobu.

Paralelní korpusy dnes existují pro mnohé páry jazyků, jakkoliv technologie jejich výstavby nikterak jednotná není, tím méně je takové i jejich skutečné využití. Od dob, kdy paralelní skutečně znamenalo bilingvní, se však pokročilo dále, zvláště díky čistě technologickým možnostem. Od těch dob, kdy jediným skutečně paralelním korpusem byla bible, odhlédneme-li od jejích různých překladů, se dnes pokročilo (přes mezifázi anglicko-francouzského parlamentního *Hansardu*) ke zdánlivě velkolepým možnostem nabízeným evropskými multilingválními soubory jako je *Europarl* (debaty evropského parlamentu, 11 jazyků, průměrně 50 mil. slov na jazyk; `http://www.statmt.org/europarl/`) nebo *The JRC-Acquis Multilingual Parallel Corpus* (22 jazyků, 636 milionů slov, zákonodárství EU, tzv. acquis communautaire; `http://langtech.jrc.it/JRC-Acquis.html`). Jejich obsahová stránka je však značně limitovaná a tím i odvozené možnosti praktického využití lingvistického.

Je tedy zřejmé, že nejde jen o prosté počty slov v dostupných překladech, ale především o *obsah textů*, resp. o to, jak velkou a jak obecnou část jazyka paralelní korpusy zachycují, a tedy o to, jaké jsou cíle tvorby paralelního korpusu a, řečeno ještě jinak, jaké jsou *potřeby, ale i možnosti* takového počínání. Odhlédneme-li od úzkého právnického využití, které nám oba velké evropské korpusy nabízejí, je zřejmě, že je třeba vážit skutečné potřeby, které zdaleka nejsou jen právnické. Dnešní situaci paralelních korpusů charakterizují především dvě věci. Na jedné straně existuje relativně hodně dvoujazyčných, v zásadě oportunních korpusů pro řadu jazyků, které jsou navzájem většinou inkompatibilní, na druhé straně vznikla už paralelně a vícekrát aspoň elementární technologie k jejich prohledávání. Komputační lingvisté se tomuto úkolu věnovali a věnují však jen do té míry, do jaké lze dělat věci automaticky (především pro strojový překlad), a to včetně metod alignmentu (zarovnávání) textů; stále však tu řada potřeb zůstává nepokrytá (viz Čermák – Rosen, 2012). V návaznosti na to tedy zůstává především na straně lingvistů úkol paralelní data shromažďovat a začít je využívat především v těch aspektech, které dosud realizovatelné nebyly. *Cíl srovnávat více jazyků* je v dnešní multilingvální Evropě velmi smysluplný, nejde však jen o příznivé politické ovzduší. Takové srovnávání může a musí v širším smyslu především naplnit a ospravedlnit staré diktum, že *jazyk je prostředek přenosu významu od myšlenky k formě*. A to se dá snadno doplnit o diktum další, totiž že *jazykové srovnávání je i mostem umožňujícím přenos významu i mezi jazyky* navzájem.

Paralelní korpusy jsou pochopitelně možné jen tam a tehdy, jsou-li dostupná podkladová, tj. překladová data mezi jazyky. Dostupnost dat je však často problém, který se nedá nijak řešit, pokud neexistují. Tento problém, který je pro jednojazyčný korpus jedním z mnoha, se v tomto případě stává zcela zásadní a primární. Pro ty jazyky, které nemají možnost využívat existence společného fondu překladů z krásné literatury (tj. překladů z nich nebo do nich), ba ani svou roli v mezinárodním kontextu (například jako jeden z oficiálních jazyků EU), se tato překážka stává zásadní natolik, že nebezpečně omezuje další růst paralelního korpusu, protože příslušná data prostě neexistují.

Cílem našeho příspěvku není věnovat se technickým aspektům výstavby paralelních korpusů, jakkoliv jsou hojné a netriviální, protože jsou vždy novátorské a realizované pro daný jazyk poprvé, a to jak v oblasti segmentace vět, tokenizace, zarovnávání textů (alignment), tak i lingvistické anotace. Každý jazyk tu vyžaduje specifické nástroje vyvinuté jen pro něj. V dalším se zaměříme proto především na obecnější aspekty netechnické.

Můžeme si pak snadno v dané situaci představit, že vznikající disciplína *srovnávací korpusová lingvistika* tu může najít dostatečnou a zásadní vzpruhu, pokud se multilingvní korpusy dostatečně rozrostou, budou usilovat o přijatelnou míru reprezentativnosti a zaměří svůj výzkum na aspekty skutečně multilingvální. Protože každé srovnání potřebuje své *tertium comparationis*, je zřejmé, že si lingvisté pak musejí být jisti tím, že srovnávání v takovém širším měřítku musí nalézt i společný širší metodologický rámec, který by měl být typologický. Ale realizace je tu dosud daleko.

2 Jazykové kontakty a překlad. Česká jazyková situace

Skutečnost, že jak bilingvní tak multilingvní korpusy jsou v zásadě podmíněny **jazykovými kontakty** a zakládají se tedy na dostupných překladech mezi jazyky a že jejich počet roste jen postupně, má své důvody i důsledky. Z kulturního a historického hlediska představuje úhrn dostupných překladů z jednoho jazyka do druhého sumu nejrůznějších nitek i proudů zájmu, ať už podmíněného dobově (jako v případě módních románů) či reálných a užitečných, které daná komunita měla a má po dané časové období, vázaných na komunitu jinou a její texty. To je hned nápadné, srovnáme-li takovou sumu přeloženého mezi dvěma malými jazyky, kam se dostalo do centra pozornosti často leccos užitečného či zajímavého, čeho si překladatelé všimli. Pokusíme-li se zobecnit, zdá se, že platí, že velikost průniku dostupných přeložených textů u více jazyků je nepřímo úměrná množství těchto jazyků; jinými slovy *počet textů sdílených mnoha jazyky se úměrně počtu jazyků snižuje*.

Takto lze nahlížet i kulturní, politické a další vlivy mezi komunitami, pokud studujeme počet, typ a rozšíření překladů mezi nimi v jejich úhrnu, a to nejen pro jeden jazyk a etnikum za ním, ale i pro komunitu větší a multilingvální, jakou je třeba Evropa. Jakkoliv existuje mnoho různých typů překladu *z* (a *do*) velkého jazyka (= *zdrojový jazyk*), jsou recipienty překladů ve většině případů, řečeno zjednodušeně, malé jazyky, tj. ty, do kterých se texty překládají (= *cílový jazyk*). To se promítá nepřímo i do skladby paralelních korpusů.

Za dané geopolitické situace se většina pozornosti, až na několik výjimek, upírá k paralelním korpusům, které se orientují na páry složené ze dvou velkých jazyků (jako je angličtina a francouzština v Hansard Corpus) či na takové páry, ve kterých je aspoň jeden z jazyků velký, jako je angličtina. Díky rozšířenosti angličtiny a některých dalších jazyků je ale také jasné, že páry dvou malých jazyků v tomto pohledu dost strádají. Přitom v kontrastu k všední praxi a jejím potřebám reálné lingvistické požadavky ukazují jinam, k potřebě srovnávání ve velkém měřítku a kvalifikovanějšímu studiu (všech druhů) jazyků i všech druhů textů. Proto je nutné, aby se rozumně shromažďovala srovnatelná data z co největšího počtu jazyků.

To platí i o **českém jazyce**. Je to slovanský jazyk, kterým mluví 10 milionů lidí, tedy jeden z těch malých jazyků. Jako typický flektivní jazyk má rysy, které se jen stěží najdou v angličtině, francouzštině, němčině či čínštině, jako je bohatá flexe o sedmi pádech, slovesný vid, volný slovosled, bohatá verbální prefixace, bohatá derivace substantiv, desubstantivní adjektiva (typ *vlakový*), množství partikulí aj., jakkoliv většinu těchto rysů má společných, ne však identických s ostatními slovanskými jazyky. Historicky jím mluví lidé ve středu Evropy, kde čeština byla vždycky jazykem na rozcestí v důsledku vlivu jiných, mezi něž patří především němčina či polština a slovenština a na druhé straně po několik desetiletí dočasně i nesousední ruština.

Čeština měla tradičně dva druhy těsných jazykových kontaktů se svými sousedy, slovanskými na jedné straně (slovenštinou a polštinou) a německým (rakouskou a německou němčinou), a oba představují nutnost ve výzkumu věnovat pozornost velmi odlišným problémům. Z nich představuje zvláště velmi dlouhý kontakt s němčinou nejen nutnost, ale i možnost, kterou lze využít zajímavěji, půjdeme-li hlouběji, za pouhé výpůjčky, totiž do sémantiky, kalků a vlivů na gramatický systém.

Všechny tyto faktory měly a mají svůj vliv, který se promítá do češtiny, v níž tento zajímavý vlivový souběh nabízí předmět jak specifického výzkumu, tak výzkumu obecného, především v typologickém rámci. Ten, obohacený o pohled zvenku, by měl být zajímavý už šíře a nejen pro české mluvčí a lingvisty, jakkoliv sem patří dnes i nejnovější vliv globální angličtiny. Odtud tedy představa a pozadí mnohojazyčného korpusu, majícího češtinu uprostřed mezi koncentricky navázanými dalšími jazyky, a tedy představa **InterCorpu**. Dodejme hned, že takto výlučné postavení češtiny nikterak neupozaďuje žádný další z participujících jazyků, které tu lze zkoumat také, dokonce i s vyloučením češtiny.

3 Projekt InterCorp a jeho povaha

Na rozdíl od jiných projektů (viz Čermák – Rosen, 2012) je InterCorp otevřený a usiluje stále o pokračující růst, kdekoliv je to možné, tj. všude tam, kde jsou k mání dostupné texty a finance k tomu potřebné. Jeho hlavní filozofie je stejná jako u velkého jednojazyčného korpusu ČNK: *v zásadě čím víc dat, tím lépe.* A protože česká data dostupná už jsou a byla k mání o něco dříve, je v zásadě třeba získávat jen nečeské překladové texty, ať už se najdou hotové (v elektronické podobě), nebo se naskenují a dodá se jim potřebná úprava; dnes se však podle potřeby získávají i některé texty, kde nová, resp. nově naskenovaná může být i čeština a zveřejněné jsou až v podobě v InterCorpu.

Seznam a počet jazyků vstupujících do InterCorpu je stále pragmaticky otevřený, tj. jediným ohledem je dostupnost textů; proto je i řada textů, které, protože jsou vždy teprve v procesu zpracování a čekají na své plné začlenění do systému a zveřejnění. Je jasné, že každý jazykový pár je odlišný (kromě společné češtiny) jak co do rozsahu, tak obsahu. Ukázalo se totiž, že původní představa, že existují překladové texty společné většině, ne-li všem relevantním jazykům, se dosud nepotvrdila; můžou však existovat texty, které dosud nebylo možné získat a zařadit.

Takový je tedy **obecný cíl** výstavby korpusu InterCorp i vlastní implementace projektu, který ho rámcuje. **Politika jeho výstavby** je jednoduchá a snad i skromná:

(1) Shromažďují se jen *současné texty*, vymezené tak, že nemají sahat před rok 1945 (i když můžou zahrnovat i texty starší, jsou-li vydané znovu po tomto datu). Tato časová hranice je stanovena vědomě: kromě klasické literatury začíná skutečná četba textů a tedy i současného jazykového úzu zhruba právě tady. Je to jinými slovy rozhodnutí nejen čistě praktické, ale i způsob, jak specifikovat distinkci synchronie – diachronie. Takováto hranice však pro některé jiné budovatele korpusů se nezdá zřejmě důležitá. Je to otevřený a obtížně řešitelný problém, který záleží na konkrétní situaci v jednotlivých jazycích, který je dobře vidět i tehdy, když zdrojový text je starší než překlad, který mohl vzniknout až po roce 1945, resp. jde o nový poválečný překlad. V takovém případě je zřejmě nejlepší prosté pragmatické řešení, které respektuje kvalitu a povahu textu.

(2) Jakkoliv by bylo ideální řešení dosáhnout u každého jazykového páru jistý druh *rovnováhy* co do počtů výchozích textů na každé straně, zůstává to z pragmatických důvodů zatím jen zbožným přáním a není ani jasné, má-li se o něj důsledně usilovat za každou cenu, tj. např. o stejný počet překladů *z* a *do* češtiny. I proto se to dosud nestalo kritériem určujícím výstavbu InterCorpu. Přesto může mít stejný počet původních textů na obou stranách páru pro jisté cíle své zřejmé přednosti.

(3) Kvůli očividnému nedostatku *textů sdílených více jazyky* bylo rozhodnuto, že do InterCorpu se začlení i některé texty, jejichž původním jazykem není ani jeden z jazyků daného páru, což je případ zvláště malých a nesousedních jazyků. Takové „třetí“ texty pak pocházejí zpravidla ze široce překládaných jazyků. Takto má povahu takového *třetího jazyka* např. v případě dostupného česko-srbského subkorpusu (na podzim 2011) 6 z 15 titulů v česko-srbské části, kdy jde zvl. o texty z angličtiny, ale i italštiny, polštiny, portugalštiny a ruštiny. Obecnou zásadou je mít přehled o častěji opakovaných překladech do různých jazyků, vybírat z něj a zajistit tak co nejširší vazbu na co nejvíce jazyků. Takto se dává přednost titulům překládaným do více jazyků. Toto rozhodnutí, tj. připustit v některých případech existenci neoriginálního jazyka na obou stranách jazykového páru, je třeba pak brát v úvahu při těch rozborech korespondence, kde na tom může záležet; obecně ho však odmítnout nelze, je to nouzový prostředek tam, kde jiná, přímá cesta není. Techniku a kritéria vyhodnocování relevantnosti tohoto druhu *nepřímé ekvivalence* při začlenění třetího jazyka je třeba teprve hledat, především v kontrastu k ekvivalenci přímé.

(4) *InterCorp* se snaží být lingvisticky co nejobecnější, aby mohl sloužit mnoha *různým cílům*: lingvistickým, nelingvistickým, akademickým, praktickým, výuce překladu aj. Proto je mj. důležité jím zachytit co nejvíce různých druhů a typů jazyka a lexikonu. Je však třeba připomenout, že vybudovat vyvážený paralelní korpus (dvoujazyčný i vícejazyčný) je mnohem těžší než korpus jednojazyčný. Důvody jsou aspoň čtyři.

(a) Některé textové typy, ale většina mluvených textů se překládají jen málokdy; sem patří většina případů novinářského jazyka, který je naopak tak důležitý v jednojazyčných korpusech. To je také důvod pro pragmatické řešení soustředit se na to, co dostupné je: proto se *InterCorp* skládá výlučně z textů *psaných*, jakkoliv, aspoň zatím teoreticky, nejsou mluvené texty vyloučené.

(b) Pokud jde o *prózu neliterární* a její převažující žánr, *žurnalistiku*, je k dispozici multilingvální zdroj (*Project Syndicate*, viz už výše, je to mezinárodní asociace novin vydávající komentáře a analýzy od předních autorů ovlivňujících veřejné mínění) a objevuje se i další slibný kandidát v tomto směru, *Presseeurope*[1], což je portál monitorující přední evropské deníky, který se v současnosti překládá do 10 jazyků včetně češtiny.

(c) Realizují se postupně snahy začlenit do InterCorpu více typů textu, zvláště z oblasti *specifického jazyka* parlamentních diskusí EU (*Europarl*[2]), dokumenty zákonů (*EUR-Lex*[3], *JRC-ACQUIS Multilingual Parallel Corpus*[4]), o kterých už byla řeč také výše, či dále různé otevřené zdroje technické literatury a softwarových manuálů (jako je *OPUS, Open Source Parallel Corpus*[5]) atd., jakkoliv jde o jazyk jen úzce zaměřený.

Volba takovýchto textů je jen pragmatická a záleží na jejich (A) existenci, (B) dostupnosti a (C) *legálních předpisech*, které jejich dostupnost regulují. To jsou však obecné otázky týkající se všech druhů textů. Rozhodně ale platí, že uživatel korpusu si vždy může svobodně vybrat určité texty a studovat či užívat je podle svých potřeb a zájmu a jen podle nutnosti může být jeho přístup omezen, například dalším heslem ap. Kvůli pragmatické povaze projektu InterCorp je obtížné plánovat nějakou definitivní podobu korpusu, který se do značné míry stále mění, resp. roste.

(d) V InterCorpu však kvantitativně převládá *nespecifický jazyk,* především beletristický, který se vnímá jako priorita zaměřená primárně na maximálně pokrytí základního lexikonu, pokud je dostupný, tj. toho, který je důležitější a obecnější než zmíněné specifické typy jazyka, protože je v komunikaci univerzální.

Takovéto teoretické a praktické ohledy lze tedy hledat za ideou velkého multilingválního korpusu s češtinou v centru. *InterCorp*[6] je součástí širšího projektu *Český národní korpus* (ČNK)[7]. Vlastní centrální představa v jádru InterCorpu je tedy lingvisticky triviální, jakkoliv se neozývá často; mít vlastní jazyk bohatě pokrytý jednojazyčným korpusem, tj. zevnitř, nemusí stačit: *jazyk se musí také studovat zvenčí, prizmatem druhých jazyků.*

[1] http://www.presseurop.eu
[2] http://www.europarl.europa.eu/
[3] http://eur-lex.europa.eu
[4] http://wt.jrc.it/lt/Acquis/
[5] http://urd.let.rug.nl/tiedeman/OPUS/
[6] http://korpus.cz/intercorp
[7] http://korpus.cz

Tento projekt je nepochybně jedinečný co do svého rozsahu, volby textů (jakkoliv dosud převládá beletrie), ale i podstatného vkladu manuální práce mnoha lidí (při narůstající kvalitě zarovnání, identifikace hranic věty a zmenšování počtu chyb). Účastníci projektu, pozvaní ke spolupráci v r. 2005 do týmu vedeného ústavem *Českého národního korpusu FFUK*, pocházejí z většiny jazykových kateder, ústavů a jejich oddělení filozofické fakulty Univerzity Karlovy v Praze a několika dalších akademických institucí, mezi které patří i spolupracovníci ze Slovenska, Polska aj., patří k nim však v neposlední řadě také početní studentští pomocníci. Současný stav zpracovávaných jazyků je 25 (plus čeština), online zveřejněných jazyků je 23 (a tedy už víc než má acquis communautaire). V současnosti se ke zpracování i studiu užívá paralelní konkordanční program PARK (= *par*alelní *k*orpus)[8], vyvinutý a stále zlepšovaný na FFUK. K InterCorpu se může zdarma přihlásit každý registrovaný uživatel ČNK[9].

Níže uvedená **tabulka** nabízí v přehledu čísla pro jednotlivé jazyky platné pro současnou verzi korpusu (čísla pro češtinu jsou vysoká proto, že se česká data v jednotlivých jazykových párech opakují). „Titulem" se zde míní především román jako dominantní typ textu, který je v InterCorpu zastoupený nejvíc. Některé jazyky však mají výhodu větší vyváženosti textů, která je zřejmější u většího počtu textů.

Písmeno *S* (± Syndicate) v jednom sloupci upozorňuje na počet titulů pocházejících z politických komentářů korpusu *Project Syndicate* (viz výše). V současnosti jeho dostupná vydání obsahují data česká, anglická, francouzská, německá, ruská a španělská z let 2000 – 2008, k nimž brzo přibudou další z nových vydání, zvláště arabská a čínská, ale taky albánská, katalánská, hindská aj.; čísla z tohoto korpusu jsou započítána do celkových čísel a jejich rozsah se pohybuje mezi 1,5 – 2 milióny slov pro daný jazyk. Celkový přehled stavu InterCorpu tedy v současnosti představuje data z března 2010 (čísla jsou udána v tisícovkách):

Jazyk (L2)	**Slovní tokeny v češtině** × 1000	**Tokeny v jiném jazyce (L2)** × 1000	Počet titulů
angličtina	4,041	4,705	S + 34
bulharština	1,057	1,049	14
dánština	80	102	4
finština	497	423	11
francouzština	2,415	3,120	S + 21
chorvatština	4,363	4,599	69
italština	2,254	2,591	26
maďarština	1,030	985	15
němčina	6,466	7,480	S + 70
nizozemština	2,448	2,046	45
litevština	318	272	7
lotyština	1,121	1,067	23
polština	2,450	2,422	40

[8] `http://korpus.cz/Park`

[9] `http://korpus.cz/english/dohody.php`

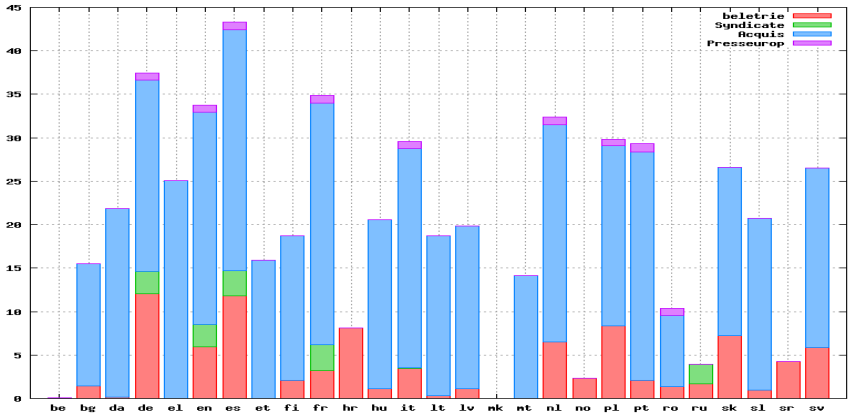
Jazyk (L2)	**Slovní tokeny v češtině** × 1000	**Tokeny v jiném jazyce (L2)** × 1000	Počet titulů
portugalština	1,261	1,436	18
rumunština	461	564	4
ruština	2,873	2,902	S + 23
slovenština	352	351	7
slovinština	813	901	15
srbština	1,129	1,209	19
španělština	7,210	8,427	S + 82
švédština	1,439	1,643	25
Celkem	44,077	49,293	572

Každý z jazykových párů je zjevně odlišný, liší se rozsahem i obsahem, což je, jak se ukazuje, přirozené a odpovídá to skutečné distribuci překladů v různých jazycích. Podle stavu dat z loňského roku je nejčastěji zastoupený titul M. Kundery *Nesnesitelná lehkost bytí* (v 9 jazycích včetně češtiny) a dalších pár (aspoň 7) je ve zpracování. Kunderův román *Žert* je zastoupený v 18 případech a po nich následuje J. K. Rowlingová se svým *Harrym Potterem a kamenem mudrců* (*Harry Potter and the Philosopher's Stone*, 14×) a J. R. R. Tolkienův *Pán prstenů* (*Lord of the Rings*, 12x). I tady jsou některé verze dosud ve stadiu zpracování a zpřístupní se časem.

K aspoň stručné demonstraci poměrně značného pokroku od jara 2010 (viz tabulku výše) si uveďme aspoň zběžně novější (plná) čísla reprezentující stav pro 22 jazyků mimo češtinu (stojící v pozadí) z května 2012 (opět v tisícovkách slov/tokenů, včetně neliterárních textů); uvádějí se napřed počty slov z vlastních knižních textů a v závorce, uvozené pomocí *vč A-*, všechny počty slov všech textů, tj. včetně textů z Acquis communautaire, jejichž kvalita i hodnota je ovšem omezená, resp. nižší.

- *angličtina 9 229 (vč A-35 031), bulharština 1 466 (vč A-16 356,) dánština 190 (vč A-23 362),*
- *finština 2 140 (vč A-20 054), francouzština 6 939 (vč A-34 682), chorvatština 8 207 (vč A-8 207), italština 4 374 (vč A-30 297), litevština 358 (vč A-20 445), lotyština 1 116 (vč A-21 526),*
- *maďarština 1 249 (vč A-21 926), němčina 16 355 (vč A-39 998), nizozemština 7 426 (vč A-33 857), norština 2 335 (vč A-2 335), polština 9 264 (vč A-31 554), portugalština 2 959 (vč A-30 220), rumunština 2 162 (vč A-10 852), ruština 4 162 (vč A-4 162), slovenština 7 408 (vč A-28 270), slovinština 1 015 (vč A-22 161), srbština 4 351(vč A-4 351), španělština 15 832 (vč A-45 198), švédština 5 997 (vč A-27 971).*

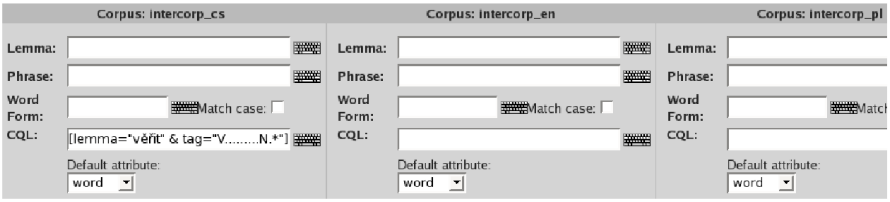
Graficky zachycuje vnitřní poměry čtyř obecných typů textů následující přehled pro 25 jazyků (včetně některých, dosud periferních „nových“):



4 Užití InterCorpu, studium jeho výsledků a důsledky

InterCorp se rozvíjí postupně dál, a to se týká nejen nárůstu a povahy jeho textů, ale i obslužného softwaru a dodávané lingvistické informace; bibliografické údaje jsou uloženy ve zvláštní databázi (`http://korpus.cz/intercorp/?req=page:info`). Software PARK (autorem je M. Štourač), fungující na pozadí korpusového manažeru Manatee (P. Rychlý) umožňuje prohledávání více jazyků zároveň a tolik textů, kolik je třeba: volí si je badatel podle nabídnutého seznamu dostupných paralelních textů, a to buď jen v jednom směru, např. slovensko-českém, směru opačném či bez rozdílu a směru. Hledat je možné, jako v každém větším korpusu a jeho manažeru, tvary i lemmata (lemmatizovaná a morfologicky značkovaná je však zatím jen část textů), popř. i kombinace či výsledky zadané výrazem v CQL (pokud je k dispozici značkování), kdy se ve výsledku zadaná forma barevně označí jen ve zdrojovém jazyce a v cílovém jazyce či jazycích (může jich být víc) se objeví jen automaticky vybraný odpovídající kontext, v němž uživatel musí sám rozhodnout v nabídnutém kontextu, co považuje za ekvivalent; výsledky je možné prohlížet v konkordanci horizontálně nebo vertikálně a listovat v nich po stránkách. Podle potřeby lze nabídnutý automatický výsledek rozšířit na potřebný větší kontext. Výsledky lze ukládat do formátu spreadsheetu (Excel), kde je lze podle potřeby i řadit.

Dotazovací tabulka, která se objeví po zadání jazyků a textů, s nimiž chceme pracovat, ukazuje příklad (ve starší verzi dat InterCorpu) na hledání českého tvaru verba *věřit* (viz zadané *word*) v korpusu anglickém a polském (zadání se pro ilustraci zde uvádí i jazyce CQL). Ekvivalenty se hledají, opět pro ilustraci, jen v Kunderově románu *Nesnesitelná lehkost bytí* a *Žert* a jeho ekvivalentech (překladech) do angličtiny, polštiny a ruštiny. Počet cílových jazyků lze ovšem libovolně zvětšovat. Srov.



Z výsledků se zde uvádějí pouze první tři nálezy. Počet tokenů (tvarů) zaznamenává software a uvádí nahoře za jménem korpusu (např. intercorp_pl (166 365 tokens) uvádí, že v polském odpovídajícím subkorpusu se našlo 166 365 odpovídajících, resp. kandidátních tvarů). Tabulka uvádí konkordanci v horizontálním zobrazení s výsledky ve sloupcích vedle sebe (lze zvolit i vertikální zobrazení ukazující každý jazyk na zvláštním řádku pod sebou).

intercorp_cs *(160320 tokens)*	intercorp_en *(192071 tokens)*	intercorp_pl *(166365 tokens)*	intercorp_ru *(162671 tokens)*
Show options Kwic			
Nevěřit (ustavičně a systematicky , bez chvile zaváhání) si vyžaduje obrovského úsilí a také tréninku , to jest častých policejních výslechů .	show context Maintaining non-belief (constantly , systematically , without the slightest vacillation) requires a tremendous effort and the proper training - in other words , frequent police interrogations .	show context Nieufność (ciągła i systematyczna , bez chwili wahania) wymaga ogromnego wysiłku i treningu , to znaczy częstych przesłuchań policyjnych .	show context Неверие (постоянное и систематическое , без тени колебания) требует колоссального усилия и тренировки , иными словами , частых полицейских допросов .
Nebylo možno nevěřit jeho upřímnému hlasu .	show context There was no doubting that forthright voice of his .	show context Nie można było nie wierzyć szczerości jego głosu .	show context В искренности его голоса сомневаться было нельзя .
Řekla jim , že to ví , ale že nevěřila , že by soudruh Jahn …	show context She said yes , she knew , but she would never have believed that Comrade Jahn …	show context Powiedziała , że wie , ale że nie wierzyła , żeby towarzysz Jahn … Spytali , czy dobrze mnie zna .	show context Она сказала им , что знает , но не могла бы поверить , что товарищ Ян …

Je nicméně důležité připomenout, že InterCorp lze nastavit libovolně jako *jakoukoliv konstelaci jazykových párů* či trojic, čtveřic, pětic jazyků aj., v nichž ani nemusí být explicitně uvedená čeština a výsledky v ní. Jednotlivé jazyky lze dokonce „odpoutat“ od ostatních a studovat je přímo v rámci větších možností, které nabízí manažer Bonito.

Uveďme si pár dalších příkladů, tentokráte už přímo lingvisticky zaměřených, které se pokusí ukázat jak možnosti hledání, tak problémy, které nalezené výsledky můžou přinášet. Podívejme se na **lexikální ekvivalenci**, a to napřed v podobě velmi jednoduchého výchozího českého lemmatu *stůl*, tedy slova v podstatě monosémního a na jeho ekvivalenty v angličtině a italštině, a to na M. Kunderově textu *Nesmrtelnost* a J. K. Rowlingové *Harry Potter and the Philosopher's Stone* (H. P. a kámen mudrců), kdy jeden výchozí text je v češtině a druhý v angličtině. Ilustruje se tu případ typické **jedno-víceznačné ekvivalence.**

Ke všem 100 výskytům českého lemmatu *stůl* dostáváme v angličtině 89 ekvivalentů (tj. 89 %) v podobě *table*, 5 jako *desk*, 1 jako *desktop* a 5 je případů, kdy se nenabízí *žádný ekvivalent* (tj. 5 %). Italské výsledky jsou pestřejší a nabízejí 83 ekvivalentů v podobě základního *tavola* (83 %), 3 jako *banco*, dvakrát *cattedra*, ale taky jednou *banchetto,*

scrivania a *scrittoio*, zatímco *žádný ekvivalent* se nedal najít v 9 případech (9 %). Na první pohled se výsledky můžou zdát poměrně jednoduché a přímočaré, ale nápadně vysoké procento žádných ekvivalentů, které je v italštině vyšší než v angličtině, nutí k zamyšlení. K nalezení odpovědi se podívejme na dva příklady, které můžou naznačovat dvě rozdílné možnosti. Prvním je česko-anglický případ založený na předpokladu, resp. implikaci (protože operace se provádějí na *stole*, a proto se *stůl* vůbec nemusí zmiňovat); druhý česko/anglicko-italský případ je vědomé vynechání anglického *table*, jakkoliv bezprostřední kontext k tomu nenabízí žádnou záminku.

Uveďme si konkrétní příklady s kontexty.

CZ *při nějaké nevinné operaci zemřela na operačním* ***stole*** *mladá pacientka kvůli nedbale provedenému uspání*
ENG *a young woman who in the course of a completely minor operation died because of carelessly administered anaesthetic*

CZ *Hagrid se k němu naklonil přes* ***stůl****.*
ENG *Hagrid leaned across the* ***table****.*
IT *Hagrid si chinò verso di lui.*

Nicméně důležitější je zřejmě se zaměřit na **rozmanitost ekvivalentů** k prostému českému *stolu*, které se tu nabízejí, totiž na 3 pozitivní ekvivalenty v angličtině a 6 v italštině. Zaměříme-li se v důsledku jemnějšího studia jejich kontextů na rozdíly v jejich zřejmě **komplementární distribuci**, máme před sebou velmi úrodnou půdu, z níž lze mj. čerpat možnosti vylepšování ne vždy vyhovujících slovníků, máme-li se omezit jen na zcela praktické využití. Taková možnost i potřeba je zjevná už z toho faktu, že většina italských ekvivalentů s velmi nízkou frekvencí výskytu se často nedá ve slovníku najít.

Výše už bylo naznačeno, že zkoumání paralelního korpusu **blízkých jazyků** má zvláštní cenu, pohlédneme-li na něj poněkud jemněji, jakoby z hloubky. Takovým případem jsou i čeština a slovenština, kde jen na rovině lexikální

(1) vedle *identifikace jasných rozdílů*, tj. většinou formálně jasných (banální *borůvky – čučoriedky*, polysémní *les – hora*, či jemněji a jen fonologicky *zájem – záujem* aj.), hrozí i

(2) *bagatelizace jemných* a často opomíjených *rozdílů*, jakási nebezpečná supergeneralizace jen zdánlivě jasných vzájemných korespondencí.

Takovým případem je do velmi vysoké míry korespondence polyfunkční formy *ale – ale* v obou jazycích (většinou konjunkce, často ale i partikule). Nahlédneme-li do česko-slovenského korpusu (majícího v dané době už pozoruhodný počet 132 textů) a podíváme se jen letmo, na pár příkladech několika textů, na slovenské korespondence českého *ale*, uvidíme, že obrovská většina slovenských ekvivalentů českého *ale* tradiční intuitivní představu jednoduchého vztahu podporuje, a tedy také *ale*. Nicméně pár příkladů nás musí znepokojit. Jednak jde o rozdíl ve variabilitě slovenských ekvivalentů *lenže, alebo, hoci, no* aj., které nejsou zase tak okrajové, a jednak o specifické odchylky v partikulární funkci této formy, srov. *Ale co vás to napadá!* a slovenský ekvivalent *Čo vám to zišlo na um!*, kde se zdá, že další možnosti ve slovenštině blokuje zároveň to, že *Čo vám to zišlo na um!* je v tomto smyslu neměnný frazém vyjadřující integrálně modalitu výchozího českého *ale*.

Takové příklady jsou jen malou ilustrací, která však chce naznačit, že je třeba je důkladně zkoumat. Lze se pak nadát, že dosavadní slovníky mezi oběma jazyky, které nejsou nikterak optimální, nahradí slovníky nové, kde se už k datům v InterCorpu bude přihlížet a dospěje se k lepším ekvivalentům s jasnou informací o jejich distribuci.

Aspoň jednu výše připomínanou možnost, kterou **multilingválnost** přináší, si přiblížíme na příkladu, do kterého se výrazně promítá už i zmíněná **typologie jazyků**. Na příkladech z devíti jazyků (na datech Kunderova *Žertu*) si lze ukázat, jak se v těchto jazycích tvoří substantivní pojmenovánání několika typů. Jde o jazyky někdy typologicky vyhraněnější (*flektivní* čeština, polština a ruština a *aglutinační* finština, a sledovaným rysem výrazně i němčina a norština) i méně vyhraněné (smíšené, zvl. francouzština a španělština), zvlášť stojí *izololační* angličtina. I když všechny jazyky jsou typologicky smíšené (obv. S převažujícím jedním rysem), pro jednoduchost se tu kromě flexe, aglutinace a izolativnosti další rysy neuvádějí (srov. VAR, varianty), i když např. španělština je z jiného hlediska flektivní taky aj. srov.

		Prosté	Derivace	Kompozice	Kolokace	Kolokace
FL	**češ**	*hodiny*	*hodinky*	*0/běžící pás*	*nákladní auto*	*toaletní stolek*
	pol	*zegar*	*zegarek*	*0/ruchoma taśma*	*0/ciężarówka*	*mały stół, toaletka*
	ruš	*časy*	*0/časy*	*0/konvejer*	*0/gruzovik*	*tualetnyj stolik*
AGL	**fin**	*kello*	*0/kello*	*liukuhihna*	*0/kuormaauto*	*pieni pöytä*
	něm	*Uhr*	*(Armband)Uhr*	*Fließband*	*0/Lastauto*	*0/Toilettetisch*
	nor	*ur*	*0/ur*	*transportbåndet*	*0/lastebil*	*0/toalettbord*
IZOL	**angl**	*clock*	*0/watch*	*0/production line*	*0/truck*	*small table*
VAR	**špan**	*reloj*	*0/reloj*	*0/silla del peluquero*	*0/camion*	*mesa pequeña*
	fr	*montre*	*0/pendule*	*0/chaîne*	*(voiture de tourisme)*	*petite table*

První dva sloupce uvádějí napřed případ prostý, resp. nederivovaný lexém (i když v češtině je to díky plurálu poněkud méně jasné) a pak lexém derivovaný. Kromě flektivní češtiny a polštiny se všude jinde neliší *hodiny* a *hodinky*. Třetí až pátý sloupec ukazují konkurenci kolokací a kompozit, zase ale jen v neflektivních jazycích (tam je všude jen kolokace); ruština tu není typická. Nicméně je jasně vidět souvislost pojmenování pomocí kolokace a izolačního typu v angličtině, i když i zde je jednoslovná výjimka (*truck*); výjimky lze však nalézt i u francouzštiny a španělštiny a jejich povahu a status může zpřesnit jen obsáhlejší analýza.

I zde musíme odhlédnout od překladatelských idiosynkratičností, kde např. ve fr. je chybné *voiture de tourisme*, nebo tam, kde překladatel neuvádí existující možnost (opět ve fr. *table de chevet*, popř. *coiffeuse*) a spokojí se s prostým deskriptivním *petite table* aj.

V širším smyslu takováto ilustrace ukazuje také, *jaký typ výsledků* jednoho druhu se pro jednotlivé jazyky dá z výzkumu čekat.

5 Výzkum a jeho možnosti

InterCorp může být a už i je užitečným zdrojem poznání, o čemž svědčí i první publikované výsledky; některé nalezly své místo v publikovaných sbornících na konferencích 2009 a 2011 (Čermák – Klégr – Corness, 2010; Čermák – Kocek, 2010; Čermák, 2011).

I když tu jde obecně o mnoho možností, výzkum multilingválního korpusu lze zjednodušeně chápat jako dvojího druhu (A) aplikovaný a (B) teoretický.

Aplikovaný výzkum (**A**) bude záviset na skutečné poptávce a mohl by být tradičně propojený hlavně s překladovými studiemi a lexikografií (Teubert, 2001; Text Corpora and Multilingual Lexicography, 2007). Specificky zajímavou možností se tu jeví studium problémů interpretace téhož textu ve více různých překladech, pokud budou k dispozici. Každý překlad je třeba chápat jako idiosynkratický mj. v tom, že zachycuje vždy jen část významu výchozího textu, čímž dospíváme ke staronové otázce, co se vlastně v překladu obvykle či vždycky ztrácí aj.

Jakkoliv se multilingvální lexikografie momentálně velké popularitě netěší (na rozdíl od terminologie, srov. *Eurodicautom,* resp. *IATE*), situace se může změnit. Uveďme jen, že by např. bylo užitečné mít slovník blízce příbuzných jazyků, jako je čeština, slovenština a polština, skandinávských či jižních románských jazyků, které by se daly užívat pro kontrolu falešných přátel, resp. mezijazykových homonym aj.

Praktický význam lze rozhodně hledat i v oblasti strojového překladu, automatického text-mining, automatické disambiguace aj.

Teoretický výzkum (**B**) může v prohloubeném srovnávání jazyků nabídnout vhledy i do oblastí nových či jen málo zkoumaných. Nicméně jde především o **srovnávací korpusovou lingvistiku**, kterou je třeba začít pěstovat, kde multilingvální korpus může přijít k užitku tím, že nabídne lepší data obecné lingvistice, typologii, pragmatice a minimálně i studiu diskurzu.

V takovémto rámci se může objevit řada obecných témat a otázek. Takto si žádá např. stará a dosud jen obecně formulovaná otázka *jazykové příbuznosti* přesnějšího poznání a nutnosti pracovat jak s menšími tak většími skupinami jazyků. Naproti tomu stojí zdánlivě nekonečná *různost* zvl. nepříbuzných jazyků, až dosud pokrývaná v podstatě jen typologií a univerzáliemi, která by si zasloužila jakékoliv smysluplné zpřesnění, zvláště např. V podloženém navržení *typologie rozdílů*, pokud je to vůbec možné.

Silnou stránkou tradičního většinového studia monolingvních korpusů je vždy nesporná opora v autentických textech a skutečných kontextech. Naproti tomu však studium bilingvních a multilingvních korpusů je odlišné v tom, že překlady prostě nejsou originální, autentické texty (a tedy vlastně ani ne přeložené kontexty, o čemž se málo mluví). Nabízí se tedy potřeba *metodologie hodnocení překladových protějšků*.

Je zřejmé, že s pohybem vzhůru, od prostých lexikálních jednotek skrze kolokace ke větám a jejich kombinacím se hodnota každého takového postupu musí nutně stávat problematičtější a čím dál interpretačně otevřenější, ba dokonce někdy spornější. Nicméně uvědomíme-li si, že základem a *východiskem všude musí zůstat význam*, zdá se, že velmi zajímavé výsledky se mají hledat mj. spíše v oněch vyšších rovinách než v těch nižších, jako je slovo. Máme-li k dispozici paralelní korpus či korpusy nabízející množství kontextů a bohatou varietu ekvivalentů hledaného prvku na škále, která se dá statisticky vyhodnocovat, máme k dispozici prostředky k dosažení mnohem lepších výsledků, než jaké dosud

přinášel starý a manuální přístup, založený na nesystematických a často i podivných příkladech s problematickým zobecněním.

6 Závěr

Je zřejmé, že možnosti, které paralelní korpusy pro komparativní studium jazyků nabízejí, se vlastně teprve otevírají, a na hodnocení toho, co teprve má přijít a přijde, jak lze doufat, je příliš brzo. Zdá se, že by mohlo být užitečné tu znovu připomenout, že daná oblast paralelních korpusů, jejich výzkumu a desiderata z toho vyplývající už jednou, aspoň zčásti shrnuta byla. Připomeňme si proto hlavní závěry z panelové diskuse, nazvané *Final Panel Discussion of the 2009 InterCorp Conference in Prague* (Čermák – Klégr – Corness, 2009). Dospěli k nim po živé diskusi účastníci konference InterCorpu, v níž se vedle hlavních a tíživých otázek a problémů začíná rysovat i pár odpovědí.

1. Úloha třetího jazyka v bilingvních korpusech: míra a metodologie
Názory a ohlasy: Třetí jazyk je nepostradatelný tehdy, není-li počet překladových textů v jazykové dvojici příliš velký. Míra jeho zastoupení se nezdá tak důležitá. Je však třeba rozlišovat co do jeho relevantnosti mezi jazykem originálního textu a překladem; konečné výsledky by se měly ověřovat proti velkému vyváženému jednojazyčnému korpusu.

2. Vyvažování dvou jazyků v paralelním korpusu
Názory a ohlasy: Je žádoucí, jakkoliv rovnováhu v počtu dostupných textů můžou narušovat pragmatické faktory.

3. Společné textové jádro pro více jazyků
Názory a ohlasy: Jako možnost je jistě žádoucí, i když dopředu ve skutečnosti nevíme, kolik uživatelů by tento rys umožňující srovnávání více než dvou jazyků mohlo užívat; lze ale snadno najít faktory, které mluví ve prospěch této možnosti.

4. Legální problémy vztahující se ke copyrightu a vlastnictví textů
Názory a ohlasy: Nemějme obavy. Žádný korpusový lingvista se ještě nedostal před soud za porušení copyrightu tím, že dal do korpusu nějaký text. Sběr paralelních korpusů by se neměl zastavit kvůli legálním formalitám; vždyť přece vždycky existuje prostředek umožňující jen omezený přístup a bude-li třeba, i přístup zaheslovaný; ani praxe dělení textů do vzorků či přeřazování částí textu v jejich sledu se nezdá užitečná.

5. Kritický počet slov či rozsah paralelního korpusu pro praktické účely, specificky v lexikografii.
Názory a ohlasy: Rozsah záleží na cíli, kdy např. v (bilingvní) lexikografii usilující o 20 tisíc lemmat je třeba miliónů slov korpusu.

6. Různé

Návrhy: paralelní korpus by měl pokud možno zahrnovat více textových typů (tj. vedle beletrie a odborných textů). Paralelní korpus je užitečný v jazykové výuce.

Literatura

BARLOW, Michael: Using Concordance Software in Language Teaching and Research. In: Proceedings of the Second International Conference on Foreign Language Education and Technology. Kasugai, Japan: LLAJ & IALL 1992, s. 365 – 373.

BARLOW, Michael: Parallel texts in linguistic analysis. In: Multilingual Corpora in Teaching and Research. Ed. S. P. Botley – T. McEnery – A. Wilson. Amsterdam: Rodopi 2000, s. 106 – 115.

BARLOW, Michael: ParaConc: Concordance software for multilingual parallel corpora. In: Proceedings from First International Workshop on Language Resources for Translation Work and Research. LREC 2002, s. 20 – 24.

Multilingual Corpora: Teaching and Research. Ed. S. Botley – A. McEnery – A. Wilson. Amsterdam: Rodopi 2000. 230 s.

Korpusová lingvistika Praha 2011 – 1 InterCorp. Ed. F. Čermák. Praha: Nakladatelství Lidové noviny 2011. 372 s.

InterCorp: Exploring a Multilingual Corpus. Ed. F. Čermák – A. Klégr – P. Corness. Praha: Nakladatelství Lidové noviny 2010. 253 s.

Mnohojazyčný korpus InterCorp: Možnosti studia. Ed. F. Čermák – F. Kocek. Praha: Nakladatelství Lidové noviny 2010. 292 s.

ČERMÁK, František – ROSEN, Alexandr: The Case of InterCorp, a multilingual parallel corpus. In: International Journal of Corpus Linguistics, 2012, roč. 17, č. 3, s. 411 – 427.

GAGE, William W.: Contrastive Studies in Linguistics: A Bibliographical Checklist. Washington D.C.: Center for Applied Linguistics 1961. 17 s.

HAMMER, John H. – RICE, Frank A.: A bibliography of contrastive linguistics. Washington, DC: Center for Applied Linguistics 1965.

JOHANSSON, Stig: Seeing through Multilingual Corpora; On the use of corpora in contrastive studies. Studies in Corpus Linguistics. John Benjamins Publishing Company 2007. 355 s.

MELAMED, Dan I.: Empirical Methods for Exploiting Parallel Texts. MIT Press 2001. 198 s.

TEUBERT, Wolfgang: Corpus Linguistics and Lexicography. In: International Journal of Corpus Linguistics, 2001, roč. 6, Special Issue, s. 125 – 153.

Text Corpora and Multilingual Lexicography. Ed. W. Teubert. University of Birmingham Benjamins Current Topics, 2007, 162 s.

VAVŘÍN, Martin – ROSEN, Alexandr: Intercorp: A Multilingual parallel Corpus. In: Trudy Meždunarodnoj konferencii Korpusnaja lingvistika 2008, Sankt-peterburgskij gosudarstvennyj univerzitet, Sankt-Peterburg, 2008, s. 156 – 162.

Building Large Corpora and Tools for Computer Lexicography

Karel Pala – Pavel Rychlý

Faculty of Informatics, Masaryk University,
Brno, Czech Republic

Abstract. This paper presents tools for building large corpora and examples of their results. It also describes several tools for computer lexicography created at Faculty of Informatics, Masaryk University.

1 Corpus tools at NLP Centre FI MU

The development of the corpus tools in the NLP Centre FI MU (previously NLP Laboratory) started in 1996, i. e. soon after its establishing. We started with using CWB (Corpus Workbench, CQP – Corpus Query Processor) from Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart[1]. It appeared that the corpus tools have had some limitations, e. g. size limits ($\approx$ 400 M tokens), or that there was no non-latin1 support. Thus we decided to improve the CQP tool and in this way the G(raphical) CQP for Czech National Corpus was developed. It was a multiplatform GUI allowing for queries by CQP.

1.1 Manatee/Bonito

The GCQP was initially used for some time but we worked on a new corpus manager named later Manatee/Bonito. It was released in 2002 and contained some new features, e. g. fast query evaluation and bigger corpus size support. It was completely built on client/server architecture where Manatee was a server and Bonito a graphical client (GUI) through which a user can query and search the particular corpus data. The first version of the Bonito was originally based on GCQP but then the second version – Bonito2 was developed, which was a web application (Rychlý, 2007).

1.2 Sketch Engine

In the year 2003 we started our cooperation with the Lexical Computing Ltd. whose result was the integration of the Sketch Engine (Kilgarriff et al., 2004) with the Manatee/Bonito. The Sketch Engine (SkE) was built in Manatee/Bonito allowing to provide word sketches – one page tables describing an individual word collocational behaviour in corpus, as well as thesaurus capturing

[1] `http://www.ims.uni-stuttgart.de`

semantically close words occuring together with a key word. Most of the current Manatee/Bonito development is related to the SkE. Since the SkE is a proprietary application NoSketch Engine[2] has also been developed and is freely accessible. The research related to the development of the SkE for various world languages is being discussed at the annual Sketch Engine Workshop (SKEW).

2 Large Corpora

2.1 Motivation

What can be a motivation for building large corpora, i. e. corpora with size exceeding 1 billion tokens? The answer is relatively simple: more data – better data. Let us give an example: noun "test":

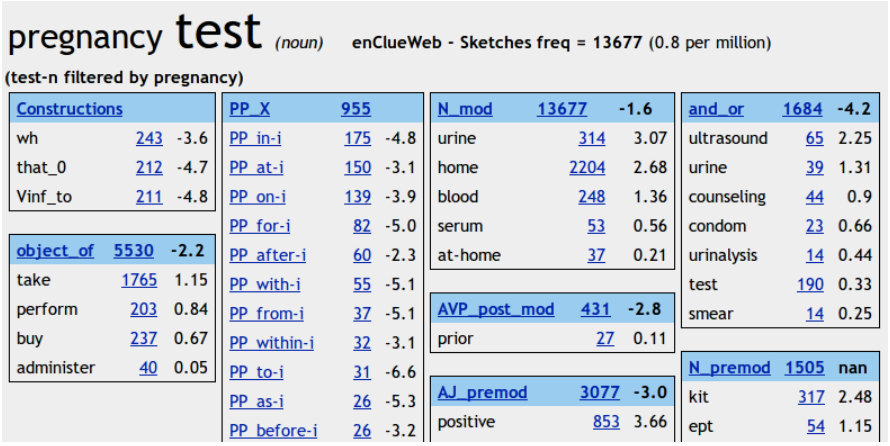
- British National Corpus
- 15 789 hits
- word sketches from the Sketch Engine
- object-of: *pass, undergo, satisfy, fail, devise, conduct, administer, perform, apply, boycott*
- modifier: *blood, driving, fitness, beta, nuclear, pregnancy*

The question is: can we freely combine any two from that list? It is hard to answer this question from BNC data. The following list contains collocations of two phrases in two different corpora:

- "blood test" in BNC
  - object-of: *order (3), take (12)*
- "blood test" in enClueWeb16
  - object-of: *order (708), perform (959), fast (173), undergo (174), administer (123), conduct (229), require (676), repeat (80), run (347), request (105), take (1215)*
- "pregnancy test" in BNC
  - 26 hits, no significant collocations
- "pregnancy test" in enClueWeb16
  - object-of: *take (1765), perform (203), buy (237), administer (40)*

The numbers are the number of hits in the respective corpus. Figure 1 shows the whole word sketch of *pregnancy test* in the enClueWeb16 corpus. We can see that we do not get any useful data about such phrases from BNC. Much bigger corpora provide valuable information on such phrases.

[2] `http://nlp.fi.muni.cz/trac/noske/`



pregnancy **test** *(noun)* **enClueWeb - Sketches freq = 13677** (0.8 per million)

(test-n filtered by pregnancy)

Constructions		
wh	243	-3.6
that_0	212	-4.7
Vinf_to	211	-4.8

object_of	5530	-2.2
take	1765	1.15
perform	203	0.84
buy	237	0.67
administer	40	0.05

PP_X	955	
PP_in-i	175	-4.8
PP_at-i	150	-3.1
PP_on-i	139	-3.9
PP_for-i	82	-5.0
PP_after-i	60	-2.3
PP_with-i	55	-5.1
PP_from-i	37	-5.1
PP_within-i	32	-3.1
PP_to-i	31	-6.6
PP_as-i	26	-5.3
PP_before-i	26	-3.2

N_mod	13677	-1.6
urine	314	3.07
home	2204	2.68
blood	248	1.36
serum	53	0.56
at-home	37	0.21

AVP_post_mod	431	-2.8
prior	27	0.11

AJ_premod	3077	-3.0
positive	853	3.66

and_or	1684	-4.2
ultrasound	65	2.25
urine	39	1.31
counseling	44	0.9
condom	23	0.66
urinalysis	14	0.44
test	190	0.33
smear	14	0.25

N_premod	1505	nan
kit	317	2.48
ept	54	1.15

Fig. 1. Word sketch for the phrase "pregnancy test" in enClueWeb16

2.2 Giga corpora

If we decide to go for building very large corpora we have to have an appropriate procedure. First, we need to think of suitable resources. One sort of the convenient single high-volume sources comes from press agencies, e. g. GigaEnglish, GigaChinese from LCD. Then it is possible to use crawled web pages; to obtain them is rather cheap and easy, for example, series WaC – Web as Corpus (English, Italian, German), and also series TenTen – 10^{10} for English, Czech, Russian... It is also possible to work with specialised corpora from web search which are smaller than the previous ones but focused on selected domains. They can be obtained via WebBootCat (Corpus Architect in SkE) (Baroni et al., 2006).

2.3 The tools and the building process for large web corpora

The first step is the web crawling or using some of the existing web collections such as ClueWeb09[3] (see below). When performing crawling, the language identification has to be done first to select a text in the desired language. Next follows page cleaning (only text) and after that removing (deduplication) the repeating parts of the crawled text. POS-tagging is usually being performed simultaneously with lemmatization as well. Then metainformation reflecting domains can be added together with the necessary encoding and searching.

We have developed a new web crawler for building text corpora named SpiderLing (Suchomel and Pomikálek, 2011), which replaced previously-used tool Heritrix. It has asynchronous communication design and allows for partial processing of the downloaded data. It is aimed to

[3] `http://www.lemurproject.org/clueweb09.php/`

improve the crawling efficiency $yield\ rate = \frac{final\ data}{downloaded\ data}$ and can be focused on the text-rich web domains.

Web pages contain a lot of text data which are not suitable for text corpora, i. e. headers, footers, menu, etc. These sections of web pages are called boilerplate. A tool (jusText) has been developed for removing boilerplate content. It is an open source instrument which can be found on `http://code.google.com/p/justext/`.

Another problem with web corpora is presence of duplicate texts or paragraphs. They are a natural result of the way web pages are created. Onion (ONe Instance ONly) is a tool for removing duplicate parts from large collection of texts. It detects and removes exact duplicate content as well as the near duplicates. It is based on computing n-grams of words and is optimized to be scalable up to terabytes of texts. Onion is an open source which can be found on `http://code.google.com/p/onion/`.

In our work, we also use the ClueWeb09 collection to build a large corpus of English. The ClueWeb09 contains about 1 billion web pages in 10 languages. It was compiled at the Carnegie Mellon University in January and February 2009. Its total size is 5 TB (compressed) and 25 TB (uncompressed); the data are in the WARC format, which is the output of web servers during page downloads. For English, it contains about 500 billion pages (1.9 TB gzipped). The resulting corpus (enClueWeb) has more than 80 billion tokens (Pomikálek et al., 2012), Table 1 lists more detailed numbers. The enClueWeb16 corpus mentioned above is the 16-billion-token part from the whole enClueWeb.

Value	Size [mil.]
word forms	68 845
numbers	1 485
alphanumeric	70 330
punctuation	9 849
others	1 810
total tokens	81 990
documents	138

Table 1. Sizes of the enClueWeb

During processing of the data from web pages to final corpus, each step means a reduction of size. The biggest reduction comes from removing boilerplate, where less than half of data are removed. Exact numbers for a 7% subcorpus of enClueWeb are listed in Table 2.

Processing of such large corpora takes a long time, even on a powerful server. To prepare all the data within a reasonable time we have experiments in parallel processing of selected steps. We have used the following configuration for our experiments:

- server: 8 × 8-core Intel Xeon X7560, 2.27 GHz CPUs, 440 GB RAM
- storage: 70 TB RAID-6 (disks: 2 TB 7200 rpm SAS 6 Gbps)

stage	word count [mil.]	% of previous step
original subcorpus	24 633	n/a
after removing boilerplate	11 363	46.1 %
after language filtering	9 586	84.4 %
after removing exact duplicates	8 983	93.7 %
after block level deduplication	7 279	81.0 %

Table 2. Size reduction during processing of 7% of enClueWeb

The results are displayed in Table 3.

Step	CPUs	Time	RAM
removing boilerplate + lang. filtering	50	108 h	< 1 GB
removing identical documents	10	9 h	5 GB
generating hashes of n-grams	10	1 h	< 1 GB
finding hashes of duplicate n-grams	1	9 h	10 GB
deduplication	1	65 h	148 GB
POS-tagging	20	44 h	< 1 GB

Table 3. Parallelization of the enClueWeb processing

3 Tools for Lexicography at NLP Centre

3.1 GDEX – Good Dictionary EXamples

The GDEX tool has been developed to help lexicographers find appropriate dictionary examples and facilitate building of the individual dictionary entries (Kilgarriff et al., 2008). To obtain good examples the following criteria are applied: short sentences with simple structure, short words, frequent phrases, easy readability and position of the keyword. GDEX also contains a sentence-ranking library and uses the combination of classifiers and operators. Its performance can be optimized from manually annotated data. The first GDEX version is available within NoSketch Engine.

3.2 DEB II platform

Lexicographical applications built at the NLP Centre FI MU are developed on the DEB II platform, i. e. they are tools designed for various kinds of dictionary editing and browsing. Some of them are tools for the development of dictionary writing systems (lexicographical workstations).

On the DEB II platform we can handle practically any dictionary data in the XML format, a common feature being a strict client-server architecture. The server part consists of:

- server side modules (servlets in the programming language Ruby),
- database backend (Oracle Berkeley DB XML or Sedna).

The client component displays:

- simple functionality,
- graphical interfaces based on Mozilla/Firefox extensions,
- web interfaces are used as a rule,
- it is freely available at `http://deb.fi.muni.cz/apt`.

3.3 DEBDict

One of the most frequently used lexicographical tools developed on the DEB II platform is a dictionary browser named DEBDict (Horák et al., 2006b), which allows one to work with six Czech main dictionaries, particularly:

- SSČ – Slovník spisovné češtiny (Dictionary of Literary Czech),
- SSJČ – Slovník spisovného jazyka českého (Dictionary of Written Czech),
- PSJČ – Příruční slovník spisovného jazyka českého (Hand Dictionary of Literary Czech),
- SČS – Slovník českých synonym (Dictionary of Czech Synonyms),
- SCS – Slovník cizích slov (Dictionary of the Words of Foreign Origin),
- SČFI – Slovník českých frazeologismů a idiomů (Dictionary of the Czech Frazeologisms and Idioms.

The DEBDict also makes other resources accessible:

- CZWN – Czech WordNet linked to Princeton WordNet,
- Slovene dictionary,
- Complex Russian Dictionary,
- CIA World Factbook,
- Seznam Encyklopedie,
- Map of Czech Republic,
- integrated morphological analyzer Ajka.

Like other tools on the DEB platform, it is based on the architecture client/server; data are stored in the XML format, and they can differ for different dictionaries. DEBDict also allows integrating dictionaries of any language, so in this sense it is language independent. Some useful functions – searching within the definitions – are at the user's disposal. Presently, the tool is regularly used by almost 1000 users from all over the world, mostly working with 6 main Czech dictionaries. It can be easily accessed and then obtained via `http://nlp.fi.muni.cz/declaration/`. An example of DEBDict window is Figure 2.

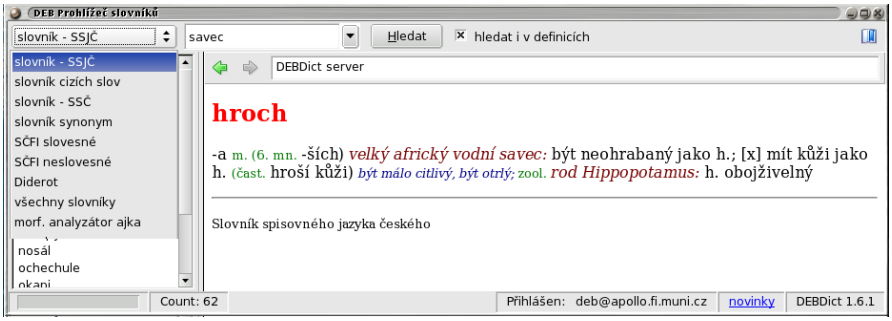


Fig. 2. DEBDict

3.4 Praled/Pralex

Lexicographical workstation for Czech

This lexicographical workstation has been developed in close cooperation with ÚJČ AV ČR according to their lexicographers requirements. The tool is being used for the creation of the Lexical Database of the Czech Language in 21st Century (Goláňová, 2011). Praled is a complex tool with various lexicographical functions:

- handling and writing dictionary definitions and their different parts,
- searching in corpus using Word Sketch Engine and selection of the relevant collocations,
- standard working with concordances and their processing,
- compiling different parts of the dictionary entry – morphology, stylistic features, etymology, pronunciation, valency, …
- providing information on frequencies in a corpus,
- management, approving and correcting of the work done by lexicographers, its revising, correcting and storing.

3.5 DEBVisDic

This tool is an editor and browser for building and handling semantic networks, mainly wordnets (Horák et al., 2006a). It provides multiple views of multiple wordnets in selected windows together with their synchronization. Main functions are related to the synset editing – writing and changing them and copying. Semantic relations – hypero/hyponymy, synonymy, antonymy, holo/meronymy, etc. can be easily handled working with hypero/hyponymy trees and their browsing. Various types of queries can be asked and the respective resulting lists obtained. The functionality of journaling is implemented indicating the actions that took place. Also, plain XML view of the particular synsets can be provided. In fact, DebVisDic has become a standard tool for handling WordNets, and at present it is used for building Nepali, Chinese, Zulu, Afrikaans, Russian, and WordNets of other languages as well.

3.6 Tools for building terminological dictionaries

They are small dictionary writing systems making it possible to build a terminological dictionary for a selected terminological domain. The first of them is named TeDi[4] (Terminological Dictionary) and is designed for preparation of a new terminological dictionary of the Czech art terms together with a multilingual terminological database (Czech, English, German, French). Presently it contains more than 1 300 entries. Two main functions are – editing and browsing. XML format is used for storing entries in the database.

3.7 TeAgro – tool for building agronomical terminology

Similarly as TeDi, the tool TeAgro has been designed for building a multilingual glossary of the agronomical terminology (Czech and English so far). It displays similar functions for editing and browsing of the agronomical terms. The entries are stored in XML format as well. It should be remarked that the DEB II platform is very versatile and using it, one can develop a lexicographical workstation of a similar sort quickly and comfortably in a relatively short time – for an experienced programmer it can be approx. a month.

3.8 Tool for the Pattern Dictionary of English verbs

It is an editor and browser for building context patterns of English verbs. The tool is based on the context pattern analysis (CPA) which was primarily developed by P. Hanks, and relies on the Theory of Norms and Exploitations (Hanks, 2004).

It associates word meaning with word use by means of analysis of phraseological patterns and collocations (BNC is used as the main resource). The "meaning" of a pattern is expressed as a set of basic implicatures, and meanings of verbs are associated with prototypical sentence contexts found in a corpus.

The goal is to develop the Pattern Dictionary of English Verbs (PDEV) – approx. 720 items have been compiled so far. PDEV is conceived as a fundamental resource for use in computational linguistics, NLP, and also for language teaching and cognitive science. Verb patterns are prepared manually for Italian and Spanish as well. The tool is integrated with the corpus manager Manatee/Bonito and Sketch Engine. See (Hanks, 2004), `http://nlp.fi.muni.cz/projects/cpa/`

"Inter-annotators" agreement for the patterns has been investigated by Holub and Cinková (Cinková et al., 2010). Relations to the Framenet and possibly to VerbNet and WordNet are going to be explored in near future. Further development of the PDEV in cooperation with University of Wolverhampton (R. Mitkov) is planned from October 2012.

3.9 Conclusions

We have demonstrated that our research within computer lexicography is focused on:

- developing and improving corpus tools (Manatee/Bonito, Sketch Engine),
- corpus building – especially very large web corpora,

[4] `http://deb.fi.muni.cz/clients-tedi.php`

- tools for building large web corpora – toolkit Spiderling-JusText, Onion, Chared,
- among others also large Web Slovak corpus (approx. 900 million tokens) was built,
- lexicographical tools – workstations, individual editors and browsers (mostly language independent),
- they are used for building lexical databases (terminological, wordnets, verbal – Verballex, PDEV),
- successful cooperation with the industry (Seznam.cz, Lexical Computing Ltd.),
- general goal: to make building resources easier and richer.

Our general goal is to develop appropriate tools for lexicographers and in this way to make building high-quality lexical resources not only easier but also richer. The presented results have shown show that we are successful in this endeavour.

Acknowledgements

This work has been partly supported by the Ministry of Education of CR within the LINDAT-Clarin project LM2010013.

References

Baroni, M., Kilgarriff, A., Pomikálek, J., and Rychlý, P. (2006). WebBootCaT: instant domain-specific corpora to support human translators. pages 247–252.

Cinková, S., Holub, M., Rychlý, P., Smejkalová, L., and Šindlerová, L. (2010). Can Corpus Pattern Analysis be used in NLP? In *Proceedings of the Text, Speech and Dialogue Conference*, pages 67–74, Berlin-Heidelberg. Springer Verlag.

Goláňová, H. (2011). Novočeský lexikální archiv a excerpce v průběhu let 1911–2011. *Slovo a slovesnost*, (4):287–300.

Hanks, P. (2004). The syntagmatics of metaphor and idiom. *International Journal of Lexicography*, 17(3):245–274.

Horák, A., Pala, K., Rambousek, A., and Povolný, M. (2006a). DEBVisDic–first version of new client-server WordNet browsing and editing tool. In *Proceedings of the Third International Wordnet Conference (GWC-06)*. Jeju Island, Korea.

Horák, A., Pala, K., Rambousek, A., and Rychlý, P. (2006b). New clients for dictionary writing on the DEB platform. In *DWS 2006: Proceedings of the Fourth International Workshop on Dictionary Writings Systems*, pages 17–23.

Kilgarriff, A., Husák, M., McAdam, K., Rundell, M., and Rychlý, P. (2008). GDEX: Automatically finding good dictionary examples in a corpus. In *Proceedings of the XIIIth EURALEX International Congress.*, pages 425–432, Barcelona. Universitat Pompeu Fabra.

Kilgarriff, A., Rychlý, P., Smrž, P., and Tugwell, D. (2004). The Sketch Engine. *Proceedings of Euralex*, pages 105–116. Available at: `http://www.sketchengine.co.uk`.

Pomikálek, J., Jakubíček, M., and Rychlý, P. (2012). Building a 70 billion word corpus of English from ClueWeb. In Calzolari, N., Choukri, K., Declerck, T., Doğan, M. U., Maegaard, B., Mariani, J., Odijk, J., and Piperidis, S., editors, *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, pages 502–506, Istanbul, Turkey. European Language Resources Association (ELRA).

Rychlý, P. (2007). Manatee/Bonito – A Modular Corpus Manager. In *Proceedings of Recent Advances in Slavonic Natural Language Processing 2007*, pages 65–70. Masaryk University, Brno.

Suchomel, V. and Pomikálek, J. (2011). Practical Web Crawling for Text Corpora. In *Proceedings of Recent Advances in Slavonic Natural Language Processing 2011*, pages 97–108, Brno, Czech Republic. Tribun EU.

Multilingual Resources with Bulgarian – Recent Developments (IMI-BAS Experience)

Ludmila Dimitrova

Institute of Mathematics and Informatics, Bulgarian Academy of Sciences,
Sofia, Bulgaria

Abstract. In this paper I make a brief overview of the multilingual resources employing Bulgarian language developed recently at the Mathematical Linguistics Department of the Institute of Mathematics and Informatics at the Bulgarian Academy of Sciences. These resources – digital corpora, mono- and bilingual lexical databases, and a Bulgarian-Polish online dictionary – are prepared under three EC projects and two bilateral academic projects. The first Bulgarian-Polish parallel, aligned and comparable corpus, Bulgarian-Slovak parallel and aligned corpus, the first Bulgarian-Polish-Lithuanian parallel, aligned and comparable corpus, Bulgarian-Polish lexical database, and their applications are briefly presented.

1 Introduction

The Department of Mathematical Linguistics at the IMI-BAS participated in three large language engineering EC projects:

- COP project 106 MULTEXT-East *Multilingual Text Tools and Corpora for Central and Eastern European Languages*, 1995–1997, coordinator Jean Véronis, CNRS[1];
- INCO Copernicus project PL96-1142 CONCEDE *Consortium for Central European Dictionary Encoding*, 1998–2000, coordinator Roger Evans, University of Brighton[2];
- 7th FP project GA 211938 MONDILEX *Conceptual Modelling of Networking of Centres for High-Quality Research in Slavic Lexicography and their Digital Resources*, 2008–2010, coordinator Ludmila Dimitrova, IMI-BAS[3].

Four multilingual corpora under these EC projects and two bilateral academic projects were developed or are still under construction:

- MULTEXT-East *Parallel and Comparable Corpora with Bulgarian*, (MULTEXT-East EC project),
- *Bulgarian-Polish Parallel and Comparable Corpora*, (Joint research project between IMI-BAS and ISS-PAS),

[1] http://aune.lpl.univ-aix.fr/projects/multext-east/

[2] http://www.itri.brighton.ac.uk/projects/concede/

[3] http://www.mondilex.org

– *Bulgarian-Slovak Parallel and Aligned Corpora*, (Joint research project between IMI-BAS and ĽŠIL-SAS),
– *Bulgarian-Polish-Lithuanian Parallel and Comparable Corpora*, (Joint research project between IMI-BAS and ISS-PAS).

2 MULTEXT-East Corpus

MULTEXT-East project[4] builds an annotated multilingual corpus: English and six languages from Central and Eastern Europe: Bulgarian, Czech, Estonian, Hungarian, Romanian and Slovene (Dimitrova et al., 1998; Dimitrova et al., 2005) composed of 3 major parts:

– Parallel Corpus
– Comparable Corpus
– Speech Corpus (small) – consists of texts comprising fourty short passages of five thematically connected sentences, each spoken by several native speakers, with phonemic and orthographic transcriptions.

MULTEXT-East Parallel Corpus

MULTEXT-East parallel corpus is a multilingual corpus, based on George Orwell's novel "**1984"** in the English original and its translations in Bulgarian, Czech, Estonian, Hungarian, Romanian and Slovene. The parallel corpus is produced as a well-structured, lemmatized, CES-corpus. The texts are automatically annotated for tokenization, sentence boundaries, and Part-of-Speech annotation, using the project tools.

An excerpt of the 3rd version of the MULTEXT-East parallel corpus – CesANA encoding (XML/TEI P4) – is presented in the table below:

Априлският ***ден бе*** *ясен и студен, часовниците биеха тринайсет часа.*
.....
<tok type=WORD from='Obg.1.1.1.1\12'>
<orth>**ден**</orth>
<disamb><base>**ден**</base><ctag>NCMS-N</ctag></disamb>
<lex><base>**ден**</base><msd>Ncms-n</msd><ctag>NCMS-N</ctag></lex>
</tok>
<tok type=WORD from='Obg.1.1.1.1\16'>
<orth>**бе**</orth>
<disamb><base>**съм**</base><ctag>VAIA3S</ctag></disamb>
<lex><base>**бе**</base><msd>Qgs</msd><ctag>QG</ctag></lex>
<lex><base>**съм**</base><msd>Vaia2s</msd><ctag>VAIA2S</ctag></lex>
<lex><base>**съм**</base><msd>Vaia3s</msd><ctag>VAIA3S</ctag></lex>
</tok>
.....

[4] http://aune.lpl.univ-aix.fr/projects/multext-east/

MULTEXT-East aligned corpus

Alignment between the English version and a translation in each of the six CEE languages ensures six pair-wise alignments that produce the aligned corpus.

For Bulgarian, the alignment is made by the Vanilla aligner; the tool has shown 6 699 bilingual links in total:

Aligned pairs	**Sentences**	%
2-2	2	0.030
2-1	23	0.345
1-2	36	0.540
1-1	6637	99.074
0-1	1	0.015

The following table presents 1-1 aligned sentences from the MTE aligned corpus:

<Obg.1.1.7.4>Още три сгради, подобни по външен вид и размери, бяха посети из **Лондон**. <Oen.1.1.9.2>Scattered about **London** there were just three other buildings of similar appearance and size.
<Obg.1.1.7.5>И дотолкова се извисяваха над околните здания, че от покрива на жилищен дом **Победа** можеха да се видят и четирите едновременно. <Oen.1.1.9.3>So completely did they dwarf the surrounding architecture that from the roof of **Victory Mansions** you could see all four of them simultaneously.

MULTEXT-East Comparable Corpus

A comparable corpus can be defined as a collection of texts composed independently in the respective languages and put together on the basis of similarity of content, domain, and communicative function. Comparable corpora can be created from a variety of sources: collections of texts distributed in electronic format (e.g. newspaper archives on CD-ROM, the Internet, etc.), or even from scanned or typewritten material. Criteria for creating comparable corpora could also be different: in dependence of the homogeneity of texts, or in terms of features such as subject domain, etc. The size of comparable corpora can vary depending on how well they meet these criteria; for example, collections of newspaper articles downloaded from the Internet can produce great comparable corpora. The Bulgarian MULTEXT-East comparable corpus is annotated manually at the paragraph level, tagged with sub-paragraph mark-up (abbreviations, dates, names), and contains two sub-corpora:

1. Bulgarian fiction – contemporary Bulgarian literature, 97 251 words in total;
2. Bulgarian newspapers – newspaper excerpts, 96 538 words in total.

The corpus of Bulgarian fiction comprises the following novels:

1. *PASSION or the Death of Alice* by Emilia Dvorianova,
2. *I Want, I Believe, I Can* by Julia Berberyan (first four chapters of the novel).

It is annotated at a paragraph level ( ), and tagged with sub-paragraph mark-up (<q rend="PRE=mdash"> </q>), (<name> </name>), (<date> </date>), etc.

The table below shows an excerpt of the novel *PASSION or The Death of Alice:*

<q rend="**PRE=mdash**">**Какво е това, дето го чете момичето, госпожо, неясно ми се вижда, да не й повлияе зле.**</q>

А тя ми отвърна като знаеща: <q rend="**PRE=mdash**"> **Философия,**
<name type="**person**">**Йо**</name>. </q>

3 Bulgarian-Polish Corpus

The first Bulgarian-Polish corpus is developed under the joint research project between IMI-BAS and ISS-PAS called "Semantics and Contrastive Linguistics with a focus on a Bilingual Electronic Dictionary", coordinated by L. Dimitrova and V. Koseska. Its most recent version contains total of approx. five million words and comprises two corpora: parallel, including aligned sub-corpus and comparable corpus (Dimitrova and Koseska, 2009).

Bulgarian-Polish Parallel Corpus

The parallel corpus contains more than three million words, mostly in literary texts – novels and short stories. A small part contains texts of official documents of the European Commission available through the Internet.
The texts are divided into two parts:

- Original Bulgarian texts with Polish translations or *vice versa;*
- Texts translated from other languages into both the Bulgarian and Polish languages.

Bulgarian-Polish Aligned Corpus

This corpus currently contains approximately 1 million words, mainly texts of Polish novels or science fiction and their Bulgarian translations:

- Stanisław Lem's *Solaris* and *Return from the Stars*
- Ryszard Kapuściński's *Another Day of Life*, *The Shadow of the Sun*, and *The Soccer War*
- Stefan Żeromski's *Ashes*.

The texts are aligned at the paragraph level (Level P) and/or at the sentence level (Level S). Two language-independent, freely available programs were used to align Bulgarian-Polish parallel texts at the Level S:

1. Memory Translation 2007, a computer aided tool TextAlign (`http://mt2007-cat.ru/index.html`)
2. Bitext Aligner/Converter, bitext2tmx aligner (`http://bitext2tmx.sourceforge.net/`).

These software packages have applications in computer-assisted translation. Both tools align bilingual texts without bilingual dictionaries, but the human editing is obligatory. The resulting aligned texts are similar.

An example of the alignment at the Level P from Stefan Żeromski's *Ashes* (vol. 1, part 1) follows:

Polish: ** **Psy ucięły. Zaraz potem drugi głos, bliższy Rafała, odpowiedział jednokrotnie tym samym sposobem.** **	**Bulgarian**: ** **Кучетата млъкнаха. Веднага след това друг глас, по-близко до Рафал, отговори еднократно по същия начин.** **
Polish: ** **Młody myśliwiec jeszcze przez czas pewien leżał na ziemi, pękając ze złości: Po chwili jednak zerwał się na równe nogi, strzepnął śnieg z siebie, odszukał w krzakach pojedynkę. Wytarł oczy i, na podobieństwo sarn skacząc przez choiny, pomknął na dół.** **	**Bulgarian**: ** **Младият ловец лежа още малко на земята, позеленял от яд. Но после изведнъж скочи на крака, изтупа снега от себе си и потърси пушката в храстите. Избърса очи и скачайки като сърна през младите елички, полетя надолу.** **

The following example shows 2-1 aligned sentences from Kapuściński's *The Soccer War:*

```
<tu tuid="0000000353">
        <tuv xml:lang="Polish">
          <seg> Myslałem, że skonał. Ale po chwili wychyliła sie twarz, szara,
sciagnieta, naiwna, wyczekujaca z pokora na nastepny akt przeznaczenia. </seg>
        </tuv>
        <tuv xml:lang="Bulgarian">
             <seg> Помислих, че това е краят, но след малко над капака се подаде
лицето му – посивяло, изопнато, очакващо с наивно покорство какво още може да
му поднесе съдбата. </seg>
        </tuv>
</tu>
```

Bulgarian-Polish Comparable Corpus

The Bulgarian-Polish comparable corpus contains texts in Bulgarian and Polish of similar sizes: these are xcerpts from newspapers, literary works, mostly modern Bulgarian and Polish literature (2nd half of the 20th c.), with the text size comparable in both languages, and available on the Internet, for example, in Bulgarian: *The Iron Oil Lamp* and *The Bells of Prespa* by Dimitar Talev, *Tobacco* and *Doomed Souls* by Dimitar Dimov; examples in Polish would be: Ryszard Kapuściński's *Imperium,* Stanisław Lem's *The Star Diaries.*

4 Bulgarian-Slovak Parallel and Aligned Corpus

The Bulgarian-Slovak corpus is currently developed under the joint research project between IMI-BAS and ĽŠIL-SAS "Electronic Corpora – Contrastive Study with Focus on Design of Bulgarian-Slovak Digital Language Resources", and coordinated by L. Dimitrova and R. Garabík. The corpus comprises two sub-corpora: parallel and aligned.

Parallel Bulgarian-Slovak/Slovak-Bulgarian Corpus

This corpus contains more than 1.2 million words in original Bulgarian novels with Slovak translations or *vice versa,* and texts of novels, fiction, and short stories in other languages translated into Bulgarian and Slovak (Dimitrova and Garabík, 2011).

Bulgarian-Slovak Aligned Corpus

The Hunalign software is used to align Bulgarian-Slovak/Slovak-Bulgarian parallel texts of the corpus at the sentence level.

The bilingual aligned corpus currently contains approximately 500 000 words in parallel texts, aligned at the sentence level, including the texts of:

- Bulgarian novels and their Slovak translations:
  - Dimitar Dimov's *Doomed Souls;*
  - Pavel Vezhinov's *The Barrier*,
- Slovak novel: Klára Jarunková's *The Silent Wolf's Brother* and its Bulgarian translation,
- Bulgarian and Slovak translations of Jaroslav Hašek's *The Good Soldier Švejk.*

The recent version of the aligned Bulgarian-Slovak/Slovak-Bulgarian corpus is available via a simple web interface at `http://korpus.sk:8090/` (Fig. 1 and Fig. 2).

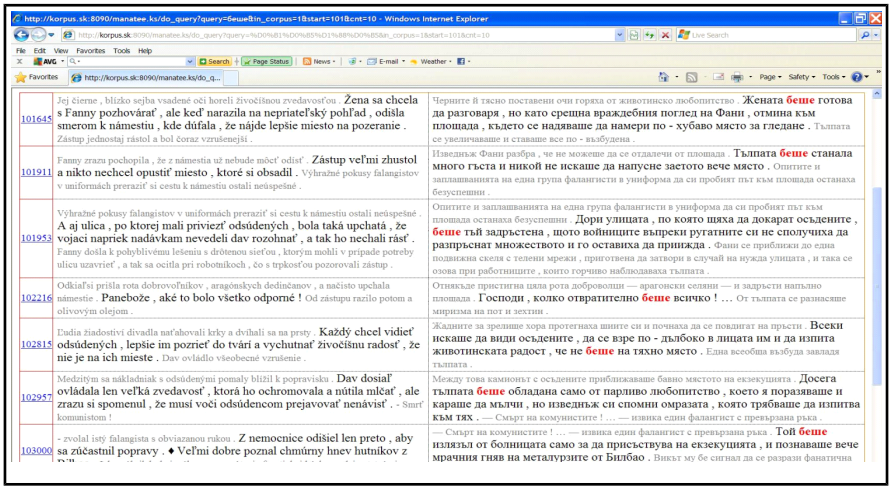


Fig. 1. Bulgarian-Slovak aligned corpus: concordances of Bulgarian verb **беше**

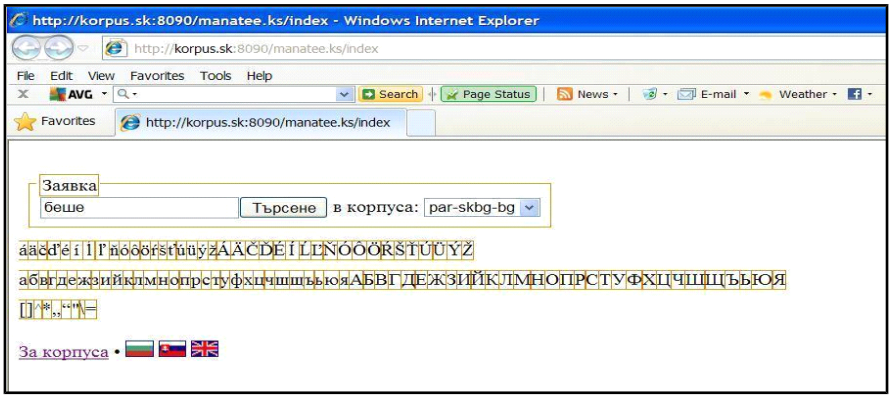


Fig. 2. Bulgarian-Slovak aligned corpus: Web search interface – a dialogue box in Bulgarian

5 Bulgarian-Polish-Lithuanian Corpus

The trilingual corpus is an experimental corpus, recently under development due to the need for research material for contrastive studies in these three languages. Currently, it contains more than three million words in two Slavic languages: Bulgarian, which belongs to the South-Slavic subgroup, Polish from the West-Slavic subgroup of the Slavic languages, and one Baltic language – Lithuanian, which belongs to the Eastern Baltic group. The corpus comprises two corpora: parallel and comparable (Dimitrova et al., 2010). Some parallel texts are aligned at Level S.

Parallel Corpus

The parallel Bulgarian-Polish-Lithuanian corpus contains more than one million words up to now. A part of the parallel corpus comprises original literary texts (fiction, novels, and short stories) in one of the three languages with translations in the other two, and texts of brochures of the European Commission which are official documents of the European Union and the European Parliament, available through the Internet. The remaining part of the parallel corpus comprises texts in other languages translated into Bulgarian, Polish, and Lithuanian.

Thus, the texts are classified as either

- Core – original literary texts (fiction, novels, and short stories) in one of the three languages with translations in the other two, aligned at the paragraph and sentence levels.
- Translations – texts in other languages translated into Bulgarian, Polish, and Lithuanian: novels translated from other languages, aligned at the paragraph level (Antoine de Saint-Exupéry's *Le Petit Prince*, Michael Bulgakov's *Master and Margarita*), and specialized texts of brochures and official documents of the European Commission, European Union, available through the Internet.

The recent result of the collaborative work is the **aligned Bulgarian-Polish-Lithuanian corpus**. At the first stage of the aligned process, the align software tool was used to align the original text, e.g. Stanislav Lem's *Solaris* in Polish, and its Bulgarian translation. At the second stage, the procedure is repeated with the input pair being the original Polish text and its Lithuanian translation. At the third stage, after a comparison of the two output aligned texts, Polish-Bulgarian and Polish-Lithuanian, the alignment ends up with a sequence of triples: a sentence in Polish and its translations in Bulgarian and Lithuanian.

The following example presents an excerpt from the texts of Stanislav Lem's *Solaris* (using TextAlign) aligned at sentence level:

```
<tu tuid="0000000011">
      <tuv xml:lang="polish">
        <seg>Widziałem już seledynowy kontur jedynego wskaźnika.</seg>
      </tuv>
      <tuv xml:lang="bulgarian">
        <seg>Вече различавах светлозелените контури на универсалния указател.</seg>
      </tuv>
      <tuv xml:lang="lithuanian">
        <seg>Jau išskyriau žalsvus universalaus indikatoriaus kontūrus.</seg>
      </tuv>
</tu>
```

The following table shows an excerpt from Bulgarian-Polish-Lithuanian aligned corpus Level P:

Bulgakov's *Master and Margarita*:

BG: Кайсиевият сок вдигна обилна жълта пяна и наоколо замириса на бръснарница. Литераторите го изпиха и веднага се разхълцаха, платиха и седнаха на една пейка с лице към езерцето и с гръб към Бронная.
PL: Morelowy napój wyprodukował obfitą żółtą pianę i w powietrzu zapachniało wodą fryzjerską. Literaci wypili, natychmiast dostali czkawki, zapłacili i zasiedli na ławce zwróceni twarzami do stawu, a plecami do Bronnej.
LT: Abrikosų gėrimas suputojo geltona puta, ir oras pakvipo kirpykla. Literatai atsigėrę tučtuojau ėmė žagsėti, užsimokėjo ir susėdo ant suolelio veidais į tvenkinį ir nugaromis į Bronaja gatvę.
RU: Абрикосовая дала обильную желтую пену, и в воздухе запахло парикмахерской. Напившись, литераторы немедленно начали икать, расплатились и уселись на скамейке лицом к пруду и спиной к Бронной. *Часть 1, Глава 1 „Никогда не разговаривайте с неизвестными" //Интернет-библиотека Алексея Комарова* — `http://ilibrary.ru//`)

Bulgarian-Polish-Lithuanian Comparable Corpus

The trilingual Bulgarian-Polish-Lithuanian comparable corpus contains literary works representing mostly modern Bulgarian, Polish, and Lithuanian literature (from the second half of the 20th century), with the text size being comparable in the three languages, as well as texts from the electronic media. The latter text type encompasses descriptions of the same event in all three languages. The English text is also included. Such texts are specified as "parallel descriptions of content".

6 Lexical Databases (LDBs)

The CONCEDE project[5] developed a so-called model for creation of standardized (according to TEI[6]) lexical databases in six European languages: Bulgarian, Estonian, Czech, Hungarian, Romanian, and Slovene (MTE-languages).

The first Bulgarian LDB (Dimitrova et al., 2002), developed for the CONCEDE project, contains more than 2 700 lexical entries from (Andreychin et al., 1994).

[5] `http://www.itri.brighton.ac.uk/projects/concede/`

[6] `http://www.tei-c.org/index.xml`

Following the monolingual CONCEDE model for LDBs, we have designed and developed a bilingual LDB (Dimitrova et al., 2009a) for supporting a Bulgarian-Polish online dictionary (Dimitrova et al., 2009b). In addition, some new tags are added for presentation of the following:

- The Bulgarian conjugation (in total 3 conjugations) – <conjugation> tag and <type> tag;
- Semantics information – `<semantic>` tag and `<type>` tag (type = 1 for verbs that mean "state", type = 2 – for "event");
- Aspect of verbs in tag `<gram>` (for perfective and imperfective aspect of verbs);
- Specific information about transitivity or intransitivity of verbs (in tag `<subc>`).

The following example shows the presentations of the dictionary entry for headword ***боря*** in a paper dictionary (Sławski, 1987):

бо'рl, -иш *vi*. niepokoić, męczyć; **~я се** borykać się, walczyć, zmagać się

and in the Bulgarian-Polish LDB:

```
<entry>
  <hw>борlя</hw>
  <conjugation><orth>-иш</orth><type>2</type></conjugation>
<semantic><orth>състояние</orth><type>1</type></semantic>
            <subc>преходен</subc>
            <pos>гл.</pos>
            <gram>несв.</gram>
            <struc type="Sense" n="1">
                        <trans>niepokoić</trans>
            </struc>
            <struc type="Sense" n="2">
                         <trans>męczyć</trans>
            </struc>
            <struc type="Derivation" n="1">
                        <orth>~я се</orth>
                        <subc>непреходен</subc>
                        <pos>гл.</pos>
                        <gram>несв.</gram>
            <struc type="Sense" n="1">
                        <trans>borykać się</trans>
            </struc>
            <struc type="Sense" n="2">
                        <trans>walczyć</trans>
            </struc>
            <struc type="Sense" n="3">
                        <trans>zmagać się</trans>
            </struc>
            </struc>
</entry>
```

7 Bulgarian-Polish Online Dictionary

The first Bulgarian-Polish online dictionary was designed to be a general-purpose dictionary oriented to the common user and available by open access via the Net.

It is realized by the web-based application supported by the Bulgarian-Polish LDB and Relational DB.

The following examples show how the Bulgarian verb ***боря, боря се /****fight, strive, struggle****/* is inserted in the database through the administrative module of the web application (especially the information about its transitivity, semantic features, and conjugation type) (Fig. 3), and further, how this information is displayed on the screen to the casual/end-user (Fig. 4).

The web-based casual/end-user interface is bilingual. The user can choose the input language (Bulgarian or Polish) with possibilities to search for translation in both directions: Bulgarian-to-Polish, or Polish-to-Bulgarian. The Bulgarian-to-Polish translation will display the whole information existing in the dictionary entry but the opposite translation will be made only from the main senses of the Bulgarian headwords (Fig. 5).

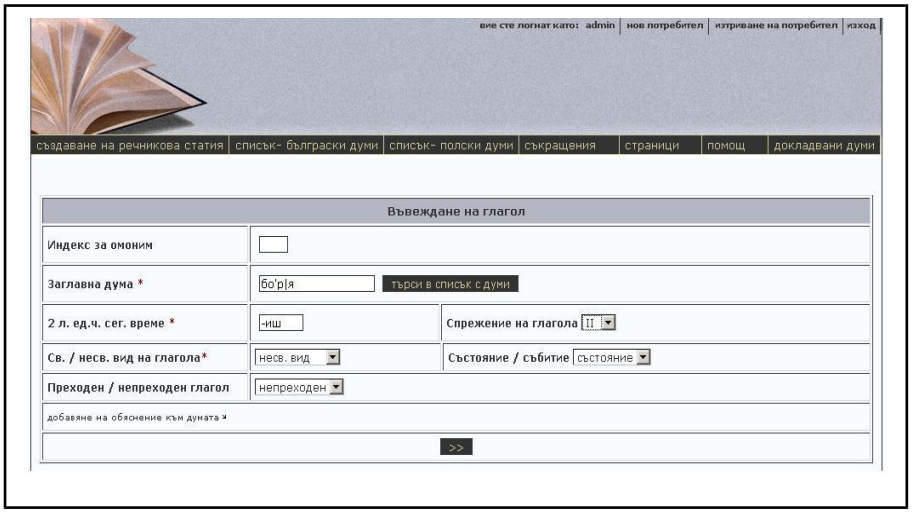


Fig. 3. Bulgarian-Polish Online Dictionary – insert a verb (**боря, боря се /**fight, strive, struggle**/**)

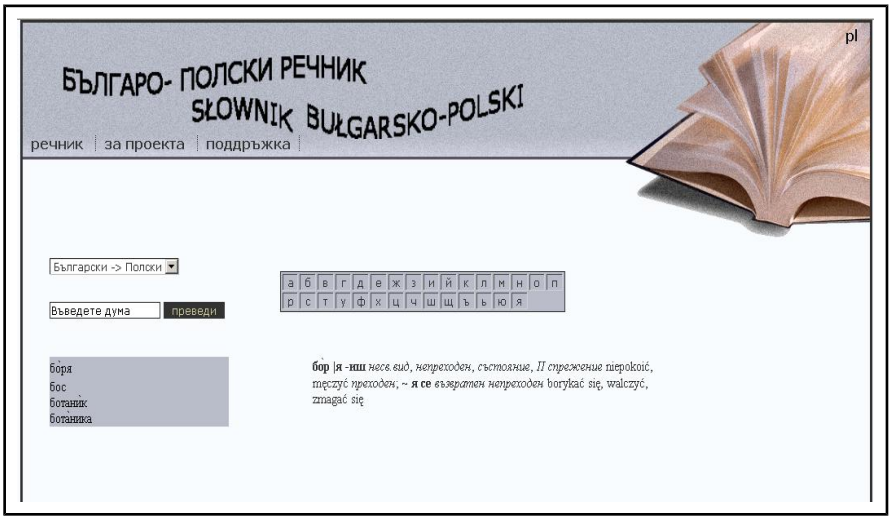


Fig. 4. Bulgarian verb "**боря, боря се**" /fight, strive, struggle/

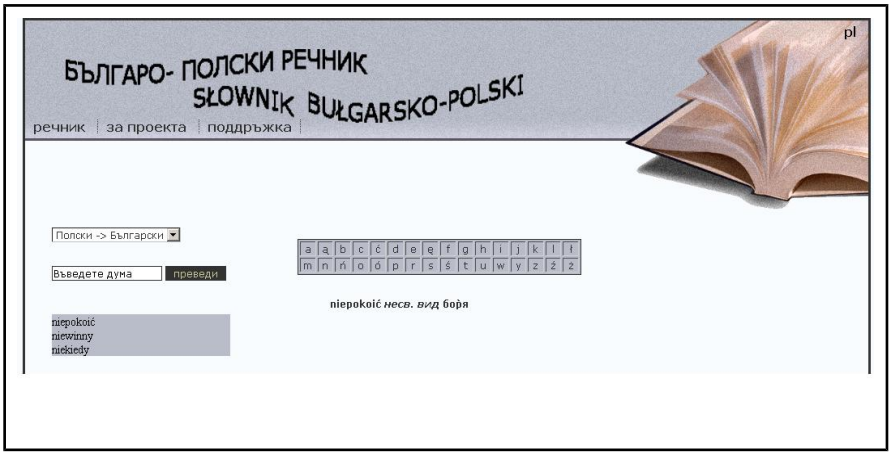


Fig. 5. Polish verb **"*niepokoić*"** /fight, strive, struggle/

8 Conclusion and Future Work

The paper presents briefly the multilingual resources with Bulgarian developed at Mathematical Linguistics Department of the IMI-BAS. These resources will be widely applicable to the contrastive studies in a multilingual context, in the field of human and machine translation, as well as in education. Future implementation will include a presentation on the web of some Bulgarian-Polish data, especially of the aligned Bulgarian-Polish corpus – a web page with an appropriate trilingual interface in Bulgarian,

Polish, and English for easy access to the corpus, is envisaged. Finally, I would like to thank all colleagues and my students with whom I worked throughout the years for the development of the Bulgarian annotated language resources.

References

Dimitrova, L., Pavlov, R., and Simov, K. (2002). The Bulgarian Dictionary in Multilingual Data Bases. *Cybernetics and Information Technologies,* 2 (2):12–15.

Dimitrova, L., Pavlov, R., Simov, K., Sinapova, L. (2005). Bulgarian MULTEXT-East Corpus – Structure and Content. *Cybernetics and Information Technologies,* 5(1):67–73.

Dimitrova, L. and Koseska. V. (2009). Bulgarian-Polish Corpus. *Cognitive Studies/Études Cognitives*, 9:133–141.

Dimitrova, L., Koseska-Toszewa, V., and Roszko, D., Roszko, R. (2010). Application of Multilingual Corpus in Contrastive Studies (on the example of the Bulgarian-Polish-Lithuanian Parallel Corpus). *Cognitive Studies/Études Cognitives,* 10:217–240.

Dimitrova, L. and Garabík, R. (2011). Bulgarian-Slovak Parallel Corpus. In Majchráková, D. and Garabík, R., editors, *Proceedings of the Sixth International Conference Natural Language Processing, Multilinguality (SLOVKO 2011)*, pages 44–50, Brno, Czech Republic. Tribun.

Dimitrova, L. (2009). From Electronic Corpora to Online Dictionaries (on the example of Bulgarian Language Resources). In Levická, J. and Garabík, R., editors, *Proceedings of the Fifth International Conference: NLP, Corpus Linguistics, Corpus Based Grammar Research. (SLOVKO 2009)*, pages 78–92, Brno, Czech Republic. Tribun.

Dimitrova, L., Erjavec, T., Ide, N., Kaalep, H. J., Petkevič, V., and Tufis D. (1998). Multext-East: Parallel and Comparable Corpora and Lexicons for Six Central and Eastern European Languages. In Boitet, Ch. and Whitelock, P., editors, *Proceedings of COLING-ACL'98: 17th International Conference on Computational Linguistics and 36th Annual Meeting of the Association for Computational Linguistics, Montréal, Québec, Canada,* pages 315–319, San Francisco, U.S.A.

Dimitrova, L., Panova, R., and Dutsova, R. (2009a): Lexical Database of the Experimental Bulgarian-Polish online Dictionary. In Garabík, R., editor, *Proceedings of MONDILEX Third Workshop Metalanguage and Encoding Scheme Design for Digital Lexicography*, pages 36–47, Brno, Czech Republic. Tribun.

Dimitrova, L., Koseska, V., Dutsova, R., and Panova, R. (2009b). Bulgarian-Polish online Dictionary – Design and Development. In Koseska, V., Dimitrova, L., and Roszko, D., editors, *Proceedings of MONDILEX Fourth Workshop Representing Semantics in Digital Lexicography,* pages 76–88, Warsaw, Poland.

Andreychin, L. et al. (1994). *Bulgarian Explanatory Dictionary / Dictionary of the Bulgarian Language*, Sofia, Bulgaria. Nauka i Izkuvstvo Publishing House.

Sławski, F. (1987). *Podręczny słownik Bułgarsko-Polski z suplementem,* Warsaw, Poland. Wydawnictwo Wiedza Powszechna.

`http://aune.lpl.univ-aix.fr/projects/multext-east/`

`http://www.itri.brighton.ac.uk/projects/concede/`

`http://www.mondilex.org`

`http://www.tei-c.org/index.xml`

Automatic Text Processing and Deeply Annotated Text Corpora of Russian: Interaction and Mutual Impact[1]

Leonid Iomdin

A. A. Kharkevich Institute for Information Transmission Problems,
Russian Academy of Sciences, Moscow, Russia

Abstract. This paper gives a brief overview of the existing annotated text corpora of Russian. The main focus, however, will be on SynTagRus, a corpus of Russian texts annotated with dependency-type syntactic structures, specific word senses, and lexical functions. It will be shown how statistical data collected from the corpus are used to improve lexical and syntactic disambiguation during automatic text parsing. Other uses of SynTagRus will be outlined, including construction of dependency parsers by machine-learning techniques and regression testing of the rule-based parser.

1 Annotated Text Corpora of Russian: What are They?

Corpus linguistics in Russia has been developing intensively and extensively over the last decade, giving rise to a rich variety of monolingual, bilingual, and multilingual corpora of texts. The most important Russian corpora are incorporated into the so-called "Национальный корпус русского языка" (National Corpus of the Russian Language), abbreviated as НКРЯ or NCRL, and available through the portal `www.ruscorpora.ru`.

In addition to this corpus, several other large and even supersize corpora are being collected, including the so-called General Internet Corpus of Russian tagged with syntactic dependencies, which will be compiled automatically by a statistical parser (see Belikov, Selegei, and Sharov, 2012) and an Open Corpus (Bocharov et al., 2012) developed manually since 2009 on the principles of crowdsourcing (`www.opencorpora.org`). Since the latter two projects are far from being completed, we will confine ourselves to the discussion of the resources constituting the National Corpus of the Russian language.

The National Corpus is currently hosted by Yandex, Russia's largest search engine (which is to be changed in the near future, as a special noncommercial partnership is expected to be founded for the purpose of managing, developing, and hosting Russian corpora). It consists of a number of independently created corpora, listed below:

1) The **main corpus** provided with morphological annotation, which is comprised of over 300 million words belonging to written texts of a variety of genres starting from the 18th century. In most cases, the annotation is morphologically ambiguous: a word may have more than one set of morphological tags corresponding to different parts of speech and/or morphological features. The main corpus contains a subcorpus of texts with resolved

[1] This work has been partially supported by the Russian Foundation of Basic Research (grant No. 10-06-00478) and a grant from the Presidium of Russian Academy of Sciences within the program of basic research in corpus linguistics. The author expresses his gratitude to both organizations.

morphological ambiguity, which counts over seven million words. Lexical ambiguity is not resolved independently; words can be considered to be lexically disambiguated only if the word senses belong to different parts of speech, as in the adjective *slepoj* 'blind' vs. the noun *slepoj* 'blind person'. The corpus now contains partial semantic annotation presented in the form of simple semantic features of words.

2) The **syntactic corpus** provided with morphological and syntactic annotation comprises about 770 000 words (over 52 000 sentences) and provides a syntactic dependency structure for all sentences. The corpus is fully disambiguated, both morphologically and syntactically, so that every word is supplied with one part-of-speech tag and a unique set of morphological features, while every sentence has only one dependency tree structure.

3) The **newspaper corpus** is built on the same principles as the main corpus and comprises the articles of seven mass media since the year 2 000 (four newspapers published in Moscow and three electronic mass media), which now counts over 170 million words.

4) Several aligned **parallel corpora** (English-Russian, Russian-English, German-Russian, Ukrainian-Russian, Russian-Ukrainian, Belarussian-Russian, Russian-Belarussian, and multilingual).

5) A **dialectal corpus** is composed of samples of dialect speech from various regions of Russia, and presented in quasi-standard orthography. The corpus disregards phonetic variation but demonstrates morphological, lexical and syntactic peculiarities of regional and dialectal usage.

6) A **poetry corpus** is primarily comprised primarily of Russian poetic works of the 18th and 19th centuries, supplemented by work of a number of 20th century poets. In addition to non-disambiguated morphological tagging built on the same principles as that of the main corpus, the texts are provided with information on the poetic meters used in them.

7) An **educational corpus**, intended for learners' of Russian offers disambiguated morphological tagging for simple prosaic texts.

8) A **corpus of Spoken Russian** consists of transcripts of samples of public and private oral speech of the 20th and 21st centuries as well as film transcripts. The corpus is supplied with morphological and partial semantic annotation.

9) An **accentological corpus** (corpus of history of Russian word stresses).

10) A **Church Slavonic corpus** comprises modern liturgical texts of the 19th and 20th centuries, as well as older religious and biblical texts. The corpus enables the search of words in three orthographic systems.

11) A **Multimedia corpus** is composed of fragments of films released since 1930 and presented as video files, audio files and textual transcripts, as well as lists of gestures present in these fragments.

2 SynTagRus treebank: Basic Facts

The Syntactic Corpus, listed as item two of the NCRL above, is in fact a subcorpus of the SynTagRus Treebank developed by the Laboratory of Computational Linguistics of the Institute of Information Transmission Problems, Russian Academy of Sciences (Boguslavsky et al., 2000; 2002). SynTagRus is created independently of all NCRL modules and is sent to the ruscorpora.ru portal in order to facilitate public access to it. Normally, the syntactic corpus is updated at this portal 3-4 times a year as SynTagRus is progressing.

The main material distinction between SynTagRus and NCRL Syntactic corpus is that the latter does not show word senses of lexical units of which corpus sentences are composed, or lexical functional annotation.

Currently, SynTagRus contains over 52 000 sentences (ca. 770 000 words) belonging to texts of a variety of genres (contemporary fiction, popular science, newspaper, journal and Wikipedia articles dated between 1960 and 2012, texts of online news etc.) and is steadily growing.

A decade ago, when the development of SynTagRus started, the authors chose a dependency-based annotation scheme, considering the fact that Russian is a language with relatively free word order. This annotation has much in common with that of the Prague Dependency Treebank (see e.g. Hajič et al., 2000), but in contradistinction to PDT, it is based on the Meaning ⇔ Text theory by Igor Meľčuk, especially as concerns the inventory of syntactic relations used (see e.g. Meľčuk, 1988).

SynTagRus (and its "mirror" in NCRL) is so far the only corpus of Russian supplied with comprehensive morphological annotation and syntactic tagging in the form of a complete dependency tree provided for every sentence. An example of such a tree is given in Fig. 1 below, which is the screenshot of the structure for the sentence

(1) *Наибольшее возмущение участников митинга вызвал продолжающийся рост цен на бензин, устанавливаемых нефтяными компаниями* (It was the continuing growth of petrol prices set by oil companies that caused the greatest indignation of the participants of the meeting):

In the dependency tree, nodes represent words (lemmas), annotated with parts of speech and morphological features, while arcs are labeled with syntactic dependency types. There are about 75 different dependency labels used in the treebank. Of these, about a half is adopted from the Meaning ⇔ Text theory, while the rest were introduced by the developers of the ETAP-3 parser (see below) and SynTagRus itself. Dependency labels used in Fig. 1 are as follows:

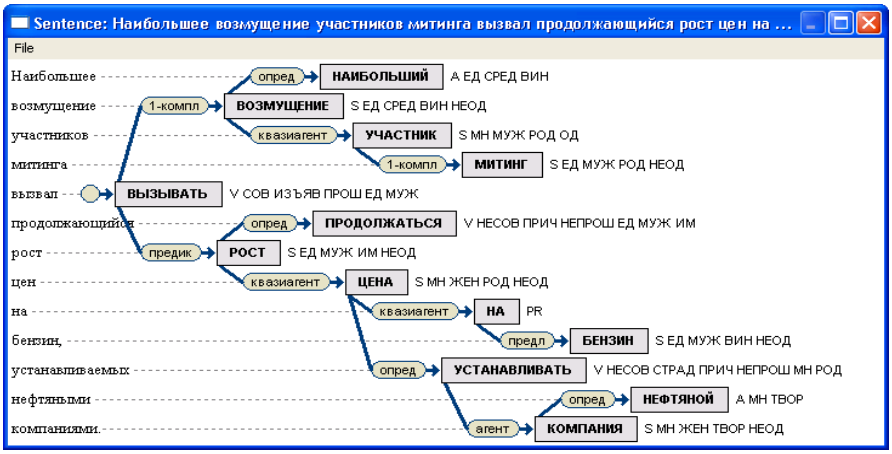


Fig. 1. A syntactically annotated sentence from the SynTagRus Treebank

1) **предик** (predicative) which, prototypically, represents the relation between the verbal predicate as head and its subject as daughter; 2) **1-компл** (first completive) which denotes the relation between a predicate word as head and its direct complement as daughter; 3) **агент** (agentive) which introduces the relation between a predicate word (verbal noun or verb in the passive voice) as head and its agent in the instrumental case; 4) **квазиагент** (quasi-agentive) which relates any predicate noun as head with its first syntactic actant as daughter whenever this latter is not eligible for being qualified as the noun's agent; 5) **опред** (modificative) which connects a noun head with an adjective/participle daughter if this latter serves as an adjectival modifier to the noun; 6) **предл** (prepositional) which accounts for the relation between a preposition as head and a noun as daughter.

In the dependency tree of SynTagRus one token normally corresponds to one node. There are, however, certain exceptions:

- Composite words like *пятидесятиэтажный* 'fifty-storeyed', where one token corresponds to two or more nodes;
- Multiword expressions like по крайней мере 'at least', where several tokens correspond to one node;
- So-called **phantom** nodes for the representation of hard cases of ellipsis, or gapping, which do not correspond to any particular token in the sentence (cf. *Я предпочитаю чай, а жена кофе* 'I prefer tea and (my) wife coffee'), which is expanded into *Я предпочитаю чай, а жена предпочитает*$_{\text{PHANTOM}}$ *кофе* 'I prefer tea and (my) wife prefers $_{\text{PHANTOM}}$ coffee.'

Importantly, dependency trees of SynTagRus allow non-projective links. Roughly, the share of sentences having such links is close to 10 % of all sentences. In most cases, such sentences have only one link that violates projectivity, although there are a few exceptions.

Syntactic annotation in SynTagRus is performed semi-automatically – every sentence is first processed by the rule-based Russian parser of an advanced NLP system, ETAP-3 (see e.g. Apresjan et al., 2003; Iomdin et al., 20*12*), and then edited manually by at least two linguist experts who handle errors of the parser as well as cases of ambiguity that cannot be reliably resolved without world knowledge. At present, the parser is also unable to automatically create adequate structures with gapping.

The editors work in a sophisticated software environment, Structure Editor, or StrEd, specifically developed for the purpose of treebank creation (see Iomdin and Sizov, 2009).

The parser now processes raw sentences without prior part-of-speech tagging, even though at present a statistical POS tagger for Russian is being developed which will be integrated shortly into the system.

Morphological annotation in SynTagRus is based on a comprehensive morphological dictionary of Russian that counts about 130 000 entries (over 4 million word forms). The ETAP-3 morphological analyzer uses this dictionary to produce morphological annotation of words belonging to the corpus, including lemma, part-of-speech tag, and additional morphological features, depending on the part of speech: (a) animacy, gender, number and case (for nouns, adjectives, numerals and participles); (b) degree of comparison and attenuation (for adjectives and adverbs); (c) short form (for adjectives and participles); (d) representation (for verbs – this is the category whose values are finiteness, infinitive, participle, and adverbial participle ("деепричастие"); (e) aspect, tense, mood, person, voice (for verbs), and (f) composite form (for nouns, adjectives and numerals): the latter is ascribed to word forms like *угле-* or *физико-* which can appear in composites like *угледобыча* 'coal extraction' or *физико-химический* 'physical and chemical'.

The current version of SynTagRus contains partial lexical functional annotation.

For collocations that could be presented with the apparatus of lexical functions of the Meaning ⇔ Text theory (see e.g. Apresjan et al., 2007), the tagging includes information on values and attributes of such lexical functions. In all, over a hundred simple and compound lexical functions are used in the annotation.

In Fig. 2, the SynTagRus structure for the sentence...

(2) *Техника осуществления подобных проектов до сих пор вызывает удивление у многих специалистов* 'The technique of implementing such projects is still surprising (lit. is still causing surprise in) many specialists'

…contains information that the verb *вызывать* 'cause' is the value of the lexical function $CausOper_1$ for the keyword *удивление* 'surprise', and the preposition *у* 'at, in' is the necessary context for such a collocation to appear in the text. For more details on lexical and functional annotation in SynTagRus (see Timoshenko et al., 2009).

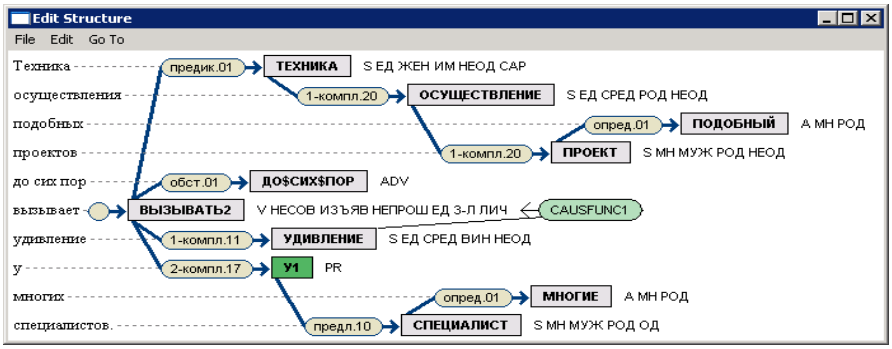


Fig. 2. A structure from the SynTagRus Treebank supplied with lexical functional annotation

Another important feature of SynTagRus is that it provides information on the lexical meaning, or word senses of words appearing in the treebank sentences. Word senses are defined according to the combinatorial dictionary of ETAP-3. This fact, regretfully, restricts the use of such information for the public; to improve the situation, the developers strive to match the word senses with those of other resources, including WordNet.

To give an example, consider two sentences that feature different word senses of the verb *толковать: толковать 1* 'interpret' (sentence 3), and *толковать 2* 'converse' (sentence 4):

(3) *Пока же исследователи по-разному* ***толкуют*** *"первоисточник", а в качестве доказательства своей правоты приводят отдельные археологические находки последних лет* 'So far, the researchers interpret differently the primary source, and as a proof of their being right they present isolated archeological findings of the recent years'

(4) ***Толкуют*** *о сооружении местного Сити, почти как в Москве, и, разумеется, с высоченным небоскребом напротив Смольного на невском правобережье.* 'They talk about the construction of a local City, almost like in Moscow and, naturally, with a very high skyscraper opposite Smolny on the Neva river bank.'

The structures for (3) and (4) are presented in Fig. 3 and 4, respectively. In addition to the ambiguous word *толковать,* we can see a few other instances of word sentence disambiguation: *пока 2* 'so far' (an adverb opposed to the conjunction *пока 1* 'while', *а 1* 'and' (a conjunction opposed to a particle), and a few others.

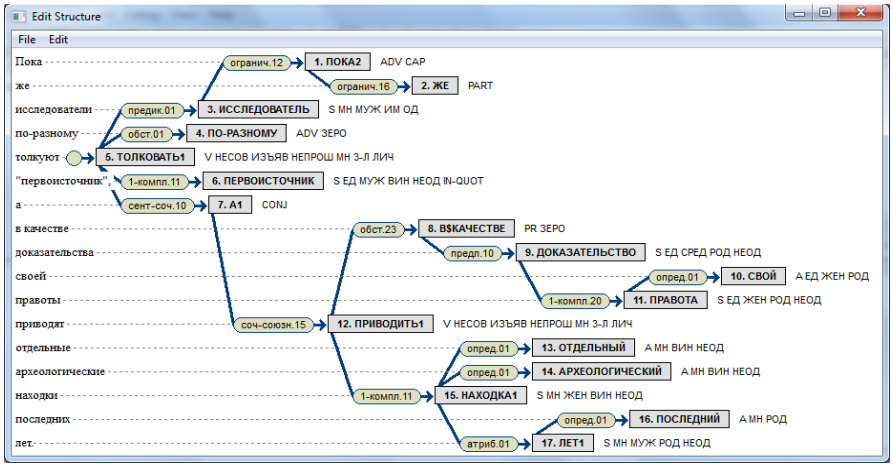


Fig. 3. The structure for sentence (3) showing word senses of polysemous words

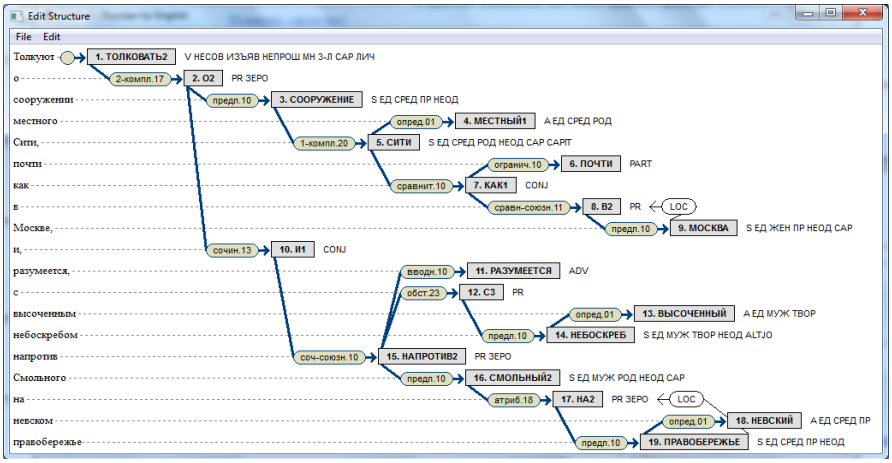
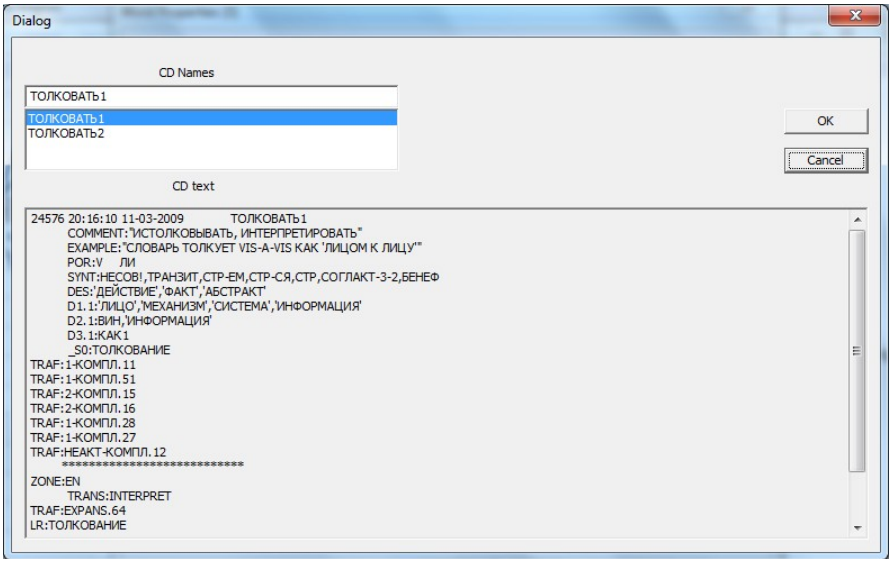
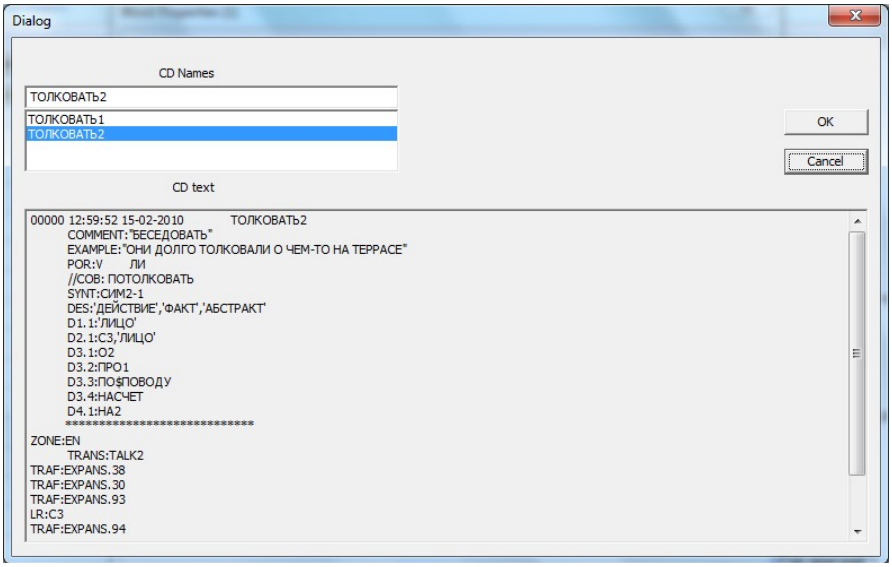


Fig. 4. The structure for sentence (4) showing word senses of polysemous words

The Structure Editor allows the developers and the editors to see the respective entries of the dictionary by clicking on the ambiguous word and verify if the parser has chosen the right word sense:





3 SynTagRus and ETAP parser: Living Together

It goes without saying that SynTagRus is primarily a valuable linguistic resource which can be used to acquire new knowledge of the language. I will briefly present an example of such knowledge.

The Russian syntax has a specific durative construction which introduces the duration of an action, process, or state by adjoining a noun phrase in the accusative to a verb denoting this action, process, or state: *Он жил здесь целый год* 'lit. 'He lived here the whole year' or *Он писал статью два месяца* lit. 'He wrote the article two months'. Russian theoretical grammars describe the construction in detail, emphasizing the fact that the verb participating in it can only stand in the imperfective voice (or else belong to a particular class of perfective perduratives like *поработать* ≈ 'work for some time' or *высидеть* ≈ 'sit for some time until the situation changes'): it is impossible to say something like **Он написал статью два месяца* ≈ 'He finished writing the article two months'. Surprisingly, a search for the durative construction in SynTagRus brought unexpected results. On the one hand, examples were found of nonperdurative perfective verbs that co-occur with the accusative noun phrase of duration (like *Он отдохнёт часик и поедет дальше* 'he will rest an hour and will then drive on'). On the other hand, there were instances where the syntactic head of the construction was not a verb at all but an adjective, which theoretical grammars disregard at all: *Абонент недоступен вот уже два часа* 'The subscriber has been unreachable for two hours now' or *Я занят весь следующий месяц* 'I will be busy the whole next month'. Moreover, there are sentences in which such an adjective is not even the main predicate of the sentence, cf. *К старой тетке // Четвертый год больной в чахотке, // Они приехали теперь* ≈ 'They now came to visit an old aunt who had been sick with tuberculosis for over three years' (Alexander Pushkin). By producing such sentences SynTagRus, helped reconsider the conditions in which the durative construction can appear. Most interestingly, this became possible only because the treebank was constructed semi-automatically: obviously, the parser could not produce anything that contradicts the grammar, so the right parses were produced by the editors, who knew the idea of the construction and manually corrected the parses produced by the computer!

In addition to this, SynTagRus is actively used as a computational resource. Being considered the gold standard, the treebank is used to collect various statistical data and to create training sets for machine learning.

To be more specific, within ETAP-3 SynTagRus provides the statistics of the different syntactic constructions, lexical co-occurrences, patterns of ambiguities etc., which is used at several points of the parsing algorithm, provided the statistical component of ETAP-3 is activated.

Further, it serves as an efficient and accurate evaluation resource, which is used to evaluate the performance of ETAP-3 parser and in this way find and resolve some of the system's bottlenecks.

Finally, it is used for regression testing of ETAP-3. This is done in the following way. From time to time (normally every other week) ETAP is run on the whole corpus. Sentences that receive parses exactly equivalent to those stored in the corpus (this constitutes between 30 and 35 percent of the bulk of the corpus) are selected as basis for regression testing. ETAP is then run on this test set to see if changes introduced in the dictionary, rules, or software affected the state of the test set. Regression testing has proven helpful in ensuring the stability of the parser and eventually improving it. Regression testing helps improve the SynTagRus itself – sometimes the discrepancies in parses detected by regression test runs point to erroneous annotation in the corpus, which is then corrected.

I will finish by mentioning the fact that SynTagRus was used to develop a successful automatic parser (Nivre, Boguslavsky, and Iomdin, 2008). This parser is expected to be used in the construction of the General Internet Corpus of Russian mentioned above. The future will show how successful this parser can be.

References

Apresjan, J., Boguslavsky, I., Iomdin, L., Lazursky, A., Sannikov, V., Sizov, V., and Tsinman, L. (2003). ETAP-3 linguistic processor: A full-fledged NLP implementation of the MTT. In *Proceedings of the First International Conference on Meaning-Text Theory*, pages 279–288, Paris, France.

Apresjan, J., Boguslavsky, I., Iomdin, L., and Tsinman, L. (2007). Lexical Functions in Actual NLP-Applications. In Wanner, L., editor, *Selected Lexical and Grammatical Issues in the Meaning–Text Theory,* pages 199–230. John Benjamins.

Belikov, V., Selegei, V., and Sharov, S. (2012). Preliminary Considerations towards Developing the General Internet Corpus of Russian / Беликов В. И., Селегей В. П., Шаров С. А. Пролегомены к проекту Генерального интернет-корпуса русского языка (ГИКРЯ). In Компьютерная лингвистика и интеллектуальные технологии. По материалам ежегодной Международной конференции «Диалог», pages 37–50, Moscow, Russia.

Bocharov, V., Alexejeva, S., Granovsky, D., Ostapuk, N., Stepanova, M., and Surikov, A. (2012). Сегментация текста в проекте «Открытый корпус» / Text Segmentation in the Open Corpora Project. In Компьютерная лингвистика и интеллектуальные технологии. По материалам ежегодной Международной конференции «Диалог», pages 51–60, Moscow, Russia.

Boguslavsky, I., Grigorieva, S., Grigoriev, N., Kreidlin, L., and Frid, N. (2000). Dependency Treebank for Russian: Concept, Tools, Types of Information. In *Proceedings of COLING-2000,* pages 987–991, Saarbrücken, Germany. Bernd Bohnet.

Boguslavsky, I., Chardin, I., Grigorieva, S., Grigoriev, N., Iomdin, L., Kreidlin L., and Frid, N. (2002). Development of a dependency treebank for Russian and its possible applications in NLP. In *Proceedings of LREC-2002,* pages 852–856.

Hajič, J., Böhmová, A., Hajičová, E., Vidová, B., and Hladká, Z. (2012). The Prague Dependency Treebank: A Three-Level Annotation Scenario. In Abeillé A., editor, *Treebanks: Building and Using Parsed Corpora*, pages 103–127, Amsterdam, Netherlands. Kluwer.

Iomdin L. and Sizov, V. (2009). Structure Editor: a Powerful Environment for Tagged Corpora. In *Research Infrastructure for Digital Lexicography: Proceedings of MONDILEX Fifth Open Workshop,* pages 1–12, Ljubljana, Slovenia.

Iomdin, L., Petrochenkov, V., Sizov, V., and Tsinman, L. (2012). ETAP parser: state of the art. *Dialog 2012. Computational Linguistics and Intellectual Technologies. International Conference,* 11(18):830–843. Moscow, Russia. RGGU Publishers.

Meľčuk, I. (1988). *Dependency Syntax: Theory and Practice.* State University of New York Press.

Nivre, J., Boguslavsky, I., and Iomdin, L. (2008). Parsing the SYNTAGRUS Treebank of Russian. In *Proceedings of the 22nd International Conference on Computational Linguistics. (Coling 2008),* pages 641–648, Manchester, UK.

Timoshenko, S., Boguslavsky, S., Iomdin, L., and Frolova, T. (2009). Development of the Russian Tagged Corpus with Lexical and Functional Annotation. In *Proceedings of Metalanguage and Encoding Scheme Design for Digital Lexicography: MONDILEX Third Open Workshop,* pages 83–90, Bratislava, Slovakia.

Počítačové kartografovanie nárečí v Slovanskom jazykovom atlase[1]

Pavol Žigo

Jazykovedný ústav Ľ. Štúra, Slovenská akadémia vied, Bratislava, Slovensko

Abstract. In the paper the author presents a proportion of Slovak dialectologists in a project of Slavic Linguistic Atlas and describes prospects of a computer support using a *MAPola* programme. The text includes samples of computer elaborations from the volume of lexical series of the Slavic Linguistic Atlas focused on Agriculture.

Významným medzinárodným projektom v oblasti lingvistickej geografie je projekt *Slovanský jazykový atlas,* zahrnujúci výsledky výskumu nárečí všetkých slovanských jazykov. Je realizáciou myšlienky, ktorú prezentovali v r. 1929 na prvom slavistickom zjazde v Prahe jazykovedci Antoine Meillet (1866 – 1936) a Lucien Tesniére (1893 – 1954). Takýto rozsiahly projekt od začiatkov predpokladal riešenie množstva teoretických aj metodologických otázok a jednotný prístup medzinárodného kolektívu bádateľov pri interpretácii množstva areálovo diferencovaných javov. Komplexné spracovanie nárečí jednotlivých slovanských jazykov nadobudlo konkrétnu realizáciu až na IV. slavistickom zjazde v Moskve v r. 1958. Na základe spoločne prijatého rozhodnutia interpretovať výsledky komplexného slavistického výskumu metódou lingvistickej geografie vznikla Medzinárodná komisia pre Slovanský jazykový atlas. Komisia vypracovala bodovú sieť, ktorú tvorí pôvodných 853 výskumných lokalít od Jadranského mora po Ural, pričom sa do projektu zahrnuli aj slovanské lokality, ktoré sú administratívne začlenené do štátnych útvarov s majoritným neslovanským obyvateľstvom. Do južnoslovanského areálu patria slovinské nárečia (18 lokalít + 3 lokality ležiace na území Talianska), chorvátske nárečia (a to aj chorvátske nárečia na území Rakúska), nárečia Bosny a Hercegoviny, srbské a čiernohorské nárečia, macedónske nárečia (a to aj na území Albánska a Grécka) a bulharské nárečia (3 lokality na území Grécka, 1 na území Turecka). Západoslovanský areál tvoria české a moravské nárečia, slovenské nárečia (z toho 3 lokality slovenských nárečí sú na území Maďarska), lužickosrbské nárečia a 98 lokalít poľských nárečí. Východoslovanskú jazykovú skupinu tvorí sieť 74 lokalít bieloruských nárečí, 121 lokalít ukrajinských nárečí (ďalších 5 lokalít ukrajinských nárečí leží na území Rumunska a 3 lokality na území Moldavska), 322 lokalít ruských nárečí (1 lokalita na území Litvy, 1 lokalita na území Lotyšska, 1 lokalita na území Estónska).

Vo všetkých lokalitách explorátori vyplnili dotazník obsahujúci 3 454 skúmaných javov z oblasti hláskoslovia, tvaroslovia, lexiky, tvorenia slov, sémantiky, prozódie a syntaxe. Pri vypĺňaní dotazníka sa uplatnili zásady osobitne fonetickej transkripcie (každému zvuku zodpovedá osobitný znak). Na zasadnutí Medzinárodnej komisie pre Slovanský jazykový atlas pri Medzinárodnom komitéte slavistov, ktoré sa konalo na 6. medzinárodnom slavistickom zjazde vo Varšave r. 1973 jeho účastníci skonštatovali, že terénny výskum vo

[1] Príspevok vznikol v rámci riešenia projektu VEGA 2/5036/25 Slovanský jazykový atlas.

všetkých lokalitách je ukončený a získaný terénny materiál možno spracúvať a pripravovať na publikovanie. Od tohto medzníka sa na základe dohody začali prípravy na vydávaní dvoch sérií Slovanského jazykového atlasu: hláskoslovno-gramatickej, v ktorej sa interpretujú javy z oblasti hláskoslovia, prozódie, tvaroslovia a skladby slovanských nárečí, v druhej – lexikálno-slovotvornej sérii sa spracúvajú javy z oblasti tvorenia slov, lexiky a sémantiky. Na základe takejto perspektívy sa aj prípravné práce skoncentrovali do dvoch sekcií s pracovnými názvami hláskoslovná a lexikálna. Osobitné postavenie má morfonologická sekcia, ktorej hlavným cieľom je príprava rekonštruovaných podôb kartografovaných štruktúr do praslovančiny, na základe čoho možno v rozsiahlych slavistických dimenziách interpretovať spoločné východiskové štruktúry sledovaných nárečových javov najmä v lexikálno-slovotvornej sérii. Bez takejto prípravy nebolo a nie je mysliteľné kartografické spracovanie veľkého množstva nárečových javov rozsiahleho skúmaného areálu[2].

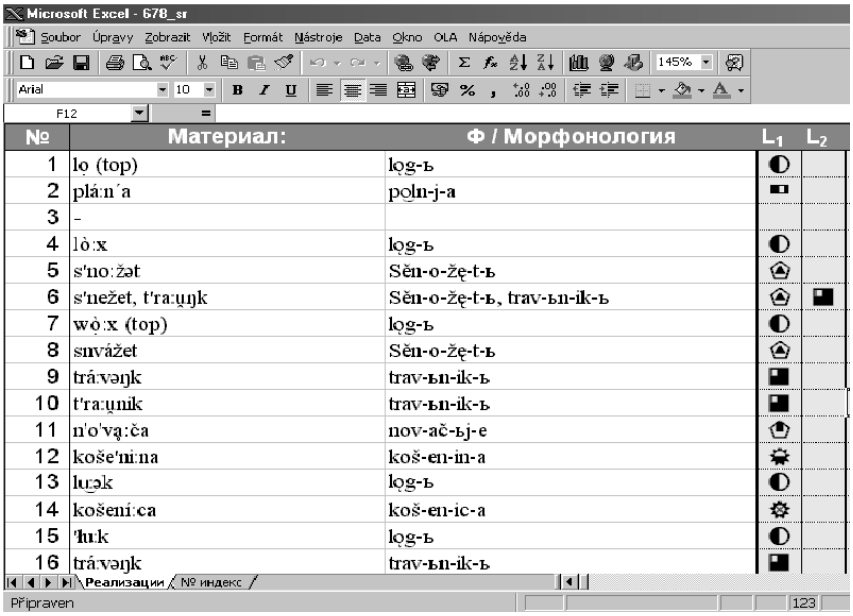
Všetky vydané zväzky sa v autorských komisiách pripravovali klasickou metódou: podkladové materiály a rukopisy sa pripravovali v strojopisnej podobe, mapy sa maľovali ručne pomocou šablónky a po niekoľkonásobných revíziách a korekciách sa rukopis odovzdával na spracovanie do vydavateľstva. Pri prepisovaní fonetickej transkripcie a prekresľovaní máp do vydavateľskej podoby vznikalo množstvo chýb, ktoré si vyžadovali ďalšie niekoľkonásobné korektúry a úpravy. Rozvoj moderných technológií a kybernetizácia priniesli množstvo pozitívnych prvkov aj do spracovania materiálov Slovanského jazykového atlasu (v jazykovednej slavistike známy pod skratkou OLA podľa podoby *Общеславянский лингвистический атлас*).

Napredovanie moderných metód viedlo dialektológov k príprave počítačovej podpory spracovania materiálov Slovanského jazykového atlasu do takej miery, že na zasadnutí Mezinárodnej komisie v slovinskom Strunjane v r. 2000 vznikla Komisia pre počítačové spracovanie materiálov Slovanského jazykového atlasu. Úlohou novovzniknutej komisie

[2] Po náročných prípravných prácach na dotvorení bodovej siete a prácach na dotazníku vyšli tieto zväzky Slovanského jazykového atlasu: v hláskoslovno-gramatickej sérii Zv. 1. Reflexy *ě. Red. B. Vidoeski – P. Ivić. Belehrad, 1988; Zv. 2a. Reflexy *ę. Red. V. V. Ivanov. Moskva, 1990; Zv. 2b. Reflexy *ǫ. Red. J. Basara. Vroclav, 1990; Zv. 3. Reflexy ьг, ьl, гъ, лъ. Red. J. Basara. Varšava, 1994; Zv. 4.a. Red. Z. Topolińska. Striednice ъ, ь. Skopje, 2004; Zv. 4.b: vokalizácia ь, ъ – chorvátska komisia. Na rozličnom stupni rozpracovania majú jednotlivé národné komisie tieto zväzky hláskoslovno-gramatickej série:, Zv. 5: metatézy likvíd – česká komisia, Zv. 6: vývin *o – ruská komisia, Zv. 7: vývin *e – ruská komisia, Zv. 8: vývin *i, *y, *u – slovinská komisia. V lexikálno-slovotvornej sérii doteraz vyšli tieto zväzky OLA: Zv. 1. Živočíšna ríša. Red. R. I. Avanesov. Moskva, 1988; Zv. 2. Chov domácich zvierat. Red. B. Falińska. Varšava, 2000; Zv. 3. Rastlinstvo. Red. A. I. Padlužny. Minsk, 2000; Zv. 8. Povolania a spoločenský život. Red. J. Basara – J. Siatkowski. Varšava, 2003, Zv. 6. Domácnosť a príprava stravy – ruská komsia; Zv. 4. Poľnohospodárstvo – slovenská komisia (poradie vydávania jednotlivých zväzkov série nemusí byť s ohľadom na náročnosť problematiky a finančné možnosti riešiteľského národného koletívu v súlade s číslovaním zväzkov). Na rozličnom stupni rozpracovania majú jednotlivé národné komisie tieto zväzky lexikálno-slovotvornej série: Zv. 5. Doprava a komunikácia. Ľudová technika. Staviteľstvo – ukrajinská komisia; Zv. 7. Odev a obuv. Hygiena a liečiteľstvo – lužickosrbská komisia. Pripravujú sa materiály z tematického okruhu Človek do 9. zväzku, ktorého prípravu prevzala poľská komisia, materiály z tematického okruhu Rodinné vzťahy do 10. zväzku pripravuje bulharská komisia.

bolo hľadať možnosti kybernetizácie materiálov OLA, zefektívniť prácu všetkých účastníkov projektu na jednotlivých pracoviskách národných komisií použitím jednotnej metódy spracovania materiálov získaných terénnym výskumom. Pozitívnu rolu zohrali v tejto fáze prípravy skúsenosti slovinských jazykovedcov v oblasti tvorby fontov fonetickej transkripcie a prípravy digitálnej matrice mapy OLA. Slovenská národná pracovná skupina vypracovala program počítačovej podpory spracovania materiálov Slovanského jazykového atlasu s názvom *MAPola 1.9*. Jeho podstatou je šablóna, ktorá obsahuje dve vzájomne prepojiteľné polia – realizačný list a číselný index. Nemennými prvkami prvého poľa – realizačného listu – je vertikálny zoznam výskumných bodov 1 – 853 v prvom stĺpci (ich názvy a administratívne začlenenie sa ako sprievodná informácia uvádzajú v uzamknutom stĺpci tabuľky), do druhého – otvoreného stĺpca tabuľky sa fonetickou transkripciou zapisujú výsledky terénneho výskumu sledovaného javu v príslušnej lokalite (označenie *Материал*, pozri obr. 1), do tretieho – otvoreného stĺpca (*Ф / Морфонология*) sa v rámci lexikálno-slovotvornej série OLA vpisujú výsledky morfonologickej rekonštrukcie sledovaného javu (pripravuje ich morfonologická sekcia a okrem etymologickej analýzy slúžia aj na zjednotenie grafickej interpretácie sledovaného javu na mape). V rámci prípravy hláskoslovno-gramatickej série atlasu možno do tohto stĺpca (fakultatívne, kvôli kontrole) uviesť striednicu sledovaného hláskoslovného javu. Do otvorených stĺpcov *L1, L2* program umiestňuje symboly, ktorými sa príslušný jav na mape kartografuje vľavo od čísla označenia skúmanej lokality (porov. detail obrázku č. 3), do stĺpcov *P1, P2* sa uvádzajú odkazové znaky, umiestnené na mape vpravo od čísla označenia skúmanej lokality (odkaz na materiál, odkaz na komentár k mape a pod.). Do stĺpca *Экстра* program umiestňuje ďalšie „ľavostranné znaky“ v prípade, že sa v skúmanej lokalite sledovaný jav vyskytuje aj v tretej, štvrtej, piatej podobe.

Realizačný list je softvérovo prepojený s listom nazvaným Číselný index (*№ индекс*). Údaje medzi týmito dvoma listami možno príkazom prenášať. V prvom riadku realizačného listu sa uvádza pozícia vybraného grafického symbolu. Označenie *L1, L2, P1, P2*, *Экстра* je prenesené z realizačného listu a znamená umiestnenie vybraného znaku na uvedenej pozícii. V druhom riadku (*Знак*) sa zo systému znakov vyberá konkrétny potrebný grafický znak (krúžok, štvorec, trojuholník, kosoštvorec..., pozri obr. 2). V treťom riadku sa v lexikálno-slovotvornej sérii OLA pre potreby morfonologickej sekcie a prípravy morfonologických legiend uvádza praslovanská – východisková – rekonštrukcia skúmaného javu. Ďalšie riadky príslušného stĺpca obsahujú čísla lokalít, ku ktorým treba príslušný symbol, resp. jav z druhého, resp. tretieho riadku stĺpca na mape zobraziť (vertikálny rozmer poľa, stĺpec, rešpektuje počet skúmaných lokalít).



Microsoft Excel - 678_sr

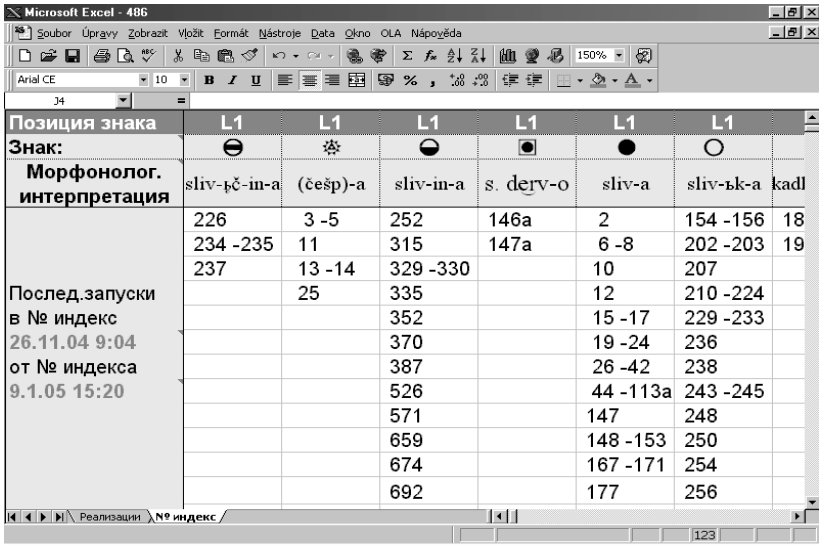
Soubor Úpravy Zobrazit Vložit Formát Nástroje Data Okno OLA Nápověda

№	Материал:	Ф / Морфонология	L_1	L_2
1	lo̯ (top)	lo̯g-ъ		
2	plá:n´a	po̯ln-j-a		
3	-			
4	lò:x	lo̯g-ъ		
5	s'no:žət	Sěn-o-žę-t-ъ		
6	s'nežet, t'ra:u̯ŋk	Sěn-o-žę-t-ъ, trav-ьn-ik-ъ		
7	wò:x (top)	lo̯g-ъ		
8	snvážet	Sěn-o-žę-t-ъ		
9	trá:vəŋk	trav-ьn-ik-ъ		
10	t'ra:u̯nik	trav-ьn-ik-ъ		
11	n'o'vą:ča	nov-ač-ьj-e		
12	koše'ni:na	koš-en-in-a		
13	lu̯ɔk	lo̯g-ъ		
14	košeni:ca	koš-en-ic-a		
15	'luk	lo̯g-ъ		
16	trá:vəŋk	trav-ьn-ik-ъ		

Реализации | № индекс

Připraven 123

Obr. 1. Realizačný list so zápisom terénneho výskumu, morfonologickou rekonštrukciou a príslušným symbolom prvých 16 lokalít (Slovinsko) z testovacej verzie mapy L 678 *lúka* 4. zväzku lexikálno-slovotvornej série *Slovanského jazykového atlasu*



Microsoft Excel - 486

Soubor Úpravy Zobrazit Vložit Formát Nástroje Data Okno OLA Nápověda

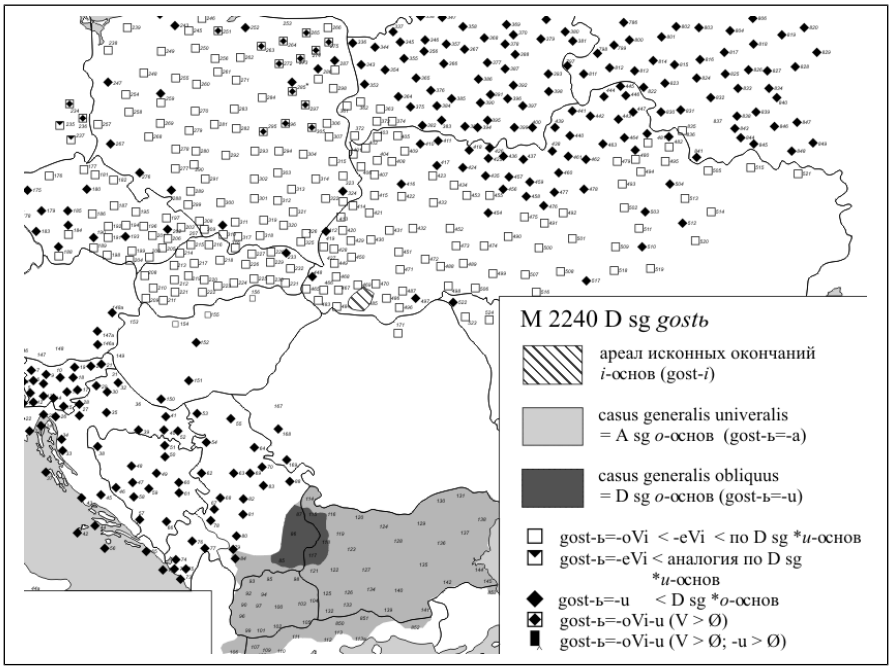
Позиция знака	L1	L1	L1	L1	L1	L1	
Знак:							
Морфонолог. интерпретация	sliv-ьč-in-a	(češp)-a	sliv-in-a	s. de̯rv-o	sliv-a	sliv-ьk-a	kadl
	226	3 -5	252	146a	2	154 -156	18
	234 -235	11	315	147a	6 -8	202 -203	19
	237	13 -14	329 -330		10	207	
Послед.запуски		25	335		12	210 -224	
в № индекс			352		15 -17	229 -233	
26.11.04 9:04			370		19 -24	236	
от № индекса			387		26 -42	238	
9.1.05 15:20			526		44 -113a	243 -245	
			571		147	248	
			659		148 -153	250	
			674		167 -171	254	
			692		177	256	

Реализации | № индекс

123

Obr. 2. Číselný index testovacej verzie mapy Sl 486 *strom slivky* so znakmi, morfonologickou rekonštrukciou tvarov a pozíciami na mape

Horizontálny rozmer číselného indexu pozostáva z takého množstva stĺpcov, koľko zistených, resp. relavantných javov treba kartografovať. Okrem vzájomného prenosu údajov z realizačného listu do číselného indexu a naopak možno na ľubovoľné požadované miesto importovať údaje z iných materiálov (súvislých textov, kartoték, indexov) alebo tieto údaje importovať do iných súborov. Túto fázu využívania programu možno označiť ako štruktúrovanú inventarizáciu nárečového materiálu, pozostávajúcu zo synchronického opisu, etymologickej rekonštrukcie, slovotvornej štruktúry na úrovni paradigmatických aj syntagmatických vzťahov.



Obr. 3. Výsek hotovej mapy s javom M 2240 – analýza substantívnej deklinácie: vývin koncoviek datívu singuláru pôvodných i-kmeňov na príklade slova *hosť*

Druhým krokom je softvérové spájanie týchto údajov z vyplnenej šablóny, ktorú tvorí realizačný list a číselný index, so súradnicami na digitálnej mape. Podstata fungovania programu spočíva v kompatibilite jednotlivých údajov šablóny a digitálnej mapy. Pred spojením údajov dvoch programov softvér poskytuje používateľovi voľbu veľkosti grafických symbolov (podľa klasického nastavenia veľkosti písma), ich farby a vzájomného usporiadania – v prípade, že sa pri jednej lokalite vyskytuje viac symbolov, možno si zvoliť ich vzájomné usporiadanie vertikálne (nad sebou) aj horizontálne (vedľa seba, pozri obr. č. 3). Tak isto možno zvoliť vzájomnú vzdialenosť symbolov, odsadenie, t. j. vzdialenosť od čísla lokality a pod. Súčasťou softvéru je algoritmus, ktorý automaticky vyrieši kolízie, ku ktorým by mohlo dôjsť vzájomným prekrývaním viacerých znakov susediacich lokalít,

resp. prekrývaním údajov na digitálnej mape (názvy miest, nežiaduce uloženie značky do mora, na rieku, hraničnú čiaru a pod.) Výsledkom interakcie spomenutých programov je nárečová mapa, ktorú možno ďalším softvérovým vybavením doplniť o izoglosy, šrafy, popisky, legendy... vo zvolenej farbe, veľkosti, type písma a pod. Údaje o symboloch, ich veľkosti, rozmiestnení a ďalšie zmeny možno využitím programu softvérovo pružne zmeniť v realizačnom liste, resp. číselnom indexe, a to buď opravou požadovaného miesta, alebo opätovnou tvorbou celej mapy. Časovo je tento proces veľmi nenáročný.

Výhodou softvérového vybavenia je najmä jeho nenáročnosť a dostupnosť východiskových produktov. Ďalšou výhodou je popri komplexnom spracovaní množstva javov na rozsiahlom území aj možnosť spracúvania jednotlivých javov na tom istom území či spracovanie ohraničeného územia. Ak by sme pri kartografovaní využili softvérovú ponuku spracovať napr. morfologický jav v riadkoch s údajmi zo slovenských nárečí, výsledkom by bola mapa so symbolmi označujúcimi v konkrétnom prípade podoby *hosťovi* (pozri na obr. 3 prázdny štvoruholník; v našom príspevku uvádzame len základný mechanizmus fungovania počítačovej podpory).

Opísaný softvér *MAPola* prezentovala počítačová sekcia Slovanského jazykového atlasu na bratislavskom zasadnutí Medzinárodnej komisie OLA v októbri 2004. Účastníci zasadnutia z 12 slovanských krajín (Slovensko, Slovinsko, Chorvátsko, Srbsko a Čierna Hora, Bosna a Hercegovina, Macedónsko, Česká republika, Poľsko, Nemecko, Ruská federácia, Ukrajina, Bielorusko, od r. 2005 aj Bulharsko) sa so softvérom *MAPola* oboznámili, zoponovali ho a prijali ako záväzné východisko prípravy ďalších zväzkov Slovanského jazykového atlasu. Počítačovú podporu *MAPola* v súčasnosti slovenská národná komisia používa pri príprave štvrtého zväzku lexikálno-slovotvornej série OLA *Poľnohospodárstvo,* pri spracúvaní podkladov do 5. zväzku hláskoslovnej série OLA (česká národná komisia) a pri príprave podkladov do zväzku, v ktorom sa bude spracúvať substantívna deklinácia (slovenská komisia). Hlavný zmysel počítačovej podpory projektu OLA spočíva v tom, že umožňuje dialektológovi syntetizujúcim spôsobom pripraviť na vydanie produkt, ktorého podoba eliminuje nežiaduce zásahy a vnášanie chýb do textovej aj kartografickej časti atlasu, t. j. minimalizuje potrebu jazykovej aj grafickej korekcie do takej miery, aby výsledky dialektologickej práce v elektronickej podobe bolo možné odovzdať priamo do tlačiarne. Uvedomujeme si, že naša metóda nepredstavuje najdokonalejšie možnosti spracovania nárečového materiálu. Vznikla z praktickej potreby a vnímame ju ako jednu z etáp, ktorá sa bude ďalej rozvíjať a v súčasnosti je jedným z prostriedkov spracúvania nárečového materiálu z celého slovanského územia, a nebola hlavným cieľom našej práce, ale prostriedkom zefektívnenia klasických metód v oblasti jazykového zemepisu. Výsledky počítačovej sekcie OLA, na ktorých má významný podiel slovenská národná komisia, vznikli paralelne s prípravou jednotlivých zväzkov v rámci tohto projektu a sú významným krokom efektivizácie dialektologickej práce a lingvistickej geografie na medzinárodnej úrovni.

Literatúra

Вопросник Общеславянского лингвистического атласа. Moskva: Nauka 1965.

HABOVŠTIAK, Anton: Slovanský jazykový atlas. In: Studia Academica Slovaca. 18. Bratislava: Alfa 1989, s. 129 – 140.

Obščeslavianskij lingvističeskij atlas. Serija fonetiko-grammatičeskaja. Zv. 1. Refleksy *ě. Red. B. Vidoeski a P. Ivić. Belehrad: Jugoslovenska Akademija Nauk 1988. 164 s.

Obščeslavianskij lingvističeskij atlas. Serija fonetiko-grammatičeskaja. Zv. 2a. Refleksy *ę. Red. V. V. Ivanov. Moskva: Nauka 1990. 178 s.

Obščeslavianskij lingvističeskij atlas. Serija fonetiko-grammatičeskaja. Zv. 2b. Refleksi *ǫ. Red. J. Basara. Vroclav: Polska Akademia Nauk 1990. 124 s.

Obščeslavianskij lingvističeskij atlas. Serija fonetiko-grammatičeskaja. Zv. 3. Refleksi ьr, ьl, rъ, lъ. Red. J. Basara. Varšava: Państwowe Wydawnictwo Naukowe 1994. 164 s.

Obščeslavianskij lingvističeskij atlas. Serija fonetiko-grammatičeskaja. Zv. 4b. Refleksi ъ, ь. Red. B. Vidoeski – P. Ivić – Z. Topolińska. Skopje: MANU 2004. 152 s.

Obščeslavianski lingvističeski atlas. Serija fonetičesko-grammatičeskaja. Zv. 4a. Refleksi ъ, ь. Red. D. Brozović. Zagreb: HAZU 2006. 164 s.

Общеславянский лингвистический атлас. Серия лексико-словообразовательная. Zv. 1. Животный мир. Red. R. I. Avanesov. Moskva: Nauka 1988. 192 s.

Общеславянский лингвистический атлас. Серия лексико-словообразовательная. Zv. 2. Животноводство. Red. B. Falińska. Varšava: Państwowe Wydawnictwo Naukowe 2000. 192 s.

Общеславянский лингвистический атлас. Серия лексико-словообразовательная. Zv. 3. Растительный мир. Red. A. I. Padlužny. Minsk: IANB 2000. 168 s.

Общеславянский лингвистический атлас. Серия лексико-словообразовательная. Zv. 8. Профессии и общественная жизнь. Red. J. Basara – J. Siatkowski. Varšava: Państwowe Wydawnictwo Naukowe 2003. 192 s.

REHUŠ, S.: MAPola. 1.9. Bratislava (elektronická verzia). 2004.

Národný korpus bakalárskych, diplomových, dizertačných, rigoróznych a habilitačných prác slovenských vysokých škôl a boj proti plagiátorstvu

Július Kravjar

Centrum vedecko-technických informácií SR, Bratislava, Slovensko

Abstract. The paper analyzes the formation and a two-year operation of the National Corpus of Bachelor's, Master's, Doctoral and Habilitation Theses of Slovak Higher Education Institutions (hereinafter HEI), which also acts as a comparative corpus for the originality check. Each type of thesis has to be submitted to the national corpus to undergo the originality check before it is defended. The originality check also includes comparison to selected Internet resources.

The first activities related to the creation of electronic repositories of higher education institution theses are rooted at the threshold of this millennium. A significant activity was the central IT development project from 2004, "Building digital academic libraries – collection and provision of access to the full texts of publications of Slovak universities (ETD.SK)", which was submitted by 16 academic libraries of 12 universities represented by the Slovak University of Prešov. This project marked the beginnings of nationwide cooperation in this field. Unfortunately, the implementation of the project encountered obstacles in the form of inadequate financial and staff resources and the main obstacle was the lack of legislative support.

Plagiarism is a phenomenon that existed in the past and will exist in the future. Slovakia with its population of 5.4 million is confronted with plagiarism of university theses in the same extent as other countries. The spreading of ICT and the low awareness of copyright and intellectual property rights plus a rapid growth in the number of HEI and students – all that contributed to the spread of unwanted species of "creativity" – plagiarism. There were no systematic measures to hinder the growth of plagiarism.

The year 2008 was a significant milestone. The Ministry of Education decided to make a systemic change: to implement a comprehensive solution for nationwide collection and processing of theses in order to create a national corpus of these theses, to improve the protection of copyright and intellectual property rights, to improve the quality of theses by checking their originality and to build a barrier against the expanding plagiarism. This time, there was no lack of legislative support.

In 2009, an amendment to the Higher Education Act was adopted, including the following crucial change: Before the defence of the thesis, the university forwards the thesis in the electronic form to the central repository. The thesis undergoes the originality check. The thesis and the relevant metadata are kept in the central repository for a period of 70 years from the date of registration. The Ministry manages the central repository; its operation is delegated to an institution directly managed by the Ministry of Education.

According to our information, the obligation to use the services of the central register of theses and dissertations together with the originality check at a national level for all higher education institutions operating in Slovakia under the Slovak legal order is a unique solution not only in Europe, but probably also in the world.

Universities do not pay for this service. The acquisition costs of the system and its operating costs have been supported by the Ministry of Education.

The theses and dissertations registered in the central repository after 31 August 2011 are publicly available from 1 September 2011 – this was made possible by another amendment of the Higher Education Act.

The Central Register of Theses of Slovak HEI and the Plagiarism Detection System became reality at the end of April 2010. Slovak HEI are obliged to use both systems. The annual increase in the central register is about 80 000 theses. At the end of April 2012, there were 189 700 theses in the central repository.

The quality of the originality check was improved by the jubilee Slovak National Corpus (SNK JÚĽŠ SAV) providing the Slovak Morphological Analyzer and Slovak Synonym Dictionary. We have already had several ideas for further cooperation in the near future. We are grateful for the cooperation of the jubilee institution, and we wish it a lot of creativity in the next decades.

1 Úvod

Príspevok analyzuje vznik a dvojročnú prevádzku národného korpusu bakalárskych, diplomových, dizertačných, rigoróznych a habilitačných prác slovenských vysokých škôl a naň naviazaného systému na odhaľovanie plagiátov. Tento národný korpus nazývame Centrálnym registrom bakalárskych, diplomových, dizertačných, rigoróznych a habilitačných prác – CRZP. Do národného korpusu, resp. do CRZP povinne prichádza každá z uvedených typov prác ešte pred jej obhajobou a podrobuje sa kontrole originality.

2 Prvé kroky budovania zbierok vysokoškolských prác

Budovanie digitálnych zbierok vlastnej akademickej produkcie a ich sprístupňovanie vo forme úplných textov digitálnych dokumentov, prístupných cez počítačovú sieť, preniká v poslednom desaťročí postupne aj do prostredia akademických knižníc SR. Medzi pilotné projekty v tejto oblasti patrili projekty zamerané na elektronické spracovanie publikačnej činnosti zamestnancov vysokých škôl a elektronické spracovanie záverečných a kvalifikačných prác. Prvé takéto projekty realizovali akademické knižnice už začiatkom milénia. Mimoriadne aktívna v tejto oblasti bola Univerzitná knižnica Technickej univerzity v Košiciach, ktorá v roku 2002 dokončila vývoj a testovanie vlastného informačného systému na evidenciu publikačnej činnosti zamestnancov, ktorý bol v roku 2006 rozšírený o systém na elektronický zber a archiváciu záverečných prác študentov. Aktívne v tejto oblasti boli aj ďalšie akademické knižnice SR, najmä Slovenská poľnohospodárska knižnica pri SPU v Nitre, Univerzitná knižnica Prešovskej univerzity, akademické knižnice Univerzity Komenského v Bratislave a iné.

Významným medzníkom v oblasti budovania digitálnych knižníc akademickej produkcie SR bol marec 2004, keď 16 akademických knižníc 12 slovenských univerzít zastúpených Prešovskou univerzitou podalo centrálny rozvojový IT projekt pod názvom *Budovanie digitálnych akademických knižníc – zber a sprístupnenie úplných textov publikácií slovenských univerzít (ETD.SK).* ETD je medzinárodne zaužívanou skratkou pre vysokoškolské záverečné práce v elektronickej forme, tzv. Electronic Theses and Dissertations. Projekt ETD.SK znamenal počiatky kooperácie v tejto oblasti na národnej úrovni so snahou nadviazať na medzinárodné aktivity v oblasti ETD. V rámci projektu boli stanovené organizačné, technické a technologické predpoklady na zber ETD, v jednotlivých knižniciach boli vybudované bazálne digitalizačné pracoviská a vytvorené dve hardvérové úložiská (Košice, Prešov). Bohužiaľ, projekt nebol vo všetkých knižniciach dostatočne realizovaný predovšetkým pre nedostatočné finančné a personálne zabezpečenie, ale najmä pre nedostatočnú legislatívnu podporu (Haľko, 2011).

Pozitívnym sprievodným efektom bola skutočnosť, že vysoké školy začali sprístupňovať záverečné a kvalifikačné práce na stránkach svojich akademických knižníc.

3 Plagiátorstvo

Plagiátorstvo je jav, ktorý existoval v minulosti a bude zrejme existovať aj v budúcnosti. V starovekom Ríme slovo plagiátor (plagiarius) označovalo únoscu, a to predovšetkým únoscu detí alebo otrokov. Za akt plagiátorstva (plagium, crimen plagii) sa potom považovalo i poskytnutie útočiska cudziemu otrokovi na úteku.

Obsah tohto slova sa časom podstatne zmenil a zachoval sa v prenesenom význame. Dnes pod „únosom“ chápeme únos názorov, nápadov a iných nehmotných bohatstiev, ktoré sa neoprávnene „privlastňujú“ a vydávajú za originálne, pôvodné. Plagiátorstvo môžeme definovať ako použitie originálnych myšlienok a tvorivých vyjadrení inej osoby s úmyslom prezentovať ich ako vlastné myšlienky a vyjadrenia (Szattler, 2007).

Pri absencii vhodných nástrojov na prevenciu a potláčanie plagiátorstva môže tento fenomén prerásť do neakceptovateľných rozmerov. Úplne eliminovať všetky druhy plagiátorstva zrejme nebude možné ani v budúcnosti, ale je potrebné stavať im bariéry na všetkých frontoch. Nesmieme ho ignorovať ani tolerovať.

Prvá vysoká škola začala používať systém na odhaľovanie plagiátov v roku 2001 a bola osamotenou bežkyňou celých sedem rokov.

4 Východiská

Slovensko s 5,4 miliónmi obyvateľov je konfrontované s plagiátorstvom vysokoškolských prác rovnako ako iné krajiny. Plagiátorstvo – tento neželaný druh „tvorivosti“ – bolo do roku 2010 na slovenskej vysokoškolskej scéne bujnejúcim živlom. Bytostne chýbalo systémové opatrenie, ktoré by bolo určitou hrádzou pre jeho ďalší rast. Prudko sa zvyšujúci

počet vysokých škôl a ich študentov, rastúca penetrácia internetu a nízke povedomie o autorských právach a o právach duševného vlastníctva – to všetko prispelo k rastu plagiátorstva. Ak porovnáme rok 1989 s rokom 2011, zistíme, že počet vysokých škôl sa strojnásobil na 39 a počet študentov vzrástol štvornásobne na 250 tisíc. Kým v roku 1989 bola penetrácia internetu nulová, v roku 2010 dosiahla 74,3 % a v roku 2011 bola už na úrovni 79,2 %.

A ako prispieva internet k šíreniu plagiátorstva? V podstate dvojako. Na internete sú voľne dostupné referáty, záverečné práce a aj odborná literatúra. Necitlivé použitie funkcií „skopíruj a vlož“ bez patričného citovania, zakomponovanie skopírovaného do textu, pod ktorý sa podpíše „autor“, to je klasický príklad plagiátorstva.

Ponuky týkajúce sa vypracovania „podkladov“ k rôznym druhom prác na objednávku nájdeme na internete bez väčšieho úsilia. Existujú webové stránky ponúkajúce vypracovanie prác so širokým záberom (seminárne, absolventské, bakalárske, diplomové, dizertačné, MBA...), pokrývajúce trhy vyše desiatky krajín, ale sú aj také, ktoré sa zameriavajú len na lokálne, resp. jazykovo a historicky príbuzné trhy. Je potrebné mať na zreteli, ako konštatuje Z. Adamová, že „vydávanie cudzieho diela za vlastné je totiž plagiátorstvo“ (Sudor, 2012).

Objednávanie bakalárskej, diplomovej alebo dizertačnej práce a platenie za jej vypracovanie nepatrí len medzi slovenské špeciality, je to globálny problém. Ak by sa zistilo ešte pred koncom štúdia, že práca bola robená na objednávku, študentovi môže byť predčasne ukončené štúdium. Ak sa to však zistí až po získaní vysokoškolského vzdelania, nič sa nestane. Slovenský zákon o vysokých školách nepozná odňatie vysokoškolského titulu. Titul zostane neporušený v rukách jeho vlastníka (Húska, 2012).

Štúdia Zavádzanie pravidiel akademickej etiky na slovenských vysokých školách (Králiková, 2009) mapuje stav akademickej etiky na slovenských vysokých školách (ďalej VŠ) a poukazuje na to, že väčšina respondentov z radov pedagogických pracovníkov mala priamu skúsenosť s podvádzaním u študentov. Najdiskutovanejšou témou v médiách bolo plagiátorstvo študentov, v rámci ktorého sa hovorilo najmä o spôsoboch podvádzania a o tom, aké nástroje vysoké školy používajú na eliminovanie plagiátorstva. K ďalším témam patrilo plagiátorstvo pedagógov, respektíve ľudí zastávajúcich vysoké verejné pozície. Neexistencia širšej diskusie o akademickej etike má aj ďalšie dôsledky. Jedným z nich je, že členovia akademickej obce a širšej verejnosti nerozumejú významu akademickej etiky a sú preto aj menej citliví na jej porušovanie. (Králiková, 2009).

Otázky týkajúce sa zberu a spracovania vysokoškolských prác v elektronickej podobe a otázky plagiátorstva boli často sa opakujúcimi témami diskusií v akademickej obci, ale bez výrazného pokroku. Semienka budúcich zmien boli zasiate v septembri 2006 na 36. plenárnom zasadnutí Slovenskej rektorskej konferencie (ďalej SRK; Zápisnica z 36. riadneho zasadnutia pléna Slovenskej rektorskej konferencie, 2006), keď boli schválené dva dokumenty týkajúce sa akademickej etiky (pozri Opatrenia na odstránenie plagiátorstva pri spracovaní a prezentovaní bakalárskych, diplomových a dizertačných písomných prác, 2006; Etický kódex zamestnancov vysokých škôl, 2006). Tieto dokumenty zaoberajúce sa etikou pedagógov a študentov sa stali dokumentmi celoštátneho významu. Ale navrhované opatrenia na zamedzenie plagiátorstvu neboli uvedené do života (Králiková, 2009). Vo februári 2008 sa SRK vrátila na svojom 43. plenárnom zasadnutí

k problému plagiátorstva a požiadala ministerstvo školstva o koordináciu činností súvisiacich s obstaraním systému na odhaľovanie plagiátov, a odporúčala, aby vysoké školy plagiátorstvo postihovali a vytvárali elektronické archívy vysokoškolských prác (Výročná správa o hospodárení Slovenskej rektorskej konferencie za rok 2008, 2009).

I keď sa pomerne veľa diskutovalo o plagiátorstve, o potrebe boja proti nemu, nakoniec riešenie prinieslo až Ministerstvo školstva Slovenskej republiky svojím rozhodnutím z roku 2008 (v tomto roku len dve VŠ využívali systém na odhaľovanie plagiátov), ktoré sa stalo záväzné pre všetky VŠ pôsobiace na Slovensku podľa slovenského právneho poriadku:

- zriadený bude Centrálny register záverečných a kvalifikačných prác, VŠ budú povinne každú záverečnú a kvalifikačnú prácu zasielať do centrálneho registra;
- práca sa v centrálnom registri bude uchovávať spolu s menom a priezviskom autora a názvom vysokej školy, ktorá prácu zaslala, počas 70 rokov odo dňa jej registrácie;
- práce pred obhajobou povinne prejdú procesom overenia originality, ktorej výstupom je Protokol o kontrole originality;
- kontrola originality sa bude robiť voči prácam v CRZP, voči vybraným internetovým zdrojom a ďalším dostupným elektronickým zdrojom.

Vytýčené úlohy smerovali k zvýšeniu kvality VŠ štúdia a zároveň k:

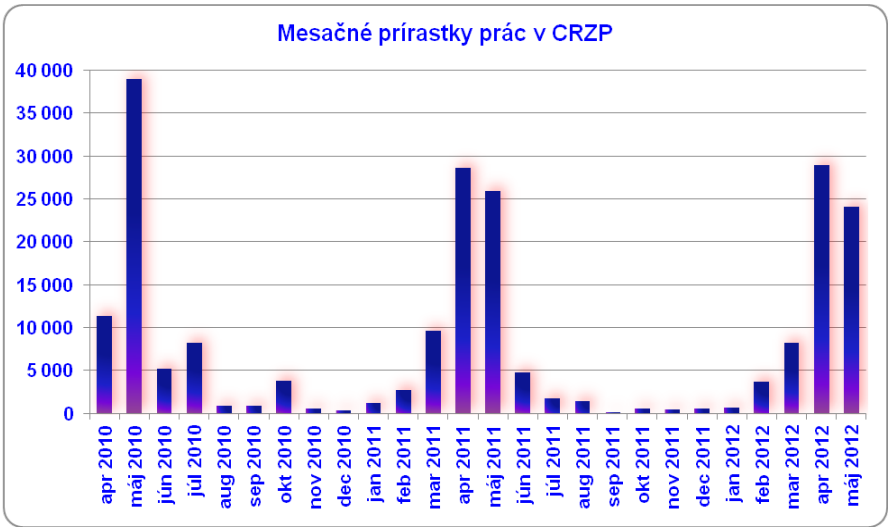
- ochrane autorských práv a práv duševného vlastníctva, k rastu povedomia o nich;
- zvýšeniu kvality prác vykonávaním kontroly originality;
- vybudovaniu národného elektronického centrálneho registra prác;
- postaveniu bariéry bujnejúcemu plagiátorstvu.

V roku 2009 bola prijatá novela zákona o vysokých školách s touto najdôležitejšou zmenou: Ministerstvo spravuje centrálny register záverečných (bakalárskych, diplomových, dizertačných), rigoróznych a habilitačných prác. Pred pripustením osoby k obhajobe záverečnej práce, rigoróznej práce alebo habilitačnej práce vysoká škola zašle túto prácu v elektronickej forme do centrálneho registra záverečných, rigoróznych a habilitačných prác a na základe informácie z centrálneho registra záverečných, rigoróznych a habilitačných prác overí mieru originality zaslanej práce. Zaslaná práca sa v centrálnom registri záverečných, rigoróznych a habilitačných prác uchováva spolu s menom a priezviskom autora a názvom vysokej školy, ktorá záverečnú, rigoróznu alebo habilitačnú prácu zaslala, počas 70 rokov odo dňa registrácie (Zákon 496/2009).

Práce, ktoré prišli do centrálneho registra (národného korpusu) po 31. 8. 2011 sú prístupné verejnosti – to umožnila ďalšia novela vysokoškolského zákona. Centrálny register bakalárskych, diplomových, dizertačných, rigoróznych a habilitačných prác slovenských vysokých škôl a systém kontroly originality sa stali skutočnosťou na konci apríla 2010. Slovenské vysoké školy sú povinné používať oba systémy. K 31. 5. 2012 bolo v registri evidovaných 213 773 prác. Hranica 200 tisíc prác bola prekročená v máji 2012.

Dodávateľom oboch systémov je firma SVOP, spol. s r. o. O schopnostiach a zručnosti spoločnosti svedčí jej víťazstvo na medzinárodnej súťaži External Plagiarism Detection Performance at PAN 2011 Lab Uncovering Plagiarism, Authorship, and Social Software Misuse Conference v Amsterdame v roku 2011. Dokázala sa presadiť v prostredí, kde sa pracovalo len s textami v angličtine, španielčine a nemčine a zvíťazila v každom z piatich sledovaných ukazovateľov. Jednou zo sledovaných úloh bolo aj odhaľovanie translačného plagiátorstva.

Sezónnosť posielania prác do CRZP dokumentuje nasledujúci graf:



5 Čo bezprostredne priniesla implementácia projektu

Pred realizáciou projektu bolo potrebné vziať do úvahy nasledujúce súvislosti a nájsť riešenie:

- rozmáhajúce sa plagiátorstvo záverečných a kvalifikačných prác,
- sporadické úsilie VŠ o kontrolu originality záverečných a kvalifikačných prác,
- neexistenciu systémových nástrojov v boji proti plagiátorstvu na národnej úrovni,
- centralizovaný prístup verejnosti k záverečným a kvalifikačným prácam,
- zvýšenie úrovne povedomia o autorských právach a právach duševného vlastníctva.

Zverejnenie oznámenia, že sa pripravuje centrálny register a kontrola originality pre záverečné a kvalifikačné práce priniesli preventívny efekt ešte pred implementáciou. Ďalšie výsledky implementácie projektu:

- zásadný prelom v boji proti plagiátorstvu na Slovensku, povinné využívanie centrálneho registra a kontroly originality (voči centrálnemu registru, voči vybraným internetovým zdrojom a ďalším elektronickým zdrojom);
- existencia a reálna prevádzka nástroja na ochranu autorských práv a na potláčanie plagiátorstva;
- jednotná metodika na zber záverečných a kvalifikačných prác (vytvorenie spoločného centrálneho registra pre všetky VŠ) a jednotný systém na kontrolu ich originality;
- zautomatizovanie zberu záverečných a kvalifikačných prác, kontroly originality a distribúcie protokolov o kontrole originality;
- Samotná existencia centrálneho registra a systému na odhaľovanie plagiátov pôsobí preventívne, a to nielen v spoločenstve študentov. Pomáha zvyšovať povedomie o autorských právach, o právach duševného vlastníctva aspoň v akademickej obci, zlepšuje prácu študentov s literatúrou, internetom a citáciami a prispieva k vyššej kvalite záverečných a kvalifikačných prác;
- „Spustenie systému malo najmä psychologický účinok – študenti boli zodpovednejší pri písaní práce a opatrnejší pri používaní zdrojov," povedal prezident Slovenskej rektorskej konferencie Libor Vozár (sme.sk, 20. 8. 2010);
- Zavedenie systému ocenil aj rektor Ekonomickej univerzity v Bratislave Rudolf Sivák, podľa ktorého pozitívne ovplyvnil najmä prístup študentov. „Práce sa robia viac samostatne, majú vyššiu kvalitu a zvýšila sa tiež miera citácie použitej literatúry," zhodnotil (sme.sk, 20. 8. 2010);
- Komplex centrálneho registra s kontrolou originality na národnej úrovni sú povinné využívať všetky VŠ na Slovensku, ktoré pôsobia podľa právneho poriadku Slovenskej republiky. Týka sa to 35 VŠ, 33 z nich aktívne využíva centrálny register a kontrolu originality. Zvyšné dve VŠ vznikli v roku 2011 a stále čakajú na svoje prvé záverečné a kvalifikačné práce;
- Všetky záverečné a kvalifikačné práce sa sústreďujú v centrálnom registri, kde sa uchovávajú počas 70 rokov odo dňa registrácie;
- Verejnosť si môže overiť podozrenie z plagiátorstva na stránke, kde sú záverečné a kvalifikačné práce zverejňované;
- Protokol o kontrole originality nie je potvrdením, že práca je originálom alebo plagiátom. Protokol je podkladovým materiálom na rozhodovanie skúšobnej komisie, je pomôckou. Upozorňuje na dokumenty, ktoré mohli uniknúť pozornosti školiteľa alebo oponenta. Protokol o kontrole originality identifikuje časti textu predloženej práce, totožné s časťami textov prác uložených v CRZP a s vybranými internetovými zdrojmi. Protokol o kontrole originality práce sa sprístupní skúšobnej komisii na vyhodnotenie a je súčasťou zápisu o záverečnej (štátnej) skúške;
- VŠ za využívanie týchto systémov neplatia; obstarávacie náklady boli pokryté zo zdrojov Ministerstva školstva, vedy, výskumu a športu Slovenskej republiky, prevádzkové náklady tiež uhrádza ministerstvo.

6 Záver

Veľké rezervy v boji proti plagiátorstvu sú vo výchove mladej generácie. Plagiátorstvu je potrebné predchádzať a procesy výchovy a vzdelávania ku kultúre nepodvádzania musia začať postupne a primerane od najskoršieho veku, pokojne od úrovne predškolskej výchovy (Skalka et. al., 2009).

Správne orientovaný a načasovaný vzdelávací proces a implementácie pokročilých technológií majú veľký potenciál v boji s plagiátorstvom. Technológie nie sú žiadnym všeliekom. Veľmi dôležitá a nezastupiteľná je úloha vzdelávania – od začiatku vzdelávacieho procesu – v úzkej súvislosti s prevenciou a odhaľovaním, s jasne definovanými pravidlami a sankciami a vo vzájomnej interakcii všetkých týchto zložiek (Kravjar, 2011a).

Implementácia centrálneho registra a kontroly originality na národnej úrovni do každodennej praxe je zrejme jedinečným riešením v Európe a pravdepodobne aj vo svete. Míľnik bol postavený. Iniciátorom, architektom, vývojárom, organizátorom a všetkým zúčastneným stranám, ktoré prispeli k životaschopnosti tohto projektu, patrí uznanie. Systém má veľký potenciál uplatnenia v mnohých oblastiach. Kontrola originality by sa mohla robiť všade tam, kde je výstupom písomná práca. Do úvahy môžu pripadať:

- seminárne a iné práce na VŠ,
- výskumné správy,
- žiadosti o projekty/granty,
- záverečné práce na zvyšovanie kvalifikácie pedagogických profesií, ale aj iných profesií,
- stredoškolské práce v rámci vyučovacieho procesu,
- stredoškolské práce v rámci mimoškolskej činnosti,
- publikačná činnosť všeobecne (nadviazať vzťahy s vydavateľstvami) atď.

Systém sa neustále vyvíja. V minulom roku získal dodávateľ systému SVOP, s. r. o. prvenstvo na medzinárodnej súťaži antiplagiátorských systémov v Amsterdame, kde bol ich algoritmus najlepší vo všetkých piatich sledovaných ukazovateľoch. Významné ocenenie získal systém aj na medzinárodnej konferencii ITAPA 2011 v Bratislave. V kategórii Nové služby sa celoplošný systém overovania originality záverečných prác pre vysoké školy na Slovensku umiestnil na druhom mieste.

K zvyšovaniu kvality kontroly originality prispel aj jubilujúci Slovenský národný korpus, oddelenie Jazykovedného ústavu Ľ. Štúra SAV, poskytnutím Morfologického analyzátora slovenského jazyka a Synonymického slovníka slovenského jazyka. Už dnes máme niekoľko námetov na ďalšiu spoluprácu v budúcnosti.

Literatúra

Antiplagiátorský systém vraj vystrašil študentov, viac citujú. In: SME, 20. 8. 2010. Dostupný z WWW: http://www.sme.sk/c/5513734/antiplagiatorsky-system-vraj-vystrasil-studentov-viac-cituju.html

Etický kódex zamestnancov vysokých škôl. Dostupný z WWW: http://old.srk.sk/zaznam/89/Eticky-kodex-zamestnancov-vysokych-sk%C3%B4l/

HAĽKO, Peter: Elektronický zber záverečných prác na Prešovskej univerzite v Prešove. In: Uninfos 2011. Univerzitné informačné systémy. Prešov: Združenie EUNIS Slovensko – Prešovská univerzita v Prešove, s. 61 – 65. Dostupný z WWW: http://www.pulib.sk/elpub2/PU/Uninfos2011/data/P11_Halko.pdf

HÚSKA, Michal: Napíšem vám diplomovku. Stačí zaplatiť... In: Hospodárske noviny, 23. 3. 2012. Dostupný z WWW: http://style.hnonline.sk/vikend/c1-55146090-napisem-vam-diplomovku-staci-zaplatit

KRÁLIKOVÁ, Renáta: Zavádzanie pravidiel akademickej etiky na slovenských vysokých školách. 2009. Dostupný z WWW: http://www.governance.sk/assets/files/publikacie/akademicka_etika.pdf

KRAVJAR, Július: Antiplagiátorský systém. In: Sborník konference Inforum 2011. Dostupný z WWW: http://www.inforum.cz/pdf/2011/kravjar-julius.pdf

KRAVJAR, Július: Barrier to thriving plagiarism. Príspevok prednesený na konferencii The 5th International Plagiarism Conference. 2012. Dostupný z WWW: http://archive.plagiarismadvice.org/documents/conference2012/finalpapers/Kravjar_fullpaper.pdf

Opatrenia na odstránenie plagiátorstva pri spracovaní a prezentovaní bakalárskych, diplomových a dizertačných písomných prác. 2006. Dostupný z WWW: http://old.srk.sk/zaznam/90/Stanovisko-plena-SRK-k-plagiatorstvu/

SKALKA, J. et al.: Prevencia a odhaľovanie plagiátorstva. Nitra: Univerzita Konštantína Filozofa 2009. 125 s. Dostupný z WWW: http://www.crzp.sk/dokumenty/prevencia_odhalovanie_plagiatorstva.pdf

SUDOR, Karol: Veľký biznis: Diplomovka na kľúč v akcii za pár stoviek eur. In: SME, 6. 4. 2012. Dostupný z WWW: http://www.sme.sk/c/6329494/velky-biznis-diplomovka-na-kluc-v-akcii-za-par-stoviek-eur.html

SZATTLER, Eduard: Právne a morálne aspekty plagiátorstva. In: Duševné vlastníctvo. Revue pre teóriu a prax v oblasti duševného vlastníctva, 2007, roč. 11, č. 1, s. 30 – 34.

Výročná správa o hospodárení Slovenskej rektorskej konferencie za rok 2008. Bratislava: Slovenská rektorská konferencia 2009. 40 s. Dostupný z WWW: http://www.srk.sk/images/stories/dokumenty/vyrocne_spravy/2008.pdf

Zákon č. 496/2009. Dostupný z WWW: http://www.portalvs.sk/files/files/zakon_496_2009.pdf

Zápisnica z 36. riadneho zasadnutia pléna Slovenskej rektorskej konferencie, ktoré sa konalo dňa 27. – 28. 9. 2006 v Trnave. 2006. Dostupný z WWW: http://www.srk.sk/images/stories/dokumenty/zapisnice/36.pdf

Maďarský národný korpus 2.
Pokus o nový korpus maďarského jazyka

Tibor Pintér

Jazykovedný ústav, Maďarská akadémia vied, Budapešť, Maďarsko

Abstract. It has been almost a decade since publishing the final version of the Hungarian National Corpus. Now, the Research Institute for Linguistics of the Hungarian Academy of Sciences is preparing an up-dated version of the national corpus. The corpus will be placed on new pillars: in means of word count as well as in means of the processes being made on the collected texts. The presentation is focused on the steps of corpus building strategies (if possible, with comparative analysis of the current and available version). The presentation is going to show the methods of text processing and problem solving which occurs in case of the foreseen amount of texts. While presenting the methods, the presentation is going to illustrate some bottlenecks of the building process. The methods and processes presented are going to cover the following issues: 1) Collection of texts, 2) Text analysis and annotation, 3) Corpus engine, 4) the GUI, 5) Other possibilities of corpus exploitation. In some parts of the presentation comparison to the available version of the Hungarian National Corpus will be provided.

1 Úvod

Oddelenie jazykovej technológie Jazykovedného ústavu Maďarskej akadémie vied (ďalej MAV) je centrálnou inštitúciou, ktorá sa venuje jazykovým technológiám v Maďarsku. Počas svojej pätnásťročnej existencie sa zapojilo do rôznych národných a medzinárodných projektov, v rámci ktorých vytvorilo mnoho užitočných aplikácií a zdrojov. Jedným z takýchto zdrojov je Maďarský národný korpus, ktorý bol utvorený ako reprezentatívny korpus maďarského jazyka a je používaný nielen v Maďarsku, ale aj v okolitých krajinách, kde žijú komunity Maďarov (o Maďarskom národnom korpuse pozri podrobnejšie Váradi, 2002; Pintér, 2007).

Jazykovedný ústav MAV naplánoval pod záštitou projektu CESAR aktualizáciu[1] mnohých maďarských zdrojov, medzi ktorými je i Maďarský národný korpus (o úlohách projektu pozri Ogrodniczuk a kol., 2012). Aktualizácia korpusu bola nevyhnutná, keďže existujúci korpus maďarského jazyka oslávil v roku 2012 (takisto ako Slovenský národný korpus) už 10 rokov svojej existencie. Počas uplynulých rokov bol maďarský korpus aktualizovaný len v istých častiach – nie však celkovo.

V čase písania tohto príspevku práce na novej verzii korpusu stále prebiehali. Niektoré práce sú už hotové, iné sú dokončené len sčasti. To je dôvodom, prečo v niektorých častiach textu používame minulý čas, kým inde prítomný a budúci čas.

[1] Súčasťou plánov je aj harmonizácia maďarských jazykových zdrojov so zdrojmi ostatných partnerov v rámci projektu CESAR.

2 Maďarský národný korpus

Korpus bol vytvorený hneď po založení Oddelenia jazykovej technológie Jazykovedného ústavu MAV medzi rokmi 1998 a 2002. Prvý korpus bol utvorený z textov pochádzajúcich z Maďarska. Texty z okolitých regiónov boli prítomné nesystematicky, len ako ukážky. Jediné zdroje zahraničných textov boli noviny Új Szó zo Slovenska a Romániai Magyar Szó z Rumunska. Texty zaradené do korpusu boli vybrané náhodne a ich výber si kládol za cieľ znázorniť varianty maďarčiny za hranicami Maďarska. Do druhej verzie korpusu boli už začlenené nasledujúce subkorpusy:

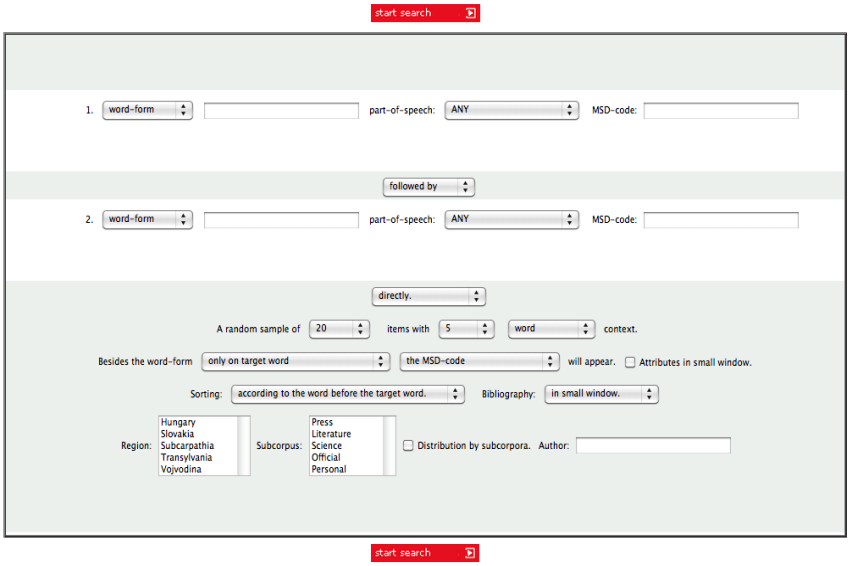
1. subkorpus maďarského jazyka na Slovensku,
2. subkorpus maďarského jazyka v Rumunsku,
3. subkorpus maďarského jazyka na Ukrajine,
4. subkorpus maďarského jazyka v Srbsku.

Publikovaniu týchto korpusov predchádzal systematický zber textov a ich spracovanie. Konečná podoba korpusu (ktorá je oficiálne treťou verziou) mala nasledujúce parametre:

	HU	SK	RO	UK	SR	TOTAL
tlač	71,0	5,7	0,7	5,5	1,5	84,5
beletria	35,5	1,4	0,4	0,8	0,2	38,2
vedecký	20,5	2,3	0,7	1,6	0,3	25,5
úradný	19,9	0,2	0,3	0,6	0,1	20,9
hovorený/ osobný	17,8	—	0,4	0,4	0,1	18,6
SPOLU	164,7	9,5	2,5	8,9	2,0	187,6

V začiatkoch budovania korpusu bolo cieľom dosiahnuť 100 miliónov slov (tokenov), čo sa z dnešného pohľadu môže javiť ako nenáročná úloha. Na reprezentovanie písaného a hovoreného úzu maďarského jazyka bolo v roku 2005 k dispozícii dostatočné množstvo tokenov – spolu 187 miliónov. Toto relatívne malé množstvo slov (pred desiatimi rokmi bolo ešte postačujúce) súviselo s technikami zberu a malým počtom textov, ktoré bolo možne získať z internetu.

Maďarský národný korpus bol prvým (a stále je jediným) maďarským korpusom, ktorý slúži ako pomôcka na posúdenie jazykových javov, variantov. Je jedinečný v mnohých ohľadoch: jednak je to korpus, ktorý je – po registrácii – dostupný každému používateľovi, jednak je sprístupnený s ergonomickým vzhľadom. Vyhľadávanie je možné pomocou dvoch grafických užívateľských rozhraní v sofistikovanejšej verzii (obrázok 1) a jednoduchšej verzii (obrázok 2).



Obr. 1. Grafické užívateľské rozhranie Maďarského národného korpusu so sofistikovanejším vyhľadávaním (`http://corpus.nytud.hu/mnsz`)

Unikátnosť korpusu je aj vo vnútornom rozložení subkorpusov. Maďarský národný korpus je vnútorne rozvrstvený z hľadiska geografického (podľa jazykových variantov – štandardov používaných v jednotlivých jazykových komunitách) aj z hľadiska žánrov – textových typov. Hoci je korpus vyvážený, v súčasnosti už nezodpovedá jazykovej realite maďarčiny a stavu korpusov vo svete.

3 Maďarský národný korpus 2

Keďže dozrel čas na prepracovanie Maďarského národného korpusu, Oddelenie jazykovej technológie Jazykovedného ústavu MAV sa podujalo uskutočniť potrebnú zmenu. Nový korpus maďarského jazyka zostane niektorými vlastnosťami spojený so svojím predchodcom, prinesie však používateľom aj nové možnosti, napríklad nové vrstvy analýzy, presnejšiu morfologickú anotáciu a nový dizajn. Cieľ korpusu zostáva nezmenený: reprezentovať maďarský jazyk na území Maďarska aj za jeho hranicami, a to v rôznych

doménach písaného a hovoreného jazyka. Cieľom nového korpusu je dosiahnuť rozsah 1 miliardy tokenov a pri analýze použiť nové anotačné vrstvy (napr. NER-tagging, presnejšiu syntax).

Obr. 2. Grafické užívateľské rozhranie Maďarského národného korpusu s jednoduchším vyhľadávaním (`http://corpus.nytud.hu/mnszbeta/`)

Práce na tvorbe korpusu môžeme rozdeliť do piatich etáp, ktoré pokrývajú aj hlavné problémové okruhy:

1. Zber textov
2. Analýza a anotácia
3. Systém spravovania korpusu – korpusový manažér (engine)
4. GUI
5. Prostriedky odvodené z korpusu

4 Problémy a riešenia

V prípravnej fáze korpusu sa vyskytujú rôzne typy problémov, nie sú to vždy len technické problémy. Napríklad pri zbere materiálu je problematické dosiahnuť, aby bolo čo najviac textov získaných legálne, na základe autorského povolenia majiteľa textu. Pri zbere tlače bolo prvoradou úlohou získať aj metadáta textov (rôznorodosť textov a snaha o autenticitu nás viedla k zberu čo najpodrobnejších dát). Pri získavaní textov do korpusu tlače sme popri textoch zbierali aj metadáta k týmto textom, avšak ani spracovanie tohto korpusu nebolo bez problémov. Keďže metadáta nie sú menej dôležité ako samotný text, považujeme za potrebné priradiť základné metadáta ku každému textu[2]. Zaujímavý je aj štatút textov v korpuse, resp. získanie príslušných licenčných práv na jednotlivé typy sprístupnenia textov: v novej verzii Maďarského národného korpusu chceme ponúknuť tri možnosti prístupu k textom:

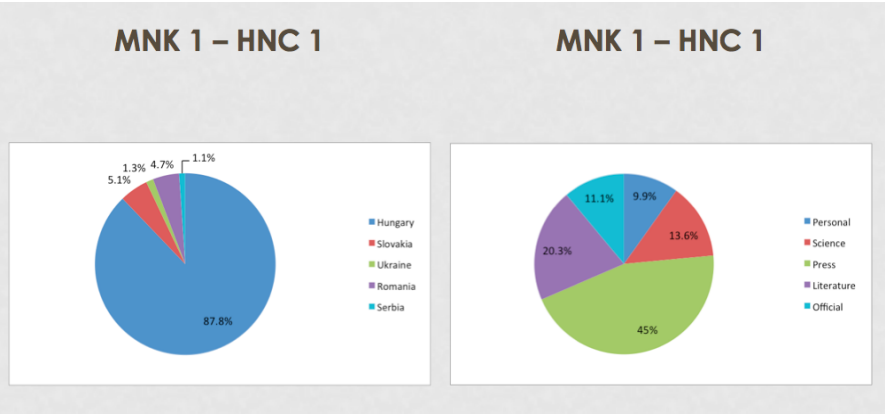
1. základný prístup k textom bude možný prostredníctvom grafického užívateľského rozhrania vo forme pseudokorpusu (ako v prípade Maďarského národného korpusu 1) – užívateľ získa prístup k textom korpusu po prihlásení sa (každý používateľ dostane užívateľské meno a heslo);
2. špecifický prístup pre tých, ktorí potrebujú väčšie množstvo textov na výskum, s možnosťou stiahnutia jedného menšieho korpusu s pomiešanými vetami či odsekmi (rozsah a vnútorné rozloženie korpusu nie je v tejto chvíli ešte stanovené). Do tohto korpusu sa dostanú iba také texty, ktorých licenčné práva na použitie umožňujú tento typ používania. Pred stiahnutím korpusu uzavrieme s používateľom zmluvu o používaní, v ktorej sa používateľ zaviaže k špecifickému spôsobu používania textov;
3. osobitný prístup ku korpusu, v ktorom budú jednotky textu v originálnom poradí, na špecifické výskumy, ktoré potrebujú väčšie množstvo textov na to, aby preukázali isté súvislosti textových jednotiek.

Pri príprave korpusu bolo potrebné získať licenčné práva k textom, ktoré budú tvoriť všetky tri typy korpusov.

Jednotlivé fázy prípravy korpusu prinášajú mnohé problémy, ktoré sa jednak viažu osobitne na náš korpus, jednak sú to problémy, ktoré sa vyskytujú pri príprave každého veľkého korpusu. V nasledujúcich častiach sa zaoberáme základnými problémami a úlohami, ktoré sa vyskytli a stále sa vyskytujú pri tvorbe Maďarského národného korpusu 2.

Otázky súvisiace s tematickým okruhom „zber textov“ sú spojené s problematikou žánrovej pestrosti. Pri úvahách o žánrovom rozložení nového korpusu sme vychádzali zo štatistík aktuálneho Maďarského národného korpusu. Pri tvorbe existujúceho korpusu bolo dôležité, aby obsahoval aj texty spoza hraníc, avšak v čase jeho tvorby išlo skôr o ukážku takéhoto typu textov ako o texty či osobitný korpus, ktorý by bol porovnateľný s maďarskou časťou (pri takom malom množstve textov je otáznou aj ich používanosť).

[2] Najväčším problémom Maďarského národného korpusu 1 (jediný dostupný korpus v čase písania príspevku) sú práve nedostatky súvisiace s metadátami.



Obr. 3. Jazykové a žánrové vrstvy v Maďarskom národnom korpuse

Zber textov ovplyvňuje viac faktorov, z ktorých najdôležitejším je formát textov, presnejšie typ súborov. Do korpusu sa dostanú iba tie texty, ktoré sa dajú spracovať bez optického rozpoznávania znakov (OCR). Takýmto spôsobom sa síce do korpusu veľké množstvo textov nedostane, avšak na druhej strane sa pri tvorbe korpusu môžeme plnšie sústrediť na spracovanie textov a databáz.

V dnešnej dobe je jedným z najväčších očakávaní „stráviteľná“ podoba textov, t. j. aby sprístupnené texty boli buď zoradené do nejakej databázy, alebo boli aspoň v takej podobe (html alebo txt formát, kde sú jednotlivé typy informácií – autor textu, názov textu, deň vydania textu atď. – na tom istom mieste), aby mohli byť spracované ďalej bez manuálnej práce. V tomto smere sa zdá, že jednotlivé žánre textov sa viažu k totožným typom (napríklad denníky a časopisy uverejnené na internete, ako aj ďalšie texty z internetu) textových súborov, čo veľmi uľahčuje ich spracovanie (napríklad od maďarského spravodajského portálu `http://index.hu` sme získali dvojgigabajtový súbor textov obsahujúcich viac než 250 miliónov slov; podobne aj texty z databáz ďalších denníkov – časť maďarského denníka na Slovensku Új Szó zo stránky `http://ujszo.com`).

Aj keď možné zdroje textov sú priamo na internete, teda prakticky v spracovateľnej podobe, zbieranie textov nie je bez problémov. Základným princípom tvorby korpusu je legálne použitie textov, oportunistické zbieranie a nelegálne sťahovanie textov neprichádza do úvahy. To znamená, že niektoré žánre sa dajú získať ľahšie (napr. beletria, texty z webových fór), avšak z právneho hľadiska sú niektoré texty priam nedosiahnuteľné. Najviac energie pri zbieraní zdrojov sme vynaložili na získanie textov z internetových denníkov. Zdá sa, že v Maďarsku ešte vydavateľstvá nedozreli na to, aby podporovali vedu a dôverovali Maďarskej akadémii vied. Situácia, žiaľ, nie je lepšia ani v iných vyspelejších krajinách. O rovnakých problémoch v korpusovom projekte SONAR píšu vo svojom príspevku O. de Clerq a M. Reynaert (2010).

Začiatkom roka 2012 sme poslali viac než päťdesiatim redakciám novín a časopisov oficiálny list z Jazykovedného ústavu MAV s prosbou o poskytnutie textov do korpusu. Žiaľ, len málo z oslovených bolo ochotných vyjsť nám v ústrety. Aj keď média a inter-

netové portály boli ochotnejšie oveľa menej, než sme očakávali, potrebné množstvo textov sa nám nakoniec podarilo zozbierať. Texty sme získali nielen z Maďarska, ale aj spoza jeho hraníc (napr. z Kanady a USA).

Beletria je v korpuse reprezentovaná Digitálnym archívom Literárneho múzea S. Petőfiho (`http://pim.hu`), ktorý zahŕňa kompletné diela súčasných spisovateľov z Maďarska aj spoza jeho hraníc. Časť beletrie sme získali z Maďarskej elektronickej knižnice Štátnej knižnice C. F. Szécsényho (`http://mek.oszk.hu`). Z uvedených prameňov sa do korpusu dostanú diela známych aj menej známych autorov.

Vedecká literatúra je vo väčšine zastúpená textami z Maďarskej elektronickej knižnice Štátnej knižnice C. F. Szécsényho, pričom materiály sme získali aj z niektorých vedecko-technických časopisov (tie sú z hľadiska štýlu zaradené do tohto typu subkorpusu).

Najmenšiu časť korpusu bude tvoriť úradný jazyk, a to pre – relatívne – málo dostupné množstvo úradného textu. Dôvodom nízkeho zastúpenia tohto typu textov je výskyt úradných textov zväčša vo formáte pdf, teda v súboroch, ktoré sú ťažšie spracovateľné.

Zvláštnu časť korpusu bude tvoriť hovorený jazyk, ktorý bude zastúpený v dvoch podobách: prepisy zvukových materiálov (napr. prepisy televíznych a rozhlasových relácií) a blogy a ostatné výdobytky WEB2. Pôjde o množinu písanej a hovorenej podoby jazyka.

Právna stránka zozbieraných textov je prvoradou vecou. Maďarský národný korpus je súčasťou projektu META-SHARE a musí mať vyriešené všetky právne náležitosti. Preto je ku každému spracovanému textu priložené aj povolenie na použitie v korpuse.

Pred tým, než sa začne vlastné spracovanie zozbieraných textov, musia texty prejsť fázami predbežného spracovania, ktoré zahŕňajú nasledujúce postupy:

– kódovanie znakov: keďže texty majú rôzne kódovania znakov, texty zjednocujeme do systému Latin-2. Znaky, pre ktoré neexistuje variant v tomto systéme, zobrazujeme ako numerické (hexa) unicode alebo html entity. Takýmto spôsobom môžeme všetky znaky kódovať do unicodu a ich zobrazenie na webe bude jednoduché. Na vyhotovenie požadovaného formátu sú potrebné rôzne konvertory, ktoré sme museli pripraviť na našom pracovisku;
– filtrovanie textov: táto fáza sa týkala len menšej časti textov, ktoré boli zozbierané z internetu, filtrovanie prebehlo iba pri tých textov, ktoré prechádzali do ďalšej fázy spracovania (obrázky, videá a iné pre nás nadbytočné elementy boli v textoch vynechané);
– filtrovanie maďarských textov: korpus bude reprezentovať maďarský jazyk a bude slúžiť ako pomôcka na výskum maďarského jazyka, z tohto dôvodu boli z neho inojazyčné texty odstránené;
– filtrovanie duplikátov: pri tvorbe korpusu sme narazili na množstvo duplicitných textov, ktoré sme pomocou heuristík a ručnou prácou odstránili;
– spracovanie XML textov: výstupom tejto fázy sú formulované XML súbory, ktoré sú štruktúrované na paragrafy a obsahujú metadáta v záhlaví súborov. Výstup tejto fázy je vstupom pre samostatnú anotačnú schému.

5 Analýza a anotácia

Z rôznych dostupných spôsobov anotácií sme používali in-line anotáciu v pôvodnom korpuse aj v novej verzii korpusu (nebolo potrebné meniť zaužívané anotačné postupy). Aplikovaná anotačná schéma korpusu sa delí na XML a in-line anotáciu, pričom XML anotácia je používaná na označenie vnútorných štruktúr textu (structure type) a in-line anotácia na označenie menších, špeciálnych častí textu (token, span). Z vnútorných štruktúr sú označené tituly, odseky, vety, slová a slovné spojenia. Špeciálne časti textu sú označené atribútmi `<lemma>` a `<postag>`. Novinkou v tejto verzii korpusu sú špeciálne anotačné vrstvy, napríklad named entity, ktoré sú označené na rovine tokenov prostredníctvom IOB anotácie.

Pri reprezentácii jednotlivých anotačných vrstiev sme určili hlavné kategórie takto: základné jednotky textu (text, odsek, veta, slovo) sú označené HTML kódmi (`<div>`, `<p>`, `<s>`, `<w>`), ich vlastnosti sú označené atribútmi. Ide o morfologickú anotáciu a unifikáciu (`<w lemma="..." msd="...">...</w>`) každej lexémy.

In-line (vnútorná) anotácia:

```
<w lemma="zeke" msd="N.ACC">zekét</w>
```

Vnútorné časti vety (pri súvetiach) sú označené konvenčnými IOB atribútmi. Tento prístup anotácie je používaný aj pri značkovaní špeciálnych jednotiek textu, ako sú NP (noun phrase), NE (named entity) a verbálne prefixy (o postupe značkovania pozri podrobnejšie Tjong Kim Sang – Veenstra, 1999).

```
<w id="1" lemma="kovács" msd="..." ne="B-PERS" neid="1">Kovács</w>
<w id="2" lemma="Pisti" msd="..." ne="I-PERS">Pisti</w>
```

NP:

```
<w id="1" np="B:head:3">A</w>
<w id="2" np="I">zöld</w>
<w id="3" np="I">zekét</w>
<w id="4" np="O">úgy</w>
<w id="5" np="O">viselte</w>
<w id="6" np="B:head:6">Károly</w>
<c id="7" np="O">.</c>
```

Verbálne prefixy:

```
<w id="1" lemma="el" vpfx="v:3">el</w>
<w id="2" lemma="fog">fog</w>
<w id="3" lemma="jön" vpfx="pre:1">jönni</w>
```

Vyššie spomenuté špeciálne anotačné vrstvy zatiaľ v Maďarskom národnom korpuse neexistujú – tieto vrstvy aplikované na maďarské texty sú označené len v niektorých špeciálnych maďarských korpusoch (Szeged Named Entity Recognition Corpus a Szeged Criminal Named Entity Corpus).

Zvolený postup anotácie má výhody i nedostatky. Najväčšou výhodou uvedenej analýzy je jej jednoduchosť. Výstupný XML súbor nepotrebuje schému, len jednoduchý DTD. Keďže plánovaný nový korpus bude mať okolo 1 miliardy slov, veľkou výhodou bude i rýchlosť konečnej analýzy. Potenciálne rozšírenie existujúcich vrstiev o nové vrstvy analýzy (čo bude pri tomto korpuse dôležité a očakávané) sa stane pri takomto rozložení ľahko zvládnuteľnou prácou. Istou výhodou je, že anotačné schémy môžu byť vnímané aj samostatne (oddelené od textu, rovnako môže byť aj text oddelený od anotácií), teda anotácia Maďarského národného korpusu môže byť vnímaná ako in-line, ale aj ako stand off.

Na začiatku práce sme presne pomenovali aj nevýhody tohto typu anotácie, ktoré však postupom času úspešne riešime. Jedným z problémov bolo napríklad spracovanie viacslovných jednotiek NE a NP.

6 Systém spravovania korpusu – korpusový manažér a grafické užívateľské rozhranie

Otázky súvisiace s použitím korpusového manažéra sú ešte stále otvorené. Pred posledným určením parametrov môžu byť na rýchle spracovanie veľkého množstva slov použité nasledujúce korpusové manažéry:

- NoSketch Engine: Manatee/Bonito
- IMS Open Corpus Workbench (CWB) – CQP
- Emdros
- XAIRA (XML Aware Indexing and Retrieval Architecture)

Vzhľadom na prednosti nástroja a potreby Maďarského národného korpusu bude pravdepodobne za korpusový manažér zvolený NoSketch Engine (dôjde k zmene v porovnaní s prvou verziou korpusu, s ktorou používatelia pracujú prostredníctvom CWB – CQP).

Jednou z posledných úloh bude zhotovenie grafického užívateľského rozhrania (GUI). Hlavné parametre rozhrania sú už síce známe, avšak musia byť ešte doplnené o ďalšie vlastnosti moderného GUI. Plocha bude interaktívna, zhotovená širšie pre rôzne potreby používateľov. V porovnaní s dnes dostupnou, používanou plochou (obrázok 1) ponúkne

nové rozhranie oveľa viac možností. Používateľ si bude môcť zvoliť z ponúk rôznych zdrojov získaných z korpusov (pravdepodobne interaktívnym spôsobom) a budú mu k dispozícii aj štatistiky generované z jednotlivých korpusov (pravdepodobne s možnosťou nastavenia vlastných parametrov).

7 Zdroje získané z korpusu

Ďalšie zdroje získané z korpusu slúžia náročnejším používateľom. Ich cieľom je pomôcť používateľovi, aby mal k dispozícii z korpusu aj zložitejšie informácie. Tak môže byť korpus používaný na čo najširšiu škálu výskumov týkajúcich sa maďarského jazyka. Najvýznamnejšou z ponúk budú dva menšie korpusy určené na stiahnutie, ktoré budú môcť byť použité na ďalšie spracovanie. Ponuku zdrojov sa snažíme zostaviť tak, aby boli užitočné pre jazykovedný výskum aj pre výskum jazykových technológií. Preto budú v ponuke aj n-gramy, frekvenčné slovníky (aj podľa subkorpusov) a pre inštitúcie a firmy, ktoré by chceli robiť na korpuse vlastné výskumy, bude pripravené aj API.

8 Záver

Počas desiatich rokoch existencie Maďarského národného korpusu prebehli v databáze menšie zmeny. Dnes už čas dozrel, aby bol tento korpus modifikovaný hlbšie. Maďarský národný korpus bol vždy korpusom, ktorý ponúkal niečo viac než ostatné maďarské korpusy. Je to korpus *par excellence* a aby si udržal svoju privilegovanú pozíciu, zmeny v ňom musia byť výrazné. Rozvoj korpusovej lingvistiky a technológií sa za uplynulých desať rokov podstatne zrýchlil. Preto musí nové „vydanie“ korpusu odrážať vývoj v týchto oblastiach spolu s požiadavkami používateľov. Veríme, že na plánované zmeny v korpuse nebudeme musieť čakať ďalších desať rokov.

Literatúra

CLERCQ De, Orphée – REYNAERT, Martin: SoNaR Acquisition Manual, version 1.0, LT3 Technical Report – LT3, 10. 2. 2010.

OGRODNICZUK, Maciej – GARABÍK, Radovan – KOEVA, Svetla – KRSTEV, Cvetana – PĘZIK, Piotr – PINTÉR, Tibor – PRZEPIÓRKOWSKI, Adam – SZASZÁK, György – TADIĆ, Marko – VÁRADI, Tamás – VITAS, Duško: Central and South-European language resources in META-SHARE. In: Infotheca, 2012, roč. 12, č. 1, s. 3 – 26.

PINTÉR, Tibor: „Határtalan“ magyar nyelv – az első, határon túli magyar nyelvváltozatokat tartalmazó strukturált magyar nyelvi korpuszról. In: Fórum Társadalomtudományi Szemle, 2007, č. 1, s. 165 – 182.

TJONG KIM SANG, Erik F. – VEENSTRA, Jorn: Representing text chunks. In: EACL, 1999, s. 173 – 179.

VÁRADI, Tamás: The Hungarian National Corpus. In: Proceedings of the 3rd LREC Conference, Las Palmas, 2002, s. 385 – 389.

Building Organized Text Corpora for Speech Technologies in the Slovak Language

Daniel Hládek – Ján Staš – Jozef Juhár

Faculty of Electrical Engineering and Informatics, Technical University of Košice, Slovakia

Abstract. In the presented paper, the whole process of the data preparation is described. The first component is the web agent that extracts raw text data from various formats and stores it in the database. This text is then segmented into tokens and sentences. The segmented text is ready for the filtration process, where unnecessary fragments, such as parts of web page user interface or textual advertisements are removed. After that, named entities in the corpus have to be identified and some of them, such as numbers or abbreviations, transcribed into the spoken form. The result is a set of data, prepared for the creation of the statistical language model.

1 Introduction

A large text database is necessary in the process of creation of the language model for *speech technologies*. This is especially true for the Slovak language that is characterized by very rich morphology and large vocabulary. A quantity of inflections and non-mandatory word order in a sentence means that the number of required word sequences in the training corpus is very high.

To overcome a data sparsity problem, domain-oriented and correctly prepared training corpus that is large enough is very important. In order to have a sufficient amount of text suitable for *statistical language modeling*, a specific set of tools for automatic *gathering*, *storing*, *sorting*, *processing*, and *filtering* have to be designed. The main aim of this text database is to provide a valuable tool for the research in the field of *computational linguistics* and *automatic speech recognition*.

Basic Terms

- **Document** – one complete unit of data containing text information;
- **Raw Data** – an unprocessed document, acquired from a source;
- **Raw Text** – document, where formatting is removed;
- **Segmented Text** – raw text, where certain parts, such as words and sentences are identified and put together;
- **Filtered Text** – segmented text, where sentences that are not suitable for language modeling are removed;

- **Transcribed Text** – text, where certain entities (such as abbreviations, numbers, dates or acronyms) are put into normalized form;
- **Training Corpus** – a collection of domain-oriented segmented, transcribed and filtered texts usable for building a language model.

The original source of text data is usually divided into documents. These documents are acquired from the source as raw, unprocessed data. Using an appropriate extraction method, formatting and other unimportant information is removed from the document and raw text can be sent for tokenization. The tokenization step identifies semantic entities in text and removes unnecessary spaces. The resulting tokenized text then can be processed as a stream of atomic units: words or named entities. By inspecting types of tokens and calculating error ratio, each sentence in the segmented text can be checked and filtered out. The last step required for creation of the training corpus is transcription. All identified entities have to be given to the spoken form. Transcription rules can take surrounding tokens into the account. In some cases, when a required context for the antecedent part of the transcription rule is too long or the condition is too difficult, the *regular expression* must be used. However the classical regular expression should be avoided, because they are difficult to read and computationally intensive. The final tokenized, filtered and transcribed text can be used as a training corpus.

There are more initiatives to build an organized database of the Slovak language for research purposes, such as the Slovak National Corpus (Horák et al., 2004; Slovak National Corpus, 2012) or Corpus of Spoken Slovak (Pleva et al., 2007). Neither of them is really suitable for creation of a training corpus that is large and flexible enough for the creation of the statistical model of the Slovak language (Juhár et al., 2012) for all domains that are required.

One of the most common systems for web page crawling and text retrieval for the purpose of building an organized database is Lucene (Hatcher and Gospodnetic, 2004). However, this one is not really focused on building a training corpus in the Slovak language. The previous work on the custom system is described in (Juhár et al., 2012; Ondáš et al., 2011).

2 Data Collection

For the training corpus creation we can use more types of electronic sources:

- **Printed Media** – paper books, newspapers, and magazines. Printed text must be scanned and processed by OCR software first. In this phase we have to focus on this source;
- **Static Electronic Sources** – electronic databases of text, such as laws or theses, available on removable media such as DVD or that can be downloaded from certain Internet site;
- **On-line Electronic Sources** – usually a web page that contains news, magazines or commercial presentations. This source seems to be the most promising, but many problems with cleaning and extracting have to be solved.

All these sources need to be gathered in one place. The best way to store, process, and sort large quantity of structured data seems to be a *relational database*.

2.1 Database Structure

All documents in the database can be stored in just one relation that have the following attributes:

1. **Document File** – location of the document file on the disk. It has been found out that storing binary data directly in the database is not efficient for this use case. Files are rather stored in dedicated directory on the disk;
2. **Extracted Text** – text that was extracted from the document file and is ready to be incorporated to the corpus;
3. **Document Source** – description of the original document location. URI string seems to be the most convenient way. This format is general enough to assign each document a unique name;
4. **Document Processing Status** – we can assign some information about the preliminary outcome of the processing. Possible states of the document are:
   (a) `processing` – document is not ready yet;
   (b) `final` – extracted text can be inserted to the corpus;
   (c) `error` – document contains error and could not be processed;
   (d) `copy` – document is a copy of some other document in the database;
   (e) `bad text` – extracted text of the document has low quality.
5. **Meta Information** – string with some additional information about document such as keywords that can help us determine document domain;
6. **Segment** – every document in the database can be assigned to a certain partition of the database dedicated to a certain domain.

2.2 Redundancy Control

The most important issue when building a big text database is control of document redundancy to restrict multiple copies of the same document in the resulting corpus. This issue has to be taken in mind during each insert or update of the database. It is possible that the same document can be stored under various URIs. On the other hand, one URL can have a document that is often updated and each time it contains different content.

That is why multiple criteria for *redundancy control* have to be used:

1. **URI-based Redundancy Control** – documents in the database should have a unique URI to ensure that form one place there will be just one document. URI of the document is usually not a very long string, so it is sufficient to implement this control as a simple unique constraint in the SQL.
2. **Document File Redundancy Control** – to ensure that the same file will occur just once in the database. It is very time-consuming to compare contents of two files. To help searching of the file copy, *hash code* of the file that is stored in the database. Hash code is a control sum of all the characters of a string, so that two different strings have different hash codes. This easily allows control duplicities of text strings, such as extracted text by assigning “unique” constraint to the document file hash column.

3. **Extracted Text Redundancy Control** – ensures that there will be no two documents with the same content. The same article can be in a PDF or HTML form and this rule should prevent this case. Again, hash code of the extracted text is stored in the database together with the unique constraint.

2.3 Raw Data Collection and Extraction

One of the most important sources is the Internet because it stores large amount of text on various themes. For that purpose a *text gathering agent* called webAgent, has been created (Hládek and Staš, 2010). The text gathering agent must fulfill tasks such as HTTP header parsing for handling redirections and MIME type resolution, and HTML parsing for encoding detection, HTML entity replacement and link extraction for further web exploration. Text gathering traverses web pages, extracts raw text, and calculates document error rate using a dictionary. Then it stores the document into relational database.

One document contains a certain amount of texts that can be used for creation of a text corpus. Process of acquisition of the text and sorting out irrelevant text is called *extraction*.

3 Data Processing

The acquired raw text data have to be processed to obtain a training corpus, ready for creation of the statistical language model. Raw text have to be *segmented*, *filtered*, and *transcribed*.

3.1 Text Segmentation

The first step is text segmentation – splitting the text into tokens, where each token expresses a single unit with certain meaning. The end of a sentence is also detected and marked as a self-standing token `</s>`. This step will make further processing much easier.

Text written by a human is ambiguous – white spaces separating words can be sometimes omitted or can be repeating. Dots can mean the end of a sentence or be a part of a number, abbreviation or date. In the formatted documents, new lines can mean the end of a sentence or be a part of lists.

For resolving this problem, a set of *context-free rules* (Thurston, 2006) to identify semantic units that should act as a single token has been designed. According to the context, the following entities are identified: words, numbers, dates, abbreviations, acronyms, lists, quotes, punctuation titles, paragraphs, sparse words, divided words, parentheses, URLs, e-mails, etc.

Let us have a look at the following example:

```
Z m l u v a
Spol. DAN s. r. o. sa zakladá 20. 4. 2004 v Ban. Bystrici.
```

Correct tokenization and sentence identification influences every next step in the process of training corpus creation. The desired of the text segmentation should look like this:

```
<s> Zmluva </s>
<s> Spol. DAN s.r.o. sa zakladá 20.4.2004 v Ban. Bystrici . </s>
```

The designed rules can be described by an antecedent and consequent part:

- **Section Heading** – a short text between two empty lines;
- **Sparse Word** – a word that has spaces inside to emphasize it;
- **Enumeration** – consists of more lines, starting with number;
- **Abbreviation** – a string, consisting of several characters finished by a dot. This entity in ambiguous – word ending with a dot does not have to be an abbreviation. To resolve if a word ending with a dot is an abbreviation, an abbreviation dictionary is used. If the word is an abbreviation, it is expanded and added to the sentence. Otherwise, it is handled like a normal word and the end of a sentence is marked. In some special cases it can begin with a capital letter or contain more parts, for example: `Z.z.` – Zbierka zákonov (Code of laws);
- **Word** – regular word;
- **Date** – entity consisting of tokens – number dot number. In the Slovak language it consists of one number with a dot expressing the day followed by a number expressing the month (for example: `10. 12.`). The year number does not have to be incorporated into the date entity because it is rewritten just like basic number. The date entity is easily distinguished from a floating point number because in the Slovak text notation with comma is used (for example: `13,4`);
- **Order Numeral** – number with dot, not matched as a date;
- **Basic Numeral** – number that does not belong to the previous entities. A basic numeral occurs usually as a string of numbers;
- **Special Symbol or Punctuation**;
- **End of Sentence** – consists of a dot that was not matched by previous entities (abbreviation, date, or order numeral);
- **E-mail or URL**.

Every other entity or character that is not matched by the previous rules can be omitted. The correctly tokenized text greatly simplifies the rest of the training corpus preparation. If entities described above are properly separated by space and unnecessary spaces are removed, it is easier to identify a token for the purpose of filtration or transcription. In most cases is not necessary to use a regular expression, and this feature makes the system faster and less error prone.

3.2 Text Filtration

The training corpus should contain text that is representative to the target domain of the designed speech recognition system. All texts that are from the web can contain a lot of irrelevant data that are not useful for the resulting corpus. The bad text contains foreign language words, parts of web page, graphical user interface, text advertisements or repeating headlines. Those parts of text that are not expected to occur in the input of the system should be removed.

An example of irrelevant text:

```
Motorky Nákladné autá Osobné autá Poľnohospodárske stroje Bicykle
Strana 1 2 3 4 5 6 7 8 9 10
To be or not to be?
. . . . . .
Sekcia domáci nábytok
Sekcia domáci nábytok
Sekcia domáci nábytok
Sekcia domáci nábytok
```

Irrelevant text is identified and removed using a set of filters. Heuristic procedure to evaluate quality of the sentence comes from the following presumptions:

- A sentence is irrelevant when it repeats too often.
- A sentence is irrelevant when it contains too many out-of-vocabulary words.
- A sentence is irrelevant when it contains too many non-word tokens.

Count-based Filter. Too often repeating sentences, such as headlines or advertisements, are not useful to the corpus. These sentences have to be removed by the sentence count filter. The counting algorithm uses *hash code of the sentence* to quickly determine how many times a sentence appeared in the corpus. A hash code is calculated for each sentence, and hash codes are counted. If the sentence occurs too many times, it is discarded.

Dictionary-based Filter. Some sentences, even those with low occurrence, can degrade value of a training text and can contain too many out-of-vocabulary (OOV) words. These sentences have to be removed by a dictionary based filter. For each sentence, an *error coefficient* E_d is calculated according to the following simple formula:

$$E_d = \frac{N_{OOV}}{N_t}, \tag{1}$$

where N_{OOV} is a number of out-of-vocabulary words in the sentence and N_t is the total number of tokens in the sentence. Using E_d, every sentence can be evaluated and those with high error can be removed.

Token-based Filter. In a similar way, non-word entities, such as numerals or some invalid strings, can be used to identify irrelevant sentence. The *error coefficient* E_t is calculated as

$$E_t = \frac{N_{nw}}{N_t}, \tag{2}$$

where N_{nw} is a number of tokens that are not words. This filter can identify sentences with too high occurrence of random characters or numbers.

3.3 Text Transcription

The last step in the training corpus preparation is text transcription. The main goal of this step is to put it in a form that is close to the spoken language. Every graphical symbol should have assigned its word form.

Text transcription examples:

```
Dvaja ľudia 2. apríla zaplatili 2,- Sk .
dvaja ľudia druhého apríla zaplatili dve koruny bodka

Podľa § 32 písm. C Z.z. ,
podľa paragrafu tridsať dva písmeno C zbierky zákonov čiarka
```

There are several types of entities in the gathered text that should have to be transcribed to a spoken form. For example, order numerals or abbreviations have to be rewritten to word form; regular words just have to be checked for correctness, converted to lowercase and added to the corpus. Also, it is useful to detect common typing errors and correct them.

Thanks to the tokenization, performed in the previous step, it is easier to detect necessary entities. In most cases, a *hash table-based transcription rule* can be used. This approach allows processing a large quantity of rules.

Transcription rules can be divided into categories according to the complexity of their antecedent part:

1. **Single Token** (the fastest) – in this case, the antecedent part of a transcription rule consists of a single token. It is not necessary to take any context into the account:
   - abbreviations;
   - long abbreviations;
   - date transcriptions;
   - acronyms;
   - dates;
   - special symbols – expansion of some special characters, such as % or §, have to be rewritten as words.
2. **Single Token and Context** – this type of transcription rule is a little more complex because it takes a surrounding context into the account. This rule still can be implemented using only a hash table. It is useful when correct form of the identified token requires declination, such as number transcription. A very hard problem is determining the correct grammatical case and gender of the numeral. This have to be found out by inspecting context of the numeral, usually the following word. An approach for the Slovak Morphological Analyzer is described in (Garabík, 2006).
3. **Regular Expression** – in this case, a token and a certain part of the surrounding tokens is used as an antecedent part of the transcription rule. The regular expression rule has to be applied to the whole sentence, and it is very computationally complex. On the other hand, it can catch very special grammatical cases and can be used for correction of the common typing errors. A more difficult analysis based on the morphological structure of the Slovak

language requires using a part-of-speech tagging system (Hládek et al., 2011; Spoustová et al., 2009), such as:

- multi-word expressions;
- spelling errors;
- named entities.

4 Experimental Results

The result of the whole process is a training corpus, ready for creating a statistical model of the Slovak language for a selected domain. The text gathering agent has downloaded documents:

- Court of Justice texts and laws – 151 340 items;
- Government and local authority web pages – 758 167 items;
- Blogs – 656 632 items;
- News – 1 881 193 items;
- Diploma theses – 797 557 items;
- Old literature – 2 168 items;
- Unsorted web pages – 5 099 102 items.

From this database, we get a 15 GB training corpus consisting of 2 thousand million tokens and 100 million of sentences. This training corpus has been used to build a language model for speech recognition tasks:

- Slovak Court of Justice (Accuracy = 95%);
- Parliament speech transcription (Accuracy = 85%);
- Broadcast news transcription (Accuracy = 90%).

5 Conclusion

In this paper, a brief overview of creation of the training corpus is described. This approach has been used to construct various statistical models of the Slovak language that have been used in a real-world applications such as transcription of dictation for the Slovak Ministry of Justice.

Acknowledgements

The research presented in this paper was supported by the Research and Development Operational Program funded by the ERDF under the project IMTS-26220220141 (50%) and IMTS-26220220155 (50%).

References

Garabík, R. (2006). Slovak Morphology Analyzer based on Levenshtein Edit Operations. In *Proc. of the 1st Workshop on Intelligent and Knowledge Oriented Technologies, WIKT'06*, pages 2–5, Slovak Academy of Sciences, Bratislava, Slovak Republic.

Hatcher, E. and Gospodnetic, O. (2004). *Lucene in Action*. Manning Publications.

Hládek, D. and Staš, J. (2010). Text Mining and Processing for Corpora Creation in Slovak Language. *Journal of Computer Science and Control Systems*, 3(1):65–68.

Hládek, D., Staš, J., and Juhár, J. (2011). A Morphological Tagger Based on a Learning Classifier System. *Journal of Electrical and Electronics Engineering*, 4(1):65–70.

Horák, A., Gianitsová, L., Šimková, M., Šmotlák, M., and Garabík, R. (2004). Slovak National Corpus. In Sojka, P., Kopeček, I., and Pala, K., editors, *Text, Speech and Dialogue, 7th International Conference, TSD'04*, pages 89–94, Berlin–Heidelberg. Springer-Verlag.

Juhár, J., Staš, J., and Hládek, D. (2012). Recent Progress in Development of Language Model for Slovak Large Vocabulary Continuous Speech Recognition. *InTech, Open Access Publishing*, pages 261–276.

Ondáš, S., Juhár, J., and Čižmár, A. (2011). Extracting Sentence Elements for the Natural Language Understanding based on Slovak National Corpus. *Springer-Verlag Berlin Heidelberg, LNCS 6800*, pages 171–177.

Pleva, M., Juhár, J., and Čižmár, A. (2007). Slovak Broadcast News Speech Corpus for Automatic Speech Recognition. In *Proc. of the 8th International Conference on Research in Telecomunication Technology, RTT'07*, pages 334–337, Liptovský Ján, Slovak Republic.

Slovak National Corpus (2012). Ľ. Štúr Institute of Linguistics, Slovak Academy of Science, Bratislava, Slovak Republic.

Spoustová, D., Hajič, J., Raab, J., and Spousta, M. (2009). Semi-Supervised Training for the Averaged Perceptron POS Tagger. In *Proc. of the 12th Conference of the European Chapter of the Association for Computational Linguistics, EACL'09*, pages 763–771, Athens, Greece.

Thurston, A. D. (2006). Parsing Computer Languages with an Automaton Compiled from a Single Regular Expression. In *Proc. of the 11th International Conference on Implementation and Application of Automata, CIAA'06*, pages 285–286, Taipei, Taiwan.

Collaboratively Developed Lexical Resources for Bulgarian with Application to Dictionaries and Reference Sources Compilation

Velislava Stoykova

Institute for Bulgarian Language, Bulgarian Academy of Sciences, Sofia, Bulgaria

Abstract. The paper presents results from several collaborative projects of the Institute for Bulgarian Language (IBL) – Bulgarian Academy of Sciences in designing available electronic text corpora and related language technology applications used for fast and up-to-date production of dictionaries and reference sources of Bulgarian language. It summarizes results of our finished and on-going projects and their approaches to develop sustainable strategy to use electronic text corpora for both academic research with theoretical value and for practical applications in compiling different types of dictionaries and reference sources.

1 Introduction

The electronic text corpora available at the Institute for Bulgarian Language (IBL) have been recently created within the framework of several collaboratively developed projects, and serve as the largest publicly available electronic lexical database for Bulgarian language. The scale and the use of this huge resource is a reliable base for both the theoretical research and practical applications which allow storage, retrieval and search procedures.

At the same time, the volume and content of stored lexica allow computer-assisted applications for lexicography where balanced corpora and related software (with representations of language-specific grammar features) are used for fast production and updating of various types of dictionaries and other lexical reference sources (including that for specialized lexica with different domain applications).

2 General purpose corpora

The tradition behind the creation and use of general purpose corpora in linguistic research and dictionaries compilation for Bulgarian language dates back to 70 years ago, and includes creation of printed index cards archives available at the IBL (which consist of almost 7 million index card files). They present text examples of manually extracted word contexts from texts of different authors, genres, styles, periods and thematic domains. The balanced approach used for designing of printed archives was later adopted for the creation of Bulgarian National Corpus – the largest electronic database for Bulgarian language.

2.1 Bulgarian National Corpus (BulNC) – design and applications

The BulNC (Koeva et al., 2010) was developed within the framework of several collaborative projects between the Department of Bulgarian Lexicology and Lexicography and the Department of Computational Linguistics at the IBL and is open for further enlargement and improvement. The BulNC is a large-scale representative on-line corpus of Bulgarian. It includes more than 320 000 000 words and utilizes software applications for storage, retrieval, and search (some of which are language-specific and still under development).

The BulNC incorporates modern original and translated Bulgarian texts from the middle of the 20th century until present days but mostly created after 1945, and hence, presents the modern Bulgarian language. Generally, the BulNC includes electronic lexicographic archive and consists of four general sub-corpora: the Bulgarian Brown Corpus, the Structural Corpus of Bulgarian Electronic Documents (2001–2009), the Structural Corpus of Bulgarian Printed Editions (1945–2009), and transcripts of spoken data. The distribution of genres of texts included in subcorpora is as follows:

Bulgarian Brown Corpus – consists of texts from the Internet (originals and translations) published within the period 1990–2005 which represents mostly fiction with the size of about 4 million words.

Structural Corpus of Bulgarian Electronic Documents (2001–2009) – includes mostly non-fiction from the following domains: economy, politics, law, medicine, sport, administration, journalism and fiction. It consists of texts of about 292 million words.

Structural Corpus of Bulgarian Printed Editions (1945–2009) – incorporates electronic versions of books (originals and translations) as well as periodicals, and includes texts from the following genres: prose, poetry, tales, plays and folklore. It consists of texts of about 28 million words.

The corpora are provided with detailed standardized metadata descriptions which include information about the title, author(s), thematic domain, genre, source, etc. of incorporated texts. The BulNC is provided, also, with facilities for advanced search which can perform various types of query search (mostly by string but also search for collocations and concordances).

The BulNC is recently used for the project *Dictionary of Bulgarian Language (DBL)* – for a compilation of the multivolume academic explanatory dictionary prepared at the IBL. Basically, the BulNC is extensively applied as a general lexical resource for word sense definitions in compilation of recently published volumes of DBL (vol. XII–XIV) and for updating of initial volumes (vol. I–V).

It was also used as a main lexical resource for compilation of various concise and updated dictionaries of Bulgarian of different lexicographic genres, recently published like the *Dictionary of New Words (2011)* and the *Dictionary of Bulgarian Synonyms with Antonyms (2012)*.

The BulNC is publicly available in two electronic versions – web-based and Sketch Engine application, and they both allow on-line search procedures for general and language-specific queries. The web-based version of BulNC is available free of charge and can be found on `http://search.dcl.bas.bg/`.

Further, we are going to compare search results to extract conceptual relations for the example word *function* (функция). We will analyse results from the BulNC (which is a general-purpose corpus) and for two specialized corpora – MathWikiBul and MathWiki – so to demon-

strate how corpora of different genres can generate different types of semantic conceptual relations.

2.2 Advanced search

The BulNC on-line search facilities allow extraction of context information representing sentence examples of a related query from all incorporated text sources. The search includes word forms, binary lexical relations (semantic relations) and feature structures (grammatical relations). The queries allow to specify word combinations, ordered words, inflectionally related words, semantically related words, part-of-speech, etc. However, the search facilities of BulNC are still under elaboration.

Nevertheless, the BulNC is successfully used for compilation of various dictionaries of Bulgarian language, mostly by searching semantically related words and extracting semantic relations. That type of search allows to generate the concordances and collocations for a related word.

Fig. 1 shows the BulNC produced concordances (all occurrences) of the word *function* (функция). The resulted word examples sentences represent quantitative contexts which can be further semantically interpreted to formulate different meanings of a word *function*.

The generated contexts from BulNC represent sentence examples on the base of which the general meanings of the word *function* can be defined. Thus, the BulNC induced contexts give common lexica definitions due to the balanced content of the corpus.

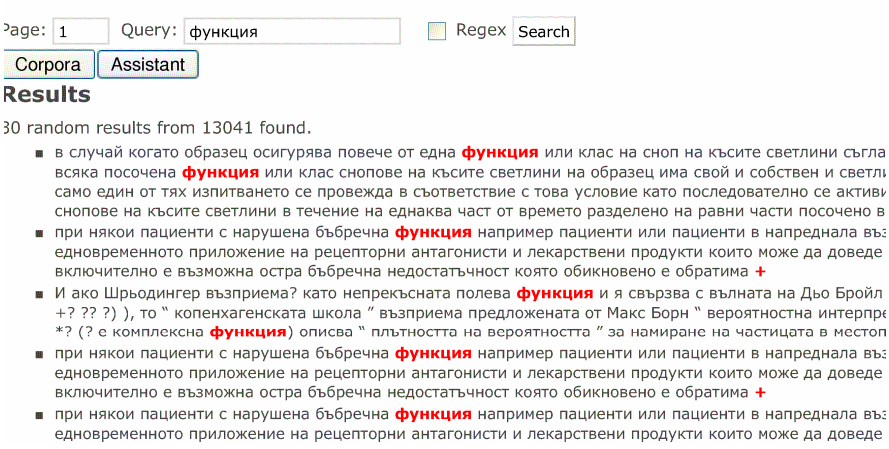


Fig. 1. The results of advanced search for the word *function* (функция) from BulNC

3 Specialized lexica corpora

The specialized lexica corpora are designed for a particular domain. Such types of corpora for Bulgarian are in a process of creation, and require special efforts to define principles to assemble the balanced content and the distribution of included subcorpora, so to represent

properly the specialized knowledge of a related domain. Specialized electronic text corpora are reliable source for extracting and updating semantic relations for terms of a particular domain in reference sources compilation. Specialized corpora can be used for preparation of both domain reference sources or study materials where a structured knowledge of related domain is presented in language-independent way.

3.1 MathWikiBul and MathWiki corpora

The corpus-based approach has been used for creation of highly structured semantically oriented and up-to-date specialized lexical resources in Bulgarian within the framework of innovative project *Conceptual Semantic Network Representation* aimed to use statistical search to extract semantic conceptual relations from specialized corpora in the domain of mathematics.

The project uses two web-based electronic text corpora MathWikiBul (in Bulgarian) and MathWiki (in English) consisting mostly of encyclopedic texts from Wikipedia (approximately 150 000 words each) in the domain of precalculus. We employ comparative corpora analysis approach to relate results from search of both corpora and both languages. The corpora are compiled and the Sketch Engine (SE) lexicographic software is used for processing.

Fig. 2. The generated keywords from MathWiki corpus

We define semantic relations by generating word contexts through extraction of keywords, word concordances, collocations and co-occurrences using different statistical approaches based on retrieval and clustering of statistically similar words (Lin, 2002). Additionally, the comparison of search results between general and specialized corpora is used to define general and specialized semantic relations by using semantic distance.

3.2 Semantic search

Semantic conceptual relations are viewed as horizontal and vertical. The horizontal relations are those of synonymy, antonymy, meronymy and show semantic similarity (Spärck Jones, 1986) or semantic distance. The vertical relations express the semantics of ordering or hierarchy by hyperonymy and hyponymy.

The SE (Kilgarriff and Rundell, 2002; Kilgarriff et al., 2004) software for processing electronic text corpora allows use of combined statistical approaches for semantically similar words extraction and comparison of results between several corpora which allows multilingual applications. It performs search for keywords, concordances, collocations and co-occurrences for a related word.

The keywords are evaluated on the base of word frequency lists and can be generated by using statistical search (Kilgarriff and Rundell, 2002) and ranking. The generated keywords from MathWiki are shown at Fig. 2. The results are presented, also, in (Stoykova and Mitkova, 2011; Stoykova and Petkova, 2012) and relate the same basic precalculus concepts like *function(s)*, *numbers*, *polynomials*, *graphs*, *equations*, etc. which were extracted with the same approach from MathWikiBul.

The results are similar since both corpora present the same specialized mathematical knowledge in different languages. However for more detailed semantic analysis a generation of concordances, collocations and co-occurrences is needed.

Concordances are quantitative contexts of a related word and give all its occurrences within the whole processed corpus. The SE generated concordances of the keyword *function* (функция) from MathWikiBul and MathWiki are shown at Fig. 3. Compared with the concordances generated by the BulNC for the same word (Fig. 1) they present in very many of their examples a specialized mathematical knowledge through formulas expressions and do not give any general meaning of the investigated word.

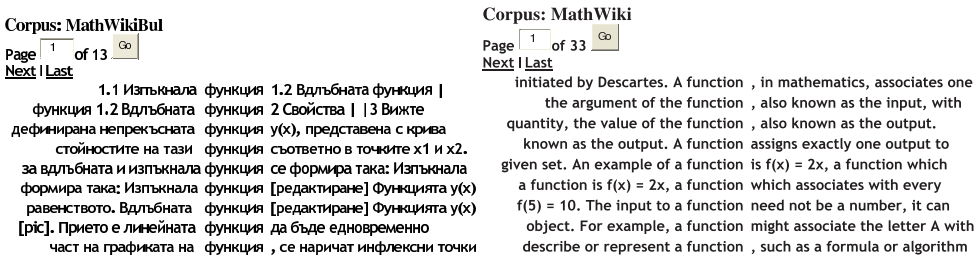


Fig. 3. Concordances of the keyword *function* (функция) from MathWikiBul and MathWiki

The basic conceptual relations definitions are generated by the use of collocations and co-occurrences of a related word. Collocations and co-occurrences are words which are most probably to be found with a related word. We use the techniques of *T-score*, *MI-score* (Church and Hanks, 1991) *MI3-score* (Oakes, 1998) incorporated in the SE.

The search results from processing MathWikiBul and MathWiki for collocations of the keyword *function* are similar for both Bulgarian and English. They present the most frequent content words which are most probably to be found with the keyword *function* and define its semantically related concepts.

The search results are shown at Fig. 4. They use *T-score* criterion for ranging the semantically related concepts but results according to *MI-score* and *MI3-score* criteria are offered as well. The results present the concepts *exponential*, *rational*, *polynomial*, *complex*, *logarithmic*, *trigonometric*, etc. as semantically related to the keyword *function*.

Collocation candidates MathWikiBul

Page 1 Go
Next >

	Freq	T-score	MI	MI3
p/n аналитична	17	4.103	7.707	15.882
p/n реална	14	3.725	7.805	15.420
p/n тригонометричните	12	3.452	8.204	15.374
p/n комплексна	11	3.302	7.872	14.791
p/n непрекъсната	10	3.135	6.834	13.478
p/n примитивна	7	2.581	5.346	10.961
p/n косинус	5	2.222	7.320	11.964
p/n холоморфна	4	1.983	6.860	10.860
p/n линейна	4	1.978	6.513	10.513
p/n обратна	4	1.975	6.320	10.320
p/n котангенс	3	1.719	7.098	10.267
p/n проста	3	1.719	7.098	10.267

Collocation candidates MathWiki

Page 1 Go
Next >

		Freq	T-score	MI	MI3
p/n	exponential	65	8.001	7.040	19.085
p/n	inverse	29	5.309	6.152	15.868
p/n	rational	24	4.804	5.689	14.859
p/n	propositional	20	4.437	7.011	15.655
p/n	polynomial	23	4.544	4.253	13.301
p/n	complex	19	3.978	3.517	12.013
p/n	logarithm	12	3.268	4.146	11.316
p/n	trigonometric	11	3.072	3.764	10.683
p/n	increasing	8	2.794	6.357	12.357
p/n	tangent	8	2.721	4.716	10.716
p/n	relation	8	2.719	4.689	10.689
p/n	continuous	7	2.605	6.004	11.618

Fig. 4. Collocation words of keyword *function* (функция) from MathWikiBul and MathWiki

3.3 Conceptual semantic hierarchy

The statistical similarity does not represent always synonymy and in the case, it expresses hierarchical conceptual relations of the basic concept *function* and the collocated concepts. The conceptual semantic term relations extracted by collocations and co-occurrences mostly represent vertical semantic relations like hyponymy or hyperonymy.

Thus, for our research results, we are using such interpretation and we define *polynomial function*, *exponential function*, and *rational function* as the most important hyponymic concepts of the hyperonym conceptual term *complex function*.

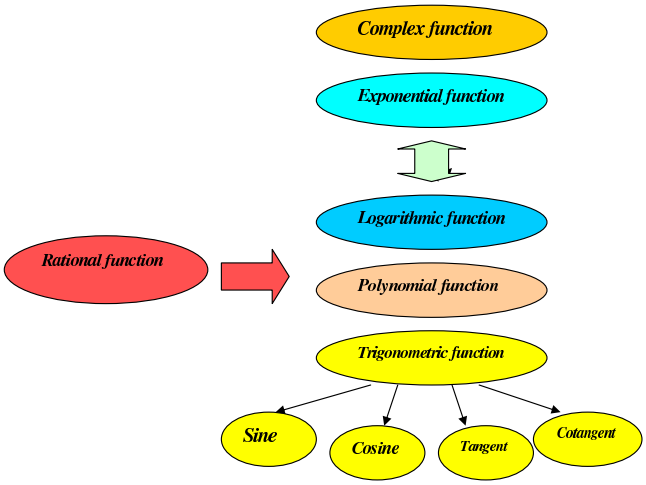


Fig. 5. Conceptual semantic hierarchy of basic concept *function*

The constructed conceptual semantic hierarchy (Fig. 5) of the basic domain concepts underlay internal domain knowledge representation in ontology-like style. The resulting semantic interpretation is language-independent and presents domain knowledge of precalculus.

4 Conclusion

The extracted terms and semantic relations show that using general and specialized text corpora and statistically-based search techniques to extract keywords, collocations and co-occurrence words are effective approach for mathematical conceptual precalculus terms extraction. The terms are evaluated on the base of their high frequency in the MathWikiBul and MathWiki corpora and their relatively low frequency in the BulNC.

The methodology is used for fast production of up-to-date terminological reference sources (like specialized dictionaries or thesauri) or building ontology (for defining the logical relations, conceptual relations or hierarchical semantic relations).

References

Church, K. and Hanks, P. (1991). Word Association Norms, Mutual Information and Lexicography. *Computational Linguistics*, (16:1):22–29.

Kilgarriff, A. and Rundell, M. (2002). Lexical Profiling Software and its Lexicographic Applications: a Case Study. In *Proceedings from EURALEX 2002*, pages 807–811.

Kilgarriff, A., Rychlý, P., Smrž, P., and Tugwell, D. (2004). The Sketch Engine. In *Proceedings from EURALEX 2004*, pages 105–116.

Koeva, S., Blagoeva, D., and Kolkovska, S. (2010). Bulgarian National Corpus Project. In *LREC 2010 Proceedings*, pages 3678–3684.

Lin, D. (2002). Automatic Retrieval and Clustering of Similar Words. In *Proceedings of the COLING-ACL*, pages 768–774.

Oakes, M. (1998). *Statistics for Corpus Linguistics*. Edinburgh University Press, Edinburgh, the United Kingdom.

Spärck Jones, K. (1986). *Synonymy and Semantic Classification*. Edinburgh University Press, Edinburgh, the United Kingdom.

Stoykova, V. and Mitkova, M. (2011). Conceptual Semantic Relationships for Terms of Precalculus Study. *WSEAS Transactions on Advances in Engineering Education*, 8(1):13–22.

Stoykova, V. and Petkova, E. (2012). Automatic extraction of mathematical terms for precalculus. *Proceedia Technology*, 1:464–468.

J A Z Y K O V E D N É
Š T Ú D I E XXXI

Rozvoj jazykových technológií a zdrojov na Slovensku a vo svete (10 rokov Slovenského národného korpusu)

Editorky
Mgr. Katarína Gajdošová
Mgr. Adriána Žáková

Obálka
Jozef Michaláč

Zodpovedný redaktor vydavateľstva VEDA
Emil Borčin

Technickí redaktori
Mgr. Radoslav Brída
RNDr. Radovan Garabík
RNDr. Ján Mášik

Prvé vydanie
Vydala VEDA,
vydavateľstvo Slovenskej akadémie vied, v Bratislave
v roku 2014
ako svoju 4074. publikáciu
z tlačových podkladov
Jazykovedného ústavu Ľ. Štúra SAV.
190 strán.
Náklad 150 výtlačkov.

ISBN 978-80-224-1391-6