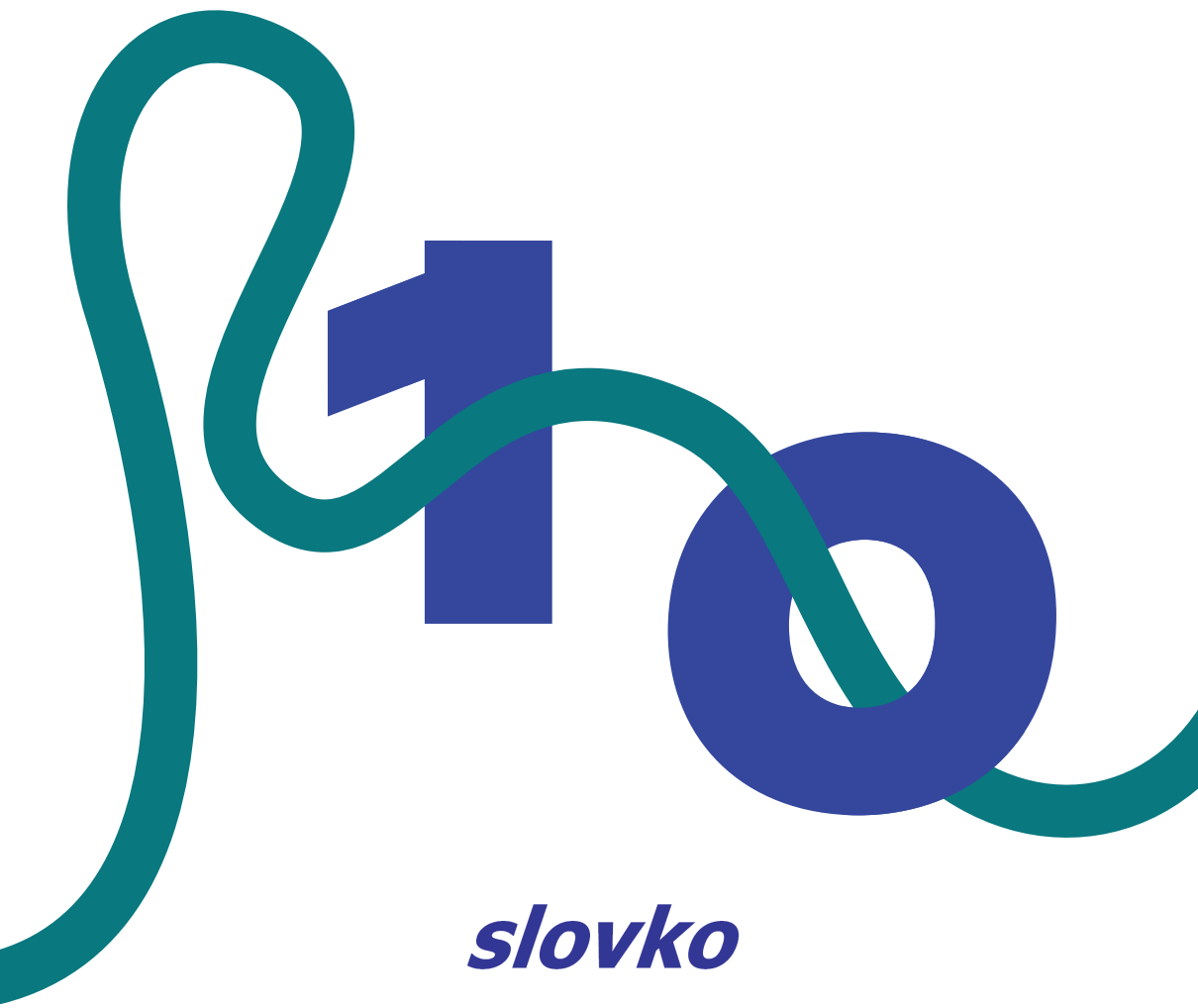
OD SLOVKA K SLOVKU
FROM SLOVKO TO SLOVKO



slovko

od Slovka k Slovku
from Slovko to Slovko

Zostavovateľka
Mária Šimková

Grafická úprava
Michal Slovák

Logo Slovka
Vladimír Benko

Návrh obálky 2001
Hana Kohútová
(s použitím grafického motívu
P. Kleeho)

Návrh obálok 2005, 2007, 2009, 2011,
2013, 2015
Vladimír Benko

ISBN
978-80-971690-4-6

Slovenčina a čeština v počítačovom spracovaní

Obsah

2001

Slovenčina a čeština v počítačovom spracovaní

Malá inventúra pred hľadaním spoločného jazyka

2001

Seminár *Slovenčina a čeština v počítačovom spracovaní* sme organizovali so zámerom zlepšiť vzájomnú informovanosť ľudí zaoberajúcich sa na Slovensku problematikou počítačov vo vzťahu k jazyku a naopak, jazyka vo vzťahu k počítačom. V tejto oblasti sú u nás značne izolované ostrovčeky aktivít prebiehajúcich v odlišných vedných odboroch a teoreticko-aplikačných kontextoch. Tieto sa zameriavajú len na neveľké úseky a komponenty odboru počítačová lingvistika, pod ktorým možno chápať spolu s automatizovaným spracovaním jazykových dát (vývoj a použitie príslušných programových nástrojov) aj teoretickú vy-
´stavbu jazykovedy v podobe modelov založených na využívaní jazykových znalostí v procedurálnej aj deklaratívnej podobe (gramatické pravidlá

a počítačové lexikóny vybavené vysoko štruktúrovanou explicitnou informáciou, ako aj procedúry túto informáciu využívajúce), alebo v podobe modelov založených na pravdepodobnosti (stochastické gramatiky). V tomto zmysle môžu byť pre nás relevantné aj aktivity v oblasti umelej inteligencie a niektoré komponenty kognitívnej lingvistiky (scenáre, rámce a prototypy). Rozšírenie záberu má svoje opodstatnenie, pretože v pozadí formalizovaných a explicitných modelov stojí zreteľ k možným aplikáciám a naopak, technické aplikácie implementujú modely a výrazne stimulujú pohyb v teoretickej jazykovede vo všeobecnosti.

Na rozdiel od Slovenska je situácia v Českej republike diametrálne odlišná. Keďže česká počítačová lingvistika je cieľavedome budovaná ako samostatný odbor už viac ako tridsať rokov, predstavuje dnes široké spektrum bádateľských a aplikovaných aktivít od formálnej (matematickej) lingvistiky cez počítačovú lexikografiu, strojový preklad, spracovanie rečového signálu až po korpusovú lingvistiku. Druhým cieľom nášho seminára je preto poskytnúť slovenskej odbornej verejnosti a študentom z lingvistických aj nelingvistických odborov ucelenejší pohľad na výsledky práce v oblasti počítačového spracovania češtiny, ktorá patrí v tomto smere medzi európsku a vo viacerých parametroch aj medzi svetovú špičku. Tento prehľad nám sprostredkujú vedúce osobnosti v danom teoretickom alebo aplikovanom úseku.

Naším seminárom by sme chceli nadviazať na vedeckú, pedagogickú a organizačnú prácu Jána Horeckého, ktorý sa od začiatku šesťdesiatych rokov 20. storočia usiloval uplatňovať princípy a metódy matematickej lingvistiky na materiáli slovenského jazyka (v jeho bibliografii možno nájsť vyše päťdesiat štúdií a článkov s danou problematikou). Tieto poznatky sprostredkoval aj svojim študentom a ašpirantom v prednáškach z matematickej lingvistiky a základov jazykovedy (skriptum *Úvod do matematickej lingvistiky,* Bratislava 1969 a učebnica *Základy jazykovedy,* Bratislava 1978). J. Horecký sa usiloval zakotviť matematickú jazykovedu aj inštitucionálne, keď v roku 1962 založil oddelenie matematickej lingvistiky a fonetiky v bývalom Ústave slovenského jazyka SAV. V oddelení, ktoré viedol do roku 1970, pripravovali mladí adepti vedy K. Buzássyová, J. Bosák a R. Gedeon pomocou diernoštítkového stroja slovník tvarov a morfém a publikovali rad štúdií z oblasti štatistickej lingvistiky. V roku 1970 sa z iniciatívy J. Horeckého uskutočnilo v Smoleniciach medzinárodné sympózium o algebraickej lingvistike, na ktorom sa zúčastnili vedúce osobnosti pôsobiace v tomto odbore (materiály zo sympózia boli uverejnené v *Recueil Linguistique de Bratislava,* vol. IV, Bratislava 1973). Matematická jazykoveda v tom čase zahŕňala aj staršiu kvantitatívnu (štatistickú) lingvistiku, zaoberajúcu sa najmä skúmaním frekvencie jazykových jednotiek v textoch, a algebraickú jazykovedu ako súhrn teórií vypracovaných s použitím matematickej logiky, teórie množín, teórie grafov a iných súčastí teoretického aparátu matematiky.

Ak sme vyššie spomínali kvantita-

tívnu analýzu textov vyúsťujúcu do tvorby abecedných a konkordančných indexov a frekvenčných slovníkov, nemôžeme nespomenúť priekopníka kvantitatívnych metód v jazykovede a literárnej vede na Slovensku G. Altmanna a tiež autora frekvenčného slovníka slovenčiny (*Frekvencia slov a tvarov*, Bratislava 1969) a publikácie *Frekvencia tvarov a konštrukcií v slovenčine* (Bratislava 1985) J. Mistríka.

Počítačovú lingvistiku chápanú podľa J. Horeckého ako aplikovaná matematická lingvistika, predstavovali v tom čase vlastne len zahraničné práce na strojovom preklade. Začiatkom sedemdesiatych rokov viaceré administratívne rozhodnutia sčasti externé (vo vzťahu k odboru), sčasti vnútorné (súvisiace s organizačnými zmenami v Jazykovednom ústave Ľudovíta Štúra) spôsobili, že sa matematická lingvistika nemohla na Slovensku rozvíjať ako samostatný odbor. J. Horecký v nasledujúcich rokoch spolupracoval na niektorých problémoch automatizovanej morfologickej analýzy slovenčiny s českými kolegami z Ústavu formální a aplikované lingvistiky na Matematicko-fyzikálnej fakulte Karlovej univerzity v Prahe.

Vybrané lingvistické problémy (napr. lematizácia, tvorba sémantických sietí, rozpoznávací slovník) boli v osemdesiatych rokoch súčasťou riešenia úlohy spracovania textu pre potreby informačných (najmä knižničných) systémov (M. Cigánik, A. Appelová).

Absencia inštitucionálnej a personálnej základne počítačovej lingvistiky sa odrazila na malej pripravenosti slovenskej jazykovedy vyrovnať sa s búrlivým zahraničným vývojom v oblasti automatizovaného spracovania prirodzeného jazyka (Natural Language Processing – NLP). Tento typ výskumu nahradil koncom šesťdesiatych rokov problematiku strojového prekladu (stiahnutú z agendy administratívnym zásahom). Bádanie sa začalo orientovať na dostatočne presný opis základných prvkov a vzťahov jazykovej štruktúry vhodný na spracovanie do podoby počítačového programu. Mohlo preto slúžiť ako vhodná základňa na etablovanie počítačovej lingvistiky ako samostatnej vednej disciplíny a nie ako aplikácie matematickej lingvistiky.

Trocha lepšie je slovenská jazykoveda pripravená na novú empirickú orientáciu jazykovedy, ktorá sa prejavuje v budovaní reprezentatívnej materiálovej základne jazykovedného výskumu. Tieto rozsiahle zdroje autentického materiálu majú podobu textových databáz – korpusov. Korpusová lingvistika, ktorá sa zaoberá problematikou budovania a využívania textových korpusov, nie je len súborom progresívnych metód zhromažďovania a spracúvania materiálu, ale predstavuje v zásade iný pohľad na jazyk a jeho jednotky. Tu treba opäť spomenúť osobnosť J. Horeckého, ktorý stál aj pri revitalizácii počítačovej lingvistiky v Jazykovednom ústave, keď v rokoch 1988 – 1989 pripravil projekt bázy dát slovenského jazyka, v rámci ktorej sa začalo uvažovať aj o budovaní korpusu. Vďaka ústretovému prístupu riaditeľa Informačného centra SAV E. Kostolanského a jeho spolupracovníkov V. Benka a C. Belicu sa uskutočnilo niekoľko pracovných stretnutí, na ktorých sa prerokúvali možnosti

vytvorenia spoločnej pracovnej skupiny počítačovej lingvistiky. V roku 1990 táto skupina (J. Horecký, V. Benko, A. Jarošová, E. Páleš) vznikla v Jazykovednom ústave Ľ. Štúra SAV a začala koncepčne pripravovať korpus a lexikálnu bázu dát. Tento zámer sa začal v roku 1992 realizovať ako grantový projekt, bol však natoľko finančne a personálne poddimenzovaný (V. Benko, A. Jarošová, M. Šimková; všetci traja s polovičnou a menšou časovou kapacitou), že sa podarilo vybudovať len interný korpus Jazykovedného ústavu Ľ. Štúra. Ide zatiaľ o pomerne nevyvážený súbor textov (bez anotácií), rozdelený na niekoľko fragmentov, ktoré doteraz nie sú spojené do jedného celku najmä preto, že používaný vyhľadávací program nedokáže naraz spracovať väčší objem textu.

Začiatkom deväťdesiatych rokov E. Páleš teoreticky a aplikačne dopracúval počítačový model slovenčiny, ktorý v roku 1994 publikoval v monografii *SAPFO – parafrázovač slovenčiny – počítačový nástroj na modelovanie v jazykovede*. Začiatok deväťdesiatych rokov priniesol aj založenie Laboratória počítačovej lingvistiky na Pedagogickej fakulte Univerzity Komenského, kde sa pripravuje morfologická databáza slovenčiny (E. Kostolanský, J. Hašanová, V. Benko) a budúci učitelia a prekladatelia dostávajú v kurze prednášok informáciu o počítačovej podpore prekladu, o jazykových zdrojoch a počítačovej podpore výučby cudzieho jazyka. Nemáme presnú predstavu o mieste lingvistickej problematiky v štruktúre učebných predmetov v našich vzdelávacích inštitúciách prírodovedného a technického smeru, ale aktívna účasť študentov Technickej univerzity v Košiciach na medzinárodných kolokviách mladých jazykovedcov, ktoré už desať rokov organizuje pracovníčka Jazykovedného ústavu Ľ. Štúra M. Nábělková, svedčia o dobrých výsledkoch košických kolegov pracujúcich v odbore umelá inteligencia. Témy viacerých diplomových prác a tímových projektov študentov Matematicko-fyzikálnej fakulty a Fakulty elektrotechniky a informatiky Univerzity Komenského nás presviedčajú o tom, že na týchto pracoviskách nám vyrastajú partneri pre efektívny dialóg.

V roku 1999 sa v Bratislave uskutočnil medzinárodný seminár o inovačných prístupoch k viacjazyčným zdrojom a jazykovým technológiám *Text Corpora and Multilingual Lexicography* (Textové korpusy a viacjazyčná lexikografia). Toto podujatie, ktoré v spolupráci s Inštitútom nemeckého jazyka v Mannheime (W. Teubert, A. Lawson) organizovali Jazykovedný ústav Ľ. Štúra (A. Jarošová) a Pedagogická fakulta Univerzity Komenského (V. Benko), bolo súčasťou medzinárodného projektu Európskej komisie Trans-European Language Resources Infrastructure – COPERNICUS (Transeurópska štruktúra jazykových zdrojov). V týchto dňoch vrcholia organizačno-legislatívne aktivity smerujúce k založeniu korpusového pracoviska v rámci Jazykovedného ústavu Ľ. Štúra.

Veľmi si vážime záujem všetkých, ktorí sem prišli prednášať aj počúvať, diskutovať a hľadať riešenia. Vitajte!

Alexandra Jarošová

Slovko

Zoznam príspevkov

2003

Automatic annotation of measure expressions for Croatian
Božo Bekavac — Krešimir Šojat

From paper to collective knowledge in a „golden translation memory" (parallel texts compilation) and corpora
Jasna Belc — Simon Bratina

Rozšírenie možností syntetizátora reči pre účely služby SMS to Voice
Sachia Daržágín — Milan Rusko — Marián Trnka

Design and expectations for a comparable Slovak-Spanish English learner corpus
Ana Díaz — Salvador Valera

Russian Lemmatisation with DATR
Roger Evans — Carole Tiberius — Dunstan Brown — Greville Corbett

Návrh morfologického tagsetu SNK
Martina Forróová — Radovan Garabík — Lucia Gianitsová — Alexander Horák — Mária Šimková

Some Aspects of Tokenization and Segmentation in Slovak National Corpus
Radovan Garabík

Počítačová lingvistika na KPI FEI TU Košice
Ján Genči

Grapheme Frequencies in Slovene
Peter Grzybek — Emmerich Kelih

Využití ČNK a PZK pro ověřování valenčních vlastností deverbativních substantiv se zabudovanou rolí
Jana Klímová — Veronika Kolářová-Řezníčková

Charakterizácia slovného bohatstva textu
Jana Kusendová

Analyse automatique du slovaque: Analyseur morphologique
Diana Lemay

Automatická analýza češtiny bez slovníku a „předvídatelnost" západních slovanských jazyků
Patrice Pognan

Praktické možnosti použití česko-slovenského TTS systému s harmonickým modelem řeči
Jiří Přibil — Anna Přibilová

Bipartite Finite State Transducers as Morphology Analyser, Synthesizer, Lemmatizer and Unknown-Word Guesser
George Totkov — Rositsa Doneva

Slovko

Príhovor

2003

Vážené dámy a páni,

dovoľte, aby som Vás všetkých pozdravil v mene Jazykovedného ústavu Ľudovíta Štúra Slovenskej akadémie vied a takisto v mene Slovenskej jazykovednej spoločnosti pri SAV.

Ono Slovko, ako túto akciu trochu roztopašne, ale s citom pre mieru nazvali organizátori, nie je, ako všetci vieme, prvé. Kým však pri premiérovej akcii tento názov figuroval iba neoficiálne, teraz je to už pomenovanie celkom oficiálne. No dnešný medzinárodný seminár sa od zakladajúceho určite nelíši len v tejto jednej veci. Kým ten prvý v roku 2001 bol usporiadaný predovšetkým so zámerom zabezpečiť vzájomnú informovanosť ľudí zaoberajúcich sa na Slovensku problematikou počítačov a jazyka, ako to vtedy sformulovala Alexandra Jarošová, dnes, ako to zasa potvrdzuje nazretie do programu, má cieľ omnoho náročnejší a sofistikovanejší. Na seminári nás čakajú vystúpenia a diskusie domácich a zahraničných účastníkov na subtílnejšie a delikátnejšie témy. Je to možné aj preto, že počítačová a korpusová lingvistika i na Slovensku medzitým postúpila o míľový krok dopredu a na slovenskej jazykovednej scéne už zaujíma svoje oprávnené miesto. Už i viacerí klasickí jazykovedci, ktorí nikdy neboli a ani nie sú orientovaní technologicky ani programátorsky, začali chápať, že príchod novej technológie do jazykovedného výskumu znamená nové možnosti a zrejme aj novú etapu v lingvistickom myslení i v oblasti aplikovanej lingvistiky, vrátane prípravy lexikografických a gramatických diel.

Želám tomuto podujatiu príjemnú pracovnú atmosféru a vydarený priebeh. Už teraz sa teším na ďalší „slovkovský“ zborník, ktorého vydanie s potešením podporíme. Považujme týmto seminár za otvorený.

Bratislava 24. októbra 2003

Slavomír Ondrejovič
riaditeľ Jazykovedného ústavu
Ľ. Štúra SAV
predseda Slovenskej jazykovednej
spoločnosti pri SAV

Slovko

Table of Contents

2005

me Frequencies for Slavic Languages
Peter Grzybek and Emmerich Kelih

DaskaL – A Web¬based Application for Foreign Language Teaching
Kjetil Raa Hauge, Svetla Koeva, Emil Doychev and Georgi Cholakov

Aspects of an XML-Based Phraseology Database Application
Denis Helic and Peter Ďurčo

VerbaLex – New Comprehensive Lexicon of Verb Valencies for Czech
Dana Hlaváčková and Aleš Horák

Orwell's 1984 – Playing with Czech and Slovak Versions
Jaroslava Hlaváčová

Czech Language Parsing Using Meta Grammar Formalism with Contextual Constraints
Aleš Horák and Vladimír Kadlec

Analysis of Rule Based Phonetic Transcription Technique Applied to Slovak Language
Jozef Ivanecký

Construction of Spoken Corpus Based on the Material from the Language Area of Bohemia
Marie Kopřivová and Martina Waclawičová

Multimedia Reading Book – Utilization of an XML Document Format and Audio Signal Processing
Marek Nagy

Morphological Idiosyncrasy in Hungarian Multiword Expressions
Csaba Oravecz, Viktor Nagy and Károly Varasdi

Valency Frames and Semantic Roles in Czech
Karel Pala

Question Answering in Polish Using Shallow Parsing
Dariusz Piechociński and Agnieszka Mykowiecka

In Search of the Best Method for Sentence Alignment in Parallel Texts
Alexandr Rosen

Word Tests for Speech Understandability Evaluation in Slovak
Milan Rusko and Marián Trnka

Bulgarian and English Semantic Dictionaries for the Purposes of Information Retrieval
Max Silberztein and Svetla Koeva

Slavic Text Taggers Project
Danko Šipka

Multi Word Named Entity Recognition in Polish Texts
Dominika Urbańska and Agnieszka Mykowiecka

Creating of Slovak Electronic Phonetic Dictionary for Use in Speech Recognition
Pavol Vančo and Marek Nagy

Russian Historical Corpora of the 18th and 19th Centuries
Victor Zakharov

Building a Pilot Spoken Corpus
Jana Zemljarič Miklavčič and Marko Stabej

Slovko

Opening Speech

2005

Ladies and gentlemen,

I almost feel like saying that Slovko has come to Bratislava once more. In the beginning, as we all well know, was Slovo – the Word. Slovko – an event that has quickly become a tradition – can be viewed as its more familiar continuation here in Bratislava. While our inaugural linguistics conference, held in 2001, focused solely on the languages of the former Czechoslovakia (i.e. the proceedings of this conference, edited by Alexandra Jarošová, were entitled Computer Processing of the Slovak and Czech languages), the subsequent meeting was enhanced by other languages and, as a glance at the programme suggests, this trend is being maintained.

For some time now, the purpose of the conference has been not only to meet the need for mutual acquaintance and briefing with regard to each other's findings. It is also that Slovakia has succeeded in creating for itself favourable conditions within corpus and computer linguistics; Slovak linguists have made a successful entry into the international framework of these branches and have contributed to their development, so that their results – I trust you will not find my words too immodest – can be treated as substantive. My claim is to be confirmed by our scientific meeting, which holds out the promise of extremely interesting papers and discussions. I hope that I will not be divulging an official secret by saying that the outcomes and the work itself of the Slovak National Corpus team have not gone unnoticed by the higher authorities, who have decided to award them the prestigious Science and Technology Prize. The team are to receive their award from the Minister of Education of the Slovak Republic tomorrow evening.

On behalf of the Ľudovít Štúr Institute of Linguistics of the Slovak Academy of Sciences, I wish you a stimulating discussion and I hope that you will find our "Central European" conference pleasant and fruitful. Now, it is my privilege to declare Slovko 2005 officially open.

Bratislava, November 8th 2005

Slavomír Ondrejovič
Director, Ľudovít Štúr Institute of Linguistics

Slovko

Table of Contents

2007

A Cross-linguistic Database of Children's Printed Words in Three Slavic Languages
Radovan Garabík, Markéta Caravolas, Brett Kessler, Eva Höflerová, Jackie Masterson, Marína Mikulajová, Marcin Szczerbiński and Piotr Wierzchoń

Effective Methods of Building Slovak-Czech Dictionary
Marek Grác

Administration Framework for the DEB Dictionary Server
Aleš Horák and Adam Rambousek

Precision of Statistical Syllable Segmentation as a Function of Training Data Quality
Jozef Ivanecký and Daniela Majchráková

Program Concorder and Jaroslav Seifert's Individual Dictionary
Ladislav Janovec and Martin Wagenknecht

Collocations in Russian. Analysis of Association Measures
Maria Khokhlova

The Role of Word Frequency Vocabularies in the Research of Psychology and Philosophy Terminological Systems
Oksana S. Kozak

Variation of Czech Lexicon as Reflected by Corpora Comparison
Michal Křen

Hyperlemma: A concept Emerging from Lemmatizing Diachronic Corpora
Karel Kučera

Semi-automatic Semantic Annotation of Slovak Texts
Michal Laclavík, Marek Ciglan, Martin Šeleng, Stanislav Krajčí, Peter Vojtek and Ladislav Hluchý

Terminology and Terminological Activities in the Present-Day Slovakia
Jana Levická

Beyond Syntactic Valence: FrameNet Markup of Example Sentences in a Slovenian-German Online Dictionary
Birte Lönneker-Rodman

Automatic Word Clustering in Russian Texts based on Latent Semantic Analysis
Olga Mitrofanova, Polina Panicheva and Vyacheslav Savitsky

Corpus Analysis of Selectional Preferences in Russian
Olga Mitrofanova, Viktoria Belik and Vera Kadina

Lexterm, an Open Source Tool for Lexical Extraction
Joaquim Moré, Mercè Vázquez and Luis Villarejo

Tools for Working with Corpus Evidence in the Lexical Database LEXIKON 21 (Program PRAMAT and the Exemplification Tool)
Zdeňka Opavská and Barbora Štěpánková

Computer Processing Derivational Relations in Czech
Karel Pala and Dana Hlaváčková

Wider Framework of the Research Plan Creation of a Lexical Database of the Czech Language of the Beginning of the 21st Century
Albena Rangelova and Jan Králík

Optimization of Russian Bilingual Dictionaries
Elizaveta Rumyantseva

Corpus of Spoken Slovak Language
Milan Rusko and Radovan Garabík

Prosody Annotation in Slovak Using Sk-ToBI
Milan Rusko, Róbert Sabo and Martin Dzúr

The Possibilities and Limits of Lexicographical Description of the Czech Lexicon in Database Form
Jindra Světlá

Automatic Term Recognition in Polish Texts
Dominika Urbańska and Dariusz Piechociński

Parallel French-Slovak Corpus
Dorota Vasilišinová and Radovan Garabík

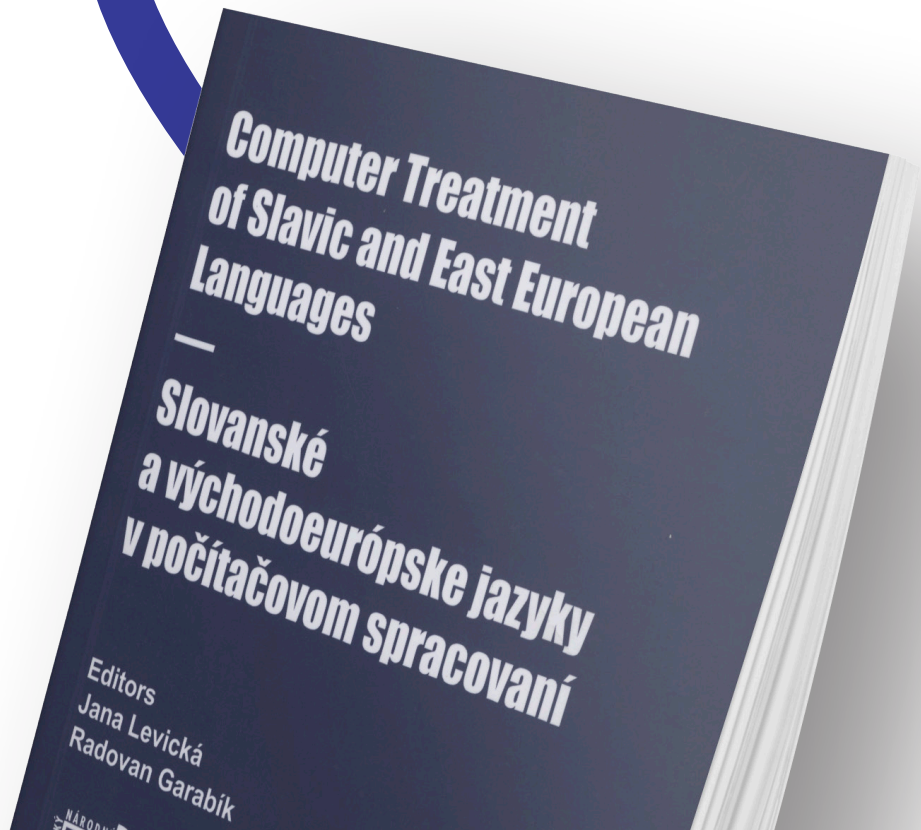
Tools for the Input of Morphological Data – L 21 Solution Proposal
Milada Voborská

Comparing Natural Language Identification Methods Based on Markov Processes
Peter Vojtek and Mária Bieliková

Spoken Corpus ORAL2006, Information It Provides and General Characteristics of Spoken Text
Martina Waclawičová

Citation Card Files, Corpora of the Past
Victor Zakharov

Povaha a úzus interjekcí: případ češtiny
František Čermák


Computer Treatment
of Slavic and East European
Languages
—
Slovanské
a východoeurópske jazyky
v počítačovom spracovaní
Editors
Jana Levická
Radovan Garabík

Slovko

Foreword

2007

The fourth edition of the biannual conference SLOVKO, focused on NLP, computational lexicography and terminology, only partially recalls on the founding seminar in 2001 when Slovak computational linguistics was literally in its infancy. The first edition only paid attention to Slovak and Czech and more or less highlighted the 30-year Czech tradition of this field, the second and third one kept enlarging in topics and countries and thus the proceedings of SLOVKO 2007 reflect the international character of this scientific event as it offers a greater and richer scope of the computational processing issues concerning not solely the Slavic languages.

Moreover, papers by Slovak authors, covering basic and applied research, indicate the revolutionary changes that Slovak computational and corpus linguistics have undergone since then. In 2001 the Ľudovít Štúr Institute of Linguistics, Slovak Academy of Sciences in Bratislava created a corpus lingustics department that has already managed to present the sixth version of a fully lemmatised and annotated general corpus containing 350 million tokens. Apart from the general corpus, the Department has been working on partial projects of parallel corpora, terminology database and at present an oral corpus.

The topics of the fourth edition include but were not limited to:

- theoretical issues of computational lexicography and terminography
- bilingual lexicography and terminography
- dictionary structure and dictionary entries compilation
- corpus development: data collection, annotation and processing, word sense disambiguation, co-occurrence analysis and pertinent collocations of lexicographical and terminographical relevance
- new methods in data extraction and terminology mining from corpora
- terminology databases and terminology management systems
- linguistic components of information systems

We would like to thank all authors for their effort and willingness to present results of their research within the SLOVKO 2007 conference. A word of appreciation and gratitude goes also to the program committee for helping to choose, evaluate and correct submitted abstracts. We hope that our conference will enhance fruitful discussions and mutual cooperation as well as projects, whose presentation we will heartily welcome at SLOVKO 2009.

Organisers

Slovko

Edičná poznámka

2007

Štvrtý ročník bienálnej konferencie SLOVKO, zameraný na NLP, komputačnú lexikografiu a terminológiu, len čiastočne pripomína zakladajúci seminár z roku 2001, keď bola slovenská počítačová lingvistika doslova v plienkach. Prvý ročník mal v centre pozornosti iba slovenčinu a češtinu a viac-menej prezentoval predovšetkým 30-ročnú českú tradíciu tohto odboru, druhý a tretí ročník sa postupne rozširovali tematicky aj teritoriálne tak, že zborník SLOVKO 2007 už odráža medzinárodný charakter podujatia, ponúkajúc širší a bohatší záber počítačového spracovania nielen slovanských jazykov.

Príspevky slovenských autorov, pokrývajúce základný aj aplikovaný výskum, poukazujú na prevratné zmeny, ktoré slovenská počítačová a najmä korpusová lingvistika za ten čas zaznamenala. Od roku 2002 sa v Jazykovednom ústave Ľudovíta Štúra SAV v Bratislave vybudovalo korpusové pracovisko, ktoré už stihlo sprístupniť odbornej verejnosť šiestu verziu lematizovaného a anotovaného všeobecného korpusu v rozsahu 350 miliónov tokenov. Popri tom oddelenie pracuje na čiastkových projektoch paralelných korpusov, terminologickej databázy a najnovšie aj hovoreného korpusu.

Hlavné témy 4. ročníka konferencie, ktorých sa zahraniční a slovenskí autori mohli pridržať, sú:

- teoretické otázky komputačnej lexikografie a terminografie
- bilingválna lexikografia a terminografia
- slovníková štruktúra a tvorba slovníkových hesiel
- tvorba korpusov: zber dát, anotácia a spracovanie, dezambiguácia, kookurenčná analýza a lexikograficky alebo terminograficky relevantné kolokácie
- nové metódy v extrahovaní dát a získavanie terminológie z korpusov
- terminologické databázy a systémy terminologického manažmentu
- lingvistické súčasti informačných systémov

Chceli by sme sa poďakovať všetkým autorom za úsilie a ochotu prezentovať výsledky svojej práce práve na konferencii SLOVKO 2007. Poďakovanie patrí zároveň vedeckému výboru za pomoc pri výbere, hodnotení a korigovaní abstraktov. Veríme, že naša konferencia napomôže vzájomnú diskusiu a nové spolupráce a projekty, ktorých prezentáciu uvítame na stretnutí SLOVKO 2009.

Organizátori

Slovko

Table of Contents

2009

Scientific Text Corpora as a Lexicographic Source
Larisa Belyaeva

The Corpus of Georgian Dialects
Marine Beridze and David Nadaraia

Corpus of Computational Linguistics Texts
Tatiana Bobkova, Mariia Kasianenko, Kuzma Lebedev, Valentyna Lukashevych, Pavlo Petrenko and Liubov Grydneva

"We Only Say We Are Certain When We Are Not": A Corpus-Based Study of Epistemic Stance
Vaclav Brezina

A Model for Corpus-Driven Exploration and Presentation of Multi-Word Expressions
Annelen Brunner and Kathrin Steyer

Text-Oriented Thesaurus Retrieval System for Linguistics
Natalia P. Darchuk and Viktor M. Sorokin

From Electronic Corpora to Online Dictionaries (on the Example of Bulgarian Language Resources)
Ludmila Dimitrova

Evaluating Grid Infrastructure for Natural Language Processing
Radovan Garabík, Jan Jona Javoršek and Tomaž Erjavec

Synset Building Based on Online Resources
Ján Genči

Shallow Ontology Based on Verba-Lex
Marek Grác

Multimodal Russian Corpus (MURCO): General Structure and User Interface
Elena Grishina

Electronic Lexical Card Index for the Ukrainian Dialects (ELCIUD)
Pavlo Grytsenko, Olena Siruk and Viktor M. Sorokin

Inflectional Entropy in Slovak
Adriana Hanulíková and Doug J. Davidson

Exploring Derivational Relations in Czech with the Deriv Tool
Dana Hlaváčková, Klára Osolsobě, Karel Pala and Pavel Šmerk

On Epistemicity, Grammatical Person and Speaker Deixis in Polish (Based on the Polish National Corpus)
Łukasz Jędrzejowski

A Russian EFL Learner Corpus from Scratch
Olga Kamshilova

Preliminary Analysis of a Slavic Parallel Corpus
Emmerich Kelih

Operators for Extending and Developing an Utterance (Based on Operators of Concessive Relation)
Jana Kesselová

Changes in Valency Structure of Verbs: Grammar vs. Lexicon
Václava Kettnerová and Markéta Lopatková

Corpus-Based Analysis of Lexico-Grammatical Patterns (on the Corpus of Letters of N. V. Gogol)
Maria Khokhlova and Victor Zakharov

'New/novelty' Concept Set Dynamics as a Marker of Lexical and Grammatical Paradigm Evolution for Psychology Sublanguage
Oksana S. Kozak

Methodological Foundations for Contrastive Model of Verb Valence
Ružena Kozmová

Dictionary of Štúr's Slovak
Ľubomír Kralčák

Annotation Procedure in Building the Prague Czech-English Dependency Treebank
Marie Mikulová and Jan Štěpánek

Automatic Analysis of Terminology in the Russian Corpus on Corpus Linguistics

Olga Mitrofanova and Victor Zakharov
Using Speech and Handwriting Recognition in Electronic School Worksheets
Marek Nagy

Composite Lexical Units as an Element of Lexicographical Historical Computer System
Irina Nekipelova

IT: Moving Towards Real Multilingualism
Antoni Oliver and Cristina Borrell

Introduction of Non-Verbal Means of Communication in the Corpus of Live Speech
Tatyana Petrova and Olga Lys

MorphCon – A Software for Conversion of Czech Morphological Tagsets
Petr Pořízka and Markus Schäfer

Recent Developments in the National Corpus of Polish
Adam Przepiórkowski, Rafał L. Górski, Marek Łaziński and Piotr Pęzik

Spoken Texts Representation in the Russian National Corpus: Spoken and Accentologic Sub-Corpora
Svetlana Savchuk

The Meaning of the Conditional Mood Within the Tectogrammatical Annotation of Prague Dependency Treebank 2.0
Magda Ševčíková

The Creation of the Morphological Ambiguity Depository in Ukrainian
Olga Shypnivska and Sergij Starykov

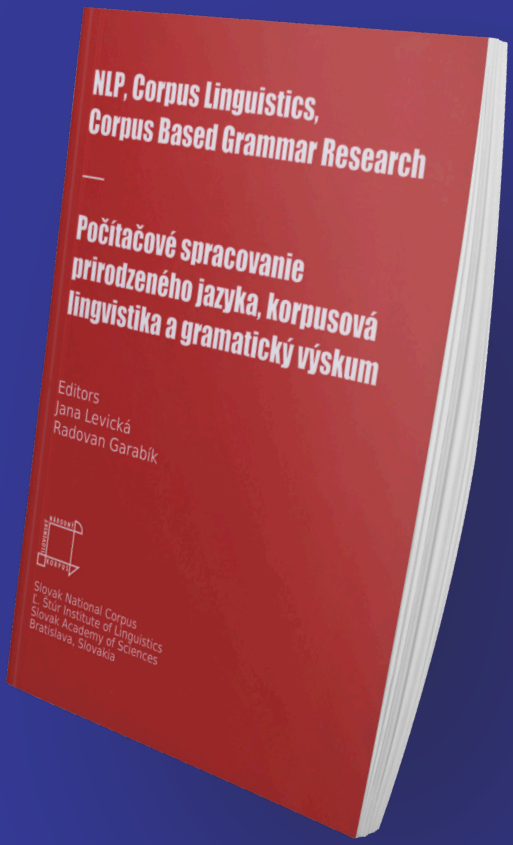
Frequency of Words and Their Forms in Contemporary Slovak Language Based on the Slovak National Corpus
Mária Šimková and Miroslav Ľos

Analysis of the Means Expressing Strong 'Necessity Not To' in English and Czech Based on General and Parallel Corpora
Renata Šimůnková

Diatheses in the Czech Valency Lexicon PDT-Vallex
Zdeňka Urešová and Petr Pajas

A Corpus of Spoken Language and Its Usefulness in the Research on Language Contact
Marcin Zabawa

Vybudování databází na základě slovníku jako korpus
Miloud Taïfi and Patrice Pognan


NLP, Corpus Linguistics,
Corpus Based Grammar Research
—
Počítačové spracovanie
prirodzeného jazyka, korpusová
lingvistika a gramatický výskum
Editors
Jana Levická
Radovan Garabík
Slovak National Corpus
Ľ. Štúr Institute of Linguistics
Slovak Academy of Sciences
Bratislava, Slovakia

Slovko (2001–2009)

Five Editions of the International Conference

2009

Slovakia cannot boast of many linguistic events that have been organised on regular basis and focusing on one specific field. This kind of symposiums is actually rather unique, in the present-day Slovak context only three of them are known nationally: onomastic conferences are held in different parts of Slovakia, Banská Bystrica hosts conferences on communication and annual Young Linguists' Symposium covers all linguistic disciplines as well as interdisciplinary areas (in 2010 its 20th edition will be held). Moreover, this seminar once saw the early presentations of some of the pioneers of Slovak computational and corpus linguistics (Emil Páleš) and also hosted the first participants of Slovko 2001 (Karol Furdík, Jozef Ivanecký). Although there has been only 5 events named Slovko, all of them of interdisciplinary nature dealing with areas of computational and corpus linguistics, the conference gained the international character and its tradition seems to be well rooted.

2001 was the year of organization of the first Slovko conference, which was held on October 26–27 (at that time still called Computer Processing of Slovak and Czech and attended exclusively by Czech and Slovak lecturers and audience). This symposium represented in the first place an event organised "with the aim to improve mutual awareness and knowledge of people in Slovakia involved in the issues of computers related to the language and vice versa: language related to computers" (A. Jarošová: Malá inventúra pred hľadaním spoločného jazyka. In: Slovenčina a čeština v počítačovom spracovaní. Bratislava: Veda 2001, p. 7)[1]. Quoting the author and the main organiser on behalf of the Ľudovít Štúr Institute of Linguistics, Slovak Academy of Sciences (the second organiser being Vladimír Benko on behalf of the Faculty of Pedagogy, Comenius University), contemporary Slovakia featured only fairly isolated islets of respective activities in different scientific disciplines and theoretically-applied contexts, out of which also efforts in the area of artificial intelligence and cognitive linguistics revealed to be relevant. However, neither results of the automatised processing of Slovak language data nor relevant language data in terms of electronic corpus of texts in Slovak language were available at that time. In comparison with the Czech Republic it was a diametrically different situation since computational linguistics there had been intentionally developed as an autonomous field for more than 30 years and the 100 million token National Corpus of the Czech language has been made available on the internet in 2000. Czech lecturers could therefore "offer the Slovak professional public as well as students of linguistic and non-linguistic disciplines rather comprehensive overview of the work results in the area of Czech computer processing; this language belongs in this respect among the European and in terms of different parameters also among world's elite" (ibid). The seminar met also another of intended aims: "carry on with the scientific, pedagogic and organisational work of Ján Horecký who had been doing his best since the beginning of the 1960s

1 Available from http://korpus.juls.savba.sk/structure/dicts.sk.html

the principles and methods of mathematical linguistics on the material of the Slovak language" (ibid) and "who witnessed the revival of the computational linguistics in the Institute of Linguistics when he initiated in 1988–1989 a project of data base of the Slovak language, within which an idea arose to start to build a corpus" (ibid, p. 9). 14 papers altogether were presented, 5 of which by Czech authors, and consequently published in the proceedings. These proceedings were – at the same time – partial result of the Ľ. Štúr Institute of Linguistics and Faculty of Pedagogy involvement in the multinational project Trans-European Language Resources Infrastructure II – TELRI II, carried out as a coordinated action within the European Commission programme INCO-COPERNICUS in 1999–2001.

The question whether the need of mutual informing of people involved in the field of computational linguistics in Slovakia and the exchange of experience with foreign specialists was only a one-time, temporal issue or whether it had a broader context, was answered in 2003 when Vladimír Benko (as a chief organiser on behalf of the Ľ. Štúr Institute of Linguistics and Faculty of Pedagogy) held an international scientific seminar on 24–25 October, this time already called Slovko, subtitled as Slavic Languages in Computer Processing. The event was attended by 54 people interested in this area, 20 out of which were from Slovakia and the rest from abroad. This edition of Slovko saw the first presentation of the Slovak National Corpus as an actually existing electronic

period 1955–2003 (in August 2003 it was made available for the public with its 26 million of tokens, in December 2003 the second version of the Corpus was released, containing 166 million of tokens) as well as the team of researchers of new department of the Ľ. Štúr Institute of Linguistics, created in mid-2002. Fledgeling corpus linguists lead by Mária Šimková presented first of all the whole of Corpus Project and then partial research solutions in segmentation, lemmatisation and morphological annotation of texts of the Slovak language. The first Slovak morphological tagset presented at Slovko was shortly afterwards publicly examined and reviewed by Slovak and foreign specialists and at the beginning of 2004 started a manual annotation of selected texts of the Slovak National Corpus.

Due to the fact that no Slovak university or college offered either an autonomous course of study or a specific seminar of computational and corpus linguistics, the Slovak National Corpus department as a chief investigator of the State Research Programme: Integrated Computational Processing of the Slovak Language for Linguistic Research Purposes organised regular lectures and seminars focused on these two disciplines. A part of it was published in the proceedings entitled Insight into the Slovak and Czech Corpus Linguistics. Ed. M. Šimková. Bratislava: Veda 2006. 208 p.[2] As a natural conclusion of these and also of some other activities was the organisation of Slovko on 10–12 November in 2005 and receiving the Slovak Academy of Sciences Prize for Building Infrastructure in Science on 11 November 2005. The third international seminar Slovko had already a programme and organizing committee and thanks to aroused interest it was also extended to computer treatment of Slavic and East European languages. It was attended by more than 60 interested researchers, which has represented the highest number of participants of one edition of Slovko. Altogether 29 papers were presented in more-or-less homogenous sessions: spoken corpora, speech analysis and synthesis, computer lexicography, parallel and historical corpora, terminology, e-learning, which were later on published in the proceedings Computer Treatment of Slavic and East European Languages (Ed. R. Garabík. Bratislava: Veda 2005. 246 p.).

Slovko in 2007 (explicitly held on October 25–27) was again organised by the Slovak National Corpus team (namely Radovan Garabík, Jana Levická, and Mária Šimková), this time as a solid fourth edition of a biannual international conference focused on NLP and computational lexicography and terminology. More than 50 participants had an opportunity to attend 37 lectures and presentations; the proceedings entitled Computer Treatment of Slavic and East European Languages 2007 (Ed. J. Levická, R. Garabík. Bratislava: Tribun 2007. 318 p.) were available upon arrival at the conference. Beside the topics that have been fundamental for every Slovko (corpus development including spoken corpora: data collection, annotation and processing), the foreground was reserved for theoret-

2 Available from http://korpus.juls.savba.sk/publications/sk.html

ical issues of computational lexicography and terminography, bilingual lexicography and terminography, co-occurrence analysis and pertinent collocations of lexicographical and terminographical relevance, new methods in data extraction and terminology mining from corpora, terminology databases and terminology management systems and so forth. Contributions of Slovak authors covering basic as well as applied research pointed to revolutionary changes that Slovak computational and corpus linguistics in particular had seen in the recent period. For the sake of the example, we can mention several results of the Slovak National Corpus team work: on January 1st 2007 began the second phase of the realisation of the State Research and Development Programme: Building the Slovak National Corpus and Integrated Computational Processing of the Slovak Language for the Linguistic Research Purposes, the 350 million token fully lemmatised and morphologically annotated general corpus was made available for public. Beside other products (database of digitised lexicographic and other linguistic works, parallel corpora), highest priority has been assigned to a partial project of the Slovak National Corpus, the Slovak terminology database, which was presented by its chief investigator Jana Levická, as well as the Slovak Spoken Corpus whose technical features were shown by (in cooperation with Milan Rusko) R. Garabík.

We decided to hold Slovko 2009 in the Congress Centre Smolenice of the Slovak Academy of Sciences to celebrate sort of a jubilee edition since it is the fifth in a row (all the previous editions took place in Bratislava but in different venues). Originally, the conference was to have been focused on parallel corpora, however, since the Institute of the Czech National Corpus, Charles University, Prague, organised a conference with the same topic on 17–19 September 2009 within their InterCorp Project, we sought to come up with a different key issue that could be of immediate interest in the Central-European context. Since the Institute of the Czech Language, Academy of Sciences of the Czech Republic ceased to organise regular conferences entitled Grammar & Corpora (hosted by František Štícha), the main conference topic was in fact right at hand: corpus linguistics, more precisely corpus based linguistic research. In this respect this year's Slovko has once again gathered more Slovak and Czech participants, which can be viewed also as a reminder of the beginnings of Slovko. Presented contributions point to other topics and, at the same time, give evidence of the progress in this field: majority of them do not deal with a development of standard corpora of written or spoken texts but rather with the tools of text analysis, building, and usage of language resources, linguistic components of information systems, computer-aided translation systems, localisation and lexicography, computer-aided language learning etc. Similarly as the organisers of the first Slovko participated in international projects, the Slovak National Corpus team and Slovak computational and corpus linguistics have

been involved in several international cooperation projects, especially in the 7th framework programme (EU FP7 INF 211983 Conceptual Modelling of Networking of Centres for High-Quality Research in Slavic Lexicography and Their Digital Resources).

Five editions of Slovko reflect the corpus and computational linguistics in Slovakia, Czech Republic, as well as in neighbouring and more distant countries, namely Austria, Belgium, Bulgaria, France, Croatia, Hungary, Germany, Norway, Netherlands, New Zealand, Poland, Russia, Slovenia, Spain, Ukraine have had their participants so far. Approximately 130 participants altogether from 17 countries have attended Slovko conference (however, some of them have participated repeatedly, on regular basis). Overall, Slovko conferences hosted almost 300 participants that could listen to and/or read nearly 150 papers. We would like to express our gratitude for excellent cooperation to all members of programme committees and abstract peer reviewers. We also thank to all participants for creating fruitful working and social atmosphere, to Czech colleagues in particular for they have been showing us support from the beginning and attended regularly all Slovko's editions. We appreciate very much the work of all members of each organising committee and hope that the flag will be handed on.

Mária Šimková
Jana Levická

Slovko (2001 - 2009)

Päť ročníkov medzinárodného podujatia

2009

V rámci lingvistických podujatí na Slovensku nemáme veľa takých, ktoré by sa uskutočňovali pravidelne a boli zamerané na jednu oblasť. Takéto podujatia sú naozaj skôr výnimočné, v súčasnom slovenskom kontexte sú širšie známe asi tri: na rôznych miestach Slovenska sa konávajú onomastické konferencie, v Banskej Bystrici konferencie o komunikácii a celú lingvistiku, ale aj interdisciplinárne oblasti zahŕňa každoročné Kolokvium mladých jazykovedcov (v r. 2010 sa už chystá 20. ročník), na ktorom sa ako mladí kedysi prezentovali aj jedni z priekopníkov slovenskej počítačovej a korpusovej lingvistiky (E. Páleš) a prví účastníci Slovka 2001 (K. Furdík, J. Ivanecký). Podujatí s názvom Slovko, ktoré sú zamerané interdisciplinárne na oblasti počítačovej a korpusovej lingvistiky, bolo zatiaľ len päť, no ukazuje sa, že sa založila dobrá tradícia s medzinárodným presahom.

Keď sa v r. 2001 pripravovalo a 26. – 27. októbra aj uskutočnilo prvé Slovko (ešte pod názvom Slovenčina a čeština v počítačovom spracovaní a za účasti výlučne českých a slovenských prednášateľov i poslucháčov), išlo predovšetkým o podujatie organizované „so zámerom zlepšiť vzájomnú informovanosť ľudí zaoberajúcich sa na Slovensku problematikou počítačov vo vzťahu k jazyku a naopak, jazyka vo vzťahu k počítačom“ (A. Jarošová: Malá inventúra pred hľadaním spoločného jazyka. In: Slovenčina a čeština v počítačovom spracovaní. Bratislava: Veda 2001, s. 7)[1]. Ako konštatovala autorka a hlavná organizátorka seminára za Jazykovedný ústav Ľ. Štúra SAV (druhým organizátorom bol V. Benko za Pedagogickú fakultu Univerzity Komenského), v tejto oblasti boli v tom čase na Slovensku značne izolované ostrovčeky aktivít prebiehajúcich v rôznych vedných odboroch a teoreticko-aplikačných kontextoch, z ktorých sa ako relevantné ukazovali aj aktivity v oblasti umelej inteligencie a kognitívnej lingvistiky. Neboli však k dispozícii výsledky automatizovaného spracovania slovenských jazykových dát ani relevantné jazykové dáta v podobe elektronického korpusu textov slovenského jazyka. Išlo o diametrálne odlišnú situáciu v porovnaní s Českou republikou, kde sa počítačová lingvistika cieľavedome budovala ako samostatný odbor viac ako tridsať rokov a národný korpus českého jazyka bol v r. 2000 zverejnený na internete v rozsahu 100 miliónov tokenov. Českí prednášatelia preto mohli „poskytnúť slovenskej odbornej verejnosti a študentom z lingvistických aj nelingvistických odborov ucelenejší pohľad na výsledky práce v oblasti počítačového spracovania češtiny, ktorá patrí v tomto smere medzi európsku a vo viacerých parametroch aj medzi svetovú špičku“ (tamže). Seminár splnil aj ďalší zo svojich cieľov: „nadviazať na vedeckú, pedagogickú a organizačnú prácu Jána Horeckého, ktorý sa od začiatku 60. rokov 20. storočia usiloval uplatňovať princípy a metódy matematickej lingvistiky na materiáli slovenského jazyka“ (tamže) a „ktorý stál aj pri revitalizácii počítačovej lingvistiky v Jazykovednom ústave, keď v rokoch 1988 – 1989 pripravil projekt bázy dát slovenského jazyka, v rámci ktorej sa začalo uvažovať aj o budovaní korpusu“ (tamže, s. 9). Na

1 Dostupný aj z http://korpus.juls.savba.sk/structure/dicts.sk.html

seminári odznelo a v zborníku z neho bolo publikovaných 14 príspevkov, z toho 5 od českých autorov. Publikovaný zborník zároveň predstavoval jeden z výsledkov účasti JÚĽŠ SAV a PdF UK v mnohonárodnom projekte Trans-European Language Resources Infrastructure II – TELRI II, ktorý sa ako súbor coordinated action uskutočnil v rámci programu Európskej komisie INCO-COPERNICUS v r. 1999 – 2001.

Otázka, či potreba vzájomnej informovanosti pracovníkov v oblasti počítačovej lingvistiky na Slovensku a výmena skúseností s odborníkmi zo zahraničia bola iba jednorazová, dočasná, alebo má širší kontext, bola zodpovedaná v r. 2003, keď V. Benko (ako hlavný organizátor za JÚĽŠ SAV i PdF UK) uskutočnil 24. – 25. októbra medzinárodný odborný seminár už pod názvom Slovko, v podtitule špecifikovaný na Slovanské jazyky v počítačovom spracovaní. Podujatia sa zúčastnilo 54 záujemcov o túto oblasť, z toho 20 zo Slovenska a 34 zo zahraničia. Na tomto Slovku sa prvýkrát prezentoval Slovenský národný korpus – ako už reálne existujúca elektronická databáza písaných textov z r. 1955 – 2003 (od augusta 2003 verejne prístupná v rozsahu 26 miliónov tokenov, v decembri 2003 bola zverejnená druhá verzia v rozsahu 166 miliónov tokenov) i ako kolektív pracovníkov nového oddelenia JÚĽŠ SAV, založeného v podstate v polovici r. 2002. Začínajúci korpusoví lingvisti pod vedením M. Šimkovej prezentovali jednak celý korpusový projekt, jednak čiastkové riešenia pri segmentácii, lematizácii a morfologickej anotácii textov slovenského jazyka. Prvá podoba slovenského morfologického tagsetu predložená na Slovku bola zakrátko osobitne verejne oponovaná domácimi aj zahraničnými odborníkmi a začiatkom r. 2004 sa rozbehla ručná morfologická anotácia vybraných textov SNK.

Vzhľadom na to, že na žiadnej vysokej škole na Slovensku neexistoval samostatný odbor a ani sa nevyučo-

val samostatný predmet počítačovej a korpusovej lingvistiky, oddelenie SNK JÚĽŠ SAV ako riešiteľ štátnej úlohy Integrated Computational Processing of the Slovak Language for Linguistic Research Purposes organizovalo pravidelné prednášky a semináre z tejto oblasti. Časť z nich bola publikovaná v zborníku Insight into the Slovak and Czech Corpus Linguistics. Ed. M. Šimková. Bratislava: Veda 2006. 208 s.[2] Prirodzeným vyústením týchto a ďalších aktivít bolo zorganizovanie Slovka 10. – 12. novembra 2005 a prevzatie Ceny SAV za budovanie infraštruktúry pre vedu 11. novembra 2005. Tretí medzinárodný seminár Slovko už bol obohatený o vedecký výbor a organizačný výbor a vzhľadom na prejavený záujem bol rozšírený na Slovanské a východoeurópske jazyky v počítačovom spracovaní. Zúčastnilo sa ho vyše 60 záujemcov, čo bol a stále je najvyšší počet účastníkov jedného Slovka. Celkovo odznelo 29 príspevkov vo viac-menej homogénnych blokoch: hovorené korpusy, analýza a syntéza reči, počítačová lexikografia, paralelné korpusy, historické korpusy, terminológia, e-learning a pod. Tieto vyšli v zborníku Computer Treatment of Slavic and East European Languages (Ed. R. Garabík. Bratislava: Veda 2005. 246 s.).

Slovko v r. 2007 (25. – 27. októbra) organizoval kolektív SNK (najmä R. Garabík, J. Levická, M. Šimková) už ako stabilizovanú 4. bienálnu medzinárodnú konferenciu so zameraním na NLP a počítačovú lexikografiu a terminológiu. Vyše 50 účastníkov si malo možnosť vypočuť 37 príspevkov a prezentácií, zborník Computer Treatment of Slavic and East European Languages (Ed. J. Levická, R. Garabík. Bratislava: Tribun 2007. 318 s.) mali všetci k dispozícii priamo na podujatí. Okrem tém, ktoré sú ako základné súčasťou každého Slovka (tvorba korpusov vrátane hovorených: zber dát, anotácia a spracovanie), sa do popredia dostali teoretické otázky komputačnej lexikografie a terminografie, bilingválna lexikografia

2 Dostupný aj z http://korpus.juls.savba.sk/publications/

a terminografia, kookurenčná analýza a lexikograficky alebo terminograficky relevantné kolokácie, nové metódy v extrahovaní dát a získavanie terminológie z korpusov, terminologické databázy a systémy terminologického manažmentu a pod. Príspevky slovenských autorov, pokrývajúce základný aj aplikovaný výskum, poukázali na prevratné zmeny, ktoré slovenská počítačová a najmä korpusová lingvistika zaznamenala za posledné obdobie. Napr. z produkcie SNK, ktoré 1. januára 2007 začalo 2. etapu riešenia štátnej úlohy Budovanie Slovenského národného korpusu a elektronizácia jazykovedného výskumu na Slovensku, bol verejnosti k dispozícii hlavný korpus v rozsahu 350 miliónov tokenov s plnou lematizáciou a morfologickou anotáciou na báze vlastného tagsetu. Okrem ďalších produktov (databáza elektronických lexikografických a iných lingvistických diel, paralelné korpusy) sa prioritným čiastkovým projektom SNK stala Slovenská terminologická databáza, ktorú na konferencii predstavila jej hlavná riešiteľka J. Levická, a budovať sa začal aj Slovenský hovorený korpus, ktorého technické parametre prezentoval (v spolupráci s M. Ruskom) R. Garabík.

Slovko 2009 sme ako polojubilejné, piate v poradí umiestnili do Kongresového centra SAV v Smoleniciach (všetky predchádzajúce boli v Bratislave, a to na viacerých miestach). Hlavným zameraním mali byť pôvodne paralelné korpusy, ale keďže Ústav Českého národného korpusu FF UK v Prahe organizoval takto profilovanú konferenciu v súlade s cieľmi svojho projektu InterCorp 17. – 19. septembra 2009, hľadali sme inú hlavnú tému, ktorá by mohla byť aktuálnou v stredoeurópskom priestore. A keďže Ústav pro jazyk český AV ČR prestal organizovať pravidelné konferencie Grammar & Corpora (pod vedením F. Štíchu), priam sa nám ponúkla ako hlavná téma korpusová lingvistika, resp. korpusovo zamerané lingvistické výskumy. V tejto súvislosti je na tohtoročnom Slovku opäť viac slovenských a českých účastníkov, čo je aj

istá pripomienka východiskovej situácie Slovka 2001. Posun v tejto oblasti nielen v našich krajinách naznačujú aj ďalšie prezentované témy: už to nie sú otázky budovania štandardného korpusu písaných či hovorených textov, ale viac nástroje textovej analýzy, tvorba a využitie jazykových zdrojov, lingvistické zložky informačných systémov, preklad s počítačovou podporou, lokalizácia a lexikografia, didaktika vyučovania cudzích jazykov s počítačovou podporou a pod. A tak ako organizátori prvého Slovka boli členmi medzinárodných projektov, aj Slovenský národný korpus a slovenská počítačová a korpusová lingvistika sa v súčasnosti prezentuje vo viacerých medzinárodných spoluprácach, predovšetkým v 7. RP (EU FP7 INF 211983 s názvom Conceptual Modelling of Networking of Centres for HighQuality Research in Slavic Lexicography and Their Digital Resources).

Päť ročníkov Slovka je odrazom vývinu počítačovej a korpusovej lingvistiky na Slovensku, v Čechách, v okolitých i vzdialenejších krajinách, z ktorých sú doteraz zastúpené Belgicko, Bulharsko, Francúzsko, Gruzínsko, Holandsko, Chorvátsko, Maďarsko, Nemecko, Nórsko, Nový Zéland, Poľsko, Rakúsko, Rusko, Slovinsko, Španielsko, Ukrajina. Zahraničných účastníkov bolo na všetkých konferenciách Slovko spolu asi 130 (nie sú to unikátne výskyty, viacerí sú našimi pravidelnými hosťami) zo 17 krajín. Spolu so slovenskými účastníkmi sme privítali na týchto podujatiach takmer 300 záujemcov, ktorí si mali možnosť vypočuť a/alebo prečítať takmer 150 príspevkov. Za vynikajúcu spoluprácu ďakujeme všetkým členom vedeckého výboru a posudzovateľom abstraktov. Ďakujeme všetkým účastníkom za vytváranie tvorivej pracovnej i spoločenskej atmosféry, osobitne českým kolegom, ktorí nás od začiatku podporovali a pravidelne sa zúčastňovali aj Slovka. Ďakujeme všetkým členom doterajších organizačných výborov a dúfame, že štafeta pôjde ďalej.

Mária Šimková
Jana Levická

Slovko

Table of Contents

2011

Czech-Slovak Parallel Corpora for MT between Closely Related Languages
Petra Galuščáková and Ondřej Bojar

Slovene Lexical Database
Polona Gantar and Simon Krek

Building Annotated Corpora without Experts
Marek Grác

The Lexicographic Representation of Czech Diatheses: Rule Based Approach
Václava Kettnerová and Markéta Lopatková

Translating prepositions from Czech into Russian: challenges for the Machine Translation
Natalia Klyueva and Naděžda Runštuková

Extracting Verbs with PP/NP Variation from the Large 3-gram Corpus
Mikhail Kopotev, Natalia Kochetkova and Eduard Klyshinsky

On the Question of Homonymy and Polysemy in the Lexicographical Practice of the Russian Language Semantic System in its Development Modeling
Irina Nekipelova

Historical and Etymological Electronic Dictionary System Modelling
Irina Nekipelova and Elvira Zarifullina

A Web-based Morphological Tagger for Bulgarian
Aleksandar Savkov, Laska Laskova, Petya Osenova, Kiril Simov and Stanislava Kancheva

Automatic Extraction of Multiword Expressions using Linguistic Constraints for Slovak LVCSR
Ján Staš, Daniel Hládek, Marián Trnka and Jozef Juhár

Common Formal Framework for Multilingual Representation of Inflectional Morphology for Two Related Slavonic Languages
Velislava Stoykova

Recent Developments in ParaSol: Breadth for Depth and XSLT based web concordancing with CWB
Ruprecht von Waldenfels

The Instrumental Environment for the Automatic Syntactical Analysis of Ukrainian
Iryna Zamaruieva and Olga Shypnivska

Slovko

Foreword

2011

After two years since the last International Conference under the umbrella name Slovko, we now meet at the 6th edition of the event which is primarily focused on Natural Language Processing (NLP) and multilingualism. According to the proposals submitted, authors of the contributions focused mainly on the issues of parallel corpora and automatic translation, corpus based grammar research and occasionally other discussed issues. There is an absence of lexicographers and terminologists, much like the e-learning subject, which seems to be less prevalent in the academic and university scene whose representatives regularly take part in the Slovko conference.

While organizing previous editions, the main theme and event venue in particular were at the centre of attention. The focus of the following edition is already taking its shape: semantic networks and multilingual dictionaries. However, under different circumstances, other questions have came up which have led to certain organizational changes. This will be reflected in the form of the published contributions in a consistent way as to formally address the adjusted requirements for publications evaluating.

The 6th conference edition will take place in a small city called Modra famous for its fine wine and blue-and-white porcelain as well as for the renowned linguist Ľudovít Štúr, author of the Slovak orthography rules and whose name our institute also bears. One of his visions was permanent Slavic collaboration and six of the Slavic nations will be present at the conference (Bulgaria, Czech Republic, Slovakia, Slovenia, Russia, and Ukraine). Naturally, our events are open to all languages (Štúr himself could speak several Slavic languages plus Latin, German, Hungarian, French, Greek, and was studying Hebrew and English) and at the conference, more languages used to be present. Each edition is different and each time it brings new preferred issues, presenters and discussants. We are able to meet scientists from other countries in various forums while dealing with projects on NLP, Language Resources, and Corpus Linguistics.

To all the participants of the 6th Conference edition, we hope you have a pleasant stay in Modra and enjoy fruitful discussions and mutual inspirations that will continue beyond this meeting and become a part of prospective cooperation.

Mária Šimková
Translated by Adriána Žáková

Slovko

Úvod

2011

Po dvoch rokoch od posledného stretnutia na medzinárodnej konferencii so zastrešujúcim názvom Slovko sa stretávame na 6. ročníku tohto podujatia, ktoré je primárne zamerané na počítačové spracovanie prirodzeného jazyka a mnohojazyčnosť. Záujem autorov príspevkov sa podľa návrhov predložených v propozíciách sústredil najmä na paralelné korpusy a otázky automatizovaného prekladu, významne sú zastúpené gramaticky orientované korpusové výskumy, ojedinele i ďalšie z ponúknutých tém. Citeľná je neprítomnosť lexikografov a terminológov, bez zastúpenia ostala aj téma e-learningu, zrejme všeobecne málo rozšírená na prevažne akademických a univerzitných pracoviskách, ktorých zástupcovia sa pravidelne zúčastňujú konferencie Slovko.

Pri organizovaní doterajších ročníkov sme si kládli predovšetkým otázku hlavnej témy a miesta konania podujatia. Zameranie nasledujúceho ročníka sa nám už črtá: sémantické siete a multilingválne slovníky. Súčasne sa však tentoraz, pod vplyvom rôznych okolností, vynorili ďalšie otázky, ktoré nás nabádajú k istým organizačným zmenám. Najviac sa to prejaví na spôsobe publikovania príspevkov tak, aby to zodpovedalo formálne nastaveným kritériám na hodnotenie publikačných výstupov.

6. ročník konferencie Slovko sme umiestnili do malého mestečka Modra, ktoré je okrem dobrého vína a kvalitnej keramiky známe aj pôsobením kodifikátora spisovnej slovenčiny Ľudovíta Štúra, ktorého meno nesie naše kmeňové pracovisko. Jednou z jeho vízií bola spolupráca slovanských národov, z ktorých šesť je zastúpených aj na tomto stretnutí (Bulharsko, Česká republika, Slovensko, Slovinsko, Rusko, Ukrajina). Naše podujatia sú, samozrejme, otvorené voči všetkým jazykom (aj sám Štúr ovládal okrem viacerých slovanských jazykov ešte latinčinu, nemčinu, maďarčinu, francúzštinu, gréčtinu, učil sa hebrejčinu a angličtinu) a na Slovku ich priebežne bývalo zastúpených viac, ale každý ročník je iný, prináša vždy aktuálne nové preferencie tém a nové zloženie prezentujúcich a diskutujúcich. S predstaviteľmi ďalších krajín sa zas stretávame na iných fórach a pri riešení projektov v oblasti počítačového spracovania jazyka, jazykových zdrojov a korpusovej lingvistiky.

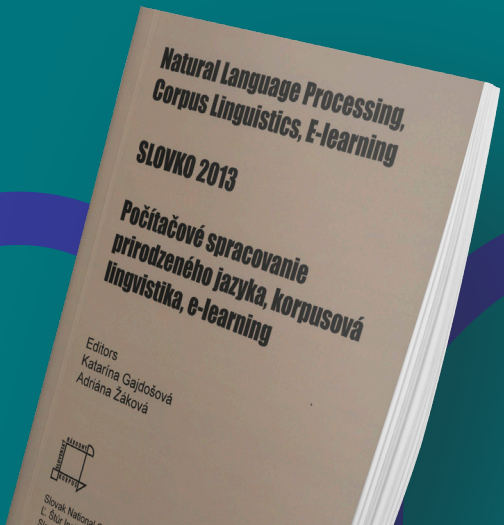
Účastníkom 6. ročníka konferencie Slovko želáme príjemný pobyt v Modre, plodné rokovania a vzájomné inšpirácie, ktoré prekročia rámec tohto stretnutia a stanú sa súčasťou perspektívnej spolupráce.

Mária Šimková

Slovko

Table of Contents

2013



The Corpus CzeSL in the Service of Teaching Czech for Foreigners – Errors in the Use of the Pronoun *který*
Andrea Hudousková

Delimitation of Participles in the Manual Morphological Annotation
Agáta Karčová

Corpus Based Identification of Czech Light Verbs
Václava Kettnerová, Markéta Lopatková, Eduard Bejček, Anna Vernerová and Marie Podobová

Agents Expressed by Prepositionless Instrumental Modifying Czech Nouns Derived from Intransitive Verbs
Veronika Kolářová

Corpus-based Online Word Formation Exercises for Advanced Learners of English – Challenges and Solutions
Grzegorz Krynicki

Experimenting with Slovak Wikipedia as a Source for Language Technologies
Michal Laclavík, Štefan Dlugolinský and Michal Blanárik

Query Interface for Diverse Corpus Types
Tomáš Machálek and Michal Křen

The Effect of Stop Words elimination on Sequence Patterns Extraction in Comparable Corpora
Dáša Munková, Michal Munk and Martin Vozár

Valency of Selected Primary Adjectives in the SYN2010 Corpus
Kateřina Najbrtová

Event Extractor: Email Events Detection and Calendar Integration
Filip Ogurčák and Michal Laclavík

Formal (Morpho)Syntax Properties of Reflexive Particles *se, si* as Free Morphemes in Contemporary Czech
Vladimír Petkevič

Introduction to Online Learning
Katarína Pišútová

Automatic Extraction of Multiword Units from Slovak Text Corpora
Ján Staš, Daniel Hládek, Jozef Juhár and Martin Ološtiak

Verb Valency and Argument Non-correspondence in a Bilingual Treebank
Jana Šindlerová, Zdeňka Urešová and Eva Fučíková

Determination of Czech BCT Prototypes on the Basis of Corpus Data
Tatiana Timoshchenko

Veni, Vidi, Vici: The Language Technology Infrastructure Landscape after CESAR
Tamás Váradi

Czech Subjectivity Lexicon: A Lexical Resource for Czech Polarity Classification
Kateřina Veselovská

A Corpus-based Analysis of the Functionality and the Meaning of Infinitive "Frustrative Construction" in Czech and Slovak
Uliana Yazhinova

Slovko

Foreword

Slovko 2013 – Natural Language Processing, Corpus Linguistics, E-learning will be again held in Bratislava. The organizers – Slovak National Corpus Department of Ľ. Štúr Institute of Linguistics, Slovak Academy of Sciences are honoured to host participant from eight countries: Bulgaria, Czech Republic, Germany, Greece, Hungary, Poland, Slovakia and Slovenia.

Over three days participants will be able to benefit from 29 presentations, including 3 plenary talks. Unfortunately, one third of submitted papers on given topics has not been recommended by the Programme Committee members. We thank to all reviewers for their constructive suggestions and their help to make the conference even more successful.

The 7th edition of the biennial conference increased the presence of the linguistically-oriented (corpus-based and corpus-driven) studies. The more technically oriented papers provide information on effectiveness of the approaches applied, experimenting and innovative methods. Latest trends and tendencies in enhancing the corpus data can be found also in the papers written by Slovak authors.

We wish all participants of the conference Slovko 2013 profitable time and positive inspiration for further cooperation in the field of natural language processing, corpus linguistics and similar research.

Mária Šimková
Translated by Adriána Žáková

2013

Úvod

Slovko 2013 – počítačové spracovanie prirodzeného jazyka, korpusová lingvistika, e-learning sa koná opäť v Bratislave a organizátori zo Slovenského národného korpusu Jazykovedného ústavu Ľudovíta Štúra Slovenskej akadémie vied vítajú na tomto podujatí účastníkov z ôsmich krajín: Bulharsko, Česká republika, Grécko, Maďarsko, Nemecko, Poľsko, Slovensko a Slovinsko.

Počas troch dní rokovania odznie celkovo 29 príspevkov, z toho 3 plenárne prednášky. Záujemcov o prezentáciu a publikovanie výsledkov svojej práce vo vymedzených tematických okruhoch bolo podstatne viac, no posudzovatelia z vedeckého výboru neodporúčali tretinu prihlásených príspevkov. Všetkým recenzentom ďakujeme za vykonanú prácu a za zvyšovanie kvality celého podujatia.

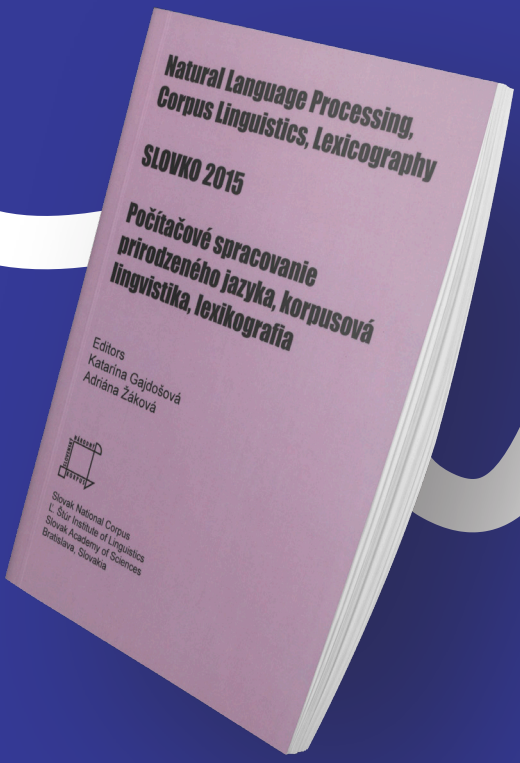
Na 7. ročníku našej bienálnej konferencie sa zvýšil podiel lingvisticky zameraných štúdií (corpus-based alebo corpus-driven), ktorých je takmer polovica. V technicky orientovaných príspevkoch prevládajú informácie o efektivite uplatňovaných postupov, experimentovanie a hľadanie inovatívnych metód. Nové oblasti výskumu a snahy o skvalitnenie korpusových dát a výstupov sa objavujú aj v príspevkoch slovenských autorov.

Všetkým účastníkom konferencie Slovko 2013 želáme užitočne strávený konferenčný čas a pozitívne prínosy z rokovaní pre ďalšie projekty v oblasti počítačového spracovania prirodzeného jazyka, korpusovej lingvistiky a súvisiacich výskumov.

Mária Šimková

Slovko

Table of Contents



2015

Slovko

Foreword

2015

Slovko 2015 – this year's edition entitled ***NLP, Corpus Linguistics, Lexicography*** – represents a follow up of previous autumn meetings in Bratislava. Organisers, both from the Slovak National Corpus of the Ľ. Štúr Institute of Linguistics, Slovak Academy of Sciences, and from the Slovak Centre of Scientific and Technical Information, are honoured to welcome participants from five countries: Austria, Czech Republic, France, Slovakia and Slovenia.

Two conference days offer 18 presentations, including two plenary talks. Not all papers registered for presentation were also published – current programme comprises also two presentations that cannot be found in the proceedings. Members of the programme committee carefully reviewed every paper sent with the registration (two reviewers for each text) and thus contributed to the overall quality of the scientific event and of this publication, for which we would like to express our sincere gratitude.

The 8th edition of the biannual conference Slovko 2015 experiences the increase of papers dealing with corpus linguistics including lexicography. On the other hand, computationally oriented papers are in a minority. There is a significant shift from presenting new written corpora and their analyses to the issues concerning the building and research of spoken, even dialect corpora. We believe that this focus of papers will also become a source of inspiration both for conference participants and readers of the proceedings in their further work in the area of NLP, corpus linguistics and related research in Slovakia and neighbouring countries.

We wish all the participants of Slovko 2015 an enjoyable stay in the Slovak Centre of Scientific and Technical Information and in Bratislava in particular to those who came from abroad. We would also like to invite you to Slovko 2017 that will be focusing, besides NLP and corpus linguistics, on computational terminology and terminography.

Mária Šimková
Translated by Jana Levická

Slovko

Úvod

2015

Slovko 2015 – tentoraz s podtitulom ***počítačové spracovanie prirodzeného jazyka, korpusová lingvistika, lexikografia*** – nadväzuje na predchádzajúce jesenné stretnutia v Bratislave. Organizátori zo Slovenského národného korpusu Jazykovedného ústavu Ľudovíta Štúra Slovenskej akadémie vied a Centra vedecko-technických informácií Slovenskej republiky vítajú na tomto podujatí účastníkov z piatich krajín: Česká republika, Francúzsko, Rakúsko, Slovensko a Slovinsko.

Počas dvoch dní rokovania odznie celkovo 18 príspevkov, z toho dve plenárne prednášky. Nie všetky príspevky prihlásené na prezentáciu sú aj publikované – v aktuálnom programe odznejú aj tri príspevky, ktoré sa v tomto zborníku nenachádzajú. Členovia vedeckého výboru starostlivo posúdili každý prihlásený článok (jeden text dvaja posudzovatelia) a prispeli tak ku kvalite celého podujatia a tejto publikácie, za čo im patrí naše poďakovanie.

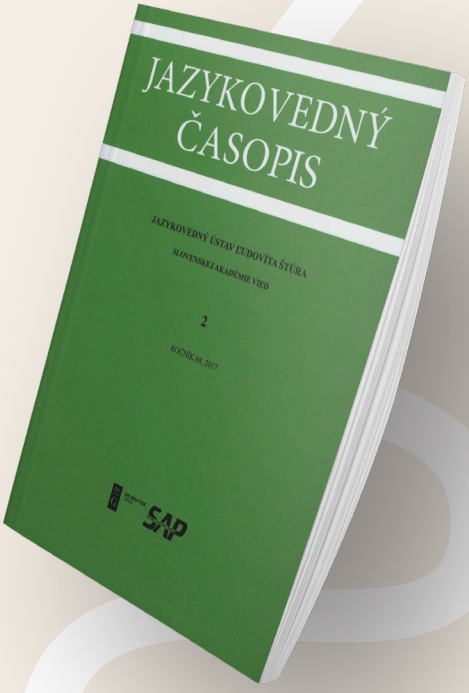
Na 8. ročníku bienálnej konferencie Slovko 2015 sa zvýšil počet príspevkov z oblasti korpusovej lingvistiky vrátane lexikografie. Technicky orientovaných príspevkov je tentoraz menej. Výrazný posun nastal od prezentácií budovania a analýz písaných korpusov smerom k riešeniu otázok tvorby a výskumov hovorených, dokonca aj nárečových korpusov. Veríme, že aj takto zamerané príspevky budú pre poslucháčov a diskutujúcich na konferencii, ako aj pre čitateľov publikovaných štúdií prínosom a inšpiráciou do ďalších prác v oblasti počítačového spracovania prirodzeného jazyka, korpusovej lingvistiky a súvisiacich výskumov na Slovensku i v okolitých krajinách.

Všetkým účastníkom konferencie Slovko 2015 želáme príjemný pobyt v Centre vedecko-technických informácií, mimobratislavským účastníkom aj v Bratislave vôbec. Zároveň vás už teraz pozývame na Slovko 2017, ktoré bude okrem počítačového spracovania prirodzeného jazyka a korpusovej lingvistiky primárne venované počítačovej terminológii a terminografii.

Mária Šimková

Slovko

Content



2017

Slovko

Slovo na úvod

2017

9. ročník bienálnej medzinárodnej konferencie **Slovko** sa koná v Bratislave 25. – 27. októbra 2017. Okrem tradičného zamerania na počítačové spracovanie prirodzeného jazyka (v písanej i hovorenej podobe) a na korpusovú lingvistiku sa osobitne venuje terminológii a e-terminológii. Tento tretí tematický okruh je v každom ročníku konferencie iný (http://korpus.sk/slovko.html) a tentoraz súvisí s riešením projektu *Analýza terminologickej práce Jána Horeckého ako inšpirácia pre terminologický manažment 21. storočia na Slovensku* (projekt VEGA 2/0114/15, zodpovedná riešiteľka Jana Levická).

Organizátori zo Slovenského národného korpusu Jazykovedného ústavu Ľ. Štúra SAV dostali na prvú výzvu vyše 90 prihlášok záujemcov o prezentáciu svojich najnovších výsledkov v uvedených oblastiach. Do stanoveného termínu prišlo 44 príspevkov, ktoré starostlivo posúdili členovia vedeckého výboru a na publikovanie vybrali 31 z nich. Recenzentom aj na tomto mieste ďakujeme za vyjadrenia, ktorými prispeli k zlepšeniu kvality publikovaných príspevkov a celého priebehu konferencie.

Osobitné poďakovanie patrí redakčnej rade a hlavnej redaktorke Jazykovedného časopisu za poskytnutie publikačného priestoru vedecko-výskumným témam, ktoré v slovenskej lingvistike nemajú dlhú tradíciu. Postupy a výsledky korpusovo a počítačovolingvisticky orientovaných výskumov v oblasti gramatiky, lexikológie, terminológie a analýzy hovorenej reči môžu byť obohatením existujúcich metodologických postupov pri poznávaní a opise jazyka a rozšírením doterajších poznatkov.

Počas rokovania odznejú 2 plenárne prednášky a 32 prezentácií od 69 autorov – 18 príspevkov je v individuálnom autorstve, 16 v spoluautorstve dvoch, troch i viacerých autorov. Spomedzi 11 zúčastnených krajín je najviac autorov z ČR (32), SR (14) a Ruska (8), ďalšie krajiny sú zastúpené menším počtom (Gruzínsko, Chorvátsko, Fínsko, Rakúsko, Ukrajina, Nemecko, Poľsko, Švédsko). Všetkým účastníkom konferencie Slovko 2017 želáme úspešný priebeh rokovania, vzájomne užitočné a obohacujúce diskusie, ako aj nadviazanie nových kontaktov na prípadné budúce spolupráce.

Nasledujúce Slovko v roku 2019 bude jubilejné 10. a plánujeme ho v samostatnom tematickom okruhu špecifickejšie zamerať na témy, ktoré sa objavujú už v tomto ročníku: dynamika jazyka a gramatické zmeny v súčasnom jazyku na báze korpusových dát.

Mária Šimková

Slovko

Foreword

2017

The 9th edition of the biannual conference **Slovko** is held in Bratislava from 25–27 October 2017. In addition to the traditional focus on natural language processing (in its written and spoken form) and on corpus linguistics it is specifically devoted to terminology and e-terminology. The third thematic area being different in each edition (http://korpus.sk/slovko.html) was chosen in accordance with the organizers' project *Analysis of terminology work of Ján Horecký as an inspiration for the terminology management of the 21st century in Slovakia* (VEGA project no 2/0114/15, coordinator Jana Levická).

Organizers from the Slovak National Corpus Department of the Ľ. Štúr Institute of Linguistics of the Slovak Academy of Sciences received upon the first call more than 90 registrations. Within the set deadline they received 44 articles presenting the latest results in the above-mentioned areas which were carefully reviewed by the programme committee members who recommended as many as 31 papers for publishing in the proceedings. We would like to express our gratitude to all reviewers who helped to improve the quality of the published papers and the conference itself.

Special thanks goes to the Editorial Board and Chief Editor of the Journal of Linguistics for providing the possibility to publish the articles on scientific research topics without a long tradition in the Slovak linguistics. Approaches and findings of the corpus and computer-oriented researches in the field of grammar, lexicology, terminology and speech analysis can contribute to the existing methodological approaches aiming at better understanding and description of the language, as well as at extending the current knowledge.

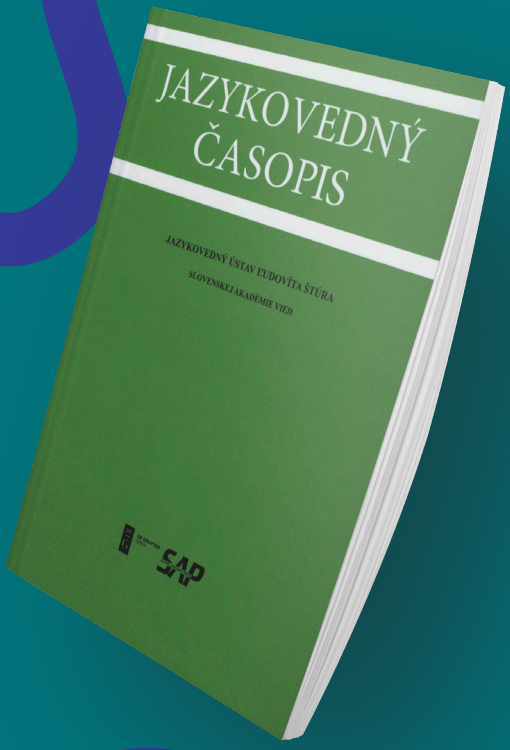
The event includes 2 plenary talks and 32 presentations by 69 authors – out of which 18 were prepared by a single author, while 16 presentations resulted from the co-authorship. from among 11 participating countries, the vast majority of authors come from the Czech Republic (32), followed by the Slovak Republic (14) and Russia (8), other countries are represented by fewer authors (Georgia, Croatia, Finland, Austria, Ukraine, Germany, Poland, Sweden). We wish all the participants of Slovko 2017 a successful conference, useful and enriching discussions as well as numerous opportunities for networking leading to future cooperation.

The following, 10th jubilee edition of Slovko – planned to be held in 2019 – will be focusing, in its third thematic area, on topics that appear already in this edition: dynamics of language and grammar changes in the contemporary language based on corpus data.

Mária Šimková
Translated by Adriána Žáková

Slovko

Contents



2019

Corpus-based/driven linguistic and DH research

Colloquial Lexemes in Journalistic Texts
Lucia Jasinská

Frequency in Corpora as a Signal of Lexicalization (On the Absolute Usage of Comparative and Superlative Adjectives)
Pavla Kochová

On the Valency of Various Types of Adverbs and Its Lexicographic Description
Jakub Sláma and Barbora Štepánková

The Synchronic Dynamics of Words Ending *in -ita/-osť*
Mária Šimková

Slovak Comparative Correlatives: New Insights
Jakob Horsch

Analysis of Verbal Prepositional "of" Structures
Marianna Hudcovičová

Temporal 'since' in Slovak: Conjunction(s) and Aspect Choice – a Corpus Study
Paula Kyselica and René M. Genis

In Which Clause Do Subordinate Conjunctions Prosodically Belong?
Zuzana Komrsková and Petra Poukarová

Russian Indefinite Pronoun *kakoi-libo*: Non-standard Usage and Changes in the Semantics
Yulia Kuvshinskaya

Ways of Automatic Identification of Words Belonging to Semantic Fields
Victor Zakharov

Analysis of the Lemma *mateřství* (Motherhood)
Zuzana Černá and Radek Čech

Corpus-supported Semantic Studies: Part/Whole Expressions in Russian
Igor Boguslavsky and Leonid Iomdin

Wackernagel's Position and Contact Position of Pronominal Enclitics in Older Czech. Competition or Cooperation?
Radek Čech, Pavel Kosek, Olga Navrátilová and Ján Mačutek

Frequency Dictionary of 16th Century Cyrillic Written Monument
Oksana Nika and Svitlana Hrytsyna

Kinship Terminology in Western Slavic Languages Based on Copora Analysis
Jana Kobzová

Gender-specific Adjectives in Czech Newspapers and Magazines
Adrian Jan Zasina

Corpus building

From the National Corpus of Polish to the Polish Corpus Infrastructure
Maciej Ogrodniczuk, Rafał L. Górski, Marek Łaziński and Piotr Pęzik

Relevant Criteria for Selection of Spoken Data: Theory Meets Practice
Marie Kopřivová, Zuzana Komrsková, Petra Poukarová and David Lukeš

The DIALEKT Corpus and Its Possibilities
Hana Goláňová and Martina Waclawičová

Annotation in the Corpus of Texts of Students Learning Slovak as a Foreign Language (ERRKORP)
Michaela Mošaťová and Katarína Gajdošová

Microsyntactic Parts of Speech in NovaMorf, a New Morphological Annotation of Czech
Vladimír Petkevič, Jaroslava Hlaváčová, Klára Osolsobě, Martin Svášek and Josef Šimandl

Improving Nominalized Adjectives Tagging
Klára Osolsobě and Hana Žižková

Modifications of the Czech Morphological Dictionary for Consistent Corpus Annotation
Jaroslava Hlaváčová, Marie Mikulová, Barbora Štěpánková and Jan Hajič

Levels of Annotation in the Slovenian Training Corpus ssj500k 2.2
Mija Bon and Polona Gantar

Creation and use of language resources

Meaning and Semantic Roles in CzEngClass lexicon
Zdeňka Urešová, Eva Fučíková, Eva Hajičová and Jan Hajič

Introducing Semantic Labels into the DeriNet Network
Magda Ševčíková and Lukáš Kyjánek

Non-systemic Valency Behavior of Czech Deverbal Nouns Based on the NomVallex Lexicon
Veronika Kolářová, Anna Vernerová and Jonathan Verner

Towards Reciprocal Deverbal Nouns in Czech: From Reciprocal Verbs to Reciprocal Nouns
Václava Kettnerová and Markéta Lopatková

Processing of Derivational Features for (Semi)Automatic Creation of Dictionary Definitions in the User Interface (CZEDD) for Learning Czech as a Second Language: Suffix *-tel* and *-ista*
Erik Citterberg and Adriana Válková

Conception and Development of an Open Database System on Historical Multilingualism in Austria
Katharina Prochazka, Ludwig Maximilian Breuer and Agnes Kim

Natural language prosessing

On Possibilities and Methods of Analysis of Thematic Expressions in Spoken Text
Petr Pořízka

Identification of Spontaneous Spoken Texts in Slovak
Róbert Sabo, Peter Krammer, Ján Mojžiš and Marcel Kvassay

Affordable Annotation of the Mobile App Reviews
Marek Grác, Markéta Masopustová and Marie Valíčková

Slovko

Foreword

"There is one part of corpus linguistics that becomes more and more significant, namely, the one that is not primarily concerned only with corpus building or the issues of representativeness, possibilities of lemmatization and morphological or other types of tagging. It comprises methods of corpus mining, i.e., methods broadening language descriptions and enhancing their quality, revealing new relations and inventing new concepts for description of language reality."

V. Cvrček and T. Kováříková

2019

Reading through the papers of anniversary SLOVKO 2019 proceedings, the 2011 quote of V. Cvrček and T. Kovaříková from the journal *Naše řeč* still proves valid. The observation, however, does not diminish the relevance of neither part of corpus-aided ways of language research. Our ambition with this years' 10th anniversary international linguistic conference SLOVKO was therefore to structure the proceedings so that they would reflect the rich thematic and methodical diversity of contemporary corpus linguistics and natural language processing.

The first group out of four covers the "corpus mining" research. Corpus-based and -driven linguistic and DH-related papers use corpora as a well-established resource for deeper understanding of language phenomena, e.g. specific grammar, lexical, or semantic issues of individual Slavic languages from synchronic or diachronic point of view. The second bulk of papers consists of those dedicated to issues of corpus and digital infrastructure building, especially possibilities of improvement of corpus tagging. The third group of papers focuses on creation and use of language resources (lexicons and databases) for various target groups, both learners and native speakers. The three contributions from the fourth group fall into the traditional research agenda of NLP: sentiment analysis, as well as identification of spoken texts or specific expressions within them.

Biannual SLOVKO conferences, as it is evident from this year's proceedings as well as from all of the previous editions, have gained considerable geographical outreach. Presentations and papers were authored by researchers coming from not only "our" area, i.e., Slovakia, Czech Republic, Hungary, Poland and Austria, but also from Slovenia, Germany, Netherlands, France, Spain, Sweden, Finland, Greece, Bulgaria, Croatia, Ukraine, Russia, Georgia, USA, and Morocco. They have contributed to the spreading and sharing of insights into language(s) enabled by various corpora, language resources and tools. It is our hope that the SLOVKO conference, including this year's 10th anniversary edition, will remain a diverse and – as we believe – inspiring meeting point for language professionals, linguists, corpus linguists or IT specialists.

Editors

Slovko

Predhovor

„Čím dál víc se hlásí o slovo ta část korpusové lingvistiky, která se primárně nezabývá pouze sestavováním korpusů, otázkami reprezentativnosti, možnostmi lemmatizace a morfologického nebo jiného tagování. Její součástí jsou způsoby vytěžování korpusů, tedy způsoby zkvalitňování a doplňování popisů jazyka, objevování nových vztahů a vynalézání nových konceptů sloužících pro popis jazykové reality."

2019

V. Cvrček – T. Kováříková

Pri pohľade na príspevky v zborníku z jubilejnej konferencie SLOVKO 2019 je tvrdenie V. Cvrčka a T. Kováříkovej, publikované v časopise Naše řeč v roku 2011, stále aktuálne. Konštatovanie českých lingvistov však nezmenšuje relevanciu žiadnej oblasti v rámci korpusových výskumov jazyka. Naším zámerom pri tohtoročnej jubilejnej desiatej medzinárodnej lingvistickej konferencii SLOVKO preto bolo usporiadať jednotlivé príspevky v zborníku tak, aby organicky odzrkadľovali bohatú tematickú a metodologickú rôznorodosť súčasnej korpusovej lingvistiky a počítačového spracovania prirodzeného jazyka.

Prvú časť tvoria štúdie z oblasti „vyťažovania korpusov". Štúdie opierajúce sa o korpusy, resp. vychádzajúce z nich, ako aj štúdie zamerané na oblasť digitálnych humanitných a spoločenských vied (digital humanities), využívajú korpusy ako východisko pri synchrónnom či diachrónnom skúmaní jazykových javov, napríklad pri riešení špecifických gramatických, lexikálnych či sémantických otázok objavujúcich sa v jednotlivých slovanských jazykoch. Druhú skupinu príspevkov tvoria tie, ktoré sa venujú otázkam budovania korpusov a digitálnych infraštruktúr, napríklad možnostiam zlepšovania korpusového značkovania. Predmetom tretej skupiny príspevkov je tvorba a využívanie jazykových zdrojov (lexikónov a databáz) rôznymi cieľovými skupinami (či už ide o tých, ktorí sa učia cudzí alebo materinský jazyk). Tri príspevky vo štvrtej skupine sa zaoberajú tradičnými výskumnými úlohami počítačového spracovania prirodzeného jazyka: analýzou sentimentu, identifikáciou hovorených prejavov či špecifických výrazov v rámci nich.

Bienálne konferencie SLOVKO, ako je zrejmé z tohtoročného zborníka, ale aj zo všetkých predchádzajúcich ročníkov, pritiahli pozornosť vedcov a bádateľov nielen z nášho geografického priestoru, t. j. zo Slovenska, Českej republiky, Maďarska, Rakúska a Poľska, ale aj zo Slovinska, Nemecka, Holandska, Francúzska, Španielska, Švédska, Fínska, Grécka, Bulharska, Chorvátska, Ukrajiny, Ruska, Gruzínska, USA či Maroka. Konferencie tak prispievajú k šíreniu a zdieľaniu poznatkov o jazyku (jazykoch), získaných vďaka rôznym korpusovým zdrojom a nástrojom. Dúfame, že konferencia SLOVKO, vrátane aktuálneho desiateho ročníka, ostane aj v budúcnosti takou rôznorodou a – ako veríme – inšpiratívnou platformou pre stretávanie sa odborníkov pracujúcich s jazykom, jazykovedcov, korpusových lingvistov a špecialistov na IT.

Editori

Vedecký výbor / Programme Committee

Anić, Ana Ostroški	2017
Böhmerová, Adela	2019
Bojar, Ondřej	2015
Cvrček, Václav	2017
Čermák, František	2009, 2011, 2013, 2015
Dimitrova, Ludmila	2011, 2013
Ďurčo, Peter	2007, 2009, 2011, 2013, 2015
Erjavec, Tomaž	2015, 2017
Garabík, Radovan	2007, 2009, 2011, 2013, 2015
Genči, Ján	2017, 2019
Grác, Marek	2017, 2019
Hajič, Jan	2007, 2009, 2011
Hanks, Patrick	2007, 2009
Hebal-Jezierska, Milena	2019
Horák, Aleš	2007, 2009, 2011, 2017, 2019
Iomdin, Leonid	2013, 2015, 2017, 2019
Ivanová, Martina	2017, 2019
Jarošová, Alexandra	2007, 2009
Jelínek, Tomáš	2017, 2019
Kopřivová, Marie	2019
Křen, Michal	2015, 2017, 2019
Laclavík, Michal	2015
Levická, Jana	2015, 2017, 2019
Lopatková, Markéta	2019
Małachowicz, Marta	2017
Múcsková, Gabriela	2017
Nábělková, Mira	2011
Ogrodniczuk, Maciej	2015, 2017
Oliva, Karel	2007, 2009, 2011, 2013
Osolsobě, Klára	2017, 2019
Pala, Karel	2007, 2009, 2011, 2013, 2015
Pekarovičová, Jana	2013
Petkevič, Vladimír	2007, 2009, 2011, 2013, 2015, 2017, 2019
Piasecki, Maciej	2015, 2017
Pognan, Patrice	2013, 2015, 2017, 2019
Przepiórkowski, Adam	2011, 2013
Schmitz, Klaus-Dirk	2007
Sokolová, Miloslava	2009
Stoykova, Velislava	2017, 2019
Šimková, Mária	2013, 2017, 2019
Štefčík, Jozef	2017
Štícha, František	2009
Uhliarik, Ivor	2017, 2019
Urešová, Zdenka	2019
Váradi, Tamás	2011, 2013, 2017, 2019
Zacharov, Viktor	2009, 2011

Organizačný výbor / Organizing Committee

Avramovová, Miroslava	2007
Benko, Vladimír	2001, 2003
Bobeková, Kristína	2019
Gajdošová, Katarína	2011, 2013, 2015, 2017, 2019
Garabík, Radovan	2005
Gaži, Gabriel	2003
Holičová, Bronislava	2003
Horák, Alexander	2005
Chochol, Martin	2003
Janočková, Nikol	2003
Jarošová, Alexandra	2001, 2005
Levická, Jana	2005, 2007, 2009, 2017, 2019
Majchráková, Daniela	2009, 2011
Majirošová, Mária	2013
Ondrejovič, Slavomír	2005
Šimková, Mária	2005, 2007, 2009, 2011, 2019
Vančo, Tomáš	2005
Vasilišinová, Dorota	2007
Zvončeková, Dáša	2007
Žáková, Adriána	2013, 2015, 2017

Autori /Authors

NÁRODNÝ SLOVENSKÝ KORPUS

Jazykovedný ústav Ľudovíta Štúra SAV Bratislava

nástroje

Speech

Data

slovníky

Corpora

Terminology

anotácia

Transcription

dáta

text

zdroje

jazyk

terminológia

Dictionaries

transkripcia

Annotation

morfológia

analyzátor

Model

Language

korpus

https://korpus.sk/slovko.html