SLOVKO 2011.
POČÍTAČOVÉ SPRACOVANIE PRIRODZENÉHO JAZYKA, MULTILINGVÁLNOSŤ

V dňoch 20. – 21. 10. 2011 sa v kongresovom centre hotela Majolika v Modre uskutočnila šiesta medzinárodná konferencia SLOVKO 2011, organizovaná oddelením Slovenského národného korpusu Jazykovedného ústavu Ľ. Štúra SAV v Bratislave a tematizovaná ako Počítačové spracovanie prirodzeného jazyka, multilingválnosť (NLP, Multilinguality). Podujatie slávnostne otvorila Nicol Janočková, zástupkyňa riaditeľa Jazykovedného ústavu Ľ. Štúra SAV v Bratislave, spolu s Radovanom Garabíkom, jedným z organizátorov podujatia.

Na konferencii vystúpilo 17 prednášajúcich z Bulharska, Českej republiky, Slovenska, Slovinska a Ukrajiny. Referujúci mohli predniesť svoj príspevok v jednom z rokovacích jazykov, ktorými boli všetky slovanské jazyky a angličtina. Okrem prednášajúcich sa podujatia zúčastnili aj hostia z Ruska a záujemcovia zo Slovenska, a tak bolo možné začuť v kuloároch debaty vo viacerých slovanských jazykoch. V zborníku tlačených príspevkov – *Natural Language Processing, Multilinguality*. (Eds. Daniela Majchráková – Radovan Garabík. Brno: Tribun 2011. 173 s. ISBN 978-80-263-0049-6), ktorý mali účastníci konferencie k dispozícii už na podujatí, je publikovaných 19 príspevkov. Texty príspevkov v zborníku sú publikované v anglickom jazyku.[1]

V správe prinášame informáciu aj o príspevkoch, ktoré sú publikované v zborníku, no ich autori sa na podujatí nezúčastnili a nemohli ich predniesť. Uvádzame ich na konci textu spolu s troma odprezentovanými príspevkami, ktoré síce na konferencii odzneli, ale v zborníku nie sú publikované. Inak sú informácie o textoch v tejto správe radené abecedne podľa priezviska autora (v prípade viacerých autorov podľa priezviska prvého) tak, ako sú usporiadané aj v zborníku.

Diagnostika chýb je neoddeliteľnou súčasťou zlepšovania kvality a odolnosti každého systému automatického rozpoznávania reči, a to najmä v prípade jazykov s limitovaným množstvom zdrojov. Štefan Beňuš z Filozofickej fakulty Univerzity Konštantína Filozofa v Nitre a Miloš Cerňak, Milan Rusko, Marián Trnka, Sachia Darjaa a Róbert Sabo z Ústavu informatiky Slovenskej akadémie vied v Bratislave skúmajú v príspevku *Semi-automatic Approach to ASR Errors Categorization in Multi-speaker Corpora* poloautomatický prístup ku kategorizácii chýb, využiteľný v databázach, ktoré obsahujú množiny identických viet produkovaných dostatočne veľkým množstvom hovoriacich. Používajú pritom matice vytvorené z usporiadaného zoznamu hovoriacich a z usporiadaného zoznamu viet. Autori navrhli algoritmus, ktorý pomocou takejto matice vyhľadáva chyby pri automatickom rozpoznávaní reči.

V príspevku *Towards a Multilingual Database of Verb-related Terminology* predstavili Katarína Chovancová a Jana Klincková z Fakulty humanitných vied Univerzity Mateja Bela v Banskej Bystrici mnohojazyčnú databázu lingvistických termínov. Databáza vzniká na domovskom pracovisku autoriek článku a na jej tvorbe sa podieľa niekoľkočlenný tím spolupracovníkov – slovakistov aj romanistov. Cieľom databázy je utvoriť zázemie pre slovenskú lingvistickú terminológiu. Koncovými používateľmi databázy majú byť všetci, ktorí pracujú

[1] V záujme zhody mien autorov príspevkov v programe konferencie, v zborníku a publikovanej správe ponechávame prepisy z cyrilského písma do latinského písma podľa anglickej normy tak, ako ho vo svojich príspevkoch uviedli ich autori. (Pozn. red.)

s prekladmi odborných lingvistických textov do románskych jazykov (francúzština, španielčina, taliančina). Pri heslovom slove v zázname sa uvádzajú informácie o jeho gramatických kategóriách, ďalej výslovnosť, skrátený termín, synonymá heslového slova, kolokácie, kontext, v ktorom sa termín používa v odbornej literatúre aj s uvedením zdroja, definícia, príklady, odbor, disciplína, teoretický smer, v ktorom sa termín používa, tematická oblasť, hodnotiace kritérium. Je tu aj možnosť uviesť jazykovú a encyklopedickú poznámku, ako aj hyperonymá, izonymá, hyponymá, antonymá a príbuzné termíny. Databáza zatiaľ nie je odbornej verejnosti prístupná.

O existencii bulharsko-slovenského paralelného korpusu informujú v článku *Bulgarian –Slovak Parallel Corpus* Ludmila Dimitrova z Inštitútu matematiky a informatiky Bulharskej akadémie vied v Sofii a Radovan Garabík z Jazykovedného ústavu Ľ. Štúra Slovenskej akadémie vied v Bratislave. Bázou korpusu je bulharská beletria preložená do slovenčiny, slovenská beletria preložená do bulharčiny a texty preložené do oboch jazykov z tretieho jazyka. Korpus je zarovnaný na úrovni viet bez použitia dvojjazyčného počítačového slovníka. Slovenská časť textov je automaticky lematizovaná a morfologicky anotovaná. Perspektívne autori korpusu počítajú s morfologickou anotáciou bulharskej časti korpusu a automatizovanou syntaktickou anotáciou oboch častí. Korpus je verejne prístupný na stránke Slovenského národného korpusu http://korpus.sk:8090/ a mohol by slúžiť ako pomôcka pri strojovom preklade a ako trénovací materiál pre štatistické modely. Už v tejto podobe je použiteľný pri výučbe v škole.

Príspevok Petra Ďurča z Filozofickej fakulty Univerzity sv. Cyrila a Metoda v Trnave a Pedagogickej fakulty Univerzity Komenského v Bratislave *The Slovak Dictionary of Collocations* je venovaný zásadám prvého slovenského slovníka kolokácií. Slovník vychádza z bázy Slovenského národného korpusu a pozostáva z 250 kolokačných profilov najfrekventovanejších slovenských podstatných mien.

Princeton WordNet je lexikálna databáza, ktorá obsahuje súbory anglických synoným spolu s ich sémantickými vzťahmi. Ondrej Dzurjuv a Ján Genči z Fakulty elektrotechniky a informatiky Technickej univerzity v Košiciach a Radovan Garabík z Jazykovedného ústavu Ľ. Štúra SAV v Bratislave sa v spoločnom príspevku *Generating Sets of Synonyms between Languages* zaoberajú niekoľkými metódami vytvárania synsetov v ďalších jazykoch s použitím anglického WordNetu a príslušného dvojjazyčného slovníka. Tieto metódy sa používajú na vytváranie slovenských synsetov a automatické generovanie slovenskej databázy WordNet.

Možnosťami česko-slovenského strojového prekladu sa v príspevku *Czech-Slovak Parallel Corpora for MT between Closely Related Languages* venovali Petra Galuščáková a Ondřej Bojar z Matematicko-fyzikálnej fakulty Univerzity Karlovy v Prahe. Za bázové dáta si zvolili primárne knihy zo slovensko-českého paralelného korpusu utvoreného v oddelení Slovenského národného korpusu, paralelný korpus Acquis JRC utvorený z voľne dostupných textov EÚ, texty z webových stránok Európskej komisie a texty oficiálneho časopisu EÚ – EurLEX, ktoré rozdelili na sety trénovacie, ladiace a testovacie v rôznych pomeroch. V príspevku načrtli postup strojového prekladu medzi dvoma blízkopríbuznými jazykmi a poukázali na nevyhnutnosť množstva relevantných dát, ktoré sú potrebné na testovanie jeho výsledkov.

Polona Gantar z Vedecko-výskumného centra Slovinskej akadémie vied a umení v Ľubľane a Simon Krek zo spoločnosti Amebis v Kamniku a Inštitútu Jožefa Stefana v Ľubľane predstavili v príspevku *Slovene Lexical Database* koncept novej lexikálnej databá-

zy slovinčiny. Jej zostavovatelia si stanovili dva ciele: vytvoriť platformu na zostavenie výkladových aj prekladových slovníkov slovinčiny a zároveň zdokonaliť nástroje počítačového spracovania slovinčiny. Databáza je usporiadaná do šiestich informačných úrovní, počnúc lexikálno-gramatickými informáciami (od morfológie po sémantiku) cez syntaktické, kolokačné a frazeologické informácie až po obohatenie databázy o príklady z korpusu. Zdrojom dát je korpus FidaPLUS vo veľkosti 620 miliónov slov. Databáza má slúžiť na automatickú dezambiguáciu slovinčiny.

Novátorský prístup k anotácii syntaktickej roviny českého jazyka ponúkol vo svojom príspevku *Building Annotated Corpora without Experts* Marek Grác z Fakulty informatiky Masarykovej univerzity v Brne. V prvej časti projektu anotácie rôznych jazykových rovín korpusu predstavuje autor možnosť anotácie syntaktickej roviny češtiny, ktorú vykonávajú nie špecificky školení anotátori, ale študenti. Autor zdôrazňuje nevyhnutnosť jednoduchého manuálu, ktorý anotátori používajú, keď overujú správnosť automatizovaného nástroja na identifikáciu menných a slovesných skupín, koordinácií a jednoduchých viet v texte. Anotátori overujú len ne/správnosť elementov, ktoré sa vzťahujú na príslušnú vetu a sú v nej značkované automatizovane. Autor dodáva, že hoci tento spôsob využitia ľudských zdrojov pri značkovaní veľkého množstva dát nie je dokonalý, je postačujúci a dostatočne ekonomický z hľadiska času a množstva označkovaných dát pre potreby vyvíjania automatizovaných nástrojov a poloautomatizovanú mnohoúrovňovú anotáciu špecifického korpusu.

Václava Kettnerová a Markéta Lopatková z Matematicko-fyzikálnej fakulty Karlovej univerzity v Prahe informovali v príspevku *The Lexicographic Representation of Czech Diatheses: Rule Based Approach* o novom spôsobe reprezentácie českých diatéz vo valenčnom lexikóne českých slovies. Diatézy rozdelili na tri typy: gramatické, syntaktické a sémantické. Autorky na základe doterajších výskumov konštatovali, že gramatické a syntaktické diatézy je možné zachytiť formálnymi syntaktickými pravidlami. Perspektívne plánujú hlbší výskum s kombináciou rôznych typov diatéz.

Strojový preklad predložiek z češtiny do ruštiny predstavili v príspevku *Translating Prepositions from Czech into Russian: Challenges for the Machine Translation* Natalia Klyueva z Matematicko-fyzikálnej fakulty Karlovej Univerzity v Prahe a Naděžda Runštuková z Filozofickej fakulty Karlovej Univerzity v Prahe. Autorky zhrnuli rozdiely v používaní predložiek v češtine a ruštine a informovali o spôsobe, akým ich strojový preklad spracúva. Úlohou výskumu je zmapovať chyby, ktoré sa vyskytujú v prekladoch predložkových spojení. Príspevok zahŕňa výskum česko-ruského paralelného korpusu, ako aj analýzu strojového prekladu systému Česílko založeného na pravidlách a strojového prekladu systému Joshua založeného na štatistike. Na základe výsledkov porovnania autorky konštatovali, že strojový preklad založený na štatistike je pri preklade predložiek presnejší. Súčasne navrhli zlepšenia schémy pre systémy strojového prekladu.

Morfologickému tageru sa v príspevku *A Web-based Morphological Tagger for Bulgarian* venoval autorský kolektív v zložení Aleksandar Savkov, Laska Laskova, Petya Osenova, Kiril Simov a Stanislava Kancheva z Inštitútu informačných a komunikačných technológií Bulharskej akadémie vied v Sofii. Autori predniesli informácie o morfosyntaktickom tageri a lematizátore pre bulharčinu. Pri jeho tvorbe využili SVM tager, morfologické slovníky bulharčiny a lingvistické pravidlá. Perspektívne chcú zlepšiť stratégiu výberu správnych tagov pri anotácii a sprístupniť jednoduchý parser.

Autorský kolektív Ján Staš, Daniel Hládek a Jozef Juhár z Fakulty elektrotechniky a informatiky Technickej univerzity v Košiciach a Marián Trnka z Ústavu informatiky Slovenskej akadémie vied v Bratislave opísali v príspevku *Automatic Extraction of Multiword Expressions Using Linguistic Constraints for Slovak LVCSR* proces automatickej extrakcie najfrekventovanejších viacslovných výrazov z korpusu textových dát z oblasti súdnictva a proces trénovania a testovania modelu slovenského jazyka s týmito výrazmi. Automatická extrakcia viacslovných výrazov v systéme bola podmienená spodobovaním a zdvojovaním hlások na hranici slov s cieľom znížiť chybovosť pri rozpoznávaní krátkych jednoslabičných slov. Výsledky modelovania jazyka pomocou viacslovných výrazov ukazujú mierne zlepšenie v presnosti rozpoznávania krátkych jednoslabičných slov na začiatku viet a po dlhých pauzách.

Velislava Stoykova z Inštitútu bulharského jazyka Bulharskej akadémie vied v Sofii analyzovala vo svojom príspevku *Common Formal Framework for Multilingual Representation of Inflectional Morphology for Two Related Slavonic Languages* možnosti spoločného skúmania dvoch príbuzných morfologických systémov – ruštiny a bulharčiny, pričom využila jazyk DATR na formulovanie pravidiel pre flektívne kategórie lexém v oboch jazykoch. Porovnávala zásady a motivácie navrhovaného kódovania, ktoré využíva nemonotónne ortogonálne sémantické siete. V príspevku ponúkla zásady spracovania flektívnej morfológie zavedením sémantickej hierarchie pomocou tradičných gramatických pravidiel.

Ludmila Dimitrova z Inštitútu matematiky a informatiky Bulharskej akadémie vied v Sofii a Violetta Koseska-Toszewa, Danuta Roszko a Roman Roszko z Inštitútu slovanských štúdií Poľskej akadémie vied vo Varšave informujú v príspevku *Bulgarian-Polish-Lithuanian Corpus – Recent Progress and Application* o projekte tvorby trilingválneho korpusu. Textovou bázou korpusu sú beletristické texty v jednom z týchto jazykov, pričom vo zvyšných troch existujú ich preklady. Okrem toho sa v korpuse nachádzajú aj preklady textov (dokumenty EÚ či beletria svetových autorov) do bulharčiny, poľštiny a litovčiny z tretieho jazyka. Zložkou porovnávacieho korpusu sú publicistické texty z internetu, existujúce vo všetkých troch jazykoch (najmä reakcie médií na medzinárodné udalosti). Autori zdôrazňujú, že predstavený korpus či akýkoľvek iný multilingválny korpus je nenahraditeľnou základňou pre tvorbu dvoj- a viacjazyčných slovníkov a porovnávacích gramatík. Využitie nachádza u prekladateľov, pri vyučovaní cudzích jazykov v odbore translatológia, ale aj pri výučbe cudzincov. Nezanedbateľným prínosom je použitie multilingválneho korpusu ako jazykového materiálu na trénovanie počítačových nástrojov na strojový preklad.

Príspevok *The Instrumental Environment for the Automatic Syntactical Analysis of Ukrainian*, ktorý predstavili Iryna Zamaruieva a Olga Shypnivska z Kyjevskej národnej univerzity Tarasa Ševčenka v Kyjeve sa zaoberá opisom inštrumentálneho prostredia pre automatickú syntaktickú analýzu ukrajinčiny. Autorky v ňom uvádzajú hlavné charakteristiky databáz na automatickú syntaktickú analýzu a všeobecné princípy automatickej syntaktickej analýzy.

V každom jazyku, ktorý používa pády, slovesá ovplyvňujú zmeny v podstatných menách a v predložkových spojeniach. V ruštine je toto variantné ovplyvňovanie vzhľadom na jej bohatú morfológiu veľmi rozšírené. Odvodzuje sa z diachrónnych procesov aj sémantických posunov v modernom ruskom jazyku. Možnosti automatickej extrakcie slovies so substantívnymi predložkovými a bezpredložkovými spojeniami z korpusu trigramov prezentovali Mikhail Kopotev z Univerzity v Helsinkách, Natalia Kochetkova z Moskovského štátne-

ho inštitútu elektroniky a matematiky v Moskve a Eduard Klyshinsky z Keldyshovho inštitútu aplikovanej matematiky Ruskej akadémie vied v Moskve v príspevku *Extracting Verbs with PP/NP Variation from the Large 3-gram Corpus*. Algoritmus extrakcie je založený na vyhľadaní trojíc sloveso – (predložka) – substantívum a ich vzájomnom automatickom porovnávaní.

Príspevok Iriny Nekipelovej zo Štátnej technickej univerzity v Iževsku *On the Question of Homonymy and Polysemy in the Lexicographical Practice of the Russian Language Semantic System in its Development Modelling* predstavuje elektronický model historicko-etymologického slovníka. Autorka v ňom navrhuje nový pohľad na diferenciáciu a koreláciu sémantickej slovotvorby, založený na histórii jazyka. Toto rozlíšenie umožňuje lepšie pochopiť homonymné a polysémické javy, identifikovať a systematizovať procesy sémantických zmien v priebehu jazykového vývinu. Lexikálno-sémantické reprezentácie slov ako jednotky lexikografického opisu sú založené na týchto procesoch.

Irina Nekipelova a Elvira Zarifullina zo Štátnej technickej univerzity v Iževsku prezentujú v príspevku *Historical and Etymological Electronic Dictionary* koncept tvorby elektronického historicko-etymologického slovníka, informujú o zložení textov uvedených v databáze, o štruktúre systému a jeho jednotlivých zložkách (modul vizualizácie, dopĺňací a vyhľadávací modul).

Slovanský paralelný korpus ParaSol vyvinutý v Berne a Regensburgu opisuje vo svojom príspevku *Recent Developments in ParaSol: Breadth for Depth and XSLT Based Web Concordancing with CWB* Ruprecht von Waldenfels z Inštitútu slovanských jazykov a literatúr Univerzity v Berne. V príspevku autor informuje o súčasnom stave výskumu so zameraním na koncepčné rozhodnutia, ktoré sa týkajú dopĺňania korpusu a užívateľského rozhrania.

Mária Šimková z Jazykovedného ústavu Ľ. Štúra Slovenskej akadémie vied v Bratislave predniesla príspevok *Ekvivalencia epistemických častíc v slovensko-českom paralelnom korpuse*, v ktorom predstavila použitie paralelného korpusu blízkopríbuzných jazykov na výskum epistemických častíc *zrejme, doista, zaiste, pravdepodobne* a spôsobu ich prekladu. Autorka vyčlenila tri skupiny použitia skúmaných epistemických častíc v oboch korpusoch: priame ekvivalenty, ktoré sú priamo využiteľné v lexikografickej praxi; posuny pri používaní, ktoré stimulujú lingvistické skúmanie a hľadanie odpovedí na otázku smeru prekladu; tzv. voľné preklady, teda rozdielne použitie v každom z jazykov, ktoré je typické pri preklade textov cez tretí jazyk a otvára otázku adekvátnosti (vernosti) prekladu, čo je využiteľné vo výučbe translatológov.

V príspevku *The Czecho-Slovak Sketching: A Compatible Word Sketch Grammar for Czech and Slovak* hovoril Vladimír Benko z Jazykovedného ústavu Ľ. Štúra Slovenskej akadémie vied v Bratislave o možnostiach tvorby pravidiel pre nástroj Sketch Engine pre potreby porovnávacích výskumov. Predstavil využiteľné slovenské a české korpusy, osobitne sa pristavil pri veľkých zdrojoch dát, ktorými sú dnes dostupné slovenské a české webové korpusy, ale zdôraznil aj obmedzenia, ktoré vyplývajú z takejto úlohy. Najväčším z nich je azda nekompatibilita medzi pravidlami tvorby gramatiky word sketchu, pričom český word sketch používa A-style sketch grammar, slovenský V-style sketch grammar. Riešením by mohlo byť inkorporovanie českého webového korpusu do slovenského Sketch Engine a prepísanie, resp. transformovanie slovenskej sketch gramatiky na tagset používaný taggerom Ajka.

Peter Baláž zo spoločnosti Edukácia@Internet predniesol príspevok *Slovak Online*, v ktorom informoval o rovnomennom projekte, jeho priebehu a výsledkoch. Spomenul spoluprácu s partnerskými domácimi i zahraničnými inštitúciami aj ďalšie projekty, ktoré nadväzu-

jú na projekt Slovak Online. Všetky informácie o projekte je možné nájsť na webovej stránke: http://slovake.eu/sk/.

Záverečné poďakovanie všetkým zúčastneným aj prednášajúcim vyjadril za Jazykovedný ústav Ľ. Štúra SAV Vladimír Benko. Osobitne poďakoval členom oddelenia Slovenského národného korpusu Jazykovedného ústavu Ľ. Štúra SAV za organizáciu príjemného podujatia v malebnom prostredí mestečka pod Karpatmi. Ako uviedla Mária Šimková zo Slovenského národného korpusu Jazykovedného ústavu Ľ. Štúra SAV v úvode konferenčného zborníka, už dnes je známe hlavné tematické zameranie ďalšej konferencie s dobre známym názvom – SLOVKO 2013. Bude zamerané na sémantické siete a multilingválne slovníky, no bude otvorené aj pre všetky tradičné témy z oblasti počítačového spracovania prirodzeného jazyka a korpusovej lingvistiky.

Katarína Gajdošová – Beáta Kmeťová – Adriána Žáková

JAZYK A DISKURZ V KULTÚRNOM A POLITICKOM KONTEXTE

Medzinárodná vedecká konferencia venovaná „Jazyku a diskurzu v kultúrnom a politickom kontexte“ v dňoch 21. – 23. septembra 2011 prebiehala tak, ako to predpovedal jej prvý rečník, a zároveň jeden z odborných garantov konferencie, riaditeľ Jazykovedného ústavu Ľ. Štúra SAV, prof. PhDr. Pavol Žigo, CSc.: „Myslím si, že obsah jednotlivých referátov bude obohatením každého účastníka konferencie, že sa na tejto konferencii schádza spoločnosť, v ktorej diskurz nebude rozptyľovaný, tak ako to bolo v klasickej latinčine a že to nebude ani hádanie ani nič, čo by túto spoločnosť delilo. Všetci sme sem prišli s tým, že sa vzájomne obohatíme obsahom jednotlivých referátov.“ Práve Jazykovedný ústav L. Štúra SAV – konkrétne Oddelenie súčasného jazyka – sa zhostil úlohy privítať štyridsať referujúcich zo Slovenska, z Česka, Poľska, Maďarska, Rakúska a Estónska v Kongresovom centre SAV v Smoleniciach.

Podujatie sa stalo zavŕšením grantového projektu VEGA – *Slovenčina ako kultúrny jav a médium kultúry* – ktorý sa realizoval pod vedením Juraja Dolníka a hoci nie fyzicky, predsa bol Juraj Dolník – ako najčastejšie citovaný autor – prítomný v pozoruhodnom množstve príspevkov, a čo je dôležité, príspevkov obsahovo rôznorodých.

Predovšetkým program prvého dňa anticipoval interdisciplinárne smerovanie celého podujatia. Prívlastok „najheterogénnejšia krajina Strednej Európy“ robí zo Slovenska výnimočne zaujímavú krajinu na výskum. A práve rozbiehajúci sa výskum *Jazyková situácia a jazyková politika na Slovensku v medzinárodnom kontexte*, sčasti nadväzujúci na projekt Eugena Paulinyho zo šesťdesiatych rokov minulého storočia, predstavil Slavomír Ondrejovič. v príspevku *Jazyková situácia a jazyková politika – druhé dejstvo*. Kládol si v ňom otázku, či pomerne silné zastúpenie menšín zaručuje rozvoj spoločenského života na Slovensku v intenciách multikulturalizmu, liberalizmu, tolerancie. Avšak, ako zdôraznil, neexistujú výskumy, ako na Slovensku jazyky žijú, ako vnímajú svoju situáciu minoritní, ale aj majoritní obyvatelia, a to najmä v súčasnej zložitej jazykovopolitickej situácii.