SLOVKO 2009
POČÍTAČOVÉ SPRACOVANIE PRIRODZENÉHO JAZYKA, KORPUSOVÁ LINGVISTIKA A GRAMATICKÝ VÝSKUM

Jubilejný 5. ročník medzinárodnej konferencie SLOVKO 2009, tento rok tematizovanej ako *Počítačové spracovanie prirodzeného jazyka, korpusová lingvistika a gramatický výskum*, sa uskutočnil v dňoch 25. – 27. novembra 2009 v Smoleniciach. Podujatie slávnostne otvorila Mária Šimková, vedúca oddelenia Slovenského národného korpusu Jazykovedného ústavu Ľ. Štúra Slovenskej akadémie vied v Bratislave, spolu s Janou Levickou, hlavnou organizátorkou konferencie. Odborné skúsenosti z oblasti korpusovej lingvistiky a gramatického výskumu na báze korpusov si na vedecké podujatie do príjemného prostredia Kongresového centra Slovenskej akadémie vied v Smoleniciach prišlo vymeniť a o svojich výsledkoch poinformovať vyše 40 prednášateľov z 13 krajín (Bulharsko, Česká republika, Francúzsko, Holandsko, Nemecko, Nový Zéland, Poľsko, Rakúsko, Rusko, Slovensko, Slovinsko, Španielsko, Ukrajina). Príspevky boli prednesené zväčša v angličtine, ale zaznela aj ruština, ukrajinčina, poľština, slovinčina, čeština či slovenčina. Na podujatí mali účastníci k dispozícii aj tlačená verziu príspevkov – zborník *NLP, Corpus Linguistics, Corpus Based Grammar Research*. Eds. J. Levická – R. Garabík. Brno: Tribun 2009. 398 s. ISBN 976-80-73399-875-2.

Prednesené príspevky v rámci vymedzených tém zahŕňali viaceré oblasti. Zastúpené boli analýzy hovorených a paralelných korpusov, počítačová lexikografia, terminológia, tvorba korpusových nástrojov, databáz, ale aj sémantický, gramatický a štatistický výskum na korpusovej báze.

Výsledky výskumu na hovorených korpusoch prezentoval Václav B ř e z i n a (Filozofická fakulta Aucklandskej univerzity na Novom Zélande) v príspevku „*We Only Say We Are Certain When We Are Not": A Corpus-Based Study of Epistemic Stance*. Modálne postoje v reči skúmal na materiáli dvoch korpusov hovorenej angličtiny – na hovorenej časti BNC (British National Corpus) a MICASE (Michigan Corpus of Academic Spoken English). Podľa autora je pri medziľudskej komunikácii paradoxom, že o našej istote hovoríme len vtedy, keď si istí nie sme. Jeho analýza štyroch slov *must*, *certain*, *sure, certainly* potvrdzuje toto tvrdenie, to však platí pravdepodobne len vtedy, ak ide o neistotu v intersubjektívnom rozhraní.

Svetlana S a v č u k (Vinogradovov inštitút ruského jazyka Ruskej akadémie vied v Moskve) informovala v príspevku *Spoken Texts Representation in the Russian National Corpus: Spoken and Accentologic Sub-Corpora* o dvoch ruských hovorených subkorpusoch. Vysvetlila spôsob značkovania, štruktúru ruských hovorených korpusov, prepis zvukových záznamov a lingvistické anotácie zozbieraných dát (metatextová, morfologická a sociologická anotácia). Pozornosť venovala aj problémom spojeným s jednotlivými výstavbovými prvkami hovorených korpusov. Pri jazyku s nefixným prízvukom, akým je ruština, zohráva dôležitú úlohu práve korpus s označenými prízvukmi, aby mohol používateľ pracovať s rozsiahlym materiálom potrebným na rôzne výskumy akcentologických špecifikácií v ruštine.

Jelena G r i š i n a (Inštitút ruského jazyka Ruskej akadémie vied v Moskve; *Multimodal Russian Corpus (MURCO): General Structure and User Interface*) predstavila časť databázy Ruského národného korpusu – MURCO. Ide o špecifickú databázu zvukových a filmových záznamov a ich prepisov, ktorá zahŕňa najmä filmové dáta z ruského hovoreného korpusu. Ruské hovorené korpusy sú sociolingvisticky, morfologicky, sémanticky, foneticky a akcento-

logicky označkované. MURCO ponúka možnosť vlastnej fonetickej anotácie filmových alebo zvukových záznamov používateľom. Perspektívne budú do databázy MURCO zahrnuté všetky filmové záznamy ruského hovoreného korpusu (5 miliónov textových jednotiek).

V príspevku *A Corpus of Spoken Language and Its Usefulness in the Research on Language Contact* priniesol Marcin Z a b a w a (Anglický inštitút Sliezskej univerzity v Katoviciach) pohľad na výskumu anglických slovných výpožičiek v poľštine v spontánnych neformálnych prehovoroch. Výskum realizoval na vlastnom hovorenom korpuse. Zvláštny dôraz kládol na stavbu korpusu – jeho zloženie, výber respondentov a neformálnych spontánnych prejavov, spôsob prepisu. Výsledky výskumu sú obmedzené rozsahom súkromného korpusu, pretože výskyt skúmaných anglických výpožičiek je veľmi nízky. Autor teda odporúča venovať sa cielene problematike inojazyčných výpožičiek v poľštine na báze veľkých hovorených korpusov utvorených z prehovorov mnohých hovoriacich z rôznych oblastí Poľska v neformálnych formách prehovoru – besedy, rozhovory a pod.

Didakticky ladený príspevok Marka N a g y a (Katedra aplikovanej informatiky Matematicko-fyzikálnej fakulty Univerzity Komenského v Bratislave; *Using Speech and Handwriting Recognition in Electronic School Worksheets*) ponúkol spôsob, ako použiť rozpoznávanie reči a písma v elektronických zošitoch pri vyučovaní detí na základnej škole. Autor analyzuje problémy rozpoznávania reči na takýto špecifický účel. Snaží sa preniesť cvičenia z tlačenej podoby cvičebníc do počítača a rozšíriť ich o nové zaujímavé funkcie. V teréne dopĺňa svoj fonetický slovník, aby ním pokryl čo najširšiu škálu výslovnostných realizácií vyučovaných hlások, resp. slov.

Príspevok *Introduction of Non-Verbal Means of Communication in the Corpus of Live Speech* od autoriek Taťjany P e t r o v e j a Oľgy L y s (Ďalekovýchodná národná univerzita vo Vladivostoku) ponúkol nahliadnutie do problematiky včleňovania živej reči do štruktúr hovorených korpusov. Autorky sa zamerali najmä na metatextové informácie, ktoré zahŕňajú neverbálne prvky prejavu – mimiku, gestá a pod. Pri včleňovaní takýchto osobitých prejavov do celku hovorených korpusov je potrebné dbať na zachytenie aj tých častí komunikácie, ktoré sa štandardne pri iných typoch prejavov nenachádzajú. Dôležité sú i tie prostriedky, ktoré neverbálne nahrádzajú časť komunikácie, a teda nie je možné ich zvukovo zachytiť. Autorky venovali pozornosť prejavom, ktoré sprevádzajú nejakú udalosť, a prejavom, ktoré sú zložené zo spomienok na nejakú udalosť.

Výskum na paralelnom česko-anglickom korpuse realizovala Renata Š i m ů n k o v á (Technická univerzita v Liberci). V príspevku *Analysis of the Means Expressing Strong ‚Necessity Not To' in English and Czech Based on General and Parallel Corpora* prezentovala výsledky štúdie založenej na korpusoch a zameranej na zistenie základných prostriedkov použitých na vyjadrenie významu deontického 'necessity not to' v angličtine a jeho pendantov v češtine. Textovou analýzou sa odlíšili rôzne sémantické hodnoty týchto výrazov. Autorka sa zároveň zaoberala tým, ako je rozsah výberu čiastkových významov ovplyvnený spoluvýskytom významov na odlišných úrovniach: lexikálnej, gramatickej, pragmatickej, ako aj typom výpovede.

S najnovším projektom 12-jazyčného paralelného korpusu oboznámil prítomných Emmerich K e l i h (Slavistický inštitút Univerzity v Grazi; *Preliminary Analysis of a Slavic Parallel Corpus*). Korpus pozostáva z prekladov ruskej novely Ako sa kalila oceľ autora Alexandra Ostrovského do 11 slovanských jazykov: slovinčiny, chorvátčiny, srbčiny, mace-

dónčiny, bulharčiny, ukrajinčiny, bieloruštiny, slovenčiny, češtiny, poľštiny a hornej lužičtiny. V príspevku sa zaoberá exploratívnou štatistickou analýzou a lingvistickou interpretáciou textovej dĺžky. Dĺžka textu je podľa autora jednoduchým, ale účinným nástrojom na zachytenie morfologických špecifikácií skúmaných jazykov.

V príspevku *Corpus of Computational Linguistics Texts* autorský kolektív Taťjana Bobková, Marija Kasjanenko, Kuzma Lebedev, Valentyna Lukaševyč, Pavlo Petrenko a Ľubov Grydneva (Laboratórium komputačnej lingvistiky Kyjevskej národnej filologickej univerzity v Kyjeve) informoval o počítačovom nástroji na spracovanie trilingválneho anglicko-rusko-ukrajinského korpusu textov z oblasti počítačovej lingvistiky. Každý subkorpus obsahuje asi 500 tisíc textových jednotiek a je zložený z lingvistických štúdií a príručiek v jednotlivých jazykoch. Na základe korpusu bol zostavený slovník v rozsahu asi 1 000 termínov z oblasti počítačovej lingvistiky. Korpus slúži na štatistický výskum používania termínov, gramatických tvarov a kolokácií v odborných textoch. Nástroj, v ktorom sa korpus používa, ponúka vyhľadávanie termínu, jeho výklad, ruské a ukrajinské preklady termínu a kontexty, v ktorých je termín vo všetkých troch subkorpusoch použitý.

V terminologicky orientovanom príspevku *IT: Moving Towards Real Multilingualism* predstavili Cristina Borrell a Antoni Oliver (Otvorená univerzita Katalánska) projekt zameraný na obohatenie tezauru Eurovoc, pozostávajúceho z termínov európskej legislatívy v oficiálnych európskych jazykoch, o termíny v katalánčine. Použitím nástroja na automatické vyhľadávanie termínov z oficiálnych dokumentov Európskej únie vyextrahovali viacjazyčný glosár obsahujúci viac ako 2 500 termínov, okrem iných aj v katalánskom jazyku. Následne vyvinuli štatistickú metódu automatického prekladu, pomocou ktorej preložili španielske termíny z Eurovocu do tohto jazyka. Zdôraznili, že použitím tejto metódy a vhodných zdrojov sa môže zefektívniť a urýchliť preklad.

Viktor Zacharov (Filologická fakulta Štátnej univerzity v Petrohrade) predniesol príspevok *Automatic Analysis of Terminology in the Russian Corpus on Corpus Linguistics*, na ktorom spolupracoval s Oľgou Mitrofanovou (Inštitút lingvistických štúdií Ruskej akadémie vied). Prezentoval výskum zameraný na vyvinutie stratégie poloautomatizovanej terminologickej extrakcie z korpusu ruských textov. Extrakcia sa vykonáva pomocou lingvistických a štatistických nástrojov, ktoré umožňujú generovať zoznamy jednoslovných a viacslovných termínov doplnené o frekvenčné údaje a lexikálno-syntaktické vzorce. Tieto vzorce sa ďalej používajú na analýzu kontextov, ktoré obsahujú definície termínov, odhaľujú vzájomné vzťahy medzi termínmi, umožňujú získavať synonymá k termínom, cudzojazyčné ekvivalenty a pod. Získané výsledky budú súčasťou prípravy ruského tezaura.

Ľubomír Kralčák (Filozofická fakulta Univerzity Konštantína Filozofa v Nitre; *Dictionary of Štúr's Slovak*) informoval o projekte tvorby Slovníka štúrovskej slovenčiny. V prvej etape prípravy sa pracuje na zhromažďovaní a elektronizácii jazykového materiálu, pričom do pramennej základne Slovníka štúrovskej slovenčiny sú zaradené periodiká (Slovenskje pohladi na vedi, umeňja a literatúru, Slovenskje národňje novini, ich literárna príloha Orol Tatránski, almanach Ňitra) a autorské tlače Ľ. Štúra, J. Franciscího, M. M. Hodžu, J. Kadavého, J. M. Hurbana, J. Dohnányho a A. Sládkoviča. Cieľom ďalšej etapy výskumu je lexikografická analýza získaného materiálu a zostavenie jadra Slovníka štúrovskej slovenčiny.

Natalia P. Darčuk a Viktor M. Sorokin (Kyjevská národná univerzita Tarasa Ševčenka v Kyjeve; *Text-Oriented Thesaurus Retrieval System for Linguistics*) predstavili

projekt tezauru a vyhľadávacieho systému, ktorého cieľom je zostavenie elektronického slovníka jazykovedných termínov. Vychádzali z najnovšej metodológie zostavovania tezauru tak, aby spojili štandardný opis termínov s ich prezentáciou na internete. Zároveň je ambíciou projektu preskúmanie teoretického modelu tezauru na zdokonalenie počítačových technológií používaných na analýzu textov z rozličných oblastí lingvistiky.

Patrice Pognan (INALCO LALIC-Certal v Paríži) vystúpil s príspevkom *Database Creation Based on a Dictionary Using a Corpus Method* v spoluautorstve s Miloud Taïfi (Univerzita Sidi Mohameda ben Abdalla vo Feze). Predstavil budovanie automatickej databázy na základe slovníka berberčiny stredného Maroka, ktorý obsahuje asi 7 200 koreňov. Databáza umožňuje automatické vyhľadávanie slovníkových štruktúr a rozširovanie týchto štruktúr o morfologické údaje. Jej súčasťou je aj vyše 13 tisíc autentických dokladov z tohto jazyka, ktoré sa stali materiálovým podkladom na vytvorenie korpusu berberčiny. Cieľom autorov projektu je vytvoriť francúzsko-berberský slovník verejne prístupný na internete.

Tvorbe databáz a nástrojom na prácu s nimi sa venovali viacerí autori. Ján Genči (Fakulta elektrotechniky a informatiky Technickej univerzity v Košiciach; *Synset Building Based on Online Resources*) opísal známe projekty WordNet a EuroWordNet, ktoré sú zamerané na vybudovanie tzv. synsetov (súborov synoným). Cieľom projektu WordNet bolo vybudovať anglickú databázu, projekt EuroWordNet je nasmerovaný na vytvorenie „národných" verzií WordNet-ových databáz zahŕňajúcich tzv. „medzijazykový index" pre sedem európskych jazykov. Slovenský jazyk nebol zahrnutý do tohto projektu, preto autor prichádza s návrhom niekoľkých prístupov využívajúcich hlavne zdroje prístupné na internete, ktoré by mohli byť použité na vytvorenie relevantných slovenských synsetov a ich pripojenie k Interlingua Index-u.

Marek Grác (Fakulta informatiky Masarykovej univerzity v Brne) v príspevku *Shallow Ontology Based on VerbaLex* dokázal, že ontológie môžu byť použiteľnými zdrojmi pri spracovaní prirodzených jazykov. Predstavil základnú ideu povrchovej ontológie pod názvom Sholva, ktorá je založená na VerbaLex-e, databáze slovesných valencií, kde každý valenčný ukazovateľ obsahuje tiež ukazovateľ EuroWordnet-u. Skúmal náročnosť vytvorenia ontológií, ktoré môžu pomôcť pri riešení reálnych problémov v syntaktických analýzach, dezambiguácii slov a strojovom preklade.

Ľubov V. Struhanec (Filologická fakulta Ternopoľskej národnej pedagogickej univerzity v Ternopoli) v prednáške *Representation of Changes of World Stock of the Ukrainian Literary Language of the 20th – Beginning of the 21st Century in the Information System „Lexika"* predstavil počítačový program, ktorý zahŕňa všetky relevantné ukrajinské slovníky vydané za posledné storočie (slovníky výkladové, prekladové, slovníky cudzích slov a neologizmov). Program *Lexika* umožňuje porovnávať registre analyzovaných slovníkov, vyhľadávať informácie o chronologických parametroch lexém, systematizovať informácie o vybranom slove a jeho fungovaní v texte, robiť výber jednotiek podľa ľubovoľnej diferenciačnej vlastnosti a pod.

Pavlo J. Hrycenko, Olena B. Siruk a Viktor M. Sorokin (Laboratórium počítačovej lingvistiky Filologického inštitútu Kyjevskej národnej univerzity Tarasa Ševčenka v Kyjeve) v príspevku *Electronic Lexical Card Index for the Ukrainian Dialects (ELCIUD)* prezentovali elektronickú lexikálnu kartotéku ukrajinských nárečí ELCIUD. Zaoberali sa zásadnými problémami, ktoré sa vyskytli pri zostavovaní kartotéky (vyhľadávanie optimálnej

štruktúry databázy a elektronických nárečových kartotečných lístkov, zachovanie možnosti editovania všetkých lingvistických informácií a i.) a eventuálnymi možnosťami rozloženia kartotéky. Perspektívne sa ELCIUD bude transformovať na počítačový lexikografický systém.

Irina N e k i p e l o v a (Štátna technická univerzita v Iževsku; *Composite Lexical Units as an Element of Lexicographical Historical Computer System*). sa venovala opisu sémantickej platnosti slov v databáze ruských rukopisných historických dokumentov Manuscript. Na základe sémantickej špecifikácie lexikálnych jednotiek je možné lepšie identifikovať ustálené slovné spojenia databázy. Výsledky tohto výskumu môžu byť podľa autorky aplikované v lexikografii, lexikológii či historickej lingvistike.

Problematike ustálených slovných spojení, resp. kolokácií sa venovali dva príspevky. Annelen B r u n n e r (Inštitút nemeckého jazyka v Mannheime) predniesla príspevok *A Model for Corpus-driven Exploration and Presentation of Multi-word Expressions*, na ktorom spolupracovala s Kathrin S t e y e r. Predstavila metódu lingvistickej interpretácie kolokačných dát, ktorá umožňuje získavanie a vizualizáciu rozmanitých konštrukčných typov viacslovných spojení. Pomocou tejto metódy sa získavajú zovšeobecnenia na niekoľkých úrovniach abstrakcie: povrchová syntaktická štruktúra, lexikálna realizácia, viacslovné pomenovanie a vyabstrahovaný vzorec viacslovného spojenia. Vo svojom výskume autorky vychádzajú z kolokačných profilov slov automaticky vygenerovaných z korpusu nemeckých písaných textov DeReKo, ktorý obsahuje 3 bilióny textových jednotiek.

Marija C h o c h l o v a a Viktor Z a c h a r o v (Filologická fakulta Štátnej univerzity v Petrohrade; *Corpus-based Analysis of Lexico-grammatical Patterns*) sa venovali extrakcii kolokácií z korpusu vytvoreného z listov N. V. Gogoľa. Vyhľadávanie kolokácií funguje na základe štatistických metód kombinovaných so syntaktickými modelmi pomocou tzv. nástroja Sketch Engine, ktorý generuje lingvistické informácie o kolokabilite slova, generuje zoznamy slov, s ktorými sa bázové slovo vyskytuje, a tým pomáha odhaliť ich vzájomné vzťahy. Pomocou tejto metódy je možné poznávať autorov jazyk a vytvárať špecifické slovníky autorovej reči.

Ďalšie príspevky boli zamerané na gramatický výskum na korpusovej báze. Łukasz J ę d r z e j o w s k i (Interdisciplinárneho centra európskych jazykov Slobodnej univerzity v Berlíne) vychádzal v príspevku *On Epistemicity, Grammatical Person and Speaker Deixis in Polish (Based on the Polish Nationaly Corpus)* zo skúšobných dát Poľského národného korpusu, na ktorom ukázal, ako je epistemológia textu závislá od informácií o gramatickej osobe a deiktikách. Korpusové dáta umožňujú preveriť také prípady epistemického kódovania, ktoré sa používajú zriedkavo. Výsledok výskumu okrem iného ukázal, že modálne pomocné sloveso *musieć* môže byť charakterizované nielen deontologicky, ale tiež epistemologicky.

Problematike vetných operátorov sa venovala Jana K e s s e l o v á (Katedra knižničných a slovakistických štúdií Fakulty humanitných a prírodných vied Prešovskej univerzity v Prešove) v príspevku *Operators for Extending and Developing an Utterance*. Autorka rámcovo rozlišuje dva typy operátorov ako prostriedkov na usúvzťažňovanie textových jednotiek: operátory na rozširovanie výpovede a operátory na rozvíjanie výpovede. V príspevku sa zamerala na operátory prípustkového vzťahu, ktoré považuje za prechodný typ. Empirickým zdrojom na lingvistickú analýzu jej bol Slovenský národný korpus.

O vzťahu frekvencie a funkcie anaforika *takový* uvažoval vo svojom referáte *Frequency, Function and Grammaticality: The Case of the Anaphoric takový* František Š t í c h a (Ústav

pro jazyk český Akadémie vied Českej republiky v Prahe). Anaforikum skúmal na báze veľkých písaných korpusov dostupných v rámci Českého národného korpusu a zamýšľal sa nad možnosťami jeho slovníkového spracovania. Pozornosť venoval aj neanaforickému – modálnemu použitiu slova *takový*. Za zaujímavú považuje perspektívnu frekvenčnú analýzu slova *takový* v jeho anaforickom a neanaforickom použití, ktorá bude tiež príspevkom k skúmaniu pravidelností v jazyku. Jazyk ich skrýva v miliónoch viet dostupných vo veľkých korpusoch, ktoré pomáhajú pri odhaľovaní jeho skrytých zákutí.

Ružena K o z m o v á (Filozofická fakulta Univerzity sv. Cyrila a Metoda v Trnave; *Methodological Foundations for Contrastive Model of Verb Valence*) sa venovala rozličným metodologickým východiskám slovesnej valencie v nemčine a slovenčine. Z porovnania pojmu valencie v slovenčine a v nemčine jej vyplynuli všeobecné rozdiely (v teoretickom základe valencie, v teórii slovesnej intencie, v samotnej terminológii, vo vetnej štruktúre), ako aj rozdiely pochádzajúce zo špecifických vlastností daných jazykov. V záverečnej časti príspevku analyzovala vetné modely slovesa *sagen*, ktoré korešponduje so slovenskými ekvivalentmi – dokonavým slovesom *povedať* a nedokonavým slovesom *hovoriť*.

Marie M i k u l o v á a Jan Š t e p á n e k (Ústav formálnej a aplikovanej lingvistiky Matematicko-fyzikálnej fakulty Karlovej univerzity v Prahe; *Annotation Procedure in Building the Prague Czech-English Dependency Treebank*) predstavili organizačné aspekty budovania veľkého syntakticky označkovaného korpusu PDT 2.0 a jeho lingvistickú analýzu. Rozlíšili niekoľko fáz spracovania anotácií, ich automatizovanú kontrolu a vyhodnocovanie úspešnosti anotátorskej práce v závislosti od množstva spracovaných dát i vyhodnocovanie úspešnosti medzi anotátormi navzájom.

Zdeňka U r e š o v á a Petr P a j a s (Ústav formálnej a aplikovanej lingvistiky Matematicko-fyzikálnej fakulty Karlovej univerzity v Prahe) sa vo svojom príspevku *Diatheses in the Czech Valency Lexicon PDT-Vallex* zamerali na formálne vyjadrenie diatézy. Ak je konkrétne sloveso použité v sekundárnej diatéze, jeho povrchové vyjadrenie sa mení. Základné formy tejto transpozície sú známe, avšak autori sa zameriavajú na ďalšie možné spôsoby opisu tohto javu s cieľom využívať dané transformácie na nasledujúce počítačové spracovanie. Príspevok sa zaoberal práve zmenami povrchových foriem v súvislosti s anotáciou Pražského závislostného korpusu a budovaním slovníka PDT-Vallex. Súčasťou príspevku je aj súbor pravidiel na opis jednotlivých pozícií. Pravidlá slúžili na formálnu kontrolu v projekte Pražského závislostného korpusu.

Markéta L o p a t k o v á a Václava K e t t n e r o v á (Ústav formálnej a aplikovanej lingvistiky Karlovej univerzity v Prahe) venovali v príspevku *Changes in Valency Structures of Verbs: Grammar vs. Lexicon* pozornosť zmenám vo valenčnej štruktúre českých slovies z lexikografického hľadiska. Rozlíšili dva typy zmien vo valenčnej štruktúre: gramatickú a sémantickú diatézu. Predstavili základnú typológiu možných zmien vo valenčnej štruktúre a navrhli spôsob reprezentácie týchto zmien vo valenčnom slovníku českých slovies VALLEX.

Magda Š e v č í k o v á (Ústav formálnej a aplikovanej lingvistiky Matematicko-fyzikálnej fakulty Karlovej univerzity v Prahe; *The Meaning of the Conditional Mood Within the Tectogrammatical Annotation of Prague Dependency Treebank 2.0*) zhodnotila súčasný pohľad na podmieňovací spôsob v češtine v jeho základnej funkcii – použitie kondicionálu v hypotetickej výpovedi. Sémantické vyjadrenie kondicionálu ilustrovala autorka na materiáli českých korpusov PDT 2.0 a SYN2005. V príspevku informuje o potrebe zaviesť hodnotu

kondicionálu aj do tektogramatickej roviny korpusu PDT 2.0; opisuje tri hodnoty gramatémy factmod ako možnosti zachytenia hypotetickej a nehypotetickej udalosti a opisuje tiež rozdiely medzi dvoma hypotetickými udalosťami. Pravidlá na ručné označkovanie týchto hodnôt uvádza autorka v článku. Perspektívne počíta s poloautomatickým prideľovaním týchto a ďalších gramatém opisujúcich kategóriu kondicionálu na tektogramatickej rovine v PDT 2.0.

Oľha O. Šypnivska (Kyjevská národná univerzita Tarasa Ševčenka v Kyjeve) a Serhij V. Starykov (Inštitút jazykovo-informačného fondu Národnej akadémie vied Ukrajiny v Kyjeve) sa v príspevku *The Creation of the Morphological Ambiguity Depository in Ukrainian* zaoberali morfologickými homonymami a ich dezambiguáciou podľa významu. Vo výskume sa pracovalo s množstvom lingvistických databáz morfologických homoným. V štúdii sa podáva konspektový opis formálnych klasifikácií ukrajinských morfologických homoným podľa odlišných typov vzorov, frekvenčnej závislosti týchto typov od množstva komponentov a gramatických vlastností. Štruktúra a princípy elektronického slovníka morfologických homoným sa prezentujú v slovníku tvarov slov a ilustrujú štylisticky diferencovanými textami korpusu.

Výsledky štatistického výskumu textov Slovenského národného korpusu prezentovali Mária Šimková a Miroslav Ľos (Jazykovedný ústav Ľ. Štúra Slovenskej akadémie vied v Bratislave). V príspevku *Frequency of Words and Their Forms in Contemporary Slovak Language Based on the Slovak National Corpus* skúmali autori absolútnu frekvenciu slov, tvarov slov a n-gramov, a to aj vzhľadom na jednotlivé štýly utvárajúce písané korpusy v Slovenskom národnom korpuse. Výsledky frekvenčných analýz autori porovnali s frekvenčnými výskumami v 70. rokoch minulého storočia a čiastočne aj s prvými verziami Slovenského hovoreného korpusu.

Dough Davidson (Inštitút Maxa Plancka pre kognitívne vedy v Lipsku) a Adriana Hanulíková (Inštitútu Maxa Plancka pre psycholingvistiku v Nijmegene) v príspevku *Inflectional Entropy in Slovak* vysvetlili, že štatistické miery frekvencií slov sú dôležitým údajom v psycholingvistických výskumoch, pomocou nich môžeme lepšie porozumieť psychologickému rozloženiu mentálnej lexiky a procesom vyberania, porozumenia a učenia sa slov. Zaoberali sa jednou zo štatistických mier – entropiou ohýbania slov, ktorej veľkosť závisí od množstva informácií prinášaných formou slova po ohýbaní a od štatistickej distribúcie paradigmy daného slova. Skúmali mieru entropie ohýbania v slovenčine a výsledky porovnávali s výsledkami pre srbčinu. Ako zdroj výskumu používali Slovenský národný korpus.

Softvérové nástroje na automatické testovanie derivácie slov v češtine predstavili Dana Hlaváčková, Karel Pala a Pavel Šmerk (Fakulta informatiky Masarykovej univerzity v Brne) a Klára Osolsobě (Filozofická fakulta Masarykovej univerzity v Brne) v príspevku *Exploring Derivational Relations in Czech with the Deriv Tool*. Niektoré typy odvodzovania v češtine majú vysokú mieru pravidelnosti a môžeme ich opísať formálnymi pravidlami. Nová verzia internetového rozhrania Deriv pracuje s morfologickým analyzátorom ajka, čo umožňuje lepšie formulovať komplexné pravidlá derivácie slov a otestovať viac komplikovaných prípadov derivačných vzťahov. Autori v rámci sémantickej klasifikácie navrhli 14 sémantických tried pre sufixy a 11 pre prefixy. Pravidlá rozhrania Deriv značne pomáhajú pri zisťovaní sémantiky derivačných vzťahov.

Petr Pořízka (Filozofická fakulta Univerzity Palackého v Olomouci) a Markus Schäfer (Matematicko-prírodovedecká fakulta Rýnskej univerzity Friedricha-Wilhelma

v Bone) v štúdii *Morph-Con – A Software for Conversion of Czech Morphological Tagsets* informovali o novom softvérovom nástroji MorpCon (Morphological Convertor), ktorý bol vyvinutý na konverziu medzi českými morfologickými tagsetmi. Jeho prvá verzia 0.1alpha umožňuje konverziu medzi pražským pozičným systémom a brnianskym atributívnym systémom. Pracuje s troma základnými formátmi dát: SimpleTag-Conversion, KWIC/Tag-Format, WPL-Format. MorphCon softvér je navrhnutý ako univerzálny konvertor: pomocou nových modulov umožňuje pridávanie ďalších tagsetov (nielen českých) a dátových formátov.

Radovan G a r a b í k (Jazykovedný ústav Ľ. Štúra Slovenskej akadémie vied v Bratislave) prezentoval za spolupracovníkov Jana Jonu J a v o r š e k a a Tomaža E r j a v c a (Inštitút Jožefa Stefana v Ľubľane) príspevok *Evaluating Grid Infrastrucure for Natural Language Processing*, v ktorom zhodnotil skúsenosti s programami na počítačové spracovanie prirodzených jazykov (NLP) použitými v prostredí Gridu. Cieľom autorov bolo spustenie softvéru pre systém automatického počítačového prekladu TectoMT a morfologické značkovanie textov Slovenského národného korpusu. Prístup sa ukázal ako vhodný pre NLP projekty náročné na výpočtový výkon. Autori navrhujú v danej problematike pokračovať v rámci virtuálnej organizácie (VO) zameranej na technológie spracovania jazykov (HLT).

Prehľadové príspevky o práci svojich pracovísk predniesli Adam Przepiórkowski a Ludmila Dimitrova. Adam P r z e p i ó r k o w s k i (Ústav informatiky Poľskej akadémie vied vo Varšave) za svojich kolegov Rafała L. G ó r s k e h o, Marka Ł a z i ń s k e h o a Piotra P ę z i k a predstavil v príspevku *Recent Developments in the National Corpus of Polish* najnovší výskum v rámci projektu Poľského národného korpusu. Zameral sa na predstavenie výskumu týkajúceho sa viacerých oblastí zdokonaľovania korpusu: vytvorenie textového kódovania XML schémy pre rôzne stupne lingvistických informácií, vytvorenie nástroja na manuálnu morfosyntaktickú anotáciu textov, ako aj zdokonalenie množstva vyhľadávacích korpusových nástrojov.

Príspevok *From Electronic Corpora to Online Dictionaries (on the Example of Bulgarian Language Resources)* Ludmily D i m i t r o v e j (Inštitút matematiky a informatiky Bulharskej akadémie vied v Sofii) informoval o projektoch inštitútu v spolupráci s inými medzinárodnými pracoviskami v rámci medzinárodných grantov, akými sú elektronický bulharský korpus, bulharská lexikálna databáza či dvojjazyčný bulharsko-poľský elektronický slovník. Rovnako sa zmienil o perspektívach počítačového výskumu: morfosyntaktické anotácie slovies, rozšírenie dvojjazyčných korpusov či nové užívateľské aplikácie v bulharsko-poľskom slovníku.

Slová slávnostného poďakovania za príjemnú vedeckú a diskusnú atmosféru počas celého podujatia, za prezentáciu významných výsledkov z oblasti korpusovej lingvistisky a gramatického výskumu na báze korpusov predniesol za organizátorov konferencie Radovan Garabík z oddelenia Slovenského národného korpusu Jazykovedného ústavu Ľ. Štúra Slovenskej akadémie vied v Bratislave. Prítomných pozval o dva roky znovu na Slovensko, na SLOVKO 2011.

Katarína Gajdošová – Agáta Karčová – Daniela Majchráková