SLOVKO 2015
POČÍTAČOVÉ SPRACOVANIE PRIRODZENÉHO JAZYKA, KORPUSOVÁ LINGVISTIKA, LEXIKOGRAFIA

Oddelenie Slovenského národného korpusu Jazykovedného ústavu Ľ. Štúra Slovenskej akadémie vied v Bratislave (ďalej SNK JÚĽŠ SAV) zorganizovalo v dňoch 21. – 22. októbra 2015 8. ročník medzinárodnej konferencie SLOVKO 2015 s podnázvom *Natural Language Processing, Corpus Linguistics, Lexicography – Počítačové spracovanie prirodzeného jazyka, korpusová lingvistika, lexikografia.* Podujatie sa konalo v priestoroch Centra vedecko-technických informácií SR (ďalej len CVTI) na Lamačskej ceste v Bratislave. Konferenciu slávnostným príhovorom otvorili Ján Turňa, riaditeľ CVTI SR, Ľuboš Svetoň, hlavný štátny radca odboru implementácie štátnej politiky, európskych a medzinárodných iniciatív výskumu a vývoja Ministerstva školstva, vedy, výskumu a športu SR, a Jana Levická za organizátorov.

Na konferencii SLOVKO 2015 odznelo 16 príspevkov od 19 autorov, väčšina prezentácií (9) bola prednesená v českom jazyku, 4 v slovenskom a 3 v anglickom jazyku. Priestor na diskusiu využili mnohí účastníci konferencie a v reakciách na jednotlivé príspevky tak odznelo viacero podnetných názorov a informácií. V zborníku tlačených príspevkov *Natural Language Processing, Corpus Linguistics, Lexicography* (Eds. K. Gajdošová – A. Žáková. Lüdenscheid: RAM-Verlag 2015. 170 p. ISBN 978-3-942303-32-3), ktorý mali účastníci k dispozícii už počas konferencie, je uverejnených 15 príspevkov v anglickom jazyku. Pristavíme sa krátko pri každom z nich, hoci jeden publikovaný príspevok nebol prezentovaný na konferencii, a spomenieme aj dve prednášky, ktoré odzneli v rámci programu vedeckého podujatia bez publikovania (V. Cvrček, A. Karčová). Príspevky radíme podľa tematických okruhov.

Počas prvého dňa medzinárodnej konferencie SLOVKO 2015 pokrylo 7 príspevkov 3 tematické okruhy: tvorba slovníkov na báze korpusových dát, korpusové nástroje a korpusové dáta slúžiace na výskum jazykových javov.

V októbri 2014 bol spustený nový slovinský portál Fran, ktorý zahŕňa rôzne typy slovníkov (všeobecné, historické, terminologické, nárečové), jazykové atlasy, slovinskú terminológiu, jazykové poradenstvo, ako aj odkazy na korpusy. Kozma Ahačič, Nina Ledinek a Andrej Perdih (Vedecko-výskumné stredisko SAZU, Ľubľana) ho predstavili v príspevku *Fran: Nová generácia slovníkového portálu slovinčiny*. Podľa slov A. Perdiha, ktorý o portáli referoval, jeho výhodou je jednoduchá navigácia, pričom pokročilý používateľ si môže zvoliť zložitejšie parametre vyhľadávania. V čase konferencie boli v procese príprav etymologický a frazeologický slovník, ktoré sú už toho času na stránke zverejnené. V blízkej budúcnosti by mali pribudnúť slovník štandardnej slovinčiny zo 16. storočia, nový slovník hovorovej slovinčiny a niekoľko nárečových slovníkov. Používateľ portálu Fran má k dispozícii množstvo informácií o vývoji a používaní slovinského jazyka od 16. storočia až po súčasnosť. Odozvou na podnet z diskusie k príspevku je anglická verzia portálu, ktorá bola spustená v apríli tohto roku. Tvorbe slovníkov na báze korpusových dát sa venujú aj na lexikografickom pracovisku v Prahe. Pre potreby tvorby *Akademického slovníka súčasnej češtiny* vyvíjajú pracovníci oddelenia súčasnej lexikológie a lexikografie Ústavu pre jazyk český Akadémie vied ČR podporný software ALEXIS, ktorý pred dvomi rokmi na konferencii SLOVKO 2013 predstavili ako novinku K. Barbierik a T. Liška. S odstupom času informovali o tom, ako sa za uplynulé ob-

dobie tento nástroj vyvinul, aké sú jeho nové možnosti, a načrtli jeho ďalší vývoj. V príspevku *DWS ALEXIS a jeho nové funkcie,* ktorý pripravil kolektív pracovníkov Kamil Barbierik, Martin Bodlák, Zuzana Děngeová, Vladimír Jarý, Tomáš Liška, Michaela Lišková, Josef Nový a Miroslav Virius (Akadémia vied ČR, Praha), poukazujú autori napríklad na možnosti editovania či funkcie webového rozhrania. V priebehu dvoch rokov sa menil aj lexikálny materiál, čo taktiež ovplyvnilo vývoj softvéru. Na slovníku prebiehajú práce on-line, pričom všetky zásahy ostávajú zaznamenané, čo prácu urýchľuje a zefektívňuje. V blízkej budúcnosti chcú pracovníci oddelenia sprístupniť softvér pre verejnosť a editačné práce tak budú viditeľné priamo v procese tvorby slovníka.

Výrazné tematické zastúpenie mali korpusové a webové nástroje, ktoré prezentovali traja pracovníci Ústavu Českého národného korpusu (ďalej ÚČNK). David Lukeš (Univerzita Karlova, Praha) predstavil dva webové nástroje AchSynku a MluvKonk, ktoré vznikli s cieľom uľahčiť prácu lingvistom a používateľom korpusov ORAL. V príspevku *Nové nástroje na prácu s korpusmi hovorenej češtiny radu ORAL: AchSynku a MluvKonk* poukazuje autor na nové možnosti, ktoré tieto dva nástroje ponúkajú. Nástroj AchSynku má doplniť chýbajúcu lematizáciu v hovorených korpusoch a možnosť vyhľadávania pomocou lemy. MluvKonk je vizualizačné prostredie, ktorého cieľom je sprehľadniť konkordanciu v nástroji KonText tak, že sa kľúčový výraz rozdelí do viacerých riadkov, pričom jednému hovoriacemu prislúcha jeden riadok. Takýto typ konkordancie pripomína štruktúru dialógu. D. Lukeš zdôraznil, že oba nástroje sú ešte len vo vývoji a spätná väzba je preto veľmi vítaná. Projekt *Universal Dependencies (UD)* sa zameriava na syntakticky anotované korpusy a má pomôcť pri skúmaní vzťahov medzi jednotlivými jazykmi. Daniel Zeman (Univerzita Karlova, Praha) predstavil v prezentácii *Slovanské jazyky v Universal Dependencies* tento nástroj ako vhodný na analýzu vzťahov medzi slovanskými jazykmi (najmä na porovnanie zvláštností zámen, číslic, číslovkových zámen, ako aj modálnych slovies, elíps, nominálnych predikátov a zvratných zámen) pomocou automaticky vygenerovaných syntaktických stromov, ktoré zaznamenávajú syntaktické a morfologické javy jednotlivých viet. Vďaka interlingvisticky konzistentnej anotácii pre viacero jazykov, ktorú autor v rámci tohto projektu vyvíja, sa tiež uľahčí vývoj viacjazyčného syntaktického analyzátora, zlepší sa medzijazykové vzdelávanie a jazykový výskum z perspektívy jazykovej typológie (slovné druhy, morfologicko-syntaktické opisy a syntagmatické vzťahy). Syntaktickej anotácii sa venovala aj Milena Hnátková (Univerzita Karlova, Praha), ktorá v príspevku *Automatická identifikácia druhu príslovkových určení v syntakticky anotovaných textoch* referovala o výsledkoch automatickej identifikácie českých časových adverbiálií vygenerovaných z korpusových dát pomocou nástroja FRANTA. Cieľom výskumu bolo vytvoriť zoznam temporálnych adverbiálií pomocou ustálených frekventovaných kolokácií. M. Hnátková oboznámila účastníkov konferencie s možnosťami vyhľadávania a výsledkami výskumu. Pomocou nástroja FRANTA a následnej ručnej úpravy dospeli pracovníci ÚČNK k novému deleniu časových adverbiálií. M. Hnátková zdôraznila, že napriek vysokým kvalitám korpusového nástroja FRANTA je stále potrebná ručná úprava výsledkov.

Skutočnosť, že korpusy sú odrazom živého jazyka a obsahujú obrovské množstvo dát, ktoré ponúkajú nespočetne veľa možností na výskum rôznych jazykových javov, demonštrovala aj Petra Poukarová (Univerzita Karlova, Praha) v príspevku *Slovné tvary a funkcie verba „myslet" v hovorenej češtine*. Pomocou dát z korpusu ORAL skúmala formy a funkcie slovesa *myslet,* jeho používanie v spontánnej hovorenej češtine, najmä tvary *já myslím, mys-*

lím, myslím, že... Po porovnaní prvého významu tohto slovesa zo *Slovníka spisovného jazyka českého* s výsledkami v korpuse ORAL prišla k zaujímavému záveru, že tento význam nemá v spontánnej reči takú prevahu, ako by sa očakávalo. Venovala sa aj porovnaniu výskytov spomínaných foriem v korpuse písaných textov SYN2000. Výsledkom výskumu je vízia využitia korpusov na prehodnotenie významov jednotlivých slov.

O tom, že niektoré zdroje Slovenského národného korpusu pomáhajú zvyšovať kvalitu vzdelávania na Slovensku, sa mohli účastníci presvedčiť vďaka prezentácii Júliusa Kravjara (Centrum vedecko-technických informácií SR, Bratislava) o slovenskom antiplagiátorskom systéme pod názvom SK ANTIPLAG: po piatich rokoch. J. Kravjar zhrnul výsledky a úspechy tohto systému a poukázal na veľmi pozitívne odozvy zo zahraničia. Prezentoval nevyhnutnosť projektu z viacerých hľadísk a príčin, ktoré majú vplyv na úroveň vzdelávania (ľahký prístup k informáciám a pod.). Pozitívne zhodnotil dôsledky využívania SK ANTIPLAGU, ako napríklad zvýšenie povedomia o plagiátorstve či zlepšenie kvality študentských prác. Veľkou výhodou oproti iným krajinám je, že od roku 2010 musia tento systém na Slovensku používať všetky vysoké školy.

Prvý deň konferencie spríjemnila účastníkom okrem obľúbených kávových prestávok aj prednáška o histórii, fungovaní a úlohách CVTI SR, ako aj prehliadka priestorov, najmä špecializovanej vedeckej knižnice a digitalizačného strediska.

Druhý deň konferencie otvoril Patrice Pognan (Národný inštitút orientálnych jazykov a civilizácií, Paríž) plenárnou prednáškou *Automatické lexikografické spracovanie jazykov s obmedzenými zdrojmi*. Autor predstavil realizáciu projektu najrozsiahlejšieho slovníka berberčiny v Maroku, ktorý vzniká na základe francúzsko-marockej spolupráce a je podporovaný vládnymi inštitúciami. Projekt súčasne vytvára vhodné prostredie aj na automatické spracovanie berberčiny, najmä v oblasti lexikografie, a predstavuje východisko pri tvorbe zdravotníckeho cilubà-francúzskeho slovníka. Výsledkom doterajšieho výskumu je 1223 strán berbersko-francúzskeho slovníka, nahrubo štruktúrovaný korpus, slovné spojenia (z nich bude vytváraná učebnica berberčiny) a 5000 slovies berberčiny.

Václav Cvrček (Univerzita Karlova, Praha) predniesol príspevok *Paradigmatické dotazy: nový typ korpusového dotazovania*, čo je téma, ktorej sa venuje už dlhšiu dobu (prvý príspevok z tejto oblasti publikoval v roku 2007). Ide o javy, ktoré sú v miernom rozpore so základnými prioritami korpusového výskumu, keďže v korpusovej lingvistike prevažuje pohľad syntagmatický. Rozlišuje dva druhy dotazov. Prvým je syntagmatický dotaz (CQL a iný dotazovací jazyk), ktorého výsledkom je množina tokenov bez ohľadu na anotácie. Druhý, paradigmatický dotaz je prienikom kombinácie minimálne dvoch čiastkových syntagmatických dotazov. Výsledkom je množina typov, množina jednotiek bez kontextu, ktoré sú abstrakciou z tokenov. Paradigmatický dotaz zatiaľ nie je implementovaný, je na to potrebná viacúrovňová anotácia, 100-percentná dezambiguácia, interface, výpočtový backhand a databáza anotačných jednotiek, okrem iného aj na efektívny spôsob tohto hľadania. V. Cvrček vyjadril nádej, že paradigmatické dotazovanie v budúcnosti určite poslúži mnohým lingvistom v ich vlastnom korpusovom výskume.

Hovoreným a nárečovým korpusom boli venované tri nasledujúce príspevky piatich autorov.

Hana Goláňová (Univerzita Karlova, Praha) v príspevku *Nárečový korpus DIALEKT* predstavila pripravovaný, na internete verejne prístupný nárečový korpus, metodiku zberu dát

nárečí, sociolingvistické parametre korpusu a dve úrovne prepisu: dialektologický a ortografický. Následne ukázala prípravu základných máp pre korpus a ich začlenenie do interaktívneho webového prostredia určeného na analýzu dát zo všetkých typov hovorených korpusov. Táto interaktívna webová aplikácia bude umožňovať prístup k lingvistickým informáciám z oboch nárečových korpusov a tradičných hovorených korpusov a obsahovať aj užitočné funkcie pre vedeckú komunitu i laikov. Katarína Gajdošová, Radovan Garabík a Mária Šimková (Slovenská akadémia vied, Bratislava) sú autormi príspevku *Korpus nárečí Slovenského národného korpusu* (ďalej KN-SNK), ktorý na konferencii prezentovala M. Šimková. Vo svojej prezentácii podala prehľad o súčasných textových zdrojoch zahrnutých v nárečovom korpuse, o tvorbe KN-SNK a jeho súčasnom stave, o formáte záznamov metadát a priblížila informácie o hovoriacich, konverziu prepisov do jednotného formátu, značkovanie a spôsob vyhľadávania v tomto korpuse v nástroji NoSketch Engine. Na záver načrtla perspektívu ďalších nárečových korpusových zdrojov, ktoré by sa mohli spracovať v SNK. Hovoreným korpusom Českého národného korpusu sa venovala Marie Kopřivová (Univerzita Karlova, Praha) v príspevku *Somatické frazémy v hovorených korpusoch: Evalvácia automatického značkovania frazém v hovorených korpusoch – na príklade somatických frazém*. Označkovanie časti hovoreného korpusu série ORALF (ORAL2006 a ORAL2008) nástrojom FRANTA, založenom na Slovníku českej frazeológie a idiomatiky, predstavila autorka spolu s kolegyňou M. Hnátkovou už na konferencii SLOVKO 2013, aktuálne sa M. Kopřivová rozhodla pozrieť na to, ako bola ich snaha o označkovanie úspešná, a zhodnotiť spoľahlivosť anotácie frazém v hovorených korpusoch. Niektoré české idiómy sa zobrazujú v rôznych dĺžkach a v rôznom slovoslede, tieto vlastnosti výrazne komplikujú ich identifikáciu. Práve somatické idiómy, ľahko vyhľadané pomocou kľúčového slova, sa ukázali ako najvhodnejšie na overenie presnosti anotácie.

Senja Pollak (Inštitút Jozefa Štefana, Ľubľana) v príspevku *Identifikácia kolokácií v korpuse: v hovorenej slovinčine* predstavila metodiku extrakcie a porovnávania slovných spojení v dvoch rôznych korpusoch toho istého jazyka, identifikovala kolokácie, ktoré sú charakteristické pre hovorenú slovinčinu vo vzťahu k referenčnému písanému korpusu slovinčiny. O výhodách a nedostatkoch tohto prístupu, ako aj o možných aplikáciách a zlepšeniach z pohľadu rozsiahlejšieho experimentu, ktorý sa plánuje v budúcnosti, sa naďalej diskutuje.

Aksana Schillova (Univerzita Karlova, Praha), doktorandka v odbore slovanská filológia na FF UK, prezentovala časť svojej dizertačnej práce *České predložky vyskytujúce sa v postpozícii k podstatnému menu (na materiáli korpusu SYN2010)*. V práci sa venuje hľadaniu postpozícií v tomto korpuse a nasledujúcej analýze ich frekvenčných zoznamov, kolokačných zoznamov a konkordancií. Osobitnú pozornosť venuje deadverbiálnym jednotkám *navzdory*, *napospas*, *vstříc*, *naproti*, ktoré sa podľa korpusu pravidelne využívajú nielen ako predložky, ale aj ako postpozície podstatného mena (alebo jeho náhrady). Načrtla problematické otázky a hlavné smery nasledujúceho výskumu českých postpozícií.

Agáta Karčová (Slovenská akadémia vied, Bratislava) sa v prezentácii *Analýza záporových adjektív s derivačným prefixom ne- na korpusovom základe* venovala negácii ako formálno-sémantickej operácii, ktorá sa spravidla uskutočňuje pomocou negačnej morfémy alebo jej ekvivalentov. Podrobnejšie sa zamerala na adjektíva, v ručne morfologicky anotovanom podkorpuse r-mak-4.0 vyselektovala antonymné dvojice adjektív s prefixom *ne-* a ďalšími predponami s negujúcim významom (*bez-*, *de-/dez-*, *proti-*, *anti-* a i.). Po analýze tohto mate-

riálu poukázala na nehomogénnosť tejto skupiny záporových adjektív. Ďalej sledovala pomocou analýzy dát z korpusov r-mak-4.0 a prim-6.1-juls-all záporové adjektíva s prefixom *ne-* a príponou *-teľný*, ktoré nemajú svoj neutrálny pozitívny náprotivok, prípadne je frekvencia antonymického páru rádovo vyššia ako frekvencia nenegovaného tvaru, ktorý preto spravidla nie je zachytený v lexikografických dielach.

Jazyk ponúka mnoho možností na skrátenie dlhších slov alebo fráz. Jednou z možností je použitie akronymu, čo je skrátený tvar viac ako dvoch slov obsahujúci iba ich počiatočné písmená. Zuzana Komrsková (Univerzita Karlova, Praha) v príspevku *Použitie akroným v rôznych komunikačných situáciách (na korpusovom základe)* porovnávala distribúciu najčastejšie používaných akroným, ako sú *BTW, IMHO, LMAO, LOL, OMG, ROFL, WTF*, rozšírených v písaných textoch aj v prirodzenej reči.

V závere medzinárodnej konferencie SLOVKO 2015 sa Mária Šimková, vedúca oddelenia Slovenského národného korpusu JÚĽŠ SAV, poďakovala organizátorkám, všetkým prednášajúcim, zúčastneným, diskutujúcim a vyjadrila nádej, že sa o dva roky opäť stretneme na konferencii SLOVKO 2017, ktorej hlavnou témou by mala byť počítačová terminológia a terminografia.

Katarína Chlpíková – Beáta Kmeťová
Jazykovedný ústav Ľ. Štúra SAV, Bratislava