Znanstvena monografija bo vsekakor zanimivo branje za vse, ki se ukvarjajo s slovenskim slovaropisjem, in tistim, ki želijo svoja spoznanja na tem področju nadgraditi. Priporočam jo tudi študentom in študentkam, ki se zanimajo za obravnavano tematiko, prednosti elektronske izdaje Pleteršnikovega slovarja pa bi morali spoznati že mladi v osnovnih in srednjih šolah ter ga pri učenju tudi uporabljati.

Anja Benko
*Filozofska fakulteta*
*Univerze v Mariboru*
*anja.benko@gmail.com*

## ***IS-LTC 2008: šesta konferenca Jezikovne tehnologije*, Ljubljana, 16.–17. oktober 2008**

Letošnja konferenca o jezikovnih tehnologijah, ki je kot del multikonference *Informacijska družba 2008* potekala na Inštitutu Jožef Stefan, je poleg 20 prispevkov obeležila 10. obletnico delovanja Slovenskega društva za jezikovne tehnologije (SDJT). Potek konference sta spremljala in koordinirala **Tomaž Erjavec** in **Jerneja Žganec Gros**, ki sta tudi poskrbela, da so udeleženci ob prihodu na konferenco dobili zbornik s prispevki. Predavanja so bila razporejena po sekcijah – prvi dan je bil namenjen predvsem predstavitvam znanstvenih spoznanj in raziskovalnih dosežkov strokovnjakov z univerz in institucij zunaj meja Slovenije, drugi dan pa so si znanstvenoraziskovalna spoznanja izmenjali slovenski raziskovalci in strokovnjaki.

Konferenca in **sekcija A** sta se začeli s predavanjem vabljenega gosta z Univerze v Trentu **Marca Baronija**. Po kratkem pregledu razvoja korpusne semantike in semantičnih modelov je predstavil svojo vizijo oz. možnost prihodnjega razvoja korpusne semantike v obliki semantičnega modela StruDEL. Ta model naj bi na osnovi shranjenih korpusnih podatkov o frekvenci, oblikoslovnih in skladenjskih oznakah, sintagmatski povezanosti besed in načinu povezave besed oz. besednih zvez hkrati izvajal več semantičnih metanalog kot predstopenj za ustvarjanje leksikalnih in leksikografskih virov: opredelil bi lahko pomen besed oz. besednih zvez v besedilnem kontekstu, v katerem so se pojavile, našel med seboj pomensko in smiselno povezane besede in besedne zveze ter opredelil pomenske odnose med njimi (npr. *pes : žival*; pomenski odnos *oskar : film* in *grammy : pesem*), na osnovi danih besed oz. parov besed pa bi lahko izpeljeval tudi različne možne sintagmatske kombinacije.

V **sekciji B** so bili združeni prispevki, povezani z morfološkim označevanjem, berljivostjo in strojnim prevajanjem. **Jernej Vičič** z Univerze na Primorskem je predstavil prevajalni sistem za strojno prevajanje sorodnih morfološko bogatih jezikov kot možen pripomoček za gospodarsko sodelovanje med

Slovenijo in Srbijo. Opozoril je na prednosti in pomanjkljivosti takšnih popolnoma avtomatiziranih prevodov in uporabljenih metod ter orisal možnosti nadaljnjega izpopolnjevanja tovrstnih prevajalnih sistemov za morfološko bogate jezike. **Damir Ćavar** z Univerze v Zadru, **Ivo-Pavao Jazbec** in **Siniša Runjaić** z zagrebškega Inštituta za hrvaški jezik in jezikoslovje so sodelovali s prispevkom o morfološkem razčlenjevalniku, označevalniku in lematizatorju za hrvaški knjižni jezik CroMo, ki naj bi dal podatke o morfoloških (predpone, pripone, koreni), morfosintaktičnih (sklon, število) in morfosemantičnih (vid, čas) značilnostih jezika. Predstavili so CroMovo zasnovo in razvoj, pri čemer so izpostavili označevalno shemo na podlagi ontologije GOLD (*General Ontology for Language Description*) in algoritem, po katerem CroMo deluje. Opisani morfološki analizator naj bi bilo mogoče uporabiti za katerikoli morfološko bogat jezik, tudi slovenščino. **Tim vor der Brück**, **Sven Hartrumpf** in **Hermann Helbig** s Fernove univerze v Hagnu so predstavili rezultate raziskovalnega dela na področju računalniškega preverjanja berljivosti nemških besedil. Tovrstno preverjanje običajno deluje na podlagi berljivostnih formul, oblikovanih na osnovi površinskih podatkov o pogostosti uporabljenih besed in dolžini stavkov (npr. berljivostna formula *Flesch Reading Ease*). Avtorji so predstavili preverjevalnik berljivosti DeLite, ki ne deluje samo na osnovi analize površinskih kazalcev, ampak tudi na osnovi avtomatske globinske skladenjsko-semantične analize kazalcev berljivosti. Uporabniki dobijo podatke o težje berljivih delih besedila in globalno oceno berljivosti, na podlagi katerih lahko oblikujejo bolj berljiv tekst. Preverjevalnik se je izkazal za orodje, ki za nemška besedila obeta dobre rezultate.

**Sekcija C** se je začela z vabljenim predavanjem **Tanje Schultz** z Univerze v Karlsruheju in Univerze Carnegie Mellon, ki je spregovorila o razvoju sistemov za govorno procesiranje in o različnih področjih njihove uporabe, kot so na primer dialoški sistemi, povzemanje govora, sistemi za priklic informacij in avtomatsko prevajanje govora. Opozorila je na nekaj splošnih tehnoloških in tehničnih zadreg, povezanih s pomanjkanjem podatkov, z jezikovnimi konvencijami in nesodelovanjem med jezikoslovci in računalniškimi strokovnjaki, na katere bi utegnili naleteti jeziki, ki tovrstnih sistemov še nimajo oblikovanih. V nadaljevanju je predstavila orodje SPICE, ki je sicer nastalo v angleškem jezikovnem okolju, vendar naj bi ga bilo mogoče hitro prilagoditi tudi značilnostim ostalih jezikov. Tudi sledeče predavanje v tej sekciji je bilo povezano z govornim procesiranjem. V okviru interdisciplinarnega projekta AvID je bila v sodelovanju Fakultete za elektrotehniko (**Rok Gajšek**, **Vitomir Štruc** in **France Mihelič**) in Oddelka za psihologijo Filozofske fakultete UL (**Anja Podlesek**, **Luka Komidar**, **Gregor Sočan**, **Boštjan Bajec** in **Valentin Bucik**) zasnovana emocionalna govorna podatkovna baza. Avtorji so opozorili predvsem na metodološke vidike zbiranja in snemanja materiala ter poudarili, da se baza oblikuje z namenom omogočanja preverjanja identitete govorcev ter opredeljevanja govorčevih spontanih

psihofizičnih stanj v določenih realnih situacijah. To so upoštevali tudi pri oblikovanju testnih pogojev za eksperimente, ki so med drugim vključevali igranje računalniških iger in adaptivne inteligenčne teste.

V **sekciji D** je bil združen sklop prispevkov o oblikoslovnem označevanju korpusov. **Primož Jakopin** in **Aleksandra Bizjak** z ZRC SAZU sta opisala stanje in dosedanje dosežke na področju oblikoslovnega označevanja v slovenskem prostoru, pri čemer sta se podrobneje osredotočila na potek in način oblikoslovnega označevanja 1,3-milijonskega besedilnega korpusa *Nova beseda*, ki je nastal na Inštitutu za slovenski jezik Frana Ramovša. **Jan Rupnik**, **Miha Grčar** in **Tomaž Erjavec** z Inštituta Jožef Stefan so predstavili rezultate poskusa izboljšanja točnosti oblikoslovnega označevanja slovenskih besedil, ki so jih dobili tako, da so združili oblikoslovni označevalnik podjetja Amebis, ki deluje na podlagi ročno zgrajenih pravil, in označevalnik TnT, ki deluje na podlagi strojnega učenja. Takšen metaoznačevalnik se je zaradi upoštevanja besedilnega konteksta izkazal za točnejšega pri dodeljevanju končnih oznak in predstavlja možnost prihodnjih izboljšav na področju oblikoslovnega označevanja. Predavanje **Željka Agića**, **Marka Tadića** in **Zdravka Dovedana** z Univerze v Zagrebu se je do neke mere dopolnjevalo s slovenskimi rezultati na področju oblikoslovnega označevanja besedil. Avtorji so namreč predstavili nov označevalnik, ki so ga razvili z namenom izboljšati natančnost oblikoslovnega označevanja hrvaških besedil, deluje pa na podlagi hrvaškega morfološkega leksikona. Rezultate so primerjali z rezultati, ki so jih dobili z uporabo statističnega označevalnika TnT, in nakazali smernice za nadaljnje delo na tem področju.

Dopoldanska **sekcija E** drugega dne konference je prinašala predvsem prispevke s področja govornih tehnologij. **Jerneja Žganec Gros**, **Aleš Mihelič**, **Mario Žganec** in **Uliana Dorofeeva** z Alpineona R&D ter **Nikola Pavešić** z Univerze v Ljubljani so se ukvarjali s postopkom za izbiro govornih segmentov pri vgrajeni polifonski združevalni sintezi govora in ugotovili, da je način povečevanja hitrosti postopka, ki so ga predlagali, primeren predvsem za vgrajene sintetizatorje govora. **Andrej Žgank**, **Marko Kos**, **Bojan Kotnik**, **Mirjam Sepesy Maučec**, **Tomaž Rotovnik** in **Zdravko Kačič** z Univerze v Mariboru so s spremembami na področju akustične segmentacije, izločanja značilk in akustičnega modeliranja uspešno izboljšali delovanje trenutno najkompleksnejšega razpoznavalnika slovenskega tekočega govora UMB Broadcast News. **Matej Grašič**, **Marko Kos** in **Zdravko Kačič** z Univerze v Mariboru so predstavili vpliv predhodne segmentacije govor/negovor na pravilno zaznavo menjave govorca, ki je potrebna v sistemu avtomatskega razpoznavanja govora. Metodo so preizkusili na slovenski govorni bazi BNSI Broadcast News. Na diskurzne označevalce so se osredotočili **Darinka Verdonik** in **Andrej Žgank** z Univerze v Mariboru ter **Agnes Pisanski Peterlin** z Univerze v Ljubljani, ki so analizirali, v kolikšni meri so rezultati korpusne analize diskurznih označevalcev odvisni od interpretacije označevalca gradiva in kako natančna je uporabljena shema za

označevanje diskurznih označevalcev v slovenščini. **Kristina Hmeljak Sangawa** z Univerze v Ljubljani in **Tomaž Erjavec** z Inštituta Jožef Stefan sta ponudila spletni japonsko-slovenski slovar kot model za učinkovito ustvarjanje referenčnega učnega gradiva z uporabo prosto dostopnih orodij in besedil, in sicer vzporednega korpusa, ki nastaja v okviru vaj iz prevajanja med slovenščino in japonščino, in vzporednih besedil s spleta.

Zaključna **sekcija F** je bila namenjena slovenskim korpusnim raziskavam. **Darja Fišer** z Univerze v Ljubljani in **Tomaž Erjavec** z Inštituta Jožef Stefan sta predstavila slovenski wordnet SloWNet, prvi prosto dostopni slovenski semantični leksikon, ki je bil avtomatsko izdelan s pomočjo prosto dostopnih korpusnih in leksikalnih virov. **Peter Holozan** z Amebisa je iz vzporednega korpusa samodejno luščil slovar z uporabo analizatorja za prevedbo v vmesni jezik in s pomenskim razdvoumljanjem analizatorja. Prikazal je najpogostejše težave, kot so neujemanje strukture, zanikanje izvornega stavka, manjkajoče fraze, in predstavil načine njihovega reševanja. **Tomaž Erjavec** in **Simon Krek** z Inštituta Jožef Stefan sta opisala jezikovne vire JOS, ki jih trenutno sestavljajo oblikoslovne specifikacije in dva korpusa, vzorčena iz korpusa FidaPLUS. »Jos100k« je enojezični vzorčeni in uravnoteženi korpus slovenskega jezika s 100.000 besedami z ročno označenimi oz. pregledanimi lemami in oblikoskladenjskimi oznakami, enomilijonski korpus »jos1M« pa je delno ročno pregledan. Z jezikovnimi viri JOS sta se ukvarjali tudi **Špela Arhar** z Amebisa in **Nina Ledinek** z ZRC SAZU, ki sta predstavili revizijo in nadgradnjo nabora oznak JOS za oblikoskladenjsko označevanje slovenščine na primeru oznak za glagolsko besedno vrsto. Končni nabor oznak je zasnovan s ciljem vzpostaviti enotni označevalni standard za slovenščino. Kot se je pokazalo na konferenci, pa na njem že temelji več jezikovnotehnoloških projektov. **Vesna Mikolič**, **Ana Beguš**, **Davorin Dukič** in **Miha Koderman** z Univerze na Primorskem so natančno opisali projekt Večjezični korpus turističnih besedil, katerega cilj je zgraditi primerljivi in delno vzporedni korpus besedil s področja turizma v slovenskem, italijanskem in angleškem jeziku, ki bo uporaben kot prevajalski vir, za jezikoslovne raziskave in raziskave turizma. Konferenco sta končala **Špela Vintar** z Univerze v Ljubljani in **Tomaž Erjavec** z Inštituta Jožef Stefan, ki sta govorila o nadgradnji in označevanju korpusa računalniških besedil iKorpus. Enote računalniškega izrazja, samodejno izluščene iz iKorpusa, sta primerjala z dosedanjim besediščem spletnega računalniškega terminološkega slovarja Islovar.

Zaključku konference je sledil tretji posvet Slovenskega društva za jezikovne tehnologije *Est modus in korpus: ni korpusov brez divizij* o vrstah korpusov in njihovi gradnji, ki ga je vodil **Marko Stabej**. Po uvodnih predstavitvah, kjer je **Nataša Logar** predstavila pregled korpusov za slovenščino, **Špela Vintar** pa pregled orodij za izdelavo korpusov, se je razvila diskusija o temi posveta, ki je bil zaokrožen z rednim letnim občnim zborom društva.

Tudi letos so visoka udeležba, zaradi katere so bile predavalnice med nekaterimi sekcijami komaj dovolj velike za vse poslušalce, in kakovostni prispevki pokazali, da je konferenca *Jezikovne tehnologije* nujno potreben ter pomemben vezni člen in spodbujevalka znanstvenoraziskovalnega delovanja na tem področju. Poleg tega je tudi dobrodošel forum za razpravo o najnovejših dosežkih na področju jezikovnih tehnologij tako za slovenske strokovnjake kot tudi za mednarodno izmenjavo izkušenj na področjih, ki so trenutno v središču zanimanja slovenskih jezikovnotehnoloških raziskav. Letos so bile take izrazite teme oblikoslovno označevanje korpusov, strojno prevajanje in govorne tehnologije.

Urška Jarnovič[1] in Mojca Stritar[2]
*Filozofska fakulteta UL*
*[1]urska.jarnovic@ff.uni-lj.si*
*[2]mojca.stritar@ff.uni-lj.si*

## 13. mednarodni kongres Evropskega združenja za leksikografijo *EURALEX 2008*, Barcelona, 15.–19. julij 2008

13. mednarodni kongres *EURALEX*[1] se je odvijal v poletni Barceloni, in sicer v kampusu Ciutadella univerze Pompeu Fabra. Pred kongresom je bila letos organizirana večdnevna delavnica *Lexicom*, o kateri je že bilo objavljeno poročilo v reviji *Jezik in slovstvo* (JiS, LIII/5, 107–109). Z organizacijskega stališča je bil to večji zalogaj kot tradicionalne enodnevne ali dvodnevne predkongresne delavnice, vendar pa sta dogodka drug drugemu postala dodana vrednost in skupaj ustvarila deset dni strokovnih predavanj, predstavitev, srečanj, skupnega praktičnega dela, pogovorov ter navezovanja in utrjevanja stikov – za to pa mora biti življenjsko zainteresirana vsaka strokovna javnost, ki želi slediti sodobnemu dogajanju in ga tudi aktivno vključevati v svoje redne dejavnosti.

Za predstavitev na kongresu je organizacijski odbor do konca novembra 2007 prejel več kot 300 prispevkov in jih razposlal v dvojni slepi recenzentski pregled. V skladu z visokimi recenzentskimi kriteriji je v ožji izbor prišla manj kot polovica prispevkov. Kongres je obiskalo preko 350 ljudi iz najrazličnejših držav, predstavitelji so obravnavali številne evropske jezike in tudi nekatere druge, na mednarodnost pa kaže tudi število jezikov, v katerih so prispevki napisani (angleščina, nemščina, francoščina, italijanščina, katalonščina in španščina).

Univerzi, ki je gostila delavnico in kongres, je ime dal **Pompeu Fabra** (1868–1948), ki je bil v začetku 20. stoletja vodilna osebnost v gibanju za standardizacijo katalonščine. Široko poznan je kot jezikoslovec, slovaropisec, teoretik knjižnega jezika,

[1] <http://www.iula.upf.edu/agenda/euralex_08/euralex01uk.htm.>