Jezikovne tehnologije: zbornik 9. mednarodne multikonference Informacijska družba IS 2006 (ur. Tomaž Erjavec in Jerneja Žganec Gros)

Konferenca Jezikovne tehnologije, ena od osmih konferenc, združenih pod imenom Informacijska družba, je 9. in 10. oktobra 2006 v prostorih Instituta Jožef Stefan potekala že petič, prvič pa je tokrat odprla vrata tudi mednarodni znanstveni srenji. Konferenčni zbornik na dobrih 280 straneh prinaša 52 prispevkov domačih ter tujih avtorjev, od tega 37 prispevkov, napisanih v angleščini (s prevodom povzetka v slovenščino) ter 15 prispevkov, napisanih v slovenščini (s prevodom povzetka v angleščino).

Prva v zborniku sta prispevka vabljenih predavateljev. **Steven Krauwer** predstavlja sociolingvistično temo *Krepitev manjših evropskih jezikov*, v kateri izpostavlja problematiko razvoja jezikovnih tehnologij za male jezike (tako imenuje jezike, ki jim manjka tehnološke podprtosti). Industrijski razvijalci jezikovnih tehnologij se zaradi ekonomskega potenciala raje odločajo za vlaganje kapitala v raziskave velikih jezikov, kar podpira tudi finančna politika Evropske unije – ta se prepogosto izogiba soočenju z večjezičnostjo kot problemom na ravni celotne Unije in po načelu subsidiarnosti problematiko nacionalnih jezikov raje obravnava kot skup problemov na nacionalnih ravneh.

Nick Campbell piše na temo *Sinteza govora in diskurzna informacija*, v kateri se osredotoča na pomen upoštevanja pragmatike pri prepoznavi ter sintezi govora. Segmenti govorne komunikacije, kot npr. smeh, mrmranje, medmeti raznih tipov itd. (tj. metabesedilne informacije – glede na tematiko pričujoče številke revije je smiselno izpostaviti, da se vprašanja o metabesedilnosti porajajo tudi znotraj področja jezikovnih tehnologij), prinašajo številne informacije o trenutnem stanju govorca in, še pomembneje, njegovem odnosu do sogovornika. Ti podatki so ključnega pomena predvsem pri razvoju aplikacij, ki so namenjene dialogu med strojem ter človekom v naravnem jeziku, saj programiranemu sogovorniku zagotavljajo višjo stopnjo komunikacijske avtentičnosti.

V nadaljevanju bodo na kratko predstavljeni prispevki, ki se ukvarjajo z jezikovnimi tehnologijami za slovenski jezik.

Članek **Jerneje Žganec Gros** (Alpineon, d. o. o., LJ), **Varje Cvetko Orešnik** ter **Primoža Jakopina** (ISJ Frana Ramovša, LJ) predstavlja gradnjo slovarja *SI-PRON*, tj. slovarja izgovarjav (tako osnovnih kot pregibnih oblik) besed, zbranih v slovarju SSKJ. Slovar *SI-PRON* je vgrajen v sintetizator govora *Proteus*, z zvočno podobo pa bodo dopolnjeni tudi geselski članki spletne različice slovarja SSKJ. **Jerneja Žganec Gros** je sodelovala tudi s **Simonom Dobriškom, Boštjanom Vesnicerjem** ter **Francetom Miheličem** (FE, LJ) pri raziskavi možnih izboljšav avtomatske prepoznave govora s samodejnim prilagajanjem govornega modela trenutni govorni seji, prav tako pa z **Alešem Miheličem** (Alpineon, d. o. o., LJ) ter srbskimi znanstveniki **Vladom Delićem, Darkom Pekarjem** in **Milanom Sečujskim** pri predstavitvi projekta *iTEMA*. Cilj projekta je razviti aplikacijo, ki bo uporabniku omogočala enostavnejši dostop do e-pošte. Sistem bo deloval na osnovi govornega vmesnika, kar je še posebej uporabno kot pomoč slepim in slabovidnim, za katere bo storitev na voljo brezplačno.

Precej različnih institucij združuje sodelovanje pri projektu *VoiceTRAN*, o katerem pišejo **Jerneja Žganec Gros, Stanislav Gruden, Aleš Mihelič** (Alpineon, d. o. o., LJ), **France Mihelič, Simon Dobrišek, Janez Žibert** (FE, LJ), **Tomaž Erjavec** (IJS, LJ), **Špela Vintar** (FF, LJ), **Tomo Korošec, Nataša Logar** (FDV, LJ) ter **Peter Holozan**

(Amebis, d. o. o., Kamnik). Govorni komunikator *VoiceTRAN* je sistem, ki združuje prepoznavo govora, strojno prevajanje ter sintezo govora. Omogoča avtomatsko simultano prevajanje med angleščino ter slovenščino, zaenkrat pa je specializiran za vojaško rabo. Strojnega prevajanja govora se dotika tudi prispevek **Darinke Verdonik** (FERI, MB). Avtorica definira nekatere za strojno prevajanje problematične elemente govorjenega jezika: obotavljanja, samopopravki, napačni začetki, premori itd. Za premoščanje težav z jezikovnimi pojavnostmi tega tipa predlaga označevanje identificiranih pragmatičnih atributov v korpusih govorjenega jezika.

Skupina raziskovalcev z mariborske FERI, **Andrej Žgank, Tomaž Rotovnik, Zdravko Kačič** ter **Mirjam Maučec,** predstavlja zasnovo ter zgradbo sistema *UMB Broadcast News*, ki je namenjen razpoznavanju tekočega spontanega govora v televizijskih in radijskih oddajah. Prvi trije našteti avtorji, poleg njih pa še **Matej Grašič, Marko Kos** ter **Damjan Vlaj,** pišejo tudi o jezikovnem viru *SloParl*, nastalem za potrebe razvoja omenjenega sistema. Baza *SloParl* vsebuje parlamentarne razprave iz slovenskega Državnega zbora, sestavljena je iz govornega korpusa (s transkripcijami) ter tekstovnega korpusa. Govorni del zajema 100 ur govorjenega materiala, tekstovni del 23 milijonov besed. **Mirjam Maučec, Janez Brest** ter **Zdravko Kačič** na drugem mestu s pomočjo korpusa *SVEZ-IJS* analizirajo vplive različnih tipov jezikovnih informacij in različnih velikosti učnih korpusov na reševanje problema razpršenosti podatkov pri statističnem strojnem prevajanju.

K ciljem avtomatske razpoznave govora je usmerjena tudi gradnja korpusa televizijskih informativnih oddaj *SiBN,* o kateri pišejo **Grega Milharčič** (FF, LJ), **Janez Žibert** ter **France Mihelič** (FE, LJ). Avtorji so korpus potrebovali za preizkus treh različnih metod statističnega jezikovnega modeliranja. Zadnji našteti avtor sodeluje tudi z **Melito Hajdinjak** (FE, LJ), s katero objavljata dva prispevka: v prvem je predstavljeno ogrodje *Paradise,* ki se uporablja za vrednotenje učinkovitosti govornih vmesnikov (tj. računalniških sistemov, ki uporabniku omogočajo, da z govorom dostopa do želenih informacij). V drugem prispevku pa opisujeta postopek in rezultate uporabe tega ogrodja za vrednotenje učinkovitosti dveh (nedograjenih) govornih vmesnikov za podajanje informacij o vremenu in vremenski napovedi.

Jasna Belc in **Miran Željko** (Služba za prevajanje, tolmačenje, redakcijo in terminologijo Vlade RS, LJ) predstavljata način, kako iz manjšega nabora dvojezičnih korpusov zgraditi večjezični korpus. Specializirani korpusi, ki jih avtorja gradita predvsem na osnovi pravnih besedil EU, so namenjeni v prvi vrsti kot pomoč za prevajanje dokumentov podobnega tipa. Z gradnjo korpusov se ukvarja tudi prispevek **Jane Zemljarič Miklavčič** (Center za slovenščino kot drugi/tuji jezik), ki predstavlja pilotsko študijo gradnje korpusa govorjene slovenščine. Med procesom gradnje so bili določeni kriteriji transkribiranja ter označevanja govorjenega materiala, preizkušen pa tudi konkordančnik, ki omogoča različne tipe iskanja po korpusu. Vzpostavljeni so torej potrebni pogoji za gradnjo večjega korpusa govorjene slovenščine, ki naj bi dopolnjeval korpus pisnih besedil *FidaPLUS*.

Mojca Stritar (Center za slovenščino kot drugi/tuji jezik, LJ) odpira vprašanja gradnje korpusa usvajanja slovenščine kot tujega jezika, tako s stališča same zasnove korpusa kot tudi sistema označevanja korpusnih besedil. Rezultat analize obstoječih možnosti gradnje so smernice za izdelavo pilotskega korpusa omenjenega tipa. Na specializiranem korpusu poljudnoznanstvenih besedil pa temelji raziskava **Agnes Pisanski Peterlin** (FF, LJ), ki predstavlja poskus uporabe neoznačenega specializiranega korpusa za obravnavo kažipotov, tj. metabesedilnih elementov, s katerimi tvorec besedila napoveduje prihajajočo vsebino ali pa se sklicuje na že povedano. Ker kažipoti niso oblikovno določljivi, pač pa le funkcijsko (glede na vlogo v besedilu), jih je avtomatsko težko identificirati. Avtorica se s tem problemom sooči ter kot rešitev podaja nekaj možnih metodoloških izboljšav.

Katarina Puc (Slovensko društvo Informatika, LJ) in **Tomaž Erjavec** (IJS, LJ) predstavljata specializirani korpus besedil s

področja informatike, korpus *DSI*. Korpus je nastal (in se dograjuje) za potrebe urejanja spletnega terminološkega slovarja *Islovar* – slovarja informatike, ki je na internetu na voljo zainteresiranim uporabnikom: v uporabo, pa tudi za lastno dopolnjevanje. Korpus *DSI* je urejevalcem slovarja v pomoč tako pri sami selekciji slovarskega izrazja kot tudi odločanju med variantnimi poimenovanji, urejanju sinonimnih nizov itd. **Tomaž Erjavec** je sodeloval tudi z **Nino Ledinek,** s katero sta pripravila prispevek o prvih rezultatih gradnje korpusa *Slovenska odvisnostna drevesnica (SDT)*. Korpus obsega del Orwellovega romana *1984* (približno 30.000 besed), ki je oblikoskladenjsko označen po vzoru PDT (Praške odvisnostne drevesnice). Korpus je bil avtomatsko predoznačen z izbranim naborom oznak, nato pa so bile oznake s pomočjo specializiranega programa ročno pregledane in popravljene. Za raziskave je – v več različnih formatih – prosto dostopen na internetu. V sodelovanju z **Bencejem Sárossyjem** pa **Tomaž Erjavec** analizira natančnost avtomatskega oblikoslovnega označevanja z označevalnikom *TnT*. Označevalnik sta avtorja testirala na korpusu *SVEZ-IJS*, analizirala napake označevanja in podala nekaj možnih načinov odprave najpogostejših.

Prispevek **Simona Kreka** (FF, LJ) ter **Adama Kilgarriffa** (Lexical Computing Ltd, Brighton) predstavlja *Word Sketches*, zmogljivo programsko orodje, nepogrešljivo predvsem za vse tipe leksikoloških oz. leksikografskih raziskav. Na osnovi izbranega besedilnega korpusa ter podatkov o slovničnih vzorcih določenega jezika program podaja različne tipe kolokacijskih informacij o besedah, primerja denimo dve podani sopomenki in prikaže razlike ter podobnosti v njuni rabi ipd. Program sta avtorja prilagodila za raziskovanje slovenskega jezika, in sicer na osnovi referenčnega korpusa *FidaPLUS*.

Podjetje Amebis, d. o. o., Kamnik predstavlja dva segmenta svoje razvojne dejavnosti. **Špela Arhar** in **Miro Romih** predstavljata programiranega sogovornika *Klepca*, ki je umetna inteligenca, s katero lahko prek interneta klepetamo (zaenkrat pisno) v slovenščini. Prispevek s pomočjo primerov realnih komunikacijskih nizov med programom ter uporabniki izpostavlja tipične probleme razvijanja tovrstnih programov (npr. potrebo po upoštevanju specifik klepetalniškega diskurza, potrebo po čim boljši antropomorfizaciji programa itd.). **Peter Holozan** pa piše o problemih, ki jih avtomatskemu stavčnemu analizatorju povzročajo enakopisne oblike, ki jih program lahko interpretira bodisi kot knjižne bodisi kot neknjižne (gre za odstope od knjižne norme, ki pa se v pisnih besedilih pojavljajo dovolj pogosto, da jih je pri avtomatski analizi potrebno upoštevati, npr. zapis zaimka *jaz* kot *jest* ipd.).

V prispevku **Mihaela Arčana** ter **Špele Vintar** (FF, LJ) je predstavljenih več metod za avtomatsko prepoznavo lastnih imen v besedilu, vse so tudi preizkušene na slovenskem ter nemškem jezikovnem gradivu. Kljub temu da je sistem za prepoznavo v začetni fazi razvoja, že izkazuje solidne rezultate za oba obravnavana jezika. **Špela Vintar** z **Darjo Fišer** (FF, LJ) ter **Ljupčom Todorovskim** (FU, LJ) predstavlja tudi možnost izboljšave slovenskega *Wordneta* s pomočjo hierarhičnega rôjenja podatkov. Iz sopomenskih nizov, pridobljenih z avtomatskim prevajanjem srbskega *Wordneta*, skušajo avtorji avtomatsko izločiti nerelevantne kandidate – na osnovi korpusnih podatkov o kontekstu rabe posamezne besede v nizu.

S čim se torej ukvarja slovenska jezikovnotehnološka stroka: veliko je gradnje specializiranih besedilnih korpusov in podobnih besedilnih zbirk: za potrebe razvoja govornih tehnologij, za slovaropisje, v izobraževalne namene, v prevajalske namene ter nenavsezadnje za potrebe slovenističnih raziskav. Preizkušajo se novi načini označevanja korpusov (oblikoskladenjsko označevanje, pragmatično označevanje). Testira in izboljšuje se obstoječa razvojnotehnološka metodologija, dopolnjujejo se obstoječi jezikovni viri, gradijo se novi. Za slovenščino se prilagaja zmogljiva programska oprema, obstoječe slovenske aplikacije se nadgrajujejo, snujejo se nove. Dogaja se torej veliko, tako na področju teoretične kot aplikativne znanosti.

Steven Krauwer, prvi vabljeni govorec konference, navaja naslednje predpogoje za

kvaliteten jezikovnotehnološki razvoj malih jezikov: na državni ravni se mora urediti izobraževanje ustreznega raziskovalnega profila (hitri tečaji računalništva za jezikoslovce ali jezikoslovja za računalničarje ne zadoščajo), raziskovalci pa morajo (z državno pomočjo, seveda) poskrbeti za prosto dostopnost, čim širšo uporabnost in izmenljivost ter sprotno nadgrajevanje jezikovnih virov različnih vrst. Situacija pri nas še ni enaka opisani, so pa nakazani koraki v pravo smer. Večji premiki se bodo najbrž zgodili, ko se bo splošneje uzavestilo še eno dejstvo, ki ga navaja **Krauwer** – če ne bomo za svoj jezik poskrbeli sami, ne bo tega namesto nas opravil nihče.

Tukaj pa je, kot vedno, nekaj prostora za izboljšave. Konferenc na temo jezikovnih tehnologij, s tem pa možnosti za predstavitev slovenskega znanstvenega dela, je po svetu ogromno, v Sloveniji pač ne, zato nekateri avtorji z izbiro angleščine kot jezika prispevka precej presenečajo. Prevodi povzetkov so korektna uredniška gesta, žal pa zgolj s tem temelji za razvoj slovenske jezikovnotehnološke znanstvene misli še niso zagotovljeni. Drugo, kljub sodelovanju med različnimi tipi institucij (le-to je vsekakor izredno razveseljivo) po pregledu prispevkov ostaja občutek, da raziskovalcem manjka splošni razvojni konsenz. Vprašanj na temo, kaj slovenski prostor v tem trenutku dejansko potrebuje, je vsekakor premalo, posledično manjkajo raziskovalno-razvojne prioritete.

Konferenčni prispevki odpirajo tudi vprašanje o vlogi jezikoslovja na področju jezikovnih tehnologij: doprinos sodelujočih jezikoslovcev, kakršen se kaže v obravnavanem zborniku, je sicer zadovoljiv, precej boljša situacija pa se obeta, ko bo zares postalo jasno, da razvoj jezikovnih tehnologij za slovenski jezik brez upoštevanja slovenističnega znanja ne more biti kvaliteten. Ugotoviti, kaj lahko k razvoju jezikovnih tehnologij doprinesejo strokovnjaki različnih področij, je naloga posameznih strok (torej strokovnjakov samih) in kaže, da med vsemi edinole slovenistika še vedno čaka, da jo bo k sodelovanju nekdo povabil.

Špela Arhar
Amebis, *d. o. o.*, *Kamnik*
spela.arhar@amebis.si

19. evropska konferenca in delavnica iz sistemske funkcijske slovnice, Saarbruecken, 23.–25. julij 2007

Podatki in načini njihove interpretacije v jezikoslovju so bili osrednja tema 19. evropske konference in delavnice iz sistemske funkcijske slovnice. Metodološko vprašanje, ki je lahko aktualno za katero koli vedo in vejo znanosti, se odpira ob računalniško podprtih pristopih, ki postajajo pomembna, če ne kar najpomembnejša gonilna sila razvoja tudi v jezikoslovju.

Tradicionalni analitični pristop v jezikoslovju je temeljil predvsem na hermenevtičnih metodah in introspekciji, saj drugih načinov analize podatkov dolgo ni bilo na voljo. Toda tak pristop ima nedvomne pomanjkljivosti, saj težko sledi znanstvenim zahtevam po objektivnosti in ponovljivosti rezultatov in je omejen na majhno število primerov. Razvoj informacijske tehnologije je po drugi strani omogočil uporabo induktivnih metod, statistično vrednotenje rezultatov in analizo velikega števila primerov: toda še vedno moramo na neki točki rezultate interpretirati, poleg tega lahko že podatki sami vključujejo interpretacijo, zlasti kadar delamo s kakor koli označenim gradivom. Tako prvi kot drugi pristop imata torej dobre in slabe plati, in eden osrednjih ciljev te konference je bil pretehtati uporabo obojih v sistemski funkcijski slovnici.