SLOVENŠČINA 2.0: "JEZIKOVNE TEHNOLOGIJE"

Raziskave in razvoj na področju jezikovnih tehnologij se danes za jezike s širokim krogom govorcev pospešeno prenašajo v komercialne sisteme, ki postajajo vse bolj razširjeni. Denimo, rešitve samodejne prepoznave govora in samodejne sinteze govora se množično vgrajujejo v cenovno ugodne programske pakete, namenjene predvsem uporabi na osebnih računalnikih in prenosnih telefonih. Evropa je danes eden najnaprednejših trgov za jezikovne tehnologije. Evropska unija si prizadeva, da so potrebna orodja in viri na razpolago za vse njene jezike, kot tudi glavne svetovne komercialne jezike, s čimer utira pot večjezikovni informacijski družbi ter enotnemu digitalnemu trgu.

Zaradi jezikovnih specifik pa jezikovnotehnoloških rešitev ni možno kar preprosto prenašati med jeziki. Obseg sistematične raziskanosti jezikov, ki se govorijo v Evropi, se od enega jezika do drugega zelo razlikuje, pri čemer je bila v sklopu projektov v Evropski uniji, pa tudi v nacionalnih in komercialnih projektih dobro raziskana le peščica jezikov (angleščina, španščina, francoščina in nemščina), nekateri pa so bili komajda obravnavani. Pogosto so bile prav nove države članice tiste, ki niso imele možnosti za razvoj jezikovnih tehnologij za svoje pisne in govorjene jezike. Tudi za slovenščino in hrvaščino so bile izvedene le posamezne fragmentarne raziskave.

Če raziskav na področju jezikovnih tehnologij za slovenski jezik ne bo dovolj, uporaba slovenščine v prenekaterih sodobnih informacijsko-komunikacijskih aplikacijah ne bo možna. Raziskave na področju jezikovnih tehnologij za slovenski jezik prinašajo nova spoznanja tako na jezikovnem kot tudi na tehnološkem področju in omogočajo konkurenčnost našega jezika v primerjavi z drugimi.

Slovensko društvo za jezikovne tehnologije vsako drugo leto organizira konferenco »Jezikovne tehnologije«, ki poteka v sklopu multikonference

»Informacijska družba« in predstavlja glavni slovenski forum za predstavitev raziskav s področja računalniške obdelave jezikovnih podatkov, ki vključujejo jezikovne tehnologije, izdelavo jezikovnih virov in korpusno jezikoslovje. Osma konferenca »Jezikovne tehnologije« je potekala 8. in 9. oktobra 2012. Na njej je bilo predstavljenih 38 prispevkov (med njimi dve vabljeni predavanji), ki so bili objavljeni v tiskanem zborniku, dostopnem tudi na spletu. V prispevkih so avtorji predstavili raznovrstne raziskave, v katerih je posebej izstopalo veliko število prispevkov o izdelavi korpusov in drugih jezikovnih virov ter jezikovnih orodij za slovenščino in hrvaščino, dobro zastopani pa so bili tudi prispevki s področja govornih tehnologij.

Z novo revijo »Slovenščina 2.0« se je pojavila priložnost, da najboljše prispevke s konference avtorji razširijo in po ponovnem recenzentskem postopku objavijo v tematski številki, ki je sedaj pred vami. Prispevke so napisali tako slovenski kot tudi tuji avtorji; še posebej nas veseli, da je med slednjimi večina avtorjev s Hrvaške, saj smo si blizu tako geografsko kot jezikovno, zato je medsebojno spremljanje raziskav in sodelovanje še toliko bolj smiselno.

Sedem prispevkov obravnava raznovrstne vidike jezikovnih tehnologij. Jan Šnajder v prispevku »Models for Predicting the Inflectional Paradigm of Croatian Words« predstavi metodo za avtomatsko določanje oblikoslovne paradigme hrvaškim besedam s pomočjo klasifikatorja, ki se na osnovi besednih in korpusnih značilk nauči predvidevati, ali je določen par lema–paradigma pravilen. Ker metoda temelji na nadzorovanem strojnem učenju, je jezikovno razmeroma neodvisna in jo je moč uporabiti tudi za druge oblikoslovno bogate jezike, kot npr. slovenščino, za katero takih metod še nismo razvili.

Področje prepoznavanja in klasifikacije imen oz. imenskih entitet je tako znanstveno kot komercialno zelo zanimivo, kar se kaže tudi v tem, da ga obravnavata kar dva prispevka. Nikola Ljubešić in soavtorji v prispevku

»Combining Available Datasets for Building Named Entity Recognition Models of Croatian and Slovene« predstavijo več modelov za prepoznavanje in klasifikacijo imen tako za hrvaški kot za slovenski jezik. Modeli se med seboj razlikujejo po tem, koliko jezikovnih orodij potrebujejo za svoje delovanje, pri čemer so najboljše rezultate dosegli z uporabo oblikoskladenjskih lastnosti besed in distribucijskih lastnosti jezika, izračunanih iz velikih neoznačenih enojezičnih korpusov. Najboljši naučeni model je skupaj s testno množico in modelom za oblikoslovno označevanje hrvaščine tudi prosto dostopen. Tadej Štajner in soavtorja v prispevku »Razpoznavanje imenskih entitet v slovenskem besedilu« obravnavajo isti problem, vendar samo za slovenščino. Pri tem uporabijo drugačen klasifikator in nekatere drugačne značilke. Tudi tu je naučeni model prosto dostopen. Posebej dobrodošlo pri obeh prispevkih pa je, da sta oba preizkušena tudi na testni množici iz drugega prispevka, s čimer lahko neposredno primerjamo kvaliteto obeh pristopov.

Trije prispevki so s področja leksikografije oz. terminologije. Darja Fišer s soavtorjema z Univerze v Wrocławu v prispevku »Grounding sloWNet on Slovene Corpus Data« predstavi metodo za dopolnjevanje slovenskega semantičnega leksikona sloWNet s pomočjo jezikovnih podatkov, pridobljenih iz jezikoslovno označenega referenčnega korpusa. Pristop, izvorno razvit za poljščino, s pomočjo preprostih statističnih metod izlušči sezname semantično podobnih besed, ki so bile nato vključene v sloWNet. Nataša Logar in soavtorici v prispevku »Terminologija odnosov z javnostmi: korpus – luščenje – terminološka podatkovna zbirka« prikažejo analizo luščenja terminoloških kandidatov, ki so jo izvedle za potrebe priprave terminološke podatkovne zbirke odnosov z javnostmi. Iztok Kosem in soavtorja v prispevku »Avtomatizacija leksikografskih postopkov« opišejo poskus uvedbe novega pristopa v proces izdelave slovarjev, pri katerem je leksikograf predvsem ocenjevalec izbir, ki jih predhodno opravi računalnik. Novost v pristopu je, da omogoča različne faze v izdelavi slovarja, pri čemer je leksikografsko delo mogoče razdeliti glede na zahtevnost opravil, kar pomeni, da je mogoče za

ocenjevanje avtomatsko izluščenih korpusnih podatkov uporabti moč množic, leksikografom pa prepustiti samo zahtevnejša opravila, kot so npr. pomenski opisi besed ipd.

V prispevku Simona Dobriška s soavtorji »Bodo pametni nadzorni sistemi prisluhnili, razumeli in spregovorili slovensko?« so zastopane tudi govorne tehnologije. Članek predstavlja trenutno stanje razvoja pametnih nadzornih sistemov in možnosti njihove uporabe za slovenski govorjeni jezik ter različne varnostno-nadzorne scenarije uporabe tovrstnih sistemov. Naslovljena so tudi širša pravna in etična vprašanja, saj je nadzor govora eno najbolj občutljivih vprašanj varstva zasebnosti.

Pred vami je šele druga številka revije »Slovenščina 2.0« in izdaja tematske številke o jezikovnih tehnologijah, torej z računalniško ozaveščenimi avtorji, je s sabo prinesla tudi tehnično izboljšavo. V večini bolj naravoslovnih revij, npr. pri Springerju, ki izdaja tudi revijo »Language Resources and Evaluation« ali ACL s »Computational Linguistics«, je LaTeX priporočeni format prispevkov. Zahvaljujoč dr. Janu Šnajderju, ki je izdelal predlogo, je sedaj možno tudi za »Slovenščino 2.0« oddajati članke v tem formatu, pri čemer se LaTeX oblikovno ujema s prispevki v Wordu.

Urednika:
Tomaž Erjavec, Jerneja Žganec Gros

Erjavec, T., Žganec Gros, J. (2013): Slovenščina 2.0: "Jezikovne tehnologije". Slovenščina 2.0, 1 (2): I-V.
URL: http://www.trojina.org/slovenscina2.0/arhiv/2013/2/Slo2.0_2013_2_01.pdf.