

## **LREC'06 – Language Resources and Evaluation Conference 2006, Genova, 24. –26. maj 2006**

Darinka Verdonik

»LREC'06 je srečanje vseh, ki verjamejo, da so jezikovni viri in njihovo vrednotenje temeljni kamen jezikovnih tehnologij, tako za pisni kot govorjeni jezik. Če želiš srečati koga s tega področja, pojdi na LREC in on/ona bo tam!« To figurativna ocena organizatorjev konference, ELRE/ELDE,<sup>1</sup> niti ni veliko pretirana – če sodimo po več kot 800 udeležencih in okoli 500 sprejetih prispevkih, kar se jih je zvrstilo samo v rednem programu konference, brez spremljajočih delavnic, gre resnično za eno osrednjih znanstvenih srečanj na področju jezikovnih virov.

Termin jezikovni viri zajema tako rekoč vsako urejeno elektronsko zbirko, ki jo lahko uporabimo kot vir informacij o jeziku, pa naj bodo to slovarji (s pomenskimi opisi, z oblikoslovnimi oznakami, fonetičnimi prepisi, dvo- ali večjezični slovarji itd.), korpusi (pisnih ali govorjenih besedil, referenčni ali specializirani, enojezični ali dvojezični (poravnani), z vključenimi oblikoslovnimi oznakami, fonetičnim prepisom, oznakami o skladijskih ali semantičnih razmerjih, pragmatičnih elementih itd.), posnetki govora s pripadajočimi transkripcijami (fonetičnimi in/ali ortografskimi, oznakami šumov, podatki o govorcih, kanalu itd.) ... Sem štejemo tudi osnovna programska orodja za zbiranje, urejanje in uporabo teh zbirk. Na LRECU-u se vsaki dve leti srečujejo strokovnjaki, ki izvirajo iz tehniških ved ali jezikoslovja, iz akademskega okolja ali gospodarstva. Jezikovni viri so podpora za razvoj različnih komponent, sistemov oz. aplikacij, ki so del razvoja t. i. e-družbe (npr. različne jezikovne storitve, e-vlada, e-oglaševanje, e-učenje ...), vendar se tukaj omejimo predvsem na njihovo uporabo v jezikoslovju. V slovenskem okolju se korpusni pristop k raziskovanju jezika šele začel razvijati, kar je omogočil zlasti referenčni korpus FIDA; zavedanje, da je lahko vpogled v vrste jezikovnih virov in pristope k raziskovanju, ki jih sodobne tehnologije omogočajo, koristen vir novih idej ali izzivov za delo jezikoslovcev z najrazličnejših področij, pa naj se ukvarjajo z glasoslovjem, skladnjo, besedotvorjem,

---

<sup>1</sup> ELRA (European Language Resources Association) je nepridobitna organizacija, ki skrbi za distribucijo jezikovnih virov. Več informacij o njej najdemo na <http://www.elra.info/>. ELDA (Evaluations and Language Resources Distribution Agency) je operativno telo ELRE, ki beleži, klasificira, zbira, validira in tudi gradi jezikovne vire. Naslov spletnih strani je <http://www.elda.org/site2006/>.

dialektologijo, semantiko, pragmatiko, slovaropisjem, prevajanjem ..., se tako v slovenski jezikoslovni javnosti šele počasi širi.

Množičnejši razvoj jezikovnih virov se v svetu začne v začetku devetdesetih let dvajsetega stoletja in po mnenju Nicolette Calzolari, vodje konference LREC'06, so bila takrat osrednja vprašanja razvoja: standardizacija jezikovnih virov, gradnja temeljnih jezikovnih virov in označevanje, distribucija jezikovnih virov. Danes se kaže tudi četrto vprašanje, ki se nanaša na metode avtomatskega pridobivanja jezikoslovnih (in drugih) informacij.

Če gledamo označevanje jezikovnih virov z vidika tradicionalnega jezikoslovja, se zdi oblikoslovno označevanje (kot je npr. vključeno tudi v slovenski referenčni korpus FIDA) bolj ali manj rešeno vprašanje. Tudi skladiščno označevanje (ki pa je za slovenščino šele v začetku) je doseglo zadovoljive rezultate. Kar se na podlagi prispevkov na konferenci LREC'06 kaže kot ospredje zanimanja jezikovnih virov danes, je po mnenju Nicolette Calzolari semantika, ki je osrednja tema tako razvoja slovarjev kot korpusov. To potrjujejo številne sekcije, ki vključujejo to temo, kot na primer Ontologije, Občutja in semantika, Vrednotenje semantike in pomenov, Semantično označeni korpusi, Sintaktični in semantični slovarji, Semantika, ontologije, semantične mreže itd.

Nove zanimive teme na področju jezikovnih virov so povezane z iskanjem, označevanjem in analizo subjektivnih elementov, tako v govorjenih kot pisnih besedilih. Gre predvsem za elemente, ki izražajo osebna mnenja, ali elemente, ki izražajo občutja. Pri slednjih nekateri vodilni avtorji s tega področja opozarjajo na terminološko razliko med emocijami (angl. emotion), ki so sicer predvsem z akustičnega vidika že dalj časa predmet zanimanja govornih tehnologij, zlasti sinteze govora, in občutji (angl. affect) v luči procesiranja govora. Že nekaj časa sta vroči temi tudi večjezikovnost, torej gradnja vzporednih korpusov in dvo- ali večjezičnih slovarjev ter orodij in algoritmov zanje, ter večmodalnost (večmodalna orodja, večmodalni korpusi ...). Precej prispevkov na konferenci se ukvarja z vrednotenjem jezikovnih virov in z razvojem orodij in metodologij za vrednotenje ter z infrastrukturo in arhitekturo jezikovnih virov. Svoj delež ohranjajo prispevki, ki se nanašajo na govor. Gradnja govornih korpusov različnega obsega in v različne namene je ta čas aktualna npr. za ruščino, portugalsščino, nizozemščino, japonsščino, italijanščino, danščino ... Seveda je govor obvezna tema v povezavi s sintezo in razpoznavo govora, prevajanjem govora in sistemi dialoga. Čeprav je konferenca posvečena jezikovnim virom, nekaj pozornosti pritegnejo tudi sistemi,

predvsem sistemi strojnega prevajanja, pridobivanja in luščenja informacij in dialoga odgovarjanja na vprašanja. Vseeno so kolegi s področja tehniških ved, s katerimi smo bili na konferenci, ugotavljali, da se jezikoslovni del publike na LREC-u v primerjavi s prejšnjimi leti krepi.

Na konferenci LREC je zastopanih okoli 50 različnih jezikov, večina evropskih, pa seveda azijskih, bližnjevzhodnih, afriških ... Slovenščino že vrsto let s po več prispevki zastopajo predvsem avtorji iz jezikovnotehnoloških centrov v Sloveniji, npr. na Institutu Jožef Stefan, Fakulteti za elektrotehniko v Ljubljani, Fakulteti za elektrotehniko, računalništvo in informatiko v Mariboru idr., redkeje pa se pojavlja kot obravnavani jezik tudi v predstavitev tujih avtorjev, predvsem po zaslugi mednarodnih projektov, v katere se vključujejo naše institucije. Letos smo avtorji prispevkov o jezikovnih virih za slovenščino predstavljali vzporedni korpus, skladiščno označeni korpus, en prispevek sega na področje semantike, dva pa na področje govornih korpusov.

Tomaž Erjavec je predstavljajal angleško-slovenski vzporedni korpus ACQUIS. Korpus vsebuje približno 10 milijonov besed, ki tvorijo angleško-slovenski pomnilnik prevodov SVEZ ACQUIS. Ta pomnilnik prevodov je bil izdelan med procesom prevajanja zakonodaje EU (ACQUIS) v slovenski jezik, v okviru prevajalske skupine SVEZ. Angleški in slovenski tekst je označen na ravni besed, lematiziran in oblikoslovno označen, vendar vse samo z avtomatskimi postopki. V raziskovalne namene je dostopen prek spletnih strani <http://nl.ijs.si/svez/>, kjer je tudi več podatkov o njem. Skupaj z Darjo Fišer je Tomaž Erjavec predstavljajal gradnjo besedne mreže WordNet za slovenski jezik. Besedna mreža je elektronska besedna zbirka, v kateri so samostalniki, glagoli, pridevniki in prislovi povezani z leksikalnimi in semantičnimi relacijami. Te vrste semantični leksikoni so pomemben vir npr. za razvoj semantičnega označevanja, avtomatskega povzemanja, avtomatskega pridobivanja podatkov ipd. WordNet za slovenski jezik vsebuje trenutno okoli 5000 konceptov, ki so bili avtomatsko prevedeni iz srbskega WordNeta s pomočjo dvojezičnega slovarja in ob podpori pojavnosti v korpusu, rezultati pa so bili ročno popravljani. S. Džeroski, T. Erjavec, N. Ledinek, P. Pajas, Z. Žabokrtsky in A. Žele so predstavljajali začetke gradnje skladiščno označenega korpusa (t.i. treebank) za slovenski jezik. Gre za del vzporednega korpusa MULTEXT-East, natančneje prvi del romana 1984. Pri tem so se naslonili na češki Prague Dependency Treebank, ki ima med podobnimi korpusi kar dolgo tradicijo. Izbira je smiselna predvsem zaradi podobnosti jezikov. Slovenski skladiščno označeni korpus obsega trenutno

2000 povedi oz. 30.000 besed. J. Gros, V. Cvetko Orešnik, P. Jakopin in A. Mihelič so predstavljali fonetični leksikon za slovenski jezik SI-PRON. Leksikon vključuje fonetične prepise za slovenske besede, ki so zbrane v SSKJ, razširjen s številnimi pregibnimi oblikami. Po obsegu je verjetno največji tovrstni leksikon za slovenski jezik, vsebuje 1,4 milijona vnosov. A. Žgank, D. Verdonik, A. Zoegling Markuš in Z. Kačič smo predstavljali bazo SINOD, ki vsebuje posnetke in transkripcije govora govorcev slovenščine, ki jim slovenščina ni materni jezik. Gre za dodatek k bazi BNSI Broadcast News, ki obsega 72 ur govora v televizijskih informativnih oddajah in okroglih mizah, transkribiranega in označenega v skladu z mednarodnimi standardi gradnje baz tipa Broadcast News. SINOD vsebuje posnetke v skupni dolžini 102 minuti. Avtorica tega poročila sem predstavljala korpus telefonskih pogovorov v turizmu TURDIS. Korpus vključuje transkripcije približno 4,5 ure telefonskih pogovorov (43.000 besed) s turistično agencijo, turistično pisarno in hotelsko recepcijo.

V zborniku z več kot 500 prispevki je seveda težko imeti pregled nad vsebino. Pri tem sta v veliko pomoč posebna knjiga povzetkov in razvrstitev člankov po sekcijah, kot so bili predstavljeni na konferenci. Sicer je res, da te razporeditve niso vedno povsem posrečene, kar pa ob toliko prispevkih organizatorjem lahko oprostimo. Zbornik izhaja tako v papirni kot elektronski obliki, verjetno pa bo počasi prevladala elektronska, saj je bilo treba letos tiskane izvode posebej naročiti. Pomemben razlog je najbrž prav obseg, ki je samo od leta 2002 do 2004 narasel s treh na šest knjig, ob toliko dodatne prtljage in prostora na knjižni polici pa človek dvakrat premisli. Elektronska verzija je tudi opremljena z avtomatskimi iskalniki po avtorjih, naslovih in ključnih besedah, kar močno pohitri dostop do iskanih vsebin.

Naslednja konferenca LREC bo leta 2008. Po napovedih se bo prvič iz evropskih držav (do zdaj je dvakrat gostovala v Španiji, po enkrat pa v Grčiji, Portugalski in Italiji) selila na afriško celino, v Marakeš v Maroku.