Pomen združevanja raziskovalnih potencialov pri preseganju jezikovnih pregrad v okviru jezikovnih tehnologij naslednjih generacij

Zdravko Kačič

Univerza v Mariboru, Fakulteta za elektrotehniko,
računalništvo in informatiko, Center za jezikovne tehnologije,
Smetanova 17, 2000 Maribor
kacic@uni-mb.si

Povzetek

V članku podajamo kratko analizo raziskovalnih aktivnosti na področju jezikovnih tehnologij v slovenskem prostoru v preteklem osemletnem obdobju ter skušamo nakazati nekaj osnovnih preprek za obsežnejše vključevanje slovenskih raziskovalnih skupin v mednarodne raziskovalne projekte. V nadaljevanju povzemamo osnovne značilnosti izvajanja raziskovalnih projektov na področju jezikovnih tehnologij v zadnjih nekaj letih v ZDA in jih primerjamo s projekti v Evropi. Nov koncept izvajanja raziskav v obliki integralnih projektov v okviru 6. okvirnega programa postavlja pred raziskovalce v slovenskem prostoru zahtevo po večjem združevanju in transparentnosti raziskovalnega dela za enakopravnejšo vključitev v te aktivnosti. Predstavljena projekta LC-STAR in TC-STAR bi lahko predstavljala vodilo za takšne aktivnosti.

1. Uvod

V zadnjih desetih letih so v evropskem prostoru potekale aktivnosti načrtne izgradnje jezikovnih virov, ki so bile uresničevane pretežno skozi projekte okvirnih programov, ki jih razpisuje Evropska unija. Četrti in peti okvirni program sta tako vključevala projekte, v okviru katerih so zgradili množico govornih in pisnih jezikovnih virov, predvsem za države članice unije, preko drugih raziskovalnih programov (npr. Copernicus, INCO Copernicus) pa so nekatere izmed teh projektov prenesli tudi na države srednje in vzhodne Evrope (npr. Onomastica, Multext-East). Prav tako so v tem obdobju v evropskem prostoru nastale prve organizirane institucionalne oblike zagotavljanja potrebnih jezikovnih virov (ELRA/ELDA) in zagotavljanja kvalitete le-teh s pomočjo ustanavljanja validacijskih centrov (na primer SPEX). Evropa si je s temi aktivnostmi v veliki meri zagotovila primat na področju gradnje predvsem večjezičnih jezikovnih virov in s tem izhodišče za uspešen razvoj tudi na področju razvoja sistemov jezikovnih tehnologij. Zaradi temeljnega pomena jezikovnih virov pri razvoju sistemov jezikovnih tehnologij, časovne zahtevnosti izgradnje virov in visoke cene se v zadnjih petih letih kažejo vse večja pričakovanja Komisije EU po zagotavljanju privatnega kapitala za izgradnjo virov in finančne podpore na nivoju posameznih držav članic unije. Takšne primere je bilo moč srečati na primer pri izvajanju projektov SpeechDat (SpeechDatII E, SpeechDat Car, Speecon ...). Ta pristop pa lahko prinese precej težav predvsem državam srednje in vzhodne Evrope, ki so bile v preteklosti največkrat slabo vključene v aktivnosti zagotavljanja osnovnih jezikovnih virov. Ti niso le komercialnega pomena, pač pa imajo velikokrat tudi nacionalni pomen. Uporaba privatnega kapitala pri izgradnji takšnih virov običajno pomeni tudi financerjevo lastništvo izdelanih jezikovnih virov. Primer so baze izgovarjav SpeechDat II (Češka, Madžarska, Slovaška, Slovenija), kjer so lastniki baz podjetja iz tujine (npr. Siemens, Philips). Ob pomanjkanju domačega kapitala lahko pomeni takšen razvoj v določeni meri odvisnost narodov od tujega kapitala na tem za posamezen narod precej občutljivem področju, ki se je doslej štelo tudi za eno od prvin zagotavljanja narodove samobitnosti.

2. Raziskovalne aktivnosti v slovenskem prostoru

V Sloveniji se med raziskovalnimi in univerzitetnimi ustanovami z razvojem jezikovnih tehnologij in izgradnjo jezikovnih virov, dosegljivih v elektronski obliki, ukvarjajo: v Ljubljani predvsem na Fakulteti za elektrotehniko (bazi izgovarjav K211D, LAPSC, Gopolis, VNTV, Luz) (Gros, 2000)(Dobrišek, 1998), Inštitutu Jožef Stefan (projekt Multext-East, IJS-ELAN, korpus FIDA) (Erjavec, 1998)(Erjavec, 1998a), Filozofski fakulteti (korpus FIDA), ZRC SAZU (besedilni korpus Nova beseda, SSKJ), Fakulteti za računalništvo in Naravoslovnotehniški fakulteti; v Mariboru na Fakulteti za elektrotehniko, računalništvo in informatiko (FERI) in v Centru za jezikovne tehnologije, ki deluje na FERI (baze izgovarjav SNABI, SpeechDat II, PoliDat, InterFace, morfološki in fonetični slovarji Onomastica, SImlex, SIflex, besedilni korpusi Večer in Delo) (Kačič, 2000)(Kaiser, 1998), (Rojc, 2002) (Hozjan, 2002), ter na Pedagoški fakulteti; med podjetji pa predvsem podjetje Amebis iz Kamnika (korpus FIDA in slovarji) (Romih, 1998) v sodelovanju z DZS (korpus FIDA in slovarji) in akademskimi institucijami ter Hermes Softlab.

Za slovenski raziskovalni prostor je na področju jezikovnih tehnologij značilen sorazmerno majhen raziskovalni potencial, ki je v preteklem obdobju deloval precej nepovezano in razpršeno. Vključenost v mednarodne projekte je bila tako bolj plod prizadevanj posameznikov in njihovih poznanstev s skupinami v tujini in manj plod organiziranega skupnega delovanja. Je pa bilo vključenosti v mednarodne projekte v preteklosti relativno malo.

Kljub temu lahko za obdobje zadnjih osmih let rečemo, da je bil na področju jezikovnih tehnologij dosežen napredek. Pri tem je opazno precejšnje nesorazmerje med izkazano znanstvenoraziskovalno aktivnostjo raziskovalcev s prispevki na najpomembnejših mednarodnih konferencah v evropskem prostoru (na primer Eurospeech 97, 99, 01, ICSLP 00, 02, LREC 98, 00, 02), kjer je bilo število publikacij slovenskih raziskovalcev velikokrat precej večje kot pa število publikacij raziskovalcev drugih srednje- in vzhodnoevropskih držav, ter vključenostjo slovenskih

raziskovalnih skupin v mednarodne projekte. Delno gre vzroke za to pripisati tudi manjši atraktivnosti slovenskega jezika zaradi njegove majhnosti (podobne razmere lahko na primer ugotavljamo pri drugih manjših jezikih).

V preteklih osmih letih (1995-2002) je Ministrstvo za šolstvo, znanost in šport Republike Slovenije financiralo 11 znanstvenoraziskovalnih projektov in tri programe, ki bi jih lahko razvrstili v področje jezikovnih tehnologij. To se morda zdi na prvi pogled veliko, vendar pa nekoliko podrobnejša analiza razkrije dokaj značilno strukturo in organiziranost za slovensko raziskovalno sfero. Omenjeni projekti so bili namreč večinoma manjši, v njih je sodelovalo majhno število strokovnjakov. Tekoči raziskovalni programi pa pokrivajo raziskovalno področje jezikovnih tehnologij večinoma le kot del svojih aktivnosti. Tudi finančni vložek je bil sorazmerno majhen, predvsem ob upoštevanju dejstva, da je problem, ki je postavljen pred raziskovalce, približno enako zahteven za vse jezike in da doseženih rešitev za druge jezike največkrat ni moč preprosto prenesti na slovenski jezik. Tako je ob takšnih razmerah in razdrobljenosti raziskovalnega potenciala težko pričakovati doseganje kritične mase raziskovalnih moči in s tem tudi mednarodno primerljivih in zanimivih rezultatov. Razdrobljenost in nepovezanost raziskav pa onemogoča tudi doseganje sinergičnega učinka med raziskovalci, kar je še kako pomembno in potrebno na izrazito interdisciplinarnem področju jezikovnih tehnologij. Res pa je tudi, da v preteklosti ni bilo zaslediti iniciativ ne s strani financerjev (ministrstev) ne s strani raziskovalcev posameznih raziskovalnih institucij po večji povezanosti raziskovalnega dela na tem področju. Doslej premalo ali skoraj nič primerov povezovanja strokovnjakov z interdisciplinarnih področij v okviru enotnih projektov se odraža tudi v samem delu raziskovalnih skupin, kjer je pogosto opazno pomanjkanje interdisciplinarnega znanja. Takšno stanje je moč opaziti tudi v prisotnosti jezikovnih tehnologij za slovenski jezik v mednarodnem prostoru, še posebej mednarodno primerljivih jezikovnih virov. Razen virov, ki so bili izvedeni v okviru mednarodnih projektov, trenutno ni razpoložljivih virov, ki bi imeli priznano mednarodno veljavo in bi bili na voljo evropski in svetovni raziskovalni skupnosti preko organizacij, ki skrbijo za zbiranje in distribucijo jezikovnih virov (npr. ELDA, TRACTOR…).

Obstoj mednarodno primerljivih jezikovnih virov za slovenski jezik predstavlja »vstopnico« slovenskim raziskovalcem v mednarodne raziskovalne projekte, še posebej projekte okvirnih programov, ki jih razpisuje Evropska unija. Pri tem je pomembno, da so jezikovni viri skladni z mednarodnimi standardi in priporočili in izvedeni v okviru različnih mednarodnih projektov ali pa verificirani pri ustreznih validacijskih centrih.

Vse to kaže na nujnost ureditve »statusa« jezikovnih virov v slovenskem prostoru. Ta je povezana predvsem z njihovo transparentnostjo, dosegljivostjo, mednarodno primerljivostjo in referenčnostjo. Korak k temu cilju pomeni mednarodna verifikacija obstoječih virov. Viri, ki so bili zgrajeni v okviru mednarodnih raziskovalnih projektov, takšno verifikacijo večinoma že imajo, npr. SpeechDat II, Onomastica, Multext-East …

V evropskem prostoru je v zadnjih nekaj letih zaslediti izrazite težnje po organiziranosti vseh vidikov znanstvenoraziskovalnega dela, združevanju vseh področij jezikovnih tehnologij pod okrilje skupnih projektov in ustanavljanju centov za validacijo vseh vrst jezikovnih virov ter institucij za njihovo zbiranje in distribucijo. Eden osnovnih namenov takšnega organiziranja znanstvenoraziskovalnega dela je tudi doseganje kritične mase raziskovalcev in sinergičnega učinka, ki predstavlja eno izmed osnovnih orientacij projektov 6. okvirnega programa Evropske unije.

3. Raziskovalne aktivnosti v ZDA in Evropi

3.1. Raziskovalne aktivnosti v ZDA

Raziskovalne aktivnosti na področju jezikovnih tehnologij potekajo v ZDA v okviru organiziranih raziskovalnih programov že več kot trideset let (Mariani, 2002). Najprej preko programov agencije DARPA (Defence Advanced Research Agency), ki še danes skrbi za izvajanje raziskav na področju razvoja osnovnih tehnologij. Za izvajanje temeljnih raziskav pa danes pretežno skrbi Nacionalna znanstvena fundacija NSF (National Science Foundation). Obe ustanovi prejemata večino finančnih sredstev za izvajanje raziskav na področju jezikovnih tehnologij iz širše tehnološke iniciative Računalništvo, informacije in komunikacije CIC (Computing, Information and Communication), ki je bila formirana leta 1998. Ta iniciativa je sledila programu Visokozmogljivo računanje in komunikacije HPCC (High Performance Computing and Communication), ki je trajal od leta 1991 do 1997. Letni proračun programa CIC je milijarda ameriških dolarjev. Sredstva iz programa CIC pa prejemajo ob že omenjenih organizacijah NSF in DARPA tudi druge nacionalne agencije, kot so: CIA, FBI, US air Force, Nacionalna agencija za varnost NSA (National Security Agency) in druge, ki pa večinoma skrbijo za razvoj aplikacij na področju jezikovnih tehnologij. V okviru agencije DARPA poteka že od leta 1984 raziskovalni program s področja jezikovnih tehnologij HLT (Human Language Technology) z letnim proračunom 34 milijonov ameriških dolarjev v letu 2000. V okviru tega programa poteka trenutno več obsežnejših pojektov, kot so: TEAM TIDES (Translingual Information Detection, Extraction & Summarization), Communicator in ROAR (Reliable Omni-Present Automatic Recognition).

Program TEAM TIDES, znotraj katerega poteka množica projektov, pokriva področje procesiranja naravnega jezika. Program bo potekal med letoma 2000 in 2005. V njem sodeluje 28 partnerjev iz univerz, raziskovalnih ustanov, podjetij in vladnih organizacij. Glavna cilja projekta sta razviti sistem avtomatskega prevajanja z vsaj 80% uspešnostjo prevajanja in sisteme avtomatskega indeksiranja in povzemanja besedil, ki bodo uporabni v praksi. Program vključuje tudi načrtovanje in izgradnjo potrebnih jezikovnih virov ter njihovo vrednotenje.

Program ROAR, ki bo potekal v letih od 2001 do 2006, pokriva področje govorne tehnologije. Glavna naloga programa je s pomočjo različnih projektov zgraditi sistem avtomatskega vodenja zapisnikov sestankov (meeting transcription). Program tako vključuje temeljne raziskave na področju avtomatskega razpoznavanja govora, vključujoč robustnost razpoznavanja v šumnih okoljih. Vključuje tudi raziskave multimodalne komunikacije. Eno pomembnih komponent programa predstavlja vrednotenje uspešnosti razvitih sistemov. V

programu tako sodelujeta tudi LDC in NIST. Bil pa je ustanovljen tudi poseben komite za vrednotenje dosežkov. Proračun projekta za obdobje šestih let je 59 milijonov ameriških dolarjev, od tega je 50% namenjenih temeljnim raziskavam. Program je odprt za tuje partnerje, ki pa si morajo zagotoviti lastna finančna sredstva za sodelovanje v programu.

Program DARPA Communicator, ki se je pričel leta 1999 in bo trajal do leta 2003, pokriva področje uporabe sistemov dialoga predvsem v vojaških aplikacijah. V programu sodeluje 18 partnerjev, od tega približno tretjina univerz. Proračun projekta je 18 milijonov ameriških dolarjev. Ciljna aplikacijska domena programa je načrtovanje potovanj in rezervacije. Partnerji projekta uporabljajo skupno zasnovo sistema, ki so jo izdelali na MIT in vključuje več različnih modulov: razumevanje govora, sledenje konteksta, samodejno tvorbo besedila, vodenje dialoga in multimodalno ter multimedijsko komunikacijo. Tudi ta program je odprt za partnerje iz tujine. Doslej so pri njegovem izvajanju sodelovali partnerji iz Švedske, Nemčije in Danske.

Skupna značilnost raziskovalnih aktivnosti, ki jih podpirajo različne fundacije v ZDA, je, da podpirajo manjše število projektov, ki pa zato vključujejo večje število partnerjev, trajajo običajno dlje kot tri leta in imajo večji proračun. Na ta način so v projekt vključeni partnerji s strokovnjaki iz vseh področij, ki jih vključuje multidisciplinarno področje jezikovnih tehnologij (avtomatska sinteza in razpoznavanje govora, procesiranje naravnega jezika), in lahko računajo na sinergični učinek pri rezultatih projekta. Hkrati so projekti dovolj veliki in večinoma trajajo dovolj dolgo, da lahko vključujejo tudi načrtovanje in razvoj potrebnih jezikovnih virov, katerih razvoj je zelo dolgotrajen in drag. Prav tako vsi programi vključujejo zelo pomembno komponento vrednotenja rezultatov projekta, ki je predvsem v ZDA zelo skrbno načrtovana, kar je posledica izkušenj prvega DARPA programa SUS (Speech Understanding Systems), ki je trajal od 1971-1976 in pri katerem je popolna neusklajenost specifikacij izvajanih projektov onemogočila kakršnokoli primerjavo in s tem povezano vrednotenje programa.

3.2. Raziskovalne aktivnosti v Evropi

Raziskovalna aktivnost je v Evropi organizirana pretežno v obliki okvirnih programov (Framework Programmes – FP). Od leta 1983 do danes je bilo tako izvedenih že pet okvirnih programov. Šesti se pričenja v letu 2003. Posamezen okvirni program običajno traja 5 let, projekti, ki se izvajajo v okvirnih programih, pa najpogosteje trajajo tri leta. Širši cilj okvirnih programov je zgraditi močno tehnološko osnovo za Evropsko unijo, dolgoročneje pa naj bi pripomogli k skladnejšemu razvoju in povečanju konkurenčnosti Evrope na svetovnem tržišču. Za razliko od večine ameriških raziskovalnih programov temelji koncept okvirnih programov predvsem na sodelovanju partnerjev znotraj posameznih projektov in ne na njihovem medsebojnem tekmovanju. Prav tako je cilj tudi razvoj sodelovanja med posameznimi projekti znotraj skupin projektov, ki pokrivajo enaka ali sorodna znanstvena področja. Posamezni projekti običajno vključujejo več partnerjev iz različnih držav, vendar zmeraj vsaj dva iz držav Evropske unije. Prav tako je prisotna težnja po uravnoteženi zastopanosti (50%-50%) partnerjev iz univerz in raziskovalnih institucij ter industrije.

V prvih štirih okvirnih programih (od 1983 do 1998) so bile raziskovalne aktivnosti na področju jezikovnih tehnologij razpršene na več različnih programov (ESPRIT, Telematics, TIDE, DRIVE, DELTA, AIM, ACTS, PECO, INCO-Copernicus, INCO-DC, INTAS…). Takšna razdrobljenost projektov po posameznih programih je nujno vodila k nesmotrnemu izkoriščanju raziskovalnih potencialov, precejšnji netransparentnosti raziskovalnih rezultatov in v dobršni meri tudi nezmožnosti izkoriščanja sinergičnega učinka. Iz teh potreb se je razvila iniciativa po ustanovitvi enovitega raziskovalnega programa jezikovnih tehnologij v letu 1998. V okviru petega okvirnega programa, ki se je pričel v letu 1999, je bilo ustanovljeno posebno akcijsko področje jezikovnih tehnologij (Human Language Technologies). Informacije o dogajanju na tem področju je moč najti na spletnih straneh HLT Central (http://www.hltcentral.org). V letu 2002 je na tej domači strani navedenih 217 projektov, od tega jih je v okviru petega okvirnega programa na področju HLT navedenih 53. Celoten proračun posameznih projektov je od 2 do 5 milijonov evrov, medtem ko delež sredstev Evropske unije običajno ne presega dveh do treh milijonov evrov.

V primerjavi z izvajanjem raziskovalnih projektov v ZDA so bili tudi v okviru petega okvirnega programa projekti mnogo manjši po obsegu, vključevali so manj partnerjev in običajno trajali krajši čas. Nekatere slabosti, ki so se pokazale pri takšni organizaciji raziskovalnih aktivnosti, predvsem razdrobljenost raziskovalnih moči in nedoseganje zadovoljivega sinergičnega učinka ter želenega prenosa rezultatov raziskovalnih projektov v prakso, bodo skušale biti presežene v okviru šestega okvirnega programa.

4. Šesti okvirni program EU in področje jezikovnih tehnologij

Izvajanje šestega okvirnega programa predvideva organizacijo tako imenovanih integralnih projektov (Integrated Projects – IP), ki bodo trajali do pet let s skupnim proračunom posameznih projektov v poprečju od 30 do 50 milijonov evrov (predviden prispevek EU). Takšni projekti bodo po obsegu, trajanju in vrednosti primerljivi s projekti, ki jih izvajajo v ZDA. Eden osnovnih namenov takšne organizacije projektov je gotovo združitev raziskovalnih potencialov evropskega prostora na posameznem raziskovalnem področju in doseganje kritične mase raziskovalnih moči, hkrati pa bo moč v mnogo večji meri računati tudi s sinergičnim učinkom znotraj projektov. Cilji projektov sedaj ne bodo več ozko usmerjeni v pokrivanje potreb posameznih ciljnih aplikacij, pač pa bodo skrbeli predvsem za nadaljnji razvoj ene ali več tehnologij – kar bo največkrat predstavljalo horizontalne nivoje organiziranosti integralnega projekta, vertikalne nivoje pa bodo predstavljali izbrani projekti razvoja ciljnih aplikacij, ki bodo v večjem ali manjšem obsegu vključevali razvite tehnologije. Obseg in trajanje projekta bosta omogočala tudi načrtovanje in izgradnjo infrastrukture (na primer jezikovnih virov) in kvalitetno vrednotenje raziskovalnih dosežkov. Veliko večji bo tudi poudarek na izvajanju temeljnih raziskav.

Projekt bo upravljal konzorcij projekta, ki se bo lahko sproti odločal, katera področja znotraj projekta zahtevajo večji delež raziskovalnih aktivnosti, in zato v času izvajanja integralnega projekta razpisal nove projekte ter povabil k sodelovanju nove partnerje. Razen partnerjev, ki bodo člani konzorcija integralnega projekta, bodo drugi partnerji običajno vključeni v projekt le krajši čas, saj bodo izvajali raziskave v okviru projektov, katerih trajanje bo običajno krajše od petih let (npr. od enega do treh let).

Očitne prednosti, ki jih ima takšen način izvajanja raziskovalnih aktivnosti v okviru šestega okvirnega programa, pa na drugi strani v mnogočem postavljajo raziskovalne skupine, ki doslej še niso sodelovale v projektih okvirnih programov, še posebej raziskovalne skupine iz držav srednje in vzhodne Evrope, v težji položaj. Veliko število projektov, sicer z manjšim proračunom, ki so jih izvajali v prejšnjih okvirnih programih, je namreč pomenilo tudi veliko število partnerjev projektov oziroma veliko število različnih »gremijev«. S tem pa tudi morda več možnosti za nove skupine, da se priključijo posameznim projektom. Večji projekti in s tem bistveno večja odgovornost članov konzorcija integralnih projektov pa vsiljujeta misel, da bodo le-ti k sodelovanju vabili in izbirali partnerje, s katerimi so v preteklosti že sodelovali in ki jim že zaupajo. Takšen pristop lahko že opazimo pri več skupinah, ki so izvajale projekte v okviru četrtega in petega okvirnega programa.

Da bi se slovenske raziskovalne skupine na področju jezikovnih tehnologij lažje in enakopravno vključile v aktivnosti prihajajočega šestega in naslednjih okvirnih programov, je nujno obdržati mednarodno primerljivost znanstvenoraziskovalnega dela (stalna prisotnost s prispevki na najpomembnejših znanstvenih konferencah) in doseči čim večjo mednarodno primerljivost ter verificiranost jezikovnih virov, ki morajo biti na voljo širši strokovni in znanstveni javnosti preko mednarodnih organizacij (npr. ELRA/ELDA).

V pripravah na izvajanje projektov v okviru šestega okvirnega programa se oblikuje nekaj iniciativ, ki bodo pokrivale veliko večino področij jezikovnih tehnologij. Ena takšnih iniciativ je tudi skupina projektov *C-STAR, ki delno potekajo že v okviru petega okvirnega programa in pomenijo v dobršni meri pripravo na izvajanje integralnega projekta v okviru šestega okvirnega programa.

4.1. Projekt LC-STAR

Projekt LC-Star (Lexica and Corpora for Speech to Speech Translation) (Höge, 2002)(Lazzari, 2000) je projekt petega okvirnega programa. Celotni proračun projekta, ki bo trajal od 2002 do 2004, je 3.3 milijona evrov. Njegov glavni cilj je zagotoviti jezikovne vire, potrebne pri razvoju tehnologije avtomatskega simultanega prevajanja govora (speech-to-speech translation), ki naj bi po nekaterih napovedih postala realnost v naslednjih petih letih. Potrebne vire predstavljajo predvsem glasoslovni slovarji (vsebujejo ortografsko in fonetično transkripcijo besed, zapisano v standardni obliki), dvojezični besedni slovarji, oblikoslovni slovarji in poravnani označeni dvojezični besedilni korpusi. Glavni cilj projekta je zagotoviti te vire za čimveč evropskih jezikov in definirati priporočila za izgradnjo in strukturo omenjenih jezikovnih virov. Cilj projekta je v tem smislu torej podoben cilju projekta SpeechDat II (Van den Heuvel, 1997) (Höge, 1999) (Moreno, 2002), katerega priporočila so danes postala neformalni standard za načrtovanje in vrednotenje baz izgovarjav. V okviru projekta bodo izvedeni jezikovni viri za naslednjih 12 jezikov: turški, ruski, italijanski, grški, španski, katalonski, nemški, klasični arabski, hebrejski, ameriška angleščina, finski in kitajski. Predviden obseg jezikovnih virov za posamezen jezik je nekaj deset tisoč enot, npr. slovar 50.000 besed, slovar 50.000 lastnih imen, zasnova slovarja, primernega za razvoj sistemov avtomatskega razpoznavanja govora s približno 100.000 besedami, in zasnova podobnega slovarja za razvoj sistemov avtomatske sinteze govora. Posebno aktivnost predstavljata vrednotenje in verifikacija izdelanih virov, ki jo bo opravljala posebej za to izbrana institucija.

Za izbrane pare jezikov bodo zgrajeni tudi eksperimentalni dvojezični jezikovni viri, katerih namen bo določiti vrsto, strukturo in obseg potrebnih jezikovnih virov. Zgrajen bo tudi eksperimentalni sistem za avtomatsko simultano prevajanje, ki bo uporabljal že obstoječe sisteme jezikovnih tehnologij (razpoznavanje, prevajanje, sinteze), ki v tem primeru predstavljajo komponente sistema avtomatskega simultanega prevajanja (speech-to-speech translation components – SST komponente). Uporabljen bo za določanje vpliva vrste, obsega in kvalitete uporabljenih jezikovnih virov na uspešnost prevajanja. Pri tem bodo v veliki meri uporabljeni izsledki in sistemi projekta Verbmobil (Wahlster, 2000). Trenutni izbrani pari jezikov so: katalonščina/ameriška angleščina, španščina/katalonščina in španščina/ameriška angleščina.

4.2. Projekt TC-STAR

Projekt TC-STAR (technology and corpora for speech to speech translation) bo predlagan kot integralni projekt v okviru 6 okvirnega programa in bo v primeru potrditve pričel z aktivnostmi leta 2003 (Höge, 2002). V okviru petega okvirnega programa že teče enoletni projekt TC-STAR_P, katerega glavni namen je priprava potrebne dokumentacije in predloga projekta TC-STAR. Pri tem bodo industrijski partnerji podali predviden razvoj tehnologije in storitev, ki bodo uporabljale SST komponente, raziskovalne skupine univerz in raziskovalne institucije predviden razvoj SST tehnologije v naslednjem petletnem obdobju, prav tako pa bo izdelan najustreznejši model vodenja in upravljanja integralnega projekta.

Predlagani čas trajanja projekta TC-STAR je 5 let. Glavni cilj projekta je preseči jezikovne pregrade in razviti tehnologijo avtomatskega simultanega prevajanja. Za dosego tega cilja so predvidena štiri akcijska področja: a) temeljne raziskave z nalogo bistvenega izboljšanja uspešnosti SST komponent (za faktor 4), b) razvojne raziskave z namenom prilagoditve SST tehnologij zahtevam področij ciljnih aplikacij in razvoj SST komponent ter sistemov na posameznih platformah, c) razvoj aplikacij na aplikacijsko odvisnih platformah ob uporabi TTS komponent in d) definicija in izgradnja jezikovnih virov ter vrednotenje SST komponent in sistemov.

Predlagan celotni proračun projekta je okrog 100 milijonov evrov, od česar naj bi evropska skupnost prispevala nekaj manj kot polovico predlaganih sredstev.

Petino sredstev pa naj bi zagotovile nacionalne organizacije (vlade) iz držav partneric projekta.

5. Potrebe po združevanju raziskovalnih potencialov v slovenskem prostoru

Dosedanja razdrobljenost in precejšnja nekoordiniranost raziskovalnih potencialov na področju jezikovnih tehnologij v slovenskem prostoru ne daje najboljših obetov za konkurenčno raziskovanje in možnosti uspešnejšega vključevanja v projekte okvirnih programov in druge mednarodne projekte. Za doseganje napredka na tem področju je nujno potrebno širše usklajevanje raziskovalnih interesov, saj so te smernice zelo močno prisotne tudi v okviru izvajanja integralnih projektov v prihajajočem šestem okvirnem programu, kar je moč zaznati tudi iz opisanih projektov LC- in TC-STAR. V prvi vrsti bi bilo zato potrebno preveriti pripravljenost skupin za bolj usklajeno in povezano raziskovalno dejavnost ter nato zbrati srednjeročne in dolgoročne usmeritve ter cilje raziskovalnih aktivnosti posameznih raziskovalnih skupin in skušati najti njihov presek in komplementarnost. S tem bi lahko prišli korak bliže k doseganju sinergičnega učinka med raziskovalnimi skupinami v slovenskem prostoru. Vsekakor pa pri tem ne gre zanemariti tudi potreb, ki se že kažejo na tem zelo zanimivem segmentu žal precej majhnega slovenskega tržišča. Zaradi majhnosti našega raziskovalnega potenciala je namreč podvajanje raziskovalnih aktivnosti verjetno zelo nesmotrno. V slovenskem prostoru obstaja kar nekaj organizacij, ki bi lahko prevzele vlogo koordinatorja ali vsaj vodile forum za usklajevanje raziskovalnih interesov ter skušale prispevati k bolj povezanemu in enotnejšemu nastopu v nacionalnem in mednarodnem merilu. V prvi vrsti bi aktivnejšo vlogo pri oblikovanju takšnega foruma lahko prevzelo Slovensko društvo za jezikovne tehnologije ali katera izmed institucij, na katerih delujejo raziskovalne skupine, ki se ukvarjajo s področjem jezikovnih tehnologij. Tako je Center za jezikovne tehnologije, ki deluje na Fakulteti za elektrotehniko, računalništvo in informatiko že od leta 1994 in ima mnogo izkušenj pri gradnji jezikovnih virov in sodelovanju v domačih in mednarodnih projektih četrtega in petega okvirnega programa, tudi pripravljen prevzeti takšno vlogo.

6. Sklep

Raziskovalne aktivnosti v okviru četrtega in petega okvirnega programa EU so bile v veliki meri usmerjene k razvoju jezikovnih virov, potrebnih za razvoj sistemov jezikovnih tehnologij, in tudi k razvoju sistemov samih. Priprave na predlaganje in izvajanje projektov v okviru šestega okvirnega programa kažejo na težnjo po združevanju vseh področij jezikovnih tehnologij in gradnji zelo zahtevnih sistemov (npr. avtomatsko simultano prevajanje govora). Pri tem predlogi v veliki meri upoštevajo že obstoječe jezikovne vire, zgrajene v predhodnih projektih, in gradijo nadaljnje raziskave na njihovi uporabi. Jeziki majhnih narodov in narodov srednje in vzhodne Evrope pa se srečujejo z nevarnostjo vse večjega izločanja iz teh dogajanj, saj bodo zaradi običajno slabo razvitih domačih raziskovalnih aktivnostih in relativno majhne udeležbe v preteklih okvirnih programih ter neobstoječih osnovnih jezikovnih virov vse težje izpolnjevali kriterije za aktivno in enakopravno vključevanje v takšne projekte. Pri tem razmere v slovenskem prostoru žal niso izjema. Precejšnja razdrobljenost raziskovalnih aktivnosti, kar gotovo velja za področje govorne tehnologije, in nepovezanost raziskovalnih skupin sta doslej onemogočali vključevanje v evropske raziskovalne projekte v večjem obsegu, saj posamezne skupine običajno niso premogle kritične mase raziskovalcev. Če pa se želimo v prihodnje pogosteje vključevati v aktivnosti šestega in prihodnjih okvirnih programov, pa bo nujnejše tesnejše sodelovanje raziskovalnih skupin, ne samo znotraj posameznih raziskovalnih področij, pač pa vseh raziskovalnih skupin in drugih ustanov, ki v slovenskem prostoru delujejo na področju jezikovnih tehnologij.

7. Viri

Höge, H., 2002. Project Proposal TC-STAR. V: *Zbornik LREC 2002*, 1:136-141.

Mariani, J., 2002. Are we losing ground to the US?. V: *HLT Central, www.hltcentral.org*.

Höge, H in drugi, 1999. SpeechDat multilingual speech databases for teleservices: Accross the finish line. V: *Zbornik Eurospeech 1999*, 6:2699-2702.

Lazzari, G., 2000. Spoken translation: challenges and opportunities. *V: Zbornik ICSLP 2000.*

Lavie, A. in drugi, 2002. Enhancing the Usability and Performance of NESPOLE! – a Real-World Speech-to-Speech Translation System. *V: Zbornik HLT 2002*..

Moreno, A. in drugi, 2002. SpeechDat across all America: SALA II, V: Zbornik LREC 2002, 1:16-20.

Van den Heuvel, H. 1997. Validation criteria for databases. *SpeechDat Technical Report, SD1.3.3*.

Wahlster, W., 2000. VERBMOBIL. Verbmobil final symposium. Saarbruecken, 2000.

Erjavec, T., Gorjanc, V., Stabej, M. (1998) Korpus FIDA Zbornik konference Jezikovne tehnologije za slovenski jezik, Ljubljana, pp. 124-127.

Erjavec, T. (1998a) The MULTEXT-East Slovene Lexicon. Zbornik konference ERK`98, Portorož, Slovenija, pp. 189-192.

Gros, J., Mihelič, F., Dobrišek, S., Erjavec, T., Žganec M. (2000) Corpora of Slovene Spoken Language for Multi-lingual Applications. Zbornik LREC'2000. Athens, str. 953-956.

Kačič, Z., Horvat, B., Zogling S. (2000) Issues in Design and Collection of Large Telephone Speech Corpus for Slovenian Language. Zbornik LREC'2000, str. 943-946.

Kaiser., J., Kačič, Z. (1998) Development of Slovenian SpeechDat Database. Proceedings of the Workshop on Speech Database Development for Central and Eastern European Languages. Granada. Spain.

Romih, M. (1998) Amebis in jezikovne tehnologije. Zbornik konference Jezikovne tehnologije za slovenski jezik, Ljubljana, pp. 29-34.

Rojc, Matej, Kačič, Zdravko, Verdonik, Darinka (2002). Design and implementation of the Slovenian phonetic and morphology lexicons for the use in spoken language applications. Zbornik: LREC'2002, vol. 4, str. 1296-1300.

Hozjan, Vladimir, Kačič, Zdravko. (2002) Objective analysis of emotional speech for English and Slovenian interface emotional speech databases. Zbornik: LREC 2002, vol. 6, str. 2019-2023.