

Jezikovni viri projekta LC-STAR

Darinka Verdonik, Matej Rojc

Fakulteta za elektrotehniko, računalništvo in informatiko
Smetanova ul. 17, 2000 Maribor
darinka.verdonik@uni-mb.si

Povzetek

V članku predstavljamo jezikovne vire, ki se razvijajo v mednarodnem projektu LC-STAR, s podrobnejšo predstavitevijo jezikovnih virov za slovenščino, ki jih v okviru projekta razvija Fakulteta za elektrotehniko, računalništvo in informatiko Univerze v Mariboru. Jezikovni viri so v osnovi dveh vrst: veliki slovarji za izboljšanje razpoznavne in sinteze govora (113.000 besed) ter poravnani slovarji fraz in slovarji besed s področja turizma za izboljšanje komponente govorno orientiranega prevajanja. Kakovost razvitih jezikovnih virov se zagotavlja z natančnimi določili in neodvisno validacijo, ki jo opravljata SPEX in CST. Razviti jezikovni viri naj bi odpravili pomanjkljivosti na tem področju. Viri bodo na voljo prek ELRE/ELDE.

1. Uvod

Namen mednarodnega projekta LC-STAR (Lexica and Corpora for Speech to Speech Translation Components) (www.lc-star.com) je razviti jezikovne vire, ki so potrebni za izboljšanje komponent strojnega simultane prevajanja govora. Te so (Hoege, 2002):

- razpoznavna govora z velikim slovarjem
- visoko kvalitetna sinteza govora
- besedilno orientirano strojno prevajanje

V projektu so se združili pomembni industrijski in akademski partnerji: Siemens AG, IBM, Nokia, NSC (Natural Speech Communication Ltd.), UPC (Universitat Politècnica de Catalunya), RWTH University of technology. Univerza v Mariboru se je projektu pridružila kot zunanji partner, ki bo zagotovil jezikovne vire po standardih projekta za slovenski jezik.

Po ugotovitvah v LC-STAR (Hartikainen et al., 2003) so glede na široko sprejete pristope k razpoznavanju govora (skriti modeli Markova) in sinteze govora (konkatenativna sinteza) zahteve glede vsebine jezikovnih virov dokaj ustaljene in jasne: osrednje načelo je zagotoviti velike slovarje besed, ki pokrivajo številna različna področja rabe, in korpuse govornih besedil. Vendar pri obstoječih virih konzorcij LC-STAR ugotavlja predvsem naslednje pomanjkljivosti:

- premajhno pokritost po različnih aplikacijskih področjih
- nezadostno prilagojenost za sintezo in razpoznavo govora
- nezadostno kvaliteto
- pomanjkanje standardov
- slabo pokritost po jezikih

Projekt LC-STAR naj bi odpravil te pomanjkljivosti jezikovnih virov za komponente strojnega simultane prevajanja govora, s tem da bodo razviti visoko kvalitetni jezikovni viri za številne svetovne jezike po skupnih, v projektu postavljenih standardih. Jeziki, pokriti v okviru projekta LC-STAR, so: turški, ruski, italijanski, grški, španski, katalonski, nemški, klasični arabski, hebrejski, ameriška angleščina, finski, kitajski, slovenski.

V nadaljevanju članka najprej na kratko pregledamo, kateri jezikovni viri za razvoj komponent sistemov strojnega simultane prevajanja govora so razviti za slovenščino, nato pa natančneje predstavljamo jezikovne vire, ki se razvijajo v okviru projekta LC-STAR, s poudarkom na slovenskih jezikovnih virih, ki nastajajo v okviru projekta na Univerzi v Mariboru.

2. Slovenski jezikovni viri, uporabni za razvoj komponent strojnega simultane prevajanja govora

Jezikovni viri, potrebni za razvoj komponent strojnega simultane prevajanja govora, so v splošnem dvoji: oblikoslovni in glasoslovni slovarji ter različni korpusi govornih besedil. Zaradi časovne in finančne zahtevnosti pridobivanja in urejanja korpusov govornih besedil se sicer velikokrat uporabljajo kar korpusi pisnih besedil, čeprav zaradi razlik med govornim in pisanim besedilom to ni najustreznejša rešitev za razvoj kvalitetnih sistemov strojnega simultane prevajanja govora.

Od dostopnih slovarjev za slovenski jezik imamo tako slovar lastnih imen Onomastica (Kačič, 1997), ki ga distribuirala ELRA/ELDA. Ta organizacija ponuja tudi govorno bazo posnetkov SpeechDat (II) (Kaiser, Kačič, 1998) za slovenski jezik, ki je prvi jezikovni vir za slovenski jezik, preverjen v mednarodnem centru za validacijo SPEX. Žal je oba jezikovna vira treba plačati. Od prosto dostopnih govornih baz je na voljo MobiLuz, govorna baza poizvedovanj o letalskih informacijah, ki vključuje tudi fonetični prepis in oblikoslovne oznake (Gros et al., 2000), v raziskovalne namene je mogoče dobiti tudi oblikoslovni slovar slovenskega jezika, ki je nastal v okviru projekta Multext-East (nl.ijs.si/ME), ta obsega 15.000 lem. Ostali jezikovni viri, ki se lahko uporabijo za razvoj sistemov strojnega simultane prevajanja slovenskega govora, so bolj ali manj narejeni za lastno uporabo v organizacijah, ki so sodelovale pri njihovem nastajanju, so različno obsežni, vsebujejo različne podatke in so različno kvalitetni. Govorne baze za slovenski jezik so tako še Snabi (Kačič, 2002), Polidat (Zoegling Markuš, Kačič, Horvat, 2000), Luz (Gros, 1996), Gopolis (Dobrišek et al., 1998), govorna zbirka vremenskih napovedi (Žibret, Mihelič, 2000). Od slovarjev je za slovenski jezik narejen še en večji oblikoslovni slovar, Simlex (Verdonik et al., 2002; Rojc, Kačič, 2003) z 20.000 lemami, od večjih glasoslovnih slovarjev pa Siflex, ki vsebuje fonetični prepis vseh besed iz Simlexa (ibid.). Drugih večjih oblikoslovnih ali glasoslovnih slovarjih ni zaslediti, obstaja pa še nekaj manjših za potrebe sintetizatorjev govora oz. za študije

samodejnega napovedovanja naglasnega mesta (Šef, 2002). Večjega, referenčnega korpusa govorjenih besedil (eno-, dvo- ali večjezičnega) za slovenski jezik nimamo, lahko pa zasledimo objavo, ki napoveduje njegovo gradnjo (Stabej, Vitez, 2000). Boljše je stanje pri pisnih korpusih (enojezična FIDA (www.fida.net), Nova Beseda (bos.zrc-sazu.si/s_beseda), paralelni korpus ELAN (nl.ijs.si/elan/) in drugi), vendar so ti, kot rečeno, lahko samo izhod v sili, kadar govorimo o razvoju komponent strojnega simultanelega prevajanja govora. Zaradi razlik med govorjenim in pisnim besedilom je strojno simultano prevajanje govora drugačna naloga kot strojno prevajanje (pisanega) besedila.

Navedeno kaže, da podobne pomanjkljivosti, kot jih ugotavljajo v LC-STAR za jezikovne vire nasploh, veljajo tudi za slovenske jezikovne vire, potrebne za razvoj komponent strojnega simultanelega prevajanja govora.

3. Jezikovni viri, razviti v okviru projekta LC-STAR

Projekt LC-STAR poteka vzporedno v dveh smereh.

1. del:

- zgraditi seznam besed, ki bo pokrival številna različna področja rabe,
- določiti standarde in zgraditi slovarje za razpoznavanje govora z velikim slovarjem in visoko kvalitetno sintezo govora; slovarji bodo vsebovali oblikoslovne oznake besed in fonetični prepis.

2. del:

- raziskati govorno orientirano prevajanje, pri tem je poudarek na preučitvi potrebnih jezikovnih virov,
- izgradnja demonstracijskega sistema strojnega simultanelega prevajanja govora za španščino in angleščino,
- specifikacija in izgradnja jezikovnih virov, potrebnih za govorno orientirano prevajanje. Ti viri bodo:
 - 3-jezični vzporedni korpus pogovorov s področja turizma z vsaj 500.000 besedami,
 - vzporedni slovarji 10.000 fraz s področja turizma,
 - oblikoslovni in glasoslovni slovarji besed iz vzporednih slovarjev fraz za 9 jezikov.

Univerza v Mariboru je kot zunanja partnerica, odgovorna za izgradnjo jezikovnih virov za slovenski jezik, udeležena predvsem v prvem delu projekta, v katerem bodo za 13 jezikov zgrajeni oblikoslovni in glasoslovni slovarji z vsaj 100.000 enotami, od tega slaba polovica lastnih imen, pa tudi v drugem delu projekta pri pripravah vzporednih slovarjev fraz in leksikonov besed iz teh fraz.

V nadaljevanju podrobneje opisujemo najprej velike slovarje besed za sintezo in razpoznavanje govora ter zatem slovarje za govorno orientirano prevajanje za slovenski jezik.

3.1. Slovarji za izboljšanje razpoznave in sinteze govora za slovenski jezik

Slovarji za izboljšanje razpoznave in sinteze govora oz. t.i. veliki slovarji so zgrajeni iz treh enot: občnih besed, lastnih imen in aplikacijskih besed, tj. besed, potrebnih za delovanje govorno vodenih aplikacij s šestih semantičnih področij (splošno, finance, potovanja, posredovanje informacij, upravljanje in storitve, telekomunikacije). Dodatno so bile k občinim besedam v celoti zbrane besedne vrste, ki imajo končno število besed (predlogi, vezniki, zaimki...) - t.i. zaprti nabori. Natančnejši prikaz velikosti posameznih enot slovarja je v tabeli 1.

ENOTA SLOVARJA	VELIKOST
občne besede	65.096
zaprti nabor	1.146
lastna imena	45.027
aplikacijske besede	6.040

Tabela 1: Velikost posameznih enot velikega slovarja LC-STAR za slovenski jezik.

3.1.1. Občne besede

Občne besede so bile izbrane iz 12-milijonskega korpusa, v katerem so bila zbrana besedila iz 6 večjih področij: šport, novice, kultura in zabava, gospodarstvo in finance, potrošniške informacije, osebne komunikacije. Glavna vira za korpus sta bila časopisa Večer in Delo, za potrošniške informacije so bila dodana besedila iz nekaterih prilog k tema časopisoma, nekaj besedil revije Gea, poljudnoznanstvena besedila s portala Svarog in nekateri priročniki z interneta. Natančna struktura korpusa je prikazana v tabeli 2.

Področje	Viri	Leto objave
Šport	<i>Večer</i> , rubrike: <i>Šport, Prosti strel</i>	1997-2002
Novice	<i>Večer</i> , rubrike: <i>Prva stran, Črna kronika, Kronika, Maribor, Maribor okolica, Zadnja stran, Zunanja politika, Doma in po svetu, Slovenska kronika, V žarišču</i>	1997-2002
Finance/ gospodarstvo	<i>Večer</i> , rubrike: <i>Gospodarstvo, Finance</i>	1997-2002
Kultura/ zabava	<i>Večer</i> , rubrike: <i>Kultura, Čitalnica, Film, Borštnikovo, Lent, Reportaže zanimivosti, Reportaža, Potovanja</i>	1997-2002
Potrošniške informacije	<i>Večer</i> , rubrike: <i>Zdravje, Zdravstvo, Univerza znanost, Računalništvo, Arhitekturna beseda, priloge Raziskovalec, Zdravje, priloga Dela Znanost, revija Gea, novice iz Svaroga, priročniki z interneta</i>	1997-2003
Osebne komunikacije	<i>Večer</i> , rubrike: <i>Pisma bralcev, Pisma, Vaši odmevi, Vaša mnenja odmevi</i>	1997-2002

Tabela 2: Viri za korpus občnih imen in leto publikacije.

Korpus je bil tokeniziran, odstranila so se ločila, cifre, najpogostejši znaki (\$, &, % itd.). Nato so bila avtomatsko odstranjena lastna imena, okrajšave in vsi singletoni (besede, ki se pojavijo samo enkrat), nato pa preostala lastna imena in okrajšave še ročno. Tako urejen korpus je

šel za t.i. čisti korpus. Velikost takega korpusa za slovenski jezik je prikazana v tabeli 3.

Področje	Velikost	Različne besede
Šport	1.888.753	42.029
Novice	2.178.834	63.912
Finance/gospodarstvo	3.411.268	62.837
Kultura/zabava	2.716.028	82.305
Potrošniške informacije	1.146.009	48.326
Osebnе komunikacije	950.179	40.500
Skupaj	12.291.071	139.645

Tabela 3: Velikost čistega korpusa.

Iz tega korpusa je bil izločen seznam besed za nadaljnjo obdelavo tako, da je bila dosežena pokritost izvornega korpusa najmanj 95 %. Zaradi varovala smo izbrali nekoliko višji odstotek. Natančni podatki o velikosti seznama občnih besed so v tabeli 4.

Področje	Različne besede	
	Št. besed	Pokritost
Šport	17.256	96%
Novice	31.479	96%
Finance/gospodarstvo	24.469	96%
Kultura/zabava	40.847	96%
Potrošniške informacije	28.418	96%
Osebnе komunikacije	23.923	96%
Skupaj	65.096	98,11%

Tabela 4: Velikost seznama občnih besed po številu besed ter pokritost izvornega korpusa po posameznih področjih in skupaj.

3.1.2. Zaprti nabori

Ker se zlasti funkcijske besede pogosto uporabljene na različnih področjih, s korpusom pa niso zajete vse, je bil dodatno k seznamu občnih besed ročno zbran seznam funkcijskih besed, ki jih je končno število. Naj poudarimo, namen tega zbiranja je bil samo izpopolniti seznam občnih besed, pridobljen iz korpusa, s funkcijskimi besedami, ki se pogosto uporabljajo. Te besede so bile zbrane v 13 skupinah glede na kasnejše oblikoslovno označevanje in so naslednje:

- predlogi
- vezniki
- členki pritrjevanja in zanikanja
- kazalni zaimki
- nedoločni, nikalni, poljubnostni, mnogostni, totalni, drugostni, istostni, celostni zaimki
- vprašalni zaimki
- osebni zaimki
- osebni svojilni zaimki
- osebni in svojilni povratni zaimki
- oziralni in oziralno poljubnostni zaimki
- pomožni glagol *biti* v vseh oblikah
- naklonski glagoli
- fleksijske končnice (ta nabor ni del slovarja, zbran je bil samo informativno)

Tako je bilo skupaj zbranih nekaj več kot 1000 besed, od teh jih je bilo nekaj več kot 700 že zbranih s korpusom. Na ta način je bilo torej dejansko dodanih nekaj več kot 300 besed h končnemu slovarju.

3.1.3. Lastna imena

V projektu LC-STAR je veliko pozornosti namenjene tudi lastnim imenom, saj so pogosto uporabljena v različnih aplikacijah, njihovo izgovorjavo pa je težko pravilno določiti, zlasti pri tujih lastnih imenih. Ker pa je zanesljiva razpoznavna in sinteza zelo pomembna za prihodnje govorne aplikacije, je nabor lastnih imen v projektu velik (Ziegenhain et al., 2004).

Nabor slovenskih lastnih imen šteje 45.027 imen in je v skladu s specifikacijami projekta razdeljen na tri večje enote: osebna imena, krajevna imena, imena organizacij. Tabela 5 prikazuje natančnejšo strukturo enot in velikost slovenskega slovarja lastnih imen.

Področje	Velikost	%	Podpodročje
Osebna imena	22.469	49,2	
Zemljepisna imena	11.828	25,9	mesta
			gore, reke in druge zemljep. enote
			glavna mesta
			velika in pomembna mesta
			pomembni turistični in drugi kulturni spomeniki
			ulice
			države
Organizacije	11.357	24,9	pridobitne in nepridobitne organizacije, podjetja
Skupaj št. vnosov	45.654	100	trgovske znamke
Skupaj št. razl. vnosov	45.027		

Tabela 5: Lastna imena: razdelitev po področjih in po velikosti.

Večbesedna lastna imena, kot so Stari_trg, Novo_mesto, štejejo kot en vnos. Nekatera lastna imena so lahko hkrati razvrščena v več področij (npr. Jelka je lahko ime trgovine ali osebno ime), takih vnosov je v slovarju čez 600. Zbrana lastna imena so takšna, kot so v rabi v slovenskem prostoru. Pri zemljepisnih lastnih imenih se pojavi precej tujih imen predvsem za kraje, za katere nimamo poslovenjenega imena, tuja imena vključena tudi z imeni večjih tujih korporacij in trgovskih znamk, veliko besed tujega izvora pa je tudi v imenih sicer slovenskih podjetij.

Viri, iz katerih so bile pridobljene besede, in vrsta virov so opisani v tabeli 6.

Področje	Viri	Vrsta vira
Osebna	Onomastica	CD

imena		
Zemljepisna imena	Krajevni leksikon Slovenije	CD
	Veliki družinski atlas	ročno
	Atlas Slovenije	ročno
	Geografski atlas sveta za šole	ročno
	Slovenija: turistični vodnik	ročno
	Interaktivni atlas Slovenije	internet
	Onomastica	CD
Organizacije	Telefonski imenik Slovenije: Pomlad 2002	CD
	Onomastica	CD
	internet, trgovine	internet, ročno

Tabela 6: Viri in vrsta virov za pridobivanje seznama lastnih imen.

3.1.4. Aplikacijske besede

Seznam aplikacijskih besed je dvojega izvora:

- števila, črke in okrajšave iz korpusa občnih besed;
- referenčni seznam 5700 besed v ameriški angleščini, preveden v slovenščino, ki pokriva 6 večjih aplikacijskih področij: globalna področja (mere, znaki, okrajšave...), finančne storitve, potovanja, posredovanje informacij, kontrola, telekomunikacije. Seznam je zagotovil konzorcij LC-STAR.

Skupaj je bilo tako zbranih 6040 besed, od teh jih je bilo 3510 že zbranih s korpusom občnih imen. S seznamom aplikacijskih besed je bilo torej dodanih 2530 besed k skupnemu končnemu naboru.

Skupaj - občne besede, zaprti nabori, lastna imena in aplikacijske besede - je bilo tako zbranih okoli 113.000 besed, ki predstavljajo vnose za veliki slovar, namenjen za izboljšanje razpoznavne in sinteze govora.

3.1.5. Oblikoslovno označevanje in fonetični prepis

Po zgoraj opisanih postopkih zbrane besede se oblikoslovno označijo in fonetično prepišejo v skladu s standardi, določenimi v projektu (Hartikainen et al., 2003).

Za fonetični prepis je uporabljena fonetična abeceda Sampa (Zemljak et al., 2002). Besede so zlogovane; znak za zlog je vezaj (-). Pri večbesednih vnosih je premor med besedami označen z višajem (#). Naglas je določen za celoten zlog, označen je z narekovajem pred zlogom ("). Vsi fonemi in znaki za zlog, premor in naglasno mesto so ločeni med seboj s presledkom. Fonetični prepisi so bili narejeni polavtomatsko ter ročno pregledani in popravljeni.

Oblikoslovne oznake so natančneje predstavljene že v (Verdonik et al., 2004), tukaj predstavljamo le glavne značilnosti.

Osrednje kategorije besed so: samostalnik, pridevnik, števnik, zaimek, glagol, pomožni glagol, prislov, veznik, predlog, členek, medmet. Za vse sklonljive kategorije (samostalnik, pridevnik, števnik, zaimek) so določeni atributi sklon, spol, število, za glagol in pomožni glagol pa poleg spola in števila tudi oseba, naklon (povedni,

pogojni, velelni, nedoločnik, deležnik), čas, način in vid. Lastna imena so kategorizirana kot samostalnik, sledi atribut za oznako vrste lastnega imena (osebno, zemljepisno, država, mesto, ulica, organizacija, trgovska znamka, turistični spomeniki). Oznak za spol, sklon ali število pri lastnih imenih ni, tudi ko so enobesedna. Samostalnikom moškega spola je v 4. sklonu pripisana oznaka za živost ali neživost, pri pridevniki in prislovih je označen atribut o stopnji, števniki so dodatno označeni kot glavni, vrstilni, množilni in ločilni, zaimki so razdeljeni na osebne, kazalne, oziralne, nedoločne, vprašalne, povratne in svojilne, prislovi pa na prislove časa, kraja in načina.

Tako urejen slovar je zakodiran v jeziku XML. Struktura slovarja je razvidna iz sledečega izseka:

```
<ENTRYGROUP orthography="Abramičeva_ulica">
<ENTRY>
  <NOM class="STR" />
  <LEMMA>Abramičeva_ulica</LEMMA>
  <PHONETIC>a - " b r a : - m i - t S E - v a # " u : - l i - t s a
  </PHONETIC>
</ENTRY>
</ENTRYGROUP>
<ENTRYGROUP orthography="agenciji">
<ENTRY>
  <NOM class="common" number="dual" gender="feminine"
  case="nominative" />
  <LEMMA>agencija</LEMMA>
  <PHONETIC>a - " g e : n - t s i - j i </PHONETIC>
</ENTRY>
<ENTRY>
  <NOM class="common" number="singular" gender="feminine"
  case="dative" />
  <LEMMA>agencija</LEMMA>
  <PHONETIC>a - " g e : n - t s i - j i </PHONETIC>
</ENTRY>
<ENTRY>
  <NOM class="common" number="dual" gender="feminine"
  case="accusative" />
  <LEMMA>agencija</LEMMA>
  <PHONETIC>a - " g e : n - t s i - j i </PHONETIC>
</ENTRY>
<ENTRY>
  <NOM class="common" number="singular" gender="feminine"
  case="locative" />
  <LEMMA>agencija</LEMMA>
  <PHONETIC>a - " g e : n - t s i - j i </PHONETIC>
</ENTRY>
<ENTRY>
  <NOM class="common" number="dual" gender="feminine"
  case="nominative" />
  <LEMMA>agencija</LEMMA>
  <PHONETIC>a - g E n - " t s i : - j i </PHONETIC>
</ENTRY>
<ENTRY>
  <NOM class="common" number="singular" gender="feminine"
  case="dative" />
  <LEMMA>agencija</LEMMA>
  <PHONETIC>a - g E n - " t s i : - j i </PHONETIC>
</ENTRY>
```

```

<ENTRY>
  <NOM class="common" number="dual" gender="feminine"
  case="accusative" />
  <LEMMA>agencija</LEMMA>
  <PHONETIC>a - g E n - " ts i : - j i </PHONETIC>
</ENTRY>
<ENTRY>
  <NOM class="common" number="singular" gender="feminine"
  case="locative" />
  <LEMMA>agencija</LEMMA>
  <PHONETIC>a - g E n - " ts i : - j i </PHONETIC>
</ENTRY>
</ENTRYGROUP>
<ENTRYGROUP orthography="aretirajo">
<ENTRY>
  <VER number="plural" gender="not_specified" person="3"
  mood="indicative" tense="present" voice="active"
  polarity="not_specified" aspect="perfect" />
  <LEMMA>aretirati</LEMMA>
  <PHONETIC>a - r E - " t i : - r a - j O </PHONETIC>
</ENTRY>
<ENTRY>
  <VER number="plural" gender="not_specified" person="3"
  mood="indicative" tense="present" voice="active"
  polarity="not_specified" aspect="imperfect" />
  <LEMMA>aretirati</LEMMA>
  <PHONETIC>a - r E - " t i : - r a - j O </PHONETIC>
</ENTRY>
</ENTRYGROUP>

```

3.2. Slovarji za govorno orientirano prevajanje za slovenski jezik

Drugi del projekta LC-STAR je orientiran v proučitev in izgradnjo jezikovnih virov, ki bi izboljšali govorno orientirano prevajanje. V skladu s tem so bili v projektu proučeni različni pristopi k strojnemu prevajanju govora in rezultati kažejo, da je mogoče razviti robustno strojno simultano prevajanje govora za majhno do srednje veliko semantično področje. Največji problemi strojnega simultanelega prevajanja govora so glede na raziskave naslednji:

- pridobivanje podatkov, ki so specifični za posamezna področja, eno- ali dvojezičnih
- robustnost komponent strojnega simultanelega prevajanja govora na napake pri razpoznavanju in značilnosti govorjenega jezika
- razvoj učinkovite enote za razpoznavo govora in govorno orientirano prevajanje (Moreno et al., 2004)

Trenutno najbolj obetaven pristop k strojnemu prevajanju govora je statistični. Obstoječi sistemi, ki temeljijo na statističnem strojnem prevajanju, uporabljajo velike dvojezične poravnane korpuse za učenje sistemov. Novejše raziskave (Koehn et al., 2003; Vogel et al., 2003) pa kažejo, da so lahko uporabni tudi kratki dvojezični poravnani izseki iz stavkov, značilnih za neko semantično področje. Na področju pogovorov v turizmu, na katero je orientiran projekt LC-STAR, so pogoste t.i. fraze, kot so "kje je...", "rad bi...", "mi lahko pomagate...". Tudi eksperimenti v okviru projekta (Ueffing, Ney, 2003) so pokazali, da bi seznam takih fraz pomagal izboljšati sisteme govorno orientiranega prevajanja, prav tako uspešnost izboljša, če so besede v takem dvojezičnem

slovarju fraz označene glede na pripadnost besedni vrsti in lematizirane.

Zaradi zgoraj navedenih rezultatov se v projektu pripravljata dve vrsti virov, namenjenih za izboljšanje komponente govorno orientiranega prevajanja za področje turizma:

- poravnani slovar 10.000 fraz v 9 jezikih (tudi slovenskem), ki se pogosto uporabljajo v turizmu; besedam so dodane oznake o besedni vrsti v skladu s shemo, določeno v prvem delu projekta, in lema;
- enojezične slovarje polnopomenskih besed, ki se pojavijo v teh frazah, urejene enako kot veliki slovarji za izboljšanje razpoznave in sinteze govora.

Vir za poravnani slovar fraz je bil referenčni seznam fraz v ameriški angleščini, ki je bil preveden v ostale jezike, tudi slovenščino. Referenčni seznam je zagotovil konzorcij LC-STAR.

4. Zagotavljanje kakovosti jezikovnih virov LC-STAR

Pomanjkanje kakovosti jezikovnih virov je bilo ugotovljeno kot ena od glavnih pomanjkljivosti obstoječih jezikovnih virov, zato je eden od namenov projekta to preseči. Poleg natančnih določil in skupnih standardov, po katerih so razviti vsi jezikovni viri v okviru projekta, konzorcij LC-STAR zagotavlja kakovost tudi z neodvisno validacijo. To opravljata nizozemski SPEX (Speech Processing EXpertise centre), validacijski center ELRE, in danski CST (Center for Sprogteknologi). S tem je zagotovljeno, da vsi partnerji v projektu dosegajo enako kvaliteto razvitih jezikovnih virov.

Validacija je delno avtomatska in delno ročna. SPEX je razvil programska orodja za avtomatsko validacijo, s katerimi se preverja ustrezna formalna struktura slovarjev. Programska orodja so na voljo partnerjem, da lahko sami preverjajo ustreznost jezikovnih virov, formalno pa preveri rezultate tudi SPEX. CST zagotavlja ročno validacijo, v kateri preverja jezikoslovno ustreznost jezikovnih virov (pravilnost pripisanih oblikoslovnih oznak, ortografskega zapisa, fonetičnega prepisa, ustreznost prevodov). Validacijo opravljajo neodvisni strokovnjaki, naravni govorniki posameznih jezikov.

5. Zaključek

V članku smo predstavili jezikovne vire, ki se razvijajo v okviru mednarodnega projekta LC-STAR, s podrobnejšo predstavitevijo slovenskih jezikovnih virov, ki jih za projekt razvija Fakulteta za elektrotehniko, računalništvo in informatiko Univerze v Mariboru.

Za slovenščino lahko podobno kot za druge jezike ugotovimo pri večini obstoječih jezikovnih virih, ki so uporabni za sisteme strojnega simultanelega prevajanja govora, pomanjkanje kakovosti, premalo široko pokritost, pomanjkljivo primernost za razvoj sinteze in razpoznave govora ter pomanjkanje standardov. Sodelovanje v projektu LC-STAR je prispevek k izboljšanju tega stanja. Razviti slovarji bodo predstavljali dobro osnovo za razvoj in izpopolnjevanje komponent strojnega simultanelega prevajanja govora.

6. Viri in literatura

- Dobrišek, S., Gros, J., Ipšič, I., Pepelnjak, K., Mihelič, F., Pavešič, N. (1998). Gopolis: slovenska podatkovna zbirka govornih poizvedovanj. V: Informacijska družba IS'1998: Jezikovne tehnologije za slovenski jezik. 105-108.
- Domača spletna stran LC-STAR: <http://www.lc-star.com/>.
- Domača spletna stran Multext-East: <http://nl.ijs.si/ME/>.
- Hartikainen, E., Maltese, G., Moreno, A., Shammass, S., Tiegenhain, U. (2003). Large Lexica for Speech-to-Speech Translation: From Specification to Creation. In Proceedings of the Eurospeech (pp. 1529--1532). Geneva.
- Hoege, H. (2002): Project Proposal TC-STAR - Make Speech to Speech Translation Real. V: Third International Conference on Language Resources and Evaluation. 136-140.
- Gros, J., F. Mihelič, S. Dobrišek, T. Erjavec, M. Žganec (2000). A phonetically and prosodically annotated Slovene speech corpus. V: IS'2000: Jezikovne tehnologije. 27-30.
- Gros, J., I. Ipšič, F. Mihelič, N. Pavešič (1996). Segmentation and labelling of Slovenian diphone inventories. COLING096. 298-303.
- Kaiser, J., Kačič, Z. (1998). Development of Slovenian SpeechDat Database. In the Proceedings of First International Conference on Language Resources & Evaluation. Granada, Spain.
- Kačič, Z. (1997). Copernicus Onomastica project COP 58: Final report. FERL.
- Kačič, Z. (2002). Pomen združevanja raziskovalnih potencialov pri preseganju jezikovnih pregrad v okviru jezikovnih tehnologij naslednjih generacij. V: Informacijska družba IS'2002: Jezikovne tehnologije. 111-115.
- Koehn, P., Knight, K. (2003). Feature-Rich Statistical Translation of Noun Phrases. 41st Annual Meeting of the ACL, Sapporo, Japan.
- Moreno, A., Conejero, D., Castell, N., Gimenez, J. (2004). Language independent specification of LR for translation. LC-STAR Project IST-2001- 32216 Deliverable D5.5.
- Rojc, M., Kačič, Z. (2003). Efficient Development of Lexical Language Resources and their Representation. International Journal of Speech Technology 3/6. 259-275.
- Stabej, M., Vitez, P. (2000). KGB (korpus govornih besedil) v slovenščini. IS'2000: Jezikovne tehnologije. 79-81.
- Šef, T. (2002). Naglaševanje nepoznanih besed pri sintezi slovenskega govora. V: Informacijska družba IS'2002: Jezikovne tehnologije. 149-152.
- Ueffing, N., Ney, H. (2003). First experimental results on baseline speech-to-speech translation systems. LC-STAR Project IST-2001- 32216 Deliverable D4.4.
- Verdonik, D., Rojc, M., Kačič, Z., Horvat, B. (2002). Zasnova in izgradnja oblikoslovnega in glasoslovnega slovarja za slovenski knjižni jezik. In Proceedings of Jezikovne tehnologije/Information Society Multi-Conference (pp. 44--48). Ljubljana, Slovenia.
- Verdonik, D., Rojc, M., Kačič, Z. (2004). Creating Slovenian Language Resources for Development of Speech-to-Speech Translation Components. In Proceedings of the LREC'04 (pp.1399--1402). Lisbon.
- Vogel, S., Zhang, Y., Huang, F., Tribble, A., Venugopal, A., Zhao, B., Waibel, A. (2003). "The CMU Statistical Machine Translation System". MT-Summit, New Orleans, USA.
- Zemljak, M., Kačič, Z., Dobrišek, S., Gros, J., Weiss, P. (2002). Računalniški simbolni fonetični zapis slovenskega govora. Slavistična revija, 50(2), 159--169.
- Ziegenhain, U. et al. (2004). Specification of corpora and word lists in 12 languages. LC-STAR Project IST-2001-32216 Deliverable D1.1.
- Zoegling Markuš, A., Kačič, Z., Horvat, B. (2000). Razvoj slovenske baze izgovorjav "POLIDAT". IS'2000: Jezikovne tehnologije. 95-98.
- Žibert, J., Mihelič, F. (2000). Govorna zbirka vremenskih napovedi. IS'2000: Jezikovne tehnologije. 108-111.