Razvoj slovenske baze izgovarjav POLIDAT

Aleksandra Zögling Markuš*, Zdravko Kačič, Bogomir Horvat****

*Center za interdisciplinarne in multidisciplinarne raziskave in študije, Univerza v Mariboru
Razlagova 22, 2000 Maribor, Slovenija
sandra.zogling@uni-mb.si

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

Povzetek

V članku so opisani cilji in naloge projekta PoliDat. Podana sta model in struktura baze za slovenski jezik. Namen projekta PoliDat je izgradnja baze izgovarjav za razvoj avtomatskih telefonskih sistemov govornega dialoga. Baza je razdeljena v dva dela: govor posnet preko mobilnega telefonskega omrežja in govor posnet preko fiksne telefonske linije. Pri izbiri govorcev so bile upoštevane zahteve za uravnoteženost po spolu, starosti in narečju govorcev. Za vsakega govorca vsebuje baza 43 izgovarjav v skupni dolžini okoli pet minut. Baza trenutno vsebuje govor 150ih govorcev za vsak del baze (mobilni, fiksni), končni cilj pa je 1000 govorcev.

Abstract

In this paper the main goals and tasks of the PoliDat project are described. The design and the structure of the database for Slovenian language are given. The goal of the PoliDat project is to develop spoken language resources for developing automatic telephone speech dialogue systems. The database is divided into two parts: the mobile telephone quality speech and the fixed telephone quality speech. Requirements for balance of gender, age, and dialect were taken in consideration for the recruitment of speakers. The database includes 43 utterances for each speaker in total length of approximately 5 minutes. At the end of the project each part of the database will contain speech of 1000 speakers, currently the database contains speech of 150 speakers for each part.

1 Uvod

V bližnji prihodnosti bo prodor jezikovnih tehnologij najbolj opazen na področju telekomunikacij, kjer že obstaja široko tržišče različnih proizvodov in storitev. Pri tem je vrednost tržišča, ki zajema telefonske storitve – te bodo v bližnji prihodnosti delno ali popolnoma avtomatizirane z uporabo sodobnih sistemov jezikovnih tehnologij, v Evropi že danes več milijard eurov na leto. Tržišče takšnih storitev bo po predvidevanjih v prihodnosti raslo s podobno hitrostjo, kot se bo razvijal trg mobilne telefonije.

Govorna tehnologija zajema širok spekter področij, katerih skupni cilj je ustvariti orodja, s katerimi bo človek lahko komuniciral s strojem na človeku najnaravnejši način – z uporabo govora. Večina obstoječih govorno krmiljenih telefonskih storitev je izvedena s pomočjo razpoznavanja izoliranih besed. Da bi povečali prijaznost takih sistemov in jim zagotovili konkurenčnost tudi v prihodnje, je potrebno razviti sisteme, ki bodo uporabljali razpoznavanje tekočega ali celo spontanega govora. Za razvoj takih sistemov pa so potrebne obsežne baze izgovarjav. V zadnjih nekaj letih smo priča pravi poplavi različnih projektov, katerih naloga je zasnova baz izgovarjav. Zastavljeni projekti so posledica novih dosežkov na področju avtomatskega razpoznavanja govora in od tod izhajajočih novih potreb. Trendi razvoja jezikovnih tehnologij gredo zelo hitro tudi v smeri mobilne telefonije.

Infrastrukturo jezikovnih tehnologij sestavljajo ustrezno dokumentirane in standardizirane baze izgovarjav, slovarji besed, leksikoni in v splošnem vso znanje o jeziku, ki je dostopno v elektronski obliki.

Zaradi vezanosti na jezik pa bo moral vsak narod sam poskrbeti za izgradnjo ustrezne infrastrukture. To je še posebej pomembno za majhne narode, saj bodo v nasprotnem prisiljeni takšne storitve uporabljati v tujem jeziku. Zaradi zapletenosti in obsežnosti problema ter vezanosti na jezik, bo razvoj kompleksnejših storitev za jezike z majhnim številom potencialnih uporabnikov za tuje proizvajalce nezanimiv. Žal je proces izgradnje jezikovne infrastrukture v veliki meri zaobšel države srednje in vzhodne Evrope, ki bodo ta zaostanek same le s težavo nadomestile.

V Sloveniji se z izgradnjo baz izgovarjav ukvarjajo predvsem na Fakulteti za elektrotehniko, računalništvo in informatiko v Mariboru in Fakulteti za elektrotehniko v Ljubljani. Na slednji so doslej zgradili bazo izgovarjav Gopolis (Dobrišek, 1998) in Luz (Gros, 2000). Na Fakulteti za elektrotehniko, računalništvo in informatiko bazo SNABI (Kačič, 2000) in SpeechDat II za slovenski

jezik (Kaiser, 1997). Sodobnim standardom in priporočilom, ki so bila izdelana v okviru projekta SpeechDat II (Winsky, 1997) in ki zagotavljajo kvaliteto baze, primerne za razvoj sistemov telefonskega govornega dialoga za delo v realnem okolju, pa v celoti zadošča trenutno le baza izgovarjav SpeechDat II za slovenski jezik. Ta vključuje govor 1000 govorcev posnet preko fiksne telefonske linije.
Izgradnja baze izgovarjav PoliDat za mobilno in fiksno telefonsko omrežje poteka v sodelovanju med Fakulteto za elektrotehniko, računalništvo in informatiko (v delo so vključeni: Center za jezikovne tehnologije, Laboratorij za digitalno procesiranje signalov in Laboratorij za digitalne in informacijske sisteme), Centrom za interdisciplinarne in multidisciplinarne raziskave in študije Univerze v Mariboru ter podjetjem Telekom Slovenije.

2 Izbira govorcev in zahtevane porazdelitve

Pri izgradnji baze Polidat želimo v čim večji meri zadostili priporočilom in kriterijem, ki so bili definirani v okviru evropskega projekta SpeechDat II (Senia 1997a, 1997b, 1997c). Pri tem je nedvomno eden najpomembnejših kriterijev narečna uravnoteženost baze. S sodelovanjem podjetja Telekom Slovenije, kot partnerja v projektu, bo v veliki meri moč izpolniti tudi ta kriterij. Večina govorcev, ki sodelujejo pri snemanju baze je izbrana izmed zaposlenih v podjetju Telekom Slovenija, ki ima poslovne enote po vsej Sloveniji. Del govorcev je izbranih tudi med zaposlenimi na fakulteti, študenti mariborske univerze in njihovimi svojci. Izbrani so glede na postavljene zahteve uravnoteženosti porazdelitve govorcev po spolu, starosti in narečju. Skupno bo v času snemanja baze za vsak del baze (mobilna, fiksna) izbranih 1400 govorcev, izmed katerih bo 1000 govorcev uporabljenih za definiranje končne baze izgovarjav.

2.1 Spol in starost

Tako kot narečno področje ima tudi spol govorca pomemben vpliv na ustrezno reprezentativnost baze, saj ženske v primerjavi z moškimi govorijo v povprečju z višjim tonom (višina in jakost).

Najboljša uravnoteženost govorcev po spolu je 50 % moških in 50 % žensk.

Tudi starost vpliva na govor. Opažene so razlike med starostnimi skupinami pri uporabi besed in skladnji, ki pa so opazne le pri spontanem govoru.
Predvidena uravnoteženost govorcev po starosti je naslednja:

- najmanj 20 % govorcev starih med 16–30 let,
- najmanj 20 % govorcev starih med 31–45 let,
- najmanj 15 % govorcev starih med 46–60 let.

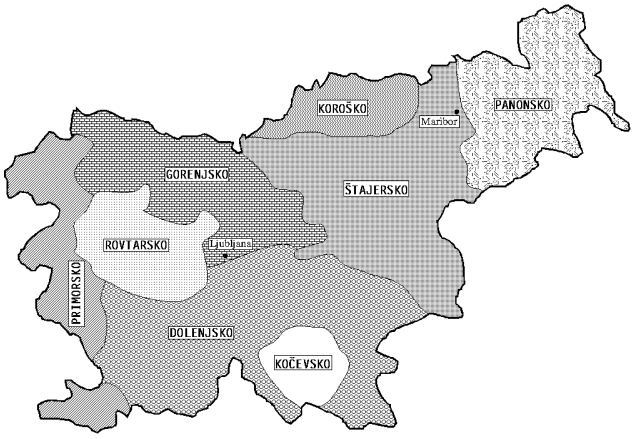
2.2 Narečje

Narečno področje govorcev ima močan vpliv na njihov govor. Pomembni so tudi drugi dejavniki, kot na primer jezikovno ozadje staršev ter kraj, kjer je govorec odraščal.

Snemanje izvajamo v okviru sedmih regij: Prekmurje, Štajerska, Koroška, Dolenjska, Notranjska, Gorenjska in Primorska. Ob tem smo upoštevali priporočila projekta Speechdat II, da naj vsako izbrano narečno področje zajema najmanj 1% populacije prebivalstva v državi. Zaradi tega smo število narečnih področij omejili na deset. Pri določitvi področij smo izhajali iz karte slovenskih narečij (Geodetski zavod Slovenije, 1993). Določili smo naslednja narečna področja: Panonsko, Štajersko, Koroško, Dolenjsko, Kočevsko, Rovtarsko, Gorenjsko, Primorsko, govor mesta Maribor in govor mesta Ljubljane. Slika 1 kaže uporabljeno narečno razdelitev slovenskega govornega prostora.
Število govorcev iz posameznega področja je določeno glede na število prebivalcev v danem področju. Tabela 1 prikazuje porazdelitev govorcev po narečjih.

Št.	Narečno področje	Število govorcev	Število govorcev (%)
1.	Panonsko	160	11,43
2.	Štajersko	310	22,14
3.	Koroško	50	3,57
4.	Dolenjsko	190	13,57
5.	Kočevsko	20	1,43
6.	Rovtarsko	60	4,28
7.	Gorenjsko	160	11,43
8.	Primorsko	110	7,86
8.	Maribor	110	7,86
10.	Ljubljana	230	16,43

Tabela 1: Porazdelitev govorcev po narečnih področjih.



Slika 1: Narečna porazdelitev slovenskega govornega prostora

2.3 Izobrazbena struktura

Kriterij izobrazbene strukture klicalcev smo razčlenili na naslednje kategorije: osnovna šola, poklicna šola, srednja šola, višja šola, visoka šola, magisterij in doktorat.

2.4 Okolje klica

Pri snemanju govora je okolje, v katerem se je nahajal govorec v času snemanja, pomemben faktor. Ker je namen baze PoliDat čim bolje odraziti realne razmere, so lahko klici opravljeni iz poljubnega okolja :

- telefonska govorilnica,
- dom,
- pisarna,
- tovarna,
- javni prostor (letališče, železniška postaja, gostinski lokal),
- cesta,
- vozilo,
- ostalo.

3 Vsebina baze

Pri definiranju korpusa je potrebno upoštevati tudi nekatere sociološke vidike, to je način govorne komunikacije, ki je najpogostejši v nekem jezikovnem okolju. Skupni okvir definiranega korpusa predpisuje število in obseg posameznih enot, ki jih morajo vsebovati korpusi za posamezna jezikovna področja. Predpisani vsebinski okvir za baze posameznih jezikov vsebuje: izolirane številke, zaporedja številk, naravna števila, denarne zneske, črkovane besede, uro, datum, odgovore da/ne, imena mest, najpogostejše ukaze pri posameznih aplikacijah, fraze, ki vsebujejo ukaze pri posameznih aplikacijah in fonetično uravnotežene stavke (F. Senia, 1997c).

Ker je baza v osnovi sestavljena iz dveh delov, mobilnega in fiksnega, smo uporabili isti korpus pri snemanju za klice preko mobilnega telefona kot tudi za klice preko vrvičnega telefona.
Baza bo vsebovala:

- 1000 označenih klicev posnetih preko mobilnega telefonskega omrežja (izbranih je 1400 govorcev)

- 1000 označenih klicev posnetih preko fiksne telefonske linije (izbranih je prav tako 1400 govorcev).

Uporabljeni korpus zajema 900 fonetično uravnoteženih stavkov različnih dolžin, 42 aplikacijskih besed (besede, ki bodo predvidoma uporabljane v različnih aplikacijah sistemov telefonskega govornega dialoga), 1500 fonetično uravnoteženih besed, številke in 150 številskih nizov– številke kreditnih kartic, 150 številskih nizov–telefonske številke in 100 različnih denarnih zneskov.
Vsak govorec ima svojo identifikacijsko številko, ki jo sestavlja 5 številk.
V času enega klica posname telefonski strežnik 43 izgovarjav klicalca, ki so sestavljene iz zgoraj omenjenih enot. Tabela 2 prikazuje opis vsebine klicev.
Posnet govor sestavlja pretežno brani tekst z vnaprej pripravljene predloge, ki jo dobi vsak govorec pred snemanjem, pa tudi spontan govor - na primer odgovor klicalca na vprašanja o klicalčevem letu rojstva, o trenutni uri itd.
Hkrati s snemanjem smo pričeli tudi z označevanjem baze, ki zajema slušno preverjanje pravilnosti izgovarjave, vnašanje podatkov iz vrnjenih vprašalnikov govorcev in označevanje motenj v govornem signalu. Motnje so lahko stacionarne ali kratkotrajne, povzroča pa jih lahko telefonski kanal ali pa govorec. Prav tako označujemo vse nepravilne, nejasne, napačne ali nepopolne izgovarjave.

Opis izgovarjave	Izgovarjave za posamezni klic
Aplikacijske besede	6 aplikacijskih besed
Zaporedje izoliranih številk	1 zaporedje 10 izoliranih številk
Vezane številke	1 številka predloge 1 telefonska številka 1 številka kreditne kartice 1 identifikacijska koda
Datumi	1 spontani datum 1 predložen datum 1 relativna oz. splošna časovna fraza
Fraze za iskanje ključnih besed	1 fraza za iskanje ključnih besed z aplikacijsko besedo
Izolirane številke	1 izolirana številka
Črkovane besede/fraze	1 spontana (ime govorca) 1 ime kraja 1 določena za dosego ustrezne pokritosti črk v korpusu
Denarni znesek	1 denarni znesek
Naravna števila	1 naravno število
Imena za imenike	1 spontano (ime govorca) 1 spontano (mesto odraščanja) 1 ime mesta (nabor 500) 1 ime podjetja/ustanove (nabor 500) 1 ime in priimek (nabor 150)
Vprašanja	1 pretežno "da" vprašanje 1 pretežno "ne" vprašanje
Stavki	9 fonetično uravnoteženih stavkov
Časovne fraze	1 spontana fraza (trenuten čas) 1 časovna fraza (besedna oblika)
Besede	4 fonetično uravnotežene besede

Tabela 2: Specifikacija vsebine baze

Vsak izmed 43 posnetkov je shranjen v svojem zapisu, ki mu je pridružen zapis s transkripcijo posnetka. Transkripcija je ortografska z dodanimi podrobnostmi o slišnih govornih in negovornih dogodkih, kot so: vzdihovanje, glasno dihanje, presluh in motnje na linijah. Vse besede, ki se pojavljajo v posneti bazi, so zbrane v fonetičnem slovarju. Fonetične transkripcije besed v slovarju so bile tvorjene s pomočjo programskega orodja za avtomatsko grafemsko-fonemsko pretvorbo (Rojc, 2000), nato pa še ročno pregledane. Transkripcije so zapisane z naborom fonetičnih simbolov SAMPA za slovenski jezik (Dobrišek, 1996).

4 Snemalno okolje in označevanje

Govor snemamo preko fiksnega telefonskega omrežja. Govorni strežnik je nameščen v Laboratoriju za digitalno

procesiranje signalov. Sestavljata ga PC–računalnik z ustrezno programsko opremo in ISDN–telefonska kartica. Na telefonsko omrežje je priključen preko linije ISDN. Uporabljena frekvenca vzorčenja je 8kHz, ob 8–bitni kvantizaciji in a-law kodiranju. Posnetki se tedensko prenašajo na delovne postaje, kjer se označujejo in dokončno obdelajo.

Označevanje posnetkov se opravlja s PoliDatLabel programskim orodjem, izdelanim na Univerzi v Mariboru. Orodje je bilo posebej razvito za označevanje baze PoliDat. Delo z orodjem PoliDatLabel je avtomatizirano do največje možne mere, kar skrajša čas označevanja in zmanjšuje število napak. Označevanje poteka v dveh fazah. Najprej zapišemo besede, nato pa dodamo oznake o slišnih dogodkih.

Posneti govorni material je zapisan v posebni drevesni strukturi baze, ki je definirana upoštevajoč možnost čim hitrejšega branja s CD-ROMa, na katerem bo baza tudi zapisana.

5. Zaključek

Baza PoliDat je namenjena razvoju in testiranju sistemov govorne tehnologije ter bo zagotavljala dovolj veliko količino podatkov za učenje sistemov avtomatskega razpoznavanja govora neodvisnega govorca za delo v realnem okolju. Izgradnja baze PoliDat pomeni pomemben prispevek k prizadevanjem slovenskih raziskovalcev za uvajanje sistemov govornih tehnologij za slovenski jezik v sodobna komunikacijska omrežja in kar je še pomembneje, ohranja stik z najnovejšim razvojem na tem, tudi za slovenski narod, zelo pomembnem področju. Trenutno namreč skupaj z bazo SpeechDat II za slovenski jezik predstavlja edino mednarodno primerljivo bazo izgovarjav za slovenski jezik, ki bo zadostila priporočilom in kriterijem gradnje baz izgovarjav, ki so bili definirani v okviru projekta SpeechDat II. To je še toliko bolj pomembno zato, ker je razvoj sodobnih sistemov govornih tehnologij mogoč le ob ustrezni infrastrukturi, ki jo predstavljajo predvsem baze izgovarjav, korpusi in slovarji. Načrtujemo, da bo ob zaključku projekta baza PoliDat dosegljiva preko mednarodnega združenja ELRA.

Viri

Dobrišek, S., Gros, J., Horvat, B., Kačič, Z,. Mihelič, F., (1996). Pobuda za standardizacijo računalniškega simbolnega fonetičnega zapisa slovenskega govora. Portorož: Zbornik ERK'96, str. 247–250.

Dobrišek, S., Gros, J., Mihelič, F., Pavešič, N. (1998) Recording and Labelling of the GOPOLIS Slovenian Speech Database. Proceedings of the First International Conference on Language Resources and Evaluation, pp. 1089-1096. Granada, Spain.

Geodetski zavod Slovenije (1993). Karta slovenskih narečij.

Gros, J., Mihelič, F., Dobrišek, S., Erjavec, T., Žganec M. (2000) Corpora of Slovene Spoken Language for Multi-lingual Applications. Zbornik Second International Conference on Language Recources and Evaluation. Athens, pp. 953-956.

Kačič, Z., Horvat, B., Zögling S. (2000) Issues in Design and Collection of Large Telephone Speech Corpus for Slovenian Language. Zbornik Second International Conference on Language Recources and Evaluation. Athens, pp. 943-946.

Kaiser J. and Kačič Z. (1997). SpeechDat Slovenian Database for the Fixed Telephone Network. University of Maribor.

Rojc M. and Kačič Z. (2000). A Computational Platform for Development of Morphologic and Phonetic Lexica, Proceedings LREC 2000, Athens, pp. 277-281.

Sania F. at al. (1997a) Environmental and speaker specific coverage for Fixed Networks. SpeechDat project, doc ref LE2-4001-SD1.2.1.

Sania F. at al. (1997b) Specification of speech database interchange format. SpeechDat project, doc ref LE2-4001-SD1.3.1.

Sania F. and J.G. van Velde (1997c). Specification of ortographic transcription and lexicon conventions. SpeechDat project, doc ref LE2-4001-SD1.3.2.

Winsky. R. (1997) Definition of Corpus, scripts and standard for Fixed Networks. SpeechDat project, doc ref LE2-4001-SD1.1.1.