Elektrotehniški vestnik 75(3): 85–90, 2008
Electrotechnical Review, Ljubljana, Slovenija

Slovenska baza BNSI Broadcast News za razpoznavanje tekočega govora

Andrej Žgank, Darinka Verdonik, Zdravko Kačič

Univerza v Mariboru, Fakulteta za elektrotehniko, računalništvo in informatiko, Laboratorij za digitalno procesiranje signalov, Smetanova ul. 17, SI-2000 Maribor, Slovenija
E-pošta: andrej.zgank@uni-mb.si

Povzetek. V članku bomo predstavili nov slovenski jezikovni vir, bazo BNSI Broadcast News, ki vsebuje posnetke televizijskih dnevnoinformativnih oddaj. Vir je namenjen razvoju razpoznavalnikov tekočega govora z velikim slovarjem besed za neomejeno domeno. Sestavljajo ga govorna baza z ročno tvorjenimi transkripcijami v obsegu 36 ur in tekstovni korpus, ki vsebuje 11 mio besed. Govorna baza je namenjena učenju akustičnih modelov, tekstovni korpus pa bo uporabljen za izdelavo jezikovnih modelov. Novi jezikovni vir je plod sodelovanja med Univerzo v Mariboru, FERI in RTV Slovenija. V članku bomo najprej predstavili postopek zajemanja gradiva in karakteristike baze. Opisali bomo potek ročnega zapisovanja govornega korpusa. Sledila bo podrobna analiza govornega in tekstovnega dela baze, ki je namenjena predstavitvi vseh lastnosti jezikovnega vira, ki vplivajo na razvoj razpoznavalnikov govora.

Ključne besede: slovenski jezikovni vir, avtomatsko razpoznavanje tekočega govora, govorni korpus, besedilni korpus, Broadcast News

The Slovenian BNSI Broadcast News database for continuous speech recognition

Extended abstract.

The paper presents a new Slovenian language resource, the Slovenian BNSI Broadcast News database. Its main goal is to produce the necessary language resources for the Slovenian large vocabulary continuous speech recognition in an unrestricted domain. The BNSI Broadcast News database is a result of cooperation between the Faculty of Electrical Engineering and Computer Science of the University of Maribor, and the Slovenian national broadcaster RTV Slovenia. The project started in 2002. The BNSI database comprises two subsets: a speech database with transcriptions and text corpus. The former will be used for acoustic model training and the latter for language modeling.

The speech database consists of two different types of news shows *TV Dnevnik* and *Odmevi* from the period 1999 – 2003. As shows were recorded from the archive, a broad spectrum of topics was covered in these news shows.

The text corpus is based on scenarios that were generated for the news shows. A larger number of different types of news shows (9) was included in the text corpus to maximize its size. The text in scenarios covers all the read speeches from news shows. The period between years 1998 – 2004 was included in the corpus. The raw format of the text corpus is RTF, which also includes several meta-information. For language modeling, an ASCII text format was generated with important meta-information placed in the header of each news show.

Manual transcriptions of news shows were produced using the Transcriber tool. A personal computer was the only tool needed. The transcription rules created by LDC [11], LIMSI [12] and COST 278 BN SIG [13] were taken as a baseline. Different language-dependent rules were added to this baseline setup. Usually, a three-phase approach was applied to produce the transcriptions. As the initial text, a selected scenario of the news show was taken. It covered approximately 40% of the spoken material. In the first step, the acoustic segmentation was performed. In the second one, the speakers were created and the missing text was transcribed. In the last phase, the acoustic background and different acoustic events were labeled. Approximately 27 hours of work were needed to produce transcription for one hour news show. Thereafter the transcription was checked by two supervisors.

The BNSI speech corpus consists of 42 shows in the total duration of 36 hours. The ratio between the training, development and evaluation set is given in Table 1. The speech corpus transcriptions have 268k words, 37k of them are different. In the speech corpus, there are 1565 speakers, 1069 of them are male and 477 female. The gender of 19 speakers is unknown. There are significantly more male speakers but, the female speakers more frequently appear in the role of reporters, which balances out this ratio. The gender ratio according to the speech duration is 53.65% versus 46.34% for male speakers. The five speakers with the longest recordings are presented in Table 3.

The ratio between f-conditions [15], which are important as speech recogniser needs acoustically homogenous speech parts, is given in Figure 1. The two largest classes belong to the read speech in studio (f0) and read or spontaneous speech with a background. The dialect of each speaker in the database was labeled with one of the 15 categories presented in Section 4. For each part of a news show the topic was classified according to one of the 25 various categories.

The text corpus consists of 600 monthly collections of news shows scenarios. A small part of the raw-text material is presented in Figure 2. There are 11M words in the text corpus, with 175k of them being different. The BNSI text corpus was compared to the Slovenian newspaper text corpus Večer (Ta-

Prejet 12. marec, 2007
Odobren 19. marec, 2008

ble 4). The analysis of word frequencies in the text corpus was performed (Figure 3). Only 1238 words occur more than 1000 times and almost 120k words occur only once. The three most frequent words (Table 5) belong to the group of functional words.

The new Slovenian language resource is being used for the development of the LVCSR speech recogniser. The Slovenian BNSI Broadcast News database will be available to the research community by ELRA/ELDA. The BNSI database was later supplemented with an additional in-house set in the total duration of another 36 hours of speech. Also, the first non-native Slovenian speech database SINOD [17] was produced as an add-on to the BNSI Broadcast News database.

Key words: Slovenian language resources, automatic continuous speech recognition, speech corpus, text corpus, Broadcast News

1 Uvod

Raziskave na področju avtomatskega razpoznavanja slovenskega govora so se začele v drugi polovici osemdesetih letih z razvojem prvih preprostih sistemov na ravni velikosti slovarja izoliranih besed. Takšni sistemi so bili v zadnjih nekaj letih tudi uspešno preneseni v vsakdanjo uporabo [1, 2]. Bistveno večji izziv je področje razpoznavanja tekočega slovenskega govora z velikim slovarjem besed za neomejeno domeno. Eden od vzrokov je kompleksna struktura slovenskega jezika. Tukaj je treba poudariti predvsem pregibnost besed, kar veča slovarsko strukturo in zamenljivost besed med seboj, ter relativno prost besedni red v stavku, kar zmanjšuje učinkovitost napovedovanja besednega reda s pomočjo n-gramskega statističnega jezikovnega modela.

Osnovna predpostavka za delo pri razpoznavanju govora je obstoj jezikovnih virov, ki so potrebni za učenje akustičnih in jezikovnih modelov. Tako je bilo za slovenski jezik izdelanih kar nekaj govornih in jezikovnih virov (1000 FDB SpeechDat(II), PoliDat, Gopolis, Vreme, SNABI [3]), vendar nobeden ni omogočal razpoznavanja tekočega govora za neomejeno domeno. To je bila glavna motivacija za razvoj novega slovenskega jezikovnega vira, baze BNSI Broadcast News, ki sledi v zadnjem času tudi čedalje pomembnejši zahtevi po mednarodni dostopnosti jezikovnega vira. V zvezi s tem so v sklepni fazi dogovori z evropsko agencijo ELDA. Po začetku projekta BNSI sta v slovenskem prostoru nastala tudi jezikovna vira SiBN (dnevnoinformativne oddaje) in SloParl (parlamentarne debate), ki pa sta za zdaj oba interne narave.

Domena baz Broadcast News (BN) je razpoznavanje govora v televizijskih (ali radijskih) informativnih oddajah [4]. Področje se je začelo razvijati leta 1996. Prva baza vrste Broadcast News je bila angleška baza 1996 HUB4, ki ji je sledila 1997 HUB4 [5]. V naslednjih letih je bilo izdelanih še kar nekaj baz za različne jezike: španski 1997 HUB4, italijanski IBNC [6], francoski [7], češki ... Specifikacije baze BNSI smo zasnovali primerljivo glede na lastnosti angleške baze 1996 HUB4 in italijanske baze IBNC, zato bomo v nadaljevanju članka slednjo uporabili za primerjavo. Baze iz domene BN omogočajo razvoj sistemov na različnih raziskovalnih področjih: razpoznavanju tekočega govora, podnaslavljanju oddaj v živo, indeksiranju, sledenju govorca ... Za nekatere jezike pa so že na voljo tudi uporabniške aplikacije [8].

Baza BNSI Broadcast News za slovenski jezik* je nastala v okviru sodelovanja med Fakulteto za elektrotehniko, računalništvo in informatiko Univerze v Mariboru in nacionalno televizijo RTV Slovenija v Ljubljani. Pogodba o izdelavi baze je bila podpisana konec leta 2002. Da bi omogočili celotno pokritje potreb po slovenskih jezikovnih virih za domeno BN, smo bazo zasnovali kot skupek dveh neodvisnih delov: govornega korpusa, sestavljenega iz posnetkov oddaj, in besedilnega korpusa, sestavljenega iz scenarijev oddaj.

Ob koncu projekta v članku predstavljamo lastnosti baze BNSI Broadcast News, ki so pomembne s stališča razvoja razpoznavalnikov govora in novosti, ki jih prinaša na področju slovenskih jezikovnih virov. Na kratko bomo predstavili tudi organizacijske vidike izdelave baze in postopek zajemanja gradiva. Več podrobnosti o tej tematiki je mogoče najti v članku [9].

2 Zajemanje gradiva in lastnosti baze BNSI

Najprej bomo predstavili govorni korpus baze BNSI, ki zajema zvočne posnetke oddaj in njihove ročno tvorjene transkripcije. Namenjen je predvsem učenju akustičnih modelov slovenskega razpoznavalnika govora. Na podlagi analize informativnih oddaj RTV Slovenija [9] smo v govorno bazo vključili dnevnoinformativne oddaje *TV Dnevnik* in *Odmevi*. Zajemali smo oddaje iz obdobja 1999–2003 in s tem zagotovili široko pokritost različnih vključenih tematik.

Drugi – neodvisni – del baze BNSI je besedilni korpus, ki je nastal na podlagi scenarijev različnih informativnih oddaj. Predvidena je predvsem njegova uporaba na področju izdelave statističnih jezikovnih modelov za razpoznavalnik govora. Ponavadi se pri takšnem postopku izdelave modelov uporabljajo besedilni korpusi, narejeni na podlagi časopisnega gradiva. Zaradi razlike med časopisnim in televizijskim medijem prihaja do opaznih razhajanj v jeziku med obema vrstama besedil, kar oteži delovanje razpoznavalnika govora. Besedilni korpus BNSI se bo tako pri izdelavi jezikovnih modelov uporabljal za utežitev časopisnega korpusa na domeno BN. Z uporabo takšnega pristopa lahko pričakujemo izboljšano delovanje razpoznavalnika govora in zmanjšanje deleža neznanih besed v testnem naboru.

*Delo je delno finančno podprla Agencija za raziskovalno dejavnost RS, po pogodbi številka P2-0069.

Pri izdelavi scenarija informativne oddaje uporabljajo na RTV Slovenija komercialno aplikacijo Avid iNEWS. Hkrati z besedilom oddaje, ki ga voditelj bere z zaslona, vsebuje scenarij tudi različne dodatne metapodatke, ki jih bomo predstavili v nadaljevanju. Da bi zagotovili čim večji obseg besedil, smo v korpus zajeli scenarije iz obdobja 1998–2004. Hkrati s scenariji oddaj RTV Dnevnik in Odmevi smo zajeli tudi različne druge oddaje, ki nastajajo v okviru uredništva informativnih oddaj: jutranja poročila (7:00, 8:00, 9:00, kratka poročila, v celoti bran govor; pogosto je ponavljanje vsebine), opoldanska poročila (po lastnostih so podobna jutranjim), popoldanska poročila (daljši prispevki, prisotnih je več govorcev, vključeno je javljanje s terena; posnetki vsebujejo tudi spontan govor, ki ni vključen v scenarij oddaje), tedenske oddaje (različne informativne oddaje, ki se pripravljajo tedensko; obravnavajo najpomembnejše dogodke zadnjega tedna, pri tem pa ponavadi vsebujejo relativno veliko spontanega govora).

Scenariji za izdelavo besedilnega korpusa so bili skopirani s strežniškega sistema RTV Slovenija v formatu RTF, ki je omogočil ohranitev vseh metapodatkov, ki so bili prisotni v njih. V besedilni korpus je hkrati z originalnim RTF-formatom vključena tudi verzija v ASCII-obliki, ki vsebuje samo čisto besedilo in jo je kot takšno že mogoče neposredno uporabljati za izdelavo statističnih jezikovnih modelov.

3 Postopek zapisovanja govornega gradiva

Hkrati z razvojem področja baz BN [4] so nastajala tudi različna priporočila za ročno zapisovanje govornega gradiva. Tako smo pri izdelavi baze BNSI upoštevali priporočila, ki so nastala v okviru organizacij LDC [11] in LIMSI [12]. Med delom smo jih dopolnjevali tudi z napotki, ki so nastali v okviru pobude COST 278 EUBN [13]. Zaradi posebnosti slovenskega jezika je bilo treba definirati tudi vrsto jezikovno odvisnih pravil, o katerih je več informacij predstavljenih v članku [9].

Za obdelavo govornega gradiva smo uporabljali orodje Transcriber [14]. Delovno mesto vsakega zapisovalca je bilo zasnovano tako, da je bil računalnik edino orodje, ki ga je zapisovalec potreboval pri svojem delu [10].

V različnih obdobjih projekta je govorno gradivo zapisovalo osem različnih zapisovalcev. Celoten proces izdelave transkripcije sta nadzirala dva nadzornika, od katerih je bil eden slovenist. Kot osnova transkripcije, ki je bila začetek dela za vsako oddajo, je bil uporabljen scenarij oddaje, ki je v povprečju pomenil približno 40 odstotkov celotnega trajanja oddaje. Osnovni postopek izdelave transkripcije lahko v grobem razdelimo na tri korake. V prvem poteka segmentacija govornega posnetka na akustično homogene dele, kar je potrebno zaradi principa učenja akustičnih modelov. V drugem koraku je zapisovalec določil posamezne govorce in njihove lastnosti ter zapisal izgovorjeno. V zadnjem, tretjem koraku je označeval zvočnego ozadje in akustične dogodke ter popravljal transkripcije težavnih odsekov. Za obdelavo 60-minutne oddaje je bilo potrebnih približno 27 ur dela.

4 Govorni korpus

Z zagotovljenim dostopom do arhiva RTV Slovenija smo lahko za zajemanje govornega gradiva uporabili originalne arhivske analogne kasete formata Beta SP. Zvok oddaj smo presneli na DAT-kasete, ki so bile izvor govornega korpusa. Zvočni signal na DAT-kasetah je bil digitaliziran s 16-bitno ločljivostjo pri 48kHz vzorčenju. Pozneje smo zvočni format spremenili v 16-bitnega pri 16kHz vzorčenju. Hkrati smo na DVD-medije zajemali tudi sliko in zvok oddaje. Pri zajemanju je bil uporabljen izgubni kodek MPEG-2. Ti posnetki so bili pomoč zapisovalcem pri izdelavi transkripcij.

nabor	dolžina (mm:ss)	# večernih	# nočnih
učni	1795:06	18	16
razvojni	178:16	2	2
testni	184:58	2	2
skupaj	2158:20	22	20

Tabela 1. Dolžina in število oddaj, vključenih v nabore govorne baze BNSI Broadcast News
Table 1. Duration and number of news shows in different BNSI Broadcast News speech database sets

Govorni korpus BNSI (tabela 1) vsebuje skupaj 42 oddaj, pri čemer je dolžina učnega nabora 30 ur. Razvojni nabor obsega tri ure gradiva in je namenjen optimizaciji parametrov razpoznavalnika govora. Vrednotenju razpoznavalnika govora je namenjen testni nabor v dolžini treh ur (IBNC: 30 ur, 150 oddaj). Transkripcije celotnega govornega korpusa baze BNSI Broadcast News vsebujejo 268k besed, od tega 37k različnih (IBNC: 318k besed, 23k različnih).

	# govorcev	moški	ženske	neznano
BNSI	1565	1069	477	19

Tabela 2. Število in spol govorcev v govorni bazi BNSI Broadcast News
Table 2. Number and gender of speakers in the BNSI Broadcast News speech database

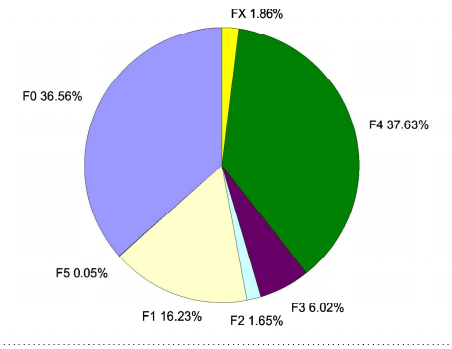
Govorni korpus baze BNSI vsebuje skupaj 1565 različnih govorcev (tabela 2), od tega jih je 1069 (68,31 %) moškega spola in 477 ženskega (IBNC: 677 govorcev). Za 19 govorcev spola kljub uporabi videoposnetkov ni bilo mogoče enoumno določiti. Praviloma gre pri tem za kratke odseke govora v kompleksnem akustičnem okolju. Število moških govorcev je sicer bistveno večje od števila

ženskih govork, vendar le-te pogosteje nastopajo v vlogi novinark, tako da pri skupni minutaži oddaj ni tako velike razlike med obema spoloma (moški spol 53,65 %, ženski spol 46,34 %). Pet govorcev z najdaljšo minutažo smo predstavili v tabeli 3.

oseba	minutaža (s)
Tomaž Ranc	3349,2
Janja Koren	2668,8
Slavko Bobovnik	2662,3
Uroš Slak	2180,4
Ksenija Horvat	2118,2

Tabela 3. Minutaža petih najpogostejših govorcev v bazi BNSI Broadcast News
Table 3. Duration of five most frequent speakers in the BNSI Broadcast News database

Vseh pet govorcev z najdaljšo minutažo v bazi BNSI Broadcast News je voditeljev informativnih oddaj RTV Slovenija. Največ govornega gradiva pripada Tomažu Rancu, skupaj je posnetih skoraj 56 minut njegovega govora. Ženska govorka z največjim obsegom govornega gradiva (približno 45 minut) je Janja Koren. Tem petim govorcem pripada skupaj 3 ure in 36 minut posnetkov.



Slika 1. Razmerje med f-razredi za bazo BNSI Broadcast News
Figure 1. F-condition ratios for the BNSI Broadcast News database

S stališča razpoznavanja govora je bistvenega pomena zagotoviti akustično homogene odseke govornega signala. Tako je bilo definiranih 7 f-razredov [15], ki jih opisujejo. Delitev govornega korpusa glede na f-razrede je predstavljena na sliki 1. Največji razred je f4 s 37,63 % (IBNC: 21,1%), ki vsebuje bran in spontan govor z zvočnim ozadjem. Ta kategorija je v bazi BNSI v primerjavi z drugimi bazami tako obsežna zaradi zelo striktnih pravil za zapisovanje, saj je bilo že zelo tiho zvočno ozadje iz montiranega videoposnetka klasificirano v razred f4. Razredu f0, ki vsebuje brani govor v studijskem okolju, pripada 36,56 % korpusa (IBNC: 57,0%). Obseg spontanega govora v studijskem okolju je 16,23 % (IBNC: 15,0%).

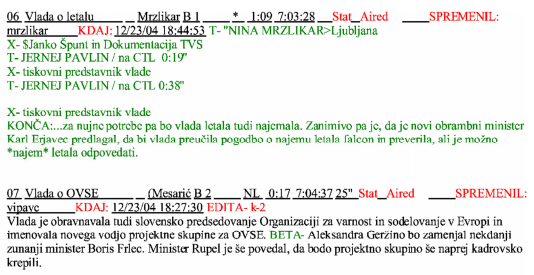
Govorcem v bazi smo pripisali lastnosti njihovega govora, in sicer je ta lahko bil bran ali spontan. Označevali smo, ali je govorec govoril v zbornem jeziku. V nasprotnem primeru smo označili narečje govorca. Slovenščina je znana po številnih narečjih, po splošni oceni naj bi jih bilo okoli 48. Za potrebe baze BNSI smo označevali 15 skupin socialnih zvrsti govora, in sicer: SV-pogovorno, JZ-pogovorno, ljubljansko, mariborsko, panonsko, štajersko, koroško, dolenjsko, primorsko, rovtarsko, gorenjsko, koroškoslovensko, porabskoslovensko, furlanskoslovensko, zdomci. Kategoriji SV-pogovorno in JZ-pogovorno označujeta skupino govorcev, ki govorijo pogovorno zvrst. Pri tem smo ločevali samo govorce iz jugozahodnih delov Slovenije in severovzhodnih delov Slovenije. Zaradi večjega števila govorcev iz obeh največjih mest Slovenije smo vključili posebno zvrstno oznako za govorce mariborskega in ljubljanskega govora. Oznake panonsko, štajersko, koroško, dolenjsko, primorsko, rovtarsko in gorenjsko smo uporabili pri govorcih, ki so govorili v narečju, in so določene glede na slovenske narečne skupine. Posebne zvrstne oznake smo dodali za govorce, ki živijo pod vsakodnevnim vplivom tujih jezikov, tj. Slovence v zamejstvu in zdomce. Označene lastnosti govorcev bo mogoče uporabiti med izdelavo akustičnih modelov.

Pri razpoznavanju govora za pregibne jezike je ena izmed možnosti, kako zožiti iskalni prostor, tudi modeliranje različnih domen. V ta namen smo na podlagi kategorij, predstavljenih v [16], definirali 25 različnih kategorij, v katere so bili glede na vsebino klasificirani prispevki med zapisovanjem. Hkrati s kodo kategorije domene je bil za vsak prispevek tvorjen tudi kratek opis vsebine prispevka.

5 Tekstovni korpus BNSI

S strežniškega sistema RTV Slovenija smo zajeli sedem letnikov scenarijev (1998–2004), ki so bili osnova za izdelavo besedilnega korpusa. Vključene informativne oddaje smo predstavili v drugem poglavju. Skupaj je v besedilnem korpusu več kot 600 mesečnih zbirk oddaj (npr. *TV Dnevnik*), kar pomeni v povprečju 7,1 zbirke oddaj na mesec. Pretvorba besedila iz scenarija in izločanje neuporabljenih delov (zaključni odlomki, tuji govor) je potekalo avtomatično, glede na oznake v RTF-formatu scenarija. Izsek iz izhodiščnega RTF-scenarija je predstavljen na sliki 2. Scenarij na sliki 2 vsebuje hkrati z branim govorom (zadnje štiri vrstice na sliki 2) tudi različne metainformacije (domena prispevka, datum oddaje, topik...), ki smo jih delno ohranili kot dodatno informacijo v tekstovnem korpusu. Tematika prispevka je v scenariju označena s kratko besedno frazo. Takšno oznako je po potrebi mogoče preslikati v 25 kategorij, ki smo jih uporabili pri akustičnih transkripcijah. Besedilo, ki se nahaja v sredinskem delu slike 2, je zaključni od-

lomek govora iz terenskega prispevka, ki je namenjen pomoči pri režiji oddaje. Takšne odseke smo praviloma izločili iz tekstovne baze, saj ne pokrivajo celotne vsebine prispevka. Pripravljenega besedilnega gradiva zaradi velikega obsega nismo dodatno ročno pregledovali, tako da so v njem ostale morebitne tipkarske napake, katerih delež pa je minimalen.


06_Vlada o letalu _ Mrzlikar B 1 * 1:09 7:03:28 Stat_Aired SPREMENIL:
mrzlikar KDAJ: 12/23/04 18:44:53 T- "NINA MRZLIKAR>Ljubljana
X- Sljanko Šput in Dokumentacija TVS
T- JERNEJ PAVLIN / na CTL 0:19"
X- tiskovni predstavnik vlade
T- JERNEJ PAVLIN / na CTL 0:38"

X- tiskovni predstavnik vlade
KONČA:...za nujne potrebe pa bo vlada letala tudi najemala. Zanimivo pa je, da je novi obrambni minister Karl Erjavec predlagal, da bi vlada preučila pogodbo o najemu letala falcon in preverila, ali je možno *najem* letala odpovedati.

07_Vlada o OVSE _ (Mesarič B 2 NL 0:17 7:04:37 25" Stat_Aired SPREMENIL:
vipavc KDAJ: 12/23/04 18:27:30 EDITA- k-2
Vlada je obravnavala tudi slovensko predsedovanje Organizaciji za varnost in sodelovanje v Evropi in imenovala novega vodjo projektne skupine za OVSE. BETA- Aleksandra Geržino bo zamenjal nekdanji zunanji minister Boris Frlec. Minister Rupel je še povedal, da bodo projektno skupino še naprej kadrovsko krepili.

Slika 2. Izsek iz scenarija oddaje, ki je bil osnova besedilnega korpusa v bazi BNSI Broadcast News
Figure 2. Show's scenario used as raw material for text corpus in the BNSI Broadcast News database

Ena bistvenih lastnosti, ki hkrati z zvrstjo besedila (npr. časopisna zvrst) določa uporabnost tekstovnega korpusa, sta njegova velikost in besedna raznolikost. Statistika za tekstovni koprus BNSI je predstavljena v tabeli 4.

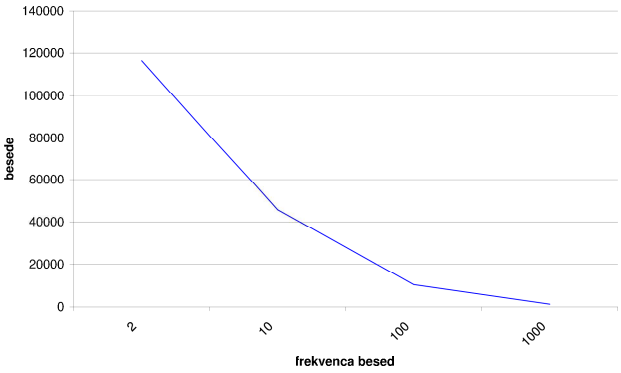
korpus	letniki	# besed	# različnih
BNSI	7	11M	175k
BNSI	1 (2001)	1,3M	70k
Večer	4	95M	736k
Večer	1 (2001)	27M	444k

Tabela 4. Velikost in besedna raznolikost tekstovnih korpusov BNSI Broadcast News in Večer
Table 4. Number of words and number of different words in the BNSI Broadcast News and Večer text corpora

Tekstovni korpus BNSI vsebuje 11 mio besed, od katerih je 175k besed različnih. Obseg smo primerjali s časopisnim korpusom Večer, kjer štirje letniki vsebujejo 95 mio besed, od tega je 736k različnih. Tekstovni korpus BNSI je sicer po svojem obsegu manjši od korpusa Večer, vendar lahko zaradi drugačne zvrsti besedila bistveno pripomore k uspešnosti jezikovnega modeliranja.

Za potrebe analize prekrivanja slovarja tekstovnega korpusa BNSI smo iz obeh korpusov izločili letnik 2001. Lastnosti za izločena letnika so prav tako predstavljene v tabeli 4. Iz obeh letnikov 2001 smo izločili slovar 64k najpogostejših besed, saj se slovar takšne velikosti tipično uporablja v razpoznavalnikih tekočega govora. Med slovarjema velikosti 64k besed je prišlo do 57,8 odstotnega prekrivanja najpogostejših besed. Ta relativno majhen obseg prekrivanja nakazuje pomanjkljivosti uporabe izključno časopisnih korpusov v jezikovnem modeliranju informativnih televizijskih oddaj. V korpusu BNSI smo analizirali tudi delež števil, saj jih je praviloma smiselno modelirati kot ločeno kategorijo. V obsegu 11 mio besed je približno 0,2 mio različnih števil.

Za dodaten prikaz raznolikosti tekstovnega korpusa BNSI smo analizirali pogostost pojavljanja posameznih besed, kar je prikazano na sliki 3.



Slika 3. Pogostost pojavljanja besed v tekstovnem korpusu BNSI Broadcast News
Figure 3. Frequency of words in the BNSI Broadcast News text corpus

Iz grafa na sliki 3 vidimo, da je velik delež besed uporabljen manj kot 10-krat. Samo enkrat je uporabljenih skoraj 120k besed. Velik delež med njimi pripada kategoriji imen, katerih nabor se zaradi dnevnoinformativne narave jezikovnega vira zelo dinamično spreminja. Zato je zelo pomembno, da v jezikovni vir zajamemo čim daljše časovno obdobje, kar smo tudi upoštevali pri izdelavi baze BNSI Broadcast News. Na drugi strani je samo 1238 besed uporabljenih več kot 1000-krat. V zadnjem koraku analize tekstovnega korpusa BNSI smo poiskali tri najpogostejše besede (tabela 5).

	beseda	frekvenca
1.	je	350k
2.	v	344k
3.	in	223k

Tabela 5. Tri najpogostejše besede v tekstovnem korpusu BNSI Broadcast News
Table 5. Three most frequent words in the BNSI Broadcast News text corpus

Kot vidimo iz tabele 5, pripadajo vse tri najpogostejše besede kategoriji funkcijskih besed. Rezultat se sklada s pričakovanji, ki temeljijo na lastnostih slovenskega jezika. Najpogostejša beseda v tekstovnem korpusu BNSI je pomožni glagol "je", ki je uporabljen kar 350k-krat.

6 Sklep

V članku smo predstavili izdelavo baze BNSI Broadcast News, ki je nov govorni in tekstovni vir na področju avtomatskega razpoznavanja tekočega slovenskega

govora. Primerjava značilnosti baze BNSI z italijansko bazo IBNC, ki je tudi nastala na podlagi arhivskih posnetkov, je pokazala podobnost obeh. Edina večja razlika, ki ni posledica razlike v formatu in jeziku dnevnoinformativnih oddaj (število oddaj, govorcev, velikost slovarja...), je večji delež razreda f4 v bazi BNSI, ki je posledica določil v navodilih za zapisovanje. Baza bo širši raziskovalni skupnosti dostopna za uporabo prek evropske agencije za distribucijo jezikovnih virov ELDA.

Predstavljeni projekt izdelave baze BNSI Broadcast News smo v nadaljevanju razširili z dodatnimi 36 urami govornega gradiva v učnem naboru. Ker pa bo le-ta namenjen izključno interni uporabi, ga nismo vključili v predstavitev baze. Kot zadnji korak projekta BNSI smo izdelali tudi prvo slovensko govorno bazo tujih govorcev za razpoznavanje tekočega govora SINOD, ki je dopolnilo baze BNSI. Podrobnosti o bazi SINOD je mogoče najti v [17].

Naslednji korak dela z bazo BNSI Broadcast News je razvoj sistema za razpoznavanje tekočega slovenskega govora z velikim slovarjem besed. Prvi – preliminarni – rezultati s področja segmentiranja zvočnega signala in s področja razpoznavanja govora so bili že predstavljeni v [18].

7 Zahvala

Avtorji članka se zahvaljujemo uslužbencem RTV Slovenija, ki so nam pomagali pri zajemanju gradiva za izdelavo baze BNSI Brodcast News. Posebna zahvala gre tudi vsem, ki so sodelovali pri zapisovanju govornega gradiva, vključenega v bazo BNSI.

8 Literatura

[1] A. Žgank, et. al, Govorno voden informacijski portal LentInfo - predhodna analiza rezultatov. *Jezikovne tehnologije 2002*, Ljubljana.

[2] G. Sket, B. Imperl, M-vstopnica - uporaba avtomatskega razpoznavanaj govora v praksi. *Jezikovne tehnologije 2002*, Ljubljana.

[3] Z. Kačič, Pomen združevanja raziskovalnih potencialov pri preseganju jezikovnih pregrad v okviru jezikovnih tehnologij naslednjih generacij. *Jezikovne tehnologije 2002*, Ljubljana.

[4] D. S. Pallett, The role of the National Institute of Standards and Technology in DARPA's Broadcast News continuous speech recognition research program. *Speech Communication*, Vol. 37, Issues 1-2, 1:3–14, 2002.

[5] D. Graff, The 1996 Broadcast News Speech and Language-Model Corpus. *Proc. DARPA Workshop on Human Language Technology*, 1996.

[6] M. Federico, D. Giordani, P. Coletti. Development and Evaluation of an Italian Broadcast News Corpus. *Proc. LREC*, Atene, Grčija, 2000.

[7] S. Galliano, et. al. The ESTER phase 2 evaluation campaign for the rich transcription of french broadcast news. *Proc. INTERSPEECH*, Lizbona, Portugalska, 2005.

[8] J. Brousseau, et. al. Automatic Closed-Caption of Live TV Broadcast News in French. *Proc. Eurospeech 2003*, Ženeva, Švica.

[9] A. Žgank, et. al. Acquisition and annotation of Slovenian broadcast news database. *Proc. LREC 2004*, Lizbona, Portugalska.

[10] A. Žgank, D. Verdonik, A. Zögling Markuš, Z. Kačič. BNSI Slovenian broadcast news database - speech and text corpus. *Proc. INTERSPEECH*, Lizbona, Portugalska, 2005.

[11] LDC domača stran, http://www.ldc.upenn.edu/Projects/-Corpus-Cookbook/transcription/broadcast-speech/-english/index.html

[12] LIMSI domača stran, http://www.etca.fr/CTA/-gip/Projets/Transcriber/fr/user.html

[13] COST 278 BN SIG domača stran, http://cost278.org/bn

[14] C. Barras, E. Geoffrois, Z. Wu, M. Liberman. Transcriber: Development and use of a tool for assisting speech corpora production. *Speech Communication*, Vol. 33, Issues 1-2, 5-22, 2001.

[15] R. Schwartz, H. Jin, F. Kubala, S. Matsoukas, Modeling those F-Conditions - or not, *Proc. DARPA Speech Recognition Workshop 1997*, pp 115-119, Chantilly, VA.

[16] R. Amaral, I. Trancoso, Topic Indexing of TV Broadcast News Programs. *Lecture Notes in Computer Science*, Springer, Vol. 2721, Jan 2003, 219–226.

[17] A. Žgank, D. Verdonik, A. Zögling Markuš, Z. Kačič. SINOD - Slovenian non-native speech database. *Proc. LREC 2006*, Genova, Italija.

[18] A. Žgank, T. Rotovnik, M. Sepesy Maučec, Z. Kačič. Osnovna zgradba razpoznavalnika slovenskega tekočega govora UMB Broadcast News, *Jezikovne tehnologije 2006*, Ljubljana.

Andrej Žgank je doktoriral leta 2003 na Fakulteti za elektrotehniko, računalništvo in informatiko Univerze v Mariboru, kjer je trenutno zaposlen kot višji raziskovalec v Laboratoriju za digitalno procesiranje signalov. Raziskovalno se ukvarja s področjem avtomatskega razpoznavnja govora in razvojem telekomunikacijskih storitev.

Darinka Verdonik je doktorirala na Filozofski fakulteti v Ljubljani (2006). Zaposlena je kot samostojna raziskovalka na Fakulteti za elektrotehniko, računalništvo in informatiko Univerze v Mariboru. Njeno raziskovalno področje so jezikovne tehnologije in pragmatično jezikoslovje.

Zdravko Kačič je redni profesor na Fakulteti za elektrotehniko, računalništvo in informatiko Univerze v Mariboru. Vodi Laboratorij za digitalno procesiranje signalov. Njegovo raziskovalno področje vključuje analizo in razčlenitev kompleksnih zvočnih scen, sisteme razpoznavanja govora in zajemanje govornih baz podatkov.