Korpus in baza Gos Videolectures

Darinka Verdonik

* Fakulteta za elektrotehniko, računalništvo in informatiko, Univerza v Mariboru
Koroška 46, 2000 Maribor
darinka.verdonik@um.si

1. Uvod

Leta 2011 je bil dokončan prvi sklop referenčnega govornega korpusa Gos. Zajetih je bilo 120 ur posnetkov govora oz. 1 mio. besed. S tem obsegom se korpus Gos uvršča na spodnjo mejo primerljivih referenčnih govornih korpusov za ostale evropske jezike, ki se v zadnjih letih v vse več jezikih bližajo obsegu 10 mio. besed.

Z namenom razširitve obstoječih gradiv govornega korpusa, hkrati pa tudi izdelave dodanega avdio-tekstovnega gradiva za razvoj avtomatskega razpoznavanja tekočega govora, se je leta 2016 začel projekt izdelave dodatnega korpusa in avdio baze h korpusu Gos, poimenovanega Gos Videolectures. Projekt[1] izvaja Fakulteta za elektrotehniko, računalništvo in informatiko Univerze v Mariboru. Zaključil se je leta 2018, celotna predvidena baza obsega ca. 14 ur govora oz. 130.000 besed.

2. Pregled stanja

Poleg referenčnega govornega korpusa Gos obstaja za slovenski jezik še nekaj večinoma manjših ali specifičnih govornih zbirk oz. korpusov.

V repozitoriju Clarin.si sta pod kategorijo »speech database« poleg Gos Videolectures na voljo še bazi SNABI (posnetih 132 govorcev, ki so prebrali vsak po 200 povedi, skupaj ca. 15.000 posnetkov) in SOFES 1.0 (transkribirani in segmentirani avdio posnetki poizvedovanj po letalskih informacijah v skupnem obsegu 10 ur). Prek ELRE je distribuirana baza BNSI Broadcast News (Žgank et al., 2004), ki zajema segmentirane avdio posnetke in transkripcije 36 ur govora informativnih TV-oddaj. Ostale govorne baze oz. korpusi niso dostopni neposredno oz. ne kot podatkovna baza. Sorodna bazi BNSI Broadcast news je baza SiBN Broadcast News (Žibert in Mihelič, 2004), ki zajema 29 ur govora informativnih TV-oddaj. Baza Sloparl (Žgank et al., 2006) vključuje 100 ur govora s transkripcijami v obliki obdelanih magnetogramov parlamentarnih razprav – gre za področno zelo specifično in zelo grobo transkribirano govorno bazo. Govorna baza projekta Translectures (Golik et al., 2013) zajema 33 ur govora za slovenščino. Narečni korpus vasi Kopriva GOKO (60 minut posnetkov) (Šumenjak, 2013) zajema posnetke narečnega govora vasi Kopriva in je dostopen prek iskalnika, ne pa kot podatkovna baza. Podobno velja za narečni korpus GOSP, ki vključuje govor vasi Osp v Slovenski Istri.

3. Korpus in avdio baza Gos Videolectures

3.1. Izbor gradiv

Gos Videolectures skuša ustrezno upoštevati tako potrebe jezikoslovja kot potrebe govornih tehnologij po jezikovnih virih. Zajema več kot 14 ur (130.000 besed transkripciji) izbranih posnetkov javnih predavanj s portala Videolectures.net. Posnetki so izbrani tako, da zastopajo različna strokovna področja in različne skupine govorce (predavateljev) glede na spol, starost in regijo.

Razlog za širitev korpusa Gos na področje javnih predavanj je aktualnost tega področja tako za jezikoslovne raziskave kot avtomatsko razpoznavanje govora. Z jezikoslovnega vidika imamo pri tem opraviti z akademskim jezikom. Gre za jezik javnega diskurza, ki ima po eni strani velik vpliv na oblikovanje slovenskega govorjenega standardnega (zbornega) jezika, po drugi strani pa skozi procese šolanja vpliva tudi na vsakdanji govorjeni jezik v visoko šolstvo vključene populacije, ki je vse večja. Ob tem ta jezik pomembno vpliva na razvoj strokovne terminologije in (bolj ali manj) uspešno širjenje slovenščine na eno najbolj zahtevnih in terminološko najbolj hitro razvijajočih se področij – znanost. Z jezikoslovnega vidika je jezik predavanj zato izredno aktualno področje.

Tudi za razvoj razpoznavanja govora so javna predavanja eno najbolj aktualnih področij za aplikacijo tehnologije: avtomatsko razpoznavanje govora v javnih predavanjih je prvi korak k avtomatskemu strojnemu prevajanju javnih predavanj, kar bi omogočilo boljšo dostopnost vsebin neslovensko govorečim (referenčni projekt s tega področja je bil Translectures – https://www.translectures.eu/). Za domače uporabnike bi z avtomatskim razpoznavanjem govora v javnih predavanjih omogočili naprednejše metode avtomatskega

[1] Projekt sta sofinancirala Ministrstvo za kulturo RS in Javna agencija za raziskovalno dejavnost Republike Slovenije iz državnega proračuna v okviru raziskovalnega programa P2-0069.

indeksiranja ter boljše avtomatsko iskanje in premikanje po vsebinah predavanj v bazah, kot je Videolectures. Izreden pomen bi imela integracija tehnologije avtomatskega razpoznavanja govora na področje javnih predavanj za gluhe in naglušne.

Tabela 1 predstavlja razporeditev izbranih posnetkov v Gos Videolectures glede na strokovno področje. Pri definiranju osnovnih področij smo izhajali iz šifrantov ved, področij in podpodročij, ki veljajo v raziskovalni dejavnosti in jih beleži ARRS (https://www.arrs.gov.si/sl/gradivo/sifranti/). ARRS deli vede na naslednje velike sklope: naravoslovje, tehnika, medicina, biotehnika, družboslovje, humanistika, interdisciplinarne raziskave. Vsaka od teh ved ima več področij in podpodročij. Precej podoben je tudi evropski šifrant raziskovalne dejavnosti CERIF, ki loči humanistične vede, družboslovje, naravoslovno-matematične vede, biomedicinske vede in tehnološke vede. OECD in EUROSTAT pa upoštevata šifrant FOS, ki loči naravoslovne vede, tehniške in tehnološke vede, medicinske in zdravstvene vede, družbene vede in humanistične vede. Na podlagi tega smo se odločili deliti izbrana predavanja na 5 področij: humanistika, družboslovje, medicina, naravoslovje/matematika ter tehnika. Pri tem smo si prizadevali kolikor mogoče enakomerno razporediti obseg gradiva po teh področjih. V končni različici v bazi vseeno nekoliko prevladuje področje humanistike, saj smo morali zaradi končnega manjšega števila besed v transkripciji, kot smo predvidevali, v prvotno enakomerno uravnotežen izbor vključiti dodatne razpoložljive posnetke. Čeprav v splošnem velja formula, da je pri govorjenju v eni minuti izgovorjenih 150 besed, je pri javnih predavanjih ta številka okrog 140 besed na minuto.

Zaradi potrebe po uravnoteženju gradiva smo se omejili tudi pri dolžini izbranih posnetkov, in sicer smo pretežno zajemali posnetke, krajše od 45 minut, kar 16 od skupno 37 posnetkov je celo krajših od 20 minut.

Področje	Št. posnetkov	Dolžina	Št. besed
Humanistika	7	4:31:03	39.871
Družboslovje	9	3:29:55	28.840
Medicina	7	2:13:59	17.721
Naravoslovje/matematika	7	2:46:31	24.202
Tehnika	7	2:31:44	21.713
Skupaj	37	15:33:12	132.347

Tabela 1: Razporeditev gradiv Gos Videolectures glede na strokovno področje.

Zajete podatke smo skušali kolikor mogoče uravnotežiti tudi po demografskih kriterijih. Korpus Gos, ki je bil pri tem izhodišče za razmislek, zajema demografske kriterije: spol, starost, dosežena izobrazba in regijski izvor, pa tudi prvi jezik (tuji govorci slovenščine) in državo bivanja (kar se nanaša na slovenske manjšine v sosednjih državah). Za podkorpus Gos Videolectures je treba te kriterije prilagoditi specifikam področja. Zajemanje govorcev iz slovenskih narodnostnih manjših in tujih govorcev slovenščine ni prednostni cilj baze. Prav tako ni smiseln kriterij o izobrazbi, saj kot predavatelji/ce večinoma nastopajo osebe z visoko izobrazbo. Kot ključni demografski kriteriji, ki jih skušamo po najboljših močeh upoštevati pri zajemanju posnetkov, tako ostanejo spol, starost in regijska pripadnost. Vendar – z izjemo spola – uravnotežanje pri tem ni bilo mogoče v celoti, saj smo lahko o starosti in regijski pripadnosti govorcev sodili le na podlagi videza, slušnega vtisa, kraja izvajanja dogodka oz. dostopnih javnih podatkov. Zanesljiv demografski podatek v bazi je zato samo podatek o spolu govorcev, kot je prikazano v tabeli 2.

Govorci	%
Moški	57 %
Ženske	43 %

Tabela 2: Razporeditev gradiv Gos Videolectures glede na spol govorcev.

Čeprav je bil izhodiščni cilj, da je število govorcev po spolu enakomerno razporejeno, je na koncu vendarle v prid moških govorcev, delno zaradi dodajanja gradiv ob koncu zaradi manjšega števila besed v transkripciji, kot je bilo predvidevano, delno pa tudi zato, ker se je pokazalo, da moški govorci v celotni bazi Videolectures.net toliko prevladujejo, da je ob upoštevanju področnih in drugih demografskih kriterijev težko popolnoma uravnovesiti obseg gradiv glede na ta kriterij.

3.2. Transkripcije

Pri transkribiranju gradiv za Gos Videolectures smo izhajali iz specifikacij transkribiranja, definiranih v korpusu Gos (Verdonik in Zwitter Vitez, 2011), in prav tako vključuje zapis na dveh nivojih: pogovornem, kjer zapišemo besede ortografsko (ne fonetično), vendar tako, kot so izgovorjene, ter standardiziranem, kjer različnim variantam neke besedne oblike pripišemo krovno standardno obliko.

Zaradi cilja, da bazo bolje prilagodimo tudi potrebam govornih tehnologij, pa vendarle vključimo nekaj manjših sprememb. O teh smo že obširno razmišljali in jih specificirali v objavah Žgank et al. (2014b) in v Verdonik (2014), zato zainteresiranega bralca napotujemo na ta vira. V splošnem gre za nekoliko natančnejše označevanje akustičnega ozadja in akustičnih dogodkov ter za nekatere posebnosti zapisovanja govora (zapisovanje dvoustničnega 'U' in člena 'ta' v pogovornem zapisu, zapisovanje neverbalnih in polverbalnih glasov, standardizacija nestandardnih polnopomenskih izrazov pa v gradivu Gos Videolectures ni bila aktualna problematika, saj se tovrstni izrazi ne pojavljajo).

Ob transkribiranju smo zabeležili tudi osnovne metapodatke, med drugim podatke o predavanju (regijo, kraj in čas, kdaj je potekalo, ter kratek opis, za kakšno predavanje gre). Za potrebe avdio procesiranja je na podlagi slušnega vtisa dodana tudi (subjektivna) informacija o kakovosti zvočnega posnetka z lestvico od 1 do 12. Za govorce smo zabeležili podatek o spolu, na podlagi dostopnih informacij pa ocenili tudi podatek o starosti (do 35 let ali nad 35 let) ter regionalni pripadnosti (jugozahodna ali severovzhodna).

Transkribiranje v pogovornem zapisu smo izvajali v orodju Transcriber 1.5.1 (Barras et al., 2000). Čeprav je na voljo novejša različica istega orodja, Transcriber AG (http://transag.sourceforge.net/), se je pri njenem testiranju pokazalo, da je nestabilna in ima preveč hroščev, starejša različica, ki teh težav nima, pa hkrati ponuja tudi vse potrebne funkcionalnosti.

Zaradi konsistentnosti zapisov je vse transkripcije pogovornega zapisa izvedel en izkušen zapisovalec. Standardizirani zapis je bil v prvem koraku avtomatsko izdelan ter nato ročno popravljen v orodju Transcriber 1.5.1, pri čemer smo hkrati izvedli tudi dodatno ročno kontrolo pogovornega zapisa in odpravili odkrite napake v zapisu.

4. Zaključek

Korpus in avdio posnetki Gos Videlectures so na voljo za raziskave prek konkordančnika NoSketchEngine pri IJS. Prav tako so dostopne izvorne datoteke, ki jih lahko uporabniki snamejo v repozitoriju CLARIN.SI. Tam so dostopni avdio posnetki v formatu wav, katerih uporaba je vezana na izvorne licence pri Videolectures.net in niso na voljo za komercialno rabo (licenca CC BY-NC-ND 4.0). Transkripcije so na voljo v treh formatih: kot TEI xml, kot vertikalna tabela Sketch Engina in kot izvorne transkripcijske datoteke, ki se lahko odprejo s programom Transcriber 1.5.1 ali drugim podobnim, ki podpira format .trs, oz. v tekstovnem urejevalniku. Dostopne so pod licenco CC-BY 4.0.

Kot smo nakazali v uvodu, v slovenščini še vedno močno zaostajamo v razpoložljivih virih za govorjeni jezik ne samo v primerjavi z velikimi evropskimi jeziki, ampak tudi v primerjavi s takimi, kjer je število govorcev podobno majhno kot pri slovenščini (npr. nizozemski, slovaški, danski, češki). Nekaj dodatnega govorjenega avdio gradiva s transkripcijami lahko v naslednjih letih pričakujemo v okviru projekta Slovenščina na dlani (http://projekt.slo-na-dlani.si/sl/), vendar bo to spet v majhnem obsegu in prilagojeno potrebam projekta, ki je usmerjen v izdelavo sodobnih učnih e-pripomočkov za učenje slovenščine v osnovnih in srednjih šolah. Potreba po novem projektu, ki bi bil usmerjen (tudi) v izgradnjo obsežnejšega kvalitetnega govornega vira, tako ostaja aktualna.

Literatura

Claude Barras, Edouard Geoffrois, Zhibiao Wu, Mark Liberman. 2000. Transcriber: Development and use of a tool for assisting speech corpora production. *Speech Communication, special issue on Speech Annotation and Corpus Tools*, 33(1–2):5–22.

Pavel Golik, Zoltan Tüske, Ralf Schlüter, Hermann Ney. 2013. Development of the RWTH transcription system for Slovenian. V: *Zbornik konference Interspeech 2013*, str. 3107-3111, Lyon, Francija.

Klara Šumenjak. 2013. Priprava gradiva in standardizacija nivojev zapisa za potrebe dialektološkega korpusa GOKO. V: A. Žele, ur., *Družbena funkcijskost jezika (vidiki, merila, opredelitve)*. Obdobja 32, str. 443–449. Ljubljana, Center za slovenščino kot drugi/tuji jezik pri Oddelku za slovenistiko Filozofske fakultete. http://www.centerslo.net/files/file/simpozij/simp32/zbornik/Sumenjak.pdf.

Darinka Verdonik. 2014. Vprašanja zapisovanja govora v govornem korpusu Gos. V: T. Erjavec, J. Žganec Gros, ur., *Jezikovne tehnologije: zbornik 17. mednarodne multikonference Informacijska družba - IS 2014*,

str. 151-156. Ljubljana, Institut Jožef Stefan. http://library.ijs.si/Stacks/Proceedings/InformationSociety/2014/2014_IS_CP_Volume-G_(LT).pdf.

Darinka Verdonik, Ana Zwitter Vitez. 2011. *Slovenski govorni korpus Gos*. Ljubljana, Trojina, zavod za uporabno slovenistiko.

Andrej Žgank, Tomaž Rotovnik, Mirjam Sepesy Maučec, Darinka Verdonik, Jani Kitak, Damjan Vlaj, Vladimir Hozjan, Zdravko Kačič, Bogomir Horvat. 2004. Acquisition and annotation of Slovenian Broadcast News database. V: *Zbornik konference LREC 2004*, str. 2103–2106. Lizbona, Portugalska..

Andrej Žgank, Tomaž Rotovnik, Matej Grašič, Marko Kos, Damjan Vlaj, Zdravko Kačič. 2006. SloParl - Slovenian parlamentary speech and text corpus for large vocabulary continuous speech recognition. V: *Zbornik konference Interspeech 2006*, str. 197-200. Pittsburgh, Pensylvania, ZDA.

Andrej Žgank, Gregor Donaj, Mirjam Sepesy Maučec. 2014. Razpoznavalnik tekočega govora UMB Broadcast news 2014: Kakšno vlogo igra velikost učnih virov? V: T. Erjavec, J. Žganec Gros, ur., *Jezikovne tehnologije – IS 2014*.

Andrej Žgank, Ana Zwitter Vitez, Darinka Verdonik. 2014. The Slovene BNSI broadcast news database and reference speech corpus GOS: towards the uniform guidelines for future work. V: *Ninth International Conference on Language Resources and Evaluation*, str. 2644-2647. Reykjavik, Islandija. http://www.lrec-conf.org/proceedings/lrec2014/index.html.

Janez Žibert, France Mihelič. 2004. Development of Slovenian broadcast news speech database. V: *Zbornik konference LREC 2004*. Lizbona, Portugalska.