NAČELA TRANSKRIBIRANJA IN OZNAČEVANJA POSNETKOV V REFERENČNEM GOVORNEM KORPUSU SLOVENŠČINE

Ana Zwitter Vitez*, Jana Zemljarič Miklavčič, Marko Stabej**, Simon Krek*****
*Trojina, zavod za uporabno slovenistiko, Škofja Loka
**Filozofska fakulteta, Ljubljana
***Amebis d. o. o., Kamnik; Institut "Jožef Stefan", Ljubljana

UDK 811.163.6'271.16:003.035

V okviru projekta Sporazumevanje v slovenskem jeziku nastaja referenčni govorni korpus slovenskega jezika, ki bo govorni vir za nekatere jezikovne priročnike in različne jezikoslovne raziskave. Zaradi praktične namembnosti govornega korpusa so ključni cilji njegove gradnje čim bolj pregledne iskalne možnosti in čim lažja berljivost transkripcij. V prispevku predstavljamo načela označevanja posnetkov ter segmentiranja in transkribiranja govora.

govorni korpus, govorjeni jezik, transkribiranje, segmentiranje

The project Communication in Slovene includes the construction of a reference corpus of spoken Slovene, which will function as a resource for certain language guides and research projects. Due to its practical goals, key aims of the corpus are straightforward search options and easy-to-read transcription. This paper presents the method to be used for the mark-up of recordings, and for segmenting and transcribing speech samples.

spoken corpus, spoken language, transcribing, segmenting

1 Referenčni govorni korpus slovenščine

Med pisnimi besedili in spontanim govorjenim diskurzom so bistvene razlike. Včasih jih opazimo ob prebiranju spletnih forumov[1] in kratkih telefonskih sporočil, ki v pisni prenosnik prenašajo govorjeno podobo jezika. V takih primerih je tvorcu sporočila popolnoma vseeno, kako ga bo napisal, važno je, da naslovnik razume njegov pomen. Pri večjih sistematiziranih zbirkah besedil pa so natančna pravila označevanja in transkribiranja posnetkov ključnega pomena, ker lahko uporabniku le tako omogočajo kvalitetno iskanje.

S tovrstnimi nalogami smo se spopadli v okviru projekta Sporazumevanje v slovenskem jeziku (SSJ; www.slovenscina.eu), kjer nastaja referenčni govorni korpus v obsegu enega milijona besed (ali približno 112 ur posnetkov). Predstavljal bo govorno komponento referenčnega korpusa slovenskega jezika in bo v projektu govorni vir za dva jezikovna priročnika:

– leksikalno podatkovno bazo (s podatki o frekvenci in pomenski strukturi ter z zgledi rabe),

– pedagoško korpusno slovnico.

Referenčni govorni korpus bo omogočal tudi druge korpusne raziskave govorjene podobe slovenskega jezika v najrazličnejših govornih situacijah, zlasti v leksikografiji, skladnji, analizi diskurza, govornih tehnologijah in sociolingvistiki, pa tudi v tradicionalnem opisnem jezikoslovju. Zato pravila gradnje in označevanja posnetkov upoštevajo tudi njegovo drugotno namembnost.

1 Na primer, ko na forumu www.cveka.com v okviru debate o mozoljih zasledimo tri sporočila istega najstnika, prvič *hwala*, drugič *tnx* in tretjič *thnks*.

V prispevku bomo po kratkem pregledu transkripcijskih praks govorjene slovenščine in ključnih mednarodnih priporočil predstavili načela označevanja in transkribiranja posnetkov pri referenčnem govornem korpusu.

2 Pregled obstoječih praks

2.1 Transkribiranje slovenskih govorjenih besedil

V slovenskem prostoru skupnega standarda transkribiranja še ni, zasledimo pa lahko tri usmeritve bistveno različnih praks transkribiranja govorjenih besedil:

1. V slovenski dialektologiji se običajno uporablja t. i. tradicionalna slovenska fonetična transkripcija oz. fonetična transkripcija OLA (Slovanski lingvistični atlas) z dodanimi različnimi diakritičnimi znaki. (Zorko, 1995).

2. V besediloslovnih in pragmatičnih raziskavah govorjene slovenščine, pa tudi v spletnih forumih, klepetalnicah, blogih in nekaterih literarnih delih se večinoma uporablja ortografski zapis pogovornega jezika z naslednjima skupnima značilnostma:

– raba slovenskega knjižnega črkopisa brez dodatnih posebnih znakov (npr. za polglasnik),

– zapis ponazarja pojave moderne vokalne redukcije in druge pogovorne in narečne prvine govorjene slovenščine (npr. maš čevle, to je zloml, jes mam, notr držim ...).

3. V jezikovnotehnološki praksi (Žgank idr., 2004, 2006, Zemljarič Miklavčič, 2007) se uporablja v glavnem poknjižen zapis govorjenega jezika, iz katerega niso več vidni pogovorni in narečni pojavi. V tej smeri je nastal tudi poskus definiranja pravil za transkribiranje govornih korpusov (Zemljarič Miklavčič, 2007).

Avtorji se zelo različno odločajo tudi o tem, s kakšnimi podatki o govorcih in posnetkih opremljajo svoje transkripcije.

2.2 Mednarodni standardi

Za referenčni govorni korpus so najbolj relevantna priporočila evropske iniciative EAGLES (Expert Advisory Group on Language Engineering Standards),[2] ki priporočajo tri ravni transkripcije:

S1 – ortografska predstavitev besedila v standardni (knjižni) normi,

S2 – fonemska predstavitev besed v citatni obliki (tj. v obliki, kot so besede izgovorjene v izolaciji),

S3 – fonetična transkripcija, ki predstavlja dejansko glasovno podobo izjave.[3]

Skupina EAGLES se zavzema tudi za označevanje identitete govorca, menjavanja govorcev in hkratnega govora. Pri zapisu govora priporočajo zapis glasovnih polleksikalnih enot (eee, mhm, aha itn.) in neleksikalnih enot (smeh, kašljanje, jok itn.), samopopravljanj, besednih fragmentov in nerazumljivih fragmentov. Na prozodični ravni priporočajo označevanje premorov.

3 Izhodišča za definiranje pravil transkribiranja

Referenčni govorni korpus slovenskega jezika predstavlja zbirko govorjenih besedil, pri kateri lahko predvidevamo zelo različne iskalce z zelo različnimi predznanji, koncepti in interesi. Zato smo pri označevanju posnetkov skušali ohraniti tiste kontekstne informacije, ki so pomembne kot potencialni iskalni pogoj za uporabnika (podatke o diskurzih, podatke o govorcu, dejansko govorjeno podobo diskurza). Hkrati pa želimo, da bi uporabnik glede na pravila zapisa govora čim hitreje našel želeno obliko.

4 Označevanje in transkribiranje posnetkov

Glede na različne metode in cilje jezikoslovnih raziskav bo lahko uporabnik po govornem

2 http://www.ilc.cnr.it/EAGLES/home.html.

3 Za fonemski in fonetični zapis priporočajo uporabo fonetične abecede SAMPA oz. X-SAMPA.

korpusu iskal z različnimi kriteriji. Zato je vsak posnetek opremljen z naslednjimi podatki:

1. podatki o govorcih,
2. podatki o diskurzu,
3. transkripcija govora s strukturo diskurza.

4.1 Podatki o govorcih

Podatki o govorcih v govornem korpusu vključujejo naslednje:

1. identifikacijsko kodo govorca,
2. spol,
3. starost (do 10, od 10 do 14, od 15 do 18, od 19 do 24, od 25 do 34, od 35 do 59, nad 60, nedoločno (ni podatka)),
4. regionalno pripadnost govorca glede na registrsko območje (če posamezen govorec zaradi daljšega bivanja na različnih območjih čuti pripadnost različnim regijam, so zabeležene vse ustrezne regije):
   - SV (MS, MB, SG, CE),
   - JZ (NM, KK, LJ, KR, GO, PO, KP),
   - ostale (Italija, Avstrija, Madžarska, tujina, nedoločno (ni podatka)).
5. izobrazba (OŠ ali manj, SŠ, višja ali visoka šola, fakulteta ali več, nedoločno (ni podatka)),
6. prvi jezik (slovenščina, angleščina, nemščina, italijanščina, madžarščina, južnoslovanski jeziki (brez slovenščine), albanščina, drugi: slovanski, germanski, romanski, neindoevropski, nedoločno).

4.2 Podatki o diskurzu

Podatki o diskurzu vključujejo naslednje:

1. identifikacijsko kodo diskurza,
2. dolžino posnetka v minutah in sekundah,
3. tip diskurza glede na kriterije, definirane v specifikacijah za zbiranje gradiva (javni informativno-izobraževalni, javni razvedrilni, nejavni nezasebni, nejavni zasebni),
4. vrsto situacije (npr. televizija, radio, osnovna šola, srednja šola, fakulteta, telefon, osebni stik),
5. opis diskurza (npr. ime televizijske hiše in ime oddaje, ime radijske postaje in ime oddaje, tip institucije, kjer poteka nejavni nezasebni diskurz itn., tip interakcije, npr. doma/družina, doma/prijatelji, delovno mesto/sodelavci itn.),
6. regijo (kjer poteka diskurz):
   a. za javni diskurz: SV Slovenija, JZ Slovenija, celotna Slovenija,
   b. za nejavni diskurz: MS, MB, SG, CE, LJ, NM, PO, KR, KK, KP, GO, Italija, Avstrija, Madžarska, Neslovenci,[4] nedoločno,
7. vir (v primerih, ko prejmemo posnetke od zunanje institucije ipd.),
8. kraj: zemljepisni kraj (mesto, naselje, vas itn.), kjer je potekal diskurz,
9. čas: datum in okvirna ura, ko je potekal diskurz,
10. udeležence: število aktivnih udeležencev,
11. opis govornega dogodka: opis najvažnejših kontekstnih značilnosti (npr. tema, namen pogovora, razmerja med udeleženci ...).

4.3 Transkripcija

4.3.1 Struktura diskurza

Diskurzi v govornem korpusu so segmentirani na manjše enote, vloge in izjave. Izjava je osnovna enota govora, ki približno ustreza pojmu povedi v pisnem jeziku in je prozodično, semantično in skladenjsko prepoznavna in analizabilna enota. Vloga pomeni govor enega govorca, dokler ga ne prekine drug govorec. Vloga je lahko sestavljena iz ene ali več izjav.

Pri hkratnem govoru dveh ali več govorcev ne označujemo natančno, kateri deli besedila so izgovorjeni hkrati. Za začetek hkratnega govora štejemo začetek izjave, v kateri se vključi drug govorec, za konec hkratnega govora

4 Posebna oznaka za diskurze, v katerih sodelujejo pretežno tuji govorci slovenščine, prilagojena določilom v specifikacijah zbiranja gradiva.

štejemo konec zadnje izjave, v kateri se pojavlja hkratni govor.

4.3.2 Zapis govora

Pri zapisu govora se je hitro pokazalo, da nekaterih ciljev (hitro in enostavno transkribiranje, dejanska podoba diskurza, avtomatsko iskanje po besednih oblikah z enako oblikoslovno in semantično vlogo, a različnimi glasovnimi podobami) ni mogoče rešiti z eno samo rešitvijo. Zato smo ustvarili dva nivoja zapisa govora: na prvem nivoju zapisa, ki ga imenujemo »pogovorni zapis«, zapišemo besede sicer ortografsko (ne fonetično!), vendar tako, kot so izgovorjene; na drugem nivoju, ki ga imenujemo »knjižni zapis«, pa »poknjižimo« zapis na tak način, da različnim variantam neke besedne oblike (npr. *mam*, *jemam*) pripišemo krovno knjižno obliko (npr. *imam*). Tako s prvim nivojem omogočimo dober vpogled v besedje in oblike govorjenega jezika, z drugim nivojem pa razširimo iskalne možnosti ter omogočimo uspešnejše nadaljnje avtomatsko označevanje besedil.

4.3.2.1 Prvi nivo zapisa govora – pogovorni zapis

Osnovni cilj pogovornega zapisa je čim bolj zvesta predstavitev glasovne podobe govora na čim bolj berljiv način. Zato govor zapisujemo v veljavnem slovenskem črkopisu, od pravopisne norme pa zapis odstopa na mestih, kjer določena izgovorjena beseda odstopa od t. i. zborne izreke.

Najpogostejša odstopanja od zborne izreke so predstavljena v naslednjih točkah:

1. Redukcije:
   a. Glasov, ki niso izgovorjeni, ne zapisujemo, npr. *tud*, *neki*, *tko*, *mam*, *čevli* ...
   b. Polglasnika ne zapisujemo posebej pri:
      i. zvočnikih r, l, m, n: *sn*, *pr*, *mislm*, *hitr*, *prjatlci* ...,
      ii. enoglasovnih predlogih, členkih ipd.: *s*, *z*, *d* ... (tudi če so izgovorjeni zložno, s polglasnikom),
      iii. enozložnih besedah: *jz*, *nč*...
   c. Polglasnik lahko zapisujemo z »e« v dvoali večzložnih besedah, npr. *kešni* (*kakšni*), razen pred zvočniki m, n, r, l (*zloml*, *mislm*, *hitr* ...).
   d. Zapisovanje oblik pomožnega glagola »biti«:
      i. redukcije »bi« v »b« zapisujemo kot samostojno besedo, npr. *ne b* (*ne bi*), *če b* (*če bi*), *pa b mene* (*pa bi mene*), *najraj b vidu* ...
      ii. redukcije in premene oblik za prihodnjik (*bom*, *boš*, *bo* ...) zapisujemo na naslednji način: *čev* (*če bo*), *navm* (*ne bom*), *nav* (*ne bo*) ...

2. Premen po zvenečnosti v pisavi ne upoštevamo (*podplutba*, *grandž scena* ...).

3. Dvoustnični v zapisujemo s črko »v« (*prov*, *nav*, *navm*, *odpravt*, *davn* ...) oz. tudi z »l«, če tako izhaja iz knjižne norme (*kosil* (v pomenu *kosilo*), *mel* (v pomenu *imel*)). Če je u samoglasniški, ga pišemo s črko »u« (*pršu*, *vidu* ...).

4. Pokrajinsko specifične foneme pišemo z najbližjimi ustreznimi črkami, odvisno tudi od izgovorjave v konkretnih primerih, npr. »*šiest*«, »*šuola*«; »*h*« za zveneči primorski h (hriem).

5. Podaljšane neleksikalne enote pri iskanju formulacije pišemo s tremi črkami, in sicer: *eee*, *eem*, *mmm* ... oziroma z nizom črk, ki najbolje ustreza dejanski izgovorjavi.

6. Prekinjene besede označimo s praznim oklepajem stično za besedo, npr. *lju()*.

7. Ločil ne uporabljamo, izjemi sta:
   a. vprašaj za vprašanja,
   b. klicaj za izrazito ukazujoč govor oz. ob zavpitju ali vzkliku, z namenom, da si uporabniki lažje predstavljajo zvočno podobo govora in lažje razumejo pomen.

8. Izjave začenjamo z malo začetnico.

9. Lastna imena:
   a. Domača lastna imena zapisujemo tako, kot so izgovorjena, vendar z veliko začetnico skladno s pravopisom, npr. *Delo*, *Brežice*. Večbesedna lastna imena

dodatno označimo z zavitimi oklepaji (npr. *[Novo mesto]*, *[Ministrstvo za kulturo Republike Slovenije]* itn.).

b. Tuja lastna imena zapisujemo tako, kot so izgovorjena, vendar z veliko začetnico, npr. *Bler*, *Hjuston*. Če so večbesedna, jih označimo z zavitimi oklepaji, npr. *[Nju Jork]*.

10. Osebne podatke, izrečene v posnetkih, anonimiziramo, tako da označimo samo vrsto podatka (npr. *[ime]*, *[priimek]* ...).

11. Kratice pišemo tako, kot so izgovorjene, vendar skupaj, če gre za eno kratico, npr. *erteve*, *teve*. Če je kratica lastno ime, jo pišemo z veliko začetnico, npr. *Sazuja*, *Tevetri* itn.

12. Pragmatično pomembne nejezikovne zvoke (npr. zvonjenje telefona, na katerega se odzove govorec), označimo z oznako »zvok«. Če zvok ne vpliva na diskurz, ga ne označujemo.

13. Prozodičnih lastnosti govora, kot so jakost, intonacija, trajanje glasov in krajši premori, ne označujemo. Daljše premore (več kot 1,5 sekunde) označimo kot prazno izjavo brez besedila in brez govorca ter dodamo oznako »premor«. Ne opisujemo posebej, kaj se je v premoru dogajalo (npr. tišina, reklame ipd.).

4.3.2.2 Drugi nivo zapisa govora – knjižni zapis

Zaradi boljših iskalnih možnosti bo uporabniku dostopen tudi drugi nivo zapisa govora, ki bo vsaki besedi pripisal najbližjo knjižno besedno obliko.

Osrednje vodilo pri tem je, da pri pretvorbi pogovornega zapisa v knjižni zapis odpravimo glasoslovne premene, ki so prisotne pri posamezni besedni obliki, da dobimo knjižno različico istega leksema. Na drugih jezikovnih ravneh besed ne spreminjamo. Če določenega leksema ni v knjižni normi, ga ohranimo v obliki, ki se pojavlja v govoru. Ker se faza drugega nivoja zapisa šele začenja, bodo natančna načela transkribiranja še dodelana.

5 Zaključek

Označevanje in transkribiranje besedil je torej zasnovano v funkciji uporabnosti govornega korpusa, ki zahteva čim lažjo berljivost transkripcij in preglednost iskalnih kriterijev pri zajetih besedilih. Seveda pa je prvi, pogovorni nivo transkribiranja, še vedno kompromis, ker ne omogoča fonetične transkripcije in prozodičnih oznak. Zato bo za raziskovanje fonetične podobe jezika v letu 2009/2010 vzpostavljen konkordančnik, ki bo omogočal kompleksnejše iskanje in pregledovanje gradiva ter povezavo transkripcij s pripadajočimi zvočnimi datotekami, ob dodatnih zagotovljenih sredstvih pa so v naslednjih letih mogoče tudi razširitve funkcionalnosti korpusa s fonetično transkripcijo (vsaj za del gradiva), oblikoslovnim in skladenjskim označevanjem ter seveda dodajanjem novega gradiva.

Literatura in viri

GARG, Saurabh, MARTINOVSKI, Bilyana, ROBINSON, Susan, STEPHAN, Jens, TETREAULT, Joel, TRAUM, David R., 2004: Evaluation of transcription and annotation tools for a multi-modal, multi-party dialogue corpus. *Proceedings of 4th Internation Conference on Language Resources and Evaluation '04*. Lizbona.

KRAJNC, Mira, 2005: *Besedilne značilnosti javne govorjene besede: Na gradivu sej mariborskega Mestnega sveta*. Maribor: Slavistično društvo.

KRANJC, Simona, 1999: *Razvoj govora predšolskih otrok*. Ljubljana: Znanstveni inštitut Filozofske fakultete.

ROHLFING, Katharina, LOEHR, Daniel, DUNCAN, Susan, BROWN, Amanda, FRANKLIN, Amy, KIMBARA, Irene, MILDE, Jan-Torsten, PARRILL, Fey, ROSE, Travis, SCHMIDT, Thomas, SLOETJES, Han, THIES, Alexandra, WLLINGHOFF, Sandra, 2006: Comparison of multimodal annotation tools – workshop report. *Gespraechsforschung - Online-Zeitschrift zur verbalen Interaktion* 7. 99–123.

SMOLEJ, Mojca, 2006: *Vpliv besedilne vrste na uresničitev skladenjskih struktur: Primer narativnih besedil v vsakdanjem spontanem govoru*. Doktorska disertacija. Ljubljana: Filozofska fakulteta.

ŠAROTAR, Dušan, 2007: *Biljard v Dobrayu*. Ljubljana: Študentska založba.

VERDONIK, Darinka, 2006: *Analiza diskurza kot podpora sistemom strojnega simultanega prevajanja govora*. Doktorska disertacija. Ljubljana: Filozofska fakulteta.

VERDONIK, Darinka, ROJC, Matej, 2006: Are you ready for a call? – Spontaneous conversations in tourism for speech-to-speech translation systems. *5th International Conference on Language Resources and Evaluation*. Genova.

VITEZ, Primož, ZWITTER VITEZ, Ana, 2004: Problem prozodične analize spontanega govora. *Jezik in slovstvo* 49/6. 3–24.

ZEMLJARIČ MIKLAVČIČ, Jana, STABEJ, Marko, 2005: Building a pilot spoken corpus. Garabik, Radovan (ur.): *Computer Treatment of Slavic and East European Languages*. Bratislava. 229–240.

ZEMLJARIČ MIKLAVČIČ, Jana, 2006: Korpus govorjene slovenščine. Erjavec, Tomaž, Žganec Gros, Jerneja (ur.): *Jezikovne tehnologije: zbornik 9. mednarodne multikonference Informacijska družba IS 2006*. Ljubljana: Institut Jožef Stefan. 124–127.

ZEMLJARIČ MIKLAVČIČ, Jana, 2007: *Načela oblikovanja govornega korpusa za slovenščino*. Doktorska disertacija. Ljubljana: Filozofska fakulteta.

ZORKO, Zinka, 1995: *Narečna podoba Dravske doline*. Maribor: Kulturni forum.

ŽGANK, Andrej, ROTOVNIK, Tomaž, VERDONIK, Darinka, KAČIČ, Zdravko, 2004: Baza Broadcast News za slovenski jezik (BNSI) in sistem za razpoznavanje tekočega govora. Erjavec, Tomaž, Žganec Gros, Jerneja (ur.): *Informacijska družba IS'2004: Jezikovne tehnologije*. Ljubljana: Institut Jožef Stefan. 94–98.

ŽGANK, Andrej, ROTOVNIK, Tomaž, GRAŠIČ, Matej, KOS, Marko, VLAJ, Damjan, KAČIČ, Zdravko, 2006: Slovenska govorna baza parlamentarnih razprav za avtomatsko razpoznavanje govora. Erjavec, Tomaž, Žganec Gros, Jerneja (ur.): *Informacijska družba IS'2006: Jezikovne tehnologije*. Ljubljana: Institut Jožef Stefan. 115–118.

ŽIBERT, Janez, MIHELIČ, France, 2004: Development, evaluation and automatic segmentation of Slovenian Broadcast News Speech Database. Erjavec, Tomaž, Žganec Gros, Jerneja (ur.): *Informacijska družba IS'2004: Jezikovne tehnologije*. Ljubljana: Institut Jožef Stefan. 94–97.

WELSH, Irvine, 1997: *Trainspotting*. Prevod: A. Skubic. Ljubljana: DZS.