Navodila za zapisovanje govora v projektu Gos Videolectures

Pripravila: Darinka Verdonik
Verzija: 04
Kraj, datum: Maribor, 21. april 2016

Vsebina

0 Uvod

Za transkribiranje uporabljamo orodje Transcriber 1.5.1.

Za vsak posnetek ustvarimo posebno evidenco s podatki o diskurzu, tako da izpolnimo podatke v tabeli **GosVideolectures-govorci-vzorec**. Podatke vnašamo ob koncu transkribiranja posameznega posnetka na podlagi vsebine v posnetku, dodatne informacije najdemo v tabeli Seznam-videov-izbor-dv-2016-final oz. na spletnem portalu Videolectures http://videolectures.net/, kjer poiščemo dotični posnetek.

1 Segmentiranje

1.1 Segmenti oz. izjave

Vsak posnetek segmentiramo na osnovne enote transkribiranja, to so segmenti. Segmenti v podkorpusu Gos Videolectures enako kot v korpusu Gos ustrezajo pojmu izjave, pri čemer razumemo izjavo kot osnovno enoto govora, ki približno ustreza pojmu (kratke) povedi v pisnem jeziku. Je semantično in skladenjsko kolikor mogoče zaključena enota, vedno pa jo zamejujejo ustrezni premori v govoru, ki omogočajo postavitev časovne meje med segmentoma na način, da v zvočnem signalu ni odrezan noben delček fonema predhodne ali naslednje izgovorjene besede. Pri segmentiranju se vedno odločimo za čim krajše enote, saj slednje omogočajo uspešnejše učenje akustičnih modelov.

1.2 Premori

Premore na začetku posnetka ali med govorom, daljše kot 1,5 sek., označimo kot prazen segment/izjavo in vanjo dodamo samo oznako »premor«.

1.3 Označevanje govorcev

1.3.1 Gos Videolectures

Za vsak segment označimo, kdo je govorec. Govorcu določimo kodo po načelu:
Ym-vlxxx – za moške govorce
Yf-vlxxx – za ženske govorke

Y pri tem predstavlja spremenljivko, za katero lahko uporabimo poljubno črko slovenske abecede, razen šumnikov, ki označuje govorca.

xxx je spremenljivka, s katero številčimo govorce v bazi Gos Videolectures od 001 do maks. 999. Pri tem vsak govorec v bazi Gos Videolectures dobi svojo številko. **Vodimo posebno evidenco v Excelu,**

v kateri zapišemo v enem stolpcu ime govorca, v drugem pa kodo, ki mu pripada. Evidenco ustvari transkriptor in jo poimenuje GosVideolectures-evidencagovorcev.xls.

Govorcem pripišemo podatek o spolu. Na podlagi vidnega vtisa opredelimo gorce po starosti na mlajše (do 35 let) ali starejše (nad 35 let). Na podlagi vtisa o jezikovnem kodu, ki ga govorec uporablja, opredelimo zvrst jezika kot prevladujoče knjižno ali prevladujoče pogovorno SV ali prevladujoče pogovorno JZ ali kot tuji govorec. Podatke vnesemo v evidenco govorca **GosVideolectures-govorci-vzorec.xls** in jo shranimo kot tekstovno datoteko z istim imenom, kot je ime posnetka, ter končnico _g1.txt ali _g2.txt, če sta govorca dva in je treba za vsakega izpolniti evidenco.

Za morebitne posamezne replike iz občinstva ne izpolnjujemo posebne evidence. Pri določanju kode takim govorcem v kodi govorca vedno uporabimo X za spremenljivko Y. Ostalo določimo enako kot pri vseh govorcih.

1.4 Menjavanje vlog

V podkorpusu Gos Videolectures ne pričakujemo veliko menjavanja vlog različnih govorcev, saj gre večinoma za govor enega govorca/predavatelja. Posledično ne pričakujemo veliko označevanja menjavanja vlog različnih govorcev ali hkratnega govora. Koliko se to vseeno pojavi, ustrezno označimo (glej zgoraj) vse dodatne govorce in zapišemo njihov govor. Hkratni govor obravnavamo enako kot v korpusu Gos, torej:

Kot začetek hkratnega govora označimo začetek izjave, v kateri se vključi drug govorec, ne glede na to, ali se vključi na začetku ali sredi izjave. Kot konec hkratnega govora označimo konec zadnje izjave, v kateri se pojavlja hkratni govor, ne glede na to, ali se hkratni govor konča že sredi izjave. Če traja hkratni govor dalj časa in lahko znotraj njega postavimo mejo med izjavami, ne da bi s tem rezali govor katerega koli od govorcev, to naredimo. Oporne signale obravnavamo enako kot hkratni govor. V hkratnem govoru ne označujemo, točno kateri deli besedila so izgovorjeni hkrati.

2 Zapisovanje govora

2.1 Prvi nivo zapisa govora – pogovorni zapis

Osrednje vodilo: Govor zapisujemo v veljavnem slovenskem črkopisu in upoštevamo veljavne strategije predstavljanja posameznih glasov z določenimi črkami. Upoštevaje omejitve, ki izhajajo predvsem iz omejenega nabora črk, pa pri tem kolikor mogoče verno predstavimo glasovno podobo govora.

Podrobna pravila zapisa:

2.1.1 Tehnikalije

Izjave začenjamo z malo, ne z veliko začetnico.

Ločil ne uporabljamo, izjema sta:

- vprašaj za vprašanja,
- klicaj za izrazito ukazujoč govor oz. ob zavpitju ali vzkliku,

z namenom, da si uporabniki lažje predstavljajo zvočno podobo govora in lažje razumejo pomen. Vprašaj in klicaj pišemo stično.

Čeprav tako baza Translectures kot baza BNSI Broadcast News vsebujeta ločila, smo v podkorpusu Gos Videolectures glede tega ohranili enako načelo kot v korpusu Gos. Dodajanje ločil govorjeno rabo na silo prilagaja standardom, ki veljajo za pisno rabo, zahteva nekaj časa, truda in dodatnega (pravopisnega) znanja transkriptorjev, hkrati pa ne pomeni posebno pomembne informacije za razpoznavanje govora. Ker je korpus Gos segmentiran na način, da je vsak segment mogoče razumeti kot zaključek povedi, je dejansko edino izpuščeno ločilo pri takem načelu transkribiranja vejica.

Besedne fragmente (prekinjene besede ipd.) označimo s praznim oklepajem stično za besedo, npr. *lju()*. Za besedne fragmente štejejo samopopravljanja – ko govorec začne izgovarjati neko besedo, pa jo sredi izgovarjanja prekine in izreče neko drugo besedo.

Če se v posnetku pojavijo osebni podatki o govorcih (ime, priimek ipd.), jih anonimiziramo tako, da označimo samo vrsto podatka, in sicer na naslednji način: *[ime], [priimek].*

Prav tako anonimiziramo osebne podatke o osebah, ki sicer niso prisotne v diskurzu, so pa vseeno omenjene, če gre za nejavne osebnosti.

Normalno, kot lastno ime, pa zapišemo imena javnih osebnosti, ki so omenjena v diskurzu, npr. imena politikov, športnikov, novinarjev in voditeljev, umetnikov in drugih kulturnih delavcev ter ostalih medijsko opaznih osebnosti.

Za vse osebne podatke, ki jih anonimiziramo, dodatno označimo tisti del zvočnega posnetka, v katerem je izgovorjen osebni podatek oz. zaporedoma več osebnih podatkov, tako da je v transkripciji zapis časovnih mej, ki določajo osebni podatek v zvočnem posnetku. V ta namen uporabimo vedno oznako »Insert background« in odkljukamo vedno samo »shh«. Isto funkcijo, »insert background«, uporabljamo tudi za označevanje zvočnega ozadja, vendar pri zvočnem ozadju izbiramo vedno samo med »music«, »speech« in »other«.

Številke vedno izpišemo z besedo.

Nerazumljive ali nerazločljivo izgovorjene besede ali daljše govorne enote označimo z oznako »neraz« (Insert event oz. Ctrl + d, v okence vpišemo *neraz*). Trajanja ne označujemo.

2.1.2 Fonetične premene

Redukcije

- Glasov, ki niso izgovorjeni, ne zapisujemo, npr. *tud, neki, tko, mam, čevli*...
- Polglasnika ne zapisujemo posebej pri:
    - zvočnikih r, l, m, n: *sn, pr, mislm, hitr, zloml, prjatlci...*
    - enoglasovnih predlogih, členkih ipd.: *s, z, d...* (tudi če so izgovorjeni zložno, s polglasnikom)
    - enozložnih besedah: *js, nč*...
- Polglasnik lahko zapisujemo z »e« v dvo- ali večzložnih besedah, npr. *kešni (kakšni),* razen pred zvočniki m, n, r, l (*zloml, mislm, hitr*...).
- Zapisovanje oblik pomožnega glagola »biti«:
    - 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...*
    - redukcije in premene oblik za prihodnjik (*bom, boš, bo*...) zapisujemo na naslednji način: *čev (če bo), navm (ne bom), nav (ne bo)*...

Premen po zvenečnosti v pisavi ne upoštevamo (*tud dobr, tud tak, grandž scena*...).

Zvočnik dvoustnični v (ni nosilec zloga) zapisujemo s črko »v« (*prov, nav, navm, odpravt, davn*...) oz. tudi z »l«, če tako izhaja iz knjižne norme (*kosil, mel*). Če je u samoglasniški, tj. je nosilec zloga (tudi če gre za predlog *v,* izgovorjen samoglasniško), ga pišemo s črko »u« (*pršu, vidu, u tem delu*...).

Analiza obstoječih zapisov v korpusu Gos je kasneje pokazala, da »tovrstno načelo govorcem slovenščine vseeno ni popolnoma domače, ko morajo zapisovati besede govorjene slovenščine, ki še nimajo ustaljenega »standarda« zapisovanja, in sicer se marsikje namesto predvidenega zapisa z 'v' ali 'l' vrine zapis z 'u' – npr. *laufati*, *šlauf* ali *genau* se v zapisu z 'u' pojavljajo celo v Besedišču in tudi po korpusu Gigafida močno prevladuje različica z 'u', čeprav bi po zgornjem pravilu pisali *lavfati, šlavf, genav*. Podobno so dvojnice lahko pri medmetih, npr. *au* in *av* (po SSKJ).

V zvezi s tem se pojavlja tudi nekaj več nedoslednosti v pogovornem zapisu korpusa Gos, kjer najdemo po večkrat tudi pogovorne zapise tipa *mau* (*malo*), *biu* (*bil*), *šou* (*šel*), *dou* (*dol*), *prou* (*prav*), *dau* (*da bo*), *nou* (*ne bo*) itd., namesto predvidenega zapisa s črko v/l. Kljub temu pa je večinsko zapis z v/l v tovrstnih vlogah prevladujoč in zdi se, da bi bilo spreminjanje načela v zapis z 'u' še bolj problematično: potem bi namreč besede, ki v glasovni podobi sledijo standardu, še vedno pisali z 'v' ali 'l', npr. *imel*, in kontrast z *meu* namesto *mel* bi verjetno vnesel še več zmede in nedoslednosti. Edina sprejemljiva sprememba tega pravila bi zato bila, da se vodi seznam besed ali oblik, za katere lahko po pisnih korpusih sledimo tendenci po pisanju s črko 'u' v teh položajih, ostale pa se še naprej pišejo z 'v' oz. 'l'. Je pa vprašanje, ali ni tako pravilo še bolj problematično s stališča doslednosti zapisovanja kot obstoječe uniformno vodilo.« (Verdonik 2014)

Glede na te ugotovitve se odločimo, da v podkorpusu Gos Videolectures ohranimo obstoječi standard zapisovanja dvoustničnega U z v ali l in smo posebej pozorni na ugotovljene nedoslednosti zgoraj.

Diftonge in druge pokrajinsko specifične foneme, ki jih ni v knjižnem jeziku, pišemo z najbližjimi ustreznimi črkami, odvisno tudi od izgovorjave v konkretnih primerih, npr. *»ej«, »ov«, »je«; »u«* ali tudi *»i«* za u s preglasom; *»h«* ali tudi *»g«* za zveneči primorski h; *»r«* za mehkonebni koroški r itd.

2.1.3 Neverbalni in polverbalni glasovi

Podaljšan polglasnik ali zvočnik *m* ali *n* in njihove kombinacije, ki pogosto zapolnjujejo premore v govoru, pišemo s tremi črkami, in sicer: *eee, eem, een, nnn, mmm...* Druge medmete zapišemo z nizom črk, ki najbolje ustreza dejanski izgovorjavi. Trajanja medmetov ne označujemo posebej.

Te vrste neverbalni in polverbalni glasovi so bili naslednja točka, kjer se je ob analizi obstoječih zapisov v korpusu Gos pokazala potreba po večjem poenotenju in bolj natančnem specificiranju zapisovanja, in sicer:

Načela zapisovanja izhajajo iz dveh stališč: način zapisa naj bi bil govorcem slovenščine čim bližji, hkrati pa naj bi omogočal največjo možno mero avtomatskega procesiranja teh izrazov v govorjenem besedilu. Načela so:

1. *izraze zapišemo raje z eno besedo kot več besedami (npr. ojoj namesto o joj),*
2. *kjer ni bistvene razlike v zvočni podobi in funkciji/pomenu, ohranimo enoten zapis za različne rabe (npr. mhm bi posamično morda zapisali tudi kot ehm, vendar je razmejitev težko objektivno določiti, zato raje ohranjamo vedno mhm),*
3. *izraze zapisujemo prednostno s tremi črkami, tako da se razlikujejo od drugih besed (npr. raje vaa kot va), razen kjer ni nevarnosti, da bi bil zapis identičen zapisu kakih drugih besed, ali če je drugačen zapis že močno uveljavljen (npr. eh),*
4. *dvoustnični U prednostno pišemo z 'v' (av, vav),*
5. *podaljševanje glasov se ne označuje z več črkami, ampak se ohranja enoten zapis (npr. vedno jee, ne jeee ali podobno),*
6. *prednost ima poslovenjen zapis (npr. jes, ne yes, okej, ne ok ali okay).*

Pri zapisovanju se tako opiramo na seznam zapisov neverbalnih in polverbalnih glasov, zapisan v prilogi teh navodil.

Kot izstopajoč neenotni zapis v korpusu Gos (Verdonik 2014) izpostavi neverbalno glasovno zanikanje. Opažene so bile naslednje različice zapisovanja tega pojava: n n, m m, a a, e e, nn, aa, mm. V govoru ti glasovi dejansko nihajo od bolj vokalnega, a-jevskega prek polglasniškega do zvočniškega m ali n. Potreben je bolj enoten in unikaten zapis. V podkorpusu Gos Videolectures uporabljamo dve različici zapisa: nn in aa.

Podoben primer je neverbalno glasovno pritrjevanje, za katerega je bil realiziran zapis mm – tega ohranimo in ga uporabljamo izključno za tovrstno rabo.

2.1.4 Pisanje skupaj in narazen

Zloženke: kadar čutimo, da gre za eno besedno enoto, jih pišemo skupaj in brez vezaja (ne glede na to, ali gre za podredno ali priredno zloženko), če ne predstavljajo ene besedne enote oz. gre za zvezo prislova in pridevnika, ju zapišemo s presledkom kot dve besedi.

Kratice:

a. Pišemo tako, kot so izgovorjene, vendar skupaj, če gre za eno kratico, npr. *erteve, teve, trr*.
b. Če je kratica lastno ime, jo pišemo z veliko začetnico, npr. *Sazuja, Tevetri* itd.

Določni člen *ta:* Določila, ki bi posebej omenjalo pisanje člena 'ta' v tipu 'ta rdeči' (kjer je 'ta' nenaglašen in izgovorjen skupaj s sledečim pridevnikom), v specifikacijah transkribiranja za korpus Gos ni bilo. Iz zapisov v korpusu je razvidna praksa, da se člen piše kot samostojna beseda. Verdonik (2014) v zvezi s tem ugotavlja: »Ob tem pa na nivoju pogovornega zapisa (kot posamezne lapsuse pa posledično tudi na ravni standardiziranega zapisa) vseeno občasno zasledimo stični zapis, zelo pogosto za zvezo *ta mali/ta mala,* npr. *tamal, tamav*, *tamalo, tamali, tamalima, tamavga, tamalga*, poleg te pa bolj kot ne posamično še za zveze *taprav/tapravo, tapravga* (*ta pravi*), *tazaden* (*ta zadnji*), *tamladi* (*ta mladi*), *taprv* (*ta prvi*), *tazadno* (*ta zadnjo*) itd.

Medtem ko je na nivoju standardiziranega zapisa res najbolj praktično in smiselno nestično pisanje, zlasti z vidika kasnejšega oblikoslovnega označevanja, izdelave besednih seznamov in iskanja po besedilu, pa bi veljalo še enkrat razmisliti o možnosti stičnega pisanja v pogovornem zapisu. S tem bi namreč omogočili avtomatsko ločevanje med rabami tipa zaimek + pridevnik (*hvala za ta lep mejl*) in rabami tipa člen + pridevnik (*je bil predračun tak da je šu tist talep lijak ven*), ki jih je mogoče zanesljivo ločevati samo ročno in s pomočjo zvočnega posnetka.«

V zvezi z določnim členom 'ta' zato v podkorpusu Gos Videolectures vzpostavimo prakso da se na ravni pogovornega zapisa piše stično (*je šu tist talep lijak ven*), na ravni standardiziranega zapisa pa kot dve besedi (*je šel tisti ta lep lijak ven*).

2.1.5 Lastna imena in tuje besede

Lastna imena:

- 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}, {Lenart v Slovenskih goricah}, {Ministrstvo za kulturo Republike Slovenije}, {Občina Starše}, {Osnovna šola Ivana Cankarja}* itd.).
- 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}, {Los Endželes}*.

Citatne besede, ki niso lastno ime: pišemo tako, kot so izgovorjene.

Lastnih besed ne označujemo dodatno (ni potrebno dodajanje dogodka Ctrl + d – Named Entities).

Zapisovanje govora v tujem jeziku (cela izjava ipd.) – ne zapišemo, dodamo pa oznako Event, tujjez.

3 Označevanje akustičnega ozadja in akustičnih dogodkov

Kadar se v ozadju govora pojavijo kakšni dalj časa (3 sek. ali več) trajajoči zvoki ali šumi, označimo tak odsek na signalu ter določimo, ali je šum v ozadju glasba, govor ali kaj tretjega (šum). Za vsako akustično ozadje označimo začetek in konec trajanja na signalu s posebno časovno oznako, in sicer tako, da uporabimo funkcijo:
Segmentation -> Insert background –> v pojavnem oknu odkljukamo »music« (glasba v ozadju), »speech« (govor v ozadju), »other« (katerikoli drug šum v ozadju). Opcije »shh« nikoli ne uporabimo za označevanje ozadja, saj je rezervirana za anonimizacijo govorcev.

Označiti moramo začetek in konec trajanja zvoka v ozadju. Ob začetku v pojavnem oknu vključimo kljukico, ob koncu pa v istem pojavnem oknu odznačimo kljukico.

Akustično ozadje označimo samo, kadar se le-to spremeni v primerjavi s tem, kakšno ozadje prevladuje v posnetku večino časa.

Kadar se med govorjenjem pojavijo kratki zvoki, te označimo brez časovne oznake, dodamo samo opis dogodka, tako da uporabimo funkcijo:
Edit -> Insert event ali Ctrl + d, nato v okence vpišemo enega od naslednjih možnih dogodkov:

smehgo -> smeh govorca
smehna -> smeh občinstva
smehob -> smeh govorcev in občinstva
glas -> zvoki, ki nastanejo z govorili, kot so zehanje, vzdih, odkašljanje, pogrkavanje ipd.
dih -> slišen vdih ali izdih med govorjenjem
zvok -> zvoki, ki ne nastanejo z govorili

Vsakemu od teh dogodkov označimo ustrezno trajanje glede na to, ali se prekriva z govorom (določimo start in end of event), samo eno besedo (apply to previous word) ali z nobeno izgovorjeno besedo (instantenious event).

Priloga: Zapisovanje najpogostejših neverbalnih in polverbalnih glasov

Seznam je narejen na podlagi korpusa Gos v obsegu 1 mio. besed, dostopnega na www.korpus-gos.net, marca 2014. Znak # pred zapisom pomeni, da zapis ni enoznačen in je lahko identičen zapisu kake druge besede, npr. veznika, členka ipd. Če pred zapisom ni znaka #, pomeni, da se mu lahko avtomatsko pripiše enaka lema, kot je obstoječi zapis, in oblikoskladenjska oznaka za medmet.

#a
aa (zanikanje)
aaa
aaaa (a z vprašalno intonacijo v vlogi vprašanja)
aam
aan
ah
aha
ahah
ahaha
ahahaha
ahja
ahjoj
ahm
ahoj
#aj
#aja
ajah
ajaj
aje
ajej
ajo
ajoj
alo
ao
aua
auva
av
#ba
bljeh
brum
bu
bvak
buf
bum
bumč
bvum
bzz
bž
ccc (tleskajoči zvok z jezikom, ki se trikrat ponovi in izraža, da nečesa ne odobravamo)
ck
dammm
dh
dum
#e
eee
eem
een
eev
eh
ehe
eheh
ehehe
ej
eje
ejo
ejoj
fuf
fuj
fuu
grr
ha
haha
hahaha
hahahaha
hajaj
#he
heh
hehe
hehehe
hej
hhh
#hi
hihi
hijaj
hijo
hjoj
hjujujujuju
hm
#ho
hoho
hohoho
hohop
hojoj
hopa
hopla
hopsasa
hov
hu
huh
huhu
#i
iii
ija
ijo
ijoj
jah
jaj
jao
jea
jee
#jej
jes
johoho
johoj
joj
jojojojojojo
joo
jov
joz
juhej
juhu
juhuhu
jupi
juu
klink
maa
mahh
mee
mh
mhm
miu
mjav
mm (pritrjevanje)
mmm
nanananananana
nee
nhn
njam
njm
nn (zanikanje)
nnn
#o
oa
oh
ohja
ohjej
oho
ohoho
ohohoho
#oj
oja
ojej
ojla
ojoj
ojojej
ojojo
ojojoj
ojojojo
ojojojoj
ojojojojoj
ola
ooa
ooo
op
opa
opala
ops
ov
ovh
paf
pavf
pff
pha
plop
pom
puf
ratatatatata
rc
rrr
ssk
sss
tada
tadadada
tadam
tarararata
taratatam
tarararan
tarararararararararara
taratataratat
tk
totrolodontodo
tp
tralala
tumbapa
tup
#u

ua
uf
uh
#uhu
uhuhu
ujej
umbapa
#uo
ups
upsala
vaa
vav
vov
zk
šink
šk
ššš
čk
čuf
ču