DOI
https://doi.org/10.18690/um.ff.4.2024.2

ISBN
978-961-286-882-6

NADGRADNJA DIGITALNE SLOVARSKE BAZE ZA SLOVENŠČINO IN SLOVENSKEGA OBLIKOSLOVNEGA LEKSIKONA SLOLEKS S PODATKI O GOVORJENI SLOVENŠČINI: NAČRTI IN CILJI

JAKA ČIBEJ, NEJC ROBIDA, SIMON KREK
Univerza v Ljubljani, Filozofska fakulteta, Ljubljana, Slovenija
jaka.cibej@ff.uni-lj.si, nejc.robida@ff.uni-lj.si, simon.krek@ff.uni-lj.si

V prispevku predstavljamo načrte in cilje za dopolnjevanje jezikovnih virov, kot sta Digitalna slovarska baza za slovenščino in Slovenski oblikoslovni leksikon Sloleks, s podatki o govorjeni slovenščini oz. natančneje o tipično govorjenem besedišču, in sicer predvsem za namene jezikovnotehnoloških potreb (npr. razpoznavalniki in sintetizatorji govora). Po kratkem pregledu sorodnih raziskav predstavimo gradivo, ki ga bomo uporabili za ta namen (korpusa GOS in JANES), ter poglavitne izzive, na katere naletimo pri vključevanju nestandardnega besedišča v obstoječe vire, ki so bili do zdaj namenjeni predvsem pisni standardni slovenščini. Poleg problematike kanoničnih oblik (npr. lavfati/laufati) naslovimo npr. tudi tematiko nestandardnih fonemov ([ˈgɾɔːza] vs. [ˈɦɾɔːza]), nestandardnih izgovorjav standardnih besednih oblik (mislim [ˈmiːslim] → [ˈmiːsləm]) ter nestandardne morfologije (Mihatov, opravičavam). Opisane izzive bomo v okviru projekta MEZZANINE opisali, rešitve pa dokumentirali v smernicah, ki bodo omogočile sistematično polnjenje obstoječih jezikovnih virov s tipično govorjeno leksiko.

Ključne besede:
Sloleks,
leksikon,
govorjena slovenščina,
nestandardno besedišče,
korpusi govorjene slovenščine

Univerzitetna založba Univerze v Mariboru

DOI
https://doi.org/
10.18690/um.ff.4.2024.2

ISBN
978-961-286-882-6

Extending the Digital Dictionary Database of Slovene and the Sloleks Morphological Lexicon of Slovene with Spoken Slovene Data: Plans and Goals

Jaka Čibej, Nejc Robida, Simon Krek
University of Ljubljana, Faculty of Arts, Ljubljana, Slovenia
jaka.cibej@ff.uni-lj.si, nejc.robida@ff.uni-lj.si, simon.krek@ff.uni-lj.si

Keywords:
Sloleks,
lexicon,
spoken Slovene,
non-standard vocabulary,
corpora of spoken Slovene

This paper presents plans and goals for extending language resources such as the Digital Dictionary Database of Slovene and the Sloleks Morphological Lexicon of Slovene with data on spoken Slovene – particularly typically spoken vocabulary – for language technology purposes (e.g. speech recognition and synthesis). After a brief overview of related work, we present the material we will use for this purpose (the GOS and JANES corpora) as well as the main challenges we encounter when incorporating non-standard vocabulary into existing resources that have so far been mainly intended for written standard Slovene. In addition to the issue of canonical forms (e.g. lavfati/laufati), we also address the issues of non-standard phonemes ([ˈgɾɔːza] vs. [ˈɦɾɔːza]), non-standard pronunciations of standard word forms (mislim [ˈmiːslim] → [ˈmiːsləm]) and non-standard morphology (Mihatov, opravičavam). The challenges will be described in the framework of the MEZZANINE project, and the solutions will be documented in guidelines that will enable the systematic extension of existing language resources with typical spoken lexis.

University of Maribor Press

1 Uvod[1]

Slovensko jezikoslovje (tudi korpusno) se je do zdaj pri gradnji jezikovnih virov v veliki meri osredotočalo na pisni jezik, deloma tudi zato, ker je pridobivanje gradiva v pisni obliki enostavnejše in hitrejše ter časovno in finančno manj zahtevno od pridobivanja zvočnih posnetkov, zamudnega transkribiranja in urejanja kompleksnih pravnih omejitev v zvezi z osebnimi podatki in avtorskimi pravicami. V zadnjem desetletju pa je konstanten napredek pri razvoju govornih tehnologij (npr. razpoznavalnikov in sintetizatorjev govora) vzrok za povečano potrebo po gradnji jezikovnih virov za govorjeno slovenščino. V slovenskem korpusnem jezikoslovju je bila ta v primerjavi s pisno (zlasti standardno) slovenščino deležna manj pozornosti: na voljo so že nekateri korpusi in podatkovne baze govorjene slovenščine – npr. GOS v1.1 (Zwitter Vitez et al. 2015), skladenjsko označeni Slovenian UD Treebank (Dobrovoljc & Nivre 2016), GOS-VL v4.2 (Verdonik et al. 2021) ter Artur v0.1 (Verdonik et al. 2022) in GOS v2.0 (Zwitter Vitez et al. 2023); zadnja dva sta bila izdelana v nedavno zaključenem projektu Razvoj slovenščine v digitalnem okolju (RSDO) –, precej manj pozornosti pa je bilo namenjene govorjeni slovenščini pri leksikonskih in leksikografskih virih, kot sta Slovenski oblikoslovni leksikon Sloleks (Čibej et al. 2022) in Digitalna slovarska baza slovenščine (Kosem et al. 2021)[2] Sloleks, ki v različici 3.0 vsebuje iztočnice, njihove pregibne oblike in podatke o njihovih izgovorjavah (v mednarodnih fonetičnih abecedah IPA in SAMPA), še ni bil razširjen s podatki, ki so tipični za govorjeno slovenščino. Pomanjkanje podatkov o npr. tipično govorjenem besedišču predstavlja oviro pri uspešni implementaciji jezikovnih tehnologij za slovenščino, kot so npr. razpoznavalniki govora; če določena beseda (npr. zrihtati, poštimati) in podatki o njenem izgovoru niso vključeni v zaledne podatkovne baze, na katere se zanaša razpoznava govora, je razpoznavalnik bodisi ne razpozna ali pa jo razpozna napačno.[3]

[1] Prispevek je nastal v okviru raziskovalnega projekta *Temeljne raziskave za razvoj govornih virov in tehnologij za slovenski jezik* (J7-4642) in raziskovalnega programa *Jezikovni viri in tehnologije za slovenski jezik* (P6-0411), ki ju financira Javna agencija za znanstvenoraziskovalno in inovacijsko dejavnost Republike Slovenije (ARIS).

[2] Tudi rastoči Slovar slovenskega knjižnega jezika (eSSKJ) se denimo opira na pisno jezikovno gradivo (npr. korpusi Gigafida 2.0, slWaC, KAS; Divjak Race in Gliha Komac 2022).

[3] Na težave pri razpoznavi tipično govorjenega besedišča smo npr. že naleteli pri projektu Online Notes, ki je namenjen razvoju sistema za samodejno prevajanje slovenskih predavanj v tuje jezike: https://www.cjvt.si/online-notes/

Če želimo tovrstne pojave vključiti v leksikon, moramo proučiti, kako sistematično navajati podatke o govorjeni slovenščini v pisni obliki tako, da so intuitivni uporabnikom in čim bolje odsevajo jezikovno rabo, hkrati pa so strojno berljivi ter neposredno uporabni za razvoj jezikovnih tehnologij. To praznino v jezikovnih podatkih nameravamo zapolniti v projektu MEZZANINE. Ena od nalog delovnega sklopa 4 je med drugim tudi leksikonska in leksikografska[4] obravnava besedišča, ki se tipično pojavlja v govorjeni slovenščini, ne (oziroma zelo redko) pa v pisni.

2 Sorodne raziskave

Z govorjenim slovenskim jezikom so se do 60. let prejšnjega stoletja še največ ukvarjali dialektologi, kot raziskovalce govorjenega jezika pa Smolej (2012) pa v svoji raziskavi zvrstnosti besedil v spontanem govoru izpostavlja predvsem Bredo Pogorelec, Jožeta Toporišiča in Borisa Urbančiča. Od raziskav iz zadnjega desetletja, se jih večina ukvarja predvsem z žanrsko analizo (Lengar Verovnik 2010, Verdonik 2017), tudi kot s predpogojem pred primerjavo govorjenega in pisnega jezika (Zwitter Vitez 2016). Precej raziskav je tudi na temo socialnih zvrsti v govorjenem jeziku (Kalin Golob 2008, 2009; Poteko 2019, Rolih 2017, Race 2021). Omeniti je treba tudi analizo razlik med govorci in govorkami glede na vrsto diskurza in druge metapodatke (Zwitter Vitez 2019, 2016). Tipično govorjeno besedišče, ki je osrednja tematika pričujočega članka, pa do zdaj še ni bilo sistematično raziskano z vidika kanoničnosti zapisa in vključevanja v jezikovne vire. Verdonik (2017) je v govoru denimo raziskovala diskurzne označevalce, kot so *ja*, *aha*, *aja*, *mhm*, *okej*, *no*, *eee*, *eem* itd., in izpostavila, da so „/o/vira /…/ tudi različni principi transkribiranja gradiva, zlasti pri segmentiranju na izjave in pravilih zapisa“ (Verdonik 2017: 95).

Obstaja že nekaj slovarskih virov, ki opisujejo tipično govorjeno besedišče (predvsem narečno): npr. slovar podjunske narečne leksike (Benko 2013), besednjak nadiškega narečja (Špehonja 2012), slovar oblačilnega izrazja v Kanalski dolini (Kenda Jež 2007),[5] a gradivo ni bilo pridobljeno korpusno in ni strojno berljivo, kanonične oblike pa so bile pripisane od zgoraj navzdol, tj. brez vpogleda, kako govorci in govorke te besede dejansko zapisujejo. Ti vpogledi so postali mogoči šele s preselitvijo oz. razširitvijo dela nestandardne komunikacije na splet, ko je tipično

[4] Še večji problem so višji nivoji obdelave s semantičnimi tehnologijami, saj še ni podatkov o pomenu tipično govorjenih besed. Tudi obogatitev DSB s pomenskimi podatki je v načrtu v projektu MEZZANINE, a se v tem prispevku omejujemo le na dopolnitev z oblikami in izgovorjavami.

[5] Celoten nabor slovarskih virov z narečno leksiko je na voljo na povezavi: http://bos.zrc-sazu.si/c/dial/

govorjena leksika postala tudi zapisana in jo lahko najdemo npr. v tvitih, e-poštnih sporočilih in podobnih žanrih v spletni slovenščini (na to temo več v Čibej 2021, Michelizza 2015 ter Zwitter Vitez in Fišer 2018).

3 Korpusni podatki kot podlaga za analizo tipično govorjene leksike

Da bi proučili probleme, na katere naletimo pri vključevanju tipično govorjenega besedišča v digitalne jezikovne vire, smo se oprli na gradivo iz dveh korpusov. Kot vir za govorjeno slovenščino smo uporabili trenutno največji in najsodobnejši vir za govorjeno slovenščino GOS 2.0[6] (Zwitter Vitez et al. 2023), ki obsega 320 ur transkribiranih posnetkov govora (uravnoteženega glede na različne tipe govornih dogodkov, npr. radijske in TV-oddaje, predavanja, zasebni pogovori, sestanki itd.; glej Robida et al. 2023: 35), zajetih med letoma 2007 in 2022, oz. približno 2,5 milijona pojavnic. Iz korpusa GOS 2.0 smo izvozili frekvenčni seznam lem, pri čemer smo izločili vse leme, ki so že vključene v Sloleks, in tiste, ki se v korpusu pojavijo le enkrat. Na ta način smo pridobili 3.879 lem, ki smo jih pregledali, označili pa smo tiste, ki so potencialno nestandardne oz. tipično govorjene (skupaj 503 leme). Na podlagi pregleda gradiva smo identificirali težave, ki jih je treba nasloviti v smernicah za vključevanje lem v jezikovne vire (več o tem v razdelku 4).

Ker iz korpusa GOS 2.0 pridobimo le transkribiran zapis govora, tj. zapis, kot so ga po določenih smernicah zapisali transkriptorji, ta ne odseva nujno zapisa, kakršnega bi uporabili govorci sami. Ker zlasti za problematiko kanoničnih oblik (glej razdelek 4.1) potrebujemo tudi podatke o dejanskem zapisu tovrstnega besedišča, smo na enak način kot pri korpusu GOS 2.0 izvozili seznam lem tudi iz korpusa spletne slovenščine JANES 1.0 (Erjavec et al. 2018); za namene preliminarne raziskave, ki jo opisujemo v tem članku, smo se omejili le na podkorpus tvitov, ki vsebuje 151 milijonov pojavnic oz. 10 milijon tvitov, ki so jih uporabniki omrežja Twitter (približno 9.000) napisali med letoma 2013 in 2017.

[6] Gos 2.0 je najnovejša različica korpusa, ki je nastala v okviru projekta Razvoj slovenščine v digitalnem okolju z združitvijo korpusov Gos 1.1 (Zwitter Vitez et al. 2015), Gos VideoLectures 4.2 (Verdonik et al. 2021) in dela govorne baze Artur v0.1 (Verdonik et al. 2022).

4 Izzivi pri vključevanju tipično govorjene leksike v jezikovne vire

V tem razdelku na kratko predstavljamo glavne izzive, na katere smo naleteli ob pregledu frekvenčnih seznamov korpusov GOS 2.0 in JANES-Tviti 1.0. Rešitve za predstavljene dileme bodo opisane v smernicah, ki so načrtovane kot rezultat projekta MEZZANINE.

4.1 Kanonične oblike tipično govorjenega besedišča

Pri tipično govorjenem besedišču naletimo na težavo že na nivoju samega zapisa; ker se beseda pojavlja le v govoru in je v standardni slovenščini ni, načeloma tudi ne obstaja dogovorna standardna oblika (zlasti v primerih, ko beseda še ni izpričana v nobenem od obstoječih jezikovnih priročnikov). V takih primerih potrebujemo kanonične oblike, tj. oblike, ki dogovorno predstavljajo tipično govorjeno besedo v pisnem jeziku. Kanonične oblike so potrebne, da je v slovarske baze lahko enotno in sistematično vneseno besedišče govorjene slovenščine (in ne pride npr. do primerov, ko se *tavžent* in *tavžnt* obravnavata kot popolnoma ločeni iztočnici). Jezikovna raba v korpusu JANES-Tviti 1.0 kaže, da enoznačnega odgovora na vprašanje kanoničnih oblik ni: že če npr. opazujemo zapisovanje sklopa *au/av* in primerjamo besedi *lavfati/laufati* in *gravžati/graužati*, vidimo, da se uporabniki_ce odločajo za zapise, ki niso vedno medsebojno konsistentni: zapis *laufati* je neprimerljivo pogostejši kot *lavfati*, po drugi strani pa je zapis *graužati* mnogo redkejši od *gravžati* (Tabela 1).

.

Tabela 1: Absolutne frekvence lem *gravžati/graužati* in *lavfati/laufati* v korpusu JANES-Tviti 1.0

Lema (1)	Frekvenca (1)	Lema (2)	Frekvenca (2)
lavfati	182	laufati	4.330
gravžati	184	graužati	27

Vir: lasten

Nadaljnja sistematična analiza načinov zapisovanja v korpusih bo pokazala, kako lahko po eni strani zagotovimo čim bolj sistematično zapisovanje tipično govorjenega besedišča v strojno berljivih jezikovnih virih za slovenščino, po drugi strani pa poskrbimo, da je tudi iskanje po jezikovnih virih za uporabnike_ce intuitivno in omogočeno tudi s pomočjo nekanoničnih oblik – načelo, ki mu je sledil tudi korpus GOS z dvema nivojema zapisa – s pogovornim in standardiziranim (glej

Verdonik in Zwitter Vitez 2020: 57–58). To odpira tudi vprašanje splošne obravnave povezav med iztočnicami v Sloleksu, ki v trenutni različici (3.0) še ni zadovoljivo razrešeno: ni npr. še nobene povezave med iztočnicama *asortiment* in *asortima*; *zajedalec* in *zajedavec* (obstaja pa npr. povezava med *volilec* in *volivec*). Na enak način težavo predstavljajo večbesedne enote, ki v trenutno različico Sloleksa še niso vključene; dilema pisanja skupaj in narazen pa se pojavi tudi v tipično govorjenem besedišču (npr. *kešpička* vs. *keš pička*; na obe različici naletimo v korpusu JANES-Tviti 1.0).

4.2 Sklopi

V govorjeni slovenščini pogosto naletimo na sklope, ki se v pisni standardni slovenščini zapisujejo narazen, precejšnja pa je tudi razlika med njihovo standardno in nestandardno izgovorjavo. Primeri, na katere naletimo v korpusu GOS 2.0, so npr. *nem* (ne bom), *toj* (to je), *vreji* (v redu), *nav* (ne bo), *daj* (da je), *nam* (ne bom), *kvauš* (kaj boš), *avte* (ali boste) in *navmo*/*naumo* (ne bomo). Upoštevati je treba, da se tovrstni primeri pojavljajo tudi v pisnem jeziku v nestandardni spletni slovenščini, kot jo izkazuje npr. korpus spletne slovenščine JANES 1.0.

> ampak sej san še enkrat no morda da razčistmo pa da **navmo** predolgo tukaj ne (GOS 2.0)
> Pa upam da se **navmo** spet zgrešil (JANES 1.0)

Sklopi so težavni za razpoznavalnike govora, saj so v zalednih podatkovnih bazah bodisi še povsem nepopisani (npr. *navmo*) ali pa so prekrivni z drugimi besednimi oblikami (*daj* kot glagol *dati*, *nem* kot pridevnik, *nam* kot zaimek). Problem predstavljajo tudi pri normalizaciji pisnih besedil v nestandardni spletni slovenščini (glej Čibej et al. 2016). Pri vključevanju tovrstnih elementov v Digitalno slovarsko bazo za slovenščino in Slovenski oblikoslovni leksikon Sloleks naletimo na več izzivov. Sloleks 3.0 še ne vsebuje podatkov o večbesednih enotah, zato dodajanje nestandardne izgovorjave [nem] za niz *ne bom* v trenutni strukturi še ni mogoče. V Sloleksu 3.0 so zgoraj omenjenim sklopom najbližje naslonske oblike zaimkov (npr. *zanj*, *nanjo*), ki so vključene kot ločene iztočnice (npr. iztočnica *name*, ki vsebuje pripadajoče oblike *name*, *nanj*, *nanjo*, *nanje* itd.). Pomembna razlika je ta, da so enobesedne različice v teh primerih prav tako standardne (*za njega* – *zanj*), v primeru *navmo* pa gre za nestandardno obliko. Digitalna slovarska baza je z vidika večbesednih enot trenutno omejena na kolokacije, sopomenske nize, frazeološke enote in podobno, ne pa na nepolnopomenske nize tipa *ali boš*, *da je* itd. Določiti je

torej treba učinkovit način, na katerega lahko v bazo in leksikon dodajamo elemente, ki so v nestandardnem jeziku enobesedni, v standardnem večbesedni, izgovorjave enobesedne oblike pa ni mogoče sestaviti iz izgovorjav posameznih besednih delov. Možna rešitev so ločene iztočnice (ustrezno označene kot nestandardne) z dodano individualno izgovorjavo, a morajo biti na voljo povezave na ustrezne enobesedne komponente. Omeniti je treba, da so tudi pri sklopih problem kanonične oblike, kar je razvidno tudi iz nekonsistentnosti, ki jih najdemo pri standardiziranih zapisih iz korpusa GOS 2.0: npr. *pauš/nau/nou/auš* vs. *mivš/nevte/čevš/navmo/davš*; raznolike zapise pa najdemo tudi v korpusu JANES-Tviti 1.0.

4.3 Nestandardne izgovorjave standardnih oblik

Trenutna različica Sloleksa vsebuje le standardne izgovorjave besed, pogosto pa v govorjeni in pisni nestandardni slovenščini naletimo na nestandardne izgovorjave tudi pri besedah, ki imajo neposredne ustreznice v standardni slovenščini. Še zlasti je to očitno pri nekaterih pogostih nepolnopomenskih besedah, npr. *lahko* in *toliko*. V korpusu GOS 2.0 npr. ob pregledu različnih oblik z normalizirano obliko lahko najdemo 54 različnih nestandardnih oblik, med njimi *loh*, *lah*, *lahk*, *lohk*, *lehko*, *lahka*, *lahku*, *lohko*, *lohku*, *leko* in *lejko*. Nekatere od izgovorjav se precej razlikujejo od standardne in zato lahko povzročajo težave pri razpoznavi govora, zato je treba leksikon z njimi ustrezno dopolniti.

Enako velja tudi za nekatere bolj sistematične in predvidljive nestandardne izgovorjave, npr. izpust končnega /i/ v deležniku na *-l* moškega spola množine – *(oni so) začéli* [zaˈʧeːl]. Te bomo pridobili z analizo strojno izluščenih najpogostejših razlik med normaliziranimi in dejanskimi pojavnicami v govorjeni transkripciji v korpusu GOS 2.0 – ker je korpus majhen, ne bomo le neposredno dodajali izgovorjav, ki jih najdemo v njem, temveč bomo na podlagi njegovih podatkov poskusili določiti tipične vzorce, ki jih lahko nato v leksikonu apliciramo na izgovorjave različnih besed, ki spadajo v enako kategorijo. Za primer: med najpogostejšimi razlikami med normalizirano in dejansko pojavnico je denimo redukcija nenaglašenega /i/ pri glagolih v prvi osebi ednine v sedanjiku, npr. *mislim* [ˈmiːslim] → [ˈmiːsləm], *sodim* [ˈsoːdim] → [ˈsoːdəm], *sovražim* [səvˈraːʒim] → [səvˈraːʒəm], *vozim* [ˈvoːzim] → [ˈvoːzəm]. Na enak način lahko nato tvorimo nestandardne izgovorjave tudi pri sorodnih oblikah, ki jih morda v korpusu GOS

2.0 ne najdemo, npr. *premislim* [pɾɛˈmiːsləm], *razsodim* [ɾaˈsːoːdəm], *zasovražim* [zasɔvˈɾaːʒəm], *prevozim* [pɾɛˈvoːzəm].

4.4 Nestandardne izgovorjave standardnih oblik

Sloleks v različici 3.0 ne vsebuje oblik z nestandardnim oblikoslovjem. Izjema so redke nestandardne oblike, ki so bile sporadično dodane v leksikon v začetni različici kot primeri nestandardnega gradiva (za ponazoritev načina, na katerega format leksikona podpira nestandardne podatke) in za namene Slogovnega priročnika v okviru projekta Sporazumevanje v slovenskem jeziku. Primeri so npr. nestandardne oblike iztočnic "hči" in "mati" (npr. tožilnik "hčero", "hči"; tožilnik "mati"), nestandardne pregibne oblike lastnoimenskih samostalnikov (npr. Shakespeare, Shakespearea/Shakespeareja) in izlastnoimenskih pridevnikov ("Shakespearjev" vs. "Shakespearov"). Tovrstne oblike so v leksikonu označene s kvalifikatorjem "nestandardno".

Ob dopolnjevanju leksikona s podatki o govorjeni slovenščini se odpira priložnost za nadaljnji razmislek o načinu opredeljevanja nestandardnosti v leksikonu, vzpostaviti pa je treba tudi enoten sistem, na katerega so nestandardne oblike povezane s standardnimi ustreznicami. Za primer: iztočnici *Voltaireov* in *Voltairov* sta v trenutni različici ločeni in povsem nepovezani, nestandardne in standardne oblike pa bo na enak način treba povezati tudi npr. v primerih z nestandardno podaljšavo osnove (npr. *Mihov* in *Mihatov*, *Mitjo*/*Mitja* in *Mitjata*).

Pregled frekvenčnih seznamov oblik in lem iz analiziranih korpusov razkrije več različnih pojavov nestandardnega oblikoslovja,[7] a so mnogo pogostejši v korpusu JANES 1.0 kot v GOS 2.0 (kar je do neke mere pričakovano, saj je JANES 1.0 neprimerljivo večji). Poleg že omenjenih podaljšav osnove s *-t* pri samostalnikih (npr. *Mihata*, *Mihatov*) najdemo tudi npr. nestandardne podaljšave kratkih nedoločnikov in namenilnikov na *-č* s *-t* (*rečt*, *tečt*, *oblečt*) ter alternativne pregibne paradigme pri samostalnikih na *-elj* (*nudelj* - tožilnik množine *nudlje*/*nudeljne*) in pri glagolih na *-ovati*/*-avati* (npr. *obupujem*/*obupavam*):

[7] Upoštevati je treba, da tudi potencialno standardne oblikoslovne variante niso vse vključene v Sloleks, npr. *stropi* vs. *stropovi*; ali pa so nestandardne oblike prekrivne s standardnimi, npr. *zrak* – *zraku* (rodilnik); dodajanje nestandardnih oblikoslovnih pojavov v leksikon bo torej potekalo vzporedno s standardnimi variantami, a se v prispevku z zgledi omejujemo le na nestandardne.

> se **opravičavam** ker sn šele zdaj prišel domov..... (JANES 1.0, 2013, komentar na novice, rtvslo.si)

> celo v kratkih rokavih se lahko sprehodiš po njej in **srečavamo** se na istih bregovih, ki pričujejo le vodo v bližini … (JANES 1.0, 2014, komentar na novice, rtvslo.si)

> Trenutno **dokončavam** piksno sonaxovega extreme 3 voska, nad katerim pa nisem pretirano navdušen. (JANES 1.0, 2009, forumsko sporočilo, avtomobilizem.com)

Na podoben način se pojavljajo tudi oblike na *-avlem, -avleš* itd.

> ne ne **zajebavlem** ovi dela za štiristo evrof ti boš delo za tisoč štiristo evrof (GOS 2.0, 2009, klic prijatelju)

> Se **opravičavlem**, če sem narobe dojela. (JANES 1.0, 2017, tvit, Twitter)

Pri teh se odpira tudi vprašanje, kako pripisati ustrezno (kanonično) normalizirano obliko in lemo. Da stvar še ni povsem razrešena, potrjujejo tudi nekonsistentnosti v transkripcijah govora za korpus JANES 1.0, kjer so normalizirane oblike za enake pojave pogosto različne od primera do primera, kot lahko vidimo v naštetih zgledih (normalizirana oblika je v oglatih oklepajih):

> gda gledan letnice rojstev teh slavnih lidi se vedno **spitavlen [spitaen]**, ka v pizdi delan s svojin življenjon :/ (JANES 1.0, 2014, tvit, Twitter)

> zakljucil sem opolnoci na stacjonu pri **Jakatu [Jakat]**;) (JANES 1.0, 2016, tvit, Twitter)

4.5 Nestandardni fonemi

Fonetični zapisi v mednarodni fonetični abecedi IPA ter njenem ekvivalentu SAMPA so v trenutni različici Sloleksa 3.0 prilagojeni standardni izgovorjavi – to velja tudi za nabor grafemov, ki jih uporabljamo za zapis izgovora. Glasov (fonemov in njihovih variant), ki jih v standardnem jeziku uporabljamo, lahko v Slovenski slovnici naštejemo 61, grafemov zanje pa 60. Vse te glasove (kot sta na primer favkalni in obstranski *d*) že imamo v naboru, z nekaterimi izjemami, npr. t. i. srednja samoglasnika *e* in *o*, ki jih navajata tudi Slovenski pravopis 2001 in 8.0, in višje polglasniško izgovorjeni kratki naglašeni *a* (SP 2001 in Slovenska slovnica), zapisan v paru s kratkim naglašenim a (['a/'ʌ]). Pri vključevanju nestandardnega in tipično govorjenega besedišča pa bo s tega vidika potreben dodaten premislek o morebitnem

vključevanju nestandardnih glasov oziroma grafemov zanje v nabor vseh simbolov, ki jih uporabljamo za zapis izgovora. V posnetkih govorjene slovenščine v korpusu GOS 2.0 na primer lahko zasledimo tudi zveneči mehkonebni pripornik [ɣ], kadar govorec_ka besedo groza izreče kot [ˈɣɾɔːza], ali pa izgovor z jezičkovim *r* (na primer [ˈmaːʀka]). Govorci_ke pogosto tuje besede berejo s tujimi glasovi in izgovora ne poslovenijo, na primer izgovor mesta München, ki ga ne preberejo poslovenjeno kot [ˈmiːŋhən], temveč z nemškim glasom *ü* ([ˈmyːŋhən]), zato se odpira vprašanje, katere tuje glasove bi bilo smiselno dodati v nabor fonemov.

Nestandardno gradivo imamo v načrtu vključiti v jezikovne vire predvsem z vidika jezikovnotehnoloških potreb, zato je treba določiti, kateri fonemi sploh (zadostno) vplivajo npr. na uspešnost razpoznave govora; obenem pa tudi, kateri fonemi so dovolj pogosto uporabljeni in regionalno razpršeni, da jih velja vključiti. Za ta namen imamo v načrtu analizo kakovosti prepoznave govora v različnih narečjih, in sicer s pomočjo posnetkov in ročnih transkripcij spletne strani narecja.si; posnetki in transkripcije vsebujejo tudi metapodatke o regijah oz. narečnih skupinah, na podlagi statistične analize pa bomo poskušali ugotoviti, katere nestandardne foneme je smiselno vključevati v nabor in kako natančno mora biti opisana glasovna raven, da dobimo najoptimalnejšo razpoznavo govora in s tem najboljše mogoče samodejno transkribiranje.

5 Zaključek

V prispevku smo na kratko predstavili preliminarno analizo izzivov vključevanja tipično govorjenega besedišča v digitalne jezikovne vire na podlagi pregleda frekvenčnih seznamov lem iz korpusov GOS 2.0 in JANES-Tviti 1.0. V okviru projekta MEZZANINE bomo sistematično in na podlagi realne jezikovne rabe v smernicah opisali probleme in rešitve, ki smo jih implementirali, zato da lahko navajamo podatke o govorjeni slovenščini v pisni obliki tako, da so intuitivni uporabnikom_cam in čim bolje odsevajo jezikovno rabo; so strojno berljivi ter neposredno uporabni za razvoj jezikovnih tehnologij (kot so npr. razpoznavalniki in sintetizatorji govora); hkrati pa ne privedejo do neobvladljivosti digitalnih jezikovnih virov.

Rezultati raziskave bodo poleg že omenjenih smernic tudi novi različici Sloleksa oz. Digitalne slovarske baze, ki bosta obogatena s podatki o govorjeni slovenščini. Na tej točki velja ponovno omeniti, da je v okviru projekta MEZZANINE načrtovana tudi leksikografska obdelava tipično govorjenega besedišča z vidika pomenskih podatkov, kar bo še dodaten doprinos k temu v slovenskem prostoru trenutno še slabo raziskanemu področju.

Literatura

Pravopis 8.0: Pravila novega slovenskega pravopisa za javno razpravo. Dostop 16. 6. 2022. na www.fran.si/pravopis8.

Anja BENKO, 2013: *Strokovni narečni slikovni slovar – podjunska narečna leksika*. Dostop 31. 8. 2023 na https://www.narecna-bera.si/

Jaka ČIBEJ et al., 2022: *Morphological lexicon Sloleks 3.0*, Slovenian language resource repository CLARIN.SI, ISSN 2820-4042, http://hdl.handle.net/11356/1745.

Jaka ČIBEJ, 2021: *Korpusna analiza in prepoznavanje regionalnih jezikovnih različic v spletni slovenščini (doktorska disertacija)*. Ljubljana: Filozofska fakulteta Univerze v Ljubljani.

Jaka ČIBEJ, Darja FIŠER, Tomaž ERJAVEC, 2016: Normalisation, Tokenisation and Sentence Segmentation of Slovene Tweets. *Proceedings of Normalisation and Analysis of Social Media Texts (NormSoMe) 2016, Language Resources and Evaluation Conference (LREC 2016)*. Portorož, Slovenia: 5–10.

Duša DIVJAK RACE, Nataša GLIHA KOMAC, 2022: Rastoči slovar slovenskega knjižnega jezika (eSSKJ): organizacija in prikaz jezikovnih podatkov. *Leksikologija i leksikografija I : zbornik radova sa međunarodnog naučnog skupa "Leksikografski postupak u različitim tipovima referentnih djela": Sarajevo, 27.–28. maja 2022. godine*. Ur. Senahid Halilović. Sarajevo: Akademija nauka i umjetnosti Bosne i Hercegovine. Dostop 18. 6. 2023 na https://publications.anubih.ba/bitstream/handle/123456789/747/3.%20Divjak%20Race%2c%20D.%3b%20Gliha%20Komac%2c%20N..pdf?sequence=19&isAllowed=y. 45–59.

Kaja DOBROVOLJC, Joakim NIVRE, 2016: The Universal Dependencies Treebank of Spoken Slovenian. *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, Portorož, Slovenia.

Tomaž ERJAVEC, Nikola LJUBEŠIĆ, Darja FIŠER, 2018: Korpus slovenskih spletnih uporabniških vsebin Janes. Ur. Darja Fišer. *Viri, orodja in metode za analizo spletne slovenščine*. Znanstvena založba Filozofske fakultete Univerze v Ljubljani.

Monika KALIN GOLOB, 2008: SMS-sporočila treh generacij. *Slovenščina med kulturami*. Ur. Miran Košuta. 283–294.

Monika KALIN GOLOB, 2009: Razpadajoči modeli: pogovorne zvrsti na javni prireditvi. *Slovenska narečja med sistemom in rabo. Obdobja 26*. Ur. Vera Smole. Ljubljana: Znanstvena založba Filozofske fakultete. 519–525.

Karmen KENDA-JEŽ, 2007: *Shranli smo jih v bančah: slovarski prispevek k poznavanju oblačilne kulture v Kanalski dolini*. Dostop 31. 8. 2023 na http://bos.zrc-sazu.si/c/dial/.

Iztok KOSEM, Simon KREK, Polona GANTAR, 2022: Semantic data should no longer exist in isolation: the Digital Dictionary Database of Slovenian. *19th EURALEX International Congress "Lexicography for Inclusion"*. Alexandroupolis, 2021.

Tina LENGAR VEROVNIK, 2010: *Radijski novinarski dvogovorni žanri kot okvir jezikovnih izbir novinarjev (doktorska disertacija)*. Ljubljana: Filozofska fakulteta Univerze v Ljubljani. Dostop 18. 6. 2023 na https://repozitorij.uni-lj.si/Dokument.php?id=114245&lang=slv.

Mija MICHELIZZA, 2015: *Spletna besedila in jezik na spletu: Primer blogov in Wikipedije v slovenščini*. Ljubljana: Založba ZRC.

Ina POTEKO, 2019: Socialnozvrstna analiza govora slovenskih govorcev na YouTubu. *Slovenski javni govor in jezikovno-kulturna (samo)zavest. Obdobja 38*. Ur. Hotimir Tivadar. Ljubljana: Znanstvena založba Filozofske fakultete. 237–246.

Duša RACE, 2021: *Pogovorni jezik: vrste in položaji (doktorska disertacija)*. Ljubljana: Filozofska fakulteta Univerze v Ljubljani.

Nejc ROBIDA, Kaja DOBROVOLJC, Luka TERČON, Darinka VERDONIK, 2023: *Gos 2.0 [Elektronski vir]: poročilo projekta Razvoj slovenščine v digitalnem okolju: aktivnost DS1.5*. Ljubljana: Univerza v Ljubljani, Center za jezikovne vire in tehnologije. Dostop 15. 6. 2023 na https://www.cjvt.si/rsdo/wp-content/uploads/sites/18/2023/06/RSDO_Kazalnik_Gos_v2.pdf.

Maša ROLIH, 2017: *Sleng in pogovorni jezik v spletni komunikaciji (doktorska disertacija)*. Koper: Fakulteta za humanistične študije.

Mojca SMOLEJ, 2012: *Besedilne vrste v spontanem govoru*. Ljubljana: Znanstvena založba Filozofske fakultete.

Nino ŠPEHONJA, 2012: *Besednjak nediško-taljansko*. Dostop 31. 8. 2023 na http://bos.zrc-sazu.si/c/Dial/Spehonja/index.html

Darinka VERDONIK et al., 2021: *Spoken corpus Gos VideoLectures 4.2 (transcription)*, Slovenian language resource repository CLARIN.SI, ISSN 2820-4042, http://hdl.handle.net/11356/1444.

Darinka VERDONIK et al., 2023: *ASR database ARTUR 1.0 (transcriptions)*, Slovenian language resource repository CLARIN.SI, ISSN 2820-4042, http://hdl.handle.net/11356/1772.

Darinka VERDONIK, 2017: Vpliv komunikacijskih žanrov na rabo diskurznih označevalcev. *Slovenske korpusne raziskave*. Ur. Špela Vintar. Ljubljana: Znanstvena založba Filozofske fakultete. Dostop 15. 6. 2023 na https://ebooks.uni-lj.si/ZalozbaUL/catalog/download/30/81/852?inline=1. 88–108.

Darinka VERDONIK, Ana ZWITTER VITEZ, 2020: *Slovenski govorni korpus Gos*. Ljubljana: Znanstvena založba Filozofske fakultete. Dostop 12. 6. 2023 na https://ebooks.uni-lj.si/ZalozbaUL/catalog/view/228/328/5306.

Ana ZWITTER VITEZ, 2019: Javni diskurz in profil govorcev: oblikoskladenjska in leksikalna analiza korpusa Gos. *Slovenski javni govor in jezikovno-kulturna (samo)zavest*. Obdobja 38. Ur. Hotimir Tivadar. Dostop 13. 6. 2023 na https://centerslo.si/simpozij-obdobja/zborniki/obdobja-38/. 255–267.

Ana ZWITTER VITEZ, 2016: Specifike govorjene slovenščine glede na formalnost sporazumevalnega položaja. *Toporišičeva obdobja*. Obdobja 35. Ur. Erika Križišnik in Miran Hladnik. Dostop 19. 6. 2023 na https://centerslo.si/simpozij-obdobja/zborniki/obdobja-35/. 351–359.

Ana ZWITTER VITEZ, Darja FIŠER, 2018: Govorne prvine v nestandardni spletni slovenščini. *Viri, orodja in metode za analizo spletne slovenščine*. Ur. Darja Fišer. Ljubljana: Znanstvena založba Filozofske fakultete. Dostop 19. 6. 2023 na https://ebooks.uni-lj.si/zalozbaul//catalog/download/111/203/2405-1?inline=1. 254–272.

Ana ZWITTER VITEZ et al., 2023: *Spoken corpus Gos 2.0 (transcriptions)*, Slovenian language resource repository CLARIN.SI, ISSN 2820-4042, http://hdl.handle.net/11356/1771.

Ana ZWITTER VITEZ, Jana ZEMLJARIČ MIKLAVČIČ, Simon KREK, Marko STABEJ, Tomaž ERJAVEC, 2021: *Spoken corpus Gos 1.1*, Slovenian language resource repository CLARIN.SI, ISSN 2820-4042, http://hdl.handle.net/11356/1438.