METODA POLAVTOMATSKEGA POPRAVLJANJA LEM IN OBLIKOSKLADENJSKIH OZNAK NA PRIMERU UČNEGA KORPUSA GOVORJENE SLOVENŠČINE ROG

Jaka ČIBEJ,[1] Tina MUNDA[1]

[1]Filozofska fakulteta, Univerza v Ljubljani

V prispevku predstavljamo postopek lematizacije in oblikoskladenjskega označevanja učnega korpusa govorjene slovenščine ROG, ki je vzorčen iz korpusa govorjene slovenščine GOS (različici 1.1 in 2.0). Pripisovanje lem in oblikoskladenjskih oznak je potekalo v več stopnjah in je za razliko od sorodnih označevalnih kampanj za slovenščino vsebovalo dodatno stopnjo, v kateri so bile leme in oblikoskladenjske oznake pred ročnim popravljanjem strojno navzkrižno primerjane z oblikami v Slovenskem oblikoslovnem leksikonu Sloleks. Predlagana metoda je znatno pospešila delo ter zmanjšala količino redundantnih pregledov in končnih stroškov. Njena prednost je tudi delitev označevalnih nalog na več sklopov s podobnimi problemi (npr. razlikovanje med imenovalnikom in tožilnikom), ob primerni pripravi podatkov pa v velikem deležu primerov od označevalcev ne zahteva poznavanja oblikoskladenjskih oznak po sistemu MTE-6. Poleg rezultatov označevanja v prispevku predstavljamo tudi poglavitne dileme, na katere naletimo pri označevanju govorjene slovenščine.

Ključne besede: lematizacija, oblikoskladenjsko označevanje, govorjena slovenščina, korpusi govorjene slovenščine

1 UVOD

Projekt MEZZANINE[1] (*Temeljne raziskave za razvoj govornih virov in tehnologij za slovenski jezik*, J7-4642), ki poteka med letoma 2022 in 2025 v sodelovanju več raziskovalnih institucij, se osredotoča na razvoj odprto dostopnih govornih virov za slovenščino, ki so ključnega pomena tako za jezikoslovne raziskave (fonetika, dialektologija, slovnica) kot tudi za jezikovne tehnologije in orodja. Projekt sestoji iz štirih poglavitnih delovnih sklopov, v okviru tega prispevka

[1]Projekt MEZZANINE: https://mezzanine.um.si/

pa sta najbolj relevantna sklopa 3 (*Segmentacija in označevanje govora*) in 4 (*Govorjena leksika*).

Med cilji sklopa 3 je npr. predvideno označevanje korpusov govorjene slovenščine na različnih ravneh, npr. z netekočnostmi in dialoškimi dejanji, v nadaljnjem koraku pa je načrtovana tudi primerjava, kako lahko označene netekočnosti izboljšajo strojno oblikoskladenjsko označevanje, lematizacijo in skladenjsko razčlenjevanje govorjenih besedil.

Istočasno poteka tudi sorodni projekt SPOT ((Dobrovoljc, 2024); *Na drevesnici temelječ pristop k raziskavam govorjene slovenščine*, Z6-4617; 2022–2024)[2], ki se osredotoča na opis skladenjskih značilnosti slovenskega govora, med njegovimi cilji pa je priprava kakovostno skladenjsko razčlenjenega korpusa govorjene slovenščine.

Aktivnosti v obeh omenjenih projektih so pokazale potrebo po korpusu govorjene slovenščine, ki bi bil ustrezno označen tudi na osnovnih označevalnih ravneh, kot sta oblikoskladnja in lematizacija. Nastali bogato označeni korpus bi nato lahko služil kot govorjeni ekvivalent pisnemu učnemu korpusu SUK 1.0.[3]

Vprašanja oblikoskladnje in lematizacije so v projektu MEZZANINE povezana s cilji sklopa 4, ki se poglavitno ukvarja z razreševanjem dilem pri vključevanju tipično govorjenega besedišča v leksikonske in leksikografske vire, npr. kako določiti kanonično obliko zapisa pri variantnih zapisih (*gravžati* oz. *graužati*) in kako določati leksikalne značilnosti pri tipično govorjenih iztočnicah (npr. *mus*, *fertik*, *oreng*). Gre za vprašanja, ki jih je treba razreševati tudi pri oblikoskladenjskem označevanju in lematizaciji govorjene slovenščine, zato se je označevanje korpusa na teh ravneh izkazalo za aktivnost, ki neposredno pomaga pri odgovorih na raziskovalna vprašanja delovnega sklopa 4.

Čeprav je bil predvideni obseg ročnega označevanja relativno obvladljiv (ROG vsebuje približno 100.000 pojavnic v primerjavi z 1 milijonom v korpusu SUK 1.0; več o gradivu, ki smo ga označevali v okviru tega prispevka, v razdelku 4), pa tovrstne označevalne kampanje glede na pretekle izkušnje kljub vsemu zahtevajo precej časovnega in finančnega vložka (več o tem v razdelku 2). Ob upoštevanju omejenih sredstev in kadrovskih zmogljivosti v projektu MEZZA-

[2]Projekt SPOT: https://spot.ff.uni-lj.si/

[3]Korpus SUK 1.0 je nastal v projektu *Razvoj slovenščine v digitalnem okolju*: https://rsdo.slovenscina.eu/ (Arhar Holdt in sod., 2023)

NINE smo zato zasnovali metodo, s pomočjo katere smo označevalni postopek, ki je bil preizkušen npr. pri označevanju učnega korpusa SUK 1.0, nekoliko preoblikovali z dodatno stopnjo predobdelave, v kateri še pred ročnim popravljanjem vse strojno pripisane oblikoskladenjske oznake in leme avtomatsko navzkrižno primerjamo s Slovenskim oblikoslovnim leksikonom Sloleks (Čibej in sod., 2022) in označevalne podatke razdelimo na več vsebinsko povezanih sklopov, ki obravnavajo podobne probleme (npr. razlikovanje med imenovalnikom in tožilnikom). S tem znatno pospešimo delo, olajšamo reševanje nalog, izboljšamo konsistentnost odločitev pri podobnih zagatah ter zmanjšamo količino redundantnih pregledov (pregledovanje nedvoumnih primerov) in končnih stroškov.

V prispevku predstavljamo novo metodo priprave podatkov, postopek in rezultate označevanja ter poglavitne dileme, na katere smo naleteli. V razdelku 2 povzamemo delo in izkušnje predhodnih označevalnih kampanj, v razdelku 3 predstavimo novo polavtomatsko metodo za popravljanje lem in oblikoskladenjskih oznak ter način kategoriziranja korpusnih pojavnic v označevalne scenarije. Nadaljujemo z opisom priprave podatkov in poteka označevanja (razdelek 4) ter povzamemo poglavitne rezultate (razdelek 5). V razdelku 6 opišemo najpogostejše označevalne dileme pri lematizaciji in oblikoskladenjskem označevanju govorjene slovenščine in v zaključku (razdelek 7) sklenemo raziskavo z načrti za prihodnje delo.

2 SORODNE RAZISKAVE

Najobsežnejši označevalni kampanji na nivoju oblikoskladenjskih oznak in lem v slovenskem prostoru sta bili izvedeni pri označevanju učnih množic JANES-Tag (Erjavec in sod., 2016b) in JANES-Norm (Erjavec in sod., 2016a) v okviru projekta JANES (Fišer in sod., 2020) ter učnega korpusa SUK 1.0 (Arhar Holdt in sod., 2023) oz. njegovih podkorpusov, npr. SentiCoref (Pori in sod., 2022).

Pri obeh kampanjah je bil osnovni postopek označevanja podoben: besedila so bila najprej strojno tokenizirana, stavčno segmentirana, oblikoskladenjsko označena in lematizirana (pri korpusih JANES-Tag in JANES-Norm tudi normalizirana), strojne oznake pa je nato ročno popravljala skupina označevalcev_k, za katerimi so oznake dokončno preverili še razsodniki. Pri kampanjah v pro-

jektu JANES je bila za označevanje uporabljena platforma WebAnno (Eckart de Castilho in sod., 2016), ki omogoča tudi večkratno označevanje istih besedil in sprejemanje končnih odločitev (kuriranje) v primerih, ko se označevalci_ke razhajajo. Pri označevanju podkorpusov korpusa SUK 1.0 so bile za to uporabljene Google Preglednice.

V obeh primerih je šlo za zelo obsežno in zahtevno kampanjo, ki je zahtevala veliko mero organizacije in tako časovnih kot tudi kadrovskih zmogljivosti: označevanje tokenizacije, stavčne segmentacije in normalizacije prvega dela korpusa JANES-Norm je npr. vključevalo skupno 11 označevalcev_k in trajalo 7 tednov (Čibej in sod., 2016) ter zahtevalo približno 270 ur označevalskega dela in dodatnih 45 ur dela pri razreševanju razhajanj. Označevanje lematizacije in oblikoskladenjskih oznak v korpusu JANES-Tag - prav tako z 11 označevalci_kami je potekalo od marca 2016 do oktobra 2016 (Čibej in sod., 2018). Popravljanje korpusa SUK (Arhar Holdt in sod., 2023) je s 24 označevalci_kami trajalo skupno 4 mesece.

K znatnemu časovnemu vložku je v obeh primerih prispevalo tudi uvajanje označevalcev_k, ki zlasti pri označevanju oblikoskladenjskih oznak po sistemu MULTEXT-East v6 (MTE-6)[4] s skupno 1.900 oznakami zahteva precej predpriprav in predstavlja strmo učno krivuljo za tiste, ki predhodno z oznakami še niso seznanjeni. Njihovo zanesljivost je bilo nato treba preveriti še z večkratnimi oznakami (npr. označevanje enakih besedil v skupinah po 3) in sprejemanjem končnih odločitev.

V vseh naštetih označevalnih kampanjah so bile popravljene posamezne zaporedne pojavnice v besedilu, kar je zlasti pri popravljanju oblikoskladenjskih oznak kognitivno zelo naporno, saj od označevalcev zahteva, da ob vsaki pojavnici mentalno preskakujejo med zelo raznolikimi problemi glede na besedno vrsto. Da bi to breme olajšali, so bili pri označevanju korpusa SentiCoref (Pori in sod., 2022) označevalci razdeljeni v več skupin, vsaka pa je označevala različne besedne vrste.

Končni rezultati najnovejše tovrstne označevalne kampanje v okviru projekta RSDO (Arhar Holdt in sod., 2023) so pokazali, da je učinkovitost strojne lematizacije in oblikoskladenjskega označevanja za slovenščino že dovolj visoka, da je mogoče namesto celostnih ročnih pregledov besedil uporabiti polavtomatske

[4] Oblikoskladenjske oznake Multext East v6: https://nl.ijs.si/ME/V6/msd/html/msd-sl.html

postopke, ki identificirajo najbolj problematična mesta. Poveden je npr. podatek, da je bilo v korpusu SentiCoref popravljenih le približno $1,3$ % vseh pojavnic v korpusu (kar je v skladu s pričakovano točnostjo lematizacijskega modela), od strojno pripisanih oblikoskladenjskih oznak pa jih je bilo popravljenih $2,9$ %. Glede na analize najpogostejših vrst popravkov približno 25 % popravkov izvira iz problemov ločevanja med občnoimenskostjo oz. lastnoimenskostjo samostalnikov (*Delo* vs. *delo*) in razdvoumljanja enakopisnih oblik (npr. tožilnik in imenovalnik pri neživih samostalnikih moškega spola).

V nadaljevanju zato predstavljamo novo metodo za predpripravo podatkov, ki upošteva zgoraj naštete ugotovitve in pri ročnem označevanju implementira polavtomatske rešitve.

3 METODOLOGIJA

Novi označevalni postopek se opira na Slovenski oblikoslovni leksikon Sloleks; v delu, opisanem v tem prispevku, smo uporabljali različico 3.0 (Čibej in sod., 2022) oz. približno 100.800 iztočnic in njihovih oblik, ki so bile ročno preverjene. Sloleks je največja odprto dostopna strojno berljiva zbirka slovenskih besed, v kateri so za vsako iztočnico (npr. *miza*) naštete njene pregibne oblike (*mize*, *mizi*, *mizo*, ...) in ustrezajoče oblikoskladenjske oznake po sistemu MTE-6 (npr. *Sozei*; samostalnik, občni, ženski spol, ednina, imenovalnik).

Metoda izhaja iz dveh poglavitnih predpostavk: (1) da strojno pripisanih lem in oznak pri določenih pojavnicah v korpusu ni treba pregledovati, saj imajo glede na leksikon enoumne oznake in leme; (2) da je pri nekaterih pojavnicah treba pregledati le leme ali pa samo oblikoskladenjske oznake, izbira potencialnih pripisanih vrednosti pa je glede na leksikon omejena. Metoda zato vsako pojavnico v korpusu navzkrižno primerja z oblikami v Sloleksu in preveri, (a) ali je oblika prisotna v leksikonu; (b) ali analizirani obliki v leksikonu pripada ena sama lema ali več; (c) ali je kombinaciji oblike in leme na podlagi leksikona mogoče pripisati nedvoumno oznako ali pa je možnosti več. Na podlagi ugotovljenih značilnosti algoritem pojavnici pripiše ustrezen označevalni scenarij iz nabora, ki ga prikazuje Tabela 1.

Tabela 1: Označevalni scenariji.

Scenarij	*Opis*	*Primer*
1.1.1	ena oblika, ena lema, ena oznaka	zdaj – zdaj – Rsn
1.1.2	ena oblika, ena lema, več možnih oznak	slik – slika – Sozdr|Sozmr
1.2	ena oblika, več možnih lem	lahko – lahek|lahko
1.2.1	ena oblika, razdvoumljena lema, ena oznaka	lahko – lahko – Rsn
1.2.2	ena oblika, razdvoumljena lema, več možnih oznak	lahko – lahek – Ppnzet|Ppnzeo|Ppnsei|Ppnset
2.1	oblike ni v leksikonu, lema pa je	/
2.2	oblike in leme ni v Sloleksu, potreben je ročen popravek	hozentregerji
0	neuvrščena pojavnica	npr. ločila

V scenarij 1.1.1 spadajo pojavnice, ki imajo v leksikonu le eno obliko z nedvoumno lemo in eno nedvoumno oblikoskladenjsko oznako (npr. oblika *zdaj* se v leksikonu pojavi le pod lemo zdaj in le z oznako *Rsn*). Pri scenariju 1.1.2 je kombinacija leme in oblike nedvoumna, razdvoumiti pa je treba oblikoskladenjsko oznako (npr. oblika *slik* nedvoumno spada pod lemo *slika*, a lahko izraža rodilnik dvojine ali pa rodilnik množine). Scenarij 1.2 vsebuje pojavnice, pri katerih je treba najprej razdvoumiti lemo in pozneje še oblikoskladenjsko oznako; scenarij 1.2.2 (ki je eno od nadaljevanj scenarija 1.2) zajema pojavnice, pri katerih je bila lema razdvoumljena, enako kot pri 1.1.2 pa ima kombinacija oblike in razdvoumljene leme lahko več oblikoskladenjskih oznak (npr. oblika *lahko* je lematizirana bodisi kot *lahek* bodisi kot *lahko*; kot pridevnik pa ima lahko glede na leksikon štiri različne oznake (*Ppnzet*, *Ppnzeo*, *Ppnsei*, *Ppnset*). Po scenariju 2.1 oblika manjka v leksikonu, pripisana lema pa obstaja (npr. če gre za zatipkano besedo ali pa legitimno varianto, ki še ni zabeležena v leksikonu). Scenarij 2.2 je edini, ki ga je v celoti treba popraviti ročno, saj v leksikonu še ni oblike in leme. Scenarij 0 vsebuje pojavnice, ki jih ni treba ročno označevati (npr. ločila).

Poleg naštetih označevalnih scenarijev je treba omeniti še podscenarije pri kategorijah 1.1.1 in 1.1.2. V vsaki sta namreč še dodatni podkategoriji M (ang. *mismatch*) in L (ang. *lowercase*), npr. 1.1.1.M, 1.1.1.L, 1.1.2.M itn.

Sozer | Sozmi | Sozmt	... , da preneha, da pač iz svoje	**diete**	Sozer	izloči meso.
Sozer | Sozmi | Sozmt	... veganstvu, vegani ne, kar se tiče	**prehrane**	Sozer	, se pravi, ne jejo, e ...
Sozer | Sozmi | Sozmt	... en majhni hobit, ki potuje skozi te	**dežele**	Sozmt	in nosi, e, prstan v Goro ...
Sozer | Sozmi | Sozmt	... bombardirali, ker je to glavna povezava železniške	**proge**	Sozer	Ljubljana-Trst.

Slika 1: Primer označevalnih nalog iz sklopa 1.1.2 (razločevanje sklona in števila pri samostalnikih ženskega spola).

Podkategorija L je glede na pogoje enaka krovni kategoriji, le da pri navzkrižnem primerjanju s Sloleksom upošteva obliko z malimi tiskanimi črkami: to je predvsem koristno za besede na začetku povedi ali izjave, katerih oblike zaradi zapisa z veliko začetnico ni mogoče neposredno najti v leksikonu.

Podkategorija M označuje primere, pri katerih je kombinaciji oblike in leme pripisana oblikoskladenjska oznaka, ki zanju v leksikonu ni predvidena. To se npr. zgodi v primerih, ko je označevalnik pripisal oznako, ki je ni v leksikonu - tak primer je npr. *samo*, ki je v Sloleksu 3.0 naveden le kot členek (L), pojavlja pa se tudi kot veznik (Vp). Podkategorija M je koristna tudi za vmesno preverjanje ustreznosti oznak - če npr. označevalci med fazo popravljanja leme spremenijo lemo iz prislovne (*odlično*) v pridevniško (*odličen*) in se strojno pripisana prislovna oznaka ne sklada s predvidenimi pridevniškimi v leksikonu. To bodisi opozarja na neustrezno izbrano oznako ali pa na pomanjkljivost v leksikonu.

Pojavnice je glede na označevalne scenarije mogoče smiselno razdeliti v različno zahtevne naloge, znotraj posameznega scenarija pa naloge razvrstiti po sklopih s podobnimi problemi (npr. glede na to, med katerimi oblikoskladenjskimi oznakami mora označevalec izbirati).

Glede na scenarij se cilji označevalne naloge nekoliko razlikujejo, v splošnem pa ena označevalna naloga po tej metodi zajema eno pojavnico s konkordančnim kontekstom ter potencialne vrednosti, ki jih je pojavnici mogoče pripisati. Slika 1 prikazuje primer sklopa nalog iz scenarija 1.1.2, v katerem mora označevalec določati, ali se samostalniki ženskega spola pojavljajo v rodilniku ednine (*Sozer*), imenovalniku množine (*Sozmi*) ali tožilniku množine (*Sozmt*). Navedene so vse izbire oblikoskladenjskih oznak iz leksikona, ciljna pojavnica pa ima prikazan še levi in desni kontekst. Pri označevanju so bile na voljo tudi nekateri drugi podatki - podrobneje jih predstavljamo v razdelku 4.

Zaradi omejenega obsega označevanja in majhnega števila označevalcev (več o tem v razdelku 4) je označevanje v našem primeru potekalo v okolju Microsoft Excel, v primeru obsežnejše kampanje pa bi bilo za ta namen z vidika uporabniške prijaznosti smiselno razviti vmesnike za označevalne platforme, kot sta npr. PyBossa[5] in LabelStudio.[6] To bi med drugim omogočalo tudi dodatno preverjanje kakovosti s sprotnim preverjanjem veljavnosti oblikoskladenjskih oznak in lem. To preverjanje smo v našem primeru opravili s postprocesiranjem označenih datotek.

3.1 Omejitve in prednosti

Metoda predpostavlja, da je korpus že ustrezno tokeniziran in segmentiran. Ker se pri tovrstnem načinu označevanja osredotočamo na pojavnice, popravljanje tokenizacijskih napak ni zelo uporabniško prijazno (označevalec lahko doda komentar, problem pa nato ročno razreši razsodnik), zato je fazo tokenizacije priporočljivo opraviti že pred razdelitvijo v označevalne scenarije.

V primeru večbesednih enot se lahko zgodi, da se posamezne pojavnice razvrstijo v različne scenarije (npr. *lindy hop*). Če so označevalne naloge razdeljene med različne označevalce, morajo biti na tovrstne primere dodatno pozorni.

Upoštevati je treba tudi, da se pri tej metodi nekatere napake lahko izmuznejo skozi sito: to je zlasti problem v primeru enakopisnic, ki so v leksikonu obravnavane kot nedvoumne, glede na jezikovno rabo v korpusu pa niso. Tak primer je npr. oblika *šalam*, ki je v leksikonu nedvoumna (*šalam* - *šala* - občni samostalnik ženskega spola, množina, dajalnik), v korpusu pa se je pojavila kot samostalnik moškega spola (*... narezano šalamo oz. šalam ...*). Ta pojav je predvidoma redek, z vse boljšo pokritostjo leksikona pa bo v prihodnje še redkejši.

Po drugi strani metoda omogoča, da preskočimo odvečno delo (npr. pregledovanje enoumnih oznak, ki lahko zajemajo tudi petino pojavnic), pri pojavnicah, ki jih je treba pregledati, pa omeji število odločitev (če gre npr. samo za razlikovanje med skloni). Namesto polnih oblikoskladenjskih oznak po sistemu MTE je na ta način mogoče za označevalce izpisati le razločevalne značilnosti (npr. *množina, tožilnik*), za katere ne potrebujejo dolgotrajnega uvajanja, zmanjša pa se tudi potreba po navzkrižnem preverjanju.

[5] PyBossa: https://docs.pybossa.com/
[6] Label Studio: https://labelstud.io/

Na ta način je lažje tudi posodabljanje označevalnih smernic, saj so vsi podobni označevalni problemi že zbrani v sklope, na podlagi katerih je mogoče za določeno dilemo doreči bolj sistematično rešitev.

4 PRIPRAVA PODATKOV IN OZNAČEVANJE

Podatki za učni korpus govorjene slovenščine so bili vzorčeni iz korpusa GOS, in sicer iz različic 1.1 (Zwitter Vitez in sod., 2021) (iz katere je bilo vzorčenih pribl. 40.000 pojavnic) in 2.0 (Zwitter Vitez in sod., 2023) (pribl. 50.000 pojavnic). Ker gre za ročne transkripcije govora, ki so bile ročno segmentirane na izjave in razdeljene na pojavnice, te pa imajo ročno pripisane tudi normalizirane oblike (npr. *pršu – prišel*), dodatnih popravkov tokenizacije na tej stopnji nismo pričakovali. Nekaj težav je predstavljala razlika v delitvi na segmente med različnimi deli korpusa GOS. Za razliko od gradiva v različici 1.1, ki je bila segmentirana na semantično relativno zaključene enote, je bil del iz različice 2.0, ki izhaja iz zbirke Artur (Verdonik in sod., 2023), za potrebe razvoja razpoznavalnika govora segmentiran po prozodičnih kriterijih (glede na premore). Takšni segmenti pogosto ne odsevajo koherentnih pomensko zaokroženih enot, širši kontekst izjave pa je nujen za ustrezno oblikoskladenjsko označevanje in lematizacijo. Pred označevanjem smo zato za namene popravljanja lem in oblikoskladenjskih oznak te segmente strojno preporazdelili glede na ločila, ki so bila postavljena med transkripcijo posnetkov, in na ta način poenotili reprezentacijo segmentov med tistimi deli, ki so bili vzorčeni iz različice 1.1, in tistimi iz različice 2.0. Kriteriji vzorčenja in postopek strojne resegmentacije so podrobneje opisani v prispevku (Verdonik in sod., 2024).

Ob pripravi podatkov smo upoštevali, da gre za razliko od predhodnih sorodnih označevalnih kampanj, ki so se osredotočale bodisi na standardno pisno ali pa (nestandardno) spletno slovenščino, pri tej kampanji za označevanje govorjene slovenščine, zato pretekli izsledki niso nujno prenosljivi. Zasnovano metodo za polavtomatsko popravljanje smo zato najprej preizkusili na prvem delu učnega korpusa, ki zajema približno 30.000 pojavnic, ki so bile vključene tudi v skladenjsko označeno odvisnostno drevesnico za slovenščino *Spoken Slovenian UD Treebank* oz. SST (Dobrovoljc in Nivre, 2016) in imajo leme in oblikoskladenjske oznake že ročno popravljene. Razdelitev že ročno označenih pojavnic na označevalne scenarije je bila pomembna predvsem zato, da je razkrila, koliko

razhajanj (in predvsem spregledanih napak) bi lahko pričakovali ob označevanju novega gradiva po polavtomatski metodi. Rezultati delitve so prikazani v Tabeli 2.

Tabela 2: Delitev podmnožice SST na označevalne scenarije.

Scenarij	*Pogostost*	*Odstotek*
1.1.1	8.300	29,12 %
1.1.2	11.047	38,76 %
1.2	6.234	21,87 %
2.2	537	1,88 %
1.1.1.L	11	0,04 %
1.1.1.M	11	0,04 %
1.1.2.L	66	0,23 %
1.1.2.M	104	0,36 %
0	2.192	7,69 %
Skupaj	28.502	100,00 %

Problematične so predvsem pojavnice iz kategorije 1.1.1.M, ki so glede na leksikon povsem nedvoumne, v resnici pa niso – teh namreč označevalci ne bi podrobno pregledovali. Nekoliko manj problematičen je scenarij 1.1.2.M (kjer imajo pojavnice nedvoumno lemo in več možnosti oblikoskladenjskih oznak, a prava ni vključena v leksikon). Gre npr. za primere tipa *gremo* v velelniškem naklonu, ki v leksikonu še ni predviden, ali pa medmete, kot je *o* (*"o, to pa ne bo šlo"*), katerega oblika je v leksikonu navedena le kot predlog ali pa kot samostalnik. Vseh tovrstnih problematičnih pojavnic je v podmnožici SST le za $0,4$ %, kar nakazuje, da je metoda dovolj točna, da je z njo mogoče pripraviti podatke tudi za označevanje preostalega dela učnega korpusa.

V Tabeli 3[7] je prikazana delitev pojavnic na scenarije še za preostala vzorca, ki sta bila vključena v učni korpus ROG (V1 – dodatnih 10.000 pojavnic iz različice 1.1 in V2 – 50.000 pojavnic iz različice 2.0).

V natančen ročni pregled so bile vključene vse naloge z izjemo scenarijev 0, 1.1.1 in 1.2.1 (več o tem v razdelku 6). Označevala sta skupno dva označevalca, ki sta sodelovala tudi pri označevalnih kampanjah v okviru projekta RSDO (Arhar Holdt in sod., 2023) in sta bila dobro seznanjena tako z označevalnimi smernicami kot

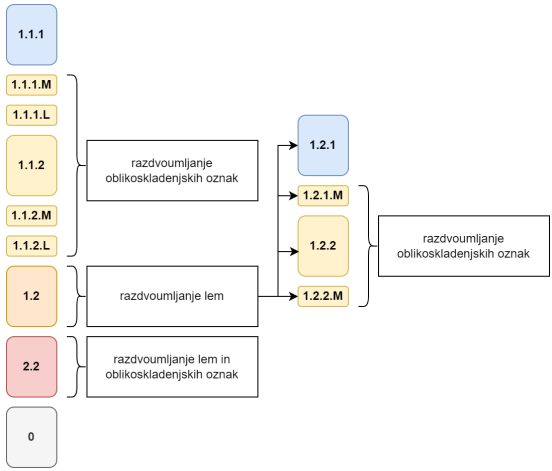
[7] Z *** so označeni nadaljevalni scenariji scenarija 1.2, v katerem najprej razdvoumimo lemo, pojavnice pa nato ponovno razdelimo na nadaljevalne scenarije.

Tabela 3: Delitev ostalih vzorcev na označevalne scenarije.

Scenarij	*Pogostost – V1*	*Delež – V1*	*Pogostost – V2*	*Delež – V2*
1.1.1	3.962	31,31 %	10.335	21,25 %
1.1.1.L	5	0,04 %	54	0,11 %
1.1.1.M	2	0,02 %	26	0,05 %
1.1.2	4.391	34,70 %	17.679	36,36 %
1.1.2.L	17	0,13 %	213	0,44 %
1.1.2.M	54	0,43 %	737	1,52 %
1.2	3.000	23,71 %	8.141	16,74 %
***1.2.1	1.543	12,19 %	3.879	7,98 %
***1.2.1.M	22	0,17 %	110	0,23 %
***1.2.2	1.369	10,82 %	4.028	8,28 %
***1.2.2.M	66	0,52 %	124	0,26 %
2.2	233	1,84 %	497	1,02 %
0	990	7,82 %	10.942	22,50 %
Celota	12.654	100,00 %	48.624	100,00 %

z oblikoskladenjskimi oznakami MTE-6. Prvi označevalec je pregledoval leme, drugi pa oblikoskladenjske oznake (v nekaterih primerih je dodatno popravil tudi leme). Slika 2 prikazuje potek označevanja. Pojavnice iz različnih scenarijev so bile vključene v različne stopnje pregleda; odvisno od scenarija je bila na koncu pregledana le oblikoskladenjska oznaka (npr. 1.1.2), lema (npr. 1.2.1) ali oboje (npr. 2.2).

Pri označevanju so bile na označevalcema v datoteki v pomoč tudi nekateri drugi podatki. Pri vsaki pojavnici, ki jo je bilo treba označiti, sta bila poleg kratkega konteksta (do 5 pojavnic levo in desno, glej Sliko 1) ločeno navedena tudi razširjeni kontekst (celoten segment iz korpusa) ter povezava na ustrezno konkordanco v korpusu Gos 2.1 v konkordančniku NoSketchEngine (Zwitter Vitez in sod., 2023). Za vsako pojavnico so bile dodane še tri povezave do posnetkov: do segmenta, v katerem je pojavnica, ter do predhodnega in naslednjega segmenta. Navedene so bile tudi vse možnosti za razdvoumljanje leme oz. oblikoskladenjske oznake, ki jih predvideva Sloleks. Ohranjen je bil tudi ID pojavnice iz korpusa, s čimer smo poskrbeli za popolno sledljivost sprememb in lažje vključevanje popravkov v končno različico korpusa.



Slika 2: Označevalni delotok za urejanje lem in oblikoskladenjskih oznak.

5 REZULTATI OZNAČEVANJA

V tem razdelku predstavljamo rezultate ročnega popravljanja lem (podrazdelek 5.1) in oblikoskladenjskih oznak (podrazdelek 5.2).

5.1 Popravki lem

Popravki lem so bili redki, saj je bilo na koncu pojavnic s spremenjeno lemo le 396 v vzorcu V2 ($0,81$ % celotnega vzorca) in 175 v vzorcu V1 ($1,38$ % celotnega vzorca).

Glede na porazdelitev popravkov po označevalnih scenarijih so bile spremembe leme najpogostejše pri scenariju 2.2 (42 % vseh popravkov lem), pri katerem niti oblike niti leme še ni v leksikonu, zaradi česar je pričakovano, da bo lematizacijski model na tovrstnih podatkih zagrešil največ napak. V vzorcu V2 je bila lema popravljena pri 164 pojavnicah (od 497 iz scenarija 2.2, torej 33 %), v vzorcu V1 pa pri 73 (od 233 iz scenarija 2.2, torej približno 31 %). V obeh vzorcih je bila torej približno tretjina neznanih pojavnic lematizirana napačno.

Določanje leme je za model problematično zlasti pri pregibnih oblikah lastnih imen (npr. *Netflix* – **Netflixu*, *Šerbi* – **Šerba*, *Lidl* – **Lidel*, *Bohinjka* – **Bohinjek*) ali pa pri pregibnih oblikah samostalnikov, pri katerih pride do podaljšave z -j

(*espe* – **espej*, *mikronivo* – **mikronivoj*. Pri neznanih pojavnicah tudi pogosto napačno presodi besedno vrsto in npr. glagol lematizira kot samostalnik (*zmučka* namesto *zmučkati*), prislov ali medmet kot glagol (*tulele* – **tuleti*, *ojojojo* – **ojojati*) ipd. S tega vidika so problematične pojavnice, ki izhajajo iz drugih jezikov in se v slovenščini pregibajo (*sitcom* – **sitec*, *solfeggio* – **solfeggiti*).

Lematizator ima težave tudi z odločanjem med občnoimenskostjo in lastnoimenskostjo (*slofit* – *Slofit*, *kliping* – *Kliping*, *covid* – *Covid*) ter z lematizacijo daljših besed, ki obsegajo 15 znakov ali več, pri katerih se zadnji del leme močno pokvari (*jezikovnotehnološki* – **jezikovokolohoški*, *knjižnojezikosloven* – **knjižnozezozoven*, *prikrojevalnica* – **prikrojalalnica*).

Med scenariji, ki so v leksikonu, so pričakovano najbolj problematične enakopisnice, tj. pojavnice iz scenarija 1.2 in njegovih podscenarijev – ti zajemajo 328 pojavnic (približno 57 % vseh popravkov lem). Med najpogostejšimi popravki so zlasti popravki med pridevniki na eni strani in prislovi na drugi, npr. *mogoč* – *mogoče*, *dober* – *dobro*, podobno tudi *ves* – *vse*, *tak* – *tako*. Z vidika govorjene slovenščine je zanimiv popravek *ti* – *te*, ki se nanaša na štajerski *te* (”*te pa si bil to samo po Sloveniji?*”), ki v leksikonu še ni zabeležen in je bil zato strojno lematiziran kot *ti* ali *ta* ter nato ročno popravljen v *te*.

V scenariju 1.1.2, pri katerem je bilo treba razdvoumljati oblikoskadenjske oznake, je bilo opravljenih le 6 popravkov lem, kar nakazuje, da je ločevanje razdvoumljanja lem in oblikoskladenjskih oznak smiselno.

5.2 Popravki oblikoskladenjskih oznak

Popravki oblikoskladenjskih oznak so bili nekoliko pogostejši kot pri lemah, a še vedno zajemajo manjšino pojavnic. V vzorcu V2 je bila oznaka spremenjena le pri 2.029 pojavnicah ($4,17$ % celotnega vzorca), v vzorcu V1 pa pri 627 pojavnicah ($4,95$ % vzorca).

Po pričakovanjih je bilo 1.782 popravkov ($67,09$ % vseh popravkov oznak) opravljenih znotraj scenarija 1.1.2 (vključno z 1.1.2.M in 1.1.2.L), pri katerem gre za razdvoumljanje slovnično enakopisnih oblik z nedvoumno lemo. Pričakovanih je tudi 578 popravkov ($21,76$ %) iz scenarija 1.2 in podscenarijev, kjer popravek leme pogosto zahteva tudi popravek oznake. Čeprav je bila v scenarij 2.2 (neleksikonske pojavnice) zajeta le manjšina popravkov (296 pojavnic oz.

$11,15$ %), pa analiza deleža popravljenih pojavnic znotraj scenarija 2.2 pokaže, da je bilo v vzorcu V2 strojno napačno označenih $37,83$ % pojavnic, v vzorcu V1 pa $46,35$ % pojavnic. Pri ostalih scenarijih je bil ta delež mnogo manjši, le okrog 7 %, kar poudarja pomen ustrezno posodobljenega leksikona za uspešno oblikoskladenjsko označevanje.

V Tabeli 4 so po pogostosti razvrščene oblikoskladenjske značilnosti strojno označenih pojavnic, pri katerih je bilo treba najpogosteje popraviti oblikoskladenjsko oznako. Po pogostosti so na prvem mestu splošni pridevniki, velja pa opazovati predvsem delež popravljenih pojavnic znotraj kategorije – v tem primeru so med najbolj problematičnimi lastnoimenski samostalniki moškega spola, pri katerih je bilo treba popraviti kar četrtino vseh pojavnic. Podobno tudi z glavnimi besednimi števniki in vprašalnimi zaimki. Zanimivo je, da so pri strojnem označevanju skoraj povsem neproblematični glagoli, pri katerih je bilo popravkov v vseh kategorijah (nedovršni, dovršni, dvovidski, pomožni) skupaj le 84, med $0,5$ in $1,3$ %.

Tabela 4: Oblikoskladenjske značilnosti najpogosteje popravljenih pojavnic (s frekvenco nad 100).

Značilnosti	*Popravljeno*	*Vse pojavnice*	*Delež*
Pp (pridevnik, splošni)	384	2.998	12,81 %
Som (samostalnik, občni, moški)	281	3.412	8,24 %
Soz (samostalnik, občni, ženski)	267	3.287	8,12 %
Rs (prislov, splošni)	261	5.103	5,11 %
Zk (zaimek, kazalni)	215	1.860	11,56 %
Zo (zaimek, osebni)	140	1.341	10,44 %
Slm (samostalnik, lastni, moški)	122	473	25,79 %
Sos (samostalnik, občni, srednji)	110	1.361	8,08 %
Kbg (števnik, besedni, glavni)	109	486	22,43 %
Vp (veznik, priredni)	106	3.265	3,25 %
Zv (zaimek, vprašalni)	103	497	20,72 %

V Tabeli 5 so prikazani najpogostejši popravki oblikoskladenjskih značilnosti (s frekvenco vsaj 50). Ti zajemajo več kot polovico vseh popravkov (53 %), skoraj tretjina (28 %) pa je zgolj razlikovanja med imenovalnikom in tožilnikom.

Tabela 5: Najpogostejši popravki oblikoskladenjskih značilnosti (s frekvenco vsaj 50).

Popravek	*Frekvenca*	*Delež*	*Primeri*
imenovalnik, tožilnik	561	21,12 %	Somei → Sometn (*stol*), Kbg-mi → Kbg-mt (*tisoč*), Zk-mei → Zk-met (*ta*)
tožilnik, imenovalnik	190	7,15 %	Sometn → Somei (*video*), Zk-set → Zk-sei (*tisto*), Kbg-mt → Kbg-mi (*devetsto*)
prislov, členek	136	5,12 %	Rsn → L (*a*)
moški, ženski	122	4,59 %	Zotmmt–k → Zotzmt–k (*jih*), Ppnmmr → Ppnzmr (*naslednjih*)
imenovalnik množine, rodilnik ednine	82	3,09 %	Sozmi → Sozer (*preiskave*), Ppnzmi → Ppnzer (*radijske*), Sosmi → Soser (*zdravila*)
splošni pridevnik, splošni prislov	80	3,01 %	Ppnsei → Rsn (*mogoče*), Ppnzet → Rsn (*primerno*)
moški, srednji	67	2,52 %	Zotmet–k → Zotset–k (*ga*), Ppnmeo → Ppnseo (*zdravim*), Kbvmei → Kbvsei (*devetnajststo*)
priredni veznik, splošni prislov	64	2,41 %	Vp → Rsn (*zato*)
občni, lastni	55	2,07 %	Somei → Slmei (*Piano*), Somem → Slmem (*Lidlu*), Sozer → Slzer (*Jute*)
vprašalni zaimek, splošni prislov	50	1,88 %	Zv-sei → Rsn (*kako*), Zv-set → Rsn (*kaj*)

6 OZNAČEVALNE DILEME

Dileme, ki so se pojavile pri ročnem pregledu korpusa ROG, so pričakovano izhajale iz razlik med govorjenim in pisnim standardnim jezikom, označevalna kampanja v tem prispevku pa omogoča prvi sistematični popis lematizacijskih in oblikoskladenjskih problemov, na katere naletimo v govorjeni slovenščini. V grobem jih lahko razdelimo na tri poglavitne skupine:

(a) Težko določljiva kanonična oblika: kot že omenjeno, v govorjeni slovenščini naletimo na tipično govorjeno besedišče, ki se v standardnem pisnem jeziku ne pojavlja, zato tudi ni opisano v obstoječih leksikografskih virih in kot tako nima enovite standardne osnovne oblike. Pri nekaterih dileme ni (npr. *mezmes*), pri drugih pa njihova izgovorjava dopušča več zapisov, npr. *oreng*, *orenk*, *orng*, *ornk*; *gravžati*, *graužati*; *hozentregar*, *hozentreger*, *hozntreger* itn. Poleg tega se v korpusu pojavljajo tudi dialektalne variante iste neuslovarjene besede, kar določanje leme še otežuje.

(b) Težko določljiva oblikoskladenjska oznaka: izkazalo se je, da v govorjenem jeziku veliko sicer standardnih besed zavzema drugačen skladenjski položaj kot v pisnem, kar postavlja pod vprašaj ohranitev oblikoskladenjske oznake, tipične za pisna besedila, npr: *a* v diskurznih markerjih 'a ne', 'a to' ("*... da so se imeli na koga obrniti, a ne.*"), kjer se odločamo med oznakama za priredni veznik in členek, in *ali* v zaključku vprašalne povedi ("*ja fajn a boste peli pri maši tudi ali?*") – tu se glede na leksikon odločamo med prirednim veznikom in prislovom. Potem so še besede, ki so tipične za govor in jih v standardnih pisnih besedilih niti ne zasledimo, npr.: nesklonljivi *ta* kot podkrepitev pridevniške besede ("*... glejte, mi eee vidimo, da prav ta pravega vira ...*"), ki ima prekrivno lemo z zaimkom 'ta', vendar ga njegova nesklonljivost razmejuje od zaimkov in bi ga lahko upravičeno uvrstili med členke; tipično štajerski nesklonljiv *te*, ki lahko nadomešča prislov 'takrat' ("*zato ker te komaj ceniš suho cesto, ko se moraš ...*") ali pa je mašilo ("*pa sva se zadnjič komaj te končno odločili, kateri film bova gledale.*"; *en* oz. *ene* v pomenu 'približno', ki sicer deloma morfološko posnema zaimkovni števnik 'en', toda zavzema zanj netipično skladenjsko vlogo: vedno določa številnik ali merni prislov ("*... Bi kar ene štiri vzel*"); samo v vlogi veznika, npr. v transkripciji "*glasbeno šolo sem naredil pet let, samo zadnjega letnika nisem, samo teorijo sem delal ...*", kjer bi vsaj eno pojavitev 'samo' lahko obravnavali kot protivni veznik – katero, ni jasno niti iz posnetka govora.

Nenazadnje sem spadajo še tuje besede, ki so se pokazale kot problematične tako za lematizacijo kot za oblikoskladenjsko označevanje že pri označevanju SUK 1.0 (Arhar Holdt in sod., 2023). Čeprav je bil ta izziv deloma razrešen, so za celovito rešitev potrebne podrobnejše analize tovrstnih besed v slovenščini. Teh besed je namreč v nadgradnjah predvsem govornih korpusov pričakovati več.

(c) Izmuzljive oblike: nekatere oblike pomotoma padejo v scenarij 1.1.1 in tako uidejo ročnemu pregledu. Do tega najpogosteje pride zaradi napak v transkripciji, npr. *uče* namesto 'uče…' (stično tropičje zaznamuje nedokončano besedo), kar v skladu s Sloleksom dobi oznako za glagol (*učiti – uče*: sedanjik, tretja oseba, množina), in *ke*, ki je lahko zatipk za 'ker' ali pa ni ustrezno normalizirano v 'ki' ali 'kaj'; lahko pa oblika pristane v tem scenariju tudi zaradi pomanjkljivosti Sloleksa, npr: *šalam*, ki se v Sloleksu pojavi le kot oblika iztočnice 'šala' v dajalniku množine, nima pa iztočnice 'šalam', kar je mišljeno npr. v "*eem, popečem na trakce narezano š… pač šalamo oziroma šalam.*"Tovrstne napake so praviloma redke, poudariti pa je treba tudi, da se s posodabljanjem leksikona vse tovrstne dileme v prihodnjih označevalnih kampanjah znajdejo v drugih scenarijih, saj npr. z dodajanjem iztočnice *šalam* v oblikoslovni leksikon oblika postane dvoumna na ravni leme, zato spada v scenarij 1.2, ne več v 1.1.1.

7 ZAKLJUČEK

V prispevku smo predstavili označevanje učnega korpusa govorjene slovenščine ROG na ravni oblikoskladenjskih oznak in lem z novo polavtomatsko metodo ter opravili prvi popis dilem. Rezultati so spodbudni zlasti ob primerjavi s pričakovanim časovnim obsegom označevanja po povsem ročni metodi, ki je bila uporabljena npr. pri označevanju korpusa SUK 1.0: glede na pretekle izkušnje namreč označevanje lem in oblikoskladenjskih oznak za vsako pojavnico vzame približno 12 sekund. V našem primeru bi pri približno 60.000 pojavnicah vzorca, skupini 6 označevalcev, 3 zbranih odgovorih na pojavnico in 10-urno tedensko kvoto kampanja trajala približno 9-10 tednov, skupaj 500 ur študentskega dela (oz. 160 ur, če bi zbirali le po en odgovor na pojavnico), pri čemer ni vštetо še delo koordinatorjev in tehnične podpore. Za označevanje korpusa ROG smo potrebovali skupaj 105 ur (25 ur za leme in 80 ur za oblikoskladenjske oznake), končni delež popravljenih pojavnic pa je primerljiv.

Metodo se lahko v prihodnje uporabi tudi za iskanje nekonsistentnosti v predhodno označenih korpusih, kot je SUK 1.0. Scenarije je mogoče spremljati tudi po posodobitvi različic leksikona, saj morebitne spremembe nakažejo potencialno nekonsistentnost v oznakah. Scenariji bi morda lahko bili uporabni tudi za natančnejšo kategorizacijo napak, ki se lahko uporabijo kot potencialne uteži za

natančnejše evalvacije lematizacijskega in oblikoskladenjskega modela (napaka v sklonu je npr. manj resna od napake v besedni vrsti).

Smiselno bi bilo razmisliti tudi o vzporednem posodabljanju leksikona in učnega korpusa in obratno, zato da preverimo, ali se pojavnice v korpusu še vedno ujemajo s stanjem v leksikonu. Posodabljanje leksikona se je izkazalo za pomembno nalogo za nadaljnje označevanje, zlasti pri kanoničnih oblikah tipično govorjenega besedišča, ki se v pisni (standardni) obliki ne pojavlja (*šravf*, *šrauf*; *orng*, *ornk*, *oreng*). To bi pripomoglo tudi k večji konsistentnosti transkripcij govora. Dileme, ki so bile identificirane med označevanjem korpusa ROG, bodo natančneje opisane v smernicah za vključevanje tipično govorjenega besedišča v digitalne jezikovne vire, kar je prav tako eden od ciljev projekta MEZZANINE (uvodni načrti za smernice so bili že predstavljeni v prispevku (Čibej in sod., 2024)).

V prihodnje bi veljalo opraviti tudi natančnejši popis in raziskavo najbolj problematičnih pojavnic za avtomatsko označevanje (tudi na podlagi učnega korpusa SUK 1.0) oz. izdelati seznam, katere pojavnice so načeloma neproblematične kljub morebitni enakopisnosti v leksikonu (npr. *kaj* vs. *kaja*, starinska beseda za kajenje).

Po zaključku označevanja na različnih ravneh bo Učni korpus ROG na voljo pod odprto licenco na repozitoriju CLARIN.SI.

ZAHVALA

Prispevek je nastal v okviru raziskovalnega projekta *Temeljne raziskave za razvoj govornih virov in tehnologij za slovenski jezik* (MEZZANINE, J7-4642), raziskovalnega projekta *Na drevesnici temelječ pristop k raziskavam govorjene slovenščine* (SPOT, Z6-4617) in raziskovalnega programa *Jezikovni viri in tehnologije za slovenski jezik* (P6-0411), ki jih financira Javna agencija za znanstvenoraziskovalno in inovacijsko dejavnost Republike Slovenije (ARIS).

Za pregled lem se avtorja iskreno zahvaljujeta Matiju Škofljancu, za dodatne predloge pri zasnovi polavtomatske metode dr. Kaji Dobrovoljc. Iskrena hvala tudi anonimnim recenzentom_kam za konstruktivne pripombe.

LITERATURA

Arhar Holdt, Š., Čibej, J., Dobrovoljc, K., Erjavec, T., Gantar, P., Krek, S., ... Žitnik, S. (2023). Nadgradnja učnega korpusa ssj550k v suk 1.0. *Razvoj slovenščine v digitalnem okolju*, 119–156.

Čibej, J., Arhar Holdt, Š., Fišer, D. in Erjavec, T. (2018). Ročno označeni korpusi janes za učenje jezikovnotehnoloških orodij in jezikoslovne raziskave. V *Viri, orodja in metode za analizo spletne slovenščine* (str. 44–73). https://ebooks.uni-lj.si/ZalozbaUL/catalog/view/111/203/2416.

Čibej, J., Fišer, D. in Erjavec, T. (2016). Normalisation, tokenisation and sentence segmentation of slovene tweets. V *Normalisation and analysis of social media texts (normsome) - lrec 2016* (str. 5–10). Portorož, Slovenia. http://www.lrec-conf.org/proceedings/lrec2016/workshops/LREC2016Workshop-NormSoMe_Proceedings.pdf#page=10

Čibej, J., Gantar, K., Dobrovoljc, K., Krek, S., Holozan, P., Erjavec, T., ... Robnik-Šikonja, M. (2022). *Morphological lexicon sloleks 3.0.* http://hdl.handle.net/11356/1745 (Slovenian language resource repository CLARIN.SI)

Čibej, J., Robida, N. in Krek, S. (2024). Nadgradnja digitalne slovarske baze za slovenščino in slovenskega oblikoslovnega leksikona sloleks s podatki o govorjeni slovenščini: načrti in cilji. *Stanje in perspektive uporabe govornih virov v raziskavah govora*, 27–40.

Dobrovoljc, K. (2024). Skladenjska drevesnica govorjene slovenščine: stanje in perspektive. *Stanje in perspektive uporabe govornih virov v raziskavah govora*, 41–62.

Dobrovoljc, K. in Nivre, J. (2016, May). The Universal Dependencies treebank of spoken Slovenian. V N. Calzolari in sod. (Ur.), *Proceedings of the tenth international conference on language resources and evaluation (LREC'16)* (str. 1566–1573). Portorož, Slovenia: European Language Resources Association (ELRA). https://aclanthology.org/L16-1248

Eckart de Castilho, R., Mújdricza-Maydt, É., Yimam, S. M., Hartmann, S., Gurevych, I., Frank, A. in Biemann, C. (2016, December). A web-based tool for the integrated annotation of semantic and syntactic structures. V *Proceedings of the workshop on language technology resources and tools for digital humanities (LT4DH)* (str. 76–84). Osaka, Japan: The COLING 2016 Organizing Committee. https://www.aclweb.org/anthology/W16-4011

Erjavec, T., Fišer, D., Čibej, J. in Arhar Holdt, Š. (2016a). *CMC training corpus janes-norm 1.2.* http://hdl.handle.net/11356/1084 (Slovenian language resource repository CLARIN.SI)

Erjavec, T., Fišer, D., Čibej, J. in Arhar Holdt, Š. (2016b). *CMC training corpus janes-tag*

1.1. http://hdl.handle.net/11356/1081 (Slovenian language resource repository CLARIN.SI)

Fišer, D., Ljubešić, N. in Erjavec, T. (2020). The janes project: language resources and tools for slovene user generated content. *Language Resources Evaluation*, *54*, 223–246. https://doi.org/10.1007/s10579-018-9425-z

Pori, E., Čibej, J., Munda, T., Terčon, L. in Arhar Holdt, Š. (2022). Lematizacija in oblikoskladenjsko označevanje korpusa senticoref. V *Konferenca jezikovne tehnologije in digitalna humanistika* (str. 162–168). Ljubljana, Slovenija. https://nl.ijs.si/jtdh22/pdf/JTDH2022_Pori-et-al_Lematizacija-in-oblikoskladenjsko-oznacevanje-korpusa-SentiCoref.pdf

Verdonik, D., Bizjak, A., Sepesy Maučec, M., Gril, L., Dobrišek, S., Križaj, J., ... Dretnik, N. (2023). *ASR database ARTUR 1.0 (transcriptions).* http://hdl.handle.net/11356/1772 (Slovenian language resource repository CLARIN.SI)

Verdonik, D., Ljubešić, N., Rupnik, P., Dobrovoljc, K. in Čibej, J. (2024). Izbor in urejanje gradiv za učni korpus govorjene slovenščine rog. V *Konferenca jezikovne tehnologije in digitalna humanistika.* Ljubljana, Slovenija.

Zwitter Vitez, A., Zemljarič Miklavčič, J., Krek, S., Stabej, M. in Erjavec, T. (2021). *Spoken corpus gos 1.1.* http://hdl.handle.net/11356/1438 (Slovenian language resource repository CLARIN.SI)

Zwitter Vitez, A., Zemljarič Miklavčič, J., Krek, S., Stabej, M., Erjavec, T., Verdonik, D., ... Dobrovoljc, K. (2023). *Spoken corpus gos 2.0 (transcriptions).* http://hdl.handle.net/11356/1771 (Slovenian language resource repository CLARIN.SI)

A METHOD FOR SEMI-AUTOMATIC CORRECTIONS OF LEMMAS AND MORPHOSYNTACTIC TAGS: THE CASE OF THE ROG TRAINING CORPUS OF SPOKEN SLOVENE

In the paper, we present the process of correcting lemmatization and morphosyntactic tags in the ROG Training Corpus of Spoken Slovene, sampled from the GOS Corpus of Spoken Slovene (versions 1.1 and 2.0). Corrections of lemmas and morphosyntactic tags were conducted in several phases and, unlike similar annotation campaigns for Slovene, included an additional preprocessing phase in which lemmas and morphosyntactic tags were automatically cross-referenced with the forms included in the Sloleks Morphological Lexicon of Slovene. This new method has significantly sped up manual work as well as reduced the number of redundant checks and final costs. Its advantage is also the fact that annotation tasks are divided into batches of similar problems (e.g. discriminating between the nominative and accusative case). With adequate data preparation, this method in a significant number of examples requires no knowledge of MTE-6 morphosyntactic tags. In addition to the results of the annotation, we also present the principal dilemmas encountered when annotating spoken Slovene.

Keywords: lemmatization, morphosyntactic tagging, spoken Slovene, spoken Slovene corpora