REPUBLIKA SLOVENIJA
MINISTRSTVO ZA KULTURO

EVROPSKA UNIJA
EVROPSKI SKLAD ZA REGIONALNI RAZVOJ
NALOŽBA V VAŠO PRIHODNOST

Leksikon besednih oblik Sloleks

Poročilo projekta Razvoj slovenščine v digitalnem okolju

Aktivnost DS1.3

Avtor: Jaka Čibej[°,□]

[°] Fakulteta za računalništvo in informatiko Univerze v Ljubljani
[□] Filozofska fakulteta Univerze v Ljubljani

Ljubljana: Center za jezikovne vire in tehnologije, Univerza v Ljubljani, 2023

Operacijo Razvoj slovenščine v digitalnem okolju sofinancirata Republika Slovenija in Evropska unija iz Evropskega sklada za regionalni razvoj. Operacija se izvaja v okviru Operativnega programa za izvajanje evropske kohezijske politike v obdobju 2014-2020.

Vsebina

1 Projektni okvir 1

2 Opis projektnih aktivnosti 1

3 Ročno popravljanje iztočnic iz Sloleksa 2.0 3

3.1 Postopek popravljanja naglašenih oblik 4

3.2 Kategorizacija iztočnic glede na vrsto izgovorjave 6

3.3 Priprava smernic za obravnavanje naglašenih oblik 7

3.3 Primerjava popravkov naglašenih oblik med različicama 2.0 in 3.0 7

3.4 Strojno generiranje zapisov IPA in SAMPA 9

4 Cevovod za strojno podprto širjenje leksikona 10

4.1 Generator oblik 10

4.2 Naglaševalnik 12

4.3 Grafemsko-fonemski pretvornik v fonetični zapis IPA/SAMPA 12

4.4 Pregibalnik 13

4.5 Pridobivanje kandidatov za nove leksikonske enote 14

4.6 Strojno generiranje novih leksikonskih enot 15

5 Programska oprema za ročni pregled in popravljanje leksikonskih enot 15

6 Ocena uspešnosti in prioritete za nadaljnji razvoj 18

7 Literatura 22

Operacijo Razvoj slovenščine v digitalnem okolju sofinancirata Republika Slovenija in Evropska unija iz Evropskega sklada za regionalni razvoj. Operacija se izvaja v okviru Operativnega programa za izvajanje evropske kohezijske politike v obdobju 2014-2020.

1 Projektni okvir

Poročilo oz. kazalnik *Leksikon besednih oblik Sloleks* je nastalo pod okriljem projekta Razvoj slovenščine v digitalnem okolju, ki sta ga med leti 2020 in 2023 sofinancirali Republika Slovenija in Evropska unija iz Evropskega sklada za regionalni razvoj (Operacija se je izvajala v okviru Operativnega programa za izvajanje evropske kohezijske politike v obdobju 2014–2020). Cilj projekta je bil zadovoljiti potrebe po računalniških izdelkih in storitvah s področja jezikovnih tehnologij za slovenski jezik za raziskovalne organizacije, podjetja in širšo javnost. Spletna stran projekta z dostopom do rezultatov: https://slovenscina.eu/.

Kazalnik se umešča v prvi projektni delovni sklop z naslovom *Jezikovni viri*. Namen delovnega sklopa je bil nadgraditi slovenske besedilne korpuse in leksikon besednih oblik. Prenovili smo učne množice in postopke za strojno označevanje sodobne slovenščine. Rezultat so osveženi in povečani jezikovni viri, ki so na voljo tako uporabniški skupnosti kot za strojno rabo. Z razvitimi postopki in orodji bo posodabljanje slovenskih korpusov v prihodnosti hitrejše in preprostejše.

Med področja, ki smo se jim na projektu posebej posvetili, sodi tudi nadgradnja leksikona besednih oblik Sloleks. Rezultati, ki so po projektu odprto dostopni za nadaljnjo rabo, predstavljajo osnovo za gradnjo jezikovnih priročnikov in tehnologij za slovenščino.

2 Opis projektnih aktivnosti

Pri aktivnosti DS1.3 smo se osredotočali na izboljšavo Slovenskega oblikoslovnega leksikona Sloleks, največje odprto dostopne zbirke oblikoslovnih in naglasnih podatkov za slovenščino. V različici 2.0 (Dobrovoljc et al. 2019), ki je bila ob začetku projektnih aktivnosti najnovejša, je Sloleks obsegal 100.803 iztočnice (npr. "miza [samostalnik ženskega spola]", "pisati [glagol, nedovršni]") s pripadajočimi besednimi oblikami (npr. "mize, mizi, mizo", "pišem, pišeš, pišejo"), ki so bile ročno pregledane. Različica je vsebovala tudi naglašene besedne oblike (npr. "míza, míze, mízi", "píšem, píšeš, píšejo"), a so bili naglasi pripisani avtomatsko. Iz naglašenih oblik so bili avtomatsko generirani tudi fonetični zapisi v mednarodni fonetični abecedi (IPA) in fonetični abecedi SAMPA. Sloleks je

Operacijo Razvoj slovenščine v digitalnem okolju sofinancirata Republika Slovenija in Evropska unija iz Evropskega sklada za regionalni razvoj. Operacija se izvaja v okviru Operativnega programa za izvajanje evropske kohezijske politike v obdobju 2014-2020.

največji odprto dostopni vir informacij o besedah v slovenskem jeziku, njihovi izgovorjavi in pregibanju, zato je bistvenega pomena za številna jezikovnotehnološka orodja in se v praksi že uporablja za oblikoslovne označevalnike, razpoznavalnike govora, kot učna množica za učenje avtomatskih naglaševalnikov ipd. Ob različnih scenarijih rabe leksikona so se pokazale tudi določene pomanjkljivosti, nekatere od najbolj perečih pa smo odpravili v okviru aktivnosti DS1.3.
Zlasti za razvoj govornih jezikovnih tehnologij je bilo ključno, da se avtomatsko pripisane naglase iz Sloleksa 2.0 ročno pregleda ter ponovno generira fonetične zapise z ustreznejšimi in jezikoslovno bolj podprtimi pravili, kar bi odpravilo napake pri podatkih o izgovoru besed. **Postopek popravljanja naglašenih oblik in fonetičnih zapisov** v okviru aktivnosti DS1.3 je predstavljen v razdelku 3.

Ker je bil Sloleks z vidika števila iztočnic vsebinsko nazadnje dopolnjen leta 2015 (v različici 1.2: Dobrovoljc et al. 2015), so v njem manjkale enote, ki so se v zadnjih letih pojavile v jezikovni rabi oz. so morda postale pogostejše in zaradi predhodne redkejše rabe v leksikon še niso bile vključene. Ročno dodajanje novih iztočnic v Sloleks (skupaj s podatki o oblikah, naglašenih oblikah in fonetičnih zapisih) je časovno potratno in neučinkovito, zato smo v okviru aktivnosti DS1.3 pripravili tudi **cevovod za strojno podprto širjenje leksikona**, ki več strojnih elementov za generiranje iztočnic poveže v zaporeden proces obdelave. Cevovod je opisan v razdelku 4.

Strojno generiranje novih oblik in iztočnic ni 100-odstotno zanesljivo, zato je pri širjenju leksikona nujno potrebna tudi rešitev za urejanje avtomatskega gradiva. V okviru dejavnosti DS1.3 smo zato razvili **nabor vmesnikov za učinkovito popravljanje strojno generiranih iztočnic** v okviru platforme PyBossa (lokalna inštalacija Centra za jezikovne vire in tehnologije UL: https://mnozicenje.cjvt.si/). Vmesniki so opisani v razdelku 5.

Pri izvedbi aktivnosti so sodelovali konzorcijski partnerji Filozofska fakulteta Univerze v Ljubljani (FF UL), Fakulteta za računalništvo in informatiko Univerze v Ljubljani (FRI UL), Znanstvenoraziskovalni center Slovenske akademije znanosti in umetnosti (ZRC SAZU) in Institut "Jožef Stefan" (IJS).

REZULTATI AKTIVNOSTI

- Novi leksikon Sloleks 3.0, ki vsebuje v celoti ročno pregledanih in jezikoslovno popravljenih obstoječih 100.000 enot, povečan pa je z 265.000 novimi strojno izdelanimi enotami in

Operacijo Razvoj slovenščine v digitalnem okolju sofinancirata Republika Slovenija in Evropska unija iz Evropskega sklada za regionalni razvoj. Operacija se izvaja v okviru Operativnega programa za izvajanje evropske kohezijske politike v obdobju 2014-2020.

dokumentirana na repozitoriju CLARIN.SI (http://hdl.handle.net/11356/1745) pod licenco CC BY-SA 4.0.

- Orodje za celovit jezikoslovni ročni pregled in popravljanje strojno pridobljenega leksikonskega gradiva:
    - Pregibalnik kot orodje je v celoti objavljen pod licenco Apache 2 na repozitorju Github: https://github.com/RSDO-DS3/SloInflector).
    - PyBossa, ki je uporabljena za popravljanje leksikonskih enot, je na voljo na naslovu: https://mnozicenje.cjvt.si/project/category/sloleks/.
    - Razviti vmesniki za specifične leksikonske naloge pa so prav tako objavljeni na repozitoriju Github pod licenco Apache 2.0: https://github.com/clarinsi/pybossa_task_presenters/tree/main/rsdo_sloleks.
- Specifikacije za obravnavo naglasne variantnosti v leksikonski bazi in uporabniškem vmesniku - vse smernice so na voljo na portalu Wiki Centra za jezikovne vire in tehnologije Univerze v Ljubljani: https://wiki.cjvt.si/shelves/oznacevanje-slovenskega-oblikoslovnega-leksikona-sloleks pod licenco CC BY-SA 4.0. Največ pozornosti smo namenili vprašanju naglaševanja lastnih imen ter zapisu izgovora kratic in besed tujega izvora na način, da je iz polfonetičnega zapisa mogoče strojno tvoriti fonetični zapis IPA in SAMPA. V smernicah so popisane tudi dileme, ki jih je treba razrešiti v prihodnjih različicah Sloleksa.

3 Ročno popravljanje iztočnic iz Sloleksa 2.0

Popravljanje strojno naglašenih oblik iz Sloleksa 2.0 je potekalo s pomočjo študentk in študentov jezikoslovnih smeri s Filozofske fakultete Univerze v Ljubljani. Pri tem je sodelovalo skupno 13 oseb, a niso vse sodelovale v celotni aktivnosti. Popravljanje je potekalo na lokalni inštalaciji množičenjske platforme PyBossa[1], ki jo vzdržuje Center za jezikovne vire in tehnologije Univerze v Ljubljani. Študentke in študenti so prejeli vsak svoje uporabniško ime in geslo za dostop do nalog. Platformo in vmesnike, ki smo jih razvili za popravljanje Sloleksa, podrobneje opišemo v razdelku 5.

[1] https://mnozicenje.cjvt.si

Operacijo Razvoj slovenščine v digitalnem okolju sofinancirata Republika Slovenija in Evropska unija iz Evropskega sklada za regionalni razvoj. Operacija se izvaja v okviru Operativnega programa za izvajanje evropske kohezijske politike v obdobju 2014-2020.

3.1 Postopek popravljanja naglašenih oblik

Sloleks 2.0 vsebuje 100.802 iztočnici različnih besednih vrst (npr. samostalniki, glagoli, pridevniki, prislovi, zaimki ipd.). Ob upoštevanjem načel dobre prakse pri množičenju smo iztočnice razdelili v več paketov, v katerih so bile iztočnice vsebinsko podobne – za označevalke in označevalce je namreč kognitivno mnogo manj naporno, če se pri določeni nalogi lahko osredotočijo le na en sam problem oz. eno samo kategorijo. Obenem so bile tudi v Sloleksu 2.0 nekatere iztočnice do določene mere že pregledane v nekaterih predhodnih projektih (npr. *Naglaševanje leksikona Sloleks*[2] v okviru razpisa CLARIN.SI 2018), a le za namene evalvacije strojnega naglaševanja. V okviru projekta RSDO smo tem iztočnicam zdaj pripisali še morebitne naglasne variante (npr. iztočnici *gejzír* smo dodali tudi varianto *gêjzir*) in tako poskrbeli, da so prav vse iztočnice iz različice 2.0 šle skozi vse nivoje jezikoslovne obravnave, hkrati pa nismo podvajali dela.

Tako smo npr. strojno preverili, katere iztočnice imajo glede na obstoječe jezikovne vire (npr. Slovenski pravopis) nepremičen naglas (npr. *skodélica*, *skodélice*, *skodélici*, *skodélico* ...), jih združili po besednih vrstah (npr. samo samostalniki z nepremičnim naglasom) in od označevalcev_k zahtevali le potrditev, ali je naglas res nepremičen (imeli pa so tudi možnost, da vpišejo pripombo, če je naglas premičen oz. če obstaja morda tudi varianten naglas), s čimer smo jim olajšali delo in pospešili postopek popravljanja ter prihranili čas za iztočnice, ki so potrebovale podrobnejšo obravnavo. Na podoben način smo v ločen paket vključili iztočnice, pri katerih smo strojno zaznali, da je naglas premičen (npr. *mêdved*, *medvéda* ...) – pri teh so morali označevalci in označevalke popraviti vse oblike. Razrez iztočnic po različnih nalogah je prikazan v Tabeli 1.

[2] https://www.clarin.si/info/storitve/projekti/#Naglasevanje_leksikona_Sloleks

Operacijo Razvoj slovenščine v digitalnem okolju sofinancirata Republika Slovenija in Evropska unija iz Evropskega sklada za regionalni razvoj. Operacija se izvaja v okviru Operativnega programa za izvajanje evropske kohezijske politike v obdobju 2014-2020.

Tabela 1: Razrez iztočnic po različnih nalogah.

Kategorija	**Število iztočnic**
Iztočnice z ročno potrjenim nepremičnim naglasom, obravnavane v sklopu predhodnega projekta CLARIN.SI 2018	14.237
Iztočnice z ročno potrjenim in popravljenim nepremičnim naglasom iz predhodnega projekta CLARIN.SI 2018	1.767
Evalvirane iztočnice iz predhodnega projekta CLARIN.SI 2018	580
Iztočnice z nepremičnim naglasom in z izkazanimi naglasnimi variantami	1.241
Iztočnice s premičnim naglasom in naglasnimi variantami	14
Občnoimenski samostalniki z nepremičnim naglasom in brez variant v oblikah	28.181
Pridevniki z nepremičnim naglasom in brez variant v oblikah	12.580
Prislovi z nepremičnim naglasom in brez variant v oblikah	2.238
Lastnoimenski samostalniki z nepremičnim naglasom in brez variant	2.943
Občnoimenski samostalniki z nepremičnim naglasom in z variantami v oblikah	94
Glagoli z nepremičnim naglasom in brez variant v oblikah	81
Občnoimenski samostalniki, ki niso vključeni v obstoječe jezikovne vire in nimajo variant v oblikah	2.003
Glagoli, ki niso vključeni v obstoječe jezikovne vire in nimajo variant v oblikah	140
Pridevniki, ki niso vključeni v obstoječe jezikovne vire in nimajo variant v oblikah	1.018
Prislovi, ki niso vključeni v obstoječe jezikovne vire in nimajo variant v oblikah	466
Večnaglasni pridevniki z nepremičnim naglasom	1.800
Večnaglasni občnoimenski samostalniki brez variantnih oblik	2.547
Pridevniki s premičnim naglasom	261
Prislovi s premičnim naglasom	75
Preostale nepregibne besedne vrste (členki, vezniki itd.)	250
Dodatni prislovi z nepremičnim naglasom	811
Dodatni pridevniki z nepremičnim naglasom	646
Preostali občnoimenski samostalniki	627
Samostalniki z nepremičnim, a spremenljivim naglasom	1.125
Glagoli z nepremičnim, a spremenljivim naglasom	311
Pridevniki z nepremičnim, a spremenljivim naglasom	729
Prislovi z nepremičnim, a spremenljivim naglasom	60
Preostali prislovi, ki niso vključeni v obstoječe jezikovne vire	167
Preostali glagoli	53
Zaimki	134
Besedni števniki	153

Operacijo Razvoj slovenščine v digitalnem okolju sofinancirata Republika Slovenija in Evropska unija iz Evropskega sklada za regionalni razvoj. Operacija se izvaja v okviru Operativnega programa za izvajanje evropske kohezijske politike v obdobju 2014-2020.

Preostali deležniški pridevniki	147
Preostali lastnoimenski samostalniki	6.657
Samostalniki s premičnim naglasom	604
Glagoli s premičnim naglasom	8.715
Iztočnice z velikimi črkami (kratice, simboli in rimski številniki)	1.298
Iztočnice, ki vsebujejo številke (nebesedni številniki)	1.257
Okrajšave	67
Preostali splošni pridevniki, ki niso vključeni v obstoječe jezikovne vire	300
Preostali svojilni pridevniki, ki niso vključeni v obstoječe jezikovne vire	4.377
Svojilni pridevniki, ki so vključeni v obstoječe jezikovne vire	48
SKUPAJ	100.802

3.2 Kategorizacija iztočnic glede na vrsto izgovorjave

V različici 2.0 so bile vse iztočnice obravnavane enako – vsem je bil strojno pripisan naglas, naglašene oblike pa so bile nato strojno pretvorjene v fonetičnih zapis IPA oz. SAMPA. Za nekatere iztočnice velja, da naglašene oblike v resnici ne morejo imeti, npr. simboli in formule (*hPa*, *mio*, *NaCl*), kratice (*RTV*, *IPI*) in tuja lastna imena (*Bordeaux*, *Shannon*). V različici 3.0 so vse ročno pregledane iztočnice označene z ustrezno kategorijo načina izgovorjave oz. tipa grafemsko-fonemske pretvorbe (Tabela 2). Velika večina ročno pregledanih iztočnic (94 %) sledi slovenskim pravilom grafemsko-fonemske pretvorbe, pri preostalih pa izgovorjav ni mogoče generirati strojno, temveč jih je treba dodati ročno.

Tabela 2: Ročno pregledane iztočnice glede na vrsto izgovora oz. grafemsko-fonemske pretvorbe.

Kategorija	**Opis in primer**	**Število iztočnic**
Slovene G2P	Beseda sledi slovenskim pravilom grafemsko-fonemske pretvorbe (*miza*, *računalnik*)	94.770
Slovene G2P with minor deviation	Beseda sledi slovenskim pravilom grafemsko-fonemske pretvorbe, a z manjšim odstopanjem (npr. če je grafem "s" izgovorjen kot "z" – *visažistka* –, kar je neobičajno).	104
Other G2P	Beseda ne sledi slovenskim pravilom grafemsko-fonemske pretvorbe (*Shakespeare*, *sommelier*).	3.053
Ambiguous G2P	Besedo je mogoče izgovoriti v skladu s slovenskimi pravili grafemsko-fonemske pretvorbe ali pa kot besedo, ki ne sledi slovenskim pravilom (npr. *Amanda* z izgovorjavama /amánda/ in /əmênda/).	68
Acronym	Beseda se izgovori kot kratica (npr. *FBI*, *RTV*).	845
Abbreviation	Beseda je okrajšava druge besede (npr. "ga." - "gospa").	70
Numeral	Beseda je glavni številnik (npr. "500", "XVI").	1.840
Symbol	Beseda je simbol (npr. "mlrd", "kg")	48

Operacijo Razvoj slovenščine v digitalnem okolju sofinancirata Republika Slovenija in Evropska unija iz Evropskega sklada za regionalni razvoj. Operacija se izvaja v okviru Operativnega programa za izvajanje evropske kohezijske politike v obdobju 2014-2020.

3.3 Priprava smernic za obravnavanje naglašenih oblik

Za označevalke in označevalce smo pripravili tudi serijo označevalnih smernic, s katerimi so se seznanili pred začetkom dela. Vse smernice so na voljo na portalu Wiki Centra za jezikovne vire in tehnologije Univerze v Ljubljani[3].

Največ pozornosti smo namenili vprašanju naglaševanja lastnih imen ter zapisu izgovora kratic in besed tujega izvora na način, da je iz polfonetičnega zapisa mogoče strojno tvoriti fonetični zapis IPA in SAMPA. V smernicah so popisane tudi dileme, ki jih je treba razrešiti v prihodnjih različicah Sloleksa. Ker je bil v času izvajanja projektne dejavnosti relevantni del novega slovenskega pravopisa (Pravopis 8.0; *O prevzemanju iz posameznih jezikov*[4]; *Glasoslovni oris*[5]) še v pripravi, smo pri pisanju smernic upoštevali izgovorjave, ki smo jih našli s pomočjo glasovnih posnetkov na spletu (npr. *MMC RTVSLO*, *YouTube*), nismo pa dodajali morebitnih oznak, ki bi nakazovale, katera različica izgovorjave je v skladu s pravopisnimi pravili. V prihodnjih različicah imamo v načrtu izgovorjave obravnavati tudi s pravorečnega vidika (o tem več v razdelku 6).

3.3 Primerjava popravkov naglašenih oblik med različicama 2.0 in 3.0

Tabela 3 predstavlja spremembe v naglašenih oblikah iztočnic, ki sledijo slovenskim pravilom grafemsko-fonemske pretvorbe, med različicama 2.0 in 3.0.

Tabela 3: Pregled potrjenih, popravljenih, dodanih in izbrisanih naglašenih oblik za iztočnice iz Sloleksa 2.0.

Besednovrstna kategorija	Potrjeno	Popravljeno	Dodano	Izbrisano
Samostalnik, občni, moški spol	264.940	18.580	7.930	1.566
Samostalnik, občni, srednji spol	131.209	4.940	1.636	34
Samostalnik, občni, ženski spol	336.833	9.832	4.122	2
Samostalnik, lastni, moški spol	42.681	25.481	588	18
Samostalnik, lastni, srednji spol	772	171	0	0
Samostalnik, lastni, ženski spol	37.540	23.546	456	0
Pridevnik, deležniški	228.456	13.043	3.991	4.457
Pridevnik, splošni	831.923	58.150	33.484	1.709

[3] https://wiki.cjvt.si/shelves/oznacevanje-slovenskega-oblikoslovnega-leksikona-sloleks
[4] https://www.fran.si/pravopis8/Poglavje/8/prevzemanje_jeziki
[5] https://www.fran.si/pravopis8/Podpoglavje/7-2

Operacijo Razvoj slovenščine v digitalnem okolju sofinancirata Republika Slovenija in Evropska unija iz Evropskega sklada za regionalni razvoj. Operacija se izvaja v okviru Operativnega programa za izvajanje evropske kohezijske politike v obdobju 2014-2020.

Pridevnik, svojilni	265.653	125.831	6.109	2.145
Glagol, glavni, dovršni	101.367	18.812	10.643	1.362
Glagol, glavni, nedovršni	77.895	8.055	3.199	40
Glagol, glavni, dvovidski	48.594	6.749	2.059	266
Glagol, pomožni	33	31	5	0
Prislov, deležje	554	64	1	0
Prislov, splošni	8.958	519	325	5
Številnik, arabski, glavni	0	0	0	0
Številnik, arabski, vrstilni	0	0	0	0
Številnik, besedni, drugo	1.566	939	0	0
Številnik, besedni, glavni	713	40	48	0
Številnik, besedni, vrstilni	12.540	715	880	0
Številnik, besedni, zaimkovni	94	1	0	0
Številnik, rimski, glavni	0	0	0	0
Številnik, rimski, vrstilni	0	0	0	0
Zaimek, celostni	470	57	5	0
Zaimek, kazalni	885	229	86	0
Zaimek, nikalni	399	6	58	0
Zaimek, nedoločni	1.390	852	223	0
Zaimek, osebni	192	88	3	0
Zaimek, povratni	65	8	0	0
Zaimek, svojilni	605	110	4	0
Zaimek, vprašalni	458	56	2	0
Zaimek, oziralni	223	271	2	0
Veznik, podredni	30	2	7	1
Veznik, priredni	26	2	4	0
Členek	70	4	4	0
Medmet	107	7	4	0
Okrajšava	0	0	0	0
Predlog, dajalnik	10	0	1	0
Predlog, imenovalnik	1	0	1	0
Predlog, mestnik	7	0	6	0
Predlog, orodnik	8	0	4	0
Predlog, rodilnik	54	5	21	0
Predlog, tožilnik	16	0	10	0
SKUPAJ	**2.397.337**	**317.196**	**75.921**	**11.605**

Sloleks 2.0 je vseboval skupno 2.792.003 naglašenih oblik. Glede na to je bilo v okviru projekta RSDO kot ustreznih potrjenih približno 86 % naglašenih oblik, 13 % pa je bilo treba ročno popraviti. Odstranjenih je bilo približno 0,4 % neustreznih naglašenih oblik, novo dodanih pa je bilo 2,7 % naglašenih oblik. Največji delež popravljenih oblik v primerjavi s potrjenimi oblikami najdemo pri lastnoimenskih samostalnikih (npr. *Novak*, *Aniston*, *Celje*) in svojilnih pridevnikih (npr. *Cerarjev*,

Operacijo Razvoj slovenščine v digitalnem okolju sofinancirata Republika Slovenija in Evropska unija iz Evropskega sklada za regionalni razvoj. Operacija se izvaja v okviru Operativnega programa za izvajanje evropske kohezijske politike v obdobju 2014-2020.

Matejin); pri teh je bilo treba popraviti do 38 % oblik, kar je pričakovano, saj se je v okviru Sloleksa pred projektom RSDO lastnoimenskim samostalnikom posvečalo najmanj pozornosti, obenem pa so lastnoimenski samostalniki po naglasu mnogo manj predvidljivi kot občni samostalniki.

3.4 Strojno generiranje zapisov IPA in SAMPA

Za strojno generiranje fonetičnih zapisov IPA in SAMPA je bil v različici 2.0 uporabljen zelo preprost grafemsko-fonemski pretvornik IPA/SAMPA na podlagi pravil, ki pa ni upošteval npr. premen po zvenečnosti/nezvenečnosti (*glasba*, *prerokba*, *iztekati*), polglasnikov pri nenaglašenih polglasniških r-jih (*vrtína*) in pri določenih končnih besednih delih (npr. -alec: *prevajalec*), dvoustničnih u-jev, zapisanih kot L (npr. *naredil*, *sfriziral*) ipd.

Za različico 3.0 smo pripravili izboljšan pretvornik IPA/SAMPA, ki je bolje podprt z jezikoslovnimi pravili in ima tudi možnost, da pri pretvarjanju v fonetični zapis upošteva dodatne značilnosti, kot je npr. oblikoskladenjska oznaka po sistemu MULTEXT-East v6[6] in pa oblikoslovni vzorec besede. Pretvornik je del Pregibalnika, ki ga podrobneje opisujemo v razdelku 4.

Pretvornik smo ročno evalvirali na množici 362 ciljno izbranih besed, ki ponazarjajo različne probleme, ki jih prejšnji pretvornik ni razrešil, npr. premena po nezvenečnosti (*predsedováti*) in zvenečnosti (*glásba*, *Kócbek*), zlivanje grafemov v en fonem (*odzív*), daljši izgovor podvojenih soglasniških fonemov (*oddája*, *podtálen*), nezveneč izgovor zvenečih nezvočnikov na koncu besede (*enáčb*, *slúžb*), polglasniki pri zlogotvornih r-jih (*rdečína*, *zavŕžen*) ipd. S pomočjo pretvornika smo generirali fonetične zapise za vse iztočnice, ki sledijo slovenskim pravilom grafemsko-fonemske pretvorbe, in evalvirali vzorec še dodatnih 100 iztočnic ter dosegli 98-odstotno točnost. Iztočnice iz različice 2.0, ki ne sledijo slovenskim pravilom grafemsko-fonemske pretvorbe, v različici 3.0 nimajo dodanih izgovorjav, saj zahtevajo posebno obravnavo v skladu s pravopisnimi pravili (lastna imena in kratice), problematiko pa je smiselno nasloviti tudi skupaj z ročnim pregledom novo dodanih iztočnic v Sloleksu 3.0, ki večinoma zajemajo lastna imena (več o tem v razdelku 4).

[6] http://nl.ijs.si/ME/V6/msd/html/msd-sl.html

Operacijo Razvoj slovenščine v digitalnem okolju sofinancirata Republika Slovenija in Evropska unija iz Evropskega sklada za regionalni razvoj. Operacija se izvaja v okviru Operativnega programa za izvajanje evropske kohezijske politike v obdobju 2014-2020.

4 Cevovod za strojno podprto širjenje leksikona

Ena od dejavnosti znotraj sklopa A1.2 je zajemala tudi pripravo cevovoda za strojno podprto širjenje leksikona. Razvito orodje, ki smo ga poimenovali Pregibalnik, je za API-klice skupaj z dokumentacijo[7],[8] na voljo na naslovu https://storitve.cjvt.si/pregibalnik, in sestoji iz treh komponent: (1) generator oblik, (2) naglaševalnik; in (3) grafemsko-fonemski pretvornik v IPO/SAMPO. Pregibalnik vse tri komponente poveže v zaporeden proces obdelave, podrobneje pa sestavo in posamezne komponente predstavljamo v nadaljevanju.

4.1 Generator oblik

Generator oblik je orodje, ki kot vhod prejme lemo (osnovno obliko besede, npr. "miza", "drevo", "testirati") in njeno oblikoskladenjsko oznako po sistemu MULTEXT-East v6 (npr. "Sozei", "Sosei", "Ggnn"). Na podlagi slovničnih značilnosti leme izbere ustrezen model za napovedovanje oblikoslovnega vzorca besede – načrta, po katerem se beseda pregiba – po sistemu, opisanem v Arhar Holdt (2021). Vsebuje šest modelov, ki so bili naučeni na iztočnicah Sloleksa 2.0: trije modeli za samostalnike ženskega, srednjega in moškega spola, en model za glagole, en model za pridevnike in en model za prislove. Pri ostalih besednih vrstah, ki so bodisi nepregibne bodisi bolj predvidljive, generator tvori oblike glede na implementirana jezikovna pravila. Potem ko model napove oblikoslovni vzorec, iz vhodne leme generira vse oblike skupaj z njihovimi oblikoslovnimi oznakami in tako ponudi celotno oblikoslovno paradigmo:

[7] https://storitve.cjvt.si/pregibalnik/docs
[8] https://storitve.cjvt.si/pregibalnik/redoc

Operacijo Razvoj slovenščine v digitalnem okolju sofinancirata Republika Slovenija in Evropska unija iz Evropskega sklada za regionalni razvoj. Operacija se izvaja v okviru Operativnega programa za izvajanje evropske kohezijske politike v obdobju 2014-2020.

Vhod:

naftaš [Somei]

Izhod:

naftaš [Somei], naftaša [Somer], naftašu [Somed], naftaša [Sometd], naftašu [Somem], naftašem [Someo]

naftaša [Somdi], naftašev [Somdr], naftašema [Somdd], naftaša [Somdt], naftaših [Somdm], naftašema [Somdo]

naftaši [Sommi], naftašev [Sommr], naftašem [Sommd], naftaše [Sommt], naftaših [Sommm], naftaši [Sommo]

Glede na preliminarne evalvacije na odprtih besednih vrstah (Tabela 4) generator oblik dosega visoko točnost, najvišjo pri prislovih (98 %), najnižjo pa pri samostalnikih moškega spola (75 %).

Tabela 4: Rezultati evalvacije generatorja oblik.

Kategorija	Točnost	Makro-F1	Referenčna točnost
Samostalnik moškega spola	75,3 %	73,7 %	27,5 %
Samostalnik ženskega spola	96,0 %	95,3 %	58,7 %
Samostalnik srednjega spola	90,9 %	90,2 %	73,4 %
Pridevniki	91,8 %	90,2 %	31,0 %
Prislovi	98,0 %	97,7 %	89,3 %
Glagoli	85,0 %	80,7 %	44,3 %

Pri samostalnikih moškega spola je točnost nižja predvsem zato, ker generator zgolj iz leme ne more ugotoviti, ali gre za živo ali neživo stvar, zato npr. žive samostalnike klasificira kot nežive (npr. *stremuh*, *melanholik*) ali obratno (npr. *dezodorant*, *Teksas*) – razlika v rezultatu je razvidna le v tožilniku ednine. Pri pridevnikih se med drugim težje odloča, ali se pridevnik na *-en* pregiba z vzorcem *-nega*, *-nemu* ali *-enega*, *-enemu* (npr. *vedoželjen* – *vedoželjnega* in ne **vedoželjenega*).

Generator oblik je za API-klice skupaj z dokumentacijo[9],[10] na voljo tudi preko ločenega API-ja na naslovu https://orodja.cjvt.si/pregibalnik/form-generator.

[9] https://orodja.cjvt.si/pregibalnik/form-generator/docs
[10] https://orodja.cjvt.si/pregibalnik/form-generator/redoc

Operacijo Razvoj slovenščine v digitalnem okolju sofinancirata Republika Slovenija in Evropska unija iz Evropskega sklada za regionalni razvoj. Operacija se izvaja v okviru Operativnega programa za izvajanje evropske kohezijske politike v obdobju 2014-2020.

4.2 Naglaševalnik

Kot del Pregibalnika smo uporabili predhodno razvito orodje za strojno naglaševanje na podlagi nevronskih mrež (Krsnik 2017) s točnostjo do skoraj 88 %. Orodje smo v okviru projekta RSDO ustrezno prilagodili oz. preoblikovali njegov format, da ga je mogoče vključiti v Pregibalnik kot komponento delotoka oz. da je mogoče do njega dostopati preko aplikacijskega programskega vmesnika (API). Naglaševalnik kot vhod prejme besedno obliko in njeno oblikoskladenjsko oznako po sistemu MULTEXT-East v6 (npr. *nuklearizacije* [Sozer]) ter vrne naglašeno obliko in njeno oblikoskladenjsko oznako (*nuklearizácije* [Sozer]); oziroma serijo besednih oblik in oblikoskladenjskih oznak, vrne pa serijo naglašenih oblik in oblikoskladenjskih oznak.

Naglaševalnik je kot ločena komponenta za API-klice skupaj z dokumentacijo[11],[12] na voljo preko API-ja na naslovu: https://orodja.cjvt.si/pregibalnik/accentuator.

4.3 Grafemsko-fonemski pretvornik v fonetični zapis IPA/SAMPA

Razvili smo izboljšan grafemsko-fonemski pretvornik v fonetični zapis IPA/SAMPA, ki je bolje podprt z jezikoslovnimi pravili in ima tudi možnost, da pri pretvarjanju v fonetični zapis upošteva dodatne značilnosti, kot je npr. oblikoskladenjska oznaka po sistemu MULTEXT-East v6 in pa oblikoslovni vzorec besede. V njem so implementirana tudi številna fonološka pravila, ki jih predhodni grafemsko-fonemski pretvornik, ki je bil uporabljen na različici Sloleks 2.0, ni upošteval, npr. premena po zvenečnosti/nezvenečnosti (*glasba*, *prerokba*, *iztekati*), polglasniki pri nenaglašenih polglasniških r-jih (*vrtína*) in pri določenih končnih besednih delih (npr. -alec: *prevajalec*), dvoustnični u-ji, zapisanih kot L (npr. *naredil*, *sfriziral*), zlivanje grafemov v en fonem (*odzív*), daljši izgovor podvojenih soglasniških fonemov (*oddája*, *podtálen*), nezveneč izgovor zvenečih nezvočnikov na koncu besede (*enáčb*, *slúžb*) ipd.

Grafemsko-fonemski pretvornik kot vhod lahko prejme nenaglašeno ali – za fonetični zapis, ki vsebuje tudi označen naglas – naglašeno obliko, opcijsko pa tudi njeno oblikoskladenjsko oznako in kodo oblikoslovnega vzorca, ki sta uporabljeni za pretvarjanje določenih grafemov v foneme na nekaterih

[11] https://orodja.cjvt.si/pregibalnik/accentuator/docs
[12] https://orodja.cjvt.si/pregibalnik/accentuator/redoc

Operacijo Razvoj slovenščine v digitalnem okolju sofinancirata Republika Slovenija in Evropska unija iz Evropskega sklada za regionalni razvoj. Operacija se izvaja v okviru Operativnega programa za izvajanje evropske kohezijske politike v obdobju 2014-2020.

predvidljivih položajih – npr. polglasnik pri samostalnikih moškega spola na -ek, pri katerih pri pregibanju pride do izpusta grafema -e: *málček* (IPA: ˈmaːlʧək), *málčka* (IPA: ˈmaːlʧka); dvoustnični u pri deležnikih na -l, npr. *oprávil* (IPA: ɔˈpraːʋiu̯), *narézal* (IPA: naˈreːzau̯).

Kot smo omenili v razdelku 2, pretvornik pri besedah, ki sledijo slovenskim pravilom grafemsko-fonemske pretvorbe, dosega do 98-odstotno točnost. Pretvornik je za API-klice kot ločena komponenta skupaj z dokumentacijo[13,14] na voljo preko API-ja na naslovu: https://orodja.cjvt.si/pregibalnik/g2p.

4.4 Pregibalnik

Pregibalnik je storitev, ki vse tri komponente (generator oblik, naglaševalnik in grafemsko-fonemski pretvornik v zapis IPA/SAMPA) združi v en delotok, ki na vhodu prejme lemo in njeno oblikoskladenjsko oznako po sistemu MULTEXT-East v6, vrne pa vse pregibane in naglašene oblike ter ustrezajoče fonetične zapise. Pregibalnik se torej lahko uporabi kot strojni način za polnjenje podatkovne baze oblikoslovnega leksikona z novimi iztočnicami in tako prihrani znatno količino časa v primerjavi z ročnim vnašanjem podatkov v bazo.

Časovno učinkovitost Pregibalnika smo evalvirali na besedah *denuklearizacija* (samostalnik ženskega spola) in *makroizvozen* (pridevnik). Pregibalnik je za generiranje vseh podatkov potreboval 3 sekunde oz. 5 sekund, ročno delo za pisanje oblik in naglaševanje pa je trajalo 9 minut in skoraj 24 minut. Pri rezultatih, ki jih generira Pregibalnik, je sicer treba upoštevati, da je podatke treba pregledati, saj morda zaradi avtomatskega postopka vsebujejo napake, vendar je tudi pri ročnem delu izkušen jezikoslovec zagrešil več napak (od preprostih zatipkov do zamenjave oblikoskladenjskih oznak). Pregibalnik torej ponuja časovno učinkovito rešitev za nadaljnje širjenje obsega leksikona.

Orodje je za API-klice skupaj z dokumentacijo[15,16] na voljo preko API-ja na naslovu: https://orodja.cjvt.si/pregibalnik. Koda Pregibalnika in posameznih komponent pa je objavljena na repozitoriju CLARIN.SI na Githubu[17].

[13] https://orodja.cjvt.si/pregibalnik/g2p/docs
[14] https://orodja.cjvt.si/pregibalnik/g2p/redoc
[15] https://orodja.cjvt.si/pregibalnik/docs
[16] https://orodja.cjvt.si/pregibalnik/redoc
[17] https://github.com/RSDO-DS3/SloInflector

Operacijo Razvoj slovenščine v digitalnem okolju sofinancirata Republika Slovenija in Evropska unija iz Evropskega sklada za regionalni razvoj. Operacija se izvaja v okviru Operativnega programa za izvajanje evropske kohezijske politike v obdobju 2014-2020.

4.5 Pridobivanje kandidatov za nove leksikonske enote

Kandidati za nove leksikonske enote so bili pridobljeni iz besedilnega korpusa Gigafida 2.0 (Krek et al. 2019). Ročno so bili pregledani seznami najpogostejših lem občnih samostalnikov ženskega, moškega in srednjega spola, pridevnikov, glagolov in prislovov (z izjemo tistih, ki so bili že vključeni v Sloleks 2.0), ki so v korpusu Gigafida 2.0 izkazovali absolutno frekvenco vsaj 5. Ker odgovor na vprašanje, ali je iztočnico treba vključiti v jezikovni vir, ni vedno enoznačen, so vsako lemo pregledali_e po trije_i označevalci_ke in s pomočjo korpusnih primerov določili_e, ali je lema ustrezna za vključitev v leksikon. Na ta način smo pridobili 50.524 kandidatov za vključitev z različnimi stopnjami potrjenostmi, kot prikazuje Tabela 5.

Tabela 5: Kandidati za iztočnice iz Gigafide 2.0 glede na stopnjo potrjenosti.

Stopnja potrjenosti	**Število iztočnic**
3 označevalci_ke	20.390
2 označevalca_ki	17.859
1 označevalec_ka	12.275

V različico 3.0 smo vključili vse iztočnice, ki so bile kot ustrezne potrjene vsaj enkrat. Razrez iztočnic po besednih vrstah prikazuje Tabela 6. Na tej točki je treba omeniti, da so zaradi narave strojnega oblikoskladenjskega označevanja in lematizacije v korpusu včasih iztočnice glede na besednovrstno kategorijo označene napačno (npr. občni samostalnik kot lastni samostalnik ali obratno). Pri potrjevanju iztočnic so imeli_e označevalci_ke navodila, naj se osredotočajo le na občnoimensko besedišče. Večina lastnoimenskih samostalnikov v tabeli je torej v resnici občnoimenskih, njihove oblikoskladenjske lastnosti pa bo treba popraviti v prihodnjih različicah.

Operacijo Razvoj slovenščine v digitalnem okolju sofinancirata Republika Slovenija in Evropska unija iz Evropskega sklada za regionalni razvoj. Operacija se izvaja v okviru Operativnega programa za izvajanje evropske kohezijske politike v obdobju 2014-2020.

Tabela 6: Število potrjenih kandidatov iz Gigafide 2.0 glede na besednovrstno kategorijo.

Besednovrstna kategorija	**Število iztočnic**
Samostalnik, občni, moški spol	7.619
Samostalnik, občni, ženski spol	4.975
Samostalnik, občni, srednji spol	1.896
Samostalnik, lastni, moški spol	3.121
Samostalnik, lastni, srednji spol	3
Samostalnik, lastni, ženski spol	616
Pridevnik, splošni	15.715
Pridevnik, svojilni	1.751
Pridevnik, deležniški	600
Prislov, splošni	7.044
Prislov, deležje	102
Glagol, glavni, dovršni	4.496
Glagol, glavni, nedovršni	2.188
Glagol, glavni, dvovidski	398

Med kandidate smo dodali tudi približno 210.000 lastnoimenskih samostalnikov iz Gigafide 2.0 – vse, ki so se v korpusu pojavili z absolutno frekvenco vsaj 10, pri čemer smo odstranili očitni šum (npr. lastna imena z malo začetnico in lastna imena, ki so vsebovala nenavadne simbole, kot so emodžiji). Ta nabor je bil preobsežen za obseg ročnega dela, ki je bil predviden v projektni dejavnosti, zato za različico 3.0 ni bil pregledan.

4.6 Strojno generiranje novih leksikonskih enot

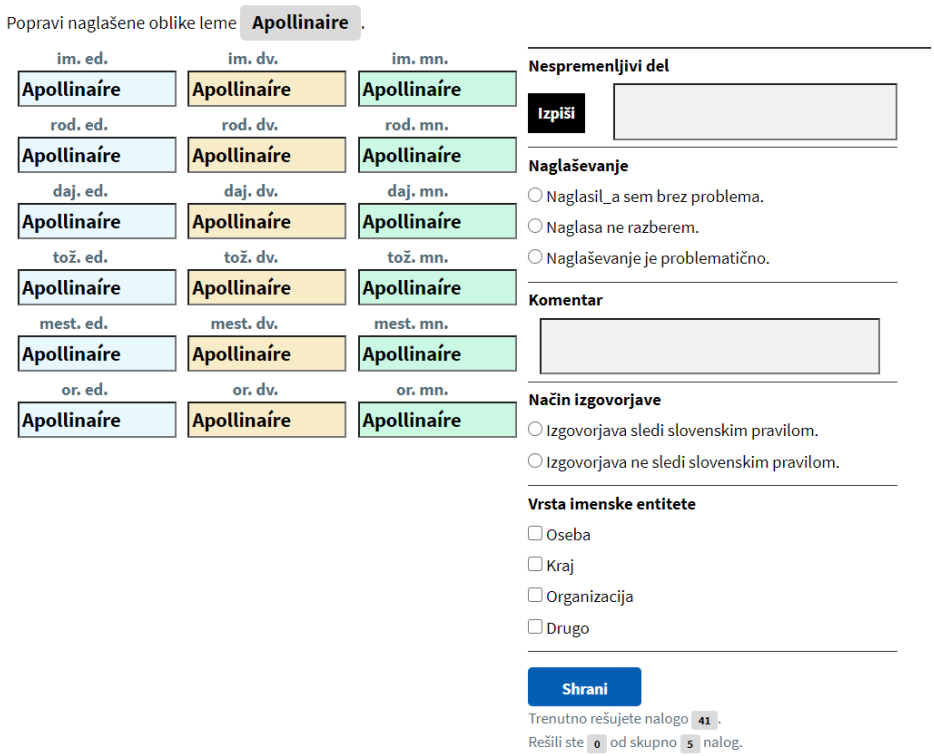
Zbranim kandidatom iz Gigafide 2.0 smo s pomočjo Pregibalnika strojno pripisali oblikoslovne paradigme, zgenerirali naglašene oblike in fonetične zapise IPA. V sklopu projekta smo na tak način pripravili nabor 265.000 strojno generiranih leksikonskih enot.

5 Programska oprema za ročni pregled in popravljanje leksikonskih enot

V projektni aktivnosti smo zagotovili tudi orodje, v katerega je mogoče uvoziti avtomatsko pripravljene podatke. Uporabniško prijazni vmesnik omogoča njihov hiter ročni jezikoslovni pregled in potrjevanje oz. popravljanje. V ta namen smo uporabili množičenjsko platformo PyBossa, odprtokodno

Operacijo Razvoj slovenščine v digitalnem okolju sofinancirata Republika Slovenija in Evropska unija iz Evropskega sklada za regionalni razvoj. Operacija se izvaja v okviru Operativnega programa za izvajanje evropske kohezijske politike v obdobju 2014-2020.

programsko opremo, ki omogoča lokalno namestitev in razvoj vmesnikov za reševanje nalog v okolju HTML/JavaScript. V slovenskem prostoru lokalno inštalacijo PyBosse[18] upravlja Center za jezikovne vire in tehnologije Univerze v Ljubljani in je bila že uspešno uporabljena za urejanje podatkov v številnih jezikoslovnih projektih, kot so *Slovar sopomenk sodobne slovenščine* (Arhar Holdt et al. 2018), *Kolokacijski slovar sodobne slovenščine* (Kosem et al. 2018a) in *Veliki slovensko-madžarski slovar* (Kosem et al. 2018b). V okviru projektne dejavnosti v projektu RSDO smo posebej za namen urejanja leksikonskih enot v Sloleksu izdelali več kot 30 različnih vmesnikov za reševanje nalog, ki se jih lahko uporabi glede na vrsto izhodiščnih podatkov oz. informacije, ki jih želimo z označevanjem oz. urejanjem pridobiti. Vmesniki so npr. ločeni glede na besedne vrste (samostalniki, pridevniki, glagoli, prislovi itd.), na način naglaševanja (premični ali nepremični naglas) in na dodatne značilnosti (npr. pri samostalnikih ločeno za lastna imena). Slika 1 prikazuje primer vmesnika za lastnoimenske samostalnike, ki prikazuje strojno naglašene oblike za iztočnico Apollinaire (lastnoimenski samostalnik ženskega spola).



Slika 1: Primer vmesnika za urejanje strojno generiranih leksikonskih enot v platformi PyBossa.

[18] https://mnozicenje.cjvt.si/

Operacijo Razvoj slovenščine v digitalnem okolju sofinancirata Republika Slovenija in Evropska unija iz Evropskega sklada za regionalni razvoj. Operacija se izvaja v okviru Operativnega programa za izvajanje evropske kohezijske politike v obdobju 2014-2020.

Naglašene oblike je mogoče poljubno spreminjati, naglašene znake pa je mogoče vnesti tudi s posebnimi simboli. naglase vstavimo tako, da pred črko, ki jo želimo naglasiti, vstavimo ustrezen simbol (za krativec vpišemo "\", za ostrivec "/", za strešico "="; vse brez narekovajev). Besedo "agencija" bi npr. naglasili tako, da bi pred črko "e" vpisali znak "/", torej "ag/encija". Ob shranjevanju podatkov platforma samodejno pretvori simbole v ustrezne znake (torej "ag/encija" – "agéncija"). V primeru lastnoimenskih samostalnikov ima večina bodisi nepremični in nespremenljivi naglas bodisi so nenaglašeni (če se ne izgovarjajo po slovenskih pravilih grafemsko-fonemske pretvorbe); da uporabniku_ci ni treba ročno popravljati vseh oblik, če so strojno neustrezno naglašene, lahko klikne na gumb »Izpiši« v razdelku »Nespremenljivi del«. V sosednji okvirček se nato izpiše del iztočnice, ki je skupen vsem oblikam (v primeru imena »Marko« bi imeli oblike Marko, Marka, Marku itd.; vsem oblikam pa je skupen »Mark«). Uporabnik_ca lahko v tem primeru naglasi le nespremenljivi del besede in z enim popravkom uredi celotno paradigmo. V ločenih razdelkih označi tudi, ali npr. naglasa ne more razbrati (v primeru nepoznanih priimkov z več možnimi variantami naglasa, npr. »Ramšak« - »Rámšak« ali »Ramšák«?) oz. ali besede ni mogoče naglasiti; v ločenem razdelku pa označi tudi, ali beseda sledi slovenskim pravilom grafemsko-fonemske pretvorbe ali ne. Opcijsko je v tem primeru dodan tudi razdelek za označevanje vrste imenske entitete – vmesniki so prilagodljivi in jih je mogoče spreminjati glede na potrebe.

Vmesnike smo razvijali sproti s popravljanjem 100.802 iztočnic iz Sloleksa 2.0 (več o tem v razdelku 3) in jih sproti izboljševali, odpravljali hrošče ipd. Vsi vmesniki so bili torej že evalvirani in preizkušeni, tako da jih je mogoče neposredno uporabiti tudi pri popravljanju novogeneriranih leksikonskih enot v različici 3.0.

Vsako nalogo so rešili po trije_i označevalci_ke. Njihovi odgovori so bili na platformi zabeleženi v formatu .JSON, nato pa je bilo preverjeno ujemanje med njimi. V primerih, v katerih se označevalci_ke niso ujemali v odločitvah, je končno odločitev sprejel razsojevalec. Po potrebi je mogoče način označevanja prilagoditi glede na želene nastavitve (npr. manj ali več označevalcev_k).

Vmesniki so odprto dostopni in na voljo pod licenco Apache 2.0 na repozitoriju Github CLARIN.SI[19]. Vse označevalne naloge, ki so bile izvedene v okviru dejavnosti DS1.3 na projektu RSDO, so na voljo na

[19] https://github.com/clarinsi/pybossa_task_presenters/tree/main/rsdo_sloleks

Operacijo Razvoj slovenščine v digitalnem okolju sofinancirata Republika Slovenija in Evropska unija iz Evropskega sklada za regionalni razvoj. Operacija se izvaja v okviru Operativnega programa za izvajanje evropske kohezijske politike v obdobju 2014-2020.

povezavi: https://mnozicenje.cjvt.si/project/category/sloleks/. S pooblaščenim dostopom je tu mogoče ustvarjati tudi nove naloge (navodila so na voljo v uradni dokumentaciji PyBosse[20]).

6 Ocena uspešnosti in prioritete za nadaljnji razvoj

Vsi zastavljeni cilji iz projektne prijave so bili v projektni aktivnosti DS1.3 doseženi, uporabljene metode pa so se izkazale za uspešne, a je še vedno precej potenciala za dodatne izboljšave in nadgradnje. Poglavitne možnosti nadaljnjega razvoja tako leksikona kot spremljajočih orodij (Pregibalnik) naštevamo v tem razdelku.

Dodatno širjenje leksikona

Leksikon Sloleks je bil v okviru projekta RSDO razširjen s približno 265.000 iztočnicami (od tega je približno 50.000 potrjenih kot ustreznih) iz korpusa pisne standardne slovenščine Gigafida 2.0. Pri tem je treba omeniti, da so na voljo tudi številni drugi korpusi, ki pokrivajo različne besedilne zvrsti – npr. korpus akademske slovenščine KAS (Žagar et al. 2022), korpus spletne slovenščine JANES (Ljubešić et al. 2017), korpus govorjene slovenščine GOS (Zwitter et al. 2021), korpus parlamentarnega govora ParlaMint (Erjavec et al. 2021), spremljevalni korpus slovenščine Trendi (Kosem et al. 2022) – in številni drugi enciklopedični viri (Wikipedija), statistični viri (npr. podatki o lastnih imenih s Statističnega urada Republike Slovenije) ter viri[21], za katere so bile pravice za uporabo odkupljene v okviru projekta RSDO (npr. Veliki splošni leksikon). Leksikon se npr. uporablja pri razvoju modelov za razpoznavo govora, zato je nujno, da je v njem vključeno raznoliko besedišče in ga je smiselno razširiti z dodatnimi viri; posebej je treba poudariti tudi spremljevalni korpus Trendi, ki vsebuje najnovejša besedila in kot tak pokriva tudi besedišče, ki šele prihaja v rabo.

Posebej je treba omeniti tudi, da trenutna različica Sloleksa ne vsebuje podatkov o npr. nestandardnih oblikah in izgovorjavah, ki so tipične zlasti za govorjeno slovenščino (npr. »naredu« namesto »naredil«), za razpoznavalnike govora pa bi bilo nujno, da se sistematično obravnava tudi ta vidik jezika.

[20] https://docs.pybossa.com/build/tutorial/
[21] https://www.cjvt.si/viri-in-orodja/slovarji-in-leksikoni/rsdo-odkup/

Operacijo Razvoj slovenščine v digitalnem okolju sofinancirata Republika Slovenija in Evropska unija iz Evropskega sklada za regionalni razvoj. Operacija se izvaja v okviru Operativnega programa za izvajanje evropske kohezijske politike v obdobju 2014-2020.

Dodatni popravki obstoječih enot in urejanje strojno generiranih enot

V okviru projekta RSDO so bile dodane številne avtomatsko generirane iztočnice, ki pa jih je treba urediti in pregledati. Določene dodatne popravke je treba vnesti tudi pri ostalih iztočnicah, saj je treba sistematično preveriti konsistentnost nekaterih pojavov v celotnem Sloleksu. Tak primer so npr. pridevniki, pri katerih je treba v korpusu preveriti, ali se v rabi stopnjujejo ali ne; ali pa samostalniki, ki se pojavljajo zgolj v množini (npr. *katakombe*). Razločiti je treba tudi nekatere enakopisnice, ki so bile strojno združene v eno samo iztočnico (npr. »leden« - »ledni« ali »ledeni«). S pomočjo korpusov je treba pregledati je tudi, ali se v jezikovni rabi pojavljajo določene variante pregibnih oblik (npr. stropi - stropovi), in iztočnice ustrezno dopolniti z njimi, obenem pa je treba izpopolniti trenutni sistem normativnih kvalifikatorjev v Sloleksu, ki je omejen na oznaki »nestandardno« in »variantno«.

Zlogovanje

V oblikoslovni leksikon je treba dodati zlogovane oblike (npr. mi-za, o-pe-ra-ci-ja, u-zur-pi-ra-ti), ki se lahko uporabijo za razvoj delilnikov besed in za druge jezikovnotehnološke aplikacije, kot so npr. sistemi zaznavanja napačnega zapisa skupaj ali narazen ("na glas" in "naglas"). Za dodajanje zlogovanih oblik je treba razviti in evalvirati zlogovalnik, ki bo dovolj natančen tudi za tvorjenje zlogovanih oblik pri novo dodanih besedah. Zlogovalnik je nato treba implementirati v cevovod Pregibalnika.

Povezane iztočnice in besedne družine

V Sloleksu 2.0 in 3.0 imajo le nekatere iztočnice navedene tudi povezane iztočnice – načeloma gre za besedotvorno povezane besede (npr. pisati – pisatelj, maček – mačka), a so trenutno navedene povsem nesistematično, med različnimi vrstami povezav pa ni nobenih razlik; npr. na enak način sta obravnavana pisatelj – pisateljica kot pisati – pisatelj (več o tem v Čibej 2021). Razviti je treba sistem, ki bo omogočal učinkovito strojno povezovanje besedotvorno povezanih iztočnic in avtomatsko tvorjenje besednih družin ter morfoloških derivacijskih drevesnic, s katerimi bo mogoče v leksikonu strojno povezovati novo dodane iztočnice. Za primer lahko vzamemo besedišče, ki je nastalo v času pandemije: *korona*, *koronski*, *protikoronski* ter *covid*, *coviden*, *postcoviden* ipd. Sistem bi tako omogočal širjenje leksikona s skupkom povezanih iztočnic, ne le s posameznimi nepovezanimi iztočnicami. S tem bi se olajšalo tudi leksikografsko delo, saj bi leksikografom omogočalo hitro obravnavo več pomensko podobnih iztočnic naenkrat.

Operacijo Razvoj slovenščine v digitalnem okolju sofinancirata Republika Slovenija in Evropska unija iz Evropskega sklada za regionalni razvoj. Operacija se izvaja v okviru Operativnega programa za izvajanje evropske kohezijske politike v obdobju 2014-2020.

Izgovorjave za tuja lastna imena in besede tujega izvora

Pri številnih tujih lastnih imenih je še vedno potrebno ročno dodajanje izgovorjav – za ustrezno razpoznavo govora je namreč nujno, da so v leksikonu navedene tako kanonične izgovorjave (oz. tiste v skladu s pravopisnimi in pravorečnimi pravili) kot tudi nestandardne izgovorjave, ki so v slovenskem prostoru zelo razširjene (npr. Erdoğan – po pravorečju in v skladu s turško izgovorjavo /êrdoan/, a v Sloveniji pogosto /êrdogan/). Tuja lastna imena so v Sloleksu 2.0 predstavljala le okrog 3 % vseh iztočnic, v različici 3.0, ki vsebuje še dodatnih 210.000 lastnih imen iz Gigafide 2.0 (v Sloleksu 2.0 je bilo lastnih imen le okrog 10.000), pa pričakujemo, da bomo naleteli na mnogo večjo količino. Poleg ročnega pregleda in kategorizacije lastnih imen glede na vrsto izgovorjave (slovenska ali neslovenska pravila grafemsko-fonemske pretvorbe) ter dodajanja samih izgovorjav je treba tudi doreči, kakšno je razmerje med pomenom, izgovorjavo in leksemom oz. kdaj tuja izgovorjava vpliva na to, da iztočnico obravnavamo kot ločen leksem (zlasti pri enakopisnicah).

Dodajanje podatkov o korpusnih frekvencah

V različicah 2.0 in 3.0 Sloleks za vse iztočnice in njihove pregibne oblike vsebuje podatke o njihovi absolutni pogostosti v korpusu Gigafida 2.0 – s temi podatki lahko npr. ugotovimo, kako pogosto se samostalnik »hči« v korpusu pojavlja v rodilniku ednine, kar je koristno za uporabnike_ce. Obenem nam to omogoča, da se v primeru, da se moramo za specifičen razpoznavalnik govora zaradi tehničnih omejitev omejiti na določen nabor najpogostejših besed, ki jih bo razpoznavalnik lahko razpoznaval, odločimo, da besede z izjemno redkimi pojavitvami izločimo. Sloleksu je zato treba dodati tudi podatke o absolutnih (in relativnih) frekvencah iz drugih korpusov; tako bo npr. pri gradnji razpoznavalnika govora specifično za akademski kontekst mogoče dati prednost besedam, ki so v akademskem diskurzu pogostejše oz. bolj pričakovane.

Popravki na nivoju naglašenih oblik

V različici 3.0 Sloleks vsebuje naglašene oblike (ročno potrjene za prvi 100.802 iztočnici in avtomatske za ostale iztočnice), iztočnice same pa nimajo pripisane naglasnih vzorcev na podoben način, kot so pripisani oblikoslovni vzorci (ki povedo, kako se beseda pregiba, npr. računalnik-0, računalnik-a, računalnik-u, računalnik-0, …). Na enak način je treba izdelati tudi sistem pripisovanja naglasnih vzorcev, s katerimi lahko kategoriziramo besede glede na to, kako se naglašujejo (npr. nepremični in nespremenljivi naglas: míza, míze, mízi, mízo …; nepremični in spremenljivi naglas: Rožlè, Rožléta,

Operacijo Razvoj slovenščine v digitalnem okolju sofinancirata Republika Slovenija in Evropska unija iz Evropskega sklada za regionalni razvoj. Operacija se izvaja v okviru Operativnega programa za izvajanje evropske kohezijske politike v obdobju 2014-2020.

Rožlétu, …; premični in spremenljivi: móž, možá, možu, možá, …). Izdelava naglasnih vzorcev v strojno berljivi obliki bo omogočila tudi izboljšave trenutnega naglaševalnika, ki je zasnovan tako, da naglašuje vsako besedno obliko posebej. Naglaševalnik na podlagi naglasnih vzorcev bo iztočnice naglaševal bolj konsistentno, prilagojen pa bo tudi dvo- in večnaglasnim besedam (npr. mákroekonómika), na katerih trenutni naglaševalnik ni bil naučen.

Kratice

Kratice (npr. RTV, FBI, SCT) so poseben primer in jih trenutna različica Pregibalnika ne podpira, saj zahtevajo posebno obravnavo; pri kraticah je namreč pregibna oblika odvisna od načina izgovorjave, ki pa ni razviden iz same leme. Kratico "OZN" npr. lahko pregibamo kot "OZN-a", če je izgovorjava /ozéêna/, ali kot "OZN-ja", če je izgovorjava /ozənəja/, s polglasnikom. Za Pregibalnik je torej treba razviti ločen modul, ki ustrezno tvori oblike in izgovorjave za kratice.

Večbesednost

V različici 3.0 Sloleks vsebuje zgolj enobesedne iztočnice, v prihodnjih različicah pa je nujno treba določiti, kako so v njem obravnavane tudi večbesedne enote, npr. večbesedne imenske entitete (Slovenj Gradec, New York) in druge večbesedne enote (streljati kozle, iti rakom žvižgat) oz. enote, ki jih glede na tokenizacijska pravila obravnavamo kot večbesedne (npr. DMA-krmilnik). Ko bo način obravnave večbesednih enot določen, bo vse, kar izpolnjuje kriterije za ločeno iztočnico, v Sloleksu vneseno pod lasten vnos.

Nova različica Sloleks 3.0 v spletnem vmesniku

Sloleks 2.0 je na voljo tudi v spletnem vmesniku[22] Centra za jezikovne vire in tehnologije Univerze v Ljubljani. Vmesnik je treba posodobiti, da bo lahko prikazoval podatke, ki so bili v Sloleks dodani v različici 3.0, npr. oblikoslovne vzorce, način izgovorjave iztočnice, status pregledanosti ipd. Po posodobitvi je treba tudi različico 3.0 omogočiti v spletnem vmesniku, da bo na voljo uporabnikom_cam za iskanje.

[22] https://viri.cjvt.si/sloleks/slv/

Operacijo Razvoj slovenščine v digitalnem okolju sofinancirata Republika Slovenija in Evropska unija iz Evropskega sklada za regionalni razvoj. Operacija se izvaja v okviru Operativnega programa za izvajanje evropske kohezijske politike v obdobju 2014-2020.

7 Literatura

Arhar Holdt 2021 = Arhar Holdt, Š. (2021). Oblikoslovni vzorci za strojno procesiranje slovenščine. In Š. Arhar Holdt (Ed.), Nova slovnica sodobne standardne slovenščine: viri in metode (pp. 87–124). Znanstvena založba Filozofske fakultete. https://e-knjige.ff.uni-lj.si/znanstvena-zalozba/catalog/view/325/477/7313-1.

Arhar Holdt et al. 2018 = ARHAR HOLDT, Špela, ČIBEJ, Jaka, DOBROVOLJC, Kaja, GANTAR, Apolonija, GORJANC, Vojko, KLEMENC, Bojan, KOSEM, Iztok, KREK, Simon, LASKOWSKI, Cyprian, ROBNIK ŠIKONJA, Marko. Thesaurus of Modern Slovene: By the Community for the Community. V: Čibej, Jaka, Vojko Gorjanc, Iztok Kosem, Simon Krek (ur.). *Proceedings of the XVIII EURALEX International Congress: Lexicography in Global Contexts.* ISBN 978-961-06-0097-8). 1. izd. Ljubljana: Znanstvena založba Filozofske fakultete. 2018, str. 401-410. https://e-knjige.ff.uni-lj.si/znanstvena-zalozba/catalog/view/118/211/3000-1.

Čibej 2021 = Čibej, J. (2021). Strojno luščenje medbesednih povezav v oblikoslovnem leksikonu Sloleks 2.0. In Š. Arhar Holdt (Ed.), Nova slovnica sodobne standardne slovenščine: viri in metode (pp. 125–159). Znanstvena založba Filozofske fakultete. https://e-knjige.ff.uni-lj.si/znanstvena-zalozba/catalog/view/325/477/7313-1.

Dobrovoljc et al. 2015 = Dobrovoljc, Kaja, Krek, Simon, Holozan, Peter, Erjavec, Tomaž, Romih, Miro, 2015, *Morphological lexicon Sloleks 1.2*, Slovenian language resource repository CLARIN.SI, ISSN 2820-4042, http://hdl.handle.net/11356/1039.

Dobrovoljc et al. 2019 = Dobrovoljc, Kaja et al., 2019, *Morphological lexicon Sloleks 2.0*, Slovenian language resource repository CLARIN.SI, ISSN 2820-4042, http://hdl.handle.net/11356/1230.

Erjavec et al. 2021 = Erjavec, Tomaž et al., 2021, *Multilingual comparable corpora of parliamentary debates ParlaMint 2.1*, Slovenian language resource repository CLARIN.SI, ISSN 2820-4042, http://hdl.handle.net/11356/1432.

Kosem et al. 2018a = KOSEM, Iztok (avtor, član uredniškega odbora), BÁLINT ČEH, Júlia (avtor, član uredniškega odbora), GORJANC, Vojko (avtor, član uredniškega odbora), KOLLÁTH, Anna (avtor, član

Operacijo Razvoj slovenščine v digitalnem okolju sofinancirata Republika Slovenija in Evropska unija iz Evropskega sklada za regionalni razvoj. Operacija se izvaja v okviru Operativnega programa za izvajanje evropske kohezijske politike v obdobju 2014-2020.

uredniškega odbora), KOVÁCS, Attila (avtor, član uredniškega odbora), KREK, Simon (avtor, član uredniškega odbora), NOVAK-LUKANOVIČ, Sonja (avtor, član uredniškega odbora), RUDAŠ, Jutka (avtor, član uredniškega odbora). Osnutek koncepta novega velikega slovensko-madžarskega slovarja. Ljubljana: Univerza v Ljubljani, Filozofska fakulteta, 2018. https://www.cjvt.si/komass/wp-content/uploads/sites/17/2020/08/Osnutek-koncepta-VSMS-v1-1.pdf.

Kosem et al. 2018b = KOSEM, Iztok, KREK, Simon, GANTAR, Polona, ARHAR HOLDT, Špela, ČIBEJ, Jaka, LASKOWSKI, Cyprian. Kolokacijski slovar sodobne slovenščine. V: FIŠER, Darja (ur.), PANČUR, Andrej (ur.). *Zbornik konference Jezikovne tehnologije in digitalna humanistika / Proceedings of the conference on Language Technologies & Digital Humanities,* 20.-21. september 2018, Ljubljana. Ljubljana: Znanstvena založba Filozofske fakultete v Ljubljani. 2018, str. 133-139, http://www.sdjt.si/wp/wp-content/uploads/2018/09/JTDH-2018_Kosem-et-al_Kolokacijski-slovar-sodobne-slovenscine.pdf.

Krek et al. 2019 = Krek, Simon et al., 2019, *Corpus of Written Standard Slovene Gigafida 2.0*, Slovenian language resource repository CLARIN.SI, ISSN 2820-4042, http://hdl.handle.net/11356/1320.

Krsnik 2017 = Krsnik, Luka. *Napovedovanje naglasa slovenskih besed z metodami strojnega učenja*: magistrsko delo: magistrski program druge stopnje Računalništvo in informatika. Ljubljana, 2017. http://eprints.fri.uni-lj.si/3978/.

Ljubešić et al. 2017 = Ljubešić, Nikola, Erjavec, Tomaž in Fišer, Darja, 2017, *Twitter corpus Janes-Tweet 1.0*, Slovenian language resource repository CLARIN.SI, ISSN 2820-4042, http://hdl.handle.net/11356/1142.

Žagar et al. 2022 = Žagar, Aleš et al., 2022, *Corpus of academic Slovene KAS 2.0*, Slovenian language resource repository CLARIN.SI, ISSN 2820-4042, http://hdl.handle.net/11356/1448.

Zwitter et al. 2021 = Zwitter Vitez, Ana, Zemljarič Miklavčič, Jana, Krek, Simon, Stabej, Marko in Erjavec, Tomaž, 2021, *Spoken corpus Gos 1.1*, Slovenian language resource repository CLARIN.SI, ISSN 2820-4042, http://hdl.handle.net/11356/1438.

Operacijo Razvoj slovenščine v digitalnem okolju sofinancirata Republika Slovenija in Evropska unija iz Evropskega sklada za regionalni razvoj. Operacija se izvaja v okviru Operativnega programa za izvajanje evropske kohezijske politike v obdobju 2014-2020.