REPUBLIKA SLOVENIJA
MINISTRSTVO ZA KULTURO

EVROPSKA UNIJA
EVROPSKI SKLAD ZA REGIONALNI RAZVOJ
NALOŽBA V VAŠO PRIHODNOST

KOST 1.0: Korpus usvajanja slovenščine kot drugega/tujega jezika

Poročilo projekta Razvoj slovenščine v digitalnem okolju

Aktivnost DS1.6

Avtorji: Mojca Stritar Kučuk[□], Špela Arhar Holdt[○,□], Iztok Kosem[○], Tina Munda[○], Luka Krsnik[○]

[○] 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

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 Špela Arhar Holdt: Projektni okvir

Poročilo oz. kazalnik *KOST 1.0: Korpus usvajanja slovenščine kot drugega/tujega jezika* 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 besedilne korpuse, ki smo se jim na projektu posebej posvetili, sodi tudi specializiran korpus, ki ga opisujemo v tem kazalniku: korpus usvajanja slovenščine kot drugega ali tujega jezika KOST. Korpus ponuja pomemben uvid v jezikovno rabo določenega tipa, poseben pa je tudi zato, ker zahteva dodatne korake pri pripravi korpusnega gradiva: vsebuje namreč informacije o popravkih jezikovnih napak. Da bi s pripravo in nadgradnjo tovrstnih korpusov lahko učinkovito nadaljevali, je bilo treba vzpostaviti protokole za kontinuirano zbiranje in procesiranje korpusnega gradiva, razvili pa smo tudi nova orodja za ročno označevanje in kategoriziranje jezikovnih popravkov. Nova orodja so odprto dostopna za nadaljnjo rabo, že med projektom pa smo jih uporabili za izboljšavo in dopolnitev obstoječih korpusov.

2 Špela Arhar Holdt, Tina Munda: Opis projektnih aktivnosti

Eden izmed ciljev projekta RSDO je bil pripraviti načrt za kontinuirano nadgrajevanje korpusov usvajanja slovenščine ter vzpostaviti delotoke in orodja za redno zbiranje in procesiranje besedil.

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.

Aktivnost se je posvečala korpusom usvajanja slovenščine, ki v primerjavi z ostalimi korpusnimi viri vsebujejo oznake jezikovnih napak oz. popravkov. Tako označena besedila so izjemnega pomena, saj omogočajo empirični uvid v opismenjevanje šolajoče se populacije, pripravo didaktičnih gradiv ter orodij, ki opismenjevanje in pismenost podpirajo in razvijajo. V projektu smo se koordinirano posvečali gradnji korpusa KOST (korpus slovenščine kot tujega in drugega jezika), ki vsebuje besedila, ki jih pišejo tujci v okviru programov Leta plus in različnih dejavnosti Centra za slovenščino kot drugi in tuji jezik. V ta besedila so bili ročno pripisani in kategorizirani jezikovni popravki. Za učinkovito pripravo in nadgradnjo korpusa smo vzpostavili ustrezne delotoke in razvili novo orodje.

Na ravni procesiranja pridobljenih korpusnih besedil smo izbrali uporabniku prijazno in zmogljivo orodje – za slovenščino smo lokalizirali in nadgradili orodje Svala, ki omogoča (a) transkripcijo besedil, ki so napisana na roko, in (b) označevanje jezikovnih popravkov, ki je bilo doslej najbolj zamudno in programsko izredno slabo podprto. Ob razvoju programa smo na novo definirali format korpusnih besedil in ga dali na voljo tudi v ostalih formatih, da je kompatibilen z ostalimi slovenskimi korpusi.

Na projektni aktivnosti je sodelovalo več raziskovalcev z Univerze v Ljubljani, pa tudi zunanjih sodelavcev, programer, lektorji, študenti. Soavtorstvo je navedeno pri vsakem posameznem projektnem rezultatu.

Delo je potekalo prek celotnega časa trajanja projekta po naslednjih korakih:

- Aktivacija mreže učiteljev slovenščine kot drugega oz. tujega jezika, lektorjev in drugih, ki so sodelovali pri testnem zbiranju gradiva v času trajanja projekta. Sodelovanje je služilo za evalvacijo delotokov. V sodelovanje smo vključili več kot 24 sodelavcev s programov, povezanih s slovenščino kot drugim oz. tujim jezikom.
- Izbira programa za označevanje in kategorizacijo jezikovnih napak. Izbrali smo odprto dostopen program, ki omogoča pregledno in učinkovito delo s podatki. Uporabniška prijaznost je zagotovljena tudi na ravni uvoza in izvoza podatkov. Projektno delo je zajemalo lokalizacijo odprto dostopnega programa Svala, ki smo ga prilagodili tako, da vsebuje predpripravljene nabore kategorij oznak za korpusa KOST in Šolar. Za testiranje programa smo označili del korpusnih besedil in glede na izkušnjo označevalcev implementirali dodatne popravke.

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.

Uporabnost orodja smo preizkusili na skupini študentov slovenistike in ocenili, do kolikšne mere je mogoče v delotoke vključiti študentsko delo.

- Priprava korpusnih gradiv za vključitev v konkordančnike. Za korpuse, ki vsebujejo jezikovne popravke, smo razvili specializirani format XML TEI, ki je skladen z ostalimi slovenskimi besedilnimi korpusi in povezljiv s formatom orodja Svala.

REZULTATI AKTIVNOSTI

- Program za označevanje jezikovnih napak oz. popravkov v korpusih usvajanja slovenščine: https://orodja.cjvt.si/svala/.
- Korpus KOST 1.0 kot baza na repozitoriju CLARIN.SI: http://hdl.handle.net/11356/1753, kjer so na voljo tudi povezave na konkordančnike.
- (Pričujoče) specifikacije za nadgradnjo korpusa KOST z opredelitvijo želene vsebine, označevanja in kategoriziranja jezikovnih napak.

3 Mojca Stritar Kučuk: Korpus slovenščine kot tujega jezika KOST

3.1 KOST 1.0

Korpus slovenščine kot tujega jezika KOST je digitalna zbirka besedil, ki so jih napisali odrasli govorci, za katere slovenščina ni prvi jezik. Ime KOST (= korpus slovenščine kot tujega jezika) ni popolnoma terminološko ustrezno, saj gre pri večjem delu vključenih besedil za slovenščino kot drugi in ne tuji jezik, vendar je bilo izbrano zaradi večje ekonomičnosti.

Korpusi usvajanja tujega jezika (angl. *learner corpora*) so v zadnjem desetletju doživeli razmah. Število tovrstnih korpusov je glede na seznam obstoječih korpusov, ki ga vodi Centre for English Corpus Linguistics, poskočilo s 73 korpusov leta 2012 na 191 korpusov leta 2022. Največ, 121, je bilo pisnih korpusov, za katere je tudi najlažje pridobivati besedila, sledili so jim govorni korpusi (44), 24 pa je bilo pisnih in govornih korpusov. Večina teh korpusov ima en ciljni jezik, dobra desetina pa jih vključuje več. Prednjači angleščina, ki je ciljni jezik v 52 % korpusov. Med ostalimi ciljnimi jeziki so arabščina, češčina, estonščina, finščina, francoščina, gelščina, hrvaščina, islandščina, italijanščina, katalonščina, kitajščina,

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.

korejščina, latvijščina, litovščina, madžarščina, nemščina, nizozemščina, norveščina, perzijščina, poljščina, portugalščina, romunščina, ruščina, španščina in švedščina.

Nekakšna zlata mera za obstoječe korpuse usvajanja jezikov, ki so v približno primerljivem sociolingvističnem položaju kot slovenščina, je pisni korpus z milijonom besed, različnimi prvimi jeziki tvorcev ter dodanimi oblikoskladenjskimi oznakami in oznakami napak.

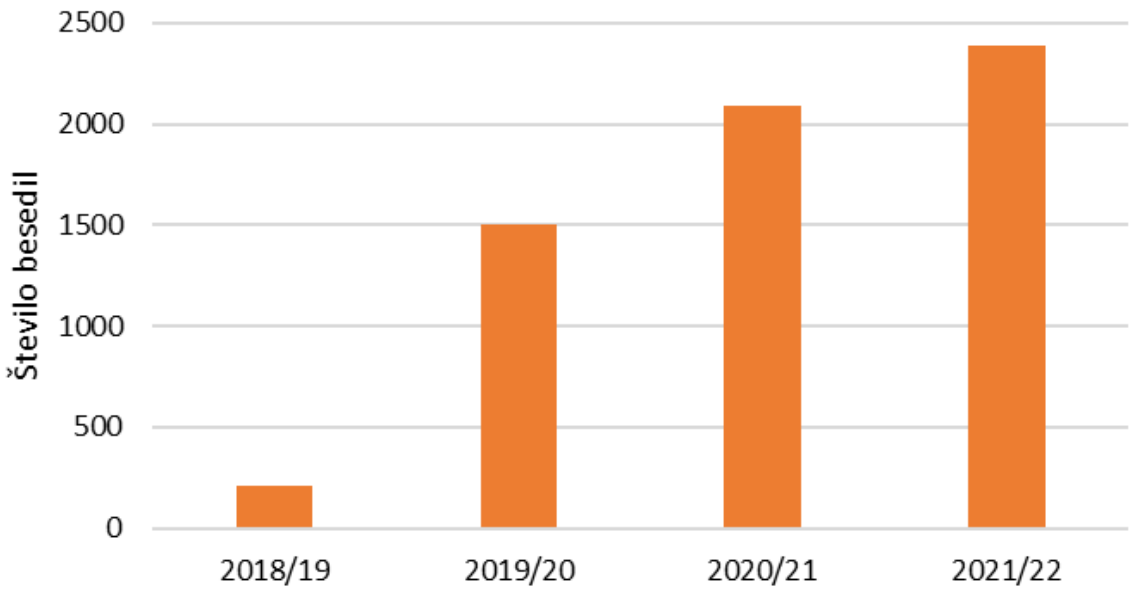
KOST 1.0 obsega 835.000 besed in se po zasnovi, vsebini, velikosti, označenosti, dostopnosti in uporabnosti lahko primerja z obstoječimi korpusi za druge jezike s sorodnim sociolingvističnim položajem, npr. češčino, hrvaščino ali latvijščino. Besedila so v glavnem nastala na lektoratih in tečajih slovenščine kot drugega oziroma tujega jezika. Večina tvorcev besedil kot prvi jezik govori srbsko, bosansko in makedonsko, vključena pa so tudi besedila govorcev drugih jezikov. Tvorci v slovenščini dosegajo različne ravni jezikovne zmožnosti, od začetnikov do izpopolnjevalcev.

3.2 Zbiranje besedil

Zbiranje besedil za KOST 1.0 poteka od leta 2019. Za zagon izdelave je bila ključna odločitev Univerze v Ljubljani, da od študijskega leta 2016/17 kot enega od ukrepov internacionalizacije univerze izvaja Leto plus. Gre za poseben modul, namenjen tujim študentom, redno vpisanim v študijske programe Univerze v Ljubljani. Hkrati z rednim študijem jim omogoča brezplačno učenje slovenščine na dveh lektoratih v obsegu treh študijskih ur na teden. Modul Leto plus je bil za zagon KOST-a idealen, saj je v njem zagotovljen dostop do večjega števila govorcev slovenščine kot tujega jezika, med 250 in 350. Gre za govorce s precej homogenim jezikovnim ozadjem, največ jih prihaja iz Bosne in Hercegovine, Srbije in Makedonije. Tudi njihova jezikovna zmožnost v slovenščini je precej primerljiva: v prvem, zimskem semestru večinoma začnejo brez predhodnega znanja slovenščine, zaradi skupne slovanske osnove pa hitro napredujejo in so ob koncu letnega semestra sposobni jezikovno (pre)živeti v slovenskem okolju, komunicirati o vsakdanjih temah in do neke mere v slovenščini opravljati redne študijske obveznosti. Dodatna prednost Leta plus je, da s temi govorci slovenščine kot tujega jezika učitelji nimajo samo stika pri pouku, pač pa od njih redno dobivajo tudi domače naloge, napisane v slovenščini. Ne nazadnje pa je za KOST pomembno tudi, da so pedagoški delavci na Letu plus večinoma redno zaposleni lektorji, tako da sodelovanje pri gradnji korpusa sodi med njihove redne delovne obveznosti.

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.

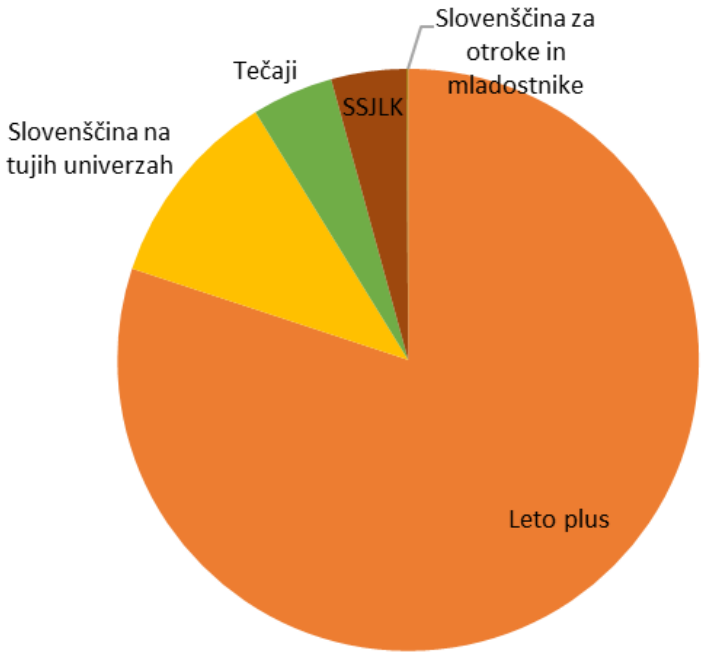
Zbiranje besedil za KOST 1.0 se je pričelo v študijskem letu 2018/19, kot prikazuje Graf 1, pa je bilo vsako leto zbrano več besedil.



Graf 1: Količina zbranih besedil po študijskih letih

Zbiranje besedil se je iz modula Leto plus razširilo še na različne programe Centra za slovenščino kot drugi in tuji jezik, kot prikazuje Graf 2. Seznam sodelavcev pri zbiranju besedil je dostopen na: https://www.cjvt.si/korpus-kost/projektna-ekipa/.

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.



Graf 2: Količina zbranih besedil po programih Centra za slovenščino kot drugi in tuji jezik

Velika večina vključenih besedil, skoraj 84 %, je napisana na računalnik. Covidno obdobje je bilo glede dostopa do takih besedil zelo produktivno, saj se je zaradi pandemije celotno poučevanje preselilo v digitalno okolje in se je občutno povečal dotok digitalno napisanih domačih nalog. Besedila, ki nastanejo na izpitih ali med poukom v razredu, pa so napisana na roko in jih je treba pretipkati. V skladu s svojimi časovnimi zmožnostmi to opravljajo učitelji ali strokovni delavci na programih.

Besedila v KOST-u so točno taka, kot jih napišejo tvorci, brez kakršnih koli jezikovnih popravkov. Edina izjema se nanaša na pravopis. Ker ta ni v ospredju raziskav pri slovenščini kot tujem jeziku in ker bi radi zvišali uspešnost morebitnega avtomatskega označevanja korpusnega gradiva v prihodnosti, v besedilih popravimo stičnost ločil ter odstranimo dvojne presledke. Zaradi lažje berljivosti pri pretipkavanju besedil, prvotno napisanih samo z velikimi tiskanimi črkami, uporabimo male tiskane črke in pri tem upoštevamo slovenska pravopisna pravila.

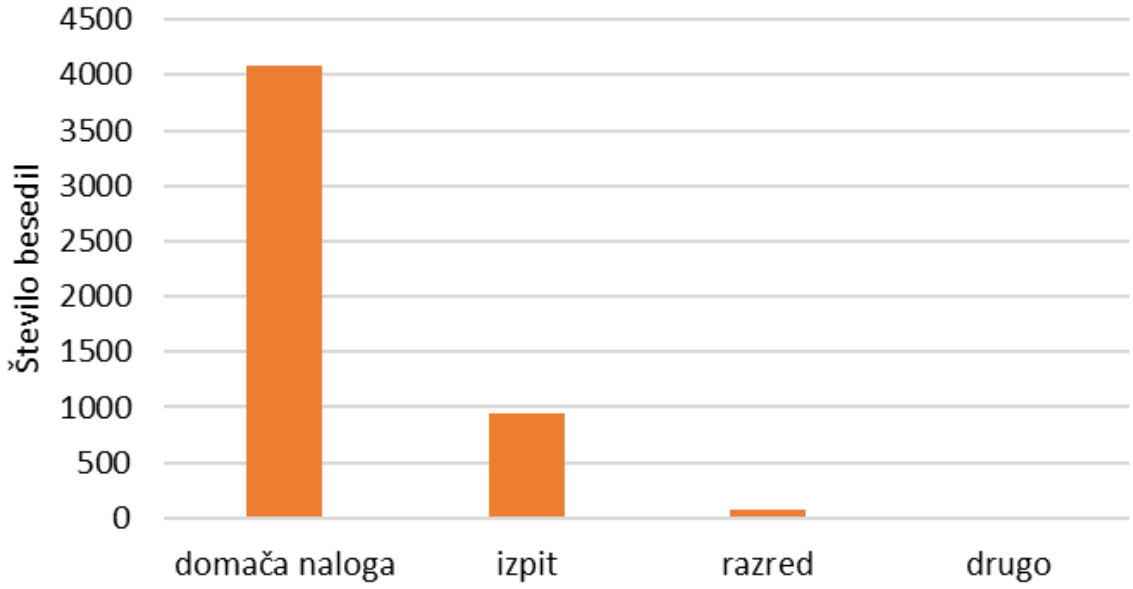
Vsako besedilo, vključeno v KOST, je shranjeno v samostojno tekstovno datoteko in poimenovano s kodo, ki izurjenemu uporabniku korpusa da nekaj osnovnih podatkov. Koda »L3-2122-121« denimo

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.

pomeni, da gre za besedilo s številko 121, ki je nastalo pri določenem učitelju v okviru programa Leto plus v študijskem letu 2021/22. Metajezikovni podatki o tvorcih in besedilih pa so zbrani v posebni Excelovi tabeli.

3.3 Vključena besedila

V KOST 1.0 je vključenih 6311 besedil oz. 1.032.012 besed. Skoraj 84 % jih je nastalo v digitalni obliki, preostala pa je bilo treba pretipkati. Graf 3 prikazuje okoliščine njihovega nastanka. Največ je t. i. domačih nalog, torej besedil, ki so jih tvorci napisali doma, brez nadzora učitelja. Pri tem je nemogoče nadzirati, ali jih dejansko napišejo sami in koliko si pomagajo z jezikovnimi viri in orodji. Tista besedila, za katera učitelj predvideva, da so bila prevedena v slovenščino s pomočjo strojnega prevajalnika, v KOST niso vključena, saj ne povedo ničesar direktnega o jezikovni zmožnosti njihovih tvorcev. Sledijo besedila z izpitov. Pri tem gre večinoma za interne izpite na tečajih ali lektoratih. V vsakem primeru so ta besedila nastala v kontroliranih okoliščinah, kar se tiče časovnih omejitev pri pisanju in uporabe zunanjih pripomočkov, ki običajno ni dovoljena. Nekaj besedil je bilo napisanih v razredu, v okviru različnih dejavnosti med poukom.

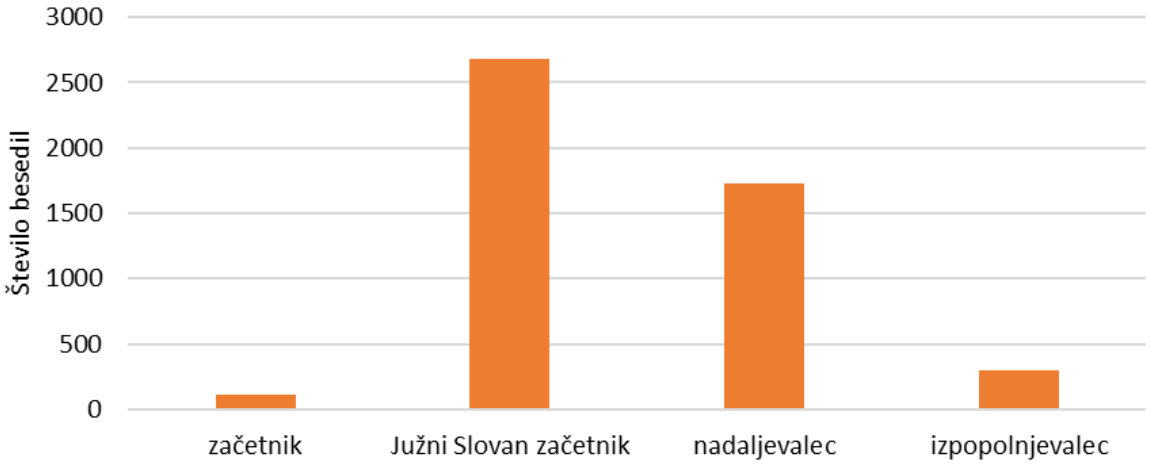


Graf 3: Okoliščine nastanka besedil, vključenih v KOST

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.

V KOST so vključene različne vrste besedil. Največ je esejev oz. spisov na različne teme (npr. o družini, prehrani, zdravju), nekaj pa je tudi praktičnega pisanja (npr. življenjepis, prošnja za delo). Ker je v praktičnih besedilih veliko osebnih podatkov, ki bi jih bilo treba zakrivati s kodami, in gre pretežno za ponavljanje ustaljenih vzorcev, le malo pa je dejanske tvorbe besedil, so v KOST vključena v manjši meri. Če so tvorci pred pisanjem dobili natančnejša navodila za pisanje, so tudi ta zabeležena med metapodatki, saj tvorci nemalokrat dobesedno ponavljajo fraze iz njih.

Besedila so razvrščena po štirih stopnjah, ki približno odslikavajo trenutno jezikovno zmožnost njihovih tvorcev, kar prikazuje Graf 4. Ta zmožnost sicer nikakor ni zanesljivo določena, temveč gre za pragmatično oceno, ki jo največkrat poda tvorčev trenutni učitelj, in je torej namenjena le okvirni orientaciji med besedili. Po tej lestvici je v KOST-u največ besedil Južnih Slovanov začetnikov, to je govorcev katerega od osrednjejužnoslovanskih jezikov (bosanščine, hrvaščine, črnogorščine, srbščine) ali makedonščine, ki so se slovensko šele začeli učiti pred največ dvema semestroma. Njihov napredek je zaradi sorodnosti izhodiščnega in ciljnega jezika običajno hiter. Kot nadaljevalci so označeni tisti, ki so se slovensko že učili pred udeležbo v programu, v okviru katerega je nastalo v korpus vključeno besedilo. Med njimi so lahko velike razlike (npr. med slovanskimi in neslovanskimi nadaljevalci). Manj je besedil izpopolnjevalcev, ki so ponavadi daljša, kompleksnejša in z manj napakami. Najmanj pa je besedil začetnikov, torej govorcev slovenščini nesorodnih jezikov v začetnih fazah učenja. Njihova besedila so tudi relativno najkrajša.

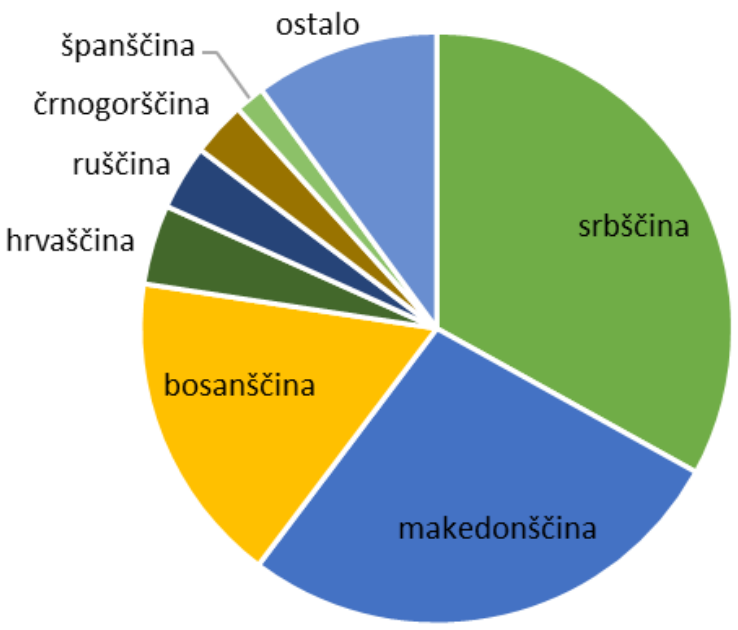


Graf 4: Štiri stopnje ocenjene jezikovne zmožnosti tvorcev besedil v slovenščini v KOST 1.0

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.

3.4 Tvorci besedil in osebni podatki

V KOST 1.0 so vključena besedila več kot 950 tvorcev, od tega je slabih 34 % moških in 66 % žensk. Govorijo 30 različnih prvih jezikov, najpogostejši med njimi so prikazani na spodnjem grafu. V skladu s populacijo na modulu Leto plus dobre tri četrtine vseh tvorcev predstavljajo govorci osrednjejužnoslovanskih jezikov in makedonščine. Ali med srbščino, bosanščino in hrvaščino prihaja do jezikovnih razlik in ali je razlikovanje med njimi za potrebe jezikovne analize sploh potrebno, bodo pokazale nadaljnje analize. Med ostalimi jeziki, ki so v KOST-u zastopani z manj tvorci, so albanščina, angleščina, francoščina, grščina, hebrejščina, italijanščina, japonščina, kirgiščina, kitajščina, korejščina, madžarščina, nemščina, nizozemščina, poljščina, romunščina, slovaščina, slovenščina (gre za tvorce iz slovenskega zamejstva), srbohrvaščina in ukrajinščina.



Graf 5: Prvi jeziki tvorcev besedil, vključenih v KOST 1.0

Ker sta ureditev pravic za uporabo podatkov in varovanje osebnih podatkov ključnega pomena, vsi tvorci, katerih besedila so vključena v KOST, podpišejo izjavo, s katero dovoljujejo vključitev svojih

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.

besedil. K izjavi sodi zbiranje osebnih podatkov, ki so nujni za analizo korpusnega gradiva: spol, starost, fakulteta, letnik in stopnja študija, izobrazba, prvi jezik in ostali jeziki, ki jih znajo govorci, ter podatki o morebitnem predhodnem učenju slovenščine ali bivanju v Sloveniji. Vsi ti podatki so v KOST-u zabeleženi kot metapodatki. Izjavo, ki so jo pravno preverili na Oddelku za upravljanje s tveganji in varstvo osebnih podatkov na Univerzi v Ljubljani, sodelujočim v podpis ponudijo njihovi učitelji. Pred podpisom jim natančno razložijo projekt in pogoje sodelovanja. Razveseljivo je, da izjavo podpiše velika večina vseh, ki jim je bila ponujena. Vse izjave so shranjene v digitalni in, če so bile podpisane na papirju, tudi tiskani obliki.

Tvorci so v korpusu anonimni. Njihova imena so nadomeščena s kodami; koda »L-hr-m-0006« denimo pomeni, da gre za tvorca moškega spola s prvim jezikom hrvaščino, ki je dobil zaporedno številko 6. Vsi osebni podatki v besedilih so nadomeščeni s kodami v oglatih oklepajih, npr. osebna imena so nadomeščena s kodo [XImeX], krajevna pa z [XKrajX]. Tako izgubimo nekaj jezikovnih informacij, denimo o pregibanju imen, vendar s tem zadostimo zahtevam po varovanju osebnih podatkov. Kjer so v besedilih lastna imena ohranjena, gre za pisanje o fiktivnih osebah.

3.5 Označevanje napak

V korpusu KOST 1.0 so na delu besedil označene jezikovne napake. Označevanje napak poteka ročno. V KOST-u 1.0 so označene na 10 % vseh besedil, kar je ustaljen delež tudi v drugih korpusih, denimo v češkem CzeSL. Vsaka napaka dobi oznako glede na spodnjo klasifikacijo (gl. Tabelo 1), pripisana pa ji je tudi popravljena oblika. Kategorije napak temeljijo na predhodni klasifikaciji, ki je bila preizkušena za poskusni korpus slovenščine kot tujega jezika (PiKUST), prilagojena za prvo verzijo korpusa usvajanja slovenščine kot prvega jezika Šolar in prilagojena tudi zahtevam označevalnega orodja Svala.

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: Kategorije napak s primeri in številom pojavitev v KOST 1.0

Krovna kategorija	**Kategorija napake**	**Število pojavitev v KOST-u**	**Primer iz KOST-a**	
			Napačna oblika	**Popravljena oblika**
Napake zapisa	Ločilo	3658	V zimskem času**,** večina ljudi uporablja	V zimskem času večina ljudi uporablja
	Črkovanje	1346	ne i**s**pušča	ne i**z**pušča
	Skupaj/Narazen	227	**ni sem** poskusil	**nisem** poskusil
	Mala/velika začetnica	460	energija **I**n paneli	energija **i**n paneli
	Krajšave	3	**in dr.**	**idr.**
Napake besedišča	Samostalnik	698	izobraževanje sem skupaj s celotnim **društvom** nadaljevala	izobraževanje sem skupaj s celotno **družbo** nadaljevala
	Glagol	883	vozila, ki **hodijo** na gas	vozila, ki **vozijo** na plin
	Zaimek	489	hodim na zmenke s **mojim** fantom	hodim na zmenke s **svojim** fantom
	Pridevnik	455	**vetrene** elektrarne	**vetrne** elektrarne
	Prislov	397	grem **doma**	grem **domov**
	Predlog	503	**sa** enom prijateljicom	**z** eno prijateljico
	Veznik	400	Moje najlepše potovanje je bilo kdaj sem bil v Kitajski.	Moje najlepše potovanje je bilo, **ko** sem bil na Kitajskem.
	Ostalo	320	**petindvajest**	**petindvajset**
Napake oblike	Samostalnik	2370	v Sloveni**u**	v Sloveni**o**
	Glagol	1263	tam živi**mo** sestra in jaz	tam živi**va** sestra in jaz
	Zaimek	523	ker sem **ih** spoznal	ker sem **jih** spoznal
	Pridevnik	916	Pohištvo je v zelo **dobremu** stanju	Pohištvo je v zelo **dobrem** stanju

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.

	Prislov	40	Želim se naučiti kar hitrije slovenščino	Želim se naučiti čim **hitreje** slovenščino
	Ostalo	94	štir**je** predavanja	štir**i** predavanja
Napake skladnje	Struktura	347	rada bi **da živim**	rada bi **živela**
	Besedni red	1781	Zdi **mi se** strogo	Zdi **se mi** strogo
	Izpuščeni jezikovni elementi	736	Na fakulteto grem avtobusom	Na fakulteto grem **z** avtobusom
	Odvečni jezikovni elementi	508	upam **se** da bom uspel	upam, da bom uspel

V Tabeli 1 ni prikazana kategorija povezanih popravkov, ki jo lahko dodamo vsem ostalim oznakam. Gre za oblike v besedilu, ki jih je treba popraviti, potem ko popravimo nekaj drugega v sobesedilu. Tipičen primer je napaka besedišča pri samostalniku v zgornji tabeli, kjer je treba ob popravku nepravilnega samostalnika popraviti tudi obliko pridevnika. To je označeno kot napaka oblike pridevnika in hkrati kot povezan popravek. Takih oznak je v KOST-u 747.

Natančnejša navodila za označevanje napak so na voljo v stalno dopolnjujočem se priročniku za označevanje napak, dostopnem na spletni strani korpusa. Z označevanjem dodatnega gradiva se namreč pojavljajo tudi nove dileme, ki jih razrešujemo sproti. Načeloma pa je osnovno vodilo, da s popravki čim manj posegamo v besedilo in popravljamo čim manj napak. Izogibamo se stilističnim popravkom in popravljamo predvsem zapis, besedišče in obliko besed, v skladnjo pa skušamo posegati čim manj. V redkih primerih, kadar napačni obliki ne znamo pripisati popravljene, to označimo s [???]. Pomembno pa je, da se uporabniki KOST-a zavedajo, da so oznake napak subjektivne. Zato bo kakršna koli poglobljena analiza napak vedno zahtevala tudi temeljit ročni pregled zadetkov.

Večina označevanja napak za KOST 1.0 je opravila urednica korpusa, v jesenskih semestrih 2021/22 in 2022/23 pa smo preverili tudi, kako enostavno in učinkovito je označevanje napak za polprofesionalne uporabnike, namreč za študente 3. letnika 1. stopnje slovenistike. 39 študentov je označilo 172 besedil. Pred tem smo načrtno izvedli le krajše usposabljanje oz. prikaz dela s Svalo, saj smo želeli preizkusiti, kako dobro se znajdejo brez podrobnejših navodil. Besedila, ki so jih označili, je nato pregledala še urednica KOST-a, študenti pa so svoje delo predstavili v okviru seminarja pri predmetu Slovenščina kot

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.

tuji jezik. Rezultati so bili pozitivni: čeprav je bilo v povprečju v njihovih besedilih 35 % neustreznih oznak, pa nobeden od označevalcev ni imel večjih težav s samo aplikacijo za označevanje Svala. Še največji izziv jim je predstavljalo združevanje besed v skupine. Študenti so tovrstno delo ocenili kot zanimivo, strokovno precej, tehnično pa manj zahtevno, razmeroma zamudno, a koristno zanje in za širšo skupnost. Izrazili so zadovoljstvo z možnostjo praktičnega, tehnično nezahtevnega dela, pri katerem so morali dejansko uporabiti tudi jezikoslovno znanje, pridobljeno pri študiju. (Arhar Holdt et al. 2022: 28, 29)

3.6 Dostop do korpusa

Korpus KOST 1.0 lahko izključno v izobraževalne in raziskovalne namene uporabljajo učitelji, študenti, raziskovalci in drugi, ki jih zanima slovenščina kot tuji jezik. Kot baza je dostopen na repozitoriju CLARIN.SI[1] pod pogoji licence CC BY-SA 4.0. Poleg tega je bil v projektu RSDO razvit format korpusov z jezikovnimi popravki, med katere spada tudi korpus KOST. Po novem je zanje na voljo specializirani XML TEI, ki omogoča vključevanje v konkordančnike, kot so noSkE Bonito, Crystal in KonText.

V prihodnosti želimo poleg obstoječega razviti specializirani konkordančnik za korpus KOST, ki bo omogočal polno izrabo bogato označenega korpusnega gradiva, vključno z možnostmi izrabe metapodatkov in sočasne vizualizacije izvornega ter popravljenega besedila. Z razvojem takega konkordančnika bo KOST postal širše uporaben jezikovni vir, zanimiv za vse, ki raziskujejo slovenščino kot drugi oz. tuji jezik, torej za pisce učnih gradiv, oblikovalce kurikulov, učitelje ali tiste, ki jih zanima jezik na splošno. Omogočal bo prepoznavo najpogostejših jezikovnih napak, značilnih za govorce določenih prvih jezikov, in s tem pripravo bolj osredotočenih učnih gradiv, pa tudi ustreznejše poudarke v samem pedagoškem procesu.

Gradivo iz korpusa KOST je bilo uporabljeno tudi pri pripravi učbenika za slovenščino kot drugi jezik za južnoslovanske govorce, v katerem je poudarjen kontrastivni vidik poučevanja (Slovenščina 1+, Stritar Kučuk in Šter 2021).

[1] http://hdl.handle.net/11356/1753

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 Špela Arhar Holdt, Iztok Kosem, Mojca Stritar Kučuk, Luka Krsnik, Tina Munda: Program Svala in novi format

4.1 Program Svala

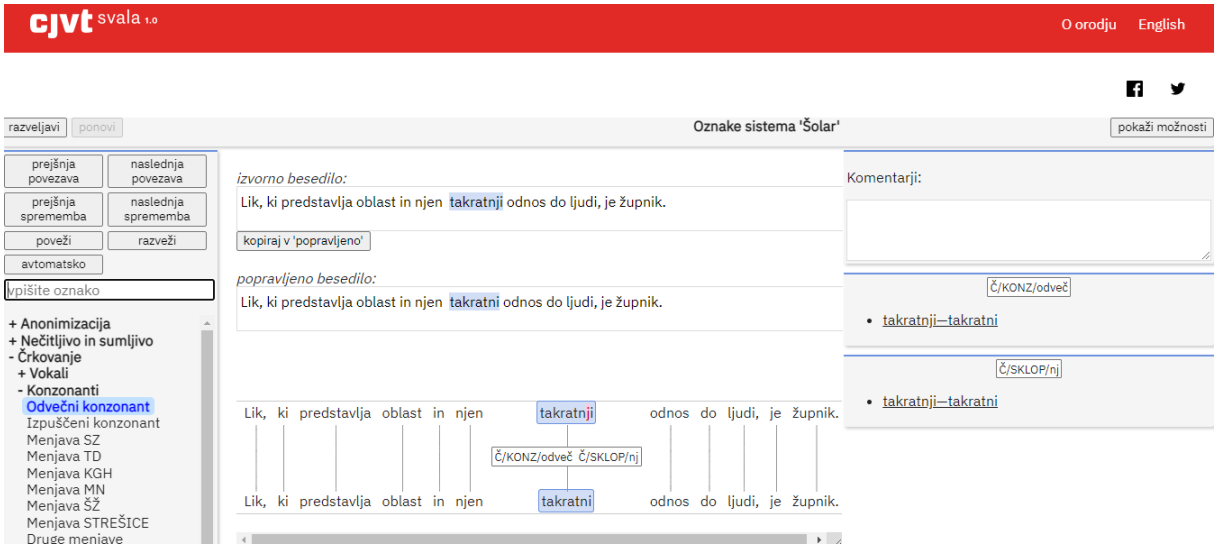
Novi program za prikaz in urejanje besedil z jezikovnimi popravki CJVT-Svala predstavlja poleg portala za zbiranje besedil prav tako pomemben napredek pri metodologiji gradnje korpusov usvajanja slovenščine in s tem pri njihovi učinkoviti nadgradnji.

Pred posvojitvijo tega odprtokodnega programa smo prekategorizacijo jezikovnih popravkov opravljali v za naše namene prilagojenem orodju Sketch Engine. Ta postopek je bil izjemno časovno obremenjujoč – korpus smo morali najprej pretvoriti v format VERT za uvoz v Sketch Engine; tam smo učiteljske jezikovne popravke označevali z novimi kategorijami (gl. razdelek 3.1), kar je spremljalo sprotno izvažanje korpusnih datotek in njihovo pretvarjanje v XML, da smo lahko nove oznake kategorij, ki se v Sketch Enginu beležijo posebej, vključili v zapis korpusnih datotek lokalno; nato je bilo treba spet opraviti pretvorbo za ponoven uvoz v Sketch Engine. (Arhar Holdt et al. 2022: 27)

Tak potek dela nas je motiviral, da smo se ozrli po praksah evropske jezikovnotehnološke skupnosti. Švedski kolegi z univerz v Göteborgu in Stockholmu so za gradnjo korpusa švedščine kot drugega ali tujega jezika zasnovali orodje Svala (Wirén 2019), ki smo ga ocenili kot dobro izhodišče za učinkovitejšo doseganje naših projektnih ciljev. Največja prednost orodja je, da združuje več korakov delotoka izgradnje korpusa usvajanja jezika, in sicer psevdomizacijo, normalizacijo in označevanje jezikovnih popravkov v besedilih (Arhar Holdt et al. 2022: 27). Poleg tega portal SweLL (Volodina et al. 2019), v katerega je orodje Svala vključeno, skrbi še za vodenje delotokov za zbiranje in urejanje korpusnega gradiva (prav tam). To orodje smo v projektu RSDO prilagodili za slovenščino in ga nadgradili, da omogoča delo s korpusoma KOST in Šolar. V luči potreb smo v CJVT-Svala prioritetno prenesli funkcionalnosti za transkripcijo, preprosto anonimizacijo in označevanje napak, naprednejše module, tj. avtomatsko podprto anonimizacijo in vodenje označevalnih delotokov, pa smo odložili na prihodnje priložnosti za razvoj (prav tam). Orodje je na voljo na spletnem mestu https://orodja.cjvt.si/svala/, v nadaljevanju pa predstavljamo njegovo delovanje in uporabo.

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.

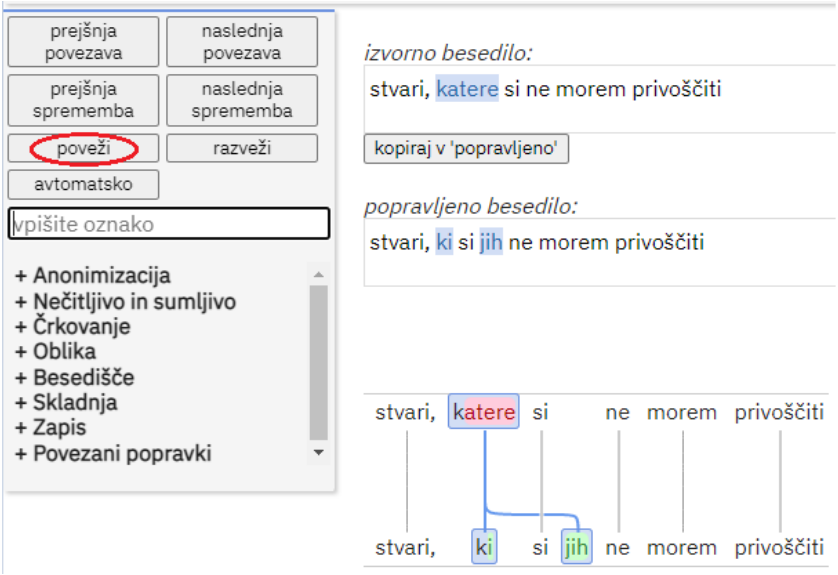
Izvorno besedilo uvozimo, prilepimo ali transkribiramo v okence pod napisom 'izvorno besedilo'. S klikom na gumb 'kopiraj v popravljeno' se izvorno besedilo prenese v okence pod napisom 'popravljeno besedilo', kamor vnesemo jezikovne popravke. V okencu na dnu strani, v t. i. grafu povezav, se eno pod drugim izpišeta izvorno in popravljeno besedilo, njuni deli pa se med seboj avtomatsko povežejo s črticami, ki jim žargonsko rečemo *špageti*. S klikom na vsako črtico je mogoče pripisati kategorijo oznak za popravke, in sicer ločeno za korpus Šolar in za korpus KOST (več o oznakah gl. razdelek 3.5). Na Sliki 8 je v meniju na levi strani zaslona prikazan sistem oznak za Šolar, pri čemer se do želene oznake lahko preklikamo prek kategorij in podkategorij ali pa v iskalno polje nad seznamom kategorij ('vpišite oznako') vpišemo del opisa popravka in avtomatsko nas bo premestilo na oznako na seznamu, ki se vsaj delno ujema z iskalnim nizom. Z izbiro oznake s seznama se ta obesi na izbran *špaget* v grafu povezav.



Slika 8: Primer izvornega in popravljenega besedila v vmesniku CJVT-Svala s sistemom oznak za Šolar

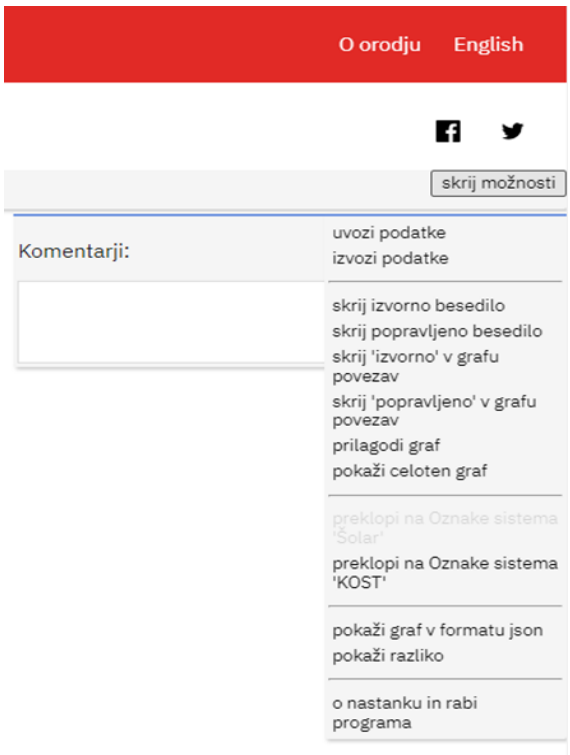
Poleg načina izpisa obeh verzij besedila in medbesedilnih povezav s pripeto oznako za jezikovno težavo so vizualna pomoč tudi barve – napaka iz izvornega besedila je obarvana rdeče, popravek napake pa zeleno. Nad iskalnim poljem v levem meniju najdemo še praktične gumbe z ukazi za premik na prejšnjo/naslednjo povezavo, prejšnjo/naslednjo spremembo, za ročno povezavo ali razvezavo neustrezno povezanih pojavnic (recimo ko se več besed v izvornem besedilu zamenja z eno v popravljenem ali obratno; gl. Sliko 9).

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.



Slika 9: Ročno povezovanje pojavnic v CJVT-Svali

Na desni strani zaslona je okno 'Komentarji', kjer lahko označevalec poda dodaten komentar, in spustni seznam z možnostmi, do katerega dostopamo s klikom na 'prikaži možnosti', kot prikazuje Slika 10.



Slika 10: Okno za komentar in možnosti v CJVT-Svali

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.

Tukaj so funkcionalnosti za uvoz/izvoz besedila, prikaz v vmesniku (skrij/pokaži izvorno besedilo/popravljeno in nastavitve prikaza grafa povezav), preklop med oznakami sistema KOST in Šolar, možnost prikaza grafa povezav v formatu JSON in povezava na uporabniški priročnik orodja (v angleščini).

V skrajnem zgornjem desnem kotu je možnost menjave jezika v angleščino in polje 'O orodju' z opisom vmesnika.

Orodje smo evalvirali na dveh nalogah: za gradnjo korpusa KOST 1.0, kjer je bil v vmesniku označen del besedil, in nadgradnjo Šolarja na tretjo različico, natančneje popravljanje in usklajevanje oznak iz druge različice glede na posodobljene smernice za označevanje.

4.2 Novi korpusni format

Poleg orodja za učinkovito urejanje besedil z jezikovnimi popravki (gl. razdelek 4.1) je še ena pomembna rešitev za tovrstne korpusne vire, zasnovana v tem projektu, standardni format za zapis korpusa z jezikovnimi popravki. Verzija Šolar 2.0 je že odsevala shemo TEI, a še ni bila povsem usklajena s TEI standardom, ki je v rabi za druge korpuse, smo po posvetovanjih v širši projektni ekipi zasnovali popolnoma nov format korpusa z jezikovnimi popravki. Najprej smo izvožene datoteke v formatu JSON, ki je privzeti datotečni format za izvoz iz Svale, pretvorili v XML po shemi TEI[2]. V svojem ustroju TEI XML predvideva element za zapis popravkov v besedilu, a smo glede na množico in vrsto podatkov, ki smo jih želeli zakodirati v korpusno datoteko, morali obstoječo strukturo ponovno premisliti. Nov format korpusa usvajanja jezika je posledica predvsem pristopa, ki ga uvaja Svala. Gre za tri izhodne datoteke JSON (nepopravljena (kost-errs) in popravljena besedila (kost-corr) ter datoteka s povezavami med vsako pojavnico iz nepopravljene in popravljene verzije datoteke, skupaj z oznakami za jezikovne popravke), ki so nato pretvorjene v XML in združene v eno datoteko XML, ki je skladna s shemo TEI.

Takšno sestavo ima tudi Šolar, od KOST pa se razlikuje po strukturi datoteke – v KOST metapodatki vsebujejo več informacij o tvorcu in okoliščinah nastanka besedila – in po tipologiji jezikovnih popravkov.

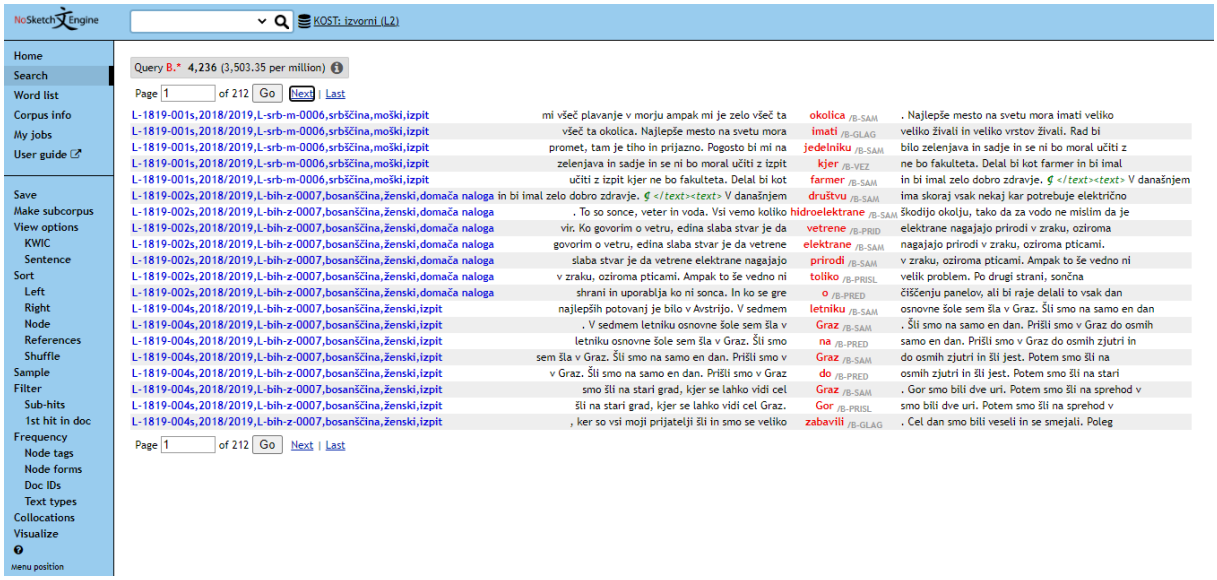
[2] https://tei-c.org/

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.

Korpus je na voljo še v bolj robustnih formatih CoNLL-U in JSON, ki sta primernejša za strojno procesiranje, ter VERT, ki ga tipično zahteva uvoz v konkordančnike (npr. Sketch Engine).

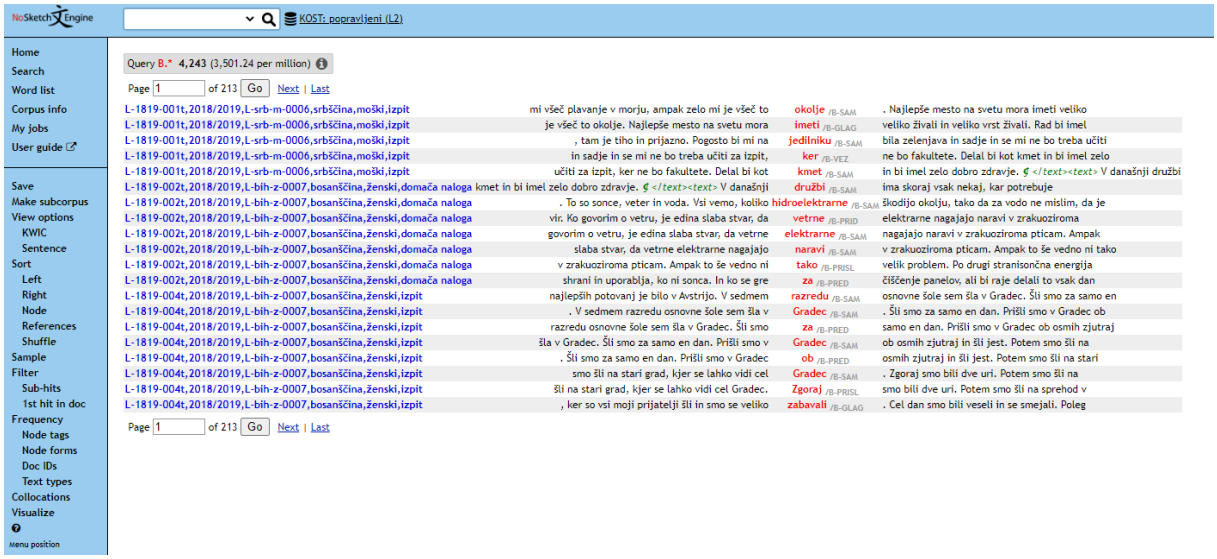
4.3 Konkordančnik

Korpus KOST 1.0 smo uvozili v konkordančnika NoSketchEngine (tako v starejši uporabniški vmesnik Bonito kot v novejšega Crystal) in KonText, ki sta del infrastrukture CLARIN. Za vsak korpus smo datoteki z izvornimi in s popravljenimi besedili uvozili ločeno, saj specializiran konkordančnik, ki bi omogočal prikaz nepopravljenih in popravljenih besedil z vsebinskimi oznakami hkrati, v našem okolju še ni bil razvit. Za celostno primerjavo besedil je tako treba odpreti obe verziji besedil (KOST: izvorni in KOST: popravljeni) v ločenih oknih. Postavitev enega okna ob drugem na zaslonu je lahko začasna rešitev za primerjanje podatkov, ne glede na to pa že sama vključitev korpusov v konkordančnike omogoča jasen pregled označenega gradiva in jezikoslovne analize.



Slika 11: Prikaz KOST: izvorni v konkordančniku NoSketchEngine

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.



Slika 12: Prikaz KOST: popravljeni v konkordančniku NoSketchEngine

5 Mojca Stritar Kučuk: Prioritete za nadaljnji razvoj

Korpus KOST želimo v prihodnjih letih **povečati** vsaj za četrtino, torej na vsaj 1.000.000 besed, kar je velikost, pogosta pri tujih korpusih usvajanja. Povečati ga nameravamo predvsem na račun nejužnoslovanskih jezikov. V ta namen bomo okrepili sodelovanje z lektorji slovenščine kot tujega jezika na univerzah v tistih državah, v katerih se slovenščino bolj množično učijo bodisi kot potomci izseljencev (npr. Argentina, ZDA) bodisi zaradi drugih razlogov, kot je sorodnost jezikov (npr. Poljska, Češka, Slovaška). Vzpostaviti želimo tudi redno pridobivanje besedil na izpitih slovenščine v izvedbi Izpitnega centra CSDTJ, predvsem na vstopni in osnovni ravni, s čimer bomo dobili vpogled v slovensko pisno produkcijo nižje izobraženih govorcev. Skupaj želimo zbrati besedila v skupnem obsegu vsaj 209.000 besed, to pomeni od 1400 do 2000 novih besedil (odvisno od njihove dolžine).

Najpomembnejši prihodnji cilj v zvezi s korpusom KOST pa je **povečati delež besedil**, na katerih so označene jezikovne napake, in čim bolj **uravnotežiti označene deleže** med različnimi prvimi jeziki tvorcev. Trenutno so napake označene na 10 % vseh besedil, kar je ustaljen delež v tujih tovrstnih korpusih. Idealno bi bilo, da bi imel vsak označeni podkorpus (npr. podkorpus govorcev srbščine,

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.

makedonščine, nemščine) vsaj 50.000 besed, vendar bomo razmerja prilagajali dejanskemu dotoku besedil. Za doseganje tega cilja pa je še bolj kot dotok novih besedil pomembno označevanje napak v besedilih, ki so že vključena v KOST. To označevanje bomo pospešili tako, da bomo vanj vključili študente slovenistike in druge jezikoslovce. Pred začetkom dela jih bomo ustrezno usposobili, njihove rezultate pa sproti nadzorovali.

Poleg tega želimo **razviti specializirani konkordančnik** za korpus KOST, ki bo omogočal polno izrabo bogato označenega korpusnega gradiva in bo po koncu projekta uporaben za vse korpuse z jezikovnimi popravki (npr. korpus lektorskih popravkov Lektor ali korpus šolskih besedil Šolar). Ta prosto dostopni konkordančnik mora omogočati nazoren, uporabniku prijazen prikaz konkordanc tako v besedilih z označenimi jezikovnimi popravki kot na neoznačenih besedilih korpusa KOST, skupaj z vsemi relevantnimi metapodatki o tvorcih in besedilih, po katerih bo mogoče statistične podatke filtrirati in izvažati. Pomembno je, da uporabnik jasno vidi tako izvorno, tvorčevo besedilo kot tudi popravljeno verzijo tega besedila, pa tudi povezave med obema besediloma skupaj z oznakami vrste jezikovnih napak. Vizualni prikaz tovrstnih povezav je posebno zahteven pri oznakah skladenjskih in drugih napak, ki se nanašajo na večje jezikovne enote, npr. besedne zveze ali stavke. Konkordančnik mora nuditi tudi seznam najpogostejših napak, prek katerega je mogoče enostavno prikazovati konkordance za posamezne napake.

6 Literatura

Arhar Holdt et al. 2022 = Arhar Holdt, Špela, Kosem, Iztok in Stritar Kučuk, Mojca. Metode in orodja za lažjo pripravo korpusov usvajanja jezika. V: Nataša Pirih Svetina, Ina Ferbežar (ur.): *Na stičišču svetov: slovenščina kot drugi in tuji jezik*. *Obdobja 41*. Ljubljana: Založba Univerze v Ljubljani, 2022. https://centerslo.si/simpozij-obdobja/zborniki/obdobja-41/.

Arhar Holdt et al. 2017 = Arhar Holdt, Špela, Kosem, Iztok, Gantar, Polona, 2017: Corpus-based resources for L1 teaching: the case of Slovene. V: Ann Marcus-Quinn, Tríona Hourigan (ur.): *Handbook on digital learning for K-12 schools*. Cham: Springer. 91–113.

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.

Granger 2008 = Granger, Sylviane: Learner corpora. V: Anke Ludeling, Merja Kyto (ur.): *Corpus Linguistics. An International Handbook*. Berlin: Mouton de Gruyter, 2008. 259–275.

Stritar Kučuk 2022 = Stritar Kučuk, Mojca, 2022: KOST med korpusi usvajanja tujega jezika. Obdobja 41: Na stičišču svetov: slovenščina kot drugi in tuji jezik. 323–334. https://centerslo.si/wp-content/uploads/2022/11/Stritar-Kucuk_Obdobja-41.pdf.

Stritar Kučuk 2022 = Stritar Kučuk, Mojca, 2020: Modul Leto plus – prvi korak do korpusa slovenščine kot tujega jezika. Zbornik konference Jezikovne tehnologije in digitalna humanistika 2020. 131–135. http://nl.ijs.si/jtdh20/pdf/JT-DH_2020_StritarKucuk_Modul-Leto-plus%e2%80%93prvi-korak-do-korpusa-slovenscine-kot-tujega-jezika.pdf.

Stritar Kučuk in Šter 2021 = Stritar Kučuk, Mojca in Šter, Helena. 2021. Slovenščina 1+. Ljubljana: Univerza V Ljubljani Filozofska Fakulteta.

Volodina et al. 2019 = Volodina, Elena, Granstedt, Lena, Matsson, Arild, Megyesi, Beáta, Pilán, Ildikó, Prentice, Julia, Rosén, Dan, Rudebeck, Lisa, Schenström, Carl-Johan, Sundberg, Gunlög, Wirén, Mats, 2019: The SweLL Language Learner Corpus: From Design to Annotation. *Northern European Journal of Language Technology* 6. 67–104.

Wirén et al. 2019 = Wirén, Mats, Matsson, Arild, Rosén, Dan, Volodina, Elena, 2019: SVALA: Annotation of Second-Language Learner Text Based on Mostly Automatic Alignment of Parallel Corpora. Inguna Skadina, Maria Eskevich (ur.): *Selected papers from the CLARIN Annual Conference 2018, Pisa, 8–10 October 2018*. Linköping: Linköping University Electronic Press. 227–239.

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.