

Polona Gantar
Univerza v Ljubljani
Filozofska fakulteta in
Fakulteta za računalništvo in informatiko

DOI: 10.4312/jis.68.4.89-108
1.01

Mija Bon
Univerza v Ljubljani
Filozofska fakulteta

Magdalena Gapsa
Univerza v Ljubljani
Fakulteta za računalništvo in informatiko

Špela Arhar Holdt
Univerza v Ljubljani
Filozofska fakulteta in
Fakulteta za računalništvo in informatiko

ŠOLAR-EVAL: EVALVACIJSKA MNOŽICA ZA STROJNO POPRAVLJANJE JEZIKOVNIH NAPAK V SLOVENSKIH BESEDILIH

Razvojni korpusi so temeljni vir za raziskave jezikovnega usvajanja, pripravo didaktičnih gradiv in razvoj orodij, ki podpirajo opismenjevanje in pismenost. Za slovenščino je na voljo razvojni korpus *Šolar*, ki vključuje 5.485 besedil, ki so jih učenke in učenci samostojno tvorili pri pouku, ter 36.570 učiteljskih popravkov. Ti popravki so izjemno dragoceni, ker omogočajo opazovanje avtentične povratne informacije v kontekstu razvoja pisnih zmožnosti. Ker pa niso dosledni in konsistentni, korpus *Šolar* ni uporaben za razvoj in evalvacijo strojnega popravljanja. Zato smo iz gradiva korpusa *Šolar* pripravili evalvacijsko množico *Šolar-Eval*, ki vsebuje 109 šolskih spisov, popravljenih dosledno in konsistentno po maksimi minimalnega jezikovnega posega, čemur se podrobneje posvečamo v prispевku. Metodologija priprave vključuje izbor gradiva, vnos in vsebinsko označevanje popravkov s pomočjo orodja CJVT Svala 1.0 in strojno jezikoslovno označevanje. *Šolar-Eval 1.0* je za raziskovalno in razvojno skupnost na voljo na repozitoriju CLARIN.SI pod odprto licenco.

Ključne besede: jezikovni popravki, korpus *Šolar*, strojno prevajanje, evalvacijska množica, minimalni jezikovni poseg

Šolar-Eval: Evaluation dataset for automated error correction in Slovene texts

Developmental corpora are an important resource for research into language acquisition, for the creation of didactic material and for the development of tools to support literacy. For the Slovenian language, the *Šolar* developmental corpus is available, which contains 5,485 texts produced independently by pupils during lessons and 36,570 corrections by teachers. These corrections are extremely valuable, as they allow the observation of authentic feedback in the context of the development of writing skills. However, as they are not consistent, the *Šolar* corpus is not suitable for the development and evaluation of systems for automated error correction. Therefore, using the material of the *Šolar* corpus we developed the *Šolar-Eval* evaluation set, which contains 109 school essays. The essays were consistently corrected according to the minimal linguistic intervention principle, which is the focus of the paper. The preparation methodology includes data selection, error correction and labelling of corrections with the use of the CJVT Svala 1.0 tool, and automated linguistic tagging of the texts. *Šolar-Eval* 1.0 is available to the research and development community on the CLARIN.SI repository under an open licence.

Keywords: language corrections, *Šolar* corpus, automated correction, evaluation set, minimal linguistic intervention

1 Uvod

Razvojni korpusi (angl. *developmental corpora*, Leech 1997: 19) so premišljeno grajene digitalne zbirke avtentičnih besedil, ki ponujajo vpogled v razvoj jezikovnih kompetenc pri mlajših naravnih govoricah in govorkah določenega jezika.¹ Pisni razvojni korpusi tipično zajemajo primere osnovnošolskega in srednješolskega pisanja, pogosto pa tudi oznake jezikovnih težav, ki se v teh besedilih pojavijo. Ti korpusi so med pomembnejšimi specializiranimi jezikovnimi viri in del temeljne jezikovne infrastrukture, saj predstavljajo empirično osnovo za raziskave s področja jezikovnega usvajanja in didaktike, za pripravo učnih gradiv, vaj, testov, učnih množic za strojno procesiranje naravnega jezika in razvoj orodij, ki opismenjevanje in pismenost podpirajo in razvijajo.

Za slovenščino je na voljo razvojni korpus *Šolar*, ki se razvija že dobro desetletje (Rozman idr. 2012; Kosem idr. 2016; Arhar Holdt in Kosem 2023) in je trenutno dostopen v različici 3.0 (Arhar Holdt idr. 2022a). Korpus *Šolar* je po obsegu in vsebini povsem primerljiv tovrstnim virom za jezike z večjim številom govorcev (pregled slednjih je na voljo v Arhar Holdt in Kosem 2023). Vsebuje 5.485 besedil (večinoma esejev, pa tudi praktično-sporazumevalnih besedil) v obsegu 1.635.407 pojavnic. V korpusu je tudi 36.570 učiteljskih jezikovnih popravkov, ki so vsebinsko kategorizirani v 180 različnih tipov (Arhar Holdt idr. 2022b). Do sedaj je bil korpus uporabljen za pripravo jezikoslovnih raziskav o značilnostih šolskega pisanja (mdr. Kosem idr. 2012; Može 2013; Arhar Holdt in Rozman 2015; Rozman idr. 2018), učnih gradiv (Ahačič 2017a, 2017b; Arhar Holdt idr. 2017), referenčnega seznama pogostih splošnih besed za slovenščino (Arhar Holdt idr. 2020), pa tudi za razvoj slovničnega pregledovanja oziroma

¹ Definicija je nekoliko poenostavljena, saj vemo, da razvoj jezikovnih kompetenc ni prisoten le pri mlajših govoricah in govorkah, prav tako ni povsem natančno govoriti (le) o naravnih govoricah, saj so v osnovnih in srednjih šolah, kjer se besedila za razvojne korpuse tipično zbirajo, tudi učenci/-ke, ki jim jezik okolja ni nujno prvi ali edini.

strojne identifikacije izbranih jezikovnih problemov (Holozan 2013, 2015; Petrič 2022; Mokotar 2023).

Konceptualna posebnost korpusa *Šolar* je, da vsebuje avtentične učiteljske popravke. Od primerljivih tujih jezikovnih virov popravke učiteljev vključuje le korpus *Chyby* (Pala idr. 2003), ki pa se za razliko od *Šolarja* posveča pisanju na univerzitetni ravni. Drugi razvojni korpusi vsebujejo jezikovne popravke, ki so jih vpisali sodelujoči raziskovalci. Odločitev za vključitev učiteljskih popravkov, s pomočjo katerih je mogoče opazovati podajanje povratne informacije v kontekstu razvoja pisnih zmožnosti, je izredno dragocena in predstavlja veliko prednost korpusa *Šolar*. Za določene naloge pa se tovrstni popravki ne obnesejo najbolje. Tipičen primer je uporaba korpusa za učenje in zlasti evalvacijo strojnih postopkov, saj bi morali biti za ta namen jezikovni popravki vneseni dosledno in konsistentno, kar pa za učiteljsko popravljanje ne velja.²

Na manjšo uporabnost korpusa *Šolar* za evalvacijo strojnega popravljanja je bilo v našem prostoru že opozorjeno (npr. Holozan 2013, 2015; Kranjc in Robnik Šikonja 2015). Dodaten problem predstavlja dejstvo, da tudi drugi slovenski korpusi s sicer dosledneje pripisanimi jezikovnimi popravki za ta namen niso uporabni: korpus lektoriranih besedil *Lektor* (Popič 2014) za razliko od *Šolarja* prinaša robustno označene popravke, v množici katerih je težko najti jezikovni problem, katerega strojno prepoznavo želimo evalvirati; ker vsebuje besedila izkušenih piscev, je določenih napak manj, kot bi jih za evalvacijo potrebovali. Podobno velja za *KOST* (Stritar Kučuk 2022), ki prinaša dokaj robustno označena besedila avtorjev, ki se slovenščine učijo kot drugega ali tujega jezika.

Ker je natančna in zanesljiva evalvacija ključna za razvoj strojnega popravljanja in sorodnih jezikovnih tehnologij za slovenščino, smo iz gradiva korpus *Šolar* pripravili evalvacijsko množico Šolar-Eval, ki jo bo v prihodnje mogoče uporabljati za ta namen. V prispevku opišemo metodologijo priprave, pri čemer se posebej posvetimo vprašanju zagotavljanja doslednosti in konsistentnosti jezikovnih popravkov. Prispevek sklenemo z opisom rezultata in napovedjo nadaljnjega dela.

2 Metodologija

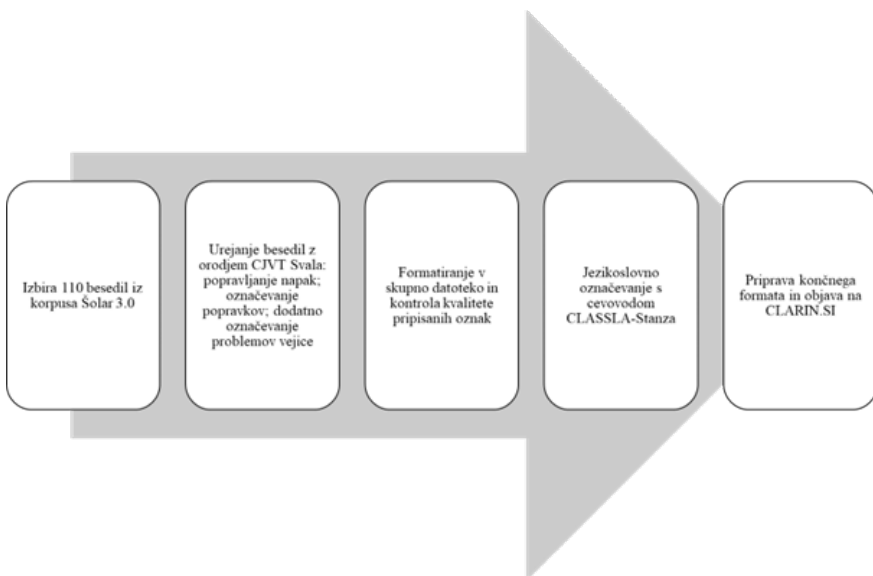
Vodilo pri pripravi evalvacijske množice je bilo, da mora vsebovati avtentična besedila, v njih pa raznolike, dobro zastopane, dosledno obravnavane in natančno

² Učiteljski popravki so prilagojeni razvojni stopnji učenca/-ke, včasih se popravek pojavi le pri prvi napaki določene vrste (z napotilom učenca/-ki, naj ostale težave poišče in popravi sam/-a), popravek lahko le grafično nakaže, da je določen del besedila treba spremeniti, pogosti pa so tudi popravki, kjer učiteljica ali učitelj namesto popravljanja posameznih napak koreniteje poseže v besedilo – spremeni ubeseditev tako, da obenem izboljša slog, odpravi vsebinsko napako in podobno (o tem tudi Tinta 2019). Posamezni popravki, zlasti na ravni skladnje in besedilne koherence v povezavi s slogovno (ne)ustreznostjo besedila, pogosto zahtevajo poseg na ravni obsežnejšega, največkrat celotnega besedila (npr. usklajevanje glagolskega časa). Nenazadnje na popravljanje vpliva tudi časovna obremenjenost učiteljev, zaradi katere se včasih osredotočajo le na izbrane jezikovne težave.

kategorizirane jezikovne težave oziroma popravke. Besedila smo izbrali iz korpusa *Šolar* 3.0 s pomočjo informacij, ki so na voljo v formatiranem frekvenčnem korpusnem izvozu, dostopnem na repozitoriju CLARIN.SI (Arhar Holdt idr. 2022c).

Za besedila, v katerih se pojavlja največ učiteljskih popravkov, smo preverili skupno pogostnost različnih vrst popravkov. Želeli smo, da se vsaka od 180 možnih kategorij pojavi vsaj 5-krat, z izjemo: oznak za nečitljiva mesta, ki smo jih v evalvacijski množici želeli čim manj; dodatnih (neobveznih) oznak za vsebinske napake in jezikovne variante; oznak za vrsto napačno stavljene vejice, ki so v korpusu *Šolar* 3.0 pripisane le sporadično (Arhar Holdt idr. 2022b). Na osnovi te analize smo izbrali 110 spisov, ki skupno pokrivajo 19,7 % vseh popravkov v korpusu *Šolar* 3.0. Eden od spisov se je kasneje zaradi težav s formatiranjem izkazal za problematičnega, zato smo ga iz obravnave izločili.

Urejanje popravkov je potekalo z orodjem CJVT Svala 1.0,³ ki omogoča pregledno sopostavitev izvornega in popravljenega besedila, povezovanje med obema ter vsebinsko označevanje vrste vsakega posameznega popravka. Pri pripravi je sodelovala ekipa petih jezikoslovcev. Delo je potekalo v treh zaporednih korakih: (a) vnos jezikovnih popravkov skladno s smernicami, (b) vsebinska kategorizacija popravkov po sistemu korpusa *Šolar* in (c) vsebinsko natančnejše označevanje problemov vejice. Po označevanju smo datoteke JSON, ki jih pridobimo iz orodja Svala, umestili v skupno datoteko, kjer smo izvedli kontrolo kvalitete, dodali metapodatke in besedila jezikoslovno označili z označevalnikom CLASSLA-Stanza (Terčon in Ljubešić 2023). Evalvacijska množica je na voljo na repozitoriju CLARIN.SI⁴ (Arhar Holdt idr. 2023a).



Slika 1: Postopek priprave evalvacijske množice *Šolar-Eval*

³ Orodje je dostopno in opisano na strani: <https://orodja.cjvt.si/svala/> (dostop 30. 11. 2023).

⁴ Povezava: <http://hdl.handle.net/11356/1902> (dostop 30. 11. 2023).

Delotoki, ki jih prikazuje slika 1, so skladni s trenutnimi standardi za pripravo jezikovnih virov za sodobno slovenščino in dobro dokumentirani, zato se jim v tem prispevku ne posvečamo natančneje. Metodološko zanimivejši je poskus konsistentnega jezikovnega popravljanja. Na pravopisni ravni, kjer je relativno jasno, kaj jezikovnega je (ne)skladno s trenutno normo, je manj težav, medtem ko so na višjih ravneh, kot je izbira besedišča, zlasti pa skladnja in besedilna koherenca, možne različne interpretacije učenčevega oziroma učenkinega pisanja, kot tudi različne rešitve za odpravo identificiranega jezikovnega problema (gl. tudi Tinta 2019).

Smernice za popravljanje smo oblikovali na štirih temeljnih raziskovalnih izhodiščih: **1. Opredelitev problema:** presoja, katere tipe jezikovnih težav je smiselno popravljati in katerih ne. **2. Pripisljivost vsebinske oznake:** preverba, ali je popravek mogoče vsebinsko kategorizirati z oznakami korpusa *Šolar*. **3. Aktualnosti specifičnih popravkov:** zagotavljanje skladnosti z aktualno jezikovno rabo in normo, npr. s preverjanjem gradiva v korpusu *Gigafida 2.0* (Krek idr. 2020), *Kolokacijskem slovarju sodobne slovenščine* (Kosem idr. 2023), *Slovarju sopomenk sodobne slovenščine* (Arhar Holdt idr. 2023) in v jezikovnih ter v pravopisnem priročniku na portalu Fran (Fran, slovarji Inštituta za slovenski jezik Frana Ramovša ZRC SAZU 2014–). **4. Korekcija doslednosti:** usklajevanje in preverjanje odločitev, tudi v smislu odstranjevanja že obstoječih popravkov, če jih ni mogoče enotno interpretirati in aplicirati na druga besedila.

Pri popravljanju smo sledili **maksimi minimalnega posega**. Pod tem izrazom razumemo dvoje: (a) v besedilo ne posegamo, če je učenčevo oziroma učenkino rešitev mogoče interpretirati tudi kot (manj) ustrezno varianto, in (b) jezikovno napako popravimo tako, da je poseg v smislu spremembe strukture stavka in z napako povezanih elementov v sobesedilu čim manjši.

3 Vnašanje popravkov v besedila množice Šolar-Eval

V nadaljevanju opišemo smernice za opredeljevanje jezikovnih popravkov pri izdelavi evalvacijske množice in konkretne jezikoslovne premisleke ter odločitve pri posameznih jezikovnih problemih. Osredotočamo se na jezikovne posege na višjih ravneh,⁵ in sicer jezikovne popravke obravnavamo v treh sklopih: 1. pomen-skem oziroma vsebinskem, kjer obravnavamo napake na ravni besedišča, večbesednih enot, pa tudi napake, ki se nanašajo na razumevanje vsebinskih dejstev; 2. slovnično-skladenjskem, kjer obravnavamo napačno rabo oblikoslovnih elementov, kot so zaimki in predlogi, na ravni skladnje pa vezljivost, besedni red in zapis premega govora; v 3. sklopu obravnavamo širok nabor t. i. slogovnih popravkov, ki zajemajo tako raven besedišča kot skladnje in pogosto segajo na področje širšega besedila.

⁵ Pri popravkih zapisa (npr. izbira velike/male začetnice, zapis skupaj/narazen), črkovanja in oblikoslovja smo sledili podatkom iz referenčnih virov in priročnikom ter primerom v Arhar Holdt idr. 2022b, popravke vejic pa natančneje obravnava prispevek Bon in Gapsa (v pripravi).

Zgledi, ki jih navajamo za pojasnjevanje popravkov, so vzeti iz korpusnih besedil in skrajšani do te mere, da še ustrezno ponazarjajo obravnavani problem. Izvirno besedilo, na katerega se nanaša popravek, in jezikovni popravki so v zgledih poudarjeni, popravki so navedeni za pokončnico. V zgledih z več napakami ponazarjamo zgolj tiste, ki so predmet obravnave, preostalih napak ne izpostavljamo oziroma puščamo izvirno besedilo.

3.1 Popravki na ravni besedišča

Popravki na ravni besedišča se tipično nanašajo na pomensko ali zvrstno neustrezno rabo besede, besedne zveze, zlasti kolokacije, napačno rabo frazeoloških enot, formulaičnih izrazov, zvez z glagoli v pomensko oslavljeni rabi ipd. Sem sodijo tudi popravki, ki se nanašajo na preverljiva vsebinska dejstva.

3.1.1 Raba pomensko ali zvrstno neustrezne besede

Na ravni besedišča smo v prvi vrsti popravljali napačno rabo besede glede na pomen, ki ga ima v povedi, in glede na izbiro zvrsti ali besedne zaznamovanosti. Pomensko neustrezno besedo smo nadomestili po principu minimalnega posega tako, da smo poiskali besednovrstno primerljiv, a pomensko ustreznejši nadomestek. V primeru odločanja med več možnostmi smo se posvetovali in medsebojno usklajevali.

- *Punca zanosi in noče povedati Adamu, ker je vedela, da bi se **odklonil**|**odpo-vedal** njej in otroku*
- *Antigona se je odločila pokopati svojega brata, zaradi predvsem božjih zakonov samo mislim, da v resnici ni **morala**|**mogla** gledati, kako ni bil pokopan.*

Za jezikovni poseg smo se odločali tudi, kadar se je v besedilu pojavljalo zaznamovano ali zvrstno neustrezno besedišče (npr. *odprla luč, runkl, koštalo, itak, prišparal, probal, za brez veze, bolj fajn, zanga, tle*). Odločitev za popravek smo v takih primerih podprli s preverbo rabe v korpusu *Gigafida* 2.0, potencialnim obstojem nevtralne ustreznice, preverjanjem stilističnih in drugih oznak v slovarskih priročnikih ter z upoštevanjem siceršnje jezikovne kompetence v širšem besedilu učenca/-ke. Zaznamovano besedišče smo ohranili, kadar se je pojavilo v narekovajih oziroma premem govoru, npr. »*Ma, to se zgodi vsakemu.*« (gl. tudi 2.3.6).

3.1.2 Večbesedne enote: kolokacijskost in frazeološkost

Kot pogost problem obravnavanih besedil se je pokazala neustrezna ali netipična raba besednih zvez, zlasti kolokacij in frazeoloških enot. Med temi primeri je bilo precej takšnih, ki jih v korpusu *Šolar* 3.0 učitelji niso popravljali, a jih je mogoče

s pomočjo sodobnih slovarskih priročnikov in korpusov prepoznati in nadomestiti z jezikovno rabo, ki je tipična, deluje naravno ter je specifična za slovenščino. Spodnji primeri ponazarjajo problem netipične rabe besed glede na kolokacijskost in besednozvezno ustaljenost:

- *Čas|Svet se je dandanes po eni strani zelo spremenil, po drugi strani pa še vedno ostaja enak*
- *ta ljubezen ne bi propadla na tragični konec|doživela tragičnega konca*

V prvem primeru gre za sovpad zvez *časi se spreminjajo* in *svet se je spremenil*, pri drugem pa za kolokacijo *tragični konec*, ki se tipično pojavlja z glagolom *doživeti*. Popravke smo izbirali tako, da je bil poseg v besedilo čim manjši, čeprav bi bilo z večjim posegom mogoče doseči enako ali celo višjo tipičnost (npr. *ta ljubezen je tragično propadla*, ki je v slovenščini prav tako pogosta kolokacija).

Podobno smo obravnavali frazeološke enote, kar nakazujejo primeri spodaj. V prvem primeru gre za križanje zvez *vzeti kaj v svoje roke* in *(pre)vzeti kaj nase*, v drugem pa za manj ustrezno izbiro glagola *skočiti* v frazeološki enoti *kaj pade komu v oči*. Ta frazem se tipično uporablja v pomenu, da kdo nekaj opazi zato, ker izstopa, ker je nekaj posebnega. Varianta z glagolom *skočiti* je v rabi veliko redkejša. Ob tem se zdi, da je avtor/-ica s prislovom *naenkrat* in z izbiro glagola *skočiti*, ki vsebuje semantično komponento trenutnosti dejanja, želel/-a izpostaviti predvsem ‚nenadnost‘ in ‚nepričakovanost‘ dejanja, kar je posledično vzrok za izbiro manj ustreznega glagola v frazeološki enoti.

- *Hamlet je probleme prevzel|vzel v svoje roke*
- *Odisej se premišljeval, se sprehajal po barki, kar naenkrat mu je v oči skočila|padla jama.*
- *brez kaj reči|besed*

Pri zvezah s pomensko oslabljenimi glagoli (*imeti*, *dati* ipd. + samostalnik)⁶ smo se za popravke odločali na podlagi pomenske interpretacije. Tipičen primer je denimo *Tudi Kreon vztraja pri svojem, da mora Antigona dobiti kazen|biti kaznovana za svojo neubogljivost*, kjer je popravek utemeljen s tem, da je zvezo *dobiti kazen* mogoče interpretirati tudi kot ‚dobiti nalog za plačilo kazni‘. Enako tudi v primeru *Urednik slov. lista Čupa mu je dal predlog|predlagal, naj piše o asimilaciji Slovencev v Trstu*.

3.1.3 Vsebinski popravki

V šolskih spisih se poleg jezikovnih pojavljajo tudi vsebinske napake, ki zadevajo faktične podatke, pa tudi besedilno koherenco in ustrezno referiranje na elemente znotraj besedila. Tipična primera sta nerazumevanje razmerij med literarnimi

⁶ Za več primerov gl. Gantar idr. 2018.

osebami (npr. oče – stric, mama – žena, sin – brat) in neustrezna poimenovanja literarnih oseb. Ker se pri tovrstnih zamenjavah poruši tudi notranja koherenca zaradi napačno vzpostavljenih referenc, smo v teh primerih besedilo popravljali, npr.

- *To kar je storil je skušal popraviti s poroko **Jurija**|**Izidorja** in Agate.*
- *Kreona, Antigoninega **očeta**|**strica**, pa težko uvrstimo pod strogo držljive ali prilagodljive ljudi.*

Na drugi strani nismo popravljali vsebinskih napak, ki so posledica morebitnega napačnega razumevanja obravnavane učne teme ali z njo povezanih dejstev. Npr. primera *Šolal se je v Pesnici za učitelja slovenščine in angleščine* za Šolar-Eval nismo popravili, za razliko od korpusa *Šolar* 3.0, kjer je bil učiteljski popravek *Šolal se je v Pesnici, nato je študiral v Mariboru in postal učitelj slovenščine in angleščine*.

3.2 Popravki na slovnični in skladenjski ravni

V nabor slovničnih popravkov sodi označevanje napačne rabe povratnosvojljnih zaimkov in povratnih glagolov ter raba predlogov v predložnih zvezah. Znotraj skladnje obravnavamo še napake na ravni stavčnih razmerij, vezljivosti, besednega reda in zapisa premega govora.

3.2.1 Povratnosvojljni zaimki in povratni glagoli

V besedilih smo popravljali napačno rabo povratnosvojljnih zaimkov, zlasti zamenjevanje z osebnosvojljnimi (*svoj* vs. *moj*, *njegov* ipd.), in povratnosvojljnih glagolov, glede katerih smo stanje preverjali v slovarskih in pravopisnem priročniku na portalu Fran ter v korpusu *Gigafida*.

- *Ampak še več nadutosti bi naj imela, da **se mu nazaj ugovarja**|**mu ugovarja***
- *Med izvidnico zagleda nemskega vojaka in lepa mlado žensko Katarino, ki **se ljubimka z njim**|**z njim ljubimka***
- *z Antonom sta **si podobnih misli**|**podobnih misli***

Glagol *ugovarjati* tipično zahteva dajalniški predmet, tj. *ugovarjati komu/čemu*, za slovenščino neustrezna pa je raba povratnosvojljnega zaimka ob njem, tj. **ugovarjati se*. Na podlagi *Slovarja sopomenk sodobne slovenščine* (Arhar Holdt idr. 2023) je mogoče predvideti ustrezne povratnosvojljne sopomenke kot npr. *upreti se in zoperstaviti se*, pri čemer bi zamenjava z ustrežnejšo sopomenko pomenila korenitejši poseg v besedilo. Hkrati je mogoče predvidevati, da je prišlo do kontaminacije z zvezo *nazaj odgovarjati* v pomenu 'ugovarjati', ki je na slovarskem portalu Fran v *Prekmurskem slovarju* označena kot narečna, v SSKJ2 pa kot nižje pogovorna. Najustreznejši popravek v smislu minimalne intervencije je torej le ohranitev glagola *ugovarjati*.

Glagol *ljubimkati* predvideva orodniško predložno dopolnilo, ne pa tudi povratnega zaimka. Predvidevati je sicer mogoče, da je učenec/-ka imel/-a v mislih glagol *poljubljati*, ki je glede na kontekst pomensko ustrežnejši, zaimek *se* ob njem pa nakazuje vzajemnost, vendar je pomensko ustrezen in minimalen poseg izpust povratnega zaimka. Enako velja za zadnji primer, kjer je opaziti kontaminacijo zveze *biti podoben/istih/enakih misli* in *biti si podoben*.

3.2.2 Predlogi

Med oblikoskladenjskimi problemi smo obravnavali tudi izbiro predlogov, kjer smo v korpusu in slovarskih virih preverjali rabo predložnih glagolskih in samostalniških zvez glede na tipičnost vezljivostnega vzorca in pomensko ustreznost, npr.

- *nekaj v njemu gnalo na takšno odločitev*|*k takšni odločitvi*
- *pripoveduje o odnosu med staršema iz otroških*|*skozi otroške oči*
- *veliko ljudi polnih s hinavščino, lažmi in prevarami* |*hinavščine, laži in prevar*

V slovenščini uporabljamo glagol *gnati* tudi za izražanje potrebe po fizični ali miselni aktivnosti, npr. *radovednost ga žene*, kot ga je uporabil/-a tudi učenec/-ka v prvem primeru, vendar pa se s predlogom *na* tipično veže v povsem drugih pomenih, npr. *gnati na vodo*; *gnati na pašo*. Poleg tega se tudi samostalnik *odločitev*, kadar se pojavlja s pomensko sorodnimi glagoli, kot npr. *spodbuditi*, *napeljati* in *prisiliti*, tipično veže s predlogom *k* in ne *na*.

3.2.3 Vezljivost

Dosledno smo obravnavali primere, kjer je učenec/-ka izpustil/-a določilo, ki ga zahteva glagol, npr.

- *Ko* |*ga* *Antigona* *že drugič hoče pokopat jo pri delu dobijo stražari*
- *Takrat mu je malo žal, vendar se sprijazni* |*s tem* *in želi vsaj to, da*

3.2.4 Besedni red

Znotraj besednorednih popravkov smo obravnavali raznovrstne probleme: besedni red znotraj naslonskega niza ali besedne zveze, stavčno sosledje znotraj povedi in členitev po aktualnosti. Pozorni smo bili predvsem na konstrukcije, kjer atipični besedni red vodi v neustrezen pomen, spominja na pogovornost ali tujejezično interferenco. Skladno z maksimo minimalnega posega smo besedni red popravljali le v primerih širšega konsenza v označevalni ekipi, sicer smo pustili izvorno različico.

- *Antigona se je odločila pokopati svojega brata, zaradi predvsem|predvsem zaradi božjih zakonov*
- *Kandor glavna oseba Sencnega plesa je pravkar pokopal svojo ženo|, ki jo je zelo ljubil, na pokopališču na gorenskem, ~~ki jo je zelo ljubil~~*
- *Oba Kreon in Antigona sta prinesla tragičen konec, saj ostati sam brez družine je težko|saj je težko ostati sam brez družine*

3.2.5 Združevanje in razdruževanje povedi

Pogosta težava šolskih spisov so zelo dolge povedi, v katerih se spričo zapletene stavčne zgradbe pojavljajo tudi različne skladijske težave. Glede na smernice smo za potencialno razdruževanje oziroma krajsanje obravnavali povedi, dolge tri ali več stavkov, pri tem pa smo upoštevali razumljivost povedi, vrsto razmerja (priredja/podredja), ustreznost veznikov glede na ponavljajoča se pri- ali podredja in morebitne slogovne posebnosti, ki jih je učenec/-ka uporabil/-a namensko, da bi ustvaril/-a živo pripoved. Nekoliko redkeje smo se odločali za združevanje povedi, kar je bilo v večini primerov povezano tudi z odpravljanjem ponavljanja, kot kaže zadnji primer spodaj.

- *V spopadu Laert zabode Hamleta z zastrupljenim mečem, a med nadaljnjim bojem si meč nenamerno zamenjata in tudi Hamlet zabode Laerta ~~in tako~~. Tako sta bila oba obsojena na propad, v|. V tem trenutku je Laert priznal svojo zmoto, da ni razmišljal s svojo glavo; in priznal, da je vse to načrtoval Klavdij.*
- *Najprej bomo pogledali Antigono. ~~Antigona~~, ki se trdno drži svojih načel, neglede na situacijo v kateri je*

3.2.6 Premi govor in dobesedni navedki

V besedilih smo skušali dosledno obravnavati zapis premega govora in dobesednih navedkov, pri čemer smo se v primeru neustrezne ali neenotne rabe narekovajev navadno odločali za preoblikovanje v nepremi govor. Rabo narekovajev za zaznamovano besedišče, kar učitelji v korpusu Šolar 3.0 pogosto popravljajo, smo po principu minimalnega posega pustili.

- *Ampak preden rečemo **ne** bom se suočil z njim |, da se ne bomo soočili z njim, ga moramo dobro analizirati.*
- *Materi pa z »ostrimi« besedami zada globoko rano*

3.3 Slogovni popravki

Kot slogovno neustrezne dele besedila smo obravnavali različne jezikovne ubeseditve, ki zadevajo tako raven besedišča, oblikoslovja, kot tudi raven besednozvezne

in besedilne skladnje ter žanrske in zvrstne problematike. Določitev obsega jezikovne intervencije na slogovni ravni besedila je bila z vidika izdelave evalvacijske množice najtrši oreh, saj se popravki na besedilni ravni nanašajo na celostno razumevanje besedila, besedilno koherenco, referiranje ipd. in posledično segajo na več označevalnih ravni.

3.3.1 Popravki na ravni stavčnih odvisnikov

Sem sodijo popravki na ravni stavčnih odvisnikov, posledično pa tudi rabe veznikov in izbire besedišča, kot kažejo spodnji primeri. Skladno s smernicami smo popravke v odstavkih zato dodajali postopoma od najmanj interpretativnih do bolj interpretativnih, pri slednjih pa smo minimalni poseg iskali v konsenzu celotne označevalne ekipe.

- *Kreon*-kljub temu, | da mu videc Terezias pove, da bo ostal sam in nesrečen, če ubije Antigono, **a**ga Kreon kot **na nek način|neke vrste diktator to ne posluša|ne posluša** in še vedno **vztraja pri tem|vztraja**, da mora biti **Antigona obsojena|obtožena** na smrt.
- Nič je ni bolelo **le, kot|bolj kot to**, da ni mogla pokopati brata
- Takšno obnašanje oziroma dejanje je prisotno še v današnjem življenju, |, kajti vsak človek, **ko mu je neka sila|ki je v stiski, da gre|ko mora** v zapor, dobi denarno kazen **in |ali** kaj podobnega |, bo vedno pripravljen izdati tudi svoje najbližje.

3.3.2 Menjavanje glagolskega časa

Po principu minimalne intervencije v besedila nismo posegali v primeru jezikovnih sredstev in skladenjskih struktur, za katere smo predvidevali, da jih je učenec/-ka uporabil/-a z namenom ustvariti določen slogovni učinek, govorjenost/diskurznost, dramatičnost ali literarnost ipd., kamor sodi zlasti uporaba diskurzivnih in nagovornih elementov, menjavanje časov, pa tudi uporaba neformalnih in metaforičnih izrazov, menjavanje žanra/registra ipd.

- *Ravnatelj Polde mu je oprostil, saj je vedel, da se je zlagal. Martin se mu zahvali in se spet pošali* (menjava preteklik – sedanjik brez popravka)
- *Čeprav je zaljubljen vanjo in ga je prevzela, se odloči da izpolni|bo izpolnil ukaz* (sedanjik namesto prihodnjika)

3.3.3 Raba pogovornih diskurzivnih označevalcev

V šolskih spisih se pojavljajo diskurzivni označevalci, ki so značilni za govorjeni jezik in zato lahko v standardnem pisnem jeziku delujejo pogovorno in s tem slogovno

neustrezno. Takšne primere smo obravnavali vsakega posebej in skupno presodili, kdaj bi bil popravek skladen z maksimo minimalnega posega. V rabo členkovnih izrazov, kot je beseda *no* v spodnjih dveh primerih, denimo nismo posegali, če je šlo za pomensko modifikacijo povedi, v spodnjem zgledu npr. za izražanje relativnosti:

- *No, pa za|Za začetek povejmo kaj o delu Romeo in Julija*
- *Ampak tudi moja bratranca imata podobno zgodbo. No, ne moremo reči podobno, ker so bile razmere nekdanj bistveno drugačne od današnjih.* (brez popravka)

3.3.4 Prekomerna ali neustrezna raba zaimkov

Poleg tipičnih težav z izbiro zaimkov (npr. *ves - cel, enak - isti, on - sam*) je v šolskem pisanju opazna prekomerna raba kazalnih in svojilnih zaimkov. V takih primerih smo ohranjali povezave na elemente predhodnega besedila, kjer je bilo to potrebno zaradi razumljivosti. Na drugi strani smo zaimke odstranjevali, če za razumljivost niso bili potrebni oz. če je njihova prekomerna raba razumljivost poslabšala.

- *S tem mislim na ~~njeno~~-bistvo ~~te~~ zgodbe, ki pa je ~~to~~, da je človek v življenju velikokrat postavljen na križišče*
- *To je ponavadi glavni povod za maščevanje in ko se zavemo, kaj smo naredili narobe, je večkrat ~~to~~ že prepozno*
- *~~To~~-Prepričanje o trajnem miru duše je zanjo vzrok, ki je „kriv“ ~~za to~~, da se od tu naprej drži svojih načel*
- *Martin Krpan hitro ustavi ~~svoj~~-avto in steče ven*
- *V njem se pojavi razkol med ~~njegovimi~~-čustvi do Katarine in ~~njegovo~~-nalogo |, ki jo ima v vlogi partizana in partizanskih ciljev*

3.3.5 Ponavljanje

Znotraj slogovnih popravkov smo obravnavali tudi različne vrste ponavljanja, ki pogosto vplivajo tudi na popravke na leksikalni in skladenjski ravni pa tudi na ravni odstavka oziroma celotnega besedila. Navajamo nekaj tipičnih primerov obravnave ponovljenih polnopomenskih besed ali zvez, kopičenja veznikov in parafraz že ubesedenega.

- *tisti teden je bil zelo zanimiv ~~teden~~.*
- *njuna ljubezen nima meja in še vedno vztrajata pri svojih čustvih, ~~ki jih čutita~~*
- *Ko ~~pa~~ dobijo kazen za to, ~~pa~~ takoj odnehajo*
- *Oba čutita ljubezen drug do drugega, ampak ju ta verski spopad |ločuje drug od drugega | med seboj ločuje.*
- *Če se vrnem na svoje razmišljanje v prejšnjem odstavku, pa ne mislim, da so absurdna dejanja v današnji družbi posledica brezčutnosti in nesmisla življenja. ~~Če se vrnem na|Na primer morilca, se~~ dandanes |se velikokrat dogajajo umori predvsem zaradi denarja.*

3.3.6 Nejasni ali nedokončani fragmenti

V redkih primerih se v besedilih pojavijo tudi nejasni ali nedokončani fragmenti, pri katerih ni mogoče z dovoljšno prepričanostjo razbrati pomena, ki ga je želel/-a ubesediti učenec/-ka. Takšne fragmente smo iz popravljenega besedila izpustili, dodatno pa smo jih označili z vsebinsko oznako »nejasno«.

- *Kreona, Antigoninega očeta, pa težko uvrstimo ~~pod~~ med strogo drgljive ali prilagodljive ljudi.*
- *Kreon misli, da je trma in da bo precej hitro zlomila. ~~Ponos ne daje, ki je drugim hlapec.~~ In Antigona bi mu naj pokazala nadutost s kršenjem njenega vega ukaza. Ampak še več nadutosti bi naj imela, da se mu nazaj ugovarja.*

4 Evalvacijska množica Šolar-Eval

Šolar-Eval 1.0 vsebuje 109 spisov iz korpusa Šolar 3.0. Kot je razvidno iz tabele 1, je v popravljeni različici več pojavníc in povedi, kar je posledica jezikovnih popravkov: deljenja predolghí povedi na krajše ter (po večini) dodajanja manjkajočih vejíc.

	Izborna besedila	Popravljená besedila
Število odstavkov	740	740
Število povedi	4.017	4.281
Število pojavníc	69.658	70.226

Tabela 1: Obseg evalvacijske množice Šolar-Eval (izborna in popravljená besedila)

Zgradbo evalvacijske množice in število pojavníc v izvornih besedilih glede na vrsto šole in razred predstavlja tabela 2.

	Gimnazije	Poklicne šole	Strokovne šole	Osnovne šole	Skupaj
1. letnik	22.737	626	5.507		28.870
2. letnik	11.807		1.212		13.019
3. letnik	7.556	5.740	369		13.665
4. letnik	6.778		1.050		7.828
6. razred				757	757
7. razred				3.120	3.120
8. razred				2.270	2.270
9. razred				129	129
Skupaj	48.878	6.366	8.138	6.276	69.658

Tabela 2: Število pojavníc v evalvacijski množici Šolar-Eval glede na vrsto šole ter razred/letnik

V besedila smo dodali 9.808 vsebinsko kategoriziranih jezikovnih popravkov. Največji delež predstavljajo popravki na ravni zapisa, kamor sodijo tudi številni popravki ločil (3.136 popravkov oziroma 31,97 %). Prav tako so pogosti popravki skladenjskih struktur (2.660 oziroma 27,12 %). Sledijo popravki besedišča (1.150 oziroma 11,73 %), oblike (961 oziroma 9,80 %) in črkovanja (958 oziroma 9,77 %). Posebna kategorija so t. i. povezani popravki, ki so posledica kakega drugega popravka, npr. popravek začetnice, ki je posledica deljenja daljše povedi na dve krajši (932 oziroma 9,50 %). Manj je popravkov, ki nakazujejo nečitljiva ali nerazumljiva mesta (11 oziroma 0,11 %). Tabela 3 predstavlja natančnejše podatke o frekvenčni zastopanosti raznovrstnih popravkov, kar prikazujejo številke v oklepajih pri posamezni kategoriji (za razlago kategorij gl. Arhar Holdt idr. 2022b).

Prvi nivo oznake	Drugi nivo oznake	Koda oznake
Zapis (3.136) Koda: Z	Ločila (2.664) Koda: LOČ	vzorec-vejica-stavki (1.181), vzorec-vejica-priredja-odvisnikov (471), vzorec-vejica-vrinjen-odvisnik (294), vzorec-vejica-stavčni-členi (173), nerazvrščeno (159), vzorec-vejica-pristavki (136), vzorec-vejica-priredja-zvez (82), vzorec-vejica-vezniki (55), vzorec-vejica-elipsa-povedka (45), vzorec-vejica-kopičenje-veznikov (24), vzorec-vejica-kot (24), vzorec-vejica-kopičenje-ločil (14), vzorec-vejica-pridevniški-niz (4), vzorec-vejica-navajanje (2)
	Mala/velika začetnica (254) Koda: MV	začetek-povedi (52), osebna-imena (43), stvarna-imena (31), narodnost (26), občna-imena (25), pridevnik-ski (19), drugo (18), zemljepisna-imena (16), premi-govor (14), hiperkorekcija-ločila (8), pridevnik-drugo (2)
	Skupaj/narazen (211) Koda: SN	skupaj-glasol (96), skupaj-predlog (45), skupaj-prislov (15), narazen-predlog (15), narazen-prislov (13), narazen-drugo (10), skupaj-drugo (9), narazen-pridevnik (8)
	Krajšave (4) Koda: KR	drugo (4)
	Števila (3) Koda: ŠTEV	drugo (3)
Skladnja (2.660) Koda: S	Odvečni jezikovni elementi (1.124) Koda: ODVEČ	zaimek-kazalni (117), ponavljanje (109), zaimek-svojilni (100), zaimek-osebni (81), veznik-pa-drugo (76), členek (65), predlog (63), samostalnik-občno-ime (61), prislov-drugo (58), samostalnik-lastno-ime (54), stavek (49), veznik-drugo (46), zaimek-drugo (43), veznik-začetek (38), glasol-biti (38), pridevnik (34), veznik-pa-vezniki (29), prislov-mera (25), veznik-dvojni (25), glasol-drugo (8), poved (5)
	Struktura (575) Koda: STR	deljenje-stavkov (347), združevanje-stavkov (74), preoblikovanje-stavka (52), ločilo-veznik (42), besedna-zveza-stavek (30), svojina-rodilnik (18), svojina-od (12)

Prvi nivo oznake	Drugi nivo oznake	Koda oznake
	Besedni red (468) Koda: BR	povedek-prislovno-določilo (103), povedek-predmet (81), členek (76), povedek-osebek (63), drugo (50), naslonski-niz-znotraj (47), znotraj-stavčnega-člena (37), naslonski-niz-prirednost-podrednost (11)
	Izpuščeni jezikovni elementi (376) Koda: IZPUST	predlog-drugo (55), glagol-bit (49), zaimek-osebni (46), samostalnik-občno-ime (43), zaimek-drugo (39), veznik-drugo (35), prislov (19), členek (18), samostalnik-lastno-ime (18), glagol-drugo (18), veznik-pa (11), predlog-povnovljen (10), pridevnik (9), stavek (6)
	Dodatne oznake (117) Koda: DOD	vsebina-drugo (96), vsebina-napake (19), pomensko-prazni (2)
Besedišče (1.150) Koda: B	Menjava prek meja besedne vrste (299) Koda: MEN	polnopomenska-v-zaimek (111), drugo (72), zaimek-v-polnopomensko (33), prislov-pridevnik-bz (24), besedna-družina (20), veznik-zaimek (19), glagol-bz (13), samostalnik-bz (7)
	Glagol (193) Koda: GLAG	drugo (118), predpona (46), moči-morati (18), naklonski (11)
	Veznik (160) Koda: VEZ	drugo (87), sprememba-odnosa (37), protivni (23), in-pa-ter (13)
	Zaimek (132) Koda: ZAIM	ki-kateri (38), drugo (33), povratna-svojlilnost (26), oziralni (20), noben (15)
	Samostalnik (115) Koda: SAM	občno-besedišče (84), napačno-lastno (25), lastno-občno (6)
	Predlog (110) Koda: PRED	glagolske-zveze (45), neglagolske-zveze (35), drugo (16), lokacijske-dvojnice (14)
	Prislov (46) Koda: PRISL	drugo (46)
	Dodatne oznake (45) Koda: DOD	zaznamovano (45)
	Pridevnik (40) Koda: PRID	drugo (40)
	Ostale besedne vrste (10) Koda: OST	drugo (10)
Oblika (961) Koda: O	Kategorialni popravki (833) Koda: KAT	sklon-drugo (114), število-em (110), število-dm (94), čas (80), spol (73), sklon-rt (64), nedoločnik-kratki (46), število-ed (43), vid (33), povratnost (32), stopnjevanje (28), nedoločnik-namenilnik (22), oblika-zaimka (18), naklon (18), določnost (17), sklon-mo (12), sklon-dm (11), nedoločnik-osebna (9), oseba (7), način (2)
	Paradigmatski popravki (98) Koda: PAR	neglagolska-končnica (28), glagolska-končnica (22), neglagolska-osnova (15), preglas-in-čč (14), glagolska-osnova (11), neobstojni-vokal (8)
	Dodatne oznake (30) Koda: DOD	besede-mati-hči (15), variante (9), besede-otrok (6)

Prvi nivo oznake	Drugi nivo oznake	Koda oznake
Črkovanje (958) Koda: Č	Konzonanti (405) Koda: KONZ	izpust (151), menjava-strešice (76), menjava-drugo (58), odveč (48), menjava-mn (23), menjava-sz (16), menjava-kgh (13), menjava-šž (10), menjava-td (10)
	Vokali (266) Koda: VOK	izpust (102), menjava-ao (55), menjava-drugo (41), menjava-ci (34), odveč (18), menjava-uo (16)
	Črkovni sklopi (143) Koda: SKLOP	lj (52), nj (31), ij (26), podvojene (16), zlog (10), premet (8)
	Ustnično-ustnični w (91) Koda: W	začetek (59), konec (21), sredina (8), v (3)
	Variantni predlogi (53) Koda: PRED	sz (48), kh (5)
Povezani popravki (932) Koda: P	Zapis (509) Koda: ZAP	mala-velika (509)
	Oblika (229) Koda: OBL	drugo (229)
	Skladnja (194) Koda: SKLA	drugo (193), osebek (1)
Nečitljivi prime-ri (11) Koda: N		nečitljivo (11)

Tabela 3: Zastopanost posameznih vrst popravkov v množici Šolar-Eval

Podobno kot velja za korpus *Šolar*, so najpogostejši med jezikovnimi popravki v Šolar-Eval popravki stave vejice, znotraj te kategorije pa so najpogostejši popravki, povezani z vejico med stavki. Glede na skupno število izstopajo tudi popravki odvečnih jezikovnih elementov, ki vključujejo odpravljanje ponavljanja, nepotrebnih zaimkov, členkov in podobno. Pri popravkih skladenjske strukture besedil prednjačita deljenje in združevanje stavkov. Jezikovni popravki besedišča se pogosto nanašajo na zamenjavo polnopomenskih besed z zaimki in obratno. Med popravki oblike so najpogostejše spremembe oblike glede na sklon in število, na ravni črkovanja pa popravki, povezani z izpustom soglasnikov in samoglasnikov ter menjavo črkovnih sklopov.

Kot omenjeno v razdelku 2, smo spise za vključitev v množico Šolar-Eval izbrali glede na zastopanost posameznih jezikovnih popravkov, in sicer tako, da je bila vsaka od kategorij v korpusu *Šolar* zastopana z vsaj petimi popravki. Kot je vidno iz tabele 3, se tudi v Šolar-Eval večina kategorij pojavi več kot 5-krat, obstajajo pa tudi izjeme, npr. redkejšje kategorije pri stavi vejice, popravki zapisa krajšav ter ločil in podobno. Uporaba množice za evalvacijo strojnega popravljanja bo pokazala, za katere kategorije popravkov je zastopanost zadovoljiva in kakšne vrste gradivo bi bilo v nadaljevanju treba še dodati.

5 Sklep in nadaljnje delo

V članku smo predstavili pripravo množice Šolar-Eval 1.0, ki vsebuje 109 spisov iz korpusa *Šolar* 3.0 s konsistentno pripisanimi in natančno vsebinsko označenimi jezikovnimi popravki. Množica je na voljo v odprtem dostopu na repozitoriju CLARIN.SI in bo koristna za raziskovalce in razvijalce, ki se ukvarjajo z razvojem strojnega popravljanja slovenskih besedil.

Naše delo poteka pod okriljem projekta Empirična podlaga za digitalno podprt razvoj pisne jezikovne zmožnosti,⁷ kjer bomo množico uporabili za evalvacijo novih odprtokodnih modelov za strojno popravljanje napak črkovanja, zapisa in oblikoslovja. V nadaljevanju želimo v množici ročno pregledati in popraviti strojno pripisane jezikoslovne oznake, kar bo omogočilo uporabo za evalvacijo strojnega označevanja besedil z jezikovnimi napakami.

Priprava Šolar-Eval je ponudila priložnost za jezikoslovne razmisleke in razvoj metodologije za dosledno (ali doslednejše) pripisovanje jezikovnih popravkov. Skladno s predhodnimi raziskavami se je potrdilo, da je usklajevanje na nižjih ravneh preprostejša naloga kot velja za višje ravni, vendar je tudi na ravni popravkov besedišča, skladnje in besedilne koherence mogoče doseči višjo stopnjo skladnosti, kot jo kažejo učiteljski popravki v korpusu *Šolar*. Zanimivo vprašanje je zlasti, ali oziroma do katere mere bi bilo smiselno k tovrstnemu (po)enotenju v jezikovnodidaktičnem smislu stremeti, koliko in kdaj pa jezikovno popravljanje in podajanje povratne informacije razumeti kot interpretativen in individualiziran proces. Slednje je toliko pomembnejše v luči novih tehnologij generativne umetne inteligence, ki ponujajo številne možnosti ne le jezikovnega popravljanja in izboljševanja, temveč tudi besedilnega soustvarjanja. Z izdelavo velikih jezikovnih modelov za slovenščino bodo tovrstna orodja vedno boljša in uporabnejša tudi za slovenske jezikovne uporabnice in uporabnike. Jezikovnodidaktična stroka je pred izzivom ugotoviti, kako lahko nove možnosti uporabimo (tudi) na način, ki bo jezikovnemu razvoju v podporo. K temu cilju seveda pripomore tudi možnost kvalitetne evalvacije strojnega popravljanja oziroma širšega poseganja v besedilo, k čemur prispevamo z delom, ki smo ga predstavili v prispevku.

Zahvala

Projekt Empirična podlaga za digitalno podprt razvoj pisne jezikovne zmožnosti v letih 2021–2024 financira Ministrstvo za kulturo Republike Slovenije. Raziskovalna programa št. P6-0411 (Jezikovni viri in tehnologije za slovenski jezik) in št. P6-0215 (Slovenski jezik – bazične, kontrastivne in aplikativne raziskave) sofinancira Javna agencija za znanstvenoraziskovalno in inovacijsko dejavnost Republike Slovenije iz državnega proračuna.

⁷ Povezava: <https://www.cjvt.si/prop/> (dostop 30. 11. 2023).

Viri

Arhar Holdt, Špela, Gantar, Polona, Bon, Mija, Gapsa, Magdalena, Lavrič, Polona in Klemen, Matej, 2023a: *Dataset for evaluation of Slovene spell- and grammar-checking tools Šolar-Eval 1.0*. Repozitorij CLARIN.SI. <http://hdl.handle.net/11356/1902>. (Dostop 6. 12. 2023.)

Arhar Holdt, Špela, Rozman, Tadeja, Stritar Kučuk, Mojca, Krek, Simon, Krapš Vodopivec, Irena, Stabej, Marko, Pori, Eva, Goli, Teja, Lavrič, Polona, Laskowski, Cyprian, Kocjančič, Polonca, Klemenc, Bojan, Krsnik, Luka in Kosem, Iztok, 2022a: *Developmental corpus Šolar 3.0*. Repozitorij CLARIN.SI. <http://hdl.handle.net/11356/1589>. (Dostop 30. 10. 2023.)

Arhar Holdt, Špela, Rozman, Tadeja, Stritar Kučuk, Mojca, Krek, Simon, Krapš Vodopivec, Irena, Stabej, Marko, Pori, Eva, Goli, Teja, Lavrič, Polona, Laskowski, Cyprian, Kocjančič, Polonca, Klemenc, Bojan, Krsnik, Luka, Žagar, Aleš in Kosem, Iztok, 2022c: *Frequency list of language problems from Šolar 3.0*. Repozitorij CLARIN.SI. <http://hdl.handle.net/11356/1716>. (Dostop 30. 10. 2023.)

Kolokacijski slovar sodobne slovenščine 2.0: <https://viri.cjvt.si/kolokacije/slv/#>. (Dostop 26. 9. 2023.)

Korpus pisne standardne slovenščine Gigafida 2.0: <https://viri.cjvt.si/gigafida/>. (Dostop 26. 9. 2023.)

Kosem, Iztok, Gantar, Polona, Krek, Simon, Arhar Holdt, Špela, Čibej, Jaka, Laskowski, Cyprian, Pori, Eva, Klemenc, Bojan, Dobrovoljc, Kaja, Gorjanc, Vojko in Ljubešič, Nikola, 2023: *Collocations Dictionary of Modern Slovene KSSS 1.0*. Repozitorij CLARIN.SI. <http://hdl.handle.net/11356/1250>. (Dostop 30. 10. 2023.)

Slovar sopomenk sodobne slovenščine 2.0: <https://viri.cjvt.si/sopomenke/slv/>. (Dostop 26. 9. 2023.)

Slovarski portal Fran. Inštitut za slovenski jezik Frana Ramovša ZRC SAZU: www.fran.si. (Dostop 26. 10. 2023.)

Literatura

Ahačič, Kozma, 2017a: *Kratkoslovnica: slovenska slovnica za osnovno šolo*. 1. izdaja. Ljubljana: Rokus Klett.

Ahačič, Kozma, 2017b: *Slovnica na kvadrat: slovenska slovnica za srednjo šolo*. 1. izdaja. Ljubljana: Rokus Klett.

Arhar Holdt, Špela in Kosem, Iztok, 2023: Šolar, the developmental corpus of slovene. Preprint. DOI: <https://doi.org/10.21203/rs.3.rs-3274669/v1>.

Arhar Holdt, Špela in Rozman, Tadeja, 2015: Možnosti uporabe podatkov iz korpusa Šolar za pripravo slovarskih priročnikov. Smolej, Mojca (ur.): *Slovnica in slovar – aktualni jezikovni opis*. 1. del. Obdobja 34. Ljubljana: Znanstvena založba Filozofske fakultete. 67–74.

Arhar Holdt, Špela, Gantar, Polona, Kosem, Iztok, Pori, Eva, Robnik Šikonja, Marko in Krek, Simon, 2023: Thesaurus of Modern Slovene 2.0. Medved, Marek idr. (ur.): eLex 2023: electronic lexicography in the 21st century (eLex 2023): proceedings of the eLex 2023 conference: [Brno], 27–29 June 2023. Brno: Lexical Computing CZ. 366–381. <https://ellex.link/ellex2023/wp-content/uploads/82.pdf>. (Dostop 30. 11. 2023.)

Arhar Holdt, Špela, Kosem, Iztok in Gantar, Polona, 2017: Corpus-based resources for L1 teaching: the case of Slovene. Marcus-Quinn, Ann in Hourigan, Triona (ur.): *Handbook on digital learning for K-12 schools*. [S. l.]: Springer. 91–113. DOI: https://doi.org/10.1007/978-3-319-33808-8_1.

Arhar Holdt, Špela, Lavrič, Polona, Roblek, Rebeka in Goli, Teja, 2022b: *Kategorizacija učiteljskih popravkov: Smernice za označevanje korpusa Šolar*. Rezultat projekta Razvoj slovenščine v digitalnem okolju. Različica 1.1. <https://wiki.cjvt.si/books/11-jezikovni-popravki-solar/page/oznacevalne-smernice>. (Dostop 30. 11. 2023.)

Arhar Holdt, Špela, Pollak, Senja, Robnik Šikonja, Marko in Krek, Simon, 2020: Referenčni seznam pogostih splošnih besed za slovenščino. Fišer, Darja in Erjavec, Tomaž (ur.): *Zbornik konference Jezikovne tehnologije in digitalna humanistika, 24.–25. september 2020, Ljubljana, Slovenija*. Ljubljana: Inštitut za novejšo zgodovino. 10–15. http://nl.ijs.si/jtdh20/pdf/JT-DH_2020_Arhar-Holdt-et-al_Referencni-seznam-pogostih-splosnih-besed-za-slovenscino.pdf. (Dostop 30. 11. 2023.)

Bon, Mija in Gapsa, Magdalena, v pripravi: Analiza napak pri rabi vejice v šolskih spisih. *Škrabčevi dnevi 13: Zbornik prispevkov s simpozija 2023*. Nova Gorica, 20. oktober 2023.

Gantar, Polona, Arhar Holdt, Špela, Čibej, Jaka, Kuzman, Taja in Kavčič, Teja, 2018: Glagolske večbesedne enote v učnem korpusu ssj500k 2.1. Fišer, Darja in Pančur, Andrej (ur.): *Zbornik konference Jezikovne tehnologije in digitalna humanistika, 20.–21. september 2018, Ljubljana, Slovenija*. Ljubljana: Znanstvena založba Filozofske fakultete. 85–92. <http://nl.ijs.si/jtdh18/JTDH-2018-Proceedings.pdf>. (Dostop 30. 11. 2023.)

Holozan, Peter, 2013: Uporaba strojnega učenja za postavljanje vejic v slovenščini. *Uporabna informatika* 21/4. 196–209.

Holozan, Peter, 2015: Možnosti uporabe jezikovnih tehnologij za določanje težav pri rabi vejice. Dobrovoljc, Helena in Lengar Verovnik, Tina (ur.): *Pravopisna razpotja: razprave o pravopisnih vprašanjih*. Ljubljana: Založba ZRC, ZRC SAZU. 77–92.

Kosem, Iztok, Arhar Holdt, Špela, Gantar, Polona in Krek, Simon, 2023: Collocations Dictionary of Modern Slovene 2.0. Medved, Marek idr. (ur.): *eLex 2023: electronic lexicography in the 21st century (eLex 2023): proceedings of the eLex 2023 conference: [Brno], 27–29 June 2023*. Brno: Lexical Computing CZ. 491–507. <https://elex.link/elex2023/wp-content/uploads/100.pdf>. (Dostop 30. 11. 2023.)

Kosem, Iztok, Rozman, Tadeja, Arhar Holdt, Špela, Kocjančič, Polonca in Laskowski, Cyprian Adam, 2016: Šolar 2.0: nadgradnja korpusa šolskih pisnih izdelkov. Erjavec, Tomaž in Fišer, Darja (ur.): *Zbornik konference Jezikovne tehnologije in digitalna humanistika, 29. september–1. oktober 2016, Filozofska fakulteta, Univerza v Ljubljani, Ljubljana*. Ljubljana: Znanstvena založba Filozofske fakultete Univerze v Ljubljani. 95–100. http://www.sdj.si/wp/wp-content/uploads/2016/09/JTDH-2016_Kosem-et-al_Solar-2-0-nadgradnja-korpusa-solskih-pisnih-izdelkov.pdf. (Dostop 30. 11. 2023.)

Kosem, Iztok, Stritar Kučuk, Mojca, Može, Sara, Zwitter Vitez, Ana, Arhar Holdt, Špela in Rozman, Tadeja, 2012: *Analiza jezikovnih težav učenecv: korpusni pristop*. Ljubljana: Trojina. DOI: <https://doi.org/10.4312/9789610603535>.

Kranjc, Anja in Robnik Šikonja, Marko, 2015: Postavljanje vejic v slovenščini s pomočjo strojnega učenja in izboljšanega korpusa Šolar. Fišer, Darja (ur.): *Zbornik konference Slovenščina na spletu in v novih medijih, Ljubljana, 25.–27. november 2015*. Ljubljana: Znanstvena založba Filozofske fakultete Univerze v Ljubljani. 38–43. <https://ebooks.uni-lj.si/ZalozbaUL/catalog/view/80/172/1807>. (Dostop 30. 11. 2023.)

Krek, Simon, Arhar Holdt, Špela, Erjavec, Tomaž, Čibej, Jaka, Repar, Andraž, Gantar, Polona, Ljubešić, Nikola, Kosem, Iztok in Dobrovoljc, Kaja, 2020: Gigafida 2.0: the reference corpus of written standard Slovene. Calzolari, Nicoletta idr. (ur.): *LREC 2020: Twelfth International Conference on Language Resources and Evaluation: May 11-16, 2020, Palais du Pharo, Marseille, France*. Paris: ELRA – European Language Resources Association. 3340–3345. <http://www.lrec-conf.org/proceedings/lrec2020/LREC-2020.pdf> (Dostop 30. 11. 2023.)

Leech, Geoffrey, 1997: Teaching and language corpora: A convergence. Wichmann, Anne idr. (ur.): *Teaching and language corpora*. London: Longman. 1–23. DOI: <https://doi.org/10.4324/9781315842677>.

Mokotar, Rok, 2023: Obvladovanje slovničnih napak v šolskih pisnih izdelkih z metodami za obdelavo naravnega jezika. Diplomsko delo. Ljubljana: Fakulteta za računalništvo in informatiko, Univerza v Ljubljani. <https://repozitorij.uni-lj.si/IzpisGradiva.php?id=144932>. (Dostop 30. 11. 2023.)

Može, Sara, 2013: Raba kratkega nedoločnika: korpusni pristop. *Slovenščina 2.0* 1/1. 155–175. DOI: <https://doi.org/10.4312/slo2.0.2013.1.155-175>.

Pala, Karel, Rychlý, Pavel in Smrž, Pavel, 2003: Text Corpus with Errors. Matoušek, Václav in Mautner, Pavel (ur.): *Text, Speech and Dialogue, Proceedings of 6th International Conference, TSD 2003, České Budějovice, Czech Republic, September 8-12, 2003*. Berlin, Heidelberg: Springer. 90–97. DOI: https://doi.org/10.1007/978-3-540-39398-6_13.

Petrič, Timotej, 2022: Predlogi jezikovnih popravkov v slovenščini z modelom SLoBERTa. Diplomsko delo. Ljubljana: Fakulteta za računalništvo in informatiko, Univerza v Ljubljani. <https://repozitorij.uni-lj.si/IzpisGradiva.php?id=139688>. (Dostop 30. 11. 2023.)

Popič, Damjan, 2014: Revising translation revision in Slovenia. Mikolič Južnič, Tamara idr. (ur.): *New Horizons in Translation Research and Education 2*. Joensuu: University of Eastern Finland, Philosophical faculty. 72–89. <https://erepo.uef.fi/handle/123456789/14340>. (Dostop 30. 11. 2023.)

Rozman, Tadeja, Arhar Holdt, Špela, Pollak, Senja in Kosem, Iztok, 2018: Kolokacije v korpusu Šolar. *Jezik in slovstvo* 63/2–3. 117–128. <https://www.jezikinslovstvo.com/pdf.php?part=2018%7C2-3%7C117-128>.

Rozman, Tadeja, Stritar, Mojca in Kosem, Iztok, 2012: Šolar – korpus šolskih pisnih izdelkov. Rozman, Tadeja idr. (ur.): *Empirični pogled na pouk slovenskega jezika*. Ljubljana: Trojina. DOI: <https://doi.org/10.4312/9789610603511>.

Stritar Kučuk, Mojca, 2022: KOST med korpusi usvajanja tujega jezika. Pirih Svetina, Nataša in Ferbežar, Ina (ur.): *Na stičišču svetov: slovenščina kot drugi in tuji jezik*. Obdobja 41. Ljubljana: Znanstvena založba Filozofske fakultete Univerze v Ljubljani. 323–334.

Terčon, Luka in Ljubešić, Nikola, 2023: CLASSLA-Stanza: The Next Step for Linguistic Processing of South Slavic Languages. Preprint, arXiv. DOI: <https://doi.org/10.48550/arXiv.2308.04255>.

Tinta, Marjana, 2019: Korpus Šolar in zanesljivost ocenjevanja šolskih pisnih nalog na različnih stopnjah šolanja. Magistrsko delo. Nova Gorica: Fakulteta za humanistiko, Univerza v Novi Gorici. <https://repozitorij.ung.si/Dokument.php?id=18748&lang=slv>. (Dostop 30. 11. 2023.)