Prihodnost korpusa Šolar

Špela ARHAR HOLDT
Univerza v Ljubljani, Filozofska fakulteta
Univerza v Ljubljani, Fakulteta za računalništvo in informatiko

Eva PORI
Univerza v Ljubljani, Filozofska fakulteta

Iztok KOSEM
Univerza v Ljubljani, Filozofska fakulteta
Univerza v Ljubljani, Fakulteta za računalništvo in informatiko
Institut »Jožef Stefan«

Povzetek

Razvojni korpusi so skrbno oblikovane digitalne zbirke avtentičnih besedil, ki omogočajo vpogled v jezikovni razvoj mlajših naravnih govorcev določenega jezika. Pisni razvojni korpusi, kakršen je za slovenščino korpus Šolar, vključujejo primere pisanja osnovnošolskih in srednješolskih učencev, pogosto skupaj s popravki jezikovnih težav, in kot taki 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. Prispevek predstavlja značilnosti slovenskega razvojnega korpusa v primerjavi s podobnimi viri za druge jezike, njegov razvojni krog in številne novosti, ki jih je k metodologiji gradnje prispevalo delo na projektu Razvoj slovenščine v digitalnem okolju. Glavne novosti so izboljšana pravna podlaga za zbiranje besedil, uporabniško prijazen portal za oddajo besedil, orodje CJVT Svala za transkripcijo, anonimizacijo in označevanje popravkov ter izboljšani korpusni format. Ob pojavu generativne umetne inteligence in jezikovnih orodij, ki uporabnicam in uporabnikom pomagajo pri pisanju in komuniciranju izpostavimo spremljanje razvoja (in morebitnega upada) jezikovnih kompetenc kot ključno za nadaljnje

delo in ponudimo strategijo prihodnjega razvoja korpusa Šolar in sorodnih podatkovnih virov.

Ključne besede: razvojni korpus, Šolar 3.0, metodologija korpusne gradnje, CJVT Svala, portal za zbiranje besedil

Abstract

Developmental corpora are carefully designed digital collections of authentic texts that provide insights into the development of younger native speakers' language skills. Written developmental corpora, such as the Šolar corpus for Slovene, include examples of writing by primary and secondary school students, often accompanied by language corrections, and as such, provide an empirical basis for research in the fields of language acquisition and didactics, for the development of teaching materials, exercises, tests, training sets for natural language processing, and for the development of tools that support and develop literacy. The paper presents the characteristics of the Slovene developmental corpus compared to similar resources for other languages, its development cycle and the many innovations of the corpus-building methodology developed under the umbrella of the Development in the Digital Environment project: an improved legal basis and a user-friendly portal for text collection, the CJVT Svala tool for transcription, anonymisation and annotation of corrections, and an enhanced corpus format. With the emergence of generative artificial intelligence and language tools that help users write and communicate, we highlight the monitoring of linguistic competencies' development (and possible decline) as crucial to future work and offer a strategy for the further development of the Šolar corpus and related data resources.

Keywords: developmental corpus, Šolar 3.0, corpus building methodology, CJVT Svala, portal for text collection

1 Uvod

Razvojni korpusi (ang. *developmental corpora*, Leech, 1997:19) so premišljeno grajene digitalne zbirke avtentičnih besedil, ki ponujajo vpogled v razvoj jezikovnih kompetenc pri mlajših naravnih govorcih

in govorkah določenega jezika.[1] V prispevku se osredotočamo na pisne razvoje korpuse, ki tipično zajemajo primere osnovnošolskega in srednješolskega pisanja, pogosto pa tudi oznake jezikovnih težav, ki se v teh besedilih pojavijo. Ti korpusi 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.

Zaradi vsega naštetega so razvojni korpusi med pomembnejšimi specializiranimi jezikovnimi viri in del temeljne jezikovne infrastrukture. Mogoče pa je predvideti, da bo zanimanje za tovrstne vire in metodologijo njihove priprave v prihodnje še naraščalo, kot posledica napredka na področju generativne umetne inteligence in raznovrstnih jezikovnih orodij, ki uporabnicam in uporabnikom pomagajo pri pisanju in komuniciranju. Po pojavu tehnologij, ki ustvarjajo besedila skupaj s piscem ali namesto njega, namreč postaja vprašanje spremljanja razvoja (in morebitnega upada) človeških jezikovnih kompetenc še bolj pereče in temeljno kot v preteklosti.

V evropskem prostoru je mogoče najti kar nekaj primerov razvojnih korpusov, ki vsebujejo pisna besedila osnovnošolcev in/ali dijakov, ne gre pa prezreti, da je takšnih virov bistveno manj od korpusov z besedili govorcev, ki se določenega jezika učijo kot drugega/tujega (ang. *learner corpora*). Za angleščino so na voljo korpusi LUCY (Sampson, 2003), LOCNESS (Granger, 1998) in obsežna zbirka novozelandskih esejev, ki jih je zbrala Parr (2010). Za nemščino so za osnovnošolsko pisanje na voljo korpusi H1, H2, E2, ERK1 (Berkling, 2016; 2018) in Litkey (Laarman-Quante idr., 2019), za srednješolsko pisanje pa korpus KoKo (Abel idr., 2014). Za italijanščino so na voljo korpus CItA (Barbagli idr., 2016), trojezični LEONIDE (Glaznieks idr., 2022) in korpusi esejev, ki so jih zbrali Marconi idr. (1993) za osnovnošolsko in Borghi (2013) za srednješolsko raven. Številni razvojni korpusi za

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

francoščino so na voljo prek portala È:CALM (Ho-Dac idr., 2020). Med novejšimi je mogoče omeniti tudi zbirke besedil za islandščino (Arnardóttir idr., 2021, Ingason idr., 2021) in DOESTE (Martins idr., 2020), ki vsebuje besedila v evropski in brazilski portugalščini. Razen zadnjih dveh in LUCY, ki so po obsegu nekoliko manjši, ter korpusa Parr, ki prinaša skoraj 21.000 esejev, zajemajo navedeni viri nekje med 2.500 in 5.000 besedil oziroma med 100.000 in 1.000.000 pojavnic. Veliko jih vsebuje tudi jezikovne oznake na osnovnih nivojih (tokenizacija, lematizacija, oblikoskladnja) ter popravke, ki so jih vnesli raziskovalci, ki so korpus gradili.

Šolar, razvojni korpus za slovenščino, je v veliki meri primerljiv, mestoma presega prakse iz tujine, mestoma pa se jim tudi odmika. Največja konceptualna razlika je v odločitvi, da se v korpus vključijo avtentični učiteljski popravki, s pomočjo katerih je mogoče opazovati podajanje povratne informacije neposredno v kontekstu razvoja pisnih kompetenc. Od tujih primerov avtentične popravke učiteljev oz. profesorjev vključuje korpus Chyby (Pala idr., 2003), ki pa se za razliko od Šolarja posveča pisanju na univerzitetni ravni. Korpus Šolar, ki nastaja in se razvija že od leta 2012 (Kosem idr., 2012; 2016), je svojo zadnjo nadgradnjo doživel leta 2023. Prenovljena različica 3.0 je izšla pod okriljem projekta Razvoj slovenščine v digitalnem okolju,[2] kjer so bili zasnovani in evalvirani postopki ter orodja za kontinuiran razvoj korpusa Šolar, posredno pa tudi drugih korpusov, ki vsebujejo jezikovne popravke.[3]

Nekatere novosti, ki so nastale na projektu Razvoj slovenščine v digitalnem okolju, so bile omenjene v prispevkih Arhar Holdt idr. (2022a) ter Arhar Holdt in Kosem (2023), vključene so tudi v projektno poročilo (Arhar Holdt idr., 2023). Vendar rezultati do sedaj še niso bili celovito in pregledno predstavljeni z vidika prispevka za raziskovalno skupnost, razvoja discipline in samega korpusa. V tem prispevku najprej predstavimo razvojni krog korpusa Šolar, sledi popis

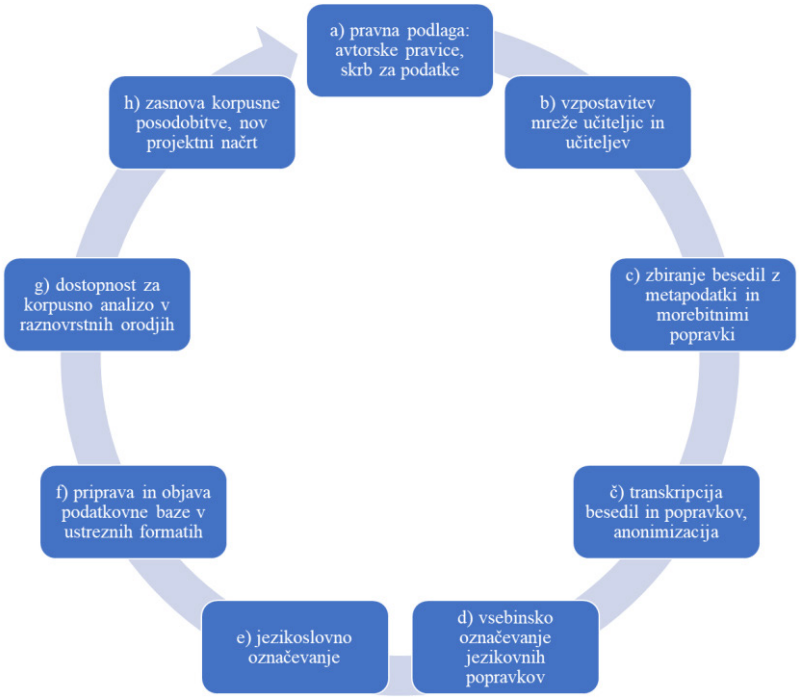
2 Cilj projekta je bil zadovoljiti potrebe po računalniških izdelkih in storitvah s področja jezikovnih tehnologij za slovenski jezik za raziskovalne organizacije, podjetja in širšo javnost. Spletna stran projekta z dostopom do rezultatov: https://slovenscina.eu/.

3 V našem prostoru gre omeniti še korpus slovenščine kot tujega jezika KOST (Stritar Kučuk, 2022) in korpus lektorskih popravkov Lektor (Popič, 2014).

postopkov in orodij za nadaljnje zbiranje, kratka predstavitev Šolarja 3.0 in njegove dostopnosti, zaključujemo pa z naborom prioritet za nadaljnje delo in strategijo za prihodnji razvoj korpusa.

2 Razvojni krog korpusa Šolar

Gradnja korpusa Šolar poteka primerljivo z drugimi korpusnimi viri, določene specifike kaže Slika 1. Zbiranje besedil poteka s pomočjo učiteljske skupnosti, besedilodajalci so učenci oz. dijaki, zato je pomemben del gradnje vzpostavitev mreže in motivacija za sodelovanje učiteljske skupnosti. Učitelji oz. učiteljice morajo urediti pogodbo s šolo, ki dovoljuje zbiranje gradiva, prav tako pa zbrati soglasja avtorjev in avtoric oz. njihovih zakonitih zastopnikov. Poskrbijo tudi za oddajo besedil in vseh želenih metainformacij o njih. Ko je gradivo zbrano, ga pretvorimo v korpusna besedila, kar vključuje



Slika 1: Razvojni krog korpusa Šolar.

transkripcijo (kadar so izvorna besedila napisana na roko, kar trenutno velja za večino primerov), anonimizacijo osebnih podatkov, ki se lahko v besedilih pojavljajo, vnos in označevanje jezikovnih popravkov, jezikoslovno označevanje in izdelavo korpusne baze v končnem formatu oz. formatih. Baza mora biti umestljiva v orodja za analizo, med katerimi so zlasti konkordančniki in druga orodja za vizualizacijo ter ekstrakcijo korpusnih podatkov. Za vse sinhrone jezikovne vire je ključno kontinuirano nadgrajevanje in posodabljanje gradiva; za razvojne korpuse, kjer so longitudinalne raziskave posebej zaželene, pa to velja še toliko bolj. Zasnova projektne nadgradnje ponuja tudi priložnost za oceno uporabljenih postopkov in popis želenih izboljšav.

Kot je popisano v Arhar Holdt idr. (2022a), so bile pri gradnji korpusa Šolar 1.0 in 2.0 na številnih mestih prisotne težave. Pri zbiranju besedil za prvo različico so učitelji in učiteljice pošiljali fizične kopije besedil učencev, njihova kakovost pa se je razlikovala glede na uporabljeni kopirni stroj. Kopirani dokumenti so bili pogosto črno-beli, kar je oteževalo razlikovanje med popravki, ki jih je opravil učitelj, in tistimi, ki je zabeležil učenec sam. Za drugo različico korpusa smo prešli na zbiranje skeniranih besedil, po možnosti barvnih, in s tem na posredovanje PDF-datotek prek spleta, še vedno pa je bilo zamudno zbiranje metapodatkov in spremljanje procesa sodelovanja učiteljskih ekip. Izjemno zamudna je bila tudi priprava korpusnih dokumentov. Zapisovalci in zapisovalke so jezikovne popravke v besedila vpisovali s pomočjo XML-oznak, kar je bilo zahtevno, nepregledno in je vodilo v številne napake. Vsebinsko kategorizacijo jezikovnih popravkov smo pri verziji 2.0 opravljali v za naše namene prilagojenem orodju Sketch Engine (Kilgarriff idr., 2004). Korpus smo morali najprej pretvoriti v format VERT za uvoz v Sketch Engine; tam smo po vsebinskih sklopih opravili revizijo oznak. Med delom smo izvažali korpusne datoteke in jih pretvarjali v format XML, da smo lahko novooznačene kategorije zapisali v korpusne datoteke, spet opravili pretvorbo in korpus uvozili nazaj v Sketch Engine. Zaradi načina dela označevalci in označevalke niso imeli pregleda nad širšim kontekstom označevanega besedila, niso mogli spreminjati

segmentacije popravkov in odpravljati težav, ki niso bile vezane na točno tisto oznako, ki so jo v določenem koraku imeli v analizi.

Po koncu projekta Razvoj slovenščine v digitalnem okolju so koraki korpusne gradnje temeljito nadgrajeni. Na voljo je spletno mesto z informacijami, prenovljenimi pogodbami in repozitorijem za oddajo besedil in metapodatkov, kar učiteljski skupnosti olajša zbiranje in posredovanje gradiva (Razdelek 3). Bistvene izboljšave so na ravni metodologije priprave korpusnih besedil (Razdelek 4): za slovenščino smo lokalizirali in nadgradili uporabniku prijazno in zmogljivo orodje Svala, ki omogoča transkripcijo besedil, označevanje jezikovnih popravkov in pregledno sočasno anonimizacijo potencialno občutljivih osebnih informacij, ki se lahko pojavljajo v besedilih. Šolar 3.0 (Razdelek 5) je na voljo z bogatejšimi jezikoslovnimi oznakami, od katerih so zlasti dragocene skladenjske, ter v novem formatu, ki je v celoti kompatibilen z ostalimi slovenskimi korpusi.

3 Zbiranje korpusnega gradiva

3.1 Pravne rešitve

Po odločitvi učiteljev za sodelovanje pri zbiranju besedil in še pred uporabo portala sledi najprej pravna ureditev sodelovanja, in sicer med raziskovalno ustanovo na eni strani ter šolo in učenci na drugi. S šolo se sklene pogodba o sodelovanju, z učenci oz. njihovimi zastopniki pa pogodba o prenosu ustreznih avtorskih pravic. Podpisane pogodbe (dva izvoda pogodbe s šolo in dva izvoda pogodbe z vsakim avtorjem oz. avtorico šolskih besedil oz. njegovim zakonitim zastopnikom) učitelj oz. učiteljica pred pričetkom sodelovanja pri zbiranju pošlje raziskovalni enoti, ki gradi korpus, kjer jih podpiše še druga stranka in po en izvod vrne na šolo.[4] Na ta način je zbiranje besedil pravno urejeno, saj brez tega zbrano gradivo ne more biti odprto dostopno za nadaljnjo rabo. Za vsa vključena besedila tako obstaja pogodba, ki opredeljuje prenos avtorskih pravic ter načine

4 Izkušnje projektnega sodelovanja s šolami so namreč pokazale, da šolski sistem preferira fizično podpisovanje, da trenutno še ni opremljen ali pa pripravljen na digitalno podpisovanje dokumentov.

hranjenja in procesiranja besedil. Pomembno pri tem pa je, da se pravne rešitve ne osredotočajo le na obdobje trajanja specifičnega projekta, npr. točno določeno šolsko leto, ker to ovira in onemogoča kontinuirano in širše zbiranje besedil. Trenutne pogodbe so na voljo kot Priloge 2–4 v (Arhar Holdt idr., 2023).

3.2 Portal za oddajo besedil

Pravni ureditvi sodelovanja sledi delo s portalom,[5] ki je razvit je z namenom, da bi oddajanje besedil – in vse korake, potrebne za sodelovanje – olajšali tako skupnosti sodelujočih učiteljev kot raziskovalcem, ki besedila za korpus pripravljajo. Pri razvoju portala je bila v ospredju želja, da bo njegova uporaba enostavna in intuitivna, hkrati pa bo vsebovala vse uporabne funkcionalnosti. Uporabniku prijazen vmesnik je osnovni pogoj za sodelovanje čim večjega števila učiteljev, katerih veščina dela z računalnikom in s tem odnos do njega se lahko močno razlikuje. Portal vzpodbuja tudi vzpostavljanje skupnosti sodelujočih besedilodajalcev; saj lahko ekipa učiteljev z iste šole s pomočjo statistik spremlja svoj napredek pri zbiranju besedil, tudi primerjalno glede na druge šole v regiji.[6]

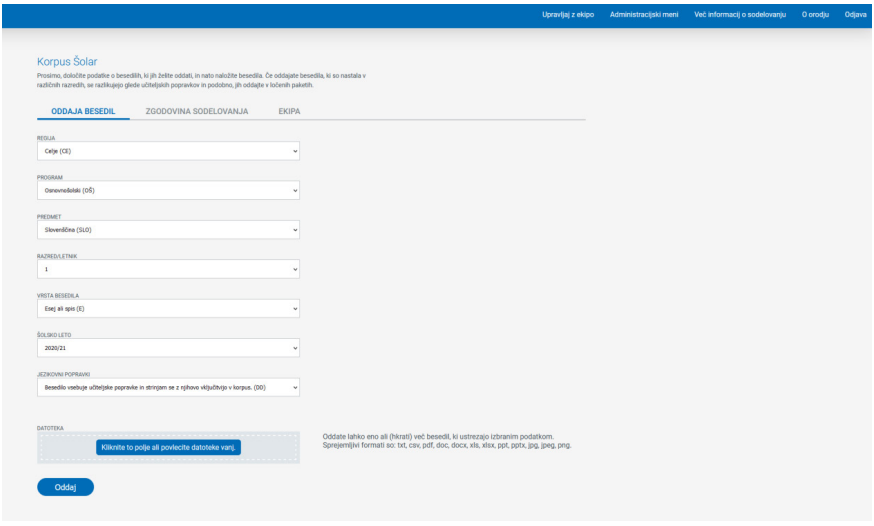
Na vstopni strani portala je na voljo povezava na spletno mesto,[7] kjer je predstavitev korpusa Šolar in pregledna navodila za sodelovanje pri njegovi gradnji. Pri prvi uporabi portala za zbiranje besedil se mora uporabnik registrirati, pri čemer posreduje svoje ime in priimek, naziv institucije, na kateri je zaposlen, e-naslov, določi pa še geslo za vstop v portal in svojo vlogo pri zbiranju: vlogo *Mentor/-ica* izbere, kdor bo zbiral in oddajal šolska besedila, vlogo *Koordinator/-ica* pa tisti, ki bo poleg zbiranja in oddajanja skrbel še za komuniciranje z vodstvom šole in znotraj skupine mentorjev, če je sodelujočih učiteljev z iste šole več. Na izbiro je še *Druga vloga*, ki pokriva opazovalce ali stranske deležnike.

5 Portal je na voljo na spletni strani https://zbiranje.cjvt.si/solar/login/.

6 Učiteljsko skupnost motivira tudi pridobitev točk za napredovanje v nazive, za kar se pripravi potrdilo o sodelovanju pri projektu, ki se na osnovi *Pravilnika o napredovanju zaposlenih na področju vzgoje in izobraževanja v nazive* (Uradni list RS, št. 54/02, 123/08, 44/09, 18/10, 113/20 http://www.pisrs.si/Pis.web/pregledPredpisa?id=PRAV4272) vrednoti na Ministrstvu za izobraževanje, znanost in šport.

7 Dostopno na https://rsdo.slovenscina.eu/zbiranje-besedil-za-korpus-solar.

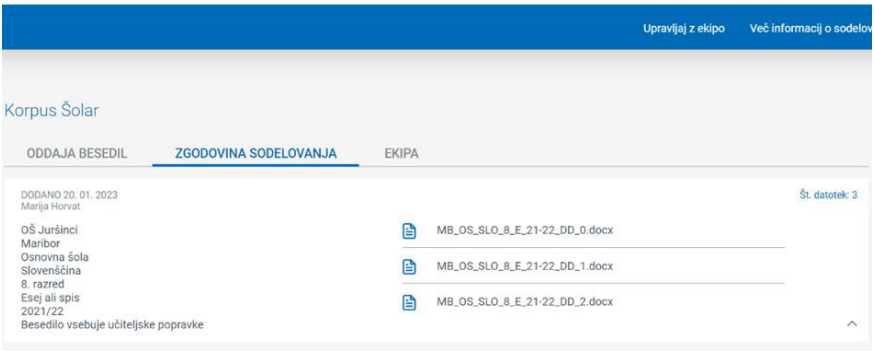
Po uspešni prijavi v portal se uporabnik znajde na strani z osrednjo funkcionalnostjo – oddajo besedil (Slika 2). S pomočjo spustnih seznamov določi vse metapodatke, ki jih potrebujemo za pripravo korpusnih besedil: regijo, v katero se uvršča šola sodelujočega učenca; šolski program (npr. osnovnošolski; splošna in strokovna gimnazija; srednje poklicno izobraževanje); predmet, pri katerem so besedila nastala; razred oz. letnik, v katerem so besedila nastala; vrsto besedila (npr. esej ali spis; praktično besedilo, napisano za oceno; šolski test); šolsko leto, in informacijo, ali besedilo vsebuje jezikovne popravke ter ali sodelujoči učitelj dovoljuje njihovo vključitev v korpus. Sledi oddaja besedil, ki ustrezajo vnesenim metapodatkom, in so lahko v formatih txt, csv, pdf, doc, docx, xls, xlsx, ppt, pptx, jpg, jpeg ali png. Pred oddajo je naložene datoteke mogoče še enkrat pregledati in jih odstraniti ali zamenjati z drugimi. Po potrditvi oddaje se izpiše obvestilo o uspešni oddaji in številu oddanih datotek.



Slika 2: Metapodatki za naložena besedila in okno za oddajo datotek v Zavihku 'Oddaja besedil'.

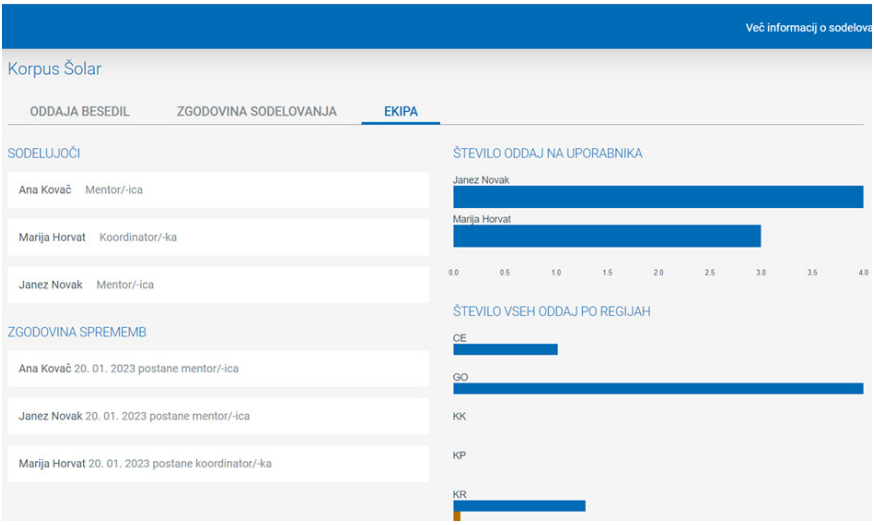
Na naslednji strani portala – v zavihku 'Zgodovina sodelovanja' se beležijo naložene in oddane datoteke uporabnika. Vidne so vse osnovne informacije o oddaji, npr. datum oddaje, ime šole, predmet ter podrobnejše informacije in pogled na naložene datoteke, katerih

imena so tvorjena iz kod vseh izbranih metapodatkov o besedilih (Slika 3).



Slika 3: Razširjen pogled na paket oddanih besedil v zavihku 'Zgodovina sodelovanja'.

V zavihku 'Ekipa' (Slika 4) so shranjeni podatki o sodelujočih članih ekipe. Na levi strani zaslona so izpisana njihova imena skupaj z vlogo, pod tem pa beležena zgodovina sprememb (npr. vlog učiteljev). Desna stran zaslona prikazuje graf s podatki o številu oddanih datotek vsakega člana ekipe in graf, ki izrisuje število vseh oddaj po regijah in vrsti šole. Podatki v obeh grafih se sproti osvežujejo.

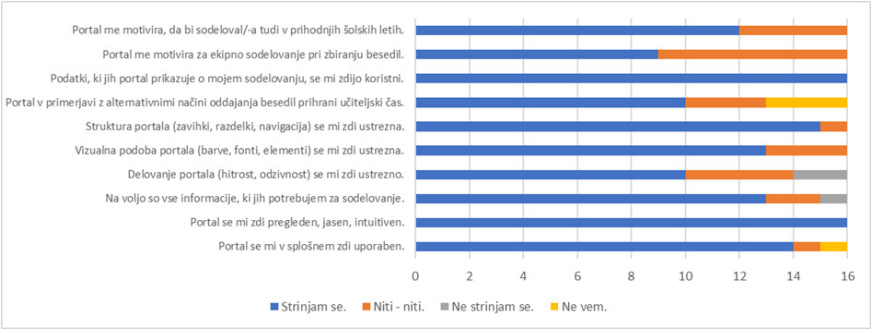


Slika 4: Podatki o sodelujočih članih ekipe v 'Zavihku Ekipa'.

V vmesniku se nahaja še meni za upravljanje z ekipo. Učitelj z vlogo koordinatorja tu najde podatke o članih ekipe v določeni instituciji, omogočeno mu je tudi ročno dodajanje novih članov. Več administratorskih možnosti imajo raziskovalci, ki koordinirajo korpusno gradnjo. Ti lahko potrjujejo in odstranjujejo uporabnike, urejajo imena sodelujočih inštitucij, posodabljajo metapodatke že oddanih vnosov in podobno.

Portal za oddajo besedil je evalviralo 16 učiteljic in učiteljev s 13 šol. Celoten vprašalnik z vsemi odgovori je na voljo kot Priloga 1 v (Arhar Holdt idr., 2023), kjer je tudi opredeljeno, katere identificirane težave so že bile odpravljene in katere čakajo na prihodnji razvoj.

Ocena posameznih strukturnih elementov portala je vključevala vrednotenje funkcionalnosti spletnega mesta z osnovnimi informacijami o sodelovanju, registraciji in prijavi v portal, vnosu podatkov o besedilih, (ne)praktičnosti načina oddaje besedil, strukture portala oz. (ne)funkcionalnosti osrednjih zavihkov. Pri podajanju splošne ocene so se evalvatorji lahko opredelili še do vizualne podobe, delovanja (odzivnosti, hitrosti) portala in motivacijskih elementov za sodelovanje. Na splošno so bili sodelujoči z zasnovo portala zadovoljni, kot kaže Slika 5, za prihodnje delo pa bodo dobrodošli zlasti razmisleki o elementih, ki spodbujajo k dolgoročnejšemu sodelovanju.

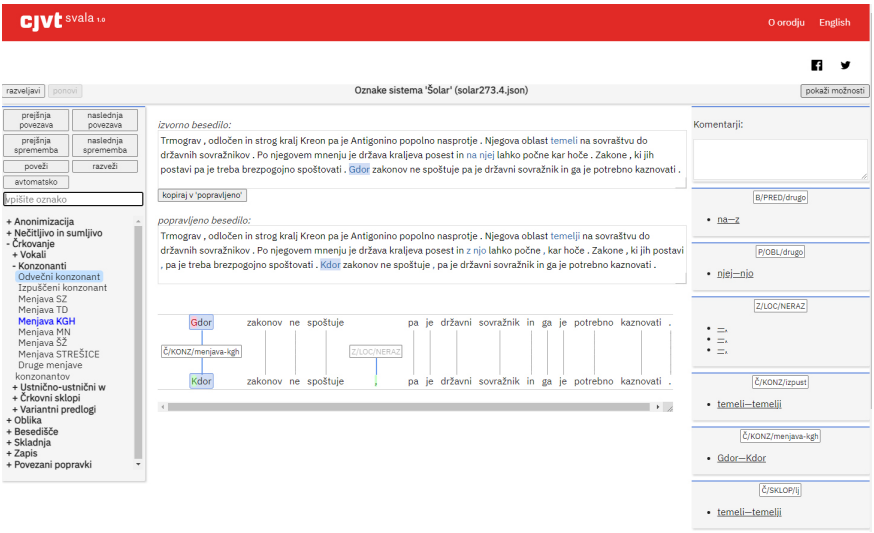


Slika 5: Učiteljska ocena funkcionalnosti na portalu za oddajo besedil.

4 Priprava korpusnih besedil

4.1 Transkripcija, anonimizacija in označevanje popravkov

Orodje CJVT Svala[8] je lokalizirana in adaptirana različica odprtodostopnega orodja Svala, ki je nastalo za pripravo korpusa švedščine kot drugega/tujega jezika (Wirén, 2019). Največja prednost orodja Svala je, da združuje več korakov priprave korpusnih besedil, in sicer transkripcijo, anonimizacijo in označevanje jezikovnih popravkov v besedilih.[9] CJVT Svala 1.0 omogoča označevanje popravkov po dveh sistemih, in sicer po sistemu označevanja korpusa Šolar (Arhar Holdt idr., 2022b) in po sistemu označevanja korpusa KOST (Stritar Kučuk, 2023). Orodje je zasnovano tako, da je mogoče dodati tudi nove označevalne sisteme.



Slika 6: Primer izvornega in popravljenega besedila v vmesniku CJVT Svala 1.0 s sistemom oznak za Šolar.

8 Orodje je prosto dostopno na https://orodja.cjvt.si/svala/, koda je na voljo na repozitoriju GitHub: https://github.com/clarinsi/swell-editor.

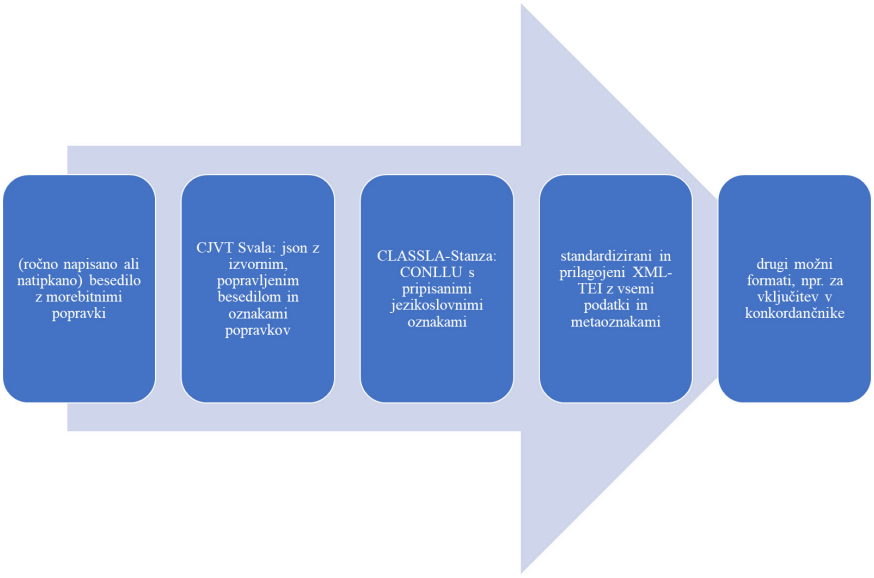
9 Portal SweLL, v katerega je izvorno orodje Svala vključeno (Volodina idr., 2019), skrbi še za vodenje delotokov za urejanje korpusnega gradiva, česar pa za slovenščino trenutno nismo aplicirali.

Način dela z novim orodjem prikazuje Slika 6. Na sliki v gornjem okencu (*izvorno besedilo*) vidimo odstavek avtentičnega besedila iz korpusa Šolar, pod katerim je različica z vpisanimi učiteljskimi popravki. Pod besediloma je t. i. graf povezav, kjer so pojavnice izvornega in popravljenega besedila medsebojno povezane. S klikom na povezavo je mogoče dodati vsebinsko kategorijo jezikovnega popravka, pri čemer se do želene oznake lahko preklikamo s pomočjo menija oznak na levi strani zaslona ali s pomočjo iskalnega okenca nad tem menijem. Primer na sliki kaže popravek besede *gdor – kdor* in pripis oznake črkovanja, specifično za problem menjave med konzonanti *k, g* in *h.* V pomoč pri označevanju so tudi barve – napaka v izvornem besedilu je obarvana rdeče, popravek pa zeleno – in gumbi z ukazi za premik na prejšnjo/naslednjo povezavo, prejšnjo/naslednjo spremembo in za ročno povezavo ali razvezavo neustrezno povezanih pojavnic.

Med urejanjem besedila je mogoče enostavno poskrbeti tudi za anonimizacijo, za kar je v sistemu Šolar predvidena posebna oznaka. Anonimizirati je mogoče s pomočjo kod, npr. *Mirko – XImeX*, ali z uporabo nadomestnih pojavnic, pri čemer je mogoče reproducirati in označiti tudi morebitne jezikovne popravke (npr. *z Mirkotom – z Markom*).

4.2 Jezikoslovno označevanje in korpusni format

Želja in potreba raziskovalne skupnosti je zagotoviti primerljivo jezikoslovno označevanje in standardizirani format temeljnih jezikovnih virov. Za specializirane korpuse, kakršen je Šolar, je ključna metodološko ustrezna povezljivost z referenčnim korpusom, pa tudi drugimi viri iz družine pedagoških korpusov, kamor sodita denimo korpus šolskih učbenikov (Kosem idr., 2022) in mladinske književnosti (Verdonik idr., 2022). Če so korpusni podatki različno označeni in v različnih formatih, so primerjave težje in manj natančne. Treba je torej načrtno skrbeti, da razvojni korpus na ravni jezikovnih oznak in formata sledi standardom, ki se vzpostavljajo v raziskovalnem prostoru, ter da se v primeru novosti tudi ustrezno posodablja.



Slika 7: Cevovod priprave korpusnih besedil za korpuse z označenimi jezikovnimi napakami.

Slika 7 prikazuje trenutni cevovod priprave korpusa Šolar in širše korpusov, ki vsebujejo jezikovne popravke. Proces se prične z besedilom, ki je bodisi ročno napisano ali natipkano. S programom CJVT Svala besedilo uredimo v dve različici, izvorno in popravljeno, ter dodamo oznake popravkov. Tako strukturirani podatki se izvozijo v formatu JSON. Naslednji korak je jezikoslovno označevanje. Trenutno najsodobnejši in najzmogljivejši označevalnik za slovenščino je Classla-Stanza (Terčon & Ljubešić 2023), ki omogoča pripis oznak na številnih nivojih. Po označevanju so datoteke na voljo v formatu CONLLU. Sledi pretvorba v XML TEI, pripravljen posebej za korpuse z jezikovnimi popravki, kjer so korpusna besedila opremljena z oznakami in metapodatki o vrsti in izvoru besedila. Skladno s praksami priprave jezikovnih virov, ki so dostopni prek repozitorija CLARIN.SI, se iz tega formata pripravi različica VERT za vključitev v konkordančnike noSketchEngine in KonText.

Za format TEI[10] smo se odločili že pri pripravi korpusa Šolar 2.0, ki je bil na voljo v različici brez vpisanih popravkov (v celoti

10 Spletna stran iniciative: https://tei-c.org/.

kompatibilen s TEI) in s popravki (prilagojeni TEI). Format, ki je na voljo od korpusa Šolar 3.0 naprej, sledi ločitvi korpusa na tri dele: (jezikoslovno označeno) izvorno besedilo, (jezikoslovno označeno) popravljeno besedilo ter oznake popravkov na spremenjenih delih posameznih povedi. Pri urejanju formata so bile odpravljene težave s segmentacijo napak, ki je predhodno dovoljevala t. i. gnezdene popravke: primere, kjer je bila poleg oznake popravka na določenem segmentu besedila prisotna dodatna oznaka popravka, ki je veljala le za manjši vsebovani del tega segmenta. Gnezdenja popravkov program Svala ne dovoljuje, zato jih tudi novi format ne predvideva. Tovrstne primere, ki so se v različici 2.0 pojavljali v približno 350 odstavkih, smo za Šolar 3.0 ročno popravili in odpravili.

5 Korpus Šolar 3.0

5.1 Sestava korpusa Šolar 3.0

Na projektu je bila pripravljena različica 3.0 korpusa Šolar,[11] ki v vseh pogledih, z izjemo vsebine, prinaša nadgradnjo v primerjavi s prejšnjimi verzijami. Korpus sestavlja 5.485 pisnih izdelkov, ki so jih pri pouku samostojno tvorili učenci slovenskih osnovnih in srednjih šol. Večinoma gre za besedila učencev 7.–9. razreda osnovne šole – vključen pa je tudi manjši vzorec besedil iz 6. razreda – in dijakov vseh letnikov srednje šole. S korpusom torej opazujemo pisno kompetenco šolajoče se populacije starosti 12–18 let.

Vsako besedilo je opremljeno z metapodatki, in sicer: vrsta šole (osnovna ali srednja), predmet, pri katerem je bilo besedilo tvorjeno, razred oz. letnik tvorca besedila, regija, v katero je šola umeščena, in datum nastanka besedila. Del korpusa (2.094 besedil) je označen z učiteljskimi popravki po sistemu oznak, ki ga podrobneje opisujemo v nadaljevanju tega razdelka. Popravki učiteljev so del izvornih pisnih izdelkov učencev, kar pomeni, da odsevajo realno sliko popravljanja šolskih spisov v izobraževalnem procesu.

V Tabelah 1, 2, 3 in 4 predstavljamo vsebino korpusa, pri čemer je vsaka tabela razdeljena v dva dela: v levem, belem delu so

11 Dostopno na http://hdl.handle.net/11356/1589.

predstavljeni podatki za celoten korpus, v desnem, osivenem delu pa podatki samo za besedila z učiteljskimi popravki. Števila in odstotki so vedno podani glede na določeno kategorijo.

Tabela 1 prikazuje razporeditev korpusnih besedil oz. števila besed glede na slovenske regije. Besedila iz severovzhodnih regij (Celje, Maribor, Murska Sobota, Slovenj Gradec) predstavljajo 23,9 % vseh besedil, besedila iz jugozahodnih regij (Gorica, Koper, Kranj, Krško, Ljubljana, Novo mesto, Postojna) pa 76,1 %. Od vseh regij ima ljubljanska regija tako največje število besedil (1495 oz. 27,3 %) kot besed (453,030 oz. 27,7 %). Najslabše zastopani regiji sta murskosoboška z 0,3 % besed in postojnska z 1,7 % besed.

Tabela 1: Število in odstotek besedil ter besed glede na regije v korpusu Šolar 3.0.

Regija	Št. besedil	Odst. besedil	Št. besed	Odst. besed	Št. popravljenih besedil	Odst. popravljenih besedil	Št. besed v poprav. besedilih	Odst. besed v poprav. besedilih
Celje	623	11,4 %	177644	10,9 %	32	0,6 %	11084	0,7 %
Maribor	271	4,9 %	71258	4,4 %	92	1,7 %	27097	1,7 %
Murska Sobota	43	0,8 %	4733	0,3 %	22	0,4 %	3223	0,2 %
Slovenj Gradec	372	6,8 %	97966	6,0 %	102	1,9 %	22313	1,4 %
Gorica	521	9,5 %	263852	16,1 %	321	5,9 %	205477	12,6 %
Koper	111	2,0 %	32898	2,0 %	74	1,3 %	21420	1,3 %
Kranj	380	6,9 %	75524	4,6 %	10	0,2 %	501	0,0 %
Krško	656	12,0 %	205366	12,6 %	147	2,7 %	40637	2,5 %
Ljubljana	1495	27,3 %	453030	27,7 %	467	8,5 %	166221	10,2 %
Novo mesto	924	16,8 %	224862	13,7 %	249	4,5 %	83798	5,1 %
Postojna	89	1,6 %	28274	1,7 %	0	0 %	0	0 %
Skupaj	5485	100 %	1635407	1907562	1516	27,6 %	581771	35,6 %

Tabela 2 prikazuje razporeditev korpusnih besedil in števila besed glede na vrsto šole. Večina besedil prihaja iz različnih vrst srednjih šol, medtem ko osnovnošolska besedila predstavljajo 19,7 % vseh korpusnih besedil oz. 16,3 % besed. Najbolj izstopajo visoki

deželi strokovnih šol in gimnazij, ki predstavljajo 41,2 % besedil in 37,5 % besed oz. 28,2 % besedil in 37,6 % besed. Delež besedil iz poklicnih šol je 9,8 % in predstavljajo 7,2 % besed.

Tabela 2: Število in odstotek besedil ter besed glede na vrsto šole v korpusu Šolar 3.0.

Vrsta šole	Št. besedil	Odst. besedil	Št. besed	Odst. besed	Št. popravljenih besedil	Odst. popravljenih besedil	Št. besed v poprav. besedilih	Odst. besed v poprav. besedilih
Osnovna šola	1081	19,7 %	267146	16,3 %	395	7,2 %	110932	6,8 %
Strokovna šola	2262	41,2 %	613483	37,5 %	574	10,5 %	186809	11,4 %
Poklicna šola	540	9,8 %	117886	7,2 %	143	2,6 %	44878	2,7 %
Gimnazija	1549	28,2 %	615067	37,6 %	404	7,4 %	239152	14,6 %
Neznano	53	1,0 %	21825	1,3 %	0	0 %	0	0 %
Skupaj	5485	100 %	1635407	100 %	1516	27,6 %	581771	35,6 %

Pregled razporeditve besedil in števila besed glede na razred osnovne šole oz. letnik srednje šole, ki ga najdemo v Tabeli 3, prikazuje dokaj uravnoteženo zastopanost. Najbolj izstopa 4. letnik s 25 % besedil oz. 27,9 % besed, kar pa je v skladu s pisno produkcijo, saj je te največ ravno v 4. letniku, ko so tudi besedila daljša. Nižjo zastopanost besedil iz 5. letnika in maturitetnega tečaja lahko pojasnimo s tem, da sta redkeje obiskana.

Tabela 3: Število in odstotek besedil ter besed glede na letnik/razred v korpusu Šolar 3.0.

razred / letnik	Št. besedil	Odst. besedil	Št. besed	Odst. besed	Št. popravljenih besedil	Odst. popravljenih besedil	Št. besed v poprav. besedilih	Odst. besed v poprav. besedilih
6. razred	208	3,8 %	45305	2,8 %	23	0,4 %	7685	0,5 %
7. razred	229	4,2 %	54433	3,3 %	92	1,7 %	22949	1,4 %
8. razred	325	5,9 %	93628	5,7 %	132	2,4 %	43505	2,7 %
9. razred	319	5,8 %	73780	4,5 %	148	2,7 %	36793	2,2 %
1. letnik	1024	18,7 %	317130	19,4 %	427	7,8 %	163610	10,0 %
2. letnik	1018	18,6 %	252775	15,5 %	236	4,3 %	108411	6,6 %
3. letnik	870	15,9 %	308496	18,9 %	252	4,6 %	99299	6,1 %
4. letnik	1373	25,0 %	456196	27,9 %	181	3,3 %	92522	5,7 %
5. letnik	86	1,6 %	21510	1,3 %	25	0,5 %	6997	0,4 %
Maturitetni tečaj	33	0,6 %	12154	0,7 %	0	0 %	0	0 %
Skupaj	5485	100 %	1635407	100 %	1516	27,6 %	581771	35,6 %

Tabela 4 predstavlja razporeditev korpusnih besedil oz. besed glede na tip besedila. Kot lahko vidimo, prevladujejo eseji (58,7 % besedil oz. 77,6 % besed), sledijo pisni izdelki, ustvarjeni pri pouku (15,0 % besedil oz. 6,9 % besed), testi (13,7 % besedil oz. 11,1 % besed) in praktična besedila, napisana za oceno (12,6 % besedil oz. 4,4 % besed).

Tabela 4: Število in odstotek besedil ter besed glede na tip besedila v korpusu Šolar 3.0.

Tip besedila	Št. besedil	Odst. besedil	Št. besed	Odst. besed	Št. popravljenih besedil	Odst. popravljenih besedil	Št. besed v poprav. besedilih	Odst. besed v poprav. besedilih
Pisni izdelek	823	15,0 %	112107	6,9 %	201	3,7 %	31988	2,0 %
Esej	3218	58,7 %	1269793	77,6 %	1280	23,3 %	547169	33,5 %
Praktično besedilo	691	12,6 %	71455	4,4 %	0	0 %	0	0 %
Test	753	13,7 %	182052	11,1 %	35	0,6 %	2614	0,2 %
Skupaj	5485	100 %	1635407	100 %	1516	27,6 %	581771	35,6 %

Korpus je bil jezikoslovno označen s cevovodom CLASSLA v1.1.1[12] na ravneh tokenizacije, stavčne segmentacije, lematizacije, oblikoskladenjskih oznak po sistemu MULTEXT-East v6,[13] odvisnostne skladnje po sistemu JOS-SYN[14] in imenskih entitet.[15] Oznake na nivoju odvisnostne skladnje in imenskih entitet predstavljajo novost v primerjavi z različico 2.0, izboljšana pa je tudi natančnost oznak na ostalih nivojih, saj so bile pripisane z izboljšanim označevalnim orodjem.

5.2 Metodologija označevanja jezikovnih popravkov

Pri procesu odločanja, v katero kategorijo popravkov spada določena težava, so nepogrešljive jasne smernice. Za korpus Šolar 3.0 smo uporabili sistem oznak, ki je bil razvit v različici korpusa 2.0, a smo ga dodatno uredili in nadgradili (Arhar Holdt idr., 2022b). V nadaljevanju predstavljamo osnovno kategorizacijo oznak za jezikovne popravke. Glavnih kategorij je sedem, te pa se hierarhično delijo na podkategorije.

Črkovanje: na to raven uvrščamo učiteljske popravke, ki se nanašajo na zapis glasu ali glasovnega sklopa v besedi. Lahko gre za odvečne, izpuščene ali zamenjane črke (*polen* [laži]* namesto *poln [laži]*; *[je] vrjel** namesto *[je] verjel*; *vstrajen** namesto *vztrajen*) ali črkovne sklope (*zberejejo** namesto *zberejo*; *sprej** namesto *sprejel*; *zastojn** namesto *zastonj*), vzglasje besed na u- oz. v- (*Vsedla** namesto *Usedla*; *uzamejo** namesto *vzamejo*) in variantne predloge (*k* [koncu]* namesto *h*).

Oblika: na ravni oblike označujemo (a) težave na ravni izbire sklona, števila, spola, recimo *[o dekletu, ki je] zanosila** namesto *zanosilo*, in kategorij drugih besednih vrst, (b) popravke besednih oblik, ki niso del standardnih paradigem, recimo *poprimiti** namesto *poprijeti* in (c) dodatne oznake, ki jih pripisujemo le v primeru, da osnovna vsebinska oznaka popravka že obstaja, dodatna oznaka pa

12 https://github.com/clarinsi/classla/
13 https://wiki.cjvt.si/books/04-oblikoskladnja-multext-east
14 https://wiki.cjvt.si/books/06-odvisnostna-skladnja-jos-syn
15 https://wiki.cjvt.si/books/08-imenske-entitete

omogoča združevanje podatkov po drugem kriteriju ali pripis dodatne (poljubne) informacije. Tukaj je denimo informacija o variantnosti besedne oblike, recimo obliki *grada* in *gradu*, ki sta glede na trenutno normo obe legitimni, a je ena nevtralnejša, na kar želi učitelj učenca opozoriti.

Besedišče: na to raven umeščamo popravke besedišča, kar vključuje menjavo ene besede z drugo, pri čemer se lahko besedna vrsta in/ali besednozvezna struktura ohrani ali spremeni. Podkategorije so razdeljene na probleme po besednih vrstah, npr. pri samostalniku ena izmed oznak obeležuje napačno lastno ime (*Lovrenc** namesto *Lovro* [*Kuhar]*), pri glagolu menjavo glagolov *moči-morati* (*[ne bi] moral* [opisati]* namesto *mogel*) ipd. Ločeni so primeri podkategorij z menjavo prek meja besedne vrste (npr. polnopomenske besede v zaimek ali obratno: *Hamlet – on*), zadnja skupina pa prinaša dodatne oznake za zaznamovanost besede, recimo *faks** namesto *fakulteta*.

Skladnja: na tej ravni označujemo popravke, ki posegajo na raven besednozvezne, stavčne in povedne sklanje, npr. popravke besednega reda (*[prepričan je da] generalove ukaze je potrebno* [upoštevati]* namesto *je generalove ukaze potrebno*), skladenjskih struktur (*truplo matere** namesto *materino truplo*), medstavčnih razmerij (*Herod je nekega dne priredil slavje. Salomi slavje* ni bilo preveč po godu.* namesto *Herod je nekega dne priredil slavje, ki Salomi ni bilo preveč po godu.*) itd. Podkategorija dodatne oznake tukaj zaobjema pleonazme, recimo *[Ko se je] vrnila nazaj** namesto *[Ko se je] vrnila*, odvečne, pomensko prazne ali vsebinsko napačne dele.

Zapis: na ravni zapisa označujemo predvsem popravke začetnic (*[v] Nemškem* [jeziku]* namesto *nemškem*) in pisanja skupaj ali narazen (*Nažalost** namesto *Na žalost*). V korpusu so označena tudi mesta napačne stave ločil, kjer prevladuje raba vejice. Skupina ločil je edina, ki ni bila v celoti ročno pregledana in kategorizirana, in sicer zaradi razširjenosti pojava.

Povezani popravki: v to kategorijo uvrščamo vse primere, ki niso samostojen popravek, ampak so posledica primarnega jezikovnega popravka, recimo popravek besedne oblike, ki je le posledica

menjave pred njo stoječega predloga. Da lahko označene podatke ustrezno statistično interpretiramo, je pomembno, da so tovrstni posegi v besedilo ločeni od primarnih popravkov učenčevih jezikovnih izbir. Povezani popravki v osnovi sledijo obstoječi tipologiji, le da del oznake ponazarja, da gre za povezan popravek.

Nečitljivi in sumljivi primeri: posebej so označeni primeri, kjer se v učenčevem besedilu ali učiteljskem popravku pojavlja nečitljiv besedilni fragment, ki ga pri transkripciji ni bilo mogoče interpretirati, recimo *§§§mor§§§*; in primeri, kjer so popravki nenavadni, kjer recimo sumimo, da je prišlo do napake pri transkripciji ali je popravek enak napaki – tem pripišemo oznako za preverbo, ki je začasna in se v končni različici korpusa ne pojavlja.

6 Dostopnost korpusa

Skladno z dobrimi praksami odprtega dostopa do jezikovnih podatkov je korpus Šolar 3.0 kot baza na voljo pod odprto licenco (CC BY-NC--SA 4.0) na repozitoriju CLARIN.SI (Arhar Holdt idr., 2022c). Vključen je tudi v konkordančnike, ki so del infrastrukture CLARIN.SI: KonText, NoSketch Engine Bonito in NoSketch Engine Crystal. Ti konkordančniki omogočajo ločen uvoz (pod)korpusov z izvornimi (‘korpus učenci’) in popravljenimi besedili (‘korpus učitelji’), nato pa v vsaki od različic napredno iskanje, prikaz in izvoz korpusnih podatkov.

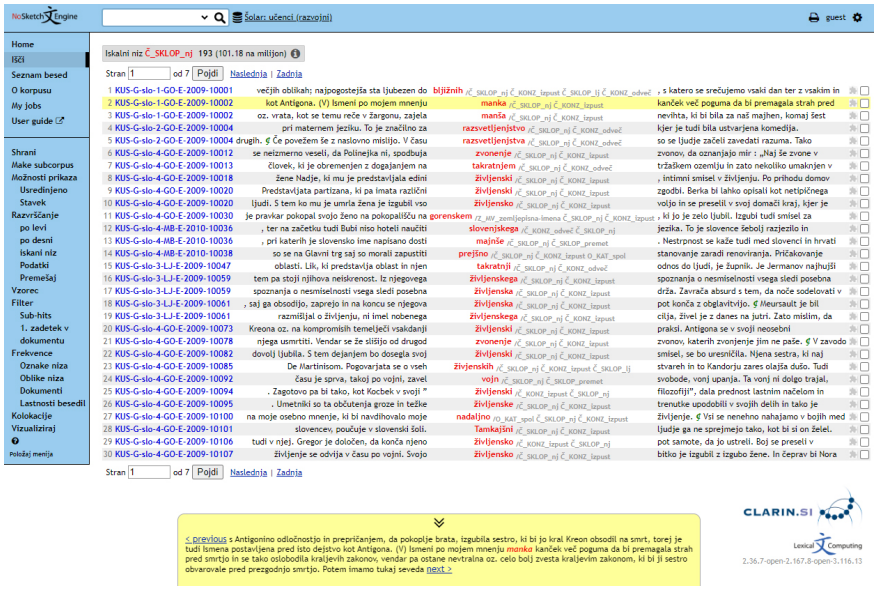
- Šolar 3.0 kot baza: http://hdl.handle.net/11356/1589
- KonText:
  - korpus učenci: https://www.clarin.si/kontext/query?corpname=solar30_orig
  - korpus učitelji: https://www.clarin.si/kontext/query?corpname=solar30_corr
- NoSketch Engine Bonito:
  - korpus učenci: https://www.clarin.si/noske/sl.cgi/first?corpname=solar30_orig&reload=1&iquery=
  - korpus učitelji: https://www.clarin.si/noske/sl.cgi/first?corpname=solar30_corr&reload=1&iquery=

- NoSketch Engine Crystal:
  - korpus učenci:
    https://www.clarin.si/ske/#dashboard?corpname=solar30_orig
  - korpus učitelji:
    https://www.clarin.si/ske/#dashboard?corpname=solar30_corr

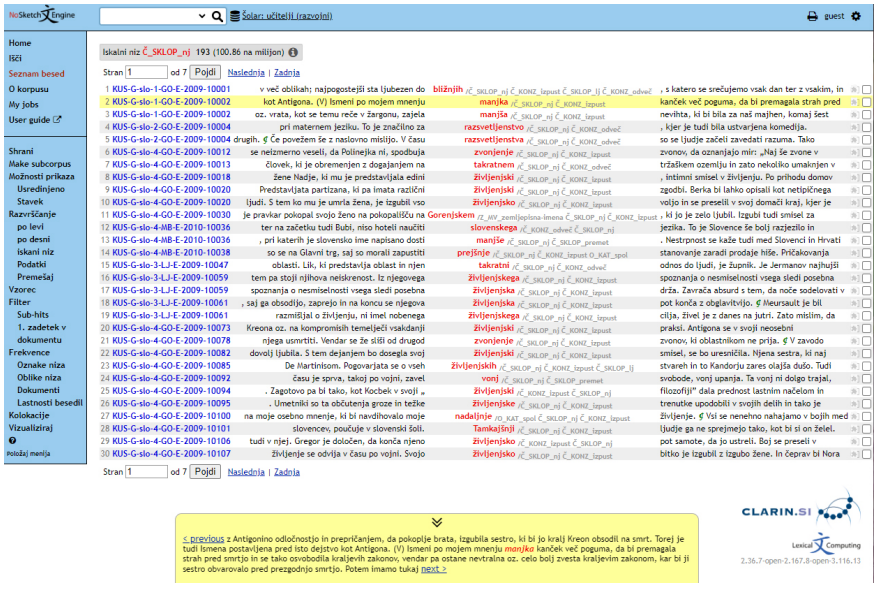
Optimalen in za bodoče delo zaželen bi bil konkordančnik, ki omogoča pregleden skupen prikaz obeh korpusnih različic, vendar že ločena umestitev v zgoraj naštete konkordančnike omogoča številne načine napredne rabe korpusnih podatkov. Osnovna zmogljivost je izdelava konkordančnega niza, pri čemer je mogoče kot parametre iskanja uporabiti raznovrstne v korpusu pripisane oznake. Na Slikah 8 in 9 je za primer prikazan vmesnik NoSketchEngine Bonito, in sicer rezultati iskanja s pomočjo oznake jezikovnega popravka, ki združuje črkovalne težave sklopa *nj*. Kot kaže slika, konkordančnik omogoča enostavno kopiranje zgledov, kar je koristno za pripravo učnih gradiv in vaj. Izvažati je mogoče konkordance, kolokacije, sezname pojavnic in oznak, pri katerih je zlasti ključna možnost primerjave podatkov z drugimi korpusi, ki so vključeni v orodje; kot je bilo omenjeno v Razdelku 4.2, je za korpus Šolar dragocena zlasti možnost primerjave z referenčnim korpusom pisne slovenščine, ki je trenutno Gigafida 2.0 (Krek idr., 2020).

Dodatne možnosti iskanja po jezikovnih oznakah in vizualizacija drevesnic, pripravljenih po sistemu odvisnostne skladnje JOS-SYN, ponuja prostodostopni program Q-CAT. Slika 10 prikazuje iskanje glagolskih oblik, ki so skladenjsko povezane (kot del povedka) z lemo "se". V izbrani povedi so z zeleno prikazana mesta, ki jih je označevalnik prepoznal kot imenske entitete, pod pojavnico so nanizane leme in oznake MSD, z rumenimi povezavami so povezani deli povedka, z zeleno deli besednih zvez (v podrednem in prirednem razmerju), z rdečo pa stavčni členi, pri čemer *ena* grobo ustreza jezikoslovni kategoriji osebka, *dve* predmeta, *tri* določil, ki opredeljujejo lastnosti, in *štiri* ostalih določil, npr. kraja in časa.[16]

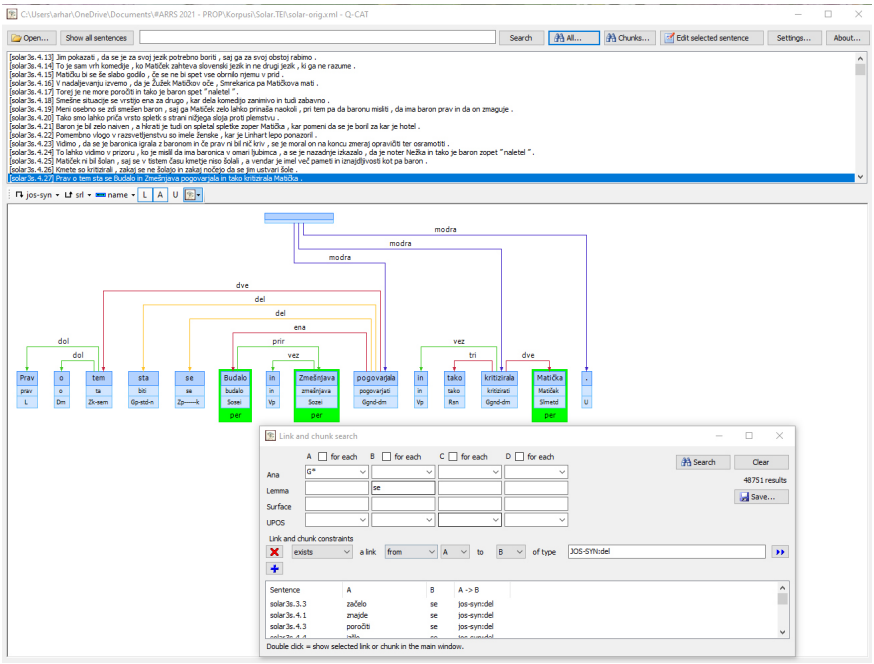
16 Označevalne smernice in predstavitev oznak: https://wiki.cjvt.si/books/06-odvisnostna-skladnja-jos-syn.



Slika 8: Prikaz Šolarja 3.0 – učenci v konkordančniku NoSketchEngine Bonito.



Slika 9: Prikaz Šolarja 3.0 – učitelji v konkordančniku NoSketchEngine Bonito.



Slika 10: Iskanje po skladenjskih oznakah in prikaz označene povedi v programu Q-CAT.

7 Sklep in nadaljnje delo

V prispevku smo predstavili namen in način gradnje razvojnega korpusa za slovenščino. Da bi s pripravo tovrstnih korpusov lahko učinkovito nadaljevali, smo vzpostavili 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 obstoječega korpusa.

Prva prioriteta za nadaljnji razvoj korpusa je njegova vsebinska nadgradnja. Poskrbeti je treba za **povečanje njegovega obsega in reprezentativnosti** po regijah, vrsti šole, razredu/letniku avtorja in predmetu, pri katerem je besedilo nastalo. Komplementarno zasnovi korpusa Šolar je treba dodati zbiranje v smer širjenja korpusne vsebine na eni strani proti pisni tvorbi **v nižjih razredih** in na drugi proti **študentskemu pisanju** (slednje je že vključeno v raziskovalni

projekt ARIS J7-3159,[17] vendar le na ravni razvoja metodologije). Želja je zagotoviti **redno korpusno posodabljanje,** kar pomeni zbiranje, vzorčenje in transkribiranje vsako tretje šolsko leto. Da bi slednje lahko uspelo, je treba **dvigniti ozaveščenost in spodbujati šole** k rednemu sodelovanju. Zahtevane kadrovske kapacitete za takšen kontinuiran razvoj so 1 FTE, ki si ga na letni ravni delita jezikoslovec, ki skrbi za zbiranje in pripravo gradiva, ter tehnični sodelavec, ki skrbi za korpusni format in dostopnost v vseh želenih orodjih.

Druga prioriteta, ki je bila vključena v projekt Nadgradnja korpusov za slovenščino kot drugi in tuji jezik KOST in KUUS,[18] je izboljšati dostopnost in povečati izrabo korpusnih podatkov. Za osnovne korpusne analize je vključitev v konkordančnike CLARIN.SI izrednega pomena, vendar obstoječa orodja ne omogočajo polne izrabe bogato označenega gradiva, ki ga prinaša korpus Šolar. V nadaljevanju je treba razviti **specializirani konkordančnik**, ki bo uporaben za vse korpuse z jezikovnimi popravki. Novi konkordančnik mora biti po zasnovi primerljiv z obstoječim, da se omogoči uporabniški prenos znanja, obenem pa mora imeti dodatne možnosti za izrabo metapodatkov, s pomočjo katerih bi lahko natančneje interpretirali posamezne rezultate iskanj po korpusu. Še bolj nujna je možnost preglednega prikazovanja jezikovnih napak skupaj s popravki, zmogljivo iskanje po izvornih in popravljenih oblikah ter klikljive statistike najpogostejših jezikovnih popravkov. Z razvojem specializiranega konkordančnika bo Šolar postal širše uporaben jezikovni vir, zanimiv 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. Za najširšo možno rabo je treba zagotoviti tudi **izobraževanja učiteljev** o rabi novega korpusa in o izrabi jezikovnotehnoloških virov pri pouku slovenščine (in širše).

Tretja prioriteta je nadaljnji razvoj metodologije zbiranja. Velik časovni prihranek bi ponudila dopolnitev delotokov z **optičnim**

17 Spletna stan projekta: https://www.cjvt.si/prop/.
18 Spletna stran projekta: https://www.cjvt.si/korpus-kost/projekti/.

branjem ročno napisanih besedil, pri čemer bodo potrebne adaptacije za šolsko rabo (kjer so v besedilih prisotne črkovalne napake in druge značilnosti pisanja, ki se razvija) ter natančno pregledovanje ter popravljanje optično prebranih rokopisov. Druga možnost za pohitritev dela je **strojno podprta identifikacija, vpis in kategorizacija učiteljskih jezikovnih popravkov** v ročno napisanih ali digitalnih besedilih – učiteljski popravki so v določeni meri predvidljivi in ponavljajoči se, kar bi bilo mogoče izkoristiti. Tretja možnost za pohitritev postopka je **vključitev množičenja z didaktično perspektivo** v proces korpusnega grajenja. Pri tem bi bilo mogoče sodelovati s predavatelji, ki poučujejo jezikovno didaktiko in sorodne predmete na terciarni stopnji in bi v transkribiranje ter označevanje popravkov vključili študente in študentke, ki se pripravljajo na podajanje jezikovne povratne informacije učencem in dijakom. Množičenje je mogoče organizirati tudi za širšo populacijo, pri čemer pa je treba zagotoviti ustrezno kontrolo kvalitete in motivacijo za sodelovanje.

V projektu smo ugotovili, da je v nadaljevanju treba nekoliko bolje urediti **pretvorbo besedil iz formata JSON,** ki ga uporablja program Svala, ter končnim želenim XML TEI. Izziv je zlasti zapisovanje ločil, ki se za označevanje v Svali ločijo od besednih pojavnic, za končni format pa jih je treba ustrezno stično oz. nestično spet urediti v izvorno obliko, ki je v šolskih besedilih lahko tudi neskladna s trenutnimi jezikovnimi pravili. Nenazadnje, evalvirati je treba, v kolikšni meri jezikovne napake v korpusnih besedilih vplivajo na **natančnost strojnega jezikovnega označevanja** na posameznih označevalnih ravninah, in zagotoviti ustrezne metodološke nadgradnje ali opozorila.

Predvsem pri ciljih, ki se vežejo na metodologijo, je treba slediti **mednarodnim iniciativam, rešitvam in dobrim praksam,** ne le na področju razvojnih korpusov, ampak širše na področju digitalne humanistike, npr. za metodologijo optičnega branja, transkribiranja itd. Stremeti je treba tudi k oblikovanju **mednarodnih standardov** za gradnjo korpusov z jezikovnimi popravki, saj bi to izboljšalo njihovo primerljivost in olajšalo samo uporabo, lažji bi bil tudi prenos znanja in rešitev. Nenazadnje, zagotoviti je treba raziskave, ki bodo

omogočile **sintetične analize in podatke** za pripravo pedagoških učnih gradiv, jezikovnih priročnikov in orodij. Velik potencial predstavlja **primerjava podatkov** iz korpusa Šolar (šolska produkcija) s podatki, ki reprezentirajo šolsko recepcijo (npr. učbeniki, mladinska književnost, uporabniško generirane spletne vsebine), na drugi strani pa primerjava šolskega pisanja v kontekstih, kjer se slovenščina poučuje kot prvi jezik v primerjavi s poučevanjem slovenščine kot drugega/tujega jezika. Kot smo izpostavili v uvodu, bo pojav **postopkov in orodij generativne umetne inteligence** brez dvoma prinesel tudi nove, še nepredvidene izzive in rešitve, zato je toliko bolj ključno, da tudi za slovenščino novim možnostim in spoznanjem karseda hitro sledimo.

Zahvala

Projekt Razvoj slovenščine v digitalnem okolju sta 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). Projekt Empirična podlaga za digitalno podprt razvoj pisne jezikovne zmožnosti (J7-3159) in program Jezikovni viri in tehnologije za slovenski jezik (P6-0411) sofinancira Javna agencija za znanstvenoraziskovalno in inovacijsko dejavnost Republike Slovenije iz državnega proračuna.

Literatura

Abel, A., Glaznieks, A., Nicolas, L., & Stemle, E. (2014). KoKo: An L1 Learner Corpus for German. In Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC 2014), Reykjavik, Iceland (pp. 2414–2421). European Language Resource Association (ELRA).

Arhar Holdt, Š., Kosem, I., & Stritar Kučuk, M. (2022a). Metode in orodja za lažjo pripravo korpusov usvajanja jezika. In N. Pirih Svetina & I. Ferbežar (ur.), *Na stičišču svetov: slovenščina kot drugi in tuji jezik, Obdobja 41* (pp. 23–30). Ljubljana: Založba Univerze v Ljubljani. https://doi.org/10.4312/Obdobja.41.2784-7152

Arhar Holdt, Š., Lavrič, P., Roblek, R., & Goli, T. (2022b). *Kategorizacija učiteljskih popravkov: Smernice za označevanje korpusa Šolar.* Različica 1.1. Rezultat projekta Razvoj slovenščine v digitalnem okolju. https://wiki.cjvt.si/books/11-jezikovni-popravki-solar/page/oznacevalne-smernice

Arhar Holdt, Š., Rozman, T., Stritar Kučuk, M., Krek, S., Krapš Vodopivec, I., Stabej, M., Pori, E., Goli, T., Lavrič, P., Laskowski, C., Kocjančič, P., Klemenc, B., Krsnik, L., & Kosem, I. (2022c). Developmental corpus Šolar 3.0. Slovenian language resource repository CLARIN.SI. http://hdl.handle.net/11356/1589

Arhar Holdt, Š., & Kosem, I. (2023). Šolar, the developmental corpus of Slovene. PREPRINT (Version 1) available at Research Square. doi: 10.21203/rs.3.rs-3274669/v1

Arhar Holdt, Š., Kosem, I., Pori, E., Munda, T., Stritar Kučuk, M., Voršič, I., Petek, T., Šek, P., & Krsnik, L. (2023). *Šolar 3.0: korpus šolskih pisnih besedil: poročilo projekta Razvoj slovenščine v digitalnem okolju: aktivnost DS1.6*. Ljubljana: Univerza v Ljubljani, Center za jezikovne vire in tehnologije, 2023. https://www.cjvt.si/rsdo/wp-content/uploads/sites/18/2023/06/RSDO_Kazalnik_Solar_v2.pdf

Arnardóttir, Þ., Xu, X., Guðmundsdóttir, D., Stefánsdóttir, L. B., & Ingason, A. K. (2021). Creating an error corpus: Annotation and applicability. In M. Monachini & M. Eskevich (Eds.), CLARIN Annual Conference Proceedings (pp. 59–63). Virtual Edition.

Barbagli, A., Lucisano, P., Dell'Orletta, F., Montemagni, S., & Venturi, G. (2016). CItA: An L1 Italian learners corpus to study the development of writing competence. In Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016), Portorož, Slovenia (pp. 88–95). European Language Resources Association (ELRA).

Berkling, K. (2016). Corpus for children's writing with enhanced output for specific spelling patterns (2nd and 3rd grade). In N. Calzolari idr. (Eds.), Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016), Portorož, Slovenia (pp. 3200–3206). European Language Resources Association (ELRA).

Berkling, K. (2018). A 2nd Longitudinal Corpus for Children's Writing with Enhanced Output for Specific Spelling Patterns and Evaluation In Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018), Miyazaki, Japan (pp. 2262–2268). European Language Resources Association (ELRA).

Borghi, C. C. (2013). *Analisi di produzioni scritte. Valutazioni e misure automatizzate di elaborati scolastici. Tesi di dottorato in pedagogia sperimentale*. Università di Roma.

Glaznieks, A., Frey, J. C., Stopfner, M., Zanasi, L., & Nicolas, L. (2022). LEONIDE: A longitudinal trilingual corpus of young learners of Italian, German and English. *International Journal of Learner Corpus Research*, *8*(1), 97–120.

Granger, S. (1998). The computer learner corpus: A versatile new source of data for SLA research. In S. Granger (Ed.), *Learner English on Computer* (pp. 3–18). Addison Wesley Longman.

Ho-Dac, L. M., Fleury, S., & Ponton, C. (2020). É:calm resource: a resource for studying texts produced by French pupils and students. In Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), Marseille, France (pp. 4327–4332). European Language Resources Association (ELRA).

Ingason, A. K., Arnardóttir, Þ., Stefánsdóttir, L. B., & Xu, X. (2021). The Icelandic Child Language Error Corpus (IceCLEC) Version 1.1, CLARIN-IS, http://hdl.handle.net/20.500.12537/133

Kilgarriff, A., Rychlý, P., Smrz, P., & Tugwell, D. (2004). The Sketch Engine. In G. Williams, & S. Vessier (Eds.), Proceedings of the Eleventh EURALEX International Congress, Lorient, France (pp. 105–116). Universite de Bretagne-sud.

Kosem, I., Stritar Kučuk, M., Može, S., Zwitter Vitez, A., Holdt, A., Š., & Rozman, T. (2012). Analiza jezikovnih težav učencev: korpusni pristop. Trojina, zavod za uporabno slovenistiko.

Kosem, I., Rozman, T., Arhar Holdt, Š., Kocjančič, P., & Laskowski, C. A. (2016). Šolar 2.0: nadgradnja korpusa šolskih pisnih izdelkov. In T. Erjavec & D. Fišer (Eds.), Proceedings of the Conference on Language Technologies & Digital Humanities, September 29th – October 1st, 2016 Faculty of Arts, University of Ljubljana, Ljubljana, Slovenia (pp. 95–100). Ljubljana University Press, Faculty of Arts. http://www.sdjt.si/wp/wp-content/uploads/2016/09/JTDH-2016_Kosem-et-al_Solar-2-0-nadgradnja-korpusa-solskih-pisnih-izdelkov.pdf

Kosem, I., Pori, E., Žagar, A., & Arhar Holdt, Š. (2022). Corpus of Slovenian textbooks ccUčbeniki 1.0, Slovenian language resource repository CLARIN.SI. http://hdl.handle.net/11356/1693

Krek, S., Arhar Holdt, Š., Erjavec, T., Čibej, J., Repar, A., Gantar, P., Ljubešić, N., Kosem, & I., Dobrovoljc, K. (2020). Gigafida 2.0: the reference

corpus of written standard Slovene. In N. Calzolari (Ed.), *LREC 2020, Twelfth International Conference on Language Resources and Evaluation, May 11–16, 2020, Palais du Pharo, Marseille, France: conference proceedings* (pp. 3340–3345). Paris: ELRA. http://www.lrec-conf.org/proceedings/lrec2020/LREC-2020.pdf

Laarmann-Quante, R., Dipper, S., & Belke, E. (2019). The making of the Litkey Corpus, a richly annotated longitudinal corpus of German texts written by primary school children. In Proceedings of the 13th Linguistic Annotation Workshop, Florence, Italy (pp. 43–55). Association for Computational Linguistics.

Leech, G. (1997). Teaching and language corpora: A convergence. In A. Wichmann, S. Fliegelstone, T. McEnery, & G. Knowles (Eds.), *Teaching and language corpora* (pp. 1–23). Routledge.

Ljubešić, N., & Dobrovoljc, K. (2019). What does Neural Bring? Analysing Improvements in Morphosyntactic Annotation and Lemmatisation of Slovenian, Croatian and Serbian. In Proceedings of the 7th Workshop on Balto-Slavic Natural Language Processing, Florence, Italy (pp. 29–34). Association for Computational Linguistics.

Marconi, L., Ott, M., Pesenti, E., Ratti, D., & Tavella, M. (1993). *Lessico elementare: dati statistici sull'italiano scritto e letto dai bambini delle elementari*. Zanichelli.

Martins, M., Janssen, M., Santos, T., Lopes, R., & Souza, T. (2020). DOESTE v0.5. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University, http://hdl.handle.net/11234/1-3262

Pala, K., Rychlý, P., & Smrž, P. (2003). Text Corpus with Errors. In V. Matoušek, & P. Mautner (Eds.), *Text, Speech and Dialogue (TSD 2003) Lecture Notes in Computer Science* (2807 vol., pp. 90–97). Springer.

Parr, J. M. (2010). A dual purpose database for research and diagnostic assessment of student writing. *Journal of Writing Research*, *2*(2), 129–150.

Popič, D. (2014). Revising translation revision in Slovenia. In T. Mikolič Južnič, K. Koskinen, & N. Kocijančič Pokorn (Eds.), New Horizons in Translation Research and Education 2 (pp. 72–89). University of Eastern Finland. https://erepo.uef.fi/handle/123456789/14340

Sampson, G. (2003). *The LUCY Corpus: Documentation.* University of Sussex. Retrieved August 15, 2023, from https://www.grsampson.net/LucyDoc.html

Stritar Kučuk, M. (2022). KOST med korpusi usvajanja tujega jezika. V N. Pirih Svetina & I. Ferbežar (ur.), *Na stičišču svetov: slovenščina kot drugi in tuji jezik*, *Obdobja 41* (str. 323–334). Ljubljana: Založba Univerze v Ljubljani. https://centerslo.si/wp-content/uploads/2022/11/Stritar-Kucuk_Obdobja-41.pdf

Stritar Kučuk, M. (2023). *KOST 1.0: Priročnik za označevanje napak, delovna verzija*. https://www.cjvt.si/korpus-kost/wp-content/uploads/sites/24/2022/04/Prirocnik-za-oznacevanje-napak-v-KOST-u-2022-04-13.pdf

Terčon, L., & Ljubešić, N. (2023). CLASSLA-Stanza: The Next Step for Linguistic Processing of South Slavic Languages. *arXiv*. doi: 10.48550/arXiv.2308.04255

Verdonik, D., Majninger, S., Dobrovoljc, K., Antloga, Š., Zögling Markuš, A., Voršič, I., Zemljak Jontes, M., Koletnik, M., Valh Lopert, A., Šek Martük, P., Kosem, I., Majhenič, S., Ferme, M., Žagar, A., Arhar Holdt, Š. (2022). Corpus of Slovenian texts for pedagogical purposes ccMAKS 1.0, Slovenian language resource repository CLARIN.SI. http://hdl.handle.net/11356/1692

Volodina, E., Granstedt, L., Matsson, A., Megyesi, B., Pilán, I., Prentice, J., Rosén, D., Rudebeck, L., Schenström, C. J., Sundberg, G., & Wirén, M. (2019). The SweLL Language Learner Corpus: From Design to Annotation. *Northern European Journal of Language Technology*, *6*, 67–104.

Wirén, M., Matsson, A., Rosén, D., & Volodina, E. (2019). SVALA: Annotation of Second-Language Learner Text Based on Mostly Automatic Alignment of Parallel Corpora. In I. Skadina, & M. Eskevich (Eds.), Selected papers from the CLARIN Annual Conference 2018 (pp. 227–239). Linköping University Electronic Press.