Šolar 2.0: nadgradnja korpusa šolskih pisnih izdelkov

**Iztok Kosem,*† Tadeja Rozman,*† Špela Arhar Holdt,*†
Polonca Kocjančič,* Cyprian Laskowski*‡**

* Zavod za uporabno slovenistiko Trojina,
Trg republike 3, 1000 Ljubljana
iztok.kosem@trojina.si, tadeja.rozman@trojina.si, spela.arhar@trojina.si,
polonca.kocjancic@guest.arnes.si, cyp@trojina.si
† Filozofska fakulteta Univerze v Ljubljani,
Aškerčeva 2, 1000 Ljubljana
‡ Center za jezikovne vire in tehnologije, Univerza v Ljubljani,
Večna pot 113, 1000 Ljubljana

Povzetek

Korpus Šolar je nastal v letih 2009-2010, vanj pa so vključena pisna besedila učencev osnovnih in srednjih šol. Korpus vsebuje skoraj milijon besed, več kot polovica besedil pa ima označene tudi jezikovne popravke učiteljev. V prispevku opisujemo prve korake projekta nadgradnje korpusa Šolar (delovno ime Šolar 2.0), katerega glavna cilja sta povečanje korpusa, kar bo omogočilo boljše posploševanje rezultatov in izvajanje dodatnih raziskav, in pa njegovo uravnoteženje, saj so v trenutni verziji korpusa nekatere slovenske regije slabo zastopane. Predstavljamo rezultate analize obstoječega stanja in zastavljene cilje zbiranja besedil po regijah in šolah. Opišemo tudi postopek digitalizacije, ki bo igral pomembno vlogo v vzpostavitvi dolgoročnega sistematičnega nadgrajevanja korpusa Šolar. V zadnjem delu predstavimo metodo, ki jo bomo uporabili pri reviziji kategorij jezikovnih popravkov.

Šolar 2.0: Increasing the Corpus of Texts Written by Native-speaker Students

The Šolar corpus was built in 2009-2010 and comprises texts written by native-speaker students attending Slovenian elementary and secondary schools. The corpus contains nearly one million words, with over half of the texts also containing teacher corrections of student errors. This paper presents the first steps of the Šolar 2.0 project that aims to expand the corpus, which will enable better generalisations of findings and additional research, and to balance it, given that several Slovenian regions (and their schools) are poorly represented in the existing version of the corpus. We present the analysis of existing corpus structure and goals that were set for further data collection in different regions and schools. Also described is the process of text digitisation, which will play an important role in setting up regular and systematic collection of new texts for the corpus after the end of the project. Finally, a method that will be used for revising the categorization of corrections is presented.

1 Uvod

Korpus šolskih pisnih izdelkov Šolar je nastal v okviru projekta Sporazumevanje v slovenskem jeziku[1] v letih 2009–2010, vsebuje pa skoraj milijon oz. natančno 967.477 besed (Rozman et al., 2012). V korpus je vključenih 2.703 pisnih besedil srednješolcev in učencev zadnjega triletja osnovnih šol (nekaj pa je tudi besedil učencev 6. razreda), ki so jih učenci in učenke samostojno napisali pri različnih predmetih v avtentičnih šolskih situacijah. Korpus Šolar je zato velika pridobitev za slovensko korpusno jezikoslovje, saj ponuja vpogled v pisanje šolajoče se mladine, torej populacije, katere jezikovna produkcija je bila doslej s korpusnim pristopom še neraziskana.

V prispevku predstavljamo projekt nadgradnje korpusa Šolar (delovno ime Šolar 2.0[2]), ki ga sofinancira Ministrstvo za kulturo RS. Projekt poteka v letih 2015–2018, glavna cilja sta povečanje in izboljšanje uravnoteženosti korpusa, ob tem pa želimo odpraviti tudi nekatere pomanjkljivosti pri označevanju jezikovnih popravkov učiteljev, ki jih trenutno vsebuje 56 % besedil v korpusu.

2 Razvojni korpusi

Razvojni korpusi (angl. *developmental corpora*; po Leech 1997:19) so korpusi, ki vsebujejo besedila mlajših maternih govorcev, tj. tistih, ki so še v procesu usvajanja maternega jezika. V primerjavi s korpusi besedil govorcev tujega jezika oz. korpusi usvajanja jezika (angl. *learner corpora*) so razvojni korpusi precej redkejši, vendar pa tako korpusi usvajanja jezika kot razvojni korpusi že dolgo igrajo pomembno vlogo v poučevanju jezika, saj predstavljajo pristop od spodaj navzgor (Osborne, 2002). Rezultati analiz korpusov usvajanja jezika pa so bili uporabljeni tudi v slovarjih, npr. v slovarju Macmillan English Dictionary for Advanced Learners. V nadaljevanju sledi kratek pregled najbolj znanih razvojnih korpusov.

Eden najbolj znanih razvojnih korpusov je CHILDES (Child Language Data Exchange System)[3], baza več kot 130 korpusov (video)posnetkov otroškega govora v 20 različnih maternih jezikih, ki se zbirajo že vse od leta 1981. Polovico korpusov predstavljajo posnetki maternih govorcev angleščine. V bazi je tudi nekaj posnetkov govora otrok z jezikovnimi težavami (npr. disleksijo), tujih govorcev in dvojezičnih otrok.

Bazi CHILDES podobna zbirka korpusov je zbirka projekta EU SACODEYL (2005-2008)[4], ki vsebuje

[1] http://www.slovenscina.eu/
[2] http://solar.trojina.si/
[3] http://childes.psy.cmu.edu
[4] http://www.um.es/sacodeyl

transkribirane (video)posnetke najstniških govorcev angleščine, francoščine, nemščine, italijanščine, litvanščine, romunščine in španščine, starih med 13 in 18 let.

Korpus COLT (Corpus of Teenage Language), ki so ga leta 1993 izdelali na Univerzi v Bergnu, vsebuje 100 posnetkov oz. 50 ur govora 31 najstnikov iz londonskih okrožij, starih med 13 in 17 let (več o projektu gl. v Stenström, Andersen in Hasund 2002). Vseh 500.000 besed v korpusu je bilo ortografsko transkribiranih in oblikoskladenjsko označenih. Korpus je del referenčnega korpusa angleščine BNC (British National Corpus).

Govorjeni jezik otrok vsebuje tudi korpus POW (Polytechnic of Wales), ki so ga izdelali med 1978 in 1984 v južnem Walesu. Korpus vsebuje 65.000 besed, posnetih pa je bilo približno 120 otrok, starih med 6 in 12 let.

Od pisnih razvojnih korpusov sta poznana predvsem korpus LUCY in korpus LOCNESS. LUCY je bil izdelan leta 2003 in je sestavljen iz treh podkorpusov – korpusa besedil objavljenih avtorjev (za pričujoči pregled ni relevanten), korpusa besedil "mlajših odraslih" in korpusa besedil otrok. Korpus mlajših odraslih sestavljajo besedila gradiv za maturo, seminarske naloge in eseji študentov prvega letnika – skupaj 48 besedil oz. 33.000 besed. V korpusu otrok je 150 besedil oz. 30.000 besed, avtorji besedil pa so otroci, stari med 9 in 12 let.

Korpus LOCNESS je bil izdelan na Univerzi v Louvainu in vsebuje pisne izdelke (argumentativne in literarne eseje) dijakov in študentov, maternih govorcev angleščine. Korpus vsebuje 324.304 besede, od tega 60.209 besed predstavljajo eseji britanskih dijakov na maturi, 95.695 besed eseji britanskih študentov, 168.400 besed pa eseji ameriških študentov.

Od "neangleških" korpusov usvajanja maternega jezika velja omeniti korpus jezika tajvanskih otrok (TCLC). Gre za korpus govorjene tajvanščine, ki je bil izdelan v obdobju med 1997 in 2000 in vsebuje 300 ur posnetkov oz. 1,6 milijona besed.

Za slovenski prostor je od tujih korpusov usvajanja maternega jezika najrelevantnejši korpus Chyby, polmilijonski korpus češčine, ki vsebuje besedila (eseje in uvode diplomskih nalog) čeških univerzitetnih študentov (Bušta et al. 2009). Povprečna dolžina besedil je od 600 do 700 besed. Korpus Chyby je eden redkih korpusov usvajanja maternega jezika, ki ima označene napake tvorcev besedil. Te temeljijo na popravkih učiteljev, klasifikacija napak, ki je bila izdelana vnaprej, pa je zaradi predpostavke, da tuji in materni govorci delajo podobne napake, podobna tistim v korpusih usvajanja češčine kot tujega jezika.

Čeprav je razvojnih korpusov manj kot korpusov usvajanja jezika in posledično obstaja tudi precej manj na razvojnih korpusih temelječih raziskav in gradiv[5], pa raziskave, kot so Andersen (1997), Stenström, Andersen in Hasund (2002), Rowland et al. (2005), Bušta et al. (2009), v slovenskem prostoru pa na korpusu Šolar temelječe raziskave Kosem et al. (2012) ter Arhar Holdt in Rozman (2015a), nakazujejo na velik potencial razvojnih korpusov pri spoznavanju jezikovne produkcije mlajših maternih govorcev, njihovih težav pri sporazumevanju ter navsezadnje pri poučevanju jezika ter izdelavi didaktičnih gradiv in orodij (Arhar Holdt et al., in print). Posledično je nadvse smiselno še naprej razvijati razvojne korpuse slovenščine, kot je Šolar, tako z vidika njihove velikosti kot tudi raznolikosti besedil v njih.

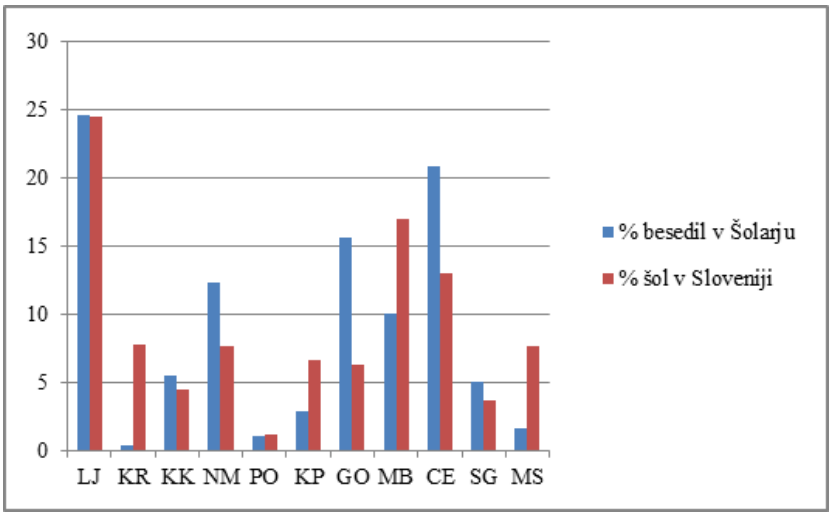
3 Velikost in uravnoteženost korpusa Šolar

Cilj projekta Šolar 2.0 je, da bi korpus Šolar povečali na približno 2 milijona besed oz. na novo dodali od 2500 do 3000 besedil. Večji korpus namreč pomeni večjo zanesljivost in uporabnost, saj po eni strani omogoča lažje pospleševanje rezultatov jezikovnih analiz, omogoča raziskave, ki jih je na manjšem gradivu težko opravljati (npr. analize o rabi polnopomenskega besedišča, kolokacijah, vezljivosti, skladnji), hkrati pa nudi večji nabor avtentičnega gradiva za pripravo npr. jezikovnih priročnikov in didaktičnih gradiv.

Povečanje je tudi priložnost za uravnoteženje korpusa. Pri gradnji prvotne verzije je bil glavni cilj doseči regijsko uravnoteženost, tj. vključitev približno 60 % besedil iz regij jugozahodnega dela Slovenije in 40 % besedil iz regij severovzhodnega dela.[6] Pri nadgradnji korpusa pa želimo bolje uravnotežiti tudi razmerje med posameznimi regijami, saj je analiza vključenega gradiva pokazala, da smo – glede na razmerje deležev šol po posameznih regijah – v korpus vključili ustrezno število besedil iz ljubljanske, krške in postojnske regije, iz slovenjgraške, novomeške, goriške in celjske je besedil nekoliko preveč, premalo pa je besedil iz kranjske, koprske, mariborske in murskosoboške regije (Graf 1).

[5] Tako za korpuse usvajanja jezika obstaja zelo obsežna literatura, kot se kaže v več kot 1.100 vnosov obsegajoči bibliografiji združenja raziskovalcev, ki se ukvarjajo z korpusi usvajanja jezika (Learner Corpus Association). Bibliografija je dostopna na http://www.learnercorpusassociation.org/resources/lcb/.

[6] Regije so določene z registrskimi območji, kot jih določa 3. člen Pravilnika o registrskih tablicah motornih in priklopnih vozil (Uradni list RS, št. 83/2006, http://www.uradni-list.si/1/objava.jsp?urlid=200683&stevilka=3637). Med JZ regije sodijo LJ, KR, KK, NM, PO, KP, GO, med SV regije pa MB, CE, SG in MS.



Graf 1: Delež besedil v Šolarju v primerjavi z deležem šol po regijah.

Poleg tega bo pri nadgradnji potrebno vključiti večje število osnovnošolskih besedil, saj je delež besedil iz osnovnih šol bistveno premajhen: v korpusu Šolar je trenutno le 18,6 % osnovnošolskih besedil, čeprav je osnovnih šol v Sloveniji približno 75 %. Zavedamo sicer se, da so šole različno velike in da deleža besedil v korpusu nima smisla pretirano uravnavati glede na število šol, vendar pa je razmerje med deležem osnovnih in deležem srednjih šol v regijah dobra orientacija pri načrtovanju nadgradnje. Idealno bi bilo, če bi po posameznih regijah lahko dosegli razmerja, v katerih bi bilo osnovnošolskih besedil od 20 % do 30 %, čeprav bo, sodeč po trenutni sestavi korpusa, to razmeroma težko doseči (Tabela 1 in Graf 2).[7]

	OŠ		**SŠ**		**skupaj**	
	Šolar	Šolar 2.0	Šolar	Šolar 2.0	Šolar	Šolar 2.0
LJ	52	954	615	369	**667**	**1323**
KR	12	306	0	117	**12**	**423**
KK	83	198	65	45	**148**	**243**
NM	40	324	295	90	**335**	**414**
PO	0	45	28	18	**28**	**63**
KP	0	270	77	90	**77**	**360**
GO	137	252	286	90	**423**	**342**
MB	0	693	271	225	**271**	**918**
CE	0	522	563	180	**563**	**702**
SG	136	153	0	45	**136**	**198**
MS	43	342	0	72	**43**	**414**
skupaj	**503**	**4059**	**2200**	**1341**	2703	5400

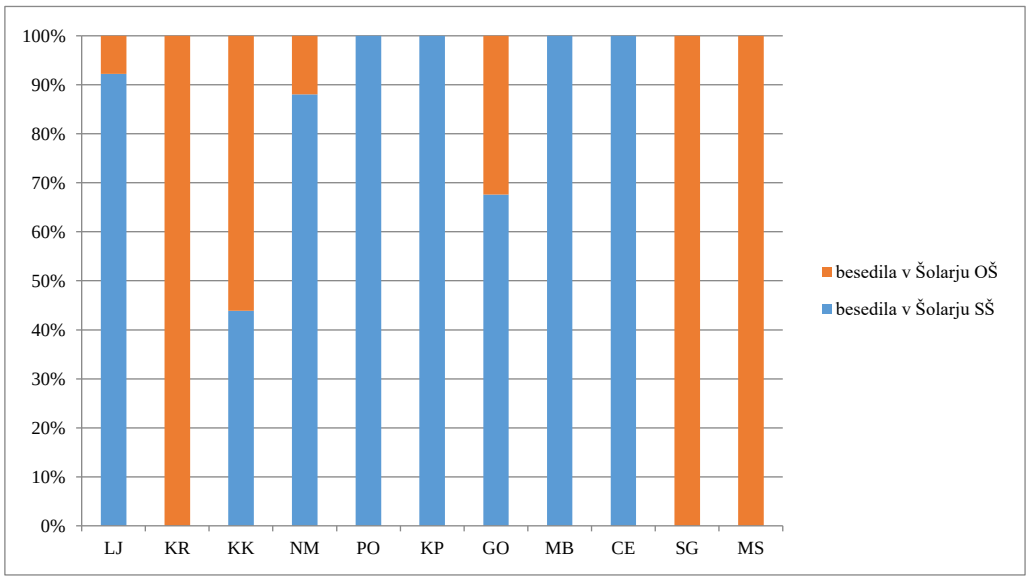
Tabela 1: Število besedil v Šolarju po stopnjah in regijah v primerjavi z idealnim številom besedil v načrtovanem Šolarju 2.0 glede na razmerja šol po stopnjah in regijah.

4 Dodajanje novih besedil

V korpus bomo skušali vključiti čim več besedil, ki smo jih prejeli pri gradnji korpusa Šolar v šolskem letu 2009/2010 in v korpus niso bila vključena. Teh besedil je 5891, kar bi bilo dovolj za načrtovano povečanje korpusa, vendar pa s temi besedili nikakor ne moremo doseči želene uravnoteženosti po regijah in stopnjah šolanja. Prednost pri vključitvi bodo zato imela besedila, ki bodo pripomogla k boljši uravnoteženosti korpusa, s šolskim letom 2016/2017 pa začenjamo tudi nov krog zbiranja besedil po šolah, ki bo bomo, tudi zaradi časa in truda, ki smo ga in ga še bomo skupaj z učitelji vložili v zbiranje besedil, skušali v čim večji meri izogniti.

[7] Dejansko bi idealna razmerja ob korpusu s 5400 besedili imela za posledico to, da bi morali odstraniti nekatera besedila iz prve verzije (zlasti iz ljubljanske in mariborske regije), čemur pa se

trajal dve leti. K sodelovanju bomo povabili osnovne šole iz vseh regij, hkrati pa bo potrebno dobiti tudi srednješolska besedila iz kranjske, slovenjgraške in murskosoboške regije, ki jih zaenkrat v Šolarju ni (Graf 2), saj jih pri gradnji korpusa Šolar sploh nismo prejeli.



Graf 2: Deleži osnovnošolskih in srednješolskih besedil po regijah.

Zbiranje besedil bo potekalo podobno kot v šolskem letu 2009/2010 (Rozman et al., 2012): potrebno bo dobiti soglasja učencev oz. njihovih staršev za vključitev besedil v korpus, učitelji in učiteljice bodo vsa besedila označili z metapodatki (gl. poglavje 3.1), nato sledi skeniranje besedil, kar je novost, saj smo pri gradnji korpusa Šolar besedila fotokopirali (gl. poglavje 4). Nova je tudi razširitev zbiranja na besedila 6. razreda, ki jih je v Šolarju le za vzorec; za to smo se odločili, da bi dobili več osnovnošolskih besedil, saj učitelji, ki učijo v zadnji triadi, učijo tudi učence 6. razredov, po drugi strani pa z nižanjem starostne meje avtorjev besedil dobimo boljši vpogled v razvoj pisne jezikovne zmožnosti. V prihodnosti bi zato bilo smiselno korpus razširiti tudi z besedili učencev na razredni stopnji, kar v okviru tega projekta sicer ni predvideno.

Novost, ki do sedaj ni bila vključena zaradi predvidenih možnih težav pri zbiranju gradiva, je tudi informacija, ali ima učenec odločbo o specifičnih učnih primanjkljajih na področju branja in pisanja. V primeru, da bomo ta podatek od učiteljev oz. učencev lahko pridobili, ga bomo vključili v različico korpusa za raziskovalce, javno pa ne bo dostopen. Čeprav na tak način zbranih besedil morda ne bo nujno dovolj za izdelavo uravnoteženega podkorpusa, pa bo zbiranje omogočilo preizkus postopka za gradnjo tovrstnega vira in prve korake pri uporabi gradiva za statistične analize jezikovnih težav in zmožnosti učencev s primanjkljaji v primerjavi s preostalo populacijo. Gradivo bo omogočilo vpogled v posredovane povratne informacije učencem in dijakom s primanjkljaji, in s tem dragoceno avtentično gradivo za izobraževanja bodočih učiteljev (Arhar Holdt in Rozman, 2015b).

4.1 Metapodatki o besedilih

Vsako besedilo mora biti označeno z naslednjimi metapodatki, sicer ni primerno za vključitev v korpus:

- šola, naslov šole in učitelj, ki je besedilo prispeval (ta podatek je shranjen v elektronskem arhivu in v korpusu ni viden);
- program: OŠ, gimnazija, SSI, SPI, PTI, NPI (trenutno so v Šolarju srednješolski programi razdeljeni na gimnazije, strokovne šole, kamor so uvrščeni programi SSI, ter poklicne šole, kamor so uvrščeni programi SPI, PTI in NPI, vendar razmišljamo, da bi zaradi večje jasnosti kategorije preimenovali, kategorijo s poklicnimi programi pa razdružili);
- razred oz. letnik (tukaj sta v obrazcu, ki ga morajo učitelji izpolniti, predvidena tudi poklicni in maturitetni tečaj);
- predmet;
- šolska situacija/besedilna vrsta: ker poimenovanja besedilnih vrst v Šolarju niso najboljša, saj povzročajo težave pri interpretaciji,[8] smo za namene zbiranja določili naslednje kategorije:[9]
    - esej/spis: eseji, spisi in ostala daljša besedila (npr. domišljijski dnevniški zapisi, lastne basni, pravljice ipd.), ki so nastala v okviru šolske naloge (testna situacija),
    - praktično besedilo: vabila, prošnje, opravičila ipd., ki so nastala v testni situaciji (so torej napisana v šoli pri pouku slovenščine za oceno),

[8] V korpusu so besedilne vrste razdeljene na: esej/spis, pisni izdelek (učna ura), test (daljše besedilo), test (odgovori na vprašanja). Kategorije so opredeljene v Rozman et al. (2012).

[9] Verjetno bomo kategorije spremenili tudi v Šolarju.

  - test: test z odgovori na (esejska) vprašanja; test mora vsebovati vsaj dve esejski vprašanji oz. vprašanji, ki zahtevata nekoliko daljši odgovor (kot test se označijo tudi testi pri slovenščini, ki vsebujejo tako vprašanja kot tvorjenje praktičnega besedila),
  - delo v razredu: vsa besedila, ki so nastala pri pouku v netestni situaciji (torej niso za oceno), učitelji dopišejo besedilno vrsto;
- regija;
- šolsko leto.

Učitelji morajo tudi s podpisom potrditi, da dovolijo vnos učiteljskih popravkov v korpus. Zbiramo sicer tako besedila s popravki kot besedila brez popravkov, kljub temu da vnos popravkov v okviru projekta ni predviden.

5 Digitalizacija

Na začetku projekta smo največjo pozornost posvetili optimizaciji procesa zbiranja in pretvorbi besedil v obliko, primerno za vključitev v korpus, saj smo pri izdelavi prve verzije korpusa opazili precej možnosti za izboljšavo in pohitritev. Med ključnimi odločitvami je bila tudi ta, da bomo vsa na roko napisana besedila digitalizirali, tako že zbrana kot tista, ki jih bomo pridobili v prihodnje. Digitalna oblika besedila, tj. sken (pa tudi že na računalniku natipkana besedila), je za nas izhodiščna, saj v primerjavi s fotokopijami omogoča enostavnejše arhiviranje, boljšo in hitrejšo organizacijo transkripcije, hitrejšo izsledljivost izvirnika in navsezadnje tudi boljšo čitljivost besedil oz. njihovih delov (pri izdelavi prve verzije so nam npr. učitelji pošiljali črno-bele kopije, na katerih so bili sicer barvni učiteljski popravki včasih težko razločljivi od učenčevega besedila).

V začetnih mesecih smo tako pripravili in preizkusili postopek digitalizacije besedil, ki so bila pridobljena v šolskem letu 2009/2010 in v korpus še niso bila vključena. V testnem obdobju smo z digitalizacijo pridobili 1651 dokumentov v formatu .pdf, ki so že primerna za vključitev v interni repozitorij in nadaljnjo obdelavo, tj. transkripcijo. Opravila, ki jih predvideva digitalizacija, so naslednja: pregled in ureditev posameznega snopiča z besedili (priprava za skeniranje, ki je odvisna tudi od lastnosti skenerja), določitev identifikacijske številke besedila, evidentiranje metapodatkov, označitev vseh strani besedila z identifikacijsko številko, priprava metapodatkov za skeniranje (t. i. »nosilna stran« ali zbirnik, ki spremlja digitalizirano skupino besedil), skeniranje in lokalno shranjevanje datotek, kompresija datotek, združevanje večstranskih izdelkov v eno datoteko ter pretvorba .tiff v .pdf. Tako pripravljene dokumente nadalje prenesemo v interni repozitorij.

Pri digitalizaciji upoštevamo tudi posebnosti besedil in določene dodatne kriterije, saj so vhodna gradiva dokaj raznolika. S posameznih šol so prišla organizirana v snopiče, ki so lahko tudi notranje razdeljeni na več podsnopičev - pri digitalizaciji to informacijo ohranjamo. Gradiva so lahko enostranska ali dvostranska, formata A4 ali A3. V veliki večini gre za fotokopije, le majhen odstotek gradiv so originali, ki so pogosto neenotnega formata. Večstranska gradiva so pogosto speta. Pri digitalizaciji je poleg naštetega eden najpomembnejših kriterijev ta, da en izdelek enega avtorja postane en dokument oziroma datoteka. To je pomembno upoštevati zlasti takrat, kadar je na eni strani več vsebinsko nepovezanih izdelkov (lahko so različni tudi avtorji) ali pa se konča en izdelek ter začne naslednji. Če je vhodno gradivo fotokopija ali besedilo brez učiteljskih popravkov, digitaliziramo sivinsko. Če gre za original z učiteljskimi popravki, pa skeniramo barvno, saj bodo informacije za transkriptorje tako popolnejše oziroma lažje dostopne. Priprava korpusa predvideva tudi anonimizacijo. Nekateri učitelji so jo izvedli že sami, preostale izdelke pa bomo anonimizirali ob transkripciji besedila. Anonimizacija zajema zakrivanje podatkov v besedilu (metapodatkov o besedilu, kot so ime in priimek avtorja in ime šole, sploh ne beležimo), ki bi lahko razkrili avtorja besedila, npr. imen in/ali priimkov avtorja, njegovih sorodnikov ali prijateljev, imena šole, kraja ipd.

Na podlagi testne digitalizacije smo opravili tudi časovne izračune in pripravili dokument s podrobnim opisom postopka ter predstavitvijo posebnih primerov. Dolgoročni načrt je, da v korpus vključimo vsa zbrana besedila, v času pisanja konferenčnega prispevka pa že pripravljamo spletni repozitorij.

6 Pripis kategorij jezikovnih popravkov

Pomemben del korpusa Šolar so označeni jezikovni popravki učiteljev, ki omogočajo različne raziskave (npr. Arhar Holdt in Rozman, 2015a; Kosem et al., 2012), prav tako pa so pomembni za izdelavo jezikovnih tehnologij, jezikovnih priročnikov in didaktičnih gradiv (npr. Pedagoški slovnični portal[10]). Pomanjkljivost, ki so jo zainteresirani uporabniki večkrat izpostavili, je odsotnost podrobnejše kategorizacije napak šolarjev. Obstoječe kategorije so trenutno namreč zelo splošne (besedišče, oblika, zapis, skladnja – pri čemer imata zapis in skladnja še podkategorije), kar otežuje direktno uporabo korpusa v razredu in urjenje orodij in programov na korpusu Šolar za namene, kot je npr. avtomatska prepoznava napak.

V korpusu Šolar trenutno najdemo 35.035 učiteljskih jezikovnih popravkov, od tega največ na ravni zapisa (61,1 %), sledijo popravki skladnje (17,7 %), besedišča (10,9 %) in oblike (10,3 %). Podrobnejša analiza popravkov je bila opravljena pri izdelavi Pedagoškega slovničnega portala (Kosem et al., 2012), v sklopu katere so bili popravki tudi ročno razvrščeni v približno 700 kategorij jezikovnih težav. Obstoječa kategorizacija se izkazuje za problematično s treh vidikov: (I) Kategorizacijo je izvajalo več označevalcev, vsak na posamezni jezikovni ravnini in po principu od spodaj navzgor. Rezultat so medravninsko deloma različni sistemi kategoriziranja, ki so do sedaj ostajali razdruženi. (II) Ker je bila kategorizacija v veliki meri ciljno usmerjena v pripravo portala, so določene vrste popravkov, npr. popravki ločil, ostali le delno obravnavani. (III) Pripisane kategorije popravkov še niso bile umeščene v XML-strukturo korpusa, torej ostajajo nedosegljive za sintetične jezikoslovne analize. Našteti problemi so botrovali odločitvi o natančnejši reviziji in nadgradnji sistema kategorij ter vpisu le-teh med oznake korpusa Šolar. Pri slednji nalogi se izkazuje za poseben izziv zagotavljanje možnosti označevanja določenega popravka z več različnimi kategorijami (npr. popravek oblike *uplivat* v *vplivati,* ki sodi tako v napake črkovanja –

[10] http://slovnica.slovenscina.eu/

popravek zapisa ụ v besednem vzglasju – kot tudi na raven morfologije – popravek kratkega nedoločnika).

Za učinkovito revizijo obstoječih kategorij je ključna uporaba sistematičnega in fleksibilnega postopka, ki omogoča pregledovanje večjih količin označenih konkordanc in sprotne prekategorizacije označenih napak v njih. V ta namen smo preizkusili orodja, izdelana bodisi predvsem za namene korpusov z jezikovnimi napakami bodisi za namene pripisovanja različnih oznak (npr. TEITOK, WebAnno). Na koncu smo se odločili za uporabo orodja Sketch Engine (Kilgarriff et al., 2004), ki podpira pregleden prikaz jezikovnih popravkov in ponuja možnost pripisovanja ter shranjevanja uporabniško določenih oznak za prikazane konkordance.

Korpus s pripisanimi revidiranimi kategorijami jezikovnih popravkov bomo v drugi polovici projekta uporabili za izdelavo učnega korpusa, namenjenega razvoju sistema za avtomatsko kategorizacijo učiteljskih popravkov in na drugi strani šolskemu pisanju prilagojenih jezikovnotehnoloških orodij, kot so denimo črkovalniki in slovnični pregledovalniki.

7 Sklep

V okviru projekta Šolar 2.0 bomo obstoječi korpus izboljšali in nadgradili, kar bo dobrodošlo tako za raziskovalce in jezikovne tehnologe kot učitelje in učence. Večji korpus, boljša uravnoteženost po regijah in stopnjah šolanja ter dodane podkategorije bodo omogočili nove vpoglede v pisno produkcijo učencev in s tem tudi izdelavo problemsko naravnanih jezikovnih priročnikov in učnih gradiv za osnovne in srednje šole, osnovanih na analizah dejanske jezikovne rabe, ki jih v slovenskem prostoru močno primanjkuje. Poleg tega želimo v okviru projekta vzpostaviti postopek zbiranja in digitalizacije besedil, ki bo omogočal dolgoročno sistematično nadgrajevanje korpusa Šolar.

Korpusi, ki bodo nastali v okviru projekta, bodo na voljo konec leta 2018, in sicer bosta korpus Šolar 2.0 in učni korpus z jezikovnimi popravki na voljo pod licenco CC BY-NC-SA 2.5 SI[11] (gre za licenco, pod katero je na voljo tudi trenutna verzija korpusa Šolar), novo zbrana besedila v korpusu Šolar pa najbrž tudi kot ločen korpus pod licenco CC BY 4.0[12].

8 Literatura

Gisle Andersen. 1997: Pragmatic markers in teenage and adult conversation. *18. konferenca ICAME*. Neobjavljeni prispevek.

Špela Arhar Holdt, Iztok Kosem in Polona Gantar. In print. Corpus-based resources for L1 teaching: The Case of Slovene. V: *Handbook on Digital Learning for K-12 Schools*. Springer.

Špela Arhar Holdt in Tadeja Rozman. 2015a. Možnosti uporabe podatkov iz korpusa Šolar za pripravo slovarskih priročnikov. V: M. Smolej, ur., *Obdobja 34: Slovnica in slovar - aktualni jezikovni opis*, 1. del, str. 67–74. Znanstvena založba Filozofske fakultete UL. http://centerslo.si/wp-content/uploads/2015/11/34_1-Arhar-Hol-Roz.pdf.

Špela Arhar Holdt in Tadeja Rozman. 2015b. Korpus Šolar: gradivni vir za raziskave pisne produkcije slovenskih učencev. *Bilten društva Bravo*, XI/23: 17-23.

Jana Bušta, Dana Hlaváčková, Miloš Jakubíček, Karel Pala. 2009. Classification of Errors in Text. V: P. Sojka, A. Horák, ur., *Proceedings of Recent Advances in Slavonic Natural Language Processing, RASLAN 2009*. str. 109–119. Brno: Masaryk University. http://nlp.fi.muni.cz/raslan/2009/papers/6.pdf.

Adam Kilgarriff, Pavel Rychlý, Pavel Smrz in David Tugwell. 2004. The Sketch Engine. V: G. Williams in S. Vessier, ur., *Proceedings of the Eleventh EURALEX International Congress, EURALEX 2004 Lorient, France July 6–10, 2004,* str. 105–116. Lorient: Universite de Bretagne - sud.

Iztok Kosem, Mojca Stritar Kučuk, Sara Može, Ana Zwitter Vitez, Špela Arhar Holdt in Tadeja Rozman. 2012. *Analiza jezikovnih težav učencev: korpusni pristop*. Trojina, zavod za uporabno slovenistiko.

Geoffrey Leech. 1997: Teaching and language corpora: A convergence. V: A. Wichmann, S. Fliegelstone, T. McEnery in G. Knowles, ur., *Teaching and language corpora,* str. 1–23. London: Longmann.

John Osborne. 2002. Top-down and bottom-up approaches to corpora in language teaching. Connor, Ulla and Thomas A. Upton (ur.): *Applied Corpus Linguistics: A Multidimensional Perspective,* str. 251–265. Amsterdam: Rodopi.

Caroline F. Rowland, Julian M. Pine, Elena V. Lieven, Anna L. Theakston. 2005. The incidence of error in young children's Wh-questions. *Journal of speech, language and hearing research* 48/2:384–404.

Tadeja Rozman, Irena Krapš Vodopivec, Mojca Stritar Kučuk in Iztok Kosem. 2012. *Empirični pogled na pouk slovenskega jezika*. Trojina, zavod za uporabno slovenistiko.

Anna-Brita Stenström, Gisle Andersen, Ingrid Kristine Hasund. 2002. *Trends in Teenage Talk: Corpus compilation, analysis and findings.* Studies in corpus Linguistics 8. Amsterdam: John Benjamins.

[11] https://creativecommons.org/licenses/by-nc-sa/2.5/si/

[12] https://creativecommons.org/licenses/by/4.0/