NINA LEDINEK – MATEJA JEMEC TOMAZIN – MITJA TROJAR – ANDREJ PERDIH – JANOŠ JEŽOVNIK – MIRO ROMIH – TOMAŽ ERJAVEC

KORPUS ŠOLSKIH BESEDIL SLOVENSKEGA JEZIKA: ZASNOVA IN GRADNJA

COBISS: 1.01
HTTPS://DOI.ORG/10.3986/JZ.28.1.07

V prispevku je predstavljen *Korpus šolskih besedil slovenskega jezika*, specializirani pisni korpus slovenščine v obsegu približno 1,8 milijona pojavnic. Korpus je bil zasnovan v okviru projekta *Franček, Jezikovna svetovalnica za učitelje slovenščine in Šolski slovar slovenskega jezika*, in sicer kot gradivska osnova za oblikovanje *Šolskega slovarja slovenskega jezika*, prvega znanstveno utemeljenega pedagoškega slovarja za slovenski jezik. Prispevek obravnava besedilnotipsko sestavo in obseg korpusa, osvetljuje tehnične postopke predpriprave besedil in njihovega jezikoslovnega označevanja ter predstavlja nabor korpusnih metapodatkov, hkrati pa pojasnjuje, v katerih formatih in pod katerimi licencami je *Korpus šolskih besedil slovenskega jezika* na voljo. Članek opozarja tudi na pravne vidike pridobivanja besedil.
Ključne besede: korpus šolskih besedil, šolski slovar, TEI, odprti dostop, urejanje avtorskih pravic

The Corpus of Slovenian School Texts: Design and Creation
This article presents the *Corpus of Slovenian School Texts*, which is a specialized corpus of written Slovenian containing around 1.8 million tokens. It was designed within the scope of the project *Franček, Language Advising Service for Teachers of Slovenian and the Slovenian School Dictionary*, and it was intended to provide language material for compilation of *Šolski slovar slovenskega jezika* (Slovenian School Dictionary), the first research-based school dictionary of Slovenian. The article discusses the text type composition and size of the corpus, sheds light on technical procedures in text preprocessing and corpus annotation, and presents the set of corpus metadata. It also explains in which formats and under what licenses the *Corpus of Slovenian School Texts* has been made available, and also draws attention to legal aspects of obtaining texts.
Keywords: school text corpus, school dictionary, TEI, open access, copyright

1 UVOD

V prispevku je predstavljen *Korpus šolskih besedil slovenskega jezika*, pisni korpus slovenščine, ki služi kot gradivska osnova za pripravo *Šolskega slovarja slovenskega jezika*. Slovenski jezikoslovci so že večkrat izrazili potrebo po kakovostnem šolskem slovarju slovenskega jezika (Weiss 1994: 350; 2001; Stabej idr. 2008; Rozman 2010; 2012; Čebulj 2013; Godec Soršak 2015; 2019; Rozman idr. 2015), na opisano vrzel v slovenskem jezikovnem opisu pa opozarja tudi *Resolucija o nacionalnem*

Prispevek je nastal v okviru projekta *Portal Franček, Jezikovna svetovalnica za učitelje slovenščine in Šolski slovar slovenskega jezika*, ki sta ga sofinancirala Republika Slovenija in Evropski socialni sklad, v okviru raziskovalnih programov P6-0038 in P5-0408, ki ju financira ARRS, ter v okviru raziskovalne infrastrukture CLARIN.SI.

programu za jezikovno politiko 2021–2025. Že pred dvema desetletjema so sicer nastali štirje slovarji za različne starostne stopnje šolarjev, tj. *Moj mali slovar* (MMS 1996), *Moj slovar* (MS 2000), *Moj prvi slovar* (MPS 2002) in *Besede nagajivke* (BN 2002), ki pa zlasti zaradi skromnega obsega (največ slovarskih sestavkov, 1.021, vsebuje MMS) ter premajhne leksikografske doslednosti in premišljenosti (Godec Soršak 2015; 2019: 280) niso povsem izpolnjevali zastavljenih ciljev.

Da bi zapolnili to vrzel v naboru slovenskih slovarjev, je na Inštitutu za slovenski jezik Frana Ramovša ZRC SAZU v okviru projekta *Spletni portal Franček, Jezikovna svetovalnica za učitelje slovenščine in Šolski slovar slovenskega jezika* (v nadaljevanju: projekt Franček) po začetnih konceptualnih pripravah (Godec Soršak 2019) na osnovi izvirnega koncepta (Petric Žižić 2020) začel nastajati *Šolski slovar slovenskega jezika*. Gre za prvi znanstveno utemeljeni pedagoški slovar za slovenski jezik, njegovi ciljni uporabniki pa so učenci od 1. do 5. razreda osnovne šole. V prvo različico slovarja je vključenih približno 2000 slovarskih sestavkov, v prihodnje pa se bo dopolnjeval z novimi sestavki. Slovar je od septembra 2021 objavljen na prosto dostopnem portalu *Franček* (www.franček.si), prvem interaktivnem pedagoškem spletnem slovarsko-slovničnem portalu za slovenščino,[1] namenjenem osnovno- in srednješolcem (Ahačič idr. 2021b; Perdih idr. 2021; Ježovnik – Kenda-Jež – Škofic 2020; Perdih 2021).

Šolski slovar slovenskega jezika gradivsko temelji na specializiranem pisnem korpusu slovenščine, poimenovanem *Korpus šolskih besedil slovenskega jezika* in namensko zgrajenem v okviru projekta Franček. V korpusu so zbrana besedila, ki v največji meri odslikavajo jezikovno realnost, relevantno za šolski pomenski opis. Vključuje sodobna besedila treh tipov: besedila šolskih učbenikov, izvirno leposlovje za otroke in šolska besedila, ki so jih oblikovali učenci.

Za slovenščino sicer že obstaja korpus *Šolar* (Kosem – Rozman – Stritar 2011; Kosem idr. 2016), ki prav tako vključuje besedila šolarjev, vendar je prvenstveno namenjen odkrivanju tipičnih napak, ki jih v knjižnem jeziku delajo šolajoči se, in težavnih mest v slovnici. *Korpus šolskih besedil slovenskega jezika* je v nasprotju s tem oblikovan za leksikografske potrebe. Od referenčnih korpusov *Gigafida 1.0* in *2.0* (Logar Berginc idr. 2020; Krek idr. 2020) se razlikuje v tem, da vključuje besedila, ki nagovarjajo ciljno skupino uporabnikov šolskega slovarja, zato so v njem v večjem deležu zastopane besedilne enote, ki izkazujejo zlasti pomene, relevantne za učence. Obenem je v korpusu veliko laže najti razumljive, nazorne

---

1 Slovenski osnovno- in srednješolci so do objave portala *Franček* lahko uporabljali zlasti slovarje, primarno namenjene odraslim rojenim govorcem jezika. Do njih so v zadnjih letih dostopali predvsem prek najpomembnejšega slovenskega slovarskega portala *Fran* (www.fran.si; Ahačič – Ledinek – Perdih 2015), ki ga omenja večina novejših šolskih učbenikov za slovenski jezik, njegovo rabo pa spodbuja tudi Zavod za šolstvo RS. Ker številne raziskave ugotavljajo, da je neprilagojenost jezikovnih virov mladim uporabnikom ena temeljnih ovir za njihovo zgodnjo rabo pri šolskem pouku (Kosem idr. 2012; Rozman idr. 2020), je bila glavna motivacija za oblikovanje portala *Franček* ravno želja po prilagoditvi obstoječih slovarskih in drugih virov šolski populaciji.

in z didaktičnega vidika ustrezne zglede rabe, ki lahko služijo kot ponazarjalno gradivo za šolski slovar.

V nadaljevanju prispevka sta podrobneje opisana zasnova in nastajanje *Korpusa šolskih besedil slovenskega jezika*. Predstavljena sta besedilnotipska sestava in obseg korpusa, opisani so tehnični postopki predpriprave besedil za korpus in njegovo jezikoslovno označevanje ter nabor korpusnih metapodatkov. Opozarjamo tudi na nekatere pravne vidike pridobivanja besedil za korpus.

2 O KORPUSU ŠOLSKIH BESEDIL SLOVENSKEGA JEZIKA

Korpus šolskih besedil slovenskega jezika je specializirani pisni korpus slovenskega jezika, primarno oblikovan za potrebe pedagoške leksikografije. Vključuje besedila, ki so izhodiščno namenjena učencem nižjih razredov osnovnih šol ali pa so jih napisali učenci, ki obiskujejo 1. do 5. razred osnovne šole.

2.1 Sestava, zapis in dostopnost korpusa

V korpus so uvrščena sodobna slovenska besedila treh tipov, in sicer:

1. šolski učbeniki za predmete, ki se poučujejo od 1. do 6. razreda osnovne šole,
2. izvirno slovensko leposlovje za otroke,
3. šolska besedila učencev in dijakov vzgojno-izobraževalnih zavodov, vključenih v projekt Franček.

Izmed učbeniških besedil so bila uporabljena le tista, ki ustrezajo veljavnim učnim načrtom. Besedila učencev so nastala v obdobju trajanja projekta Franček, torej med letoma 2017 in 2021, večina med letoma 2018 in 2020. Leposlovje za otroke vključuje sodobna izvirna slovenska otroška in mladinska literarna dela uveljavljenih avtorjev; od del, ki so nastala že v preteklih desetletjih, so bila v korpus vključena tista, ki sodijo med klasiko slovenske otroške in mladinske literature. Takih del je v korpusu malo, gre pa za pravljice ali zbirke (ljudskih) pravljic, kot so *Babica pripoveduje*, *Dvanajst ujcev*, *Hvaležni medved*, ali avtorska dela kot *Živilja in škarjice*, *Kdo je napravil Vidku srajčico* ipd. Načrtovano je bilo, da bi korpus vseboval tudi otroško periodiko (otroške in mladinske revije, kot so *Ciciban*, *Pil*, *Moj planet*), vendar zaradi avtorskopravnih omejitev vključitev teh besedil ni bila mogoča.

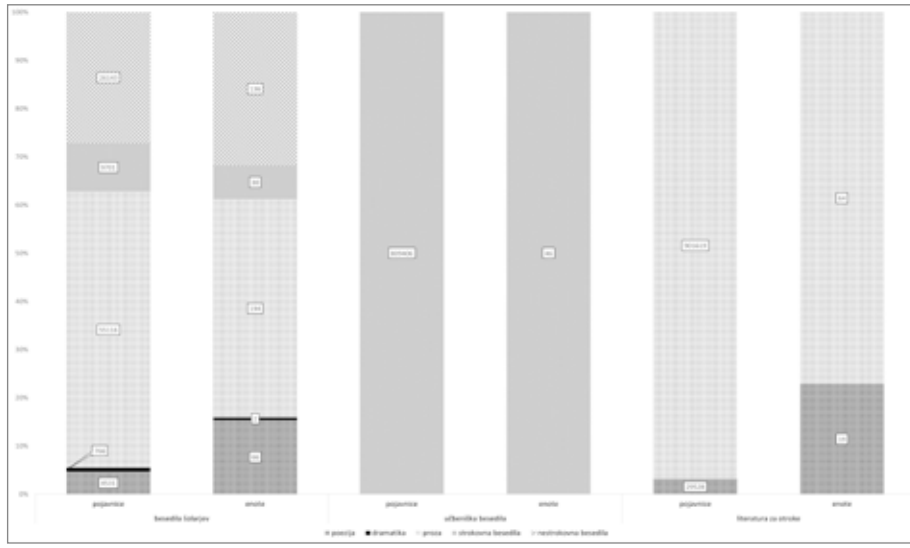
Korpus je zapisan v dveh formatih, in sicer v označevalnem jeziku XML v skladu s smernicami iniciative TEI (TEI Consortium 2017[2]), poleg tega pa tudi v t. i. vertikalni obliki, primerni za vključitev v konkordančnike, kot sta (No)SketchEngine (Rychlý 2007; Kilgarriff idr. 2014) ali KonText (Machálek 2020). Obe različici se zaradi avtorskopravnih omejitev razlikujeta ne le v zapisu, ampak tudi po vsebinski plati. Kot odprto dostopna podatkovna zbirka, zapisana v označevalnem jeziku XML TEI, je na voljo manjši del celotnega korpusa. Sestavlja ga 428 be-

2 https://github.com/clarinsi/TEI-schema

sedil (96.257 pojavnic, 7161 povedi), ki so jih oblikovali učenci in dijaki vzgojno-izobraževalnih zavodov. Na voljo je na repozitoriju raziskovalne infrastrukture CLARIN.SI pod licenco CC-BY 4.0 (Ahačič idr. 2021a).[3]

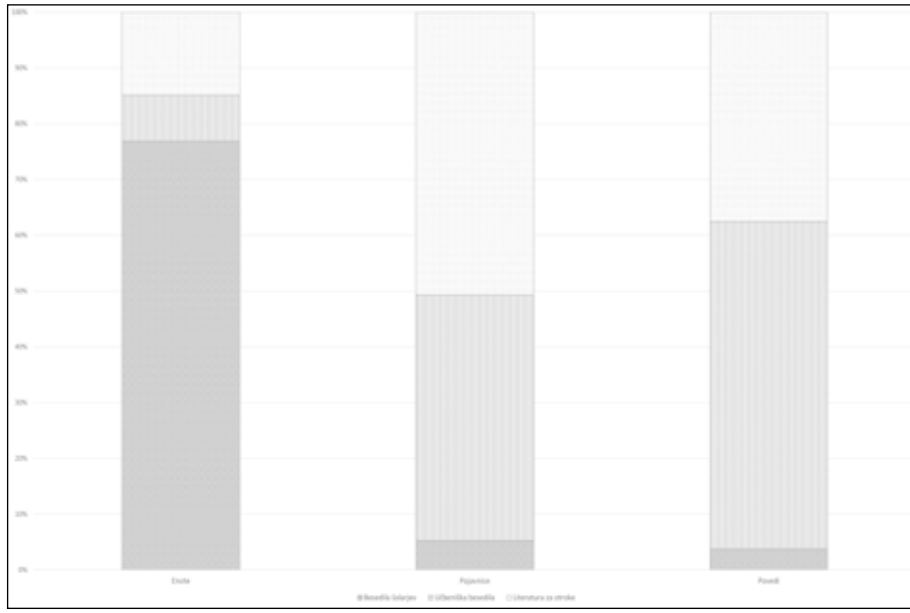
Celoten *Korpus šolskih besedil slovenskega jezika*, ki obsega 557 besedil (1.836.810 pojavnic, 191.779 povedi), je zapisan v vertikalni obliki. Uporabnikom je na voljo za iskanje prek konkordančnikov NoSketch Engine in KonText v okviru raziskovalne infrastrukture CLARIN.SI,[4] in sicer pod oznako SBSJ. Najobsežnejši del korpusa predstavljajo leposlovna besedila za otroke (83 besedil, 931.147 pojavnic, 72.090 povedi), sledijo učbeniška besedila, ki obsegajo 46 učbenikov (809.406 pojavnic, 112.528 povedi), najmanj obsežen del korpusa pa predstavlja že omenjenih 428 besedil šolarjev vzgojno-izobraževalnih zavodov, vključenih v projekt Franček. Razmerja v obsegu posameznih tipov besedil v *Korpusu šolskih besedil slovenskega jezika* so prikazana v spodnjih grafih in preglednicah, in sicer obseg posameznega podkorpusa glede na število vključenih besedilnih enot, število pojavnic in število povedi (graf 1), delež vključenih enot in pojavnic glede na tip besedila in njegove podzvrsti (graf 2) ter absolutna in relativna frekvenca pojavnic glede na oznake besednih vrst v celotnem korpusu ter znotraj posameznih podkorpusov (preglednica 1 oz. graf 3).

Graf 1: Velikost posameznega podkorpusa (metapodatek *tip besedila*)[5] glede na število vključenih besedilnih enot, število pojavnic in število označenih povedi

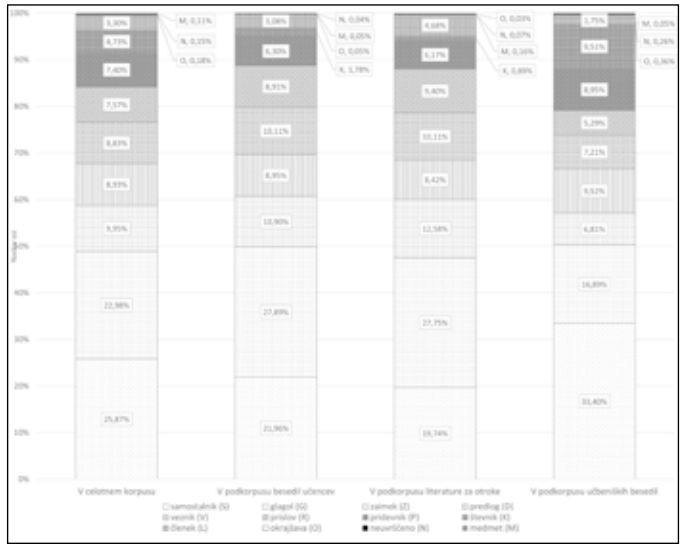


3 http://hdl.handle.net/11356/1413
4 https://www.clarin.si/info/konkordance/
5 Za pregled metapodatkov prim. 2.2.2.

Graf 2: Delež pojavnic in besedilnih enot glede na tip besedila in njegove (pod)zvrsti



Graf 3: Relativna frekvenca pojavnic glede na besedno vrsto (brez ločil (U))



Preglednica 1: Absolutna in relativna frekvenca pojavnic v celotnem korpusu ter znotraj posameznega podkorpusa (metapodatek *tip besedila*) (frekvenca ločil (U) ni upoštevana)[6]

		Absolutna in relativna frekvenca			
Besedna vrsta	**Oznaka**	**Celoten korpus**	**Besedila šolarjev**	**Učbeniška besedila**	**Literatura za otroke**
samostalnik	S	383.781 25,87 %	17.636 21,96 %	148.196 33,40 %	217.949 19,74 %
glagol	G	340.967 22,98 %	22.401 27,89 %	208.343 16,89 %	110.223 27,75 %
zaimek	Z	147.681 9,95 %	8.752 10,90 %	94.475 6,81 %	44.454 12,58 %
predlog	D	132.530 8,93 %	7.193 8,95 %	63.245 9,52 %	62.092 8,42 %
veznik	V	131.037 8,83 %	8.121 10,11 %	75.890 7,21 %	47.026 10,11 %
prislov	R	112.292 7,57 %	7.159 8,91 %	70.606 5,29 %	34.527 9,40 %
pridevnik	P	109.774 7,40 %	5.058 6,30 %	46.293 8,95 %	58.423 6,17 %
številnik	K	70.141 4,73 %	1.427 1,78 %	6.687 9,51 %	62.027 0,89 %
členek	L	48.999 3,30 %	2.471 3,08 %	35.126 1,75 %	11.402 4,68 %
okrajšava	O	2.603 0,18 %	40 0,05 %	242 0,36 %	2.321 0,03 %
neuvrščeno	N	2.233 0,15 %	29 0,04 %	529 0,26 %	1.675 0,07 %
medmet	M	1.570 0,11%	38 0,05 %	1.185 0,05 %	347 0,16 %
Skupaj		**1.483.608**	**80.325**	**750.817**	**652.466**

2.2 Predpriprava in normalizacija besedil ter jezikoslovno označevanje korpusa

Zbiranje besedil za *Korpus šolskih besedil slovenskega jezika* je potekalo v več fazah. Učbeniška besedila in literarna besedila, namenjena otrokom, smo pridobili od založb, besedila učencev in dijakov pa smo zbirali preko namensko oblikovanega spletnega vmesnika (postopek zbiranja in pretvorbe je natančneje opisan v razdelku 3.2), zato so bila zapisana v različnih izhodiščnih formatih (.pdf, .docx in .indd). Postopek pretvorbe besedil je bilo treba v prvi fazi oblikovanja korpusa prilagoditi njihovemu izhodiščnemu zapisu.

Besedila, pridobljena od založb, so bila najprej ročno pregledana, pri čemer smo določili njihove strukturne dele (kolofoni, seznami uporabljene strokovne li-

6 Absolutna frekvenca pojavnic, označenih z oznako U (ločila), je sicer 353.202 v celotnem korpusu ter 15.932, 180.330, 156.940 znotraj posameznih podkorpusov, kot si sledijo v preglednici.

terature, seznami uporabljenih kratic in krajšav, seznami naslovov slik in grafov, paginacija, glave in noge besedil, daljši tujejezični deli besedil ipd.), ki so bili v nadaljnjih fazah pretvorbe iz korpusa sistematično izločeni. S tem smo želeli zmanjšati možnost, da bi bile v korpusu kot visoko frekventne zastopane besede oz. leme, ki v besedilih za otroke vsebinsko niso relevantne (npr. *založba*, *kolofon*, *foto, Rokus* ipd.). Besedila različnih formatov so bila pretvorjena v enotni golobesedilni format s kodiranjem UTF-8. Pri strojni pretvorbi je (lahko) prihajalo do napak: zaradi likovnih in grafičnih elementov, ki pogosto dopolnjujejo otroška in mladinska besedila, je bilo takih napak več kot pri pretvorbi besedil za odrasle. Besedila so bila zaradi tega pred jezikoslovnim označevanjem korpusa ročno pregledana in popravljena. Do napak, ki bi lahko pomembno vplivale na uspešnost strojnega oblikoskladenjskega označevanja in lematizacije, je prihajalo zlasti zaradi napačne segmentacije besedila na povedi in odstavke. Ročno so bile npr. popravljene napake v strojni pretvorbi v primerih, ko je bila ena poved razdeljena v dva različna odstavka, zaradi česar statistični označevalnik analiziranih enot najverjetneje ne bi ustrezno prepoznal in bi lahko prihajalo do napak pri pripisu oblikoskladenjske oznake in leme. Pri pregledu so bili, če tak postopek pri strojni pretvorbi ni bil uspešen, iz besedil ročno odstranjeni kolofoni, odvečna paginacija, glave in noge besedil ipd.

Pri pretvorbi leposlovja za otroke v golobesedilni format so bili praviloma izgubljeni deli besedil, ki so predstavljali sestavne dele grafičnih elementov (ilustracij, fotografij, stripovskih elementov). Grafični elementi se v procesu optične razpoznave znakov (OCR) namreč niso pretvorili v besedilo. Le manjši del slikovnega gradiva, ki je vsebovalo tudi besedilo (npr. slike grafitov), je bil ročno prepisan. Nekaj literarnih besedil se v korpusu pojavi več kot enkrat (npr. ista pesem je lahko vključena v več različnih zbirk ali učbenikov), ker pa je delež takih enot zelo majhen, se za postopek deduplikacije korpusa nismo odločili.

Vsa besedila za *Korpus šolskih besedil slovenskega jezika* so bila po pretvorbi v golobesedilni format avtomatsko tokenizirana, lematizirana in oblikoskladenjsko označena po označevalnem modelu JOS[7] (Erjavec – Krek 2008; Erjavec idr. 2010), in sicer z označevalnikom Obeliks[8] (Grčar – Krek – Dobrovoljc 2012). Za označevanje s tem označevalnikom smo se odločili, da bi zaradi načrtovanih primerjalnih analiz dosegli čim večjo kompatibilnost oznak nastajajočega korpusa s korpusom *Gigafida 1.0*, torej z najaktualnejšim referenčnim korpusom, ki je bil za slovenščino na voljo v času zasnove Korpusa šolskih besedil slovenskega jezika.[9]

---

7 http://nl.ijs.si/jos/josMSD-sl.html.

8 http://razclenjevalnik.slovenscina.eu/Programska_oprema.aspx.

9 Korpus *Gigafida 2.0*, danes največji in najnovejši referenčni korpus za slovenščino, je bil javnosti predstavljen sredi leta 2019.

2.3 Izbira metapodatkov za korpus

V korpus vključena besedila so dokumentirana z naslednjimi metapodatki:

1. *avtor besedila* (pri besedilih šolarjev ima atribut vrednost »ni podatka«),
2. *naslov besedila*,
3. *leto objave besedila*,[10]
4. *založnik besedila* (pri besedilih šolarjev ima atribut vrednost »ni podatka«),
5. *tip besedila* – možne vrednosti so: »besedila šolarjev«, »učbeniška besedila«, »literatura za otroke«,
6. *zvrst besedila* – možni vrednosti sta: »umetnostna besedila«, »neumetnostna besedila«; glede na izbrano vrednost atributa je mogoče izbrati še vrednosti atributa *podzvrst besedila*:[11]
   - pri umetnostnih besedilih so možne vrednosti: »poezija«, »proza«, »dramatika«,
   - pri neumetnostnih besedilih so možne vrednosti: »strokovna besedila«, »nestrokovna besedila«,
7. *razred*[12] – npr. »1. razred«, »2. razred« ipd. (pri literaturi za otroke ima atribut vrednost »ni podatka«),
8. *šolski predmet* – npr. »matematika«, »slovenščina«, »zgodovina«, »geografija« ipd. (pri literaturi za otroke ima atribut vrednost »ni podatka«),
9. *spol avtorja besedila* – možne vrednosti atributa so: »ženski«, »moški«, »drugo« (pri literaturi za otroke in pri učbeniških besedilih ima atribut vrednost »ni podatka«).

Besedila, vključena v korpus, so besedilnotipsko raznolika. V vertikalni različici korpusa je zato dodan atribut *oznaka*, ki je dejansko sestavljen iz dveh ali treh različnih že vključenih metapodatkov. Atribut je zamišljen kot privzeti atribut za prikaz podatkov o besedilu pri prikazu konkordance v konkordančniku (No)Sketch Engine. Atribut *oznaka* sestavljajo naslednji tipi metapodatkov:

1. pri učbeniških besedilih: tip besedila, razred in šolski predmet,
2. pri literaturi za otroke: naslov besedila in leto njegove objave,
3. pri besedilih šolarjev: tip besedila, razred in šolski predmet.

---

10 Gre za leto objave konkretne izdaje besedila, ne pa za podatek o prvi objavi besedila sploh.

11 Učbeniška besedila imajo vedno atribut *neumetnostna*, in sicer *strokovna*; literatura za otroke je vedno označena z atributom *umetnostna* (v korpusnem gradivu se pri tem tipu besedila pojavljajo le tista z atributoma *pesniška* in *prozna*).

12 Atributa *razred* in *šolski predmet* pri besedilih, označenih z atributom *besedila šolarjev*, pomenita razred oz. šolski predmet, pri katerih je besedilo nastalo, pri besedilih, označenih z atributom *učbeniška besedila*, pa razred in šolski predmet, pri katerih se učbenik uporablja pri pouku. Podatek je zanimiv zlasti z vidika učbeniških besedil, saj lahko z iskanjem po učbeniških besedilih za posamezni razred npr. spremljamo, katera strokovna poimenovanja se v besedilih za posamezni razred pojavljajo.

Ob gradnji korpusa oz. njegovih različic smo zgolj za tehnične namene uporabljali še interne metapodatke, npr. identifikacijsko številko posameznega besedila, različico korpusa (v skladu s projektno prijavo so nastale štiri vmesne različice korpusa, ena za vsako koledarsko leto trajanja projekta Franček) in dostopnost (odprti dostop : prosti dostop), ki pa zaradi nerelevantnosti v končno različico korpusa niso vključeni. Z vidika jezikoslovnih raziskav besedil bi bil zanimiv še kak dodaten metapodatek, npr. o tem, v katerem delu Slovenije živi (oz. od kod izvira) in koliko je star šolar, ki je avtor besedila, vendar pa vseh tovrstnih podatkov v korpus ne moremo vključiti hkrati, saj bi bili posamezniki lahko prepoznani, ščitijo pa jih zakonski predpisi o varovanju zasebnosti mladoletnih oseb.

Metapodatke za posamezna pridobljena besedila smo v podatkovno zbirko v tabelaričnem formatu vnašali ročno, le metapodatki besedil šolarjev so bili v podatkovno zbirko vneseni avtomatsko (gl. razdelek 3.2).

3 PRIDOBIVANJE BESEDIL

Besedila smo pridobili na dva načina. Učbeniška besedila in izvirno leposlovje za otroke smo prejeli od večjih slovenskih založnikov (zlasti Mladinske knjige in založbe Rokus Klett), besedila šolarjev smo zbirali s pomočjo učiteljev v vzgojno-izobraževalnih zavodih, sodelujočih v projektu Franček.

3.1 Pridobivanje gradiva pri založbah

Pri oblikovanju strategije pridobivanja učbeniških in leposlovnih besedil smo morali izhodiščno odgovoriti na vprašanje, od koga bomo besedila poskusili pridobiti: ali jih bomo iskali pri avtorjih, ki jih omenjajo učni načrti, šolski bralni seznami (npr. za bralno značko) in druge bralne spodbude, ali pri založbah, ki s svojo uredniško politiko pomembno vplivajo na slovensko mladinsko literaturo. Zaradi učinkovitejše organizacije dela smo za besedila najprej zaprosili založbe. Te so bile pripravljene sodelovati, vendar smo morali za leposlovje za otroke sami pridobiti vsa potrebna dovoljenja avtorjev, saj založbe v pogodbah z avtorji praviloma ne urejajo prenosa materialnih avtorskih pravic na tretje osebe.

Veljavni zakon o zaščiti avtorske in sorodnih pravic (ZASP, Ur. l. RS 16/07, 68/08, 110/13, 56/15, 63/16 in 59/19)[13] namreč omejuje razpolaganje založb s pravicami avtorjev tudi tako, da je treba natančno zapisati, katere pravice se v kolikšnem obsegu, za kateri namen in na kakšen način prenašajo na založbo.[14] To pomeni, da založbe uredijo prenose za svoje potrebe, torej za objavo v reviji ali učbeniku oziroma za izdajo knjige, ne smejo pa besedil brez novega dovoljenja avtorjev odstopiti tretjim osebam, torej niti raziskovalnim ustanovam. To je od

13 Povezava na čistopis iz leta 2007: http://pisrs.si/Pis.web/pregledPredpisa?id=ZAKO403.

14 Pred leti se je v tovrstnih besedilih pogosto pojavila dikcija: »[...] vse pravice enkrat za vselej prenaša [...]«

sodelavcev projekta Franček terjalo pridobivanje soglasij posameznih avtorjev, dedičev ali njihovih zakonitih zastopnikov, le manjši del besedil je že bil v javni lasti.[15] Nekateri avtorji so svoja dela izdali v samozaložbi in smo lahko vse pravice uredili neposredno z njimi.

Pridobivanje soglasij je trajalo približno tri mesece. V tem času smo imeli s posameznimi avtorji več usklajevalnih sestankov. Mnogo avtorjev se je na dopis, ali dovolijo uporabo besedil, odzvalo z naklonjenostjo, manjši del avtorjev pa je kljub pojasnilom dovoljenje za vključitev besedil v korpus odklonil. Pogodbe o odstopu pravic so pripravili pravniki, specializirani za avtorsko pravo, poleg tega pa so sodelovali še pravniki, ki so pripravljali dokumente o varovanju osebnih podatkov in politiko zasebnosti. Čeprav si raziskovalne infrastrukture, kot je CLARIN.SI ali na evropski ravni CLARIN ERIC, prizadevajo za odprtost raziskovalnih podatkov tudi s pripravo standardiziranih pogodb in dogovorov za odstop pravic v raziskovalne namene, je bila za pripravo dokumentov pri konkretnem projektu vseeno potrebna pomoč zunanjih pravnih svetovalcev.

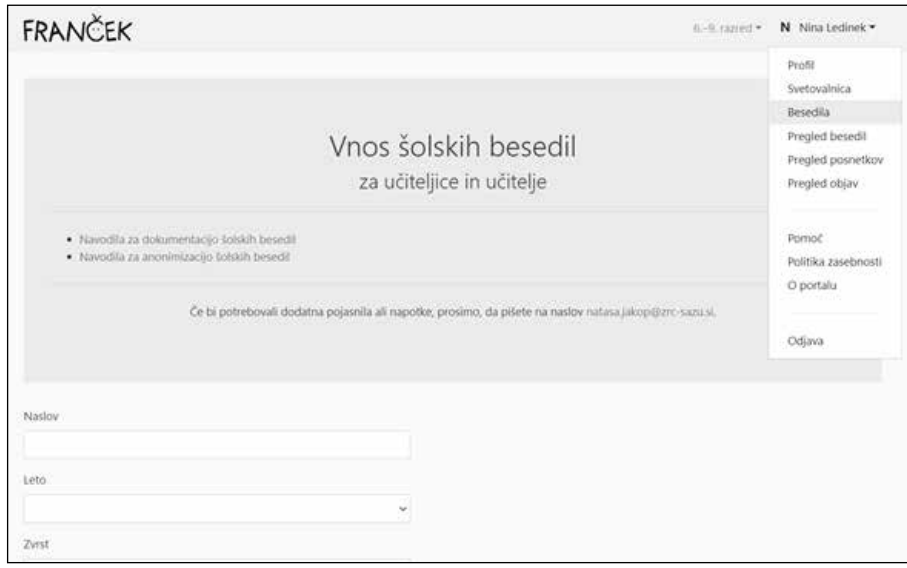
Razpisni pogoji projekta Franček so zahtevali, da so vsi viri, ki nastanejo v času projekta, dostopni pod licenco CC BY 4.0, ki omogoča nadaljnjo uporabo in predelavo ob priznanju avtorstva. Izjemo predstavljajo tisti viri, ki so bili ob začetku projekta že dostopni pod strožjimi pogoji in za katere v okviru projekta nismo pridobili soglasja imetnikov materialnih avtorskih pravic za bolj odprto uporabo. Pridobljena korpusna besedila bi za jezikoslovje sicer predstavljala zelo dragocene raziskovalne podatke, ker pa ne gre samo za raziskovalne podatke, veljavna zakonodaja ščiti avtorje z določili, ki od raziskovalcev zahtevajo natančen opis namena uporabe, kar lahko vsaka projektna skupina opredeli samo za točno določen projekt. Glavno oviro za odprti dostop, ki bi omogočal rabo tudi v naslednjih projektih in raziskavah, torej predstavlja nadaljnji namen uporabe besedil.

3.2 Pridobivanje besedil učencev

Besedila šolarjev smo zbirali s pomočjo učiteljev v vzgojno-izobraževalnih zavodih (zaposlenih v 23 osnovnih in srednjih šolah iz različnih delov Slovenije, ki poučujejo v 645 oddelkih). Besedila šolarjev so dveh tipov. Prvi tip predstavljajo besedila, ki so jih napisali učenci od 1. do 5. razreda osnovne šole. Drugi tip besedil so napisali učenci višjih razredov osnovne šole in srednješolci, pri čemer gre za besedila, ki nagovarjajo otroke v prvih razredih osnovne šole. Za pridobivanje besedil šolarjev smo vzpostavili spletni vmesnik, prek katerega so jih sodelujoči učitelji lahko oddali. Spletni vmesnik je učiteljem olajšal oddajanje besedil, raziskovalcem pa sistemsko zbiranje metapodatkov.

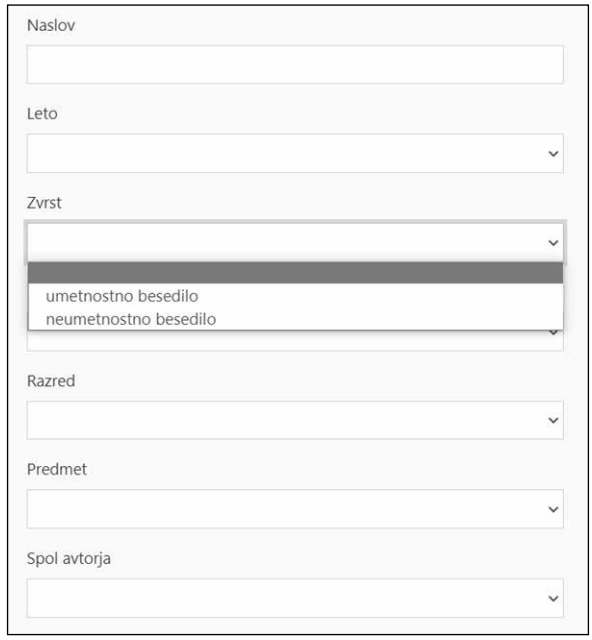
---

15 Delo postane javna last 70 let po avtorjevi smrti.



Slika 1: Spletni vmesnik za oddajo besedil šolarjev

Pred oddajo so učitelji v besedilih popravili večje slovnične in pravopisne napake (npr. napačno zapisane besede), saj bi te lahko vplivale na uspešnost postopkov avtomatske lematizacije in oblikoskladenjskega označevanja (torej tudi na samo besedišče in iskanje po korpusu). Podatki o napakah učencev v korpusu niso dokumentirani, saj korpus ni namenjen detekciji napak, ampak je bil zasnovan za leksikografske namene. Besedila šolarjev so učitelji ob oddaji tudi anonimizirali. Anonimizacija je obsegala nadomeščanje podatkov o imenih in priimkih nejavnih osebnosti, nadomeščanje naslovov, imen vzgojno-izobraževalnih in drugih podobnih zavodov ter nadomeščanje zemljepisnih imen krajev, občin, regij ipd., in sicer so učitelji lastna imena zamenjevali z nadomestnim besedilom *[ime]*, *[priimek]*, *[naslov]*, *[šola]*, *[kraj]*, *[občina]* oz. *[regija]*. Besedila šolarjev so učitelji pred oddajo tudi ustrezno dokumentirali: vsakemu so dodali metapodatke, skladne z vnaprej določenim naborom metapodatkov, pripisanim drugim besedilom v korpusu (gl. razdelek 2.2.2). Da bi bili metapodatki poenoteni, so pri večini predvidenih atributov s spustnega seznama možnih vrednosti izbrali tisto, ki je bila za konkretno besedilo najbolj ustrezna. Le izjemoma (npr. pri metapodatku *naslov besedila*) so metapodatek v vnosni obrazec spletnega vmesnika vpisali brez omejitve. Da bi lahko šolsko besedilo oddali, so morali navesti vse potrebne metapodatke besedila, hkrati pa zagotoviti, da so besedilo anonimizirali v skladu z navodili. Potrditi so morali tudi, da so bila za vsa oddana besedila podana vsa potrebna soglasja zakonitih zastopnikov. Da bi bil postopek oddaje besedil čim bolj enostaven, so bila v okviru vmesnika na voljo tudi navodila za anonimizacijo ter dokumentiranje in oddajanje besedil.



Slika 2: **Vnosni obrazec za dokumentiranje besedil šolarjev**

Skupno je bilo zbranih 428 krajših besedil šolarjev, ki so v povprečju obsegala 17 povedi oz. 225 pojavnic. Metapodatki besedil so bili iz spletnega vmesnika avtomatsko uvoženi v podatkovno zbirko v tabelarični obliki, v kateri smo v času priprave korpusa za druge tipe besedil ročno zapisovali metapodatke o zbranih korpusnih besedilih.

Posebno zahteven je bil postopek zbiranja vseh pravnih soglasij za zbiranje besedil šolarjev. S sodelujočimi vzgojno-izobraževalnimi zavodi smo podpisali pogodbe, s katerimi so lahko učitelji zbirali besedila šolarjev, še prej pa smo morali z njihovo pomočjo pridobiti soglasja zakonitih zastopnikov šolarjev, ki so oddali svoja besedila. Pri tem smo skrbno pazili, da so bila navodila za anonimizacijo natančno predstavljena tako učiteljem kot zakonitim zastopnikom, ki so se tako seznanili z načinom varovanja pravic otrok pri vseh opisanih postopkih.

4 Zaključek

Korpus šolskih besedil slovenskega jezika je specializirani pisni korpus slovenskega jezika v obsegu približno 1,8 milijona pojavnic, namensko oblikovan v okviru projekta Franček za potrebe pedagoške leksikografije. V korpus so vključena besedila, ki nagovarjajo učence nižjih razredov osnovnih šol, in sicer učbeniška besedila (46 enot, 809.406 pojavnic) in izvirno slovensko leposlovje za otroke (83 enot, 931.147 pojavnic), ter besedila, ki so jih oblikovali šolarji (428 enot, 96.257 pojavnic). Del korpusa

je dostopen kot odprto dostopna podatkovna zbirka, zapisana v označevalnem jeziku XML, skladnem s specifikacijami iniciative za zapis korpusnih besedil TEI, celoten korpus pa je za raziskovanje uporabnikom na voljo prek konkordančnikov NoSketch Engine in KonText v okviru raziskovalne strukture CLARIN.SI. Da bi korpus še naprej lahko služil kot gradivska osnova za sodobne pedagoške slovarje slovenščine, razmišljamo o njegovi posodobitvi, v prihodnosti pa načrtujemo tudi raziskave, v katerih bomo *Korpus šolskih besedil slovenskega jezika* primerjali z aktualnim referenčnim korpusom slovenščine, zlasti z vidika njune uporabe v leksikografiji.

VIRI IN LITERATURA

Ahačič idr. 2021a = Kozma Ahačič – Simon Atelšek – Tomaž Erjavec – Peter Holozan – Nataša Jakop – Mateja Jemec Tomazin – Janoš Ježovnik – Nina Ledinek – Andrej Perdih – Miro Romih – Mitja Trojar, *Corpus of Slovenian school texts SBSJ 1.0*, Slovenian language resource repository CLARIN.SI, 2021, http://hdl.handle.net/11356/1413.

Ahačič idr. 2021b = Kozma Ahačič – Janoš Ježovnik – Nina Ledinek – Andrej Perdih – Špela Petric Žižić – Duša Race, Priprava jezikovnih podatkov za pedagoški portal o slovenščini Franček, *Philological Studies* 19.1 (2021), 203–224.

Ahačič – Ledinek – Perdih 2015 = Kozma Ahačič – Nina Ledinek – Andrej Perdih, Portal Fran – nastanek in trenutno stanje, v: *Slovnica in slovar – aktualni jezikovni opis (1. del)*, ur. Mojca Smolej, Ljubljana: Znanstvena založba Filozofske fakultete, 2015 (Obdobja 34), 57–66.

BN 2002 = Tatjana Kokalj, *Besede nagajivke: učni pripomoček za učence od 2. do 5. razreda devetletne osnovne šole*, Trzin: Založba Izolit, 2002.

Čebulj 2013 = Monika Čebulj, *Raba slovarja v 1. in 2. triletju osnovne šole*, diplomsko delo, Univerza v Ljubljani, Pedagoška fakulteta, 2013, http://pefprints.pef.uni-lj.si/1854/1/Čebulj-za_oddajo_(1).pdf.

Erjavec – Krek 2008 = Tomaž Erjavec – Simon Krek, The JOS morphosyntactically tagged corpus of Slovene, v: *Proceedings of the Sixth International Conference on Language Resources and Evaluation, LREC'08*, Pariz: ELRA, 2008.

Erjavec idr. 2010 = Tomaž Erjavec – Darja Fišer – Simon Krek – Nina Ledinek, The JOS linguistically tagged corpus of Slovene, v: *Proceedings of the Seventh International Conference on Language Resources and Evaluation, LREC'10*, Valetta: ELRA, 2010.

Godec Soršak 2015 = Lara Godec Soršak, Slovenski otroški šolski slovar, v: *Slovnica in slovar – aktualni jezikovni opis (1. del)*, ur. Mojca Smolej, Ljubljana: Znanstvena založba Filozofske fakultete, 2015 (Obdobja 34), 243–250.

Godec Soršak 2019 = Lara Godec Soršak, *Zasnova šolskega slovarja za otroke v 1. in 2. vzgojno-izobraževalnem obdobju*, doktorska disertacija, Univerza v Ljubljani, Filozofska fakulteta, 2019.

Grčar – Krek – Dobrovoljc 2012 = Miha Grčar – Simon Krek – Kaja Dobrovoljc, Obeliks: statistični oblikoskladenjski označevalnik in lematizator za slovenski jezik, v: *Zbornik Osme konference Jezikovne tehnologije*, ur. Tomaž Erjavec – Jerneja Žganec Gros, Ljubljana: Institut Jožef Stefan, 2012, 89–94.

Ježovnik – Kenda-Jež – Škofic 2020 = Janoš Ježovnik – Karmen Kenda-Jež – Jožica Škofic, Reduce, Reuse, Recycle: Adaptation of Scientific Dialect Data for Use in a Language Portal for Schoolchildren, v: *Proceedings of XIX EURALEX Congress: Lexicography for Inclusion, Vol. I.*, ur. Zoe Gavriilidou – Maria Mitsiaki – Asimakis Fliatouras, [Poznań : European Association for Lexicography], 2020, 31–37.

Kilgarriff idr. 2014 = Adam Kilgarriff – Vít Baisa – Jan Bušta – Miloš Jakubíček – Vojtěch Kovář – Jan Michelfeit – Pavel Rychlý – Vít Suchomel, The Sketch Engine: ten years on, *Lexicography* 1 (2014), 7–36.

Kosem – Rozman – Stritar 2011 = Iztok Kosem – Tadeja Rozman – Mojca Stritar, How do Slovenian primary and secondary school students write and what their teachers correct: a corpus

of student writing, v: *Proceedings of the Corpus Linguistics 2011 conference, 20-22 July 2011*, Birmingham: University, 2011, http://www.birmingham.ac.uk/research/activity/corpus/publications/conference-archives/2011-birmingham.aspx.

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

Kosem idr. 2016 = Iztok Kosem – Tadeja Rozman – Špela Arhar Holdt – Polonca Kocjančič – Cyprian Adam Laskowski, Šolar 2.0: nadgradnja korpusa šolskih pisnih izdelkov, v: *Zbornik konference Jezikovne tehnologije in digitalna humanistika 2016*, ur. Tomaž Erjavec – Darja Fišer, Ljubljana: Znanstvena založba Filozofske fakultete, 2016, 95–100, http://www.sdjt.si/wp/wp-content/uploads/2016/09/JTDH-2016_Kosem-et-al_Solar-2-0-nadgradnja-korpusa-solskih-pisnih-izdelkov.pdf.

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

Logar Berginc idr. 2020 = Nataša Logar Berginc – Miha Grčar – Marko Brakus – Tomaž Erjavec – Špela Arhar Holdt – Simon Krek, *Korpusi slovenskega jezika Gigafida, KRES, ccGigafida in ccKres: gradnja, vsebina, uporaba,* 1. e-izdaja, Ljubljana: Znanstvena založba Filozofske fakultete, 2020, https://doi.org/10.4312/9789610603542.

Machálek 2020 = Tomáš Machálek, KonText: Advanced and Flexible Corpus Query Interface, v: *LREC 2020: Twelfth International Conference on Language Resources and Evaluation: May 11–16, 2020, Marseille, France*, ur. Nicoletta Calzolari, Paris: ELRA – European Language Resources Association, 2020, 7003–7008, https://www.aclweb.org/anthology/2020.

MMS 1996 = Majda Bitenc – Majda Starovašnik – Marija Ajdovec – Dijana Korošec, *Moj mali slovar*, Kranj: Osnovna šola Franceta Prešerna, 1996.

MPS 2002 = Damjana Šubic – Breda Sivec, *Moj prvi slovar*, Ljubljana: DZS, 2002.

MS 2000 = Barbara Hanuš – Irena Šimenc Mihalič – Damjana Šubic, *Moj slovar*, Ljubljana: DZS, 2000.

Perdih 2021 = Andrej Perdih, Indikatorji pri homografih na portalu Franček, *Jezikoslovni zapiski* 27.2 (2021), 7–21.

Perdih idr. 2021 = Andrej Perdih – Kozma Ahačič – Janoš Ježovnik – Duša Race, Building an Educational Language Portal Using Existing Dictionary Data, *Jazykovedný časopis* 72.2 (2021), 568–578.

Petric Žižić 2020 = Špela Petric Žižić, Tipologija razlag v Šolskem slovarju slovenskega jezika, *Slavistična revija* 68.3 (2020), 391–409.

Rozman 2010 = Tadeja Rozman, *Vloga enojezičnega slovarja slovenščine pri razvoju jezikovne zmožnosti*, doktorska disertacija, Univerza v Ljubljani, Filozofska fakulteta, 2010.

Rozman 2012 = Tadeja Rozman, Jezikovni pouk slovenščine: model (za) nove generacije, v: *Slavistika v regijah – Koper*, ur. Boža Krakar Vogel, Ljubljana: Zveza društev Slavistično društvo Slovenije – Znanstvena založba Filozofske fakultete, 2012 (Zbornik Slavističnega društva Slovenije 23), 219–225.

Rozman idr. 2015 = Tadeja Rozman – Iztok Kosem – Nataša Pirih Svetina – Ina Ferbežar, Slovarji in učenje slovenščine, v: *Slovar sodobne slovenščine: problemi in rešitve*, ur. Vojko Gorjanc – Polona Gantar – Iztok Kosem – Simon Krek, Ljubljana: Znanstvena založba Filozofske fakultete, 2015, 67–74.

Rozman idr. 2020 = Tadeja Rozman – Irena Krapš Vodopivec – Mojca Stritar – Iztok Kosem, *Empirični pogled na pouk slovenskega jezika*, Ljubljana: Znanstvena založba Filozofske fakultete, 2020.

Rychlý 2007 = Pavel Rychlý, Manatee/Bonito – A Modular Corpus Manager, v: *Proceedings of the First Workshop on Recent Advances in Slavonic Natural Language Processing (RASLAN 2007)*, ur. Petr Sojka – Aleš Horák, Brno: Masaryk University, 2007, 65–70.

Stabej idr. 2008 = Marko Stabej – Tadeja Rozman – Nataša Pirih Svetina – Nina Modrijan – Boštjan Bajec, *Jezikovni viri pri jezikovnem pouku v osnovni in srednji šoli: končno poročilo z rezultati*

dela, Ljubljana: Pedagoški inštitut, 2008, https://www.trojina.si/wp-content/uploads/2019/08/StabejRozman.pdf.

TEI Consortium 2017 = *TEI P5: guidelines for electronic text encoding and interchange*, TEI Consortium, http://www.tei-c.org/Guidelines/P5/.

Weiss 1994 = Peter Weiss, Katere slovarje smemo pričakovati po izidu Slovarja slovenskega knjižnega jezika, *Jezik in slovstvo* 39.7–8 (1994), 346–350.

Weiss 2001 = Peter Weiss, Slovenski šolski slovar, v: *Sodobna slovenska narečna poezija. Ciril Kosmač in razvoj povojne slovenske proze*, ur. Zoltan Jan, Ljubljana: Zavod Republike Slovenije za šolstvo, 2001 (Zbornik Slavističnega društva Slovenije 11), 179–188.

SUMMARY

The Corpus of Slovenian School Texts: Design and Creation

The Corpus of Slovenian School Texts is a specialized corpus of written Slovenian containing around 1.8 million tokens, which has been designed specifically for pedagogical lexicography as part of the project Franček, Language Advising Service for Teachers of Slovenian and the Slovenian School Dictionary. It contains texts intended for students in lower primary-school grades and texts composed by students in primary-school grades 1–5. The corpus has been automatically lemmatized and morphosyntactically tagged using the JOS model. It contains three types of texts: school textbooks for subjects taught in primary-school grades 1–6 (809,406 tokens; around 44% of all tokens in the corpus), original Slovenian fiction for children (931,147 tokens; around 51% of all tokens in the corpus), and school texts written by students in primary and secondary schools (96,257 tokens; around 5% of all tokens in the corpus). The part of the corpus that contains texts composed by students is encoded in the XML TEI markup language and is available in the CLARIN.SI research infrastructure repository as an open-access database (CC-BY 4.0). The entire corpus is available to users for research purposes in the NoSketch Engine in KonText concordancers, again using the CLARIN.SI research infrastructure.