Gradnja Korpusa študentskih besedil KOŠ

Tadeja Rozman,* Špela Arhar Holdt‡†

* Fakulteta za upravo, Univerza v Ljubljani
Gosarjeva ulica 5, 1000 Ljubljana
tadeja.rozman@fu.uni-lj.si
‡Filozofska fakulteta, Univerza v Ljubljani
Aškerčeva ulica 2, 1000 Ljubljana
Fakulteta za računalništvo in informatiko, Univerza v Ljubljani
Večna pot 113, 1000 Ljubljana
spela.arharholdt@ff.uni-lj.si

1 Uvod

Korpusi avtentičnih besedil šolajoče se populacije so tako v svetu kot pri nas pomemben vir informacij o jezikovni zmožnosti oseb, ki v procesu izobraževanja svojo jezikovno zmožnost še razvijajo, hkrati pa so tudi pokazatelj jezikovnih in didaktičnih praks v izobraževalnih okoljih. Ti viri so zato pomembni za jezikovno didaktiko, pripravo k uporabnikom usmerjenih jezikovnih priročnikov in gradiv, kot tudi za razvoj različnih jezikovnotehnoloških orodij. Korpusno jezikoslovje v svetovnem merilu sicer večjo pozornost namenja razvoju in analizi korpusov usvajanja tujih jezikov,[1] v slovenskem prostoru pa imamo po vzoru tovrstnih korpusov zgrajen tudi *Korpus šolskih pisnih izdelkov Šolar* (Rozman et al., 2012) oziroma razširjeno verzijo *Šolar 2.0* (Kosem et al., 2016). Vsebuje besedila, napisana pri pouku v tretjem triletju osnovnih šol in v srednjih šolah, del korpusa pa tudi avtentične učiteljske popravke, ki so s hierarhično zasnovanim sistemom oznak (Arhar Holdt et al., 2018) kategorizirani glede na vrsto jezikovnega problema. Slovenščina je tako eden redkih jezikov, ki ima tovrstne podatke za prvi jezik, a le na omejeni šolski populaciji, zato v okviru projekta *Empirična podlaga za digitalno podprt razvoj pisne jezikovne zmožnosti (ARRS, J7-3159)*[2] pripravljamo širitev korpusa s študentskimi besedili, na začetku v obliki pilotnega korpusa študentskih besedil.

2 Namen korpusa

Gradnja Korpusa študentskih besedil KOŠ je v prvi vrsti namenjena pridobivanju empiričnih podatkov o pisni jezikovni zmožnosti študentske populacije, pa tudi analitičnemu vpogledu v procese razvoja strokovnega pisanja. Temeljna jezikovna (normativna, besedilna, pragmatična) znanja naj bi dijaki sicer usvojili že do konca srednje šole, na fakultetah pa naj bi se to znanje nadgrajevalo z usvajanjem terminologije in stilističnih značilnosti strokovnih besedil. Vsaj na nejezikoslovnih študijskih smereh, kjer jezikovna izobrazba ni cilj, ampak je dobro jezikovno znanje le temelj za uspešno profesionalno delovanje, nadaljnje razvijanje jezikovne zmožnosti načeloma poteka hkrati ob usvajanju strokovnega znanja, torej ob recepciji strokovnih del ter s pisanjem npr. seminarskih nalog, esejev, raziskovalnih poročil ter pripravo govornih nastopov, sodelovanjem v strokovnih debatah ipd. Ob tem naj bi študenti uzaveščali procese razumevanja in pisanja, se ukvarjali z razumljivostjo in sprejemljivostjo besedil ter rabo strokovnega besedišča, po potrebi pa tudi odpravljali pravopisne in slovnične pomanjkljivosti. Vendar pedagogi opažamo, da obstajajo velike razlike med jezikovnimi zmožnostmi študentov, profesorji stroke pa se z reševanjem jezikovnih težav lahko ukvarjajo le v omejenem obsegu. Zdi se, da so tudi pristopi pedagogov k ozaveščanju o jezikovnih izbirah različni, ne samo zaradi različnega jezikovnega znanja, ampak tudi pogledov na smiselnost tovrstne povratne informacije, pisnih akademskih praks ipd., pa tudi pomanjkanja didaktičnih usmeritev.

Potreba po razvoju sporazumevalne zmožnosti v slovenskem strokovnem jeziku je bila prepoznana že pri pripravi *Resolucije o Nacionalnem programu za jezikovno politiko 2014–2018*,[3] tedaj določeni jezikovnonačrtovalni cilji jezikovne ureditve visokega šolstva in znanosti pa se v aktualni *Resoluciji o Nacionalnem programu za jezikovno politiko 2021–2025*[4] niso bistveno spremenili. Dokument tako določa, da je na visokošolski strokovni ravni treba omogočati učenje strokovne slovenščine ter na podlagi raziskav in analiz

[1] Več o korpusih usvajanja tujega jezika in gradnji korpusa usvajanja slovenščine kot tujega jezika gl. npr. Stritar Kučuk (2020).
[2] https://www.cjvt.si/prop/
[3] https://www.uradni-list.si/glasilo-uradni-list-rs/vsebina/2013-01-2475?sop=2013-01-2475
[4] https://www.uradni-list.si/glasilo-uradni-list-rs/vsebina/2021-01-1999?sop=2021-01-1999

strokovno-znanstvenega pisanja na visokošolski ravni izdelati učni načrt za strokovno-znanstveno pisanje za uvodni predmet v prvem letniku prvostopenjskih programov. Na podlagi teh določil, zapisanih že v predhodni resoluciji, je bil z namenom pridobivanja empiričnih podatkov o strokovno-znanstvenem pisanju leta 2019 izdelan *Korpus akademske slovenščine KAS*, tj. korpus diplomskih, magistrskih in doktorskih del (Erjavec et al., 2021), objavljenih na Nacionalnem portalu odprte znanosti.[5] V korpusu so torej zbrana strokovna študentska besedila po zaključenih stopnjah visokošolskega in univerzitetnega študija, mentorirana in v veliki meri tudi lektorirana, tako da je s stališča analiz pisne jezikovni zmožnosti študentske populacije in za analizo procesa razvoja strokovnega pisanja le deloma uporaben. Korpus KOŠ bi v perspektivi lahko odpravil vrzel korpusnih podatkov med Šolarjem in KAS-om ter ponudil osnovo za raziskave, katera temeljna znanja je potrebno (bolje) nasloviti na predhodnih stopnjah in katera na terciarni stopnji, kjer se razvoj pisnih jezikovnih zmožnosti nadaljuje na kompleksnejših besediloslovnih ravneh. Širša slika razvojnega loka bi omogočila, da opismenjevanje bolje usmerimo proti končnemu cilju, ki je polnomočno in samostojno (čeprav v skladu s sodobnimi praksami tehnološko in podatkovno podprto) pisanje različnih vrst besedil za različne sporazumevalne namene, kar je pomembno tudi za uspešno poklicno delovanje.

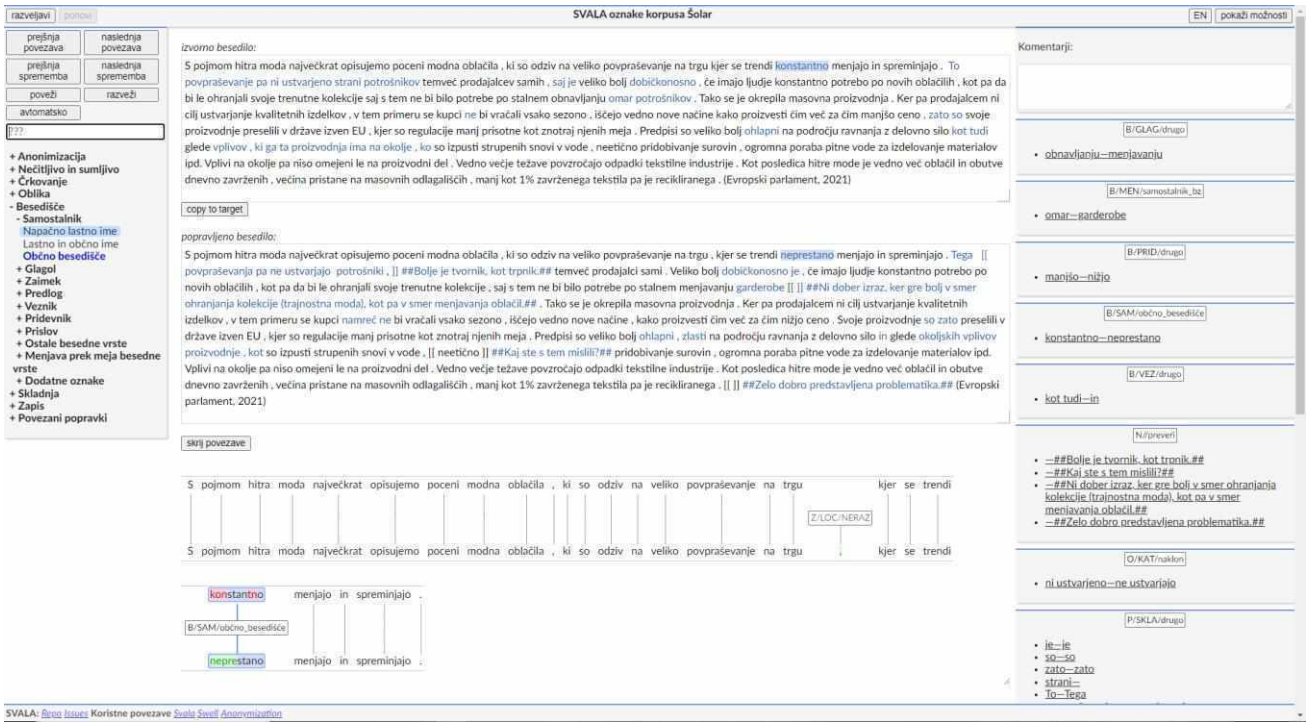
3 Zasnova korpusa

V okviru projekta je predvidena priprava pilotnega korpusa, ki bo objavljen kot podatkovna baza na repozitoriju CLARIN.SI. Gradnja korpusa poteka v študijskem letu 2021/22 in se bo predvidoma končala jeseni 2022, besedila pa bodo zbrana po metodologiji priprave korpusa Šolar, ki vključuje: pravno ureditev odprtega dostopa do rezultatov (priprava in podpis pogodb o prenosu pravic in dovoljenja za uporabo pravic), beleženje vseh relevantnih metapodatkov (program, letnik, področje študija, tip besedila, morebitno večavtorstvo, ob oddanih več verzijah istega besedila tudi oznaka prvotne in spremenjenih verzij), vsaj delna vključitev profesorskih jezikovnih popravkov, zapis v združljivem formatu in strojno označevanje.

Jezikovne popravke bomo v korpus beležili z orodjem *Svala* (Wirén, 2019), ki omogoča pregledno sopostavitev izvornega ter popravljenega besedila, psevdonimizacijo tistih delov besedila, ki bi lahko razkrili avtorstvo ali kake druge občutljive osebne informacije, ter označevanje in vsebinsko kategorizacijo jezikovnih popravkov. Orodje je bilo na projektu *Razvoj slovenščine v digitalnem okolju*[6] prilagojeno za delo s slovenskima korpusoma KOST (Stritar Kučuk, 2020) in Šolar in kot tako podpira označevanje s sistemom oznak korpusa Šolar (Arhar Holdt et al., 2018). Te oznake bomo uporabili tudi za korpus KOŠ (gl. sliko 1), predvideno pa je, da bo za študentska besedila sistem označevanja treba deloma prilagoditi. Pričakovati namreč je (in do sedaj zbrano gradivo to potrjuje), da so zaradi žanrske specifike študentskih besedil, ki jih pregledujejo profesorji nejezikoslovci, popravki redko tudi konkretni predlogi pravilnih jezikovnih izbir, ampak da gre bolj za usmeritve profesorjev, ki v svojih komentarjih študente le na splošno opozarjajo na jezikovne napake in se v večji meri posvečajo stilistiki strokovnih besedil, ustrezni rabi terminologije, citiranju, razumljivosti pisanja, argumentaciji ipd. Vsa korpusna besedila (z označenimi popravki in brez) bomo nato strojno označili na ravneh stavčne segmentacije, tokenizacije, lematizacije, oblikoskladnje, skladnje in imenskih entitet z označevalnikom CLASSLA StanfordNLP (Ljubešić in Dobrovoljc, 2019), ki se v času pisanja povzetka prav tako razvija na omenjenem projektu.

Besedila zbiramo na prvostopenjskih študijskih programih na dveh fakultetah, za vključitev v korpus pa so potencialno relevantna vsa besedila, ki so jih študenti oddali pedagogom v študijskem procesu na fakulteti in niso napisana na roko. Besedila zato zbiramo prek učiteljev, saj bomo tako z večjo gotovostjo prejeli avtentična besedila, ki se realno pišejo v študijskem okolju, predvidoma pretežno seminarske naloge, eseje, poročila, povzetke strokovnih člankov, daljše (esejske) odgovore na vprašanja, morda pa tudi dispozicije in osnutke diplomskih del. Besedila, povezana s pripravo zaključnih del, so s stališča ugotavljanja zmožnosti oblikovanja daljšega strokovnega besedila po končanem izobraževanju zelo dragocena, tudi zaradi vpogleda v mentorske komentarje in popravke, a se trenutno zdi vključitev teh besedil v korpus problematična s stališča anonimizacije, saj so zaključna dela praviloma prosto objavljena na spletu in zlahka povezljiva z osnutki, avtorji in mentorji.

[5] https://openscience.si/
[6] https://www.slovenscina.eu/



Slika1: Preizkus metodologije vpisovanja popravkov v testni različici lokaliziranega programa Svala.

4 Nadaljnji koraki

Na projektu *Empirična podlaga za digitalno podprt razvoj pisne jezikovne zmožnosti* želimo zagotoviti pilotni korpus v obsegu 200.000 pojavnic, ob gradnji pa pripraviti tudi oceno prenosljivosti metodologije Šolar na pisno produkcijo študentov in specifikacije nadaljnjega razvoja korpusa študentskega pisanja, tj. opredelitev želenega obsega, strukture glede na regionalno zastopanost, vrsto in področje izobraževanja ter tipologije popravkov. V tem okviru pripravljamo tudi krajši anketni vprašalnik za univerzitetne pedagoge, s katerim želimo pridobiti dodatne podatke o tem, kakšne so prakse podajanja povratnih informacij študentom, ter tako čim učinkoviteje zasnovati zbiranje in beleženje tega gradiva. Do sedaj zbrano gradivo po pričakovanjih nakazuje, da so prakse precej raznolike in da se v mnogočem razlikujejo od podajanja informacij profesorjev slovenščine, ki so zabeležene v korpusu Šolar.

Pod okriljem projekta bomo sicer zbrano korpusno gradivo uporabili za pilotne kvantitativne in kvalitativne jezikoslovne analize študentskega pisanja. Analize se bodo osredotočile na tipične težave pisanja in trende opozarjanja na jezikovno neustrezne ali manj ustrezne ubeseditve, kar vključuje podajanje povratne informacije z vnosom rešitve, opisna priporočila ali grafično nakazovanje mesta težave, kot morebitne druge načine. Rezultate bomo primerjali s frekvenčno urejenimi seznami jezikovnih zadreg v korpusu Šolar. Izsledki bodo predvidoma že nakazali obrise razvoja pisne jezikovne zmožnosti na prehodu iz srednješolskega v univerzitetno izobraževanje, morebitne primanjkljaje temeljnega jezikovnega znanja ter kako je mogoče učni proces z empiričnimi podatki najbolje podpreti.

Zahvala

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 raziskovalno dejavnost Republike Slovenije iz državnega proračuna.

Literatura

Špela Arhar Holdt, Polona Lavrič, Rebeka Roblek in Teja Goli. 2018. Kategorizacija učiteljskih popravkov: Smernice za označevanje korpusa Šolar 2.0. V: *1.0. Kazalnik projekta Nadgradnja korpusa Šolar*. https://solar.trojina.si/wp-content/uploads/2022/05/Smernice-za-oznacevanje-korpusa-Solar-2.0-v1.0.pdf

Tomaž Erjavec, Darja Fišer in Nikola Ljubešić. 2021. The KAS corpus of Slovenian academic writing. V: *Lang Resources & Evaluation 55*, 551–583. https://doi.org/10.1007/s10579-020-09506-4

Iztok Kosem, Tadeja Rozman, Špela Arhar Holdt, Polonca Kocjančič in Cyprian Adam Laskowski. 2016. Šolar 2.0: nadgradnja korpusa šolskih pisnih izdelkov. V: *Zbornik konference Jezikovne tehnologije in digitalna*

humanistika, 95–100. Znanstvena založba Filozofske fakultete, Ljubljana. http://www.sdjt.si/wp/wp-content/uploads/2016/09/JTDH-2016_Kosem-et-al_Solar-2-0-nadgradnja-korpusa-solskih-pisnih-izdelkov.pdf

Nikola Ljubešić in Kaja Dobrovoljc. 2019. What does Neural Bring? Analysing Improvements in Morphosyntactic Annotation and Lemmatisation of Slovenian, Croatian and Serbian. V: *Proceedings of the 7th Workshop on Balto-Slavic Natural Language Processing, BSNLP@ACL 2019*, pages 29–34. https://aclanthology.org/W19-3704.pdf

Tadeja Rozman, Mojca Stritar Kučuk in Iztok Kosem. 2012. Šolar – korpus šolskih pisnih izdelkov. V: T. Rozman, ur., I. Krapš Vodopivec, M. Stritar, I. Kosem: *Empirični pogled na pouk slovenskega jezika*, 15–35. Ljubljana: Trojina, zavod za uporabno slovenistiko.

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

Mats Wirén, Arild Matsson, Dan Rosén in Elena Volodina. 2019. SVALA: Annotation of Second-Language Learner Text Based on Mostly Automatic Alignment of Parallel Corpora. V: *Selected papers from the CLARIN Annual Conference 2018, Pisa, 8–10 October 2018*, pages 227–239. https://ep.liu.se/en/conference-article.aspx?series=ecp&issue=159&Article_No=23