Simpozij OBDOBJA 41

Špela Arhar Holdt, Iztok Kosem, Mojca Stritar Kučuk

Metode in orodja za lažjo pripravo korpusov usvajanja jezika

objavljeno v:
Nataša Pirih Svetina, Ina Ferbežar (ur.): *Na stičišču svetov: slovenščina kot drugi in tuji jezik*. *Obdobja 41*. Ljubljana: Založba Univerze v Ljubljani, 2022.
https://centerslo.si/simpozij-obdobja/zborniki/obdobja-41/

© Univerza v Ljubljani, Filozofska fakulteta, 2022.

Obdobja (e-ISSN 2784-7152)

Univerza v Ljubljani
Filozofska fakulteta

CENTER ZA SLOVENŠČINO
KOT DRUGI IN TUJI JEZIK

METODE IN ORODJA ZA LAŽJO PRIPRAVO KORPUSOV USVAJANJA JEZIKA

Špela Arhar Holdt
Filozofska fakulteta in Fakulteta za računalništvo in informatiko, Univerza v Ljubljani, Ljubljana
spela.arharholdt@ff.uni-lj.si

Iztok Kosem
Filozofska fakulteta in Fakulteta za računalništvo in informatiko, Univerza v Ljubljani, Ljubljana
iztok.kosem@ff.uni-lj.si

Mojca Stritar Kučuk
Filozofska fakulteta, Univerza v Ljubljani, Ljubljana
mojca.stritarkucuk@ff.uni-lj.si

DOI:10.4312/Obdobja.41.23-30

Prispevek predstavi težka mesta izdelave korpusov usvajanja tujega in maternega jezika, kot so transkribiranje besedil, anonimizacija, ročno označevanje in vsebinsko kategoriziranje popravkov, v nadaljevanju pa novo prosto dostopno orodje, ki ponuja rešitev za opisane metodološke izzive. Orodje, ki temelji na švedskem programu Svala, smo prilagodili za slovenščino, ga nadgradili, da omogoča delo s korpusoma Šolar in KOST, ter evalvirali s pomočjo dejanske korpusne gradnje.

korpusi usvajanja jezika, Svala, KOST, Šolar

This article highlights the challenges of creating learner and developmental text corpora that feature error corrections: transcription and anonymization of texts, and manual annotation and categorization of corrections. It presents a new freely available tool that offers a solution to these challenges. Based on the Swedish Svala software, the tool has been adapted for Slovenian, modified to work with the Šolar and KOST corpora, and evaluated as part of an actual corpus creation process.

learner corpus, developmental corpus, Svala, KOST, Šolar

1 Uvod

Korpusi usvajanja tujega in maternega jezika (angl. *learner* in *developmental corpora*, Leech 1997:19) vsebujejo besedila avtorjev in avtoric,[1] ki usvajajo oz. se učijo določen jezik, pri čemer so v ta besedila pogosto vključene tudi oznake

1 V prispevku uporabljamo moški slovnični spol kot nevtralno in vključujočo izbiro na mestih, kjer bi navedba več oblik otežila branje.

jezikovnih težav in popravkov (Granger 2008). Tovrstni korpusi predstavljajo empirično osnovo za raznovrstne raziskave s področja jezikovne didaktike, za pripravo učnih gradiv, vaj, testov, učnih množic za strojno procesiranje naravnega jezika itn. V mednarodnem prostoru nastajajo že od osemdesetih let prejšnjega stoletja in predvsem za bolj razširjene jezike kot druge jezike so dandanes na voljo številni viri, gradiva in raziskave. Za slovenščino velja omeniti tri sorazmerno mlade vire: korpus slovenščine kot tujega jezika KOST (Stritar Kučuk 2020), korpus pisanja slovenskih osnovnošolcev in dijakov Šolar (Kosem idr. 2016), na primerljiv način pa je osnovan tudi korpus lektorskih popravkov Lektor (Popič 2014).

Z razvojem korpusnega jezikoslovja se razvija tudi metodologija priprave korpusov usvajanja jezika, ki je dolga desetletja veljala za zapleteno in skorajda mučno počasno. Rešitev za mnoge metodološke težave je ponudilo orodje Svala (Wirén 2019; Volodina idr. 2019), ki ga je izdelala ekipa raziskovalnega centra Språkbanken za pripravo korpusa švedščine kot drugega ali tujega jezika. Pri projektu Razvoj slovenščine v digitalnem okolju (RSDO)[2] smo orodje prilagodili za slovenščino, ga nadgradili, da omogoča delo s korpusoma KOST in Šolar, nato pa evalvirali s pomočjo dejanske korpusne gradnje. V prispevku opisujemo metodološke izzive in olajšave, ki jih prinaša novo orodje. To bo ob koncu projekta skupnosti odprto na voljo za uporabo.

2 Glavni izzivi gradnje korpusov usvajanja jezika

Priprava korpusov z označenimi jezikovnimi popravki je zapletena in počasna, saj poleg običajnih korakov priprave korpusnih besedil, kot so pravno urejeno pridobivanje besedil, strojno označevanje in formatiranje, zahteva tudi dodatne korake: od transkribiranja (pogosto ročno napisanih) besedil, anonimizacije do ročnega označevanja in vsebinskega kategoriziranja jezikovnih popravkov. Do nedavno so raziskovalci in raziskovalke, ki so gradili tovrstne korpuse, za naštete naloge iskali in prilagajali orodja, specializirana za kak drug namen, ter se spopadali z zapleteno metodologijo. Ta je pogosto vodila v napake pri ročnem delu in zahtevala redno tehnično podporo.

Prvi izziv je prenos besedil v digitalno obliko, ustrezno za nadaljnjo obravnavo. Če se v korpus vključujejo besedila, ki so jih avtorji in avtorice že izhodiščno napisali v digitalni obliki (npr. s katerim od urejevalnikov besedil), je ta korak sorazmerno preprost. Nasprotno pa velja za besedila, napisana na roko, npr. besedila, nastala pri pouku v šoli, na jezikovnih tečajih, lektoratih ipd. Pri transkribiranju ročno napisanih besedil je potrebna velika natančnost, da se v digitalno obliko prepišejo tudi jezikovne težave, pri čemer je prisotna tudi določena mera subjektivne interpretacije, npr. ali gre pri določenem zapisu za napako črkovanja ali ne.

Ker so korpusi usvajanja jezika dragoceni za različne namene, težimo k njihovi odprti dostopnosti, slednja pa je pogojena z anonimnostjo avtorjev in avtoric besedil. Ta je zagotovljena na ravni metapodatkov, kjer avtorstva ne navajamo, pogosto pa se

2 Projektna spletna stran: https://www.slovenscina.eu/ (dostop do vseh spletnih strani, navedenih v prispevku: 19. 5. 2022).

navajajo druge informacije, nujne za ustrezno interpretacijo rezultatov, kot so vrsta šole, razred in regija pri korpusih usvajanja prvega jezika ter prvi jezik ali država izvora pri korpusih usvajanja drugega in tujega jezika. Zato je toliko bolj pomembno, da so besedila anonimizirana tudi z vidika vsebine, ki lahko zlasti pri določenih besedilnih vrstah razkriva različne osebne informacije. Zapletenost postopka anonimizacije je različna: od preprostega nadomeščanja občutljivih informacij z enoznačno kodo do psevdonimizacije z generičnimi nadomestnimi informacijami (npr. zamenjava lastnih imen z generičnimi imeni), pri čemer je mogoče na različne načine prenesti tudi morebitne jezikovne napake.

V korpusih usvajanja jezika dragoceno informacijo prinašajo oznake jezikovnih težav. Tipično jih pripravijo jezikoslovci in jezikoslovke, ki korpus gradijo, lahko pa se vključi avtentično povratno informacijo učiteljev in učiteljic, vključenih v pedagoški proces. V prvem primeru je potrebna celovita obravnava zbranih besedil, njihovo popravljanje in vpis popravkov po izbranem sistemu, v drugem je treba zagotoviti predvsem natančen prepis, ki je zahteven zlasti pri ročno napisanih besedilih, v katerih učiteljski popravki niso vedno jasni, nedvoumni in enostavno pretvorljivi v digitalno obliko. Jezikovne popravke vsebinsko kategoriziramo po izbranem sistemu označevanja, ki segajo od robustnih do podrobnejših. Ne glede na izbrani sistem zahteva odločanje o tem, v katero vsebinsko kategorijo določena jezikovna težava spada, jasne smernice, izredno dragocena za konsistentnost pa je tudi možnost celovitega pregledovanja že označenega gradiva ter – v primeru spremenjenih odločitev – enostavnega in hitrega popravljanja obstoječih oznak.

3 Dosedanja gradnja korpusa KOST

Pri gradnji korpusa KOST[3] nam je bila v veliko pomoč izkušnja s pilotnim korpusom usvajanja slovenščine kot tujega jezika PiKUST (Stritar 2012). Rešitev izziva, kako zagotoviti dostop do zadostne količine specifičnih besedil in poskrbeti za pravne vidike dostopa, pa nam je ponudil modul Leto plus, v okviru katerega se vsako leto slovenščino uči veliko število mednarodnih študentov in študentk Univerze v Ljubljani.[4] V KOST vključujemo njihove pisne domače naloge, besedila, ki nastajajo na lektoratu, in besedila, napisana na pisnih izpitih. Sodelujoči pred tem podpišejo izjavo, s katero dovoljujejo vključitev svojih besedil v korpus (prim. Stritar Kučuk 2020: 133).

Od začetka epidemije covida-19 in selitve visokošolskega pouka v razne oblike na daljavo je večina besedil, ki jih dobivamo za korpus, napisana v digitalni obliki. Za vključitev v korpus jih shranimo v goli besedilni obliki, vsako v posebni datoteki, evidenco o njih pa vodimo v enotni Excelovi tabeli. Bistveno več dela je z besedili, ki so bila napisana na roko in jih je treba digitalizirati. Pretipkavanje opravljamo zaposleni na programu Leto plus (prim. Stritar Kučuk 2020: 133). V zadnjem času si

3 Spletna stran: https://www.cjvt.si/korpus-kost/.
4 Spletna stran: https://www.uni-lj.si/studij/leto-plus/.

pri tem pomagamo s fotografiranjem in pretvorbo besedila na sliki v digitalni zapis.[5] Odvisno od čitljivosti rokopisa na sliki pri tem sicer pogosto prihaja do napak, vendar je za tistega, ki tipka, popravljanje že digitalno napisanega besedila pogosto manj obremenjujoče od pretipkavanja v celoti, poleg tega pa je lahko tako bolj pozoren na to, da digitalni zapis ustreza originalu. Na roko napisanih besedil je v KOST-u razmeroma malo – v času pisanja tega prispevka 13 %. Njihovo vključevanje v korpus pa je neizbežno, saj je vsaj pri tistih besedilih, ki so bila na roko napisana v izpitnih pogojih, bistveno večji nadzor nad zunanjimi okoliščinami tvorjenja.

Najzahtevnejši in hkrati najpomembnejši del priprave KOST-a je označevanje jezikovnih napak, ki poteka ročno in v skladu z vnaprej določeno klasifikacijo napak. V tem prispevku puščamo ob strani razmislek o kategorijah napak, omenimo le, da so razvrščene v 23 kategorij. Vsaki napačni obliki v besedilu pripišemo oznako napake in zraven navedemo popravljeno obliko. Napako je torej treba najprej prepoznati, jo klasificirati in popraviti. Zato želimo, da je vsako besedilo dostopno v dveh oblikah: izvirni in popravljeni.

4 Dosedanja gradnja korpusa Šolar

Zgodovina gradnje korpusa Šolar, ki vsebuje pisna besedila slovenskih osnovnošolcev in dijakov, priča o večplastnosti težav, s katerimi se srečajo izdelovalci tovrstnih korpusov. Pri korpusu Šolar se zbiranje besedil niti ni izkazalo za preveč problematično, saj so bili učitelji in učiteljice pripravljeni pomagati, še več, pri izdelavi prve različice, ki je vsebovala skoraj milijon besed, nam je zaradi časovnih in finančnih omejitev uspelo v korpus vključiti samo 2703 od 8594 zbranih besedil (Rozman idr. 2012). Odločitev je bila tudi metodološka, saj smo večjo pozornost posvečali uravnoteženosti korpusa, tako na ravni regijske zastopanosti kot zastopanosti različnih predmetov in nivojev izobraževanja. Zlasti regijsko zastopanost smo pri izdelavi druge različice (Kosem idr. 2019), ki vsebuje 1,63 milijona besed, še izboljšali.[6]

Za razliko od korpusa KOST je bila velika večina besedil za korpus Šolar napisanih na roko, kar je pomenilo veliko količino pretipkavanja. Pri tem je pomembno, da je pri korpusu Šolar pretipkavanje vključevalo tudi beleženje učiteljskih jezikovnih popravkov in njihovo kategorizacijo, kar je postopek upočasnjevalo in je zahtevalo tudi dodatno usposabljanje kadra. Precejšnjo oviro, ki se je pojavila pri prvem zbiranju, je predstavljalo dejstvo, da so učitelji pošiljali kopije besedil, kakovost kopij pa je bila odvisna od kakovosti uporabljenega kopirnega stroja. Poleg tega so bile kopije črno-bele in nemalokrat ni bilo jasno, kdaj je nekaj popravil učitelj in kdaj že učenec sam (Arhar Holdt idr. 2017: 98). Ta težava je bila odpravljena pri postopku zbiranja besedil za drugo različico korpusa, saj smo prešli na metodo skeniranja besedil in pošiljanje (barvnih) PDF datotek.

Medtem ko nam za razliko od korpusa KOST ni bilo treba jezikovno popravljati besedil, saj smo zgolj beležili učiteljske popravke, pa se je za zahtevno izkazalo

5 To omogoča Googlova funkcija Google Lens, prim. https://lens.google/.
6 Spletna stran projekta Šolar 2.0: https://solar.trojina.si/.

kategoriziranje napak, deloma zaradi podrobne klasifikacije (Kosem idr. 2020), za še zahtevnejše pa kasnejše popravljanje kategorij zaradi usklajevanja in na mestih poenostavitve kategorij (Arhar Holdt idr. 2018). Ravno tu smo namreč naleteli na pomanjkanje dveh rešitev: na eni strani standardnega formata za zapis korpusov z jezikovnimi popravki in na drugi strani orodja za učinkovit prikaz in urejanje besedil z jezikovnimi popravki. Za naše namene smo uporabili posebej prilagojeno orodje Sketch Engine, kar pa je imelo za posledico zelo zamuden postopek prekategorizacije: korpus smo morali najprej pretvoriti v format VERT in ga uvoziti v Sketch Engine, sledilo je označevanje obstoječih napak z novimi kategorijami. Pri tem smo redno izvažali korpusne datoteke in jih pretvarjali v format XML, da smo lahko novo-označene kategorije (te se namreč v Sketch Enginu beležijo ločeno) zapisali v korpusne datoteke, ter spet opravili pretvorbo in korpus uvozili v Sketch Engine. Ta izkušnja nas je spodbudila, da smo se povezali z mednarodnimi raziskovalci, ki so se srečevali s podobnimi izzivi. Na ta način smo izvedeli za prizadevanja švedskih kolegov in njihov program Svala, ki ga predstavljamo v nadaljevanju.

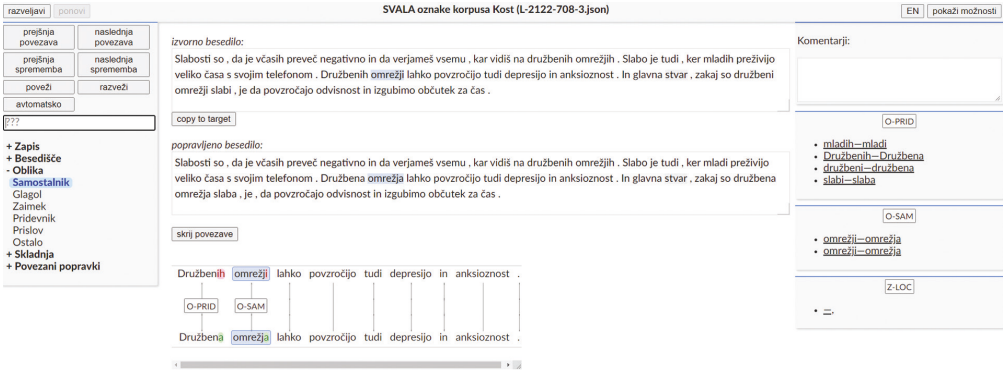
5 Program Svala in njegova adaptacija za slovenščino

Program Svala (Wirén 2019) je orodje za psevdonimizacijo, normalizacijo (popravljanje) in označevanje popravkov v besedilih učečih se. Program je del platforme SweLL (Volodina idr. 2019), ki poleg naštetega omogoča tudi vodenje delotokov za zbiranje in urejanje korpusnega gradiva. Swell in Svala sta bila razvita ob upoštevanju značilnosti gradnje korpusov z označenimi jezikovnimi popravki in potreb razvojne skupnosti, pri čemer so avtorji sledili naslednjim razvojnim načelom (Wirén 2019: 228): 1) enovito okolje, v katerem je mogoče opraviti različne korake dela, s čimer se izognemo pretvarjanju med različnimi vmesnimi formati podatkov; 2) preprost končni format, zlahka pretvorljiv v formate, ki jih zahtevajo tipična orodja za delo s podatki; 3) intuitiven uporabniški vmesnik, v katerem je mogoče posamezne korake dela opraviti ločeno ali povezano in ki nudi avtomatsko povezovanje ter pregledno vizualno informacijo o povezavah med izvornim in popravljenim besedilom; 4) administrativna podpora, s katero je mogoče delegirati delo, spremljati časovnico in statistike delotoka, vključno s preverjanjem ujemanja med označevalci.

Načrt, da se program Svala lokalizira in prilagodi za slovenščino, je bil vključen v nacionalni projekt RSDO, ki se med letoma 2020 in 2023 posveča razvoju jezikovnih virov in tehnologij za sodobno slovenščino. Projekt vključuje skrb za preprosto in učinkovito nadgrajevanje referenčnih in nekaterih specializiranih besedilnih korpusov, med katerimi sta tudi KOST in Šolar. Pri projektu smo se glede na potrebe raziskovalne skupnosti odločili, da bomo prioritetno prenesli modul, ki omogoča transkripcijo, preprosto anonimizacijo in označevanje napak, za kasnejši razvoj pa pustili naprednejše (avtomatsko podprto) anonimiziranje in vodenje označevalnih delotokov. Preizkusna različica prilagojenega programa je na spletni strani https://svala.cjvt.si/, končna pa predvidoma na https://orodja.cjvt.si/svala.

Največja prednost programa je, da združuje več korakov priprave korpusnih besedil, kar je prikazano na Sliki 1. Izvorno besedilo uvozimo, prilepimo iz drugega

programa ali transkribiramo v vrhnje okence vmesnika (napis *izvorno besedilo*). Od tam ga preprosto kopiramo v spodnje okence (napis *popravljeno besedilo*) in v duplikat vnesemo jezikovne popravke, nato pa program avtomatsko poveže dele izvornega in popravljenega besedila. Povezave, ki so pregledno prikazane v obliki črtic med besedami pod obema okencema – žargonsko jim pravimo »špageti« –, je mogoče ročno popraviti, pri čemer je omogočeno združevanje besed v skupine, ko se oznaka nanaša na več besed (npr. pri napakah besednega reda v KOST-u: *zdi mi se* → *zdi se mi*) ali ko se več besed zamenja z eno samo ali obratno (npr. popravki oziralnih zaimkov: *stvari, katere si ne morem privoščiti* → *stvari, ki si jih ne morem privoščiti*), kot tudi enosmerne povezave, kadar popravek prinese v besedilo dodatno besedo ali katero od obstoječih odstrani (npr. manjkajoči morfem *se*: *po kosilu učim* → *po kosilu se učim*).



Slika 1: Primer izvornega in popravljenega besedila iz korpusa KOST v preizkusnem vmesniku Svala.cjvt.si.

S klikom na vsako od povezav oz. črtic je popravku mogoče določiti vsebinsko kategorijo, pri čemer lahko v slovenski različici programa trenutno izbiramo med sistemoma označevanja KOST in Šolar. Na Sliki 1 je v levem meniju prikazan sistem označevanja KOST, po katerem je mogoče oznake iskati z upoštevanjem dvostopenjske hierarhije ali pa s pomočjo iskalnega okenca. Orodna vrstica zgoraj levo omogoča hitro preklikavanje med vnesenimi jezikovnimi popravki in kontrolo označevalnega procesa (povezovanje besed v skupine in njihovo razdruževanje, razveljavitev in ponovna uveljavitev pripisa oznake ipd.). Program izvorno besedilo, popravljeno besedilo, povezave in oznake popravkov beleži v formatu JSON, ki si ga je mogoče ogledati v vmesniku ali ga izvoziti.

V sklopu projekta RSDO smo program uporabili in s tem preizkusili na dveh nalogah: za pripravo korpusa KOST in za nadgradnjo korpusa Šolar 2.0 v novo različico, ki bo dostopna v poenostavljenem in prečiščenem formatu. Izkušnje z uporabo Svale so bile izredno pozitivne: za KOST je bilo do oddaje prispevka označenih več kot 500 besedil. Glavnino označevanja je opravila urednica KOST-a. Navodila za označevanje so dokumentirana v priročniku, ki se sproti dopolnjuje

(Stritar Kučuk 2022). V jesenskem semestru 2021/22 pa smo preverili tudi, kako enostavna in učinkovita je uporaba Svale za polprofesionalne uporabnike, namreč za študente 3. letnika 1. stopnje slovenistike. 19 študentov je označilo 83 besedil. Pred tem smo načrtno izvedli le krajše usposabljanje oz. prikaz dela s Svalo, saj smo želeli preizkusiti, kako dobro se znajdejo brez podrobnejših navodil. Besedila, ki so jih označili, je nato pregledala še urednica KOST-a, študenti pa so svoje delo predstavili v okviru seminarja pri predmetu Slovenščina kot tuji jezik. Rezultati so bili zelo pozitivni: čeprav je bilo v povprečju v njihovih besedilih 35 % neustreznih oznak, pa nobeden od označevalcev ni imel večjih težav s samo aplikacijo za označevanje. Še največji izziv jim je predstavljalo združevanje besed v skupine. Sicer pa so vsi študenti pri predstavitvi svojega seminarskega dela izrazili zadovoljstvo z možnostjo praktičnega, tehnično nezahtevnega dela, pri katerem so morali dejansko uporabiti tudi jezikoslovno znanje, pridobljeno pri študiju.

6 Zaključek in nadaljnje delo

V prispevku smo predstavili slovensko različico programa Svala, ki predstavlja prelomni korak na področju razvoja korpusov usvajanja jezika. Slovenska skupnost je ena prvih, ki je program prilagodila nacionalnim razvojnim potrebam. Prve izkušnje kažejo prijetno uporabniško izkušnjo, izjemen časovni prihranek in povišano kakovost rezultatov. Program je prosto in odprto na voljo za nadaljnjo uporabo in lahko služi za gradnjo ne le korpusov usvajanja jezika, ampak tudi raznovrstnih specializiranih jezikovnih virov, kjer sta dragoceni poravnanost besedil in vsebinsko označevanje jezikovnih povezav ali sprememb. V tej luči ga želimo po koncu projekta predstaviti širši javnosti in zbrati povratne informacije skupnosti o tem, kako program razvijati v prihodnje.

Pri tem je mogoče izpostaviti nadaljnje delo znotraj projekta RSDO, kjer želimo razviti celoten cevovod za pripravo korpusov usvajanja jezika. Na eni strani razvijamo portal, prek katerega bodo sodelujoči lahko hitro in enostavno oddajali besedila učečih se. Sledi delo s Svalo, čemur bo sledil korak strojnega jezikoslovnega označevanja na različnih ravneh (trenutno so predvidene: tokenizacija, segmentacija, lematizacija, oblikoskladnja, skladnja in imenske entitete), nato pa pretvorba v formata VERT in TEI, ki ju tipično zahteva umestitev korpusa v konkordančnike. Zadnji cilji projekta pa je zasnovati nov, uporabniško prijazen konkordančnik, ki se bo v večji meri osredotočal na temeljno dodano vrednost korpusov KOST in Šolar: jezikovne popravke.

Literatura

ARHAR HOLDT, Špela, KOSEM, Iztok, GANTAR, Polona, 2017: Corpus-based resources for L1 teaching: the case of Slovene. Ann Marcus-Quinn, Tríona Hourigan (ur.): *Handbook on digital learning for K-12 schools*. Cham: Springer. 91–113.

ARHAR HOLDT, Špela, LAVRIČ, Polona, ROBLEK, Rebeka, GOLI, Teja, 2018: *Kategorizacija učiteljskih popravkov: Smernice za označevanje korpusa Šolar 2.0.* V1.0. Kazalnik projekta Nadgradnja korpusa Šolar. https://solar.trojina.si/wp-content/uploads/2022/05/Smernice-za-oznacevanje-korpusa-Solar-2.0-v1.0.pdf

GRANGER, Sylviane, 2008: Learner corpora. Anke Ludeling, Merja Kyto (ur.): *Corpus Linguistics. An International Handbook*. Berlin: Mouton de Gruyter. 259–275.

KOSEM, Iztok, ARHAR HOLDT, Špela, STRITAR KUČUK, Mojca, KREK, Simon, KRAPŠ VODOPIVEC, Irena, STABEJ, Marko, PORI, Eva, GOLI, Teja, LAVRIČ, Polona, LASKOWSKI, Cyprian, KOCJANČIČ, Polonca, KLEMENC, Bojan, ROZMAN, Tadeja, 2019: *Developmental corpus Šolar 2.0*. Slovenian language resource repository CLARIN.SI. http://hdl.handle.net/11356/1214

KOSEM, Iztok, ROZMAN, Tadeja, ARHAR HOLDT, Špela, KOCJANČIČ, Polonca, LASKOWSKI, Cyprian Adam, 2016: Šolar 2.0: nadgradnja korpusa šolskih pisnih izdelkov. Tomaž Erjavec, Darja Fišer (ur.): *Zbornik konference Jezikovne tehnologije in digitalna humanistika*. Ljubljana: Znanstvena založba Filozofske fakultete. 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

KOSEM, Iztok, STRITAR KUČUK, Mojca, MOŽE, Sara, ZWITTER VITEZ, Ana, ARHAR HOLDT, Špela, ROZMAN, Tadeja, 2020: *Analiza jezikovnih težav učencev: korpusni pristop*. Ljubljana: Znanstvena založba Filozofske fakultete. https://e-knjige.ff.uni-lj.si/znanstvena-zalozba/catalog/view/229/329/5311-1

LEECH, Geoffrey, 1997: Teaching and language corpora: A convergence. Anne Wichmann, Steven Fligelstone, Tony McEnery, Gerry Knowles (ur.): *Teaching and language corpora*. London: Longmann. 1–23.

POPIČ, Damjan, 2014. Revising translation revision in Slovenia. Tamara Mikolič Južnič, Kaisa Koskinen, Nike Kocijančič Pokorn (ur.): *New horizons in translation research and education 2*. Joensuu: University of Eastern Finland. http://epublications.uef.fi/pub/urn_isbn_978-952-61-1657-0/urn_isbn_978-952-61-1657-0.pdf

ROZMAN, Tadeja, STRITAR, Mojca, KOSEM, Iztok, 2012: Šolar – korpus šolskih pisnih izdelkov. Tadeja Rozman, Irena Krapš Vodopivec, Mojca Stritar, Iztok Kosem (ur.): *Empirični pogled na pouk slovenskega jezika*. Ljubljana: Trojina, zavod za uporabno slovenistiko.

STRITAR KUČUK, Mojca, 2012: *Korpusi usvajanja tujega jezika*. Ljubljana: Zveza društev Slavistično društvo Slovenije.

STRITAR KUČUK, Mojca, 2020: Modul Leto plus – prvi korak do korpusa slovenščine kot tujega jezika. Darja Fišer, Tomaž Erjavec (ur.): *Zbornik konference Jezikovne tehnologije in digitalna humanistika 2020*. Ljubljana: Inštitut za novejšo zgodovino. 131–135. http://nl.ijs.si/jtdh20/pdf/JT-DH_2020_StritarKucuk_Modul-Leto-plus%e2%80%93prvi-korak-do-korpusa-slovenscine-kot-tujega-jezika.pdf

STRITAR KUČUK, Mojca, 2022: *KOST, Korpus slovenščine kot tujega jezika: 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

VOLODINA, Elena, GRANSTEDT, Lena, MATSSON, Arild, MEGYESI, Beáta, PILÁN, Ildikó, PRENTICE, Julia, ROSÉN, Dan, RUDEBECK, Lisa, SCHENSTRÖM, Carl-Johan, SUNDBERG, Gunlög, WIRÉN, Mats, 2019: The SweLL Language Learner Corpus: From Design to Annotation. *Northern European Journal of Language Technology* 6. 67–104.

WIRÉN, Mats, MATSSON, Arild, ROSÉN, Dan, VOLODINA, Elena, 2019: SVALA: Annotation of Second-Language Learner Text Based on Mostly Automatic Alignment of Parallel Corpora. Inguna Skadina, Maria Eskevich (ur.): *Selected papers from the CLARIN Annual Conference 2018, Pisa, 8–10 October 2018*. Linköping: Linköping University Electronic Press. 227–239.

Projekt Razvoj slovenščine v digitalnem okolju sofinancirata Republika Slovenija in Evropska unija iz Evropskega sklada za regionalni razvoj. 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.