KOST 2.0: PREDSTAVITEV KORPUSA IN POTEK OZNAČEVANJA JEZIKOVNIH NAPAK

Mojca STRITAR KUČUK
Univerza v Ljubljani, Filozofska fakulteta

Članek predstavlja korpus slovenščine kot tujega jezika KOST in potek označevanja jezikovnih napak v njem. Uvodni predstavitvi nadgrajene različice korpusa in novega spletnega konkordančnika sledi opis kriterijev za izbor besedil, v katerih so označene napake, pri čemer so poglavitni prvi jezik tvorca besedila in okoliščine nastanka besedil. Označevanje napak, ki temelji na načelu minimalnega popravka, je potekalo z večjo ekipo označevalcev, sodelovanje študentov slovenistike pa se je izkazalo kot manj uspešno. V zadnjem delu prispevka je predstavljen še potek dela z orodjem za označevanje napak Svala.

Ključne besede: slovenščina kot neprvi jezik, korpus usvajanja jezika, označevanje jezikovnih napak, orodje Svala

1 UVOD

Pisni korpus slovenščine kot tujega jezika KOST od leta 2019 nastaja na Filozofski fakulteti Univerze v Ljubljani. Prva različica korpusa, KOST 1.0, je bila objavljena spomladi 2023, njena zasnova in vsebina sta podrobno dokumentirani v Stritar Kučuk, 2024.[1] Jeseni 2023 je bil objavljen povečani in dodatno označeni KOST 2.0.[2] Ker gre za zbirko besedil neprvih govorcev slovenščine, za katera so značilne pogoste jezikovne napake, je bil eden od pomembnih vidikov snovanja in gradnje tega korpusa označevanje teh napak (Granger, 2003). V tem prispevku bosta zato predstavljena nadgrajeni KOST in jezikoslovni pogled na nekatere vidike označevanja napak v njem.

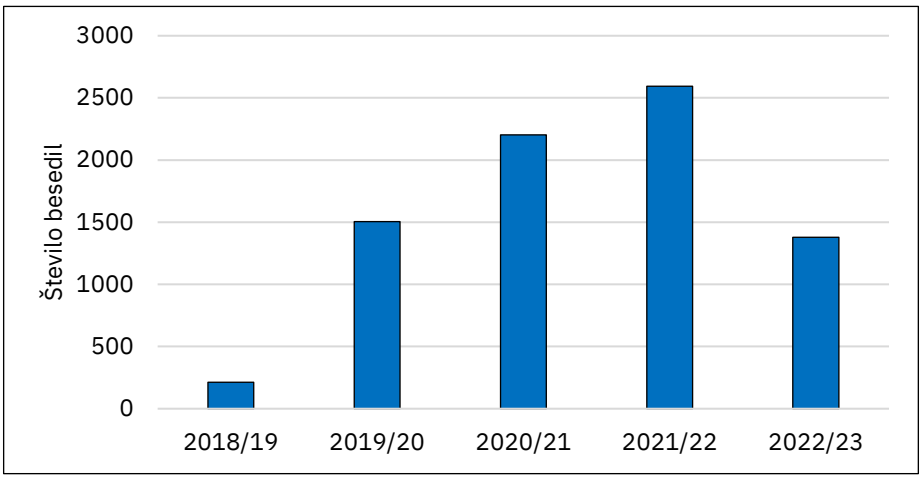
[1] Korpus KOST 1.0 je bil razvit v okviru projekta Razvoj slovenščine v digitalnem okolju, ki sta ga financirala Ministrstvo za kulturo Republike Slovenije in Evropski sklad za regionalni razvoj.

[2] Nadgradnja v KOST 2.0 je potekala s sredstvi Ministrstva za kulturo Republike Slovenije, prim. https://www.cjvt.si/korpus-kost/.

2 KOST 2.0

Trenutna različica korpusa, KOST 2.0, obsega 8347 besedil oz. 1.514.476 pojavnic. Zbiranje besedil se je pričelo v študijskem letu 2018/19 (Grafikon 1).[3]

Grafikon 1: Količina zbranih besedil po študijskih letih.



Ker je za vključitev pisnih besedil v korpus nujno, da njihovi tvorci podpišejo soglasje,[4] so bila vsa besedila za KOST pridobljena v različnih programih za organizirano poučevanje slovenščine kot drugega oz. tujega jezika, v katerih smo lahko dobili tudi ta soglasja: dobre tri četrtine v modulu Leto plus,[5] preostanek pa na različnih tečajih in lektoratih Centra za slovenščino kot drugi in tuji jezik v Sloveniji in po svetu.[6] Pri pridobivanju besedil je sodelovalo več kot 40 pedagoških in drugih sodelavcev teh programov.

V KOST vključena besedila so napisali 1303 različni tvorci, od tega sta dve tretjini žensk in tretjina moških. Vsi so anonimizirani, njihovi osebni podatki, ki se pojavijo v besedilih, pa so zakriti s kodami, npr. »Moje ime je [XImeX]

[3] Številka za študijsko leto 2022/23 je nižja, saj se je zbiranje besedil zaključilo sredi leta.

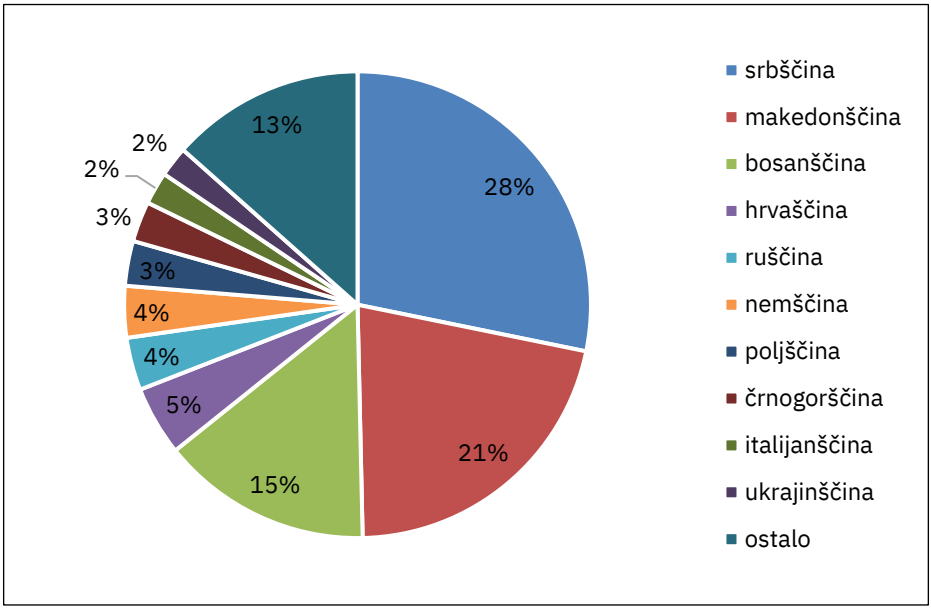
[4] V prispevku je moški spol uporabljen kot slovnično nevtralen.

[5] Prim. https://www.uni-lj.si/studij/leto-plus/.

[6] Prim. https://centerslo.si/.

[XPriimekX], in sem študentka prvega letnika«. Tvorci govorijo več kot 30 različnih prvih jezikov, največ jih je z južnoslovanskega govornega območja (Grafikon 2).

Grafikon 2: Delež tvorcev besedil glede na njihov prvi jezik.



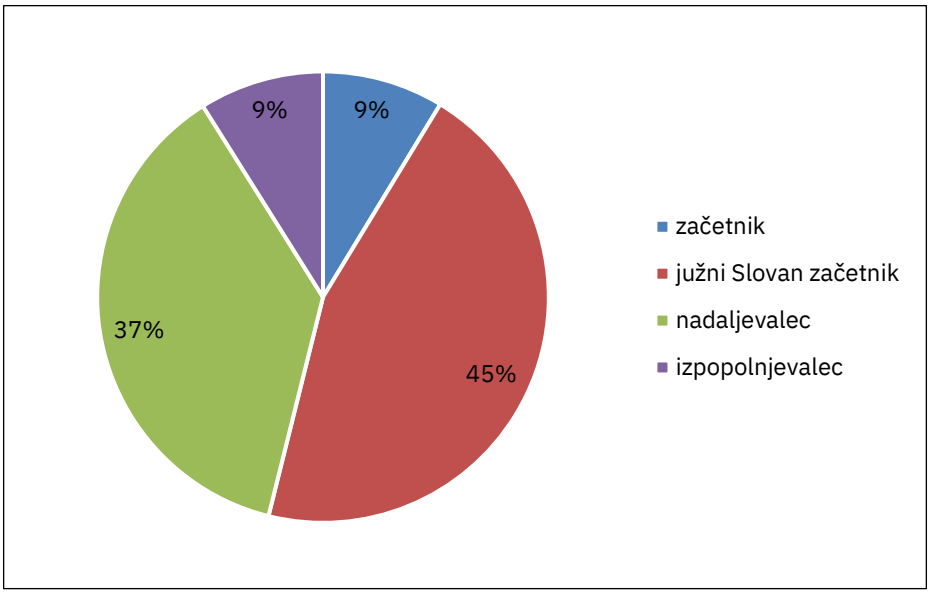
Med jeziki, navedenimi v kategoriji »ostalo« na Grafikonu 2, so španščina, nizozemščina, angleščina, slovaščina, češčina, kitajščina, slovenščina (v zamejstvu), madžarščina, japonščina, francoščina, romunščina, bolgarščina, portugalščina in grščina. Po dva ali en govorec govori albansko,[7] hebrejsko, igbojsko, indonezijsko, kirgiško, kirunsko, korejsko, švedsko in turško.

Poleg prvega jezika je za tiste, ki raziskujejo slovenščino kot neprvi jezik, pomemben podatek o tvorčevi trenutni jezikovni zmožnosti (Grafikon 3). Ta je zabeležena ob vsakem besedilu in označena s štirimi stopnjami: začetnik, nadaljevalec, izpopolnjevalec, južni Slovan začetnik (prim. Stritar Kučuk,

[7] Govorci albanščine predstavljajo eno od večjih priseljenskih skupin v Sloveniji (Razpotnik, 2024). Ker pa se ne udeležujejo organiziranih oblik jezikovnega poučevanja na Filozofski fakulteti Univerze v Ljubljani, graditelji KOST-a do besedil, ki jih morda pišejo v slovenščini, nimamo dostopa.

2024, str. 100–101). Vendar je treba upoštevati, da gre zgolj za približno oceno, namenjeno grobi orientaciji med besedili.

Grafikon 3: Deleži besedil glede na ocenjeno stopnjo jezikovne zmožnosti tvorcev.



Gradivo iz korpusa KOST je lematizirano in oblikoskladenjsko označeno, dostopno je v konkordančnikih noSketch[8] in KonText[9] ter v repozitoriju Clarin.[10] Izpostaviti pa moram novo razviti spletni konkordančnik,[11] ki omogoča uporabniku prijazen prikaz konkordanc tako v besedilih z označenimi jezikovnimi napakami kot na neoznačenih besedilih, skupaj z relevantnimi metapodatki. Nazorno so prikazane povezave med izvornim, tvorčevim besedilom in njegovo popravljeno verzijo ter oznake vrste jezikovnih napak. Poleg tega je mogoče iskanje po tipih napak (Slika 1), po okolici in po seznamih. Konkordančnik je prosto dostopen[12] in z manjšimi prilagoditvami

[8] Prim. https://www.clarin.si/ske/#dashboard?corpname=kost20_orig.

[9] Prim. https://www.clarin.si/kontext/query?corpname=kost20_orig.

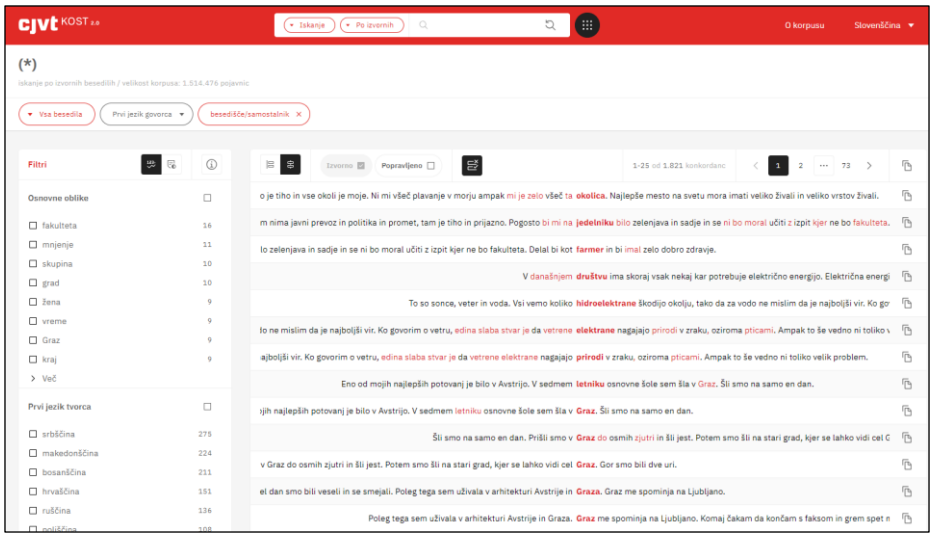
[10] Prim. http://hdl.handle.net/11356/1887.

[11] Prim. https://kost.cjvt.si/.

[12] Prim. https://github.com/TheRokUrbanc/Kost-CJVT.

uporaben tudi za druge korpuse, ki vsebujejo jezikovne popravke. Med drugim je uporabljen tudi za slovenski korpus šolskih besedil Šolar 3.0.[13]

Slika 1: Prikaz iskanja po napakah besedišča pri samostalniku.



3 KRITERIJI ZA IZBIRO BESEDIL ZA OZNAČEVANJE JEZIKOVNIH NAPAK

Označevanje jezikovnih napak v besedilih KOST-a sem zasnovala in usklajevala urednica korpusa in avtorica tega prispevka. Sprva sem besedila, na katerih sem označila napake, izbirala naključno. S povečanjem količine označenih besedil in vključevanjem drugih označevalcev pa je izbor začel potekati po dveh kriterijih. Oba sta neposredno povezana z izkušnjami in potrebami učiteljev slovenščine kot drugega jezika, ki so ena od pomembnejših skupin uporabnikov KOST-a.

3.1 Prvi jezik tvorca

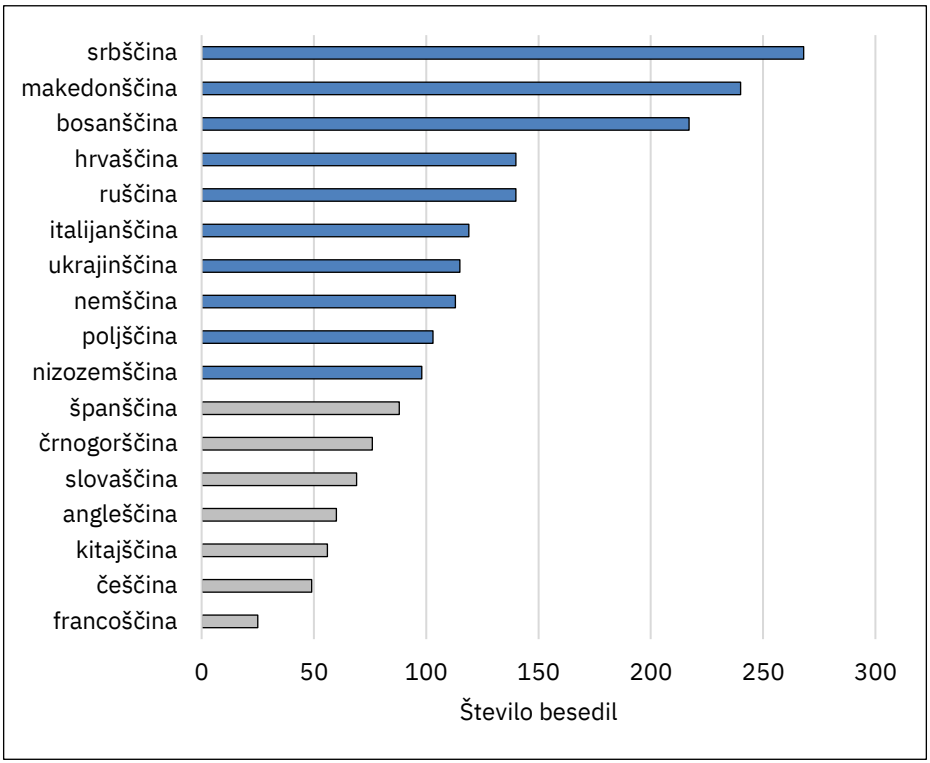
Prvi jezik tvorca je poleg predznanja ključni dejavnik, na podlagi katerega organizatorji jezikovnega poučevanja za slovenščino kot drugi jezik razvrščajo udeležence v skupine (Klinar in sod., 2022, str. 187), saj vpliva na didaktiko poučevanja, npr. pri odpravljanju posledic negativnega jezikovnega prenosa.

[13] Prim. https://solar.cjvt.si/.

Zato je to, podobno kot v številnih obstoječih korpusih usvajanja (prim. Mikelić Preradović, 2020; Rosen in sod., 2020; Volodina in sod., 2019; Gilquin, 2015, str. 17; Hammarberg, 2010), tudi najpomembnejši kriterij v korpusu KOST. Pri izbiri besedil za označevanje napak smo se omejili na jezike, za katere imamo dostop do zadostnega števila besedil, po možnosti so čim bolj raznoliki, hkrati pa so relevantni z vidika organiziranega poučevanja slovenščine.

Kot je razvidno z Grafikona 4, je trenutno največ označenih besedil govorcev srbščine, makedonščine in bosanščine. Da bi bile mogoče čim bolj zanimive in zanesljive primerjave tudi za govorce jezikov iz drugih jezikovnih skupin, je bil postavljen cilj vsaj sto označenih besedil za govorce vsakega prvega jezika. Le s približno enako velikimi podkorpusi bodo namreč ti uravnoteženi in s tem primerljivi. Pri tem je uravnoteženo število besedil zgolj pragmatičen, najlažje avtomatsko dostopen podatek. Sam po sebi še ne zagotavlja primerljivih velikosti podkorpusov, saj so besedila lahko zelo različnih dolžin. Na Grafikonu 4 so s sivo barvo označeni prvi jeziki, pri katerih želimo pridobiti še dodatna besedila in pri katerih bo to v prihodnosti zaradi razmeroma stalnega dotoka besedil predvidoma izvedljivo. Če se bo občutno povečal delež besedil govorcev še katerega prvega jezika, pri katerem trenutno nimamo večje količine zbranih tekstov, pa bo seveda vključen v označeni del korpusa.

Grafikon 4: Velikost podkorpusov z označenimi jezikovnimi napakami glede na prvi jezik tvorca.

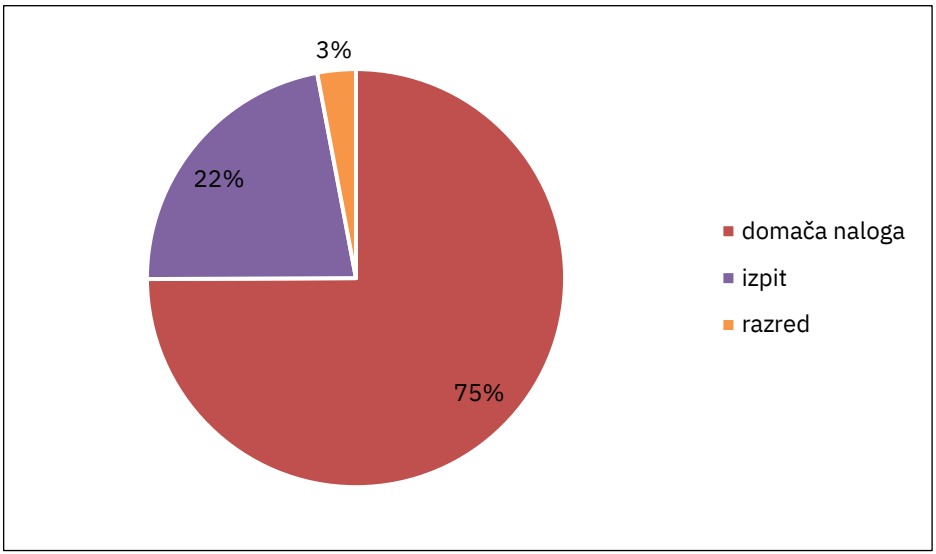


3.2 Okoliščine nastanka besedil

Drugi kriterij za izbor tekstov, ki se je pokazal kot ključen med samim procesom označevanja in ga graditelji ostalih primerljivih korpusov izrecno ne izpostavljajo (prim. Mikelić Preradović, 2020; Rosen in sod., 2020; Volodina in sod., 2019; Hammarberg, 2010), so okoliščine njihovega nastanka. Skoraj tri četrtine v KOST vključenih besedil so domače naloge (Grafikon 5), ki jih tvorci napišejo doma, brez nadzora učitelja. Večinoma so napisane v digitalni obliki in s tem enostavne za vključitev v digitalno zbirko besedil. Bistveno manj je besedil z izpitov, ki nastajajo v kontroliranih pogojih, na internih izpitih ali testih na tečajih ali lektoratih slovenščine, in besedil, napisanih v okviru različnih dejavnosti med samim jezikovnim poukom. Čeprav so takšna besedila napisana na roko in jih je za KOST treba ročno digitalizirati, jim

dajemo prednost pri označevanju napak. Pri teh besedilih so namreč okoliščine, v katerih so nastala, transparentne in enake za vse. V besedilih prvega tipa posebej pri manj motiviranih tvorcih skozi leta opažamo naraščanje uporabe strojnih prevajalnikov in drugih jezikovnih orodij (Stritar Kučuk, 2024, str. 98), zato se postavlja vprašanje, koliko odražajo dejansko jezikovno zmožnost njihovih tvorcev v slovenščini (Stritar Kučuk, 2023c). Tudi zaradi tega v označeni del KOST-a ne vključujemo doma napisanih besedil, ki so popolnoma brez napak ali z zelo malo napakami, saj je pri njih velika verjetnost, da so bila napisana z (izključno) uporabo strojnih prevajalnikov.

Grafikon 5: Deleži besedil v KOST-u glede na okoliščine nastanka.



4 POSTOPEK OZNAČEVANJA JEZIKOVNIH NAPAK

Za razliko od večine korpusnih oznak je oznake jezikovnih napak v korpuse treba dodati ročno. To ne velja samo za KOST; po dostopnih podatkih še ni korpusa usvajanja za druge jezike, v katerem bi ta proces zadovoljivo potekal avtomatsko (prim. Rosen in sod., 2020). Takšno označevanje je zamudno, zato so napake običajno označene samo na delu korpusa. V KOST-u 2.0 je to ok. 1740 besedil oziroma slabih 24 % vseh besedil.

4.1 Načelo minimalnega popravka

Vsaka neustrezna pojavitev v korpusu dobi oznako glede na taksonomijo napak (Tabela 1), pripisana pa ji je ciljna hipoteza oz. popravljena/normalizirana oblika (Lüdeling, Hirschmann, 2015, str. 143). Pri tem se držimo načela minimalnega popravka (Volodina in sod., 2019, str. 81): s popravki čim manj posegamo v besedilo in popravljamo čim manj napak, pa čeprav popravljena verzija ne zveni stoodstotno tako, kot bi jo tvorili domači govorci ciljnega jezika (Granger in sod., 2022, str. 3). Popravljamo predvsem napake na ravni zapisa, besedišča in oblike besed. V skladnjo skušamo posegati čim manj, prav tako se izogibamo stilističnim popravkom.

Tabela 1: Taksonomija jezikovnih napak v korpusu KOST.

Krovna kategorija	*Kategorija napake*	*Primer iz KOST-a*[14]
Napake zapisa	Ločilo	Naslednje jutro**,** babica se ne počuti zelo dobro > Naslednje jutro se babica […]
	Črkovanje	**uporavljam** slovnice > uporabljam slovnice
	Skupaj oz. narazen	sem **več krat** padla > sem večkrat padla
	Mala oz. velika začetnica	vidiš Trst in **Hrvaško** obalo > […] in hrvaško obalo
	Krajšave	s sladkim sadjem, z jogurtom, **i. t. n.** > […] z jogurtom, itn.
Napake besedišča	Samostalnik	Potovale smo v **Venecijo** > […] v Benetke
	Glagol	Radijo **slušamo** > Radio poslušamo
	Zaimek	odločil sem se da nadaljujem **moj** študij > […] nadaljujem svoj študij
	Pridevnik	**amerkanski** zdravniki > ameriški zdravniki
	Prislov	**Izprva** sem bil bolj skeptičen > Sprva […]
	Predlog	grem **v** sprehod > grem na sprehod
	Veznik	Upam da ne bo težav naslednje leto **kdaj** pridem > […] leto, ko pridem
	Ostalo	**Več** znam govoriti Englesko > Že znam […]

[14] Vsi primeri v članku so iz KOST-a. V tej tabeli je napaka je odebeljena, za znakom > pa je navedena popravljena oblika.

Napake oblike	Samostalnik	sem bila v samoizolaciji 28 **dna** > [...] dni
	Glagol	sem še **živil** > sem še živel
	Zaimek	Po **mom** mnenju > Po mojem mnenju
	Pridevnik	Film mi je bil všeč ker ni bil **težki** > [...] težek
	Prislov	**najraji** igram šah > najraje [...]
	Ostalo	Čez **štirimi** leta > Čez štiri leta
Napake skladnje	Struktura	Okoli 8 **mi je prišel** fant > Okoli 8 je k meni prišel fant
	Besedni red	**Tudi rada grem** v nakupovalno središče > Rada grem tudi v nakupovalno središče
	Izpuščeni jezikovni elementi	Druge jezike učimo, ker je to zanimivo > Druge jezike **se** učimo [...]
	Odvečni jezikovni elementi	priporočam da **si** ga obiščete > priporočam da ga obiščete

Kadar se zgodi, da napačni pojavitvi ne znamo pripisati popravljene, to označimo s tremi vprašaji v oglatem oklepaju: *[???]*. Takih primerov je malo, v KOST-u 2.0 le 160, npr. *To mi izboljša kakovost življenja, da lahko izgubim svoje obveznosti in ostanem v **čaravni** sreči*.

Pri označevanju napak se srečamo tudi z vprašanjem, kateri je ciljni jezik korpusa oz. norma, h kateri naj bi težili tvorci v korpus vključenih besedil. Čeprav se pri korpusu slovenščine odgovor zdi preprost, se zaplete npr. pri pogovornih izrazih. Lahko gre samo za vprašanje zapisa, npr. *tut* namesto *tudi*, *zgubiti* nam. *izgubiti*, ali dejanske leksikalne izbire, npr. *štamprl* namesto *šilce*. Ker je načeloma ciljni jezik tistih, ki se učijo slovenščino kot drugi oz. tuji jezik v eni od organiziranih oblik poučevanja, standardna oz. knjižna slovenščina, so tovrstne pojavitve v KOST-u označene kot napake (podobno je v češkem korpusu CzeSL, prim. Rosen in sod., 2020, str. 68). Pogovorno zaznamovano besedišče, npr. *faks, bus, flet*, pa dopuščamo v manj formalnih besedilnih vrstah, npr. v elektronskih pismih.

4.2 Označevalci jezikovnih napak

V uvodnih fazah sem jezikovne napake v besedilih iz KOST-a označevala sama, sčasoma pa se je ekipa označevalcev širila. V označevanje so se vključili študenti tretjega letnika slovenistike, ki so v študijskih letih 2021/22, 2022/23 in 2023/24 v okviru seminarja pri predmetu Slovenščina kot tuji jezik označili

manjšo količino korpusnih besedil. V treh letih je sodelovalo 54 študentov, ki so skupaj označili 197 besedil. Čeprav so delo ocenili pozitivno – kot zanimivo, a razmeroma zahtevno – se tovrstni način pridobivanja označenih korpusnih tekstov ni izkazal za učinkovitega. V besedilih, ki so jih označili, je bilo zaradi njihove tendence k pretiranemu (stilističnemu) popravljanju besedil, slabega znanja slovenskega pravopisa in slovnice ter nepozornosti in naglice pri delu toliko neustreznih oznak, da jih ni bilo mogoče vključiti neposredno v označeni KOST, pač pa jih je bilo treba pozorno pregledati in/ali na novo označiti (prim. Stritar Kučuk, 2023b).

Nazadnje je jezikovne napake v besedilih za KOST označevalo 12 sodelavcev: profesorjev slovenistike, ki poučujejo slovenščino kot drugi jezik, in študentov slovenistike, ki so se med seminarskim delom pokazali kot uspešnejši. Pred začetkom dela so bili vsi deležni individualnega usposabljanja, med označevanjem pa so si pomagali s priročnikom za označevanje napak (Stritar Kučuk, 2023a), v katerem so natančnejša navodila za označevanje, odločanje v primeru dvoumnosti in primeri neustreznih oznak napak. Pokazalo se je, da je pri tovrstnem delu nujna določena mera izkušenj s slovenščino kot neprvim jezikom. Zaradi konsistentnosti oznak je bilo nujno tudi, da sem vsa označena besedila še enkrat pregledala urednica KOST-a in poenotila ter popravila oznake.

4.3 Potek dela z orodjem za označevanje napak

Označevanje jezikovnih napak v tujih korpusih poteka s pomočjo različnih orodij oz. aplikacij, ki so bodisi razvite za potrebe točno določenega korpusa (za češki korpus CzeSL so razvili program Feat, Rosen in sod., 2020, str. 173–174) bodisi gre za prilagojene aplikacije obstoječih korpusov. Napake v hrvaškem korpusu CrolTec so denimo označene v okolju TEITOK, ki je uporabljen tudi v portugalskem korpusu učečih se COPLE2 (Mikelić Preradović, 2020, str. 910). Za ta način smo se odločili tudi pri KOST-u. Napake označujemo v aplikaciji Svala,[15] ki je bila razvita za potrebe švedščine kot tujega jezika v okviru projekta SweLL (prim. Wirén in sod., 2018), nato pa v okviru projekta Razvoj slovenščine v digitalnem okolju prilagojena

[15] Trenutno v različici 1.1, prim. https://orodja.cjvt.si/svala/.

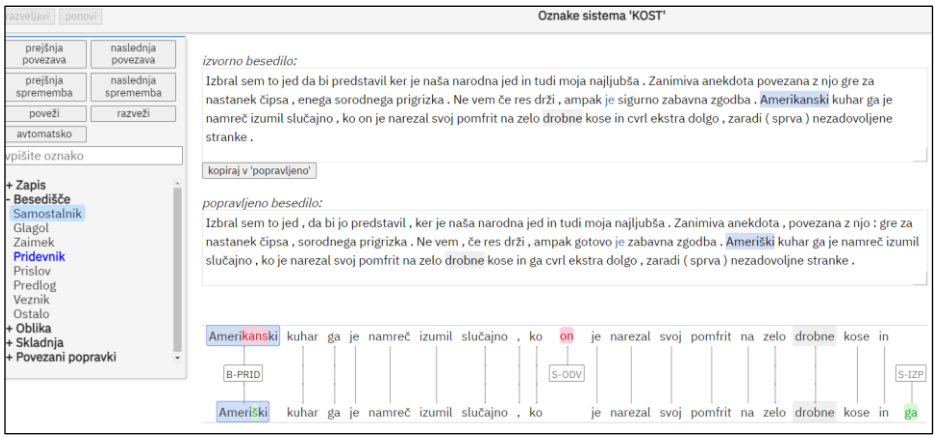
slovenskim specifikam (prim. Arhar Holdt in sod., 2022). Svala omogoča jasno vizualizacijo, dodajanje metajezikovnih oznak ter vzporeden prikaz izvornih in popravljenih besedil, poleg tega pa je odprto dostopna in uporabna za več projektov.[16]

Označevanje besedil za KOST s pomočjo orodja Svala poteka po naslednjem postopku:

- Označevalec dobi besedila, ki jih mora označiti, vsako v posebni tekstovni datoteki (format *txt*), v anonimizirani obliki in brez kakršnih koli metapodatkov. Ime datoteke je enako kodi besedila.
- V Svali označevalec označi vsak odstavek besedila posebej. Če gre za daljše besedilo, v katerem tvorec ni naredil odstavkov, ga zaradi preglednejšega prikaza v konkordančniku segmentira na odstavke po smislu.
- Vsakega od odstavkov označevalec posebej vnese v aplikacijo Svala. V zgornje okence (»izvorno besedilo«) skopira izvorno besedilo, v spodnje okence (»popravljeno besedilo«) pa vnese popravljeno verzijo besedila.
- Kjer je mogoče, orodje Svala avtomatsko poveže izvirno in popravljeno besedilo. Povezave med obema, t. i. špageti, so vidne v najnižjem okencu (Slika 2). Povezave je mogoče ročno spreminjati, dodajati, združevati ali razdruževati, označevalec pa vsaki povezavi določi eno ali več oznak napak iz izbirnika na levi.
- Označeno besedilo oz. posamezni odstavek označevalec izvozi v formatu *json*, pri čemer ohrani izvorno ime datoteke, in ga pošlje uredniku korpusa.

[16] Poleg KOST-a je bila zaenkrat uporabljena pri označevanju korpusa Šolar, uporablja pa se tudi pri nastajajočem korpusu revidiranih besedil z območja jezikovnega stika med slovenščino in italijanščino STIKit.

Slika 2: Primer označevanja napak v orodju Svala.



Kot je bilo že omenjeno, označevalec v Svali vidi celotni odstavek, pri določanju tipov napak pa celo samo eno poved. To je tudi eden od razlogov, zakaj v KOST ne vključujemo besedil, v katerih tvorec ne postavlja ločil. Zelo dolg odstavek lahko za Svalo predstavlja preveliko obremenitev. Tak je primer naslednjega kratkega besedila z le dvema končnima ločiloma, ki ni vključeno v označeni KOST:

> Včeraj smo imeli naš naslednj profesorica slovenčina svoj ime je [XImeX], ona prihaja iz [XKrajX] ima 32 let v svoj prostem času druži se z svoj pes, svoj najljbši potovanje je bilo na Praga, pošluša glasbo Rock ona tudi priporočela Batista Cadillac jaz osebno sem ga šlišala in mi je bilo všeč (dobro okus za glasba ima). Svenina nimara gavori slovensko, hrvasko in malo španščina in svoj najljbši film je bilo Wolf Street Film.

5 ZAKLJUČEK

Korpus slovenščine kot tujega jezika KOST 2.0 je sodoben jezikovni vir, ki je s svojimi parametri primerljiv z obstoječimi korpusi za druge svetovne jezike (Stritar Kučuk, 2022), zaradi bogatih metaoznak in tudi zaradi novorazvitega, uporabniku prijaznega konkordančnika pa je na voljo za širok spekter jezikoslovnih analiz in aplikacij. Ob tem se morajo uporabniki podatkov zavedati, da označevanje jezikovnih napak v besedilih neprvih govorcev slovenščine odpira številna jezikoslovna vprašanja – jezikovnotehnološka so v tem prispevku seveda puščena ob strani. Odgovori na ta vprašanja niso

samoumevni ali enoznačni. Sama se z njimi srečujem že od snovanja poskusnega korpusa slovenščine kot tujega jezika (prim. Stritar, 2012) naprej. Ali gre pri rabi neustreznega glagolskega vida, npr. glagola *počakati* nam. *čakati* v primeru *Nisem mogla veliko s njim in sem ga veliko počakala*, za napako oblike ali besedišča? Izbrati bi bilo mogoče oboje, a ker je nujno, da so oznake konsistentne, se je treba odločiti za eno možnost, v tem primeru je bilo izbrano besedišče. Tovrstne dileme se ob označevanju napak pojavljajo tako rekoč na dnevni ravni. O njihovih rešitvah je nemalokrat mogoče dvomiti, a celo v primerih, ko niso najboljše, je treba v zvezi z njimi vztrajati (prim. Lüdeling, Hirschmann, 2015, str. 144). Le konsistentno označevanje omogoča razmeroma enostavno izvedljive spremembe in izboljšave korpusnega gradiva v prihodnosti. Vsakič znova pa se tudi potrjuje, da je označevanje jezikovnih napak kompleksno delo, ki zahteva označevalca z izkušnjami s področja slovenščine kot neprvega jezika in s sposobnostjo sprejemanja pragmatičnih rešitev.

6 LITERATURA

Arhar Holdt, Š., Kosem, I. in Stritar Kučuk, M. (2022). Metode in orodja za lažjo pripravo korpusov usvajanja jezika. V N. Pirih Svetina, I. Ferbežar (ur.), *Na stičišču svetov* (str. 23–30). Založba Univerze. https://doi.org/10.4312/Obdobja.41.23-30

Gilquin, G. (2015). From design to collection of learner corpora. V S. Granger, G. Gilquin, F. Meunier (ur.), *The Cambridge Handbook of Learner Corpus Research* (str. 9–34). Cambridge University Press. https://doi.org/10.1017/CBO9781139649414

Granger, S. (2003). Error-tagged learner corpora and CALL: A promising synergy *CALICO Journal,* 20(3), 465–479.

Granger, S., Swallow, H. in Thewissen, J. (2022). *The Louvain Error Tagging Manual Version 2.0*. Centre for English Corpus Linguistics, Université catholique de Louvain. Pridobljeno januarja 2024, https://cdn.uclouvain.be/groups/cms-editors-cecl/Granger%20et%20al._Error%20tagging%20manual_v2.0_2022.pdf

Hammarberg, B. (2010). *Introduction to the ASU Corpus: A longitudinal oral and written text corpus of adult learner Swedish with a corresponding part from native Swedes.*http://www.diva-portal.org/smash/get/diva2:778204/FULLTEXT01.pdf

Klinar, M., Pisek, S., Stritar Kučuk, M. in Šter, H. (2022). Poučevanje slovenščine za redno vpisane tuje študente na Univerzi v Ljubljani. V N. Pirih Svetina, I. Ferbežar (ur.), *Na stičišču svetov* (str. 185–194). Založba Univerze v Ljubljani. https://doi.org/10.4312/Obdobja.41.185-194

Lüdeling, A. in Hirschmann, H. (2015). Error annotation systems. V S. Granger, G. Gilquin, F. Meunier (ur.), *The Cambridge Handbook of Learner Corpus Research* (str. 133–157). Cambridge University Press. https://doi.org/10.1017/CBO9781139649414

Mikelić Preradović, N. (2020). Označavanje pogrešaka u CroLTeC-u (računalnom učeničkom korpusu hrvatskog kao stranog jezika). *Rasprave: Časopis Instituta za hrvatski jezik i jezikoslovlje,* 46(2), 899–920.

Razpotnik, B. (2024). Vsak sedmi prebivalec Slovenije se je v Slovenijo priselil. Statistični urad Republike Slovenije. https://www.stat.si/StatWeb/news/Index/9999

Rosen, A., Hana, J., Hladká, B., Jelínek, T., Škodová, S. in Štindlová, B. (2020). *Compiling and annotating a learner corpus for a morphologically rich language: CzeSL, a corpus of non-native Czech*. Karlova univerza.

Stritar, M. (2012). *Korpusi usvajanja tujega jezika*. Zveza društev Slavistično društvo Slovenije.

Stritar Kučuk, M. (2022). KOST med korpusi usvajanja tujega jezika. V N. Pirih Svetina, I. Ferbežar (ur.), *Na stičišču svetov* (str. 323–334). Založba Univerze v Ljubljani. https://doi.org/10.4312/Obdobja.41.323-334

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

Stritar Kučuk, M. (2023b). Error annotation in Slovene learner corpus KOST – why L1 students can(not) do the job. *CLARC 2023: Jezik i jezični podaci*. https://uniri-my.sharepoint.com/:w:/g/personal/bperak_uniri_hr/EdBOkvsg4vJOrVeHTkQw3uYB16acgdyFh2g5S5fpdXqhYA?rtime=XEOZp-9w3Eg

Stritar Kučuk, M. (2023c). Neizbežno spodbudna ovira: vpliv strojnega prevajanja na pisno produkcijo v slovenščini kot drugem jeziku. V M. Leskovec, I. Samide (ur.), *Z jeziki danes za jutri: Aktualni izzivi poučevanja jezikov, literatur in kultur. Zbornik povzetkov prispevkov konference* (str. 66). Založba Univerze v Ljubljani.

Stritar Kučuk, M. (2024). Prvi korpus slovenščine kot tujega jezika KOST 1.0. V Š. Arhar Holdt, S. Krek (ur.), *Razvoj slovenščine v digitalnem okolju* (str. 93–117). Založba Univerze. https://doi.org/10.4312/9789612972561

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

Wirén, M., Matsson, A., Rosén, D. in Volodina, E. (2018). SVALA: Annotation of Second-Language Learner Text Based on Mostly Automatic Alignment of Parallel Corpora. *Selected papers from the CLARIN Annual Conference 2018* (str. 227–239).

KOST 2.0: CORPUS PRESENTATION AND LANGUAGE ERROR TAGGING WORKFLOW

This article presents the corpus of Slovene as a foreign language KOST and the process of tagging language errors in it. An introductory presentation of the upgraded version of KOST and the new online concordancer is followed by a description of the criteria for selecting texts for error tagging. These include the learner's first language and the preference for texts written under exam conditions. Finally, a linguistic view of the process of error tagging is presented, based on the principle of minimal correction. The process of working with a larger team of annotators is described, with an attempt to include university students of Slovene studies. The final section of the paper presents the workflow with the Svala editor tool.

Keywords: Slovene as a non-native language, learner corpus, error tagging, Svala tool