Simpozij OBDOBJA 41

Mojca Stritar Kučuk

KOST med korpusi usvajanja tujega 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/

Obdobja (e-ISSN 2784-7152)

Univerza v Ljubljani
Filozofska fakulteta

CENTER ZA SLOVENŠČINO
KOT DRUGI IN TUJI JEZIK

KOST MED KORPUSI USVAJANJA TUJEGA JEZIKA

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

DOI:10.4312/Obdobja.41.323-334

V prispevku je predstavljen pisni korpus usvajanja slovenščine kot tujega jezika KOST, poudarek pa je na njegovem položaju med obstoječimi korpusi tega tipa, zgrajenimi za druge ciljne jezike. Glede na sorodni sociolingvistični položaj je KOST mogoče primerjati s slabo desetino med več kot 190 korpusi usvajanja tujega jezika. Ugotovimo lahko, da je KOST s svojo zasnovo, trenutno velikostjo skoraj 835.000 besed, delno označenimi jezikovnimi napakami in prostim dostopom do podatkov s temi korpusi popolnoma primerljiv in kot tak uporaben vir za raznovrstne raziskave.

korpus usvajanja tujega jezika, slovenščina kot drugi in tuji jezik, klasifikacija napak

This article presents the written Slovenian learner corpus KOST, focusing on its position among other learner corpora for other target languages. In terms of the sociolinguistic position of the target language, KOST can be compared with approximately one-tenth of more than 190 learner corpora. With its design, current size of almost 835,000 words, partially tagged language errors, and free access to data, KOST is fully comparable to these corpora and thus a useful resource for various forms of language research.

learner corpus, Slovenian as a second and foreign language, error classification

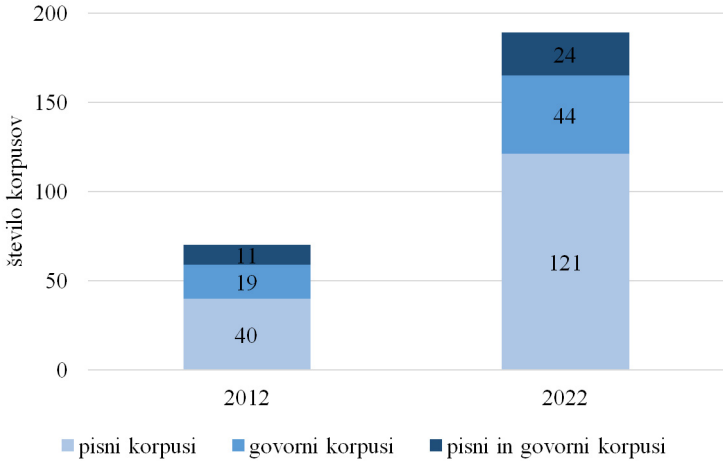
1 Uvod

Jezikovni korpusi so že vrsto let temeljni jezikovni vir. Poleg splošnih ali referenčnih korpusov so v zadnjem desetletju razmah doživeli tudi korpusi usvajanja tujega jezika (angl. *learner corpora*). Število tovrstnih korpusov za vse svetovne jezike je glede na seznam obstoječih korpusov (Centre for English Corpus Linguistics 2022) poskočilo s 73 korpusov leta 2012 (Stritar 2012: 26–45) na 191 korpusov leta 2022. V prispevku bom predstavila izbor obstoječih korpusov, ki so za slovensko situacijo bolj relevantni, nato pa se bom osredotočila na korpus slovenščine kot tujega jezika KOST.[1]

2 Obstoječi korpusi usvajanja tujega jezika

Glede prenosnika besedil, ki sestavljajo korpuse usvajanja tujega jezika, se razmerja v zadnjem desetletju niso pretirano spremenila (Slika 1). Še vedno je največ pisnih korpusov, za katere je tudi najlažje pridobivati besedila.

1 Prim. https://www.cjvt.si/korpus-kost/ (dostop 26. 4. 2022).



Slika 1: Obstoječi korpusi glede na prenosnik.[2]

Večina trenutnih korpusov usvajanja tujega jezika ima en ciljni jezik, dobra desetina, natančneje 23, pa jih vključuje več. Seveda prednjači angleščina; leta 2012 je bilo za angleščino kot ciljni jezik 64 % vseh korpusov, deset let kasneje pa je njihov delež vendarle nekoliko upadel na 52 %. Ker pa ima angleščina kot globalno dominantni jezik specifično sociolingvistično situacijo, neprimerljivo z večino drugih jezikov, bom v tej analizi nanjo usmerjene korpuse usvajanja pustila ob strani. Med preostalimi ciljnimi jeziki so: arabščina, češčina, estonščina, finščina, francoščina, gelščina, hrvaščina, islandščina, italijanščina, katalonščina, kitajščina, korejščina, latvijščina, litovščina, madžarščina, nemščina, nizozemščina, norveščina, perzijščina, poljščina, portugalščina, romunščina, ruščina, španščina in švedščina. Tudi na korpuse jezikov, ki imajo večje število domačih govorcev in so pogostejši cilj učenja tujega jezika po svetu, se v tem prispevku ne bom osredotočala, saj so za slovensko situacijo manj relevantni. Omenim naj le, da za španščino kot tuji jezik obstaja 15 korpusnih projektov, za nemščino 13, za francoščino 10, za italijanščino 9, za kitajščino in portugalščino pa po dva.[3]

Za nas so zanimivi korpusi slovanskih jezikov, ker so sorodni slovenščini in so v primerjavi z globalno bolj prisotnimi jeziki manj razširjeni med tujimi govorci. Osnovni podatki o njih so prikazani v Tabeli 1 (Boyd idr. 2014; Gajdošová 2021: 10; Mikelić Preradović idr. 2015; Mikelić Preradović 2020: 902–904; Rakhilina idr. 2016; Rosen 2017; Zasina idr. 2020).

2 Prim. Boyd idr. 2014; Gajdošová 2021; Mikelić Preradović idr. 2015; Mikelić Preradović 2020; Rakhilina idr. 2016; Rosen 2017; Zasina idr. 2020.

3 Večjezični korpusni projekti v te številke niso zajeti.

Ime korpusa	Ciljni jezik	Velikost	Opomba in povezava
CroLTeC (CROatian Learner TExt Corpus)	hrvaški	1 milijon besed	Označeni so popravki, ki so jih v besedilih naredili sami tvorci. V spletnem vmesniku je mogoče dostopati tudi do celotnih besedil. http://teitok.clul.ul.pt/croltec/
CzeSL (The Corpus of Czech as a Second Language)	češki	2 milijona besed (neoznačeni korpus), 1 milijon besed (označeni korpus)	http://utkl.ff.cuni.cz/learncorp/
Merlin	češki, nemški, italijanski	80.000 besed	Besedila so natančno umeščena na lestvici SEJO od A1 do C1. http://www.merlin-platform.eu
ERRKORP	slovaški	137.000 besed	https://korpus.sk/errkorp.html
PoLKo	poljski	11.000 besed	http://slawistyka.uw.edu.pl/pl/the-polish-learner-corpus/
RLC (The Russian Learner Corpus)	ruski	730.000 besed	Vključuje longitudinalni podkorpus akademskega pisanja, ki spremlja pisno produkcijo istih tvorcev skozi štiri leta. V korpusu ločujejo med ruščino kot drugim in kot dediščinskim jezikom (angl. *heritage language*). http://web-corpora.net/RLC

Tabela 1: Najpomembnejši korpusi usvajanja za slovanske jezike.

Za vse slovanske korpuse velja, da imajo več kot en izhodiščni jezik, tvorci pa so bodisi na različnih stopnjah jezikovne zmožnosti bodisi podatek o tem ni dostopen. Vsi korpusi so pisni razen RLC, ki vključuje tudi govorni del. Večina je lematizirana in oblikoskladenjsko označena, na vsaj delu besedil pa so tudi ročno označene napake. Sicer so bolj dodelani in bolje dokumentirani korpusi CzeSL, CroLTeC in RLC, medtem ko sta slovaški in poljski korpus še v začetnih fazah.

Od neslovanskih jezikov pa si oglejmo korpuse iz nekaterih baltskih ali skandinavskih držav, ki jih prikazuje Tabela 2 (Darǧis idr. 2020; Hammarberg 2010; Stritar 2012: 42–43).

Ime korpusa	Ciljni jezik	Velikost	Prenosnik	Opomba in povezava
ASU (Andrespråkets strukturutveckling)	švedski	493.000 besed	pisni, govorni	Longitudinalni korpus.
ASK (Norsk andrespråkskorpus)	norveški	770.000 besed	pisni	Glavni zgled pri zasnovi poskusnega slovenskega korpusa PiKUST. https://www.nb.no/sprakbanken/en/resource-catalogue/oai-clarino-uib-no-ask/
NORINT	norveški	104.000 besed	govorni (pisni)	Vključuje intervjuje s tujejezičnimi govorci, pogovore med njimi, posnetke njihovega glasnega branja in tudi nekaj pisnih besedil. Govorci so vsaj na stopnji B1. https://www.hf.uio.no/iln/english/about/organization/text-laboratory/projects/norint/index.html
IceL2EC (The Icelandic L2 Error Corpus)	islandski	125.000 besed	pisni	https://repository.clarin.is/repository/xmlui/handle/20.500.12537/106
ICLFI (International Corpus of Learner Finnish)	finski	ni podatka	pisni	Besedila so zbrali učitelji finščine z vsega sveta. https://www.kielipankki.fi/corpora/iclfi/
EIC (Estonian Interlanguage Corpus)	estonski	1 milijon besed	pisni	Vključuje podkorpus ruščine kot prvega jezika in referenčni podkorpus argumentativnih časopisnih člankov. Besedila so bila zbrana na izpitih, jezikovnih tečajih in na srednješolski olimpijadi estonščine kot drugega jezika. https://evkk.tlu.ee/vers1//?language=en
ESAM	litovski, latvijski	52.000 besed	pisni	http://esam.korpuss.lv/
LAVA (Learner Corpus of Latvian)	latvijski	190.000 besed	pisni	http://lava.korpuss.lv/en/

Tabela 2: Najpomembnejši korpusi usvajanja za nekatere baltske in skandinavske jezike.

Tudi ti korpusi so zasnovani podobno kot za slovanske jezike. Dobro dokumentiran je estonski EIS, kar ne čudi, saj je Estonija na področju jezikovnih tehnologij za

estonščino precej razvita (Krek 2012: 96–98). Kot bo razvidno iz nadaljevanja, pa je po vsebini KOST-u precej soroden latvijski korpus LAVA, saj gre za besedila tujih študentov na latvijskih univerzah, ki so v prvem ali drugem semestru učenja jezika (Darģis idr. 2020).

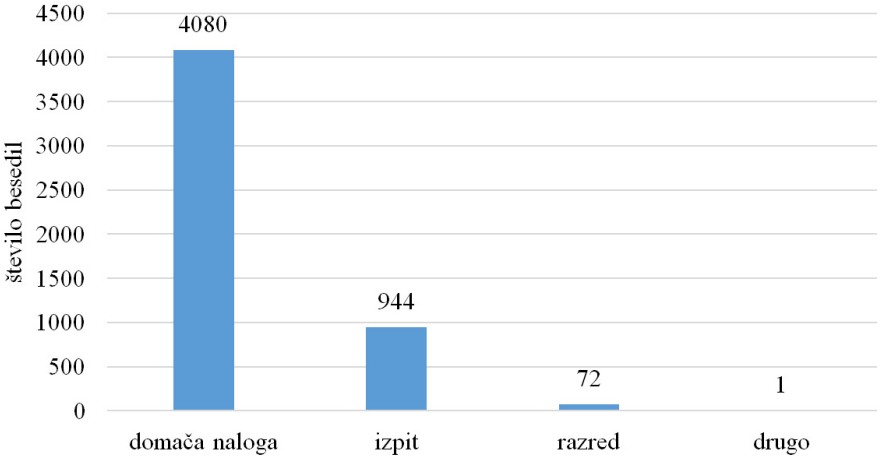
3 KOST

Pisni korpus slovenščine kot tujega jezika KOST nastaja od leta 2019 na Filozofski fakulteti Univerze v Ljubljani. Besedila pridobivamo predvsem med udeleženci lektoratov slovenščine v modulu Leto plus[4] (skoraj 87 % besedil) in med udeleženci različnih programov Centra za slovenščino kot drugi in tuji jezik (6,4 % besedil iz programa STU, 4,3 % iz drugih programov). Pri zbiranju je do sedaj sodelovalo 27 različnih učiteljev slovenščine in nekaj nepedagoških sodelavcev. Postopek zbiranja besedil je bil že opisan (prim. Stritar Kučuk 2020), zato se bom osredotočila na sestavo KOST-a po treh letih gradnje. Kot lahko vidimo iz prejšnjega poglavja, je nekakšna zlata mera za obstoječe korpuse usvajanja jezikov, ki so v približno primerljivem sociolingvističnem položaju kot slovenščina, pisni korpus z milijonom besed, različnimi prvimi jeziki tvorcev ter dodanimi oblikoskladenjskimi oznakami in oznakami napak. Oglejmo si, kako smo se temu približali s korpusom KOST.

3.1 Vključena besedila

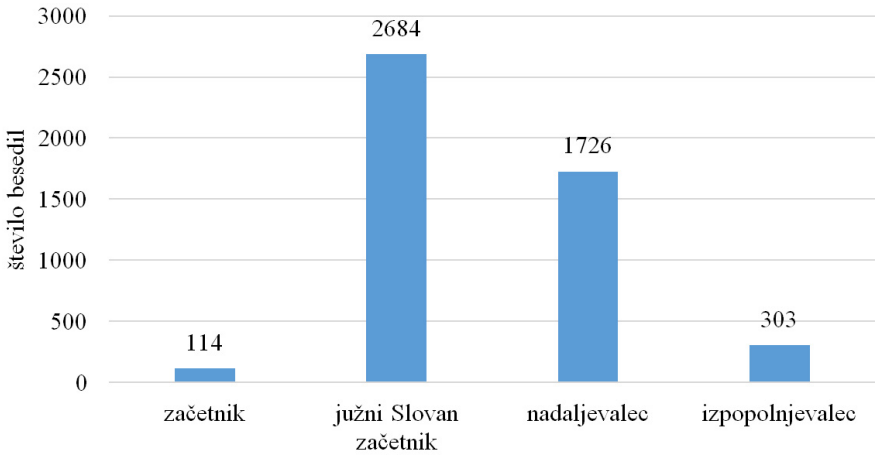
V času oddaje prispevka je v KOST-u 834.977 besed, torej 5100 besedil s povprečno dolžino okoli 164 besed. Skoraj 87 % jih je že nastalo v digitalni obliki, preostala pa je bilo treba pretipkati. Slika 2 prikazuje okoliščine njihovega nastanka. Največ je t. i. domačih nalog, torej besedil, ki so jih tvorci napisali doma, brez nadzora učitelja. Sledijo besedila z izpitov. Pri tem gre večinoma za interne izpite na tečajih ali lektoratih, pomembno pa je vedeti, da so ta besedila nastala v kontroliranih okoliščinah, kar se tiče časovnih omejitev pri pisanju in uporabe zunanjih pripomočkov, ki običajno ni dovoljena. Nekaj besedil je bilo napisanih v razredu, torej v okviru različnih dejavnosti med poukom.

4 Gre za lektorate slovenščine kot tujega jezika, namenjene redno vpisanim mednarodnim študentom v prvem letu njihovega študija na Univerzi v Ljubljani.



Slika 2: Okoliščine nastanka besedil, vključenih v KOST.

Besedila so razvrščena po štirih stopnjah (Slika 3), ki približno odslikavajo trenutno jezikovno zmožnost njihovih tvorcev.[5] Največ je besedil začetnikov, to je govorcev katerega od osrednjejužnoslovanskih jezikov (bosanščine, hrvaščine, črnogorščine, srbščine) ali makedonščine, ki so se slovensko šele začeli učiti pred največ dvema semestroma. Kot nadaljevalci so običajno označeni tisti, ki so se slovensko že učili pred udeležbo v programu, v okviru katerega je nastalo v korpus vključeno besedilo. Med njimi so lahko velike razlike (npr. med slovanskimi in neslovanskimi nadaljevalci). Manj je besedil izpopolnjevalcev, ki so po navadi daljša, kompleksnejša in z manj napakami. Najmanj pa je besedil začetnikov, torej govorcev slovenščini nesorodnih jezikov v začetnih fazah učenja. Njihova besedila so tudi relativno najkrajša.

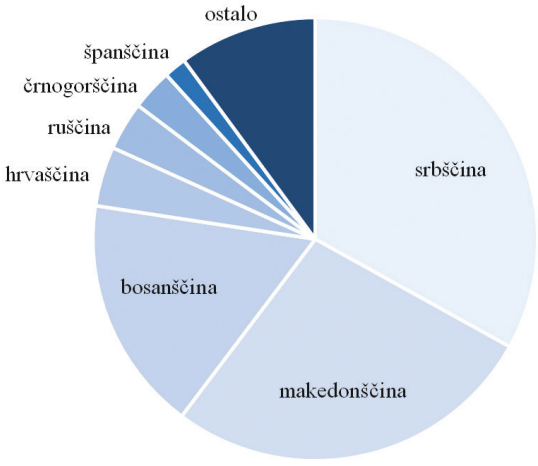


Slika 3: Okvirna stopnja, na kateri so napisana besedila, vključena v KOST.

5 Zavedati se je treba, da ta zmožnost nikakor ni zanesljivo določena, temveč gre za pragmatično oceno, ki jo največkrat poda tvorčev trenutni učitelj, in je torej namenjena le okvirni orientaciji med besedili.

3.2 Tvorci besedil

V KOST so vključena besedila 752 tvorcev, od tega je slabih 35 % moških in 65 % žensk. Govorijo 27 različnih prvih jezikov, najpogostejši med njimi so prikazani na Sliki 4.[6] V skladu s populacijo na modulu Leto plus (prim. Stritar Kučuk 2020: 132) dobre tri četrtine vseh tvorcev predstavljajo govorci osrednjih južnoslovanskih jezikov in makedonščine. Med preostalimi jeziki, ki so v KOST-u zastopani z manj kot desetimi tvorci, so: albanščina, angleščina, francoščina, grščina, hebrejščina, italijanščina, japonščina, kirgiščina, kitajščina, korejščina, madžarščina, nemščina, nizozemščina, poljščina, romunščina, slovaščina, slovenščina,[7] srbohrvaščina in ukrajinščina.



Slika 4: Prvi jeziki tvorcev, katerih besedila so vključena v KOST.

Večina pregledanih obstoječih korpusov besedila vključuje oportunistično in se ne obremenjuje z uravnoteženostjo podkorpusov glede na izhodiščni jezik ali različne stopnje jezikovne zmožnosti tvorcev. V KOST-u za zdaj izrazito izstopajo trije podkorpusi, ki skupaj predstavljajo več kot tri četrtine celote: za srbščino, bosanščino in makedonščino. Ali med prvima dvema sploh prihaja do jezikovnih razlik in ali je razlikovanje med njima za potrebe jezikovne analize sploh potrebno, bodo pokazale nadaljnje analize. Seveda pa si želimo še povečati deleže ostalih prvih jezikov, predvsem tistih, katerih govorci se bolj množično učijo slovenščino kot tuji ali drugi jezik. Zaradi tega smo že okrepili sodelovanje z lektorji slovenščine kot tujega jezika na univerzah v tistih državah, v katerih se slovenščino bolj množično učijo bodisi kot potomci izseljencev (npr. Argentina, ZDA) bodisi zaradi drugih razlogov, kot je sorodnost jezikov (npr. Poljska, Češka, Slovaška). Žal smo pri količini besedil, ki jih dobivamo za KOST, odvisni predvsem od zunajjezikovnih dejavnikov, na katere največkrat ne moremo vplivati.

6 Podatki o prvem jeziku so navedeni, kot so jih napisali sami tvorci.
7 Gre za tvorce iz slovenskega zamejstva.

Za vsakega tvorca so v KOST-u zabeleženi osnovni metapodatki: spol in leto rojstva, zaposlitveni status (študent, zaposlen ipd.), podatki o trenutnem šolanju (ime in vrsta šole, letnik ipd.), zadnja zaključena stopnja izobrazbe, prvi jezik in ostali jeziki, ki jih po samooceni zna. Dodani so tudi podatki o predhodnem učenju slovenščine in morebitnem življenju v slovenskem okolju.

Tvorci so v korpusu anonimni. Njihova imena so nadomeščena s kodami; koda »L-hr-m-0006« denimo pomeni, da gre za tvorca moškega spola s prvim jezikom hrvaščino. Vsi osebni podatki v besedilih so nadomeščeni s kodami v oglatih oklepajih, npr. osebna imena so nadomeščena s kodo [XImeX], krajevna pa z [XKrajX]. Tako sicer izgubimo nekaj dragocenih jezikovnih informacij, denimo o pregibanju imen, vendar s tem zadostimo zahtevam po varovanju osebnih podatkov. Kjer so v besedilih lastna imena ohranjena, gre večinoma za pisanje o fiktivnih osebah in krajih.

3.3 Označevanje napak v KOST-u

Eden od osrednjih namenov KOST-a je ponuditi vpogled v slovenščino tujejezičnih govorcev. Kar se tiče korpusnega jezikoslovja, je precej enotno sprejeta pot do tega označevanje jezikovnih napak (prim. Granger 2003: 466). To pomeni, da v besedilih tujih govorcev poiščemo in označimo vse napake ter jim pripišemo popravljene oblike. S tem dobimo hiter dostop do pregledne statistike najpogostejših napak, v končni obliki KOST-a pa bo uporabnik lahko iskal po izvornih oz. napačnih in popravljenih oblikah. Vsako tako označeno besedilo bo tudi videl v obeh oblikah.

Označevanje napak poteka ročno. Trenutno so napake označene na 10 % vseh besedil, kar je ustaljen delež tudi v drugih korpusih, denimo v češkem CzeSL (Rosen 2017), medtem ko ima hrvaški CroLTeC označenih 24 % besedil (Mikelić Preradović 2020: 902–904). Vseeno je naš cilj, da bi bili v označenem korpusu čim bolj uravnoteženi deleži med različnimi prvimi jeziki tvorcev, torej bomo število označenih besedil še povečali.

Več o programu za označevanje napak in samem procesu označevanja lahko preberete v prispevku Arhar Holdt, Kosem, Stritar Kučuk v tem zborniku. Tu pa si na kratko oglejmo še kategorije napak. Te temeljijo na predhodni klasifikaciji, ki je bil preizkušena za poskusni korpus slovenščine kot tujega jezika (PiKUST, prim. Stritar 2012: 154–155) in prilagojena za prvo verzijo korpusa usvajanja slovenščine kot prvega jezika Šolar (prim. Kosem idr. 2012: 34). Za KOST (Tabela 3) je bila klasifikacija prilagojena tudi zahtevam označevalnega orodja Svala.[8]

8 Prim. https://svala.cjvt.si/.

Krovna kategorija	Kategorija napake	Primer iz KOST-a	
		Napačna oblika	Popravljena oblika
Napake zapisa	ločilo	*V zimskem času**,** večina ljudi uporablja*	*V zimskem času večina ljudi uporablja*
	črkovanje	*ne i**s**pušča*	*ne i**z**pušča*
	skupaj/narazen	***ni sem** poskusil*	***nisem** poskusil*
	mala/velika začetnica	*energija **I**n paneli*	*energija **i**n paneli*
	krajšave	***in dr.***	***idr.***
Napake besedišča	samostalnik	*izobraževanje sem skupaj s celotnim **društvom** nadaljevala*	*izobraževanje sem skupaj s celotno **družbo** nadaljevala*
	glagol	*vozila, ki **hodijo** na gas*	*vozila, ki **vozijo** na plin*
	zaimek	*hodim na zmenke s **mojim** fantom*	*hodim na zmenke s **svojim** fantom*
	pridevnik	***vetrene** elektrarne*	***vetrne** elektrarne*
	prislov	*grem **doma***	*grem **domov***
	predlog	***sa** enom prijateljicom*	***z** eno prijateljico*
	veznik	*Moje najlepše potovanje je bilo **kdaj** sem bil v Kitajski.*	*Moje najlepše potovanje je bilo, **ko** sem bil na Kitajskem.*
	ostalo	***petindvajest***	***petindvajset***
Napake oblike	samostalnik	*v Sloveni**ju***	*v Sloveni**jo***
	glagol	*tam živi**mo** sestra in jaz*	*tam živi**va** sestra in jaz*
	zaimek	*ker sem **ih** spoznal*	*ker sem **jih** spoznal*
	pridevnik	*Pohištvo je v zelo **dobremu** stanju*	*Pohištvo je v zelo **dobrem** stanju*
	prislov	*Želim se naučiti kar **hitrije** slovenščino*	*Želim se naučiti čim **hitreje** slovenščino*
	ostalo	*štir**je** predavanja*	*štir**i** predavanja*
Napake skladnje	struktura	*rada bi **da živim***	*rada bi **živela***
	besedni red	*Zdi **mi se** strogo*	*Zdi **se mi** strogo*
	izpuščeni jezikovni elementi	*Na fakulteto grem avtobusom*	*Na fakulteto grem **z** avtobusom*
	odvečni jezikovni elementi	*upam **se** da bom uspel*	*upam, da bom uspel*

Tabela 3: Klasifikacija napak z ilustrativnimi primeri iz korpusa KOST.

V tabeli ni prikazana kategorija povezanih popravkov, ki jo lahko dodamo vsem ostalim oznakam. Gre za oblike v besedilu, ki jih je treba popraviti, potem ko popravimo nekaj drugega v sobesedilu (prim. Stritar 2012: 164–165). Tipičen primer je napaka besedišča pri samostalniku v zgornji tabeli, kjer moramo ob popravku nepravilnega samostalnika popraviti tudi obliko pridevnika. To je označeno kot napaka oblike

pridevnika in hkrati kot povezan popravek. Naknadna analiza napak v KOST-u pa bo pokazala, ali kategorijo povezanih popravkov sploh potrebujemo.

Natančnejša navodila za označevanje napak, odločanje v primeru dvoumnosti ipd. so preobširna za omejitve tega prispevka in so na voljo v stalno dopolnjujočem se priročniku (prim. Stritar Kučuk 2022). Z označevanjem dodatnega gradiva se namreč pojavljajo tudi nove dileme, ki jih razrešujemo sproti. Načeloma pa se pri označevanju držimo pravila, da s popravki čim manj posegamo v besedilo in popravljamo čim manj napak. Izogibamo se stilističnim popravkom in popravljamo predvsem zapis, besedišče in obliko besed, v skladnjo pa skušamo posegati čim manj (prim. Granger idr. 2022: 3). V redkih primerih, kadar napačni obliki ne znamo pripisati popravljene, to označimo s [???]. Pomembno pa je, da se uporabniki KOST-a zavedajo, da so oznake napak subjektivne. Zato bo kakršnakoli poglobljena analiza napak vedno zahtevala tudi temeljit ročni pregled zadetkov.

3.4 Dostop do KOST-a

KOST je za zdaj dostopen v aplikaciji Sketch Engine[9] v lematizirani obliki, brez dostopa do metapodatkov ali oznak napak.

4 Sklep

Korpus KOST je po prvih treh letih nastajanja suveren predstavnik korpusov usvajanja tujega jezika. Po zasnovi, vsebini, velikosti, označenosti, dostopnosti in uporabnosti se lahko primerja z obstoječimi korpusi za druge jezike s sorodnim sociolingvističnim položajem, npr. češčino, hrvaščino ali latvijščino. Nadaljnje napore bomo morali usmeriti še v povečanje podkorpusov za nejužnoslovanske prve jezike, v skladu s tem pa bomo morali bolj uravnotežiti označenost napak med različnimi podkorpusi. Tako pripravljen korpus bo lahko jezikovni vir za raznovrstne teoretične in bolj v prakso usmerjene jezikovne raziskave, npr. o najpogostejših jezikovnih napakah pri govorcih različnih prvih jezikov ali o vplivu jezikovnega prenosa na usvajanje slovenščine. Že v neuravnoteženi obliki, v kakršni je trenutno, pa je uporaben vir za različne, bolj specifično usmerjene raziskave. Med drugim je bilo gradivo iz KOST-a uporabljeno v učbeniku za slovenščino kot drugi jezik za južnoslovanske govorce, v katerem je poudarjen kontrastivni vidik poučevanja (prim. Stritar Kučuk, Šter 2021).

9 Prim. https://app.sketchengine.eu/ (dostop 21. 4. 2022).

Literatura

BOYD, Adriane, HANA, Jirka, NICOLAS, Lionel, MEURERS, Detmar, WISNIEWSKI, Katrin, ABEL, Andrea, SCHÖNE, Karin, ŠTINDLOVÁ, Barbora, VETTORI, Chiara, 2014: The MERLIN Corpus: learner language and the CEFR. *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*. Reykjavik: European Language Resources Association (ELRA). 1281–1288.

Centre for English Corpus Linguistics, 2022: *Learner Corpora around the World*. Louvain-la-Neuve: Université catholique de Louvain. https://uclouvain.be/en/research-institutes/ilc/cecl/learner-corpora-around-the-world.html

DARĢIS, Roberts, AUZIŅA, Ilze, LEVĀNE-PETROVA, Kristīne, KAIJA, Inga, 2020: Quality Focused Approach to a Learner Corpus Development. Nicoletta Calzolari idr. (ur.): *Proceedings of the 12th International Conference on Language Resources and Evaluation (LREC)*. Marseille: European Language Resources Association. 392–396. http://lava.korpuss.lv/publicatoins/LREC2020-Dargis.pdf (dostop 20. 4. 2022)

GAJDOŠOVÁ, Katarína, 2021: Čo možno nájsť v pilotnej verzii akvizičného korpusu errkorp? Katarína Gajdošová, Natália Kolenčíková (ur.): *Varia XXX: Zborník abstraktov z XXX. kolokvia mladých jazykovedcov (3.-5.11.2021)*. Bratislava: Slovenská jazykovedná spoločnosť pri Jazykovednom ústave Ľudovíta Štúra SAV. 10.

GRANGER, Sylviane, 2003: Error-tagged learner corpora and CALL: A promising synergy. *CALICO Journal* XX/3. 465–479.

GRANGER, Sylviane, SWALLOW, Helen, THEWISSEN, Jennifer, 2022: *The Louvain Error Tagging Manual: Version 2.0*. Louvain: Centre for English Corpus Linguistics, Université catholique de Louvain. https://cdn.uclouvain.be/groups/cms-editors-cecl/Granger%20et%20al._Error%20tagging%20manual_v2.0_2022.pdf

HAMMARBERG, Björn, 2010: *Introduction to the ASU Corpus: A longitudinal oral and written text corpus of adult learner Swedish with a corresponding part from native Swedes*. Stockholm: Stockholm University. https://www.diva-portal.org/smash/get/diva2:778204/FULLTEXT01.pdf (dostop 13. 4. 2022)

KOSEM, Iztok, STRITAR, Mojca, MOŽE, Sara, ZWITTER VITEZ, Sara, ARHAR HOLDT, Špela, ROZMAN, Tadeja, 2012: *Analiza jezikovnih težav učencev: Korpusni pristop*. Ljubljana: Trojina, zavod za uporabno slovenistiko.

KREK, Simon, 2012: Od SSKJ do spletnega portala standardne slovenščine. *Jezik in slovstvo* LIV/3–4. 95–113.

MIKELIĆ PRERADOVIĆ, Nives, 2020: Označavanje pogrešaka u CroLTeC-u (računalnom učeničkom korpusu hrvatskog kao stranog jezika). *Rasprave Instituta za hrvatski jezik i jezikoslovlje* XLVI/2. 899–920.

MIKELIĆ PRERADOVIĆ, Nives, BERAĆ, Monika, BORAS, Damir, 2015: Learner Corpus of Croatian as a Second and Foreign Language. Kristina Cergol Kovačević, Sanda Lucija Udier (ur.): *Multidisciplinary Approaches to Multilingualism*. Frankfurt am Main: Peter Lang. 107–126.

RAKHILINA, Ekaterina, VYRENKOVA, Anastasia, MUSTAKIMOVA, Elmira, LADYGINA, Alina, SMIRNOV, Ivan, 2016: Building a learner corpus for Russian. Elena Volodina, Gintarė Grigonytė, Ildikó Pilán, Kristina Nilsson Björkenstam, Lars Borin (ur.): *Proceedings of the joint workshop on NLP for Computer Assisted Language Learning and NLP for Language Acquisition*. Umeå: SLTC.

ROSEN, Alexandr, 2017: Introducing a corpus of non-native Czech with automatic annotation. Piotr Pęzik, Jacek Tadeusz Waliński (ur.): *Language, Corpora and Cognition*. Frankfurt am Main, Bern, Bruxelles, New York, Oxford, Warszawa, Wien: Peter Lang.

STRITAR, 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*. Ljubljana: Inštitut za novejšo zgodovino. 131–135.

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 (dostop 21. 4. 2022)

STRITAR KUČUK, Mojca, ŠTER, Helena, 2021: *Slovenščina 1+: Slovnične tabele in vaje za južnoslovanske govorce slovenščine kot drugega jezika*. Ljubljana: Znanstvena založba Filozofske fakultete.

ŠTINDLOVÁ, Barbora, ŠKODOVÁ, Svatava, ROSEN, Alexandr, HANA, Jirka, 2013: A learner corpus of Czech: Current state and future directions. *Proceedings of The Learner Corpus Research 2011*. Louvain-La-Neuve: Presses universitaires de Louvain.

ZASINA, Jan Adrian, KACZMARSKA, Elżbieta, 2020: Infrastructure of the Polish Learner Corpus PoLKo. *TaLC 2020*. https://www.researchgate.net/publication/342888260_Infrastructure_of_the_Polish_Learner_Corpus_PoLKo