Prvi korpus slovenščine kot tujega jezika KOST 1.0

Mojca STRITAR KUČUK
Univerza v Ljubljani, Filozofska fakulteta

Povzetek

Prispevek predstavlja prvi korpus slovenščine kot drugega oz. tujega jezika KOST 1.0. Gre za približno milijonski pisni korpus besedil neprvih govorcev slovenščine, ki se slovensko učijo v različnih programih Univerze v Ljubljani. Vključena besedila so v glavnem različni spisi oz. eseji, ki so bili večinoma napisani kot domača naloga, manjši del besedil pa je nastal v izpitnih okoliščinah, torej pod strožjim nadzorom. Tvorci besedil, ki so v korpusu anonimni, so večinoma naravni govorci katerega od južnoslovanskih jezikov. Posebnost korpusov usvajanja jezika so oznake jezikovnih napak. V KOST-u so te razvrščene v 23 kategorij v skladu z vnaprej določeno taksonomijo. Oznake napak in popravkov so bile v korpusna besedila dodane ročno v posebej za to razviti aplikaciji Svala. KOST 1.0 je dostopen kot baza v repozitoriju Clarin, pa tudi v konkordančnikih NoSketchEngine in KonText, podatki iz njega pa so bili že uporabljeni pri pripravi sodobnih učnih gradiv za slovenščino kot drugi jezik.

Ključne besede: korpus usvajanja tujega jezika, slovenščina kot drugi jezik, zbiranje korpusnih besedil, označevanje jezikovnih napak

Abstract

This paper presents the first learner corpus of Slovene as a second or foreign language KOST 1.0, a written corpus with approximately one million tokens. The texts were written by non-native speakers of Slovene studying Slovene in various programmes at the University of Ljubljana. The texts are mainly essays written as homework, while a smaller part of the texts were written under exam conditions, i.e. under stricter supervision. The authors of the texts, anonymised in the corpus, are mostly native speakers of a South Slavic

language. A special feature of learner corpora is the language error annotation. In KOST, these errors are classified into 23 categories according to a predefined taxonomy. The error tags and the normalised version of the texts were added manually in a specially developed application Svala. KOST 1.0 is available as a database in the Clarin repository, as well as in the NoSketch-Engine and KonText concordancers. Its data have already been used in the preparation of modern teaching materials for Slovene as a second language.

Keywords: learner corpus, Slovene as a second language, collection of corpus texts, error annotation

1 Uvod

Korpusi usvajanja tujega jezika (angl. *learner corpora*) so v dobi digitalnega jezikoslovja ključen jezikovni vir za raziskovalce, učitelje in vse ostale, ki jih zanima določen jezik kot neprvi jezik. Do nedavnega za slovenščino tovrstnih korpusov usvajanja ni bilo razen nekaj manjših poskusov bolj pilotne narave (prim. poskusni korpus PiKUST, Stritar, 2012). V okviru projekta Razvoj slovenščine v digitalnem okolju pa je bil v začetku leta 2023 objavljen prvi korpus slovenščine kot tujega jezika, KOST 1.0. Gre za digitalno zbirko pisnih besedil odraslih govorcev, za katere slovenščina ni prvi jezik. Ime KOST (= korpus slovenščine kot tujega jezika) ni popolnoma terminološko ustrezno, saj je za tvorce večjega dela vključenih besedil slovenščina drugi in ne tuji jezik (Pirih Svetina, 2005), vendar je bilo izbrano zaradi večje ekonomičnosti in lažje zapomnljivosti. V tem prispevku bodo predstavljeni zasnova korpusa in osnovni podatki o njem, opisala pa bom tudi potek označevanja napak, ki so jih pri pisanju korpusnih besedil naredili njihovi tvorci. Prav to je namreč tisti element, po katerem se korpusi usvajanja najbolj ločijo od splošnih korpusnih virov, zato mu je bilo med procesom gradnje korpusa posvečene veliko pozornosti.

2 KOST 1.0

V zadnjem desetletju so korpusi usvajanja tujega jezika doživeli razmah. Njihovo število je glede na seznam obstoječih korpusov

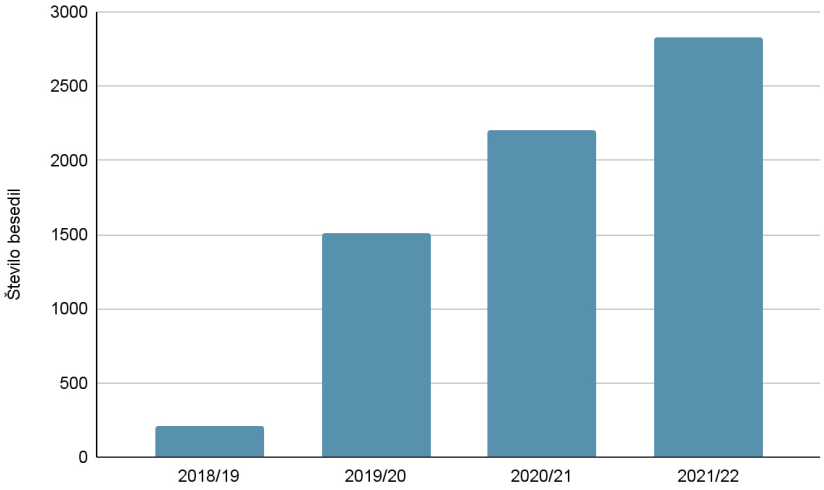
(Centre for English Corpus Linguistics, 2023) poskočilo s 73 korpusov leta 2012 na 191 korpusov leta 2022 (Stritar Kučuk, 2022). Največ, 121, je bilo pisnih korpusov, za katere je tudi najlažje pridobivati besedila, sledili so jim govorni korpusi (44), 24 pa je bilo pisnih in govornih korpusov. Večina teh korpusov ima en ciljni jezik, torej jezik, ki se ga »nekdo uči z namenom, da bi ga obvladal bodisi kot svoj prvi, drugi ali tuji jezik« (Pirih Svetina, 2005). Dobra desetina pa vključuje več ciljnih jezikov. Angleščina je ciljni jezik v dobri polovici korpusov, med ostalimi jeziki pa so še 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. Vsi ti korpusi za slovenske razmere seveda niso relevantni. Za nas je bolj zanimiv vpogled v zasnovo korpusov slovanskih jezikov, npr. hrvaškega CroLTeC (Mikelić Preradović, 2020), češkega CzeSL (Rosen, 2017), ruskega RLC (Rakhilina idr., 2016), korpuse skandinavskih jezikov, npr. švedskega SweLL (Volodina idr., 2019), in korpuse baltskih jezikov, npr. latvijskega LAVA (Darģis idr., 2020). Groba analiza teh korpusov pokaže, da je 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 (Stritar Kučuk, 2022). Kot bo razvidno iz nadaljevanja, KOST 1.0 tem standardom ustreza tako po velikosti kot po tipu besedil in raznovrstnosti njihovih tvorcev.

2.1 Besedila

KOST 1.0 obsega 6311 besedil oz. 1.032.012 besed. Zbiranje besedil se je začelo v okviru modula Leto plus,[1] ki ga Univerza v Ljubljani izvaja kot enega od ukrepov internacionalizacije. Ta modul tujim študentom, redno vpisanim v študijske programe Univerze v Ljubljani, omogoča brezplačno učenje slovenščine. Tako imamo torej dostop

1 https://www.uni-lj.si/studij/leto-plus/

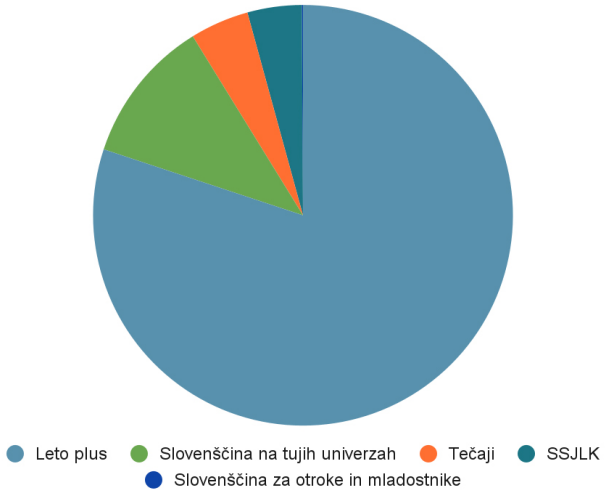
do večjega števila govorcev slovenščine kot drugega jezika in njihovih besedil, ki jih pišejo kot domače naloge ipd. na lektoratih. Zbiranje teh besedil za KOST se je pričelo v študijskem letu 2018/19, kot prikazuje Grafikon 1, pa je bilo nato vsako leto zbranih več besedil.[2]



Grafikon 1: Količina zbranih besedil po študijskih letih.

Zbiranje besedil se je iz modula Leto plus, iz katerega je bilo do sedaj pridobljenih več kot 75 % vseh besedil, razširilo še na različne programe Centra za slovenščino kot drugi in tuji jezik (Grafikon 2): lektorate slovenščine v okviru programa Slovenščina na tujih univerzah,[3] tečaje slovenščine za odrasle[4] in otroke oz. mladostnike[5] ter Seminar slovenskega jezika, literature in kulture.[6] Pri celotnem pridobivanju besedil je sodelovalo več kot 24 učiteljev, lektorjev in drugih sodelavcev teh programov.

2 Prikazani so samo podatki do vključno študijskega leta 2021/22, saj je bil KOST 1.0 zaključen s temi besedili. Zbiranje besedil intenzivno poteka tudi v nadaljnjih študijskih letih.
3 https://centerslo.si/na-tujih-univerzah/
4 https://centerslo.si/tecaji-za-odrasle/
5 https://centerslo.si/za-otroke/
6 https://centerslo.si/seminar-sjlk/



Grafikon 2: Deleži vključenih besedil glede na program, v okviru katerega so nastala.

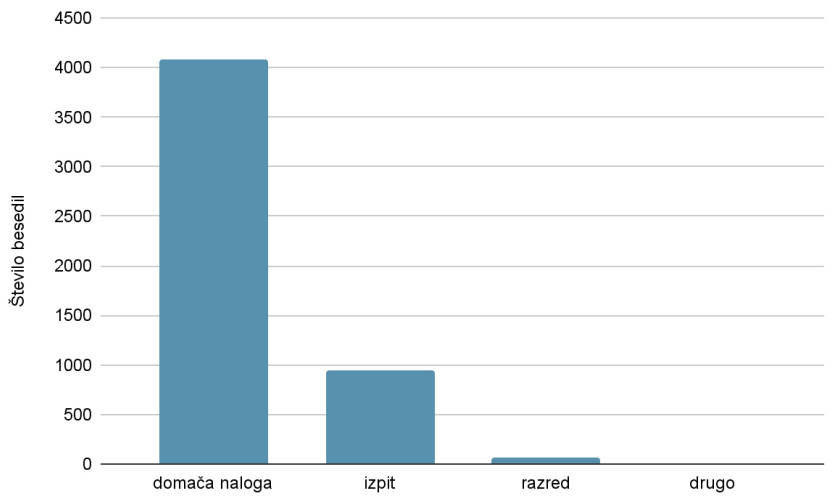
Vsako besedilo, vključeno v KOST, je poimenovano s kodo, ki izurjenemu uporabniku korpusa da nekaj osnovnih podatkov: koda L3-2122-121 denimo pomeni, da gre za besedilo, ki je nastalo pri učitelju s kodo L3 v okviru programa Leto plus v študijskem letu 2021/22, besedilo pa ima zaporedno številko 121. Poleg tega je vsako besedilo opremljeno z bogatimi metajezikovni podatki o njihovih tvorcih, okoliščinah nastanka in podobno. Zbrani so v posebni Excelovi tabeli, ki je kasneje pretvorjena v ustreznejše korpusne formate.

2.1.1 Okoliščine nastanka besedil

Velika večina vključenih besedil, skoraj 84 %, je bila napisana na računalnik. Kovidno obdobje od pomladi 2020 naprej je bilo glede dostopa do takih besedil zelo produktivno, saj se je zaradi pandemije celotno poučevanje preselilo v digitalno okolje in se je občutno povečal dotok digitalno napisanih domačih nalog. Vendar je pri tovrstnih besedilih zaradi lahkega dostopa do strojnih prevajalnikov

in drugih jezikovnih pripomočkov na spletu več dvomov glede tega, kako verodostojno odražajo jezikovno zmožnost tvorcev. S tega vidika so zanesljivejša – pa četudi do neke mere manj avtentična – besedila, ki nastajajo na izpitih ali med poukom v razredu in so napisana na roko. Ta besedila je za korpus treba pretipkati, kar so v skladu s svojimi časovnimi zmožnostmi opravili učitelji ali strokovni delavci na programih.

Besedila tvorci pišejo v različnih situacijah in o različnih temah, za KOST pa je najpomembnejše razlikovanje med okoliščinami njihovega nastanka – ali gre za pisanje s časovno omejitvijo in nadzorom učitelja glede rabe različnih jezikovnih pripomočkov ali ne (Grafikon 3). Največ je domačih nalog, ki so jih tvorci napisali doma, brez nadzora učitelja. Sledijo jim besedila z izpitov, ki so nastala v kontroliranih okoliščinah; v tem primeru gre izključno za interne izpite na tečajih ali lektoratih slovenščine. Nekaj besedil pa je bilo napisanih v razredu, v okviru različnih dejavnosti med poukom. Tudi ta besedila so večinoma bila napisana na roko, vendar z manj strogim nadzorom glede rabe pripomočkov in časovne omejitve.



Grafikon 3: Okoliščine nastanka besedil, vključenih v KOST.

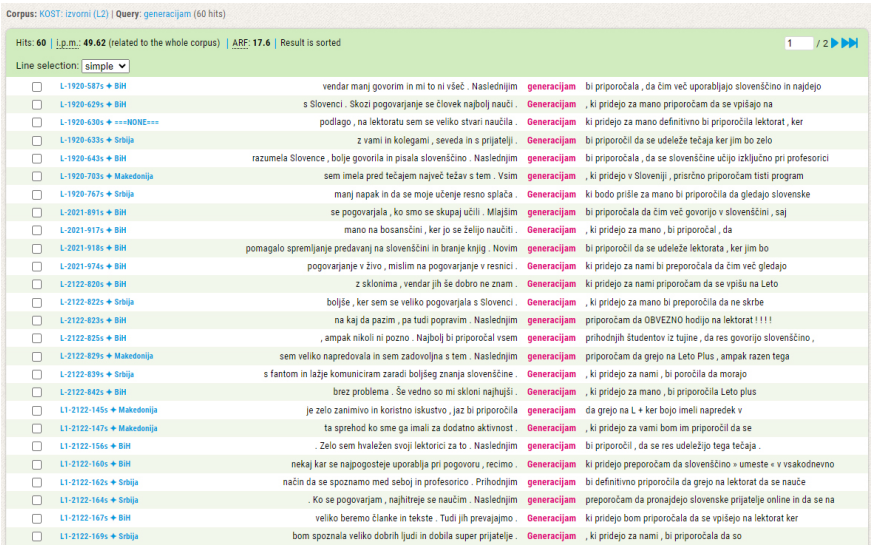
2.1.2 Vrste besedil

V KOST so vključene različne vrste besedil. Največ je esejev oz. spisov (npr. o družini, prehrani, zdravju) in poročil o različnih dejavnostih (npr. o ogledu filma, obisku muzeja, izletu po Sloveniji). Če so tvorci pred pisanjem dobili natančnejša navodila za pisanje, so ta zabeležena med metapodatki, saj tvorci nemalokrat dobesedno ponavljajo celotne fraze ali besedne zveze iz njih, s tem pa lahko navodila vplivajo na frekvenco določenih pojavnic v korpusu. Kot primer si poglejmo naslov *Moje Leto plus*, ki ga študenti Leta plus večkrat dobijo v pisnem izpitu ob zaključku drugega semestra. Spremlja ga podrobno navodilo, ki med drugim vključuje:

> V besedilu komentirajte:
> - lektorat slovenščine,
> - dodatne dejavnosti (kaj se vam je zdelo najbolj zanimivo; kaj koristnega ste dobili od vsake dejavnosti, kaj bi v zvezi s tem priporočili generacijam, ki pridejo za vami),
> - svoje učenje slovenščine (ali ste zadovoljni s svojim napredkom, kaj vam je najbolj pomagalo pri učenju, kaj bi priporočili generacijam, ki pridejo za vami), [podčrtala M. S. K.]
> - svoj študij (kaj ste pričakovali pred prihodom, kako ste zadovoljni),
> - svoje življenje v Ljubljani (kakšne težave ste imeli, kako se počutite kot študent).

V KOST 1.0 je vključenih 229 besedil s tem naslovom, zato ni presenetljivo, da najdemo 60 konkordanc za iskanje *generacijam*, ki vključujejo različne izpeljave zgoraj podčrtane fraze (Slika 1). Brez tega navodila je manj verjetno, da bi se ta fraza tako pogosto pojavljala v slovenščini kot neprvem jeziku.

Načeloma se v KOST-u izogibamo praktičnim besedilom, kakršna sta življenjepis ali prošnja za delo, saj vključujejo veliko osebnih podatkov, ki jih s pravnega vidika ne smemo prikazovati in jih moramo zakriti s kodami, kar pa precej zmanjša berljivost. Vprašljiv je tudi jezikovni vidik teh besedil, saj gre v njih pretežno za ponavljanje ustaljenih sporazumevalnih vzorcev, manj pa je dejansko



Slika 1: Konkordance za iskanje *generacijam* v korpusu KOST 1.0.

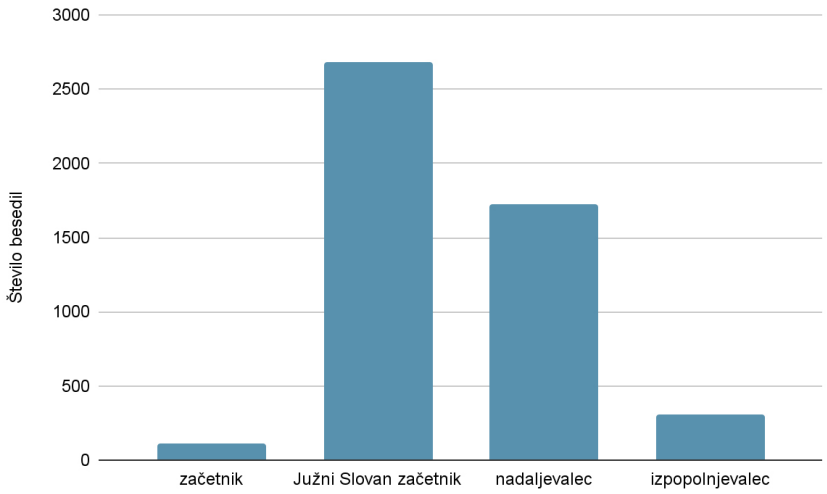
samostojne uporabe jezika. Od praktičnih besedil je zato v KOST vključenih še največ različnih e-pisem, ki so sicer napisana po navodilih, a gre vendarle za več samostojnega pisanja. Takšno navodilo je lahko na primer:

> Napišite e-pošto profesorju ali profesorici. Napišite mu/ji 2–3 vprašanji v zvezi s predavanji, izpiti, gradivom … Vprašanja naj bodo povezana, besedilo naj bo logično. Besedilo ustrezno začnite in zaključite.

2.1.3 Stopnja jezikovne zmožnosti

Besedila, vključena v KOST, so označena s štirimi stopnjami, ki odslikavajo trenutno jezikovno zmožnost njihovih tvorcev (Grafikon 4). Ta ni zanesljivo določena po vnaprej opredeljenih lestvicah, kakršna je lestvica SEJO (Kovačič idr., 2011). Gre zgolj za pragmatično oceno, namenjeno okvirni orientaciji med besedili, ki jo največkrat poda tvorčev trenutni učitelj. Po tej lestvici je v KOST-u največ besedil Južnih Slovanov začetnikov, se pravi govorcev katerega od osrednjejužnoslovanskih

jezikov (bosanščine, črnogorščine, hrvaščine, srbščine) ali makedonščine, ki so se slovensko šele začeli učiti pred največ dvema semestroma. Njihov napredek je zaradi sorodnosti izhodiščnega in ciljnega jezika običajno hiter. Kot nadaljevalci so označeni tisti, ki so se slovensko že učili pred udeležbo v programu, v okviru katerega je nastalo v korpus vključeno besedilo, zato že tvorijo kompleksnejša besedila. Med njimi so lahko velike razlike (npr. med slovanskimi in neslovanskimi nadaljevalci). Manj je besedil izpopolnjevalcev, ki so ponavadi 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.



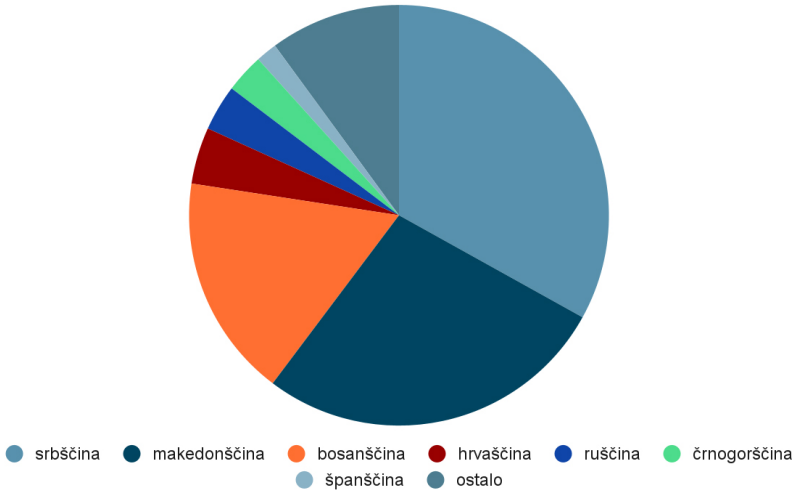
Grafikon 4: Štiri stopnje ocenjene jezikovne zmožnosti tvorcev besedil v slovenščini v KOST 1.0.

2.2 Tvorci besedil

V KOST 1.0 so vključena besedila več kot 950 tvorcev, od tega je slabih 34 % moških in 66 % žensk. V korpusu so anonimni. Njihova imena so nadomeščena s kodami; koda L-hr-m-0006 denimo pomeni, da gre za tvorca moškega spola s prvim jezikom hrvaščino, ki je dobil zaporedno številko 6.

2.2.1 Prvi jezik

Tvorci besedil, vključenih v KOST, govorijo 30 različnih prvih jezikov. Najpogostejši med njimi so prikazani na Grafikonu 5. V skladu s populacijo na modulu Leto plus (Stritar Kučuk, 2020) dobre tri četrtine vseh tvorcev predstavljajo govorci osrednjejužnoslovanskih jezikov (bosanščine, črnogorščine, hrvaščine in srbščine) in makedonščine. Nekoliko več je še govorcev ruščine in španščine. Med jeziki, ki so v KOST-u zastopani z manj tvorci, pa so albanščina, angleščina, francoščina, grščina, hebrejščina, italijanščina, japonščina, kirgiščina, kitajščina, korejščina, madžarščina, nemščina, nizozemščina, poljščina, romunščina, slovaščina, slovenščina[7] in ukrajinščina. Pri beleženju podatkov o metajeziku tvorca sledimo temu, kar je kot svoj prvi oz. materni jezik navedel sam tvorec. Zato imamo med prvimi jeziki denimo tudi srbohrvaščino.



Grafikon 5: Prvi jeziki tvorcev besedil, vključenih v KOST 1.0, glede na število tvorcev.

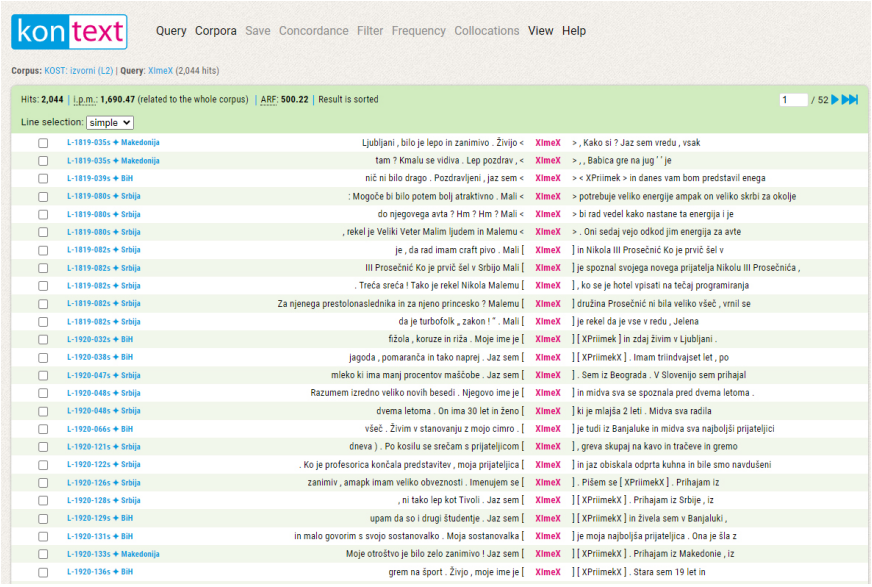
7 Gre za tvorce iz slovenskega zamejstva, pri katerih se srečujemo tudi z vprašanjem, ali naj jih sploh upoštevamo kot govorce slovenščine kot neprvega jezika. Odločitev je vsakokrat individualna.

2.2.2 Varovanje osebnih podatkov

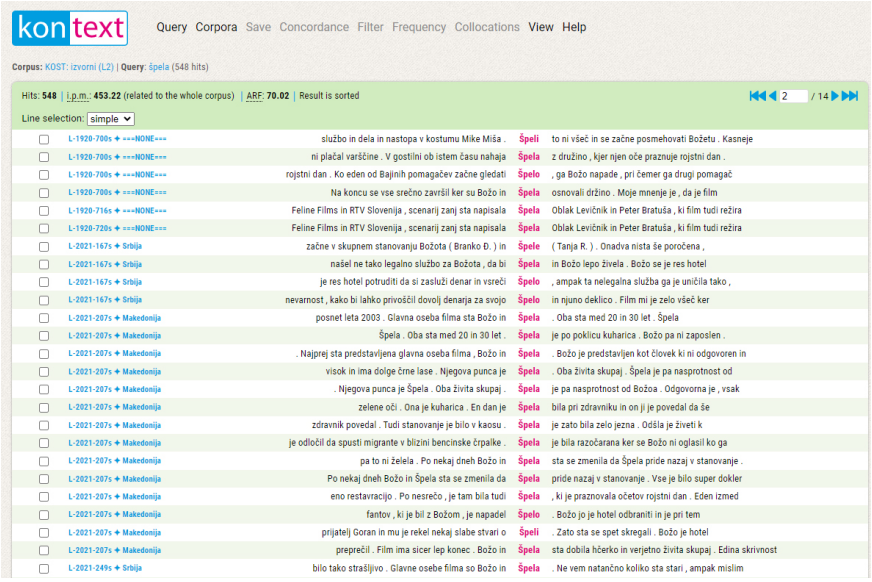
Ker sta ureditev pravic za uporabo podatkov in varovanje osebnih podatkov ključnega pomena, vsi tvorci, katerih besedila so vključena v KOST, podpišejo izjavo, s katero dovoljujejo vključitev svojih besedil. V izjavi dobimo tudi osebne podatke, ki so nujni za analizo korpusnega gradiva: spol, starost, fakulteta, letnik in stopnja študija, izobrazba, prvi jezik in ostali jeziki, ki jih znajo govorci, ter podatki o morebitnem predhodnem učenju slovenščine ali bivanju v Sloveniji. Vse to je v KOST-u zabeleženo kot metapodatek.

Izjavo, ki so jo pravno preverili na Oddelku za upravljanje s tveganji in varstvo osebnih podatkov na Univerzi v Ljubljani, sodelujočim v podpis ponudijo njihovi učitelji. Pred podpisom jim natančno razložijo o projektu in pogojih sodelovanja. Razveseljivo je, da izjavo podpiše velika večina vseh, ki jim je bila ponujena. Vse izjave so shranjene v digitalni in, če so bile podpisane na papirju, tudi tiskani obliki.

Če se v besedilih pojavijo osebni podatki, so nadomeščeni s kodami v oglatih oklepajih. Osebna imena so denimo nadomeščena s kodo [XImeX], krajevna pa z [XKrajX]. S tem zadostimo zahtevam po varovanju osebnih podatkov, a izgubimo jezikovne informacije o pregibanju teh imen, saj je koda enaka za vse sklonske oblike (Slika 2). Primanjkljaj vendarle ni prevelik, saj so lastna imena v besedilih ohranjena, kadar gre za pisanje o znanih osebnostih ali fantazijskih osebah. Na Sliki 3 so prikazane konkordance za lemo *Špela*. Pri tem gre v veliki večini primerov za enega od likov iz filma *Kajmak in marmelada*, ki je pogosta tema pisanja študentov v modulu Leto plus.



Slika 2: Prikaz zakritih osebnih imen v korpusu KOST 1.0.



Slika 3: Prikaz iskanja za lemo *Špela* v korpusu KOST 1.0.

3 Označevanje jezikovnih napak v korpusu KOST 1.0

Besedila so v KOST vključena taka, kot so jih napisali tvorci. To je samoumevno izhodišče, ki se ga držijo v vseh korpusih usvajanja. Nekateri gredo pri tem še korak dlje: v hrvaškem korpusu CroLTec označujejo naknadne popravke, ki so jih v svojih besedilih naredili tvorci, npr. ko so prečrtali del besedila ali pa ga naknadno dodali (Mikelić Preradović, 2020). Tega v KOST-u ne označujemo, ampak ohranjamo besedila v izvirnem digitalnem čistopisu. Vse jezikovne popravke, ki jih naredimo, označimo s posebnimi oznakami za jezikovne napake – z eno izjemo, ki se nanaša na nekatera pravopisna oz. tehnična vprašanja. V besedilih namreč popravimo stičnost ločil in odstranimo dvojne presledke. To je po eni strani povezano s postopkom tokenizacije korpusnih besedil v aplikaciji Svala (prim. razdelek 3.2), ki zaradi zahtev same aplikacije poteka tako, da bi se vse morebitne posebnosti pri stičnosti ločil v vsakem primeru izgubile, po drugi strani pa zapisovanje ločil niti ni v ospredju raziskav pri slovenščini kot neprvem jeziku. Čeprav imajo tvorci v KOST vključenih besedil dejansko nemalokrat težave pri zapisovanju ločil, kar naj bi bila posledica njihove navajenosti na elektronsko komunikacijo (Poteko, 2023), izguba tega podatka vendarle nima večjega vpliva na uporabnost podatkov iz KOST-a.

Najbolj se uporabnost korpusov usvajanja torej poveča, če so v njih označene jezikovne napake, ki jih pri tvorjenju v ciljnem jeziku delajo tvorci. Označene so v večini obstoječih korpusov, ki presegajo zgolj pilotske poskuse oz. manjše priložnostne raziskave. Zato smo kmalu po začetku gradnje korpusa KOST v njem začeli označevati jezikovne napake. Natančno opredeljevanje, kaj je napaka, je za namen tega prispevka nerelevantno, v grobem naj zadostuje, da so napake pojavitve v besedilu, ki so nenamerno odklonske in jih njihovi tvorci sami ne morejo popraviti (James, 1998).

V vseh obstoječih korpusih usvajanja označevanje napak poteka ročno, kar pomeni, da je relativno zamudno in počasno. Napake so potemtakem redko označene na celotnem korpusnem gradivu. V KOST-u 1.0 so označene na 10 % vseh besedil, kar je ustaljen delež tudi v drugih korpusih, denimo v češkem CzeSL (Rosen, 2017).

3.1 Orodje za označevanje napak

V okviru projekta Razvoj slovenščine v digitalnem okolju smo za ročno označevanje korpusov z označenimi jezikovnimi napakami oz. popravki razvili oz. prilagodili novo računalniško orodje. Lokalizirali smo odprto dostopni švedski program Svala (Wirén idr., 2019) in ga prilagodili, da vsebuje predpripravljene nabore kategorij oznak za korpusa KOST in Šolar (več o aplikaciji Svala je objavljeno v prispevku Arhar Holdt, Kosem, Pori v tej publikaciji). Z označevanjem gradiva za korpus KOST 1.0 smo orodje Svala[8] uspešno evalvirali.

Večino označevanja napak za KOST 1.0 sem opravila sama kot urednica korpusa. Poseben preizkus uporabnosti Svale pa je bilo delo s skupino polprofesionalnih uporabnikov, študentov 3. letnika 1. stopnje slovenistike na Filozofski fakulteti Univerze v Ljubljani, ki so besedila tujih govorcev za KOST označevali pri izbirnem predmetu Slovenščina kot drugi in kot tuji jezik v zimskem semestru študijskega leta 2021/22 in v zimskem semestru študijskega leta 2022/23. V prvem letu je sodelovalo 19, v drugem pa 20 študentov. Označili so 172 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, sem nato pregledala, študenti pa so svoje delo predstavili v okviru seminarja pri predmetu Slovenščina kot tuji jezik. S študentskega gledišča so bili rezultati pozitivni: tovrstno delo so v anonimni anketi ocenili kot zanimivo, strokovno precej, tehnično pa manj zahtevno, razmeroma zamudno, a koristno zanje in za širšo skupnost. Izrazili so zadovoljstvo z možnostjo praktičnega, tehnično nezahtevnega dela, pri katerem so morali dejansko uporabiti tudi jezikoslovno znanje, pridobljeno pri študiju. Manj zadovoljujoči so bili rezultati za sam korpus. V povprečju je bilo v besedilih študentov 35 % neustreznih oznak, ki so bile v največji meri posledica površnega dela, slabega znanja pravopisa in oblikoslovja ter pretiranega popravljanja besedil (Stritar Kučuk, 2023b).

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

3.2 Taksonomija napak

V Svali je vsako besedilo popravljeno oz. normalizirano, vsaka napaka pa dobi oznako glede na taksonomijo napak (gl. Tabelo 1). Ta temelji na klasifikaciji, ki je bila preizkušena za poskusni korpus slovenščine kot tujega jezika PiKUST (Stritar, 2012), prilagojena prvi verziji korpusa usvajanja slovenščine kot prvega jezika Šolar (Kosem idr., 2012) in prilagojena tudi zahtevam označevalnega orodja Svala (Arhar Holdt idr., 2022).

Tabela 1: Kategorije napak v korpusu KOST 1.0.

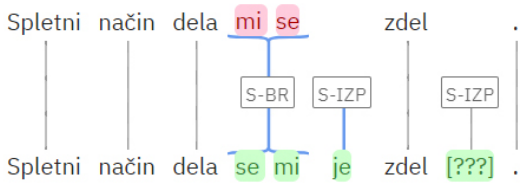
Krovna kategorija	Kategorija napake	Oznaka
Napake zapisa	Ločilo	Z-LOC
	Črkovanje	Z-CRK
	Skupaj/narazen	Z-SN
	Mala/velika začetnica	Z-MV
	Krajšave	Z-KR
Napake besedišča	Samostalnik	B-SAM
	Glagol	B-GLAG
	Pridevnik	B-PRID
	Zaimek	B-ZAIM
	Prislov	B-PRISL
	Predlog	B-PRED
	Veznik	B-VEZ
	Ostalo	B-OST
Napake oblike	Samostalnik	O-SAM
	Glagol	O-GLAG
	Pridevnik	O-PRID
	Zaimek	O-ZAIM
	Prislov	O-PRISL
	Ostalo	O-OST
Napake skladnje	Struktura	S-STR
	Besedni red	S-BR
	Izpuščeni jezikovni elementi	S-IZP
	Odvečni jezikovni elementi	S-ODV
Dodatna oznaka: Povezani popravek		POV

Orodje Svala je dovolj fleksibilno, da omogoča različne kombinacije: oznake napak se lahko nanašajo na eno besedo ali na večji del besedila, eno oznako je mogoče pripisati tudi več delom besedila, ki ne stojijo skupaj. Napačna pojavitev v korpusu lahko dobi več hkratnih oznak napak. Oznako napake pa lahko pripišemo tudi pojavitvi, ki je v normaliziranem besedilu ni mogoče navesti, kot je v primeru odvečnega dela besedila (Slika 4, primeri S-ODV).

Slika 4: Primer izvornega in popravljenega besedila iz KOST-a z označenimi različnimi tipi napak.

Natančna navodila za označevanje napak so na voljo v stalno dopolnjujočem se priročniku za označevanje napak (Stritar Kučuk, 2023a). Z označevanjem dodatnega gradiva se namreč pojavljajo nove dileme, ki jih razrešujemo sproti. V priročniku so posebej izpostavljeni primeri, ki bi jih lahko umestili v več kategorij, in primeri, ki jih označevalci napak večkrat neustrezno označijo. Načeloma pa je osnovno vodilo označevanja, da s popravki čim manj posegamo v besedilo in ravnamo po načelu minimalnega popravka (Volodina idr., 2019): besedilo spremenimo, čim manj je mogoče, in popravimo kar najmanj napak, da bo normalizirano besedilo slovnično ustrezno, razumljivo in sprejemljivo za domačega govorca slovenščine. Popravljamo predvsem zapis, besedišče in obliko besed, v skladnjo skušamo posegati čim manj, predvsem pa se izogibamo stilističnim popravkom. Uporabniki KOST-a pa se morajo zavedati, da so oznake napak do neke mere vedno subjektivne. Zato kakršna koli poglobljena analiza napak zahteva tudi temeljit ročni pregled zadetkov.

Kadar napačni obliki ne znamo pripisati popravljene oz. bi to zahtevalo preveč označevalčeve interpretacije, to označimo s [???] (Slika 5). Takih primerov je razmeroma malo, v KOST-u 1.0 84.

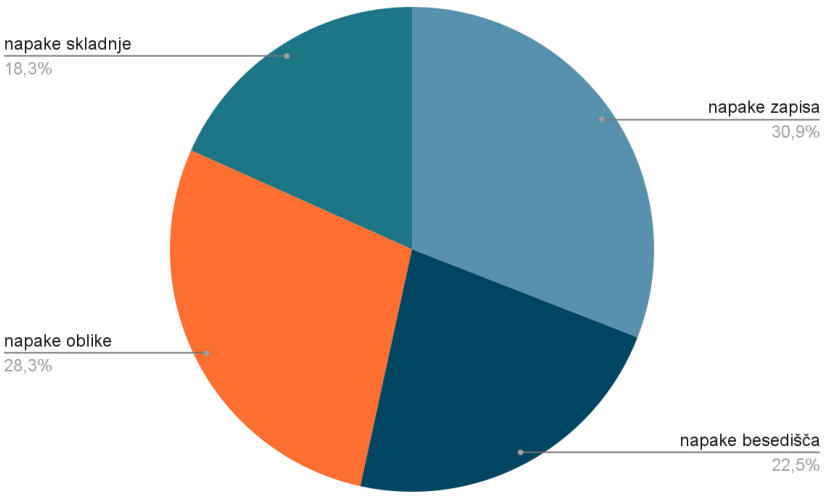


Slika 5: Primer oznake za izpuščeni del besedila, ki ga v KOST-u ne znamo ustrezno popraviti.

3.3 Napake v korpusu KOST 1.0

Čeprav je bilo v označevanje napak v korpusnih besedilih že od začetka vloženega veliko dela, pa KOST 1.0 zaradi tehničnih omejitev obstoječih konkordančnikov ni dostopen v obliki, ki bi omogočala širšo uporabnost teh oznak. Zato je tukaj vsaj osnovna statistika pogostnosti oznak po kategorijah napak.

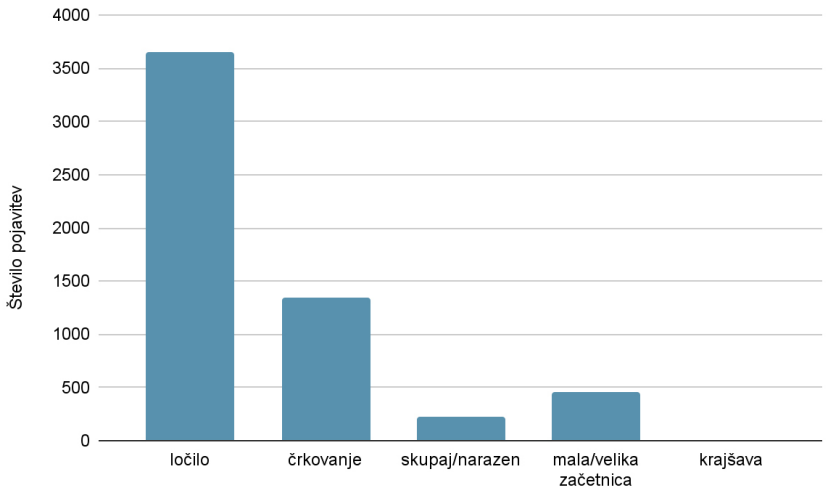
Štiri osnovne kategorije napak so med seboj približno uravnotežene (Grafikon 6). Prednjačijo napake zapisa, najmanj pa je napak



Grafikon 6: Pogostnost osnovnih tipov napak v korpusu KOST 1.0.

skladnje. Pri tem je treba upoštevati, da se napake zapisa praviloma nanašajo samo na eno besedo, napake skladnje pa na več besed, kar verjetno vpliva na to, da je njihovih pojavitev manj.

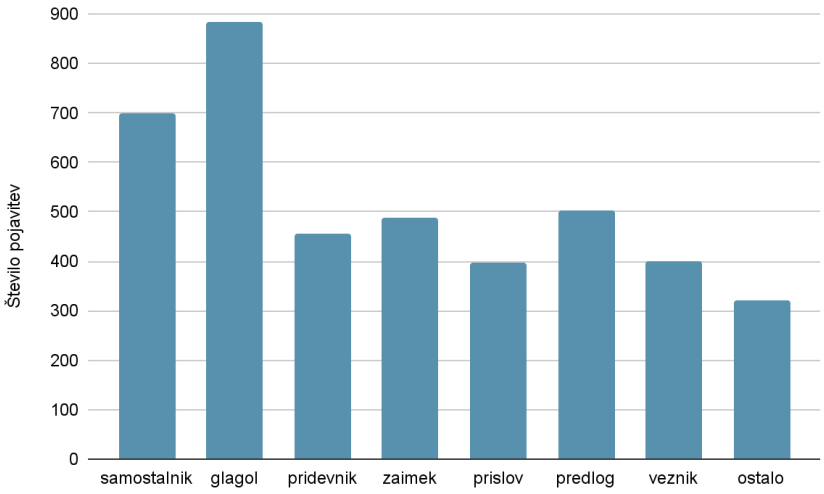
Vpogled v kategorije napak na drugi ravni pokaže, da je pri napakah zapisa (Grafikon 7) največ napak ločil (npr. *Všeč mi je ker je hiša > Všeč mi je, ker je hiša*), kar je tudi daleč najpogostejša med vsemi kategorijami napak. V veliki meri gre za postavljanje vejic. Veliko je tudi napak črkovanja oz. neustrezne pisne realizacije fonemov (npr. *v autobusu > v avtobusu*), sledita jim napačna raba male oz. velike začetnice (npr. *praznujemo Božič > praznujemo božič*) in pisanje skupaj oz. narazen (npr. *naj bolj > najbolj*), medtem ko je kategorija napak krajšav (npr. *in dr. > idr.*) skrajno redka.



Grafikon 7: Podtipi napak zapisa po pogostnosti v korpusu KOST 1.0.

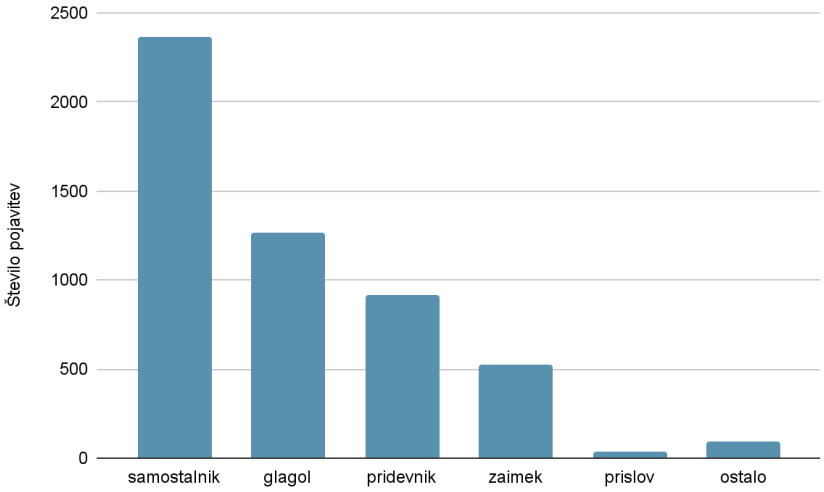
Napake besedišča (Grafikon 8), pri katerih gre za neustrezno leksikalno izbiro, so najpogostejše pri glagolih (npr. *sem se zelo težko naučila na mir > sem se zelo težko navadila na mir*). Sledijo jim samostalniki (npr. *kadiranje > kajenje*). Napake besedišča pri pridevnikih (npr. *družbena oseba > družabna oseba*), zaimkih (npr. *pri enem prijatelju > pri nekem prijatelju*), prislovih (npr. *grem doma > grem*

domov), predlogih (npr. *sa enom prijateljicom > z eno prijateljico*), veznikih (npr. *od kdaj sem > odkar sem*) in ostalih besednih vrstah (npr. *petindvajest > petindvajset*) pa so po pogostnosti bolj ali manj izenačene.



Grafikon 8: Podtipi napak besedišča po pogostnosti v korpusu KOST 1.0.

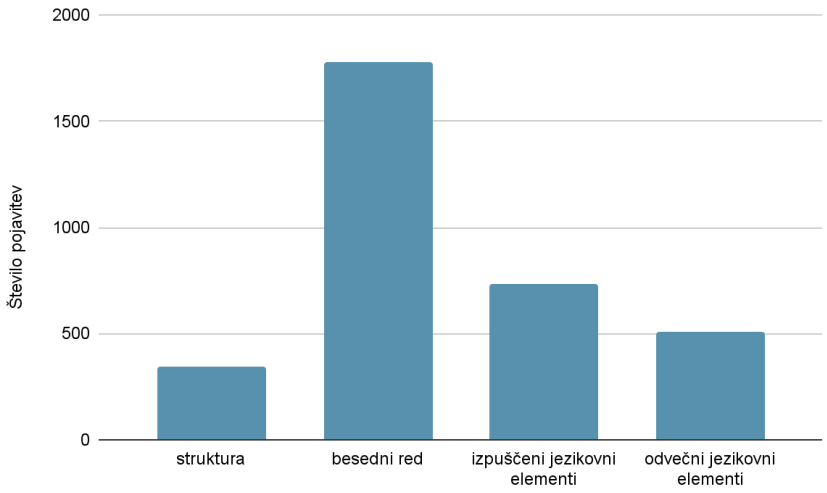
Pri napakah oblike (Grafikon 9), ki se nanašajo na pregibanje besed, je največ napak samostalnika (npr. *v Sloveniju > v Slovenijo*), kar je druga najpogostejša med vsemi kategorijami napak. Sledijo jim glagoli (npr. neuporaba dvojine pri povedku v primeru *sestra in jaz delamo > sestra in jaz delava*), pridevniki (npr. *v zelo dobremu stanju > v zelo dobrem stanju*) in zaimki (npr. *sem ih spoznal > sem jih spoznal*). Nekaj je tudi oblikoslovnih napak prislovov (npr. *hitrije > hitreje*) in številnikov (npr. *štirje predavanja > štiri predavanja*).



Grafikon 9: Podtipi napak oblike po pogostnosti v korpusu KOST 1.0.

Pri napakah skladnje (Grafikon 10) je izrazito največ napak besednega reda (npr. *zdi mi se > zdi se mi*), ki je tretja najpogostejša kategorija napak. Napak strukture je razmeroma malo (npr. *rada bi da živim > rada bi živela*), zanimivo pa je, da je izpuščenih jezikovnih elementov (npr. zaradi uporabe brezpredložnega orodnika v primeru *grem avtobusom > grem z avtobusom*) nekoliko več kot odvečnih delov besedila (npr. *upam se da bom uspel > upam, da bom uspel*).

Kategorija povezanih popravkov se v KOST-u 1.0 pojavi 747-krat. V resnici še vedno ne vemo, ali se bo ta kategorija izkazala za uporabno pri analizi ali ne. O tem se bomo lahko odločili šele, ko bomo označeno gradivo začeli zares analizirati.



Grafikon 10: Podtipi napak skladnje po pogostnosti v korpusu KOST 1.0.

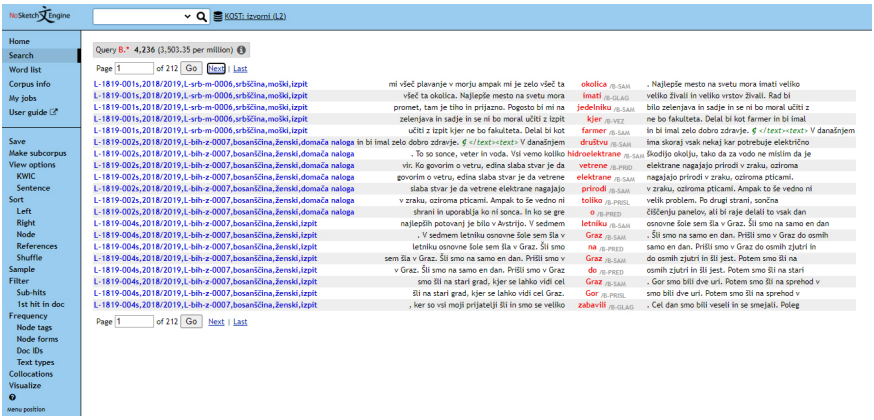
4 Dostop do korpusa KOST 1.0

Korpus KOST 1.0 je kot baza dostopen na repozitoriju Clarin.si[9] pod pogoji licence CC BY-SA 4.0. Izključno v izobraževalne in raziskovalne namene ga lahko uporabljajo učitelji, študenti, raziskovalci in drugi, ki jih zanima slovenščina kot tuji jezik. Na voljo je tudi v bolj robustnih formatih CoNLL-U in JSON ter VERT.

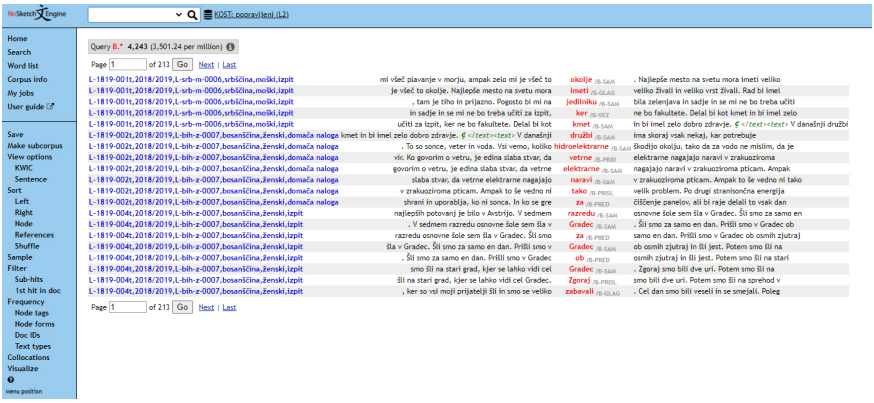
V projektu RSDO je bil razvit format korpusov z jezikovnimi popravki, ki je skladen z ostalimi slovenskimi korpusi in povezljiv s formatom orodja Svala. Tri izhodne datoteke JSON – nepopravljena in popravljena besedila ter datoteka s povezavami med vsako pojavnico iz nepopravljene in popravljene verzije datoteke, skupaj z oznakami za jezikovne popravke – so pretvorjene v XML in združene v eno datoteko XML, ki je skladna s shemo TEI.[10] KOST 1.0 je torej vključen tudi v konkordančnika NoSketchEngine[11] in KonText,[12] ki

9 http://hdl.handle.net/11356/1753
10 https://tei-c.org
11 https://www.clarin.si/noske/run.cgi/corp_info?corpname=kost10_orig&struct_attr_stats=1
12 https://www.clarin.si/kontext/query?corpname=kost10_orig

sta del infrastrukture CLARIN. Za vsak korpus sta datoteki z izvornimi in s popravljenimi besedili uvoženi ločeno (Slika 6, Slika 7). To je seveda le začasna rešitev za pregledovanje podatkov, ki pa je vendarle že omogočila prve jezikoslovne analize. Med drugim je bilo gradivo iz KOST-a uporabljeno pri pripravi učbenika za slovenščino kot drugi jezik za južnoslovanske govorce, v katerem je poudarjen kontrastivni vidik poučevanja (Stritar Kučuk in Šter, 2021, Stritar Kučuk idr., 2023).



Slika 6: Prikaz izvornega besedila v konkordančniku NoSketchEngine.



Slika 7: Prikaz popravljenega besedila v konkordančniku NoSketchEngine.

5 Pogled naprej

Kot je bilo že omenjeno, brskanje po oznakah napak v KOST-u za povprečnega uporabnika še ni mogoče, saj konkordančniki, v katere je vključen, tega ne dovoljujejo. Zato je prvi naslednji korak razvoj specializiranega konkordančnika, ki bo omogočal polno izrabo bogato označenega korpusnega gradiva, vključno z iskanjem po posameznih kategorijah napak ter možnostmi izrabe metapodatkov in sočasne vizualizacije izvornega ter popravljenega besedila. Poleg tega želimo KOST povečati, predvsem na račun nejužnoslovanskih jezikov, in vzpostaviti redno pridobivanje besedil še iz drugih virov, denimo z izpitov slovenščine v izvedbi Izpitnega centra CSDTJ.[13] Z besedili, ki jih napišejo kandidati na izpitih predvsem na vstopni in osnovni ravni, bomo dobili vpogled v slovensko pisno produkcijo nižje izobraženih govorcev, med katerimi se mnogi slovenščine načrtno ne učijo, temveč jo zgolj usvajajo iz okolja. Najpomembnejši prihodnji cilj v zvezi s korpusom KOST pa je povečati delež besedil, na katerih so označene jezikovne napake, in čim bolj uravnotežiti označene deleže med različnimi prvimi jeziki tvorcev.

S takim razvojem bo KOST postal širše uporaben jezikovni vir, zanimiv za vse, ki raziskujejo slovenščino kot drugi oz. tuji jezik. Omogočal bo prepoznavo najpogostejših jezikovnih napak, značilnih za govorce določenih prvih jezikov, in pripravo bolj osredotočenih učnih gradiv, pa tudi ustreznejše poudarke v samem pedagoškem procesu.

Zahvala

Projekt Razvoj slovenščine v digitalnem okolju, ki je podprl razvoj korpusa KOST, sta med leti 2020 in 2023 sofinancirali Republika Slovenija in Evropska unija iz Evropskega sklada za regionalni razvoj (Operacija se je izvajala v okviru Operativnega programa za izvajanje evropske kohezijske politike v obdobju 2014–2020).

13 https://centerslo.si/izpiti/

Literatura

Arhar Holdt, Š., Kosem, I., & Stritar Kučuk, M. (2022). Metode in orodja za lažjo pripravo korpusov usvajanja jezika. V Pirih Svetina, N., Ferbežar, I. (ur.), *Simpozij Obdobja 41: Na stičišču svetov: Slovenščina kot drugi in tuji jezik* (str. 23–30). Založba Univerze v Ljubljani. https://doi.org/10.4312/Obdobja.41.2784-7152

Centre for English Corpus Linguistics. (2023). *Learner Corpora around the World*. https://uclouvain.be/en/research-institutes/ilc/cecl/learner-corpora-around-the-world.html

Darģis, R,, Auziņa, I., Levāne-Petrova, K., & Kaija, I. (2020). Quality Focused Approach to a Learner Corpus Development. V Calzorali, N., idr. (ur.), *Proceedings of the 11th International Conference on Language Resources and Evaluation (LREC)* (str. 392–396). http://lava.korpuss.lv/publicatoins/LREC2020-Dargis.pdf

Granger, S. (2008). Learner corpora. V Ludeling A., Kyto, M. (ur.), *Corpus Linguistics. An International Handbook* (str. 259–275). Mouton de Gruyter.

James, C. (1998). *Errors in Language Learning and Use: Exploring Error Analysis*. Longman. https://doi.org/10.4324/9781315842912

Kosem, I., Stritar, M., Može, S., Zwitter Vitez, A., Arhar Holdt, Š., & Rozman, T. (2012). *Analiza jezikovnih težav učencev: Korpusni pristop*. Trojina, zavod za uporabno slovenistiko.

Kovačič, I., idr. (2011). *Skupni evropski jezikovni okvir: učenje, poučevanje, ocenjevanje*. Ministrstvo RS za šolstvo in šport, Urad za razvoj šolstva.

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

Pirih Svetina, N. (2005). *Slovenščina kot tuji jezik*. Izolit.

Poteko, I. (2023). Sporazumevalne navade in jezikovne izbire študentk in študentov v sms-ih in sporočilih iz mobilnih aplikacij. V Vogel, J. (ur.), *59. seminar slovenskega jezika, literature in kulture: Slovenski jezik, literatura, kultura in digitalni svet(ovi)* (str. 105–114). Založba Univerze v Ljubljani.

Rakhilina, E., idr. (2016). Building a learner corpus for Russian. *Proceedings of the joint workshop on NLP for Computer Assisted Language Learning and NLP for Language Acquisition*. SLTC.

Rosen, A. (2017). Introducing a corpus of non-native Czech with automatic annotation. *Language, Corpora and Cognition*. Peter Lang.

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

Stritar Kučuk, M. (2020). Modul Leto plus – prvi korak do korpusa slovenščine kot tujega jezika. V Fišer, D., Erjavec, T. (ur.), *Zbornik konference Jezikovne tehnologije in digitalna humanistika 2020* (str. 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, M. (2022). KOST med korpusi usvajanja tujega jezika. V Pirih Svetina, N., Ferbežar, I. (ur.), *Simpozij Obdobja 41: Na stičišču svetov: Slovenščina kot drugi in tuji jezik* (str. 23–30). Založba Univerze v Ljubljani. https://doi.org/10.4312/Obdobja.41.2784-7152

Stritar Kučuk, M. (2023a). *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. V *CLARC 2023: Jezik i jezični podaci: Knjižica sažetaka*. https://uniri-my.sharepoint.com/:w:/g/personal/bperak_uniri_hr/EdBOkvsg4vJOrVeHTkQw3uYB16acgdyFh2g5S5fpdXqhYA?rtime=RLP28Kne20g

Stritar Kučuk, M., & Šter, H. (2021). *Slovenščina 1+: Slovnične tabele in vaje za južnoslovanske govorce slovenščine kot drugega jezika*. Znanstvena založba Filozofske fakultete.

Stritar Kučuk, M., Pisek, S., & Šter, H. (2023). *Slovenščina 1+: Besedila in besedišče za južnoslovanske govorce slovenščine kot drugega jezika 1.1*. Založba Univerze.

Volodina, E., Granstedt, L., Matsson, A., Megyesi, B., Pilán, I., Prentice, J., Rosén, D., Rudebeck, L., Schenström, C., Sundberg, G., & 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., & Volodina, E., (2018). SVALA: Annotation of Second-Language Learner Text Based on Mostly Automatic Alignment of Parallel Corpora. V *Selected papers from the CLARIN Annual Conference 2018. Linköping Electronic Conference Proceedings* 159 (str. 227–239).