# Modul Leto plus – prvi korak do korpusa slovenščine kot tujega jezika

**Mojca Stritar Kučuk**

Oddelek za slovenistiko, Filozofska fakulteta, Univerza v Ljubljani
Aškerčeva 2, 1000 Ljubljana
mojca.stritarkucuk@ff.uni-lj.si

## 1. Uvod

Korpusi usvajanja tujega jezika so plodno raziskovalno in aplikativno področje.[1] Največkrat je sicer vezano na angleščino kot tuji jezik, obstajajo pa tudi korpusi za druge jezike, kot so hrvaščina (Mikelić Preradović et al., 2015), češčina (Hana et al., 2010), ruščina (Kutuzov in Kunilovskaya, 2014), nemščina (Reznicek et al., 2012), švedščina (Hammarberg, 2010) ali arabščina (Alfaifi et al., 2014). O korpusih usvajanja slovenščine kot tujega jezika je bilo že precej govora, vendar je bilo v glavnem omejeno na teoretične razmisleke in pilotne projekte (Stritar, 2012). Do konkretne izdelave korpusa usvajanja slovenščine kot tujega oz. drugega jezika pa še ni prišlo, predvsem zaradi finančnih oz. človeških razlogov. Izpeljava takih, manjših projektov, ki so zanimivi za ožjo skupino uporabnikov, je namreč nemalokrat odvisna od prizadevnosti posameznikov.

V zadnjem letu se je projekt izdelave korpusa usvajanja slovenščine kot drugega oz. tujega jezika končno začel. Čeprav ne čisto terminološko ustrezno – ne nazadnje gre pri večjem delu vključenih besedil za slovenščino kot drugi in ne tuji jezik –, smo korpusu zaradi ekonomičnosti nadeli poimenovanje KOST (= korpus slovenščine kot tujega jezika). V tem prispevku bodo predstavljeni prvi koraki na poti do pravega, splošneje uporabnega korpusa.

## 2. Zametek projekta: modul Leto plus

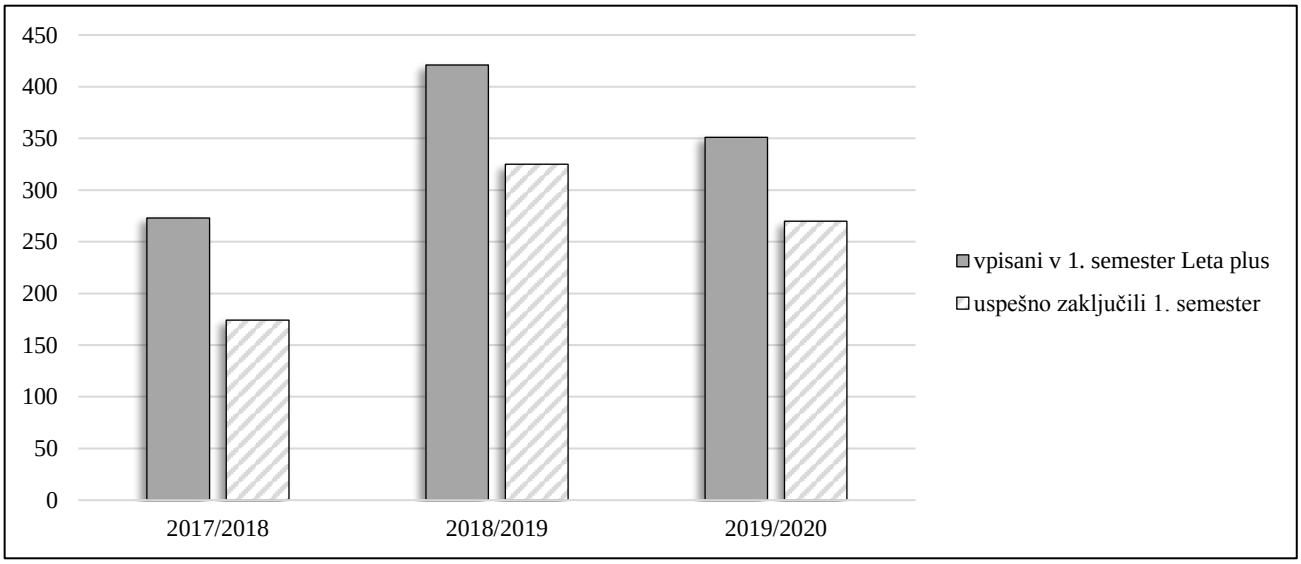
Za zagon izdelave KOST-a je ključna odločitev Univerze v Ljubljani, da je v okviru projekta *Izboljšanje procesov internacionalizacije slovenskega visokega šolstva* v študijskem letu 2016/17 pilotno izvedla Leto plus. Gre za poseben modul, namenjen tujim študentom, redno vpisanim v študijske programe Univerze v Ljubljani. V nadaljnjih študijskih letih se Leto plus[2] izvaja kot eden od ukrepov internacionalizacije univerze, financirano pa je iz razvojnega stebra univerzitetnega financiranja.

Najpomembnejša stvar, ki jo Leto plus ponuja redno vpisanim tujim študentom, je, da jim hkrati z rednim študijem omogoča brezplačno učenje slovenščine na dveh lektoratih v obsegu treh študijskih ur na teden. To pomeni okoli 90 kontaktnih ur v predavalnici, pridružuje pa se jim še okoli 30 ur dodatnih projektov, ki so izpeljani izven predavalnice ali pa jih študenti izvedejo sami.

Modul Leto plus je za zagon KOST-a idealen, saj imamo v njem zagotovljen dostop do večjega števila govorcev slovenščine kot tujega jezika (slika 1).

[1] Za rastoči seznam obstoječih korpusov prim. https://uclouvain.be/en/research-institutes/ilc/cecl/learner-corpora-around-the-world.html.

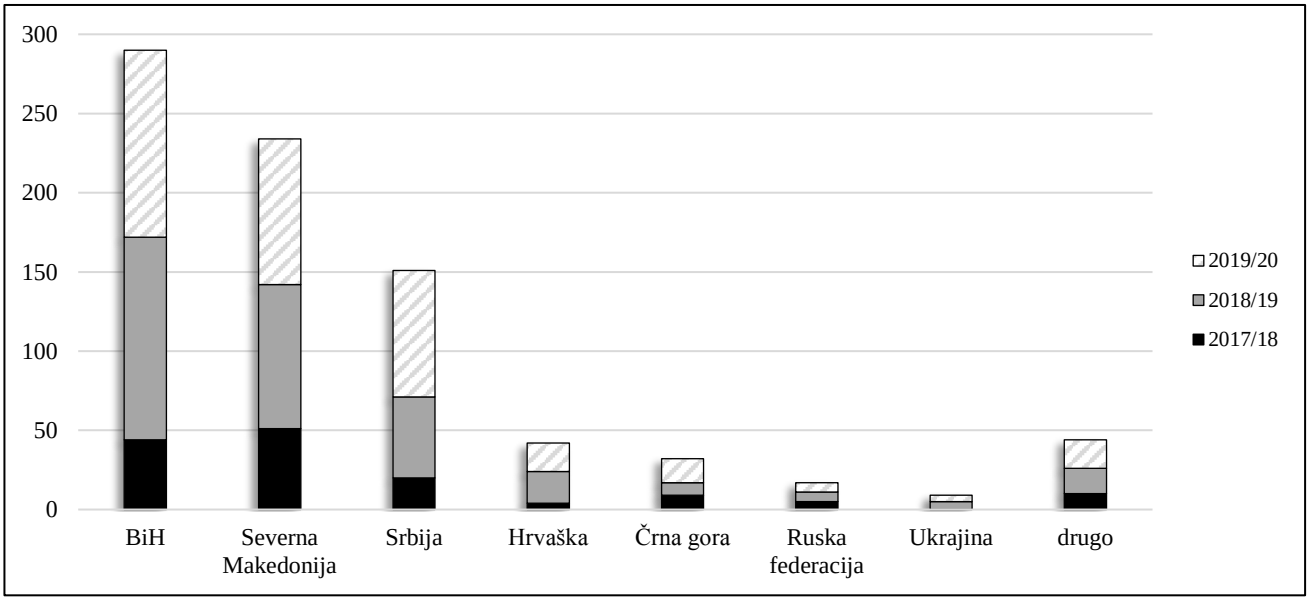
[2] https://www.uni-lj.si/studij/leto-plus.




Slika 1. Število tujih študentov v modulu Leto plus.[3]

Gre za govorce s precej homogenim jezikovnim ozadjem. Kot kaže slika 2, jih največ prihaja iz Bosne in Hercegovine, Srbije in Makedonije. Tudi njihova jezikovna zmožnost v slovenščini je precej primerljiva: v prvem, zimskem semestru večinoma začnejo brez predhodnega znanja slovenščine, zaradi skupne slovanske osnove pa hitro napredujejo in so ob koncu letnega semestra sposobni jezikovno (pre)živeti v slovenskem okolju, komunicirati o vsakdanjih temah in do neke mere v slovenščini opravljati redne študijske obveznosti. Dodatna prednost Leta plus je, da s temi govorci slovenščine kot tujega jezika nimamo samo stika, pač pa od njih redno tudi dobivamo besedila v slovenščini. Za pristop k izpitu morajo namreč oddajati pisne domače naloge.

Ne nazadnje je za KOST pomembno tudi, da smo pedagoške delavke na Letu plus večinoma redno zaposlene lektorice, tako da izdelava korpusa sodi med naše redne delovne obveznosti.




Slika 2. Udeleženci Leta plus po državah. Podatki so za zimski semester.

[3] Podatki na grafu so varljivi: največ študentov je bilo sicer v študijskem letu 2018/19, vendar le zaradi tega, ker vpis v modul ni bil omejen. Interes za Leto plus dejansko narašča; v študijskem letu 2019/20 je bilo zavrnjenih več kot sto študentov, ker zanje ni bilo mogoče zagotoviti mest na lektoratu.

# 3. Zbiranje besedil v okviru Leta plus

Čeprav dobra praksa uči, da je treba zasnovo jezikovnih korpusov načrtovati pred začetkom same gradnje, smo se izdelave KOST-a lotili oportunistično. Leto plus je namreč popoln vir za pridobivanje korpusnih besedil, ki bi ga bilo škoda ne izkoristiti.

Z zbiranjem gradiva sem pričela v študijskem letu 2018/19, v študijskem letu 2019/20 pa smo pri zbiranju že sodelovale vse redno zaposlene lektorice na Letu plus. Za uskladitev zbirke besedil in podatkov o tvorcih skrbim avtorica tega prispevka. Do sedaj je bilo zbranih skoraj 1500 pretežno krajših besedil, ki jih je napisalo skoraj 290 tvorcev. Pri enotnem zbiranju in shranjevanju besedil smo sprejeli nekaj odločitev, podrobneje pojasnjenih v nadaljevanju.

## 3.1. Dovoljenje za uporabo besedil in osebni podatki

Ker sta ureditev pravic za uporabo podatkov in varovanje osebnih podatkov ključnega pomena, so vsi študenti dobili v podpis izjavo, s katero dovoljujejo vključitev besedil, ki jih pišejo pri predmetih v okviru modula Leto plus v tekočem študijskem letu, v KOST. K izjavi sodi zbiranje osebnih podatkov, 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. Ti podatki bodo vključeni v glave besedil v KOST-u. V korpusu bodo tvorci anonimni, osebni podatki, ki se pojavljajo v besedilih, pa bodo odstranjeni oz. prekriti s kodami.

Izjavo, ki so jo pravno preverili na Oddelku za upravljanje s tveganji in varstvo osebnih podatkov na Univerzi v Ljubljani, smo študentom na Letu plus v podpis ponudile njihove lektorice. Pred podpisom smo jim natančno razložile projekt in pogoje sodelovanja. Razveseljivo je, da so v preteklih dveh študijskih letih izjavo podpisali vsi, ki jim je bila ponujena.

Vse podpisane izjave so digitalizirane in shranjene v obeh oblikah, papirni in digitalni.

## 3.2. Podatki o besedilih

Kar se tiče prvega jezika tvorcev besedil, oportunistično zberemo vsa besedila, ki jih dobimo. V skladu s študentsko populacijo Leta plus so zaenkrat prvi jeziki skoraj treh četrtin tvorcev srbščina, bosanščina ali hrvaščina. Pri označevanju podatka o prvem jeziku tvorca upoštevamo tisto, kar sami napišejo v prej omenjeni izjavi. Šele analiza zbranega korpusnega gradiva bo pokazala, ali prihaja med govorci teh treh jezikov do pomembnejših razlik pri usvajanju slovenščine. Prvi jezik za veliko skupino, to je kar četrtino govorcev, je makedonščina. Sicer pa so med zbranimi besedili še besedila govorcev s prvimi jeziki češčino, hebrejščino, italijanščino, madžarščino, nemščino, ruščino, slovaščino in ukrajinščino.

Študenti na lektoratu slovenščine tvorijo različnejše vrste besedil. Največ je esejev oz. spisov na različne teme (npr. o družini, prehrani, zdravju), nekaj pa je tudi praktičnega pisanja (npr. življenjepis, prošnja za delo). Ker je v praktičnih besedilih veliko osebnih podatkov, ki bi jih bilo treba zakrivati, in le malo dejanske tvorbe besedil, smo se odločili, da jih v KOST vključujemo v manjši meri.

Tri četrtine do sedaj zbranih besedil so študenti napisali doma kot domačo nalogo. Pri tem je nemogoče nadzirati, ali jih dejansko napišejo sami in koliko si pomagajo z jezikovnimi viri in orodji. V KOST bodo vključena vsa, če pa tisti, ki pripravlja vnos, sumi, da gre za uporabo spletnega prevajalnika ali pomoč rojenega govorca slovenščine, to zabeleži kot posebno opombo. Preostanek predstavljajo besedila, napisana na izpitu, torej pod bolj nadzorovanimi pogoji tvorjenja.

Pomembna je tudi jezikovna zmožnost tvorcev besedil. Večina študentov začne brez predznanja slovenščine, ker so z južnoslovanskega govornega področja, pa je njihov napredek hiter. V korpus zaenkrat vključujemo njihova besedila, ki jih napišejo na izpitu ob koncu prvega semestra in dobijo pozitivno oceno, ter besedila iz drugega semestra učenja slovenščine. Če je to njihovo prvo leto učenja, stopnjo jezikovne zmožnosti označimo kot »Južni Slovan začetnik«. Tako zbranih besedil je trenutno dobrih 70 %, preostalo pa so besedila nadaljevalcev in izpopolnjevalcev, ki bi jih v korpusu sicer želeli imeti več. V nadaljnjih fazah gradnje bomo v KOST predvidoma vključili tudi besedila začetnikov iz drugih govornih področij.

### 3.3. Način zbiranja besedil

Velika večina besedil, ki jih oddajajo študenti na Letu plus, je napisana na računalnik. Študenti jih oddajajo po elektronski pošti ali prek spletne učilnice na Filozofski fakulteti. Digitalna oblika zbiranje besedil seveda zelo olajša. Letni semester 2019/20 je bil glede dostopa do takih besedil zelo produktiven, saj se je zaradi pandemije celotno poučevanje preselilo v digitalno okolje.

Besedila, ki nastanejo na izpitu ali med lektoratom v razredu, pa so napisana na roko in jih je treba pretipkati. V skladu s svojimi časovnimi zmožnostmi to opravljamo lektorice Leta plus, saj predvidevamo, da bi se med besedili, ki nastanejo kot domače naloge, in tistimi, ki so napisana v bolj stresnih izpitnih okoliščinah, lahko pokazale tudi jezikovne razlike. Zaenkrat je med zbranimi besedili takih, ki so bila napisana na roko, slaba tretjina. Če bodo prve raziskave na gradivu pokazale, da razlik ni, in če se bo izkazalo, da nam za to delo primanjkuje časa, pa se bomo omejili samo na digitalno oddano gradivo.

### 3.4. Priprava besedil za vključitev v KOST

Besedila bodo v KOST-u točno taka, kot jih napišejo študenti, brez kakršnih koli jezikovnih popravkov. Ob tem si dovoljujemo izjemo. Ker pravopis ni v ospredju raziskav pri slovenščini kot tujem jeziku in ker bi radi s tem zvišali uspešnost morebitnega avtomatskega označevanja korpusnega gradiva v prihodnosti, v besedilih popravimo stičnost ločil ter odstranimo dvojne presledke. Zaradi lažje berljivosti smo se odločili tudi, da pri pretipkavanju besedil, prvotno napisanih samo z velikimi tiskanimi črkami, uporabimo male tiskane črke in pri tem upoštevamo slovenska pravopisna pravila, pa čeprav jih tvorec besedila morda ne bi.

Že v fazi zbiranja besedil smo se povezali s Centrom za jezikovne vire in tehnologije, kjer so nam pomagali pri nekaterih tehničnih vidikih. Vsako besedilo, vključeno v KOST, je shranjeno v samostojno tekstovno datoteko. V Excelovi tabeli so zbrani metajezikovni podatki o tvorcih in besedilih. Na podlagi te tabele in tekstovnih datotek z besedili bo zgeneriran korpus, primeren za analizo in objavo na spletu.

## 4. Nadaljnji koraki

Z začetkom študijskega leta 2020/21 smo zbiranje gradiva za KOST razširili še na druge institucije, neposredno povezane s slovenščino kot tujim jezikom. Najprej je to seveda Center za slovenščino kot drugi in tuji jezik (CSDTJ) Filozofske fakultete Univerze v Ljubljani,[4] na katerem imajo stalen dotok besedil tujejezičnih govorcev slovenščine zagotovljen v programih Tečaji slovenščine, Izpitni center, Seminar slovenskega jezika, literature in kulture, Slovenščina na tujih univerzah ter Slovenščina za otroke in mladostnike. Ker so programi CSDTJ za razliko od Leta plus manj vezani na govorce iz držav nekdanje Jugoslavije, računamo, da bo z vključitvijo teh besedil KOST zajemal bistveno več prvih jezikov. Če bo mogoče, bomo njihove deleže uravnotežili, tako da KOST ne bo samo korpus južnoslovanskih govorcev slovenščine kot tujega jezika.

KOST bo najprej dostopen v surovi, neoznačeni obliki. Vendar je največja dodana vrednost korpusov usvajanja jezika v označenosti napak. Jezikoslovne dileme ob tem so bile do neke mere rešene že v nekaterih predhodnih raziskavah in v obstoječih jezikovnih virih (Kosem et al., 2012, korpus lektorskih popravkov Lektor),[5] za tehnološki vidik pa še iščemo najustreznejšo rešitev, ki bi omogočala učinkovito dodajanje oznak napak v korpusna besedila. Spomladi 2020 je bil KOST vključen v uspešno prijavo na razpis Razvoj slovenščine v digitalnem okolju.[6] S tem so bila pridobljena sredstva za razvoj orodja za označevanje jezikovnih napak oz. njihovih popravkov v besedilih ter za razvoj novega konkordančnika, ki bo omogočal napredno iskanje po jezikovnih napakah in popravkih, njihovo vizualizacijo in izvažanje korpusnih podatkov. Šele ko bo poskrbljeno za to, se bomo lahko lotili zamudnega označevanja besedil in spotoma ocenili še, kolikšen delež korpusnih besedil je smiselno in izvedljivo označevati. Predvidoma bo delno označeni KOST javno dostopen na spletu leta 2022.

[4] https://centerslo.si/.
[5] http://korpus-lektor.net.
[6] https://www.gov.si/zbirke/javne-objave/javni-razpis-razvoj-slovenscine-v-digitalnem-okolju-jezikovni-viri-in-tehnologije/.

Zaradi stalnega dotoka besedil je KOST lahko zbirka, ki raste in se dopolnjuje, posebno dodano vrednost pa vidimo tudi v možnosti longitudinalnih raziskav. Kar nekaj govorcev slovenščine kot tujega jezika namreč prehaja med različnimi programi na Letu plus in CSDTJ, vse od začetne stopnje učenja do najvišjega, izpopolnjevalnega nivoja. Kako zagotoviti sledljivost in hkrati tvorcev tovrstnih besedil, ostaja vprašanje, ki ga bomo morali še rešiti.

## 5. Literatura

Abdullah Alfaifi, Eric Atwell in Ibraheem Hedaya. 2014. Arabic Learner Corpus (ALC) v2: A New Written and Spoken Corpus of Arabic Learners. V: *Proceedings of the Learner Corpus Studies in Asia and the World*, Kobe. https://www.dropbox.com/s/646ysp9q9h7353v/ALFAIFI%20LCSAW2014.pdf?dl=0.

Björn Hammarberg. 2010. *Introduction to the ASU Corpus: a longitudinal oral and written text corpus of adult learner Swedish with a corresponding part from native Swedes. Version 2010-11-16.* Stockholm University, Stockholm. http://su.diva-portal.org/smash/get/diva2:778204/FULLTEXT01.pdf.

Jirka Hana, Alexandr Rosen, Svatava Škodová in Barbora Štindlová. 2010. Error-tagged Learner Corpus of Czech. V: *Proceedings of the Fourth Linguistic Annotation Workshop*, str. 11–19, ACL, Uppsala.

Iztok Kosem, Mojca Stritar, Sara Može, Ana Zwitter Vitez, Špela Arhar Holdt in Tadeja Rozman. 2012. *Analiza jezikovnih težav učencev: korpusni pristop*. Trojina, zavod za uporabno slovenistiko, Ljubljana.

Andrey Kutuzov in Maria Kunilovskaya. 2014. Russian Learner Translator Corpus: Design, Potential and Application. V: *Text, Speech and Dialogue*, str. 315–323. Springer International Publishing.

Nives Mikelić Preradović, Monika Berać in Damir Boras. 2015. Learner Corpus of Croatian as a Second and Foreign Language. Multidisciplinary Approaches to Multilingualism. Ur. Kristina Cergol Kovačević, Sanda Lucija Udier. Peter Lang, Frankfurt am Main. 107–126.

Marc Reznicek, Anke Lüdeling, Cedric Krummes, Franziska Schwantuschke, Maik Walter, Karin Schmidt, Hagen Hirschmann in Torsten Andreas. 2012. *Das Falko-Handbuch. Korpusaufbau und Annotationen Version 2.01.* https://www.linguistik.hu-berlin.de/de/institut/professuren/korpuslinguistik/forschung/falko/FalkoHandbuchV2/.

Mojca Stritar. 2012. *Korpusi usvajanja tujega jezika*. Zveza društev Slavistično društvo Slovenije, Ljubljana.