Matej Klemen, Špela Arhar Holdt, Senja Pollak, Iztok Kosem,
Damjan Huber, Mateja Lutar

Korpus učbenikov za učenje slovenščine kot drugega in 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

KORPUS UČBENIKOV ZA UČENJE SLOVENŠČINE KOT DRUGEGA IN TUJEGA JEZIKA

Matej Klemen
Filozofska fakulteta, Univerza v Ljubljani, Ljubljana
matej.klemen@ff.uni-lj.si

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

Senja Pollak
Institut »Jozef Stefan«, Ljubljana
senja.pollak@ijs.si

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

Damjan Huber
Filozofska fakulteta, Univerza v Ljubljani, Ljubljana
damjan.huber@ff.uni-lj.si

Mateja Lutar
Filozofska fakulteta, Univerza v Ljubljani, Ljubljana
mateja.lutar@ff.uni-lj.si

DOI:10.4312/Obdobja.41.165-174

V prispevku prikažemo, kako je potekalo oblikovanje korpusa učbenikov za učenje slovenščine kot drugega in tujega jezika – KUUS, ki je nastal kot vzporedni projekt priprave stopenjskih beril na Centru za slovenščino kot drugi in tuji jezik. KUUS v trenutni različici vključuje 17 učbenikov, obsega 691.003 pojavnice oz. 491.022 besed in je skladno z načeli priprave tovrstnih jezikovnih virov opremljen z metapodatki in oznakami, ki omogočajo uporabo jezikovnih podatkov za različne namene. Predstavimo metodološke odločitve, ki smo jih sprejeli pri pripravi korpusa, trenutno različico korpusa in prvi primer uporabe korpusnih podatkov. Opišemo, kako smo podatke uporabili za pripravo pogostnostnih seznamov besed, ki so prvi korak do korpusno podprtega nabora jedrnega besedišča za slovenščino kot drugi ali tuji jezik in omogočajo primerjavo z drugimi seznami besed. Prispevek zaključimo z načrti za nadaljnji razvoj korpusa in seznamov.

slovenščina kot drugi in tuji jezik, korpus učbenikov, KUUS, seznam besed, *Skupni evropski jezikovni okvir*

This article describes the creation of a corpus of textbooks for learning Slovenian as a second and foreign language. The KUUS corpus was created as a parallel project for developing graded readers at the Center for Slovenian as a Second and Foreign Language. In its current version, KUUS includes seventeen textbooks, comprises 691,003 tokens or 491,022 words, and, in line with the principles of preparing language resources of this kind, is equipped with metadata and annotations that allow the linguistic data to be used for various purposes. The methodological decisions made in preparing the corpus, the current version of the corpus, and a first example of the use of corpus data are presented. The paper describes how the data were used to compile word frequency lists, which are the first step toward a corpus-based core vocabulary for Slovenian as a second or foreign language and allow comparison with other word lists. The article concludes with plans for further development of the corpus and lists.

Slovenian as a second and foreign language, textbook corpus, KUUS, word list, *Common European Framework of Reference for Languages*

1 Uvod

V prispevku predstavljamo nova odprto dostopna jezikovna vira, namenjena za podporo učenju slovenščine kot drugega in tujega jezika (SDTJ): korpus učbenikov za učenje slovenščine kot drugega in tujega jezika – KUUS, ki vsebuje učbenike za učenje SDTJ, in korpusno osnovan seznam jedrnega besedišča za stopnje A1, A2 ter B1 po *Skupnem evropskem jezikovnem okviru* (SEJO). Delo, ki v letu 2022 poteka s finančno pomočjo slovenske infrastrukture CLARIN.SI,[1] se je vzpostavilo leta 2019 ob pripravi stopenjskih beril[2] na Centru za slovenščino kot drugi in tuji jezik (CSDTJ). Poleg pedagoških in drugih strokovnih izkušenj pripravljavcev stopenjskih beril in preverjanja razumevanja vsebine prebranega ter neznanih besed pri testnih bralcih smo v želji po čim širšem konsenzu glede stopnje posameznega berila, mdr. tudi predvidene receptivne leksikalne zmožnosti bralcev stopenjskih beril, želeli upoštevati besedišče, vključeno v učbenike za SDTJ.

Različni seznami besed za posamezne stopnje imajo pri učenju tujih jezikov dolgo tradicijo. Za angleščino je bil npr. vpliven zlasti Westov seznam General Service List iz leta 1953, ki je v zadnjem času doživel revizijo (Browne, Culligan, Phillips 2013; Brezina, Gablasova 2015). Za SDTJ so besedni seznami na različne načine vključeni v jezikovne dokumente, npr. v *Preživetveno raven za slovenščino* (Pirih Svetina idr. 2004; Pirih Svetina 2016), *Sporazumevalni prag za slovenščino* (Ferbežar idr. 2004) itn. Obstoječi nabori besed so bili pripravljeni kot konsenz sestavljavcev posameznih dokumentov, seznam, ki ga predstavljamo v prispevku, pa temelji na korpusnem pristopu in v enem dokumentu združuje besedišče za različne stopnje. Razumemo ga kot osnovo za pripravo seznama jedrnega besedišča po stopnjah SEJO, pri čemer »jedro« razumemo kot sporazumno sprejeto, stabilno, vendar nadgradljivo izhodišče za učenje.

1 https://www.clarin.si/info/storitve/projekti/#Projekti_ki_jih_podpira_CLARINSI
2 Več o stopenjskih berilih v prispevku Klemen, Lojk v tem zborniku.

V nadaljevanju najprej prikažemo zasnovo in sestavo korpusa KUUS ter izpostavimo odločitve, ki smo jih sprejeli pri korpusni gradnji. Nato predstavimo uporabo korpusa za izdelavo izhodiščnega seznama besedišča po stopnjah SEJO A1, A2 in B1. Seznami temeljijo na korpusnem gradivu, ki smo ga primerjali z Referenčnim seznamom pogostega splošnega besedišča za slovenščino (Pollak idr. 2020), v določeni meri pa tudi ročno pregledali. Tako korpus KUUS kot korpusno pripravljeni seznami besedišča, opremljeni z vsemi relevantnimi informacijami, so javnosti od jeseni 2022 na voljo na repozitoriju CLARIN.SI pod licenco ACA ID-BY-NC-INF-NORED 1.0.[3] Prispevek strnemo z napovedjo nadaljnjega dela, ki vključuje korpusno nadgradnjo ter evalvacijo seznama besedišča v sodelovanju s širšo strokovno skupnostjo.

2 Priprava korpusa KUUS

Korpus zajema 17 učbenikov za učenje SDTJ: dva sta namenjena najstnikom, preostali pa odraslim (Tabela 1). Obsega 691.003 pojavnice oz. 491.022 besed. V korpus so vključeni učbeniki, ki so bili med letoma 2002 in 2022 izdani na CSDTJ in so bili umeščeni na lestvice SEJO (Lutar 2017, 2019) oz. je stopnja označena v učbeniku.[4] Gre za učbenike, ki se trenutno uporabljajo pri poučevanju SDTJ pri otrocih, mladostnikih in odraslih (Knez idr. 2021: 261–262, 342–343; Kavčič 2021: 26). V korpus nismo vključili vseh učbenikov CSDTJ. Izostal je učbenik za neopismenjene otroke *Križ kraž* na stopnji A1, saj v učbeniku razen navodil skoraj ni besedil (ta so skupaj z didaktičnimi navodili za izvedbo pouka in dejavnosti pri njem vključena v priročnik za učitelja). Prav tako nismo vključili verzije učbenika *A, B, C ... 1, 2, 3, gremo* iz leta 2011, ki je prilagojena za govorce albanščine, saj se v slovenskem delu besedila učbenik za govorce albanščine in splošni učbenik ne razlikujeta. Izbor učbenikov, vključenih v KUUS, pokriva različne stopnje SEJO, obsega precejšnji del aktualnih učbenikov za učenje SDTJ in zajema glavnino učbeniške produkcije CSDTJ.

V trenutno verzijo korpusa so vključeni le učbeniki, ne pa tudi (vsebinsko oz. strukturno drugačni) delovni zvezki. Nekateri učbeniki (*A, B, C ... 1, 2, 3, gremo*, *A, B, C ... gremo*, *Gremo naprej*, *Naprej pa v slovenščini*, *S slovenščino nimam težav*, *Mozaik slovenščine*), ki so predvideni za krajše, tj. do 80-urne tečaje, imajo del, namenjen (predvsem oblikoslovnim) vajam, kot se sicer pojavljajo v delovnih zvezkih. Ker so te vaje del enotne knjige, naslovljene kot *učbenik*, smo jih vključili v korpus.

Besedila so bila iz formata PDF ali DOC pretvorjena v format TXT. Ročno so bili odstranjeni deli, katerih predvideni sprejemnik ni učenec SDTJ oz. niso neposredno namenjeni pedagoškemu procesu: uvod, natančnejše vsebinsko kazalo oz. tabele ob koncu učbenika, kolofon, viri slik in besedil. Izbrisano je bilo besedilo, ki se pojavlja v glavi in nogi strani (npr. *1. enota Na tečaju slovenščine*; *2. enota Tristo petinšestdeset*

3 KUUS na http://hdl.handle.net/11356/1696, seznami besedišča pa na http://hdl.handle.net/11356/1697.

4 Vključuje tudi še neizdan učbenik z delovnim naslovom *Mozaik slovenščine*. Edini učbenik, ki nima oznake stopnje po SEJO in je vključen v korpus, je učbenik, namenjen pripravi na izpit iz znanja slovenščine na srednji in visoki ravni, *Pot do izpita iz znanja slovenščine*. Izpiti na srednji in visoki ravni so se izvajali do leta 2015 in so bili primerljivi z ravnjo B2 oz. C1.

dni; *3. enota A je daleč?*), razen številk strani. Če so v učbeniku tujejezična navodila, so bila označena s posebnimi kodami (npr. *XXXListen_and_write_down.XXX*). Besedila so bila pregledana in očiščena napak, ki so nastale pri pretvarjanju besedil v golo besedilno obliko.[5] Pogostejše napake pri pretvorbi so bile: napačen zapis *č*, *š*, *ž*, zapis besed s kombinacijo velikih in malih črk, napačno zapisana nestična končna ločila, male začetnice npr. na začetku povedi. Deljene besede so bile popravljene v nedeljene. V nekaterih učbenikih je bilo dopisano besedilo, ki je zaradi specifičnih pisav ali postavitev pri pretvorbi izpadlo.

Vsak od učbenikov je opremljen s podatki o naslovu, podnaslovu, avtorjih, letnici prve izdaje in izdaje, vključene v korpus, naslovniku, stopnji in predvidenem številu ur, v katerih naj bi učbenik predelali.

Naslov učbenika	**Stopnja po SEJO**
A, B, C ... 1, 2, 3, gremo	A1
A, B, C ... gremo	A1
Gremo naprej	A2
Naprej pa v slovenščini	B1
Slovenska beseda v živo 1a	A1
Slovenska beseda v živo 1b	A2
Slovenska beseda v živo 2	B1
Slovenska beseda v živo 3a	B2–C1
Slovenska beseda v živo 3b	C1
Slovenščina ekspres 1	A1
S slovenščino nimam težav	A2–B1
Jezikovod	C1
S slovenščino po svetu	C1
Čas za slovenščino 1	A1
Čas za slovenščino 2	A2
Mozaik slovenščine	A2
Pot do izpita iz znanja slovenščine	brez oznake

Tabela 1: Učbeniki, vključeni v korpus KUUS.

3 Priprava seznamov besedišča po stopnjah SEJO

3.1 Zasnova seznama besedišča in primerjava s seznamom pogostega splošnega besedišča

Brezina in Gablasova (2015) sta se pri oblikovanju seznama jedrnega besedišča za (britansko) angleščino odločila za leme za razliko od predhodnega Westovega seznama, ki je urejen po besednih družinah. Tej odločitvi sledimo tudi pri oblikovanju seznamov

5 Besedila so pregledali in uredili Matej Klemen, Teja Koren in Katja Krajnc.

jedrnega besedišča za SDTJ, saj je pri korpusno osnovanem pristopu urejanje po lemah in drugih korpusnih oznakah naravni prvi korak, ki omogoča napredno izrabo obstoječih orodij in postopkov, pa tudi primerjavo z drugimi korpusno pripravljenimi seznami, tudi s seznamom splošnega besedišča za slovenščino (Pollak idr. 2020).

Odločili smo se, da s korpusnim pristopom pripravimo izhodišče seznamov besedišča za spodnje tri stopnje SEJO, saj imamo do stopnje B1 razmeroma velik nabor učbenikov (12 od 17 vključenih v korpus), kar omogoča primerjavo med njimi. Za obravnavo besedišča na višjih stopnjah bo treba trenutno metodologijo nadgraditi oz. prilagoditi.

Sporazumevalni prag za slovenščino (Ferbežar idr. 2004), ki opisuje znanje uporabnika SDTJ na stopnji B1, v dodatku navaja seznam okoli 3900 besed. Dejanski obseg slovarja uporabnika SDTJ na stopnji B1 bi bilo treba potrditi z empiričnimi raziskavami in analizo korpusa usvajanja SDTJ (Stritar Kučuk 2020). Glede na tuje študije pa lahko sklepamo, da so številke za SDTJ podobne, čeprav med jeziki prihaja do razlik,[6] in pričakujemo, da bomo v seznamih, pripravljenih na podlagi KUUS, identificirali podobno število besed.

Korpusna besedila smo uvozili v orodje Sketch Engine (Kilgarriff idr. 2014), nato pa za vsakega od vključenih učbenikov izvozili besede[7] in informacijo o njihovi pogostnosti. Za vsako besedo smo izračunali relativno pogostnost v posameznem učbeniku in podatke združili v enoten tabelaričen izpis. V tabelo smo dodali pogostnost besede v celotnem korpusu, podatek, v koliko učbenikih od 17 se pojavlja in kolikokrat se pojavlja v učbenikih na posamezni stopnji SEJO. Tabela vsebuje 23.068 besed različnih besednih vrst.

Podatke v tabeli smo primerjali z besediščem, ki se pojavlja na Referenčnem seznamu pogostih splošnih besed (Pollak idr. 2020). Ta je bil pripravljen s prekrivanjem najpogostejših 10.000 lem glede na besedno vrsto iz štirih slovenskih besedilnih korpusov: uravnoteženega korpusa pisne slovenščine Kres, korpusa govorjene slovenščine GOS, korpusa računalniško posredovane komunikacije Janes ter korpusa šolske pisne produkcije Šolar 2.0, vsebuje pa 4768 pogostih splošnih lem (Arhar Holdt idr. 2020).

Rezultate primerjave prikazuje Tabela 2. Besede, ki se pojavljajo tako na seznamu pogostega splošnega besedišča kot v korpusu KUUS, so dobre kandidatke za vključitev na sezname jedrnega besedišča, označenega s stopnjami SEJO. Prekrivnost je precej visoka: samo 166 besed je takšnih, ki so na seznamu pogostih splošnih besed, ne pa v korpusu KUUS. Te bo treba natančneje pregledati, saj lahko nakazujejo

6 Če povprečimo rezultate Miltonovih testov (2010: 225) v zvezi z obsegom slovarja pri uporabnikih angleščine, francoščine in grščine kot tujega jezika, ugotovimo, da naj bi na stopnji A1 poznali 1247, na stopnji A2 1962 in na stopnji B1 2870 besed oz. poenostavljeno okoli 2000 besed od 5000 najpogostejših v jeziku za dosego stopnje A2 in 3000 za dosego stopnje sporazumevalnega praga (prav tam: 226). V projektu Kelly so pri pripravi seznamov besedišča za devet jezikov (arabščino, kitajščino, angleščino, grščino, italijanščino, norveščino, poljščino, ruščino in švedščino) za vsako stopnjo SEJO identificirali približno 1500 besed (Charalabopoulou idr. 2012: 49, 51).

7 Natančneje: lempose, tj. kombinacije leme in oznake besedne vrste, npr. *prijatelj-s*.

(vsebinske ali žanrske) vrzeli trenutnih učbenikov za učenje SDTJ.[8] V tretji skupini so besede, ki so v korpusu KUUS, ne pa na seznamu pogostih splošnih besed. Teh je po pričakovanjih precej, saj je seznam pogostih splošnih besed metodološko sorazmerno strogo zamejen (Arhar Holdt idr. 2020: 12–13). Ta del podatkov zato zahteva ročni pregled in odločitve, katere besede oz. kategorije besed vključevati med kandidate za sezname jedrnega besedišča in katerih ne.

Tip podatkov	**Št. besed**	**Primeri besed (po 20 lem z oznako »glagol«)**
Beseda se pojavlja tako na seznamu pogostega splošnega besedišča kot v korpusu KUUS.	4603	biti, imeti, iti, delati, govoriti, poslušati, jesti, prositi, priti, vedeti, piti, dobiti, želeti, gledati, prebrati, pisati, brati, dopolniti, učiti, videti
Beseda se pojavlja na seznamu pogostega splošnega besedišča, ne pa tudi v korpusu KUUS.	166	znebiti, žrtvovati, kršiti, obžalovati, pobiti, ubijati, zgrešiti, smrdeti, razpasti, pozivati, sramovati, umoriti, častiti, zmesti, zadati, poglobiti, odmevati, izvleči, dopovedati, planiti
Beseda se pojavlja v korpusu KUUS, ne pa tudi na seznamu pogostega splošnega besedišča.	18.465	telefonirati, kolesariti, tuširati, rezervirati, *želite, *povežite, parkirati, fotografirati, rolati, *označite, prestopiti, *obkrožite, *grem, *čaj, *izvolite, šolati, *glagoli, kartati, *mmm, sklanjati

Tabela 2: Primerjava besedišča v korpusu KUUS in na seznamu pogostih splošnih besed (z zvezdico * so označene napačno lematizirane besede).

3.2 Predrazvrstitev besedišča na stopnje SEJO

Na osnovi celovitega prvega pregleda podatkov smo določili robustne številčne kriterije, s katerimi smo besede v tabeli opremili z izhodiščno oznako stopnje SEJO (Tabela 3). Kriteriji se upoštevajo zaporedno: najprej se preveri kriterije za A1-jedro, ustrezajoče besedišče se označi, sledi preverba kriterija za A1-širše in tako dalje. Pri pripravi kriterijev smo upoštevali, da je za stopnjo B1 na voljo manj učbenikov kot za A1 in A2 ter da se v gradivu pojavlja tudi učbenik, ki zajema dve stopnji (A2–B1, gl. Tabelo 1). V Tabeli 3 ločeno navajamo število označenih besed, ki se pojavijo tudi na seznamu pogostih splošnih besed (*da*), in tistih, ki se ne (*ne*).

8 Hiter pregled pokaže, da gre predvsem za besedišče z negativno konotacijo, kakršno se pojavlja denimo v novicah črne kronike.

Oznaka	Kriterij za besedo	Št. besed	Primeri besed (po 20 lem z oznako »glagol«)
A1-jedro	Beseda se pojavi v vseh petih učbenikih stopnje A1.	330 (da) 66 (ne)	biti, imeti, iti, delati, govoriti, poslušati, jesti, prositi, priti, vedeti, piti, dobiti, želeti, gledati, prebrati, pisati, brati, dopolniti, učiti, videti
A1-širše	Beseda se pojavi v štirih, treh ali dveh učbenikih stopnje A1.	652 (da) 624 (ne)	morati, kupiti, živeti, pogovarjati, poznati, moči, študirati, začeti, narediti, znati, potrebovati, potovati, ogledati, hoteti, vzeti, ukvarjati, plačati, obleči, najti, voziti
A2	Beseda se pojavi največ v enem učbeniku stopnje A1, na stopnji A2 se pojavi v petih, štirih, treh ali dveh učbenikih. (*) Izjema: Če se beseda na stopnji A2 pojavi v dveh učbenikih, od katerih je eden od teh učbenik z oznako A2–B1, se umesti na stopnjo B1.	998 (da) 773 (ne)	uporabljati, zdeti, smeti, postati, ostati, odločiti, pripraviti, vpisati, meniti, naučiti, skrbeti, ugotoviti, trajati, zgoditi, preživeti, poskusiti, primerjati, spomniti, roditi, skuhati
B1	Beseda se v učbenikih stopnje A1 ne pojavi, na stopnji A2 se pojavi v največ enem učbeniku, na stopnji B1 se pojavi v enem ali dveh učbenikih. Beseda mora imeti v celotnem korpusu pogostnost vsaj 2. Na to stopnjo se uvrstijo tudi izjeme (*).	1231 (da) 1798 (ne)	vplivati, doseči, obstajati, izražati, pojaviti, privoščiti, spodbujati, pojavljati, dajati, storiti, ohraniti, opraviti, povzročati, odkriti, ustrezati, upoštevati, navesti, spreminjati, povezovati, odločati
B2&C1	Beseda se v učbenikih stopnje A1, A2, B1 ne pojavi, ampak se pojavi bodisi v učbeniku z oznako B2–C1 ali pa vsaj enem od učbenikov na stopnji C1.	863 (da) 9056 (ne)	pripisati, opozarjati, zavzemati, lotiti, braniti, naleteti, omogočiti, napovedovati, potegniti, pripomoči, približati, prikazovati, pripadati, temeljiti, odrezati, preiti, obsoditi, opredeliti, prepustiti, vmešavati
POZOR	Primeri, ki se po kriterijih ne umestijo v nobeno od skupin: imajo atipično distribucijo, so na prehodih med stopnjami ipd.	529 (da) 6148 (ne)	zaigrati, opremiti, brskati, razviti, oblikovati, izstopiti, dejati, izpisati, označevati, odpirati, zvedeti, ločevati, zaščititi, položiti, javiti, izpostaviti, liti, odrasti, reagirati, nasprotovati

Tabela 3: Kriteriji za izhodiščno razvrščanje na stopnje SEJO s številom besed in primeri.

3.3 Ročni pregled in prvi rezultati

Besede, ki so dobile oznake A1, A2, B1 in hkrati niso del referenčnega seznama splošnega pogostega besedišča, smo ročno pregledali in vsebinsko kategorizirali. Določen delež besed smo prepoznali kot relevantne kandidate za vključitev na sezname jedrnega besedišča, označenega s stopnjami SEJO. Med temi besedami je tudi za učbenike tipična jezikoslovna terminologija oz. metajezik, ki se na obravnavanih stopnjah SEJO lahko pojavlja le terminološko (npr. *poved*, *pogojnik*, *modalen*) ali pa tako terminološko kot v splošnem pomenu (npr. *tvorba*, *nedoločen*, *pomensko*). Kot relevantne za seznam smo prepoznali tudi različne simbole in okrajšave (npr. *št.*, *tj.*, *prof.*). Med kategorijami, ki smo jih prepoznali kot nerelevantne za končni seznam, so: napačno označeni ali lematizirani primeri (npr. *dokončajte*, *samostalniki*, *nedov*), lastnoimenski samostalniki (npr. *Tone*, *Primožič*, *Krško*) in številniki, ki niso sistematično reprezentirani, zato jih je smiselno na seznam dodati posebej. V določenih primerih smo ob pregledu besede v kontekstu odkrili, da zaradi napačne oznake, enakopisnosti ali večpomenskosti sodi na stopnjo, višjo od B1. Tem smo pripisali posebno oznako, da bodo ustrezno obravnavane kasneje. Rezultate predstavlja Tabela 4.

Pripisana stopnja	**Število kandidatk za jedrno besedišče**	**Število kandidatk za seznam terminoloških oz. metajezikovnih izrazov**	**Skupno število kandidatk**
A1-jedro	344	6	350
A1-širše	840	24	864
A2	1433	18	1451
B1	2518	90	2608
SKUPAJ	5135	138	5273

Tabela 4: Število besed, predoznačenih s stopnjo SEJO A1, A2 ali B1.

Kot kaže Tabela 4, je izhodišče za jedrno besedišče po stopnjah SEJO glede obsega primerljivo z mednarodnimi praksami. Upoštevati je treba tudi, da je metodologija v prvem koraku namensko popustljiva: raje obdržimo kako besedo preveč kot premalo. Pri nadaljnjem delu bomo na seznam dodali morebitne dodatne kandidatke iz kategorije POZOR, hkrati pa odstranili morebitne redundantne skupine besed, kot so npr. izlastnoimenski svojilni pridevniki.

4 Zaključek in nadaljnje delo

V prispevku smo predstavili nova jezikovna vira za poučevanje SDTJ, korpus KUUS in izhodiščni seznam jedrnega besedišča za stopnje SEJO A1–B1. V prihodnje želimo tako korpus kot seznam še nadgrajevati. V korpus bi bilo mogoče dodati delovne zvezke, morda tudi priročnike za učitelje, pri čemer bi bilo smiselno poskrbeti za označevanje različnih razdelkov korpusnega gradiva (npr. ločiti navodila, naloge, razlage itn. (Koren 2018)). Razmisliti bi bilo treba tudi o vključitvi učbenikov drugih

izdajateljev, s čimer bi lahko pridobili več učbenikov na posamezni stopnji in zmanjšali morebitni avtorski vpliv na izbiro vsebine. Ker učbeniki, izdani pri drugih založbah, niso nujno umeščeni na SEJO oz. ni vedno jasno, kako je bila stopnja določena, bi bilo treba v primeru te odločitve ustrezno nadgraditi metodologijo gradnje korpusa in na njem temelječih rezultatov.

Sezname besedišča bomo v naslednjem koraku pripravili v obliko za celovit ročni pregled, ki ga bomo opravili v sodelovanju skupine strokovnjakov, ki poučujejo SDTJ. Pri pregledovanju bomo sopostavili obstoječe oznake ter mnenja strokovne skupnosti, na kateri stopnji bi določeno besedo pričakovali. Končni seznami za stopnje A1, A2 in B1 bodo pripravljeni z ročnim pregledom in v iskanju širšega konsenza skupnosti. Srednjeročni načrt je dopolniti seznam s slovničnimi in pomenskimi informacijami, npr. oblikoslovnimi paradigmami (vključno s pogostnostjo oblik, saj pogostnost leme ne pove dosti o rabi posamezne oblike), pomenskimi indikatorji, besednimi zvezami ipd.

Ob ročnem pregledu seznama besedišča bomo natančneje ocenili njegov domet. Zanima nas predvsem, kolikšen delež besedil v (različnih) slovenskih korpusih pokriva seznam iz KUUS in kje so vrzeli, ki bi jih bilo dobro nasloviti pri nadaljnjem razvoju seznama in učnega gradiva v širšem smislu. Natančneje bomo primerjali podobnosti in razlike z Referenčnim seznamom pogostih splošnih besed in primerljivih obstoječih seznamov za slovenščino ter posledično po potrebi dopolnili tudi te.

Uporabno vrednost tako pripravljenega seznama vidimo mdr. kot pomoč pri pripravi učnih gradiv in jezikovnih testov, ki bodo lahko zvesteje odsevali jezikovno realnost uporabnikov SDTJ, ne pa izhajali le iz izkušenj in pričakovanj sestavljavcev. Uporabni pa bodi tudi pri razvoju orodij za luščenje dobrih zgledov, avtomatsko ocenjevanje različnih značilnosti pisne produkcije učečih se in drugih jezikovnih tehnologij.

Literatura

ARHAR HOLDT, Špela, POLLAK, Senja, ROBNIK ŠIKONJA, Marko, KREK, Simon, 2020: Referenčni seznam pogostih splošnih besed za slovenščino. Darja Fišer, Tomaž Erjavec (ur.): *Zbornik konference Jezikovne tehnologije in digitalna humanistika*. Ljubljana: Inštitut za novejšo zgodovino. 10–15. http://nl.ijs.si/jtdh20/pdf/JT-DH_2020_Arhar-Holdt-et-al_Referencni-seznam-pogostih-splosnih-besed-za-slovenscino.pdf

BREZINA, Vaclav, GABLASOVA, Dana, 2015: Is There a Core General Vocabulary? Introducing the New General Service List. *Applied Linguistics* XXXVI/1. 1–22.

BROWNE, Charles, CULLIGAN, Brent, PHILLIPS, Joseph, 2013: *The New General Service List*. http://www.newgeneralservicelist.org

CHARALABOPOULOU, Frieda, GAVRILIDOU, Maria, JOHANSSON KOKKINAKIS, Sofie, VOLODINA, Elena, 2012: Building Corpus-Informed Word Lists for L2 Vocabulary Learning in Nine Languages. Linda Bradley, Sylvie Thouësny (ur): *CALL: Using, Learning, Knowing: EUROCALL Conference, Gothenburg, Sweden, 22-25 August 2012, Proceedings*. Dublin: Research-publishing.net. 49–53.

FERBEŽAR, Ina, KNEZ, Mihaela, MARKOVIČ, Andreja, PIRIH SVETINA, Nataša, SCHLAMBERGER BREZAR, Mojca, STABEJ, Marko, TIVADAR, Hotimir, ZEMLJARIČ MIKLAVČIČ, Jana, 2004: *Sporazumevalni prag za slovenščino*. Ljubljana: Center za slovenščino kot drugi/tuji jezik pri Oddelku za slovenistiko Filozofske fakultete, Ministrstvo RS za šolstvo, znanost in šport.

KAVČIČ, Patricija, 2021: *Učenje in poučevanje slovenščine kot drugega in tujega jezika v Sloveniji*. Diplomsko delo. Ljubljana: Filozofska fakulteta.

KILGARRIFF, Adam, BAISA, Vít, BUŠTA, Jan, JAKUBÍČEK, Miloš, KOVÁŘ, Vojtěch, MICHELFEIT, Jan, RYCHLÝ, Pavel, SUCHOMEL, Vít, 2014: The Sketch Engine: ten years on. *Lexicography* 1. 7–36.

KNEZ, Mihaela, FERBEŽAR, Ina, KERN ANDOLJŠEK, Damjana, STABEJ, Marko, 2021: *Evalvacija modelov učenja in poučevanja slovenščine kot drugega jezika za učence in dijake, ki jim slovenščina ni materni jezik. Zaključno poročilo*. Ljubljana: Center za slovenščino kot drugi in tuji jezik.

KOKKINAKIS, Sofie Johansson, VOLODINA, Elena, 2011: Corpus-based approaches for the creation of a frequency based vocabulary list in the EU project KELLY–issues on reliability, validity and coverage. Iztok Kosem, Karmen Kosem (ur.): *Electronic lexicography in the 21st century: new applications for new users. Proceedings of eLex* 2011. Ljubljana: Trojina. 129–139.

KOREN, Teja, 2018: *Priprava poskusnega korpusa učbenikov za učenje slovenščine kot drugega in tujega jezika*. Magistrsko delo. Ljubljana: Filozofska fakulteta.

LUTAR, Mateja, 2017: Učbeniki Centra za slovenščino kot drugi in tuji jezik in Skupni evropski jezikovni okvir. *Stephanos* 3. 45–53.

LUTAR, Mateja, 2019: Učbeniki za poučevanje slovenščine kot drugega in tujega jezika. Prispevek na 3. mednarodni znanstveni konferenci Slavistični znanstveni premisleki, Slovenščina kot drugi in tuji jezik v izobraževanju, Maribor, 20. 11. 2019.

MILTON, James, 2010: The development of vocabulary breadth across the CEFR levels. Inge Bartning, Maisa Martin, Ineke Vedder (ur.): *Communicative Proficiency and Lingustic Development: intersections between SLA and language testing research*. Eurosla. 211–231.

PIRIH SVETINA, Nataša, 2016: *Preživetvena raven za slovenščino: za potrebe programa Opismenjevanje v slovenščini za odrasle govorce drugih jezikov*. Ljubljana: Center za slovenščino kot drugi in tuji jezik. https://centerslo.si/wp-content/uploads/2016/07/IC_Prezivetvena_2016.pdf

PIRIH SVETINA, Nataša, RIGLER ŠILC, Katarina, LAVRIČ, Marjana, FERBEŽAR, Ina, JERMAN, Tanja, 2004: *Preživetvena raven v slovenščini*. Krakov: TAiWPN Universitas.

POLLAK, Senja, ARHAR HOLDT, Špela, KREK, Simon, ROBNIK-ŠIKONJA, Marko, 2020: *Reference List of Slovene Frequent Common Words*. Slovenian language resource repository CLARIN.SI. http://hdl.handle.net/11356/1346

Skupni evropski jezikovni okvir: učenje, poučevanje, ocenjevanje, 2011. Ljubljana: Ministrstvo RS za šolstvo in šport, Urad za razvoj šolstva.

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

Delo podpira iniciativa CLARIN.SI. Projekt Empirična podlaga za digitalno podprt razvoj pisne jezikovne zmožnosti (J7-3159) in programa Jezikovni viri in tehnologije za slovenski jezik (P6-0411) ter Tehnologije znanja (P2-0103) sofinancira Javna agencija za raziskovalno dejavnost Republike Slovenije iz državnega proračuna.