Nina Ledinek – Mitja Trojar

Zasnova in oblikovanje *Korpusa znanstvenih besedil sodobne slovenščine*

Cobiss: 1.01
https://doi.org/10.3986/JZ.31.2.06

V prispevku predstavljamo *Korpus znanstvenih besedil sodobne slovenščine*, specializirani pisni korpus slovenščine, ki obsega 33.604.256 pojavnic iz 884 znanstvenih in strokovnih besedil zlasti s področij družboslovja in humanistike, nastalih predvsem med letoma 2000 in 2023. Osredotočamo se na prikaz besedilnotipske sestave korpusa, tehničnih postopkov predpriprave korpusnih besedil, korpusnega označevanja, formatov zapisa korpusnih besedil in dostopnosti korpusa. Predstavljamo tudi motivacijo za izgradnjo korpusa in njegovo aplikativno vrednost, pri čemer skušamo opredeliti specifike in prednosti *Korpusa znanstvenih besedil sodobne slovenščine* glede na druge slovenske korpuse, ki vključujejo strokovna in znanstvena besedila.
Ključne besede: korpus znanstvenih besedil, specializirani korpus, korpusno označevanje, format CoNLL-U

The Concept and Design of the *Corpus of Scientific Texts of Contemporary Slovenian*

This article presents the *Corpus of Scientific Texts of Contemporary Slovenian*, a specialized written corpus of Slovenian comprising 33,604,256 tokens from 884 scholarly texts, primarily in the social sciences and the humanities, published mainly between 2000 and 2023. We focus on describing the text-type composition of the corpus, the technical procedures used in preprocessing corpus texts, corpus annotation, text encoding formats, and corpus accessibility. We also discuss the rationale for constructing the corpus and its practical applications, aiming to outline the specific characteristics and advantages of the *Corpus of Scientific Texts of Contemporary Slovenian* in comparison with other Slovenian corpora that include specialized texts.
Keywords: corpus of scientific texts, specialized corpus, corpus annotation, CoNLL-U format

Nina Ledinek ▪ ZRC SAZU, Inštitut za slovenski jezik Frana Ramovša, Ljubljana ▪ nina.ledinek@zrc-sazu.si ▪ https://orcid.org/0000-0003-1068-3856
Mitja Trojar ▪ ZRC SAZU, Inštitut za slovenski jezik Frana Ramovša, Ljubljana ▪ mitja.trojar@zrc-sazu.si ▪ https://orcid.org/0000-0003-3334-2413

Prispevek je nastal v okviru projekta *eSSKJ in korpus – na poti k najsodobnejšim jezikovnim podatkom*, ki ga je financiralo Ministrstvo za kulturo RS, in programa P6-0038, ki ga financira ARIS. Temelji na raziskovalnih podatkih, ki jih hranita Založba ZRC in repozitorij Clarin.si ter so javno dostopni na povezavah https://omp.zrc-sazu.si/zalozba, https://ojs.zrc-sazu.si/ in https://www.clarin.si/info/o-projektu/.

1 Uvod

Zaradi pospešenega razvoja zunajjezikovne stvarnosti se slovenščina zlasti v poimenovalnih prvinah spreminja bistveno hitreje kot nekoč. Potreba po zanesljivosti in verodostojnosti podatkov je toliko bolj pomembna pri obravnavi strokovne, (izhodiščno) terminološke leksike, saj znanstveno védenje hitro napreduje, se nadgrajuje ter s tem postaja vedno bolj specializirano in težje dostopno splošnim uporabnikom jezika. Po drugi strani smo zaradi popularizacije znanosti in zlasti vpliva medijev priča tudi nasprotnemu pojavu – vedno hitreje naraščajočemu fondu determinologizirane leksike, tj. leksike, ki je prvotno del terminologije posameznih strok, z rabo v besedilih, namenjenih širšemu krogu naslovnikov, pa postane del splošnega besedišča (gl. Žagar Karer 2011; Žagar Karer – Ledinek 2021). In kot determinologizirana je prvotno terminološka leksika prikazana tudi v splošnih razlagalnih slovarjih.

Da bi lahko za leksikografske in terminografske potrebe v gradivsko analizo vključili tudi rabo terminov (in leksemov), kot se pojavljajo v strokovnih in znanstvenih člankih uveljavljenih slovenskih periodičnih publikacij ter v strokovnih in znanstvenih monografijah,[1] smo na Inštitutu za slovenski jezik Frana Ramovša ZRC SAZU (v nadaljevanju: ISJFR) zasnovali *Korpus znanstvenih besedil sodobne slovenščine*, ki zajema prav ta segment rabe strokovnega jezika in dopolnjuje obstoječo slovensko jezikovno infrastrukturo.

Izdelava novega korpusa strokovnih in znanstvenih besedil je z vidika slovarskega opisa za slovensko jezikovno infrastrukturo pomembna, saj je delež stvarnih besedil v slovenskih jezikovnih korpusih, ki so osnovna podlaga za oblikovanje opisa v novem temeljnem enojezičnem splošnem razlagalnem slovarju slovenščine eSSKJ, skromen, zato obstoječi korpusni viri ne odsevajo realne pomenske razplastenosti slovenskega knjižnega jezika, omenjeno stanje pa negativno vpliva tudi na leksikalni opis (Krvina – Petric Žižić 2024). Korpus Gigafida (Krek idr. 2020; Logar idr. 2020), ki kot najsodobnejši referenčni korpus služi kot temeljna gradivska osnova za slovar eSSKJ, na primer v svoji najnovejši različici 2.0 vključuje le 3,8 % stvarnih besedil, pri čemer je bilo zgolj 65 stvarnih besedil (5 %) objavljenih po letu 2010 (zadnje je iz leta 2018), medtem ko je stvarnih besedil, nastalih med letoma 1991 in 2009, 1214 (95 %)[2] (Krek idr. 2019).

---

1 Specializirani korpus seveda ne more biti osnova za oblikovanje geslovnika za splošni razlagalni slovar, ampak služi kot gradivski vir zlasti za iskanje ponazarjalnega gradiva pri delno determinologizirani leksiki. Terminološki svetovalci, ki sodelujejo pri pripravi eSSKJ, namreč mnogokrat opozarjajo na (prevelike) stvarne napake pri zgledih z delno determinologizirano leksiko, pridobljenih iz referenčnega korpusa. Tudi sicer se v distribuciji in rabi (nedeterminologiziranih) leksemov v različnih tipih besedil (lahko) kažejo pomembne razlike, zato je za ustrezen jezikovni opis smiselno izhodiščno analizirati podatke iz različnih gradivskih virov, nato pa jih sistematično interpretirati in prikazati v skladu s konkretnim slovarskim konceptom.

2 Dostopno na https://viri.cjvt.si/gigafida/System/About.

V tujini specializirani korpusi, ki združujejo več strokovnih področij, nastajajo zlasti v kontekstu terminologije (terminografije) in prevodoslovja (zato so ti specializirani korpusi pogosto tudi večjezični). Eden zgodnejših projektov je bila izdelava specializiranega korpusa s strokovnimi besedili s področja prava, ekonomije, medicine, varstva okolja in informatike, ki je nastal v 2. polovici 90. let 20. stoletja na Univerzitetnem inštitutu za uporabno jezikoslovje na Univerzi Pompeu Fabra v Barceloni. Ta večjezični korpus je bil razvit kot podpora raziskovalnim usmeritvam Inštituta in je bil namenjen raziskovanju strokovnega jezika ter razvoju jezikovnotehnoloških aplikacij (Bach idr. 1997).[3] Za pripravo terminološke podatkovne baze EcoLexicon je bil na primer izdelan korpus z besedili različnih znanosti o okolju: biologije, meteorologije, ekologije, prave varstva okolja itd. (León Araúz idr. 2018). Pregled korpusne infrastrukture na repozitorijih evropske raziskovalne infrastrukture Clarin[4] kaže, da večina strokovnih in znanstvenih korpusov evropskih jezikov vključuje besedila, nastala na prehodu v novo tisočletje. Praviloma gre za jezikovne vire, ki vključujejo zlasti zaključna dela na različnih nivojih visokošolskega izobraževanja.

Podobna je situacija za slovenščino, saj so že pred pripravo *Korpusa znanstvenih besedil sodobne slovenščine* obstajali obsežnejši specializirani korpusi strokovnih in znanstvenih besedil, npr. korpus akademske slovenščine KAS (Erjavec idr. 2016) in korpus znanstvenih objav s portala Openscience.si OSS (Žagar idr. 2023).[5] Obstajajo tudi specializirani korpusi, nastali v okviru terminografskih projektov (npr. korpus besedil odnosov z javnostmi KoRP (Logar 2013), korpus turističnih besedil Turk (Mikolič 2013), korpus besedil s področja bibliotekarstva (Kanič 2011), specializirani korpusi, ki nastajajo na Oddelku za terminologijo ISJFR). Praviloma zajemajo strokovna besedila enega področja, besedila v njih pa so lahko besedilnovrstno in celo zvrstno precej heterogena (npr. korpusi vključujejo znanstvene in strokovne članke, magistrska in doktorska dela, lahko pa tudi brošure, publicistična in oglaševalska besedila). Težava je, da se večina od njih gradivsko ne posodablja (redno). Specializirani korpusi, ki zajemajo veliko količino besedil, nastalih v pedagoškem procesu, predvsem diplomskih in magistrskih del, in vsebujejo manjši delež znanstvenih in strokovnih člankov ter znanstvenih in strokovnih monografij, so za leksikografsko in terminološko delo ter pri jezikovni analizi sicer uporabni, vendar pa bi bila njihova uporabnost še bistveno večja, če bi vključevali večji delež besedil osrednjih strokovnih besedilnih vrst. Novi korpus omenjeni primanjkljaj zmanjšuje.

---

3 Približno v tem času so se po vzoru načel za gradnjo referenčnih korpusov začela oblikovati tudi načela za gradnjo specializiranih korpusov (gl. Pearson 1998; Bowker – Pearson 2002).

4 Dostopno na https://www.clarin.eu/resource-families/corpora-academic-texts.

5 Dostopno na https://www.clarin.si/ske/#open.

Izdelava novega korpusa znanstvenih besedil je torej za slovensko jezikovno infrastrukturo velikega pomena. Je namreč ključen dopolnilni gradivski vir za izdelavo novejših splošnih razlagalnih slovarjev, ki pomembno dopolnjuje že obstoječo korpusno infrastrukturo. Zaradi ustrezne metajezikovne dokumentiranosti omogoča tudi izdelavo podkorpusov, ki bodo pripomogli k naprednemu raziskovanju jezika posameznih strok in strokovne slovenščine sploh. Korpus je zamišljen tudi kot potencialno izhodišče za oblikovanje specializiranih korpusov, ki se lahko uporabijo kot temelj za pripravo terminoloških slovarjev različnih strokovnih področij.

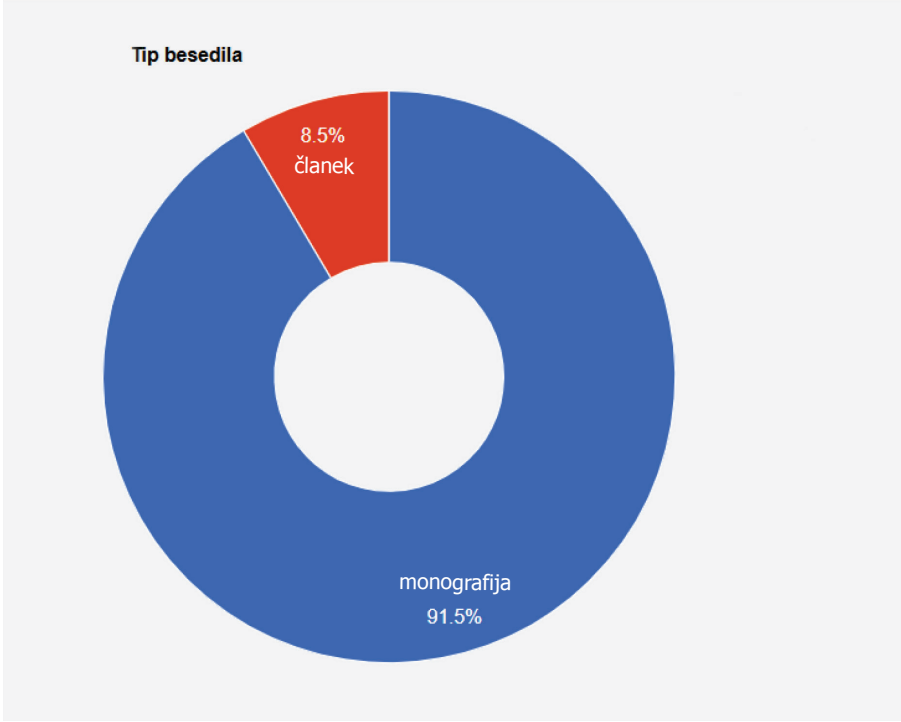
2 O *Korpusu znanstvenih besedil sodobne slovenščine*

2.1 Pridobivanje besedil, besedilnotipska sestava korpusa in njegov obseg

Korpus znanstvenih besedil sodobne slovenščine[6] (Erjavec idr. 2023) je specializirani pisni korpus slovenščine, ki je bil oblikovan med marcem in oktobrom 2023. Ker je bilo za izvedbo aktivnosti na voljo le šest mesecev, smo se morali pri gradnji korpusa obsegovno omejevati, zlasti ker smo, da bi povečali kvaliteto korpusnega vira, veliko pozornosti namenili fazi predpriprave besedil (glej razdelek 2.2).

Korpus je vsebinsko oblikovan v skladu z gradivskimi potrebami zlasti pri leksikografskem in terminografskem opisu slovenske leksike. Vanj je vključenih 884 besedil, in sicer 373 strokovnih in znanstvenih monografij ter 511 strokovnih in znanstvenih člankov v skupnem obsegu 33.604.256 pojavnic. Gre za strokovna in znanstvena besedila zlasti s področja družboslovja in humanistike (npr. jezikoslovje, muzikologija, arheologija, geografija, filozofija, zgodovina, sociologija, migracijske študije), saj je takih besedil, zlasti ko ne gre za zaključna dela na različnih nivojih visokošolskega izobraževanja, v obstoječih primerljivih korpusih nekoliko manj. Monografski del korpusa je bistveno obsežnejši, saj obsega 30.750.533 pojavnic (91,5 %), strokovni in znanstveni članki prinašajo 8,5-% delež vseh pojavnic (2.853.723 pojavnic).

6 Dostopno na https://www.clarin.si/repository/xmlui/handle/11356/1872.

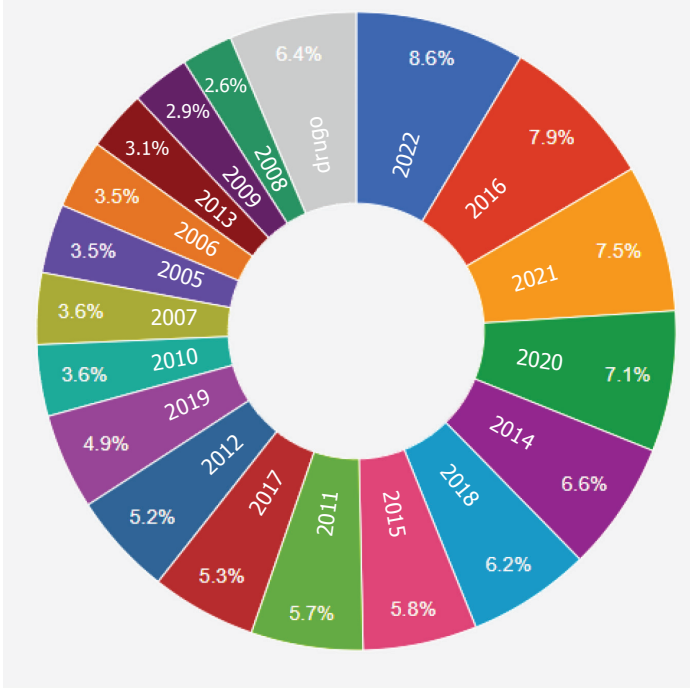


Slika 1: Razmerje med deležem monografskih besedil in deležem besedil znanstvene periodike v korpusu glede na število pojavnic[7]

Korpus vključuje besedila, nastala zlasti med letoma 2000 in 2023. Le 1,7 % korpusnih besedil je nastalo pred letom 2000, pri čemer je eno besedilo – najstarejše – iz leta 1983, preostalih 14 besedil, objavljenih pred letom 2000, pa je s konca 90. let 20. stoletja. 24,2 % korpusnih besedil je nastalo v letih 2000–2010, preostala besedila (74,1 %) so nastala med letoma 2011 in 2023.[8] Kot kaže slika 2, so besedila z vidika časa nastanka v zadnjih 25 letih zastopana razmeroma uravnoteženo.

7 Vir slike: https://www.clarin.si/ske/#text-type-analysis?corpname=kzb10&wlminfreq=1&wlicase=1&include_nonwords=1&showresults=1&wlnums=frq&wlattr=text.type.

8 Zgolj kot zanimivost omenimo, da dejstvo, da so najmlajša korpusna besedila iz leta 2023, najverjetneje pomeni, da pri nastajanju teh besedil veliki jezikovni modeli še niso igrali pomembnejše vloge. Morda ločnica jezikovno za zdaj ni zelo pomembna, utegne pa razlikovanje postati relevantno v prihodnosti.



Slika 2: **Razmerja med deleži besedil v korpusu, nastalimi v posameznem letu**[9]

Glede na kratek čas za izvedbo aktivnosti smo se odločili za pridobivanje besedil, ki v avtorskopravnem smislu nimajo omejitev glede vključitve v odprto dostopne podatkovne zbirke.[10] Da bi delo v tehnološkem smislu poenostavili (uporaba enotnega formata in podobnega oblikovanja besedil v okviru posamezne zbirke, periodične publikacije ipd.), smo za korpus izbrali besedila le enega založnika. Korpus zato vključuje monografije, ki so bile izdane pri Založbi ZRC, oz. članke iz periodičnih publikacij, ki jih (so)izdajajo inštituti ZRC SAZU.

Pri izboru besedil za korpus smo poleg uporabniških potreb (leksikografsko delo, terminografsko delo, raziskave leksike specializiranih področij in jezikoslovne raziskave) zaradi strojnega izvažanja podatkov do določene mere upoštevali tudi stopnjo in natančnost metajezikovne dokumentiranosti besedil (glede na predvideno tipologijo metapodatkov, ki omogoča raziskovanje besedil glede na potencialne podkorpuse korpusa). Izkazalo se je namreč, da so nekatera besedila

9 Vir podatkov: https://www.clarin.si/ske/#text-type-analysis?corpname=kzb10&tab=basic&filter=containing&wlattr=text.year&wlminfreq=1&include_nonwords=1&itemsPerPage=50&showresults=1&chartitems=18&cols=%5B%22frq%22%5D&showtimelines=0&diaattr=&showtimelineabs=0&timelinesthreshold=5&wlsort=frq.

10 Rezultati so morali biti zaradi zahtev financerja odprto dostopni.

metajezikovno dokumentirana pomanjkljivo (glej razdelek 2.2.2). Besedila v formatu .pdf smo po izboru strojno izvozili s platform za brezplačno objavljanje strokovnih recenziranih revij in monografij (Open Journal Systems – OJS,[11] Open Monograph Press – OMP[12]) Založbe ZRC skupaj z relevantnimi metapodatki, ki so bili na voljo v formatih .xml in .html.

2.2 Tehnična predpriprava besedil, jezikoslovno označevanje korpusa in njegova metadokumentacija

Besedila, ki so bila strojno izvožena s platform Open Journal Systems in Open Monograph Press Založbe ZRC, so bila v naslednji fazi dela iz formata .pdf strojno pretvorjena v golobesedilni format .txt s kodiranjem UTF-8. V največji možni meri so bili iz izvoženih besedil najprej strojno odstranjeni strukturni deli, ki bi v korpusu lahko predstavljali korpusni šum (npr. kolofoni, kazala, seznami uporabljene strokovne literature ipd.). Razvit je bil tudi strokovnim in znanstvenim besedilom prilagojen protokol za strojno pretvorbo besedil, ki bo zagotavljal čim enostavnejšo in čim učinkovitejšo pretvorbo besedil iz formata .pdf v format .txt pri nadaljnjih nadgradnjah korpusa. Da bi zagotovili kar najvišjo stopnjo natančnosti strojnega jezikoslovnega označevanja korpusa, so bila avtomatsko pretvorjena besedila po strojnem pregledu še dodatno ročno pregledana, napake pri strojni pretvorbi pa popravljene. Ročno so bili v celoti pregledani in popravljeni vsi znanstveni in strokovni članki v korpusu ter približno tretjina monografij. Preostali dve tretjini monografij sta bili ročno pregledani le deloma oz. z manj natančnim pregledom.

Osnovni namen ročnega pregleda in popravljanja je bil, da se iz besedil odstrani dele, ki so z vidika gradnje korpusa manj funkcionalni, torej ne predstavljajo »besedila« v smislu sklenjenega niza (vsaj nekaj) povedi, ampak so del strukturnih delov člankov in monografij, ki so za (jezikoslovno) analizo in interpretacijo podatkov manj relevantni, po drugi strani pa lahko predstavljajo težave pri strojnem označevanju besedil, pri katerem je zaradi pomanjkanja konteksta možnost napak velika. Ročni pregled so opravljali študenti, ki so pregledali vsako stran vsakega besedila in izločali vnaprej določene tipe strukturnih enot, ki so v besedilu ostale po strojni pretvorbi in strojnem odstranjevanju delov, ki bi lahko predstavljali korpusni šum. Sistematično so bili npr. izločeni zapisi v glavi in nogi, številke strani, tujejezični naslovi besedil (kot prevodi slovenskih), nazivi avtorjev, njihovi kontaktni podatki, seznami recenzentov, tujejezični izvlečki in povzetki, daljši, v besedilo neintegrirani deli besedila, zapisani z nelatinično pisavo, kolofoni, seznami literature na koncu prispevkov, besedilo opomb, če je vključevalo le navedbo spletnih strani ali bibliografske podatke, kataložne navedbe,

11 Dostopno na https://ojs.zrc-sazu.si/.
12 Dostopno na https://omp.zrc-sazu.si/.

popisi zgolj naštevalne narave ipd. Pri ročnem pregledu so bile popravljene tudi napake pri kodiranju znakov, ki so se pojavljale le lokalno,[13] napake pri segmentiranju besedila na odstavke in povedi itd. – vse z namenom, da se poveča kvaliteta korpusnega vira.

2.2.1 Jezikoslovno označevanje korpusa

Strojno in ročno pregledana ter urejena besedila, zapisana v golobesedilnem formatu .txt, so bila v naslednji fazi dela segmentirana na odstavke in povedi, tokenizirana, lematizirana ter jezikoslovno označena. Za označevanje smo uporabili orodje CLASSLA[14] (Ljubešić idr. 2024), različico 1.2.0. Korpus je bil jezikoslovno označen po uveljavljenih standardih: oblikoskladenjsko je označen po modelu MULTEXT-East v6,[15] skladenjsko je označen glede na smernice za skladenjsko označevanje modela JOS,[16] na voljo so tudi jezikoslovne oznake imenskih entitet,[17] zato je z vidika označevanja primerljiv z drugimi korpusi in omogoča vključitev v obširnejše korpusne podatkovne zbirke, npr. v korpus metaFida.[18]

2.2.2 Metadokumentiranje korpusa

V korpus vključena besedila so dokumentirana z naslednjimi tipi metapodatkov:

- avtor,
- naslov,
- leto objave,
- založnik,
- tip,
- vir,
- strokovno področje,
- ključne besede,
- DOI/URL,
- identifikacijska številka.

Poleg navedenih tipov metapodatkov interna dokumentacija vključuje še nekaj dodatnih metapodatkovnih tipov, ki so oblikovalcem korpusa v pomoč pri njegovih nadgradnjah, npr. metapodatek o tem, v kateri različici korpusa je bilo besedilo dodano v korpus, v javno objavljeni dokumentaciji pa ti metapodatki niso na voljo.

Podatki so bili v podatkovno zbirko metapodatkov večinoma uvoženi strojno, so pa bili ročno preverjeni, deloma tudi ročno dopolnjeni (npr. pri atributih

---

13 Tovrstne napake bolj sistemske narave so bile odpravljene strojno.
14 Dostopno na https://github.com/clarinsi/classla/.
15 Dostopno na https://nl.ijs.si/ME/V6/msd/html/msd-sl.html.
16 Dostopno na https://nl.ijs.si/jos/bib/jos-skladnja-navodila.pdf.
17 Dostopno na https://nl.ijs.si/janes/wp-content/uploads/2017/09/SlovenianNER-eng-v1.1.pdf.
18 Dostopno na https://www.clarin.si/repository/xmlui/handle/11356/1775.

strokovno področje, tip). Obvezni metapodatki, ki so vključeni v dokumentacijo vseh besedil, so: avtor, naslov, leto objave, založnik, tip (možni vrednosti atributa sta *monografija* in *članek*), identifikacijska številka ter DOI/URL (podatek uporabnika vodi do mesta, kjer je besedilo objavljeno v formatu .pdf). Neobvezni metapodatki prinašajo informacije o viru (metapodatek pojasnjuje, v kateri periodični publikaciji je bil strokovni ali znanstveni članek iz korpusa objavljen, vrednosti atributa so zato npr. *Jezikoslovni zapiski*, *Arheološki vestnik*, *De musica disserenda*, *Slovenski jezik*) in o strokovnem področju. Zgolj izjemoma pri kakem od člankov niso navedene ključne besede.

Ena od težav, s katerimi smo se soočali pri pridobivanju metapodatkov za korpusna besedila, je bila, da metapodatki na izvornih platformah OJS in OMP niso bili vedno ustrezno navedeni.[19] To pomeni, da so lahko bili ali pomanjkljivi (npr. niso bili navedeni pri vseh običajno rabljenih podatkovnih tipih) ali napačni (npr. podatek o jeziku, v katerem naj bi bil napisan članek, ni odražal realnega stanja). V nekaterih primerih so bili uporabljeni podatkovni tipi nerazlikovalni v smislu, da je bil isti podatkovni tip (npr. <Title>, <Subject and Keywords>, <Description>) uporabljen za vnos vsebinsko istih podatkov v slovenskem in kakem tujem jeziku, pri čemer ni bilo razvidno, kje se konča navedba podatkov v enem jeziku in od kod naprej so istovrstni podatki (npr. ključne besede) navedeni v tujem jeziku. Pri podatkovnih tipih, ki vsebujejo podatke v obliki povedi oz. krajših besedil, npr. izvleček besedila, naslov besedila, dilem praviloma ni bilo, pojavljale so se zlasti pri določanju jezika ključnih besed. Za razdvoumljanje smo si zato pomagali s posebno hevristiko, in sicer smo na podlagi preverjanja, ali se ključna beseda pojavlja v slovenskih oz. tujejezičnih leksikalnopodatkovnih zbirkah, ugibali, za kateri jezik najverjetneje gre, hkrati pa smo preverjali tudi, ali se morebiti prevodni ustreznik iskane besede (tako v slovenskem kot v tujem jeziku) že pojavlja v nizu navedenih ključnih besed. Z opisanim mehanizmom smo večino dilem glede uporabljenega jezika lahko razrešili strojno, v nekaterih primerih pa je bilo potrebno ročno razdvoumljanje ključnih besed glede na jezik. Vsi vnosi ključnih besed so bili po končanem postopku strojnega razdvoumljanja jezika še ročno preverjeni.

Posebno zahtevno nalogo je predstavljala tudi metadokumentacija besedil z vidika strokovnega področja, na katero sodijo. V mnogih primerih dilem ni bilo. Zlasti pri besedilih, v katerih so bili opisani rezultati meddisciplinarnega raziskovalnega dela, pa področja mnogokrat ni bilo mogoče povsem nedvoumno določiti. V teh primerih podatka o strokovnem področju v metadokumentacijo (zaenkrat) nismo vnesli. Podobno velja npr. za zbornike in druge publikacije, ki so vključevale besedila z različnih strokovnih področij, saj so te bibliografske enote v korpusu obravnavane kot ena monografska publikacija. Tipologija uporabljenih oznak za

19 Zdi se, da se vnašalci podatkov ne zavedajo nujno, da so metapodatki lahko uporabljeni za različne namene in da je pravilen vnos zato zelo pomemben.

strokovna področja se sicer ujema s tipologijo, ki je bila za označevanje strokovnih področij razvita pri klasifikaciji področij na Slovenskem terminološkem portalu[20] (Jemec Tomazin 2024; Jemec Tomazin – Romih 2023) in temelji na dveh uveljavljenih klasifikacijah, in sicer Evropski klasifikaciji raziskovalne dejavnosti CERIF[21] ter klasifikaciji Eurovoc.[22] Na Slovenskem terminološkem portalu je objavljena tabela, ki prikazuje pretvorbo med navedenimi klasifikacijami.[23]

2.3 Zapis korpusa in njegova dostopnost

Avtomatsko jezikoslovno označena korpusna besedila so bila v naslednji fazi priprave korpusa zapisana v skladu z najaktualnejšimi standardi za zapis korpusnih podatkovnih zbirk. Korpus je na voljo v dveh oblikah: kot odprto dostopna podatkovna zbirka z licenco CC BY-SA 4.0, zapisana v formatu CoNLL-U[24] in z metapodatki dokumentirana v datoteki formata .tsv, ter v t. i. vertikalnem formatu .vert, potrebnem za vključitev korpusa v napredne konkordančnike tipa Sketch Engine (Rychlý 2007; Kilgarriff idr. 2014) in KonText (Machálek 2020). Zaradi podrobne dokumentiranosti korpusa in premišljene vključitve korpusnih metapodatkov korpus omogoča sistematične in celovite jezikoslovne raziskave in izrabo v različne namene. Uporabnikom je dostopen prek raziskovalne infrastrukture Clarin.si (Erjavec idr. 2014; Erjavec idr. 2024), in sicer tako za prenos podatkov[25] kot prek konkordančnikov NoSketch Engine[26] in KonText.[27]

3 Uporabnost korpusa in načrti za njegovo nadgradnjo

Korpus znanstvenih besedil sodobne slovenščine zmanjšuje primanjkljaj na področju opremljenosti slovenske korpusne infrastrukture s strokovnimi in znanstvenimi besedili, zlasti takimi, ki niso zaključna dela na različnih stopnjah študija, pri čemer zagotavlja (boljšo) gradivsko podporo tako za oblikovanje temeljnega enojezičnega razlagalnega slovarja slovenščine eSSKJ kot za jezikoslovne raziskave strokovnega jezika. Hkrati je lahko izhodišče za oblikovanje specifičnih podkorpusov za potrebe oblikovanja terminoloških slovarjev posameznih strokovnih področij. Obstoječi referenčni korpusi namreč ravno na področju stvarne literature nudijo zgolj pomanjkljivo gradivsko podporo leksikografskemu delu (kot rečeno, najnovejši korpus Gigafida 2.0 vključuje manj kot 4 % stvarnih besedil, po drugi strani pa skupno

20 Dostopno na https://terminoloski.slovenscina.eu/.
21 Dostopno na https://www.arrs.si/sl/gradivo/sifranti/sif-cerif-cercs.asp.
22 Dostopno na https://eur-lex.europa.eu/browse/eurovoc.html?locale=sl.
23 Dostopno na https://terminoloski.slovenscina.eu/pomoc.
24 Dostopno na https://universaldependencies.org/format.html.
25 Dostopno na http://hdl.handle.net/11356/1872.
26 Dostopno na https://www.clarin.si/ske/#dashboard?corpname=kzb10.
27 Dostopno na https://www.clarin.si/kontext/query?corpname=kzb10.

približno 65 % časopisnih in revijalnih besedil ter 28 % besedil s spleta). Novi korpus bo zato neposredno prispeval h kvalitetnejšemu leksikalnemu opisu knjižne slovenščine. Kot izhodišče za oblikovanje specializiranih korpusov je primeren tudi za pripravo terminoloških slovarjev za posamezna strokovna področja.

Ob tem je posebno pomembno poudariti, da korpus vključuje sodobna slovenska strokovna in znanstvena besedila, nastala zlasti po letu 2010 (takih korpusnih besedil je v podatkovni bazi približno 75 %), pri čemer v prihodnje načrtujemo tudi njegove nadgradnje. Prednost korpusa glede na druge primerljive korpuse je tudi to, da vključuje znanstvene in strokovne monografije ter znanstvene in strokovne članke uveljavljenih strokovnjakov – večino podobnih obstoječih podatkovnih zbirk namreč v največji meri sestavljajo diplomska, magistrska in doktorska dela, torej zaključna dela avtorjev, ki strokovnjaki za posamezna področja bodisi šele postajajo ali so praviloma na začetku karierne poti. Za primerjavo povejmo, da v korpusu OSS kot najbolj primerljivem korpusu skupni delež takih besedil presega 80 %, znanstvenih monografij je v korpusu 0,2 %, izvirnih in preglednih znanstvenih člankov pa skupaj približno 10 %.[28]

Da bi lahko korpus še naprej služil svojemu načrtovanemu namenu, v prihodnosti načrtujemo še naslednje korake:

(1) Gradivske nadgradnje korpusa, ki bodo zagotavljale, da bodo za leksikografsko in terminografsko delo ter jezikoslovne analize na voljo besedilnovrstno raznolika najsodobnejša strokovna in znanstvena besedila. Načrtujemo razširitev korpusa z besedili drugih strokovnih področij in tudi izvoz besedil z drugih platform za brezplačno objavljanje strokovnih recenziranih revij in monografij. Besedila bodo morala biti tudi v prihodnje natančno metajezikovno dokumentirana, da bodo omogočala oblikovanje podkorpusov in s tem napredne jezikoslovne analize.

(2) Označevanje korpusnih besedil v skladu z najnovejšimi standardi jezikoslovnega označevanja korpusov (kar po potrebi vključuje tudi pretvorbo oznak oz. novo označevanje obstoječega korpusa glede na uveljavljene novejše standarde označevanja), pri čemer bodo morali biti avtomatski protokoli strojne pretvorbe v golobesedilni format in postopki jezikoslovnega označevanja besedil redno posodabljani.

(3) Zagotavljanje odprte dostopnosti nadgrajenega korpusnega vira, ki bo omogočala izmenljivost podatkov in vključitev korpusa v druge korpusne podatkovne zbirke, pri čemer bo treba zagotavljati tudi ustrezen zapis korpusa v skladu z morebitnimi novimi standardi zapisa korpusnih podatkov.

---

28 Dostopno na https://www.clarin.si/ske/#text-type-analysis?corpname=oss10&tab=basic&filter=containing&wlattr=text.type&wlminfreq=1&include_nonwords=1&itemsPerPage=50&showresults=1&chartitems=20&cols=%5B%22frq%22%5D&showtimelines=0&diaattr=&showtimelineabs=0&timelinesthreshold=5&wlsort=frq.

4 Zaključek

Korpus znanstvenih besedil sodobne slovenščine je specializirani pisni korpus slovenščine, ki obsega 884 besedil s področij družboslovja in humanistike, nastalih predvsem med letoma 2000 in 2023, in sicer 373 strokovnih in znanstvenih monografij ter 511 strokovnih in znanstvenih člankov v skupnem obsegu 33.604.256 pojavnic. Korpus je vsebinsko oblikovan v skladu z gradivskimi potrebami zlasti pri leksikografskem in terminografskem opisu slovenske leksike. Obstoječi splošni (referenčni) korpusi za slovenščino namreč vsebujejo (pre)malo specializiranih (strokovnih) besedil, specializirani korpusi pa pogosto premalo monografskih publikacij ter znanstvenih in strokovnih člankov. Pri pripravi korpusa je bila posebna pozornost posvečena opremljanju besedil z ustreznimi metapodatki (npr. podatek o avtorju besedila, letu objave, strokovnem področju besedila) ter tehnični predpripravi besedil (iz njih so bili najprej strojno odstranjeni tisti deli, ki bi lahko predstavljali korpusni šum in praviloma niso zanimivi za jezikoslovno analizo, npr. kolofoni, besedila pa so bila nato še ročno pregledana). Besedila so bila nato segmentirana na povedi, tokenizirana, lematizirana ter jezikoslovno označena. Korpus je dostopen kot odprto dostopna podatkovna zbirka z licenco CC BY-SA 4.0 v formatih CoNLL-U in .vert ter prek konkordančnikov noSketch Engine in KonText na repozitoriju Clarin.si.

V prihodnosti načrtujemo nadgradnjo korpusa z dodajanjem novih strokovnih in znanstvenih besedil z različnih strokovnih področij. Tudi pri nadgradnjah bo treba skrbno pripraviti korpusne metapodatke in upoštevati najnovejše standarde jezikoslovnega označevanja. *Korpus znanstvenih besedil sodobne slovenščine* bo tako ponujal verodostojne podatke o rabi slovenskega jezika v strokovnih in znanstvenih besedilih ter omogočal izdelavo zanesljivih leksikografskih in terminografskih opisov slovenščine.

Literatura

Bach idr. 1997 = Carme Bach – Roser Saurí Colomer – Jorge Vivaldi – Maria Teresa Cabré, *El corpus de l'IULA: descripció*, Barcelona: Universitat Pompeu Fabra, Institut Universitari de Lingüística Aplicada, 1997, http://hdl.handle.net/10230/1299.

Bowker – Pearson 2002 = Lynne Bowker – Jennifer Pearson, *Working with specialized language: a practical guide to using corpora*, London – New York: Routledge, 2002.

Erjavec idr. 2014 = Tomaž Erjavec – Jan Jona Javoršek – Simon Krek, Raziskovalna infrastruktura CLARIN.SI, v: *Zbornik 17. mednarodne multikonference Informacijska družba – IS 2014: zvezek G*, ur. Tomaž Erjavec – Jerneja Žganec Gros, Ljubljana: Institut Jožef Stefan, 2014, 19–24.

Erjavec idr. 2016 = Tomaž Erjavec – Darja Fišer – Nikola Ljubešić – Nataša Logar – Milan Ojsteršek, Slovenska akademska besedila: prototipni korpus in načrt analiz, v: *Zbornik konference Jezikovne tehnologije in digitalna humanistika, 29. september–1. oktober 2016*, ur. Tomaž Erjavec – Darja Fišer, Ljubljana: Znanstvena založba Filozofske fakultete, 2016, 58–64.

Erjavec idr. 2023 = Tomaž Erjavec – Mateja Jemec Tomazin – Nina Ledinek – Andrej Perdih – Miro Romih – Mitja Trojar – Luka Romih, Corpus of scientific texts of contemporary Slovenian KZB 1.0, *Slovenian language resource repository CLARIN.SI*, 2023, http://hdl.handle.net/11356/1872.

Erjavec idr. 2024 = Tomaž Erjavec – Nikola Ljubešić – Katja Meden – Taja Kuzman – Cyprian Adam Laskowski – Jan Jona Javoršek – Simon Krek – Mateja Jemec Tomazin – Jakob Lenardič, CLARIN.SI, the Slovenian node of CLARIN: ten years on, v: *CLARIN Annual Conference Proceedings*, ur. Vincent Vandeghinste – Thalassia Kontino, Barcelona, 2024, 76–80.

Jemec Tomazin 2024 = Mateja Jemec Tomazin, Slovenski terminološki portal, v: *Zbornik konference Jezikovne tehnologije in digitalna humanistika: 19.–20. september 2024*, ur. Špela Arhar Holdt – Tomaž Erjavec, Ljubljana: Inštitut za novejšo zgodovino, 2024, 546–556.

Jemec Tomazin – Romih 2023 = Mateja Jemec Tomazin – Miro Romih, Slovenski terminološki portal: nova priložnost za urejanje slovenske terminologije, v: *Razvoj slovenščine v digitalnem okolju*, ur. Špela Arhar Holdt – Simon Krek, Ljubljana: Založba Univerze, 2023, 211–247, https://ebooks.uni-lj.si/ZalozbaUL/catalog/view/522/852/9443.

Kanič 2011 = Ivan Kanič, Slovenski besedilni korpus bibliotekarstva – najsodobnejša slovaropisna podpora bibliotekarski terminologiji, v: *Knjižnica: odprt prostor za dialog in znanje: zbornik referatov, Strokovno posvetovanje Zveze bibliotekarskih društev Slovenije, Maribor, 20.-22. oktober 2011*, ur. Melita Ambrožič – Damjana Vovk, Ljubljana: Zveza bibliotekarskih društev Slovenije, 2011, 297–312.

Kilgarriff idr. 2014 = Adam Kilgarriff – Vít Baisa – Jan Bušta – Miloš Jakubíček – Vojtěch Kovář – Jan Michelfeit – Pavel Rychlý – Vít Suchomel, The Sketch Engine: ten years on, *Lexicography* 1.1 (2014), 7–36, DOI: https://doi.org/10.1007/s40607-014-0009-9.

Krek idr. 2019 = Simon Krek – Špela Arhar Holdt – Jaka Čibej – Andraž Repar – Nikola Ljubešić, *Specifikacije izdelave korpusa Gigafida 2.0, v1.0, 13. 6. 2019*, https://www.cjvt.si/gigafida/wp-content/uploads/sites/10/2019/06/Gigafida2.0_specifikacije.pdf.

Krek idr. 2020 = Simon Krek – Špela Arhar Holdt – Tomaž Erjavec – Jaka Čibej – Andraž Repar – Polona Gantar – Nikola Ljubešić – Iztok Kosem – Kaja Dobrovoljc, Gigafida 2.0: the Reference Corpus of Written Standard Slovene, v: *Proceedings of the Twelfth Language Resources and Evaluation Conference*, ur. Nicoletta Calzolari idr., Marseille: European Language Resources Association, 2020, 3340–3345.

Krvina – Petric Žižić 2024 = Domen Krvina – Špela Petric Žižić, Razmerje med sestavo korpusov (žanrska uravnoteženost in reprezentativnost) in njihovo zanesljivostjo pri izdelavi splošnega razlagalnega slovarja, *Slovenski jezik – Slovene Linguistic Studies* 16 (2024), 149–176, DOI: https://doi.org/10.3986/16.1.07.

León Araúz idr. 2018 = Pilar León Araúz – Antonio San Martín – Arianne Reimerink, The EcoLexicon English Corpus as an open corpus in Sketch Engine, v: *Proceedings of the 18th EURALEX International Congress*, ur. Jaka Čibej idr., Ljubljana: Euralex, 2018, 893–901.

Ljubešić idr. 2024 = Nikola Ljubešić – Luka Terčon – Kaja Dobrovoljc, CLASSLA-Stanza: the next step for linguistic processing of South Slavic languages, v: *Zbornik konference Jezikovne tehnologije in digitalna humanistika, 19.–20. september 2024*, ur. Špela Arhar Holdt – Tomaž Erjavec, Ljubljana: Inštitut za novejšo zgodovino 2024, 251–274.

Logar 2013 = Nataša Logar, *Korpusna terminografija: primer odnosov z javnostmi*, Ljubljana: Trojina, zavod za uporabno slovenistiko – Fakulteta za družbene vede, 2013.

Logar idr. 2020 = Nataša Logar – Miha Grčar – Marko Brakus – Tomaž Erjavec – Špela Arhar Holdt – Simon Krek, *Korpusi slovenskega jezika Gigafida, KRES, ccGigafida in ccKRES: gradnja, vsebina, uporaba*, Ljubljana: Znanstvena založba Filozofske fakultete, 2020, DOI: https://doi.org/10.4312/9789610603542.

Machálek 2020 = Tomáš Machálek, KonText: advanced and flexible corpus query interface, v: *Proceedings of the Twelfth Language Resources and Evaluation Conference*, ur. Nicoletta Calzolari idr., Marseille: European Language Resources Association, 2020, 7003–7008.

Mikolič 2013 = Vesna Mikolič, *Turistični terminološki slovar – predstavitev izhodišč in učinkov projekta ter opis slovarja TURS kot glavnega rezultata projekta*, elaborat, Koper: Univerza na Primorskem, Znanstveno-raziskovalno središče, 2013.

Pearson 1998 = Jennifer Pearson, *Terms in Context*, Amsterdam – Philadelphia: John Benjamins, 1998.

Rychlý 2007 = Pavel Rychlý, Manatee/Bonito – A Modular Corpus Manager, v: *Proceedings of Recent Advances in Slavonic Natural Language Processing, RASLAN 2007*, ur. Petr Sojka – Aleš Horák, Brno: Masarykova univerza, 2007, 65–70.

Žagar idr. 2023 = Kristjan Žagar – Marko Ferme – Milan Ojsteršek – Mateja Jemec Tomazin – Tomaž Erjavec, Corpus of scientific texts from the Open Science Slovenia portal OSS 1.0, *Slovenian language resource repository CLARIN.SI*, 2023, http://hdl.handle.net/11356/1774.

Žagar Karer 2011 = Mojca Žagar Karer, *Terminologija med slovarjem in besedilom: analiza elektrotehniške terminologije*, Ljubljana: Založba ZRC, ZRC SAZU, 2011.

Žagar Karer – Ledinek 2021 = Mojca Žagar Karer – Nina Ledinek, Med terminologijo in splošno leksiko: determinologizacija in z njo povezane slovaropisne ter uporabniške dileme, *Slovenski jezik – Slovene Linguistic Studies* 13 (2021), 41–60, DOI: https://doi.org/10.3986/sjsls.13.1.03.

Summary

The Concept and Design of the *Corpus of Scientific Texts in Contemporary Slovenian*

The *Corpus of Scientific Texts of Contemporary Slovenian* comprises 33,604,256 tokens extracted from 884 scholarly texts, predominantly from the social sciences and humanities, published primarily between 2000 and 2023. The corpus includes 373 published volumes and 511 articles. The majority of the material comes from disciplines such as linguistics, musicology, archaeology, geography, philosophy, history, sociology, and migration studies.

Published volumes constitute the significantly larger portion of the corpus, containing 30,750,533 tokens (91.5%), whereas articles account for 8.5% (2,853,723 tokens). The texts were initially harvested from the Založba ZRC publishing house platforms Open Journal Systems and Open Monograph Press and subsequently converted from PDF format to plain text format (.txt) encoded in UTF-8. Structural components likely to introduce corpus noise—such as tables of contents, bibliographies, lists of abbreviations, and extended foreign-language passages—were removed automatically to the greatest extent possible.

Following manual and automated verification, the texts in plain text format were segmented into sentences, tokenized, lemmatized, and linguistically annotated (morphosyntactic, syntactic, and named entity annotation). Each text in the corpus is accompanied by a rich set of metadata, including author, title, year of publication, publisher, text type, source, field, keywords, DOI/URL, and a unique text identifier. The corpus is available as an open-access data resource under the CC BY-SA 4.0 license and can be accessed via the noSketch Engine and KonText concordancers in the Clarin.si repository.

The *Corpus of Scientific Texts of Contemporary Slovenian* addresses the notable gap in corpora of scholarly texts, particularly in the social sciences and humanities. It provides a robust foundational resource for the development of a comprehensive monolingual explanatory dictionary of Slovenian, as well as for linguistic research on specialized language.