Urednika:
ŠPELA ARHAR HOLDT
SIMON KREK

RAZVOJ SLOVENŠČINE V DIGITALNEM OKOLJU

Univerza *v Ljubljani*

Kataložni zapis o publikaciji (CIP) pripravili v
Narodni in univerzitetni knjižnici v Ljubljani
COBISS.SI-ID 182678275
ISBN 978-961-297-256-1 (PDF)

www.slovenščina.eu
sporazumevanje

Razvoj slovenščine v digitalnem okolju

Urednika:

Špela Arhar Holdt in Simon Krek

Univerza v Ljubljani
Filozofska fakulteta

Razvoj slovenščine v digitalnem okolju

Zbirka: Sporazumevanje (e-ISSN 2738-4527)
Urednika zbirke: Špela Arhar Holdt, Vojko Gorjanc
Urednika publikacije: Špela Arhar Holdt, Simon Krek
Recenzenta: Monika Kalin Golob, Simon Šuster
Tehnično urejanje: Jure Preglau
Prelom: Aleš Cimprič
Oblikovanje naslovnice: Kofein dizajn
Založila: Založba Univerze v Ljubljani
Izdala: Znanstvena založba Filozofske fakultete Univerze v Ljubljani
Za založbo: Gregor Majdič, rektor Univerze v Ljubljani
Za izdajatelja: Mojca Schlamberger Brezar, dekanja Filozofske fakultete UL

Ljubljana, 2023
Prva izdaja, e-izdaja
Publikacija je brezplačna.
Publikacija je dostopna na: https://ebooks.uni-lj.si/ZalozbaUL

Projekt Razvoj slovenščine v digitalnem okolju sta med leti 2020 in 2023 sofinancirali Republika Slovenija in Evropska unija iz Evropskega sklada za regionalni razvoj. Pripravo posameznih prispevkov je finančno podprla Javna agencija za znanstvenoraziskovalno in inovacijsko dejavnost Republike Slovenije (ARIS) preko raziskovalnega programa P6-0411 in projektov J7-4642, J6-2581, J7-3159 in CRP V5-2297.

Kazalo vsebine

Uvodnik

Strateško načrtovan in neprekinjen razvoj jezikovnih virov, tehnologij in storitev je ključnega pomena za vsak jezik oz. jezikovno skupnost – je temeljni pogoj, da se lahko posameznice in posamezniki nemoteno vključujemo v nove načine komunikacije, dela in preživljanja prostega časa v sodobni družbi. Za slovenščino sta premišljenost in usklajenost še toliko pomembnejši, saj je razvojna naloga enaka, raziskovalno-razvojna skupnost, ki se ji posveča, pa manjša kot pri jezikih z več govorci.

Načrtovanje digitalne infrastrukture za sodobno slovenščino v tem trenutku še ni optimalno, pozitivno pa je, da se problematiki na nacionalni ravni posveča vedno več pozornosti. Primer dobre prakse raziskovalno-razvojnega projekta, ki je povezal deležnike v slovenskem prostoru in združil znanja različnih raziskovalnih inštitucij ter jezikovnotehnoloških podjetij, je Razvoj slovenščine v digitalnem okolju (RSDO), ki sta ga med leti 2020 in 2023 financirala Ministrstvo za kulturo Republike Slovenije in Evropski sklad za regionalni razvoj.

Na projektu smo odpravili nekatere pereče vrzeli na področju odprto dostopnih virov, tehnologij in storitev za sodobno slovenščino. Prenovili smo učne množice in postopke za strojno označevanje sodobne slovenščine, osvežili in povečali temeljne jezikovne vire in nadgradili metodologijo za njihovo bodočo gradnjo. Velik del projekta je bil posvečen razvoju govorne baze in govornih tehnologij, zlasti razpoznave govora za slovenščino, ter semantičnim virom in tehnologijam, kjer so bile aktivnosti izdelava osrednje digitalne slovarske baze, baze znanja ter virov in postopkov za različne semantične naloge. Nadgradili smo metodologijo strojnega prevajanja in zasnovali ter vzpostavili portal za urejanje slovenske terminologije.

Monografija, ki je pred vami, vključuje osem poglavij, ki jih je pripravilo dvajset avtorjev in avtoric s petih različnih inštitucij: Univerze

v Ljubljani (Filozofska fakulteta in Fakulteta za računalništvo in informatiko), Univerze v Mariboru (Fakulteta za elektrotehniko, računalništvo in informatiko), ZRC SAZU (Inštitut za slovenski jezik Frana Ramovša), Inštituta »Jožef Stefan« in jezikovnotehnološkega podjetja Amebis, d. o. o., Kamnik. Prva prispevka se posvečata pripravi virov za razvoj govornih tehnologij: **Darinka Verdonik** se osredotoči na prihodnji razvoj govornih korpusov, zlasti z vidika njihove karseda učinkovite gradnje za različne uporabniške potrebe, **Mitja Trojar** in **Andreja Bizjak** pa predstavita načela za zapis govora in izvedbo transkribiranja pri izdelavi govorne baze Artur. Sledita poglavji, ki se ukvarjata s korpusi, ki vsebujejo jezikovne popravke in so zlasti pomembni za področje jezikovnega izobraževanja: **Špela Arhar Holdt, Eva Pori** in **Iztok Kosem** predstavijo strategijo za prihodnost korpusa Šolar, ki vsebuje besedila osnovnošolskih in srednješolskih učencev; **Mojca Stritar Kučuk** pa prvi korpus slovenščine kot drugega oz. tujega jezika KOST 1.0. Zadnji predstavljeni jezikovni vir je slovenski učni korpus SUK, katerega kompleksno sestavo in ročno pregledane jezikoslovne oznake predstavljajo **Špela Arhar Holdt, Jaka Čibej, Kaja Dobrovoljc, Tomaž Erjavec, Polona Gantar, Simon Krek, Tina Munda, Nejc Robida, Luka Terčon** in **Slavko Žitnik. Slavko Žitnik** je tudi avtor prvega od dveh poglavij, ki se posvečata orodjem za obdelavo naravnega jezika – predstavi ogrodje za demokratizacijo obdelave naravnega jezika ANGLEr, vključno s podatkovnim modelom, **Aleš Žagar** in **Marko Robnik-Šikonja** pa meta-povzemalnik, ki izbira med štirimi različnimi modeli povzemanja, razvitimi za slovenščino. Monografijo zaključita **Mateja Jemec Tomazin** in **Miro Romih** z opisom Slovenskega terminološkega portala, ki ponuja uporabniku prijazen načine za urejanje slovenske terminologije.

Monografija je namenjena študentom in študentkam, predavateljem in predavateljicam, raziskovalcem in raziskovalkam, razvijalcem in razvijalkam ter vsem, ki bi radi bolje razumeli namen, sestavo in način gradnje predstavljenih projektnih rezultatov. Pomemben doprinos dela je, da pogled usmerja v prihodnost in opredeljuje korake, ki so pred nami. Zato bo uporabno branje tudi za pripravljavce in pripravljavke nacionalnih razvojnih strategij in druge področne

odločevalce in odločevalke. Urednika se prijazno zahvaljujeva vsem sodelujočim za kakovostne prispevke, recenzentoma **Moniki Kalin Golob** in **Simonu Šustru** pa za hitro branje in konstruktivne komentarje. Lepo vabljeni k branju!

Špela Arhar Holdt in Simon Krek

Zbiranje gradiv za govorne korpuse med Scilo in Karibdo

Darinka VERDONIK
Univerza v Mariboru, Fakulteta za elektrotehniko, računalništvo in informatiko

Povzetek

Govorni korpusi niso pomembni samo za tehnološki razvoj, ampak tudi za sodobno jezikoslovje. Ker zahtevajo velik časovni vložek, mora biti njihovo načrtovanje toliko bolj premišljeno. V prispevku se osredotočamo na prihodnji razvoj govornih korpusov in iščemo odgovor na vprašanji: Kdo so uporabniki govornih korpusov in kakšne so njihove potrebe po gradivih? Katere so prakse zbiranja gradiv za govorne korpuse in kako lahko sinergično naslovimo čim več različnih potreb z enotnim virom? Med bolj aktivnimi uporabniki govornih korpusov so mnoge jezikoslovne discipline kot tudi govorne in semantične tehnologije. V obstoječih slovenskih govornih korpusih že obstaja večja količina gradiv za medijski, parlamentarni in akademski govor, manjka pa avtentičnih vsakdanjih govornih interakcij, kjer bi bila potrebna bolj podrobna regionalna pokritost, visoka kvaliteta posnetkov in zajem videa, kjer je mogoče. V Sloveniji je problem nekontinuirano snemanje v izredno kratkih časovnih obdobjih, pri čemer se veliko sredstev izgublja za koordiniranje množice sodelavcev ter ni časa za podrobno načrtovanje in pripravo orodij za bolj učinkovito delo.

Ključne besede: govorni viri, razpoznavanje govora, snemanje, uporabniki

Abstract

Speech corpora are important for technological development and for modern linguistics. They require a large investment of time, therefore their planning must be all the more thoughtful. In this paper, we focus on the future development of Slovenian speech corpora and seek answers to the following questions: Who are the users of speech corpora and what are their needs for data? What are the practices of collecting data for speech

corpora and how can we synergistically address as many different needs as possible with a single source? Among the more active users of speech corpora are many linguistic disciplines as well as speech and semantic technologies. In the existing Slovenian speech corpora, there is already a large amount of media, parliamentary and academic speech. There is a lack of authentic everyday speech interactions, which would require more detailed regional coverage, high quality recordings and video capture where possible. In Slovenia, the problem is non-continuous recording in extremely short periods of time where a lot of resources are wasted on coordinating a multitude of collaborators while there is no time for detailed planning and preparation of tools for more efficient work.

Keywords: speech resources, speech recognition, recording, users

1 Uvod

Slovenščina se po jezikovnotehnološki podprtosti uvršča na rep držav s fragmentarno tehnološko podporo. Z vidika pripravljenosti na digitalno prihodnost je primerljiva z bolgarskim, slovaškim, hrvaškim, baskovskim, velškim, galicijskim in islandskim jezikom (Giagkou idr., 2023: 81). Na tehnološko podprtost jezikov seveda vplivajo razni socioekonomski in politični dejavniki in razumljivo je, da se po podprtosti tehnologij slovenski jezik nikoli ne bo mogel primerjati z nemškim, francoskim ali španskim jezikom, da angleškega ne omenjamo; vsekakor pa je treba ohranjati prizadevanja, da postane naš jezik digitalno podprt vsaj primerljivo zahodnoslovanskim in skandinavskim jezikom. Tehnološka oziroma digitalna podprtost jezika vključuje širok spekter jezikovnih tehnologij in virov, od katerih so bili mnogi podprti v projektu Razvoj slovenščine v digitalnem okolju.[1] V tem prispevku se osredotočamo samo na področje govornih virov, natančneje govornih korpusov oz. govornih baz. Za slovenščino sta na tem področju potekali do zdaj dve večji kampanji, obe zelo kratkoročni. V okviru projekta Sporazumevanje v slovenskem jeziku, ki ga je v obdobju 2008–2013 omogočilo Ministrstvo

1 https://slovenscina.eu

za izobraževanje, znanost in šport ob podpori sredstev iz Evropskega socialnega sklada, je v letih 2009–2010 nastal referenčni govorni korpus Gos (Verdonik idr., 2013) v obsegu 112 ur/1 mio. besed. Sledilo je desetletno zatišno obdobje z minimalnimi vlaganji v govorno infrastrukturo do leta 2020, ko je Ministrstvo za kulturo s pomočjo sredstev iz Evropskega sklada za regionalni razvoj spodbudilo projekt Razvoj slovenščine v digitalnem okolju, v katerem je v dveh letih in pol nastala govorna baza in korpus Artur v obsegu 1000 ur, kjer pa ne gre več samo za korpusne podatke, ampak je polovica gradiva po pisni predlogi govorjen in posnet govor, slabih dvesto ur pa ostaja brez transkripcij, samo s posnetki. S pomočjo gradiv iz Arturja je v okviru projekta Razvoj slovenščine v digitalnem okolju tudi referenčni govorni korpus Gos zrasel za več kot dvakrat, na 300 ur oz. 2,4 mio. besed.

Govorni viri niso pomembni samo za tehnološko podprtost jezika (predvsem avtomatsko razpoznavanje govora), čeprav je ta danes v središču pozornosti in nas upravičeno skrbi. Enako pomembni so za sodobno jezikoslovno znanost. Spoznavanje svojega jezika, instrumenta, prek katerega komuniciramo in se povezujemo v skupnost(i), je eden temeljnih humanističnih postulatov. Čeprav ne prinaša neposrednih ekonomskih učinkov, pomeni opazovanje jezika, človeške komunikacije in interakcije preusmeritev pozornosti nazaj k človeku in k temu, kar nas povezuje. Pomeni vrnitev znanosti nazaj k njenim primarnim izhodiščem, stran od prevladujoče kapitalistične ideologije, v kateri tudi znanost vse bolj pristaja v vlogo orodja za dodatno gospodarsko rast, ki dolgoročno izčrpava tako planet kot človeka. Najlažje, najbolj zanesljivo in najbolj široko dostopno lahko jezik in komunikacijo opazujemo prav v govornih virih, skozi posnetke govora v številnih vsakdanjih situacijah, ki smo jim izpostavljeni ali v njih aktivno sodelujemo. Korpusno jezikoslovje že več desetletij aktivno uporablja korpusne podatke v slovaropisju in slovnici (Adolphs in Carter, 2013). Tudi dialektologija v svojih raziskavah vse pogosteje posega po korpusnih podatkih (Goláňová idr., 2013; Šumenjak, 2012). Jezikoslovne discipline, ki so bolj povezane s sociološkimi (sociolingvistika, etnografija komunikacije, konverzacijska analiza) ali kognitivnimi

disciplinami (pragmatično jezikoslovje), prav tako temeljijo vedno več svojih raziskav na korpusnih podatkih (Aijmer in Rühlemann, 2015; govorni viri v okviru TalkBanka[2]), enako številne discipline, povezane z različnimi zdravstvenimi stanji ali razvojem jezika (CHILDES[3], PhonBank[4]) (MacWhinney, 2018). Podobno velja za uporabno jezikoslovje oz. bolj specifično za učenje jezika (CLARIN L2 Learner Corpora[5]). Korpusni podatki so lahko v pomoč tudi fonetičnim/fonološkim disciplinam, vključno s pravorečjem (Verdonik, 2021).

Nadaljnji razvoj govornih korpusov za slovenščino je torej ključen tako za razvoj njene tehnološke podprtosti kot tudi za razvoj slovenskega jezikoslovja. Prvi naslednji mejniki so 5 in 10 mio. besed v referenčnem govornem korpusu slovenščine ter dodatni področno specializirani in na višjih jezikovnih ravneh označeni (manjši) govorni korpusi. Ker pa govorimo o virih, ki zahtevajo velik časovni vložek, je toliko večja potreba po natančnem premisleku o njihovih potencialnih uporabnikih, njihovih potrebah, najbolj učinkovitih načinih zbiranja gradiv in ovirah pri tem, da lahko poteka razvoj v smeri, ki je najbolj smiselna in združuje čim več zaželenih učinkov. Zato sta vprašanji, ki ju naslavljamo v tem prispevku: Kdo so uporabniki govornih korpusov in kakšne so njihove potrebe po gradivih? Katere so prakse zbiranja gradiv za govorne korpuse in kako lahko sinergično naslovimo čim več različnih potreb z enotnim virom?

2 Vzorčni tuji modeli in obstoječi govorni korpusi za slovenščino

Govorni korpusi obstajajo za večino evropskih jezikov. Vse bolj intenzivno se govorni viri (ne samo korpusni, ampak tudi kot baze govora s posnetki po pisnih predlogah) razvijajo tudi za druge jezike s premalo jezikovnimi viri, t. i. »under-resourced languages« (npr. centralni kurdski jezik – Veisi idr., 2022; lugandski jezik – Mukiibi idr., 2022; švicarska nemška narečja – Plüss idr., 2022). Po drugi strani

2 https://www.talkbank.org
3 https://childes.talkbank.org
4 https://phon.talkbank.org
5 https://www.clarin.eu/resource-families/L2-corpora

so jeziki velikih jezikovnih skupnosti, kot so v Evropi angleška, nemška, francoska ali španska, tisti, ki pogosto služijo kot zgled za ostale jezike. V tem razdelku bomo podrobneje pogledali angleški govorni korpus, kjer je govorna komponenta British National Corpusa že od začetkov korpusnega jezikoslovja pogost referenčni vir za ostale jezike. Tehnološko dokaj zadovoljivo je med evropskimi jeziki podprta še nemščina, kjer je med govornimi korpusi najbolj prepoznaven korpus FOLK. Kot tretji primer bomo izbrali korpus, ki je slovenščini primerljiv po socio-ekonomskem statusu države, po številu govorcev in je prav tako slovanski jezik, to je slovaški govorni korpus.

2.1 Vzorčni tuji korpusi

British National Corpus (BNC) je bil pionir ne samo kot pisni, ampak tudi kot govorni korpus, saj je že tri desetletja nazaj, 1994, izdal govorno komponento v obsegu 4,2 milijona besed – t. i. BNC1994. Takrat je bil to eden prvih javno dostopnih korpusov svoje vrste. BNC1994 vključuje demografsko uravnotežen in besedilnovrstno uravnotežen del ter skuša biti reprezentativen za govorjeno britansko angleščino. Predstavljal je pomemben vir za raziskave v različnih jezikoslovnih disciplinah, od slovnice (Rühlemann, 2006; Smith, 2014) do sociolingvistike (McEnery, 2005; Säily, 2011; Xiao in Tao, 2007), konverzacijske analize (Rühlemann in Gries, 2015) in pragmatike (Wang, 2005; Capelle idr., 2015; Hatice, 2015), pa tudi za raziskave učenja jezika (Alderson, 2007; Flowerdew, 2009) in drugo. Love idr. (2017) navajajo kot razloge za njegovo popularnost, da obsega ortografsko zapisane podatke v velikem obsegu, da gre za splošen, reprezentativen vzorec govorjenih besedil in predvsem da je javno dostopen. Skozi čas pa je postajalo vedno bolj problematično, da se za raziskave današnje govorjene angleščine uporablja več kot dve desetletji staro gradivo. V obdobju od 2012 do 2016 je bil zato govorni del BNC nadgrajen s Spoken BNC2014, ki pa vsebuje samo demografsko uravnotežen del s posnetki v neformalnih kontekstih, ne pa tudi besedilnovrstno uravnoteženega dela. Kot razlog za to navajajo avtorji (Love idr., 2017), da po njihovem

opažanju obstaja večja potreba in zahteva po gradivu iz konverzacije in da imajo raziskovalci, ki želijo raziskovati britansko angleščino v specifičnih kontekstih, svoje lastne, specializirane korpuse oz. so takšni korpusi javno izdani (npr. BASE – korpus britanske govorjene akademske angleščine). Spoken BNC2014 obsega 11,5 milijona besed, 1251 posnetkov in sodelujočih 668 govorcev. Gre za vsakdanji neformalni govor, govorci pa so uravnoteženi glede na spol, starost, socio-ekonomski status in regijo. Za uporabnike je na voljo prek konkordančnika Sketch Engine.

Nemški govorni korpus FOLK (Schmidt, 2014) podobno kot angleški BNC izhaja iz potrebe po odprto dostopnih virih, ki so bili v času zasnove korpusa za nemščino redki in omejeni na specifične situacije, ne pa reprezentativni. Namenjen je tako za raziskovalne potrebe kot tudi za uporabo v šolskem okolju (Schmidt, 2016). Sledi ciljem, da pokrije širok nabor govornih interakcij v zasebnih, institucionalnih (predvsem interakcije v izobraževanju ter v delovnem okolju) in javnih situacijah (mediji). Kontrolirati skušajo tudi demografske kriterije, kot so regija, spol in starost govorcev. Da dokumentirajo komunikacijske prakse, vedno posnamejo in vključijo celotno interakcijo, ne samo izbranih segmentov. Ker so vidne oblike komunikacije pogosto enako pomembne kot slišne, skušajo v zadnjem času vedno, kjer je mogoče, zajeti tudi video posnetek, ne samo avdio. Projekt se je začel leta 2008 (Schmidt, 2016). V prvi izdaji je korpus obsegal 1 mio. besed (Schmidt, 2014), ker pa gre za dolgoročni načrt, se korpus ves čas dograjuje. Julija 2022 (verzija 2.18) je korpus FOLK obsegal 3,2 milijona pojavnic oz. 336 ur posnetkov, od tega 151 ur z videom (Schmidt, 2023). Korpus FOLK je za uporabnike dostopen prek konkordančnikov DGD (Datenbank für Gesprochenes Deutsch).[6]

Tudi korpus govorjene slovaščine s-hovor sodi v sklop t. i. velikih reprezentativnih govornih korpusov (Garabík, 2023). Prvič je bil izdan decembra 2008 in se od takrat ves čas nadgrajuje. Trenutna različica s-hovor-7.0 obsega 851 ur posnetkov oziroma 7,8 mio.

6 https://dgd.ids-mannheim.de/DGD2Web/jsp/Welcome.jsp

pojavnic.[7] Približno tretjino posnetkov za korpus je prispeval slovaški Nacionalni institut spomina (Nation's Memory Institute – UPN). 4,2 mio. pojavnic so posnetki iz drugih virov, poleg medijev in parlamenta je zelo veliko tudi terenskih posnetkov, pri čemer upoštevajo osrednje demografske kriterije (spol, starost, izobrazbo, regijo izvora in skladnost s standardnim jezikom), pa tudi vrsto diskurza (Garabík in Rusko, 2007). Korpus je osredotočen na splošni govorjeni jezik in ne vključuje dialektalnega govora. Za uporabnike je dostopen prek konkordančnika Sketch Engine.

2.2 Slovenski govorni korpusi

Slovenci smo potrebe po reprezentativnih govornih korpusih hitro zaznali (Stabej in Vitez, 2000), do prve izvedbe pa je prišlo desetletje kasneje (Verdonik idr., 2013). V letu 2023 je bil izdan še en pomemben govorni vir: govorna baza in korpus Artur (Verdonik idr., 2023a; Verdonik idr., 2023b), katerega cilj je bil zagotoviti gradiva za razvoj avtomatskega razpoznavanja govora za slovenščino. Gradiva iz Arturja so bila uporabljena tudi za nadgradnjo referenčnega govornega korpusa Gos v različico 2.x (Verdonik idr., 2023c), kjer so bili združeni obstoječi viri z namenom zagotavljanja nadgradnje reprezentativnega korpusa za jezikoslovne raziskave. Izdelan je bil tudi prenovljen uporabniško prijazen konkordančnik,[8] ki omogoča uporabo korpusa tudi v šoli oz. nasploh zunaj raziskovalne sfere. V Tabeli 1 so predstavljene osnovne informacije o govorni bazi in korpusu Artur. Kot vidimo iz nje, približno polovica baze vključuje posnetke branja povedi po pisnih predlogah. Čeprav gre za demografsko uravnotežen nabor velikega števila govorcev, pa besedila izhajajo iz pisnih virov (Žganec Gros idr., 2022). Od preostale polovice precejšen delež nima transkripcij, ampak samo posnetke. Največji del na novo zbranih posnetkov z ročno narejenimi kvalitetnimi zapisi govora tako obsega gradivo iz Državnega zbora Republike Slovenije, torej parlamentarni govor. Preostanek se deli dokaj enakomerno na

7 https://korpus.sk/en/corpora-and-databases/snc-corpora/publicly-available-snc-corpora/corpus-of-spoken-slovak/

8 https://viri.cjvt.si/gos/

javni govor in nejavni govor, pri čemer nejavni govor v veliko primerih ni interakcija, ampak razlaganje ali opisovanje po vnaprej določenih vsebinskih iztočnicah. Poleg teh večjih sklopov vključuje Artur še nekaj manjših, prilagojenih potrebam razvoja tehnologij. Na novo pridobljeno gradivo je torej z vidika potreb jezikoslovja zelo omejeno, njegova bistvena prednost v primerjavi s posnetki iz prvega vala snemanja v letu 2010 pa so kvalitetni avdio posnetki, ki omogočajo raziskave in razvoj na podlagi analize ali procesiranja avdio signala.

Tabela 1: Osnovni podatki o govorni bazi in korpusu Artur.

	Št. govorcev	Št. posnetkov	Trajanje v urah
Brane povedi	884	257.942	485
Črkovanje	345	676	10,5
Studijski posnetki za sintezo	1	10.109	27
Pogovori/opisovanje	263 (181 trans.)	301 (210 trans.)	94 (61 trans.)
Pametni dom (za avtomatsko razpoznavanje govora)	148 (148 trans.)	195 (189 trans.)	7,5 (7 trans.)
Opis obraza (za avtomatsko razpoznavanje govora)	125 (86 trans.)	125 (86 trans.)	10 (6 trans.)
Mediji, javni dogodki	811 (240 trans.)	400 (100 trans.)	207 (62 trans.)
Parlament	158	2799	201 (vse trans.)
Skupaj	**2222 (1586 trans.)**	**286.064**	**1067 (884 trans.)**

V letu 2023 je bila izdana tudi nadgrajena različica korpusa Gos, ki vključuje zbir vsega, kar je bilo od njegove prve izdaje na voljo pod ustrezno licenco in je bilo mogoče smiselno vključiti v reprezentativni govorni korpus, ne da se pretirano poruši uravnoteženost gradiv. Korpus Gos 2.x tako obsega sledeče vire in vsebine:

- Gos 1.1: 1 mio. besed/112 ur; avtentični posnetki, izogibanje branemu govoru, vsebuje besedilnovrstno in demografsko uravnotežen del po vzoru BNC; posnetki so pogosto slabše kvalitete, v zasebnem delu je izredno veliko segmentov s prekrivanjem govora dveh ali več govorcev;
- GosVL: 180.000 besed/22 ur; 55 predavanj ali delov predavanj, izbranih s portala Videolectures.net z upoštevanjem

uravnoteženosti po vedah in demografskih značilnosti govorcev, kolikor je bilo o njih mogoče sklepati iz posnetkov in na spletu dostopnih podatkov;

- Artur (1,2 mio. besed/185 ur):
  - javni govor, 422.000 besed/62 ur,
  - nejavni govor, 324.000 besed/61 ur,
  - parlamentarni govor, 450.000 besed/62 ur.

3 Uporabniki govornih korpusov in njihove potrebe po gradivih

V tem razdelku skušamo odgovoriti na vprašanje, kdo so uporabniki govornih korpusov in kakšne so njihove potrebe po gradivih. S pomočjo pregleda literature v mednarodnem prostoru in posebej tudi v slovenskem prostoru bomo ugotavljali, v katerih disciplinah pogosto posegajo po korpusnih podatkih, analizirali gradiva obstoječega referenčnega govornega korpusa Gos 2.x in ugotavljali, kje so pomanjkljivosti, ki jih je treba nasloviti ob prihodnjih nadgradnjah korpusa.

3.1 Uporabniki

Love idr. (2017) navajajo kot pomembne uporabnike govornega korpusa BNC slovnico, sociolingvistiko, konverzacijsko analizo, pragmatiko in učenje jezika kot drugega jezika. Schmidt (2016) posveti posebno pozornost šolskemu okolju in izobraževanju kot sicer neraziskovalnemu, a enako zainteresiranemu uporabniku govornih korpusov. Večinoma specializirani govorni korpusi v okviru projekta TalkBank opozorijo na uporabnike iz psihologije (razvoj govora) in medicine (npr. raziskave govora pri osebah z demenco, poškodbami desne hemisfere, travmatološkimi poškodbami možganov, afazijo, logopedskimi težavami). Tudi za potrebe dialektologije se večinoma razvijajo specializirani korpusi (Goláňová idr., 2013; Šumenjak, 2012). Fonetika in fonologija sta precej specifičen uporabnik, ki bolj kot same korpuse potrebujeta določene jezikovnotehnološke servise za avtomatsko predpripravo korpusnih podatkov za analizo, kot

jih ponuja na primer WebMAUS.[9] Na drugi strani so velik in zelo aktiven uporabnik govornih korpusov tehnologije: avtomatsko razpoznavanje govora (Gril idr., 2021), klasifikacija (Vlaj in Žgank, 2023) in prepoznavanje govorcev (Ljubešić in Rupnik, 2022), procesiranje govorjenega jezika (Lee idr., 2021), govorjeni sistemi dialoga (Chen idr., 2021) itd.

Med razpoložljivimi govornimi korpusi je poleg referenčnih kar nekaj korpusov specializiranih, pri čemer prevladujejo korpusi parlamentarnega govora (Ogrodniczuk idr., 2020) in govor v akademskem okolju (Verdonik, 2018; korpus MICASE[10]), verjetno predvsem zaradi lahke dostopnosti tovrstnih podatkov v primerjavi z drugimi področji. Mednarodno eden najbolj pogosto procesiranih govornih korpusov je Switchboard (Godfrey in Hollimann, 1993)[11], ki vsebuje nekoliko specifično izzvane interakcije med dvema neznancema na eno od tem, ki so bile pripravljene vnaprej, torej delno simulirano, in ne avtentično govorno situacijo.

V slovenskem okolju je tradicija raziskovanja govorne interakcije v jezikoslovju šibka in raziskave v primerjavi s pisnim jezikom redke. Tehnološki uporabniki so v slovenskem prostoru morda nekoliko bolj aktivni uporabniki govornih korpusov in baz kot jezikoslovci. V letu 2023 je bila konferenca Slavistični znanstveni premisleki, ki jo organizira Oddelek za slovanske jezike in književnosti Univerze v Mariboru, posvečena tematiki infrastrukture za raziskave govora. Raziskovalci iz slovenskega prostora, ki so se odzvali, so naslavljali vprašanja govorne infrastrukture z vidika sociolingvistike, leksike, skladnje, jezikovnih tehnologij, dialektologije, pragmatike in učenja drugega jezika, med specializiranimi področji pa so med drugim izstopali parlamentarni govor, govorjeni jezik v literaturi, v gledališču, na radiu in televiziji (Krajnc Ivič, 2023). V primerjavi z mednarodnim prostorom v slovenskem ni zaznati uporabnikov specializiranih korpusov s področja razvoja govora, čeprav je to raziskovalno področje aktivno (Marjanovič Umek idr., 2006), in ne iz logopedije in drugih

9 https://clarin.phonetik.uni-muenchen.de/BASWebServices/interface/WebMAUSBasic
10 https://quod.lib.umich.edu/cgi/c/corpus/corpus?page=home;c=micase;cc=micase
11 https://catalog.ldc.upenn.edu/LDC97S62

disciplin, povezanih z medicino. O uporabi korpusov v šolstvu in splošni javnosti je po drugi strani kar nekaj razmislekov (Logar idr., 2023), kar se kaže tudi skozi tovrstni javnosti prilagojene konkordančnike, tudi za korpus Gos.[12]

3.2 Potrebe uporabnikov

Potrebe uporabnikov so tukaj obravnavane z vidika, da skušamo z enim osrednjim referenčnim korpusom zadovoljiti potrebe čim več različnih disciplin. Čeprav so potrebe včasih kontradiktorne, je v primeru manjših skupnosti to edini način, da se zagotovi gradivo za različne uporabnike, saj je razpoložljivih finančnih sredstev malo, potreben finančni in časovni vložek pa velik.

Potrebe uporabnikov v zvezi z govornim korpusom lahko razdelimo v več ravni. Prva se nanaša na vrste situacij, ki so zajete v govorni korpus. V ta namen lahko ločujemo specializirane in referenčne korpuse. Specializirani govorni korpusi za slovenščino zelo dobro pokrivajo parlamentarni govor (Pančur idr., 2020), deloma akademski govor (Verdonik, 2018), ostalih vrst govora pa tako rekoč ne, z izjemo majhnega, 1 uro trajajočega korpusa govora Koprive na Krasu (Šumenjak, 2012) kot do zdaj edinega primera dialektološkega korpusa za slovenščino. Znotraj referenčnega korpusa Gos je sicer še dokaj obsežno zastopan tudi akademski govor, vendar samo v javnih situacijah. Tudi medijski govor je široko zastopan v referenčnem korpusu Gos, pa tudi s specializiranimi viri (npr. BNSI Broadcast News, Žgank idr., 2005). Pomemben potencialni uporabnik govornih virov v slovenskem prostoru je dialektologija. V referenčnem govornem korpusu Gos ločevanje med dialektološkim in nedialektološkim gradivom ni vzpostavljeno. Ne v prvem ne v drugem snemalnem valu na terenu snemanje ni bilo osredotočeno samo na urbana središča, ampak je potekalo mešano po vaseh in mestih v vseh slovenskih regijah. Tudi sicer ni jasnih podatkov, kakšne so v manjših urbanih središčih razlike v govoru med mestom in okoliškimi vasmi. V referenčnem korpusu Gos najdemo tako posamezne primere zelo

12 https://viri.cjvt.si/gos/

narečnega govora (tudi iz zamejstva v vseh treh sosednih državah), vendar so to samo posamični naključno vključeni narečni govori. Kriteriji za zajem gradiv so se namreč v prvem snemalnem valu ravnali po registrskih enotah, v drugem pa po statističnih regijah. Čeprav je podobna praksa običajna (Love idr., 2017), bi jo bilo v prihodnje smiselno ponovno premisliti tudi z dialektološkega vidika.

Kot vidimo na primeru Spoken BNC2014, je ključen segment govornih korpusov za mnoge jezikoslovne discipline vsakdanja govorna interakcija v zasebnih situacijah. Za slovenščino je lahko ta segment še dodatno pomemben zaradi velike dialektalne razpršenosti. Te vsebine so bile v slovenskem govornem korpusu Gos kvalitetno pokrite v prvem valu snemanja, v drugem pa veliko manj zaradi zahtev tehnologij po visoko kakovostnih posnetkih brez prekrivanja govora. V drugem valu snemanja je tako veliko vsebin celo kar monoloških in niso primerne za raziskave interakcije. Po drugem valu snemanj torej beležimo v referenčnem govornem korpusu slovenščine pomanjkanje posnetkov avtentičnih vsakdanjih nejavnih in institucionalnih govornih interakcij. Razmisliti je treba tudi o morebitni vključitvi govornih situacij, ki do zdaj niso bile zajete v korpus Gos, najdemo pa zainteresirane raziskovalce (npr. gledališki govor, dramatika), ter posebno pozornost posvetiti vprašanjem otroškega in mladostniškega govora ter govora neprvih govorcev slovenščine.

Naslednje vprašanje zajemanja gradiv je odločitev o tem, kje se vključeni posnetek začne in konča. Kot vidimo pri Schmidtu (2023), obstajajo argumenti, da se vključijo posnetki celotne interakcije od začetka do konca. V prvi izdaji korpusa Gos ta praksa ni bila dosledno upoštevana, veliko bolj v drugem valu zbiranja gradiv. To je mogoče v primeru javne in institucionalne komunikacije, kjer imajo dogodki jasne začetke in konce. V vsakdanji zasebni komunikaciji pa začetki in konci niso nujno jasni, predvsem pa so ob snemanju začetki lahko obremenjeni z razlaganjem namena snemanja, nameščanjem naprav, podpisovanjem strinjanja in uvodno nervozo govorcev zaradi snemanja. Družabni dogodki lahko potekajo tudi več ur, z vmesnimi premori, spremembami prisotnih govorcev ipd. V primeru nejavnih

terenskih posnetkov odgovor, kaj šteje kot celotna interakcija, tako ni vedno enoumen.

Tretji vidik potreb uporabnikov glede gradiv se nanaša na modalnosti zajema in tehnično kvaliteto gradiv. V prvem valu snemanja za korpus Gos je bila tehnična kvaliteta v terenskih posnetkih pogosto nizka, prednost se je dajalo avtentičnosti situacije, čim manjši invazivnosti snemalnih naprav in preprostosti njihove uporabe. Takšni posnetki ne zadoščajo potrebam disciplin, kjer se procesira ali analizira avdio signal, zato je zahteva po višji kvaliteti posnetkov tako rekoč nujna. Posebej težavno je vprašanje hkratnega govora, ki je v vsakdanji interakciji široko prisoten, po drugi strani pa so segmenti s hkratnim govorom neprimerni za avdio procesiranje. V drugem valu snemanja je bil tako isti terenski pogovor ločen v dva posnetka, za vsakega govorca en. Slabosti takega načina sta pogosto prisoten presluh in fizična ločenost transkripcij. Na ta način gradiva niso primerno pripravljena za pragmatične, diskurzne in sociolingvistične analize. Pri snemalni tehnologiji je nadalje treba nasloviti tudi vprašanje video zajema. Mnogi govornokorpusni centri (prim. Schmidt, 2023) že nekaj časa zajemajo tudi video, ne samo avdia. Govorna komunikacija ni samo slušna, ampak v večini primerov tudi vidna in ob odsotnosti vizualnega dela ne moremo celostno analizirati govora, hkrati pa je za vizualne podatke vedno bolj zainteresirana tudi tehnologija (npr. za razvoj pogovornih agentov z vizualnim vmesnikom).

Četrti vidik se nanaša na vprašanja metapodatkov o govorcih in posnetih situacijah. Ta vprašanja so že bila podrobno obravnavana v Verdonik (2022). Vsekakor je treba upoštevati, da imajo discipline, kot so sociolingvistika, pragmatika, analiza diskurza, tudi dialektologija, potrebo po čim bolj natančnem opisu konteksta, zato je lahko možnost, da se kontekst vsakega posnetka opiše v nekaj stavkih, zelo dobrodošel dodaten podatek, čeprav ni strukturiran. Podobni opisi bi se lahko dodajali tudi za govorca, saj imajo discipline, kot je dialektologija ali sociolingvistika, potrebo po čim bolj natančni predstavitvi govorca in njegove podvrženosti različnim jezikovnim vplivom. Metapodatke o govorcih in posnetkih, ki jih popišemo, moramo pri tem ločevati od kategorij, ki jih postavimo kot kriterije za

zajem. Medtem ko so popisani podatki idealno čim bolj podrobni, so kriteriji za zajem v demografsko uravnoteženem delu korpusa praviloma: spol, starost, izobrazba, regija, prvi jezik. V angleškem Spoken BNC2014 najdemo kot kriterij še socio-ekonomski status, kar v slovenskem okolju že ob začetkih govornega korpusnega jezikoslovja ni bilo prepoznano kot primeren kriterij za naše okolje. Pač pa bi bilo glede na široko narečno razpršenost smiselno razmisliti o tem, ali je treba pri kriterijih za zajem določiti regijo govorcev bolj podrobno kot samo na ravni statističnih regij.

Nazadnje je vprašanje tudi, kakšne so potrebe glede zapisa govora. V slovenskih govornih korpusih je vzpostavljena praksa dvojnega ortografskega zapisa, pogovornega in standardiziranega, primerljivo kot v nemškem (Schmidt, 2023) ali slovaškem korpusu (Garabík, 2023). Vprašanja smotrnosti takšnega zapisovanja so naslovljena v Verdonik (v tisku) in končno priporočilo je, da se s tem nadaljuje. Nekatere discipline imajo potrebo po bolj natančnih, fonemskih ali fonetičnih zapisih, zlasti dialektologija, slovaropisje, fonetika in fonologija. Na tak način je seveda mogoče zapisati le manjši del gradiv, na primer učni korpus, kar bi bil smiseln korak v prihodnosti zlasti z namenom, da se vzpostavi servis za avtomatsko pretvorbo ortografskega v fonetični zapis, kot jo omogoča na primer WebMAUS. V zapisih se ves čas ohranjajo tudi zabeležke o nekaterih osnovnih neverbalnih dogodkih med govorom, kot so daljši premori, smeh, nerazumljiv govor, govor v tujem jeziku ipd.

Iz zgornjega pregleda je vidno, da so potrebe nekaterih uporabnikov govornih korpusov do določene mere nasprotne ena drugi. Največja težava je po eni strani potreba po kvalitetnem zvoku in videu z malo ali nič hkratnega govora, po drugi pa potreba po posnetkih avtentičnih govornih dogodkov. Z namestitvijo govorcev v studio, snemanjem na ločene kanale, govorjenjem »na ukaz« zagotovimo visoko kvaliteto posnetkov, a postavimo govorce v stresno in nenaravno situacijo, za katero ne moremo trditi, da so se govorci obnašali v njej enako, kot bi se v avtentičnem okolju. Avtentično okolje pa je lahko polno hrupov in šumov, govorci se vmes premikajo po prostoru, namestitev snemalne opreme v domače okolje govorcev je hkrati

tudi veliko večji vdor v zasebno življenje govorcev kot snemanje v studiu. Za doseganje kompromisa je zato potreben zelo premišljen izbor vsake snemalne situacije in govorcev, to pa ob intenzivnih snemalnih kampanjah, omejenih s kratkimi časovnimi roki, ni mogoče.

4 Prakse zbiranja gradiv za govorne korpuse

Na podlagi pregleda tuje literature o govornih korpusih ter informacij in izkušenj pri projektih, v katerih so se snemala gradiva za slovenske govorne vire, v tem razdelku predstavljamo možne načine pridobivanja gradiv in probleme, ki jih imamo pri tem v Sloveniji.

Gradiva za govorne korpuse prihajajo iz dveh bistveno različnih virov: prvi so že obstoječi, večinoma javno tako ali drugače predvajani ali dostopni posnetki, kot so posnetki medijskih hiš, na internetu, v parlamentu, v okviru različnih javnih dogodkov ipd. Za te posnetke je treba doseči dogovore z nosilci avtorskih pravic, ki so lahko institucije (npr. državna RTV, državni zbor, Arnes), podjetja (komercialne radijske in TV-postaje), posamezniki (medijski posnetki, če govorci niso prenesli pravic na medij, predvsem pa različni javni dogodki) ali tudi mednarodne korporacije (npr. Youtubova licenca). Če posamezen akter ne vidi jasnega lastnega interesa za sodelovanje, je velika možnost, da dogovarjanje ni uspešno. Za gradiva, ki jih uspemo pridobiti in skleniti dogovor z nosilci avtorskih pravic, je treba za vsak posamezen vir izvesti test zakonitega interesa. Govor sam po sebi je namreč bibliometrični podatek (Data Protection Working Party, 2003), in tudi če govorci v njem ne navajajo osebnih podatkov, kot sta ime in priimek, je treba zagotoviti ravnanje z gradivi skladno z zakonodajo.

Drugi sklop posnetkov so terenski posnetki vsakdanjih govornih interakcij. Kot smo videli v Razdelku 3.2, je za te posnetke interes zelo velik. Tradicionalno to poteka tako, da se angažirajo študenti, ki vsak prek svoje lastne socialne mreže nagovorijo govorce, da jih smejo posneti, in vsak govorec podpiše v ta namen pripravljeno izjavo, s katero se zagotovijo vse potrebne pravice za nadaljnje deljenje posnetkov (to vključuje dovoljenje za snemanje in uporabo

posnetka, privolitev v obdelavo osebnih podatkov z informacijami o obdelavi osebnih podatkov ter dovoljenje za uporabo avtorskih pravic). V drugem snemalnem valu v Sloveniji sta večji del terenskega snemanja prevzela najeta zunanja izvajalca, saj samo s pomočjo študentov v takšnem obsegu in kratkem časovnem roku ni bilo izvedljivo. Zunanji izvajalci so na primer podjetja, ki se ukvarjajo z avdio in/ali video produkcijo in pogosto tudi že imajo lastne sezname kontaktov ljudi, ki jih angažirajo kot govorce. Tretji način snemanja je, da povabimo govorce v studio in se tam pogovarjajo. Takim načinom se izogibamo, saj težko pridobimo ljudi iz oddaljenih krajev in bi bila potrebna večja finančna nagrada govorcem, kot jo običajno omogočajo razpoložljiva sredstva. Zanimiva alternativna možnost je mobilni studio, na primer ustrezno preurejen in opremljen kombi – tak način se je med drugim uporabljal pri snemanju branega govora za bazo Artur. V prihodnje bi bilo smiselno raziskati še tehnološko podprte pristope prek spletnih platform. Za množičenje v Sloveniji sicer ne obstaja že vzpostavljena skupnost, na katero bi se lahko obrnili, in dosedanji poskusi množičenja niso dali spodbudnih rezultatov (npr. na platformi Mozzilla CommonVoice so do julija 2023 zbrali samo 14 ur slovenskega govora, čeprav je bila platforma postavljena že leta 2017). Je pa vsekakor dandanes možnost, da govorci uporabijo lastne snemalne naprave (npr. pametne telefone), veliko bolj dostopna kot včasih in je lahko pridobivanje posnetkov na način, da jih govorci sami oddajo na neko spletno mesto, zanimiva rešitev. Tako v primeru množičenja kot v primeru snemanja v studiu je ključna težava motiviranje govorcev. Plačevanje honorarja govorcem za vsak oddan posnetek pomeni velik finančni vložek, ki lahko prestavlja tudi izredno obsežno administrativno delo in velike stroške oglaševanja, če gre za kratkoročno, intenzivno snemalno kampanjo.

Primerjava z vzorčnimi tujimi govornimi korpusi (gl. Razdelek 2.1) pokaže bistveno razliko s slovensko prakso: pri vseh navedenih tujih govornih korpusih gre za dolgoročne projekte, medtem ko smo v Sloveniji vse dosedanje gradivo v govornih korpusih posneli skupaj v dobrih treh letih, v dveh izredno intenzivnih snemalnih valih z dolgim vmesnim obdobjem brez financiranja in brez kakršnega koli

signala, kdaj se bo financiranje ponovno nadaljevalo. Na tak način v delo ni mogoče učinkovito vključevati študentov, kar je škoda med drugim za študijski proces na vsebinsko povezanih študijskih smereh. Finančni stroški so znatno višji zaradi izredno zahtevnega koordiniranja. Časa za podrobno načrtovanje snemanja in transkribiranja ter preučitev in pripravo podpornih orodij in okolij, ki bi lahko pohitrili delo in zmanjšali potreben čas za izvedbo posameznih korakov izdelave govornega korpusa, pa je premalo.

5 Diskusija in zaključek

V prispevku smo izhajali iz stališča, da je nadaljnji razvoj govornih korpusov za slovenščino ključen tako za razvoj njene tehnološke podprtosti kot tudi za razvoj slovenskega jezikoslovja. Zastavili smo si vprašanji, kdo so uporabniki govornih korpusov in kakšne so njihove potrebe glede gradiv ter katere so prakse zbiranja gradiv za govorne korpuse in kako lahko sinergično naslovimo čim več različnih potreb z enotnim virom. Kot bolj aktivne uporabnike govornih korpusov smo prepoznali jezikoslovne discipline leksikologijo, slovnico, sociolingvistiko, dialektologijo, konverzacijsko analizo, pragmatiko, uporabno jezikoslovje (učenje jezika kot tujega jezika); med tehnološkimi vedami predvsem govorne in semantične tehnologije; posamično so uporabniki tudi nekatere družboslovne (npr. razvojna psihologija) in naravoslovne discipline (logopedija ipd.); in šolstvo ter drugi neakademski uporabniki. Nekatere od navedenih disciplin potrebujejo specializirane korpusne vire, kljub temu pa smo skozi prispevek iskali možnosti pokrivanja potreb čim več disciplin skozi enoten, skupen govorni korpus, ki se lahko po potrebi deli na specializirane podenote. Ugotavljali smo, da v obstoječih slovenskih govornih korpusih že obstaja večja količina gradiv za medijski, parlamentarni in akademski govor. V prihodnje smo priporočali preusmeritev v terenski zajem avtentičnih vsakdanjih govorjenih situacij, kjer bi bilo smiselno zagotavljati bolj podrobno zastavljeno regionalno pokritost posnetkov, višjo kvaliteto posnetkov, kot je bila v prvem snemalnem valu, in vključen zajem videa, kjer koli bo mogoče. Pri

praksah zbiranja gradiv smo ugotavljali kot ključni problem nekontinuirano delo oz. snemanje v izredno kratkih časovnih obdobjih ter opozorili, da se tako veliko sredstev izgublja za koordiniranje množice sodelavcev in da ni zadosti časa za podrobno načrtovanje in pripravo orodij, s katerimi bi lahko bilo delo bolj učinkovito in bolj kvalitetno.

V naslovu smo nakazali, da vidimo snemanje gradiv za govorne korpuse v Sloveniji kot plutje med Scilo in Karibdo. Med obema grozečima skalama barka slovenskih govornih korpusov pluje večkrat: prvič, ko je ujeta v kratke roke in omejena finančna sredstva, za katere so pričakovanja po končnem obsegu gradiv izredno visoka; drugič, ko skuša ustreči včasih tudi precej nasprotujočim si željam različnih uporabnikov; tretjič, ko se sooča z nedostopnostjo in omejitvami že obstoječih posnetkov v različnih institucijah. Upajmo, da je barka svoje Scile in Karibde srečno preplula in da jo v prihodnje čakajo mirnejše vode v obliki dolgoročnega, stabilnega načrtovanja, kjer bo mogoče kontinuirano in premišljeno nadgrajevati govorne korpuse ter tako za ista ali celo manjša sredstva doseči več in boljše.

Zahvala

Prispevek je nastal v okviru raziskovalnega projekta ARRS Temeljne raziskave za razvoj govornih virov in tehnologij za slovenski jezik (J7-4642).

Literatura

Adolphs, S., Carter, R. (2013). *Spoken Corpus Linguistics: From Monomodal to Multimodal*. Routledge.

Aijmer, K., Rühlemann, C. (ur.) (2015). *Corpus Pragmatics: A Handbook*. Cambridge University Press. https://doi.org/10.1017/CBO9781139057493

Alderson, C. J. (2007). Judging the frequency of English words. *Applied Linguistics*, 28(3), 383–409. https://doi.org/10.1093/applin/amm024

Cappelle, B., Dugas, E., Tobin, V. (2015). An afterthought on let alone. *Journal of Pragmatics*, 80, 70–85. https://doi.org/10.1016/j.pragma.2015.02.005

Chen, N., You, C., Zou, Y. (2021). Self-Supervised Dialogue Learning for Spoken Conversational Question Answering. *Proceedings of the Interspeech 2021*, 231–235. https://doi.org/10.21437/Interspeech.2021-120

Data Protection Working Party. (2003). *Working document on biometrics*. Article 29 of Directive 95/46/EC. https://www.google.com/url?sa=t&rct=j&q=&esrc=s&source=web&cd=&ved=2ahUKEwjn_Km6kcCAAxUhS_EDHTOCAggQFnoECB0QAQ&url=https%3A%2F%2Fec.europa.eu%2Fjustice%2Farticle-29%2Fdocumentation%2Fopinion-recommendation%2Ffiles%2F2003%2Fwp80_en.pdf&usg=AOvVaw0NtFl7DWh5OLKSW3ZrVQik&opi=89978449

Flowerdew, J. (2009). Corpora in language teaching. V M. H. Long, C. J. Doughty (Eds.), *The Handbook of Language Teaching* (pp. 327–350). Wiley-Blackwell. https://doi.org/10.1002/9781444315783.ch19

Garabík, R. (2023). Corpus of Spoken Slovak. V M. Krajnc Ivič (ur.), *Infrastruktura za raziskave govora v humanistiki in jezikovnih tehnologijah: zbornik povzetkov* (pp. 5–6). 6. mednarodna znanstvena konferenca Slavistični znanstveni premisleki, Maribor, Slovenija. Univerza v Mariboru, Univerzitetna založba. https://doi.org/10.18690/um.ff.5.2023

Garabík, R., Rusko, M. (2007). Corpus of Spoken Slovak Language. V J. Levická, R. Garabík (ur.), *Computer Treatment of Slavic and East European Languages,* Zbornik konference Slovko 2007 (pp. 222–236). Brno: Tribun.

Giagkou, M., Lynn, T., Dunne, J., Piperidis, S., Rehm, G. (2023). European Language Technology in 2022/2023. V G. Rehm, A. Way (ur.), *European language Equality: A Strategic Agenda for Digital Language Equality*. Springer. https://doi.org/10.1007/978-3-031-28819-7

Godfrey, J. J., Holliman, E. (1993). *Switchboard-1 Release 2 LDC97S62*. Web Download. Linguistic Data Consortium. https://doi.org/10.35111/sw3h-rw02

Goláňová, H., Waclawičová, M., Komrsková, Z., Lukeš, D., Kopřivová, M., Poukarová, P. (2017). DIALEKT: nářeční korpus, verze 1 z 2. 6. 2017. Praha: ÚČNK FF UK. http://www.korpus.cz

Gril, L., Sepesy Maučec, M., Donaj, G., Žgank, A. (2021). Avtomatsko razpoznavanje slovenskega govora za dnevnoinformativne oddaje. Slovenščina 2.0, 9(1), 60–89. https://revije.ff.uni-lj.si/slovenscina2/article/view/9899/9554

Hatice, C. (2015). *Impoliteness in Corpora: A Comparative Analysis of British English and Spoken Turkish*. Sheffield: Equinox.

Krajnc Ivič, M. (ur.). (2023, 18. in 19. maj). *Infrastruktura za raziskave govora v humanistiki in jezikovnih tehnologijah: zbornik povzetkov.* 6. mednarodna znanstvena konferenca Slavistični znanstveni premisleki, Maribor, Slovenija. Univerza v Mariboru, Univerzitetna založba. https://doi.org/10.18690/um.ff.5.2023

Lee, H., Yun, J., Choi, H., Joe, S., Gwon, Y.L. (2021). Enhancing Semantic Understanding with Self-Supervised Methods for Abstractive Dialogue Summarization. *Proceedings of the Interspeech 2021.* 796–800, doi: 10.21437/Interspeech.2021-1270

Ljubešić, N., Rupnik, P. (2022). The ParlaSpeech-HR benchmark for speaker profiling in Croatian. V D. Fišer, T. Erjavec (ur.), *Jezikovne tehnologije in digitalna humanistika: zbornik konference,* 117–123. Inštitut za novejšo zgodovino. https://nl.ijs.si/jtdh22/pdf/JTDH2022_Proceedings.pdf

Logar, N., Gorjanc, V., Arhar Holdt, Š. (2023). Korpus Gigafida 2.0: mnenje uporabnikov. *Jezik in slovstvo* 68(2), 75–91. https://doi.org/10.4312/jis.68.2.75-91

Love, R., Dembry, C., Hardie, A., Brezina, V., McEnery, T. (2017). The Spoken BNC2014: Designing and building a spoken corpus of everyday conversations. *International Journal of Corpus Linguistics,* 22(3), 319-344. https://doi.org/10.1075/ijcl.22.3.02lov

MacWhinney, B. (2019). Understanding spoken language through TalkBank. *Behavior Research Methods,* 51, 1919–1927. https://doi.org/10.3758/s13428-018-1174-9

Marjanovič Umek, L., Kranjc, S., Fekonja, U., Saksida, I. (ur.). (2006). *Otroški govor: razvoj in učenje*. Izolit.

McEnery, T. (2005). *Swearing in English: Bad Language, Purity and Power from 1586 to the Present*. New York, NY: Routledge.

Mukiibi, J., Katumba, A., Nakatumba-Nabende, J., Hussein, A., Meyer, J. (2022). The Makerere Radio Speech Corpus: A Luganda Radio Corpus for Automatic Speech Recognition. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, (pp. 1945–1954). Marseille, France: European Language Resources Association.

Ogrodniczuk, M., Osenova, P., Erjavec, T., Fišer, D., Ljubešić, N., Çöltekin, Ç., Kopp, M., Meden, K. (2022). ParlaMint II: the show must go on. V

D. Fišer idr. (ur.), *Proceedings of the ParlaCLARIN III,* 1–6. http://www.lrec-conf.org/proceedings/lrec2022/workshops/ParlaCLARINIII/pdf/2022.parlaclariniii-1.1.pdf

Pančur, A. Erjavec, T., Ojsteršek, M., Šorn, M., Blaj Hribar, N. (2020). *Slovenian parliamentary corpus (1990-2018) siParl 2.0.* Slovenian language resource repository CLARIN.SI. http://hdl.handle.net/11356/1300

Plüss, M., Hürlimann, M., Cuny, M., Stöckli, A., Kapotis, N., Hartmann, J., Ulasik, M. A., Scheller, C., Schraner, Y., Jain, A., Deriu, J., Cieliebak, M., Vogel, M. (2022). SDS-200: A Swiss German Speech to Standard German Text Corpus. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, (pp. 3250–3256). Marseille, France: European Language Resources Association.

Rühlemann, C. (2006). Coming to terms with conversational grammar: 'Dislocation' and 'dysfluency'. *International Journal of Corpus Linguistics*, 11(4), 385–409. https://doi.org/10.1075/ijcl.11.4.03ruh

Rühlemann, C., Gries, S. (2015). Turn order and turn distribution in multiparty storytelling. *Journal of Pragmatics*, 87, 171–191. https://doi.org/10.1016/j.pragma.2015.08.003

Säily, T. (2011). Variation in morphological productivity in the BNC: Sociolinguistic and methodological considerations. *Corpus Linguistics and Linguistic Theory*, 7(1), 119–141. https://doi.org/10.1515/cllt.2011.006

Schmidt, T. (2014). The Research and Teaching Corpus of Spoken German — FOLK. *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, 383–387, Reykjavik, Iceland. European Language Resources Association (ELRA).

Schmidt, T. (2016). Construction and Dissemination of a Corpus of Spoken Interaction – Tools and Workflows in the FOLK project. *Journal for Language Technology and Computational Linguistics*, 31(1), 127–154.

Schmidt, T. (2023). FOLK – Das Forschungs- und Lehrkorpus für Gesprochenes Deutsch. *Korpora Deutsch als Fremdsprache,* 3(1), 166–169. https://doi.org/10.48694/kordaf.3737

Smith, A. (2014). Newly emerging subordinators in spoken/written English. *Australian Journal of Linguistics*, 34(1), 118–138. https://doi.org/10.1080/07268602.2014.875458

Stabej, M., Vitez, P. (2000). KGB (korpus govorjenih besedil) v slovenščini. V T. Erjavec, J. Gros (ur.), *Informacijska družba IS'2000, Jezikovne tehnologije* (pp. 79–81). Inštitut Jožef Stefan.

Šumenjak, K. (2012). Zasnova dialektološkega korpusa na primeru govora Koprive na Krasu. V B. Krakar Vogel (ur.), *Slavistika v regijah – Koper* (pp. 73–78). Zbornik 23. Slovenskega slavističnega kongresa, Zveza društev Slavistično društvo Slovenije.

Verdonik, D. (2018). Korpus in baza Gos Videolectures. V D. Fišer, A. Pančur (ur.), *Zbornik konference Jezikovne tehnologije in digitalna humanistika*, 265–268. Znanstvena založba Filozofske fakultete. http://www.sdjt.si/wp/wp-content/uploads/2018/09/JTDH-2018_Verdonik_Korpus-in-baza-Gos-Videolectures.pdf

Verdonik, D. (2021). Govorni viri za pravorečje. V T. Mirtič, M. Snoj (ur.), *1. slovenski pravorečni posvet* (pp. 120–132). Slovenska akademija znanosti in umetnosti. https://www.sazu.si/uploads/files/publikacije21/Rared2RAZPRAVE.pdf

Verdonik, D., Kosem, I., Zwitter Vitez, A., Krek, S., Stabej, M. (2013). Compilation, transcription and usage of a reference speech corpus: the case of the Slovene corpus GOS. Language Resources and Evaluation, 47(4), 1031–1048.

Verdonik, D., Bizjak, A., Žgank, A., Bernjak, M., Antloga, Š., Majhenič, S., Čakš, P., Pucer, M., Cvetko, M., Zelenik, M., Pavlič, J., Dobrišek, S., Križaj, J., Strle, G., Ivanovska, M., Grm, K., Bajec, M., Lebar Bajec, I., Jelovšek, T., Lokovšek, J., Longyka, J., Trojar, M., Žganec Gros, J., Mihelič, A., Vesnicer, B., Dretnik, N., Bordon, D. (2023a). ASR database ARTUR 1.0 (audio). Slovenian language resource repository CLARIN.SI. http://hdl.handle.net/11356/1776

Verdonik, D., Bizjak, A., Žgank, A., Bernjak, M., Antloga, Š., Majhenič, S., Čakš, P., Pucer, M., Cvetko, M., Zelenik, M., Pavlič, J., Dobrišek, S., Križaj, J., Strle, G., Ivanovska, M., Grm, K., Bajec, M., Lebar Bajec, I., Jelovšek, T., Lokovšek, J., Longyka, J., Trojar, M., Žganec Gros, J., Mihelič, A., Vesnicer, B., Dretnik, N., Bordon, D. (2023b). *ASR database ARTUR 1.0 (transcriptions)*. Slovenian language resource repository CLARIN.SI. http://hdl.handle.net/11356/1772

Verdonik, D., Zwitter Vitez, A., Zemljarič Miklavčič, J., Krek, S., Stabej, M., Erjavec, T., Verdonik, D., Potočnik, T., Sepesy Maučec, M., Majhenič, S., Žgank, A., Bizjak, A., Gril, L., Dobrišek, S., Križaj, J., Bajec, M., Lebar Bajec, I., Jelovšek, T., Trojar, M., Bernjak, M., Dretnik, N., Strle, G., Dobrovoljc, K., Ljubešić, N., Rupnik, P. (2023c). *Spoken corpus Gos 2.0 (transcriptions).* Slovenian language resource repository CLARIN.SI. http://hdl.handle.net/11356/1771

Vlaj, D., Žgank, A. (2023). Acoustic Gender and Age Classification as an Aid to Human–Computer Interaction in a Smart Home Environment. *Mathematics*, 11(1). https://doi.org/10.3390/math11010169

Žganec Gros, J., Vesnicer, B., Dobrišek, S. (2022). A method for selection of phonetically balanced sentences in read speech corpus design. *Proceedings of the 30th European Signal Processing Conference (EUSIPCO 2022)* (pp. 1136-1139). Belgrade, Serbia: EURASIP. https://eurasip.org/Proceedings/Eusipco/Eusipco2022/pdfs/0001136.pdf

Žgank, A., Verdonik, D., Zögling Markuš, A., Kačič, Z. (2005). BNSI Slovenian broadcast news database - speech and text corpus. *Interspeech Lisboa 2005: proceedings of the 9th European conference on speech communication and technology* (pp. 1537-1540). Bonn: Universität, Institut für Kommunikationsforschung und Phonetik.

Wang, S. (2005). Corpus-based approaches and discourse analysis in relation to reduplication and repetition. *Journal of Pragmatics*, 34(4), 505–540. https://doi.org/10.1016/j.pragma.2004.08.002

Xiao, R., Tao, H. (2007). A corpus-based sociolinguistic study of amplifiers in British English. *Sociolinguistic Studies*, 1(2), 231–273. https://doi.org/10.1558/sols.v1i2.241

Transkribiranje govora pri izdelavi govorne baze Artur: od pogovornih k standardiziranim zapisom

Mitja TROJAR
ZRC SAZU, Inštitut za slovenski jezik Frana Ramovša

Andreja BIZJAK
Univerza v Mariboru, Fakulteta za elektrotehniko, računalništvo in informatiko

Povzetek

Prispevek predstavlja načela za zapis govora pri izdelavi govorne baze Artur in opis izvedbe transkribiranja govora v projektu RSDO. Opisana so načela za zapis govora za pripravo pogovornih zapisov in praktični vidiki njihove priprave. Sledi opis priprave standardiziranih zapisov, ki so bili pripravljeni z ročnim popravljanjem avtomatskih pretvorb pogovornih zapisov. Prispevek zaokrožuje opis izzivov pri izdelavi pogovornih in standardiziranih zapisov ter priporočila za podobne projekte v prihodnosti.

Ključne besede: govorjeni jezik, transkripcije, pogovorni zapisi, standardizirani zapisi, govorna baza Artur

Abstract

This chapter presents principles of transcribing speech in the making of the Artur speech database and a description of speech transcription in the project Development of Slovene in a Digital Environment. It includes a description of principles used in the creation of orthographic transcriptions as well as its practical aspects, which is followed by an account of the making of standardised transcriptions, which were created by making manual corrections to automatic conversions of orthographic transcriptions. The chapter concludes with a presentation of challenges encountered in

the making of orthographic and standardised transcriptions and with recommendations for similar future projects.

Keywords: spoken language, transcriptions, orthographic transcriptions, standardised transcriptions, Artur speech database

1 Uvod

Delovni sklop 2 projekta RSDO[1] je bil namenjen razvoju govornih tehnologij za slovenščino, pri čemer je bil osnovni cilj projekta izdelava razpoznavalnika za slovenščino.[2] Za razvoj strojnega razpoznavanja govora je bilo treba zagotoviti dovolj veliko bazo transkribiranih posnetkov avtentičnega govora v raznolikih komunikacijskih okoliščinah. V ta namen je bila ustvarjena govorna baza Artur (**A**vtomatsko **r**azpoznavanje govora **R**azvoj slovenščine v digi**t**alnem okolj**u**), ki skupaj vsebuje 1094 ur posnetega govora. Visoko kakovost transkripcij smo v projektu RSDO poskušali zagotoviti z dvotirnim načinom transkribiranja: najprej so bili izdelani t. i. pogovorni zapisi, ki so bili nato pretvorjeni v t. i. standardizirane zapise. Odločitev za dvotirni način transkribiranja govora je bila sprejeta iz dveh razlogov. Prvi je ta, da je bil dvotirni način uporabljen že pri izdelavi govornega korpusa Gos (Verdonik in Zwitter Vitez, 2011). Primerljiva metodologija transkribiranja je omogočila razširitev korpusa Gos z izborom posnetkov in transkripcij, ki so nastali v projektu RSDO (gl. Gos 2.0).[3] Drugi razlog je ocena, da bi izdelava samo standardiziranih zapisov preveč obremenila transkriptorje, kar bi po predvidevanjih precej

1 Projekt Razvoj slovenščine v digitalnem okolju 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).

2 Aktivnosti izdelave govorne baze je koordinirala Darinka Verdonik (Fakulteta za elektrotehniko, računalništvo in informatiko Univerze v Mariboru), v projektnem sklopu pa so sodelovale še 3 partnerske ustanove iz znanstvenoraziskovalnega okolja (Univerza v Ljubljani, Institut »Jožef Stefan« in ZRC SAZU), 2 partnerja iz gospodarstva (Slovenska tiskovna agencija, d. o. o., Alpineon, d. o. o.), zunanji izvajalci (Fixmedia, d. o. o., Kreativist, d. o. o., Akademija INT, d. o. o., TAIA, d. o. o.) ter študentke in študenti Univerze v Ljubljani in Univerze v Mariboru.

3 https://viri.cjvt.si/gos/

povečalo število napak v transkripcijah (gl. Verdonik, Trojar in Bizjak, 2023a).

V prispevku je najprej predstavljena struktura baze Artur, opis delotoka, opis načel, po katerih so bili pripravljeni pogovorni in standardizirani zapisi, nakazane pa so tudi težave, s katerimi smo se srečali pri gradnji baze, ter priporočila za gradnjo primerljivih baz v prihodnosti.

2 Struktura govorne baze Artur in opis delotoka njene izgradnje

Govorna baza Artur je sestavljena iz 4 sklopov:[4]

1. **Brani govor** (573 ur posnetkov): posnetki prebranih povedi, zajetih iz dela korpusa Gigafida 2.0, ki je dostopen pod ustrezno licenco (CC BY); povedi je bralo pribl. 1000 govorcev (pribl. 30 minut govora na posameznega govorca), ki so ustrezali vnaprej določenim demografskim kriterijem (uravnoteženost govorcev po spolu, starosti, statistični regiji stalnega bivališča in prvem jeziku);[5] za ta sklop baze Artur transkripcije niso bile izdelane, ker so funkcijo transkripcij (standardiziranega zapisa) opravljale povedi iz pisnega korpusa.[6]
2. **Javni govor** (208 ur posnetkov): posnetki javnih dogodkov, ki vključujejo novinarske konference, okrogle mize, intervjuje, nagovore, predavanja, seminarje, posvete in moderirane pogovore, ki so v času pandemije covida-19 večinoma potekali prek spleta. Transkripcije (pogovorni in standardizirani zapisi) so bile pripravljene za 62 ur posnetkov.

4 Spodaj navedeni podatki o bazi Artur in opis njene strukture so povzeti po Verdonik, Bizjak in Dobrišek (2023).

5 Poleg opisanega branja povedi sklop branega govora vsebuje še posnetke govora enega samega (izšolanega profesionalnega) govorca v obsegu 50 ur (za razvoj avtomatske sinteze govora) in posnetke črkovanj (v obsegu 10 ur).

6 Povedi so bile izbrane tako, da izbor povedi odraža dejansko distribucijo trifonov v slovenskih povedih, dobljeni nabor povedi pa je bil v nadaljevanju avtomatsko in ročno prefiltriran tako, da so bile izločene povedi, ki so vsebovale besede, katerih zapis se bistveno razlikuje od zapisa po slovenskem črkopisu (zlasti citatne besede, npr. *pole position*), kratice, jezikovne napake, ali pa so bile povedi kako drugače neustrezne (žaljiv govor, nepopolne povedi ipd.). Za podrobnejši opis izbora povedi gl. Žganec Gros in Vesnicer (2021) in Žganec Gros idr. (2023).

3. **Nejavni govor** (112 ur posnetkov): posnetki govora v nejavnih govornih položajih, in sicer gre za tri tipe govornih dogodkov: proste dialoge med sogovornikoma, proste monologe ter razlaganje in opisovanje. Govorci so bili izbrani po istih demografskih kriterijih kot v sklopu branega govora. V nejavni govor so vključeni tudi posnetki, namenjeni razvoju dveh specializiranih razpoznavalnikov govora, ki vključujejo naslednje govorne dogodke: opis obraza ter brane in spontane ukaze za upravljanje pametnega doma. Transkripcije (pogovorni in standardizirani zapisi) so bile pripravljene za 74 ur posnetkov.
4. **Parlamentarni govor** (201 ura posnetkov): posnetki javnih sej Državnega zbora Republike Slovenije iz dveh sklicev med letoma 2010 in 2018. Vsaka datoteka vsebuje govor enega samega govorca. Govor posameznega govorca je lahko zajet na več posnetkih, vendar v celoti ne presega 3,5 ure. Za pripravo pogovornih zapisov so bili uporabljeni zapisi sej, ki jih pripravljajo v Državnem zboru. Študentke in študenti so zapise sej uredili in popravili tako, da so ustrezali standardom za pogovorni zapis (gl. spodaj). Transkripcije (pogovorni in standardizirani zapisi) so bile pripravljene za 201 uro posnetkov.

Transkripcije (pogovorni in standardizirani zapisi) so bile pripravljene za javni, nejavni in parlamentarni govor oziroma za skupno 337 ur posnetega govora. S pogovornimi in standardiziranimi zapisi je torej opremljena pribl. tretjina posnetkov v govorni bazi Artur, s čimer so bili projektni cilji doseženi in celo preseženi.

Časovno najzahtevnejši fazi delotoka sta bili priprava pogovornih in standardiziranih zapisov. Delotok je bil razdeljen v več faz, ki so se pri javnem, nejavnem in parlamentarnem govoru zvrstile v podobnem zaporedju: oddaja izvornih avdio posnetkov, pridobljenih od različnih virov ali posnetih na terenu, validacija posnetkov in odobritev ustreznih posnetkov glede na tehnično kakovost in vsebinsko ustreznost (npr. odsotnost sovražnega govora), oddaja soglasij govorcev in dokumentacije z metapodatki o posnetkih in govorcih, ročna priprava pogovornih zapisov, validacija pogovornih zapisov, avtomatska

pretvorba standardiziranih zapisov iz pogovornih zapisov, pregled in ročno popravljanje avtomatsko tvorjenih standardiziranih zapisov. Ročnemu pregledu so sledili še avtomatski pregledi, napake, ki so bile z njimi odkrite, pa so bile popravljene ročno ali avtomatsko (z izdelavo ustreznih skriptov). Na osnovi tako pridobljenih standardiziranih zapisov in avdio posnetkov je bil razvit splošni razpoznavalnik govora za slovenščino ter dva domensko specifična razpoznavalnika.

Pri načrtovanju delovnega procesa je pomembno zagotoviti transparentnost in sledljivost posameznih korakov, kar sodelujočim omogoča, da so seznanjeni, kateri posnetki ali zapisi posnetkov so že validirani, kateri so odobreni, zavrnjeni ali trenutno še v obdelavi. Delotok je bil zasnovan v vseh fazah priprave dvonivojsko (vsaka mapa je imela svojo kopijo kot varnostni arhiv z omejenimi pooblastili za spreminjanje njene vsebine). Tovrstna sledljivost je namreč ključnega pomena pri morebitnem kasnejšem iskanju izvora ali tipa napak, hkrati pa prepreči izgubo datotek. Zasnova delotoka mora biti tudi dovolj fleksibilna, da omogoča naknadno dodajanje novih faz, če se zanje kadar koli med pripravo baze pojavi potreba, kot se je izkazalo pri Arturju (npr. naknadno dodana faza prenosa ločil in velikih začetnic iz standardiziranih v pogovorne zapise).

Pri izdelavi govorne baze Artur je bila za obdelavo in shranjevanje datotek uporabljena oblačna platforma Nextcloud.

3 Načela, uporabljena pri izdelavi pogovornih in standardiziranih zapisov

Cilj pogovornega zapisa je, da »čim bolj olajša avtomatsko fonemsko-grafemsko pretvorbo in silabizacijo. V kombinaciji s standardiziranim zapisom je zasnovan tako, da omogoča čim boljšo ekstrakcijo novih kandidatov za oblikoslovno-fonetični leksikon, ki tako ali drugače odstopajo od normirane rabe.«[7] (Verdonik in Bizjak, 2023)

7 Opozoriti velja, da je v projektu RSDO pogovorni zapis služil le kot sredstvo za doseganje vmesnega cilja, tj. izdelave standardiziranega zapisa. Pri izdelavi razpoznavalnika, ki je bil končni cilj projekta, je bil namreč uporabljen samo standardizirani zapis. Pogovorni zapis je torej imel izrazito pomožno vlogo in ni nadomestek za (znanstveno/jezikoslovno) fonetično transkripcijo govora.

Govor je zapisan v slovenskem črkopisu v skladu z veljavnimi načeli, po katerih se glasovi zapisujejo s črkami. Pri tem se upoštevajo omejitve, ki izhajajo predvsem iz omejenega nabora črk, da bi karseda natančno predstavili glasovno podobo govora (Verdonik in Bizjak, 2023). Novost v pogovornem zapisu je npr. poseben znak za polglasnik (@), ki ga do Arturja v govornih korpusih za slovenščino ni bilo.

Pogovorni zapis poleg zapisa govora vključuje še segmentacijo govora, označevanje menjavanja govorcev, označevanje akustičnega ozadja (npr. prisotnost šuma ali glasbe), akustičnih dogodkov (npr. nenadni krajši zvoki, kašljanje, glasni vdihi) ter osnovnih neverbalnih značilnosti (npr. smeh ali premor). Pri izdelavi baze Artur so ga v orodju Transcriber 1.5.1. ročno pripravili zunanji izvajalci in študenti, koordinator transkribiranja pa je naključne dele pogovornih zapisov validiral in po potrebi popravil. Glede na izkušnje z Arturjem se je izkazalo, da je za 1 uro posnetka govora potrebnih okrog 20 ur dela za zapis, segmentiranje in označevanje govora (prim. Verdonik, Trojar in Bizjak, 2023a). Povedano še drugače, za pripravo pogovornega zapisa 2 oz. 3 minut govora je v povprečju potrebna ena ura dela.

Eden od prvih korakov pri izdelavi pogovornega zapisa je segmentiranje govora, ki je bilo pri pripravi govorne baze Artur delno prilagojeno za potrebe razvoja splošnega razpoznavalnika govora. Glede na to naj segmenti ne bi bili (pre)dolgi, tj. trajajoči več kot 10 sekund. Poleg tega smo upoštevali, da lahko mejo segmenta določimo le tam, kjer je v govoru dovolj premora, tj. vsaj 0,2 sekunde, ne da bi odrezali del predhodnega ali del naslednjega fonema. Glavni usmeritvi pri postavljanju meja med segmenti sta bila (1) kratek premor v govoru in (2) dolžina segmenta, ki ne sme biti predolga (Verdonik in Bizjak, 2023). Pri tako prilagojenem načinu segmentiranja označeni segmenti ne sovpadajo vedno s stavki oz. izjavami kot semantično in skladenjsko zaključenimi enotami, kar je bilo identificirano kot težava na višjih ravneh označevanja.

Navodilo za daljše premore, trajajoče več kot 1,5 sekunde, govor v tujem jeziku in nerazumljiv govor je, da se jih označi kot prazen segment ali izjavo brez govorca. Če je nerazumljiva zgolj posamezna beseda ali kratka fraza, se vstavi oznako *neraz.* Hkratni govor, ki se

pojavi v začetku ali ob koncu segmenta, ko govorca govorita hkrati, se ustrezno označi in se ga, če je razumljiv, tudi zapiše. Ob vsaki menjavi govorcev je treba paziti, da se menjava ustrezno označi.

Akustično ozadje se označi, kadar se v ozadju govora nenadoma pojavijo dalj časa trajajoči zvoki (najmanj 3 sekunde), ter določi, ali je šum v ozadju govor, glasba ali kaj drugega (npr. aplavz, zvonjenje telefona, prometni hrup). Kadar pa se med govorom pojavijo krajši zvoki (pribl. do ene besede), se vstavijo kot akustični dogodek (npr. zehanje, kihanje, vdih, izdih).

Besedni fragmenti (prekinjene besede, samopopravki) so označeni s praznim oklepajem stično za besedo, npr. *dru()*. Če se v govoru pojavijo osebni podatki o govorcih, ki niso javne osebnosti (npr. ime in priimek), se jih s piskom anonimizira. Številke (tudi vrstilni števniki) se izpišejo z besedo znotraj oglatih oklepajev, npr. *[tretje]*. Datumi se zapisujejo znotraj zavitih oklepajev, npr. *{peti osmi dva tisoč devet}*.

Novost v jezikovni bazi Artur v primerjavi s preteklimi govornimi korpusi pri nas je uvedba nekaterih dodatnih znakov za foneme, od katerih po pogostosti izstopa @ za polglasnik, omenimo pa še *$g* za zveneči *h* in *$r* za mehkonebni *r*. Nova je tudi vpeljava ločil in velikih začetnic v pogovorni in standardizirani zapis.

Redukcije glasov so v pogovornih zapisih upoštevane, saj se neizgovorjeni glasovi ne zapisujejo, npr. *tud* (Verdonik in Bizjak, 2023), premene po zvenečnosti pa se niso zapisovale, saj smo predvidevali, da bi bili najeti zunanji izvajalci ali študenti pri njihovem zapisu preveč nedosledni. Drugače je pri parlamentarnem govoru, saj je bil ta zapisan še pred uvedbo skupnih smernic za govorno bazo Artur. Premene po zvenečnosti so v parlamentarnem govoru zapisane, čeprav ne dosledno, npr. *gdo*, različen pa je tudi zapis dvoustničnega *u̯*, ki je praviloma zapisan s črko *u*, npr. *obraunavau*, in zapis kratic, ki so mestoma zapisane z veliki črkami, npr. *ZOFI*.

S ciljem čim bolj poenotenega zapisa neverbalnih in polverbalnih izrazov (npr. *eee, hm, uh, šššš*) so bile dopolnjene smernice za njihov zapis in razširjen seznam identificiranih neverbalnih in polverbalnih izrazov. Določili smo, da jih prednostno zapisujemo z največ

eno besedo, in če le gre, s tremi črkami, zelo podoben izraz pa zapišemo vedno na isti način, brez variacij (Verdonik, Bizjak 2023). Na začetku vedno dodamo znak #, npr. *#eem.*

V spodnji tabeli je navedenih nekaj smernic za izdelavo pogovornega in standardiziranega zapisa, pripravljenih za izgradnjo govorne baze Artur (Verdonik in Bizjak, 2023), skupaj s konkretnimi primeri iz iste baze.

Tabela 1: Smernice za pripravo pogovornega in standardiziranega zapisa po posameznih problemskih sklopih.

Sklop	Pogovorni zapis	Primer	Standardizirani zapis	Primer
Redukcije	Glasov, ki niso izgovorjeni, ne zapisujemo.	mamo, tko	Uporabljamo nereducirane oblike, skladno s pravopisno normo.	imamo, tako
	Redukcijo pomožnega glagola *bi* v *b* zapisujemo kot samostojno besedo, redukcije in premene oblik za prihodnjik pa kot: *čev* (*če bo*), *navm* (*ne bom*), *nav* (*ne bo*).	ne b navm	Standardizirane oblike zanikanega pomožnega glagola, ki so v pogovornem zapisu zapisane kot ena beseda, npr. *navm* pišemo z znakom + in stično: *ne+bom*.	ne bi ne+bom
	Polglasnik vedno zapisujemo z znakom @.	misl@m, fil@m, z@, @ldje, j@t, p@r	Posebnega znaka za polglasnik ne uporabljamo. Polglasnik se zapisuje skladno s pravopisno normo.	mislim, film, z, ljudje, iti, pri
Premene po zvenečnosti	Premen po zvenečnosti, razen pri predlogih *s/z* in *k/h*, v pisavi ne upoštevamo.	tud, fizka, j@z	Premene po zvenečnosti se načeloma ne upoštevajo oz. zapis besed sledi pravopisni normi.	tudi, fizika, jaz
Dvoustnični *u̯* in samoglasnik *u*	Dvoustnični *u̯*, ki ni nosilec zloga, v neknjižnih oblikah zapisujemo s črko *v*.	šov, prov	Dvoustnični *u̯* se zapisuje skladno s pravopisno normo, tj. s črkama *v* in *l*. Zapis se ravna po pravopisni normi, reducirane oblike deležnikov na *-il*, *-al*, *-el* se pišejo s samoglasnikom.	šel, prav
	Če dvoustnični *u̯* nastopi v knjižni besedni obliki, izgovorjeni skladno s standardom, ohranimo knjižni zapis.	bil, gledal		bil, gledal
	Če je glas *u* samoglasniški, tj. je nosilec zloga, ga pišemo s črko *u*.	odloču, padu, izpelu		odločil, padel, izpeljal
	Enako velja za predlog *v*, izgovorjen kot samoglasniški *u*.	u sobi	Predlog *v* se zapisuje skladno s pravopisno normo, tj. vedno s črko *v*.	v sobi

Sklop	Pogovorni zapis	Primer	Standardizirani zapis	Primer
Narečno specifični glasovi	Diftonge in druge pokrajinsko specifične foneme, ki jih v knjižnem jeziku ni, pišemo z najbližjimi ustreznimi črkami.	guvurim, tku, gučali, fseh, fsekakor	Oblike besed s pokrajinsko specifičnimi variantami fonemov oz. fonemi se nadomeščajo z ustreznimi knjižnimi oblikami besed.	govorim, tako, gučali, vseh, vsekakor
	Zveneči primorski *h* lahko zapišemo tudi z znakom *$g*, mehkonebni koroški *r* pa z znakom *$r*.	knji$g	Narečnim oblikam, ki nimajo ustreznih oblik v knjižnem jeziku, se priredi standardizirana oblika, ki sledi pravilom slovenskega črkopisa. Pri standardizaciji se prednostno uporablja standardizirane oblike iz narečnih slovarjev.	knjig
Lastna imena, citatne in tuje besede	Domača lastna imena zapisujemo skladno s pravopisom, tuja lastna imena pa tako, kot so izgovorjena.	Avstro-Ogrske, R@dovlci Mark Kjub@n, Zum, Heri Poter	Domača lastna imena zapisujemo skladno s pravopisom. Tuja lastna imena se prav tako zapisujejo v skladu s pravopisom, tj. bodisi podomačeno bodisi citatno (če konkretno lastno ime še ni podomačeno v pravopisnih priročnikih, se prednostno uporabi citatno obliko tujega lastnega imena).	Avstro-Ogrske, Radovljici Mark Cuban, Zoom, Harry Potter
	Citatne besede in besede oz. kratke fraze v tujem jeziku se pišejo, tako kot so izgovorjene.	fen, pojnt, trejler komon sens, riz@ning, Vourld of Vorkreft	Citatne besede in besede oz. kratke fraze v tujem jeziku se praviloma pišejo citatno, lahko pa tudi podomačeno, če je podomačeni zapis že uveljavljen oz. registriran v korpusih in slovarjih slovenskega jezika.	fan, point, trailer common sense, reasoning, World of Warcraft
Ločila Pisanje skupaj, narazen ali z vezajem	Ločila uporabljamo v njihovi skladenjski rabi in skladno s pravopisom. Tako zapisujemo tudi besede skupaj, narazen ali z vezajem.	pisiar-testi	Ločila uporabljamo v njihovi skladenjski rabi in skladno s pravopisom. Tako zapisujemo tudi besede skupaj, narazen ali z vezajem.	PCR-testi
Člen *ta* Kratice	Izjema so določni člen *ta*, ki ga pišemo stično, in kratice, ki jih pišemo tako, kot so izgovorjene, z malimi črkami in skupaj. Če je kratica lastno ime, jo pišemo z veliko začetnico. Okrajšav ne uporabljamo.	tamali tapravi A@g@r@f@t@c@p@p@-ja ajti-podjetjem Estea-jem	Določni člen ta pišemo z znakom + in stično. Kratice se pišejo skladno s pravopisom, tj. z vezajem med osnovo in končnico. Tvorjenke s kraticami se pišejo skladno s pravopisom.	ta+mali ta+pravi AGRFT CPP-ja IT-podjetjem STA-jem

Tabela 1 nakazuje razmerje med pogovornimi in standardiziranimi zapisi: standardizirani zapis je zapis govora, pri katerem se govorjeni jezik zapiše tako, kot bi bil zapisan v pisnem knjižnem jeziku. Standardizirani zapis lahko nastane na podlagi predhodnega zapisa govora,[8] ki se ga prilagodi (spremeni) tako, da nastali zapis (z možnimi predvidenimi odstopanji) ustreza pravilom slovenskega pravopisa, ki veljajo za pisni knjižni jezik. V standardiziranem zapisu oblike besed, značilne za govorjeni jezik, nastopajo v obliki, določeni za pisni knjižni jezik, besedilo je smiselno členjeno na povedi, stavke in besede (npr. besede v naslonskem nizu so ločene s presledki kot v knjižnem jeziku), uporabljena so ustrezna ločila.[9] Besedam, ki (še) niso registrirane v jezikovnih priročnikih za pisni knjižni jezik (zlasti Slovenski pravopis, SSKJ2, eSSKJ)[10] oz. niso zastopane v pisnih korpusih slovenskega jezika (predvsem Gigafida 2.0),[11] se priredi oblika, ki bi jo besede pričakovano imele, če bi se uporabljale v pisnem knjižnem jeziku.[12]

Končni validaciji pogovornih zapisov je v projektu RSDO sledila faza avtomatske pretvorbe pogovornih zapisov v standardizirane. Ročno preverjanje in popravljanje tako tvorjenih standardiziranih zapisov je v primeru zahtevnejših in nejasnih delov zahtevalo poslušanje posnetka in primerjavo z ustreznim segmentom v pogovornem zapisu.[13]

8 V projektu RSDO je standardizirani zapis nastal na podlagi pogovornega zapisa, načeloma pa bi lahko nastal tudi na podlagi npr. (dialektološke) fonetične transkripcije govora. Možno je seveda tudi, da bi standardizirani zapis nastal kot prvi zapis govora (tj. brez predhodnega pogovornega zapisa ali fonetične transkripcije).

9 Največja odstopanja od pravil in konvencij knjižnega jezika se pojavljajo na ravni besednega reda (ta se ni popravljal, ker se za razvoj razpoznavalnika zahteva, da so besede v transkripciji sinhronizirane z ustreznimi signali na posnetku) in zgradbe besed in povedi (ponavljanja besed, nedokončane besede in povedi, očitni lapsusi in drugi pojavi, značilni za govorjeni jezik (npr. nedoločni členi), niso bili izločeni), predvsem pa zaradi doslednega beleženja neverbalnih in polverbalnih glasov ter akustičnega ozadja in akustičnih dogodkov (smeh, vzdihi, hrup itd.). Ena od maloštevilnih izjem od pravopisnih pravil pri zapisovanju leksemov je zapis reducirane oblike veznikov *k@* (standardizirano v: *ke*), ki ji v knjižnem jeziku ustreza več veznikov, npr. *ker*, *ko*, *ki*. Za takšna odstopanja smo se odločili, ker smo poskušali zagotoviti, da bi bila pretvorba pogovornih zapisov v standardizirane čim manj odvisna od subjektivne interpretacije pripravljavca.

10 https://fran.si/

11 https://viri.cjvt.si/gigafida/

12 Oziroma se uporabi že standardizirano obliko, če gre npr. za narečne besede, ki so že registrirane v narečnih in/ali zgodovinskih slovarjih.

13 V idealnem primeru bi pregledovalec avtomatskih pretvorb v standardizirani zapis ob validaciji vsake transkripcije poslušal celotni avdio posnetek govora. V praksi bi to še precej bolj obremenilo pregledovalca in znatno podaljšalo pregledovanje standardiziranih zapisov.

Avtomatska pretvorba je bila sestavljena iz petih korakov: tokenizacije, pretvorbe v slovarske začetnice (ang. *truecasing*), prevoda, pretvorbe v besedilne začetnice (ang. *detruecasing*) in detokenizacije. Prevod je bil izveden z uporabo prevajalskega in jezikovnega modela, naučenega na bazi Gos VideoLectures 4.2 (Verdonik et al., 2021).

Avtomatske pretvorbe v standardizirane zapise so vsebovale napake, ki so bile ob pregledu odstranjene. V Tabeli 2 spodaj so vzporedno prikazani pogovorni zapisi, avtomatske pretvorbe v standardizirane zapise in ročno popravljeni standardizirani zapisi.

Tabela 2: Primerjava pogovornih zapisov, avtomatskih pretvorb v standardizirane zapise in ročno popravljenih standardiziranih zapisov.

	Pogovorni zapis	Avtomatska pretvorba v standardizirani zapis	Ročno popravljeni standardizirani zapis
1.	Čeprav jo bol uporablajo v@ ljudskem zdravilstvu, se danes spet več uporablja kot fčasih.	Čeprav jo bolj uporabljajo v ljudskem zdravilstvu, se danes spet več uporablja kot včasih.	Čeprav jo bolj uporabljajo v ljudskem zdravilstvu, se danes spet več uporablja kot včasih.
2.	Uporablamo tuji za astmo, plučni katar ...	Uporabljamo tudi za astmo, plučni katar	Uporabljamo tudi za astmo, pljučni katar ...
3.	Kakš@n nevljud@n?	Kakšen nevljudn?	Kakšen nevljuden?
4.	Če misl@n, da mi je tu nekaj stoplo v glavo, nekšne, nekšna ... črna energija.	Če mislim, da mi je to nekaj stoplo v glavo, nekakšne, neka ...črna energija	Če mislim, da mi je tu nekaj stopilo v glavo, nekšne, nekšna ... črna energija.
5.	To jaz bi rejs reko, to ne vem, kak so to bli vzgojeni, doma, f šoli, kje drugje, to so ... #puf.	To jaz bi res rekel, to ne vem kako so to bili vzgojeni, doma, v šoli, kje drugje, to so ...puf	To jaz bi res rekel, to ne vem, kako so to bili vzgojeni, doma, v šoli, kje drugje, to so ... #puf.

Zgledi v zgornji tabeli kažejo, da je bila avtomatska pretvorba v splošnem koristna in je pregledovalcu olajšala delo: pogosto nadaljnji popravki povedi niso bili potrebni (gl. zgled 1). Zgleda 2 in 3 kažeta, da so bile avtomatske pretvorbe v povedih včasih samo deloma ustrezne (zlasti znak @ za polglasnik je bil v pretvorbah pogosto samo izpuščen, ne pa tudi nadomeščen z ustrezno črko, tj. *e*). Zgled 4 kaže, da je bil leksem *nekšen* (po SSKJ2 narečno ‘nekak,

nekakšen') pretvorjen v leksem *nekakšen*, ki je bil nato ročno popravljen nazaj v *nekšen*. Pogosto je avtomatska pretvorba povzročala napake tako, da so bila določena ločila izbrisana (zlasti tri pike, gl. zgleda 2 in 5) in/ali premeščena na drugo mesto. Tovrstne napake je bilo treba odpraviti ročno. Pomembno je poudariti tudi to, da avtomatska pretvorba načeloma ni prizadela stave ločil in velikih začetnic (razen v zgoraj opisanih primerih, ko so bila ločila izbrisana), zato je bilo treba napačno rabljena ločila in velike začetnice (tj. napake transkriptorjev, ki so pripravljali pogovorne zapise) popravljati vzporedno v pogovornih in standardiziranih zapisih.

Osnovno načelo pri popravljanju avtomatskih pretvorb je bilo, da se segmentov govora v transkripcijah ne spreminja, segmenti v pogovornih in standardiziranih zapisih se morajo torej natančno ujemati. Prav tako pripravljavec standardiziranih zapisov načeloma ni spreminjal akustičnega ozadja in akustičnih dogodkov (smeh, odkašljanje, vdih, izdih, drugi zvoki), je pa lahko opozoril na napake pri njihovem označevanju v pogovornih zapisih. V pogovornih in standardiziranih je bilo dovoljeno uporabljati omejen nabor ločil: piko, vejico, klicaj, vprašaj, podpičje, opuščaj (kot del besede, pisan stično z besedo ali sredi besede), znak za *in* (*&*), narekovaje, dvopičje, tri pike, vezaj.[14] Ločila, uporabljena v določenem segmentu pogovornega zapisa, so morala biti uporabljena tudi v ustreznem segmentu standardiziranega zapisa. Oznake besednih fragmentov so v standardiziranem zapisu ohranjene takšne, kot se pojavijo v pogovornem zapisu (se jih torej ne spreminja). Prav tako se pri pripravi standardiziranih zapisov ne spreminjajo oznake anonimiziranih osebnih podatkov in oznaka za nerazumljiv govor (*neraz*). Tudi številke in datumi so načeloma ohranjeni tako, kot so bili zapisani v pogovornem zapisu, torej številke znotraj oglatih oklepajev in datumi znotraj zavitih oklepajev.

V Tabeli 3 so prikazani zgledi pogovornih in ustreznih standardiziranih zapisov iz vseh treh sklopov baze Artur, za katere so bili pripravljeni tako pogovorni kot standardizirani zapisi.

14 Pri vezaju je prišlo do odstopanja od pravopisne norme, ker program Transcriber 1.5.1 ne omogoča razlikovanja med vezajem in pomišljajem, zato je bil vezaj uporabljen tudi namesto pomišljaja.

Tabela 3: Primerjava pogovornih zapisov in ročno popravljenih standardiziranih zapisov (primeri iz javnega, nejavnega in parlamentarnega govora).

Sklop baze Artur	Pogovorni zapis	Ročno popravljeni stand. zapis
Javni govor	Nekoč je bil človk navaj@n na trpljenje in je z večjo lahkoto šel skozi življenje. Sodob@n človk veliko hitreje podleže težavam, ker nanje ni pripravlj@n.	Nekoč je bil človek navajen na trpljenje in je z večjo lahkoto šel skozi življenje. Sodoben človek veliko hitreje podleže težavam, ker nanje ni pripravljen.
Javni govor	Če gledamo starostno st@nd@rdizirano stopnjo, s k@tero mi primerjamo bremena med državami, je to okol [dvejs@t] na [sto tisoč]. S tako stopnjo smo mi začel v@ začetku [še@s@tih] let prejšnj@ga stoletja. In zakaj so p@ršli h n@m?	Če gledamo starostno standardizirano stopnjo, s katero mi primerjamo bremena med državami, je to okoli [dvajset] na [sto tisoč]. S tako stopnjo smo mi začeli v začetku [šestdesetih] let prejšnjega stoletja. In zakaj so prišli k nam?
Nejavni govor	#Eee, mogoče na konci pri@m na isti nivo sposobnosti oziroma še na večji nivo, zato k@r bon se mogo san navučiti uravnavate moč, #e, k@r mi ne dobena elektronika pomagala, al pa kupiš novejši motor,	#Eee, mogoče na koncu pridem na isti nivo sposobnosti oziroma še na večji nivo, zato ker bom se mogel sam naučiti uravnavati moč, #e, ker mi ne nobena elektronika pomagala, ali pa kupiš novejši motor,
Nejavni govor	Ge sem se včakala, ja, jaz s@m šla f penzijo. #Eee, s@m tudi v Muri napredovala potem, sledi s@m šla f kontorlo, s@m f kontroli delala, takrat se mi je to, tudi lepo bilo, ne. #Eee v@ ... Pri plači se mi je poznalo pa še ovačik, s@m z veseljem to delo delala pa opravljala ga, s@m kontrolirala izdelke.	Ge sem se včakala, ja, jaz sem šla v penzijo. #Eee, sem tudi v Muri napredovala potem, sledi sem šla v kontrolo, sem v kontroli delala, takrat se mi je to, tudi lepo bilo, ne. #Eee v ... Pri plači se mi je poznalo pa še ovačik, sem z veseljem to delo delala pa opravljala ga, sem kontrolirala izdelke.
Parlamentarni govor	torej u bistvu fsa kritika ki je bla danes usmerjena u predlok poslanske skupine SDS je na nek način usmerjena v DESUS @k tuki jz z DESUSA opozarjam da so u tej koaliciji torej to kr vi zagovarjate to kr vi zagovarjate kje() kar je eden vaših temelnih točk programa kot stranke u bistvu rezon detre DESUSA	torej v bistvu vsa kritika, ki je bila danes usmerjena v predlog poslanske skupine SDS, je na neki način usmerjena v Desus, ker tukaj jaz iz Desusa opozarjam, da so v tej koaliciji, torej to, kar vi zagovarjate, to, kar vi zagovarjate, kje() kar je ena vaših temeljnih točk programa kot stranke, v bistvu raison d'être Desusa,
Parlamentarni govor	Js bi reku seveda potem ko je bilo potrebno dobiti soglasje za poroštvo tm se je pa pol krepko upela politika pa da ne rečem konkretno tudi nas držauni zbor. Ampak poglejte tudi u takrat naprej ko je biu dejansko #eee poroštveni zakon	Jaz bi rekel, seveda, potem ko je bilo potrebno dobiti soglasje za poroštvo, tam se je pa pol krepko vpela politika, pa da ne rečem konkretno tudi nas, državni zbor. Ampak poglejte, tudi od takrat naprej, ko je bil dejansko, #eee, poroštveni zakon

V splošnem je mogoče reči, da je bil z vidika priprave standardiziranih zapisov pričakovano najmanj problematičen oz. zahteven javni govor (gl. prva dva primera v Tabeli 3). V njem se je namreč pojavljalo zelo malo narečnih in pogovornih besed (oz. besed, ki niso zajete v splošnih slovarjih slovenskega jezika). Večji izziv je predstavljal nejavni govor, v katerem smo identificirali večjo pojavnost narečnih besed, ki se praviloma uvrščajo med težavnejše primere standardizacije.[15] Parlamentarni govor ni bil problematičen v smislu zahtevnosti standardizacije, saj gre za govor v javnem formalnem govornem položaju, za katerega izrazito narečna leksika ni značilna. Parlamentarni govor je bil najzahtevnejši v smislu vloženega časa zaradi nihajoče kakovosti pogovornih zapisov (gl. razdelek 4).

Težavnejše primere standardizacije besed je pripravljavec standardiziranega zapisa sproti beležil in oblikoval predloge za standardizirani zapis (včasih po posvetu s kolegi dialektologi). Te je pregledala in potrdila skupina treh strokovno usposobljenih projektnih sodelavcev.

V spodnji Tabeli 4 so navedeni izbrani primeri s Seznama težavnejših primerov standardiziranega zapisa v bazi Artur (Verdonik, Trojar in Bizjak, 2023b: 14–24).

Tabela 4: Izbor primerov s Seznama težavnejših primerov standardiziranega zapisa v bazi Artur.

Primer	Predlog za standardizirani zapis
ajnfah, ajnfoh	ajnfah
bohlonaj, boglonaj, bohloni	boglonaj
cajt, cet	cajt
dugi (‘dolg’)	dugi
fancy, fensi	fensi
gniliti (‘gniti’, narečno)	gniliti
jajčka (‘jajce’), ž. sp.	jajčka

15 Pri težavnejših primerih se je pripravljavec standardiziranega zapisa posvetoval s kolegi dialektologi. Prim. npr. besedo ***ovačik*** v 2. primeru nejavnega govora v Tabeli 3, ki se ne pojavi v nobenem slovarju na portalu Fran (se pa v Pleteršnikovem slovarju pojavita besedi ***ovače*** in ***ovači***, slednja pa je zabeležena tudi v Slovarju stare knjižne prekmurščine).

Primer	Predlog za standardizirani zapis
kejpop	K-pop
mezmes (‘vmes’)	mezmes
obično	obično
parajt, berajt (‘pripravljen’)	berajt
ušeta (‘ušesa’)	ušeta
vjutro	vjutro

V grobem je mogoče težavnejše primere standardiziranega zapisa razdeliti v tri skupine, in sicer na narečne besede (npr. *mezmes*), pogovorne besede, ki niso vezane na eno narečje ali manjšo skupino narečij (npr. pokrajinskopogovorne in tudi splošnoslovenske pogovorne besede, npr. *ajnfah*, *cajt*), in prevzete besede (npr. *K-pop*).

4 Težave pri izdelavi pogovornih in standardiziranih zapisov, rešitve zanje in priporočila za prihodnje projekte

Najprej velja opozoriti, da je govorna baza Artur nastajala v času pandemije covida-19, ko so bili medosebni stiki in javni dogodki močno omejeni ali celo prepovedani. Omejitev gibanja znotraj občinskih meja je pomenila dodatno oviro pri snemanju in vključevanju govorcev iz različnih regij, na samo kakovost govora pa je vplivalo tudi nošenje obraznih mask. Količino takšnih posnetkov smo zato močno omejili, metapodatek o nošenju mask pa sproti beležili.

Pri izdelavi tako obsežne govorne baze je ključno, da koordiniranje aktivnosti poteka ažurno, da je komuniciranje med deležniki periodično in da se sproti iščejo rešitve za morebitne probleme. S tem se optimizira čas v zaključnih fazah priprave govorne baze, ko se ponovno izvedejo avtomatske validacije podatkov in na njihovi osnovi časovno zelo potratni ročni popravki identificiranih napak.

Pri pripravi pogovornih zapisov za govorno bazo Artur je bilo zelo problematično pogosto menjavanje transkriptorjev in njihovo časovno zamudno uvajanje. Vsak izmed njih je namreč tvoril drug tip napak, kar je bilo treba vsakič znova identificirati. Iskanje in uvajanje

zanesljivega transkriptorja tako ostaja ena od zahtevnejših nalog, ki jo je zaradi neželene fluktuacije sodelavcev pri zapisovanju govora treba večkrat ponoviti.

V nadaljevanju so navedene in opisane nekatere najpogostejše napake, identificirane pri pripravi pogovornih zapisov. Zaradi hitrega in močno strnjenega govora posameznih moderatorjev so meje segmentov transkriptorji postavili preblizu predhodnih fonemov ali tistih, ki so jim sledili. Ko je moderator govoril neprekinjeno več kot 10 sekund, je bilo treba mejo postaviti v tistem delu signala, ko je zajel sapo. Posamezne posnetke radijskega govora smo prejeli že obrezane in v njih ob menjavi govorcev ni bilo vidnih premorov, kar je dodatno otežilo proces segmentiranja. Zaradi prilagoditve segmentacije tehničnim zahtevam za razvoj razpoznavalnika transkriptorji meja segmentov pogosto niso postavili glede na semantično-skladenjski vidik, temveč glede na premore kot prozodično značilnost.

Segmenti pri hkratnem govoru so bili včasih predolgi, zapisi pa nenatančni. V tem smislu smo zaznali nedosleden zapis opornih signalov, kot sta *ja* in *mhm,* saj je njihov natančen zapis pri hkratnem govoru časovno zelo zamuden. Precej popravkov je bilo potrebnih tudi zaradi napačnih označevanj menjav govorcev.

Mestoma zvočna ozadja in zvočni dogodki niso bili označeni ali pa so bili neustrezno označeni zgolj kot deli segmenta in ne segmenti kot celota. Daljši premori so bili pogosto izpuščeni.

Posamezne besede, ki so bile izgovorjene, niso bile zapisane ali pa so bile zapisane napačno. Gre za tip napake, ki ga je izjemno težko odkriti z avtomatskimi preverjanji in ki zahteva veliko dodatnega časa za ročno preverjanje. Če želimo v prihodnje doseči čim višjo kakovost zapisa izrazito narečnega govora, kar zahteva dodatno natančnost pri poslušanju, ga mora zapisati ustrezno usposobljen dialektolog.

Ugotovili smo, da je v tako obsežni bazi, kot je Artur, dosledno zapisovanje polglasnikov (z znakom @) skoraj nemogoče doseči, poleg tega je precej primerov, pri katerih različni transkriptorji različno interpretirajo slišani glas. Podobno je pri fonemu *v*, katerega zapis

je lahko nedosleden, npr. *vprašal* namesto *fprašal*. Dodatno smo v pogovornih zapisih zaznali rabo neustreznih črk, ki niso del slovenskega črkopisa, npr. *q* in *y*.

Izkazalo se je, da je težko ohranjati doslednost pri zapisu neverbalnih in polverbalnih glasov. Pojavili so se neenotni zapisi z eno, dvema ali tremi črkami; včasih so transkriptorji pred ali za njimi vstavili vejico, drugič ne; pogosto na začetku segmenta niso zapisali velike začetnice ali pa so pri označevanju izpustili znak #.

Zaradi prevelikega števila napak pri vstavljanju oklepajev so bili ti pred javno objavo govorne baze iz nje odstranjeni. V oglatih oklepajih so se poleg številnikov pojavljali tudi samostalniki (npr. *tretjina*), pridevniki (npr. *drugi ljudje*) in nedoločni členi (npr. *en lep dan*). Mestoma so bili datumi namesto v zavitih oklepajih zapisani v oglatih.

Zaradi pomanjkljive jezikoslovne izobrazbe transkriptorjev in njihove pogoste fluktuacije so se v zapisih pojavljale različne pravopisne napake, zlasti pri veliki začetnici, zapisu skupaj ali narazen in ločilih, najpogosteje pri vejici in vezaju. Kot posebno problematičen se je tu izkazal parlamentarni govor: na osnovi transkripcij, ki jih izdelujejo v Državnem zboru RS, so pogovorne zapise namreč pripravljali (popravljali) študentje nejezikoslovnih smeri. To je praviloma vodilo do zelo velikega števila napak v pogovornih zapisih (gl. zgleda iz parlamentarnega govora v Tabeli 3, v katerih manjka večina ločil). Slednje pomeni izjemno povečano obremenitev za pripravljavca standardiziranega zapisa, ker mora večino popravkov vnašati dvakrat, tj. hkrati v standardizirani in pogovorni zapis. Težava je bila razrešena tako, da so bila ločila in velike začetnice pri parlamentarnem govoru naknadno avtomatsko (s posebnim skriptom) prenesena iz standardiziranih v pogovorne zapise. Ključno je spoznanje, da je nadzor nad kakovostjo pogovornih zapisov bistvenega pomena; če so namreč pogovorni zapisi kakovostni, pomeni to precej manj oz. hitrejše delo za pripravljavca standardiziranega zapisa.[16] Ker se z vstavljanjem ločil govor dodatno interpretira, je neizbežno, da ločila

16 Pogovorni zapisi, ki so jih pripravljali zunanji izvajalci (podjetja), so se praviloma izkazali za precej kakovostnejše od tistih, ki so jih pripravljali študentje. Za prihodnje projekte se zato priporoča najemanje zunanjih izvajalcev.

skladno s pravopisno normo vstavljajo ustrezno usposobljeni strokovnjaki z jezikoslovno izobrazbo.

Med težavami pri pripravi pogovornih zapisov za bazo Artur je bil tudi zapis dialogov, ki so bili pri nejavnem govoru z namenom zagotavljanja višje kakovosti zvoka posneti 2-kanalno preko dveh mikrofonov. Isti dialog je bilo tako treba zapisati dvakrat, vsakič za drugega govorca, kar pomeni višje finančne stroške. Nastopi pa lahko še dodatna težava, ko se pri hkratnem govoru v ozadju sliši govor drugega govorca.

Dodatni časovno zahteven izziv je bil velik obseg zelo kratkih posnetkov, trajajočih tudi manj kot eno minuto, in zapisov zanje, ki jih je bilo treba vsakič znova prenesti, preimenovati, obdelati in shraniti.

Po pripravi standardiziranih zapisov je bilo sprva načrtovano preverjanje konsistentnosti popravkov (npr. konsistentnosti uporabe izbranih rešitev na Seznamu težavnejših primerov standardiziranega zapisa v bazi Artur, gl. Tabelo 4). Za tovrstno preverjanje in popravljanje napak je zmanjkalo časa, bi se pa mu bilo smiselno posvetiti v prihodnje, saj v standardiziranih zapisih prihaja do nedoslednosti.

Pri izdelavi baze je bil za transkribiranje uporabljen program Transcriber 1.5.1. Njegova odlika je izjemna stabilnost, je pa rokovanje z njim časovno zelo zamudno. Gre za to, da je treba pogovorni zapis in ustrezni standardizirani zapis odpreti vsakega v svojem oknu, nato pa še posebej odpreti zvočno datoteko s posnetkom govora ter ročno vsakič posebej nastaviti kodiranje (na UTF-8) in morebitne dodatne nastavitve.[17] To se je v projektu RSDO izkazalo za večjo pomanjkljivost. Pri pripravi standardiziranega zapisa je bilo namreč pregledanih 2871 parov datotek parlamentarnega govora,

17 Pri programu Transcriber 1.5.1 je zelo moteče in zamudno tudi to, da program pri uporabi smernih tipk na tipkovnici ne omogoča prehajanja kurzorja s sredine izbranega segmenta na sredino segmenta, ki leži tik nad ali tik pod njim. Pri uporabi smernih tipk ↑ in ↓ je prehod med segmenti namreč mogoč le na konec višje ali nižje ležečega segmenta, pri uporabi smernih tipk ← in → pa mora kurzor prepotovati celotno pot do začetka ali konca trenutno izbranega segmenta in šele nato do želenega mesta sredi višje/nižje ležečega segmenta. V praksi to največkrat pomeni veliko zamudnega klikanja z miško za postavljanje kurzorja na ustrezno mesto oz. za premikanje med segmenti. Delo s Transcriberjem je zamudno tudi zato, ker morata biti pogovorni in ustrezni standardizirani zapis odprta vsak v svojem oknu (ker gre za dve ločeni datoteki), pri čemer je treba vsakemu segmentu v pogovornem zapisu ročno poiskati ustrezni vzporedni segment v standardiziranem zapisu.

100 parov datotek javnega govora in 375 parov nejavnega govora (skupno 3346 parov datotek oz. 6692 datotek s standardiziranimi in pogovornimi zapisi). Ob predpostavki, da priprava delovnega okolja za 1 par datotek vzame 2 minuti (tj. odpiranje Transcriberja, odpiranje pogovornega in standardiziranega zapisa, odpiranje avdio datoteke, nastavljanje ustreznih nastavitev in vpisovanje podatkov v evidenco, ker se evidenca o opravljenem delu ni vodila avtomatsko), je bilo zgolj za pripravo na delo (!) potrebnih 111,53 ure ali 14,87 delovnega dne (po 7,5 ure). Ta čas ne vključuje dejanskega popravljanja transkripcij, poleg tega sem ni vključeno še naknadno popravljanje datotek (npr. po pregledu ujemanja v številu pojavnic).

5 Zaključek

Na osnovi izkušenj pri pripravi govorne baze Artur za prihodnje projekte priporočamo prehod na orodje OrthoNormal[18] ali kako drugo primerljivo orodje za transkribiranje, ki omogoča več fleksibilnosti pri delu (podpora za transkribiranje velikega števila kratkih posnetkov, hitro in avtomatsko prehajanje med njimi, avtomatski vzporedni prikaz segmentov v pogovornem in standardiziranem zapisu itd.). Idealno bi bilo vzpostaviti lastno spletno okolje z vgrajenimi rešitvami za obdelavo zvočnih posnetkov, ročno transkribiranje, avtomatsko razpoznavanje govora (avtomatsko generirane transkripcije), projektno vodenje in s črkovalnikom (ki bi preverjal konsistentnost popravkov) ipd. Takšno spletno okolje bi na enem mestu podpiralo izdelavo govornih baz, vanj pa bi integrirali tudi rezultate projekta RSDO: avtomatski razpoznavalnik bi transkriptorju predpripravil transkripcije (npr. pogovorne in/ali standardizirane zapise), ki bi bile že ustrezno segmentirane, transkriptor pa bi jih po potrebi le popravil v skladu s strokovnimi oz. projektnimi zahtevami. Okolje bi samodejno beležilo spremembe in vodilo evidence o opravljenem delu, kar bi bistveno povečalo informiranost vodje projekta in sodelavcem olajšalo vodenje evidenc v projektu. Zasnova in realizacija tovrstnega spletnega okolja bi seveda

18 https://exmaralda.org/de/orthonormal-de/

zahtevali stabilno financiranje, ki bi segalo onkraj sporadičnih triletnih projektov. Šele tako široko zastavljeno dolgoročno zbiranje posnetkov govora bi omogočilo tudi kakovostnejše slovnične in leksikalne opise govorjenega slovenskega jezika, ki je bil v dosedanjih raziskavah izrazito prešibko zastopan.

Projekt RSDO je kot primer dobre prakse omogočil spoznanje in utemeljil zavedanje o tem, da je v slovenskem prostoru na področju procesiranja govora in jezikovnih tehnologij nujno tudi interdisciplinarno povezovanje čim več institucij tako iz akademsko-raziskovalnega okolja kot tudi iz gospodarstva.

Literatura

Verdonik, D., & Bizjak, A. (2023). *Pogovorni zapis in označevanje govora v govorni bazi Artur projekta RSDO*. https://dk.um.si/Dokument.php?lang=slv&id=170009&dn

Verdonik, D., Trojar, M., & Bizjak, A. (2023a). Prednosti in slabosti dvotirnega zapisovanja govora v slovenskih govornih virih = Advantages and Disadvantages of Two-level Speech Transcription in the Slovenian Speech Resources. *Infrastruktura za raziskave govora v humanistiki in jezikovnih tehnologijah: zbornik povzetkov*, 111-114. https://press.um.si/index.php/ump/catalog/book/774

Verdonik, D., Trojar, M., & Bizjak, A. (2023b). *Standardizirani zapis v govorni bazi Artur projekta RSDO*. Univerza v Mariboru. https://dk.um.si/Dokument.php?id=170007&lang=slv

Verdonik, D., Bizjak, A., & Dobrišek, S. (2023). *Opis govorne baze Artur projekta RSDO*. Univerza v Mariboru. https://dk.um.si/IzpisGradiva.php?id=85199

Verdonik, D., Potočnik, T., Sepesy Maučec, M., Erjavec, T., Majhenič, S., & Žgank, A. (2021). *Spoken corpus Gos VideoLectures 4.2 (transcription).* Slovenian language resource repository CLARIN.SI. http://hdl.handle.net/11356/1444

Verdonik, D., & Zwitter Vitez, A. (2011). *Slovenski govorni korpus Gos*. Trojina, zavod za uporabno slovenistiko.

Žganec Gros, J., & Vesnicer, B., (2021). Izbor fonetično uravnoteženih besedilnih predlog za bazo branega govora. V T. Mirtič in M. Snoj (ur.), *1. slovenski pravorečni posvet* (pp. 111–119). Slovenska akademija

znanosti in umetnosti. https://www.sazu.si/uploads/files/publikacije21/Rared2RAZPRAVE.pdf

Žganec Gros, J., Vesnicer, B., Mihelič, A., Trojar, M., Dobrišek, S., Bizjak, A., & Verdonik, D. (2023). Izbor povedi za govorno bazo Artur v projektu Razvoj slovenščine v digitalnem okolju. Projektno poročilo DS2-2.1.1. https://dk.um.si/IzpisGradiva.php?id=85200

Prihodnost korpusa Šolar

Špela ARHAR HOLDT
Univerza v Ljubljani, Filozofska fakulteta
Univerza v Ljubljani, Fakulteta za računalništvo in informatiko

Eva PORI
Univerza v Ljubljani, Filozofska fakulteta

Iztok KOSEM
Univerza v Ljubljani, Filozofska fakulteta
Univerza v Ljubljani, Fakulteta za računalništvo in informatiko
Institut »Jožef Stefan«

Povzetek

Razvojni korpusi so skrbno oblikovane digitalne zbirke avtentičnih besedil, ki omogočajo vpogled v jezikovni razvoj mlajših naravnih govorcev določenega jezika. Pisni razvojni korpusi, kakršen je za slovenščino korpus Šolar, vključujejo primere pisanja osnovnošolskih in srednješolskih učencev, pogosto skupaj s popravki jezikovnih težav, in kot taki predstavljajo empirično osnovo za raziskave s področja jezikovnega usvajanja in didaktike, za pripravo učnih gradiv, vaj, testov, učnih množic za strojno procesiranje naravnega jezika in razvoj orodij, ki opismenjevanje in pismenost podpirajo. Prispevek predstavlja značilnosti slovenskega razvojnega korpusa v primerjavi s podobnimi viri za druge jezike, njegov razvojni krog in številne novosti, ki jih je k metodologiji gradnje prispevalo delo na projektu Razvoj slovenščine v digitalnem okolju. Glavne novosti so izboljšana pravna podlaga za zbiranje besedil, uporabniško prijazen portal za oddajo besedil, orodje CJVT Svala za transkripcijo, anonimizacijo in označevanje popravkov ter izboljšani korpusni format. Ob pojavu generativne umetne inteligence in jezikovnih orodij, ki uporabnicam in uporabnikom pomagajo pri pisanju in komuniciranju izpostavimo spremljanje razvoja (in morebitnega upada) jezikovnih kompetenc kot ključno za nadaljnje

delo in ponudimo strategijo prihodnjega razvoja korpusa Šolar in sorodnih podatkovnih virov.

Ključne besede: razvojni korpus, Šolar 3.0, metodologija korpusne gradnje, CJVT Svala, portal za zbiranje besedil

Abstract

Developmental corpora are carefully designed digital collections of authentic texts that provide insights into the development of younger native speakers' language skills. Written developmental corpora, such as the Šolar corpus for Slovene, include examples of writing by primary and secondary school students, often accompanied by language corrections, and as such, provide an empirical basis for research in the fields of language acquisition and didactics, for the development of teaching materials, exercises, tests, training sets for natural language processing, and for the development of tools that support and develop literacy. The paper presents the characteristics of the Slovene developmental corpus compared to similar resources for other languages, its development cycle and the many innovations of the corpus-building methodology developed under the umbrella of the Development in the Digital Environment project: an improved legal basis and a user-friendly portal for text collection, the CJVT Svala tool for transcription, anonymisation and annotation of corrections, and an enhanced corpus format. With the emergence of generative artificial intelligence and language tools that help users write and communicate, we highlight the monitoring of linguistic competencies' development (and possible decline) as crucial to future work and offer a strategy for the further development of the Šolar corpus and related data resources.

Keywords: developmental corpus, Šolar 3.0, corpus building methodology, CJVT Svala, portal for text collection

1 Uvod

Razvojni korpusi (ang. *developmental corpora*, Leech, 1997:19) so premišljeno grajene digitalne zbirke avtentičnih besedil, ki ponujajo vpogled v razvoj jezikovnih kompetenc pri mlajših naravnih govorcih

in govorkah določenega jezika.[1] V prispevku se osredotočamo na pisne razvoje korpuse, ki tipično zajemajo primere osnovnošolskega in srednješolskega pisanja, pogosto pa tudi oznake jezikovnih težav, ki se v teh besedilih pojavijo. Ti korpusi predstavljajo empirično osnovo za raziskave s področja jezikovnega usvajanja in didaktike, za pripravo učnih gradiv, vaj, testov, učnih množic za strojno procesiranje naravnega jezika in razvoj orodij, ki opismenjevanje in pismenost podpirajo.

Zaradi vsega naštetega so razvojni korpusi med pomembnejšimi specializiranimi jezikovnimi viri in del temeljne jezikovne infrastrukture. Mogoče pa je predvideti, da bo zanimanje za tovrstne vire in metodologijo njihove priprave v prihodnje še naraščalo, kot posledica napredka na področju generativne umetne inteligence in raznovrstnih jezikovnih orodij, ki uporabnicam in uporabnikom pomagajo pri pisanju in komuniciranju. Po pojavu tehnologij, ki ustvarjajo besedila skupaj s piscem ali namesto njega, namreč postaja vprašanje spremljanja razvoja (in morebitnega upada) človeških jezikovnih kompetenc še bolj pereče in temeljno kot v preteklosti.

V evropskem prostoru je mogoče najti kar nekaj primerov razvojnih korpusov, ki vsebujejo pisna besedila osnovnošolcev in/ali dijakov, ne gre pa prezreti, da je takšnih virov bistveno manj od korpusov z besedili govorcev, ki se določenega jezika učijo kot drugega/tujega (ang. *learner corpora*). Za angleščino so na voljo korpusi LUCY (Sampson, 2003), LOCNESS (Granger, 1998) in obsežna zbirka novozelandskih esejev, ki jih je zbrala Parr (2010). Za nemščino so za osnovnošolsko pisanje na voljo korpusi H1, H2, E2, ERK1 (Berkling, 2016; 2018) in Litkey (Laarman-Quante idr., 2019), za srednješolsko pisanje pa korpus KoKo (Abel idr., 2014). Za italijanščino so na voljo korpus CItA (Barbagli idr., 2016), trojezični LEONIDE (Glaznieks idr., 2022) in korpusi esejev, ki so jih zbrali Marconi idr. (1993) za osnovnošolsko in Borghi (2013) za srednješolsko raven. Številni razvojni korpusi za

1 Definicija je nekoliko poenostavljena, saj vemo, da razvoj jezikovnih kompetenc poteka skozi celo življenje (ni prisoten le pri mlajših govorcih in govorkah), prav tako ni povsem natančno govoriti (le) o naravnih govorcih, saj so v osnovnih in srednjih šolah, kjer se besedila za razvojne korpuse tipično zbirajo, tudi avtorice in avtorji, ki jim jezik okolja ni nujno prvi ali edini.

francoščino so na voljo prek portala È:CALM (Ho-Dac idr., 2020). Med novejšimi je mogoče omeniti tudi zbirke besedil za islandščino (Arnardóttir idr., 2021, Ingason idr., 2021) in DOESTE (Martins idr., 2020), ki vsebuje besedila v evropski in brazilski portugalščini. Razen zadnjih dveh in LUCY, ki so po obsegu nekoliko manjši, ter korpusa Parr, ki prinaša skoraj 21.000 esejev, zajemajo navedeni viri nekje med 2.500 in 5.000 besedil oziroma med 100.000 in 1.000.000 pojavnic. Veliko jih vsebuje tudi jezikovne oznake na osnovnih nivojih (tokenizacija, lematizacija, oblikoskladnja) ter popravke, ki so jih vnesli raziskovalci, ki so korpus gradili.

Šolar, razvojni korpus za slovenščino, je v veliki meri primerljiv, mestoma presega prakse iz tujine, mestoma pa se jim tudi odmika. Največja konceptualna razlika je v odločitvi, da se v korpus vključijo avtentični učiteljski popravki, s pomočjo katerih je mogoče opazovati podajanje povratne informacije neposredno v kontekstu razvoja pisnih kompetenc. Od tujih primerov avtentične popravke učiteljev oz. profesorjev vključuje korpus Chyby (Pala idr., 2003), ki pa se za razliko od Šolarja posveča pisanju na univerzitetni ravni. Korpus Šolar, ki nastaja in se razvija že od leta 2012 (Kosem idr., 2012; 2016), je svojo zadnjo nadgradnjo doživel leta 2023. Prenovljena različica 3.0 je izšla pod okriljem projekta Razvoj slovenščine v digitalnem okolju,[2] kjer so bili zasnovani in evalvirani postopki ter orodja za kontinuiran razvoj korpusa Šolar, posredno pa tudi drugih korpusov, ki vsebujejo jezikovne popravke.[3]

Nekatere novosti, ki so nastale na projektu Razvoj slovenščine v digitalnem okolju, so bile omenjene v prispevkih Arhar Holdt idr. (2022a) ter Arhar Holdt in Kosem (2023), vključene so tudi v projektno poročilo (Arhar Holdt idr., 2023). Vendar rezultati do sedaj še niso bili celovito in pregledno predstavljeni z vidika prispevka za raziskovalno skupnost, razvoja discipline in samega korpusa. V tem prispevku najprej predstavimo razvojni krog korpusa Šolar, sledi popis

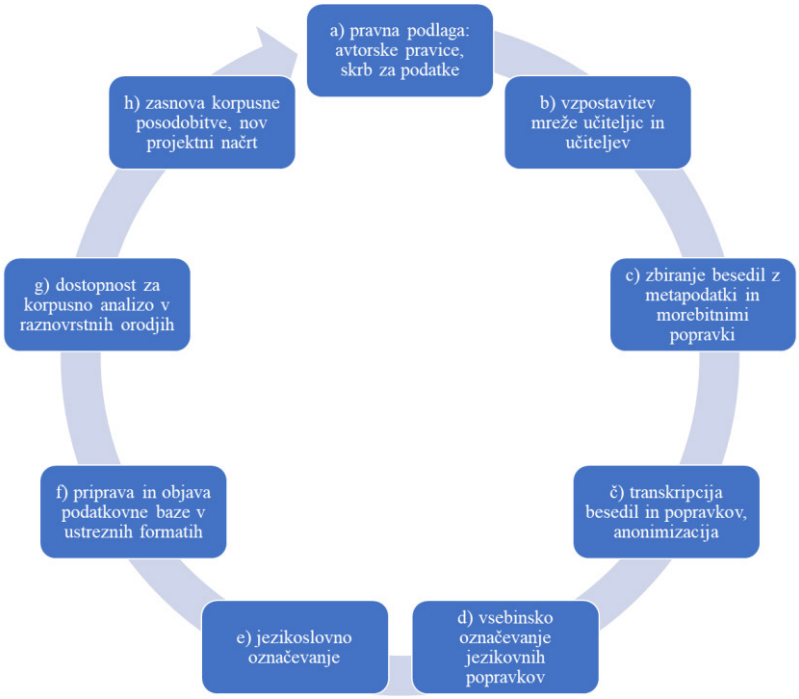
2 Cilj projekta je bil zadovoljiti potrebe po računalniških izdelkih in storitvah s področja jezikovnih tehnologij za slovenski jezik za raziskovalne organizacije, podjetja in širšo javnost. Spletna stran projekta z dostopom do rezultatov: https://slovenscina.eu/.

3 V našem prostoru gre omeniti še korpus slovenščine kot tujega jezika KOST (Stritar Kučuk, 2022) in korpus lektorskih popravkov Lektor (Popič, 2014).

postopkov in orodij za nadaljnje zbiranje, kratka predstavitev Šolarja 3.0 in njegove dostopnosti, zaključujemo pa z naborom prioritet za nadaljnje delo in strategijo za prihodnji razvoj korpusa.

2 Razvojni krog korpusa Šolar

Gradnja korpusa Šolar poteka primerljivo z drugimi korpusnimi viri, določene specifike kaže Slika 1. Zbiranje besedil poteka s pomočjo učiteljske skupnosti, besedilodajalci so učenci oz. dijaki, zato je pomemben del gradnje vzpostavitev mreže in motivacija za sodelovanje učiteljske skupnosti. Učitelji oz. učiteljice morajo urediti pogodbo s šolo, ki dovoljuje zbiranje gradiva, prav tako pa zbrati soglasja avtorjev in avtoric oz. njihovih zakonitih zastopnikov. Poskrbijo tudi za oddajo besedil in vseh želenih metainformacij o njih. Ko je gradivo zbrano, ga pretvorimo v korpusna besedila, kar vključuje



Slika 1: Razvojni krog korpusa Šolar.

transkripcijo (kadar so izvorna besedila napisana na roko, kar trenutno velja za večino primerov), anonimizacijo osebnih podatkov, ki se lahko v besedilih pojavljajo, vnos in označevanje jezikovnih popravkov, jezikoslovno označevanje in izdelavo korpusne baze v končnem formatu oz. formatih. Baza mora biti umestljiva v orodja za analizo, med katerimi so zlasti konkordančniki in druga orodja za vizualizacijo ter ekstrakcijo korpusnih podatkov. Za vse sinhrone jezikovne vire je ključno kontinuirano nadgrajevanje in posodabljanje gradiva; za razvojne korpuse, kjer so longitudinalne raziskave posebej zaželene, pa to velja še toliko bolj. Zasnova projektne nadgradnje ponuja tudi priložnost za oceno uporabljenih postopkov in popis želenih izboljšav.

Kot je popisano v Arhar Holdt idr. (2022a), so bile pri gradnji korpusa Šolar 1.0 in 2.0 na številnih mestih prisotne težave. Pri zbiranju besedil za prvo različico so učitelji in učiteljice pošiljali fizične kopije besedil učencev, njihova kakovost pa se je razlikovala glede na uporabljeni kopirni stroj. Kopirani dokumenti so bili pogosto črno-beli, kar je oteževalo razlikovanje med popravki, ki jih je opravil učitelj, in tistimi, ki je zabeležil učenec sam. Za drugo različico korpusa smo prešli na zbiranje skeniranih besedil, po možnosti barvnih, in s tem na posredovanje PDF-datotek prek spleta, še vedno pa je bilo zamudno zbiranje metapodatkov in spremljanje procesa sodelovanja učiteljskih ekip. Izjemno zamudna je bila tudi priprava korpusnih dokumentov. Zapisovalci in zapisovalke so jezikovne popravke v besedila vpisovali s pomočjo XML-oznak, kar je bilo zahtevno, nepregledno in je vodilo v številne napake. Vsebinsko kategorizacijo jezikovnih popravkov smo pri verziji 2.0 opravljali v za naše namene prilagojenem orodju Sketch Engine (Kilgarriff idr., 2004). Korpus smo morali najprej pretvoriti v format VERT za uvoz v Sketch Engine; tam smo po vsebinskih sklopih opravili revizijo oznak. Med delom smo izvažali korpusne datoteke in jih pretvarjali v format XML, da smo lahko novooznačene kategorije zapisali v korpusne datoteke, spet opravili pretvorbo in korpus uvozili nazaj v Sketch Engine. Zaradi načina dela označevalci in označevalke niso imeli pregleda nad širšim kontekstom označevanega besedila, niso mogli spreminjati

segmentacije popravkov in odpravljati težav, ki niso bile vezane na točno tisto oznako, ki so jo v določenem koraku imeli v analizi.

Po koncu projekta Razvoj slovenščine v digitalnem okolju so koraki korpusne gradnje temeljito nadgrajeni. Na voljo je spletno mesto z informacijami, prenovljenimi pogodbami in repozitorijem za oddajo besedil in metapodatkov, kar učiteljski skupnosti olajša zbiranje in posredovanje gradiva (Razdelek 3). Bistvene izboljšave so na ravni metodologije priprave korpusnih besedil (Razdelek 4): za slovenščino smo lokalizirali in nadgradili uporabniku prijazno in zmogljivo orodje Svala, ki omogoča transkripcijo besedil, označevanje jezikovnih popravkov in pregledno sočasno anonimizacijo potencialno občutljivih osebnih informacij, ki se lahko pojavljajo v besedilih. Šolar 3.0 (Razdelek 5) je na voljo z bogatejšimi jezikoslovnimi oznakami, od katerih so zlasti dragocene skladenjske, ter v novem formatu, ki je v celoti kompatibilen z ostalimi slovenskimi korpusi.

3 Zbiranje korpusnega gradiva

3.1 Pravne rešitve

Po odločitvi učiteljev za sodelovanje pri zbiranju besedil in še pred uporabo portala sledi najprej pravna ureditev sodelovanja, in sicer med raziskovalno ustanovo na eni strani ter šolo in učenci na drugi. S šolo se sklene pogodba o sodelovanju, z učenci oz. njihovimi zastopniki pa pogodba o prenosu ustreznih avtorskih pravic. Podpisane pogodbe (dva izvoda pogodbe s šolo in dva izvoda pogodbe z vsakim avtorjem oz. avtorico šolskih besedil oz. njegovim zakonitim zastopnikom) učitelj oz. učiteljica pred pričetkom sodelovanja pri zbiranju pošlje raziskovalni enoti, ki gradi korpus, kjer jih podpiše še druga stranka in po en izvod vrne na šolo.[4] Na ta način je zbiranje besedil pravno urejeno, saj brez tega zbrano gradivo ne more biti odprto dostopno za nadaljnjo rabo. Za vsa vključena besedila tako obstaja pogodba, ki opredeljuje prenos avtorskih pravic ter načine

4 Izkušnje projektnega sodelovanja s šolami so namreč pokazale, da šolski sistem preferira fizično podpisovanje, da trenutno še ni opremljen ali pa pripravljen na digitalno podpisovanje dokumentov.

hranjenja in procesiranja besedil. Pomembno pri tem pa je, da se pravne rešitve ne osredotočajo le na obdobje trajanja specifičnega projekta, npr. točno določeno šolsko leto, ker to ovira in onemogoča kontinuirano in širše zbiranje besedil. Trenutne pogodbe so na voljo kot Priloge 2–4 v (Arhar Holdt idr., 2023).

3.2 Portal za oddajo besedil

Pravni ureditvi sodelovanja sledi delo s portalom,[5] ki je razvit je z namenom, da bi oddajanje besedil – in vse korake, potrebne za sodelovanje – olajšali tako skupnosti sodelujočih učiteljev kot raziskovalcem, ki besedila za korpus pripravljajo. Pri razvoju portala je bila v ospredju želja, da bo njegova uporaba enostavna in intuitivna, hkrati pa bo vsebovala vse uporabne funkcionalnosti. Uporabniku prijazen vmesnik je osnovni pogoj za sodelovanje čim večjega števila učiteljev, katerih veščina dela z računalnikom in s tem odnos do njega se lahko močno razlikuje. Portal vzpodbuja tudi vzpostavljanje skupnosti sodelujočih besedilodajalcev; saj lahko ekipa učiteljev z iste šole s pomočjo statistik spremlja svoj napredek pri zbiranju besedil, tudi primerjalno glede na druge šole v regiji.[6]

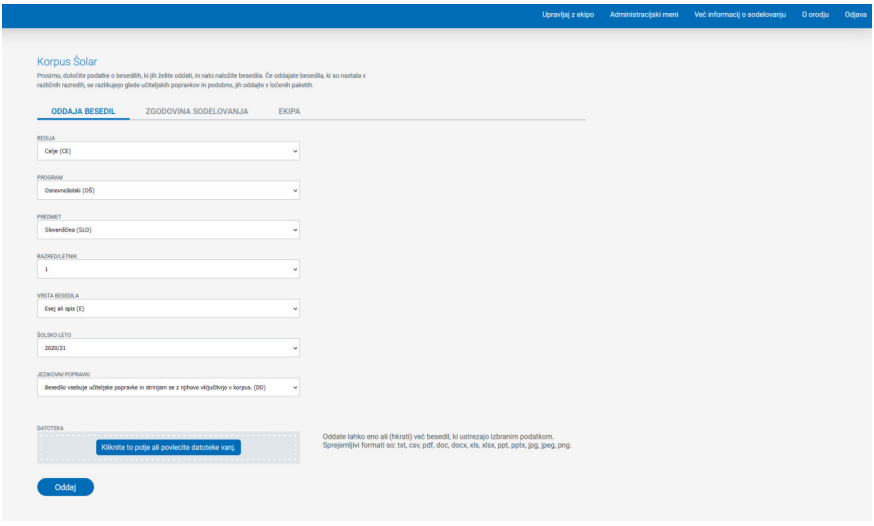
Na vstopni strani portala je na voljo povezava na spletno mesto,[7] kjer je predstavitev korpusa Šolar in pregledna navodila za sodelovanje pri njegovi gradnji. Pri prvi uporabi portala za zbiranje besedil se mora uporabnik registrirati, pri čemer posreduje svoje ime in priimek, naziv institucije, na kateri je zaposlen, e-naslov, določi pa še geslo za vstop v portal in svojo vlogo pri zbiranju: vlogo *Mentor/-ica* izbere, kdor bo zbiral in oddajal šolska besedila, vlogo *Koordinator/-ica* pa tisti, ki bo poleg zbiranja in oddajanja skrbel še za komuniciranje z vodstvom šole in znotraj skupine mentorjev, če je sodelujočih učiteljev z iste šole več. Na izbiro je še *Druga vloga*, ki pokriva opazovalce ali stranske deležnike.

5 Portal je na voljo na spletni strani https://zbiranje.cjvt.si/solar/login/.

6 Učiteljsko skupnost motivira tudi pridobitev točk za napredovanje v nazive, za kar se pripravi potrdilo o sodelovanju pri projektu, ki se na osnovi *Pravilnika o napredovanju zaposlenih na področju vzgoje in izobraževanja v nazive* (Uradni list RS, št. 54/02, 123/08, 44/09, 18/10, 113/20 http://www.pisrs.si/Pis.web/pregledPredpisa?id=PRAV4272) vrednoti na Ministrstvu za izobraževanje, znanost in šport.

7 Dostopno na https://rsdo.slovenscina.eu/zbiranje-besedil-za-korpus-solar.

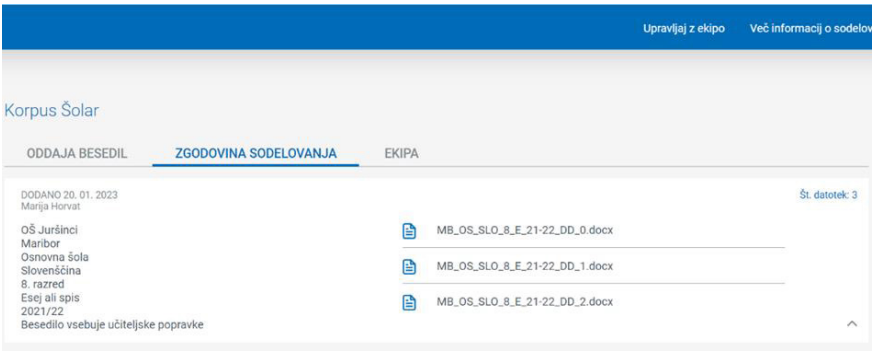
Po uspešni prijavi v portal se uporabnik znajde na strani z osrednjo funkcionalnostjo – oddajo besedil (Slika 2). S pomočjo spustnih seznamov določi vse metapodatke, ki jih potrebujemo za pripravo korpusnih besedil: regijo, v katero se uvršča šola sodelujočega učenca; šolski program (npr. osnovnošolski; splošna in strokovna gimnazija; srednje poklicno izobraževanje); predmet, pri katerem so besedila nastala; razred oz. letnik, v katerem so besedila nastala; vrsto besedila (npr. esej ali spis; praktično besedilo, napisano za oceno; šolski test); šolsko leto, in informacijo, ali besedilo vsebuje jezikovne popravke ter ali sodelujoči učitelj dovoljuje njihovo vključitev v korpus. Sledi oddaja besedil, ki ustrezajo vnesenim metapodatkom, in so lahko v formatih txt, csv, pdf, doc, docx, xls, xlsx, ppt, pptx, jpg, jpeg ali png. Pred oddajo je naložene datoteke mogoče še enkrat pregledati in jih odstraniti ali zamenjati z drugimi. Po potrditvi oddaje se izpiše obvestilo o uspešni oddaji in številu oddanih datotek.



Slika 2: Metapodatki za naložena besedila in okno za oddajo datotek v Zavihku 'Oddaja besedil'.

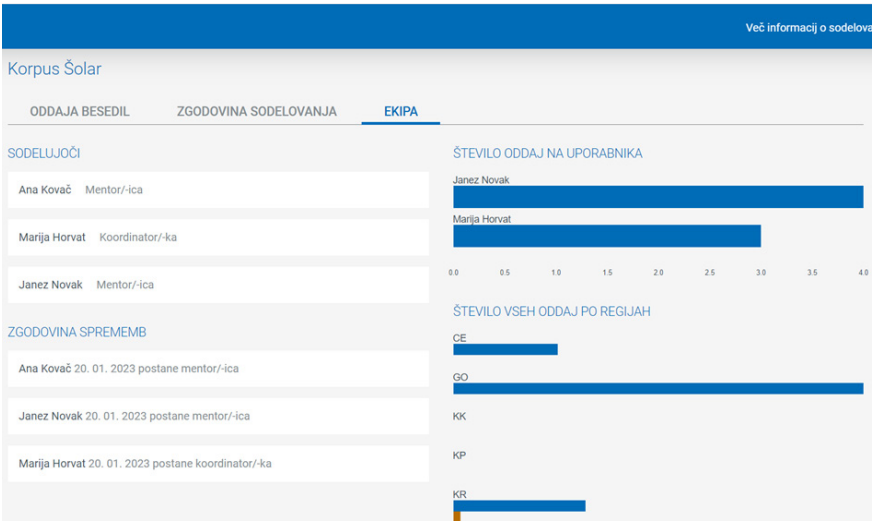
Na naslednji strani portala – v zavihku 'Zgodovina sodelovanja' se beležijo naložene in oddane datoteke uporabnika. Vidne so vse osnovne informacije o oddaji, npr. datum oddaje, ime šole, predmet ter podrobnejše informacije in pogled na naložene datoteke, katerih

imena so tvorjena iz kod vseh izbranih metapodatkov o besedilih (Slika 3).



Slika 3: Razširjen pogled na paket oddanih besedil v zavihku 'Zgodovina sodelovanja'.

V zavihku 'Ekipa' (Slika 4) so shranjeni podatki o sodelujočih članih ekipe. Na levi strani zaslona so izpisana njihova imena skupaj z vlogo, pod tem pa beležena zgodovina sprememb (npr. vlog učiteljev). Desna stran zaslona prikazuje graf s podatki o številu oddanih datotek vsakega člana ekipe in graf, ki izrisuje število vseh oddaj po regijah in vrsti šole. Podatki v obeh grafih se sproti osvežujejo.

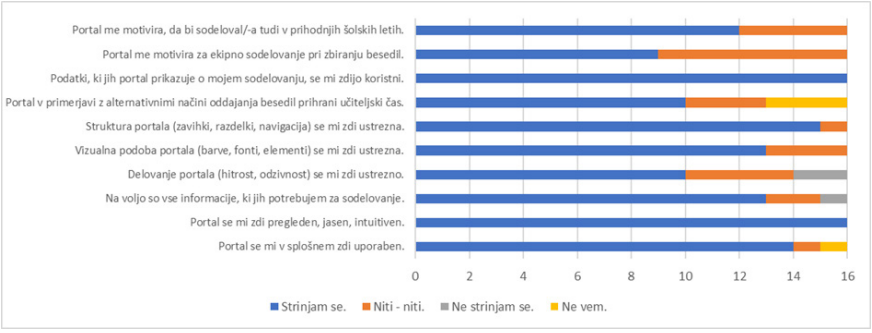


Slika 4: Podatki o sodelujočih članih ekipe v 'Zavihku Ekipa'.

V vmesniku se nahaja še meni za upravljanje z ekipo. Učitelj z vlogo koordinatorja tu najde podatke o članih ekipe v določeni instituciji, omogočeno mu je tudi ročno dodajanje novih članov. Več administratorskih možnosti imajo raziskovalci, ki koordinirajo korpusno gradnjo. Ti lahko potrjujejo in odstranjujejo uporabnike, urejajo imena sodelujočih inštitucij, posodabljajo metapodatke že oddanih vnosov in podobno.

Portal za oddajo besedil je evalviralo 16 učiteljic in učiteljev s 13 šol. Celoten vprašalnik z vsemi odgovori je na voljo kot Priloga 1 v (Arhar Holdt idr., 2023), kjer je tudi opredeljeno, katere identificirane težave so že bile odpravljene in katere čakajo na prihodnji razvoj.

Ocena posameznih strukturnih elementov portala je vključevala vrednotenje funkcionalnosti spletnega mesta z osnovnimi informacijami o sodelovanju, registraciji in prijavi v portal, vnosu podatkov o besedilih, (ne)praktičnosti načina oddaje besedil, strukture portala oz. (ne)funkcionalnosti osrednjih zavihkov. Pri podajanju splošne ocene so se evalvatorji lahko opredelili še do vizualne podobe, delovanja (odzivnosti, hitrosti) portala in motivacijskih elementov za sodelovanje. Na splošno so bili sodelujoči z zasnovo portala zadovoljni, kot kaže Slika 5, za prihodnje delo pa bodo dobrodošli zlasti razmisleki o elementih, ki spodbujajo k dolgoročnejšemu sodelovanju.

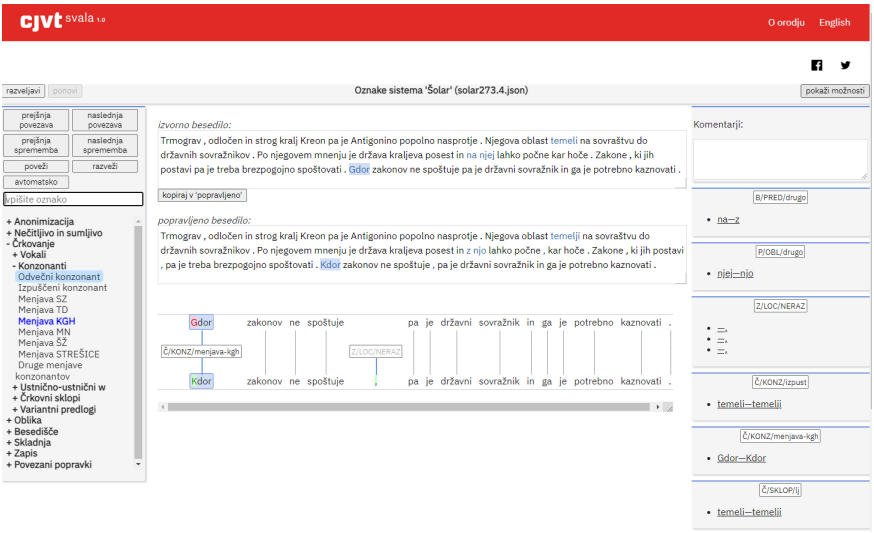


Slika 5: Učiteljska ocena funkcionalnosti na portalu za oddajo besedil.

4 Priprava korpusnih besedil

4.1 Transkripcija, anonimizacija in označevanje popravkov

Orodje CJVT Svala[8] je lokalizirana in adaptirana različica odprtodostopnega orodja Svala, ki je nastalo za pripravo korpusa švedščine kot drugega/tujega jezika (Wirén, 2019). Največja prednost orodja Svala je, da združuje več korakov priprave korpusnih besedil, in sicer transkripcijo, anonimizacijo in označevanje jezikovnih popravkov v besedilih.[9] CJVT Svala 1.0 omogoča označevanje popravkov po dveh sistemih, in sicer po sistemu označevanja korpusa Šolar (Arhar Holdt idr., 2022b) in po sistemu označevanja korpusa KOST (Stritar Kučuk, 2023). Orodje je zasnovano tako, da je mogoče dodati tudi nove označevalne sisteme.



Slika 6: Primer izvornega in popravljenega besedila v vmesniku CJVT Svala 1.0 s sistemom oznak za Šolar.

8 Orodje je prosto dostopno na https://orodja.cjvt.si/svala/, koda je na voljo na repozitoriju GitHub: https://github.com/clarinsi/swell-editor.

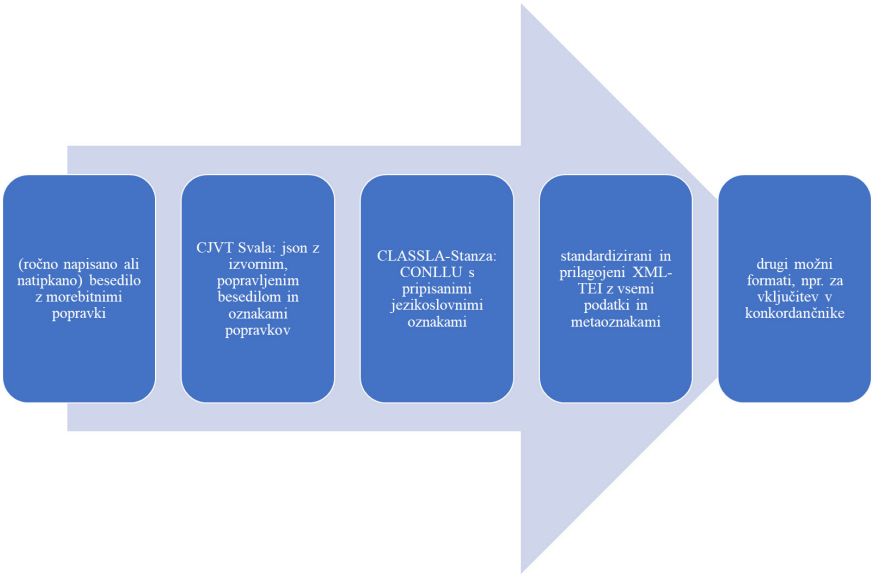
9 Portal SweLL, v katerega je izvorno orodje Svala vključeno (Volodina idr., 2019), skrbi še za vodenje delotokov za urejanje korpusnega gradiva, česar pa za slovenščino trenutno nismo aplicirali.

Način dela z novim orodjem prikazuje Slika 6. Na sliki v gornjem okencu (*izvorno besedilo*) vidimo odstavek avtentičnega besedila iz korpusa Šolar, pod katerim je različica z vpisanimi učiteljskimi popravki. Pod besediloma je t. i. graf povezav, kjer so pojavnice izvornega in popravljenega besedila medsebojno povezane. S klikom na povezavo je mogoče dodati vsebinsko kategorijo jezikovnega popravka, pri čemer se do želene oznake lahko preklikamo s pomočjo menija oznak na levi strani zaslona ali s pomočjo iskalnega okenca nad tem menijem. Primer na sliki kaže popravek besede *gdor – kdor* in pripis oznake črkovanja, specifično za problem menjave med konzonanti *k, g* in *h.* V pomoč pri označevanju so tudi barve – napaka v izvornem besedilu je obarvana rdeče, popravek pa zeleno – in gumbi z ukazi za premik na prejšnjo/naslednjo povezavo, prejšnjo/naslednjo spremembo in za ročno povezavo ali razvezavo neustrezno povezanih pojavnic.

Med urejanjem besedila je mogoče enostavno poskrbeti tudi za anonimizacijo, za kar je v sistemu Šolar predvidena posebna oznaka. Anonimizirati je mogoče s pomočjo kod, npr. *Mirko – XImeX*, ali z uporabo nadomestnih pojavnic, pri čemer je mogoče reproducirati in označiti tudi morebitne jezikovne popravke (npr. *z Mirkotom – z Markom*).

4.2 Jezikoslovno označevanje in korpusni format

Želja in potreba raziskovalne skupnosti je zagotoviti primerljivo jezikoslovno označevanje in standardizirani format temeljnih jezikovnih virov. Za specializirane korpuse, kakršen je Šolar, je ključna metodološko ustrezna povezljivost z referenčnim korpusom, pa tudi drugimi viri iz družine pedagoških korpusov, kamor sodita denimo korpus šolskih učbenikov (Kosem idr., 2022) in mladinske književnosti (Verdonik idr., 2022). Če so korpusni podatki različno označeni in v različnih formatih, so primerjave težje in manj natančne. Treba je torej načrtno skrbeti, da razvojni korpus na ravni jezikovnih oznak in formata sledi standardom, ki se vzpostavljajo v raziskovalnem prostoru, ter da se v primeru novosti tudi ustrezno posodablja.



Slika 7: Cevovod priprave korpusnih besedil za korpuse z označenimi jezikovnimi napakami.

Slika 7 prikazuje trenutni cevovod priprave korpusa Šolar in širše korpusov, ki vsebujejo jezikovne popravke. Proces se prične z besedilom, ki je bodisi ročno napisano ali natipkano. S programom CJVT Svala besedilo uredimo v dve različici, izvorno in popravljeno, ter dodamo oznake popravkov. Tako strukturirani podatki se izvozijo v formatu JSON. Naslednji korak je jezikoslovno označevanje. Trenutno najsodobnejši in najzmogljivejši označevalnik za slovenščino je Classla-Stanza (Terčon & Ljubešić 2023), ki omogoča pripis oznak na številnih nivojih. Po označevanju so datoteke na voljo v formatu CONLLU. Sledi pretvorba v XML TEI, pripravljen posebej za korpuse z jezikovnimi popravki, kjer so korpusna besedila opremljena z oznakami in metapodatki o vrsti in izvoru besedila. Skladno s praksami priprave jezikovnih virov, ki so dostopni prek repozitorija CLARIN.SI, se iz tega formata pripravi različica VERT za vključitev v konkordančnike noSketchEngine in KonText.

Za format TEI[10] smo se odločili že pri pripravi korpusa Šolar 2.0, ki je bil na voljo v različici brez vpisanih popravkov (v celoti

10 Spletna stran iniciative: https://tei-c.org/.

kompatibilen s TEI) in s popravki (prilagojeni TEI). Format, ki je na voljo od korpusa Šolar 3.0 naprej, sledi ločitvi korpusa na tri dele: (jezikoslovno označeno) izvorno besedilo, (jezikoslovno označeno) popravljeno besedilo ter oznake popravkov na spremenjenih delih posameznih povedi. Pri urejanju formata so bile odpravljene težave s segmentacijo napak, ki je predhodno dovoljevala t. i. gnezdene popravke: primere, kjer je bila poleg oznake popravka na določenem segmentu besedila prisotna dodatna oznaka popravka, ki je veljala le za manjši vsebovani del tega segmenta. Gnezdenja popravkov program Svala ne dovoljuje, zato jih tudi novi format ne predvideva. Tovrstne primere, ki so se v različici 2.0 pojavljali v približno 350 odstavkih, smo za Šolar 3.0 ročno popravili in odpravili.

5 Korpus Šolar 3.0

5.1 Sestava korpusa Šolar 3.0

Na projektu je bila pripravljena različica 3.0 korpusa Šolar,[11] ki v vseh pogledih, z izjemo vsebine, prinaša nadgradnjo v primerjavi s prejšnjimi verzijami. Korpus sestavlja 5.485 pisnih izdelkov, ki so jih pri pouku samostojno tvorili učenci slovenskih osnovnih in srednjih šol. Večinoma gre za besedila učencev 7.–9. razreda osnovne šole – vključen pa je tudi manjši vzorec besedil iz 6. razreda – in dijakov vseh letnikov srednje šole. S korpusom torej opazujemo pisno kompetenco šolajoče se populacije starosti 12–18 let.

Vsako besedilo je opremljeno z metapodatki, in sicer: vrsta šole (osnovna ali srednja), predmet, pri katerem je bilo besedilo tvorjeno, razred oz. letnik tvorca besedila, regija, v katero je šola umeščena, in datum nastanka besedila. Del korpusa (2.094 besedil) je označen z učiteljskimi popravki po sistemu oznak, ki ga podrobneje opisujemo v nadaljevanju tega razdelka. Popravki učiteljev so del izvornih pisnih izdelkov učencev, kar pomeni, da odsevajo realno sliko popravljanja šolskih spisov v izobraževalnem procesu.

V Tabelah 1, 2, 3 in 4 predstavljamo vsebino korpusa, pri čemer je vsaka tabela razdeljena v dva dela: v levem, belem delu so

11 Dostopno na http://hdl.handle.net/11356/1589.

predstavljeni podatki za celoten korpus, v desnem, osivenem delu pa podatki samo za besedila z učiteljskimi popravki. Števila in odstotki so vedno podani glede na določeno kategorijo.

Tabela 1 prikazuje razporeditev korpusnih besedil oz. števila besed glede na slovenske regije. Besedila iz severovzhodnih regij (Celje, Maribor, Murska Sobota, Slovenj Gradec) predstavljajo 23,9 % vseh besedil, besedila iz jugozahodnih regij (Gorica, Koper, Kranj, Krško, Ljubljana, Novo mesto, Postojna) pa 76,1 %. Od vseh regij ima ljubljanska regija tako največje število besedil (1495 oz. 27,3 %) kot besed (453,030 oz. 27,7 %). Najslabše zastopani regiji sta murskosoboška z 0,3 % besed in postojnska z 1,7 % besed.

Tabela 1: Število in odstotek besedil ter besed glede na regije v korpusu Šolar 3.0.

Regija	Št. besedil	Odst. besedil	Št. besed	Odst. besed	Št. popravljenih besedil	Odst. popravljenih besedil	Št. besed v poprav. besedilih	Odst. besed v poprav. besedilih
Celje	623	11,4 %	177644	10,9 %	32	0,6 %	11084	0,7 %
Maribor	271	4,9 %	71258	4,4 %	92	1,7 %	27097	1,7 %
Murska Sobota	43	0,8 %	4733	0,3 %	22	0,4 %	3223	0,2 %
Slovenj Gradec	372	6,8 %	97966	6,0 %	102	1,9 %	22313	1,4 %
Gorica	521	9,5 %	263852	16,1 %	321	5,9 %	205477	12,6 %
Koper	111	2,0 %	32898	2,0 %	74	1,3 %	21420	1,3 %
Kranj	380	6,9 %	75524	4,6 %	10	0,2 %	501	0,0 %
Krško	656	12,0 %	205366	12,6 %	147	2,7 %	40637	2,5 %
Ljubljana	1495	27,3 %	453030	27,7 %	467	8,5 %	166221	10,2 %
Novo mesto	924	16,8 %	224862	13,7 %	249	4,5 %	83798	5,1 %
Postojna	89	1,6 %	28274	1,7 %	0	0 %	0	0 %
Skupaj	5485	100 %	1635407	1907562	1516	27,6 %	581771	35,6 %

Tabela 2 prikazuje razporeditev korpusnih besedil in števila besed glede na vrsto šole. Večina besedil prihaja iz različnih vrst srednjih šol, medtem ko osnovnošolska besedila predstavljajo 19,7 % vseh korpusnih besedil oz. 16,3 % besed. Najbolj izstopajo visoki

deželi strokovnih šol in gimnazij, ki predstavljajo 41,2 % besedil in 37,5 % besed oz. 28,2 % besedil in 37,6 % besed. Delež besedil iz poklicnih šol je 9,8 % in predstavljajo 7,2 % besed.

Tabela 2: Število in odstotek besedil ter besed glede na vrsto šole v korpusu Šolar 3.0.

Vrsta šole	Št. besedil	Odst. besedil	Št. besed	Odst. besed	Št. popravljenih besedil	Odst. popravljenih besedil	Št. besed v poprav. besedilih	Odst. besed v poprav. besedilih
Osnovna šola	1081	19,7 %	267146	16,3 %	395	7,2 %	110932	6,8 %
Strokovna šola	2262	41,2 %	613483	37,5 %	574	10,5 %	186809	11,4 %
Poklicna šola	540	9,8 %	117886	7,2 %	143	2,6 %	44878	2,7 %
Gimnazija	1549	28,2 %	615067	37,6 %	404	7,4 %	239152	14,6 %
Neznano	53	1,0 %	21825	1,3 %	0	0 %	0	0 %
Skupaj	5485	100 %	1635407	100 %	1516	27,6 %	581771	35,6 %

Pregled razporeditve besedil in števila besed glede na razred osnovne šole oz. letnik srednje šole, ki ga najdemo v Tabeli 3, prikazuje dokaj uravnoteženo zastopanost. Najbolj izstopa 4. letnik s 25 % besedil oz. 27,9 % besed, kar pa je v skladu s pisno produkcijo, saj je te največ ravno v 4. letniku, ko so tudi besedila daljša. Nižjo zastopanost besedil iz 5. letnika in maturitetnega tečaja lahko pojasnimo s tem, da sta redkeje obiskana.

Tabela 3: Število in odstotek besedil ter besed glede na letnik/razred v korpusu Šolar 3.0.

razred / letnik	Št. besedil	Odst. besedil	Št. besed	Odst. besed	Št. popravljenih besedil	Odst. popravljenih besedil	Št. besed v poprav. besedilih	Odst. besed v poprav. besedilih
6. razred	208	3,8 %	45305	2,8 %	23	0,4 %	7685	0,5 %
7. razred	229	4,2 %	54433	3,3 %	92	1,7 %	22949	1,4 %
8. razred	325	5,9 %	93628	5,7 %	132	2,4 %	43505	2,7 %
9. razred	319	5,8 %	73780	4,5 %	148	2,7 %	36793	2,2 %
1. letnik	1024	18,7 %	317130	19,4 %	427	7,8 %	163610	10,0 %
2. letnik	1018	18,6 %	252775	15,5 %	236	4,3 %	108411	6,6 %
3. letnik	870	15,9 %	308496	18,9 %	252	4,6 %	99299	6,1 %
4. letnik	1373	25,0 %	456196	27,9 %	181	3,3 %	92522	5,7 %
5. letnik	86	1,6 %	21510	1,3 %	25	0,5 %	6997	0,4 %
Maturitetni tečaj	33	0,6 %	12154	0,7 %	0	0 %	0	0 %
Skupaj	5485	100 %	1635407	100 %	1516	27,6 %	581771	35,6 %

Tabela 4 predstavlja razporeditev korpusnih besedil oz. besed glede na tip besedila. Kot lahko vidimo, prevladujejo eseji (58,7 % besedil oz. 77,6 % besed), sledijo pisni izdelki, ustvarjeni pri pouku (15,0 % besedil oz. 6,9 % besed), testi (13,7 % besedil oz. 11,1 % besed) in praktična besedila, napisana za oceno (12,6 % besedil oz. 4,4 % besed).

Tabela 4: Število in odstotek besedil ter besed glede na tip besedila v korpusu Šolar 3.0.

Tip besedila	Št. besedil	Odst. besedil	Št. besed	Odst. besed	Št. popravljenih besedil	Odst. popravljenih besedil	Št. besed v poprav. besedilih	Odst. besed v poprav. besedilih
Pisni izdelek	823	15,0 %	112107	6,9 %	201	3,7 %	31988	2,0 %
Esej	3218	58,7 %	1269793	77,6 %	1280	23,3 %	547169	33,5 %
Praktično besedilo	691	12,6 %	71455	4,4 %	0	0 %	0	0 %
Test	753	13,7 %	182052	11,1 %	35	0,6 %	2614	0,2 %
Skupaj	5485	100 %	1635407	100 %	1516	27,6 %	581771	35,6 %

Korpus je bil jezikoslovno označen s cevovodom CLASSLA v1.1.1[12] na ravneh tokenizacije, stavčne segmentacije, lematizacije, oblikoskladenjskih oznak po sistemu MULTEXT-East v6,[13] odvisnostne skladnje po sistemu JOS-SYN[14] in imenskih entitet.[15] Oznake na nivoju odvisnostne skladnje in imenskih entitet predstavljajo novost v primerjavi z različico 2.0, izboljšana pa je tudi natančnost oznak na ostalih nivojih, saj so bile pripisane z izboljšanim označevalnim orodjem.

5.2 Metodologija označevanja jezikovnih popravkov

Pri procesu odločanja, v katero kategorijo popravkov spada določena težava, so nepogrešljive jasne smernice. Za korpus Šolar 3.0 smo uporabili sistem oznak, ki je bil razvit v različici korpusa 2.0, a smo ga dodatno uredili in nadgradili (Arhar Holdt idr., 2022b). V nadaljevanju predstavljamo osnovno kategorizacijo oznak za jezikovne popravke. Glavnih kategorij je sedem, te pa se hierarhično delijo na podkategorije.

Črkovanje: na to raven uvrščamo učiteljske popravke, ki se nanašajo na zapis glasu ali glasovnega sklopa v besedi. Lahko gre za odvečne, izpuščene ali zamenjane črke (*polen* [laži]* namesto *poln [laži]*; *[je] vrjel** namesto *[je] verjel*; *vstrajen** namesto *vztrajen*) ali črkovne sklope (*zberejejo** namesto *zberejo*; *sprej** namesto *sprejel*; *zastojn** namesto *zastonj*), vzglasje besed na u- oz. v- (*Vsedla** namesto *Usedla*; *uzamejo** namesto *vzamejo*) in variantne predloge (*k* [koncu]* namesto *h*).

Oblika: na ravni oblike označujemo (a) težave na ravni izbire sklona, števila, spola, recimo *[o dekletu, ki je] zanosila** namesto *zanosilo*, in kategorij drugih besednih vrst, (b) popravke besednih oblik, ki niso del standardnih paradigem, recimo *poprimiti** namesto *poprijeti* in (c) dodatne oznake, ki jih pripisujemo le v primeru, da osnovna vsebinska oznaka popravka že obstaja, dodatna oznaka pa

12 https://github.com/clarinsi/classla/
13 https://wiki.cjvt.si/books/04-oblikoskladnja-multext-east
14 https://wiki.cjvt.si/books/06-odvisnostna-skladnja-jos-syn
15 https://wiki.cjvt.si/books/08-imenske-entitete

omogoča združevanje podatkov po drugem kriteriju ali pripis dodatne (poljubne) informacije. Tukaj je denimo informacija o variantnosti besedne oblike, recimo obliki *grada* in *gradu*, ki sta glede na trenutno normo obe legitimni, a je ena nevtralnejša, na kar želi učitelj učenca opozoriti.

Besedišče: na to raven umeščamo popravke besedišča, kar vključuje menjavo ene besede z drugo, pri čemer se lahko besedna vrsta in/ali besednozvezna struktura ohrani ali spremeni. Podkategorije so razdeljene na probleme po besednih vrstah, npr. pri samostalniku ena izmed oznak obeležuje napačno lastno ime (*Lovrenc** namesto *Lovro* [*Kuhar]*), pri glagolu menjavo glagolov *moči-morati* (*[ne bi] moral* [opisati]* namesto *mogel*) ipd. Ločeni so primeri podkategorij z menjavo prek meja besedne vrste (npr. polnopomenske besede v zaimek ali obratno: *Hamlet – on*), zadnja skupina pa prinaša dodatne oznake za zaznamovanost besede, recimo *faks** namesto *fakulteta*.

Skladnja: na tej ravni označujemo popravke, ki posegajo na raven besednozvezne, stavčne in povedne sklanje, npr. popravke besednega reda (*[prepričan je da] generalove ukaze je potrebno* [upoštevati]* namesto *je generalove ukaze potrebno*), skladenjskih struktur (*truplo matere** namesto *materino truplo*), medstavčnih razmerij (*Herod je nekega dne priredil slavje. Salomi slavje* ni bilo preveč po godu.* namesto *Herod je nekega dne priredil slavje, ki Salomi ni bilo preveč po godu.*) itd. Podkategorija dodatne oznake tukaj zaobjema pleonazme, recimo *[Ko se je] vrnila nazaj** namesto *[Ko se je] vrnila*, odvečne, pomensko prazne ali vsebinsko napačne dele.

Zapis: na ravni zapisa označujemo predvsem popravke začetnic (*[v] Nemškem* [jeziku]* namesto *nemškem*) in pisanja skupaj ali narazen (*Nažalost** namesto *Na žalost*). V korpusu so označena tudi mesta napačne stave ločil, kjer prevladuje raba vejice. Skupina ločil je edina, ki ni bila v celoti ročno pregledana in kategorizirana, in sicer zaradi razširjenosti pojava.

Povezani popravki: v to kategorijo uvrščamo vse primere, ki niso samostojen popravek, ampak so posledica primarnega jezikovnega popravka, recimo popravek besedne oblike, ki je le posledica

menjave pred njo stoječega predloga. Da lahko označene podatke ustrezno statistično interpretiramo, je pomembno, da so tovrstni posegi v besedilo ločeni od primarnih popravkov učenčevih jezikovnih izbir. Povezani popravki v osnovi sledijo obstoječi tipologiji, le da del oznake ponazarja, da gre za povezan popravek.

Nečitljivi in sumljivi primeri: posebej so označeni primeri, kjer se v učenčevem besedilu ali učiteljskem popravku pojavlja nečitljiv besedilni fragment, ki ga pri transkripciji ni bilo mogoče interpretirati, recimo *§§§mor§§§*; in primeri, kjer so popravki nenavadni, kjer recimo sumimo, da je prišlo do napake pri transkripciji ali je popravek enak napaki – tem pripišemo oznako za preverbo, ki je začasna in se v končni različici korpusa ne pojavlja.

6 Dostopnost korpusa

Skladno z dobrimi praksami odprtega dostopa do jezikovnih podatkov je korpus Šolar 3.0 kot baza na voljo pod odprto licenco (CC BY-NC--SA 4.0) na repozitoriju CLARIN.SI (Arhar Holdt idr., 2022c). Vključen je tudi v konkordančnike, ki so del infrastrukture CLARIN.SI: KonText, NoSketch Engine Bonito in NoSketch Engine Crystal. Ti konkordančniki omogočajo ločen uvoz (pod)korpusov z izvornimi (‘korpus učenci’) in popravljenimi besedili (‘korpus učitelji’), nato pa v vsaki od različic napredno iskanje, prikaz in izvoz korpusnih podatkov.

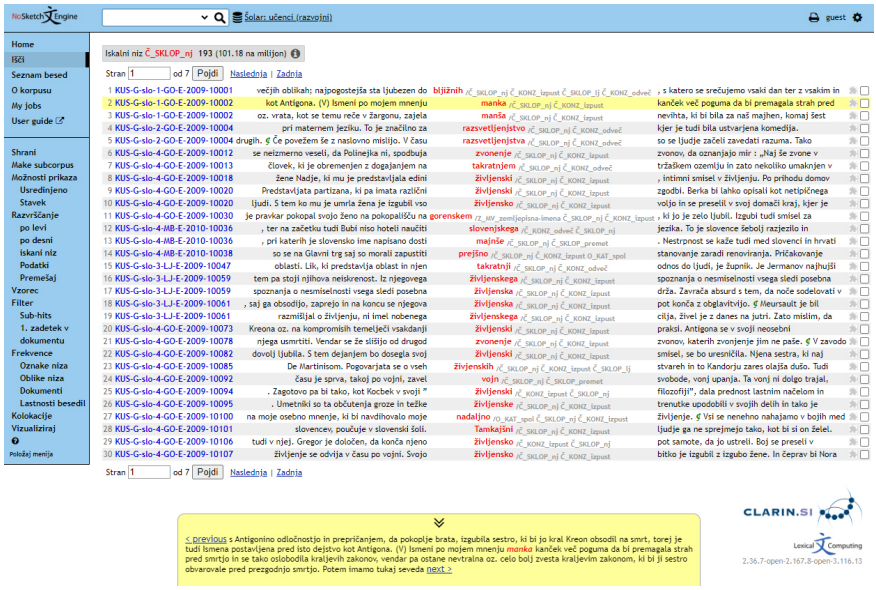
- Šolar 3.0 kot baza: http://hdl.handle.net/11356/1589
- KonText:
  - korpus učenci: https://www.clarin.si/kontext/query?corpname=solar30_orig
  - korpus učitelji: https://www.clarin.si/kontext/query?corpname=solar30_corr
- NoSketch Engine Bonito:
  - korpus učenci: https://www.clarin.si/noske/sl.cgi/first?corpname=solar30_orig&reload=1&iquery=
  - korpus učitelji: https://www.clarin.si/noske/sl.cgi/first?corpname=solar30_corr&reload=1&iquery=

- NoSketch Engine Crystal:
  - korpus učenci:
    https://www.clarin.si/ske/#dashboard?corpname=solar30_orig
  - korpus učitelji:
    https://www.clarin.si/ske/#dashboard?corpname=solar30_corr

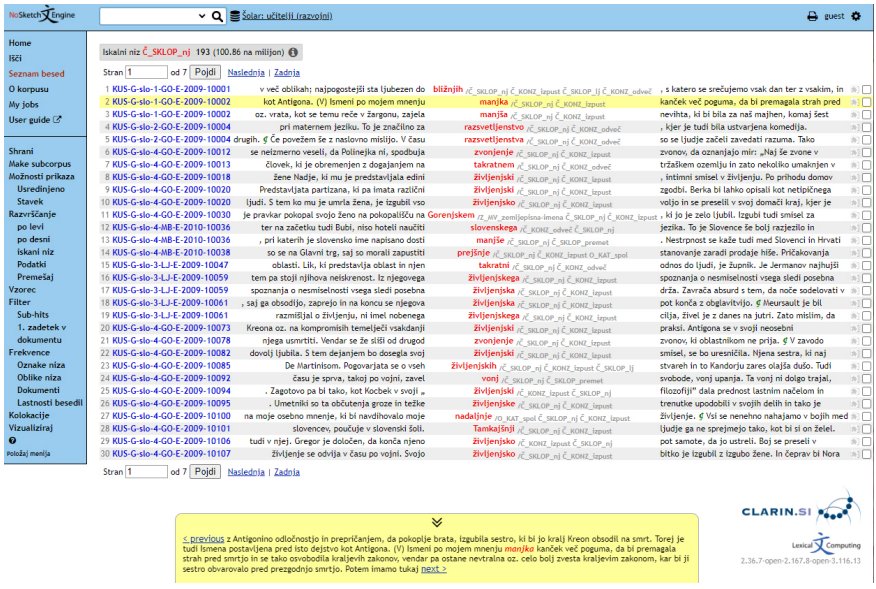
Optimalen in za bodoče delo zaželen bi bil konkordančnik, ki omogoča pregleden skupen prikaz obeh korpusnih različic, vendar že ločena umestitev v zgoraj naštete konkordančnike omogoča številne načine napredne rabe korpusnih podatkov. Osnovna zmogljivost je izdelava konkordančnega niza, pri čemer je mogoče kot parametre iskanja uporabiti raznovrstne v korpusu pripisane oznake. Na Slikah 8 in 9 je za primer prikazan vmesnik NoSketchEngine Bonito, in sicer rezultati iskanja s pomočjo oznake jezikovnega popravka, ki združuje črkovalne težave sklopa *nj*. Kot kaže slika, konkordančnik omogoča enostavno kopiranje zgledov, kar je koristno za pripravo učnih gradiv in vaj. Izvažati je mogoče konkordance, kolokacije, sezname pojavnic in oznak, pri katerih je zlasti ključna možnost primerjave podatkov z drugimi korpusi, ki so vključeni v orodje; kot je bilo omenjeno v Razdelku 4.2, je za korpus Šolar dragocena zlasti možnost primerjave z referenčnim korpusom pisne slovenščine, ki je trenutno Gigafida 2.0 (Krek idr., 2020).

Dodatne možnosti iskanja po jezikovnih oznakah in vizualizacija drevesnic, pripravljenih po sistemu odvisnostne skladnje JOS-SYN, ponuja prostodostopni program Q-CAT. Slika 10 prikazuje iskanje glagolskih oblik, ki so skladenjsko povezane (kot del povedka) z lemo "se". V izbrani povedi so z zeleno prikazana mesta, ki jih je označevalnik prepoznal kot imenske entitete, pod pojavnico so nanizane leme in oznake MSD, z rumenimi povezavami so povezani deli povedka, z zeleno deli besednih zvez (v podrednem in prirednem razmerju), z rdečo pa stavčni členi, pri čemer *ena* grobo ustreza jezikoslovni kategoriji osebka, *dve* predmeta, *tri* določil, ki opredeljujejo lastnosti, in *štiri* ostalih določil, npr. kraja in časa.[16]

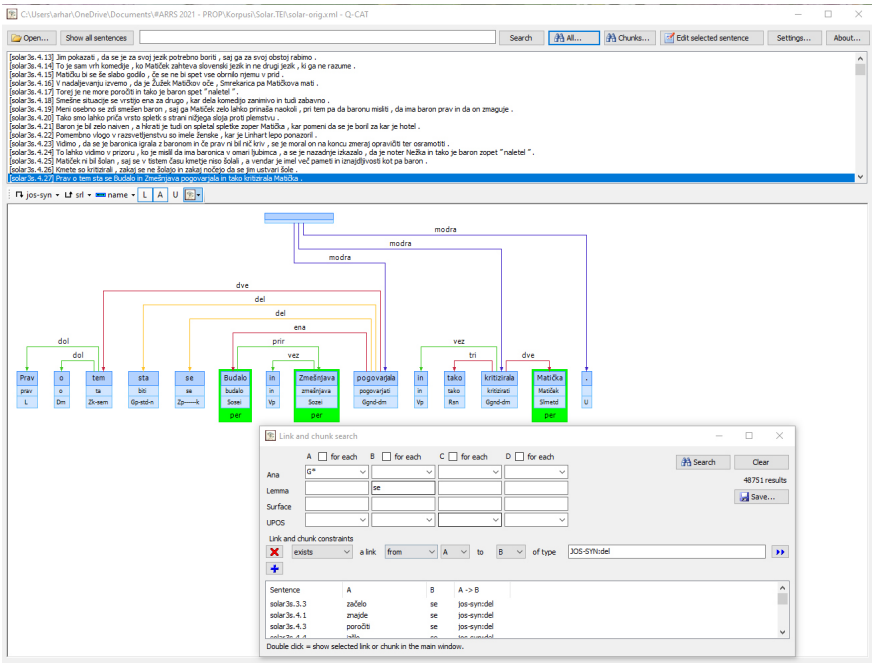
16 Označevalne smernice in predstavitev oznak: https://wiki.cjvt.si/books/06-odvisnostna-skladnja-jos-syn.



Slika 8: Prikaz Šolarja 3.0 – učenci v konkordančniku NoSketchEngine Bonito.



Slika 9: Prikaz Šolarja 3.0 – učitelji v konkordančniku NoSketchEngine Bonito.



Slika 10: Iskanje po skladenjskih oznakah in prikaz označene povedi v programu Q-CAT.

7 Sklep in nadaljnje delo

V prispevku smo predstavili namen in način gradnje razvojnega korpusa za slovenščino. Da bi s pripravo tovrstnih korpusov lahko učinkovito nadaljevali, smo vzpostavili protokole za kontinuirano zbiranje in procesiranje korpusnega gradiva, razvili pa smo tudi nova orodja za ročno označevanje in kategoriziranje jezikovnih popravkov. Nova orodja so odprto dostopna za nadaljnjo rabo, že med projektom pa smo jih uporabili za izboljšavo in dopolnitev obstoječega korpusa.

Prva prioriteta za nadaljnji razvoj korpusa je njegova vsebinska nadgradnja. Poskrbeti je treba za **povečanje njegovega obsega in reprezentativnosti** po regijah, vrsti šole, razredu/letniku avtorja in predmetu, pri katerem je besedilo nastalo. Komplementarno zasnovi korpusa Šolar je treba dodati zbiranje v smer širjenja korpusne vsebine na eni strani proti pisni tvorbi **v nižjih razredih** in na drugi proti **študentskemu pisanju** (slednje je že vključeno v raziskovalni

projekt ARIS J7-3159,[17] vendar le na ravni razvoja metodologije). Želja je zagotoviti **redno korpusno posodabljanje,** kar pomeni zbiranje, vzorčenje in transkribiranje vsako tretje šolsko leto. Da bi slednje lahko uspelo, je treba **dvigniti ozaveščenost in spodbujati šole** k rednemu sodelovanju. Zahtevane kadrovske kapacitete za takšen kontinuiran razvoj so 1 FTE, ki si ga na letni ravni delita jezikoslovec, ki skrbi za zbiranje in pripravo gradiva, ter tehnični sodelavec, ki skrbi za korpusni format in dostopnost v vseh želenih orodjih.

Druga prioriteta, ki je bila vključena v projekt Nadgradnja korpusov za slovenščino kot drugi in tuji jezik KOST in KUUS,[18] je izboljšati dostopnost in povečati izrabo korpusnih podatkov. Za osnovne korpusne analize je vključitev v konkordančnike CLARIN.SI izrednega pomena, vendar obstoječa orodja ne omogočajo polne izrabe bogato označenega gradiva, ki ga prinaša korpus Šolar. V nadaljevanju je treba razviti **specializirani konkordančnik**, ki bo uporaben za vse korpuse z jezikovnimi popravki. Novi konkordančnik mora biti po zasnovi primerljiv z obstoječim, da se omogoči uporabniški prenos znanja, obenem pa mora imeti dodatne možnosti za izrabo metapodatkov, s pomočjo katerih bi lahko natančneje interpretirali posamezne rezultate iskanj po korpusu. Še bolj nujna je možnost preglednega prikazovanja jezikovnih napak skupaj s popravki, zmogljivo iskanje po izvornih in popravljenih oblikah ter klikljive statistike najpogostejših jezikovnih popravkov. Z razvojem specializiranega konkordančnika bo Šolar postal širše uporaben jezikovni vir, zanimiv za pisce učnih gradiv, oblikovalce kurikulov, učitelje ali tiste, ki jih zanima jezik na splošno. Omogočal bo prepoznavo najpogostejših jezikovnih napak, značilnih za govorce določenih prvih jezikov, in s tem pripravo bolj osredotočenih učnih gradiv, pa tudi ustreznejše poudarke v samem pedagoškem procesu. Za najširšo možno rabo je treba zagotoviti tudi **izobraževanja učiteljev** o rabi novega korpusa in o izrabi jezikovnotehnoloških virov pri pouku slovenščine (in širše).

Tretja prioriteta je nadaljnji razvoj metodologije zbiranja. Velik časovni prihranek bi ponudila dopolnitev delotokov z **optičnim**

17 Spletna stan projekta: https://www.cjvt.si/prop/.
18 Spletna stran projekta: https://www.cjvt.si/korpus-kost/projekti/.

branjem ročno napisanih besedil, pri čemer bodo potrebne adaptacije za šolsko rabo (kjer so v besedilih prisotne črkovalne napake in druge značilnosti pisanja, ki se razvija) ter natančno pregledovanje ter popravljanje optično prebranih rokopisov. Druga možnost za pohitritev dela je **strojno podprta identifikacija, vpis in kategorizacija učiteljskih jezikovnih popravkov** v ročno napisanih ali digitalnih besedilih – učiteljski popravki so v določeni meri predvidljivi in ponavljajoči se, kar bi bilo mogoče izkoristiti. Tretja možnost za pohitritev postopka je **vključitev množičenja z didaktično perspektivo** v proces korpusnega grajenja. Pri tem bi bilo mogoče sodelovati s predavatelji, ki poučujejo jezikovno didaktiko in sorodne predmete na terciarni stopnji in bi v transkribiranje ter označevanje popravkov vključili študente in študentke, ki se pripravljajo na podajanje jezikovne povratne informacije učencem in dijakom. Množičenje je mogoče organizirati tudi za širšo populacijo, pri čemer pa je treba zagotoviti ustrezno kontrolo kvalitete in motivacijo za sodelovanje.

V projektu smo ugotovili, da je v nadaljevanju treba nekoliko bolje urediti **pretvorbo besedil iz formata JSON,** ki ga uporablja program Svala, ter končnim želenim XML TEI. Izziv je zlasti zapisovanje ločil, ki se za označevanje v Svali ločijo od besednih pojavnic, za končni format pa jih je treba ustrezno stično oz. nestično spet urediti v izvorno obliko, ki je v šolskih besedilih lahko tudi neskladna s trenutnimi jezikovnimi pravili. Nenazadnje, evalvirati je treba, v kolikšni meri jezikovne napake v korpusnih besedilih vplivajo na **natančnost strojnega jezikovnega označevanja** na posameznih označevalnih ravninah, in zagotoviti ustrezne metodološke nadgradnje ali opozorila.

Predvsem pri ciljih, ki se vežejo na metodologijo, je treba slediti **mednarodnim iniciativam, rešitvam in dobrim praksam,** ne le na področju razvojnih korpusov, ampak širše na področju digitalne humanistike, npr. za metodologijo optičnega branja, transkribiranja itd. Stremeti je treba tudi k oblikovanju **mednarodnih standardov** za gradnjo korpusov z jezikovnimi popravki, saj bi to izboljšalo njihovo primerljivost in olajšalo samo uporabo, lažji bi bil tudi prenos znanja in rešitev. Nenazadnje, zagotoviti je treba raziskave, ki bodo

omogočile **sintetične analize in podatke** za pripravo pedagoških učnih gradiv, jezikovnih priročnikov in orodij. Velik potencial predstavlja **primerjava podatkov** iz korpusa Šolar (šolska produkcija) s podatki, ki reprezentirajo šolsko recepcijo (npr. učbeniki, mladinska književnost, uporabniško generirane spletne vsebine), na drugi strani pa primerjava šolskega pisanja v kontekstih, kjer se slovenščina poučuje kot prvi jezik v primerjavi s poučevanjem slovenščine kot drugega/tujega jezika. Kot smo izpostavili v uvodu, bo pojav **postopkov in orodij generativne umetne inteligence** brez dvoma prinesel tudi nove, še nepredvidene izzive in rešitve, zato je toliko bolj ključno, da tudi za slovenščino novim možnostim in spoznanjem karseda hitro sledimo.

Zahvala

Projekt Razvoj slovenščine v digitalnem okolju 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). Projekt Empirična podlaga za digitalno podprt razvoj pisne jezikovne zmožnosti (J7-3159) in program Jezikovni viri in tehnologije za slovenski jezik (P6-0411) sofinancira Javna agencija za znanstvenoraziskovalno in inovacijsko dejavnost Republike Slovenije iz državnega proračuna.

Literatura

Abel, A., Glaznieks, A., Nicolas, L., & Stemle, E. (2014). KoKo: An L1 Learner Corpus for German. In Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC 2014), Reykjavik, Iceland (pp. 2414–2421). European Language Resource Association (ELRA).

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

Arhar Holdt, Š., Lavrič, P., Roblek, R., & Goli, T. (2022b). *Kategorizacija učiteljskih popravkov: Smernice za označevanje korpusa Šolar.* Različica 1.1. Rezultat projekta Razvoj slovenščine v digitalnem okolju. https://wiki.cjvt.si/books/11-jezikovni-popravki-solar/page/oznacevalne-smernice

Arhar Holdt, Š., Rozman, T., Stritar Kučuk, M., Krek, S., Krapš Vodopivec, I., Stabej, M., Pori, E., Goli, T., Lavrič, P., Laskowski, C., Kocjančič, P., Klemenc, B., Krsnik, L., & Kosem, I. (2022c). Developmental corpus Šolar 3.0. Slovenian language resource repository CLARIN.SI. http://hdl.handle.net/11356/1589

Arhar Holdt, Š., & Kosem, I. (2023). Šolar, the developmental corpus of Slovene. PREPRINT (Version 1) available at Research Square. doi: 10.21203/rs.3.rs-3274669/v1

Arhar Holdt, Š., Kosem, I., Pori, E., Munda, T., Stritar Kučuk, M., Voršič, I., Petek, T., Šek, P., & Krsnik, L. (2023). *Šolar 3.0: korpus šolskih pisnih besedil: poročilo projekta Razvoj slovenščine v digitalnem okolju: aktivnost DS1.6*. Ljubljana: Univerza v Ljubljani, Center za jezikovne vire in tehnologije, 2023. https://www.cjvt.si/rsdo/wp-content/uploads/sites/18/2023/06/RSDO_Kazalnik_Solar_v2.pdf

Arnardóttir, Þ., Xu, X., Guðmundsdóttir, D., Stefánsdóttir, L. B., & Ingason, A. K. (2021). Creating an error corpus: Annotation and applicability. In M. Monachini & M. Eskevich (Eds.), CLARIN Annual Conference Proceedings (pp. 59–63). Virtual Edition.

Barbagli, A., Lucisano, P., Dell'Orletta, F., Montemagni, S., & Venturi, G. (2016). CItA: An L1 Italian learners corpus to study the development of writing competence. In Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016), Portorož, Slovenia (pp. 88–95). European Language Resources Association (ELRA).

Berkling, K. (2016). Corpus for children's writing with enhanced output for specific spelling patterns (2nd and 3rd grade). In N. Calzolari idr. (Eds.), Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016), Portorož, Slovenia (pp. 3200–3206). European Language Resources Association (ELRA).

Berkling, K. (2018). A 2nd Longitudinal Corpus for Children's Writing with Enhanced Output for Specific Spelling Patterns and Evaluation In Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018), Miyazaki, Japan (pp. 2262–2268). European Language Resources Association (ELRA).

Borghi, C. C. (2013). *Analisi di produzioni scritte. Valutazioni e misure automatizzate di elaborati scolastici. Tesi di dottorato in pedagogia sperimentale*. Università di Roma.

Glaznieks, A., Frey, J. C., Stopfner, M., Zanasi, L., & Nicolas, L. (2022). LEONIDE: A longitudinal trilingual corpus of young learners of Italian, German and English. *International Journal of Learner Corpus Research*, *8*(1), 97–120.

Granger, S. (1998). The computer learner corpus: A versatile new source of data for SLA research. In S. Granger (Ed.), *Learner English on Computer* (pp. 3–18). Addison Wesley Longman.

Ho-Dac, L. M., Fleury, S., & Ponton, C. (2020). É:calm resource: a resource for studying texts produced by French pupils and students. In Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), Marseille, France (pp. 4327–4332). European Language Resources Association (ELRA).

Ingason, A. K., Arnardóttir, Þ., Stefánsdóttir, L. B., & Xu, X. (2021). The Icelandic Child Language Error Corpus (IceCLEC) Version 1.1, CLARIN-IS, http://hdl.handle.net/20.500.12537/133

Kilgarriff, A., Rychlý, P., Smrz, P., & Tugwell, D. (2004). The Sketch Engine. In G. Williams, & S. Vessier (Eds.), Proceedings of the Eleventh EURALEX International Congress, Lorient, France (pp. 105–116). Universite de Bretagne-sud.

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

Kosem, I., Rozman, T., Arhar Holdt, Š., Kocjančič, P., & Laskowski, C. A. (2016). Šolar 2.0: nadgradnja korpusa šolskih pisnih izdelkov. In T. Erjavec & D. Fišer (Eds.), Proceedings of the Conference on Language Technologies & Digital Humanities, September 29th – October 1st, 2016 Faculty of Arts, University of Ljubljana, Ljubljana, Slovenia (pp. 95–100). Ljubljana University Press, Faculty of Arts. http://www.sdjt.si/wp/wp-content/uploads/2016/09/JTDH-2016_Kosem-et-al_Solar-2-0-nadgradnja-korpusa-solskih-pisnih-izdelkov.pdf

Kosem, I., Pori, E., Žagar, A., & Arhar Holdt, Š. (2022). Corpus of Slovenian textbooks ccUčbeniki 1.0, Slovenian language resource repository CLARIN.SI. http://hdl.handle.net/11356/1693

Krek, S., Arhar Holdt, Š., Erjavec, T., Čibej, J., Repar, A., Gantar, P., Ljubešić, N., Kosem, & I., Dobrovoljc, K. (2020). Gigafida 2.0: the reference

corpus of written standard Slovene. In N. Calzolari (Ed.), *LREC 2020, Twelfth International Conference on Language Resources and Evaluation, May 11–16, 2020, Palais du Pharo, Marseille, France: conference proceedings* (pp. 3340–3345). Paris: ELRA. http://www.lrec-conf.org/proceedings/lrec2020/LREC-2020.pdf

Laarmann-Quante, R., Dipper, S., & Belke, E. (2019). The making of the Litkey Corpus, a richly annotated longitudinal corpus of German texts written by primary school children. In Proceedings of the 13th Linguistic Annotation Workshop, Florence, Italy (pp. 43–55). Association for Computational Linguistics.

Leech, G. (1997). Teaching and language corpora: A convergence. In A. Wichmann, S. Fliegelstone, T. McEnery, & G. Knowles (Eds.), *Teaching and language corpora* (pp. 1–23). Routledge.

Ljubešić, N., & Dobrovoljc, K. (2019). What does Neural Bring? Analysing Improvements in Morphosyntactic Annotation and Lemmatisation of Slovenian, Croatian and Serbian. In Proceedings of the 7th Workshop on Balto-Slavic Natural Language Processing, Florence, Italy (pp. 29–34). Association for Computational Linguistics.

Marconi, L., Ott, M., Pesenti, E., Ratti, D., & Tavella, M. (1993). *Lessico elementare: dati statistici sull'italiano scritto e letto dai bambini delle elementari*. Zanichelli.

Martins, M., Janssen, M., Santos, T., Lopes, R., & Souza, T. (2020). DOESTE v0.5. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University, http://hdl.handle.net/11234/1-3262

Pala, K., Rychlý, P., & Smrž, P. (2003). Text Corpus with Errors. In V. Matoušek, & P. Mautner (Eds.), *Text, Speech and Dialogue (TSD 2003) Lecture Notes in Computer Science* (2807 vol., pp. 90–97). Springer.

Parr, J. M. (2010). A dual purpose database for research and diagnostic assessment of student writing. *Journal of Writing Research*, *2*(2), 129–150.

Popič, D. (2014). Revising translation revision in Slovenia. In T. Mikolič Južnič, K. Koskinen, & N. Kocijančič Pokorn (Eds.), New Horizons in Translation Research and Education 2 (pp. 72–89). University of Eastern Finland. https://erepo.uef.fi/handle/123456789/14340

Sampson, G. (2003). *The LUCY Corpus: Documentation.* University of Sussex. Retrieved August 15, 2023, from https://www.grsampson.net/LucyDoc.html

Stritar Kučuk, M. (2022). KOST med korpusi usvajanja tujega jezika. V N. Pirih Svetina & I. Ferbežar (ur.), *Na stičišču svetov: slovenščina kot drugi in tuji jezik*, *Obdobja 41* (str. 323–334). Ljubljana: Založba Univerze v Ljubljani. https://centerslo.si/wp-content/uploads/2022/11/Stritar-Kucuk_Obdobja-41.pdf

Stritar Kučuk, M. (2023). *KOST 1.0: Priročnik za označevanje napak, delovna verzija*. https://www.cjvt.si/korpus-kost/wp-content/uploads/sites/24/2022/04/Prirocnik-za-oznacevanje-napak-v-KOST-u-2022-04-13.pdf

Terčon, L., & Ljubešić, N. (2023). CLASSLA-Stanza: The Next Step for Linguistic Processing of South Slavic Languages. *arXiv*. doi: 10.48550/arXiv.2308.04255

Verdonik, D., Majninger, S., Dobrovoljc, K., Antloga, Š., Zögling Markuš, A., Voršič, I., Zemljak Jontes, M., Koletnik, M., Valh Lopert, A., Šek Martük, P., Kosem, I., Majhenič, S., Ferme, M., Žagar, A., Arhar Holdt, Š. (2022). Corpus of Slovenian texts for pedagogical purposes ccMAKS 1.0, Slovenian language resource repository CLARIN.SI. http://hdl.handle.net/11356/1692

Volodina, E., Granstedt, L., Matsson, A., Megyesi, B., Pilán, I., Prentice, J., Rosén, D., Rudebeck, L., Schenström, C. J., 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. (2019). SVALA: Annotation of Second-Language Learner Text Based on Mostly Automatic Alignment of Parallel Corpora. In I. Skadina, & M. Eskevich (Eds.), Selected papers from the CLARIN Annual Conference 2018 (pp. 227–239). Linköping University Electronic Press.

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

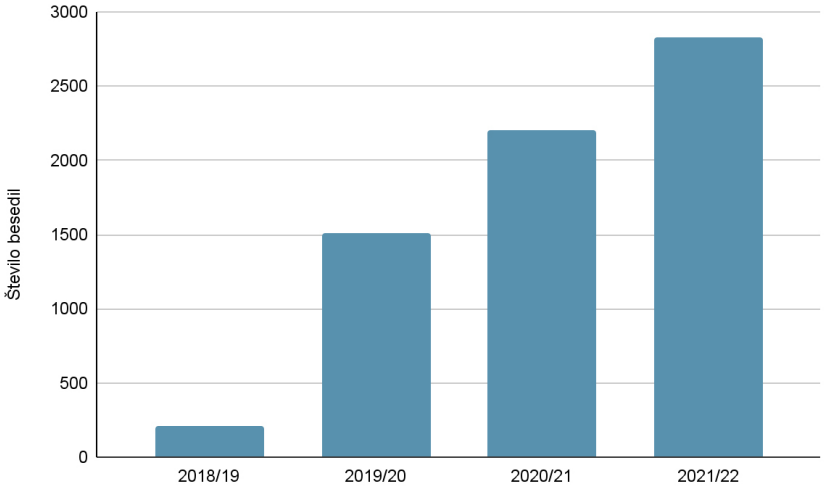
(Centre for English Corpus Linguistics, 2023) poskočilo s 73 korpusov leta 2012 na 191 korpusov leta 2022 (Stritar Kučuk, 2022). Največ, 121, je bilo pisnih korpusov, za katere je tudi najlažje pridobivati besedila, sledili so jim govorni korpusi (44), 24 pa je bilo pisnih in govornih korpusov. Večina 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/

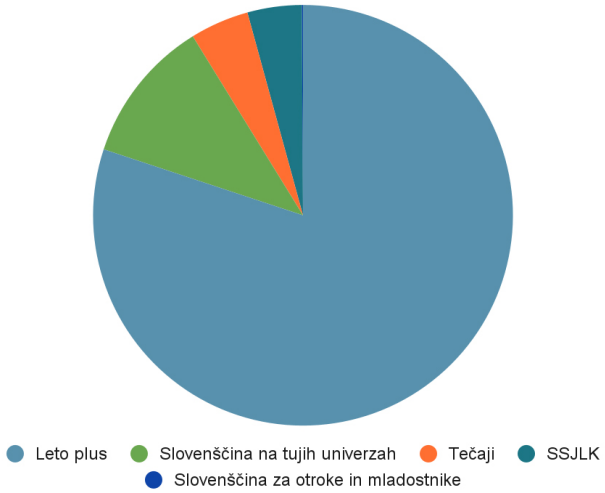
do večjega števila govorcev slovenščine kot drugega jezika in njihovih besedil, ki jih pišejo kot domače naloge ipd. na lektoratih. Zbiranje teh besedil za KOST se je pričelo v študijskem letu 2018/19, kot prikazuje Grafikon 1, pa je bilo nato vsako leto zbranih več besedil.[2]



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.

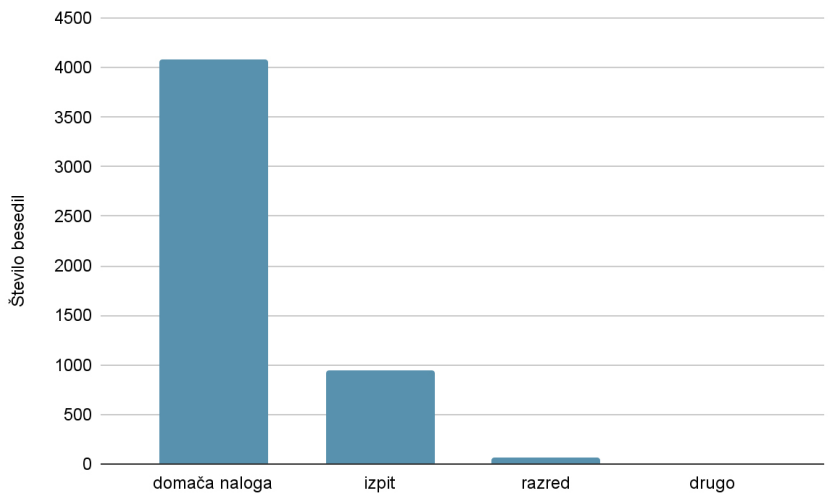
Vsako besedilo, vključeno v KOST, je poimenovano s kodo, ki izurjenemu uporabniku korpusa da nekaj osnovnih podatkov: koda L3-2122-121 denimo pomeni, da gre za besedilo, ki je nastalo pri učitelju s kodo L3 v okviru programa Leto plus v študijskem letu 2021/22, besedilo pa ima zaporedno številko 121. 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.

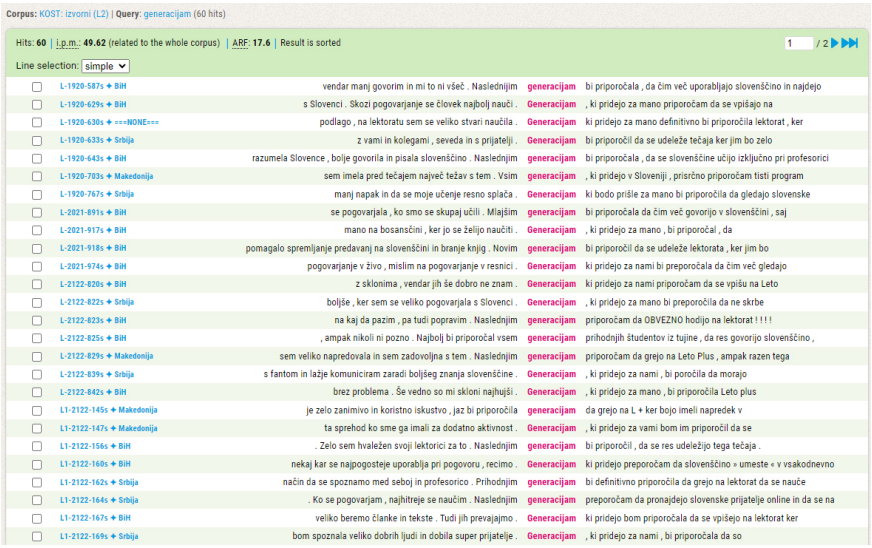
2.1.2 Vrste besedil

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.

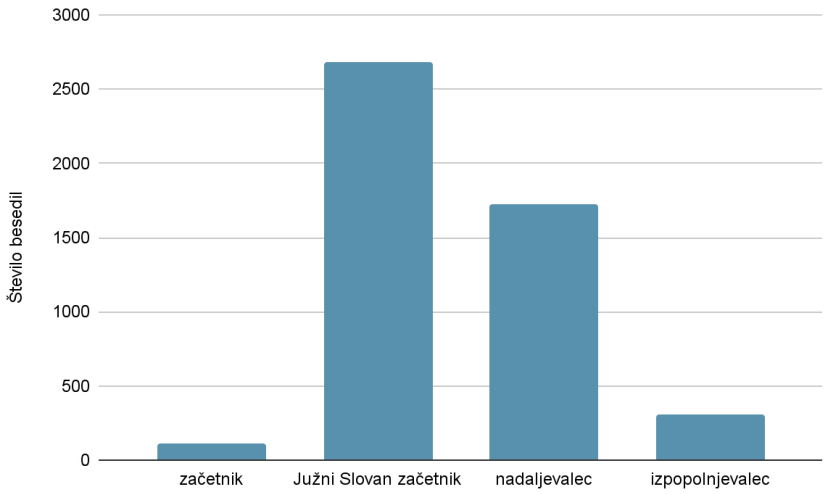
samostojne uporabe jezika. Od praktičnih besedil je zato v KOST vključenih še največ različnih e-pisem, ki so sicer napisana po navodilih, a gre vendarle za več samostojnega pisanja. Takšno navodilo je lahko na primer:

> 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.



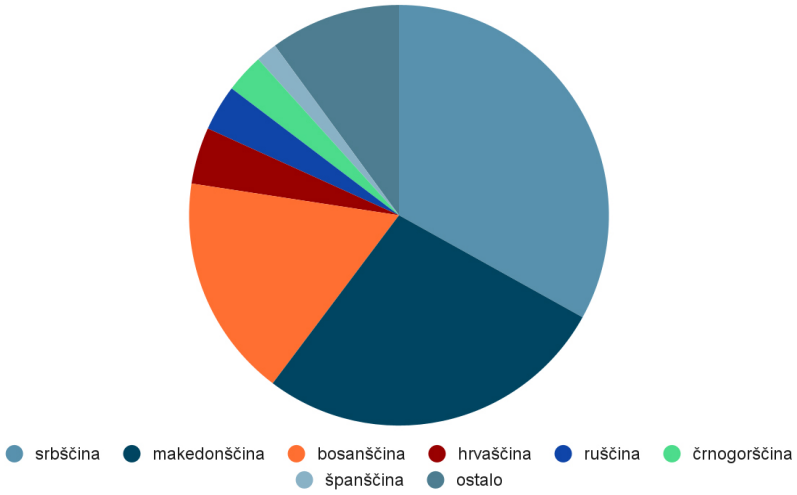
Grafikon 4: Štiri stopnje ocenjene jezikovne zmožnosti tvorcev besedil v slovenščini v KOST 1.0.

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.

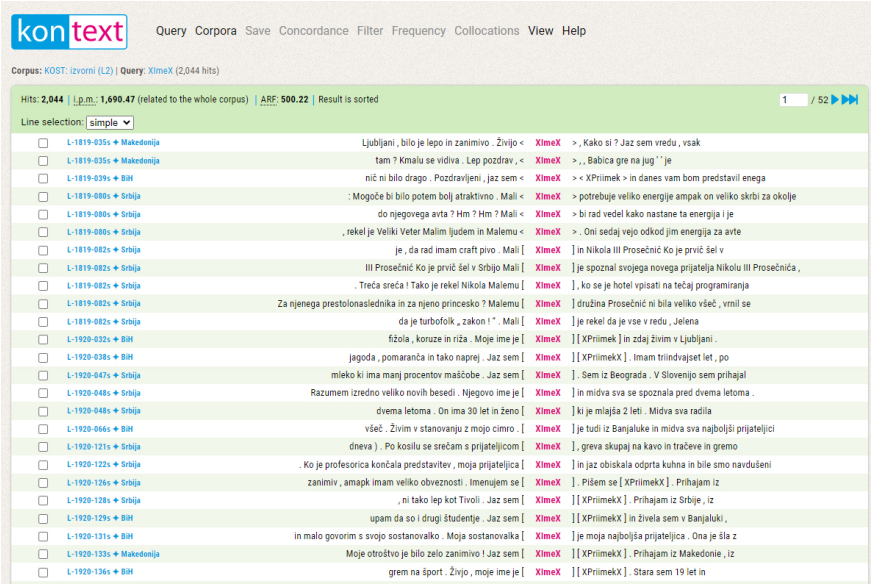
7 Gre za tvorce iz slovenskega zamejstva, pri katerih se srečujemo tudi z vprašanjem, ali naj jih sploh upoštevamo kot govorce slovenščine kot neprvega jezika. Odločitev je vsakokrat individualna.

2.2.2 Varovanje osebnih podatkov

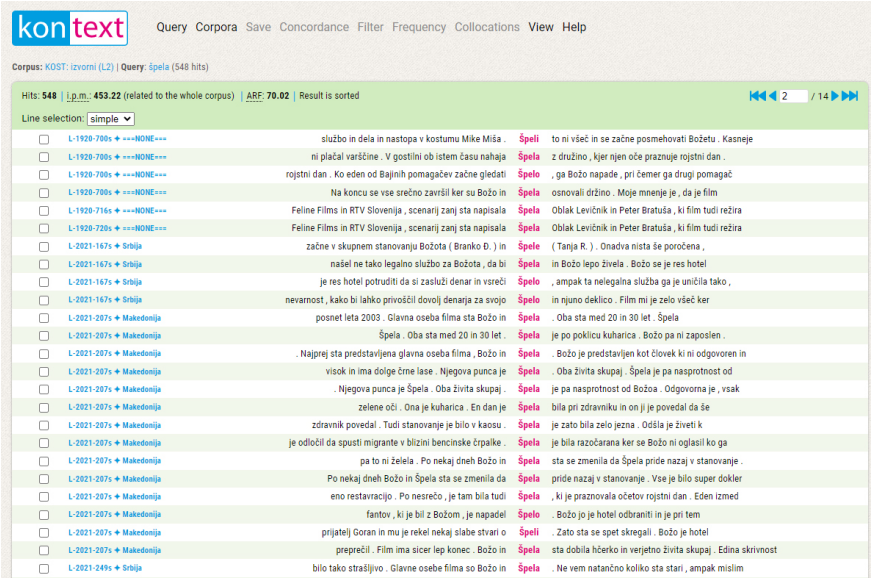
Ker sta ureditev pravic za uporabo podatkov in varovanje osebnih podatkov ključnega pomena, vsi tvorci, katerih besedila so vključena v KOST, podpišejo izjavo, s katero dovoljujejo vključitev svojih besedil. V izjavi dobimo tudi osebne podatke, ki so nujni za analizo korpusnega gradiva: spol, starost, 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.

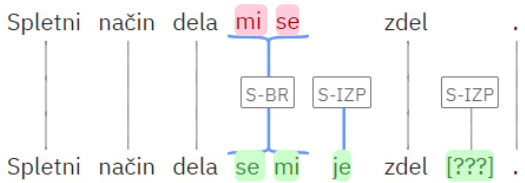
Krovna kategorija	Kategorija napake	Oznaka
Napake zapisa	Ločilo	Z-LOC
	Črkovanje	Z-CRK
	Skupaj/narazen	Z-SN
	Mala/velika začetnica	Z-MV
	Krajšave	Z-KR
Napake besedišča	Samostalnik	B-SAM
	Glagol	B-GLAG
	Pridevnik	B-PRID
	Zaimek	B-ZAIM
	Prislov	B-PRISL
	Predlog	B-PRED
	Veznik	B-VEZ
	Ostalo	B-OST
Napake 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.

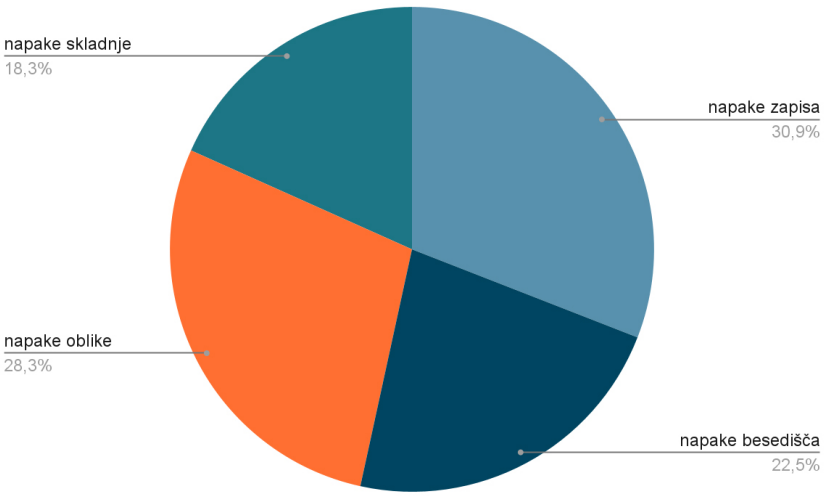


Slika 5: Primer oznake za izpuščeni del besedila, ki ga v KOST-u ne znamo ustrezno popraviti.

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.

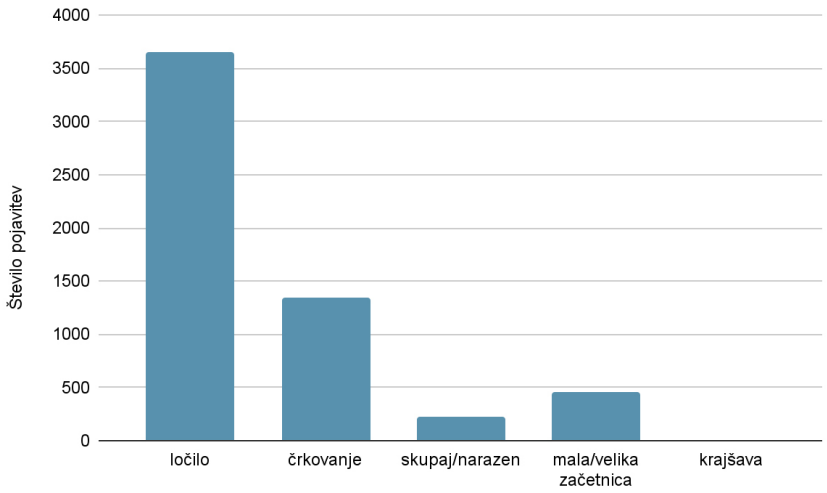
Štiri osnovne kategorije napak so med seboj približno uravnotežene (Grafikon 6). Prednjačijo napake zapisa, najmanj pa je napak



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.

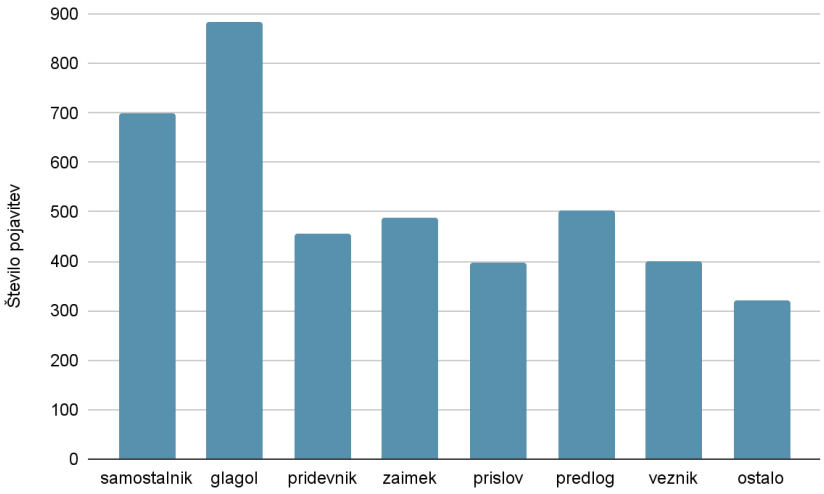
Vpogled v kategorije napak na drugi ravni pokaže, da je pri napakah zapisa (Grafikon 7) največ napak ločil (npr. *Všeč mi je ker je hiša > Všeč mi je, ker je hiša*), kar je tudi daleč najpogostejša med vsemi kategorijami napak. V veliki meri gre za postavljanje vejic. Veliko je tudi napak črkovanja oz. 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.

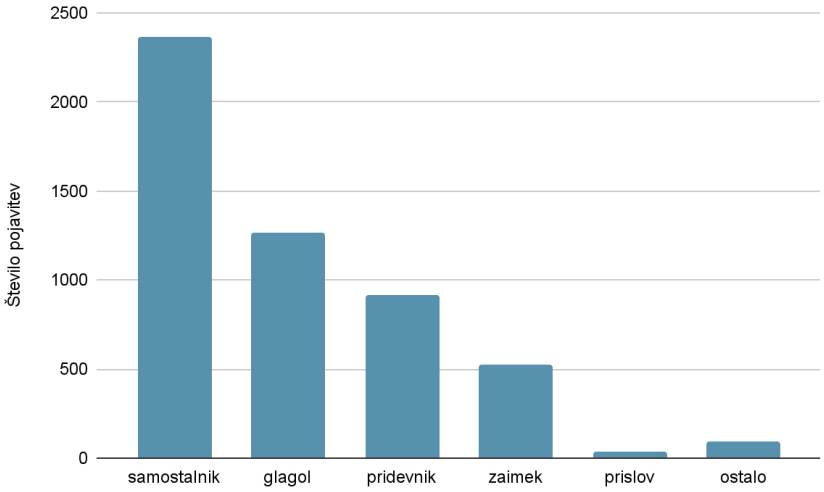
Napake besedišča (Grafikon 8), pri katerih gre za neustrezno leksikalno izbiro, so najpogostejše pri glagolih (npr. *sem se zelo težko naučila na mir > sem se zelo težko navadila na mir*). Sledijo jim samostalniki (npr. *kadiranje > kajenje*). Napake besedišča pri pridevnikih (npr. *družbena oseba >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

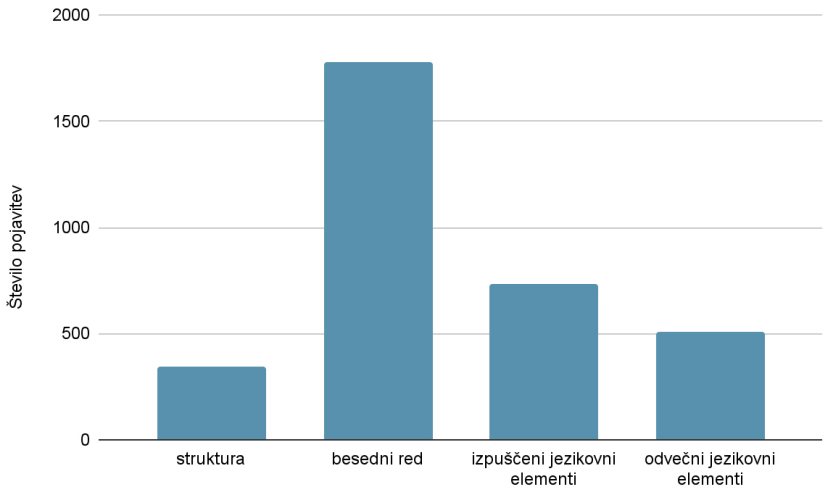
Pri napakah oblike (Grafikon 9), ki se nanašajo na pregibanje besed, je največ napak samostalnika (npr. *v Sloveniju > v Slovenijo*), kar je druga najpogostejša med vsemi kategorijami napak. Sledijo jim glagoli (npr. neuporaba dvojine pri povedku v primeru *sestra in jaz delamo > sestra in jaz delava*), 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.

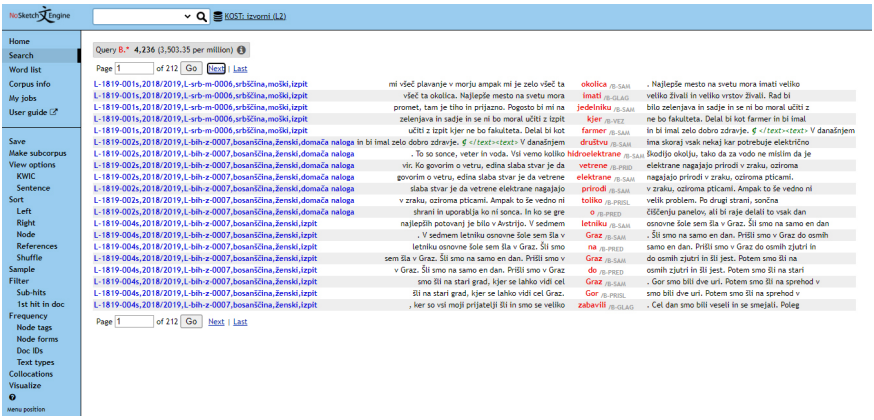
4 Dostop do korpusa KOST 1.0

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.

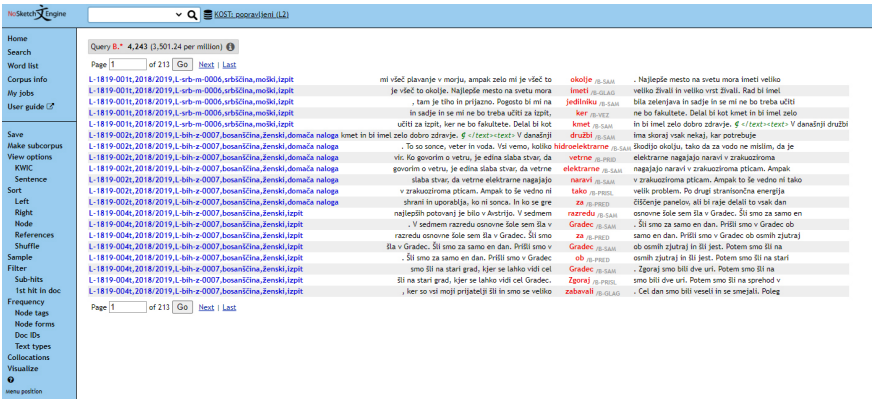
V projektu RSDO je bil razvit format korpusov z jezikovnimi popravki, ki je skladen z ostalimi slovenskimi korpusi in povezljiv s formatom orodja Svala. Tri izhodne datoteke JSON – nepopravljena in popravljena besedila ter datoteka s povezavami med vsako pojavnico iz nepopravljene in popravljene verzije 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).

Nadgradnja učnega korpusa ssj550k v SUK 1.0

Špela ARHAR HOLDT
Univerza v Ljubljani, Fakulteta za računalništvo in informatiko
Univerza v Ljubljani, Filozofska fakulteta

Jaka ČIBEJ
Univerza v Ljubljani, Filozofska fakulteta

Kaja DOBROVOLJC
Univerza v Ljubljani, Fakulteta za računalništvo in informatiko
Institut »Jožef Stefan«

Tomaž ERJAVEC
Institut »Jožef Stefan«

Polona GANTAR
Univerza v Ljubljani, Filozofska fakulteta

Simon KREK
Univerza v Ljubljani, Filozofska fakulteta
Institut »Jožef Stefan«

Tina MUNDA
Univerza v Ljubljani, Fakulteta za računalništvo in informatiko
Institut »Jožef Stefan«

Nejc ROBIDA
Univerza v Ljubljani, Filozofska fakulteta
Univerza v Ljubljani, Fakulteta za računalništvo in informatiko

Luka TERČON
Univerza v Ljubljani, Filozofska fakulteta

Slavko ŽITNIK
Univerza v Ljubljani, Fakulteta za računalništvo in informatiko

Povzetek

V prispevku predstavljamo nadgradnjo učnega korpusa, ki je namenjen učenju strojnih postopkov za jezikoslovno označevanje besedil v sodobni standardni slovenščini. Nova različica korpusa ssj500k, ki smo ga preimenovali v SUK (*slovenski učni korpus*), prinaša nova besedila in nove ročno pregledane jezikoslovne oznake različnih vrst. Korpus smo povečali s petsto tisoč na več kot milijon pojavnic z vključitvijo treh odprto dostopnih jezikovnih virov, ki vsak na svoj način odpravljajo predhodno identificirane pomanjkljivosti ssj500k: SentiCoref 1.0, ELEXIS-WSD za slovenščino in iz korpusa Gigafida 2.0 pripravljena množica Ambiga. Pregledovanje jezikoslovnih oznak opišemo po ravneh: tokenizacija, stavčna segmentacija, lematizacija, oblikoskladnja MULTEXT-East, oblikoslovje ter skladnja Universal Dependencies, skladnja JOS-SYN, udeleženske vloge, imenske entitete in koreference. Za vse ravni smo posodobili označevalne smernice, ki so pregledno zbrane in na voljo za nadaljnje delo. Na podatkih korpusa SUK smo naučili novo različico strojnega označevalnika CLASSLA-Stanza, ki dosega presežne vrednosti za vse evalvirane ravni. Z bogatim naborom ročno pregledanih jezikoslovnih oznak predstavlja učni korpus SUK enega temeljnih jezikovnih virov za sodobno slovenščino, zato zahteva neprestano posodabljanje in nadgrajevanje, kar predstavimo v zaključnem poglavju s smernicami za nadaljnji razvoj.

Ključne besede: učni korpus, ssj500k, SUK, jezikoslovno označevanje, označevalne smernice

Abstract

In this paper, we present an upgrade to the training corpus for linguistic annotation of modern standard Slovene. The new version of the ssj500k corpus, renamed to SUK, introduces both new texts and new manually reviewed linguistic tags of various types. The corpus has been expanded from 500,000 to over a million tokens by incorporating three openly accessible language resources, each addressing the previously identified shortcomings of ssj500k: SentiCoref 1.0, ELEXIS-WSD for Slovene, and a dataset prepared from the Gigafida 2.0 corpus called Ambiga. We describe the linguistic annotation process at various levels: tokenization, segmentation, lemmatization, MULTEXT-East morphology, Universal Dependencies

syntax, JOS-SYN syntax, semantic role labelling, named entity recognition, and coreference resolution. We have updated annotation guidelines, which are systematically compiled and available for further work. Using the SUK corpus data, we trained a new version of the automatic tagger CLASSLA-Stanza, which achieves outstanding results for all evaluated levels. With its manually-reviewed linguistic tags, the SUK corpus is foundational for modern Slovene, requiring ongoing improvements, which we detail in the final section with future development guidelines.

Keywords: training corpus, ssj500k, SUK, linguistic annotation, annotation guidelines

1 Uvod

Učni korpusi (ang. *training corpora*) so premišljeno grajene besedilne množice z zanesljivimi (tipično ročno pripisanimi ali pregledanimi) dodatnimi informacijami, ki se uporabljajo pri nadzorovanem strojnem učenju postopkov za obdelavo naravnega jezika. Ti postopki so lahko različni, med najbolj ključnimi za nadaljnje delo z jezikovnimi podatki pa je jezikoslovno označevanje: delitev besedila na gradnike (besede oz. pojavnice, večbesedne enote, povedi) in pripis jezikoslovnih informacij tem gradnikom. Učni korpusi za jezikoslovno označevanje zato spadajo v temeljno digitalno infrastrukturo določenega jezika in kot taki zahtevajo kontinuiran razvoj in nadgrajevanje.

Za nadzorovano učenje strojnega jezikoslovnega označevanja besedil v sodobni standardni slovenščini[1] se v našem prostoru že več kot desetletje razvija učni korpus, ki je bil do nedavnega poimenovan ssj500k (Krek idr., 2020a). Ta je vseboval 27.829 povedi (oz. približno 500.000 pojavnic, ki so korpusu dale ime), označenih na različnih jezikovnih ravneh, od segmentacije, tokenizacije, lematizacije, oblikoslovja in oblikoskladnje prek odvisnostne skladnje,

1 Za označevanje nestandardne slovenščine so na voljo učni korpusi iz zbirke Janes (Čibej idr., 2018); nedavna nadgradnja množic Janes-Tag in Janes-Norm je predstavljena v poročilu Arhar Holdt idr. (2023). Za označevanje starejše slovenščine pa je na voljo učni korpus goo300k (Erjavec, 2015).

imenskih entitet in večbesednih enot do udeleženskih vlog. Pod okriljem projekta Razvoj slovenščine v digitalnem okolju (RSDO)[2] je bil učni korpus nadgrajen z novimi besedili in oznakami, zaradi spremembe obsega pa smo ga preimenovali v SUK, *slovenski učni korpus*.

Nadgradnja korpusa predstavlja pomemben razvojni korak ne le v smislu prenove jezikovnega vira, pač pa tudi z vidika metodologije označevanja. Za vse ravni jezikovnih oznak, ki smo jih pripisovali in pregledovali, so bile posodobljene označevalne smernice, ki so po koncu projekta urejeno zbrane ter objavljene in tako na voljo za nadaljnje nadgradnje.[3] SUK 1.0, ki je pod odprto licenco na voljo na repozitoriju CLARIN.SI (Arhar Holdt idr., 2022), je bil že pod okriljem projekta uporabljen za izboljšavo strojnega označevalnika za slovenščino.

Pripravo korpusa smo z vidika projektnih ciljev predstavili v poročilu (Arhar Holdt idr., 2023), dela na posameznih označevalnih ravninah se dotikajo tudi nekateri prispevki, ki jih navajamo v nadaljevanju. V tem prispevku želimo raziskovalno-razvojni skupnosti jedrnato in celovito predstaviti nadgradnjo učnega korpusa ssj500k v SUK in s tem omogočiti njegovo usklajeno nadaljnje nadgrajevanje. Najprej predstavimo nabor besedilnih množic, s katerimi smo korpus nadgradili, sledi opis ročnega označevanja oz. pregledovanja oznak po jezikovnih ravninah in primerjava predhodne korpusne sestave z novo. Prispevek zaključimo s podatki o izboljšavah strojnega označevalnika, ki služijo kot ocena korpusne nadgradnje, ter smernicami za nadaljnje delo.

2 Metodologija

2.1 Povečanje korpusnega obsega

Korpus ssj500k (v različici 2.3: Krek idr., 2021) obsega 27.829 povedi in je v celoti ročno pregledan na ravni tokenizacije, stavčne segmentacije, oblikoskladenjskih oznak in lem. Večbesedne enote so

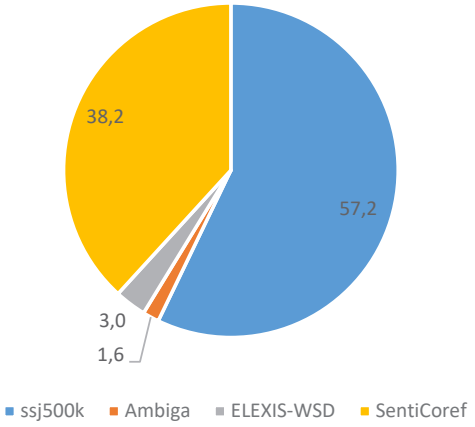
2 Spletna stran projekta z dostopom do rezultatov: https://slovenscina.eu/.
3 Smernice so dostopne na https://wiki.cjvt.si/shelves/jezikoslovno-oznacevanje-korpusov.

označene in pregledane pri 13.511 povedih, skladnja JOS-SYN pri 11.411 povedih, imenske entitete pri 9.488 povedih, skladnja UD pri 8.000 povedih in udeleženske vloge (SRL) pri 5.501 povedi (Krek idr., 2020a, Tabela 1). Ena od prioritet nadaljnjega razvoja je bilo povečanje razpoložljivega gradiva za višje označevalne ravni, v analizah korpusne sestave pa sta bili identificirani tudi potreba po povečanju s korpusnimi besedili, ki omogočajo označevanje prek meja povedi, ter dopolnitvi korpusa za boljšo zastopanost oblikoskladenjskih oznak in dvoumnih besednih oblik (Arhar Holdt in Čibej, 2021, 49–50).

Z upoštevanjem identificiranih potreb in v želji po učinkoviti izrabi že obstoječega gradiva smo za povečanje korpusa izbrali tri odprto dostopne jezikovne vire: (a) **SentiCoref 1.0** (Žitnik idr., 2022) je korpus besedil s slovenskih novičarskih portalov, ki je za namene analize sentimenta opremljen z oznakami imenskih entitet in koreferenc. Korpus odgovarja na potrebe po vključitvi gradiva za označevanje prek meja povedi, prinaša pa tudi vključitev novega označevalnega nivoja, ki spada na področje semantike – koreferenc. (b) **ELEXIS-WSD za slovenščino** (Martelli idr., 2021) je slovenski del 10-jezičnega vzporednega korpusa, ki vsebuje 2.024 povedi iz Wikipedijinih člankov. Korpus vsebuje ročno pripisane oznake za razdvoumljanje pomenov (ang. *word-sense disambiguation*) in kot tak ob korpusu SentiCoref predstavlja drugo izhodišče za strojno učenje na semantični ravni. (c) Iz korpusa **Gigafida 2.0** (Krek idr., 2020b) je bila pripravljena množica **Ambiga,** nabor 603 povedi, ki vsebujejo v predhodnem učnem korpusu nezastopane oblikoskladenjske oznake in pojavnice, identificirane kot problematične za strojno označevanje, npr. enakopisne zaimke, redke dvojinske oblike in podobno.

Novi učni korpus SUK tako sestavljajo množice ssj500k 2.3 (586.187 pojavnic oz. 57,2 %), SentiCoref 1.0 (391.962 pojavnic oz. 38,2 %), ELEXIS-WSD (31.233 pojavnic oz. 3 %) in Ambiga (16.257 pojavnic oz. 1,6 %), kot predstavlja Graf 1. Ko so bile množice za povečanje korpusnega obsega določene, je sledil strojni pripis jezikovnih oznak in njihov celoviti ročni pregled na ravni tokenizacije, stavčne segmentacije, lem in oblikoskladenjskih oznak, za izbrane dele korpusa pa še pripis in urejanje oznak na višjih označevalnih

ravneh. V nadaljevanju predstavljamo delo z oznakami, in sicer ločeno po označevalnih ravninah.



Graf 1: Besedilna sestava učnega korpusa SUK 1.0.

2.2 Segmentacija, tokenizacija, lematizacija, oblikoskladnja MULTEXT-East

Osnovni nivoji korpusnega označevanja: segmentacija, tokenizacija, lematizacija in oblikoskladnja po sistemu MULTEXT-East (žargonsko tudi MSD; ang. *morpho-syntactic description*) so bili ročno pregledani na celotnem gradivu, ki predstavlja nadgradnjo učnega korpusa (512.588 besednih pojavnic).

SentiCoref 1.0, največja izmed novih množic nadgrajenega učnega korpusa, je bil označen po fazah: (a) tokenizacija, lematizacija in segmentacija na povedi z orodjem CLASSLA-Stanza[4] (verzija 0.0.11), (b) ročni pregled teh treh ravni, (c) strojno oblikoskladenjsko označevanje po sistemu MULTEXT-East v6 z istim orodjem, (č) ročni pregled oblikoskladenjskih oznak. Ročni pregled je temeljil na uveljavljenih smernicah[5] in je potekal v spletnem okolju *Google Preglednice* (ang. *Google Sheets*). Tokenizacijo, lematizacijo in

4 https://github.com/clarinsi/classla
5 https://wiki.cjvt.si/books/04-oblikoskladnja-multext-east/page/oznacevalne-smernice, gl. Različica 1.0.

segmentacijo je pregledovalo 9 študentov, medtem ko je pri ročnem pregledu MSD-oznak sodelovalo 24 študentov jezikoslovnih smeri v razponu približno štirih mesecev, kar predstavlja eno najobširnejših tovrstnih označevalnih akcij v našem prostoru. Pregledovanje je potekalo po principu trojnega ujemanja: vsako pojavnico so neodvisno drug od drugega pregledali 3 študenti – oznake, ki so jih enotno izbrali vsi trije označevalci, so bile sprejete, oznake, pri katerih je prišlo do neujemanja, pa so bile znova pregledane v fazi kuracije (za natančnejši popis metodologije gl. Pori idr., 2022). Pri pregledu MSD-jev se je množica ELEXIS-WSD pridružila SentiCorefu (ostale ravni so bile pregledane predhodno), pri Ambigi pa je označevanje vseh štirih ravni zaradi omejenega obsega poteklo v enem koraku.

V nadaljevanju predstavljamo izzive, ki smo jih identificirali v označevalni kampanji, in rešitve, ki so vključene v nadgrajene smernice.[6] Gre za težje in mejne primere, ki so bili v predhodnih označevalnih smernicah slabše zastopani ali pa sploh niso bili, ali pa je pri pregledovanju teh pogosto prišlo do neupoštevanja smernic in s tem nedoslednosti. Dileme smo analizirali, tudi s pomočjo že označenih podatkov v ssj500k, in jih po kuraciji, kolikor je bilo mogoče, uskladili.

Prekrivnost samostalnikov v slovenskih stvarnih lastnih imenih z občnoimenskimi: Pravilo, da samostalnikom, ki so del stvarnih lastnih imen in so prekrivni z občnoimenskimi samostalniki, pripišemo občnoimenskost in jih lematiziramo z malo začetnico, je bilo v obstoječih smernicah sicer obravnavano, a označevalcem ni bilo intuitivno. Gre za primere tipa podjetje *Iskra* (lema: iskra, MSD: Sozei), časnik *Delo* (lema: delo, MSD: Sosei). Vendar to pravilo velja le za samostalnike, ne pa tudi za druge besedne vrste in ne za primere, kjer nesamostalniška besedna vrsta nastopa kot samostalnik, npr. stranka *Zares* (MSD: Slzei, lema: Zares).

Pridevniki iz osebnih in zemljepisnih lastnih imen: Pri izlastnoimenskih svojilnih pridevnikih, ki zaznamujejo vrsto in ne prave svojine ter tudi že prehajajo v zapis z malo začetnico, se je pri določanju leme

6 https://wiki.cjvt.si/books/04-oblikoskladnja-multext-east/page/oznacevalne-smernice, gl. Različica 2.0.

izkazala za težjo odločitev med malo in veliko začetnico. V obstoječih smernicah ni bilo jasnega razlikovanja med to kategorijo pridevnikov in pravimi svojilnimi pridevniki. Tako smo v nadgrajenih smernicah dodatno pojasnili obravnavo izlastnoimenskih pridevnikov: (a) pri pridevnikih iz osebnih lastnih imen imamo poleg teh, ki izražajo pravo svojino (*Pahorjeva* (lema: Pahorjev, MSD: Psnzei) *[mlada struja]*) še tiste, ki zaznamujejo vrsto in jih v rabi pogosto najdemo zapisane z malo začetnico; te primere lematiziramo z malo začetnico (*[zdravljenje] parkinsonove* (lema: parkinsonov) *[bolezni]*); (b) pri pridevnikih iz stvarnih lastnih imen (*Magov [novinar]*, *Delova [dopisnica]*) smo opredelili načelo lematizacije, in sicer z malo začetnico lematiziramo tiste, ki v referenčnem korpusu Gigafida 2.0 izkazujejo svojilno rabo (*Magov* [novinar]; lema: magov (prek mag = čarovnik), medtem ko primere, kjer je svojina konceptualno sicer možna, vendar v rabi ni izkazana, lematiziramo z veliko začetnico (*Delova [dopisnica]*; lema: Delov).

Tuja stvarna lastna imena: Tu so izziv predstavljali primeri dveh tipov: (a) tuja stvarna lastna imena iz slovenščini sorodnih jezikov, ki se v slovenskih besedilih zaradi morfološke podobnosti pregibajo po slovenskih vzorcih (npr. hrvaška imena: *Zagrebačka banka*, *Večernji list*) in (b) deli tujih stvarnih lastnih imen, ki so prevzeti v slovenščino in so pomensko prekrivni z izvorno tujo besedo (npr. *leasing*, *holding*) ali pa so oblikovno prekrivni s slovenskimi samostalniki, a si s tujo besedo ne delijo pomena, pa tudi besedni vrsti v obeh jezikih nista nujno isti (npr. *trans*, *global*). Odločili smo se, da bomo tako v primerih tipa (a) kot (b) upoštevali prekrivnost s slovenskim občnim samostalnikom, če je zadovoljeno vsaj enemu izmed dveh meril: 1) potencialno prekriven samostalnik kot del tujega lastnega imena se v rabi pregiba; 2) tuj samostalnik je prevzet, kar potrjujejo referenčni priročniki za slovenščino (npr. *[Hypo] Leasing*; lema: leasing, MSD: Somei; *[Infond] Holding*; lema: holding, MSD: Somei). Pomenska prekrivnost besede v enem in drugem jeziku ni bila nujen pogoj za uvrstitev tovrstnih primerov med občnoimenske samostalnike (*[Trade] Trans [Invest]*; lema: trans, MSD: Somei; *[Prevent] Global*; lema: global, MSD: Somei). Kot velja pri obravnavi (delov) stvarnih imen, ki jih sestavljajo neizpodbitno slovenske besede, tudi

v tujih stvarnih lastnih imenih prekrivnost iščemo le pri samostalnikih. To velja posebej izpostaviti, saj so v jezikih, sorodnih slovenščini, lahko tudi nesamostalniške besedne vrste oblikovno podobne slovenskim in se kot take lahko tudi pregibajo. Pri teh besedah je lema enaka obliki, MSD-oznaka pa 'neuvrščeno' (*Večernji* (lema: Večernji, MSD: Nj) *list* (lema: list, MSD: Somei), *Zagrebačka* (lema: Zagrebačka, MSD: Nj) *banka* (lema: banka, MSD: Sozei).

Ločevanje pridevnikov od prislovov: Obravnavali smo vprašanje, kateri besedni vrsti pripada oblika besede, ki je enaka prislovu in pridevniku, ko je ta beseda (a) v vlogi povedkovega določila (npr. *[... bi bilo] smotrno, [da bi ...]*) ali (b) v strukturi z nedoločnikom (npr. *[O tem ni] mogoče [sklepati.]*). Predhodne smernice tega niso obravnavale, kar se je odražalo tudi v korpusu ssj500k, kjer tovrstni primeri niso bili enotno označeni. Po pregledu in analizi pojavitev tovrstnih primerov v korpusu SentiCoref smo oblikovali pravilo, da besedi v obeh naštetih skladenjskih vlogah pripišemo pridevniško lemo in MSD-oznako, če v stavku ni izpustljiva (je obvezna, da je stavek koherenten), in nasprotno – prislovno lemo in oznako, če je stavek koherenten tudi brez nje (npr. *[O tem ni] mogoče* (lema: mogoče, MSD: Ppnsei) *[sklepati.]* > *O tem ni sklepati.**; *Mogoče* (lema: mogoče, MSD: Rsn) *[ste ga vznemirili.]* > *Vznemirili ste ga.*).

Predložne prislovne zveze: Podobno kot pri prejšnji dilemi je bila težava pri razlikovanju med pridevnikom in prislovom v prislovnih zvezah s predlogom (npr. *na novo*, *v živo*). Tudi tovrstni primeri so bili v korpusu ssj500k označeni neenotno in po analizi smo določili, da nepredložnemu delu v predložnih prislovnih zvezah pripišemo pridevniško lemo in MSD-oznako (*[na] novo* (lema: nov, MSD: Ppnset)).

Nesklonljivi prilastki: V obstoječih smernicah je bilo pravilo, da nesklonljive prilastke (npr. *solo*, *neto*, *bruto*) označimo kot samostalnike, kadar so sklonljivi, in kot pridevnike, kadar niso, vendar kriterij sklonljivosti ni bil jasno opredeljen. Tako smo oblikovali pravilo, da določen primer označimo kot samostalnik, če v referenčnem korpusu najdemo potrditev, da se lahko pregiba kot samostalnik (npr. *pop*, *elektro*), in kot pridevnik, če te potrditve ni (npr. *neto*, *repro*).

2.3 Oblikoslovje in skladnja po sistemu UD

Universal Dependencies (UD) je označevalna shema, ki si prizadeva za mednarodno oz. medjezično usklajeno slovnično označevanje besedil na oblikoslovni in skladenjski ravni, da bi pospešila razvoj večjezičnih jezikovnih tehnologij na eni strani in kontrastivnih jezikoslovnih analiz na drugi (de Marneffe idr., 2021). V zbirko več sto korpusov, označenih s to shemo, je bila leta 2015 priključena tudi univerzalna odvisnostna drevesnica za pisno slovenščino, drevesnica SSJ (Dobrovoljc idr., 2017), ki je ob prvi objavi vsebovala 8.000 razčlenjenih povedi korpusa ssj500k (primer na Sliki 1), v projektu RSDO pa smo jo bistveno nadgradili tako z vidika obsega kot z vidika dokumentiranosti smernic in infrastrukturne podpore za njeno nadaljnjo analizo (Dobrovoljc in Ljubešić, 2022; Dobrovoljc idr., 2023).

Jedrne smernice sheme UD, kakršne so dokumentirane na uradni spletni strani projekta,[7] za vsako izmed predlaganih »univerzalnih« oznak (17 besednih vrst, 24 oblikoskladenjskih lastnosti, 37 odvisnostnih skladenjskih relacij) podajajo razmeroma splošno opredelitev s ponazoritvami na nekaj izbranih primerih v različnih jezikih, način prenosa teh smernic na svoje konkretne jezikovne podatke pa je prepuščen avtorjem drevesnic za posamezne jezike. Ker za slovenščino ob nastanku prvotne drevesnice SSJ te smernice niso bile sistematično dokumentirane, je bil prvi korak znotraj projekta RSDO zato namenjen izčrpnemu popisu smernic UD za slovenščino, tako na spletni strani projekta (v angleščini) kot v obliki samostojnega priročnika v slovenščini.[8] Slednji poleg velikega števila ponazoritev prototipičnih in mejnih primerov vsake oznake vsebuje tudi ločeno poglavje s smernicami za označevanje kompleksnejših skladenjskih struktur (npr. elipse, primerjave, poudarjalni členki, besedilni povezovalci ...). Pri tem smo poleg opisa prvotnih smernic uvedli tudi nekaj manjših izboljšav na mestih, kjer je bila prvotna označenost korpusa SSJ-UD nedosledna ali

7 https://universaldependencies.org/

8 https://wiki.cjvt.si/books/07-universal-dependencies/page/oznacevalne-smernice, gl. Različica 1.0.

neustrezna glede na splošne, jezikovno univerzalne smernice. To pa ne velja za vse identificirane neskladnosti, saj nekatere predstavljajo precejšen odmik od doslej uveljavljenih označevalnih praks v slovenskem prostoru in bi jih bilo zato smiselno najprej nasloviti s širšo strokovno diskusijo. Tovrstna mesta smo popisali v ločeni prilogi[9] h krovnim smernicam.

Ker sta si označevalna sistema JOS in UD na ravni pripisovanja besednih vrst in drugih oblikoslovnih lastnosti precej podobna, so bila že ob nastanku prvotne odvisnostne drevesnice UD za slovenščino izdelana podrobna pravila za preslikavo oblikoskladenjskih oznak JOS v besedne vrste in oblikoskladenjske lastnosti sistema UD,[10] s katerimi je bil v celoti pretvorjen tudi učni korpus ssj500k. Na enak način smo z avtomatsko pretvorbo v univerzalne oblikoslovne oznake (besedne vrste in druge oblikoskladenjske lastnosti) pretvorili tudi novi učni korpus SUK z ročno pripisanimi oblikoskladenjskimi oznakami JOS. Ker se pretvorbena pravila v času od nastanka prejšnjih različic korpusov niso spremenila, smo v okviru projekta RSDO pretvorbo opravili zgolj na novo dodanih besedilih korpusa SUK in opravili ustaljeni ročni pregled povedi z glagolom *biti* za razdvoumljanje med pojavitvami pomožnega in glavnega glagola (po en označevalec na primer).

Poleg zgoraj opisanega označevanja celotnega korpusa SUK na oblikoslovni ravni smo prvotni korpus ssj500k oz. SSJ v obsegu 8.000 povedi dodatno povečali še za 5.435 novih ročno razčlenjenih povedi v obliki dvofazne označevalne kampanje. V prvi fazi razširitve so označevalci ročno pregledali 3.411 polpretvorjenih povedi korpusa ssj500k, ki zaradi omejene natančnosti pretvorbenih pravil v času nastanka prvotnega korpusa SSJ-UD niso bile javno objavljene, pri čemer so se označevalci osredotočili predvsem na pripisovanje novih oz. manjkajočih povezav (22.377 oz. 23,5 % vseh pojavnic). V drugi fazi širitve je bil skladenjsko razčlenjen še podkorpus ELEXIS-WSD, ki vsebuje 2.024 povedi, in sicer z ročnim pregledom vseh strojno

9 https://wiki.cjvt.si/books/07-universal-dependencies/page/oznacevalne-smernice, gl. Različica 1.0 – Priloga.

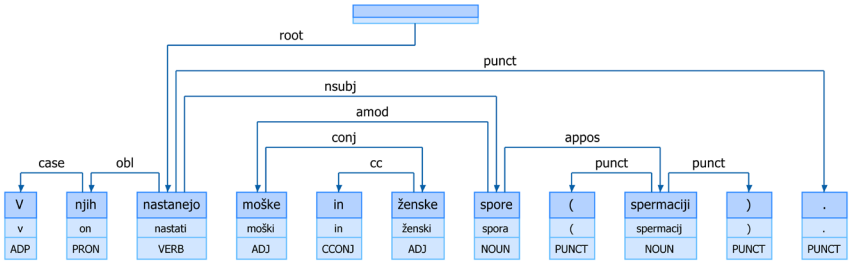
10 https://github.com/clarinsi/jos2ud

pripisanih razčlemb orodja CLASSLA-Stanza. V obeh fazah so vsako poved pregledali 2–3 neodvisni označevalci in končni kurator, pri čemer smo za označevanje uporabili orodje Q-CAT (Brank, 2022), ki odslej podpira tudi uvoz datotek v formatu CoNLL-U, za kuracijo pa spletno platformo WebAnno, ki jo vzdržuje CLARIN.SI. Pred objavo je bila glede na nekoliko spremenjene izhodiščne smernice in druge identificirane nedoslednosti s hevrističnimi poizvedbami izboljšana tudi označenost prvotne drevesnice SSJ.

Rezultati vseh zgoraj opisanih aktivnosti so objavljeni kot del novega referenčnega učnega korpusa za slovenščino SUK 1.0, s čimer se je količina učnih podatkov tako na oblikoslovni kot skladenjski ravni skoraj podvojila (gl. Tabelo 1). Univerzalno skladenjsko razčlenjeni del korpusa SUK je bil po standardni delitvi na učno, validacijsko in testno množico obenem objavljen tudi kot del skupne mednarodne zbirke drevesnic UD v2.10 – kot nova, razširjena in izboljšana različica drevesnice SSJ. Nova različica SSJ v primerjavi s prvotno vsebuje skoraj enkrat večje število pojavnic (126.427, +89,9 %), s čimer se korpus SSJ po številu pojavnic danes umešča v zgornjo osmino vseh drevesnic UD po svetu. Z razširitvijo je drevesnica SSJ postala tudi bolj raznolika, saj se vsi trije podkorpusi (izvorne povedi iz ssj500k, nove povedi iz ssj500k, nove povedi iz ELEXIS-WSD) med seboj razlikujejo tako z vidika vrste vsebovanih besedil kot njihove skladenjske kompleksnosti.

Drevesnica SSJ, tj. univerzalno oblikoskladenjsko razčlenjeni podkorpus korpusa SUK, je bila kot samostojna podatkovna množica že integrirana v številna orodja in spletne portale po svetu,[11] po njej pa je mogoče brskati tudi s pomočjo lokalno razvitega orodja Q-CAT (Slika 1) in spletnega vmesnika Drevesnik, ki sicer trenutno omogočata zgolj prikaz univerzalnih besednih vrst in odvisnostnih skladenjskih relacij, ne pa oblikoslovnih lastnosti tipa Case=Nom.[12]

11 https://universaldependencies.org/tools.html
12 https://orodja.cjvt.si/drevesnik/



Slika 1: Primer označene povedi po shemi Universal Dependencies v orodju Q-CAT.

2.4 Skladnja po sistemu JOS-SYN

Sistem JOS-SYN, ki je bil zasnovan v projektu Jezikoslovno označevanje slovenščine (Erjavec idr., 2010), sledi spoznanjem slovenskega jezikoslovja (zlasti slovnici Toporišič, 2004), obenem pa temeljnim idejam, ki jih zarisujejo obstoječi uveljavljeni sistemi odvisnostnega označevanja. Ključna lastnost sistema je, da upošteva informacije, ki jih prinašajo oblikoskladenjske oznake JOS oz. njihova sodobna različica MULTEXT-East v6 (Erjavec, 2012). Na skladenjski ravni tako dodajamo samo informacije, ki jih še ni pokrila oblikoskladnja, kar omogoči robusten, intuitiven in hitro razložljiv označevalni sistem.[13]

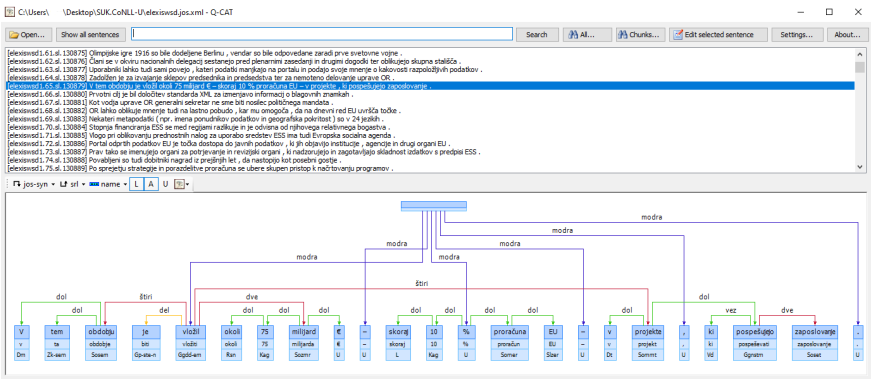
Skladenjska raven JOS-SYN je bila dobro zastopana že v prejšnji različici učnega korpusa: v ssj500k je bilo s tem sistemom označenih 11.411 povedi v 617 besedilih s skupnim obsegom 235.864 pojavnic (Krek idr., 2020a, 25–26). Na teh podatkih je že bil naučen skladenjski razčlenjevalnik za slovenščino, ki je dosegal 90,43 % za pravilno določeno mesto povezave oz. 87,52 % za pravilno določena mesto in tip povezave (Dobrovoljc idr., 2012). Cilj nove označevalne kampanje je bil označiti 2.024 novih povedi ELEXIS-WSD, s tem povečati obseg učnega gradiva, pri tem pa natančneje oceniti ter nadgraditi označevalne smernice.[14] Kampanja je trajala približno šti-

13 Sistem oznak je predstavljen na strani https://wiki.cjvt.si/books/06-odvisnostna-skladnja-jos-syn/page/predstavitev-oznak.

14 https://wiki.cjvt.si/books/06-odvisnostna-skladnja-jos-syn/page/oznacevalne-smernice, gl. Različica 1.0.

ri mesece, dva meseca za intenzivno označevanje in dva meseca za pripravo analiz in nadgradnjo smernic.

Povedi ELEXIS-WSD, v katerih je že bila ročno popravljena tokenizacija, segmentacija, lematizacija ter oblikoskladnja MULTEXT-East, smo najprej strojno skladenjsko označili z orodjem CLASSLA-Stanza (verzija 1.1.0), nato pa sta dva jezikoslovca s pomočjo orodja Q-CAT (Brank, 2022) ročno pregledala vsako od povedi in popravila strojno pripisane skladenjske oznake (Slika 2). Nejasnosti in neskladja v označevalnih rešitvah smo beležili in naslavljali sproti ob delu. Težja mesta označevanja, ki so izvirala iz nejasnosti označevalnih smernic ali novoodkritih označevalnih zadreg, smo jezikoslovno analizirali, poiskali rešitve in posodobili smernice. Poleg vrzeli v smernicah smo med delom identificirali tudi mesta, kjer so podatki v ssj500k označeni neskladno. Za določene vrste težav, ki jih navajamo v nadaljevanju, smo skladenjske oznake v podatkih ssj500k posodobili, nekaj usklajevanja bo treba še opraviti v prihodnje, nekatere težave pa se propagirajo z nižjih ravni, čemur se bo treba posvetiti v nadaljnjih projektih.



Slika 2: Označevanje odvisnostne skladnje JOS-SYN v programu Q-CAT.

Da bi označevalne smernice postale preproste za nadaljnje nadgrajevanje in uporabo, smo jih oblikovno in vsebinsko poenostavili, strukturo nadgradili in zagotovili dodatne zglede označevanja (več o tem v Arhar Holdt idr., 2023). V smernice smo dodali nova poglavja,

ki natančneje pojasnjujejo označevanje izbranih pojavov. Nova je denimo obravnava simbolov in ločil, ki nadomeščajo besede (npr. *% ° $* za besede *odstotek, stopinja, dolar*), znake *+ & /* - v pomenu veznikov 'in', 'ali' (npr. *srčno-žilna bolezen*), znak / v pomenu 'na' (*6 mg/kg*), znaka - in – v pomenu 'od'–'do', 'proti' (*v sezoni 2006–07*) ter znak - pri povezovanju kratic in števil v podredne zveze (*16-tonski*). Ti elementi pri predhodnem označevanju niso bili vpeti v besednozvezno skladnjo, zaradi česar je razpadla drevesnica vseh povedi, ki so jih vsebovale. Nove smernice, ki ločujejo besednozvezno povezljive znake od nepovezljivih, za povezljive pa jasno prikažejo načine povezovanja, so skladnejše s primerljivimi sistemi, tudi UD za slovenščino. Ker gre za večjo spremembo sistema označevanja, smo pregledali in uskladili obravnavo tovrstnih elementov tudi v ssj500k.

Obširnejša dopolnitev smernic je bila pripravljena tudi za obravnavo lastnih imen in tujejezičnih elementov. Problematiko smo strukturirali na ožje vsebinske sklope, za vsakega pripravili opis, temelječ na analizah predhodnega označevanja, pa tudi posebna opozorila, kjer je v preteklosti prihajalo do zmede. Navodila za označevanje lastnih imen so bila predhodno precej skopa, posledično pa je v ssj500k opaziti velike neskladnosti označevanja, tako pri določanju, ali zvezo obravnavati kot slovensko ali kot fragment v tujem jeziku (glede na smernice se fragmenti v tujem jeziku tipično ne povezujejo v drevesnico), kot tudi odločanje, kaj je jedro pri zvezah, ki prihajajo iz tujega jezika. Precej nedoslednosti je najti pri povezovanju tujejezičnih členov tipa *de, la, the*, za katera velja posebna obravnava, vendar jih označevalci težko prepoznavajo, kadar gre za manj znane tuje jezike. Najtrši oreh pri označevanju pa so tujejezična stvarna lastna imena, kjer naj bi označevanje sledilo odločitvam na ravni oblikoskladnje, vendar tudi tam smernice niso optimalne (Pori idr., 2022).

Od sprememb je možno izpostaviti še nekaj takšnih, ki so vezane na označevanje specifičnih struktur (za referenco gl. nove smernice[15]). V poglavje, ki se posveča označevanju struktur tipa *nujno je*,

15 https://wiki.cjvt.si/books/06-odvisnostna-skladnja-jos-syn/page/oznacevalne-smernice, gl. Različica 2.0

smo dodali obravnavo struktur *treba je,* saj je za označevalce koristno na enem mestu videti, da so pridevniki v takšnih primerih z glagola *biti* vezani s povezavo DOL, prislovi pa s TRI. Na podoben način smo v poglavje *Polstavčni desni prilastki*, ki se je predhodno osredotočalo na pridevniške, deležniške in nedoločniške polstavke (povezava DOL), dodali še primer obravnave deležijskih desnih prilastkov (povezava TRI). Nadgradili smo poglavje o prilastkovih odvisnikih, ki zdaj vsebuje tudi navodila za označevanje t. i. nepravih odvisnikov, prilastkovih odvisnikov v povedih s pristavki ter primerov tipa *dovolj star, da*. Nenazadnje, pojasnili smo navodila za označevanje osebka pri pasivnih strukturah s *se* (npr. *v hudih primerih se daje adrenalin*).

Posodobili smo dve mesti smernic, kjer se nahajajo vnaprej pripravljeni (zaključeni) seznami besed, ki jih označujemo po določenih pravilih, in sicer informativni seznam zvez, ki jih povezujemo s povezavo SKUP,[16] ter seznam členkov, ki jih povezujemo v besedne zveze kot določujoči element. Oba seznama smo posodobili na osnovi analiz predhodnega označevanja in pojavnosti obravnavanih jezikovnih elementov v referenčnem korpusu, upoštevali pa smo tudi označevalne prakse pri skladenjskem sistemu UD za slovenščino.

Pri preverbah že označenega gradiva smo identificirali tudi nedoslednosti, ki ne izvirajo nujno iz nejasnosti smernic in bi jih bilo treba v nadaljnjih projektih sistematično nasloviti in odpraviti. Poleg že omenjene težave z označevanjem (zlasti tujejezičnih stvarnih) lastnih imen so se kazala neujemanja pri povezovanju členkov in prislovov (npr. *vsaj, izključno*) in slovničnih besed, ki lahko nastopajo kot različne besedne vrste (npr. *niti, razen*), povezovanju pridevnikov, kadar modificirajo številnike (npr. *dodatnih 400 milijonov*), označevanju pridevniške in samostalniške vezljivosti, latinskih poimenovanj, citatov in drugih fragmentov (npr. pri zvezah s *pa tudi*), ločevanjem med osebkom in povedkovim določilom; predmetom in prislovnim določilom; oznakami TRI in ŠTIRI ter prilastkovimi in drugimi odvisniki (npr. v stavkih s *ko, preden, dokler*). Za urejanje doslednosti so ključni zlasti problemi, ki se lahko propagirajo na višje

16 Besede, ki imajo variantni zapis skupaj ali narazen, večbesedne veznike in podobne večbesedne enote.

ravni (udeleženske vloge) ali so posledica nerešenih vprašanj na nižjih ravneh (oblikoskladnja).

2.5 Udeleženske vloge po sistemu SRL

Označevanje korpusa s semantičnimi kategorijami izhaja iz potrebe po strojnem procesiranju jezikovnih podatkov, ki so semantične narave, in zadeva različne možnosti njihove uporabe, kot je razvoj sistemov za luščenje informacij, sistemov za odgovarjanje na vprašanja, izboljšava delovanja skladenjskih razčlenjevalnikov, strojnih prevajalnikov ipd.

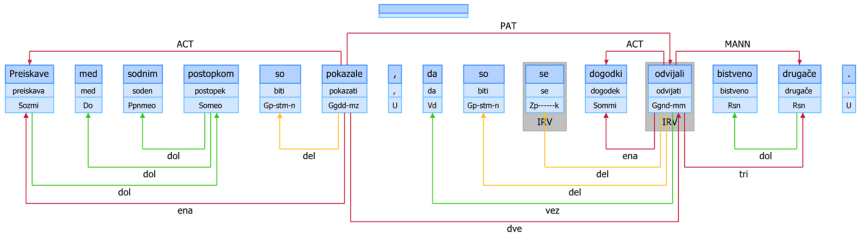
Celotni del semantično označenega korpusa SUK predstavlja **podkorpus SRL**, ki vsebuje dva dela. Korpus **SRL-ssj500k** vsebuje 9.724 ročno označenih povedi iz priprave predhodne različice učnega korpusa (ssj500k 2.3) in povedi, ki so bile v ssj500k 2.3 označene na morfološki in skladenjski ravni, niso pa bile označene na semantični ravni. **SRL-WSD** predstavlja korpus ELEXIS-WSD, ki vsebuje 2.024 povedi. Razen že predhodno ročno pregledanih povedi sta bila korpusa najprej avtomatsko označena na semantični ravni s pomočjo SRL parserja (Björkelund idr., 2009), korpus SRL-WSD pa tudi na morfološki in skladenjski ravni z orodjem CLASSLA-Stanza po sistemu JOS in UD. V označevalno kampanjo na semantični ravni je bilo skupaj vključenih 11.748 povedi, od tega je bilo 5.501 povedi ponovno pregledanih, 6.247 povedi pa je bilo najprej avtomatsko označenih, nato pa ročno pregledanih. Odločitve so bile na koncu usklajene v celotnem podkorpusu SRL učnega korpusa SUK.

Pri snovanju slovenskega modela za semantično označevanje (Krek idr., 2016) smo se glede na analizo označevalnih sistemov odločili, da bomo izhajali iz funkcijskega generativnega pristopa Praške odvisnostne drevesnice (ang. *Prague Dependency Treebank*, PDT; Mikulová idr., 2006).[17] Z vidika optimizacije pomenske razdrobljenosti, upoštevanja slovenskih specifik in prekrivnosti oznak med posameznimi sistemi smo nabor ustrezno zreducirali, kot je opisano v

17 Strnjen pregled vseh semantičnih oznak, njihov opis in zgledi je na voljo na tej povezavi: https://wiki.cjvt.si/books/10-udelezenske-vloge-srl/page/predstavitev-oznak.

Arhar Holdt idr. (2023). Podroben opis semantičnih oznak in pravila za njihovo uporabo vsebujejo Smernice za semantično označevanje učnega korpusa,[18] ki so bile v okviru projekta RSDO nadgrajene in posodobljene glede na vsebinske analize.

Vse povedi v na novo označenem in popravljenem predhodno že označenem korpusu je označila podiplomska študentka slovenistike na podlagi prve različice smernic in na podlagi sprotnih konzultacij in navodil. Celoten na novo označen korpus je nato pregledala soavtorica tega prispevka s pomočjo sistematičnih in ciljnih preverjanj. Za označevanje je bilo uporabljeno orodje Q-CAT (Brank, 2022), kot prikazuje Slika 3.



Slika 3: Prikaz semantične označevalne ravni v orodju Q-CAT.

Izhodišče semantičnega označevanja je predstavljal posamezni glagol v vseh svojih pojavitvah znotraj vnaprej določenih pomenskih skupin, npr. glagoli govorjenja, premikanja, kognitivnih procesov ipd., kar je omogočilo prepoznavanje tipičnih udeleženskih vlog, ki se povezujejo s posameznimi pomeni glagolov znotraj skupnega pomenskega polja. Z označevanjem smo začeli pri pogostejših glagolih (*biti, imeti, morati, iti, začeti, vedeti*) ter nadaljevali z upoštevanjem sorodnih pomenskih skupin, npr. glagolov rekanja (*povedati, reči, praviti, govoriti*). Na koncu smo označili glagole z zgolj eno pojavitvijo v povedi (pribl. 1200). Na ta način smo v največji možni meri zajeli povedi, za katere je bilo mogoče izpeljati čim bolj sistematične in usklajene jezikovne rešitve.

18 https://wiki.cjvt.si/books/10-udelezenske-vloge-srl/page/oznacevalne-smernice, gl. Različica 1.0.

V procesu označevanja je bil korpus nadgrajen tudi z vsebinskega vidika, pri čemer dodana vrednost temelji na jezikoslovnem premisleku že obstoječih odločitev v skladu z novimi spoznanji pri izdelavi semantičnih virov, analize vezljivostnih vzorcev pri izdelavi Vezljivostnega leksikona (Gantar, 2021; Gantar, 2023) in na upoštevanju potreb jezikovnotehnološke skupnosti.

V približno 75 % korpusa so popravljena in poenotena razmerja med udeleženci pri glagolih rekanja po načelu: REC = naslovnik glagolskega dejanja, RESLT = konkretni končni rezultat ali "izdelek" glagolskega dejanja (npr. izjava sama, ki jo največkrat uvaja odvisni stavek), PAT = vsebina ali tema glagolskega dejanja.

Druge pomembne vsebinske izboljšave korpusa temeljijo na analizi nekaterih problematičnih skladenjskih struktur in poenotenju odločitev v povezavi z označevanjem skladenjskega nivoja. Sem sodi poenotenje in uskladitev opredeljevanja razmerja med udeleženci v skladenjsko enakovrednih povedih tipa: *kdo ali kaj je kdo ali kaj*. Na podlagi smernic predhodnega semantičnega označevanja učnega korpusa ssj500k smo z udeleženske vlogo ACT, ki v splošnem zajema vršilce in pobudnike dejanja, označevali samostalnike v imenovalniku, ki nastopajo kot osebki glagola *biti*; samostalniška povedkova določila ob glagolu *biti* pa kot prizadeto (PAT): *območje medenice(ACT)* je središče telesa*(PAT)*; *problem begunc*ev*(ACT)* je stvar države*(PAT)*. Glede na omenjena izhodišča smo že na ravni prvotnega označevanja tu predvidevali največ neenotnosti na pomenskem nivoju in odstopanja med skladenjskim in pomenskim nivojem, predvsem zaradi težav pri odločanju o izhodišču in določilu stavka na pomenski ravni in o polno- oz. nepolnopomenski vlogi glagola *biti*, ki odloča med osebkovo in povedkovodoločilno vlogo na skladenjski ravni. V zvezi s tem smo pri nadgradnji korpusa sprejeli odločitev, da v skladenjsko enakovrednih povedih pomenska interpretacija sledi pravilu: kar izvem novega = prizadeti (PAT) udeleženec, o komer ali čemer izvem kaj novega = nosilni udeleženec (ACT). To v veliki meri ustreza označevanju na skladenjski ravni, kjer se temu, kar je na pomenski ravni aktant, pripisuje odvisni del povedka (povezava *dol*), temu, kar na

pomenskem nivoju opredeljujemo kot prizadeto, pa je na skladenjski ravni tipično pripisan osebek (povezava *ena*):

Dogodek v Ankaranu(dol-ACT) je bila dramatična nesreča(ena-PAT).
Gostja večera(dol-ACT) bo Desa Muck(ena-PAT).
Večina potnikov(dol-ACT) so bile ženske(ena-PAT).

Označevanje je v skladu z zgornjimi odločitvami dosledno izpeljano na pribl. 90 % povedi združenega korpusa SRL, medtem ko je smiselnost poenotenja z razmerij na skladenjski ravni (tj. ena-ACT; dva-PAT) eden od jezikoslovnih premislekov, ki terjajo širši jezikovni konsenz.

Z omenjeno vsebinsko nadgradnjo so povezane tudi odločitve pri drugih udeleženskih vlogah glagola *biti* po sistemu: *biti* + samostalnik = PAT: *dogodek(ACT) je bil nesreča(PAT)*; biti + pridevnik = RESLT: *je osamljena(RESLT)*; biti + prislov = MANN: *bo toplo(MANN)*. Popravki so bili izvedeni tudi na predhodno ročno že označenih povedih, s čimer smo želeli doseči enotnost označevanja pri nekaterih najpogostejših semantičnih vzorcih.

Prav tako so bile deloma poenotene odločitve, aplicirane na korpusne povedi v približno 80 %, pri razumevanju agentnih in deagentnih rab. Pri označevanju smo sledili pomenski interpretaciji izhodiščnega udeleženca kot vršilca dejanja (ACT), ki mu praviloma ni mogoče dodati še enega vršilca, ne da bi se pri tem spremenil pomen: *dogodki(ACT) so se odvijali bistveno drugače – *ACT je odvijal dogodke …*, in pravilu, da morajo ostati udeleženske vloge v agentnih in deagentnih strukturah nespremenjene, kjer prihaja do diskrepance med skladenjskim in pomenskim nivojem: *stvar(PAT-ena) je malce bolj zapletena – zgodbo(PAT-dve) sta sami(ACT-tri) zapletli.* Pri nadaljnji nadgradnji učnega korpusa bi bilo smiselno upoštevati tudi neenotnosti v pomenski interpretaciji, ki niso bile sistematično odpravljene, npr. *med njimi so se širile govorice(ACT)* : *potem je začela širiti govorice(PAT)*.

Nadaljnje izboljšave korpusa vidimo na več ravneh: z aktualizacijo semantičnih oznak glede na označevalni sistem PDT (opisano v

Arhar Holdt idr. (2023, 44–45)); z nadgradnjo korpusa z naborom semantičnih oznak glede na jezikoslovne analize, ki zahtevajo konsenz tudi na drugih označevalnih ravneh; ter z nadgradnjo korpusa s semantičnimi kategorijami, ki se oblikujejo znotraj pobud za povezovanje konceptov na medjezikovni ravni (npr. UniDive,[19] ELEXIS[20]).

2.6 Imenske entitete

Imenske entitete (ang. *named entities*; NE) so samostalniki in samostalniške besedne zveze, ki identificirajo neko osebo (oznaka PER), lokacijo (oznaka LOC), organizacijo (oznaka ORG) ali drug edinstven objekt v realnem prostoru in času (oznaka MISC). Tem standardnim oznakam se pridružuje še kategorija svojilnih pridevnikov, izpeljanih iz osebnega lastnega imena (oznaka DERIV-PER), npr. [*Obamova*] DERIV-PER *izvolitev*), ki se je kot odgovor na potrebo po celovitejši anonimizaciji osebnih podatkov pokazala kot nepogrešljiva. Imenske entitete so na ortografski ravni pogosto izražene z veliko začetnico (npr. *Slovenska tiskovna agencija*) ali kratico (npr. *STA*), vendar pa velika začetnica in kratica ne označujeta samo imenskih entitet (npr. *BDP*). Identifikacija imenskih entitet v besedilu je pomembna za odkrivanje koreferenčnosti, analiziranje sentimenta, ekstrakcijo informacij, povezav in dogodkov ter druge naloge, povezane s procesiranjem naravnega jezika.

V projektu RSDO so bile imenske entitete ročno pregledane v korpusih SentiCoref 1.0 in ELEXIS-WSD, tj. v 20.166 povedih oz. 96,31 % novega gradiva. SentiCoref je že vseboval strojno pripisane oznake, entitete, ki se pojavljajo v koreferenčnih verigah, pa so bile tudi ročno pregledane, medtem ko je bil ELEXIS-WSD predoznačen v projektu, z orodjem CLASSLA-Stanza. Pri ročnem pregledu obeh korpusov smo sledili predhodno uveljavljenim označevalnim smernicam.[21] Kampanja pregledovanja je potekala v spletnem orodju INCEpTION (Klie idr., 2018), ki je preprosto za uporabo in nudi

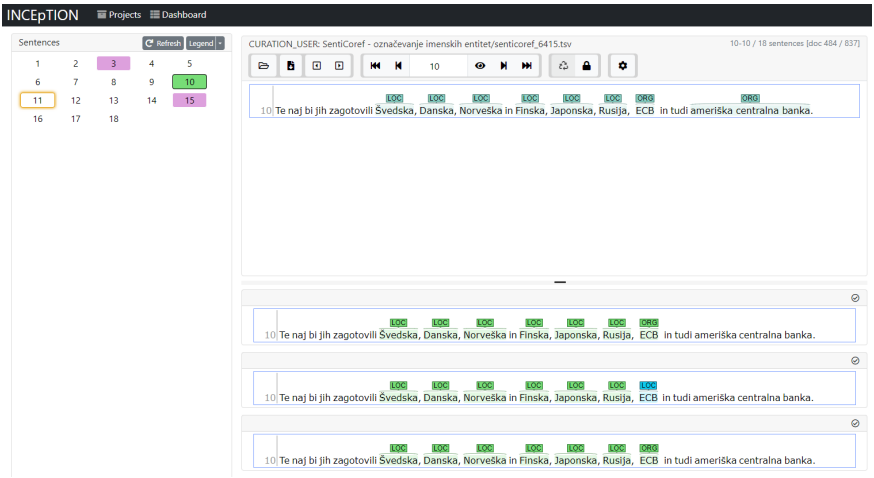
19 https://www.cost.eu/actions/CA21167/
20 https://elex.is/
21 https://wiki.cjvt.si/books/08-imenske-entitete/page/oznacevalne-smernice, gl. Različica 1.1.

dober pregled nad že opravljenim delom. Gradivo so pod vodstvom koordinatorja pregledovale tri študentke jezikoslovnih smeri. Vsako poved so pregledale vse tri študentke, neujemanja med pripisanimi oznakami pa je v fazi kuracije posebej obravnaval koordinator in tem primerom tudi pripisal končne oznake (Slika 4).

Pri označevanju se je pojavil pomislek glede obravnave ženskih oblik priimkov, ki so tvorjeni iz moških priimkov in so z oblikovnega vidika svojilni pridevniki (npr. *Kresalova*). Po tem kriteriju bi jim morali prisoditi oznako DERIV-PER, a smo tovrstnim primerom pripisali oznaka PER, saj pomensko delujejo kot osebno lastno ime, poleg tega pa so oblikoskladenjske lastnosti zabeležene na nivoju oblikoskladnje.

Kot problematično se je izkazalo tudi določanje začetka imenske entitete v primerih, ko je prvi del uradnega imena organizacije zapisan z malo začetnico, ker ga pisec besedila dojema kot vrstno poimenovanje (npr. *občina Gornja Radgona*). Obveljalo je splošno pravilo, po katerem je glavni kazalnik, da celotno enoto označimo kot imensko entiteto, velika začetnica ([*Občina Gornja Radgona*] ORG). V določenih primerih pa lahko kot imensko entiteto obravnavamo tudi primere, ki so zapisani z malo začetnico, a vsebujejo vse sestavine uradnega imena te institucije. Tak primer je [*ameriška centralna banka*]ORG, uradno slovensko poimenovanje pa je *Ameriška centralna banka*. Če je institucija zapisana kot parafraza uradnega imena, ne glede na to, ali je zapisana z malo ali veliko začetnico, je ne označimo kot imensko entiteto, npr. *Karavanški predor*, saj je uradno ime *predor Karavanke*. Posebna problematika označevalnega sistema je predvsem predpostavka, da avtorji besedil vedno upoštevajo pravopisna pravila in se tudi pozanimajo o uradni obliki imena določene institucije.

Pri obravnavi dilem se je tudi izkazalo, da bi poleg obstoječih oznak potrebovali še oznako za pridevnike iz stvarnih lastnih imen (npr. *Mercatorjev*), za katere bi po vzoru DERIV-PER lahko uvedli oznako DERIV-ORG. Enako velja za svojilne pridevnike iz entitet z oznako LOC (npr. *Lunin*), ki bi jim lahko pripisali oznako DERIV-LOC. Uvedba novih kategorij bi bil radikalnejši poseg v obstoječe smernice, kar bi bilo v prihodnje smiselno temeljiteje premisliti.



Slika 4: Prikaz faze kuracije v orodju INCEpTION: v spodnjih treh vrsticah vidimo odločitve pregledovalk, v zgornji vrstici pa je prikazana končna odločitev kuratorja.

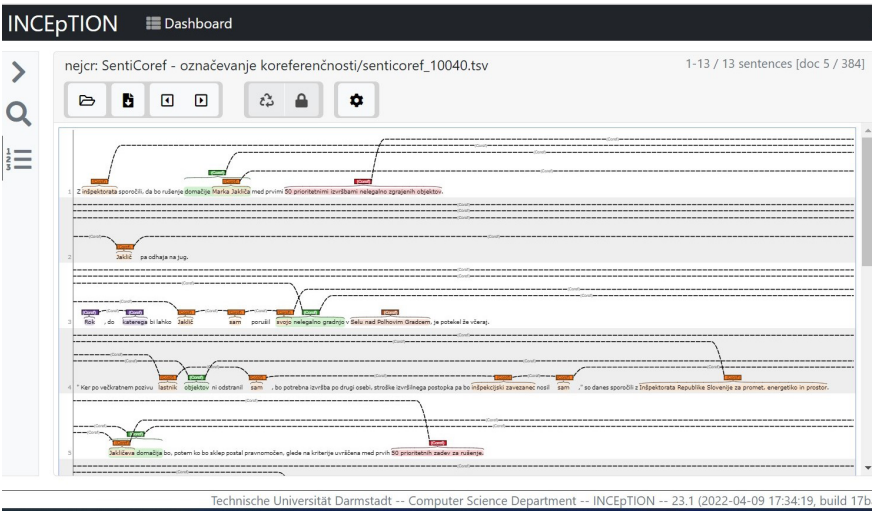
2.7 Koreference

»Odkrivanje koreferenčnosti je ena izmed treh ključnih nalog ekstrakcije informacij iz besedil, kamor spadata še prepoznavanje imenskih entitet in ekstrakcija povezav« (Žitnik in Bajec, 2018). V okviru projekta RSDO smo jo iskali v 837 besedilih korpusa SentiCoref 1.0. Besedila so obsegala 18.142 povedi oz. 391.962 pojavnic. Za iskanje koreferenc so najprimernejša krajša zaključena besedila, zato smo za to nalogo izbrali množico SentiCoref, drugo gradivo v učnem korpusu je namreč razdeljeno na odstavke ali še krajše enote.

V izbranih besedilih so predhodno že označili koreference (Žitnik in Bajec, 2018), vendar se je izkazala potreba po nadgradnji označevalnega sistema za slovanske jezike, saj ti referenčnost pogosto izražajo tudi morfemsko. Kot osnovo novega označevalnega sistema smo uporabili označevalne smernice ReLDI: Uputstvo za anotiranje koreferenci (interni dokument za projektno rabo), ki so v sklopu iniciative ReLDI 2008 nastale za potrebe srbskega jezika. Smernice smo prevedli v slovenščino, jih uredili in prilagodili, pri čemer je bila najpomembnejša odločitev označevalcev, da za razliko od srbske kampanje na ravni koreferenc ne označujemo skladenjskih značilnosti – za

slovenščino so te v korpusu SUK namreč dosledneje in celoviteje določene pri oblikoskladenjskih in skladenjskih oznakah. Smernice smo skupno pripravljali in dopolnjevali v spletnem urejevalniku *Google Dokumenti*, končna različica pa je na voljo na portalu Wiki CJVT.[22]

Kampanja označevanja koreferenčnosti je bila, kot kampanja označevanja imenskih entitet, opravljena na platformi INCEpTION (Slika 5). Gradivo sta pregledala dva raziskovalca, eden pa je kampanjo tudi koordiniral. Osnovne dileme so bile večinoma razrešene v uvajalni fazi, nekatere tudi pozneje med sprotno komunikacijo ob problemih pri označevanju samih besedil. Pomemben del uvajalne faze je bilo na primer poenotenje in določitev jasnejše terminologije v smernicah. Določili smo razmerja med termini *entiteta*, *koreferenčnost*, *koreferenca* in *omenitev*. Prav tako smo iz izvornih smernic odstranili del gradiva, ki je primerjalo označevalni sistem z alternativnimi pogledi na koreferenčnost, in poskrbeli za natančno členjenost poglavij ter pravilno označenost zgledov. Poenostavljene in eksplicitne smernice so izboljšale komunikacijo med označevalcema in sam označevalni sistem.



Slika 5: Označevanje koreferenc v orodju INCEpTION.

22 https://wiki.cjvt.si/books/09-odkrivanje-koreferencnosti/page/oznacevalne-smernice, gl. Različica 1.6.

Ob nadaljnjem označevanju se je v praksi izrazila pomembna konceptualna dilema označevanja koreferenčnosti, kadar so povezave med posameznimi omenitvami v besedilu zanikane ali pa je o povezavi posamezne omenitve z antecendentom pisec besedila izrazil dvom. Take primere najdemo predvsem pri novinarskih prispevkih, katerih temelj je naklonski členek *naj*, saj s pogojnikom (kondicionalom) izražajo konstanten dvom o resničnosti povezave med vršilcem dejanja in samim dejanjem. Pravilo, da se koreferenčnosti ob dvomu o povezavi znotraj samega besedila ali njenem zanikanju ne označuje, se je izkazalo za neizvedljivo in je posledično povzročalo precej težav. Pri nadaljnjem razvoju označevalnega sistema bo ta izziv treba upoštevati in sprejeti drugačne smernice ali pa te natančneje določiti s primeri konkretnih besedil, ne samo posameznih povedi. Nekaj težav je povzročal tudi vrstni red označevanja posameznih omenitev, saj je lahko v enem stavku samo ena koreferenca na posameznega referenta, zato smo določili, da imajo samostalniki prednost pred na primer zaimki. Pri veriženju vseh izpeljanih pridevnikov in lastnoimenskih samostalnikov pa je prihajalo do obsežnega kopičenja koreferenčnih povezav in s tem drobljenja, ki bi lahko povzročilo poznejšo slabšo ekstrakcijo informacij naučenih modelov.

Dele izvornih smernic, ki za slovenščino niso bili relevantni, smo umestili na konec dokumenta in zapisali posebno opozorilo, da v naši označevalni kampanji niso bili upoštevani. Ta del navodil smo v dokumentu vseeno ohranili, saj je v njih mnogo primerov povedi, označenih s koreferencami.

V nadaljnjih kampanjah bi bilo smiselno evalvirati uspešnost in ustreznost posameznih označevalnih odločitev in smernice še enkrat posodobiti. Trenutne zglede v smernicah je treba nadomestiti in dopolniti z realnimi zgledi iz korpusnih besedil, saj označevalna praksa razkrije številne izzive, na katere teoretične smernice ne dajejo natančnih odgovorov.

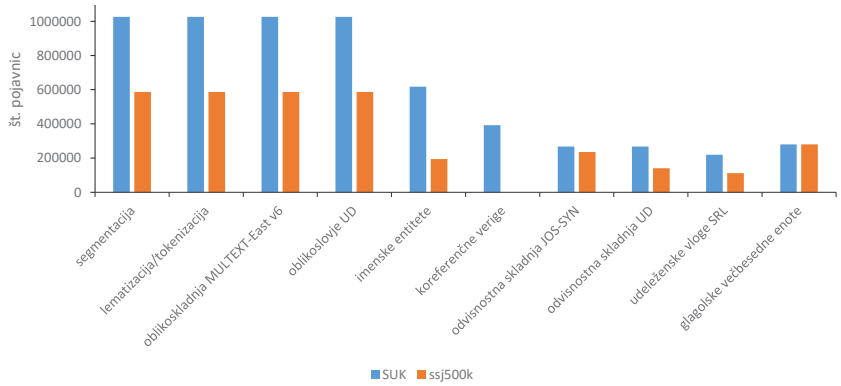
3 Kvantitativni pregled korpusa

Vseh 1.025.639 pojavnic novega učnega korpusa SUK je označenega in ročno pregledanega na ravni stavčne segmentacije, tokenizacije, lematizacije in oblikoskladenjskih oznak. Skoraj dve tretjini korpusa vsebujeta oznake imenskih entitet, dobrih 38 % celotnega korpusa pa je označenega na nivoju koreferenc. Približno četrtina korpusa vsebuje oznake odvisnostne skladnje po sistemih JOS-SYN in UD, približno petina pa oznake udeleženskih vlog SRL. Z oznakami glagolskih večbesednih enot je označenih 27 % gradiva, vse še iz ssj500k. Natančni podatki po označevalnih nivojih so prikazani v Tabeli 1.

Tabela 1: Količina pregledanega gradiva v SUK po označevalnih nivojih.

Označevalni nivo	Pojavnice	Povedi	Besedila	% celotnega SUK
Segmentacija	1.025.639	48.594	2.908	100
Lematizacija/tokenizacija	1.025.639	48.594	2.908	100
Oblikoskladnja MULTEXT-East v6	1.025.639	48.594	2.908	100
Oblikoslovje UD	1.025.639	48.594	2.908	100
Imenske entitete	617.832	29.654	1.336	60,24
Koreferenčne verige	391.962	18.142	837	38,22
Odvisnostna skladnja JOS-SYN	267.097	13.435	618	26,04
Odvisnostna skladnja UD	267.097	13.435	618	26,04
Udeleženske vloge SRL	219.216	11.748	598	21,37
Glagolske večbesedne enote	280.522	13.511	754	27,35

Označenost SUK-a v primerjavi s ssj500k je predstavljena v Grafu 2.



Graf 2: Primerjava označenega gradiva v ssj500k in SUK po označevalnih nivojih.

4 Kodiranje korpusa

Tako kot ssj500k je tudi SUK kodiran v formatu XML s shemo, ki sledi priporočilom TEI,[23] vendar po nadgrajeni kodirni shemi, ki jo priporoča CLARIN.SI.[24] Ker je SUK sestavljen iz več podkorpusov, ki imajo različne metapodatke o besedilih in ravni označevanja, je korpus oblikovan kot krovna datoteka TEI s kolofonom in povezavami na posamezne datoteke podkorpusov. Vsak podkorpus nato vsebuje razdelke z označenimi besedili.

Slika 6 prikazuje začetek podkorpusa SentiCoref, kjer vrhnji zamejuje podkorpus, gnezdeni pa prvo besedilo. V elementu <bibl> so podani metapodatki besedila, nato pa sledi začetek prvega odstavka in nato prve povedi. Prvi dve besedi sestavljata imensko entiteto tipa organizacija (<seg type="name" subtype="org">), besedi pa sta označeni z MULTEXT-East oblikoskladenjskimi oznakami in oblikoslovnimi lastnostmi po sistemu Universal Dependencies ter s svojo lemo.

23 https://tei-c.org/release/doc/tei-p5-doc/en/html/index.html
24 https://github.com/clarinsi/TEI-schema, gl. tudi https://www.clarin.si/repository/xmlui/page/data#tei

```
<div xmlns="http://www.tei-c.org/ns/1.0" xml:id="senticoref" xml:lang="sl">
   <div xml:id="senticoref1" xml:lang="sl">
      <bibl corresp="#senticoref1">
         <title>Zakaj se hrana draži?</title>
         <note type="handle">http://hdl.handle.net/11356/1285</note>
         <note type="source">SentiCoref 1.0</note>
         <note type="url">http://www.24ur.com/novice/gospodarstvo/zakaj-se-hrana-drazi.html</note>
         <note type="main_url">www.24ur.com</note>
         <note type="keywords">iztok, jarc, michel, barnier, hrana, podražitev</note>
         <note type="author">STA / V.L.</note>
         <date>2007-09-03</date>
      </bibl>
      <p xml:id="senticoref1.1">
         <s xml:id="senticoref1.1.1">
            <seg type="name" subtype="org" xml:id="senticoref1.1.1.ne1">
               <w ana="mte:Ppnzei" msd="UPosTag=ADJ|Case=Nom|Degree=Pos|Gender=Fem|Number=Sing"
                  lemma="evropski" xml:id="senticoref1.1.1.t1">Evropska</w>
               <w ana="mte:Sozei" msd="UPosTag=NOUN|Case=Nom|Gender=Fem|Number=Sing"
                  lemma="komisija" xml:id="senticoref1.1.1.t2">komisija</w>
            </seg>
```

Slika 6: Primer zapisa korpusa v TEI.

Kompleksnejše označevalne ravni, kot sta skladnja ali koreferenčnost, so kodirane v elementih <linkGrp> znotraj svoje povedi, ta element pa vsebuje skupino povezav, ki med seboj povežejo ustrezne elemente, lahko pa tudi podajo funkcijo povezave, kar se uporablja pri skladenjskih povezavah.

Kot ilustrira Slika 7 na primeru označevanja koreferenc, lahko poved vsebuje tudi segmente, ki združujejo večbesedne izraze (<seg type="coref">), ti segmenti, ali pa segmenti za imenske entitete, pa so nato prek povezav združeni v koreferenčno verigo.

Zapis XML oz. TEI je zelo ekspresiven, datoteke je tudi mogoče formalno validirati, vendar pa je za učinkovito uporabo takšnega kodiranja potrebna ustrezna programska oprema in poznavanje standarda XML, pa tudi zapis TEI je kompleksen in zahteva privajanje. V računalniškem jezikoslovju se je v zadnjih letih uveljavil bistveno bolj enostaven zapis CoNLL-U, ki je bil razvit v sklopu projekta Universal Dependencies, zato je korpus dostopen tudi v takšnem zapisu, ki pa sicer ne zajema kompleksnejših vrst oznak, kot so koreference.

CoNLL-U je enostaven tabelaričen format, namenjen zapisu oznak jezikoslovnih ravni do odvisnostne skladnje, dopušča pa tudi pripisovanje enostavnih metapodatkov povedim, odstavkom in besedilom. V stolpcu ‘ostalo’ je mogoče dopisati poljubne atribute

```
   <seg type="coref" xml:id="senticoref1.1.1.phr9-1">
      <w ana="mte:Ppnmeid"
         msd="UPosTag=ADJ|Case=Nom|Definite=Def|Degree=Pos|Gender=Masc|Number=Sing"
         lemma="francoski" xml:id="senticoref1.1.1.t17">francoski</w>
      <w ana="mte:Somei" msd="UPosTag=NOUN|Case=Nom|Gender=Masc|Number=Sing"
         lemma="kolega" xml:id="senticoref1.1.1.t18">kolega</w>
   </seg>
</seg>
<pc ana="mte:U" msd="UPosTag=PUNCT" lemma="." xml:id="senticoref1.1.1.t19">.</pc>
<linkGrp type="COREF">
   <link target="#senticoref1.1.1.nel #senticoref1.1.2.nel #senticoref1.1.3.nel"/>
   <link target="#senticoref1.1.1.phr52-1 #senticoref1.1.3.phr52-2 #senticoref1.1.11.phr52-3"/>
```

Slika 7: Primer zapisa koreferenčnih verig v TEI.

posameznim pojavnicam, z uporabo zapisa IOB pa tudi lastnosti niza pojavnic, kar je uporabljeno za označevanje imenskih entitet. Slika 8 ilustrira, ravno tako na primeru začetka korpusa SentiCoref, zapis oznak v formatu CoNLL-U, pri čemer smo izpustili nekaj za ilustracijo nepomembnih stolpcev.

```
# newdoc id = senticoref1
# title = Zakaj se hrana draži?
# handle = http://hdl.handle.net/11356/1285
# source = SentiCoref 1.0
# url = http://www.24ur.com/novice/gospodarstvo/zakaj-se-hrana-drazi.html
# main_url = www.24ur.com
# keywords = iztok, jarc, michel, barnier, hrana, podražitev
# author = STA / V.L.
# date = 2007-09-03
# newpar id = senticoref1.1
# sent_id = senticoref1.1.1
# text = Evropska komisija mora narediti analizo vzrokov rasti cen hrane , menita kmetijski minister Jarc in njegov francoski kolega .
1       Evropska        evropski        ADJ     Agpfsn  Case=Nom|Degree=Pos|Gender=Fem|Number=Sing_     NER=B-org
2       komisija        komisija        NOUN    Ncfsn   Case=Nom|Gender=Fem|Number=Sing                  NER=I-org
3       mora            morati          VERB    Vmpr3s  Aspect=Imp|Mood=Ind|Number=Sing|Person=3...     NER=O
```

Slika 8: Zapis korpusa v formatu CoNLL-U.

Ker ima SUK dva zapisa skladnje (UD in JOS-SYN), ima vsak skladenjsko označeni podkorpus dve različici CoNLL-U datotek, eno s skladnjo UD in angleškimi oznakami MULTEXT-East, drugo pa s skladnjo JOS in slovenskimi oznakami MULTEXT-East.

5 Dostopnost korpusa

Korpus SUK je dostopen za prevzem na repozitoriju jezikovnih virov raziskovalne infrastrukture CLARIN.SI (Arhar Holdt idr., 2022) pod licenco CC BY-SA 4.0, ki dovoljuje uporabo za poljubne namene, vključno s komercialnimi, vendar pod pogojem, da se prizna avtorstvo korpusa in, če se korpus nadgradi, da je nadgrajeni korpus na voljo pod enakimi pogoji kot izvirnik.

Vnos v repozitoriju vsebuje dve stisnjeni datoteki, in sicer SUK.TEI.zip (20 MB, ki se razširi v 198 MB) s korpusom, kodiranim v TEI, in SUK.CoNLL-U.zip (23 MB, razširjen v 169 MB) s korpusom v formatu CoNLL-U.

Korpus je za pregledovanje in analizo dostopen tudi prek konkordančnikov CLARIN.SI, tj. noSketch Engine (Rychlý, 2007) in KonText (Machálek, 2020), pri čemer so povezave do konkordančnikov na voljo prek vnosa v repozitoriju.

6 Ocena uspešnosti označevanja in novi označevalni modeli

Učna množica SUK je bila v okviru projekta RSDO uporabljena za učenje novih jezikovnih modelov za označevanje besedil. Pri tem smo uporabili označevalno orodje CLASSLA-Stanza,[25] ki je bilo v pretekli različici že naučeno na učnem korpusu ssj500k (Ljubešić in Dobrovoljc, 2019; Terčon in Ljubešic, 2023).

Kot pripravo na učenje modelov smo najprej izvedli delitev korpusa SUK na učno, validacijsko in testno podmnožico v razmerju 8 : 1 : 1.[26] Z uporabo učne in validacijske množice smo naučili modele za štiri ravni slovničnega označevanja: oblikoskladenjsko označevanje, lematizacija, skladenjsko razčlenjevanje (tako po sistemu JOS-SYN kot tudi po sistemu Universal Dependencies) in označevanje udeleženskih vlog. Naučene modele smo nato ovrednotili na testni množici. Rezultati evalvacije modelov so prikazani v Tabeli 2.[27]

Za vsak model so podane vrednosti izražene z oceno F1 v obliki odstotka, pri čemer je za oblikoskladenjsko označevanje prikazana ocena F1 za oznake vseh treh sistemov (MULTEXT-East v6, UD

25 https://pypi.org/project/classla/

26 Izvorna koda za proces delitve in vse nastale podmnožice so dostopne na https://github.com/clarinsi/suk-split.

27 Tabela prikazuje rezultate modelov, ki so pri učenju in evalvaciji za napovedovanje oznak uporabljali tudi novo različico slovenskega oblikoslovnega leksikona Sloleks 3.0 (Čibej idr., 2022). Celoten proces učenja in evalvacije modelov, vključno z uspešnostjo modelov, ki niso uporabljali leksikona, je podrobneje opisan na GitHub repozitoriju https://github.com/clarinsi/classla-training.

besedne vrste in UD lastnosti), za skladenjsko razčlenjevanje pa je prikazan F1 za splošno uveljavljeno oceno LAS (ang. *Labeled Attachment Score*), ki se pogosto uporablja za vrednotenje uspešnosti odvisnostnega označevanja (Nivre in Fang, 2017).

Tabela 2: Uspešnost modelov za vsako označevalno raven.

Označevalna raven	F1
Oblikoskladenjsko označevanje	97,08
Lematizacija	98,97
UD-skladnja	90,57
Skladnja JOS-SYN	93,89
Udeleženske vloge	76,24

Novi označevalni modeli večinoma dosegajo F1 vrednosti nad 90, z izjemo modela za označevanje udeleženskih vlog, za katerega je bilo v učni množici na voljo najmanj učnih podatkov od vseh zgoraj omenjenih označevalnih ravni. Po pregledu uspešnosti modela pri napovedovanju posameznih udeleženskih vlog se je izkazalo, da lahko pričakujemo precej višjo natančnost napovedovanja pri vlogah, ki so bistveno pogostejše (npr. ACT – F1 87,76 in PAT – F1 86,37), medtem ko pri redkejših model dosega manjše vrednosti (npr. ACMP – F1 36,36). Uspešnost in analiza najpogostejših napak modela za skladenjsko razčlenjevanje po sistemu UD sta podrobneje predstavljena v Dobrovoljc idr. (2023), podatki za skladnjo JOS-SYN pa v Arhar Holdt idr. (2023, 28).

Vsi omenjeni jezikovni modeli so že vključeni v najnovejšo različico orodja CLASSLA-Stanza 2.1 kot privzeti jezikovni modeli za označevanje standardne slovenščine.

7 Sklep in nadaljnje delo

V prispevku smo predstavili nadgradnjo učnega korpusa ssj500k v SUK 1.0. Korpus je temelj za učenje jezikoslovnega označevanja sodobne slovenščine, zato ga je nujno kontinuirano izboljševati in razvijati metodologijo njegove priprave. Specifične identificirane težave

in prioritete za vsako posamezno označevalno ravnijo so popisane v Arhar Holdt idr. (2023), tu pa navajamo splošne smernice nadaljnjega razvoja učnega korpusa SUK.

Prva razvojna prioriteta je **izboljševanje kakovosti korpusa in označevalnih smernic za vse vključene jezikovne ravnine.** Pri analizah označenosti korpusa ssj500k in delu z novimi podatki so se razkrile označevalne nedoslednosti, posredno pa tudi šibka mesta označevalnih smernic. Deloma so bile težave odpravljene, mestoma pa zahtevajo dodatno delo in širši strokovni konsenz o pripisovanju določenih jezikoslovnih kategorij na posameznih ravninah. Premišljeno usklajevanje potrebujemo tako znotraj korpusa, kjer se določene težave propagirajo med jezikovnimi ravninami, kakor tudi med korpusom SUK in drugimi jezikovnimi viri (leksikonom Sloleks, drugimi učnimi korpusi ipd.).

Druga prioriteta je **povečevanje korpusa**. Po gradnji korpusa SUK se potrjuje, da sta strojna lematizacija in pripisovanje oblikoskladenjskih oznak že dovolj natančna, da celostni ročni pregledi oznak niso več smiselni. V prihodnje bi se bilo bolje omejiti na ročno preverbo pri težavnih kategorijah, ki jih je mogoče predhodno (pol) avtomatsko identificirati. Za višje ravni je treba stremeti k povečanju deleža ročno označenega gradiva, za vse ravni pa zagotoviti ciljne in celostne analize mest, kjer se strojni označevalnik moti, in v primeru redkosti ali razpršenosti jezikovnih pojavov dodati ustrezno izbrane dodatne povedi ali besedila.

V povezavi s prejšnjo točko je treba izboljšati **žanrsko reprezentativnost korpusa** oz. vključiti domene sodobne standardne slovenščine, ki se glede na raziskave jezikovnih značilnosti pomembneje razlikujejo od splošne rabe, npr. pravna in uradovalna besedila, znanstveni jezik in pisanje učečih se. Kot omenjeno, trenutno poteka razvoj učnih korpusov za označevanje nestandardne slovenščine, govorjenega jezika in starejših besedil ločeno, zato je treba pri gradnji zagotoviti metodološko skladnost, npr. vključenost primerljivih oznak in pregledno, transparentno žanrsko specifično nadgrajevanje smernic. Na drugi strani je zlasti za semantično in diskurzno raven mogoče izbrati in dodati **nove vrste oznak**, saj so poleg koreferenčnosti relevantni tudi podatki za pomensko razdvoumljanje,

detekcijo metafor, ugotavljanje mnenj, detekcijo sovražnega govora in podobno.

Nujno je zagotoviti kontinuiran razvoj **orodij in spletnih servisov za označevanje, analizo in vizualizacijo** jezikoslovno označenih podatkov. Podatke višjih označevalnih ravni (skladnja, SRL) je trenutno mogoče vizualizirati v programu Q-CAT, ki omogoča tudi napredno iskanje; prav tako je mogoče po oznakah iskati v konkordančnikih, ki jih ponuja infrastruktura CLARIN.SI. Vendar sta vizualizacija in izvoz označenih podatkov v obeh orodjih omejena, zato niso enostavno, pregledno dostopni. Poseben izziv so oznake na ravneh, ki presegajo meje povedi, kot so koreferenčne verige. V nadaljevanju je treba razviti možnosti za uporabniku prijazno sočasno pregledovanje in izvažanje bogato označenih podatkov v berljivem formatu za jezikoslovne analize in druge nadaljnje rabe.

Učne množice in označevalne sheme je potrebno usklajevati s standardizacijskimi pobudami v **mednarodnem prostoru** ter sodelovati pri njihovem nastajanju. Mednarodno standardizirane učne množice omogočajo, da se za slovenščino razvijajo orodja v okviru mednarodnih konzorcijev in da je slovenščina del večjezikovnih učnih in evalvacijskih množic.

Glavni namen korpusa je razvoj strojnih označevalnikov, pri čemer je treba zagotoviti tudi njihovo **kakovostno in transparentno vrednotenje.** Na to potrebo odgovarja portal SloBENCH,[28] ki vsebuje evalvacijske množice za različne naloge procesiranja naravnega jezika. Ogrodje omogoča, da vsakdo gradi lastne modele in preko portala odda napovedi, kjer se avtomatsko izvede vrednotenje in objavi v izbrani lestvici. Trenutno je v ogrodju SloBENCH na voljo 8 javnih lestvic. Za naloge prepoznavanja imenskih entitet, odkrivanja koreferenčnosti in razčlenjevanja po sistemu UD so bile že pripravljene evalvacijske množice, ki so skladne z metodologijo priprave korpusa SUK. Za nadaljnji razvoj bi bilo potrebno v avtomatske sisteme vrednotenja vključiti tudi ostale ravni, npr. JOS-SYN in udeleženske vloge. Zaradi zagotavljanja čimbolj transparentnega vrednotenja je potrebno dodatno pripraviti označene korpuse, ki ne bodo javno

28 https://slobench.cjvt.si/

dostopni in bodo namenjeni izključno vrednotenju.

Nenazadnje je mogoče omeniti tudi potrebo po **diseminaciji korpusa SUK** ne le v domači in mednarodni razvojni skupnosti, ampak tudi v drugih strokah, zlasti jezikoslovju. Ročno pregledane oznake takšnega obsega in raznolikosti, kot jih prinaša SUK, so redek in trenutno neizkoriščen potencial za kvantitativne in kvalitativne empirične analize jezikovnih pojavov, s tem pa za posodobitev jezikovnega opisa, ki ga v prostoru nujno potrebujemo in bi imel pozitiven povratni vpliv tudi na jezikoslovno označevanje sodobne slovenščine.

Zahvala

Projekt Razvoj slovenščine v digitalnem okolju 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). Program Jezikovni viri in tehnologije za slovenski jezik (P6-0411) sofinancira Javna agencija za znanstvenoraziskovalno in inovacijsko dejavnost Republike Slovenije iz državnega proračuna.

Literatura

Arhar Holdt, Š., Bordon, D., Čibej, J., Dobrovoljc, K., Gantar, P., Lenardič, J., Munda, T., Pori, E., Robida, N., Terčon, L., in Žitnik, S. (2023). *Slovenski učni korpus: Množici SUK 1.0 in Janes-Tag 3.0: Poročilo projekta Razvoj slovenščine v digitalnem okolju*.

Arhar Holdt, Š., in Čibej, J. (2021). Analize za nadgradnjo učnega korpusa ssj500k. V Š. Arhar Holdt (Ur.), *Nova slovnica sodobne standardne slovenščine: viri in metode* (1. izd.). Znanstvena založba Filozofske fakultete.

Arhar Holdt, Š., Krek, S., Dobrovoljc, K., Erjavec, T., Gantar, P., Čibej, J., Pori, E., Terčon, L., Munda, T., Žitnik, S., Robida, N., Blagus, N., Može, S., Ledinek, N., Holz, N., Zupan, K., Kuzman, T., Kavčič, T., Škrjanec, I., … Zajc, A. (2022). *Training corpus SUK 1.0*. http://hdl.handle.net/11356/1747

Björkelund, A., Hafdell, L., in Nugues, P. (2009). Multilingual Semantic Role Labeling. V J. Hajič (Ur.), *Proceedings of the Thirteenth Conference*

on Computational Natural Language Learning (CoNLL 2009): Shared Task (str. 43–48). Association for Computational Linguistics. https://aclanthology.org/W09-1206

Brank, J. (2022). *Q-CAT Corpus Annotation Tool 1.4*. http://hdl.handle.net/11356/1684

Čibej, J., Gantar, K., Dobrovoljc, K., Krek, S., Holozan, P., Erjavec, T., Romih, M., Arhar Holdt, Š., Krsnik, L., in Robnik-Šikonja, M. (2022). *Morphological lexicon Sloleks 3.0*. http://hdl.handle.net/11356/1745

Čibej, J., Holdt, Š. A., Erjavec, T., in Fišer, D. (2018). *Ročno označeni korpusi Janes za učenje jezikovnotehnoloških orodij in jezikoslovne raziskave*. https://api.semanticscholar.org/CorpusID:165686166

de Marneffe, M.-C., Manning, C. D., Nivre, J., in Zeman, D. (2021). Universal Dependencies. *Computational Linguistics*, *47*(2), 255–308. https://doi.org/10.1162/coli_a_00402

Dobrovoljc, K., Erjavec, T., in Krek, S. (2017). The Universal Dependencies Treebank for Slovenian. V T. Erjavec, J. Piskorski, L. Pivovarova, J. Šnajder, J. Steinberger, in R. Yangarber (Ur.), *Proceedings of the 6th Workshop on Balto-Slavic Natural Language Processing* (str. 33–38). Association for Computational Linguistics. https://doi.org/10.18653/v1/W17-1406

Dobrovoljc, K., Krek, S., in Rupnik, J. (2012). Skladenjski razčlenjevalnik za slovenščino. V T. Erjavec in J. Žganec Gros (Ur.), *Zbornik Osme konference jezikovne tehnologije: Zvezek C* (str. 42–47). Institut »Jožef Stefan«.

Dobrovoljc, K., in Ljubešić, N. (2022). Extending the SSJ Universal Dependencies Treebank for Slovenian: Was It Worth It? V S. Pradhan in S. Kuebler (Ur.), *Proceedings of the 16th Linguistic Annotation Workshop (LAW-XVI) within LREC2022* (str. 15–22). European Language Resources Association. https://aclanthology.org/2022.law-1.3

Dobrovoljc, K., Terčon, L., in Ljubešić, N. (2023). Universal Dependencies za slovenščino: Nove smernice, ročno označeni podatki in razčlenjevalni model. *Slovenščina 2.0: empirične, aplikativne in interdisciplinarne raziskave*, *11*(1), 218–246. https://doi.org/10.4312/slo2.0.2023.1.218-246

Erjavec, T. (2012). MULTEXT-East: morphosyntactic resources for Central and Eastern European languages. *Language Resources and Evaluation*, *46*(1), 131–142. https://doi.org/10.1007/s10579-011-9174-8

Erjavec, T. (2015). The IMP historical Slovene language resources. *Language Resources and Evaluation*, *49*(3), 753–775. https://doi.org/10.1007/s10579-015-9294-7

Erjavec, T., Fišer, D., Krek, S., in Ledinek, N. (2010). The JOS Linguistically Tagged Corpus of Slovene. V N. Calzolari, K. Choukri, B. Maegaard, J. Mariani, J. Odijk, S. Piperidis, M. Rosner, in D. Tapias (Ur.), *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*. European Language Resources Association (ELRA). http://www.lrec-conf.org/proceedings/lrec2010/pdf/139_Paper.pdf

Gantar, P. (2021). Strojno berljiv Vezljivostni leksikon slovenskih glagolov. V Š. Arhar Holdt (Ur.), *Nova slovnica sodobne standardne slovenščine: viri in metode* (1. izd., str. 259–297). Znanstvena založba Filozofske fakultete.

Gantar, P. (2023). Analiza udeleženskih vlog s skladenjskega, pomenskega in leksikalnega vidika. V M. Smolej in M. Schlamberger Brezar (Ur.), *Prispevki k preučevanju slovenske skladnje* (1. izd., str. 77–97). Založba Univerze v Ljubljani. https://doi.org/10.4312/9789612970987

Klie, J.-C., Bugert, M., Boullosa, B., de Castilho, R., in Gurevych, I. (2018). The INCEpTION Platform: Machine-Assisted and Knowledge-Oriented Interactive Annotation. V D. Zhao (Ur.), *Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations* (str. 5–9). Association for Computational Linguistics. https://aclanthology.org/C18-2002

Krek, S., Arhar Holdt, Š., Erjavec, T., Čibej, J., Repar, A., Gantar, P., Ljubešić, N., Kosem, I., in Dobrovoljc, K. (2020). Gigafida 2.0: the reference corpus of written standard Slovene. V N. Calzolari (Ur.), *LREC 2020 : Twelfth International Conference on Language Resources and Evaluation* (str. 3340–3345). ELRA - European Language Resources Association.

Krek, S., Dobrovoljc, K., Erjavec, T., Može, S., Ledinek, N., Holz, N., Zupan, K., Gantar, P., Kuzman, T., Čibej, J., Arhar Holdt, Š., Kavčič, T., Škrjanec, I., Marko, D., Jezeršek, L., in Zajc, A. (2021). *Training corpus ssj500k 2.3*. http://hdl.handle.net/11356/1434

Krek, S., Erjavec, T., Dobrovoljc, K., Gantar, P., Arhar Holdt, Š., Čibej, J., in Brank, J. (2020). The ssj500k training corpus for Slovene language processing. V D. Fišer in T. Erjavec (Ur.), *Jezikovne tehnologije in digitalna humanistika: zbornik konference* (str. 23–33). Inštitut za novejšo zgodovino.

Krek, S., Gantar, P., Dobrovoljc, K., in Škrjanec, I. (2016). Označevanje udeleženskih vlog v učnem korpusu za slovenščino. V T. Erjavec in D. Fišer (Ur.), *Zbornik konference Jezikovne tehnologije in digitalna humanistika* (str. 106–110). Znanstvena založba Filozofske fakultete.

Ljubešić, N., in Dobrovoljc, K. (2019). What does Neural Bring? Analysing Improvements in Morphosyntactic Annotation and Lemmatisation of Slovenian, Croatian and Serbian. V T. Erjavec, M. Marcińczuk, P. Nakov, J. Piskorski, L. Pivovarova, J. Šnajder, J. Steinberger, in R. Yangarber (Ur.), *Proceedings of the 7th Workshop on Balto-Slavic Natural Language Processing* (str. 29–34). Association for Computational Linguistics. https://doi.org/10.18653/v1/W19-3704

Machálek, T. (2020). KonText: Advanced and Flexible Corpus Query Interface. V N. Calzolari, F. Béchet, P. Blache, K. Choukri, C. Cieri, T. Declerck, S. Goggi, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, in S. Piperidis (Ur.), *Proceedings of the Twelfth Language Resources and Evaluation Conference* (str. 7003–7008). European Language Resources Association. https://aclanthology.org/2020.lrec-1.865

Martelli, F., Navigli, R., Krek, S., Tiberius, C., Kallas, J., Gantar, P., Koeva, S., Nimb, S., Pedersen Sandford, B., Olsen, S., Langements, M., Koppel, K., Üksik, T., Dobrovoljc, K., Ureña-Ruiz, R.-J., Sancho-Sánchez, J.-L., Lipp, V., Váradi, T., Györffy, A., … Munda, T. (2021). Designing the ELEXIS Parallel Sense-Annotated Dataset in 10 European Languages. V I. Kosem in M. Cukr (Ur.), *eLex 2021 Proceedings: Proceedings of the eLex 2021 conference* (str. 377–395). Lexical Computing CZ, s.r.o.

Mikulová, M., Bémová, A., Hajič, J., Hajičová, E., Havelka, J., Kolářová, V., Kučová, L., Lopatková, M., Pajas, P., Panevová, J., Razímová, M., Sgall, P., Štěpánek, J., Urešová, Z., Veselá, K., in Žabokrtský, Z. (2006). *Annotation on the tectogrammatical layer in the Prague Dependency Treebank: Annotation manual*.

Nivre, J., in Fang, C.-T. (2017). Universal Dependency Evaluation. V M.-C. de Marneffe, J. Nivre, in S. Schuster (Ur.), *Proceedings of the NoDaLiDa 2017 Workshop on Universal Dependencies (UDW 2017)* (str. 86–95). Association for Computational Linguistics. https://aclanthology.org/W17-0411

Pori, E., Čibej, J., Munda, T., Terčon, L., in Arhar Holdt, Š. (2022). Lematizacija in oblikoskladenjsko označevanje korpusa SentiCoref. V D. Fišer in T. Erjavec (Ur.), *Jezikovne tehnologije in digitalna humanistika: zbornik konference*. Inštitut za novejšo zgodovino.

Rychlý, P. (2007). Manatee/Bonito-A Modular Corpus Manager. V P. Sojka in A. Horák (Ur.), *Recent Advances in Slavonic Natural Language Processing, RASLAN* (str. 65–70). Masaryk University.

Tercon, L., in Ljubesic, N. (2023). CLASSLA-Stanza: The Next Step for Linguistic Processing of South Slavic Languages. *ArXiv*, *abs/2308.04255*. https://api.semanticscholar.org/CorpusID:261881905

Toporišič, J. (2004). *Slovenska slovnica*. Obzorja.

Žitnik, S., in Bajec, M. (2018). Odkrivanje koreferenčnosti v slovenskem jeziku na označenih besedilih iz coref149. *Slovenščina 2.0: Empirične, Aplikativne in Interdisciplinarne Raziskave*, *6*(1), 37–67.

Žitnik, S., Blagus, N., in Bajec, M. (2022). Target-level sentiment analysis for news articles. *Knowledge-Based Systems*, *249*, 108939. https://doi.org/https://doi.org/10.1016/j.knosys.2022.108939

Zasnova splošnega ogrodja in podatkovnega modela za obdelavo naravnega jezika – ANGLEr

Slavko ŽITNIK
Univerza v Ljubljani, Fakulteta za računalništvo in informatiko

Povzetek

S področjem analize naravnega jezika se danes ne ukvarjajo le strokovnjaki. Obdelava naravnega jezika se uporablja na najrazličnejših vsakodnevnih področjih, kot je na primer analiza sentimenta v ocenah izdelkov ali avtomatsko uvrščanje bančnih transakcij. Zaradi uspešnega razvoja področja v zadnjih letih bi vse več uporabnikov želelo analizirati besedila, vendar jim primanjkuje tehničnega znanja. V razdelku zato pregledamo obstoječa ogrodja za demokratizacijo obdelave naravnega jezika in predlagamo lastno splošno ogrodje ANGLEr, vključno s celostnim podatkovnim modelom. Ogrodje bi omogočilo raziskovalcem enostavno vključitev novih metod v ogrodje ali njihovo prototipiranje oz. uporabo v okviru predvidenega cevovoda. Tehnično nevešči uporabniki pa bi lahko preko grafičnega vmesnika dodajali komponente in izvajali analize besedil za svoje namene. Ključna prednost predlaganega ogrodja v primerjavi z obstoječimi je (a) enoten, razširljiv, hierarhičen, verzioniran, odprt, tehnološko agnostičen in enostavno razumliv podatkovni model ter (b) šibko sklopljeno ogrodje, temelječe na osnovi Docker vsebnikov in spletnih tehnologijah. Odprtokodno ogrodje ponuja tudi širše možnosti monetizacije storitev ali razvoj tržnice modulov. Vsekakor pa bi krovna organizacija oz. skupnost morala skrbeti za smer razvoja ogrodja in aktivno promocijo uporabe.

Ključne besede: ANGLEr, demokratizacija, obdelava naravnega jezika, ogrodje, podatkovni model

Abstract

Natural language processing is no longer confined to professionals only; its techniques are now used in various areas such as sentiment analysis in product review or automatic classification of bank transactions. Due to the latest advancements in the field, more users want to analyze texts but they lack technical knowledge. In this chapter we review existing frameworks for democratization of the natural language processing and propose a new general framework ANGLEr along with a comprehensive data model. The framework would enable researchers that develop new techniques to easily incorporate their methods into the framework, prototype them or use them within a larger workflow. Technically unsavvy users on the other side would use a user interface, add components and process texts for their needs. The key advantages of the proposed framework compared to existing ones are (a) a unified, extensible, hierarchical, versioned, open, technologically agnostic and easily comprehensible data model, and (b) a loosely-coupled framework based on Docker containers and Web technologies. The opensource framework offers multiple possibilities of monetization and development of a marketplace of additional modules. We strongly believe that one organization or community would need to take care of future developments and promote the usage of the framework.

Keywords: ANGLEr, democratization, natural language processing, framework, data model

1 Uvod

V zadnjem času opažamo porast interesa za različna splošna ogrodja, ki omogočajo uporabo naprednih metod za obdelavo naravnega jezika za tehnično manj vešče uporabnike. Tehnično vešči uporabniki si lahko brez težav najdejo ustrezno orodje, v katerega lahko vključijo najboljše raziskovalne rezultate iz odprtokodnih orodij in jih uporabljajo za svoje namene. Primeri takšnih so na primer LangChain,[1] Hugging Face Enterprise Hub[2] in podobni. Za manj vešče

1 https://www.langchain.com
2 https://huggingface.co/enterprise

uporabnike so namenjene t.i. *low-code/no-code* platforme, ki omogočajo, da lahko s pomočjo vizualnih gradnikov zgradijo aplikacijo za lastne namene.[3] Takšne platforme pa so navadno preveč splošne, zaprte in ne omogočajo uporabe najnovejših raziskovalnih izsledkov. Nekje vmes obstajajo splošna ogrodja za obdelavo naravnega jezika, ki se razvijajo že vsaj od leta 2002, odkar je bilo ogrodje UIMA (Ferrucci in Lally, 2004) preneseno 6000-krat in orodje GATE (Cunningham, 2002) več kot 400.000-krat. Uporabniki obeh orodij so različnih profilov – raziskovalci, jezikoslovci, gospodarstvo ali razvijalci programske opreme. Za računalniško manj vešče uporabnike takšna orodja ponujajo grafični vmesnik, ki omogoča enostavno gradnjo cevovodov za obdelavo naravnega jezika.

Namen razdelka je pregledati ogrodja, ki omogočajo demokratizacijo uporabe metod umetne inteligence, natančneje metod obdelave naravnega jezika, pri čemer omogočajo raziskovalcem, da svoja orodja in modele enostavno vgradijo v takšna orodja in jih ponudijo v uporabo javnosti. Na podlagi pregleda predlagamo novo ogrodje ANGLEr – *A Next-Generation Natural Language Exploratory Framework*, ki bi raziskovalcem na področju obdelave naravnega jezika (ONJ) omogočilo enostavno implementacijo naprednih orodij v sistem, kjer bi lahko hitro pokazali praktično uporabo svojih metod in omogočili uporabo ostalim. Primerljiva ogrodja, ki so na voljo do sedaj, je bodisi težko splošno razširjati ali pa jih je težje uporabljati. V prispevku identificiramo ključne komponente takšnega ogrodja in predlagamo izboljšave glede na slabosti obstoječih ogrodij.

Glavni doprinosi predlaganega ogrodja ANGLEr so naslednji:

a) Razširljiva in porazdeljena arhitektura na osnovi vsebnikov Docker, ki omogoča enostavno uporabo in dodajanje orodij v sistem.
b) Splošen in odprt podatkovni model, ki omogoča shranjevanje podatkov in kompatibilnost z različnimi orodji.
c) Grafični vmesnik, ki pohitri proces izgradnje cevovoda obdelave naravnega jezika za uporabnike brez tehničnega znanja.

3 Primer no-code platforme za obdelavo besedil: https://monkeylearn.com

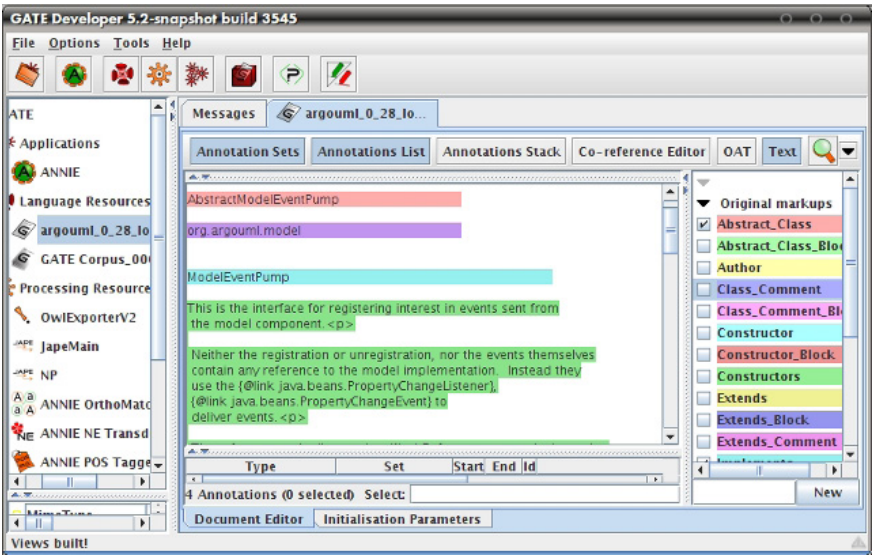
V nadaljevanju predstavimo obstoječa ogrodja in jih primerjamo med seboj (Razdelek 2). Posebej pregledamo uporabljene podatkovne modele in predlagamo lastnega (Razdelek 3). Na podlagi pregleda zasnujemo predlog novega splošnega ogrodja ANGLEr (Razdelek 4) in tudi predstavimo primer uporabe s pomočjo zaslonskih mask grafičnega vmesnika (Razdelek 5). Sklepne ugotovitve in ključna vprašanja za nadaljnji razvoj predstavimo v Razdelku 6.

2 Pregled obstoječih ogrodij

Na voljo je že kar nekaj razvitih ogrodij, ki omogočajo obdelavo besedil. Podrobneje smo pregledali obstoječa ogrodja in izbrali tista, ki so najbolj uporabna za uporabnike z omejenim tehničnim predznanjem. Analiza prednosti in slabosti posameznih nam je omogočila definirati seznam lastnosti, ki jih mora nasloviti novo ogrodje, kar prikazujemo v Tabeli 1. Orodja primerjamo predvsem glede na (a) grafični vmesnik, ki vpliva na uporabnost za nove uporabnike, (b) podatkovni model, ki vpliva na razširljivost, in (c) programski jezik ali okolje za razvoj novih vtičnikov, kar je pomembno za raziskovalce in razvijalce.

2.1 General Architecture for Text Engineering (GATE)

GATE (Cunningham, 2002) je eno izmed najstarejših in najbolj uporabljenih ogrodij za ONJ (Slika 1). Načrtovan je bil z namenom uporabe enotnega podatkovnega modela, ki naj bi podpiral širok nabor orodij za ONJ, vključno z grafičnim vmesnikom. GATE ponuja za delo in pripravo cevovodov za ONJ grafični vmesnik, ki pa je zastarel in ni bil posodobljen že več kot 10 let. Program je sicer enostavno namestiti, teče kot aplikacija Java na osebnem računalniku, vendar vmesnik ni intuitiven za nove uporabnike. Ogrodje omogoča razvoj dodatnih vtičnikov, ki pa morajo biti pripravljeni v programskem jeziku Java. Vsebuje tudi obstoječ nabor predpripravljenih vtičnikov, ki pa ne vsebujejo naprednejših metod. Zaradi vsega tega je ogrodje težko prilagoditi za uporabo novejših algoritmov.



Slika 1: Grafični vmesnik orodja GATE.

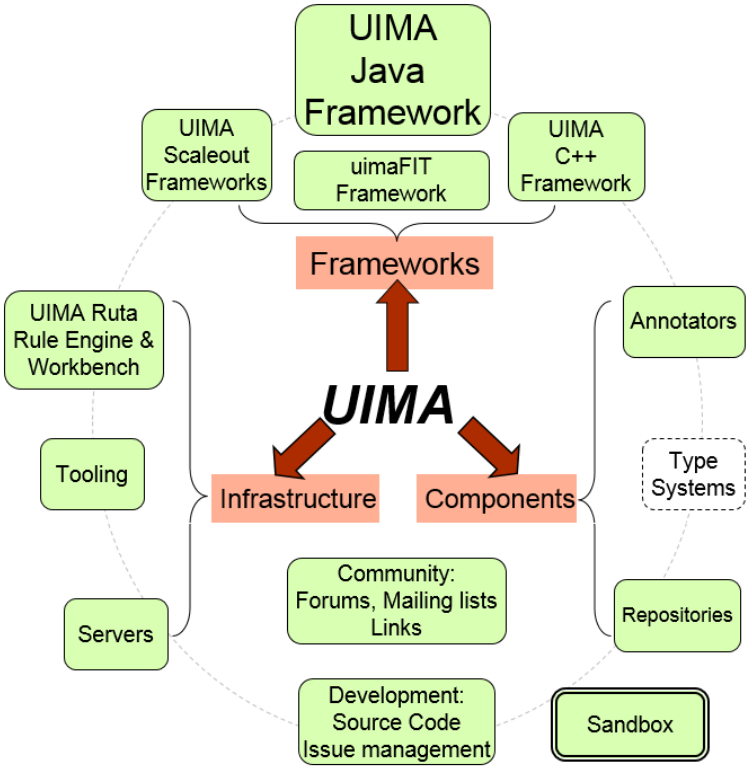
Orodje razvija Univerza Sheffield in podjetje OntoText. Skupaj ponujajo nadgrajeno ogrodje kot oblačno storitev in iskalnik GATE Mimir. Prav tako imajo objavljen seznam vtičnikov, ki so v tej storitvi na voljo.[4]

2.2 Unstructured Information Management Applications (UIMA)

UIMA (Ferrucci in Lally, 2004) je splošno ogrodje za ekstrakcijo informacij iz nestrukturiranih dokumentov, kot so besedila, slike, e-pošta in podobno (Slika 2). Je odprtokodni Apache projekt s standardizirano implementacijo tehničnega standarda UIMA, ki ga je pripravila standardizacijska organizacija OASIS. Vključuje tudi splošen podatkovni model, ki je sposoben hraniti podatke vseh razvitih komponent. Ogrodje ponuja tudi grafični vmesnik za nekatere splošne komponente. Vmesnik ni del enotne aplikacije in ne omogoča enostavnega načina kombiniranja orodij med seboj. Primarni način uporabe ogrodja še vedno zahteva urejanje XML opisov za označevanje

4 https://cloud.gate.ac.uk/shopfront

dokumentov, kar omejuje uporabnost za nove uporabnike in upočasnjuje razvoj. Nove komponente so lahko razvite v programskem jeziku C++ ali Java. Ogrodje pa omogoča, da so aplikacije UIMA razvite kot ločene komponente, kot na primer »identifikacija jezika« → »jezikovnoodvisno razčlenjevanje« → »prepoznavanje stavkov« → »prepoznavanje imenskih entitet.« Vsaka izmed komponent mora implementirati določene vmesnike ogrodja in definirati samoopisne metapodatke s pomočjo datotek XML. Namen ogrodja je, da upravlja s komponentami in med njimi posreduje podatke. Vsaka izmed komponent lahko teče na ločenem strežniku kot spletna storitev ali je replicirana na gruči strežnikov.



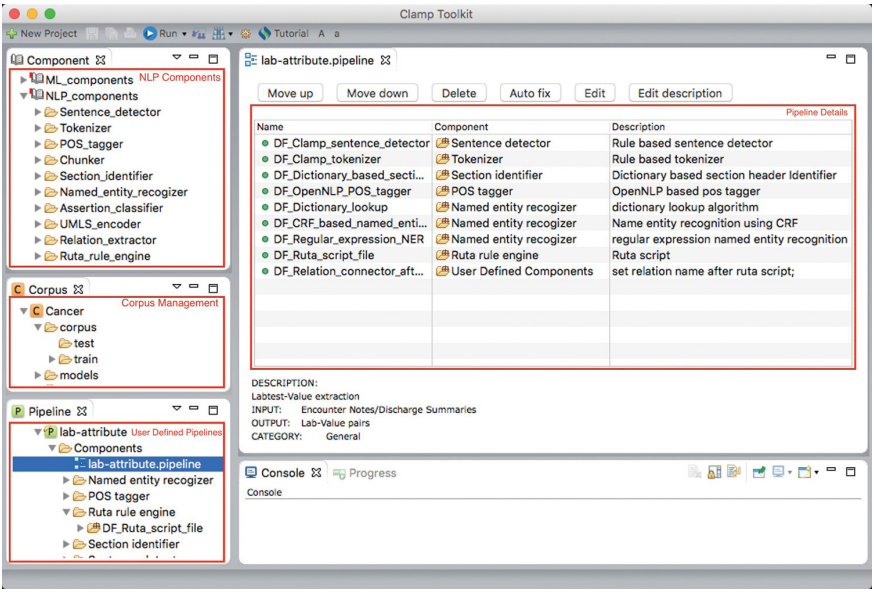
Slika 2: Tipi projektov v okolju UIMA.

Ogrodje UIMA omogoča konfiguracijo in zagon cevovodov za komponente za označevanje. Uporabniki lahko razvijajo svoje lastne

komponente ali nastavljajo in uporabljajo obstoječe. Nekaj komponent je na voljo kot del glavnega projekta, večina preostalih pa je dostopnih preko različnih repozitorijev na internetu. Platforma GitHub beleži več kot 900 repozitorijev, ki uporabljajo UIMA Java SDK. Novejše izvedbe med komponentami podpirajo tudi komunikacijo REST. Pomembno je omeniti, da je bil sistem IBM Watson, ki je osvojil tekmovanje Jeopardy v letu 2013, razvit na osnovi arhitekture UIMA.

2.2.1 Clinical Language Annotation, Modeling, and Processing (CLAMP)

Skupina raziskovalcev na področju obdelave biomedicinskih besedil (Soysal idr., 2018) je razvila orodje CLAMP, ki temelji na ogrodju UIMA. Orodje vključuje 10 namenskih UIMA komponent, kot na primer za prepoznavanje stavkov, kratic, imenskih entitet, povezav s shemo UMLS (Bodenreider, 2004) ali orodje za izvajanje pravil. Poleg tega so razvili tudi grafični vmesnik, ki temelji na ogrodju Eclipse (Slika 3). Orodje za njihove namene omogoča tudi označevanje besedil in

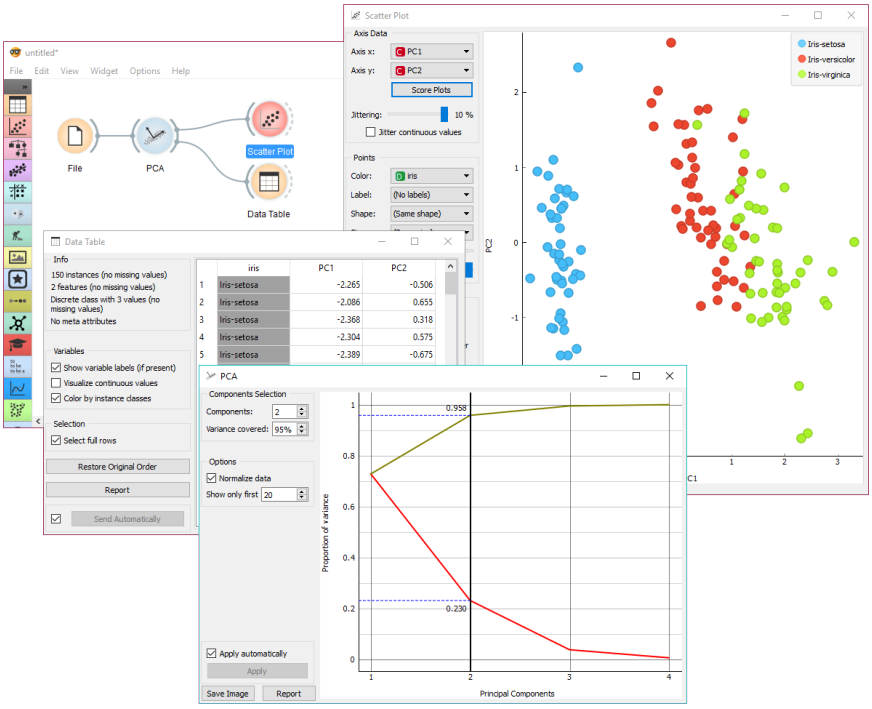


Slika 3: Prikaz grafičnega vmesnika CLAMP.

učenje modelov strojnega učenja. Ogrodje grafičnega vmesnika je zastarelo, težko razširljivo in po naših izkušnjah neintuitivno. Kljub temu je za začetnika bolj primerno kot vmesnik GATE. Celoten projekt je odprtokoden, vendar ga uporablja le manjša množica raziskovalcev na ožjem domenskem področju in tako ni prilagojen za splošno uporabo.

2.3 Orange - Data Mining Fruitful and Fun

Orodje Orange Data Mining (Demšar idr., 2013) se aktivno razvija in redno posodablja na Univerzi v Ljubljani in je eno izmed večkrat nagrajenih orodij za demokratizacijo umetne inteligence (Slika 4). Prihodnje izdaje predvidevajo tudi oblačno verzijo.



Slika 4: Primer uporabniškega vmesnika Orange.

Glavna lastnost orodja po presoji razvijalcev je intuitiven uporabniški vmesnik, primeren tudi za netehnične uporabnike. Orodje

ponuja velik nabor vtičnikov za naloge strojnega učenja in vizualizacije podatkov. Poleg tega lahko razvijalci uporabijo Orange kot programsko knjižnico. Orange je primarno namenjeno obdelavi relacijskih podatkov za klasično strojno učenje. Poleg tega pa ponuja dve razširitvi za obdelavo besedil (Razdelek 2.3.1 in 2.3.2). Največji problem pri ONJ v Orangu je tabelarična predstavitev podatkov, ki ni najbolj primerna za obdelavo besedil. Po drugi strani pa takšna predstavitev omogoča neposredno interoperabilnost z množico obstoječih orodij. Razvijalci lahko tudi razvijejo svoj vtičnik, vendar so pri tem precej vezani na programski jezik Python, specifično verzijo paketa Orange in arhitekturne značilnosti paketa, namesto šibko sklopljene komunikacije. V primeru šibke sklopljenosti bi lahko bili posamezni deli ogrodja razviti povsem ločeno med seboj, pri čemer bi jih povezoval le način medsebojne komunikacije, ki bi ga lahko vsaka komponenta implementirala na svoj način.

2.3.1 Vtičnik Orange text mining

Vtičnik Orange text mining (Slika 5) se lahko namesti neposredno preko uporabniškega vmesnika Orange. Vtičnik poleg tabelaričnega formata implementira podatkovna tipa Dokument in Korpus, vendar se zdi, da je tabelarični format še vedno glavni tip predstavitve podatkov. Na voljo pa je sicer tudi komponenta za pretvorbo podatkov med tipi. Vtičnik ponuja nekaj pogosto uporabljenih metod za predprocesiranje besedil (tokenizacija, lematizacija, filtriranje besed ipd.). Poleg tega vsebuje tudi nekaj naprednejših orodij, kot so *similarity hash* (pretvorba dokumentov v podobnostne vektorje), analiza sentimenta, določanje čustev v čivkih in podobno.[5]

5 https://orange3-text.readthedocs.io/en/latest

Slika 5: Vtičniki Orange text mining (levo) in Textable (desno).

2.3.2 Textable

Textable je ločena veja projekta Orange (Slika 5), ki jo razvija podjetje LangTech v sodelovanju z oddelkom za jezik in informacijske znanosti Univerze v Lausanne. Zdi se, da se ne posodablja redno.[6] Orodje definira uporabo tekstovnih datotek in ima podporo za datoteke tipa JSON in spletne URL vire. Verzija 3 podpira tudi različne nivoje besedilnih segmentov. Več različnih podatkovnih tipov sicer omogoča implementacijo več različnih algoritmov, vendar pa se zaradi tega izgubi velika mera kompatibilnosti med orodji, kar nas lahko omejuje pri gradnji kompleksnejših cevovodov za analizo besedil. S stališča podpore analizam orodje ponuja osnovno obdelavo besedil

6 http://textable.io

(npr. osnovno predprocesiranje, konkordance, kolokacije, matrike dokument-termin, lematizacijo ali oblikoslovno označevanje), povezavo z zunanjimi knjižnicami (npr. NLTK, Pattern, GenSim) in uvoz vsebin (npr. HTML, CSV, XML).

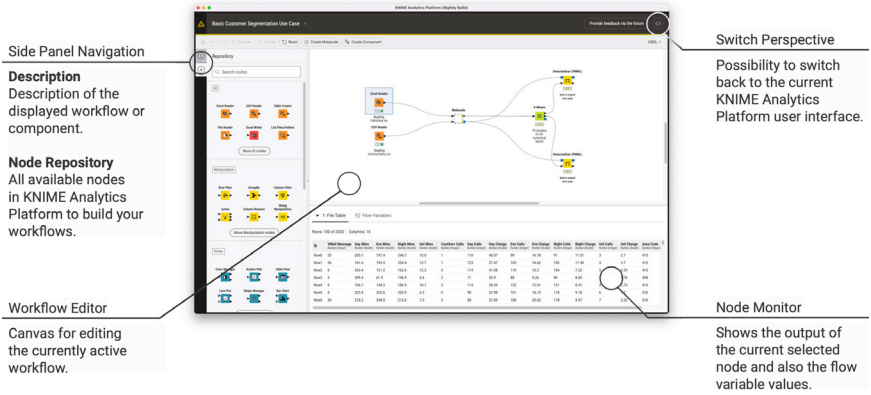
2.4 KNIME Analytics Platform

Platforma za podatkovne vede KNIME[7] je odprtokodno ogrodje, ki uporabnikom omogoča analizo in vizualizacijo podatkov brez programiranja. Njen *low/no-code* grafični vmesnik (Slika 6) ponuja množico orodij za začetnike in bolj napredne uporabnike. Platformo sicer razvija podjetje, ki prodaja tudi plačljive napredne storitve (npr. delo v oblaku, zasebni prostor, specializirane vtičnike, spletne storitve za integracijo z drugimi sistemi in podporo).

Platforma se zdi podobna orodju Orange, le da komponente komunicirajo preko spletnih storitev. Tržnica komponent vsebuje več tisoč orodij, ki jih je možno namestiti enostavno s *povleci-spusti* neposredno iz spletnega brskalnika. Komponente vsebujejo tudi najbolj napredne implementacije algoritmov (npr. uporaba modelov BERT ali pogovornih velikih jezikovnih modelov). Ravno tako kot v Orange lahko nekdo shrani postopek analize (angl. *workflow*) – tudi ti so na voljo na spletni tržnici, kjer jih je na voljo več kot 18 tisoč.

Platforma je na voljo kot namestitveni paket za več operacijskih sistemov in se namesti neposredno na osebni računalnik. Na voljo je tudi strežniška namestitev za napredne uporabnike. V primerjavi z novejšimi tehnologijami, platforma v ozadju uporablja ogrodje Eclipse, ki mu je potrebno slediti tudi pri razvoju lastnih komponent in ga obvezno uporabljati kot razvojno okolje. Osnovni programski jezik je tako Java. Navodila za razvijalce jasno definirajo strukturo projekta in tudi podatkovni model. Ogrodje kot osnovni tip podatkov predvideva *BufferedDataTable*, ki je splošna predstavitev tabelaričnih podatkov, ki jih lahko razvijalec prilagodi po lastnih željah.

7 https://www.knime.com



Slika 6: Grafični vmesnik platforme KNIME.

2.5 Programske knjižnice za obdelavo naravnega jezika

Obstaja mnogo programskih knjižnic za ONJ. Knjižnice načeloma omogočajo uporabo metod ONJ neposredno iz programskih jezikov in ne ponujajo grafičnih vmesnikov. Nekatere knjižnice se razvijajo že dalj časa, zelo popularne pa so nastale v zadnjih nekaj letih. Namenjene so predvsem ekspertom/razvijalcem in niso primerne za začetnike. Kljub temu predstavljajo pomemben del ekosistema, zato omenjamo nekaj glavnih.

Ena izmed bolj splošnih in starejših knjižnic je OpenNLP,[8] ki podpira razvoj aplikacij v programskem jeziku Java. Podobno obstaja kar nekaj knjižnic za programski jezik Python, npr. NLTK (Bird in Loper, 2004) ali Gobbli (Nance in Baumgartner, 2021), katere namen je poenostaviti uporabo globokega učenja za ONJ. Precej popularna je Stanza (Qi idr., 2020), naslednica tradicionalne knjižnice Stanford Core NLP (Manning idr., 2014), ki se razvija na Univerzi Stanford in podpira več različnih jezikov. V gospodarstvu večinoma tradicionalno uporabljajo Spacy,[9] ki vsebuje tudi kar nekaj novejših prednaučenih modelov.

Novejše knjižnice se osredotočajo predvsem na uporabo globokih nevronskih mrež in (velikih) jezikovnih modelov. Huggingface[10]

8 https://opennlp.apache.org
9 https://spacy.io
10 https://huggingface.co

se je začela razvijati kot knjižnica in sedaj predstavlja ekosistem, ki ponuja korpuse, prednaučene modele, knjižnico, repozitorij z gostovanjem in enostavno ogrodje *Spaces* za osnovno grafično preskušanje modelov. Podobno se knjižnica LangChain[11] usmerja na velike jezikovne modele, večinoma prilagojene za pogovore, pri čemer ločen projekt LangFlow[12] zagotavlja tudi grafične komponente za razvite modele.

Pisanje programov je precej počasnejše kot konstrukcija cevovodov preko grafičnega vmesnika, ki ima na voljo predpripravljene komponente. To je ena izmed ključnih funkcionalnosti za večjo uporabnost našega predlaganega ogrodja. Podobno pa se kaže tudi pri novejših programskih knjižnicah, ki želijo razvijalcem omogočiti čimhitrejše prototipiranje rešitev z uporabniki in zato ponujajo komponente za določen tip problemov ONJ.

2.6 Primerjava pregledanih ogrodij

Tabela 1 prikazuje primerjavo pregledanih ogrodij glede na izbrane dimenzije.

Tabela 1: Primerjava različnih ogrodij za obdelavo naravnega jezika.

Ogrodje	**Grafični vmesnik**	**Enotni podatkovni model**	**Programski jezik vtičnikov**
GATE	Da (OS)	Da (presplošen)	Java
UIMA	Da (Omejeno)	Da	Java ali C++
Orange /z vtičniki	Da (OS)	Različen (transformacije na voljo)	Python
KNIME	Da (OS)	Da (tabelaričen)	Java (Eclipse)
Knjižnice ONJ	Ne	Različno	Odvisno od knjižnice
ANGLEr	**Da (Splet)**	**Da (verzionirano)**	**Poljuben (Docker)**

Edini tip orodij, ki je najbolj različen od naših zahtev, so programske knjižnice. Vsa ostala štiri ogrodja so na voljo kot aplikacije,

11 https://www.langchain.com/
12 https://github.com/logspace-ai/langflow

ki jih je potrebno namestiti na operacijski sistem, medtem ko mi predlagamo spletno aplikacijo, ki teče v vsebnikih Docker in ne zahteva posebne namestitve. GATE, UIMA in KNIME omogočajo spletne storitve, ki se jih lahko namesti v oblačno infrastrukturo, vendar do neke mere zahtevajo uporabo predvidenega programskega jezika za izdelavo razširitev. V našem primeru predlagamo definicije programskih vmesnikov in enotnega podatkovnega modela za komunikacijo med posameznimi komponentami. Takšna arhitektura je tudi zelo šibko sklopljena in razširljiva in omogoča, da lahko vsakdo uporablja poljubno tehnologijo za razvoj dodatnih vtičnikov. Orange in KNIME ponujata prijazen uporabniški vmesnik z velikim naborom izdelanih komponent. Oba ponujata tudi možnost shranjevanja in deljenje cevovodov. KNIME uporablja ogromna odprtokodna skupnost, ki v ogrodje vključuje tudi najnovejše algoritme. Oba projekta se izvajata pod okriljem organizacij, ki že dalj časa vzdržujeta in posodabljata ogrodji. Slednje je predvsem pomembno, saj je potrebno ogrodja ves čas posodabljati, da delujejo z aktualnimi sistemi, zaradi česar lahko takšni projekti ostajajo aktivni.

3 Podatkovni modeli

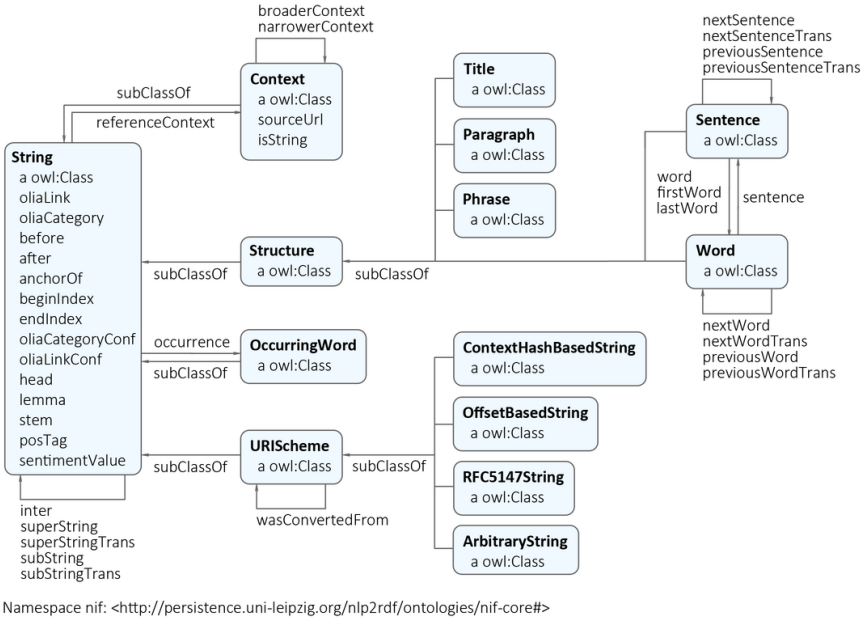
Podatkovni model omogoča predstavitev podatkov v določeni obliki. Podobno so definirane (ali standardizirane) predstavitve oznak v učnih korpusih (npr. TEI XML), ki omogočajo, da si lahko širša množica raziskovalcev med seboj deli podatke in jih razume. Podobno je pri razvoju programske opreme, saj želimo doseči, da bi algoritmi lahko »razumeli« vhode in izhode med seboj, brez da bi morali razvijalci prilagajati izhode predhodnih algoritmov v njihove ustreznice za vhode za druge algoritme. Zaradi tega morajo biti podatkovni modeli čimbolj splošni, razširljivi, razumljivi in enostavni za uporabo, kar pa je morda nemogoče doseči.

V nadaljevanju predstavimo podatkovne modele predhodno pregledanih ogrodij in standardiziran model NIF ter na podlagi ugotovitev zasnujemo podatkovni model ANGLEr. Poleg omenjenih obstaja še nekaj standardov, kot sta na primer CDA+GrAF (Meystre idr., 2012) ali

International Standard for a Linguistic Annotation Framework (Ide in Romary, 2004).[13] Prvi združuje dva standarda (ONJ in analiza grafov), drugi pa predstavlja mednarodno sprejet standard ISO. Ker nista širše uporabljena v praksi, jih ne omenjamo posebej.

3.1 Podatkovni model NLP Interchange Format (NIF)

Podatkovni model NLP Interchange Format (NIF; Hellmann idr., 2012) je nastal z namenom, da predstavi alternativo centraliziranim rešitvam (UIMA, GATE) in omogoči razvoj raznovrstnih, porazdeljenih in šibko sklopljenih aplikacij ONJ, ki sodelujejo med seboj. Verzija NIF 2.0 (Hellmann idr., 2013) definira semantično shemo v obliki ontologije RDF/OWL, s katero je možno opisovati posamezne označene dele besedila oz. prepoznane koncepte. Model se nanaša predvsem na fiksne predstavitve v besedilih, kot so naslov, odstavek, besedna zveza, stavek ali beseda (Slika 7).



Slika 7: Predstavitev podatkovnega modela (sheme ontologije) NIF 2.0.

13 https://www.cs.vassar.edu/~ide/papers/ISO+24612-2012.pdf

Jedro ontologije NIF predstavljajo viri *Strings as RDF*, katerih namen je, da označujejo izbrane dele besedila. Naslavljanje delov (tj. zaporedja črk) je implementirano v okviru definicije naslova URI posameznega objekta. Shemo je uporabljalo nekaj projektov na področju ekstrakcije informacij (npr. prepoznavanje imenskih entitet), standardizirana in vključena je bila v okviru projektov W3C ter uporabljena v nekaj delih izzven skupnosti (Sherif, 2014).

Podatkovni model se očitno ne razvija in se je nekako ustavil pri podpori za osnovna orodja.[14] To je možno zaradi več razlogov: (a) skupnost ni širše sprejela modela, (b) model je presplošen in ne omogoča širšega nabora predstavitve podatkov ONJ brez lastnih razširitev ali (c) raziskovalci ne poznajo semantičnih predstavitev RDF. Zagotovo bi moralo biti podprtih več predstavitev ONJ, vendar je verjetno na manjšo sprejetost najbolj vplivala »zima« semantičnega spleta. Raziskovalno področje semantičnega spleta namreč po 2013 ni več beležilo velikega zanimanja, kar se v zadnjih letih spreminja. S pojavom WikiData, smernic Evropske komisije glede podatkovnih prostorov in interoperabilnosti so opisi podatkov s pomočjo ontologij zelo uporabni za takšne naloge. Trenutno (v letu 2023) se prijavlja tudi akcija COST, ki bi nadaljevala razvoj semantičnih standardov za ONJ. Čeprav so funkcionalnosti semantičnega spleta zagotovo dodana vrednosti oznakam ONJ, lahko poleg oznak predstavljajo nepotrebno režijo. Zaradi tega pri podatkovnem modelu ANGLEr predlagamo predstavitev, ki se lahko enostavno poveže s semantično shemo.

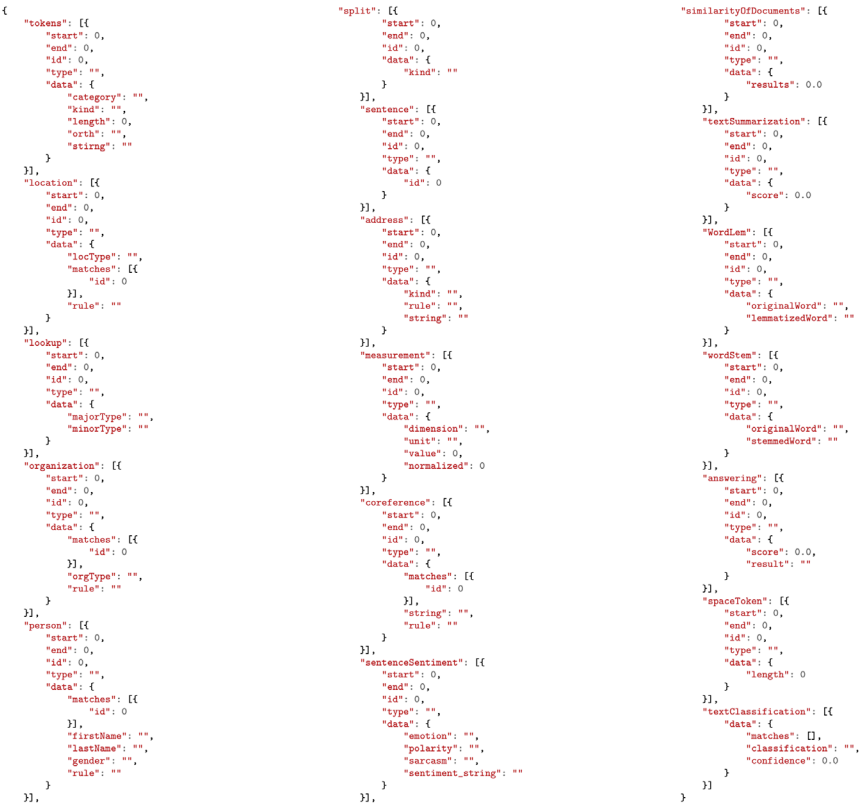
3.2 Podatkovni model GATE

Orodje GATE implementira podatkovni model v razredih programskega jezika Java, pri čemer je najvišje v hierarhiji tip razred Korpus (*Corpora*[15]). Primeri množice dokumentov, ki lahko predstavljajo korpus, so lahko tipov *TikaFormat*, *JsonDocuments*, *EmailDocuments*, *UIMADocument* ali *XMLDocument*. Vsak tip dokumenta mora omogočati tudi podporo za serializacijo v in iz zapisa GATE XML.

14 https://github.com/NLP2RDF
15 https://jenkins.gate.ac.uk/job/gate-core/javadoc/index.html

Predstavitve dokumentov so lahko pretvorjene v poljuben interni format, ki ga uporabljajo algoritmi v ogrodju GATE in se delijo na (a) vsebino, (b) oznake in (c) značilke. Vsaka oznaka, ki lahko vsebuje poljubno število značilk, je definirana z začetno točko, končno točko in tipom. Točka (*Node*) je definirana z unikatnim id-jem in odmikom v besedilu.

Poleg preddefiniranih predstavitev lahko vsakdo izdela svojo shemo[16] ali uporabi obstoječo. Slika 8 prikazuje preddefinirane tipe oznak. Menimo, da se veliko lastnosti med oznakami ponavlja in bi lahko bile organizirane bolj po skupinah/tipih in/ali hierarhično. Vsaka od oznak vsebuje tudi dodatne specifične atribute v obliki



```
{
    "tokens": [{
        "start": 0,
        "end": 0,
        "id": 0,
        "type": "",
        "data": {
            "category": "",
            "kind": "",
            "length": 0,
            "orth": "",
            "stirng": ""
        }
    }],
    "location": [{
        "start": 0,
        "end": 0,
        "id": 0,
        "type": "",
        "data": {
            "locType": "",
            "matches": [{
                "id": 0
            }],
            "rule": ""
        }
    }],
    "lookup": [{
        "start": 0,
        "end": 0,
        "id": 0,
        "type": "",
        "data": {
            "majorType": "",
            "minorType": ""
        }
    }],
    "organization": [{
        "start": 0,
        "end": 0,
        "id": 0,
        "type": "",
        "data": {
            "matches": [{
                "id": 0
            }],
            "orgType": "",
            "rule": ""
        }
    }],
    "person": [{
        "start": 0,
        "end": 0,
        "id": 0,
        "type": "",
        "data": {
            "matches": [{
                "id": 0
            }],
            "firstName": "",
            "lastName": "",
            "gender": "",
            "rule": ""
        }
    }],
    "split": [{
        "start": 0,
        "end": 0,
        "id": 0,
        "data": {
            "kind": ""
        }
    }],
    "sentence": [{
        "start": 0,
        "end": 0,
        "id": 0,
        "type": "",
        "data": {
            "id": 0
        }
    }],
    "address": [{
        "start": 0,
        "end": 0,
        "id": 0,
        "type": "",
        "data": {
            "kind": "",
            "rule": "",
            "string": ""
        }
    }],
    "measurement": [{
        "start": 0,
        "end": 0,
        "id": 0,
        "type": "",
        "data": {
            "dimension": "",
            "unit": "",
            "value": 0,
            "normalized": 0
        }
    }],
    "coreference": [{
        "start": 0,
        "end": 0,
        "id": 0,
        "type": "",
        "data": {
            "matches": [{
                "id": 0
            }],
            "string": "",
            "rule": ""
        }
    }],
    "sentenceSentiment": [{
        "start": 0,
        "end": 0,
        "id": 0,
        "type": "",
        "data": {
            "emotion": "",
            "polarity": "",
            "sarcasm": "",
            "sentiment_string": ""
        }
    }],
    "similarityOfDocuments": [{
        "start": 0,
        "end": 0,
        "id": 0,
        "type": "",
        "data": {
            "results": 0.0
        }
    }],
    "textSummarization": [{
        "start": 0,
        "end": 0,
        "id": 0,
        "type": "",
        "data": {
            "score": 0.0
        }
    }],
    "WordLem": [{
        "start": 0,
        "end": 0,
        "id": 0,
        "type": "",
        "data": {
            "originalWord": "",
            "lemmatizedWord": ""
        }
    }],
    "wordStem": [{
        "start": 0,
        "end": 0,
        "id": 0,
        "type": "",
        "data": {
            "originalWord": "",
            "stemmedWord": ""
        }
    }],
    "answering": [{
        "start": 0,
        "end": 0,
        "id": 0,
        "type": "",
        "data": {
            "score": 0.0,
            "result": ""
        }
    }],
    "spaceToken": [{
        "start": 0,
        "end": 0,
        "id": 0,
        "type": "",
        "data": {
            "length": 0
        }
    }],
    "textClassification": [{
        "data": {
            "matches": [],
            "classification": "",
            "confidence": 0.0
        }
    }]
}
```


Slika 8: Predstavitev podatkovnega modela GATE v formatu JSON. Model je iz zapisa v razredih Java izluščil Nik Hrovat (2022).

16 https://gate.ac.uk/sale/tao/splitch5.html#x8-840005.3

slovarja (*data*), pri čemer so predvidene specifične oznake za vsak tip posebej, dodatne pa se lahko dodaja poljubno.

3.3 Podatkovni model UIMA

UIMA uporablja standardizirano predstavitev, ki jo je pripravila standardizacijska skupina OASIS. Zadnja verzija standarda je bila objavljena v letu 2008, pri čemer je UIMA 1.0 dosegel skladnost s standardom v letu 2009. V primerjavi z GATE je opis predstavitev precej splošen in zelo kompleksen za nekoga, ki npr. razvija orodje ONJ in bi želel, da je njegov algoritem kompatibilen s podatkovnim modelom. Poleg tega specifikacija definira tudi komunikacijski protokol in procese za sistem ONJ.

UIMA zelo specifično definira vse vidike ONJ (v formatu XML), kar je seveda zelo primerno za veliko ogrodje. Kljub temu so razvite komponente precej stare, pri čemer se nekateri projekti trudijo približati


```
{
    "tokens": [{
        "id": 0,
        "begin": 0,
        "end": 0,
        "data": {
            "string": ""
        }
    }],
    "name": [{
        "id": 0,
        "begin": 0,
        "end": 0,
        "data": {
            "string": ""
        }
    }],
    "personTitle": [{
        "id": 0,
        "begin": 0,
        "end": 0,
        "data": {
            "kind": ""
        }
    }],
    "government": [{
        "id": 0,
        "begin": 0,
        "end": 0,
        "data": {
            "string": ""
        }
    }],
    "email": [{
        "id": 0,
        "begin": 0,
        "end": 0,
        "data": {
            "string": ""
        }
    }],
    "sentence": [{
        "id": 0,
        "begin": 0,
        "end": 0,
        "data": {
            "string": ""
        }
    }],
    "location": [{
        "id": 0,
        "begin": 0,
        "end": 0,
        "data": {
            "kind": ""
        }
    }],
    "link": [{
        "id": 0,
        "begin": 0,
        "end": 0,
        "data": {
            "type": "",
            "string": ""
        }
    }],
    "coreference": [{
        "id": 0,
        "begin": 0,
        "end": 0,
        "data": {
            "matches": [{
                "id": 0
            }],
            "score": 0.0
        }
    }],
    "documentsSimilarity": [{
        "id": 0,
        "begin": 0,
        "end": 0,
        "data": {
            "documentId": 0,
            "result": 0.0
        }
    }],
    "summarization": [{
        "id": 0,
        "begin": 0,
        "end": 0,
        "data": {
            "sentences": []
        }
    }],
    "stemmer": [{
        "id": 0,
        "begin": 0,
        "end": 0,
        "end": 0,
        "data": {
            "original": "",
            "stemmed_word": ""
        }
    }],
    "sentenceSentiment": [{
        "id": 0,
        "begin": 0,
        "end": 0,
        "data": {
            "emotion_name": "",
            "sentiment_string": ""
        }
    }],
    "classification": [{
        "id": 0,
        "begin": 0,
        "end": 0,
        "data": {
            "topic": "",
            "confidance": 0.0
        }
    }],
    "lemma": [{
        "id": 0,
        "begin": 0,
        "end": 0,
        "data": {
            "original": "",
            "lemmatized_word": ""
        }
    }],
    "answer": [{
        "id": 0,
        "begin": 0,
        "end": 0,
        "data": {
            "result": 0.0,
            "sentence": ""
        }
    }]
}
```


Slika 9: Predstavitev podatkovnega modela UIMA v formatu JSON. Model je iz implementacije izluščil Nik Hrovat (2022).

UIMA bližje ciljnim uporabnikom (npr. uimaFIT[17]). Komponente so razvite ločeno in vsaka zahteva specifičen postopek namestitve. Verjetno so tudi zaradi tega (in podpore novejšim modelom) trenutno bolj popularne platforme, kot so HuggingFace, LangChain in podobne.

3.4 Podatkovni model Orange

Osnovni tip predstavitve podatkov v ogrodju Orange so datoteke tipa *.tab*, kar predstavlja tabelaričen format za standardne naloge strojnega učenja. Za namene ONJ vtičnik za delo z besedili uporablja enak format, kjer vsaka vrstica predstavlja dokument obdelave z različnimi lastnostmi, kot so besedilo, razred in podobno. Uporabnik lahko sicer v sistem uvozi tudi podatke iz preddefiniranih formatov, kot so *.txt*, *.docx*, *.odt*, *.pdf*, *.xml* ali *.conllu*. Korpus besedil pa se lahko interaktivno ustvari tudi v orodju. Enaka funkcionalnost je na voljo tudi v Textable. Poleg osnovnih vtičnikov ta omogoča tudi uvoz podatkov iz nekaterih javnih virov – The Guardian, NY Times, Pubmed, Twitter in Wikipedija.

Rezultati vtičnikov so vrnjeni kot značilke v dodatnih stolpcih. Na primer, vtičnik za vektorske vložitve doda stolpec [*embedding-length*] za podatke o vsakem dokumentu. Nekateri vtičniki pa znotraj posameznega stolpca definirajo lastne predstavitve. Na primer, vtičnik *vreča besed* ustvari stolpec, ki vključuje besede in njihove frekvence v sledeči obliki: »beseda1=3, beseda2=123, beseda3=66, ...« Za transformacijo podatkov se lahko izdela nov vtičnik, ki se ga lahko nato vključi v cevovod obdelave, kar omogoči uporabo ogromnega števila že razvitih vtičnikov na področju strojnega učenja.

Za izdelavo popolnoma drugačne predstavitve pa lahko razvijalci razširijo Python programski razred *Orange.data.Storage* in definirajo lastno predstavitev podatkov.

3.5 Podatkovni model KNIME

Podobno kot orodje Orange (Razdelek 3.4) tudi orodje KNIME definira tabelarično strukturo kot osnoven tip predstavitve

17 https://github.com/apache/uima-uimafit

podatkov.[18] Uporablja se *BufferedDataTable,* ki je razred v programskem jeziku Java. Vsak vtičnik lahko sprejme več takšnih tabel, ki niso nujno vse shranjene v pomnilniku, kar omogoči obdelavo večjih količin podatkov. Ogrodje ponuja nekaj tipov predstavitev podatkov (razred *DataCell*), vendar lahko vsakdo predstavitve razširi za namene lastne uporabe.

3.6 Podatkovni model Stanza

Skupina Stanford NLP Group ima dolgo tradicijo ponujanja orodij ONJ. Do nedavnega so večino svojih algoritmov ponujali preko ogrodja Stanford Core NLP (Manning idr., 2014), ki je kot osnovno predstavitev uporabljal hierarhično urejeno predstavitev v obliki slovarjev (*HashMap* v programskem jeziku Java). S pojavom širše uporabljanih knjižnic v programskem jeziku Python so svoje ogrodje napisali na novo – Stanza in predstavili enostavnejši podatkovni model.

Podatkovni model Stanza je zelo jasno predstavljen v njihovih navodilih za uporabo.[19] Podobno kot pri Core NLP je tudi tu model osnovan na slovarjih, ki lastnosti predstavljajo kot ključ-vrednost (Slika 10). Model se lahko enostavno razširja, vendar je jasno definiran za osnovne potrebe ekstrakcije informacij, kot je prepoznava imenskih entitet ali lematizacija.

```
// Document
{
    "text": "The raw text.",
    "sentences": [{Sentence1}, {Sentence2}, ...],
    "entities": [{Span1}, {Span2}, ...],
    "num_tokens": 51,
    "num_words": 23
}

//Sentence
{
    "doc": Document, //back-pointer
    "text":    "The raw text",
    "dependencies": [
        {"head w.": "go", "rel.": "subj", "dep. w.": "home"},
        ...
    ],
    "tokens": [{Token1}, {Token2}, ...],
    "words": [{Word1}, {Word2}, ...],
    "entities": [{Span1}, {Span2}, ...],
    "sentiment": "positive",
    "constituency":    {ParseTree}
}
```

```
//Token
{
    "id": {"start-idx": 1, "end-idx": 4},
    "text":    "The",
    "misc": "Custom annotation value",
    "words": [{Word1}, {Word2}, ...],
    "start_char": 24,
    "end_char": 43,
    "ner": "B-ORG"
}

//Word
{
    "id": 3,
    "text": "The",
    "lemma": "the",
    "upos": "NOUN",
    "xpos": "NNP",
    "feats": "Gender=Fem|Person=3",
    "head": 2, //syntactic head word
    "deprel": "nmod", //relation to head word
    "deps": "head-word + deprel",
    "misc": "custom value",
    "parent": {Token}
}
```

```
//Span
{
    "doc": {Document},
    "text": "The Batman",
    "tokens": [{Token1}, {Token2}, ...],
    "words": [{Word1}, {Word2}, ...],
    "type": "PERSON",
    "start_char": 21,
    "end_char": 43
}

//ParseTree
{
    "label": "Noun",
    "children": [{ParseTree1}, {ParseTree2}, {ParseTree3}, ...]
}
```

Slika 10: Predstavitev podatkovnega modela Stanza v formatu JSON.

18 https://docs.knime.com/latest/analytics_platform_new_node_quickstart_guide/index.html#_number_formatter_node_implementation

19 https://stanfordnlp.github.io/stanza/data_objects.html

3.7 Predlog podatkovnega modela ANGLEr

Glede podatkovnih modelov, ki smo jih predstavili v predhodnih razdelkih, menimo, da nekateri predstavijo zanimive rešitve, vendar niso splošni, razširljivi ali enostavni za uporabo. Na primer, GATE definira nekaj osnovnih tipov za osnovno ONJ, za naprednejše analize pa bi bilo v bistvu potrebno model narediti praktično od začetka v programskem jeziku Java. UIMA nasprotno sicer definira zelo splošno in kompleksno predstavitev na podlagi standardizacije, ki pa jo zaradi tega uporablja manj uporabnikov. Orange in KNIME pričakujeta uporabo tabelaričnega formata, ki je prisiljen način predstavitve podatkov zaradi zagotavljanja čim večje skladnosti z zgodovinskim razvojem vtičnikov. Stanza sicer predstavi jasen in razumljiv podatkovni model, vendar vključuje predstavitve le za osnovne metode ONJ.

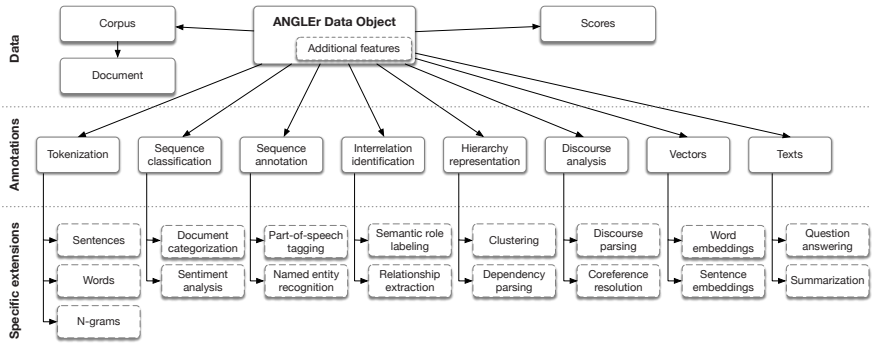
Cilj našega predloga je podpreti čim večje število orodij za analizo jezika. Poleg tega želimo pripraviti uporabniški vmesnik tako, da ga lahko razumejo tudi uporabniki, ki se ne spoznajo na princip delovanja programa. Uporabniški vmesnik mora biti sicer dovolj zmogljiv, da tudi naprednejšim uporabnikom omogoča enostavno in hitro prototipiranje. Tako lahko uporabniki hitro preizkusijo več različnih metod za analizo jezika in njihovih parametrov in takoj vidijo, kako spremembe vplivajo na pridobljene rezultate. Na podlagi pregleda obstoječih orodij smo ugotovili, da sta najbolj ključna elementa ogrodja za splošno ONJ podatkovni model in uporabniški vmesnik. Dober podatkovni model nam mora omogočati, da podpira predstavitve čim več metod ONJ, ki jih lahko zaradi tega enostavneje vključujemo v cevovode obdelav.

Menimo, da je pri snovanju podatkovnega modela potrebno definirati minimalne nabore predstavitev za posamezne naloge ONJ, ki jih sicer lahko specifične metode razširijo. Hierarhična struktura modela pa prispeva k lažji preglednosti, razumevanju in razširjanju modela. Glede na nam trenutno znane metode ONJ in uporabo v praksi, menimo, da bi zgornji nivo podatkovnega modela moral podpirati naslednje kategorije predstavitev podatkov:

- **Korpus**: Kot v večini modelov bi korpus moral vsebovati dokumente, ki so nižje-nivojska predstavitev podatkov različnih tipov, dolžin in metapodatkov. Vsak dokument pa bi vseboval določeno besedilo.
- **Razčlenjevanje**: Besedilni podatki so pogosto predstavljeni kot stavki, besede, besedne zveze, n-grami, ... Takšen tip bi moral podpirati vse različne členitve dokumentov.
- **Uvrščanje zaporedij**: Menimo, da je zaporedje splošen termin, ki predstavlja seznam členov, besed, stavkov, ... Primeri nalog, ki uporabljajo takšne predstavitve, so na primer analiza sentimenta ali kategorizacija besedil. Pri takšnih nalogah pa je ključno vedeti vsaj oznako uvrščanja.
- **Označevanje zaporedij**: Ta tip bi moral omogočati množico oznak za vsak del izbrane razčlenitve dokumentov. Primeri nalog so na primer oblikoslovno označevanje ali prepoznavanje imenskih entitet.
- **Prepoznavanje povezav**: Takšen tip bi moral omogočati hranjenje različnih tipov povezav med dvema ali več objekti. Uporabljal bi se na primer za podobnost med dokumenti, označevanju udeleženskih vlog ali prepoznavanju semantičnih relacij (»Janez« → »živi v« → »Boston«).
- **Analiza diskurza**: Tip bi moral omogočati skupno naslavljanje različnih predstavitev, kot je na primer zaporedje omenitev (tj. specifičen tip razčlenjevanja) za odkrivanje koreferenčnosti. Vsako zaporedje bi definiralo specifično oznako, ki bi veljala za izbrane objekte.
- **Hierarhična predstavitev**: Poleg osnovne predstavitve nekatera orodja ustvarjajo lastne hierarhične predstavitve znotraj izbranega besedila. Primer je na primer (plitko) označevanje drevesnic (angl. *dependency parsing*) ali gručenje podatkov.
- **Besedila**: Osnovna predstavitev podatkov, ki bi bila namenjena surovi predstavitvi ali generiranim rezultatom analiz. Dokument lahko sestoji iz različnih tipov besedil. Pri odgovarjanju na vprašanja so to na primer besedilo, vprašanje ali odgovor; pri prevodih besedila v različnih jezikih ali pri povzemanju različni povzetki.

- **Vektorji**: V času uporabe vektorskih predstavitev je potrebno zagotoviti, da so lahko različne razčlenitve besedil (tj. dokumenti, stavki, besede) predstavljeni kot vektorji.
- **Ocene**: Rezultati vrednotenj specifičnih algoritmov za namene hranjenja rezultatov analiz, vključno z metapodatki. Oznake, generirana besedila ali drugi rezultati v predstavitvi podatkov bi se referencirali na določen objekt ocen. Tako bi se omogočilo, da podatkovni model hrani osnovni korpus z oznakami in oznake več različnih algoritmov za iste naloge (npr. tri označevanja imenskih entitet treh različnih algoritmov ali njihovih različnih konfiguracij).
- **Dodatne lastnosti**: Vsak podatkovni tip bi moral poleg zahtevanih lastnosti omogočati še možnost hranjenja dodatnih lastnosti (ključ/vrednost). Takšne metapodatke bi lahko uporabljali algoritmi (npr. čas označevanja, ime algoritma). Poleg tega bi se po širši sprejetosti dodatnih lastnosti lahko le-te vključevalo kot obvezne za izpeljanke določenih tipov v naslednjih verzijah podatkovnega modela.

Na podlagi zgornjih definicij na Sliki 11 predstavljamo splošen podatkovni model ANGLEr. Krovni objekt vsebuje (a) korpus, (b) ocene in (c) visokonivojske objekte označevanj. Slednji objekti so hierarhični na način, da njihove izvedenke podedujejo obvezne lastnosti in dodatno uvajajo svoje. To omogoča interoperabilnost med različnimi



Slika 11: Hierarhija tipov podatkovnega modela ANGLEr.

metodami ONJ in različnimi verzijami podatkovnega modela. Kljub temu pa lahko katerokoli orodje obstoječim tipom dodaja še lastne metapodatke, ki so morda specifični le zanj (glej Dodatne lastnosti v zgornjem seznamu).

Slika 12 prikazuje praktičen primer strukture podatkovnega modela v formatu JSON. Različni deli so med seboj naslovljeni preko id-jev komponent. Nekatere komponente v podatkovnem modelu se lahko uporabijo neposredno, kot so definirane, ostale pa se lahko specializira glede na izbran algoritem. Na primer, predlogo za analizo diskurza se lahko razširi za namene odkrivanja koreferenčnosti ali analizo markerjev diskurza.

```
//ANGLEr Data Object
{
  "corpora": [{Corpus1}, {Corpus2}, ...],
  "scores": [{Scores1}, {Scores2}, ...],
  "annotations": [{Annotations1}, {Annotations2}, ...],
  "features": {"version": "1.0"} //Custom attributes
}

//Corpus
{
  "id": "1",
  "name": "Best NER-KB corpus",
  "documents": [{Document1}, {Document2}, ...],
  "features": {"url": "http://corpus-1.0.si"} //Custom attributes
}

//Document
{
  "id": "1-23", //Corpus id + Document id
  "text": "Best Story! Once upon a time, ...", //text used for analysis
  //Structured text parts that can be used by algorithms by key
  "text-parts": {
    "title": "Best Story!",
    "content": "Once upon a time, ...",
    "likes": 34
  }
  "features": {"length": 320, "separator": " "} //Custom attributes
}

//Scores
{
  "predicted_annotations_id": 42,
  "true_annotations_id": 31,
  "scores": {
    "F1": 0.45,
    "P": 0.88,
    "R": 0.91
  },
  "features": {"time": 2353242362} //Custom attributes
}
//Tokenization template
{
  "id": 33, //Used for algorithm input selection
  "annotation-type": "TOKENIZATION",
  "algorithm": "rule-based-1",
  "type": "SENTENCE", //WORDS, NGRAMS, WORD-PARTS, ...
  "documents": [
    {"doc_id": "1-23",
      "tokens": [{"id": 1,  "start_idx": 0, "end_idx": 342,
      "text": "Tralala, hopsasa."}, ...]
    }, ...
  ],
  "features": {"time": 2353242362} //Custom attributes
}
//Sequence classification template
{
  "id": 36, //Used for algorithm input selection
  "annotation-type": "SEQUENCE-CLASSIFICATION",
  "algorithm": "Best NN categorizer",
  "corpus_id": "1",
  "sequences": [
    {"id": 1, "class": "POSITIVE"}, ...
  ],
  "features": {"time": 2353242362} //Custom attributes
}
//Sequence annotation template
{
  "id": 39, //Used for algorithm input selection
  "annotation-type": "SEQUENCE-ANNOTATION",
  "algorithm": "CRF annotator",
  "tokenization_id": "1",
  "annotations": [
    {"doc_id": "1-23", "tags": ["O", "ORG", "O", ...]}, ...
  ],
  "features": {"annotation_time": "20sec"} //Custom attributes
}
```

```
//Interrelation identification template
{
  "id": 41, //Used for algorithm input selection
  "annotation-type": "INTERRELATION-IDENTIFICATION",
  "algorithm": "Intraverse relation extractor",
  "tokenization_id": "3",
  "relations": [
    {"relation": "employed_at", ...} //This object is specified on lower levels
  ],
  "features": {"annotation_time": "20sec"} //Custom attributes
}
//Hierarchy representation template
{
  "id": 43, //Used for algorithm input selection
  "annotation-type": "HIERARCHY-REPRESENTATION",
  "algorithm": "Agglomerative clusterer",
  "tokenization_id": null, //One of the following set only!
  "corpus_id": "1"
  "hierarchies": [
    {"parent_id": "34", "descendants": [{"id": "44", "relation": "nsubj"}, ...]}
    //This objects are specified on lower levels
  ],
  "features": {"max-hierarchy": 13} //Custom attributes
}
//Discourse analysis template
{
  "id": 45, //Used for algorithm input selection
  "annotation-type": "DISCOURSE-ANALYSIS",
  "algorithm": "Coref resolver SkipCor v2",
  "tokenization_id": 1, //These are sentences, mentions, ...
  "chains": [
    {"id1": "34", "id2": "45", "type": "anaphora"} //This object is specified on lower levels
  ],
  "features": {"singletons-num": 61} //Custom attributes
}
//Vectors template
{
  "id": 47, //Used for algorithm input selection
  "annotation-type": "VECTORS",
  "algorithm": "Doc2Vec",
  "tokenization_id": null, //One of the following set only!
  "corpus_id": "1"
  "vectors": [
    {"id": "78", "vec": [2.000, 3.241, 4.267, 5.987, ...]}
  ],
  "features": {"vector-dim": 350} //Custom attributes
}
//Texts template
{
  "id": 49, //Used for algorithm input selection
  "annotation-type": "TEXTS",
  "algorithm": "AnglerSummarizer",
  "corpus_id": "1",
  "doc-text-part-inputs": {
    "question": "question-part",
    "content": "text-part"
  }, //This object is specified on lower levels
  "texts": [
    {"doc-id": "78", "value": "No."}
  ],
  "features": null //Custom attributes
}
```

Slika 12: Predstavitev uporabe predloga podatkovnega modela ANGLEr v formatu JSON.

3.7.1 Verzioniranje podatkovnega modela

Hierarhična predstavitev podatkovnega modela izboljšuje obratno združljivost v primeru dodajanja novih tipov v model. V takem primeru so nove metode v primerjavi z obstoječimi združljive preko starševskih predstavitev podatkov. Prav tako se lahko zagotovi, da je nek objekt na nižjem nivoju lahko obdelan z algoritmom, ki deluje na višjenivojski predstavitvi.

Podatkovni model mora obstajati neodvisno od programskega paketa in biti na voljo v svojem odprtokodnem repozitoriju z jasno zgodovino sprememb. Tako se ga lahko uporablja v različnih programskih jezikih ali pa ga povzamejo tudi druga orodja. Verzioniranje mora slediti smernicam SemVer.[20]

4 Predlog arhitekture ogrodja ANGLEr

Med pregledom obstoječih ogrodij smo odkrili dobre in slabe prakse, o katerih smo lahko sklepali glede na sprejetost orodja v praksi. Namen arhitekture ANGLEr je zagotoviti razširljivo in skalabilno ogrodje, v katerega bi lahko napredni razvijalci (in raziskovalci) enostavno integrirali svoja orodja. To pomeni s čim manj branja dokumentacije o delovanju ostalih delov ogrodja. Netehnični uporabniki pa morajo imeti možnost enostavne namestitve in uporabe ogrodja (lahko tudi preko oblačne storitve). Poleg verzioniranega podatkovnega modela (Razdelek 3.7) menimo, da je pomembno zagotoviti (a) razširljivo arhitekturo, ki je šibko sklopljena preko programskih vmesnikov API, in (b) uporabniški vmesnik na podlagi vtičnikov.

Arhitektura sistema mora biti zasnovana tako, da bo mogoče podpreti vsa orodja, ki jih potrebujemo za analizo besedila. Poleg tega želimo arhitekturo narediti dovolj splošno, da bo kasneje mogoče dodati tudi nova orodja, ki si jih trenutno še nismo zamislili.

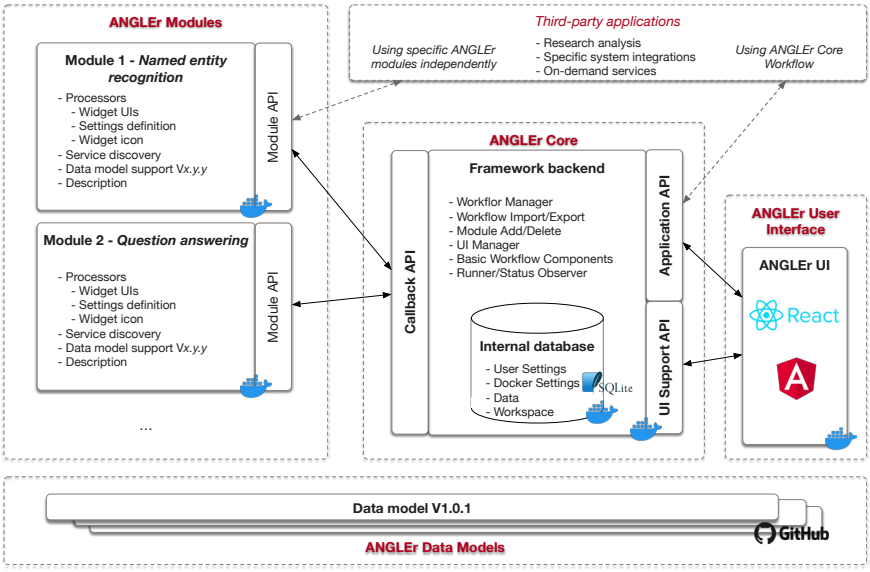
Orodje bo na osebnem računalniku (ali v oblaku) teklo kot storitev s spletnim vmesnikom.

Slika 13 prikazuje visokonivojsko predstavitev arhitekture ogrodja ANGLEr. Arhitektura sestoji iz naslednjih osnovnih gradnikov:

20 https://semver.org/

(a) moduli ANGLEr, (b) ANGLEr Core, (c) ANGLEr UI in (d) podatkovni modeli ANGLEr. S terminološkega vidika je pomembno razumeti predvsem koncepta ANGLEr modulov:

- ***Processor***: To je algoritem, ki sprejema in vrača podatke preko specifične verzije podatkovnega modela ANGLEr. V cevovodu/delotoku predstavlja en korak oz. komponento na grafičnem vmesniku.
- ***Module***: Modul predstavlja programski paket na osnovi Docker vsebnikov, ki vključuje enega ali več procesorjev (tj. implementacij specifičnih algoritmov ONJ). Vsak izmed njih pa mora omogočati integracijo z osnovnim gradnikom ANGLEr Core (tj. definicijo nastavitev, grafičnega vmesnika, ...) preko programskega vmesnika *Module API* (Razdelek 4.1). Modul je lahko vzpostavljen na ločeni/oddaljeni infrastrukuri in ga lahko neposredno prek programskega vmesnika poleg ogrodja ANGLEr uporabljajo tudi druge aplikacije (*Third-party applications*).



Slika 13: Predlog visoko-nivojske arhitekture ogrodja ANGLEr.

Podatkovni modeli ANGLEr (Razdelek 3.7) so na voljo v ločenem repozitoriju. Vsaka namestitev ogrodja ANGLEr uporablja specifičen podatkovni model, ki je obratno združljiv s starejšimi verzijami.

Osrednja arhitekturna komponenta ANGLEr Core zagotavlja glavne funkcionalnosti, ki orkestrirajo ostale komponente med seboj. Implementira upravljanje z delotoki, nameščanje/odstranjevanje modulov, komunikacijo z okoljem Docker, shranjevanjem podatkov v interno podatkovno bazo in programske vmesnike za komunikacijo. Programski vmesnik *Callback API* omogoča komunikacijo z vsemi »nameščenimi« oz. identificiranimi komponentami za ONJ. Vmesnik *Application API* omogoča programsko upravljanje s celotnim ogrodjem, kot to omogoča Orange preko knjižnice Python. Tako bo lahko nekdo npr. preko grafičnega vmesnika pripravil delotok, ga shranil, nato pa ga bo lahko integriral v nek drug sistem, ki bo delotok zaganjal preko programskega vmesnika. Vmesnik *UI Support API* pa implementira akcije, ki so vezane na nastavitve grafičnega vmesnika.

Komponenta *ANGLEr user interface* zagotavlja grafični vmesnik za delo z orodjem. Kot vse komponente je tudi ta šibko sklopljena z ostalimi deli ogrodja, kar omogoča, da bi lahko nekdo za delo z ogrodjem razvil svoj lasten vmesnik. Tako se lahko v prihodnosti za ogrodje razvije več programskih vmesnikov z različnimi funkcionalnostmi (npr. za različne interesne skupine uporabnikov), če bi bilo to potrebno. Komponenta nudi osnovne grafične elemente za *ANGLEr Core*, kot so dostop do nastavitev, menijev, delo z delotoki (tj. zaganjanje, ustavljanje, shranjevanje, odpiranje obstoječih, ...) in vključevanje grafičnih vmesnikov posameznih modulov. Specifični uporabniški vmesniki posameznih modulov (npr. ikone komponent/procesorjev, njihove nastavitve, vizualizacije) so implementirane neposredno v posameznih modulih in le prikazane v skupnem uporabniškem vmesniku.

Orodja bodo delovala kot samostojni programi, ki bodo izpostavili REST vmesnik, preko katerega bo glavni program komuniciral z njimi. Na ta način je implementacija vsakega orodja lahko povsem neodvisna od implementacije ostalih orodij in glavnega programa.

Orodja bodo praviloma pognana znotraj docker vsebnikov, s čimer bomo zagotovili delovanje orodij na vseh podprtih sistemih. Arhitektura omogoča tudi uporabo orodij, ki so pognana na oddaljenih računalnikih, saj vsa komunikacija poteka preko spletnih storitev. To omogoča, da orodja, ki za svoje delovanje zahtevajo grafično kartico (ali drugo zmogljivo opremo), tečejo na oddaljenem strežniku z ustrezno strojno opremo, uporabljamo pa jih na svojem lokalnem računalniku. Na takšen način se lahko razvije tudi tržnica storitev, kjer bi razvijalci/raziskovalci nudili svoje algoritme skupnosti.

4.1 Programski vmesnik Module API

Naloga vmesnika REST *Module API* je, da povezuje glavni program z vsakim izmed vsebnikov, ki poganjajo orodja. ANGLEr orodja uporabljajo enoten *ANGLEr Data Object*, ki je primerek podatkovnega modela ANGLEr, zato je preko vmesnika dovolj le definirati »naslovni prostor« podatkov za vhode in izhode algoritmov (tj. *inputs* in *outputs* na Sliki 14). Vmesnik mora podpirati naslednje:

- Pridobitev podatkov o modulu in orodjih, ki jih ponuja vsebnik.
- Spletno točko za delo z orodjem (tj. zagon, ustavitev).
- Spletno točko, na katerem je na voljo spletni vmesnik za konfiguracijo orodja.
- Pridobitev stanja orodja, kar pomeni seznam nalog, ki tečejo, in njihov napredek.
- Pridobivanje dnevniških zapisov.

Slika 14 prikazuje zahtevana polja pri komunikaciji z moduli ANGLEr ob pridobivanju podatkov o modulu in orodjih (tj. »*service discovery*«), za točki */about* in */processors.* Točka */docs* pa mora vračati HTML dokumentacijo o modulu. Ti podatki omogočijo delu ANGLEr Core, da na podlagi podanega osnovnega naslova modula samodejno prepozna funkcionalnosti in implementirane dele modula, ki jih lahko vključi v osnoven grafični vmesnik. Moduli so tako lahko pripravljeni v popolnoma drugih tehnologijah, kot jih uporablja osnovno ogrodje ANGLEr. Vizualizacije posameznega modula so prikazane

kot vsebniki na upravljalski strani ogrodja ANGLEr, kar pomeni, da je tudi tehnologija razvoja grafičnih elementov lahko popolnoma druga. Poleg tega lahko razvijalci uporabijo kako ogrodje za enostavno vizualizacijo svojih rezultatov, kot je na primer Gradio.[21]

```
// Endpoint: /about GET
{
  "UUID": "81a03d1a-0844-11ed-861d", //Identifier of the module.
  "name": "Best NER processors", //Name of the module.
  "version": "v1.2.9", //Module version.
  "data_model": "v1.0.1", //Version of the data model used.
  "desc": null, //(optional) Module description.
  "authors": "Jane Doe", //(optional) List of authors.
  "organisation": "University of Ljubljana", //(optional) Organisation of the authors.
  "url": "best-ner.github.io", //(optional) URL address of the page about the module.
  "docs_url": //(optional) A web page containing the documentation.
}

// Endpoint: /processors GET
[
  {
    "name": "CRF-based NER tagger" //Name of a processor.
    "short_name": "CRF-NER", //(optional) Short version of the name.
    "settings_endpoint": "/crf-ner/settings", //(optional) A module-relative address of a page containing processor settings.
    "ui_endpoint": "/ner-visualizations", //(optional) A module-relative address of a page for visualization.
    "icon": "/crf-ner/icon.ico", //A module-relative address of the icon to be used to represent the  processor.
    "category": "Semantic analysis", //A category of the processors menu that should  contain this processor (i.e., widget).
    "run_endpoint":  "/crf-ner/run", //Starting a processor. Data (i.e., current ANGLEr data object) is sent as payload.
    "status_endpoint":  "/crf-ner/status", //(optional) Returning statuses READY, PROCESSING (XX \%)
    "stop_endpoint":  "/crf-ner/stop", //Stopping/canceling current processing
    "inputs": ["TOKENIZATION", "WORDS"], //A list of ANGLEr data object types (from a hierarchy) that need to exist in the current data object for running this processor.
    "outputs": ["NER-ANNOTATION"], //A list of ANGLEr data object types (from a hierarchy, as high-level as possible) representing the processor's outputs.
    "docs_url": //(optional) A web page containing the documentation.
  },
  ... // Other available processors
]
```

Slika 14: Predlog definicije programskega vmesnika ANGLEr module API.

4.2 Arhitektura Docker

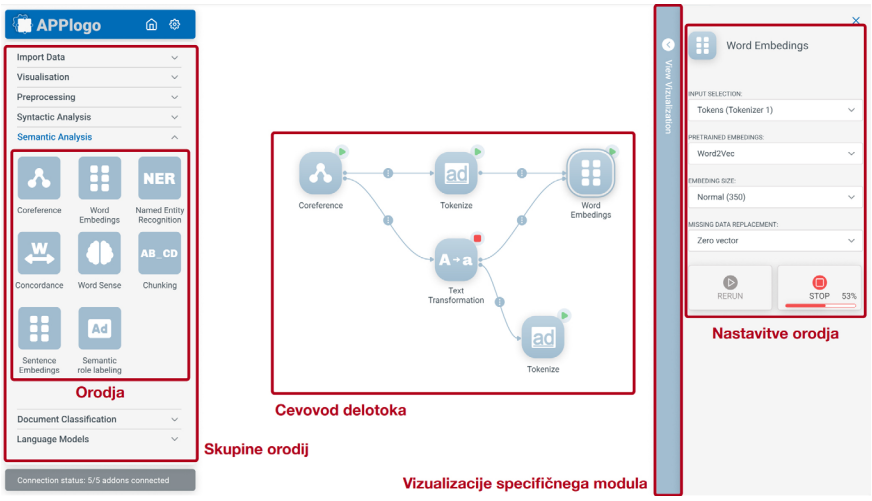
Uporaba arhitekture na osnovi tehnologije vsebnikov – Docker ni obvezna. Nekdo, ki bo razvijal svoje module, lahko na strežnikih storitve zaganja neposredno, brez vsebnikov, saj je potrebno za komunikacijo med gradniki zagotoviti le skladno implementacijo programskega vmesnika.

S stališča interoperabilnosti med sistemi, enostavne namestitve in zagona pa bodo vsi arhitekturni deli na voljo prek Docker slik. Docker slike lahko vsebujejo prednameščene vse odvisnosti, ni potrebe po posebni pripravi okolja za zagon in tudi razširjanje je enostavnejše, saj ostalim le delimo Docker sliko. V primerih uporabe modulov za več uporabnikov je možno Docker slike enostavno horizontalno in vertikalno razširjati. Osnoven sistem je načrtovan kot enouporabniški. Za ponujanje več uporabnikom pa lahko zanje tečejo različni primerki osnovnega ogrodja, ki pa vsi uporabljajo iste module.

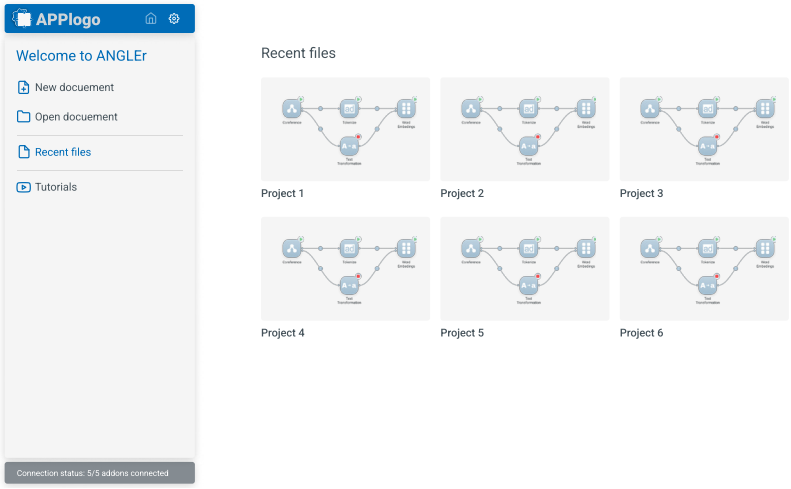
21 https://gradio.app

5 Predlog grafičnega vmesnika ANGLEr

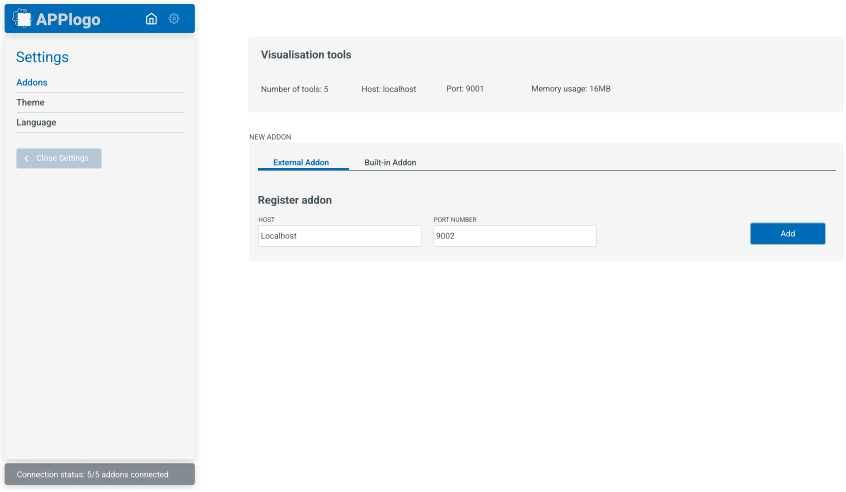
Grafični vmesnik mora omogočati poljubno povezovanje orodij s pomočjo kreiranja delotokov v obliki cevovodov procesiranja. Pomembno je definirati tudi, da se morajo komponente zaganjati zaporedno (razen ob vejitvah, kjer lahko analize tečejo vzporedno). Slika 15 prikazuje predlog splošnega grafičnega vmesnika orodja ANGLEr. Na levem delu je na voljo seznam vključenih orodij v sistem, ki jih lahko uporabnik povleče na delovno površino (tj. delotok) in povezuje med seboj v cevovod procesiranja. Na desni strani lahko upravlja z nastavitvami/parametri posameznih modulov in njihovimi vizualizacijami. Za boljšo ponazoritev možnosti je več primerov vmesnikov prikazanih na Slikah 16, 17, 18, 19 in 20.



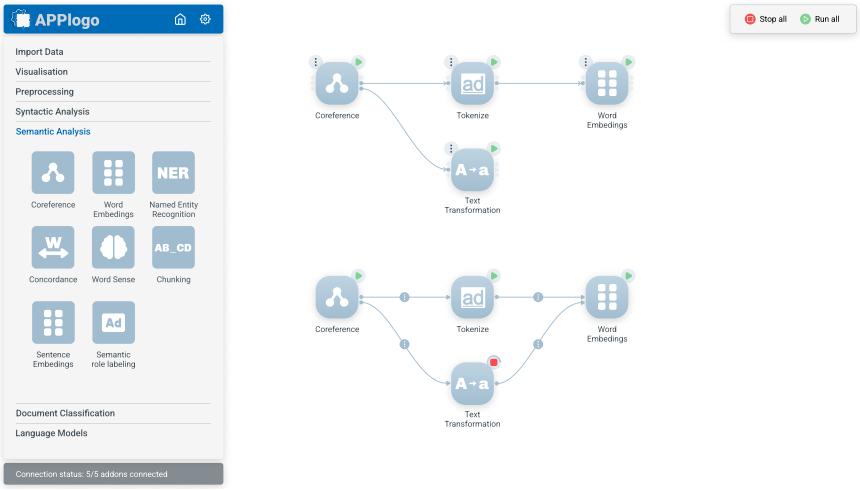
Slika 15: Predlog splošnega grafičnega vmesnika ogrodja ANGLEr.



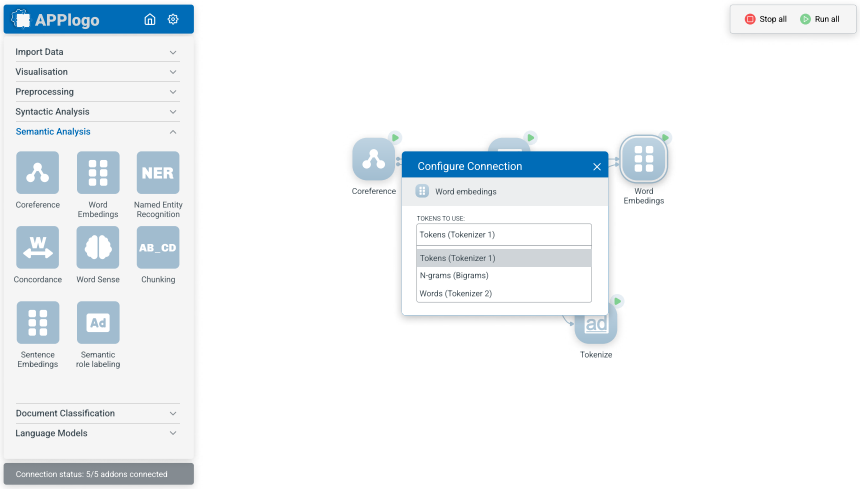
Slika 16: Predlog splošnega grafičnega vmesnika ogrodja ANGLEr. Začetni zaslon.



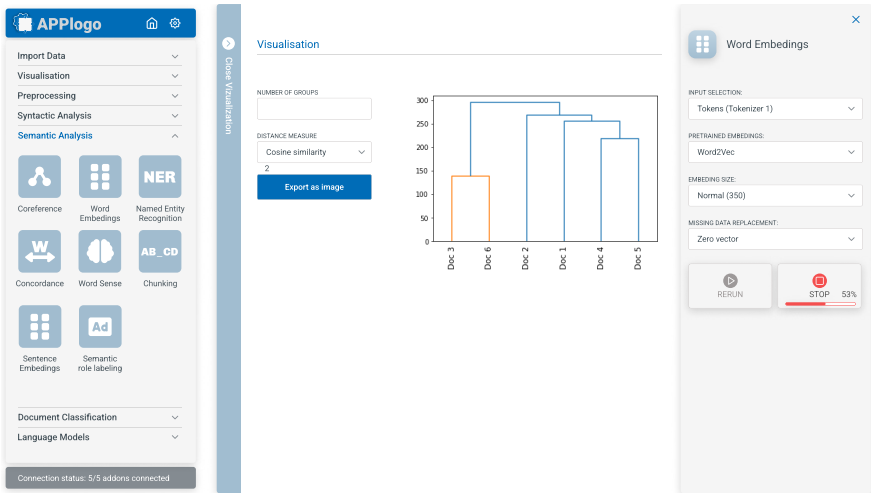
Slika 17: Predlog splošnega grafičnega vmesnika ogrodja ANGLEr. Glavne nastavitve za pregled in dodajanje modulov v sistem. Za dodajanje modula je potreben le URL naslov do njegovega programskega vmesnika REST.



Slika 18: Predlog splošnega grafičnega vmesnika ogrodja ANGLEr. Pregled delotoka, ki omogoča zaganjanje celotnega postopka, pregled posameznih procesorjev ali njihovo ločeno zaganjanje.



Slika 19: Predlog splošnega grafičnega vmesnika ogrodja ANGLEr. Povezovanje procesorjev med seboj in izbira vhodov glede na podatke v podatkovnem modelu.



Slika 20: Predlog splošnega grafičnega vmesnika ogrodja ANGLEr. Pregled vizualizacije izbranega procesorja.

6 Sklep

V tem prispevku smo pregledali obstoječa splošna ogrodja za obdelavo naravnega jezika in na podlagi tega pripravili načrt izgradnje novega splošnega in razširljivega podatkovnega modela ter ogrodja za analizo besedil – ANGLEr. Raziskovalni skupnosti in splošno zainteresirani javnosti bi takšno orodje za demokratizacijo obdelave naravnega jezika precej koristilo. Orodje bi omogočalo enostaven zagon najnovejših algoritmov s strani tehnično neveščih uporabnikov. Glede na pregledane uspešne primere bi bilo potrebno zagotoviti stalne posodobitve ogrodja in zagotavljanje primerov, novih modulov skupnosti, ki bi pri tem tudi sodelovala.

Menimo, da bi predlog izboljšal ponudbo na področju uporabe metod za obdelavo naravnega jezika. Ogrodje bi omogočilo tudi hitro prototipiranje in hiter dostop do prikaza uporabnosti metod za raziskovalce. V primerjavi z obstoječimi pristopi sta ključni izboljšavi, ki jih uvaja ANGLEr, (a) definicija enotnega, razširljivega, enostavnega in verzioniranega podatkovnega modela ter (b) vključevanje modulov le na podlagi jasno definiranega programskega

vmesnika API. Slednje omogoča, da so lahko nova orodja pripravljena v poljubni tehnologiji in da razvijalcem ni potrebno uporabljati predizbranih programskih jezikov, ogrodij ali poznati širše arhitekture ogrodja, kot je to potrebno pri najuspešnejših ogrodjih Orange in KNIME.

Ključna vprašanja, ki še dodatno poudarijo pomembnost izgradnje predlaganega ogrodja so naslednja:

- **Ali je potrebno predstaviti naloge obdelave naravnega jezika v celostnem podatkovnem modelu?** Obstaja več standardov za opisovanje podatkov in kar nekaj načinov hranjenja podatkov. Zaradi tega je pri razvijanju metod vedno potrebno prilagajati vhode in izhode, pri čemer ne obstajajo standardni postopki za transformacijo med predstavitvami. Veliko iniciativ za predstavitev podatkov je zamrlo, zato smo sledili uspešnim primerom in jih izboljšali, da je podatkovni model celosten in enostaven, kolikor je le možno.
- **Ali je potrebno (spet) razviti še eno ogrodje za obdelavo naravnega jezika?** Tehnologija se ves čas spreminja. Zaradi tega je šibko sklopljena arhitektura pravi odgovor na dolgoročno vzdrževanje sistema. Kljub temu pa je zelo pomembna tudi promocija ogrodja med zainteresiranimi deležniki. Trenutno se zdi spletni vmesnik in rešitev z vsebniki prava pot, vendar se lahko v prihodnosti spremeni tudi to. Srednja pot bi bila integracija orodij v obstoječe ogrodje (npr. Orange ali KNIME), vendar s tem postanemo zelo odvisni od ciljnega sistema, poleg tega pa je potrebno komponente ves čas prilagajati glede na razvoj osnovnega sistema.
- **Kdo so ciljni uporabniki in ali bi plačali za uporabo predlaganega sistema?** Trenutno veliko število raziskovalcev in jezikoslovcev uporablja orodja, ki jim za njihove namene omogočajo obdelavo besedil. Verjetno bi bil najboljši način slediti praksam GATE ali SketchEngine, pri čemer bi se začetni razvoj zgodil v okviru večjega raziskovalnega projekta. Celotno ogrodje bi seveda moralo ostati odprtokodno, tako da bi uporabniki lahko plačevali zgolj za storitev. Poleg tega pa bi lahko omogočili tržnico

orodij, saj bi podobno lahko raziskovalci na svojih strežnikih gostili svoja orodja in jih ponujali uporabnikom.

- **Kdo bi skrbel za dolgoročni razvoj takšnega ogrodja?** Kljub temu, da bi ogrodje uporabljala skupnost in razvijala odprtokodne module, bi moral nekdo skrbeti, posodabljati in usmerjati razvoj ogrodja. Glede na pregledana ogrodja je očitno, da »preživijo« le tista, za katera nekdo aktivno skrbi, organizira delavnice, izobraževanja (npr. Orange, KNIME, GATE). Možna »skrbnika« sta dveh tipov – raziskovalna skupina ali podjetje. Razvoj takšnega ogrodja lahko prevzame raziskovalna skupina, ki razvija tehnologije za obdelavo naravnega jezika, ima dovolj projektov, da financira tudi razvoj takšnega ogrodja, ki ga uporablja tudi za lastno promocijo. Podobno bi bilo v primeru podjetij, ki bi tržilo razvoj tehnologij za obdelavo naravnega jezika in bi ogrodje uporabljalo kot eno izmed svojih glavnih infrastrukturnih komponent. Glede na to, da bi bilo ogrodje odprtokodno, bi podjetje lahko vzporedno razvijalo tudi plačljive nadgradnje/module, ki bi bili tržno zanimivi.

Zahvale

Projekt Razvoj slovenščine v digitalnem okolju 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).

Literatura

Apache (2010). *Apache OpenNLP.* http://opennlp.apache.org

Bird, S. & Loper, E. (2006). *NLTK: the natural language toolkit*. In Proceedings of the COLING/ACL 2006 Interactive Presentation Sessions (pp. 69-72).

Bodenreider, O. (2004). *The unified medical language system (UMLS): integrating biomedical terminology*. Nucleic acids research, 32 (suppl_1), D267-D270.

Cunningham, H. (2002). *GATE, a general architecture for text engineering.* Computers and the Humanities, 36, 223-254.

Demšar, J., Curk, T., Erjavec, A., Gorup, Č., Hočevar, T., Milutinovič, M., ... & Zupan, B. (2013). *Orange: data mining toolbox in Python*. The Journal of machine Learning research, 14(1), 2349-2353.

Ferrucci, D., & Lally, A. (2004). *UIMA: an architectural approach to unstructured information processing in the corporate research environment*. Natural Language Engineering, 10(3-4), 327-348.

Hellmann, S., Lehmann, J., & Auer, S. (2012). *Linked-data aware uri schemes for referencing text fragments*. In Knowledge Engineering and Knowledge Management: 18th International Conference, EKAW 2012, Galway City, Ireland, October 8-12, 2012. Proceedings 18 (pp. 175-184). Springer Berlin Heidelberg.

Hellmann, S., Lehmann, J., Auer, S., & Brümmer, M. (2013). *Integrating NLP using linked data*. In The Semantic Web–ISWC 2013: 12th International Semantic Web Conference, Sydney, NSW, Australia, October 21-25, 2013, Proceedings, Part II 12 (pp. 98-113).

Hrovat, N. (2022). Zasnova ogrodja za izvajanje metod za procesiranje naravnega jezika. [Diplomsko delo, Univerza v Ljubljani] Repozitorij Univerze v Ljubljani. https://repozitorij.uni-lj.si/IzpisGradiva.php?lang=slv&id=141460

Ide, N., & Romary, L. (2004). *International standard for a linguistic annotation framework.* Natural language engineering, 10(3-4), 211-225.

Manning, C. D., Surdeanu, M., Bauer, J., Finkel, J. R., Bethard, S., & McClosky, D. (2014). *The Stanford CoreNLP natural language processing toolkit*. In Proceedings of 52nd annual meeting of the association for computational linguistics: system demonstrations (pp. 55-60).

Meystre, S. M., Lee, S., Jung, C. Y., & Chevrier, R. D. (2012). *Common data model for natural language processing based on two existing standard information models: CDA+GrAF.* Journal of biomedical informatics, 45(4), 703-710.

Nance, J., & Baumgartner, P. (2021). *gobbli: A uniform interface to deep learning for text in Python*. Journal of Open Source Software, 6(62), 2395.

Qi, P., Zhang, Y., Zhang, Y., Bolton, J., Manning, C.D. (2020). *Stanza: A Python Natural Language Processing Toolkit for Many Human Languages*. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations (pp. 101-108).

Sherif, M. A., Coelho, S., Usbeck, R., Hellmann, S., Lehmann, J., Brümmer, M., & Both, A. (2014). *NIF4OGGD-NLP Interchange Format for Open German Governmental Data*. In LREC (pp. 3524-3528).

Soysal, E., Wang, J., Jiang, M., Wu, Y., Pakhomov, S., Liu, H., & Xu, H. (2018). *CLAMP–a toolkit for efficiently building customized clinical natural language processing pipelines*. Journal of the American Medical Informatics Association, 25(3), 331-336.

Slovenski meta-povzemalnik

Aleš ŽAGAR
Univerza v Ljubljani, Fakulteta za računalništvo in informatiko

Marko ROBNIK-ŠIKONJA
Univerza v Ljubljani, Fakulteta za računalništvo in informatiko

Povzetek

Povzemanje besedil je pomembna naloga obdelave naravnega jezika, zato so raziskovalci v zadnjih letih razvili različne pristope, od sistemov, ki temeljijo na pravilih, do nevronskih mrež, ki v zadnjem času prevladujejo. Žal ne obstaja en sam model ali pristop, ki bi dobro deloval na vseh vrstah besedil, zato predlagamo meta-model, ki priporoča najprimernejši povzemalnik za določeno besedilo. Predlagani meta-model uporablja polno povezano nevronsko mrežo, ki analizira vhodno vsebino in napove, kateri povzemalnik bi bil za dano vhodno besedilo najboljši v smislu ocene ROUGE. Meta-model izbira med štirimi različnimi modeli povzemanja, razvitimi za slovenščino, pri čemer uporabi različne lastnosti vhodnega besedila, zlasti njegovo dokumentno predstavitev Doc2Vec. Štirje uporabljeni slovenski povzemalniki naslavljajo različne izzive, povezane s povzemanjem besedil v jeziku z manj viri. Delovanje predlaganega modela SloMetaSum ovrednotimo avtomatsko, njegove dele pa tudi ročno. Rezultati kažejo, da sistem uspešno avtomatizira korak ročne izbire najboljšega modela za dano besedilo.

Ključne besede: povzemanje besedil, jeziki z manj viri, meta-model, slovenski jezik

Abstract

Text summarization is an essential task in natural language processing, and researchers have developed various approaches over the years, ranging from rule-based systems to neural networks. However, there is no single

model or approach that performs well on every type of text. We propose a system that recommends the most suitable summarization model for a given text. The proposed system employs a fully connected neural network that analyzes the input content and predicts which summarizer should score the best in terms of ROUGE score for a given input. The meta-model selects among four different summarization models, developed for the Slovene language, using different properties of the input, in particular its Doc2Vec document representation. The four Slovene summarization models deal with different challenges associated with text summarization in a less-resourced language. We evaluate the proposed SloMetaSum model performance automatically and parts of it manually. The results show that the system successfully automates the step of manually selecting the best model.

Keywords: text summarization, low-resource languages, meta-model, Slovene language

1 Uvod

Povzemanje besedila izbere bistvene informacije v dokumentu ali zbirki dokumentov ter jih predstavi na kratek in koherenten način. Kljub dolgotrajnim prizadevanjem raziskovalcev na področju obdelave naravnega jezika (NLP) je povzemanje besedil še vedno zahtevna naloga. Z eksplozivno rastjo digitalnih informacij postaja povzemanje velikih količin besedil v krajšo in bolj obvladljivo obliko vse pomembnejše.

Obstajata dva glavna pristopa k povzemanju besedil: ekstraktivni in abstraktivni. Ekstraktivno povzemanje izbere podmnožico stavkov ali besednih zvez iz izvornega besedila, ki najbolje predstavljajo vsebino. Izbrani stavki se združijo v povzetek. V nasprotju s tem abstraktivno povzemanje ustvarja nove stavke, ki zajamejo pomen izvirnega besedila. Ekstraktivno povzemanje je preprostejše in hitrejše od abstraktivnega povzemanja, vendar lahko privede do povzetkov, ki vsebujejo odvečno in ponavljajočo se vsebino. Abstraktivno povzemanje je zahtevnejše in zahteva naprednejše tehnike

obdelave naravnega jezika, vendar lahko ustvari podobne povzetke kot človek.

Tehnologija povzemanja besedil je v zadnjih letih doživela velik razvoj z arhitekturo nevronskih mrež transformer in na njej osnovanih velikih vnaprej naučenih jezikovnih modelih, kot sta T5 (Raffel idr., 2020) in GPT-3 (Brown idr., 2020). Tako so nastali modeli za povzemanje, katerih povzetki so zelo podobni tistim, ki jih je napisal človek, z malo ponovitvami in netočnostmi, predvsem pri povzemanju novic. Ti modeli so sposobni obdelovati vedno daljše vsebine, kar omogoča izdelavo povzetkov za daljša besedila. Posledično so lahko najsodobnejši samodejni povzetki jasni in enostavni za razumevanje.

V morfološko bogati slovenščini je povzemanje besedil zaradi omejene razpoložljivosti virov in podatkov ter raziskav še večji izziv kot v angleščini. Naučili smo štiri slovenske modele za povzemanje besedil z različnimi lastnostmi na različnih učnih podatkih.[1] Naši štirje modeli zajemajo dva ekstraktivna povzemalnika (eden temelji na enostavnem izboru stavkov s pogostostjo besed, drugi pa na grafu), abstraktivni povzemalnik, ki temelji na modelu T5, in hibridni ekstraktivno-abstraktivni model. Na splošno se najbolje obnese slovenski transformerski model, ki temelji na modelu T5, vendar se ne nujno dobro posploši za vse vrste vhodnih besedil. Zato se ukvarjamo s problemom meta-povzemalnika, ki ugotavlja, kateri model povzemanja je najprimernejši za določeno besedilo glede na dolžino in žanr besedila.

Rezultat raziskave je nov slovenski sistem za povzemanje (imenovan SloMetaSum), ki ga sestavljajo ekstraktivni, abstraktivni in hibridni povzemalniki ter meta-model, ki med njimi izbira. Predlagani meta-sistem je sestavljen iz polno povezane nevronske mreže, ki analizira vhodno vsebino in priporoča najprimernejši model povzemanja za določeno besedilo. V ta namen SloMetaSum uporablja vektorsko predstavitev dokumentov Doc2Vec (Le in Mikolov, 2014) in napoveduje ocene ROUGE za vsakega od povzemalnikov. S kombiniranjem večih pristopov lahko sistem učinkovito ustvari kakovostne

1 V sklopu projekta RSDO: https://www.cjvt.si/rsdo/.

povzetke, ki so informativni in lahko razumljivi za več vrst besedil, ne glede na njihovo dolžino in žanr.[2]

Prispevki naše raziskave so naslednji:

- Razvili smo štiri modele za povzemanje, ki lahko učinkovito povzemajo besedila različnih dolžin in žanrov, zaradi česar so vsestransko uporabni.
- Uspešno smo naslovili izzive slovenskega jezika z malo viri in ustvarili visoko učinkovite modele za povzemanje slovenskih besedil.
- Ustvarili smo meta-model, ki na podlagi parametrov, kot so dolžina, zapletenost, raven abstrakcije in predvideni primer uporabe, priporoči najprimernejši model povzemanja za določeno besedilo.

Preostali del prispevka je razdeljen na šest razdelkov. V Razdelku 2 predstavimo sorodne raziskave. V Razdelku 3 so opisani nabori podatkov. V Razdelku 4 opisujemo osnovne povzemalnike in meta-model. V oddelku 5 predstavljamo empirično ovrednotenje in predstavimo ugotovitve. Razdelek 6 vsebuje zaključke in priporočila za nadaljnjo delo.

2 Sorodna dela

Zgodnji pristopi k povzemanju besedil so temeljili na statističnih frekvencah besed, položaju stavkov in stavkih, ki vsebujejo ključne besede (Nenkova in Vanderwende, 2005). Cilj teh pristopov je bil iz besedila izluščiti pomembne stavke ali besedne zveze in z njihovim povezovanjem ustvariti povzetek. Abstraktivne metode so vključevale brisanje manj pomembnih besed iz besedila za oblikovanje povzetka (Knight in Marcu, 2002).

Metode, ki temeljijo na grafih, so priljubljen pristop k povzemanju besedil. Pri tem pristopu je dokument predstavljen kot graf, kjer so povedi vozlišča, povezave pa predstavljajo odnose med njimi.

2 Demonstracijska predstavitev je na voljo na spletni strani https://slovenscina.eu/en/povzemanje. Repozitoriji kode so na voljo na naslovih https://github.com/azagsam/metamodel in https://github.com/clarinsi/SloSummarizer.

Graf se nato uporabi za izdelavo povzetka z izbiro najpomembnejših stavkov. Ta metoda je bila raziskana v več delih (Mihalcea in Tarau, 2004; Erkan in Radev, 2004).

S pojavom nevronskih mrež se je povečalo zanimanje za razvoj tehnik abstraktivnega povzemanja. Prvi nevronski abstraktivni sistemi so uporabljali arhitekturo rekurentnih nevronskih mrež, kot je LSTM (See idr., 2017; Nallapati idr., 2016). Danes najsodobnejši modeli za abstraktivno povzemanje uporabljajo arhitekturo transformer (Zhang idr., 2020; Lewis idr., 2020). Ta arhitektura intenzivno uporablja mehanizem samopozornosti za selektivno osredotočanje na pomembne dele besedila. Modeli s to arhitekturo lahko v primerjavi s prejšnjimi metodami ustvarijo bolj tekoče in koherentne povzetke.

Čeprav je bilo za povzemanje besedil predlaganih več pristopov, so mnogi omejeni na določene žanre besedil. V tem delu je naš cilj zgraditi sistem za povzemanje, ki lahko obravnava raznovrstna besedila različnih žanrov, ki nastopajo v realnosti. To vključuje besedila različnih dolžin, tematik in slogov, s ciljem izdelati povzetke, ki zajamejo najpomembnejše informacije v besedilu. Želimo razviti robusten in prilagodljiv model, ki se lahko nauči kakovostno povzemati besedila različnih vrst.

3 Učne množice

V tem razdelku opisujemo nabore podatkov, ki smo jih uporabili v naši raziskavi. V nadaljevanju podajamo kratek opis podatkovnih množic, v Tabeli 1 pa njihove statistične lastnosti.

Množica STA (splošni novinarski članki Slovenske tiskovne agencije) obsega 366.126 dokumentov. Kot približek povzetka smo uporabili prvi odstavek vsakega članka, saj ta množica ne vsebuje ročno napisanih človeških povzetkov. Naša izbira povzetka sledi pogosti praksi pri povzemanju besedil, zlasti v jezikih, ki nimajo namenskih učnih množic za povzemanje novic.

AutoSentiNews (Bučar, 2017) je podoben nabor besedil kot STA; sestavljen iz 256.567 člankov iz slovenskih novičarskih portalov

24ur, Dnevnik, Finance, RTVSlo in Žurnal24. Povzetki so izdelani iz prvega odstavka na enak način kot v podatkovni zbirki STA.

Učna množica SURS je manjša zbirka finančnih novic Statističnega urada Slovenije in obsega 4.073 dokumentov.

Korpus slovenskih akademskih besedil KAS (Žagar idr., 2022) je sestavljen iz diplomskih, magistrskih in doktorskih del, napisanih med letoma 2000 in 2018, ki so zbrana v digitalnih knjižnicah slovenskih visokošolskih zavodov ter dostopna na portalu odprte znanosti.[3] Korpus vsebuje človeške povzetke akademskih besedil.

Nabor podatkov CNN/Daily Mail (Hermann idr., 2015) je namenjen povzemanju besedil. Vsebuje s strani človeka ustvarjene izvlečke povzetkov iz novic na spletnih straneh CNN in Daily Mail. Korpus ima 286.817 učnih parov, 13.368 validacijskih parov in 11.487 testnih parov. Izvorni dokumenti imajo v povprečju 766 besed, povzetki pa 53 besed. Zbirko podatkov smo prevedli v slovenščino z uporabo strojnega prevajanja (Lebar Bajec idr., 2022).

Tabela 1: Korpusi in učne množice, uporabljene za učenje modela za predstavitev dokumentov Doc2vec in meta-modela.

Množica	Število dokumentov
STA	334.696
AutoSentiNews	256.567
SURS	4.073
CNN/Daily Mail	311.672
KAS	82.308
Skupno	**677.644**

4 Povzemalni modeli in meta-model

V tem razdelku opisujemo sestavne dele našega sistema SloMetaSum, ki ga sestavljajo štirje povzemalniki, sistem za predstavitev dokumentov in meta-model.

3 http://openscience.si/

4.1 Povzemalni modeli

Izdelali smo štiri povzemalnike, ki jih v nadaljevanju na kratko opišemo.

Osnovni povzemalnik (Nenkova in Vanderwende, 2005) za izbiro najbolj informativnih stavkov uporablja preprost pristop s frekvenco besed. Model **Grafi**, zasnovan na grafih (Žagar in Robnik-Šikonja, 2021), se zgleduje po algoritmu TextRank (Mihalcea in Tarau, 2004) in za razvrščanje stavkov uporablja ocene centralnosti stavkov. Oba modela spadata med ekstraktivne metode in se lahko uporabljata na dokumentih poljubne velikosti. V nasprotju z izvornim pristopom TextRank smo za numerično predstavitev stavkov uporabili kodirnik stavkov LaBSE (Feng idr., 2022), ki temelji na arhitekturi transformer. Model abstraktivnega povzemanja **T5-članki** uporablja vnaprej naučen slovenski model SloT5 (Ulčar in Robnik-Šikonja, 2023) in je prilagojen na strojno prevedenem naboru člankov CNN/Daily Mail (Hermann idr., 2015) z uporabo slovenskega sistema strojnega prevajanja (Lebar Bajec idr., 2022). Model **Hibrid-dolga** je kombinacija modela, ki temelji na grafu, in modela T5-članki. Najprej sestavi kratko besedilo z združevanjem najbolj informativnih stavkov (ekstraktivni korak). V naslednjem, abstraktivnem koraku pa se ti stavki povzamejo s povzemalnikom T5-članki.

4.2 Predstavitev dokumentov z modelom Doc2Vec

Za izbiro najprimernejše metode povzemanja za določeno besedilo mora meta-model pridobiti informacije o različnih lastnostih besedila. Za predstavitev dokumentov uporabimo model Doc2Vec in ga naučimo na slovenskih besedilih iz Tabele 1 (brez povzetkov). V koraku predobdelave smo odstranili visokofrekvenčne besede, ki ne prispevajo k pomenu dokumentov, kot so zaimki, vezniki itd.; da bi dodatno zmanjšali število različnih besed, smo celotno zbirko besedil lematizirali.

4.3 Meta-model

Naš meta-model je sestavljen iz polno povezane nevronske mreže, naučene za napovedovanje ocen ROUGE povzetkov. Za učno

množico smo naključno izbrali 93.419 primerov iz celotne zbirke surovih besedil. Najprej je vsak od naših štirih povzemalnikov pripravil povzetek za vse primere. Izračunali smo ocene ROUGE med referenčnimi in izdelanimi povzetki. ROUGE (Recall-Oriented Understudy for Gisting Evaluation) je metrika, ki se najpogosteje uporablja za ocenjevanje samodejno ustvarjenih povzetkov. Kakovost povzetka meri s številom prekrivajočih se enot (n-gramov, zaporedij besedil itd.) med povzetki, ki jih je ustvaril človek, in povzetki, ki so jih ustvarili sistemi za povzemanje. ROUGE ni ena sama metrika, ampak družina metrik. Najpogosteje se uporabljata ROUGE-N in ROUGE-L. Prva meri prekrivanje n-gramov (običajno unigramov in bigramov), druga pa meri najdaljšo skupno zaporedje besed v obeh povzetkih. Kot vhodni podatek za naš meta-model uporabljamo štiri ocene F1 ROUGE (ROUGE-1, ROUGE-2, ROUGE-L, ROUGE-LSum), ki kažejo, kako dobri so ustvarjeni povzetki. Podatke smo razdelili na učno, validacijsko in testno množico v razmerju 90 : 5 : 5.

Velikosti obeh podatkovnih množic sta predstavljeni v Tabeli 2. V Tabeli 3 so predstavljene povprečne vrednosti ROUGE naših povzetkov za dolga in kratka besedila. Povzemalniki, ki so specializirani za kratka besedila, dosegajo boljše rezultate na kratkih besedilih in obratno.

Tabela 2: Število učnih primerov za model predstavitve besedil in meta-povzemalnik.

Model	Velikost učne množice
Doc2Vec	677.644
Meta-model	93.419

Tabela 3: Ocene povzetkov, izražene z mero ROUGE, za dolga in kratka besedila. Najboljši rezultati za kratka in dolga besedila so v krepkem tisku.

	T5-članki	Osnovni	Grafi	Hibrid-dolga
Kratka	**14,01**	13,11	13,15	12,55
Dolga	10,51	13,12	**17,71**	17,59

5 Rezultati

V tem razdelku predstavimo načrt evalvacije modelov in rezultate. Rezultate predstavitvene metode Doc2Vec in povzemalnikov podamo v ločenih podrazdelkih.

5.1 Doc2Vec

Za učenje modela za predstavitev dokumentov Doc2Vec smo uporabili naslednje hiperparametre: največja dovoljena velikost slovarja je 100 000, velikost vektorja za predstavitev besed je 256, velikost okna konteksta je 5, najmanjša frekvenca besede, ki se vključi v slovar, je 1, skupno število epoh za učenje modela pa je 5.

Model Doc2Vec smo ocenili ročno in z avtomatskimi merami. Pri ročni analizi smo za vsakega od nekaj naključno izbranih vzorcev z uporabo kosinusne podobnosti pregledali 3 najbolj podobne vrnjene dokumente in opazovali, ali se teme dokumentov prekrivajo. Teme dokumentov so si bile v večini primerov podobne. Na podlagi tega smo sklepali, da model deluje v skladu s pričakovanji. Avtomatsko ocenjevanje je bilo del celotnega cevovoda, v katerem so bili hiperparametri modela prilagojeni za optimizacijo funkcije izgube meta-modela.

5.2 Meta-model

Naši končni rezultati so predstavljeni v Tabeli 4. Predlagani mehanizem za izbiro povzemalnih modelov smo primerjali s tremi osnovnimi mehanizmi. Mehanizem *Povprečje* vzame napovedi za vsak model povzemanja in jih povpreči. Vedno izbere model z najvišjim številom točk. Mehanizem *Drevo* uporablja regresijsko drevo; z uporabo iskanja po mreži hiperparametrov je najmanjše število vzorcev, potrebnih za razdelitev notranjega vozlišča, 100. Mehanizem *Gozd* uporablja naključni gozd; eksperimentirali smo s podobnimi vrednostmi kot pri mehanizmu Drevo in določili število dreves na 300.

Naš najboljši mehanizem za izbor povzemalnikov je nevronska mreža z dvema skritima nivojema. Skrita nivoja vsebujeta po 1024

nevronov, med postopkom učenja pa smo uporabili 10 % primerov iz učne množice za validacijo. Aktivacijska funkcija, uporabljena za ta model, je popravljena linearna enota (ReLU). Uporabili smo strategijo zgodnje ustavitve učenja s parametrom potrpežljivosti 2. Funkcija izgube, uporabljena za ta model, je povprečna kvadratna napaka.

Meta-model se je prenehal učiti po 7 epohah in na testni množici dosegel skoraj 15 točk nad srednjo osnovno vrednostjo. Opazili smo, da izbira različnih hiperparametrov ne vpliva bistveno na rezultate. Eksperimentirali smo z različnimi velikostmi skritih plasti, številom enot in aktivacijskimi funkcijami. Preizkusili smo tudi različne velikosti največjega besedišča in oken modela Doc2Vec. Navajamo samo vrednosti najboljšega modela.

Na splošno se je ta model izkazal za najučinkovitejšega med preizkušenimi mehanizmi za izbor najprimernejšega povzemalnika. Veliko število nevronov v skritem sloju je verjetno prispevalo k njegovi boljši učinkovitosti, saj omogoča večjo stopnjo kompleksnosti pri predstavitvi podatkov v modelu.

Tabela 4: Rezultati štirih mehanizmov za izbor povzemalnikov na testni množici. Meta-model-osnovni je pokazal znatno izboljšanje v primerjavi z metodami Povprečje in Drevo. Izrecno kodiranje dolžine besedil ali uravnoteženje podatkovne množice nista izboljšali rezultatov.

Model	Srednja kvadratna napaka
Povprečje	84,493
Drevo	81,631
Gozd	74,975
Meta-model-osnovni	**70,066**
Meta-model+dolžina	70,146
Meta-model+uravnoteženje	79,044

Nadalje smo preizkusili dve različici meta-modela. Meta-model+dolžina doda še en vhodni nevron, ki izrecno kodira vhodno dolžino. Ugotovili smo, da to ne izboljša modela; domnevamo, da so akademska besedila različnih žanrov, kar naša predstavitev dokumentov že dobro pokriva. Poskušali smo tudi uravnotežiti podatke, saj prvotni nabor podatkov vsebuje razmerje 1 : 5 med dolgimi in

kratkimi besedili, kar povečuje morebitno težavo pretiranega prilagajanja kratkim besedilom. Zmanjšali smo število kratkih besedil v učni množici, da smo dobili uravnoteženo učno množico s 16,932 povzetki za naš uravnoteženi meta-model. Rezultat tega uravnoteženja je slabši model, vendar še vedno boljši od osnovnega modela Povprečje.

Tabela 5 prikazuje frekvence, kolikokrat je meta-model priporočil vsak model od 1.000 vzorcev iz testne množice. Vidimo lahko, da je bil model T5-članki priporočen največkrat, in sicer 595-krat od 1.000 vzorcev. Model Hibrid-dolgi je bil priporočen 254-krat, sledi mu model Osnovni, ki je bil priporočen 80-krat. Model Graf je bil priporočen najmanjkrat, in sicer 71-krat od 1000 vzorcev.

Tabela 5: Pogostost priporočil meta-modela za vsakega od osnovnih povzemalnikov na vzorcu 1.000 povzetkov iz testne množice.

Model	Frekvenca
T5-članki	595
Hibrid-dolga	254
Osnovni	80
Grafi	71
Total	1.000

Tabela 6: Klasifikacijski rezultat meta-modela, izražen z natančnostjo, priklicem in mero F1 za vsako metodo ter število primerov v testni množici (podpora). Primerjamo uspešnost meta-modela pri izbiri modelov T5-članki, Hibrid-dolga, Osnovni in Grafi.

Povzemalnik	Natančnost	Priklic	Mera F1	Podpora
T5-članki	0,33	0,11	0,16	1.069
Hibrid-dolga	0,25	0,34	0,29	817
Osnovni	0,28	0,10	0,15	1.196
Grafi	0,38	0,67	0,48	1.589

Glede na Tabelo 6 je meta-model najbolj učinkovito priporočal povzemalnik, ki temelji na grafu, z rezultatom F1 0,48, natančnostjo 0,38 in priklicem 0,67. Za model Hibrid-dolga je dosegel oceno

F1 0,29, z natančnostjo 0,25 in priklicem 0,34, pri modelu T5-članki pa F1 0,16, natančnost 0,33 in priklic 0,11. Najnižji rezultat je dosegel pri metodi Osnovni z F1 0,15, natančnostjo 0,28 in priklicem 0,15. Večinski klasifikator ima na testni množici klasifikacijsko točnost 0,34. Klasifikacijska točnost meta-modela pa je bila 0,34.

5.3 Meta-model proti ostalim

V Tabeli 7 so predstavljeni končni rezultati ocenjevanja, pridobljeni s poskusi na testni množici. Omeniti velja, da je predlagani meta-model presegel vse druge modele pri vseh ocenah ROUGE. Ta rezultat poudarja učinkovitost in superiornost meta-modela pri izbiri najprimernejšega povzemalnika za določeno besedilo. Rezultat tudi kaže, da je avtomatizacija postopka izbire najboljšega modela za povzemanje koristna in odpravlja potrebo po ročni izbiri povzemalnika.

Tabela 7: Uspešnost na testni množici za vse modele. Meta-model doseže najboljše rezultate v vseh treh kategorijah.

Povzemalnik	ROUGE-1	ROUGE-2	ROUGE-L
T5-članki	19,01	5,61	13,52
Grafi	19,47	5,52	12,50
Hibrid-dolga	18,55	5,42	11,73
Osnovni	18,86	5,04	12,25
Meta-model	**20,38**	**5,85**	**13,67**

6 Zaključki

V sestavku predlagamo nov povzemalni sistem, ki vsebuje dva ekstraktivna, abstraktivni, hibridni in meta-model povzemanja. Novost je meta-model, ki je sestavljen iz polno povezane nevronske mreže, ki analizira vhodno vsebino in priporoča najprimernejši model povzemanja zanjo. Pristop deluje na kratkih in dolgih besedilih različnih žanrov in omogoča učinkovito in uspešno izdelavo kakovostnih povzetkov za slovenska besedila.

Čeprav predlagani model SloMetaSum predstavlja inovativno rešitev problema izbire najprimernejšega povzemalnika za dano besedilo, ni brez slabosti. Ena glavnih pomanjkljivosti je zanašanje na oceno ROUGE kot edino merilo za izbiro modelov. Čeprav je ROUGE pogosto uporabljena metrika na področju povzemanja besedil, ne odraža vedno dobro kakovosti povzetka ter ne zajame njegove koherentnosti in berljivosti. Druga morebitna slabost je omejen obseg študije, ki se osredotoča izključno na slovenski jezik. Čeprav so štirje modeli povzemanja, razviti za slovenščino, pomemben prispevek k temu področju, se morda ne bodo v enaki meri posplošili na druge jezike z manj viri, saj je za to potreben dober sistem za strojno prevajanje.

V prihodnjem delu bi bilo smiselno sistem razširiti na druge jezike. Sistem bi bilo smiselno primerjati z najnovejšimi povzemalnimi pristopi za velike jezike, predvsem angleščino. Poleg samodejnega ocenjevanja kakovosti sistema bi bilo koristno izvesti tudi uporabniške študije, da bi ocenili njegovo uporabnost in učinkovitost v realnih scenarijih. Podrobneje bi lahko analizirali delovanje sistema pri povzemanju novic, akademskih člankov in drugih vrst realnih vsebin, kjer se povzemanje pogosto uporablja.

Zahvala

Delo sta podprla Ministrstvo za kulturo Republike Slovenije preko projekta Razvoj slovenščine v digitalnem okolju (RSDO) in Javna agencija za znanstvenoraziskovalno in inovacijsko dejavnost Republike Slovenije (ARIS) preko raziskovalnega programa P6-0411 in projektov J6-2581, J7-3159 in CRP V5-2297. Projekt Razvoj slovenščine v digitalnem okolju 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).

Literatura

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., idr. (2020). *Language models are few-shot learners*. arXiv. https://arxiv.org/abs/2005.14165

Bučar, J. (2017). *Automatically sentiment annotated Slovenian news corpus AutoSentiNews 1.0* [Data set]. Slovenian language resource repository CLARIN.SI. http://hdl.handle.net/11356/1109

Erkan, G. in Radev, D. R. (2004). Lexrank: Graph-based lexical centrality as salience in text summarization. *Journal of Artificial Intelligence Research, 22,* 457–479.

Feng, F., Yang, Y., Cer, D., Arivazhagan, N. in Wang, W. (2022). Language-agnostic BERT sentence embedding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 878–891.

Hermann, K. M., Kocisky, T., Grefenstette, E., Espeholt, L., Kay, W., Suleyman, M., in Blunsom, P. (2015). *Teaching machines to read and comprehend*. In *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 1*, 1693-1701.

Knight, K. in Marcu, D. (2002). Summarization beyond sentence extraction: A probabilistic approach to sentence compression. *Artificial Intelligence, 139(1),* 91–107.

Le, Q. in Mikolov, T. (2014). Distributed representations of sentences and documents. In *International Conference on Machine Learning*, 1188–1196.

Lebar Bajec, I., Repar, A., Bajec, M., Bajec, Ž. in Rizvič, M. (2022). *NeMo neural machine translation service RSDO-DS4-NMT-API 1.0*. Slovenian language resource repository CLARIN.SI. http://hdl.handle.net/11356/1739

Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V. in Zettlemoyer, L. (2020). BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 7871–7880.

Mihalcea, R. in Tarau, P. (2004). TextRank: Bringing order into text. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, 404–411.

Nallapati, R., Zhou, B., dos Santos, C., Gulcehre, Ç. in Xiang, B. (2016). Abstractive text summarization using sequence-to-sequence RNNs and beyond. In *Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning*, 280–290.

Nenkova, A. in Vanderwende, L. (2005). *The impact of frequency on summarization*. Tech. rep., Microsoft Research. https://www.academia.edu/21603307/The_impact_of_frequency_on_summarization

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W. in Liu, P. J. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research, 21(140),* 1–67.

See, A., Liu, P. J. in Manning, C. D. (2017). Get to the point: Summarization with pointer-generator networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers),* 1073–1083.

Ulčar, M. in Robnik-Šikonja, M. (2023). Sequence to sequence pretraining for a less-resourced Slovenian language. *Frontiers in Artificial Intelligence, 6*.

Žagar, A., Kavaš, M., Robnik-Šikonja, M., Erjavec, T., Fišer, D., Ljubešić, N., Ferme, M., Borović, M., Boškovič, B., Ojsteršek, M. in Hrovat, G. (2022). *Corpus of academic Slovene KAS 2.0* [Dataset]. Slovenian language resource repository CLARIN.SI. http://hdl.handle.net/11356/1448

Žagar, A. in Robnik-Šikonja, M. (2021). Unsupervised approach to multilingual user comments summarization. In *Proceedings of the EACL Hackashop on News Media Content Analysis and Automated Report Generation*, 89–98.

Zhang, J., Zhao, Y., Saleh, M. in Liu, P. (2020). Pegasus: Pre-training with extracted gap-sentences for abstractive summarization. In *International Conference on Machine Learning*, 11328–11339.

Slovenski terminološki portal – nova priložnost za urejanje slovenske terminologije

Mateja JEMEC TOMAZIN
ZRC SAZU, Inštitut za slovenski jezik Frana Ramovša

Miro ROMIH
Amebis, d. o. o., Kamnik

Povzetek

Slovenski terminološki portal je enotno mesto za urejanje terminologije v slovenskem jeziku, hkrati pa so vsa razvita orodja in programske rešitve ponujeni tudi v odprtem dostopu na repozitoriju Clarin.si ter GitHub Clarin.si, kjer so na voljo za lastno namestitev. V poglavju je predstavljena zasnova enotnega terminološkega portala in vseh storitev, ki so na voljo neregistriranim in registriranim uporabnikom. Slovenski terminološki portal je možno uporabljati v slovenskem in angleškem jeziku, zato je dostopen tudi neslovensko govorečim, prav tako je omogočena tudi povezava z drugimi sorodnimi portali, kar predstavlja sinergijski učinek zbrane terminologije na enem mestu. Pri urejanju terminologije je pomembna pridobitev luščilnik terminoloških kandidatov iz lastnih besedil, kar uporabnikom daje še večjo možnost nadzora nad rezultati luščenja. Pomemben del je tudi svetovanje, kjer lahko vsi uporabniki, ki želijo odgovor o terminu, ki ga v slovarjih še niso našli, dobijo pri skupini strokovnjakov, omogočena pa je tudi namestitev orodja za samostojno svetovanje.

Ključne besede: terminološki portal, luščenje terminologije, terminološki vir, terminološko svetovanje

Abstract

The Slovenian terminology portal is a central place for the compilation of terminology in Slovenian, while all developed tools and software solutions are also offered in open access on the Clarin.si and Clarin.si GitHub repository, where they are available for self-installation. This chapter presents the structure of the unified terminology portal and all services available to both non-registered and registered users. The Slovenian terminology portal can be used in both Slovenian and English, so it is also accessible to non-Slovenian speakers, and it also provides a link to other related portals, which has the synergistic effect of summarising terminology in one place. An important feature of terminology editing is the acquisition of a paring of terminology candidates from users' own texts, which gives users even more control over the paring results. Another important feature is the consultation service, where all users who want an answer to a term not yet found in the dictionaries can get it from a group of experts, and it is also possible to install a self-consultation tool.

Keywords: terminology portal, term extraction, terminological source, terminological consulting

0 Uvod

Posebni sklop projekta Razvoj slovenščine v digitalnem okolju je bil namenjen slovenski terminologiji in njeni uporabi v novo razvitih jezikovnotehnoloških orodjih. Osnovno izhodišče pri načrtovanju samostojnega skupnega mesta za upravljanje slovenske terminologije je bilo prepoznavanje potreb jezikovnega uporabnika, natančneje specializiranega uporabnika terminologije, hkrati pa omogočiti dostop do relevantnih terminoloških podatkov tudi programom za strojno učenje. V ta namen je bilo na novo oblikovano specializirano mesto, ki ponuja več storitev, hkrati pa omogoča razvoj novih orodij ter zlasti pridobivanje izboljšanih terminoloških podatkov v prihodnje. Pomembni lastnosti terminologije sta verodostojnost in zanesljivost, zato smo snovalci želeli uporabnikom nameniti mesto, kjer bodo takšno terminologijo našli, skupinam področnih strokovnjakov pa omogočiti specializirano orodje za samostojno oblikovanje novih

terminoloških virov. Ob tem smo vsaki skupini ponudili, da na različnih stopnjah vključuje druge uporabnike terminološkega portala prek komentarjev oz. povezovanja pridobljenih podatkov z urejevalniki besedil, prevajalniki in lastnimi terminološkimi podatki. Začetna ideja je bila enotno dostopno mesto za vse terminološke podatke v slovenščini, vendar se je izkazalo, da (naj)večjo oviro še vedno predstavlja strukturiranost in dokumentiranost terminoloških virov ter pomanjkljivost metapodatkov, hkrati pa se v strokovni javnosti med avtorji še vedno pojavljajo zadržki do odprte uporabe podatkov in njihovega spreminjanja, zlasti če je bil terminološki dogovor o določenem terminu težko dosežen. Po dobrih desetih mesecih uporabe Slovenskega terminološkega portala, ki deluje na naslovu https://terminoloski.slovenscina.eu/, lahko ugotovimo, da so uporabniki zadržani do objave svojih projektov v odprtem dostopu, zato bi bilo smiselno jezikovno skupnost usposobiti za rabo in predstaviti prednosti objave manjših zaključenih terminoloških projektov. Terminološki portal je tudi edina storitev, ki je bila v okviru projekta razvita do stopnje TRL 8 in je namenjena končnim uporabnikom.

1 Izhodišča

Sodobna slovenska terminologija nastaja na mnogih strokovnih področjih, kar dokazujejo predvsem različni terminološki viri, zlasti mnogi novi terminološki slovarji, poleg tega pa je terminološka veda tudi živahno raziskovalno področje. O terminologiji kot naboru terminov, ki označujejo pojme v pojmovnem sistemu, govorimo vedno v okviru določenega strokovnega in znanstvenega področja. V zadnjih desetih letih se je okrepila dejavnost terminološkega svetovanja (Žagar Karer, Fajfar 2023), ki dopolnjuje primanjkljaje na strokovnih področjih, ki še nimajo svojih terminoloških slovarjev ali drugih virov, v katerih strokovnjaki ali drugi jezikovni uporabniki dobijo zanesljive informacije o najprimernejšem poimenovanju pojma v slovenščini. Izdelava terminološkega vira ne poteka samo v organizirani skupini, ki jo vodi terminolog (najpogosteje jezikoslovec, ki je specializiran za terminologijo), ampak se potrebe po urejanju terminologije

pojavljajo ob različnih priložnostih. Pogosto so urejeni terminološki seznami tudi eden od rezultatov raziskovalnih projektov, nastajajo pa tudi kot projektno delo v času študija, mnogi terminološki projekti se celo začnejo zaradi želje po poenotenju terminologije pred začetkom nove raziskave. *Slovenski terminološki portal* na enem mestu omogoča različne storitve. Izdelava terminološkega vira, še posebej terminološkega slovarja, je namreč veliko lažja s primernim urejevalnim orodjem, zato terminološki portal ponuja orodja za vse faze izdelave terminološkega vira (Fajfar, Žagar Karer 2015).

Na osnovi analize drugih spletišč[1] s terminološkimi vsebinami in spletnih mest, ki omogočajo izmenjavo terminoloških podatkov, smo pri zasnovi sodobnega terminološkega portala upoštevali naslednja splošna izhodišča:

- združiti čim več specializiranih enojezičnih in večjezičnih terminoloških virov na enem mestu, pri tem upoštevati enotno shemo in oblikovati mednarodno primerljive metapodatke, kar ustreza načelu *interoperabilnosti* pri pripravi raziskovalnih podatkov,
- vzpostaviti urejene in čim bolj standardizirane terminološke vire, ki ponujajo primerljiv nabor standardiziranih informacij, kar ustreza načelu *ponovne uporabljivosti*,
- ponuditi prostodostopno infrastrukturo za uporabo in izdelavo novih terminoloških virov, kar ustreza načelu *dostopnosti*,
- področnim strokovnjakom, prevajalcem, lektorjem, študentom in drugim uporabnikom ponuditi zmogljiva, a preprosta orodja, kar prav tako ustreza načelu *dostopnosti*,
- ponuditi možnost izvoza in izmenjave odprtodostopnih podatkov vsem registriranim uporabnikom terminološkega portala, kar ustreza načelu *dostopnosti*,
- omogočiti programsko povezljivost podatkov na terminološkem portalu za potrebe zunanjih spletnih storitev (npr. pomnilnikov prevodov), kar ustreza načelu *najdljivosti*,

1 V okviru projekta je Terminološka sekcija Inštituta za slovenski jezik Frana Ramovša ZRC SAZU analizirala 18 terminoloških spletišč po Evropi. Nekatera so uradna mesta institucij, ki se ukvarjajo terminologijo določenih jezikov, nekatera so spletna mesta, ki jih upravljajo posamezniki in nudijo različne terminološke informacije. Poleg dostopnosti in odprtosti podatkov smo preverjali tudi obseg informacij in povezanost z drugimi spletišči.

- in končno ponuditi odprti dostop do razvite programske kode in posameznih modulov na javnem repozitoriju, kar je cilj odprte znanosti.

1.1 Analizirana terminološka mesta

Pred začetkom projekta smo analizirali prosto dostopna terminološka spletišča in ocenili njihovo predstavitev podatkov, preglednost, možnost uporabe drugih jezikov, posodobljenost, pomoč uporabniku ter možnost lokalnega shranjevanja virov, kar predstavlja Tabela 1. Že pred začetkom projekta so bila analizirana slovenska spletišča s terminološkimi podatki (Jemec Tomazin, Žagar Karer 2020).[2]

Tabela 1: Pregledana terminološka spletišča.

Ime portala (Država)	Spletni naslov	Opis
FranceTerme – terminološka zbirka v okviru ministrstva za kulturo (Francija)	http://www.culture.fr/franceterme	zbirka okoli 8000 terminov, na začetni strani je poleg iskalnika seznam zadnjih objavljenih terminov in blog s terminološkimi vsebinami
Le grand dictionnaire terminologique (GDT) - terminološka zbirka urada za francoski jezik Quebec (Kanada)	http://gdt.oqlf.gouv.qc.ca/	zbirka skoraj 720.000 terminov, na začetni strani osnovni iskalnik, podatki o dostopnih zbirkah in blog s terminološkimi zanimivostmi, ki kažejo odzivnost na aktualno terminologijo (npr. v času covida-19)
Termium Plus® – terminološko mesto kanadske vlade (Kanada)	https://www.btb.termiumplus.gc.ca/tpv2alpha/alpha-fra.html?lang=fra	terminološka baza z več milijoni terminov v angleščini, francoščini, španščini in portugalščini, začetna stran hkrati omogoča osnovno in napredno iskanje z izbiro iskalnih pogojev, dodane so povezave na slovarje in glosarje prevajalskega oddelka ter iskalnik s povezavo na uradne vladne strani

2 Med slovenskimi spletnimi mesti, ki imajo veliko terminoloških podatkov, je tudi miranov.site/terminologija, ki vsebuje povezave in vhodno mesto za iskanje po več slovarjih hkrati. Na strani so dostopne in nadgrajene zlasti tiste informacije, ki so bile del spremljevalnih spletnih strani Večjezične terminološke baze Evroterm pred prenovo.

Ime portala (Država)	Spletni naslov	Opis
Cercaterm – osrednja ustanova za terminologijo v katalonščini (Termcat) (Katalonija, Španija)	https://www.termcat.cat/en	obsega okoli 250.000 terminov, na vstopni strani je ponujeno osnovno iskanje, posebne podstrani so na voljo za posamezna področja (medicina, pravo itd.), vključene so novice o terminoloških zanimivostih, bližnjice do družbenih omrežij
Struna – terminološka baza Instituta za hrvatski jezik (Hrvaška)	http://struna.ihjj.hr/en/	obsega okoli 250.000 vnosov, osnovna stran ponuja osnovno iskanje in nima drugih informacij
Hrvatski terminološki portal – skupno zbirno mesto hrvaške terminologije, ki ga upravlja Institut za hrvatski jezik (Hrvaška)	http://nazivlje.hr/	osnovno iskanje in razdelitev na štiri osnovne vire, filtriranje rezultatov glede na spletno stran, ki je vključena v iskalnik
Téarma.ie – uradno mesto za irsko terminologijo (Irska)	https://www.tearma.ie/	obsega okoli 325.000 vnosov v irščini in angleščini, na osnovni strani je dostopno osnovno iskanje, zadnji dodani terminološki vnosi, razdelitev področij
Termportalen (Norveška)	https://oda.uib.no/app?ma=term&mg=aterm	zelo osnovno iskalno orodje, ni povezav z drugimi portali, definicije v nekaterih slovarjih so tudi v angleščini, ni podatka o obsegu
FAO Term Portal - Združeni narodi terminologija s področja hrane in kmetijstva	http://www.fao.org/faoterm	uradna terminologija ZN, iskanje je omogočeno po vseh uradnih jezikih, število zadetkov omejeno na 50
Tepa Term Bank – terminološka baza Finskega centra za terminologijo (Finska)	https://termipankki.fi/tepa/en/	obsega okoli 365.000 vnosov, na prvi strani omogočeno osnovno iskanje, razdeljena po virih, vključuje tudi IATE
VALTER Government Termbank (Finska)	https://valter.sanakirja.fi/legal/info/about	prva stran ne vključuje iskalnika, ampak opis vira in preusmeritev na stran s področnimi glosarji
Rikstermbanken (Švedska)	http://www.rikstermbanken.se/	obsega okoli 250.000 vnosov, prosto dostopna, na prvi strani dostopno osnovno iskanje

Ime portala (Država)	Spletni naslov	Opis
IATE – Uradna evropska terminologija institucij EU	https://iate.europa.eu/home	obsega prek 700.000 vnosov, osnovno iskanje ne deluje brez izbire jezika
NATOTerm (NATO)	https://nso.nato.int/natoterm/Web.mvc	iskanje po terminologiji Natovih dokumentov, ponuja oznake zanesljivosti, vključena je tudi neveljavna terminologija, na prvi strani je omogočeno osnovno iskanje, ni podatka o obsegu baze
Evronim – terminološka baza po vzoru slovenskega Evroterma (Srbija)	http://prevodjenje.mei.gov.rs/evronim/index.php?jezik=srpl	terminološka baza nastaja v okviru približevanja Srbije EU in obsega okoli 58.000 vnosov, na prvi strani je omogočeno osnovno iskanje, vključeni so tudi slovenski ustrezniki
Termportal[3] (Nemčija)	http://www.termportal.de/de/	*seznami virov po področjih
DIN – terminološka baza Nemškega urada za standardizacijo (Nemčija)	https://www.din.de/de	osnovno iskanje terminov/ključnih besed po standardih
EuroTermBank[4]	https://www.eurotermbank.com/	obsega prek 625.000 večjezičnih vnosov, pri osnovnem iskanju ni treba najprej izbrati jezika, rezultati zadetkov so samo termini, brez podatkov o področjih

2 Terminološki portal

Pri zasnovi novega terminološkega portala smo torej izhajali iz rešitev obstoječih terminoloških portalov, tako domačih (npr. Terminologišče, Termania, Evroterm) kot tudi številnih tujih (Hrvatski terminološki portal, EuroTermBank). Terminološki portal z vsemi funkcionalnostmi je na voljo uporabnikom kot samostojna prostodostopna spletna storitev, v katero je vključeno večje število terminoloških

3 V času analize v 2020 je portal deloval, nekatere povezave niso bile več aktivne, v letu 2023 je ukinjena celotna stran, tudi domena ne obstaja več.

4 Projekt je sicer namenjen povezovanju terminoloških baz, ki vsebujejo terminologijo, nastalo in harmonizirano zlasti ob širitvi EU 2004, se pa dopolnjuje in vsaka država je ob predsedovanju povabljena, da se lahko vključijo tudi viri in posledično dopolni večjezična baza z uradnim jezikom vsakokratne predsedujoče države.

virov. V načrtu je tudi povezava terminoloških slovarjev s portalov Termania in Terminologišče, za katero je izdelana osnovna programska rešitev, vendar je še v fazi preizkušanja. Tako kot celoten terminološki portal so v obliki samostojne spletne storitve z možnostjo lastne namestitve ponujeni tudi nekateri specializirani moduli, npr. modul za luščenje terminoloških kandidatov iz besedil na spletnem mestu GitHub Clarin.si. Vsi ustvarjeni podatkovni viri in razvita programska koda so dostopni pod licencama CC BY-SA-4.0 in Apache 2 in shranjeni na repozitorijih Clarin.si ter GitHub.

Za pozitivno uporabniško izkušnjo je poleg funkcionalnosti zelo pomemben tudi videz. Ne samo barve, pisave, velikosti in oblike posameznih grafičnih elementov na strani, pomembna je tudi njihova razporeditev in povezave med njimi ter posameznimi spletnimi stranmi. Pri izdelavi terminološkega portala smo upoštevali pravila odzivnega oblikovanja za uporabo na različnih mobilnih napravah in prilagajanje strani uporabnikom s posebnimi potrebami.

Izhajali smo iz mednarodno priznanih in uveljavljenih smernic za oblikovanje spletnih vsebin, ki omogočajo dostopnost ranljivim skupinam: Web Accesibility Initiative W3C – Strategies, standards, resources to make the Web accessible to people with disabilities, dostopno na https://www.w3.org/WAI/.

Terminološki portal deluje v spletnih brskalnikih Google Chrome, Mozilla Firefox, Microsoft Edge in Safari.

2.1 Splošno o metajeziku terminološkega portala

Za lažjo uporabo terminološkega portala so definirani termini, ki so uporabljeni v luščilniku in urejevalniku. Portal ima na voljo tudi angleški vmesnik, da ga lahko spremljajo tudi neslovensko govoreči uporabniki.

Opredelitev osnovnih terminov v slovenščini z dodanimi angleškimi termini

definicija = strukturiran daljši opis pojma, ki določi nadrejeni pojem ali pojmovno skupino, navadno v eni povedi in brez velike začetnice na začetku ter končnega ločila na koncu, ang. **definition**

luščenje terminologije = identifikacija eno- in večbesednih terminoloških kandidatov v specializiranem korpusu s pomočjo namenskega programa, ang. **extraction of terminology**

podpodročje = ožje zamejena strokovna dejavnost, v kateri se uporablja terminologija, predstavljena v terminološkem viru, ang. **sub-domain**

področje = strokovna dejavnost, v kateri se uporablja terminologija, predstavljena v terminološkem viru, ang. **domain**

področna oznaka = oznaka, ki izbrano skupino pojmov znotraj terminološkega vira uvrsti na določeno ožje področje, ang. **domain label**

pojasnilo = zunajjezikovne okoliščine, ki niso del pojma, ponujajo pa dodatne informacije v zvezi z njim, npr. letnica sprejema pravnega predpisa, ang. **qualifier**

povezani termin = termin, ki je pojmovno blizu definiranemu terminu, npr. nadrejeni termin ali istovrstni termin, ang. **related term**

sinonim = termin v razmerju do drugega termina, ki označuje isti pojem, ang. **synonym**

slovar = jezikovni vir, v katerem so besede praviloma razvrščene po abecedi in definirane, ang. **dictionary**

slovarski sestavek = celotno besedilo vnosa, ki zajema termin, definicijo, tujejezični termin, sinonim, morebitne povezane termine; ang. **entry**

specializirani korpus = korpus besedil z določenega strokovnega področja, ki je osnova za luščenje terminoloških kandidatov, ang. **specialized corpus**

termin = beseda ali besedna zveza, ki označuje pojem, ang. **term**

terminološki kandidat = beseda ali besedna zveza, ki se v strokovnih in znanstvenih besedilih določenega področja pogosto pojavlja, zato je uvrščena med rezultate luščenja terminologije, ang. **term candidate**

tuji termin = termin, ki v drugem jeziku, npr. angleščini, hrvaščini, označuje isti pojem kot v slovenščini, ang. **foreign term**

2.2 Oblikovanje terminološkega vira

Delo praviloma poteka v manjši skupini strokovnjakov nekega področja, npr. informatikov, botanikov, davčnih strokovnjakov, lahko pa področni strokovnjak tudi sam izdela terminološki vir.

Pred začetkom dela je treba najprej določiti naslovnika terminološkega vira, kar določa tudi informacije, ki so v takšen vir vključene. Posamezne storitve na terminološkem portalu so bile oblikovane prav z namenom pomoči takšnim skupinam in posameznikom, da bi si lahko sami pripravili seznam terminoloških kandidatov ter ga nato dopolnjevali, dodajali želene informacije v urejevalniku ter terminološki vir na koncu tudi objavili, s čimer bi postal dostopen vsem uporabnikom portala. Shema slovarskega sestavka je dovolj široka, da lahko vsaka skupina strokovnjakov poljubno dodaja tiste elemente, ki jih želi in potrebuje v svojem terminološkem viru, hkrati pa sledi osnovnim načelom terminografske prakse (Košmrlj Levačič 2007).

2.3 Vključeni terminološki viri in varovanje avtorskih pravic

Čeprav je Slovenski terminološki portal zasnovan kot odprto mesto za urejanje terminologije, seveda ne spreminja obstoječih pravic avtorjev v že nastalih virih, zato so terminološki viri, ki so del portala, odprti in jih lahko registrirani uporabniki prenesejo na svoje računalnike, agregirani terminološki viri drugih mest pa ohranjajo licenco, ki jo imajo na izvirnih mestih.

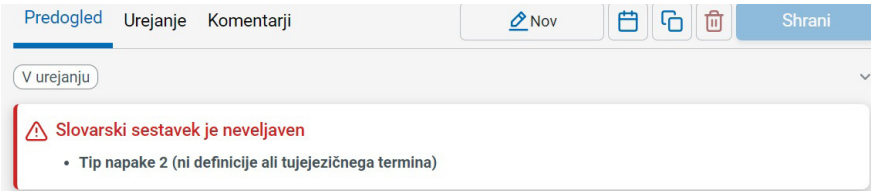
Vsi viri, ki so bili pridobljeni v okviru projekta RSDO, so bili pretvorjeni v enotno strukturo, saj je tako olajšano prikazovanje, urejanje, pretvarjanje, uvažanje, izvažanje in izmenjevanje podatkov.

2.4 Uporabniki

Uporabniki lahko uporabljajo iskanje po terminoloških virih brez registracije. Za uporabo drugih storitev – luščenja, urejanja terminoloških virov, komentiranja in zastavljanja terminoloških vprašanj v svetovalnici – pa je ob prvem obisku potrebna registracija, ob vseh naslednjih pa prijava. Vsak uporabnik lahko kadarkoli izbriše svoj

račun, ob registraciji pa se mora strinjati s pogoji uporabe in politiko zasebnosti, ki sta sestavni del pojavnega polja ob registraciji.

Vsak registrirani uporabnik lahko sam ustvari in ureja enega ali poljubno mnogo terminoloških virov. Da bi na portalu zagotovili čim bolj kakovostne podatke, je glede na uredniško politiko portala poleg samodejnega preverjanja predvideno tudi potrjevanje virov pred objavo na zunanjem delu terminološkega portala, ki ga opravi skrbnik terminoloških virov oziroma slovarjev. Samodejno preverjanje vključuje formalno/tehnično preverjanje in samodejna opozorila uredniku določenega terminološkega vira o morebitnih nepravilnostih (npr. manjkajočih vnosih ali identičnih definicijah), kar kaže Slika 1. Pogoj za objavo je, da ima vsak slovarski sestavek najmanj 2 informaciji, poleg slovenskega termina, ki je obvezna sestavina vseh slovarskih sestavkov, tudi definicijo ali tujejezični termin. S tem je onemogočena objava enojezičnih seznamov terminoloških kandidatov, ki bi drugim uporabnikom ne dala dovolj informacij, saj ne bi bilo dodanih podatkov o področju, ne bilo jasno, ali je seznam že prečiščen, kateri viri so bili vključeni pri luščenju ipd., seveda pa si lahko tisti uporabniki, ki so takšne sezname pripravili, sezname izvozijo lokalno in nadaljujejo delo z njimi.



Slika 1: Primer samodejnega opozorila v slovarskem sestavku.

Skrbnik terminoloških virov lahko v skladu s politiko terminološkega portala še ročno potrdi in dokončno objavi terminološki vir, kar je ena od možnih nastavitev celotnega urejanja portala.

2.5 Klasifikacije področij

Terminologija je veljavna znotraj področja, zato je bilo veliko truda vloženega v pripravo čim bolj uporabnega nabora področij, ki tvorijo

pomembne metapodatke pri posameznem terminološkem viru. Obstajajo različne klasifikacije področij, vendar se je v praksi pokazalo, da nobena od njih ni dovolj preprosta in blizu uporabnikom, ki ne poznajo podrobnosti različnih klasifikacijskih sistemov, po drugi strani pa tudi ne pokrivajo vseh potreb za gradnjo organizirane terminološke zbirke. Za Slovenski terminološki portal je bil pripravljen samostojen seznam področij, pripravljen po vzoru univerzalnega decimalnega klasifikacijskega sistema,[5] ki ima pripisana tudi ustrezna področja v klasifikacijah CERIF[6] in Eurovoc,[7] zato je omogočena preslikava.

Uporabnikom je poleg predlaganega seznama podpodročij omogočeno tudi dodajanje novih z namenom čim bolj jasne umestitve terminoloških virov. Osnovno področje mora biti obvezno izbrano iz spustnega seznama vključenih področij, podpodročja pa lahko registrirani uporabniki prosto vpisujejo, če želenega na spustnem seznamu predlaganih podpodročij ne najdejo. Nova podpodročja mora potrditi administrator portala, s čimer se preprečuje množenje sorodnih oznak za isto podpodročje. Uvrstitev terminološkega vira v področje predlaga urednik oz. administrator terminološkega vira, ki ima vse pravice urejanja v tem viru. Pred potrditvijo skrbnika terminoloških virov in objavo na zunanjem delu terminološkega portala je lahko opravljen še vsebinski pregled slovarskih sestavkov in reševanje morebitnih komentarjev, če jih uredniki terminološkega vira zastavijo skrbniku terminoloških virov ali administratorju portala. Tudi brez potrditve pa lahko urednik terminološkega vira svoj vir vedno uporablja v zasebnem delu terminološkega portala ali ga izvozi in shrani na svojem računalniku.

Na zunanjem delu terminološkega portala je uporabnikom vedno na voljo le zadnja verzija objavljenega terminološkega vira.

5 Univerzalna decimalna klasifikacijsa (UDK) je mednarodno enotno normativno orodje za vsebinsko označevanje dokumentov in iskanje informacij, ki ga uporabljajo knjižnice in arhivi. V slovenskem jeziku ga urejajo v Narodni in univerzitetni knjižnici.

6 CERIF je Evropska klasifikacija raziskovalne dejavnosti, vendar še vedno odraža tradicionalno delitev znanstvenih področij, zato je manj zanesljiva za določanje novih interdisciplinarnih področij.

7 Eurovoc je večjezični in multidisciplinarni tezaver EU, ki je v osnovi razdeljen na 21 področij, ta pa so še dodatno razdeljena na 127 podpodročij, nekatera se pojavljajo pri več osnovnih področjih, kar spet uporabniku brez dobrega poznavanja strukture klasifikacijskega sistema oteži izbiro najprimernejše področne oznake.

Kadar so v terminološkem viru dodane večje vsebinske dopolnitve, je smiselno terminološki vir izvoziti in ga tudi kot novo verzijo naložiti v repozitorij Clarin.si, kar omogoča sledljivost sprememb in poveča zanesljivost terminološkega vira.[8]

Po desetih mesecih uporabe lahko ugotovimo, da je registriranih uporabnikov razmeroma malo, prav tako objavljenih virov, čeprav vse storitve portala delujejo, na portalu so sproti objavljeni tudi novi odgovori na terminološka vprašanja iz Terminološke svetovalnice ISJFR ZRC SAZU.

2.6 Uporabniške vloge

Terminološki portal je tudi po zaključku projekta vodena in vsebinsko ter tehnično podprta storitev. V nadaljevanju so opisane posamezne vloge na terminološkem portalu.

- **Administrator ali skrbnik portala** – glavni urednik vseh vsebin na portalu, ki ima vse pravice dopolnjevanja, spreminjanja, brisanja.
- **Skrbnik terminoloških virov** oz. **skrbnik slovarjev** – glavni urednik vseh terminoloških virov, ki lahko dovoli objavo terminološkega vira na zunanjem delu terminološkega portala, dodaja uporabnike, da smejo urejati terminološki vir, ureja podpodročja.
- **Skrbnik svetovalnice** – glavni urednik svetovalnice, ki vodi svetovalnico in dodeljuje terminološka vprašanja posameznim svetovalcem (samo pri samostojnih namestitvah terminološkega portala, ki niso povezani s Terminološko svetovalnico Inštituta za slovenski jezik Frana Ramovša ZRC SAZU).
- **Svetovalec** – član svetovalnice, ki je odgovoren za reševanje poslanega terminološkega vprašanja in pripravo odgovora, ki ga pošlje v potrditev skrbniku svetovalnice (samo pri samostojnih namestitvah terminološkega portala, ki niso povezani s Terminološko svetovalnico Inštituta za slovenski jezik Frana Ramovša ZRC SAZU).

8 Takšen primer je tudi korpus OSS, korpus vseh besedil, ki so vključena v Nacionalni portal odprte znanosti Openscience.si in je letno nadgrajen s prirastom zaključnih diplomskih, magistrskih in doktorskih del, zato bo v repozitorij Clarin.si v pomladanskih mesecih vključena verzija s prirastom preteklega leta.

Vloge, ki jih lahko pridobijo vsi registrirani uporabniki terminološkega portala, če želijo sodelovati pri izdelavi terminološkega vira:

- **Administrator terminološkega vira** – registrirani uporabnik, ki v urejevalniku ustvari nov terminološki vir in ima vse pravice pri urejanju tega vira, k sodelovanju lahko povabi tudi druge registrirane uporabnike, ureja metapodatke in lahko vse podatke tudi izbriše.
- **Urednik terminološkega vira** – registrirani uporabnik, ki ga administrator terminološkega vira povabi k izdelovanju in urejanju slovarskih sestavkov, terminološkega vira ne more izbrisati, lahko pa izbriše posamezne slovarske sestavke.
- **Strokovni in jezikovni pregled** – administrator terminološkega vira lahko določi pravico urejanja vsebine posameznih slovarskih sestavkov tudi registriranim uporabnikom samo za jezikovni ali strokovni pregled, pri čemer lahko ti sodelavci urejajo samo vsebino slovarskih sestavkov, ne morejo pa urejati metapodatkov terminološkega vira ali brisati slovarskih sestavkov.

Registrirani uporabnik

Uporabnik, ki v obrazec vnese veljavni e-naslov, si izbere uporabniško ime in geslo ter potrdi politiko zasebnosti in pogoje uporabe, s tem pa pridobi pravico uporabe storitev luščenja, urejanja, uvoza in izvoza terminoloških virov, zastavljanja terminoloških vprašanj in komentiranja.

Neregistrirani uporabnik

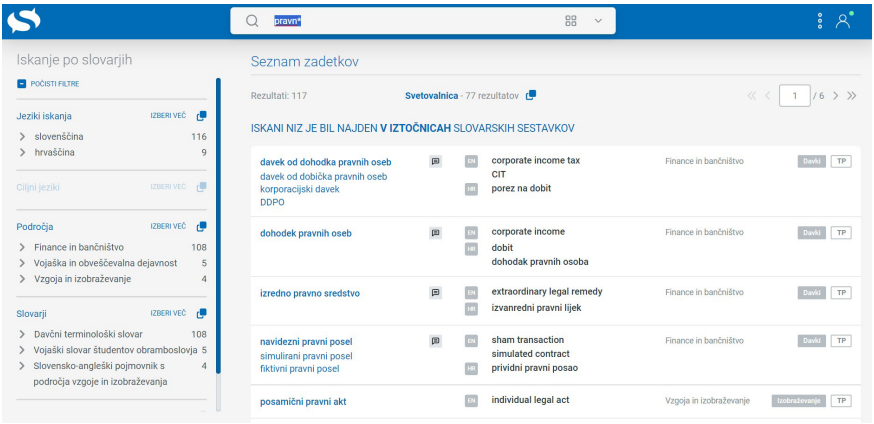
Uporabnik terminološkega portala, ki lahko poizveduje po terminih in brska med objavljenimi terminološkimi odgovori, ne more pa uporabljati drugih storitev.

Uporabniki se na terminološkem portalu identificirajo z uporabniškim imenom in e-naslovom ter geslom. Hranjenje osebnih podatkov je urejeno s Politiko zasebnosti in Pogoji uporabe, ki jih mora vsak novi uporabniki potrditi ob registraciji.

3 Iskanje

Iskanje informacij po terminoloških virih je osnovna funkcija terminološkega portala, ki jo uporablja večina neprijavljenih uporabnikov. Pri razvoju iskalnika je bilo najpomembnejše vodilo enostavnost in hitrost iskanja ter pregleden prikaz podatkov o najdenih zadetkih.

Najpogostejši način iskanje je *vpis niza znakov* (Slika 2) v iskalno polje, ki ponudi prikaz zadetkov v terminoloških virih in v svetovalnici z osnovnimi podatki. Iskanje pri malih in velikih črkah je poenostavljeno tako, da začetnica vpisane črke ne vpliva na zadetke. Iskalnik vedno najde obe.



Slika 2: Primer iskanja z vnosom niza znakov.

3.1 Osnovno iskanje

Osnovno iskanje je privzeta oblika iskanja, zato smo to iskanje naredili čim bolj preprosto. Iskalnik ima klasično iskalno polje, v katero se vpiše niz znakov, ki ga želimo najti. Iskani niz se išče ločeno po slovenskih terminih, tujih terminih in drugih elementih slovarskega sestavka.

Iskanje besed in besednih zvez

Poleg iskanja ene besede je v terminologiji pogosto iskanje večbesednih terminov. Če bi v iskalno polje napisali zvezo *davčna odločba*,

iskalnik najde vse tiste slovarske sestavke, kjer v določenem elementu hkrati nastopata besedi *davčna* in *odločba*. Vrstni red besed ni pomemben – besedi sta lahko ena za drugo ali pa je med njima še kakšna druga beseda. Presledek v tem primeru torej predstavlja operator IN.

Če želimo najti besedno zvezo, kjer morata besedi stati skupaj in v točno določenem vrstnem redu, vpišemo iskalni pogoj *"davčna odločba"*. Dvojna narekovaja torej določata iskanje po besedni zvezi.

Iskanje s posebnimi ukaznimi znaki

Omogočena je tudi uporaba nadomestnih znakov – vprašaja (?) in zvezdice (*). **?** nadomešča natanko en poljuben znak, ***** pa nadomešča od nič do neskončno poljubnih znakov.

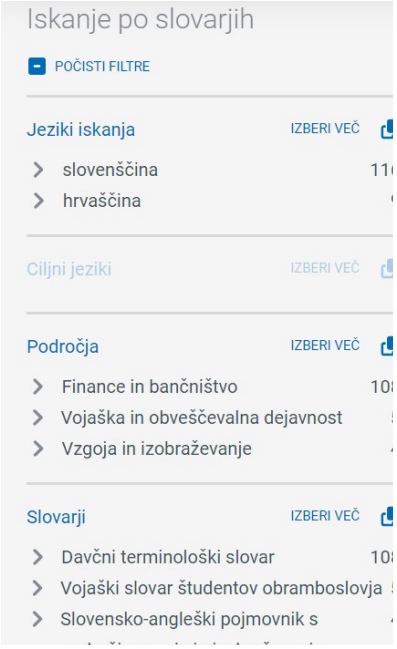
Drugi znaki, kot so klicaj (*!*) (za izločanje niza), ključnik (*#)* (za od do) in nekateri drugi, pri iskanju po analizi terminoloških portalov v Sloveniji in tujini niso zares pogosto uporabljeni, zato jih pri iskanju nismo omogočili.

Nadomestna znaka ? in * je mogoče uporabiti tudi pri iskanju po besednih zvezah.

3.2 Napredno iskanje

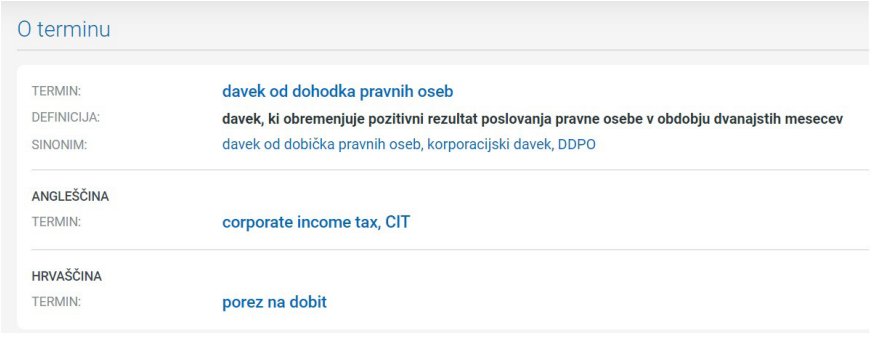
Napredno iskanje poleg osnovnega iskalnega polja za vnos iskalnega niza omogoča še iskanje po jezikih, področjih, slovarjih in tudi virih.

Zadetke je mogoče tudi filtrirati v polju na levi strani in izbrati le določeno vrsto zadetkov. Ne glede na vrsto iskanja je iskalnik prosto dostopen vsem uporabnikom, torej za uporabo iskalnika ni potrebna registracija. Prikazujejo se podatki iz vseh vključenih virov na portal in vseh z API-jem povezanih spletnih mest (Slika 3).



Slika 3: Navigacijsko okno s podatki o številu zadetkov po slovarjih, področjih in virih, ki omogoča filtriranje zadetkov.

Ko na seznamu zadetkov izberemo enega od zadetkov, se prikaže celoten slovarski sestavek (Slika 4).



Slika 4: Prikaz slovarskega sestavka v viru na Slovenskem terminološkem portalu.

Na voljo so tudi zadetki v svetovalnici (Slika 5). Vsi podatki o številu zadetkov veljajo za iskani niz znakov.

Slika 5: Prikaz zadetka v svetovalnici.

Viri, ki bodo s portalom samo povezani prek API-ja za prenos podatkov, se bodo prikazovali na izvirnem mestu.[9]

3.3 Prikaz in razvrščanje zadetkov

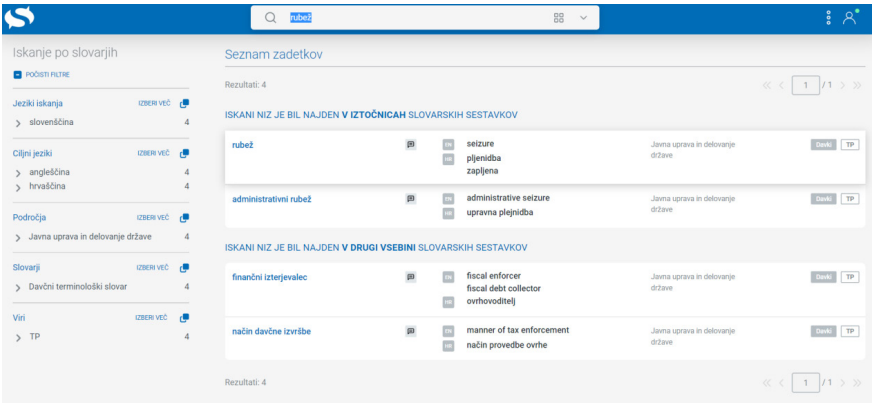
Zadetki iskanja se razvrščajo po naslednjem vrstnem redu (Slika 6):

1. Slovarski sestavki, kjer je slovenski termin enak iskalnemu nizu.
2. Slovarski sestavki, kjer slovenski termin vsebuje iskalni niz.
3. Slovarski sestavki, kjer je tuji termin enak iskalnemu nizu.
4. Slovarski sestavki, kjer tuji termin vsebuje iskalni niz.
5. Slovarski sestavki, kjer sinonimi vsebujejo iskalni niz.
6. Slovarski sestavki, kjer drugi elementi vsebujejo iskalni niz.

V okviru vsake od naštete podmnožice zadetkov se zadetki uredijo po abecedi slovenskih terminov.

Če iskanje ni uspešno, ker iskani niz ne ustreza nobenemu slovarskemu sestavku, iskalnik opozori: *Ni zadetkov.*

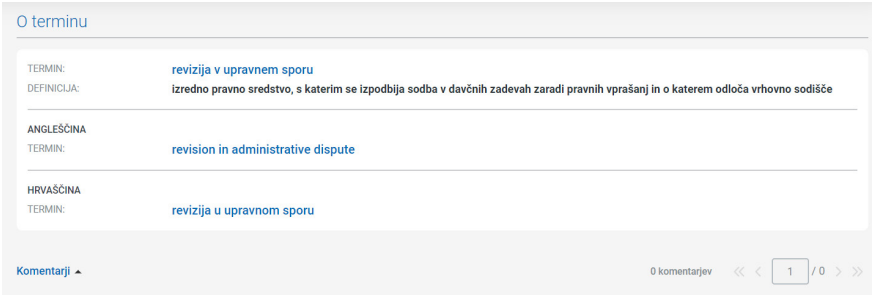
9 Storitev je na voljo v testnem okolju in še ni objavljena na Slovenskem terminološkem portalu.

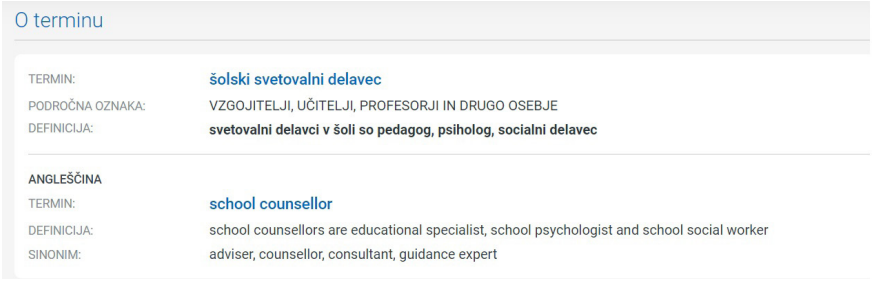


Slika 6: Prikaz razvrščanja zadetkov.

3.4 O terminu

Izbira enega termina na seznamu zadetkov prikaže vse podatke o izbranem terminu. Obseg podatkov je odvisen od izbranega terminološkega vira, ker nimajo vsi termini na terminološkem portalu enakovrstnih informacij (Slika 7, Slika 8).

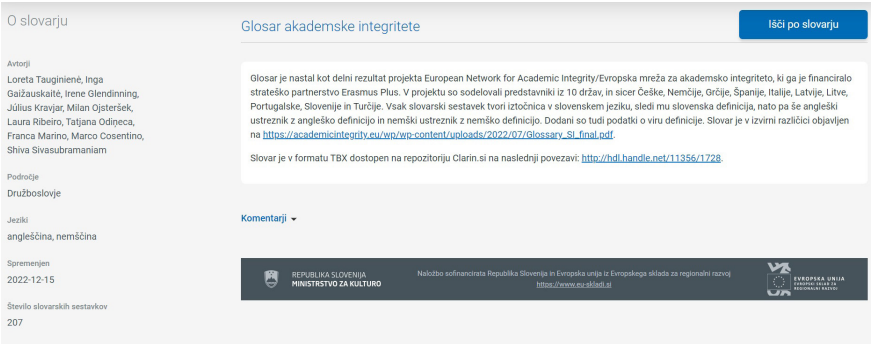




Sliki 7 in 8: Prikaz dveh terminov z različnimi vrstami podatkov.

3.5 O slovarju

Osnovne informacije o terminološkem viru so na voljo v razdelku *O slovarju*, ki vsebuje podatke, katera področja pokriva terminološki vir, kateri jeziki so vključeni, kdo so njegovi avtorji, koliko slovarskih sestavkov ima, kdaj je bil nazadnje spremenjen, pa tudi opis, kako je terminološki vir nastajal (Slika 9). Registrirani uporabniki lahko administratorju terminološkega vira oddajo tudi svoj komentar. Več informacij omogoča boljšo dokumentiranost terminološkega vira, kar je zlasti pomembno pri primerjavi in poznejših analizah.



Slika 9: Primer opisa terminološkega vira, ki je objavljen na Slovenskem terminološkem portalu (*Glosar akademske integritete*).

4 Luščenje

Osnova za izdelavo vsakega terminološkega vira je seznam terminov ali *geslovnik*. Oblikovanje geslovnika lahko poteka na več načinov. Izbira terminov je odvisna od avtorjev oz. urednikov slovarja. Registrirani uporabniki imajo na terminološkem portalu tri osnovne možnosti:

a) Termine lahko v urejevalnik dodajajo neposredno. Pri tem se opirajo na svoje znanje, izkušnje in poznavanje strokovnega področja. Takšen način je primeren zlasti za področja, ki še nimajo ustreznih strokovnih in znanstvenih besedil v slovenščini, predloge poimenovanj, ki so lahko že relativno uveljavljeni, pa bi radi razširili v širši strokovni skupnosti.

b) Druga možnost je ročno izpisovanje terminov iz strokovnih besedil in prav tako vnašanje neposredno v urejevalnik, kar je smiselno pri virih, ki se jih ne da optično prebrati ali še niso digitalizirani, prinašajo pa pomembne podatke.
c) Tretja možnost je uporaba samodejnega postopka luščenja terminoloških kandidatov iz strokovnih in znanstvenih besedil in priprava geslovnika iz seznama rezultatov. To metodo omogoča tudi terminološki portal s posebnim modulom za luščenje, ki je opisan v nadaljevanju. Uporabnik lahko pripravi svoj specializirani korpus strokovnih besedil ali izbere področje v že oblikovanem in označenem korpusu OSS, v katerem so združena besedila, objavljena na Nacionalnem portalu odprte znanosti. S pomočjo orodja za luščenje je mogoče hitro izluščiti terminološke kandidate, kar poenostavi in skrajša uvodno fazo izdelave terminološkega vira.

Luščilnik za svoje delovanje uporablja temeljne jezikovne tehnologije za slovenščino, in sicer tokenizacijo, lematizacijo in oblikoskladenjsko označevanje, pri statističnem vrednotenju terminološkosti pa se za primerjavo uporabljajo lematizirani frekvenčni seznami (n-grami) referenčnega korpusa Gigafida 2.0. Pri izboljšavi metod luščenja so bile uporabljene tudi nevronske metode, ki so izbirale terminološke kandidate glede na sobesedilo, v katerem se pojavlja (Hong Hanh idr. 2022). Kakovost izluščenih terminoloških kandidatov je odvisna od kakovosti in uravnoteženosti izbranih besedil. Besedila korpusa OSS so že predhodno oblikoskladenjsko označena, lastna uporabnikova pa je treba še označiti, zato lahko postopek luščenja traja dlje.

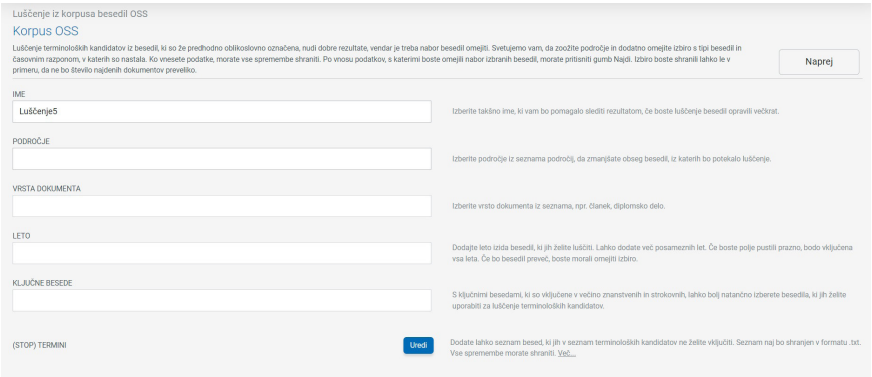
Modul za luščenje terminoloških kandidatov je na voljo v okviru terminološkega portala, kot programska koda in samostojno orodje pa je dostopen tudi na repozitoriju GitHub.[10]

10 Dostopen na https://github.com/clarinsi/rsdo_luscilnik.

4.1 Postopek luščenja

V nadaljevanju je natančneje opisan postopek luščenja na terminološkem portalu:

1. Uporabnik pripravi nabor besedil, iz katerih želi pridobiti terminološke kandidate za določeno strokovno področje. Pri tem ima dve možnosti:
   - **Izbira besedil iz korpusa OSS**, ki je dostopen tudi na konkordančnikih v okviru Clarin.si. Ta korpus se bo predvidoma dopolnjeval enkrat letno. Besedila je mogoče izbirati po področju, vrsti dokumenta, letu in ključnih besedah, saj je največje dovoljeno število besedil za luščenje omejeno (Slika 10).

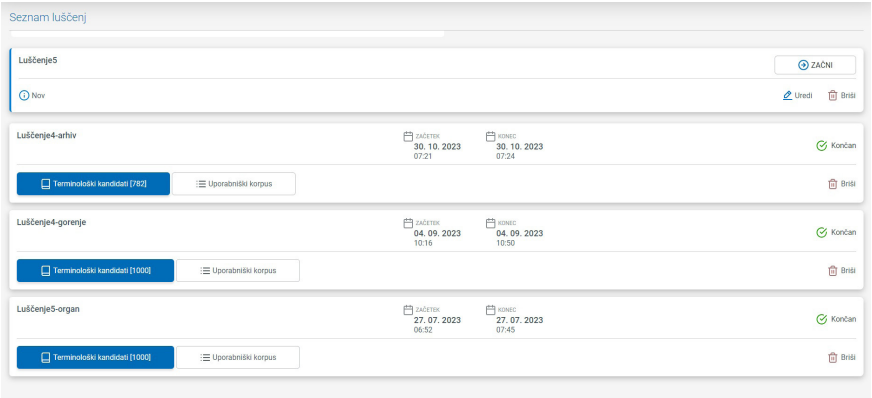


Slika 10: Prikaz uporabe korpusa OSS.

Pred začetkom luščenja in izbiro gumba Naprej je treba izbrati področje s seznama, prav tako se lahko na spustnem seznamu izbere vrsta dokumenta, npr. *izvirni znanstveni članek*, *univerzitetni učbenik z recenzijo* itd., dodatno izbiro lahko uporabnik omeji z leti objave in ključnimi besedami. Po dodanih kriterijih uporabnik dobi omejeno število dokumentov.

Primer izbiranja: V spustnih menijih izberemo področje *Varnost, zaščita in reševanje* in vrsto dokumenta *izvirni znanstveni članek* in želimo najti vse dokumente, ki ustrezajo tema dvema pogojema. Izberemo ukaz Najdi v desnem kotu spodaj. Iskalnik nam ponudi določeno število dokumentov, ki jih shranimo.

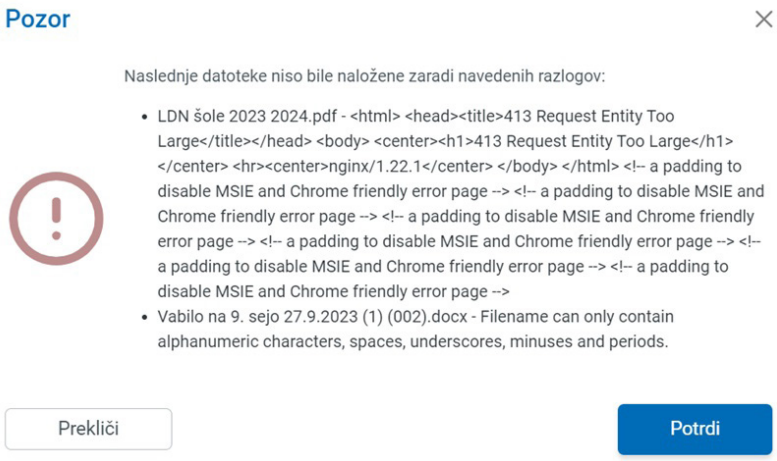
Luščenje se pojavi na seznamu luščenj (Slika 11) in lahko izberemo ukaz *Začni.*



Slika 11: Seznam luščenj.

Ker je luščenje postopek, ki zahteva več spominskega prostora, registrirani uporabnik dobi sporočilo, da bo o zaključku luščenja obveščen na e-naslov, ki ga je navedel ob registraciji. Z izbiro gumba *Razumem* se postopek luščenja začne izvajati in je tudi ob večjem številu podatkov navadno zaključen v nekaj urah.

- **Nalaganje lastnih datotek** za luščenje je omogočeno v nekaterih pogostih formatih (.pdf, .docx, .txt). Kljub dovoljenim formatom pa luščilnik najbolje deluje, če so vsebine v čisti besedilni obliki, pri drugih formatih so mogoče napake pri pretvorbi, kar zazna že program ob nalaganju dokumentov in na to opozori uporabnika (Slika 12).



Slika 12: Neuspešno naloženi lastni dokumenti.

Po nalaganju dokumentov uporabnik izbere ukaz *Naprej* in nato v seznamu osnovnih luščenj ukaz *Začni*, po koncu luščenja pa dobi obvestilo na e-naslov.

2. **Avtomatsko luščenje**

Rezultat luščenja sta seznam terminoloških kandidatov (v kanonični obliki, torej osnovni obliki, navadno v imenovalniku ednine). Kanonična oblika termina je pomembna, da strokovnjak, ki navadno ni tudi jezikoslovno usposobljen, lahko čim prej začne z urejanjem terminološkega vira. Uporabljeni kanonizator omogoča dobro prepoznavo slovenske terminologije in besednih vrst.[11] S statističnimi podatki (npr. TF, TF-IDF) lahko uporabnik filtrira in ureja zadetke, dodan pa je tudi konkordančnik, razvit na osnovi konkordančnika, ki ga uporablja korpus Gigafida, v katerem uporabnik lahko preverja pojavljanje terminoloških kandidatov in primere rabe v posameznem besedilu.

Hibridni sistem luščenja je zasnovan na podlagi prepoznavanja izbranih besednovrstnih vzorcev in statističnega razvrščanja kandidatov. Preizkusili in razvili smo naslednje metode:

11 Posamezne napake se pojavljajo pri prepoznavi 2. ženske sklanjatve, vendar je ta tudi v terminologiji redkejša, zato so pomanjkljivosti obvladljive.

- Klasične statistične mere stabilnosti besednih zvez (npr. MI, PMI), mere za primerjavo specializiranega in referenčnega korpusa (npr. LUIZ-CF), statistične mere specifičnosti za izbrano strokovno področje glede na širšo zbirko besedil (npr. TF-IDF) in druge klasične mere za luščenje terminoloških kandidatov (C/NC value). Za primerjavo z referenčnim korpusom luščilnik interno uporablja korpus Gigafida 2.0, ki je označen z istimi orodji oz. s frekvenčnim seznamom, izdelanim na podlagi tega korpusa (frekvenčni seznam n-gramov lem). Za primerjavo s širšim korpusom besedil so uporabljena besedila Nacionalnega portala odprte znanosti.
- Naprednejše metode z uporabo kontekstualnih besednih vložitev, kjer je bilo preizkušeno nenadzorovano učenje s primerjavo kontekstualnih vektorjev splošnega in domenskega korpusa ter nadzorovano učenje s prepoznavanjem terminov v učnih korpusih.

Različne mere ter njihove kombinacije so bile sistematično ovrednotene na testnih množicah, da smo lahko identificirali optimalno kombinacijo mer za končni sistem. Izluščenim terminom je pripisana kanonična oblika. Pri tem smo uporabili oblikoskladenjske oznake samega korpusa za iskanje termina v imenovalniku ter ustrezne zunanje jezikovne vire (npr. Sloleks oz. sezname terminov, ki so že vključeni v repozitorij Clarin.si).

V korpusu besedil smo razpoznali tudi stavke, ki lahko uporabniku pomagajo kot osnova pri tvorjenju definicije. Metode temeljijo na razpoznavanju dobrih primerov glede na informacijo o dolžini stavka, informacijo o mestu termina v stavku in skladenjskih informacijah, značilnih za strokovna in znanstvena besedila.

3. Seznam terminoloških kandidatov (s pripadajočimi podatki) in oba korpusna konkordančnika (za OSS in korpus lastnih besedil) sta v razdelku za luščenje znotraj terminološkega portala na voljo uporabniku za nadaljnje delo in uporabo (npr. izvoz v standardni format).

Modul za luščenje in označevanje terminov v besedilih je na voljo vsem registriranim uporabnikom. Vse uporabnike, ki bodo izdelovali svoj terminološki vir, spodbujamo k uporabi modula za luščenje.

4.2 Seznam luščenj

Osnovna stran luščenja vsebuje seznam luščenj, ki so v enem od štirih stanj – **v urejanju**, **v obdelavi**, **prekinjeno** ali **končano**. Vsak uporabnik lahko shrani podatke za največ pet luščenj v seznamu. Ob zahtevi po novem luščenju mora uporabnik izbrisati eno od luščenj, da lahko začne novo.

4.3 Dodajanje in urejanje luščenj

Pri dodajanju ali urejanju novega luščenja je treba vnesti nekaj obveznih podatkov. Med obvezne podatke sodi ime luščenja, prednastavljeno je poimenovanje *Luščenje 1*, *Luščenje 2* ..., pred luščenjem pa je treba naložiti želena besedila ali določiti parametre, na osnovi katerih bodo izbrana besedila iz korpusa OSS. Med opcijske podatke sodi vnos seznam neželenih terminov, t. i. stop seznam, ki naj jih program v vsakem primeru izloči, zato jih v končnem seznamu terminoloških kandidatov ni. Stop seznam lahko vključuje tudi termine, ki jih želi uporabnik v vsakem primeru vključiti in zato ne potrebuje še (dodatne) potrditve z luščilnikom.

5 Urejanje

Terminološki viri nastajajo v različnih okoljih. Ustvarjajo jih strokovnjaki v javnih in zasebnih ustanovah, raziskovalci, prevajalci, študenti in drugi. Spletni urejevalnik za izdelavo terminoloških virov je namenjen lažjemu delu takšnih skupin, ob tem pa jim je prek terminološkega portala na voljo tudi pomoč skrbnikov portala glede strukture terminološkega vira in metodologije dela. Pri zasnovi in izdelavi urejevalnika smo upoštevali, da uporabniki niso profesionalni sestavljalci terminoloških virov, zato smo oblikovali orodje, ki

je enostavno za uporabo, a kljub temu ponuja vse potrebne funkcije za urejanja terminološkega vira.

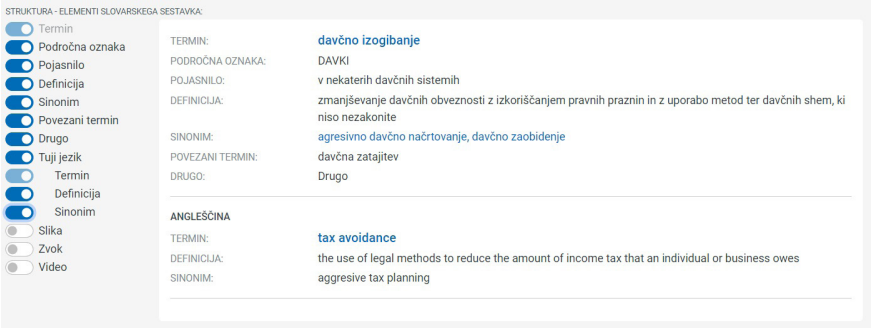
Vsi terminološki viri imajo enako strukturo zapisa slovarskega sestavka. Za nove terminološke vire, ki bodo nastali na portalu, bo mogoče prilagajati elemente slovarskega sestavka od najbolj preprostega z iztočnico v slovenskem jeziku in definicijo ali tujim terminom do kompleksnejšega z večjim številom elementov.

Omogočena sta uvoz in izvoz podatkov iz in v običajne datotečne formate ter večuporabniško delo z različnimi vlogami uporabnikov ter dodajanje večpredstavnih podatkov in povezav na zunanje vire.

Urejanje terminoloških virov je omogočeno samo registriranim uporabnikom.

5.1 Struktura slovarskega sestavka

Vsi elementi slovarskega sestavka so izbirni z izjemo termina v slovenskem jeziku, ki mu mora biti dodana definicija ali tuji termin (Slika 13). Na sliki je prikazan slovarski sestavek, ki ima izbrane skoraj vse elemente, ki so na voljo. Omogočeno je namreč tudi dodajanje multimedijskih elementov, ki še dodatno lahko dopolnijo opis posameznega pojma.



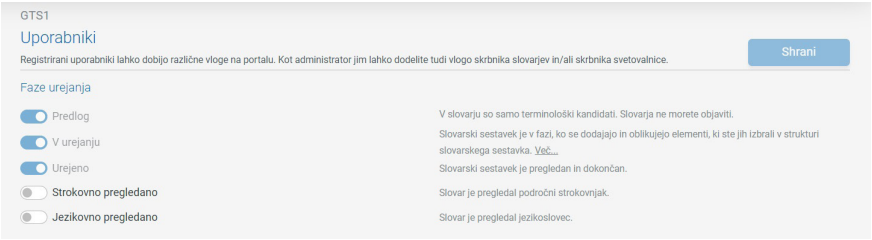
Slika 13: Struktura slovarskega sestavka z izbranimi elementi.

Pri strukturi velja opozoriti na dve posebnosti, eno je kategorija *povezani termin*, ki označuje sorodni, vendar ne isti pojem, npr. *davčna optimizacija* označuje uporabo zakonitih možnosti za

zmanjševanje davčne osnove, da bi davčni zavezanec plačal čim nižji davek, *davčno izogibanje* pa označuje izkoriščanje pravnih praznin, pri čemer je sicer osnovni namen plačilo čim nižjega zneska davka. Uporabnik lahko izbere tudi vrsto povezave med termini, in sicer *sorodni*, *ožji*, *širši*, pri čemer mora izbirati med termini, ki jih je vključil v svoj terminološki vir. Druga posebnost je polje *Drugo*, ki nima določene vsebine, uporabniki pa lahko vanj vključijo tisto vsebino, ki se jim zdi potrebna, npr. navajajo vire, primere rabe, opozorila o terminološkem dogovoru itd. Pri uporabi luščenja terminoloških kandidatov, so v tem polju izpisani tudi dobri primeri rabe, ki vključujejo termin.

5.2 Faze urejanja

Urejanje terminoloških virov je sestavljeno iz več faz (Slika 14). Na Slovenskem terminološkem portalu lahko administrator terminološkega vira aktivno vpliva na zadnji dve fazi.



Slika 14: Prikaz faz urejanja terminološkega vira.

- **Predlog:** V terminološkem viru so samo terminološki kandidati, ki jih je uporabnik iz luščilnika uvozil v novi terminološki vir. Tega seznama ni mogoče objaviti, ampak je treba dodati vsaj še en podatek (definicijo ali tujejezični termin) pri vsakem slovarskem sestavku.
- **V urejanju:** Slovarski sestavki v terminološkem viru so v fazi, ko se dodajajo in oblikujejo elementi, ki jih je administrator terminološkega vira izbral v strukturi slovarskega sestavka. V tej fazi

lahko sodelujejo vsi registrirani uporabniki, ki jih je administrator terminološkega vira povabil v skupino.

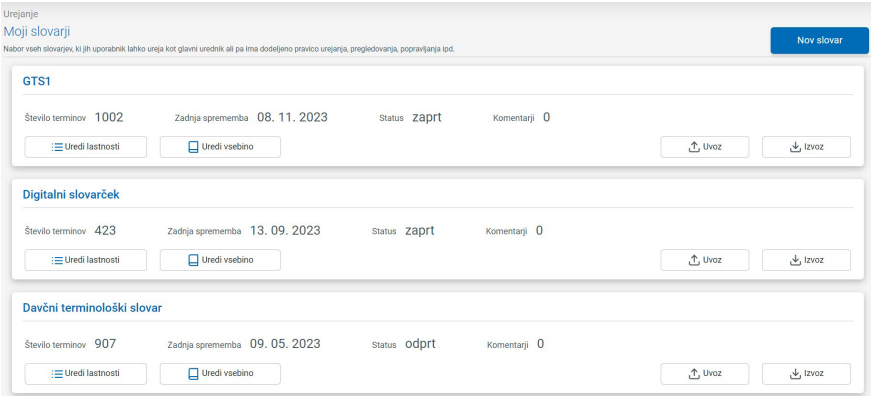
- **Urejeno:** Slovarski sestavki v terminološkem viru so pregledani in dokončani. Administrator terminološkega vira ga želi objaviti, zato pošlje sporočilo skrbniku terminoloških virov, če to funkcijo omogoča nastavitev na nivoju portala.

Ponujeni sta še dve možnosti, ki ju administratorji terminološkega vira lahko izberejo in uporabijo:

- **Strokovno pregledano:** Terminološki vir je pregledal področni strokovnjak ali recenzent, ki ni avtor definicij.
- **Jezikovno pregledano:** Terminološki vir je pregledal jezikoslovec in odpravil morebitne jezikovne pomanjkljivosti.

5.3 Nov slovar

Registrirani uporabnik lahko ureja vse terminološke vire, ki so zbrani v zavihku Moji slovarji (Slika 15). Poimenovanje *slovar* kot najpogosteje uporabljeno predstavlja sinonim terminološkemu viru. Vidni so vsi terminološki viri, ki jih je uporabnik sam začel izdelovati ali pa mu jih je drug uporabnik dovolil urejati. Viri so razvrščeni padajoče po datumu zadnje spremembe, uporabnik pa lahko vidi, koliko slovarskih sestavkov vsebujejo, število oddanih komentarjev in status, pri



Slika 15: Prikaz terminoloških virov posameznega uporabnika.

čemer *odprt* pomeni, da je terminološki vir viden vsem, tudi neregistriranim uporabnikom terminološkega portala, *zaprt*, da je viden samo tistim registriranim uporabnikom, ki lahko ta vir urejajo, in *v odpiranju*, ki kaže na to, da je skupina zaključila z delom in čaka na potrditev skrbnika terminoloških virov, da bo objavljen na javnem delu terminološkega portala.

Registrirani uporabnik se lahko odloči tudi za ustvarjanje novega slovarja z izbiro gumba desno zgoraj *Nov slovar*, pri čemer mora vpisati nekatere podatke. Vsa polja, ki zahtevajo obvezne podatke, so označena z zvezdico.

Sistemski podatki, potrebni tudi za dokumentiranost in izmenjavo metapodatkov z drugimi povezanimi mesti, so:

- **Naslov terminološkega vira**: Uporabnik mora vnesti celoten naslov terminološkega vira, kakor bo zabeležen v bibliografskih podatkih.
- **Angleški naslov terminološkega vira**: Celotni naslov terminološkega vira v angleščini smiselno dopolnjuje podatke pri izmenjavi informacij z drugimi terminološkimi portali, zato je predvideno, da ga vnese administrator terminološkega vira sam in se ne generira strojno.
- **Skrajšani naslov terminološkega vira:** Unikatno **kratko ime** terminološkega vira, ki lahko obsega največ 15 znakov, olajša prepoznavanje informacij na seznamu iskalnih zadetkov. Če uporabnik podatkov ne vpiše, se ti dodajo samodejno.
- **Avtor terminološkega vira:** Terminološki viri so verodostojnejši, če so znani avtorji vnosov, zato naj uporabnik doda vse avtorje z imenom in priimkom. Pri virih, ki nastajajo povsem na novo, so to administrator terminološkega vira in drugi dodani uporabniki s pravicami, pri dovoljeni predelavi že obstoječih virov, pa je treba navesti tudi avtorje izvirnika. Podatek sicer ni obvezen, poveča pa verodostojnost vira.
- **Področje:** Uporabnik mora iz spustnega seznama izbrati eno osnovno **področje,** na katero sodi terminološki vir, npr. *fizika*, *šport in igre*, *vzgoja in izobraževanje.*

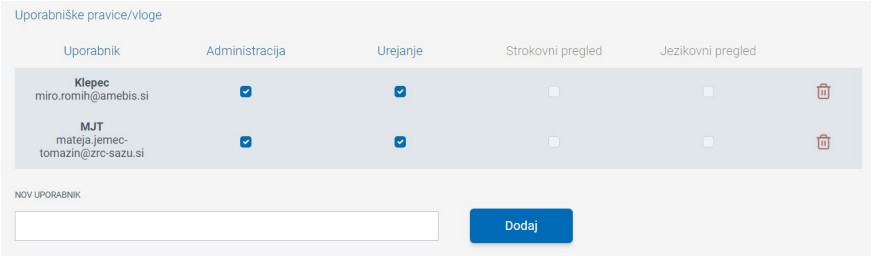
- **Podpodročja:** Uporabnik s spustnega seznama lahko izbere eno ali več ožjih **podpodročij,** ki bolj natančno določajo vsebino terminološkega vira, npr. *atletika*, *smučanje*, *alpinizem*, če je nadrejeno področje *šport*. Če uporabnik ustreznega podpodročja ne najde, ga lahko doda (vnesti je treba podatke tako v slovenščini kot angleščini), vendar mora novo podpodročje najprej potrditi administrator portala, ki lahko preveri, ali podatek že obstaja.
- **Struktura slovarskega sestavka:** Uporabnik v shemi z izbiro elementov določi, katere informacije bo vseboval terminološki vir.
- **Tuji jeziki:** Če je uporabnik v shemi določil, da bodo v terminološkem viru tudi termini v tujem jeziku, mora izbrati jezike v spustnem seznamu, sicer portal javlja napako *Niste vpisali tujih jezikov.*

Ob ustvarjanju slovarja se v Osnovnih podatkih odpre tudi polje, ki je namenjeno prostemu opisu vsebinske zasnove terminološkega vira in kjer naj bo na kratko opisan koncept in način izdelave terminološkega vira.

Dodano je tudi polje za vnos oznake ISSN, kar omogoča tudi povezavo z metrikami, ki se uporabljajo pri vrednotenju raziskovalnega dela.

Dodajanje sodelavcev terminološkega vira

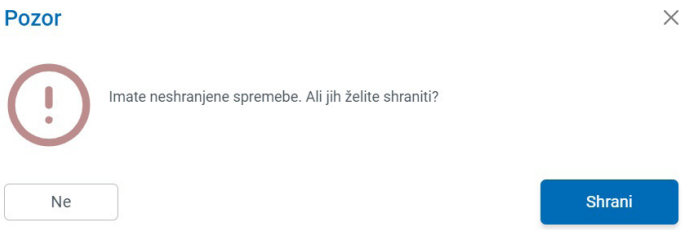
Pri izbiri posameznega terminološkega vira lahko uporabnik med *Lastnostmi* doda tudi druge registrirane uporabnike portala in jim dodeli pravice (Slika 16). Poznati mora njihovo uporabniško ime. Seznam uporabnikov drugim registriranim uporabnikom seveda ni dostopen, zato je predviden osebni stik med avtorji novega terminološkega vira, če pa bi posamezniki imeli tehnične ali vsebinske težave, se lahko po pomoč obrnejo tudi na administratorja portala. Če uporabniško ime ni pravilno vpisano, se izpiše obvestilo. Uporabniške vloge lahko dodaja tudi skrbnik terminoloških virov.



Slika 16: Uporabniške vloge v terminološkem viru.

Shranjevanje podatkov

Uporabniki morajo biti pozorni, da spremembe podatkov ves čas shranjujejo, sicer se lahko spremembe izgubijo. Vsak uporabnik je opozorjen, če želi zamenjati stran pred shranjevanjem (Slika 17).



Slika 17: Opozorilo o neshranjenih podatkih.

Področne oznake

Področne oznake so namenjene lažji organizaciji pojmov zlasti pri večjih terminoloških virih. Področne oznake so podatki, ki se razlikujejo med posameznimi terminološkimi viri. Smiselno je, da so te oznake čim krajše, navadno enobesedne ali dvobesedne, lahko so tudi v obliki krajšav. Znotraj terminološkega vira je smiselno imeti enak način poimenovanja oznak, npr. *agrarna geografija*, *demogeografija*, *klimatogeografija.* Odsvetovano je krajšanje oznak, zlasti če je okrajšava lahko enaka za več sorodnih terminov, npr. *fiz(ika)*, *fiz(iologija).*

Poleg podatkov, ki jih pripravijo uporabniki, so samodejno generirani naslednji podatki:

- datum in čas ustvarjanja terminološkega vira (zabeležen je tudi datum zadnje spremembe),
- ID terminološkega vira – enotna oznaka oblike »XX9999« (npr. »TP0012«), pri čemer je »XX« oznaka portala, »9999« pa zaporedna številka (ID) terminološkega vira na portalu.

6 Svetovanje

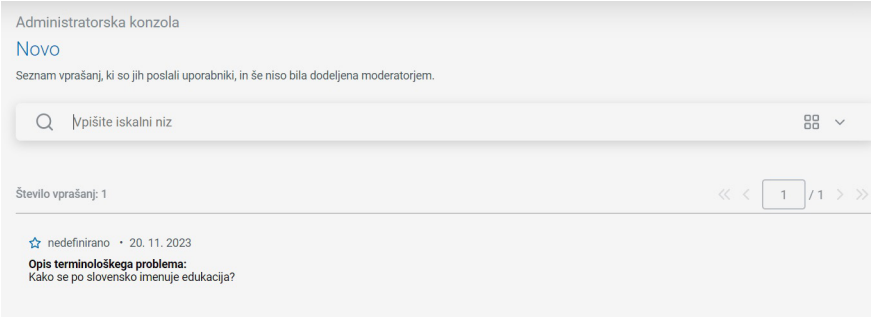
Iskanje po objavljenih terminoloških odgovorih, ki nastajajo v Terminološki svetovalnici Inštituta za slovenski jezik Frana Ramovša ZRC SAZU, je na voljo vsem uporabnikom. Nova terminološka vprašanja lahko zastavijo samo registrirani uporabniki. Svetovanje je namenjeno pomoči pri terminoloških dilemah na področjih, kjer še ni izdelanih terminoloških virov ali pa odgovora ne morete najti. Svetovanje na Slovenskem terminološkem portalu izvajajo sodelavci Terminološke sekcije Inštituta za slovenski jezik Frana Ramovša ZRC SAZU.

Na voljo je tudi samostojna namestitev modula za svetovanje na portalu GitHub.[12] V tem primeru je treba določiti člane svetovalnice, ki opravijo svetovanja, pri čemer je začetni postopek enak.

6.1 Pošiljanje vprašanj

Pri postavljanju vprašanj o najprimernejšem terminu so zelo pomembni posredovani podatki. Dovolj podatkov o terminološkem problemu omogoči svetovalcem, da svetujejo, katera rešitev bi bila najbolj ustrezna z vidika terminološke vede. Zato je zelo pomembno, da pri zastavljanju vprašanja v obrazec vnesete čim več podatkov o terminu, predvsem kaj ta pomeni, zelo koristni pa so tudi podatki o besedilih, v katerih se pojavlja, morebitne že obstoječe poimenovalne rešitve (ko več terminov označuje isti pojem), tujejezični ustrezniki itd. Na Slovenskem terminološkem portalu vprašanje dobijo sodelavci Terminološke sekcije ISJFR ZRC SAZU, pri samostojni namestitvi pa obvestilo prejme skrbnik svetovalnice (Slika 18).

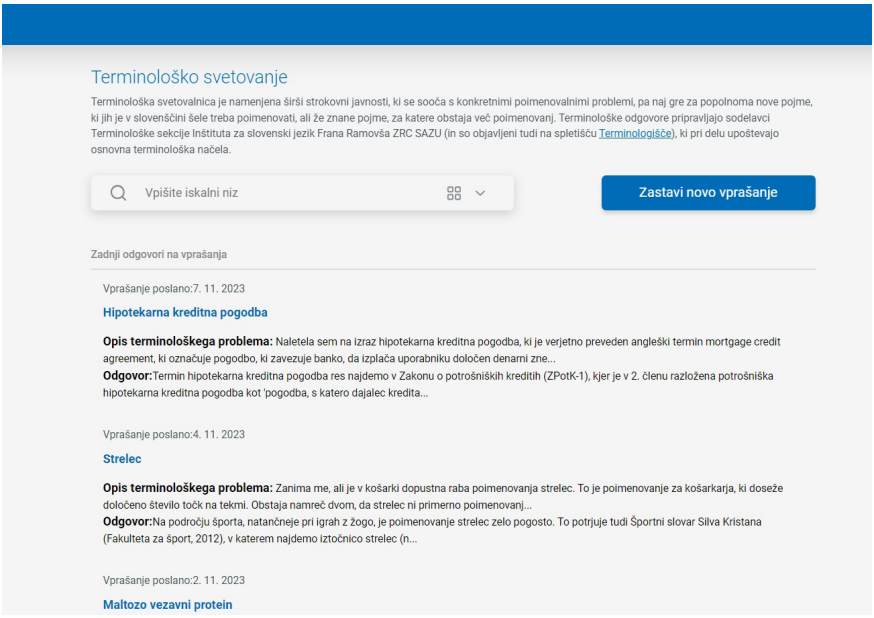
12 Na https://github.com/clarinsi/rsdo_term_portal.



Slika 18: Novo vprašanje v nabiralniku skrbnika svetovalnice.

6.2 Objava odgovorov

Odgovor na vprašanje je vedno objavljen v Terminološki svetovalnici Inštituta za slovenski jezik Frana Ramovša ZRC SAZU in prikazan tudi na Slovenskem terminološkem portalu (Slika 19). Na e-naslov, ki je naveden ob registraciji, uporabnik prejme odgovor, preden je objavljen.



Slika 19: Prikaz povzetkov odgovorov terminološkega svetovanja.

Pri samostojni namestitvi se lahko skrbnik svetovalnice odloči, da bodo odgovori objavljeni takoj, brez uporabnikove potrditve ali odziva, če je morda prišlo do komunikacijskega šuma.

Upravljanje svetovalnice je mogoče različno demokratično voditi, načeloma pa mora skrbnik svetovalnice potrditi odgovor pred objavo. Orodje je bilo razvito z namenom, da bi lahko sodelovali različni strokovnjaki tudi na daljavo, pri tem pa bi bil še vedno posameznik v vlogi skrbnika svetovalnice tisti, ki potrjuje končne odgovore.

7 Administracija

Portal je administriran in odkrite napake se odpravljajo v rednih časovnih intervalih.

7.1 Osnovne nastavitve

Administratorski del portala omogoča dodeljevanje pooblastil izbranim uporabnikom, ki jih administrator portala določi za administratorje terminoloških virov, skrbnike svetovalnice, lahko pa dodeljuje tudi vloge znotraj posameznih terminoloških virov, če administrator terminološkega vira zaprosi za pomoč.

7.2 Povezave s portali

Slovenski terminološki portal je javno dostopno spletno mesto. Programska koda je na voljo tudi na repozitoriju GitHub Clarin.si, zato je omogočena tudi lastna namestitev terminološkega portala. Če se organizacija odloči za to, potem lahko poveže svoj novi portal z drugimi terminološkimi portali, tudi s Slovenskim terminološkim portalom, in ob soglasju administratorjev teh portalov prikazuje tudi njihove terminološke vire. Pri tem mora vnesti podatke o povezanem portalu.

Dvočrkovne oznake portalov so specifične in unikatne za vsak posamezni portal. Med terminološkimi viri na povezanem terminološkem portalu lahko izberete vse terminološke vire, lahko pa samo nekatere, zlasti terminološke vire tistih področij, ki jih na novem terminološkem portalu ni.

Pri izdelavi terminološkega portala in njegovih vsebin so sodelovali:

- Vodja vsebinske zasnove: Mateja Jemec Tomazin, ZRC SAZU
- Vodja tehnične zasnove in izvedbe: Miro Romih, Amebis
- Oblikovalec: Samo Kramberger

Sodelujoče ustanove in sodelavci:

- **Amebis:** Miro Romih, Anton Romšak, Luka Romih, Jure Artiček, Miha Stele, Aljaž Grilc
- **Institut Jožef Stefan:** Senja Pollak, Hanh Thi Hong Tran, Vid Podpečan, Matej Martinc, Andraž Repar, Marko Pranjić, Nada Lavrač, Andraž Pelicon, Tomaž Erjavec
- **UL FF:** Špela Vintar
- **UL FRI:** Boštjan Slivnik, Danijel Skočaj, Marko Robnik Šikonja, Vladimir Batagelj (sicer UL FMF), Ivan Bratko, Peter Rogelj (sicer UP FAMNIT),
- **UL FU:** Polonca Kovač, Maja Klun, Jernej Podlipnik (odvetnik in predavatelj), Andreja Kostelec (samostojna raziskovalka in predavateljica), Nika Hudej (svetovalka US RS)
- **UM FERI:** Marko Ferme, Kristjan Žagar, Ivan Kovačič, Klemen Kac, Milan Ojsteršek, Damjan Strnad, Matjaž Divjak, Marko Bizjak, Matjaž Debevec, Ines Kožuh, Irena Lovrenčič Držanič
- **ZRC SAZU:** Mateja Jemec Tomazin, Simon Atelšek, Tanja Fajfar, Karmen Nemec, Jera Sitar, Mitja Trojar, Mojca Žagar Karer

Literatura

Fajfar, T., Žagar Karer, M. (2015). Pojmovni pristop k izdelavi terminološkega slovarja. V M. Smolej (ur.), *Slovnica in slovar – aktualni jezikovni opis* (209–216). Obdobja 34. https://centerslo.si/wp-content/uploads/2015/11/34_1-Fajfar-Zag-Kar.pdf

Hong Hanh, T., Martinc, M., Repar, A., Doucet, A., Pollak, S. (2022). A Transformer-based sequence-labeling approach to the Slovenian cross-domain automatic term extraction. V D. Fišer, T. Erjavec (ur.), *Zbornik konference Jezikovne tehnologije in digitalna humanistika* (196–204). https://nl.ijs.si/jtdh22/pdf/JTDH2022_Tran-et-al_A-Transformer-

based-Sequence-labeling-Approach-to-the-Slovenian-Cross-domain-Automatic-Term-Extraction.pdf

Jemec Tomazin, M., Žagar Karer, M. (2020). Slovenska terminološka spletišča in njihove zasnove. *Rasprave Instituta za hrvatski jezik i jezikoslovlje*, *46*(2), 693–716.Košmrlj-Levačič, B. (2007). *O terminih z vidika terminografske prakse*. V I. Orel (ur.), *Razvoj slovenskega strokovnega jezika* (583–598). Obdobja 24. https://centerslo.si/wp-content/uploads/2015/10/24-KosmrljLevacic.pdf

Žagar Karer, M., Fajfar, T. (2023). Terminological problems of terminology users: analysis of questions in terminological counselling service on the Terminologišče website. *Terminology*, *29*(1),78–102. https://doi.org/10.1075/term.21046.zag

Univerza *v Ljubljani*