KORPUS CVET 1.0: IZDELAVA, OPIS IN ANALIZA ZBIRKE STAREJŠIH BESEDIL V VERSKI PERIODIKI

Diana KOŠIR,[1] Tomaž ERJAVEC,[2]
[1]Inštitut za jezikoslovne študije ZRS Koper
[2]Odsek za tehnologije znanja, Institut "Jožef Stefan"

V prispevku je predstavljen proces izdelave in jezikoslovnega označevanja korpusa CVET 1.0, ki vsebuje besedila patra Hijacinta Repiča v starejšem slovenskem jeziku, objavljena v verskem glasilu *Cvetje z vertov sv. Frančiška* v obdobju 1881–1916. Besedila so bila v obliki PDF pridobljena s portala dLib, urejena v urejevalniku Word in nato pretvorjena v zapis TEI. Starejše besedje je bilo z odprtokodnim orodjem za normalizacijo avtomatsko posodobljeno, kar olajša iskanje po korpusu in nadaljnjo analizo gradiva. V članku so izpostavljene nekatere napake, ki so nastale pri posodabljanju in bodo v naslednji verziji korpusa ročno popravljene. Posodobljena besedila so bila nato še avtomatsko jezikoslovno označena z oblikoskladnjo in skladnjo po sistemu Universal Dependencies. Zapis TEI smo pretvorili v več izvedenih formatov in zbirko objavili pod odprto licenco na repozitoriju in konkordančnikih CLARIN.SI, ki so primerni za jezikoslovne analize gradiva. V drugem delu prispevka je prikazan primer analize avtorjevega pripovednega stila, opravljene s konkordančnikom noSketch Engine, ki temelji na frekvenčnih spremenljivkah najpogostejših in najmanj pogostih besed ter ključnih besed.

Ključne besede: starejša slovenščina, verski tisk, TEI, normalizacija, stilistična analiza, leksika

1 UVOD

Sredi 19. stoletja se je v slovenskem kulturnem prostoru postopoma vzpostavil periodični verski tisk z namenom poduka in verske vzgoje bralske publike. Izhajala so naslednja verska glasila: *Krščanski detoljub* (Ljubljana), *Angelček: otrokom učitelj in prijatelj* (Ljubljana), *Družinski prijatelj, poučno-zabavni list s podobami za slovenske družine* (Trst), *Drobtinice* (Gradec, Maribor, Ljubljana) in *Cvetje z vertov sv. Frančiška* (Gorica, Kamnik).

To je bilo tudi obdobje narodne prebuje s čitalniškim in taborskim gibanjem, prosvetno-kulturnimi in gospodarskimi društvenimi organizacijami, zato je

(predvsem) na obrobjih slovenskega kulturnega prostora slovenski tisk zaradi jezikovnega elementa imel tudi identifikacijsko in narodnopovezovalno vlogo (Perenič, 2012; Košir, 2022). Jezik lahko razumemo kot enega od ključnih gradnikov tako osebne kot nacionalne identitete, preko posameznikovega odnosa do (maternega) jezika pa se kaže njegova narodna zavest (prim. Nećak-Lük in sod., 1998, str. 77; Mikolič, 2000, str. 180–182).

Podobnega formata kot Slomškove *Drobtinice* in z naklado nekaj tisoč izvodov je leta 1880 na Goriškem začelo izhajati *Cvetje z vertov sv. Frančiška*. Tudi vsebinsko sta bili publikaciji sorodni – *Drobtinice* so vsebovale nabožna, življenjepisna, leposlovna, vzgojno-izobraževalna in občasno tudi poljudnostrokovna besedila (Ulčnik, 2010, str. 690), dočim je bilo v *Cvetju* leposlovja izrazito manj, poudarek je bil na življenjepisih svetnikov in nabožnih besedilih, na zgodovini reda in misijonih, platnice pa so bile posvečene jezikoslovju. Oba urednika, Slomšek in Škrabec, sta gojila poseben odnos do maternega jezika in pomembno prispevala k razvoju poenotene knjižne norme v 19. stoletju, kar se je podobno odražalo v njunih tiskovinah. Slomšek je bil jezikovno sicer bolj odprt in strpnejši pri počasi vpeljujoči se rabi l. 1851 sprejetih Svetčevih »novih oblik«, pa vendar je l. 1852 uredništvo v *Drobtinicah* objavilo, da bodo v primeru jezikovne neustreznosti prispevke popravili in dodelali; uredniška politika je kljub sprejemanju leksikalne različnosti kazala na težnjo po jezikovni enotnosti almanaha (Ulčnik, 2010, str. 687–688, 695–696). Tudi jezikoslovec Škrabec si je (nemara celo strožje) prizadeval za jezikovno izpiljenost besedil in poenoteno leksikalno rabo, »[z]ato je pa presedel dostikrat skoro po cele noči, da je prelil in predelal spise drugih po svojih zahtevah, zlasti pa, da je svoje lastne, dostikrat precej obširne sestavke, svojim nazorom primerno priredil« (Kunstelj v Korošak, 2001, str. 19). Za razliko od Slomškovega tiska je notranjost Škrabčevega *Cvetja*, z izjemo znamenitih jezikoslovnih razprav s platnic, še neraziskana.

Poleg urednika so besedila za revijo prispevali različni frančiškanski sobratje, navadno podpisani z inicialkami – tudi p. Hijacint (Anton) Repič (P. H. R.) iz koprskega samostana sv. Ane.

Projekt izdelave korpusa CVET 1.0 (Košir in Erjavec, 2024) z opusom objav patra Hijacinta Repiča v *Cvetju* je del soavtoričine doktorske raziskave, ki

odpira vprašanje prisotnosti slovenstva v slovenskem delu Istre pred 1. sv. vojno ter vloge, ki jo je imel samostan sv. Ane za jezikovno in kulturno utrjevanje Slovencev na Koprskem, s tem pa prinaša nova spoznanja na področju zgodovinske sociolingvistike, literarne vede, literarne pragmatike, uporabnega jezikoslovja in narodnega vprašanja. Metodološko raziskava sega še na področje digitalne humanistike z metodami korpusnega jezikoslovja, korpusne stilistike in računalniške stilometrije.

V nadaljevanju bo najprej predstavljen postopek priprave in označevanja korpusa CVET 1.0, nato pa bo prikazan praktični primer uporabe konkordančnikov, ki so dostopni v okviru slovenske raziskovalne infrastrukture CLARIN.SI za stilistično korpusno analizo Repičevega pripovednega sloga. Osebni slog odseva avtorjevo osebnost oz. jezikovne navade (Leech in Short, 2007, str. 23), tudi preference, saj gre za subjektivno motivirane posamične izbire iz celotnega jezikovnega repertoarja, ki se zgodijo glede na piščev namen, *kako* želi predati določeno sporočilo bralcu (Biber in Conrad, 2009, str. 144).

2 PRIPRAVA IN OBJAVA GRADIVA

Verska revija *Cvetje z vertov sv. Frančiška* je izhajala med letoma 1880 in 1944. Od začetka do leta 1915 jo je urejal eden od njenih pobudnikov p. Škrabec, nasledil ga je p. Evstahij Brlec. Goriška knjižnica Franceta Bevka je v sodelovanju s Škrabčevo knjižnico Frančiškanskega samostana Kostanjevica v Novi Gorici v letih 2017 in 2018 izpeljala projekt digitalizacije revije, ki je v formatih PDF in TXT dostopna na portalu Digitalna knjižnica Slovenije (dLib.si).

Prvi pregled Repičevih člankov v *Cvetju z vertov sv. Frančiška* je pripravil p. Bruno Korošak (2006). Celoten fond do l. 1918 (Repičeva smrt) je bil ponovno pregledan in dopolnjen. Ugotovljeno je bilo, da je p. Repič za revijo med letoma 1881 in 1916 prispeval 140 člankov.

2.1 Priprava gradiva in metapodatkov

V prvi fazi priprave gradiva smo iz dLib izbrali posamezne številke revije, v katerih se pojavljajo članki p. Repiča, in te članke iz PDF prekopirali v

urejevalnik Word. Ker so nekateri članki izšli po delih v več številkah revije, je skupno število (delnih) člankov, in s tem tudi datotek Word, 230.

V urejevalniku Word smo nato članke oblikovno poenotili in ročno popravili. Ti popravki so zajeli poenotenje zapisa ločil glede na sedanjo pravopisno normo (levostičnost končnih in nekaterih nekončnih ločil; prost zapis večpičja v tropičje; poenotena oblika narekovajev) in odpravljanje napak, ki so nastale pri OCR (tj. pretvorbi faksimila v besedilo) v datotekah PDF. Izvirno smo ohranili nekatere posebnosti stave ločil (npr. zapis zadnjega narekovaja pred/za končnim ločilom), zapis naglasnih znamenj (npr. pomensko razlikovalna raba krajše oblike povratnega osebnega zaimka “se” in predloga “sè”) in delitev na odstavke.

Vzporedno z urejanjem datotek v programu Word smo za vsako besedilo vnesli njegove metapodatke v razpredelnico Excel, in sicer: ime datoteke Word (ki je obenem identifikator besedila), avtor (vedno sicer p. Repič, vendar mestoma tudi z zaznamkom, po katerem avtorju je delo povzeto oz. prevod dela katerega avtorja je), naslov članka, mesto objave (leto, letnik, številka), URL izvornega mesta objave v dLib in na katerih straneh se članek pojavi.

2.2 Zapis TEI

V naslednjem koraku smo gradivo iz formatov Word in Excel pretvorili v zapis, ki je bolj primeren za hrambo kot tudi za nadaljnje pretvorbe v formate, namenjene posameznim orodjem, in sicer v XML shemo skladno s priporočili iniciative za kodiranje besedil TEI (TEI Consortium, 2020), bolj natančno, v shemo, ki jo raziskovalna infrastruktura CLARIN.SI priporoča za korpuse, deponirane v repozitoriju infrastrukture.[1]

Datoteke Word smo najprej pretvorili v osnovni TEI, in sicer z uporabo standardnih skript XSLT za pretvorbo v in iz dokumentov TEI,[2] medtem ko smo razpredelnico Excel z metapodatki o posameznih besedilih shranili kot datoteko TSV, ki je primerna za nadaljnje avtomatske obdelave. Nato smo z namensko skripto združili vsako besedilno datoteko TEI z njenimi

[1] https://github.com/clarinsi/TEI-schema
[2] https://github.com/TEIC/Stylesheets

metapodatki in jo formirali v eno datoteko TEI (vrhnji element <TEI>), ki vsebuje kolofon (element <teiHeader>) in besedilo (<text>). Začetek ene od teh datotek ilustrira slika 1.

Slika 1: Primer začetka zapisa posameznega besedila v formatu TEI.

```
<TEI xmlns="http://www.tei-c.org/ns/1.0" xml:id="CVET-1887_7_2_43-46" xml:lang="sl">
 <teiHeader>
   <fileDesc>
     <titleStmt>
       <title>Vzroki in koristi terpljenja pobožnih duš. (1887) [CVET]</title>
       <author>Repič, Hijacint</author>
     </titleStmt>
     <editionStmt><edition>1.0</edition></editionStmt>
     <extent><measure unit="words">1203</measure></extent>
     <publicationStmt>
       <distributor>CLARIN.SI</distributor>
       <idno type="handle">http://hdl.handle.net/11356/1226</idno>
       <date when="2024-05-07"> 7. maj 2024</date>
     </publicationStmt>
     <sourceDesc>
       <bibl type="article">
         <author>Repič, Hijacint</author>
         <title level="a">Vzroki in koristi terpljenja pobožnih duš.</title>
         <bibl type="monogr">
           <title level="j">Cvetje z vertov sv. Frančiška</title>
           <biblScope unit="volume">7</biblScope>
           <biblScope unit="number">2</biblScope>
           <biblScope unit="page" from="11" to="14">11-14</biblScope>
           <date when="1887">1887</date>
           <idno type="URI" subtype="URN">https://www.dlib.si/.../PDF</idno>
         </bibl>
       </bibl>
     </sourceDesc>
```

Celoten korpus je formiran kot en dokument XML s krovno datoteko (element <teiCorpus>), ki vsebuje kolofon korpusa in povezave (elementi XInclude) na besedilne datoteke korpusa. Kolofon korpusa poda korpusne metapodatke, tj. naslov, odgovorne osebe, velikost, založnika, distributerja, licenco, skupni opis vira in projekta ter seznam uporabljenih jezikov (slovenščina, za metapodatke tudi angleščina).

2.3 Posodabljanje besed

Korpus vsebuje dela, ki so nastala proti koncu devetnajstega in v začetku dvajsetega stoletja, ko so se marsikatere besede pisale drugače, kot pa je to v sodobni normi. Arhaične oblike besed otežijo iskanje po korpusu, odvisno od

raziskave pa tudi njihovo nadaljnjo analizo. Poleg tega bi jezikoslovno označevanje zbirke izvornih besedil imelo dosti napak, saj orodja, ki so na voljo za tako označevanje, delujejo dobro le na sodobni standardni slovenščini.

Zaradi teh razlogov smo besede v korpusu najprej avtomatsko posodobili, za kar smo uporabili odprtokodno orodje za normalizacijo cSMTiser[3] (Scherrer in Ljubešić, 2016), ki temelji na principu statističnega strojnega prevajanja in orodju Moses (Koehn, 2010). cSMTiser smo naučili posodabljanja na gajičnem delu ročno posodobljenega korpusa slovenščine goo300k (Erjavec, 2016), podobno, kot smo že pred tem naredili za posodabljanje zbirke slovenskih romanov v okviru korpusa ELTeC (Schöch in sod., 2021), za korpus starejše slovenske proze PriLit (Žejn in Erjavec, 2021) ali za zbirko slovenskih pregovorov (Babič in Erjavec, 2022). S tako izšolanim orodjem smo nato posodobili besede vseh besedil v korpusu. Pri tem velja opomba, da izvorne besedne oblike s tem niso izgubljene, pač pa so posodobljene oblike, kjer se razlikujejo od izvirnih, pripisane le-tem.

V postopku avtomatskega posodabljanja starejšega jezika so bile denimo ustrezno posodobljene besede, ki imajo polglasnik zapisan z »e« (npr. *miloserčnost* > *milosrčnost, serce* > *srce*, *smert* > *smrt*), končnica -vec > -lec (*bravec* > *bralec*), premena -t- > -d- v korenu (npr. *britkosti* > *bridkosti*), -z- > -g- v korenu (npr. *druzega*, *vbozih* > *drugega*, *ubogih*), izpad -i- v *stariši* > *starši* (mn.) in *kristijani* > *kristjani*, izpad -t- v *bogatstvo* > *bogastvo*, zapis v- > u- pri nekaterih glagolih (npr. *vmreti*, *vmerl* > *umreti*, *umrl*), obliko *apostelјnov* > *apostolov*, in nekatere funkcijske besede: zaimek *gdo* > *kdo*, veznik *temuč* > *temveč*, veznik in členek *toraj*/*torej* > *torej*, členek *vže* > *že*, predlog *sė* > *s*, *ž* > *z*. Nekatere starejše oblike besed so ostale neposodobljene (npr. predlog *mej* > med, *blager* > blagor). Po drugi strani pa dela avtomatsko posodabljanje tudi napake. Določene besede so bile posodobljene v sicer obstoječo, a neustrezno besedo (npr. *čiger* > *čigra* nam. *čigar*), po nepotrebnem posodobljene v napačen zapis (npr. *vbogajme* > *ubogajme*) oz. nove skovanke (npr. *keterim* > *ketim* nam. *katerim*). Spodnja tabela 1 prikazuje nekatere tovrstne napake, ki smo jih identificirali v korpusu, seveda pa je takih napak po vsej verjetnosti še več.

[3]https://github.com/clarinsi/csmtiser

Tabela 1: Nekatere prezrte starejše oblike in napake, ki so nastale pri avtomatskem posodabljanju leksike.

Izvorna beseda	*Posodobljena beseda*	*Korigirana posodobljena beseda*
čiger	čiger (rod. mn. od ‘čigra’)	čigar
sobrat	zbrat	sobrat
keterim	ketim	katerim
marsiketerim	marsiketim	marsikaterim
neketerim	neketim	nekaterim
kesneje	kosno	kasneje
mariveč	mareveč	marveč
gvardijan	gvardian	gvardijan
vmerje	umerje	umrje
(v) tempeljnu	(v) tempelju	(v) templju
(pred) vratmi	(pred) vratami	(pred) vrati
vbogajme	ubogajme	vbogajme
tolikanj	tolikanj	toliko / tako
(s) tebo	(s) tebo	(s) tabo / teboj
obena	obena	nobena
česer	česer	česar
mej	mej	med
ke	ke	ko
jeli	jeli	jeli = začeli; je li = ali je
namestu	namestu	namesto
ničemernost	ničemernost	nečimrnost
apostelj	apostelj	apostol
kolikanj	kolikanj	koliko
odpuščenje	odpuščenje	odpuščanje
razodevenje	razodevenje	razodetje
vsegamogočni	vsegamogočni	vsemogočni
ražalil	ražalil	razžalil
marcija	marcija	marca
preskerbel	preskrbel	preskrbel/priskrbel/poskrbel (različen pomen)
verni	vrni	verni/vrni (različen pomen)

2.4 Jezikoslovno označevanje

Na osnovi avtomatsko posodobljenih besed smo nato korpus jezikoslovno označili. Tu smo uporabili odprtokodno orodje CLASSLA[4] (Ljubešić in Dobrovoljc, 2019), s katerim smo dodali naslednje jezikoslovne oznake v besedilo, npr. za besedno obliko "*bi*":

- lemo oz. osnovno obliko besede, tu "biti";
- oblikoskladenjsko oznako po priporočilih MULTEXT-East (Erjavec, 2012), »Va-c«, ki se razveže v oblikoskladenjske lastnosti "Verb Type=auxiliary VForm=conditional", pri čemer obstaja tudi ekvivalentna oznaka v slovenščini, tu »Gp-g« in njena razvezava v "glagol vrsta=pomožni oblika=pogojnik";
- oblikoskladenjske lastnosti po sistemu Universal Dependencies za slovenski jezik (Dobrovoljc et al., 2017), tu "UPosTag=AUX Mood=Cnd VerbForm=Fin". Te oznake so sicer podobne oznakam MULTEXT-East, vendar z drugače definiranim naborom lastnosti in vrednosti, občasno se pa od njih tudi sistemsko razlikujejo, kot je tudi primer za "bi";
- odvisnostna skladenjska razčlenitev povedi po sistemu Universal Dependencies za slovenski jezik.

Posodobljena in jezikoslovno označena različica korpusa je bila formirana kot svoj korpus; format označenega besedila je ilustriran v sliki 2.

[4]https://github.com/clarinsi/classla

Slika 2: Primer zapisa TEI označene povedi v korpusu.

```
<s xml:id="CVET-1887_7_2_43-46.1.s1">
   <w xml:id="CVET-1887_7_2_43-46.1.s1.t1" ana="mte:Ncmpn"
      msd="UPosTag=NOUN|Case=Nom|Gender=Masc|Number=Plur" lemma="vzrok">Vzroki</w>
   <w xml:id="CVET-1887_7_2_43-46.1.s1.t2" ana="mte:Cc" msd="UPosTag=CCONJ" lemma="in"
      >in</w>
   <w xml:id="CVET-1887_7_2_43-46.1.s1.t3" ana="mte:Ncfpn"
      msd="UPosTag=NOUN|Case=Nom|Gender=Fem|Number=Plur" lemma="korist">koristi</w>
   <w xml:id="CVET-1887_7_2_43-46.1.s1.t4" ana="mte:Ncnsg"
      msd="UPosTag=NOUN|Case=Gen|Gender=Neut|Number=Sing" norm="trpljenja"
      lemma="trpljenje">terpljenja</w>
   <w xml:id="CVET-1887_7_2_43-46.1.s1.t5" ana="mte:Agpfpg"
      msd="UPosTag=ADJ|Case=Gen|Degree=Pos|Gender=Fem|Number=Plur" lemma="pobožen"
      >pobožnih</w>
   <w join="right" xml:id="CVET-1887_7_2_43-46.1.s1.t6" ana="mte:Ncfpg"
      msd="UPosTag=NOUN|Case=Gen|Gender=Fem|Number=Plur" lemma="duša">duš</w>
   <pc xml:id="CVET-1887_7_2_43-46.1.s1.t7" ana="mte:Z" msd="UPosTag=PUNCT">.</pc>
   <linkGrp corresp="#CVET-1887_7_2_43-46.1.s1" targFunc="head argument" type="UD-SYN">
      <link ana="ud-syn:root"
         target="#CVET-1887_7_2_43-46.1.s1 #CVET-1887_7_2_43-46.1.s1.t1"/>
      <link ana="ud-syn:cc"
         target="#CVET-1887_7_2_43-46.1.s1.t3 #CVET-1887_7_2_43-46.1.s1.t2"/>
      <link ana="ud-syn:conj"
         target="#CVET-1887_7_2_43-46.1.s1.t1 #CVET-1887_7_2_43-46.1.s1.t3"/>
      <link ana="ud-syn:nmod"
         target="#CVET-1887_7_2_43-46.1.s1.t3 #CVET-1887_7_2_43-46.1.s1.t4"/>
```

Kot je razvidno iz primera, so povedi označene z elementom <s>, besede z <w> ter ločila s <pc>. Iztočnična oblika besede je podana kot vrednost atributa @lemma, oznake MULTEXT-East kot vrednosti atributa @ana, lastnosti Universal Dependencies kot @msd, posodobljena oblika besede pa kot vrednost @norm. Skladenjska analiza je podana v elementu <linkGrp>, in sicer za vsako skladenjsko odvisnost (element <link>) kot povezava med dvema pojavnicama (prek sklica na njuna identifikatorja v atributu @target), pri čemer je oznaka skladenjske odvisnosti podana kot vrednost atributa @ana.

V različici korpusa, ki vsebuje posodobljena in jezikovno označena besedila, je dopolnjen tudi kolofon s taksonomijo skladenjskih oznak Universal Dependencies in z opisom uporabljenih orodij.

2.5 Objava in velikost korpusa

Različico 1.0 izdelanega korpusa, poimenovanega “CVET”, smo objavili v repozitoriju CLARIN.SI (Košir in Erjavec, 2024) pod odprto licenco Creative

Commons, priznanje avtorstva (CC BY).

Korpus je za prevzem na voljo v štirih stisnjenih datotekah. Vsaka vsebuje direktorij, v njem pa datoteke za eno od variant korpusa:

- Cvet.TEI: korpus v zapisu TEI, kot je predstavljen v razdelku 2.2. Poleg korenske XML datoteke korpusa vsebuje še 230 XML besedil z metapodatki in direktorij s shemo CLARIN.SI TEI;
- Cvet.TEI.ana: korpus z jezikoslovno označenimi besedili;
- Cvet.txt: korpus brez oznak (navadno besedilo), avtomatsko pretvorjen iz TEI.ana, in razpredelnica z metapodatki posameznih besedil; format je primeren za neposreden uvoz v razna orodja. Za namene raznovrstnih raziskav je ta varianta na voljo v kombinacijah sledečih različic:
    - izvorno besedilo (npr.: "*Ti pa nikaker ne poslušaj hudobe,*"); posodobljeno besedilo ("Ti pa nikakor ne poslušaj hudobe,"); lematizirano besedilo ("ti pa nikakor ne poslušati hudoba,");
    - besedilo z izvorno kapitalizacijo (kot zgoraj);
    - besedilo v malih črkah ("ti pa nikaker ne poslušaj hudobe,");
    - besedilo z izvorno stičnostjo in besedilo razdeljeno po pojavnicah ("Ti pa nikaker ne poslušaj hudobe ,");
- korpus v t. i. vertikalnem formatu, ki je primeren za uvoz v spletne konkordačnike; dodana je tudi konfiguracijska datoteka korpusa, kot se uporablja v konkordančnikih CLARIN.SI.

Repozitorijski vnos je povezan s konkordančniki CLARIN.SI, in sicer noSketch Engine in KonText, kar omogoča poizvedbe in analize korpusa brez znanja programiranja in z različnimi vizualizacijami in možnostjo shranjevanja rezultatov poizvedb.

Korpus vsebuje 230 besedil oz. 3.228 odstavkov in 10.109 (avtomatsko določenih) povedi. Pojavnic je skupno 212.703, od tega 176.700 besed (če štejemo vse pojavnice, ki niso označene kot ločilo) oz. 175.907 (kot besede prešteje konkordančnik, ki npr. cifer ne šteje kot besede). Avtomatsko je bilo posodobljenih 13.033 oz. 7,4 % besed.

3 STILISTIČNA ANALIZA GRADIVA

Stilistika je opredeljena kot jezikoslovno preučevanje sloga oz. namenske specifične rabe jezika, ki nakazuje odnos med kreativnim dosežkom (učinkom) in jezikovno manifestacijo (Leech in Short, 2007, str. 11, 55), na kar lahko vplivajo nejezikovne spremenljivke, kot so žanr, avtor, zgodovinsko obdobje ipd. (Jeffries in McIntyre, 2010, str. 1). Korpusna stilistika je razumljena kot aplikacija teorij, modelov in metod stilistike v analizi korpusa, pri čemer je poudarek na razločevanju vzorcev v jezikovni rabi s preučevanjem velikih količin jezikovnih podatkov (McIntyre, 2015, str. 61).

Pri stilistični analizi opazujemo sestavine sloga (angl. *features of style*, Leech in Short, 2007), ki so jezikovni elementi, ki v danem korpusu izstopajo – glede na ozadje pričakovanj bralca pojavnost nekega jezikovnega elementa povzroči presenečenje (notranji odklon, angl. *internal deviation*), in pomembno prispevajo k oblikovanju sloga. Te jezikovne entitete imenujemo slogovni označevalci (angl. *style markers*), ki so v določenih kontekstih najbolj ali najmanj frekvenčni (Enkvist, 1964, str. 34 v Leech in Short, 2007, str. 59). Te sestavine je pred slogovno analizo treba izbrati, kriterije za kvantitativno analizo pa po Leechu in Shortu določa preplet treh konceptov: devianca (angl. *deviance*), prominenca (angl. *prominence*) in književna relevanca (angl. *literary relevance*). Devianca je statistični pojem, povezan s kvantifikacijo, ki kaže na odklon frekvence rabe določenega jezikovnega (ali slogovnega) elementa od normalne frekvence; prominenca je psihološki pojem, intuitivni ekvivalent devianci, in je odvisen od bralčevega odziva na besedilo, ta pa je pogojen z vrsto dejavnikov (občutek za slog, bralne izkušnje, razpoloženje, koncentracija itd.), zato se zdi prominenca bolj subjektivna, čeravno je tudi tu možno izmeriti frekvenco. Književna relevanca je opredeljena kot »umetniško motiviran odmik od relativne norme« (angl. *foregrounding*), ki je kvalitativen (kreativna raba jezika v nasprotju s konvencionalno rabo) ali kvantitativen (devianca vsled nepričakovane frekvenčnosti) (Onič, 2014, str. 184). Pri naši analizi bomo opazovali posamezne jezikovne prvine, ki jih Leech in Short (2007, str. 61–63) uvrščata v skupini leksikalnih in slovničnih kategorij.

V digitalni humanistiki se je kot kvantitativna in statistična analiza avtorjevega stila uveljavila stilometrija, primerjalna metoda, ki »meri« podobnosti in

razlike med besedili na različnih jezikovnih ravninah (Žejn, 2020). Pri stilističnih in stilometričnih analizah gre najpogosteje za analize na leksikalni ravni, pri čemer so med najpogostejšimi preučevanimi kategorijami (spremenljivkami) naslednje: najpogostejših 100 besed, distribucija pogostosti besedišča, raznolikost (gostota) besedišča in *hapax legomena* oz. *hapax dislegomena* (izrazi, ki se v danem kontekstu pojavijo samo enkrat oz. dvakrat), povprečna dolžina besed oz. povedi, besedni in črkovni n-grami, pogostost najpogostejših besednih kolokacij (zaporedje dveh, treh, štirih itd. besed), najpogostejših funkcijskih besed (zaimki, pomožni glagoli, predlogi, vezniki, členki, medmeti), ki same navadno nimajo leksikalnega pomena, izražajo pa slovnični odnos z drugimi besedami v stavku oz. odnos ali razpoloženje govorca; ritmično zaporedje naglašenih in nenaglašenih zlogov, razporeditev ločil idr. (Mosteller in Wallace, 1964; Hoover, 2002, 2003; Grieve, 2007; Eder, 2011; Žejn, 2020 idr.).

Za našo kvalitativno in kvantitativno analizo smo izbrali tri kategorije (spremenljivke), povezane s frekvenčnostjo: najpogostejših 100 besed, najmanj pogoste besede in ključne besede (v vseh treh primerih bodo iskane leme, v oklepaju bodo z oznako izv. pripisane izvorne oblike besed, če se te razlikujejo od lem). Orodje noSketchEngine omogoča funkcijo iskanja ključnih besed (*keywords*), ki po frekvenci rabe odstopajo navzgor ali navzdol glede na referenčni korpus, seznam pa zajame polnopomenske in nepolnopomenske (funkcijske) besede. Primerjali bomo seznam ključnih besed glede na izbrana referenčna korpusa starejšega gradiva (IMP 1.1, Erjavec, 2014; PriLit 1.0, Žejn in Erjavec, 2021). Pogostost rabe poda vpogled v gostoto besedišča in vsebinsko označi kontekst(e) besedil v korpusni zbirki, ključne besede pa razodevajo avtorjevo pomensko polje in skozi opazovanje njihovih pomenov in kontekstov rabe razkrivajo piščev spoznavni in literarni svet (prim. Mikolič, 2020, 2022). Pri interpretaciji dobljenih kvantitativnih podatkov je treba upoštevati tudi subjektivni interpretativni okvir raziskovalca.

3.1 Analiza avtorjevega pripovednega sloga z uporabo konkordančnikov

Med najpogostejšimi 100 besedami (lemami) v korpusu CVET so: a) glagoli *biti*, *imeti*, *moči*, *hoteti*, *reči*, *morati*, *priti*, *storiti*, *prositi*, *govoriti*, *videti*, *delati*,

vedeti, *ljubiti*; b) samostalniki *Bog*, *svet*, *duša*, *človek*, *brat*, *Frančišek*, *gospod*, *dan*, *življenje*, *greh*, *srce* (izv. *serce*), *oče*, *volja*, *beseda*, *Jezus*, *milost*, *otrok*, *Kristus*, *križ*, *ljubezen*, *mati*, samostalniški zaimki *jaz*, *ti*; c) pridevniki *božji*, *dober*, *velik*, svojilni zaimki *moj*, *tvoj*, *svoj*, *njegov*, *naš*; č) prislovi *jako*, *dobro*; d) vezniki *da*, *in*, *ter*, *v*, *ako* oziralni in vprašalni zaimki v vlogi veznika *kateri* (izv. *keteri*), *kakor* (izv. *kaker*), *kar*, *kaj*, *kako*; e) členki *tudi*, *naj*, *še*, *ne*, *le*, *samo*. Lema *nič* se pojavi kot samostalnik (*Oh saj gre vse sčasoma v nič!*), sam. zaimek (*Zato se ne veseli v ničemer drugem, kar je pod nebom!*) in prislov (*Keder se nič več ne želi, takrat se čuti pravo, popolno veselje*), zaimek *vsak* v pridevniški (*Priporočeval se jima je serčno vsako jutro in vsak večer*) in samostalniški rabi (*Ako je vsacemu dovoljeno storiti si iz sukna tako ali tako obleko, bo li Bog imel menj pravic do svojega?*).

Med glagoli se poleg tistih, ki so pričakovani glede na vsakdanje sporazumevalne okoliščine, pogosto pojavijo glagoli rekanja (*reči*, *prositi*, *govoriti*, nadalje tudi *praviti*, *odgovoriti*, *povedati*) in modalni glagoli (prvo število so pojavitve, drugo pa % besedil v korpusu, kjer se lema pojavi): *moči* (597 = 75 %), *hoteti* (577 = 70,87 %) in *morati* (465 = 66,52 %). Očitna prisotnost naklonskih glagolov zasluži natančnejšo analizo kontekstov rabe. Primeri, kot so *Naposled, ker brez milosti božje ne moreš nič dobrega storiti, prosi Boga, da ti bode milostljiv / Ali Bog te hoče poskušati sè skušnjavami, je li res, da ga ljubiš ali ni / Nadalje se moramo vdati božji volji, ako imamo dušne ali telesne pogreške*, sodijo v žanrski okvir nabožnega vzgojnopoučnega članka. Analizo bi na tem mestu lahko nadgradili in poglobili z opazovanjem a) glagolskega naklona (razmerje med rabo povednega, velelnega in pogojnega naklona), kar bi denimo služilo opazovanju razmerja med prevladujočo sporočevalno oz. vplivanjsko vlogo besedil, in b) glagolske osebe in števila (1. os. mn., 2. os. ed. ali mn., 3. os. ed. ali mn.), pri čemer bi opazovali frekvenco posrednega in neposrednega nagovarjanja naslovnika (bralca).

Religiozni diskurz opredeljuje moralno-vrednostno razmerje dobro : slabo in med najpogostejšimi besedami so predvsem te, ki semantično sodijo v vrednostno oznako »dobro« (*ljubiti*, *Bog*, *duša*, *srce*, *milost*, *ljubezen*, *dober*, *dobro*).

Večina od prvih 100 besed se pojavi v več kot 60 % vseh besedil, kar pomeni,

da je njihova zastopanost razpršena skozi celoten korpus.

Po pogostosti izstopa raba zaimka *kateri*, ki se pojavi kar 1.822-krat in v 90,87 % besedil. Izrazita je vezniška raba, kjer uvaja podredje z oziralnim odvisnikom, v današnji knjižni slovenščini bi ga nadomestil zaimek *ki* (*Ali ti, moj Jezus, keteremu je vse odkrito, in keteri si vstvaril vse v meri in številu …*; *On ne ve, da so britkosti božje šibe polne ljubezni, da so mile kazni, pripravljene edino mojim izvoljenim, ketere čistim v ognju britkosti, kaker se čisti v goreči peči zlato*). Tudi sicer lahko ob hitrem pregledu gradiva ugotovimo, da so v Repičevih tekstih pogoste dolge povedi s podredji, kar bi bilo vredno natančnejše statistične analize.

Med načinovnimi prislovi po frekvenčnosti izstopa *jako* (342-krat; ob 11 pojavitvah sopomenke *zelo*; npr. *To prepričanje nas jako tolaži, ker vemo, da nas Bog more tako lehko poklicati k sebi, ko smo v sreči in zdravju, kaker ko smo v občnih nesrečah*). Glede na konkordance in metapodatke o izvirnih besedilih, ki nam jih posreduje korpus, lahko ugotovimo, da se starejša oblika ‘jako’ pojavlja skozi celotno gradivo (v 57,83 % besedil), novejša oblika ‘zelo’ pa ob njej nekoliko bolj konsistentno po letu 1913. Ker se prislov *jako*, kot bomo videli v nadaljevanju, pojavi tudi med ključnimi besedami glede na referenčni korpus starejših besedil, ga lahko označimo kot značilno potezo patrovega pripovednega sloga.

Med najmanj pogostimi besedami (ena do dve pojavitvi, korpus takih lem navaja 4.656) so: nepričakovane tvorjenke: manjšalnice (*človeček*, *stvarca*), *kletvina* (ob *kletev*), *pekočina*, *bogatin* (ob *bogataš* in posamostaljenem pridevniku *bogati* (mn.)), *tolažilo* (ob *tolažba*), *zaveržek*, *pomoček*, kroatizem *škatulje* (mn.), *nasladnost* (ob *naslada*), *ostrašen* (ob *prestrašen*), modifikacija glagolskega vida (*peljavati*, *zmaščevati*, *oveseliti*), onomatopoetični glagol *zberbrati* (= na hitro, nerazločno oz. nepremišljeno izreči, zmoliti); deležnik na -č: *kretajoč (se)*, *pazeč*, deležnik na -ši: *znebivši*, *zbudivši*; členek *znabiti*, besedna zveza *čudapoln* mir, *prvak* apostolov, *zvezdoznanka*, *zgoja* (ob *vzgoja*), *zgojitelj*, *razjasnjenje*, *život* (ob *telo*), *nesposobnost*, *neizmernost*, *gotovost*, *enakoličnost* itd. Nepričakovane oz. z vidika rabe redke dvojnice lahko razumemo kot odraz tedaj še nepoenotene pisne norme (npr. *vzgoja* ali *zgoja*), nekatere glagolske oblike in izvirne tvorjenke, ki lahko glede na dani

kontekst okrepijo pomen, pa kažejo na bogato besedišče in piščevo ustvarjalno rabo jezika. Skozi uporabo ekspresivne, slogovno zaznamovane leksike se pisec hkrati lahko čustveno razodeva, izraža naklonjenost oz. nenaklonjenost.

Ključne besede v fokusnem korpusu CVET so bile izluščene skozi primerjavo z izbranim korpusom starejše pripovedne proze PriLit (Žejn in Erjavec, 2021). Med 100 ključnimi besedami se za korpus CVET pojavijo naslednje leme, ki smo jih zaradi večje preglednosti razdelili na a) leksikalne in b) funkcijske besede:
a) samostalniki: *Frančišek*, *Monald* (izv. tudi *Monaljd*), zemljepisna imena *Asiz* (= Assisi), *Koper*, *Padova*; *redovnik*, *frančiškan*, *sobrat*, *tretjerednik*, *zavetnica*, *voditelj*, *častilec* (izv. *častivec*), *posvetnjak*, *vodilo*, *miloščina*, *milosrčnost* (izv. *miloserčnost*), *sočutje*, *blager* (= blagor; 3-krat samostalniška raba v pomenu sreča, blagoslov), *način* (prim. *na ta(k)/en/pervi/drugi/vsak način*), *sredstvo*, *oblika* (prim. *živeti po obliki sv. evangelija* = živeti skladno z nauki in zgledi iz evangelija), *kesanje*, *naslada*, *pogrešek*, *vdanost*, izglagolska samostalnika *zatajevanje*, *občevanje*; pridevniki: *redoven*, *blažen*, *brezmadežen*, *blagoslovljen*, *presladek*, *izreden*, *nepopisljiv*, (Marija) *Porcijunkuljska* (= nanaša se na cerkev Marije Angelske pri Assisiju, t. i. Porcijunkula), *vsakovrsten* (izv. *vsakoversten*), *brezštevilen*; glagoli: *pridigati*, *oznanjevati*, *občevati* (= biti v stiku, sporazumevati se), *spolnjevati* (redko *izpolnjevati*), *spodbujati*, *rabiti*, *zanemarjati*, *zoperstavljati se*, *vničiti*; prislovi: *jako*, *kesneje** (= kasneje), *naposled*;

b) vezniki: *koliker* (= kolikor), *čiger** (= čigar),[5] *keteri** (= kateri), *mariveč** (= marveč); členki: *nikaker* (= nikakor), *seveda* (pisano tudi *se ve da*), *potemtakem*, *blager* (26-krat členkovna raba).

Ugotovimo lahko, da leksikalne besede nedvomno pripadajo religijskemu (krščanskemu) diskurzu, nekatere med njimi še podrobneje usmerijo na red sv. Frančiška Asiškega (npr. *Frančišek*, *Asiz*, *frančiškan*, *vodilo*,

[5]Z znakom (*) so označene starejše oblike besed, pri katerih je v postopku avtomatskega posodabljanja prišlo do napak. Te so bile pojasnjene v poglavju 2.3.

Porcijunkuljska, *redovnik*, *redoven*, *sobrat*, *tretjerednik*, *voditelj*) in na lokalno okolje, od koder je pisec p. Repič prihajal (*bl. Monald Koprski*, *Koper*).

Med glagolskimi ključnimi besedami so pogosti glagoli rekanja, vezani na širjenje krščanske vere (liturgija, pastorala), ob tem pa tudi nekaj takšnih, ki v danih kontekstih pomenijo dejanja kršenja krščanskih načel.

Pridevniki in pomenski presežniki *blažen*, *brezmadežen*, *blagoslovljen*, *presladek*, *izreden*, *nepopisljiv*, *vsakovrsten*, *brezštevilen* ter členek *blagor* imajo ob sopojavitvi s pomensko pozitivnimi samostalniki močan pozitivni naboj, izražajo odobravanje in občudovanje koga/česa. V vsakem primeru pa z vidika intenzitete jezika omenjene jezikovne prvine okrepijo pomen ubesedenega (Mikolič, 2020). V smer krepitve argumenta vodi tudi pogosta raba prislova *jako* (redko *zelo*) in pomensko nasprotna si členka *nikaker* in *seveda*.

Seznam prvih 100 ključnih besed z referenčnim korpusom IMP (Erjavec, 2014) poda zelo podobne rezultate. Različno se denimo pojavijo še leme bogo* (za *Bog*), *tretjerednik* in *tretjerednica*, gvardian* (za *gvardijan*), *habit*, *častivec*, *ljubivec*, veznik *ke* (tudi *kè*, v pomenu ‘ko’ oz. ‘če’ v pogojnih odvisnikih, tudi za izražanje želje).

4 SKLEP

V korpusu CVET so zbrana vsa besedila patra Hijacinta Repiča, ki so izvirni in prevodni zapisi po tujejezičnih predlogah (vir avtor konsistentno navaja v samem besedilu oz. opombah). Prikazana korpusna analiza s konkordančniki CLARIN.SI je bila usmerjena v raziskovanje avtorjevega pripovednega stila. Uporabljene so bile funkcije, ki jih ponuja prosto dostopni konkordančnik noSketchEngine: seznam najpogostejših in najmanj pogostih besed (lem) ter seznam ključnih besed (lem) glede na izbrani referenčni korpus. Analiza frekvenčnosti besedja je podala nekatere pričakovane rezultate, kot so glagoli vezani na vsakdanje praktično-sporazumevalne okoliščine, pri tem pa so po pogostnosti izstopali modalni glagoli. Žanrsko besedila v glavnem sodijo med nabožne vzgojno-poučne članke, pri čemer bi bilo z vidika osebnega stila nadalje zanimivo raziskati, kakšno je razmerje med sporočanjsko in

vplivanjsko vlogo besedil v odnosu do vsebine (vzgojno-moralne problematike), z drugimi besedami, pri katerih temah je diskurz pripovedovalca odločnejši, strožji, koliko odkrito apologizira, po drugi strani pa je v nekaterih pogledih v primerjavi z drugimi verskimi pripovedniki nemara tudi milejši (znano je, da so se frančiškani, tedaj najštevilčnejši red na Slovenskem, uprli nekaterim rigoroznim verskim praksam med kleriki (Čebulj, 1922)), ter kakšen učinek takšen diskurz doseže pri bralcu.

Analiza najpogostejših samostalnikov in pridevnikov je smiselna, saj skoznjo vstopimo v avtorjevo pomensko polje in spoznamo vsebinski okvir besedil, z nadaljnjo raziskavo kontekstov rabe posamezne leksike v korpusnem gradivu pa dostopamo do vseh pojavitev v gradivu in lažje izluščimo specifike: dobesedni oz. preneseni pomen, ustaljene oz. izvirne metafore, prispodobe idr.

Opazovali smo tudi najpogostejše prislove, zaimke in členke, ki so prav tako lahko stilni označevalci, ki zaznamujejo tipično skladnjo (dolge povedi, veliko podredij), ali z vidika prepričljivosti argumentacije krepijo oz. šibijo pomen sporočila (Mikolič, 2020). Z vidika jezikovne pragmatike bi bila zanimiva tudi analiza diskurznih označevalcev v navezavi na žanr besedil: nekatera po vsebinski zgradbi in zaradi neposrednega nagovarjanja naslovnika spominjajo na pridižna besedila, prisotnost homiletičnih pripovednih prvin pa bi lahko ugotavljali tudi z analizo diskurznih označevalcev, značilnejših za govorjeni diskurz.

Korpus CVET 1.0 je prva tovrstna jezikovna zbirka besedil Škrabčevega verskega glasila *Cvetje z vertov sv. Frančiška*, velja pa omeniti, da je bilo besedje *Cvetja* sporadično zajeto že v Pleteršnikovem Slovensko-nemškem slovarju (1894–1895). Korpusno gradivo, označeno z metapodatki o avtorstvu in objavi (leto, letnik, zvezek), zato lahko služi za komparativne sinhrono-diahrone raziskave normiranja slovenskega jezika na prelomu 19. in 20. stoletja v periodičnem tisku ter za raziskave udejanjanja knjižne norme, ki jo je urednik Škrabec podrobno predstavil skozi svoje jezikoslovne znanstvene razprave na platnicah, v sami vsebini *Cvetja*.

V naslednji verziji korpusa bomo odpravili neposodobljene in napačno

posodobljene besede, in sicer s temeljitim pregledom napak, ki mu bodo sledile ročne korekture. Takšen korpus bo ne samo omogočil boljše analize, pač pa bi lahko služil tudi kot dodatna učna množica za cSMTiser, s čimer bi odpravili napake, ki se trenutno pojavljajo pri posodabljanju.

5 LITERATURA

Babič, S. in Erjavec, T. (2022). Izdelava in analiza digitalizirane zbirke paremioloških enot. *Conference on Language Technologies & Digital Humanities Ljubljana*, 2022. Pridobljeno 10. maja 2024, https://nl.ijs.si/jtdh22/pdf/JTDH2022_Babic_Erjavec_Izdelava-in-analiza-digitalizirane-zbirke-paremioloskih-enot.pdf

Biber, D. in Conrad, S. (1998). *Register, Genre, and Style*. Cambridge University Press.

Čebulj, R. (1922). *Janzenizem na Slovenskem in frančiškani: inavguralna disertacija*. Frančiškanska provincija Slovenije.

Dobrovoljc, K., Erjavec, T. in Krek, S. (2017). The Universal Dependencies Treebank for Slovenian. *Proceedings of the 6th Workshop on Balto-Slavic Natural Language Processing*. Association for Computational Linguistics, str. 33–38, doi:10.18653/v1/W17- 1406.

Eder, M. (2011). Style-Markers in Authorship Attribution. A Cross-Language Study of the Authorial Fingerprint. *Studies in Polish Linguistics, 6*(1), 99–114. Pridobljeno 15. maja 2024, https://ejournals.eu/en/journal/studies-in-polish-linguistics/article/style-markers-in-authorship-attribution-a-cross-language-study-of-the-authorial-fingerprint

Erjavec, T. (2014). *Digital library and corpus of historical Slovene IMP 1.1*. Slovenian language resource repository CLARIN.SI. http://hdl.handle.net/11356/1031

Grieve, J. (2007). Quantitative Authorship Attribution: An Evaluation of Techniques. *Literary and Linguistic Computing, 22*(3), 251–270. https://doi.org/10.1093/llc/fqm020

Hoover, D. L. (2002). Frequent Word Sequences and Statistical Stylistics. *Literary and Linguistic Computing, 17*(2), 157–180. https://doi.org/10.1093/llc/17.2.157

Hoover, D. L. (2003). Multivariate Analysis and the Study of Style Variation. *Literary and Linguistic Computing, 18*(4), 341–360. https://doi.org/10.1093/llc/18.4.341

Jeffries, L. in McIntyre, D. (2010). *Stylistics*. Cambridge University Press.

Koehn, P. (2010). *Statistical Machine Translation*. Cambridge University Press.

Korošak, B. (2006). *Anton Hijacint Repič (1863-1918), frančiškan pri sveti Ani v Kopru*. Branko.

Košir, D. (2022). Čitalniško gibanje na zahodnem in vzhodnem robu slovenskega kulturnega prostora. V M. Jesenšek (Ur.), *Čitalništvo in bralno društvo pri Mali Nedelji* (str. 66–92). Univerza v Mariboru, Univerzitetna založba. https://doi.org/10.18690/um.ff.1.2022

Košir, D. in Erjavec, T. (2024). *Corpus of texts by Hijacint Repič in "Cvetje z vertov sv. Frančiška" CVET 1.0*. Slovenian language resource repository CLARIN.SI. http://hdl.handle.net/11356/1226

Leech, G. N. in Short, M. (2007). *Style in Fiction: A Linguistic Introduction to English Fictional Prose*. 2nd ed. Pearson Longman. Pridobljeno 10. maja 2024, https://sv-etc.nl/styleinfiction.pdf

Ljubešić, N. in Dobrovoljc, K. (2019). What Does Neural Bring? Analysing Improvements in Morphosyntactic Annotation and Lemmatisation of Slovenian, Croatian and Serbian. *Proceedings of the 7th Workshop on Balto-Slavic Natural Language Processing*. Association for Computational Linguistics, str. 29–34, doi:10.18653/v1/W19-3704.

McIntyre, D. (2015). Towards an integrated corpus stylistics. *Topics in Linguistics, 16*(1), 59–68. https://doi.org/10.2478/topling-2015-0011

Mikolič, V. (2000). Povezanost narodne in jezikovne zavesti. *Jezik in slovstvo, 45*(5), 173–186. Pridobljeno 15. maja 2024, http://www.dlib.si/?URN=URN:NBN:SI:DOC-5BD79SLS

Mikolič, V. (2020). *Izrazi moči slovenskega jezika*. Annales ZRS; Slovenska matica.

Mikolič, V. (2022). *Ali bereš Cankarja?*. Slovenska matica.

Mosteller, F. in Wallace, D. L. (1964). *Inference and Disputed Authorship: The Federalist*. Addison-Wesley. Pridobljeno 20. maja 2024, https://archive.org/details/inferencedispute00most/page/n3/mode/2up

Nećak-Lük, A. in sod. (1998). *Medetnični odnosi v slovenskem etničnem prostoru*. Izsledki projekta. Inštitut za narodnostna vprašanja.

Onič, T. (2014). Univerzalnost literarnega sloga: vpogled v grafični roman. *Primerjalna književnost (Ljubljana), 37*(3), 179–198. Pridobljeno 15. maja 2024, https://ojs-gr.zrc-sazu.si/primerjalna_knjizevnost/article/view/6296/5954

Perenič, U. (2012). Čitalništvo v perspektivi družbenogeografskih dejavnikov. *Slavistična revija, 60*(3), 365–382. Pridobljeno 15. maja 2024, https://srl.si/ojs/srl/article/view/COBISS_ID-50413154

Scherrer, Y. in Ljubešić, N. (2016). Automatic Normalisation of the Swiss German ArchiMob Corpus Using Character-Level Machine Translation. *Proceedings of the 13th Conference on Natural Language Processing* (KONVENS 2016), str. 248–255.

Schöch, C., Erjavec, T., Patraș, R. in Santos, D. (2021). Creating the European Literary Text Collection (ELTeC): Challenges and Perspectives. *Modern Languages Open*. DOI: 10.3828/mlo.v0i0.364.

Ulčnik, N. (2010). Slomškove Drobtinice. *Studia Historica Slovenica, 10*(2-3), 683–703. Pridobljeno 10. maja 2024, https://shs.zgodovinsko-drustvo-kovacic.si/sites/default/files/shs2010_2-3.pdf

Žejn, A. (2020). Računalniško podprta stilometrična analiza pripovedne literature Janeza Ciglerja in Christopha Schmida v slovenščini. *Fluminensia: časopis za filološka istraživanja, 32*(2), 137–158. https://doi.org/10.31820/f.32.2.5

Žejn, A. in Erjavec, T. (2021). *The corpus of older Slovenian narrative prose PriLit 1.0*. Slovenian language resource repository CLARIN.SI. http://hdl.handle.net/11356/1319

CORPUS CVET 1.0: CREATION, DESCRIPTION AND ANALYSIS OF A COLLECTION OF OLDER TEXTS IN RELIGIOUS PERIODICALS

The paper presents the process of creation and linguistic tagging of the CVET 1.0 corpus, which contains the texts of Father Hijacint Repič in the older Slovenian language, published in the religious journal *Cvetje z vertov sv. Frančiška* in the period 1881–1916. The texts were obtained in PDF format from the dLib portal, edited in the Word editor and then converted to TEI. Older words were automatically updated using an open-source normalisation tool, which facilitates corpus search and further analysis of the material. The article points out some errors that occurred during normalisation, which will be corrected manually in the next version of the corpus (e.g. *keterim* > ketim* > katerim; *kesneje* > kosno* > kasneje; *sobrat* > zbrat* > sobrat). The updated texts were then automatically linguistically annotated, including morphosyntactic annotations as well as morphological and syntactic annotations according to the Universal Dependencies Formalism for Slovenian. We converted the TEI-encoded versions into various formats and published the collection under an open licence in the CLARIN.SI repository and concordancers suitable for linguistic analysis of the material. The second part of the paper presents an example of the analysis of the author's narrative style performed with noSketchEngine, based on the frequency variables of the most and least frequent words and keywords.

Keywords: historical Slovenian language, religious texts, TEI, normalisation, stylistic analysis, lexis