Citiranje jezikoslovnih podatkov v slovenskih znanstvenih objavah: stanje in priporočila

Darja Fišer*†, Jakob Lenardič*, Tomaž Erjavec†
* Oddelek za prevajalstvo, Filozofska fakulteta, Univerza v Ljubljani
Aškerčeva 2, 1000 Ljubljana
darja.fiser@ff.uni-lj.si
jakob.lenardic@ff.uni-lj.si
† Odsek za tehnologije znanja, Institut »Jožef Stefan«
Jamova cesta 39, 1000 Ljubljana
tomaz.erjavec@ijs.si

Povzetek

Odprta znanost temelji na prosto in odprto dostopnih znanstvenih publikacijah in podatkih. Slednji omogočajo preverjanje rezultatov predhodnih raziskav in njihovo nadgrajevanje, v kontekstu jezikovnih tehnologij in ročno označenih jezikovnih virov pa tudi šolanje novih orodij za procesiranje besedil. Vendar pa je, tako kot za znanstvene objave, tudi za podatke pomembno, da so korektno citirani, saj šele to omogoča ponovljivost raziskav, citati pa so tudi najpomembnejši pokazatelj zanimivosti in koristnosti delovanja znanstvenikov in pomembno vplivajo na njihovo možnost pridobivanja projektov in zaposlitev. V prispevku obravnavamo stanje citiranja jezikoslovnih podatkov, predvsem korpusov, v slovenskih znanstvenih publikacijah. Izvedli smo pregled večjega števila slovenskih revij in zbornikov in kvantitativno ter kvalitativno analizirali rezultate. Izsledke povzamemo in po ti. »austinskih načelih«, pokažemo, kaj je bilo že narejenega v sklopu raziskovalne infrastrukture CLARIN.SI ter predlagamo smernice za citiranje znanstvenih podatkov in načine za njihovo implementacijo.

Linguistic data citation in Slovene scientific publications: analysis and recommendations

Open science is based on freely and openly available scientific publications and data. The latter enable the verification and improvement of previous research. In the context of language technologies and manually annotated language resources, they also enable training of new text processing tools. However, just like scientific publications, research data need to be properly cited, as only this makes reproducibility of experiments possible and is the most important indicator of how interesting and useful researchers' work is in the community and plays a major role in their success with research grant proposals and career trajectory. In this paper, we survey the landscape of linguistic data (corpora) citation in Slovene scientific publications. The investigation was performed on key Slovene linguistic journals and proceedings with the results analysed both quantitatively and qualitatively. Our findings are organized according to the Austin Principles of data citation, where we present the developments in this field within the CLARIN.SI research infrastructure and propose recommendations for linguistic data citation as well as suggest solutions for their implementation.

1. Uvod

Odprti dostop do znanstvenih publikacij in podatkov pospešuje inovacije, spodbuja sodelovanje, zmanjšuje podvajanje dela in omogoča dograjevanje predhodnih rezultatov raziskav ter vključevanje državljanov in družbe (European Commission, 2012). Odprti dostop do rezultatov raziskav predvidevajo *Resolucija o nacionalnem programu za jezikovno politiko 2014–2018*,[1] *Nacionalna strategija odprtega dostopa do znanstvenih objav in raziskovalnih podatkov v Sloveniji 2015-2020*[2] ter *Akcijski načrt izvedbe nacionalne strategije odprtega dostopa do znanstvenih objav in raziskovalnih podatkov v Sloveniji 2015—2020*.[3]

V Sloveniji imamo na področju jezikovnih virov že dolgo tradicijo odprtih podatkov. Že od nastanka so bili odprto dostopni npr. jezikovni viri projektov MULTEXT-East[4], JOS[5] in SSJ,[6] leta 2013 pa je bila ustanovljena raziskovalna infrastruktura za jezikovne vire in orodja CLARIN.SI, v sklopu katere je bil vzpostavljen certificiran repozitorij, ki arhivira prek sto odprto dostopnih jezikovnih virov.

Med poglavitnimi cilji *Akcijskega načrta* je izvedba pilotnega programa *Odprti dostop do raziskovalnih podatkov v letih 2017—2020*, katerega namen je izboljšati dostop do raziskovalnih podatkov, mdr. z uvedbo novega sistema za vrednotenje raziskovalnih podatkov, v skladu s katerim bodo raziskovalni podatki, shranjeni v pooblaščenem podatkovnem središču, ki so prestali presojo pomena za znanost priznani kot znanstvena objava. Dobra praksa doslednega citiranja raziskovalnih podatkov je pomembna, ker zagotavlja in spodbuja transparentnost znanstvenega dela in posledično deluje kot ključni vzvod tovrstnega sistema vrednotenja. Raziskovalni podatki so v najboljšem primeru shranjeni v certificiranih repozitorijih (npr. repozitorij infrastrukture CLARIN.SI),[7] kar je skladno z *Akcijskim načrtom*, saj repozitoriji zagotavljajo tako trajni in transparentni dostop kot tudi jasno dokumentacijo za določen vir.

V pričujočem prispevku nas ne zanimajo jezikovni viri sami ali njihova dostopnost, niti ne njihova uporaba v raziskovalni skupnosti, temveč kako se le-ta citira v znanstvenih člankih slovenskih publikacij. Kot smo zapisali pred desetimi leti:

[1] http://www.pisrs.si/Pis.web/pregledPredpisa?id=RESO91#
[2] http://www.mizs.gov.si/delovna_podrocja/direktorat_za_znanost/sektor_za_znanost/strategije_s_podrocja_znanosti/nacionalna_strategija_odprtega_dostopa_do_znanstvenih_objav_in_raziskovalnih_podatkov_v_sloveniji_2015_2020/
[3] http://www.mizs.gov.si/fileadmin/mizs.gov.si/pageuploads/Znanost/doc/Odprti_dostop/Akcijski_nacrt_-_POTRJENA_VERZIJA.pdf
[4] http://nl.ijs.si/ME/
[5] http://nl.ijs.si/jos/
[6] http://www.slovenscina.eu/
[7] http://www.clarin.si/

»*Citiranje je še posebej pomembno, ker je merljiv kazalec raziskovalne uspešnosti, zato bi se tudi moralo dosledno izvajati. Žal pa to ni v navadi pri citiranju publikacij o jezikovnih virih: vse prepogosto se nek vir omeni samo po imenu ali pa se v najboljšem primeru doda njegov spletni naslov, namesto da bi se v virih citiralo publikacijo, kjer je vir prvotno opisan*« (Erjavec, 2009).

Od takrat se je stanje spremenilo, tako da je sedaj mogoče citirati ne samo publikacije o izdelavi nekega vira, pač pa tudi vir sam, saj npr. repozitorij CLARIN.SI za vsak vnesen vir na samem vrhu njegove spletne strani točno navaja, kako naj se ga citira. Dostop do podatkov v certificiranih repozitorijih, kot je CLARIN.SI, je v skladu s t.i. Austinskimi načeli za ustrezno citiranje v jezikoslovju, ki so povzeti v dokumentu *The FORCE11 Joint Declaration of Data Citation Principles*.[8] Poleg tega, da so natančna navodila za citiranje jezikovnih virov skladna z drugo točko Austinskih načel (*Credit and Attribution*, »Priznanje zaslug in avtorstva«), so transparentni metapodatki in stalni spletni identifikatorji, ki jih repozitoriji nudijo za vsak vir, ključnega pomena za zagotavljanje odprtega dostopa in s tem interoperabilnost, trajnost in preverljivost podatkov.

Neposredni povod za pričujoč prispevek je bilo zasedanje Interesne skupine za jezikoslovne podatke, ki je potekalo v sklopu plenarnega sestanka »Research Data Alliance« v Berlinu 22. 3. 2018. Na sestanku je bilo odprtih več vprašanj o citiranju raziskovalnih podatkov v jezikoslovju, kar nam je zbudilo zanimanje, kakšno je stanje na tem področju v Sloveniji. Prispevek ima sledečo strukturo: v 2. razdelku podamo pregled mednarodnih načel in praks pri citiranju znanstvenih podatkov v jezikoslovju, 3. razdelek analizira stanje v izbranih slovenskih publikacijah, 4. razdelek predlaga smernice za boljšo prakso na tem področju, zadnji razdelek pa zaključi in poda smernice za nadaljnje delo.

2. Mednarodna načela citiranja podatkov v jezikoslovju

Odprta znanost, odprti podatki in citiranje le-teh je v svetu trenutno v središču pozornosti, saj so obstoječe prakse tudi mednarodno zastarele, manj v naravoslovju in posebej računalništvu, mnogo bolj pa v humanistiki in jezikoslovju; tako npr. relativno nova »Splošna pravila za oblikovanje jezikoslovnih prispevkov« (Haspelmath, 2014) citiranja podatkov sploh ne omenjajo.

Obširen pregled pomena odprte znanosti, odprtih podatkov in potrebe po korektnem citiranju v jezikoslovju je podan v Berez-Kroeker et al. (2018), ki je rezultat iniciative, v kateri je sodelovalo 41 jezikoslovcev in drugih znanstvenikov. Prispevek najprej osmisli odprte raziskovalne podatke in ponovljivost raziskav, tako na splošno kot v jezikoslovju, nato pa poda pregled trenutnega stanja v jezikoslovju, kar se tiče transparentnosti uporabljenih virov in raziskovalnih metodologij. Avtorji ugotavljajo, da je po eni strani nemogoče uveljaviti ponovljivost raziskav brez primernega citiranja virov, po drugi pa, da je stanje v jezikoslovju še vedno zelo nezadovoljivo. Nato sledijo ugotovitve avtorjev glede potrebe po mehanizmih, ki bi ovrednotila tudi »delo na podatkih« pri zaposlovanju in napredovanju znanstvenikov, in nujnosti po korenitem premiku v omogočanju ponovljivosti raziskav v jezikoslovju, kar naj bi dosegli skozi izobraževanje, promocijo in razvoj ustreznih politik. Strinjajo se, da bi zbiralci podatkov za svoje delo morali dobiti primerno priznanje avtorstva, posebej takrat, ko so izdelani podatki dostopni, ponovno uporabni in jih je mogoče citirati. Prispevek zaključijo priporočila za konkretne dejavnosti, ki bi jih morali izvesti jezikoslovci, oddelki, sveti in založniki. Te dejavnosti so v veliki meri osredotočene na zagotovitev odprtih podatkov oz. izobraževanje, kako se upravlja s podatki, da sploh lahko postanejo odprti, kot tudi, kako primerno ovrednotiti to delo. Zadnje priporočilo pa je neposredno posvečeno boljšemu citiranju raziskovalnih podatkov, kjer avtorji svetujejo urednikom ter založnikom znanstvenih revij in knjig uvedbo konkretnih politik tako za izmenjavo podatkov kot za njihovo citiranje, pri slednjem tako, da razvijejo formate za citiranje jezikoslovnih podatkov.

3. Analiza citiranja objav v slovenskih znanstvenih publikacijah

3.1. Izbor gradiva in zasnova analize

Za pričujoči prispevek smo pregledali ključne slovenske revije in zbornike za področje jezikoslovja in ugotavljali, v kolikšni meri in na kakšen način avtorji prispevkov omenjajo oz. navajajo jezikovne vire. Naj poudarimo, da nas v tej raziskavi ni zanimalo, kateri jezikovni viri so v objavljenih raziskavah uporabljeni in citirani, temveč, kako jih avtorji navajajo.

Pri revijah smo analizirali navodila za avtorje in izdane številke za zadnjih pet let (2013-2017), pri zbornikih pa navodila za avtorje oz. predloge prispevkov ter celoten zbornik zadnje edicije konference. Med zborniki smo v študijo zajeli *JTDH 2016* in *Obzorja 2016*, med revijami pa: *Linguistica*, *Jezik in slovstvo*, *Jezikoslovni zapiski*, *Slavistična revija*, *Slovene Linguistic Studies* in *Slovenščina 2.0*

Skupaj je bilo pregledanih 751 znanstvenih prispevkov, od katerih jih vire omenja 133 oz. dobrih 17 %. Navedbe virov v pregledanih prispevkih ločujemo na naslednje kategorije:

- **Povezava[9] na vir v besedilu prispevka (največkrat v opombi)**. Zgled takega citiranja je v Žele (2014), kjer je povezava na korpus *Gigafida* podana v opombi. Prispevki v tovrstni kategoriji ne navajajo ključne publikacije o viru, t.j. Logar et al. (2012).
- **Povezava na vir v bibliografiji**. Zgled takega citiranja je v Ribič (2016), kjer je povezava na korpus *Gigafida* podana v končnem seznamu virov. Prispevki v tovrstni kategoriji ne navajajo ključne publikacije o viru.
- **Povezava na vir v besedilu prispevka (največkrat v opombi) kot tudi v bibliografiji**. Zgled takega citiranja je v Žele (2015), kjer je povezava na korpus *Gigafida* podana večkrat v opombah ter v končnem seznamu virov. Prispevki v tovrstni kategoriji ne navajajo ključne publikacije o viru.
- **Publikacija o viru**. Zgled takega citiranja je v Verdonik in Sepesy Maučec (2013), kjer je za korpus *OPUS OpenSubtitles* navedena ključna publikacija o viru, t.j. Tiedemann (2009).

[8] https://www.force11.org/datacitationprinciples

[9] V to kategorijo vključujemo tudi navedbe stalnih spletnih identifikatorjev, kot so handle in DOI.

- **Povezava na vir v besedilu prispevka in publikacija o viru**. Zgled takega citiranja je v Bálint Čeh in Kosem (2017). Avtorja podajata povezavo na korpus *Gigafida* v opombi in navajata ključno publikacijo, t.j Logar et al. (2012).
- **Kombinacija različnih načinov navajanja virov**. Zgled takega navajanja je v Ljubešić et al. (2013), kjer je za označevalnik JOS navedena ključna publikacija (Erjavec et al., 2010), za korpus ssj500k pa zgolj povezava med besedilom.[10]
- **Brez navedbe vira**. Zgled takega citiranja je v Vidovič Muha (2015). Avtorica se sklicuje na uporabo označevalnika JOS, vendar ne podaja niti povezave na vir niti ne navaja njegove ključne publikacije (t.j. Erjavec et al., 2010).

3.2. Pregled navodil avtorjem

V tem razdelku podajamo kratek pregled navodil za avtorje, saj je od teh navodil močno odvisno, kako bodo avtorji navajali vire. Za revije *Jezikoslovni zapiski*, in *Slovene Linguistic Studies* ter za zbornik *Obdobja* navodil avtorjem na njihovih spletnih straneh nismo našli.

Najbolj podrobna navodila za navajanje virov podaja revija *Slovenščina 2.0*,[11] ki ločuje navajanje korpusov, spletnih strani in spletnih virov:

Korpus:
- Gigafida. Dostopno prek: http://www.gigafida.net (datum dostopa).
- Cambridge English Corpus.e Dostopno prek: http://www.cambridge.org/gb/elt/catalogue/subject/item2701617/Cambridge-International-Corpus/?site_locale=en_GB (datum dostopa).

Spletna stran:
- OpenWebSpider. Dostopno prek: http://www.openwebspider.org/ (datum dostopa).
- Creative Commons. Dostopno prek: http://creativecommons.org/ (datum dostopa).

Spletni vir:
- Pew Research Center (2010): Americans Spending More Time Following the News ? Ideological News Sources: Who Watches and Why. Dostopno prek: http://www.people-press.org/ (datum dostopa).
- TEI Consortium, ur. (2011): TEI P5: Guidelines for Electronic Text Encoding and Interchange: Version 1.9.1. Dostopno prek: http://www.tei-c.org/Guidelines/P5/ (datum dostopa).
- Scott, M. (2008): WordSmith Tools: Version 5. Dostopno prek: http://www.lexically.net/downloads/version5/HTML/index.html (datum dostopa).

Jezik in slovstvo avtorje poziva,[12] da vire in literaturo navajajo ločeno, kar se nam zdi dobra praksa, saj s tem avtorjem med drugim sporočajo, da je uporaba in navajanje virov pomemben sestavni del znanstvenega prispevka. Dodatno velja dodatno omeniti, da poziv k ločenemu navajanju jezikovnih virov omogoča bralcem lažji dostop in preveritev citiranih podatkov, ki podpirajo neko znanstveno trditev, kar je skladno z npr. austinskimi načeli (glej razdelek 4) . Podrobneje ta revija načina za navajanje jezikovnih virov sicer ne definira, iz primera za navajanje spletnih strani pa lahko sklepamo, da jezikovne vire v elektronski obliki enači s spletnimi stranmi, saj kot primer navajanja spletnih strani navaja korpus *FidaPLUS*:

- Korpus slovenskega jezika FidaPLUS: <http://www.fidaplus.net>. (Dostop dan. mesec. leto.)

Na podoben način jezikovne vire obravnava revija *Linguistica*[13]:

- Le dictionnaire de la zone. 20 May 2010. http://www.dictionnairedelazone.fr/.

Slavistična revija[14] v navodilih za oblikovanje seznama literature uvaja zelo neeksaktno navajanje spletnih virov, brez navedbe spletnih povezav, verzij oz. datuma dostopa:

- Lemma (Lexikographie). Wikipedia: Die freie Enzyklopädie.
- Primož JAKOPIN, 1980: Zgornja meja entropije pri leposlovnih besedilih v slovenskem jeziku: Doktorska disertacija. Ljubljana. Na spletu.

Pri prvem primeru ni jasno, na katero različico se referenca nanaša, saj je Wikipedija kolaborativen projekt, kjer uredniki gesla lahko ves čas spreminjajo, bi bilo nujno treba dodati datum dostopa. Pri drugem primeru pa ni jasno, ali gre za referenco na doktorsko disertacijo kot publikacijo al za jezikovni vir, ki je bil v okviru disertacije razvit. Prav tako referenca ne vsebuje spletne povezave, zato bralec do vira ne more dostopati. Tovrstna praksa ne spodbuja preverljivosti in ponovljivosti raziskav ter priznavanja zaslug avtorjem virov, zato bi jo bilo pomembno čim prej izboljšati, še posebej, ker gre za jezikoslovno revijo, ki se v sistemu vrednotenja znanstvenih objav uvršča v sam vrh.

Revija *Slovene Linguistic Studies* posebej za navajanje elektronskih virov ne podaja navodil.

Podobno zbornik *JTDH*[15] v predlogi prispevkov sicer vsebuje primer dodajanja hiperpovezav v opombe in navaja načine navajanja različnih tipov enot bibliografije, a med njimi ni primerov za citiranje jezikovnih virov. Glede na to, da gre za vodilno konferenco za področje jezikovnih virov in tehnologij, bi konferenca nujno morala posvečati več pozornosti ozaveščanju in usmerjanju avtorjev prispevkov za ustrezno citiranje jezikovnih virov.

3.3. Kvantitativna analiza

Glede na podatke v Tabeli 1 vsebuje 17.7 % vseh pregledanih objav (vsaj eno) navedbo jezikovnega vira, načini navajanja pa so zelo raznoliki in razpršeni. Izrazito prevladuje navajanje povezave na vir v bibliografiji, česar se poslužuje četrtina vseh prispevkov, v katerih so bili viri uporabljeni. Dvakrat redkejša je praksa navajanja ključne publikacije o uporabljenem viru, ki je v trenutno veljavnem sistemu, ki seveda ni popoln in ni (primarni) cilj znanstvenega udejstvovanja, je pa kljub vsemu zelo pomemben za pridobivanje zaposlitev in projektov, za vrednotenje znanstvene uspešnosti edini način citiranja, ki avtorjem vira prinaša točke. Precej pogosto je kombiniranje več različnih načinov navajanja virov v istem prispevku (19 %), kar kaže na neupoštevanje navodil avtorjem oz. na pomanjkljiva navodila.

V Tabeli 2 navajamo rezultate analize za posamezne revije, ki smo jih vključili v raziskavo. Najvišji delež prispevkov, ki omenjajo jezikovne vire, vsebuje revija *Slovenščina 2.0* (97 %), najnižjega pa revija *Linguistica* (4 %), kar posredno tudi odraža programsko usmeritev revij. Po nenavajanju uporabljenih virov izrazito izstopa *Slavistična revija*, v kateri pri več kot treh četrtinah (78 %) prispevkov, ki rabo virov omenjajo, teh virov nikjer ne

[10] http://www.slovenscina.eu/tehnologije/ucni-korpus
[11] http://slovenscina2.0.trojina.si/si/oddaja-prispevkov/
[12] https://www.jezikinslovstvo.com/02.php
[13] https://revije.ff.uni-lj.si/linguistica/about/submissions#authorGuidelines
[14] https://srl.si/navodila_guidelines.pdf
[15] http://www.sdjt.si/wp/dogodki/konference/jtdh-2018/#navodila

citirajo. Glede na to, da gre za vodilno jezikoslovno revijo v našem prostoru, ki je uvrščena tudi na seznam ARRS revij posebnega pomena, bi še posebej to uredništvo revije moralo skrbeti za visok nivo raziskovalne kulture v slovenskem jezikoslovju in ustrezno citiranje raziskovalnih podatkov od avtorjev izrecno zahtevati v navodilih za avtorje.

Vseh objav	751	100,0 %
Objave z omembo vira	133	17,7 %
Hiperpovezava na vir v besedilu prispevka	13	9,7 %
Hiperpovezava na vir v bibliografiji	33	24,8 %
Hiperpovezava na vir v besedilu in v bibliografiji	14	10,5 %
Publikacija o viru	16	12,0 %
Hiperpovezava na vir v besedilu prispevka in publikacija o viru	8	6,0 %
Kombinirano	25	18,8 %
Brez	25	18,8 %

Tabela 1: Pregled distribucije različnih načinov navajanja virov v analiziranih publikacijah.

Najbolj homogeno navajanje virov je v *Slavistični reviji*, kjer smo identificirali le dva različna načina (povezava na vir v besedilu ali v bibliografiji), najbolj heterogeno pa v *Jezikoslovnih zapiskih*, kjer najdemo vse načine navajanja virov, razen kombiniranega. Najvišji delež navedbe vira v obliki hiperpovezave na spletno stran vira najdemo v reviji *Linguistica* (67 %), najvišji delež citiranja ključnega prispevka o viru pa pripada reviji *Slovenščina 2.0* (18 %).

Od posameznih načinov navajanja virov je navajanje povezav na vir v besedilu prispevka (največkrat v opombah) najpogostejši način navajanja virov v vseh revijah, razen v reviji *Slovenščina 2.0*, kjer je nekoliko pogostejše citiranje ključne publikacije o viru. Tega načina se sicer v manjšem številu prispevkov poslužujejo samo še v revijah *Jezik in slovstvo* in *Jezikoslovni zapiski*.

V Tabeli 3 navajamo rezultate za konferenci, ki smo ju vključili v raziskavo. V zborniku *JTDH 2016* so viri omenjeni v 93 % vključenih prispevkov, kar je glede na področje konference razumljivo. V tem zborniku naletimo na izrazito velik delež prispevkov (46 %), v katerih avtorji uporabljajo različne kombinacije navajanja virov. To je verjetno odraz heterogene raziskovalne skupnosti, ki se predstavlja na tej konferenci, in pomanjkljivih navodil avtorjem ter manj rigoroznega uredniškega in tehničnega pregleda končnih različic oddanih prispevkov.

V zborniku *Obdobja*, ki je bil posvečen Jožetu Toporišiču, je tovrstnih prispevkov 19 %. Glede na to, da je bila ta edicija simpozija tematsko vezana na jezikovni opis slovenščine, se zdi ta rezultat nizek. Vendar je po drugi strani občutno višji kot v programsko sorodnih revijah, predstavljenih v Tabeli 2, kar morda nakazuje spremembe sestave oz. praks tudi v tej skupnosti, saj so revije tradicionalno konzervativnejše in spremembe, do katerih v raziskovalni skupnosti prihaja, absorbirajo nekoliko kasneje od konferenc.

3.4. Kvalitativna analiza

V tem razdelku navajamo zanimivejše pojave, na katere smo naleteli pri kvalitativnem pregledu gradiva. Najprej predstavljamo nekatere primere dobrih praks, nato pa analiziramo identificirane problematične primere navajanja virov. Kot zgleden primer citiranja virov navajamo Logar et al. (2014) v reviji *Slovenščina 2.0*, ki za isti vir navaja tako ključno publikacijo o viru v bibliografiji kot tudi povezavo na vir v besedilu prispevka v sprotnih opombah, ki so prikazane na dnu relevantne strani prispevka. Na ta način bralcu omogočimo, da neposredno dostopa tako do vira kot tudi do publikacije o njem, prav tako pa avtorjem vira ustrezno priznamo zasluge in avtorstvo ter zagotovimo citiranost.

Naslednji zgleden primer citiranje virov, ki prav tako prihaja iz revije *Slovenščina 2.0*, je Arhar Holdt in Dobrovoljc (2016), ki v bibliografiji za vir navede stalni spletni identifikator handle v repozitoriju CLARIN.SI:

- `Krek, S., Erjavec, T., Dobrovoljc, K., Može, S., Ledinek, N. in Holz, N. (2015): Training corpus ssj500k 1.4. Dostopno prek: http://hdl.handle.net/11356/1052.`

Navajanje handlov je pomembno, ker bralcu zagotavlja, da bo lahko dostopal do vira, četudi se sam naslov spletne strani spremeni. Prav tako pa handle bralcu omogoča dostop do podrobnejšega opisa jezikovnega vira, ki je bil uporabljen v raziskavi, do njegovih metapodatkov, za prosto dostopne vire pa tudi do vira samega. S tem je močno izboljšana preverljivost in ponovljivost raziskav, spodbuja pa tudi nadaljnje razširitve in izboljšave raziskav ter maksimizira izrabo jezikovnega vira, izdelava katerega je zahtevala finančni in časovni vložek.

Pri pregledu smo naleteli tudi na problematične načine citiranja, ki jih uvrščamo v naslednje kategorije:

- Nekonsistentno navajanje istega vira: V *Slavistični reviji* je isti vir navajan zelo različno. Npr. v Meterc (2013) in Jakop (2014):
  - `Gigafida, korpus slovenskega jezika. Ur. Filozofska fakulteta Univerze v Ljubljani. Ljubljana: FF. Splet.`
  - `Korpus GigaFida. Na spletu.`
- Nekonsistentno navajanje različnih virov istega tipa: V reviji *Slovene Linguistic Studies* je v Štumberger (2015) za *Sloleks* navedena hiperpovezava v opombi, nemška leksikalna vira pa sta vključena v bibliografijo:

	Jezik in slovstvo		*Slavistična revija*		*Jezikoslovni zapiski*		*SLS*		*Slo 2.0*		*Linguistica*	
Vse objave	157	100%	180	100%	115	100%	26	100%	45	100%	134	100%
Objave z omembo vira	11	7%	14	8%	20	17%	8	31%	34	76%	6	4%
Povezava na vir v besedilu prispevka	0	0%	1	7%	3	15%	2	25%	5	15%	0	0%
Povezava na vir v bibliografiji	4	36%	2	14%	7	35%	2	25%	5	15%	4	67%
Povezava na vir v besedilu in bibliografiji	2	18%	0	0%	2	10%	3	38%	3	9%	1	17%
Publikacija o viru	2	18%	0	0%	1	5%	0	0%	6	18%	0	0%
Povezava na vir in publikacija o viru	0	0%	0	0%	1	5%	0	0%	6	18%	0	0%
Kombinirano	0	0%	0	0%	0	0%	1	13%	8	24%	0	0%
Brez	3	27%	11	79%	6	30%	0	0%	1	3%	1	17%

Tabela 2: Pregled praks navajanja jezikovnih virov v ključnih slovenskih znanstvenih revijah za področje jezikoslovja za obdobje 2013-2017. SLS je okrajšava za revijo *Slovene Linguistic Studies*, Slo 2.0 pa za revijo *Slovenščina 2.0*.

	Zbornik JTDH		*Zbornik Obdobja*	
Vse objave	30	100%	64	100%
Objave z omembo vira	28	93%	12	19%
Povezava na vir v besedilu prispevka	2	7%	0	0%
Povezava na vir v bibliografiji	2	7%	7	58%
Povezava na vir v besedilu in bibliografiji	3	11%	0	0%
Publikacija o viru	6	21%	1	8%
Povezava na vir in publikacija o viru	1	4%	0	0%
Kombinirano	13	46%	2	17%
Brez	1	4%	2	17%

Tabela 3: Pregled praks navajanja jezikovnih virov v ključnih konferenčnih zbornikih za področje jezikoslovja za leto 2016.

  - OWID, Online-Wortschatz-Informationssystem Deutsch des Instituts für deutsche Sprache, Mannheim, (13. 3. 2008)
  - Klappenbach, Ruth, Steinitz, Wolfgang (ur.). 1967 (1964). Wörterbuch der deutschen Gegenwartssprache. 1. Band. Berlin: Akademie-Verlag. http://www.dwds.de/ (1. 7. 2008, 27. 3. 2015).

- Neustrezne hiperpovezave: V reviji *Jezik in slovstvo* smo opazili neustrezno navajanje povezav na vire. Npr. v Polajnar (2013) ni hiperpovezave na osnovno stran vira, ampak na podstran:
  - Gigafida: http://www.Gigafida.net/Support/About

  V *Slavistični reviji* smo opazili neustrezno navajanje povezav na vire. Npr. v Fabčič (2014) je korpus *FidaPLUS* v bibliografiji naveden brez povezave:
  - FidaPLUS – Korpus slovenskega jezika. Na spletu.

  Ker je korpuse mogoče naložiti na različne konkordančnike, kar lahko privede tudi do razlik v rezultatih, je za zagotavljanje preverljivosti in ponovljivosti raziskav v referenci nujno potrebno navesti natančno povezavo, ki je bila v raziskavi uporabljena.[16]

  V reviji *Slovenščina 2.0* smo opazili nenatančo navajanje hiperpovezave do korpusa. Npr. v Arias-Badia et al. (2014), kjer je za španski korpus navedena generična povezava na konkordančnih SketchEngine:
  - SWC = Spanish Web Corpus. Available at: www.sketchengine.co.uk (20 October 2014).

  Tovrstno navajanje referenc na korpuse je med jezikoslovci precej razširjeno, je pa problematično iz več razlogov. Ne samo, da ne priznava avtorstva korpusa, temveč resno zavira preverljivost in ponovljivost raziskav, saj iz reference sploh ni razvidno, za katero različico korpusa konkretno gre, saj po eni strani obstaja več spletnih korpusov španščine, ki so jih ustvarili različni avtorji, po drugi pa so bili številni med njimi bili izdelani v več različicah in vsebujejo različno gradivo. Ko smo korpus želeli preveriti v konkordančniku SketchEngine, na katerega nas raferenca napoti, ga nismo našli, saj konkordančnik na dan preverjanja[17] ponuja dva španska korpusa tega tipa: Spanish Web Corpus oz. SpanishWaC (Sharoff 2006) in Spanish Web 2011 oz. esTenTen11 (Kilgarriff in Renau 2013). Tu je potrebno poudariti, da odgovornost za ustrezno navajanje virov ne leži samo na strani avtorjev prispevkov, temveč tudi avtorjev virov, ki bi vsem uporabnikom prvi morali zagotoviti ustrezno spremno dokumentacijo o korpusu, vključno z navodili za citiranje, tako ključnega prispevka o viru kot tudi navajanje korpusa v konkordančniku in korpusa kot podatkovno zbirko. Veliko razvijalcev virov tega še vedno ne omogoča, zato je ozaveščanje nujno potrebno tudi pri tej ciljni skupini.

4. Diskusija

Kot je pokazala analiza, je trenutno stanje na področju navajanja virov v slovenskem jezikoslovju vse prej kot idealno, saj so navodila avtorjem za področje elektronskih jezikovnih virov zelo raznolika, ponekod zastarela, pri precejšnjem številu revij in zbornikov pa celo manjkajo. Posledično so tudi prakse navajanja virov tako med kot tudi znotraj posameznih znanstvenih publikacij zalo heterogene. Še bolj pa je zaskrbljujoč podatek, da skoraj petina objavljenih prispevkov v uglednih znanstvenih revijah in zbornikih uporabljenih virov sploh ne navaja.

Da bi skušali prispevati k izboljšanju stanja, v nadaljevanju prispevka oblikujemo priporočila, ki temeljijo na mednarodnih iniciativah in predlogih, kako izboljšati citiranost raziskovalnih podatkov. Konkretno sledimo osmim načelom »austinskih principov« citiranja podatkov v jezikoslovju (Berez-Kroeker et al., 2017). Za vsako od načel podamo ime in prevod definicije, nakar ga umestimo v Slovenijo z analizo stanja in predlogi za ukrepe, kako jih realizirati.

4.1. Pomembnost

Podatki bi morali biti legitimen rezultat raziskav in jih je obvezno citirati. Citati podatkov bi za merjenje raziskovalčeve znanstvene uspešnosti morali biti enako pomembni, kot so to citati objav.

Rezultati analize so pokazali, da je to načelo v Sloveniji z manjšimi izjemami zelo slabo zastopano. Za njegovo udejanjanje sta ključna dva ukrepa. Prvi je izobraževanje, predvsem študentov, kjer njihovi profesorji oz. mentorji vztrajajo pri korektnem citiranju podatkov v seminarskih nalogah, zaključnih delih in znanstvenih objavah. Drugi ukrep bi, kot predlagajo Berez-Kroeker et al. (2018), morali izvesti uredniški odbori revij in programski odbori konference tako, da bi v navodila za avtorje dodali navodila za ustrezno citiranje jezikoslovnih podatkov, tako kot so jih predhodno za spletne vire. Posebej poudarjamo, da je dobrim praksam navajanja raziskovalnih podatkov v slovenskih publikacijah že posvečen priročnik *Priprava raziskovalnih podatkov za odprt dostop* (Štebe et al., 2015). Avtorji priporočajo, »da se v seznamu uporabljene literature podatke navaja s *polno navedbo avtorja oz. avtorjev, naslova, mesta dostopa do podatkov in stalnega identifikatorja*, skladno z oblikovnimi zahtevami znanstvene revije« (2015: 13; naš poudarek).

[16] Repozitorij *CLARIN.SI* rešuje ta problem tako, da je v navodilih za navajanje virov, ki so podani kot prva informacija v glavi vnosa za posamezen vir, jasno izpostavljeno, za katero različico vira gre in ali je ta različica dostopna preko konkordančnika. Za starejše različice repozitorij opozori o morebitni zastarelosti podatkov. Primerjaj npr. vnos za drugo različico korpusa *Gos VideoLectures (Transcriptions)* (Verdonik et al., 2017), ki je dostopna preko konkordančnika *KonText*, s prvo (Verdonik et al., 2016), ki preko taistega konkordančnika ni dostopna.

[17] https://www.sketchengine.eu [15. 4. 2018]

Tu so ključni naslovniki *Slavistična revija* kot revija s posebnega seznama ARRS, konferenca oz. monografija *Obzorja*, kot tudi pričujoča konferenca *Jezikovne tehnologije in digitalna humanistika*, ki bi na tem področju morala orati ledino. Ni odveč omeniti, da je citiranje obvezna sestavina v uvodu omenjeni Nacionalni strategiji odprtega dostopa in njenem Akcijskem načrtu, izpostavljeno pa je tudi v raznih drugih razpravah o odprtih podatkih v Sloveniji, vključno z nalogami financerja in uredništev revij.

4.2. Priznanje zaslug in avtorstva

Citiranje podatkov bi moralo služiti priznavanju znanstvenih zaslug, normativnega ter pravnega avtorstva vsem, ki so prispevali k njihovi izdelavi.

Za priznavanje znanstvenih zaslug je v Sloveniji merodajen SICRIS, ki se za štetje citatov zanaša na Web of Science in SCOPUS. Vplivanje na štetje citatov znanstvenih podatkov je tako izven dometa pričujočega članka.

Lahko pa v Sloveniji vplivamo na to, kako se točkujejo objave znanstvenih podatkov. Trenutno v sistemu COBISS že obstaja rubrika »2.20 Zaključena znanstvena zbirka podatkov ali korpus«, vendar ima takšen vnos priznanih samo 5 točk. Bistveno bolje so lahko točkovane objave pod to rubriko v primerih, ko je vir podatkov naveden v seznamu »Zaključene znanstvene zbirke podatkov, ki se upoštevajo pri kategorizaciji znanstvenih publikacij (BIBLIO-D)«.[18] Trenutno je na tem seznamu samo Arhiv družboslovnih podatkov (ADP). Pomembne objave v ADP tako privzeto dobijo 30 točk (Vončina, 2016), če so deponirani podatki s strani komisije ADP ocenjeni kot zelo pomembni.

Za jezikoslovne podatke bi bilo potrebno tudi repozitorij CLARIN.SI uvrstiti na seznam BIBLO-D, kar pa bi poleg samega predloga komisiji ARRS zahtevalo tudi bolj podrobna navodila za vnašanje virov, kot tudi ustanovitev komisije za vrednotenje vnesenih virov. Vse to pa seveda tudi zahteva precejšen vložek dela in s tem financiranja CLARIN.SI.

4.3. Dokazi

V znanstvenih objavah, kadarkoli in kjerkoli neka trditev sloni na podatkih, bi morali biti ti podatki ustrezno citirani.

Podobno kot za 1. načelo (pomembnost) je tudi tu ključno izobraževanje, navodila za avtorje in uredniška politika publikacij.

4.4. Nedvoumna identifikacija

Citiranje podatkov naj bi vsebovalo trajno metodo identifikacije, primerno za strojno obdelavo, mednarodno edinstveno in široko sprejeto v skupnosti.

Ta pogoj je v veliki meri že realiziran v sklopu repozitorija CLARIN.SI. Vsak vir ima trajni identifikator PID (*persistent identifier*) po sistemu »handle«, na vrhu strani pa je jasno napisano, kako naj se vir citira, pri čemer navedek vsebuje tudi identifikator handle. Repozitorij CLARIN.SI tudi podpira izvoz metapodatkov po shemi Dublin Core, ki jih žanje več agregatorjev: CLARIN VLO,[19] OpenAIRE,[20] re3data[21] in OAI.[22]

V jezikoslovju je poleg navajanja vira kot podatkovne zbirke pomembno tudi navajanje poizvedbe v konkordančniku. Konkordančniki CLARIN.SI, kot tudi konkordančniki projekta Sporazumevanja v slovenskem jeziku (torej konkordančnik za Gigafido,[23] Kres[24], itd.) so vsi narejeni po principu REST, da torej URL poizvedbe zadošča za ponovno in enako poizvedbo.[25] Z drugimi besedami, če po poizvedbi in prikazu rezultatov shranimo URL rezultata, je možno ta URL shraniti, in prek njega ponovno dobiti iste rezultate. Tu velja še opomba, da so takšni URL-ji tipično zelo dolgi in zato neprimerni ali vsaj težavni za citiranje. Vendar pa za krajšanje URL-je obstaja več spletnih storitev, od katerih je posebej zanimiva shortref.org[26], ki jo ponuja češki LINAT/CLARIN. Za razliko od drugih krajševalnikov ponuja shortref.org opis poizvedbe, vrne pa trajni identifikator po sistemu *handle*.

4.5. Dostop

Citiranje podatkov naj bi pripomoglo k dostopu do samih podatkov in do povezanih metapodatkov, dokumentacije, programske opreme in drugih materialov, ki so potrebni, da tako ljudje kot računalniki te podatke informirano uporabljajo.

Ta zahteva je tudi že v veliki meri realizirana v sklopu repozitorija CLARIN.SI, saj vsak vnos vsebuje tako metapodatke kot tudi same podatke, ki so pred vključitvijo v repozitorij preverjeni s strani urednikov.

4.6. Trajnost

Enoznačni identifikatorji in metapodatki, ki opisujejo podatke, morajo biti trajni, celo bolj kot sami podatki.

Repozitorij CLARIN.SI je del slovenske in evropske infrastrukture, domuje pa na Institutu »Jožef Stefan«, ki ima visoko razvito računalniško infrastrukturo. Oboje v največji možni meri ponuja garancijo za dolgotrajnost (meta)podatkov, deponiranih v repozitoriju. K trajnosti metapodatkov pa prispeva tudi že omenjeno dejstvo, da se le-ti redno izvažajo v več spletnih agregatorjev. CLARIN.SI izvaja tudi redne testiranje skladnosti in povezljivosti podatkov.

4.7. Specifičnost in preverljivost

Citiranje podatkov naj bi pripomoglo identifikaciji, dostopu in preverjanju specifičnih podatkov, ki podpirajo neko trditev. Citiranje ali metapodatki citiranja naj bi vsebovali podatke o izvoru in stabilnosti v zadostni meri, da omogočijo preverbo, da je specifičen časovni okvir, različica ali del podatkov, ki so bili naknadno prevzeti, enak kot podatki, ki so bili izvorno citirani.

Tudi tu CLARIN.SI v veliki meri zadošča temu načelu. Vnosi v repozitorij se ne spreminjajo, v primeru dopolnjenih ali popravljeni podatkov se ti vpišejo v nov vnos, vendar z medsebojno povezavo med starim in novim vnosom. Velja posebej poudariti, da je nadzor nad različnimi verzijami, ki ga repozitorij CLARIN.SI

[18] http://home.izum.si/COBISS/bibliografije/Kateg-znan-zbirke.html
[19] https://vlo.clarin.eu/
[20] https://www.openaire.eu/
[21] https://www.re3data.org/repository/r3d100011922
[22] http://www.language-archives.org/archive/clarin.si
[23] http://www.gigafida.net/
[24] http://www.korpus-kres.net/
[25] Seveda, če se medtem ni spremenil korpus.
[26] http://shortref.org/

omogoča, eno izmed priporočil ustreznega digitalnega skrbništva jezikovnih podatkov (npr. Štebe et al., 2015: 6). Mnogo virov je zapisanih po priporočilih TEI, ki tipično vsebujejo bogate metapodatke, s katerimi je mogoče podrobno določiti želene izseke virov.

4.8. Interoperabilnost in fleksibilnost

Metode za citiranje podatkov naj bi bile fleksibilne v zadostni meri, da omogočajo različne prakse med skupnostmi, vendar se ne smejo razlikovati v tolikšni meri, da ogrozijo interoperabilnost praks citiranja podatkov med skupnostmi.

Repozitorij mdr. navaja naslov in avtorje vsakega vira ter na vrhu dostopne strani vira točno definira, kako je vir potrebno citirati.

5. Zaključki

V prispevku smo predstavili rezultate študije, s katero smo preverjali stanje citiranja jezikoslovnih podatkov, predvsem korpusov, v najpomembnejših slovenskih znanstvenih revijah in zbornikih, ki so bili objavljeni v zadnjih petih letih. Izvedli smo pregled navodil za avtorje ter kvantitativno in kvalitativno analizirali obseg in način navajanja virov, s katerim smo pokazali, da stanje ni zavidljivo in si je zato potrebno prizadevati za ozaveščanje, izobraževanje in podporo v skupnosti.

Po opravljeni analizi ugotavljamo, da na jezikovnih virih temelji manj kot petina vseh objavljenih prispevkov, kar je glede na stopnjo razvitosti in razpoložljivosti jezikovnih virov za slovenščino malo in kaže na izključenost skupnosti, ki vire razvija, iz »mainstream« jezikoslovne raziskovalne skupnosti pri nas. Kjer pa so bili v raziskavi uporabljeni, pa jih v skoraj petini prispevkov avtorji sploh ne navajajo. To kaže na pomanjkanje ozaveščenosti jezikoslovcev o pomenu navajanja vseh virov v znanstvenem publiciranju.

S prispevkom, v katerem smo predlagali načela za ustrezno citiranje digitalnih jezikovnih virov, ki temeljijo na mednarodnih poročilih, smo storili prvi korak v tej smeri. Brez tega onemogočamo preverljivost, ponovljivost in nadgrajevanje prejšnjih raziskav, ki so osnovni temelji odprte znanosti. Korektno citiranje jezikovnih virov pa je pomembno tudi zato, ker je v njihov razvoj potrebno vložiti izjemno veliko truda in časa, znanstveni citati pa so najpomembnejši indikator znanstvene uspešnosti.

Ozaveščanje in izobraževanje bi bilo potrebno začeti že v okviru univerzitetnih študijskih programov in poskrbeti za ustrezne smernice za navajanje virov tudi v tem kontekstu. Na področju ozaveščanja skupnosti aktivnih raziskovalcev pa bi z izobraževalnimi dogodki in spletnimi gradivi veliko lahko pripomogla nacionalna raziskovalna infrastruktura CLARIN.SI.

Čim prej bi bilo potrebno vzpostaviti dialog s knjižničarji in uredništvi, ki imajo neposreden stik z raziskovalci in tako tudi veliko moč pri promoviranju dobrih praks citiranja jezikovnih virov, zaradi česar so eni najpomembnejših akterjev pri vzpostavljanju in zagotavljanju dobrih praks za citiranje.

Prav tako pa je nujno potrebno poskrbeti tudi za ozaveščanje razvijalcev virov, ki lahko k ustreznemu citiranju veliko pripomorejo tako, da ustrezno deponirajo in dokumentirajo svoje vire. Za odprto znanost namreč še zdaleč ni dovolj, da nek vir obstaja in je dostopen, temveč mora biti tudi opremljen z vso potrebno spremno dokumentacijo, med katero vključujemo tudi navodila za citiranje. Opuščanje teh praks že na prvem koraku zavira ustrezno citiranje, avtorji pa pri tem pogosto ostajajo nemočni. K temu bi lahko z nudenjem ustrezne dokumentacije, izobraževanj in tehnične podpore veliko doprinesla nacionalna raziskovalna infrastruktura CLARIN.SI.

V prihodnje bi bilo zanimivo raziskavo razširiti na jezikovne vire s področja eksperimentalnega in računalniškega jezikoslovja, ki jezikovne vire uporabljajo kot podatkovne množice, zaradi česar se njihovi interesi, pa tudi potrebe razlikujejo od skupnosti, ki smo se jim posvetili v tej raziskavi. Prav tako načrtujemo dodatno analizo praks navajanja primerov v slovenskih znanstvenih publikacijah s področja jezikoslovja.

Zahvala

Avtorji se anonimnim recenzentom zahvaljujejo za zelo koristne pripombe. Raziskava, opisana v prispevku, je bila opravljena v okviru raziskovalnih infrastruktur za jezikovne vire in orodja CLARIN.SI in CLARIN ERIC.

6. Literatura

Špela Arhar Holdt in Kaja Dobrovoljc. 2016. Vrednost korpusa *Janes* za slovensko normativistiko. *Slovenščina 2.0*, 2: 1–37. http://slovenscina2.0.trojina.si/arhiv/2016/2/Slo2.0_2016_2_02.pdf. Zadnji dostop 10.4.2018.

Špela Arhar Holdt, Kaja Dobrovoljc in Iztok Kosem. 2016. Predstavitveni portal spletnih jezikovnih virov za slovenščino. V: T. Erjavec in D. Fišer, ur., *Zbornik konference Jezikovne tehnologije in digitalna humanistika*, str. 27–31. http://www.sdjt.si/wp/wp-content/uploads/2016/09/JTDH-2016_Arhar-et-al_Predstavitveni-portal-spletnih-jezikovnih-virov-za-slo.pdf. Zadnji dostop 20.4.2018.

Blanca Arias-Badia, Elisenda Bernal in Araceli Alonso. 2014. An online Spanish Learners' dictionary: the Daele project. *Slovenščina 2.0*, 2: 53–71. http://slovenscina2.0.trojina.si/arhiv/2014/2/Slo2.0_2014_2_05.pdf. Zadnji dostop 10.4.2018.

Júlia Bálint Čeh in Iztok Kosem. 2017. Prvi koraki do novega velikega slovensko-madžarskega slovarja: analiza relevantnih dvojezičnih virov. *Slovenščina 2.0,* 2. http://slovenscina2.0.trojina.si/arhiv/2017/2/Slo2.0_2017_2_06.pdf. Zadnji dostop 17.8.2018.

Andrea L. Berez-Kroeker, Lauren Gawne, Gary Holton, Susan Smythe Kung, Peter Pulsifer in Lauren B. Collister. The Data Citation and Attribution in Linguistics Group, & the Linguistics Data Interest Group. 2017. The Austin Principles of Data Citation in Linguistics (Različica 0.1). http://site.uit.no/linguisticsdatacitation/austinprinciples/. Dostop 15.4.2018.

Andrea L. Berez-Kroeker, Lauren Gawne, Susan Smythe Kung, Barbara F. Kelly, Tyler Heston, Gary Holton, Peter Pulsifer, David I. Beaver, Shobhana Chelliah, Stanley Dubinsky, Richard P. Meier, Nick Thieberger, Keren Rice in Anthony C. Woodbury. 2018. Reproducible research in linguistics: A position statement on data citation and attribution in our field. *Linguistics*, 56(1): 1–18. https://doi.org/10.1515/ling-2017-0032.

European Commission. 2012. Towards better access to scientific information: Boosting the benefits of public investments in research. http://ec.europa.eu/research/science-society/document_library/pdf_06/era-communication-towards-better-access-to-scientific-information_en.pdf. Dostop 13.8.2018.

Tomaž Erjavec. 2009. Odprtost jezikovnih virov za slovenščino. V: M. Stabej, ur., *Simpozij OBDOBJA 28*. http://centerslo.si/wp-content/uploads/2015/10/28-Erjavec.pdf. Dostop 13.8.2018.

Tomaž Erjavec, Darja Fišer, Simon Krek in Nina Ledinek. 2010. The JOS Linguistically Tagged Corpus of Slovene. V: *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*. http://www.lrec-conf.org/proceedings/lrec2010/summaries/139.html. Zadnji dostop 17.8.2018.

Melanija Larisa Fabčič. 2014. Mentalna podoba človeka v slovenskih, nemških in madžarskih primerjalnih frazemih. *Slavistična revija*, 62(2): 195–215. https://srl.si/sql_pdf/SRL_2014_2_05.pdf. Zadnji dostop 10.4.2018.

Martin Haspelmath. 2014. The Generic Style Rules for Linguistics. *Zenodo*. https://doi.org/10.5281/zenodo.253501.

Nataša Jakop. 2014. Leksikalizacija prostorskih razmerij v slovenščini: jezikovnopragmatični vidik. *Slavistična revija*, 62(3): 353–362. https://srl.si/sql_pdf/SRL_2014_3_08.pdf. Zadnji dostop 20.4.2018.

Adam Kilgarriff in Irene Renau. 2013. esTenTen, a vast web corpus of Peninsular and American Spanish. *Procedia-Social and Behavioral Sciences*, 95, 12-19. https://doi.org/10.1016/j.sbspro.2013.10.617.

Nikola Ljubešić, Marija Stupar, Tereza Jurić in Željko Agić. 2013. Combining Available Datasets for Building Named Entity Recognition Models of Croatian And Slovene. *Slovenščina 2.0*, 2. http://slovenscina2.0.trojina.si/arhiv/2013/2/Slo2.0_2013_2_03.pdf. Dostop 13.8.2018.

Nataša Logar Berginc, Miha Grčar, Marko Brakus, Tomaž Erjavec, Špela Arhar Holdt in Simon Krek. 2012. *Korpusi slovenskega jezika Gigafida, KRES, ccGigafida in ccKRES: gradnja, vsebina, uporaba*. Ljubljana: Trojina, zavod za uporabno slovenistiko; Fakulteta za družbene vede. https://www.fdv.uni-lj.si/docs/default-source/zalozba/pages-from-logar-et-al---korpusi.pdf?sfvrsn=2. Zadnji dostop 17.8.2018.

Nataša Logar, Polona Gantar in Iztok Kosem. 2014. Collocations and examples of use: a lexical-semantic approach to terminology. *Slovenščina 2.0*, 1: 41–61. http://slovenscina2.0.trojina.si/arhiv/2014/1/Slo2.0_2014_1_03.pdf. Zadnji dostop 10.4.2018.

Matej Meterc. 2013. Antonimija enako motiviranih paremioloških enot (primeri iz slovenščine in slovaščine). *Slavistična revija*, 61(2): 361–376. https://srl.si/sql_pdf/SRL_2013_2_02.pdf. Zadnji dostop 10.4.2018.

Janja Polajnar. 2013. Neprodani in trdni. Ja, seveda, potem pa svizec ... Osamosvajanje oglasnih sloganov v slovenskem jeziku. *Jezik in slovstvo*, 58(3): 3–19. https://www.jezikinslovstvo.com/pdf.php?part=2013|3|3%E2%80%9319. Zadnji dostop 10.4.2018.

Janja Ribič. 2016. Ujemanje med povedkom in osebkom v kopulativnih stavkih. *Jezik in slovstvo*, 61(2): 139–147. https://www.jezikinslovstvo.com/pdf.php?part=2016|2|139%E2%80%93147. Zadnji dostop 10.4.2018.

Jörg Tiedemann. 2009. News from OPUS – A Collection of Multilingual Parallel Corpora with Tools and Interfaces. V: Nicolov, N., Bontcheva, K., Angelova, G., Mitkov, R. (ur.): *Recent Advances in Natural Language Processing*, 5: 237–248. Amsterdam, Philadelphia: John Benjamins. http://stp.lingfil.uu.se/~joerg/published/ranlp-V.pdf. Zadnji dostop 17.8.2018.

Ada Vidovič Muha. 2015. Propozicija v funkcijski strukturi stavčne povedi – vprašanje besednih vrst (poudarek na povedkovniku in členku). *Slavistična revija*, 63(4): 389–406. https://srl.si/sql_pdf/SRL_2015_4_04.pdf. Zadnji dostop 10.4.2018.

Serge Sharoff. 2006. Creating general-purpose corpora using automated search engine queries. *WaCky! Working papers on the Web as Corpus*. Gedit, Bologna, http://wackybook.sslmit.unibo.it/pdfs/sharoff.pdf. Dostop 13.8.2018.

Janez Štebe, Sonja Bezjak in Irena Vipavc Brvar. 2015. Priprava raziskovalnih podatkov za odprt dostop. Priročnik za raziskovalce. Fakulteta za družbene vede, Založba FDV. https://www.dlib.si/details/URN:NBN:SI:DOC-06SLBVXX.

Saška Stumberger. 2015. Slovaropisna obravnava novejše leksike. *Slovene Linguistic Studies*, 10: 153–166. https://kuscholarworks.ku.edu/bitstream/handle/1808/18316/08_Stumberger.pdf. Zadnji dostop 15.4.2018.

Darinka Verdonik in Mirjam Sepesy Maučec. 2013. *Slovenščina 2.0*, 1. http://slovenscina2.0.trojina.si/arhiv/2013/1/Slo2.0_2013_1_06.pdf. Dostop 13.8.2018.

Darinka Verdonik, Tomaž Potočnik, Mirjam Sepesy Maučec in Tomaž Erjavec. 2016. *Spoken corpus Gos VideoLectures 1.0 (transcription)*. Slovenian language resource repository CLARIN.SI, http://hdl.handle.net/11356/1069.

Darinka Verdonik, Tomaž Potočnik, Mirjam Sepesy Maučec in Tomaž Erjavec. 2017. *Spoken corpus Gos VideoLectures 2.0 (transcription)*, Slovenian language resource repository CLARIN.SI, http://hdl.handle.net/11356/1158.

Mira Vončina. 2016. Zaključena znanstvena zbirka podatkov – primeri katalogizacije in Sicris vrednotenja. Delavnica ADP, 26.10. https://www.adp.fdv.uni-lj.si/adp_delavnica_okt2016/presentations/2016_MiraVoncina_Znanstvena_zbirka_podatkov.pdf. Zadnji dostop 7.9.2018.

Andreja Žele. 2014. Členki tudi kot vnašalniki novih prostorskih razmerij v obstoječe sporočilo. *Slavistična revija*, 62(3): 321–330. https://srl.si/sql_pdf/SRL_2014_3_05.pdf. Zadnji dostop 10.4.2018.

Andreja Žele. 2015. Konverzija v slovenščini. *Jezik in slovstvo*, 60(2): 65–77. https://www.jezikinslovstvo.com/pdf.php?part=2015|2|65%E2%80%9377. Zadnji dostop 15.4.2018.