# CITIRANJE JEZIKOVNIH PODATKOV V SLOVENSKIH ZNANSTVENIH OBJAVAH V OBDOBJU 2013–2019

Jakob LENARDIČ
Univerza v Ljubljani, Filozofska fakulteta

Tomaž ERJAVEC
Institut Jožef Stefan

Darja FIŠER
Univerza v Ljubljani, Filozofska fakulteta, Institut Jožef Stefan

*Lenardič, J., Erjavec, T., Fišer, D. (2020): Citiranje jezikovnih podatkov v slovenskih znanstvenih objavah v obdobju 2013–2019. Slovenščina 2.0, 8(1): 1–34.*

*DOI:* https://doi.org/10.4312/slo2.0.2020.1.1-34

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 ter pomembno vplivajo na njihovo priznanost in s tem možnost pridobivanja projektov ter zaposlitev. V prispevku najprej predstavimo ti. »austinska načela« citiranja jezikovnih podatkov in opišemo tovrstne aktivnosti v sklopu infrastrukture CLARIN.SI. Nato analiziramo stanje citiranja jezikovnih podatkov, predvsem korpusov, v šestih vodilnih slovenskih jezikoslovnih znanstvenih revijah (*Jezik in slovstvo*, *Slavistična revija*, *Slovenščina 2.0*, *Linguistica*, *Slovene Linguistic Studies* in *Jezikoslovni zapiski*) ter v zbornikih dveh znanstvenih konferenc z jezikoslovno tematiko (*Jezikovne tehnologije in digitalna humanistika* ter *Obdobja*) za obdobje zadnjih sedmih let, tj. 2013–2019. Pregledali smo 1.074 znanstvenih objav in kvantitativno ter kvalitativno analizirali rezultate. S kvantitativnega vidika pokažemo, da v celotnem obdobju zgolj dobra

četrtina pregledanih člankov vključuje rabo virov ter da je v poznejšem obdobju (2018–2019) raba virov v objavah več kot dvakrat pogostejša kot v zgodnejšem obdobju (2013–2017). Načine navajanja virov razvrstimo v pet kategorij (npr. *navajanje hiperpovezave na vir v besedilu* ter *navajanje ključne publikacije o viru*); pokažemo, da je raba posameznega načina v veliki meri odvisna od navodil avtorjem za posamezno publikacijo. S kvalitativnega vidika se osredotočamo predvsem na vire z vnosom v repozitoriju raziskovalne infrastrukture CLARIN.SI, za katere pokažemo, da so z redkimi izjemami neustrezno citirani. Izsledke povzamemo in po ti. »austinskih načelih« pokažemo, kaj je bilo že narejenega v sklopu infrastrukture CLARIN.SI ter predlagamo smernice za citiranje jezikoslovnih podatkov in načine za njihovo implementacijo.

**Ključne besede:** odprta znanost, citiranje raziskovalnih podatkov, jezikovni viri, austinska načela, slovenske revije in zborniki

## 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] Med poglavitnimi cilji *Akcijskega načrta* je bil pilotni program *Odprti dostop do raziskovalnih podatkov v letih 2017—2020*, ki sicer v tem obdobju nikoli ni bil celotno izveden, si je pa prizadeval 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

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

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.

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,[7] v sklopu katere je bil vzpostavljen certificiran repozitorij, ki trenutno arhivira več kot sto petdeset odprto dostopnih jezikovnih virov.

V pričujočem prispevku nas zanima, kako se uporaba jezikovnih virov citira v znanstvenih člankih vodilnih slovenskih jezikoslovnih publikacij. O stanju citiranja jezikovnih virov smo pred več kot desetimi leti zapisali naslednje:

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 repozitorij CLARIN.SI ponuja dolgoročno hrambo virov, ob tem pa za vsak vnesen vir na samem vrhu njegove spletne strani navaja, kako naj se ga citira. Dostop do podatkov v certificiranih repozitorijih, kot je CLARIN.SI, je v skladu s ti. austinskimi načeli (Berez-Kroeker idr., 2017) za ustrezno citiranje v jezikoslovju, ki so povzeti v dokumentu *The FORCE11 Joint Declaration of Data Citation Principles* (Data Citation Synthesis Group, 2017).[8] Poleg tega, da natančna navodila za citiranje jezikovnih virov sledijo drugi točki 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

4 http://nl.ijs.si/ME/

5 http://nl.ijs.si/jos/

6 http://www.slovenscina.eu/

7 https://www.clarin.si/repository/xmlui/

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

zagotavljanje odprtega dostopa in s tem interoperabilnosti, trajnosti in preverljivosti podatkov.

Pričujoči prispevek je razširitev konferenčnega prispevka Fišer, Lenardič in Erjavec (2018), v katerem smo pregledali, kako so slovenski raziskovalci citirali jezikovne vire v šestih vodilnih jezikoslovnih revijah (*Jezik in slovstvo*, *Slavistična revija*, *Slovenščina 2.0*, *Slovene Linguistic Studies*, *Linguistica* in *Jezikoslovni zapiski*) ter zbornikih dveh znanstvenih konferenc z jezikoslovno tematiko (*Jezikovne tehnologije in digitalna humanistika* ter *Obdobja*), ki so izšli v obdobju med 2013 ter 2017. V tem prispevku pregled razširimo z analizo objav v zadnjih dveh letih, tj. 2018 in 2019.

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 idr. (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 s podatki upravlja, da sploh lahko postanejo odprti, kot tudi, kako to delo primerno ovrednotiti. Zadnje priporočilo pa je neposredno posvečeno boljšemu citiranju raziskovalnih podatkov, kjer avtorji prispevka urednikom ter založnikom znanstvenih revij in knjig svetujejo uvedbo konkretnih politik tako za izmenjavo podatkov kot za njihovo citiranje, pri slednjem tako, da izoblikujejo 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 sedem let (2013–2019), pri zbornikih pa navodila za avtorje oz. predloge prispevkov ter štiri zbornike dveh konferenc z jezikoslovno tematiko. Med zborniki smo v študijo zajeli *Jezikovne tehnologije in digitalna humanistika (JTDH) 2016*, *JTDH 2018*, *Obdobja 2016* ter *Obdobja 2019*, med revijami pa *Linguistica*, *Jezik in slovstvo*, *Jezikoslovni zapiski*, *Slavistična revija*, *Slovene Linguistic Studies* in *Slovenščina 2.0*.[9]

9 V primeru revije *Linguistica* nismo upoštevali 58. zvezka iz leta 2018, saj so vsi članki v njem napisani v francoščini, nihče od avtorjev pričujočega prispevka pa ni kompetenten govorec tega jezika.

Analiza citiranja je bila opravljena dvakrat (Tabela 1). Prvi pregled, ki smo ga že predstavili v Fišer, Lenardič in Erjavec (2018), je bil opravljen za obdobje 2013–2017, v katerem smo pregledali 751 znanstvenih prispevkov, od katerih jih vire omenja 133 oz. dobrih 17 %. Temu smo nato dodali še pregled za obdobje 2018–2019, v katerem smo pregledali 323 objav, od katerih jih vire omenja 155 oziroma 48 %. Skupaj smo torej pregledali 1.074 znanstvenih objav iz obdobja 2013–2019, od katerih jezikovne vire navaja 288 člankov (27 %).

Navedbe virov v pregledanih prispevkih ločujemo na naslednje kategorije:

**Povezava**[10] **na vir v besedilu prispevka** (največkrat v opombi). Zgled takega citiranja je v Žele (2014) v *Slavistični reviji*, kjer avtorica povezavo na korpus *Gigafida* podaja v opombi. Prispevki v tovrstni kategoriji ne navajajo ključne publikacije o viru, ki bi v tem primeru bila Logar Berginc idr. (2012).

**Povezava na vir v bibliografiji**. Zgled takega citiranja je v Trivunović (2019) v reviji *Jezikoslovni zapiski*, kjer avtorica navaja korpusa *Gigafida 2.0* in *IMP* tako, da podaja hiperpovezavo v končnem seznamu virov, ne navaja pa ključnih publikacij o viru, se pravi Logar Berginc idr. (2012) za *Gigafida 2.0* ter Erjavec (2015a) za *IMP*.

**Povezava na vir v besedilu prispevka kot tudi v bibliografiji**. Zgled takega citiranja je Žele (2018) v reviji *Jezik in slovstvo*, kjer avtorica podaja povezavo na korpus *Gigafida* v opombi[11] ter v končnem seznamu virov. Tudi prispevki v tej kategoriji ne navajajo ključne publikacije o viru.

**Publikacija o viru**. Zgled takega citiranja je v Ljubešić, Miličević Petrović in Samardžić (2019) v *Slavistični reviji*, kjer avtorji za orodje *TweetCat* navajajo ključno publikacijo o viru, tj. Ljubešić, Fišer in Erjavec (2014).

**Povezava na vir v besedilu prispevka in publikacija o viru**. Zgled takega citiranja je v Bálint Čeh in Kosem (2017). Avtorja podajata povezavo na

10 V to kategorijo vključujemo tudi navedbe stalnih spletnih identifikatorjev, kot sta handle in DOI.

11 Pri tem prispevku je zanimivo, da avtorica za korpus *Gigafida* navaja hiperpovezavo na demonstracijsko različico korpusa, tj. http://demo.gigafida.net, ki pa že od leta 2014 ni dostopna, medtem ko v bibliografiji navaja običajnejšo povezavo na spletno stran http://www.gigafida.net.

korpus *Gigafida* v opombi in navajata ključno publikacijo, tj. Logar Berginc idr. (2012).

**Kombinacija različnih načinov navajanja virov**. Zgled takega navajanja je v Žitnik in Bajec (2018) v *Slovenščina 2.0*. Avtorja navajata korpus *ssj500k 1.4* z medbesedilno navedbo avtorjev korpusa (tj. Krek idr. 2015) ter s trajnim identifikatorjem v pripadajočem zapisu v bibliografiji – v razdelku 3.4 bomo videli, da je to najustreznejši način navedbe jezikovnega vira v primeru, da ima vir vnos v jezikovnem repozitoriju. Nadalje pa Žitnik in Bajec (2018) navajata ključni publikaciji za orodji SkipCor (tj. Žitnik, Šubelj in Bajec, 2014) ter nutIE (Žitnik idr., 2017), vendar za vira ne podajata povezav, saj nobeno orodje ni prosto dostopno.

**Brez navedbe hiperpovezave na vir ali ključne publikacije o viru.** 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 (tj. Erjavec idr., 2010).

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

| | 2013–2017 | | 2018–2019 | | Σ | |
|---|---|---|---|---|---|---|
| Vseh objav | 751 | | 323 | | 1.074 | |
| Objave z omembo vira | 133 | 100 % | 155 | 100 % | 288 | 100 % |
| Povezava na vir v besedilu prispevka | 13 | 10 % | 11 | 7 % | 24 | 8 % |
| Povezava na vir v bibliografiji | 33 | 25 % | 39 | 25 % | 72 | 25 % |
| Povezava na vir v besedilu in v bibliografiji | 14 | 11 % | 17 | 13 % | 31 | 11 % |
| Publikacija o viru | 16 | 12 % | 11 | 7 % | 27 | 9 % |
| Povezava na vir v besedilu prispevka in publikacija o viru | 8 | 6 % | 13 | 8 % | 21 | 7 % |
| Kombinirano | 25 | 18 % | 24 | 15 % | 49 | 17 % |
| Brez | 25 | 18 % | 40 | 25 % | 65 | 23 % |

### 3.2 Pregled navodil avtorjem

Ker smo sklepali, da je od navodil avtorjem, ki so jih pripravili uredniški odbori revij in programski odbori konferenc, močno odvisno, kako bodo avtorji navajali vire, jih povzemamo v tem razdelku. Za revijo *Jezikoslovni zapiski* ter za zbornik *Obdobja* navodil avtorjem na njihovih spletnih straneh nismo našli.

Najpodrobnejša navodila za navajanje virov podaja revija *Slovenščina 2.0,*[12] ki ločuje navajanje korpusov, spletnih strani in spletnih virov:

Korpus:

- Gigafida. Dostopno prek: http://www.gigafida.net (datum dostopa).
- Cambridge English Corpus. 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,[13] da vire in literaturo navajajo ločeno, kar se nam zdi dobra praksa, saj s tem avtorjem med drugim sporočajo, da sta uporaba in navajanje virov pomemben sestavni del znanstvenega prispevka. Dodatno velja omeniti, da poziv k ločenemu navajanju jezikovnih virov omogoča bralcem lažji dostop in preverjanje citiranih podatkov, ki podpirajo neko znanstveno trditev, kar je skladno tudi z austinskimi načeli (glej razdelek 4). Podrobneje ta revija načina

12 http://slovenscina2.0.trojina.si/si/oddaja-prispevkov/

13 https://www.jezikinslovstvo.com/02.php

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*:[14]

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

*Slavistična revija*[15] v navodilih za oblikovanje seznama literature uvaja zelo nenatančno navajanje spletnih virov, brez navedbe spletnih povezav, verzij ali 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 vira se referenca nanaša, saj je Wikipedija kolaborativen projekt, kjer uredniki gesla lahko ves čas spreminjajo, zato bi bilo nujno treba dodati datum dostopa. Pri drugem primeru pa ni jasno, ali gre za referenco na doktorsko disertacijo kot publikacijo ali za jezikovni vir, ki je bil v okviru disertacije razvit. Predvsem pa nobena od referenc 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 v slovenskem okolju uvršča v sam vrh.

Revija *Slovene Linguistic Studies*[16] posebej za navajanje elektronskih virov ne podaja navodil.

Podobno zbornik *JTDH*[17] v predlogi prispevkov sicer vsebuje primer dodajanja hiperpovezav v opombe in navaja načine navajanja različnih tipov enot

14 https://revije.ff.uni-lj.si/linguistica/about/submissions#authorGuidelines

15 https://srl.si/navodila_guidelines.pdf

16 https://sjsls.byu.edu/guidelines-for-contributors/

17 http://www.sdjt.si/wp/dogodki/konference/jtdh-2018/#navodila

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 288 (27 %) 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 (17 %), kar kaže na neupoštevanje navodil avtorjem oz. na pomanjkljiva navodila.

**Tabela 2:** *Pregled praks navajanja jezikovnih virov v ključnih slovenskih znanstvenih revijah za področje jezikoslovja za obdobje 2013–2017 ter za obdobje 2018–2019*

| | *Jezik in slovstvo* | | | | | | *Slavistična revija* | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 2013–17 | | 2018–19 | | Σ | | 2013–17 | | 2018–19 | | Σ | |
| Vse objave | 157 | 100 % | 52 | 100 % | 209 | 100 % | 180 | 100 % | 78 | 100 % | 258 | 100 % |
| Omemba vira | 11 | 7 % | 17 | 33 % | 28 | 13 % | 14 | 8 % | 39 | 50 % | 53 | 21 % |
| URL na vir v besedilu | 0 | 0 % | 0 | 0 % | 0 | 0 % | 1 | 7 % | 2 | 5 % | 3 | 6 % |
| URL na vir v biblio. | 4 | 36 % | 6 | 34 % | 10 | 36 % | 2 | 14 % | 4 | 10 % | 6 | 11 % |
| URL na vir v besedilu in biblio. | 2 | 18 % | 3 | 18 % | 5 | 18 % | 0 | 0 % | 0 | 0 % | 0 | 0 % |
| Publikacija o viru | 2 | 18 % | 2 | 12 % | 4 | 14 % | 0 | 0 % | 3 | 8 % | 3 | 6 % |
| URL na vir in publ. | 0 | 0 % | 2 | 12 % | 2 | 7 % | 0 | 0 % | 0 | 0 % | 0 | 0 % |
| Kombinirano | 0 | 0 % | 1 | 6 % | 1 | 4 % | 0 | 0 % | 2 | 5 % | 2 | 4 % |
| Brez | 3 | 27 % | 3 | 18 % | 6 | 21 % | 11 | 79 % | 28 | 72 % | 39 | 73 % |

| | *Slovenščina 2.0* | | | | | | *Slovene Linguistic Studies* | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 2013–17 | | 2018–19 | | Σ | | 2013–17 | | 2018–19 | | Σ | |
| Vse objave | 45 | 100 % | 16 | 100 % | 61 | 100 % | 26 | 100 % | 9 | 100 % | 35 | 100 % |
| Omemba vira | 34 | 76 % | 9 | 56 % | 43 | 70 % | 8 | 31 % | 5 | 56 % | 13 | 37 % |
| URL na vir v besedilu | 5 | 15 % | 0 | 0 % | 5 | 12 % | 2 | 25 % | 0 | 0 % | 2 | 15 % |
| URL na vir v biblio. | 5 | 15 % | 1 | 11 % | 6 | 14 % | 2 | 25 % | 3 | 60 % | 5 | 38 % |
| URL na vir v besedilu in biblio. | 3 | 9 % | 1 | 11 % | 4 | 9 % | 3 | 37 % | 0 | 0 % | 3 | 23 % |
| Publikacija o viru | 6 | 18 % | 1 | 11 % | 7 | 16 % | 0 | 0 % | 0 | 0 % | 0 | 0 % |
| URL na vir in publ. | 6 | 18 % | 1 | 11 % | 7 | 16 % | 0 | 0 % | 0 | 0 % | 0 | 0 % |
| Kombinirano | 8 | 23 % | 4 | 45 % | 12 | 28 % | 1 | 13 % | 1 | 20 % | 2 | 15 % |
| Brez | 1 | 3 % | 1 | 11 % | 2 | 5 % | 0 | 0 % | 1 | 20 % | 1 | 8 % |
| | *Linguistica* | | | | | | *Jezikoslovni zapiski* | | | | | |
| | 2013–17 | | 2018–19 | | Σ | | 2013–17 | | 2018–19 | | Σ | |
| Vse objave | 134 | 100 % | 24 | 100 % | 158 | 100 % | 115 | 100 % | 52 | 100 % | 167 | 100 % |
| Omemba vira | 6 | 4 % | 12 | 50 % | 18 | 11 % | 20 | 17 % | 17 | 33 % | 37 | 22 % |
| URL na vir v besedilu | 0 | 0 % | 0 | 0 % | 0 | 0 % | 3 | 15 % | 2 | 12 % | 5 | 13 % |
| URL na vir v biblio. | 4 | 66 % | 3 | 25 % | 7 | 39 % | 7 | 35 % | 10 | 58 % | 17 | 46 % |
| URL na vir v besedilu in biblio. | 1 | 17 % | 5 | 42 % | 6 | 33 % | 2 | 10 % | 2 | 12 % | 4 | 11 % |
| Publikacija o viru | 0 | 0 % | 0 | 0 % | 0 | 0 % | 1 | 5 % | 0 | 0 % | 1 | 3 % |
| URL na vir in publ. | 0 | 0 % | 0 | 0 % | 0 | 0 % | 1 | 5 % | 1 | 6 % | 2 | 5 % |
| Kombinirano | 0 | 0 % | 0 | 0 % | 0 | 0 % | 0 | 0 % | 1 | 6 % | 1 | 3 % |
| Brez | 1 | 17 % | 4 | 33 % | 5 | 28 % | 6 | 30 % | 1 | 6 % | 7 | 19 % |

V Tabeli 2 navajamo rezultate analize za posamezne revije, ki smo jih vključili v raziskavo. Za celotno obdobje pregleda 2013–2019 vsebuje najvišji delež prispevkov, ki omenjajo jezikovne vire, revija *Slovenščina 2.0* (70 %), najnižjega pa revija *Linguistica* (11 %). Pomembno je, da je pri vseh revijah (razen pri *Slovenščina 2.0*) v drugem obdobju pregleda – tj. 2018–2019 – bistveno večji delež objav, ki temeljijo na rabi jezikovnih virov, kot v prvem obdobju pregleda 2013–2017. Največjo rast opazimo pri reviji *Linguistica,* v kateri v

prvem obdobju 2013–2017 zgolj 4 % vseh objav vključuje rabo jezikovnih virov, medtem ko v drugem obdobju 2018–2019 ta delež zraste na polovico vseh prispevkov. Podobno opazimo pri *Slavistični reviji*, kjer v obdobju 2013–2017 rabo virov vključuje zgolj 8 % objav, medtem ko je ta delež v 2018–2019 tudi tu polovica vseh objav. Čeprav se zdi, da je ta dvig vsaj delno posledica dejstva, da so posamezne številke pregledanih revij v drugem obdobju v splošnem bolj osredotočene na empirično jezikoslovje kot v prvem,[18] je porast verjetno vseeno indikator vedno močnejše težnje k rabi jezikovnih virov, saj smo ga identificirali pri skoraj vseh vodilnih slovenskih jezikoslovnih revijah, ki so programsko precej raznolike.

Najbolj homogeno navajanje virov je v reviji *Linguistica*, kjer smo identificirali zgolj dva različna načina citiranja (od šestih možnih): povezava na vir zgolj v bibliografiji ter povezava na isti vir tako v besedilu kot v bibliografiji. Najbolj heterogeni načini citiranja so v revijah *Jezikoslovnih zapiskih* ter *Slovenščina 2.0*, kjer najdemo vseh pet načinov navajanja virov. Najvišji delež navedbe vira v obliki hiperpovezave na spletno stran vira, podane v bibliografiji, najdemo v reviji *Jezikoslovni zapiski* (46 %), najvišji delež citiranja ključnega prispevka o viru pa pripada reviji *Slovenščina 2.0* (16 %). Od posameznih načinov navajanja virov je navajanje povezav na vir v bibliografiji prispevka najpogostejši način navajanja virov v vseh revijah, razen v reviji *Slovenščina 2.0*, kjer je nekoliko pogostejše kombinirano citiranje.

Po nenavajanju uporabljenih virov izrazito izstopa *Slavistična revija*, v kateri avtorji pri skoraj tri četrtini (73 %) prispevkov, ki rabo virov omenjajo, teh virov nikjer ne citirajo. Zanimivo je tudi, da delež člankov v *Slavistični reviji* brez ustrezne navedbe ostaja približno enak v obeh analiziranih obdobjih, kar poleg tega, da skoraj vsaka revija dosledno ohranja svoj prevladujoči način citiranja v obeh obdobjih (npr. kombinirano citiranje pri *Slovenščina 2.0* ter navedbe povezav zgolj v bibliografiji in nič primerov objav z navedbo med besedilom pri *Jeziku in slovstvo*), daje misliti, da so načini citiranja

18 Na primer 59. številka revije *Linguistica* iz 2018, ki je bila edina pregledana številka te revije za obdobje 2018–2019, se je osredotočala na nemško aplikativno jezikoslovje, ki pogosto temelji na rabi jezikovnih virov, medtem ko so se domala vse številke v obdobju 2013–2017 osredotočale na različne veje teoretičnega jezikoslovja (npr. generativna slovnica), ki so inherentno manj vezane na preveritvi jezikovnih podatkov v sodobnih korpusih.

precej odvisni od navodil avtorjem v posamezni reviji. V prejšnjem razdelku smo na primer videli, da *Slavistična revija* podaja zelo nespecifična navodila za navajanje jezikovnih virov, ki zahtevajo precej neuporabno oznako *Na spletu* namesto dejanskih hiperpovezav, kar tako najbrž botruje visokemu deležu neustreznih oziroma manjkajočih navedb (o problematiki hiperpovezav glej tudi razdelek 3.4). 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.

**Tabela 3:** *Pregled praks navajanja jezikovnih virov v ključnih konferenčnih zbornikih za področje jezikoslovja*

| | *Obdobja* | | | | | | *JTDH* | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 2016 | | 2019 | | Σ | | 2016 | | 2018 | | Σ | |
| Vse objave | 64 | 100 % | 56 | 100 % | 120 | 100 % | 30 | 100 % | 36 | 100 % | 66 | 100 % |
| Omemba vira | 12 | 19 % | 23 | 41 % | 35 | 29 % | 28 | 93 % | 33 | 92 % | 61 | 92 % |
| URL na vir v besedilu | 0 | 0 % | 1 | 4 % | 1 | 3 % | 2 | 7 % | 6 | 18 % | 8 | 13 % |
| URL na vir v biblio. | 7 | 58 % | 13 | 56 % | 20 | 57 % | 2 | 7 % | 0 | 0 % | 2 | 3 % |
| URL na vir v besedilu in biblio. | 0 | 0 % | 2 | 9 % | 2 | 6 % | 3 | 11 % | 3 | 9 % | 6 | 10 % |
| Publikacija o viru | 1 | 8 % | 1 | 4 % | 2 | 6 % | 6 | 21 % | 4 | 12 % | 10 | 16 % |
| URL na vir in publ. | 0 | 0 % | 2 | 9 % | 2 | 6 % | 1 | 4 % | 7 | 21 % | 8 | 13 % |
| Kombinirano | 2 | 17 % | 2 | 9 % | 4 | 11 % | 13 | 46 % | 13 | 39 % | 26 | 43 % |
| Brez | 2 | 17 % | 2 | 9 % | 4 | 11 % | 1 | 4 % | 0 | 0 % | 1 | 2 % |

V Tabeli 3 navajamo rezultate za konferenci, ki smo ju vključili v raziskavo. V zborniku *JTDH 2016* so viri omenjeni v 93 % vključenih prispevkov, v zborniku *JTDH 2018* pa v 92 % prispevkov, kar je glede na področje konference razumljivo. V obeh edicijah zbornika naletimo na izrazito velik delež prispevkov (vse skupaj 43 %), v katerih avtorji uporabljajo različne kombinacije navajanja virov. To je verjetno odraz heterogene raziskovalne skupnosti, ki se predstavlja na tej konferenci, ter heterogenosti dostopnosti virov – npr., nekateri viri so deponirano v trajnih repozitorijih, ki beležijo tudi ključne

publikacije, medtem ko drugi niso dostopni in so tako navedeni zgolj s ključno publikacijo.

V zborniku *Obdobja* je prispevkov z omembo jezikovnega vira leta 2016 19 %, leta 2019 pa 41 %. Glede na to, da sta bili ediciji simpozija iz 2016 ter 2019 tematsko vezani na jezikovni opis sodobne slovenščine, se zdi ta rezultat nizek. Vendar sta odstotka kljub temu tako v 2016 kot tudi v 2019 občutno višja od večine programsko sorodnih revijah v istem obdobju, 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 Berginc, Gantar in Kosem (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, saj bralcu zagotavlja, da bo lahko dostopal do vira, četudi se naslov njegove 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.

Za obdobje 2018–2019 smo v vseh šestih revijah v Tabeli 2 posebej opazovali, kako avtorji navajajo jezikovne vire, ki imajo vnos v repozitoriju CLARIN.SI, ki avtorjem omogoča, da se na vire sklicujejo z navedbo handlov. Tovrstnih člankov je 19 (od 155, se pravi dobrih 12 %), pri čemer gre v večini primerov za avtocitate. Na tem mestu jih za vsako revijo posebej navedemo skupaj z viri, ki so citirani:

- *Jezik in slovstvo* (2 objavi):
  - Rozman idr. (2018), ki se sklicujejo na korpus *Šolar* (Rozman idr., 2013);
  - Zwitter Vitez (2018), ki se sklicuje na korpus govorjene slovenščine *GOS* (Zwitter Vitez idr., 2013).
- *Slavistična revija* (3 objave):
  - Dobrovoljc (2018a), ki se sklicuje na korpus *GOS*;
  - Orel (2019), ki se sklicuje na korpus *GOS* ter na družino korpusov stare slovenščine *IMP* (Erjavec, 2014);
  - Marvin idr. (2019), ki se sklicujejo na leksikalno bazo *SNABI* (Kačič idr. 2002) ter na korpus *ccKres* (Logar Berginc idr., 2013).
- *Slovene Linguistic Studies* (2 objavi):
  - Jelovšek in Erjavec (2019), ki navajata referenčni korpus stare slovenščine *goo300k 1.2* (Erjavec, 2015b);
  - Krvina (2019), ki uporablja korpus *IMP* ter korpus *ssj500k 2.2* (Krek idr., 2019).
- *Slovenščina 2.0* (5 objav):
  - Žitnik in Bajec (2018), ki navajata korpus *ssj500k 1.4* (Krek idr., 2015);

    - Arhar Holdt in Čibej (2018), ki navajata morfološki leksikon *Sloleks* (Dobrovoljc idr., 2019);
    - Dobrovoljc (2018b), ki uporablja korpus *GOS*;
    - Pisanski Peterlin in Mikolič Južnič (2018), ki uporabljata *GOS*;
    - Pori in Kosem (2018), ki uporabljata vir *Leksikalna baza za slovenščino* (Gantar idr., 2012).
- *Jezikoslovni zapiski* (6 objav)*:*
    - Stramljič Breznik (2018), ki uporablja leksikon *Sloleks*;
    - Furlan (2018), ki uporablja korpus *IMP*;
    - Uhlik in Žele (2018), ki uporabljata *IMP*;
    - Atelšek (2019), ki uporablja *IMP*;
    - Hudeček in Mihaljević (2019), ki uporabljata spletni hrvaški korpus *hrWaC* (Ljubešić in Klubička, 2016);
    - Trivunović (2019), ki navaja korpus *IMP*.
- *Linguistica* (1 objava):
    - Petrič (2019), ki navaja korpus *GOS*.

Zgolj v treh (16,7 %) od zgoraj navedenih 18 objav je jezikovni vir naveden s povezavo handle na CLARIN.SI vnos, v katerem so, kot rečeno, sistematsko beleženi pomembni metapodatki, kot je avtorstvo vira. Te tri objave so:

- Jelovšek in Erjavec (2019) v *Slovene Linguistic Studies*;
- Žitnik in Bajec (2018) v *Slovenščina 2.0* ter
- Arhar Holdt in Čibej (2018) v *Slovenščina 2.0*.

Preostali avtorji običajno navajajo omenjene jezikovne vire zgolj s hiperpovezavo na projektno stran (namesto na trajni vnos v repozitoriju) ali na konkordančnik, s katerim so iskali po viru. Najpogosteje je neustrezno naveden korpus *IMP*, ki je namesto z ustreznejšo navedbo na Erjavec (2014) običajno citiran zgolj s podano hiperpovezavo na (ne nujno trajno) projektno stran, ki

jo gosti Inštitut »Jožef Stefan«. Variante slednjega (manj ustreznega) citiranja v bibliografiji med drugim opazimo pri Furlan (2018), Uhlik in Žele (2018) ter Atelšek (2019) v *Jezikoslovnih zapiskih*:

- IMP: korpus starejše slovenščine <http://nl.ijs.si/imp/, dostop xx.yy.201z>

Že prej smo omenili, da neoptimalno navajanje tovrstnih virov verjetno nastaja kot delna posledica obstoječih navodil avtorjem. To je predvsem razvidno iz prispevka Marvin idr. (2019) v *Slavistični reviji*, v katerem se avtorji sicer sklicujejo na leksikalno bazo *SNABI* z ustrezno navedbo Kačič idr. (2002) v repozitoriju CLARIN.SI, vendar je v bibliografiji namesto ustrezne povezave handle (tj. http://hdl.handle.net/11356/1051) navedena že prej omenjena oznaka *Na spletu* na sledeč način (Marvin idr., 2019, str. 549):

- Zdravko Kačič, Bogomir Horvat, Aleksandra Markuš Zögling, Robert Veronik, Matej Rojc, Andrej Žgank, Mirjam Sepesy Maučec in Tomaž Rotovnik, 2002: SNABI Database for Continuous Speech Recognition 1.2. Slovenian language resource repository CLARIN.SI. Na spletu.

Da tovrstna navedba brez handla nastaja neposredno zaradi navodil avtorjem, potrjuje konferenčna objava Marvin idr. (2018) iz zbornika *JTDH 2018*, ki je predhodna različica objave Marvin idr. (2019) v *Slavistični reviji*. Namreč, konferenčna objava Marvin idr. (2018) vsebuje isto referenco na jezikovni vir avtorjev Kačič idr. (2002), citiran na enak način kot v zgornjem zgledu iz *Slavistične revije*, vendar tokrat z navedbo ustreznega handla, tj. http://hdl.handle.net/11356/1051. Zaradi tovrstnih zgledov predlagamo, da bi bilo v obstoječih revijah treba navodila avtorjem razširiti – zdi se, da bi do velikega napredka prišlo že samo, če bi uredništvo nekaj besed namenilo dobri praksi citiranja, predvsem pomembnosti navajanja virov s trajnimi identifikatorji (če le-ti obstajajo) in hkratno navedbo avtorjev vira (o pomembnosti vključitve te informacije glej razdelek 4.2), saj tovrstna praksa trenutno ni opisana v navodilih nobene izmed pregledanih publikacij.

Poleg problematičnega citiranja virov s trajnim vnosom v repozitoriju smo naleteli tudi na druge težave, 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:
    - 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 *Jezikoslovni zapiski* 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

Zastareli viri: Rath (2019) v *Slovene Linguistic Studies,* Stopar in Ilc (2019) v *Slavistični reviji* ter *Kulčar* (2018) v *Jezikoslovnih zapiskih* navajajo jezikovne zglede iz korpusa FidaPLUS, kar je nenavadno, saj je ta korpus že od leta 2012 v celoti vključen v Gigafida. Rath (2019) za ta korpus navaja hiperpovezavo http://www.fidaplus.net/, ki pa ne deluje že od leta 2016.

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. Velja poudariti, da repozitorij *CLARIN.SI* rešuje ta problem tako, da je v navodilih za navajanje virov, ki so podana kot prva informacija v glavi vnosa za posamezen vir, jasno izpostavljeno, za katero različico vira gre in ali je ta različica dostopna preko spletnega konkordančnika. Za starejše različice repozitorij opozori o morebitni zastarelosti podatkov. Kot primer navajamo vnos za drugo različico korpusa *Gos VideoLectures (Transcriptions)* (Verdonik idr., 2017), ki je dostopna preko

konkordančnika *KonText*,[19] medtem ko prva različica (Verdonik idr., 2016) preko tega konkordančnika ni dostopna.

V reviji *Slovenščina 2.0* smo opazili nenatančno navajanje hiperpovezave do korpusa. Npr. v Arias-Badia, Bernal in Alonso (2014), kjer je za španski korpus navedena generična povezava na konkordančnik 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 izdelani v več različicah in vsebujejo različno gradivo. Ko smo korpus želeli preveriti v konkordančniku SketchEngine, na katerega nas referenca napoti, ga nismo našli, saj konkordančnik na dan preverjanja[20] ponuja dva španska korpusa tega tipa: Spanish Web Corpus oz. Spanish WaC (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 podatkovne zbirke. 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

19 https://www.clarin.si/kontext/first_form?corpname=gos_vl

20 https://www.sketchengine.eu (15. 4. 2018)

zelo 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 »austinskim načelom« citiranja podatkov v jezikoslovju (Berez-Kroeker idr., 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 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 bi njihovi profesorji oz. mentorji morali vztrajati pri korektnem citiranju podatkov v seminarskih nalogah, zaključnih delih in znanstvenih objavah. Drugi ukrep bi, kot predlagajo Berez-Kroeker idr. (2018), morali izvesti uredniški odbori revij in programski odbori konferenc 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, Bezjak in Vipavc Brvar, 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, str. 13; naš poudarek).

Tu so ključni naslovniki *Slavistična revija* kot revija s posebnega seznama ARRS, konferenca oz. monografija *Obdobja*, kot tudi konferenca *Jezikovne tehnologije in digitalna humanistika*, ki bi na tem področju v skladu s svojim poslanstvom morala orati ledino, podobno, kot to že vrsto let počne mednarodna konferenca LREC v okviru iniciative *Identificiraj, opiši in deli*

*svoj jezikovni vir*, ki je od 2010 obvezni del postopka oddaje konferenčnega prispevka in so jo že prevzele tudi nekatere druge mednarodne konference. Postopek identifikacije poteka s pomočjo uporabe trajnega enkratnega identifikatorja International Standard Language Resource Number (ISLRN, www.islrn.org). Od 2014 imajo avtorji prispevkov na konferenci LREC poleg objave prispevka tudi možnost objave razvitega jezikovnega vira v repozitoriju »LRE Map«.[21] 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 in 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, in sicer prek predlogov raziskovalni infrastrukturi Osrednjih specializiranih informacijskih centrov (OSIC).[22] 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)«.[23] Trenutno je na tem seznamu samo Arhiv družboslovnih podatkov (ADP). Pomembne objave v ADP tako 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 OSIC zahtevalo

21 http://lremap.elra.info/

22 https://www.arrs.si/sl/infra/osic/predstavitev.asp

23 http://home.izum.si/COBISS/bibliografije/Kateg-znan-zbirke.html

tudi podrobnejša navodila za vnašanje virov, kot tudi ustanovitev komisije za vrednotenje vnesenih virov. Vse to pa seveda tudi zahteva precejšen vložek dela in s tem financiranje CLARIN.SI.

#### 4.3 Dokazi

*V znanstvenih objavah bi morali biti podatki ustrezno citirani povsod, kjer neka trditev sloni na podatkih.*

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 prav tako podpira izvoz metapodatkov po shemi Dublin Core, ki jih žanje več agregatorjev: CLARIN VLO,[24] OpenAIRE,[25] re3data[26] in OAI.[27]

Večina ostalih ponudnikov jezikovnih virov v Sloveniji, kot npr. slovarski portal Fran[28] na ZRC SAZU, Termania[29] podjetja Amebis, d.o.o, ali stran z viri CJVT[30] ne ponujajo trajnih identifikatorjev. Izjema tu sta digitalna knjižnica dLib[31] NUK, kjer je vsaki publikaciji pripisan trajni identifikator po sistemu URN, ter digitalna knjižnica Sistory[32] INZ, ki, tako kot CLARIN.SI, tudi uporablja sistem handle.

24 https://vlo.clarin.eu/
25 https://www.openaire.eu/
26 https://www.re3data.org/repository/r3d100011922
27 http://www.language-archives.org/archive/clarin.si
28 http://www.fran.si/
29 https://www.termania.net/
30 https://viri.cjvt.si/
31 http://www.dlib.si/
32 https://www.sistory.si/

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,[33] Kres,[34] itd.) so vsi narejeni po principu REST, da torej URL poizvedbe zadošča za ponovno in enako poizvedbo.[35] Z drugimi besedami, URL, ki ga dobimo po poizvedbi in prikazu rezultatov, je mogoče 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,[36] ki jo ponuja češki LINDAT/CLARIN. Za razliko od drugih krajševalnikov ponuja shortref.org opis poizvedbe, kot skrajšani URL pa vrne 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 lahko 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

33 http://www.gigafida.net/

34 http://www.korpus-kres.net/

35 Seveda, če se medtem ni spremenil korpus.

36 http://shortref.org/

agregatorjev. CLARIN.SI izvaja tudi redno 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. Posebej velja poudariti, da je nadzor nad različnimi verzijami, ki ga repozitorij CLARIN.SI omogoča, eno izmed priporočil ustreznega digitalnega skrbništva jezikovnih podatkov (npr. Štebe, Bezjak in Vipavc Brvar, 2015, str. 6). Številni viri so zapisani 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 bi to ogrozilo interoperabilnost praks citiranja podatkov med skupnostmi.*

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

## 5 ZAKLJUČEK

V prispevku smo predstavili rezultate razširjene študije, s katero smo preverjali stanje citiranja jezikovnih podatkov, predvsem korpusov, v najpomembnejših slovenskih znanstvenih revijah in zbornikih, ki so bili objavljeni v dveh različnih obdobjih, ki skupaj zajemata zadnjih sedem let. V raziskavi smo kvantitativno in kvalitativno analizirali obseg in način navajanja virov, izvedli pa smo tudi pregled navodil za avtorje znanstvenih revij in zbornikov, vključenih v raziskavo. Rezultati študije kažejo, 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 nevključenost skupnosti, ki vire razvija, v osrednjo jezikoslovno raziskovalno skupnost pri nas. V prispevkih, ki temeljijo na uporabi jezikovnih virov, pa viri v skoraj petini primerov sploh niso ustrezno citirani. 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 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 ostajajo daleč najpomembnejši indikator znanstvene uspešnosti.

Z ozaveščanjem in izobraževanjem 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.

Videli smo, da trenutna navodila avtorjem v vseh pregledanih revijah bistveno premalo natančno opisujejo dobro prakso citiranja jezikovnih virov in da v veliki meri ravno to vodi do neoptimalnega stanja, ki smo ga opisali za obdobje 2013–2019. Posledično bi bilo potrebno vzpostaviti dialog z 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. Ta proces se je že začel v okviru RDA Node Slovenia.[37] Projekt vodi ADP, ki se mdr. trudi za »razvoj pilota politike znanstvenih založb glede obveznosti predaje raziskovalnih podatkov k objavljenim znanstvenim člankom in aktivna promocija obveze za dostop, ocenjevanje in citiranje raziskovalnih podatkov«. Za

37 https://www.rd-alliance.org/groups/rda-slovenia

te namene je projekt že vzpostavil smernice (Štebe, Dolinar in Bezjak, 2019), kot tudi dialog z nekaterimi raziskovalnimi revijami v Sloveniji z namenom, da služijo kot piloti za vzpostavljanje boljših navodil avtorjev, ki bodo mdr. poudarjale pomembnost navajanja virov s trajnimi identifikatorji, kot je povezava handle, ter navedbo avtorjev, pri čemer je pričujoča revija eden od teh pilotov, predvidene spremembe v navodilih za avtorje pa so bile tudi predstavljene na konferenci »Znanstvene revije Slovenije in raziskovalni podatki«.[38]

Poleg sprememb navodil za avtorje revij in konferenčnih zbornikov bi bilo prav tako nujno 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 prispevkov, ki tovrstne nepopolne vire uporabljajo, pa so pri tem nemočni. Tudi k temu bi lahko z nudenjem ustrezne dokumentacije, izobraževanj in tehnične podpore veliko doprinesla nacionalna raziskovalna infrastruktura CLARIN.SI.

V prihodnje bi bilo raziskavo zanimivo 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.

**Zahvala**

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

## LITERATURA

Arhar Holdt, Š. in Dobrovoljc, K. (2016). Vrednost korpusa *Janes* za slovensko normativistiko. *Slovenščina 2.0, 4*(2), 1–37. doi: 10.4312/slo2.0.2016.2.1-37

Arhar Holdt, Š. in Čibej, J. (2018). Morphological Patterns in the Sloleks Lexicon of Slovene: An Initial Set of Patterns for Nouns. *Slovenščina 2.0, 6*(2), 33–66. doi: 10.4312/slo2.0.2018.2.33-66

38 https://www.adp.fdv.uni-lj.si/dogodki/znanstvene-revije-slovenije-raziskovalni-podatki/

Arias-Badia, B., Bernal, E. in Alonso, A. (2014). An online Spanish Learners' dictionary: the Daele project. *Slovenščina 2.0, 2*(2), 53–71. doi: 10.4312/slo2.0.2014.2.53-71

Atelšek, S. (2019). Navajanje prevzetih jezikoslovnih terminov in celovitost pojmovnih skupin v Cigaletovi *Znanstveni terminologiji* (1880). *Jezikoslovni zapiski, 25*(1), 67–82. doi: 10.3986/jz.v25i1.7566

Bálint Čeh, J. in Kosem, I. (2017). Prvi koraki do novega velikega slovensko-madžarskega slovarja: analiza relevantnih dvojezičnih virov. *Slovenščina 2.0, 5*(2), 113–150. doi: 10.4312/slo2.0.2017.2.113-150

Berez-Kroeker, A. L., Gawne, L., Holton, G., Smythe Kung, S., Pulsifer, P. in Collister, L. B. (2017). The Data Citation and Attribution in Linguistics Group, & the Linguistics Data Interest Group. The Austin Principles of Data Citation in Linguistics (Version 0.1). Dostopno prek http://site.uit.no/linguisticsdatacitation/austinprinciples

Berez-Kroeker, A. L., Gawne, L., Smythe Kung, S., Kelly, B. F., Heston, T., Holton, G., Pulsifer, P., Beaver, D. I., Chelliah, S., Dubinsky, S., Meier, R. P., Thieberger, N., Rice, K. in Woodbury, A. C. (2018). Reproducible research in linguistics: A position statement on data citation and attribution in our field. *Linguistics, 56*(1), 1–18. doi: 10.1515/ling-2017-0032

Data Citation Synthesis Group. (2014). *Joint Declaration of Data Citation Principles*. Martone, M. (ur.). San Diego CA: FORCE11. doi: 10.25490/a97f-egyk

Dobrovoljc, K. (2018a). Raba tipično govorjenih diskurznih označevalcev na spletu. *Slavistična revija, 66*(4), 497–513. Dostopno prek https://srl.si/ojs/srl/article/view/2018-4-1-6

Dobrovoljc, K. (2018b). Formulaicity in Slovene. *Slovenščina 2.0, 6*(2), 67–95. doi: 10.4312/slo2.0.2018.2.67-95

Dobrovoljc, K., Krek, S., Holozan, P., Erjavec, T., Romih, M., Arhar Holdt, Š., Čibej, J., Krsnik, L. in Robnik-Šikonja, M. (2019). *Morphological lexicon Sloleks 2.0*. Slovenian language resource repository CLARIN.SI. Dostopno prek http://hdl.handle.net/11356/1230

Erjavec, T. (2009). Odprtost jezikovnih virov za slovenščino. V M. Stabej (ur.), *Simpozij OBDOBJA 28*. Dostopno prek http://centerslo.si/wp-content/uploads/2015/10/28-Erjavec.pdf

Erjavec, T., Fišer, D., Krek, S. in Ledinek, N. (2010). The JOS Linguistically Tagged Corpus of Slovene. V *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*. Dostopno prek http://www.lrec-conf.org/proceedings/lrec2010/summaries/139.html

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

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

Erjavec, T. (2015b). *Reference corpus of historical Slovene goo300k 1.2*. Slovenian language resource repository CLARIN.SI. Dostopno prek http://hdl.handle.net/11356/1025

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

Fišer, D., Lenardič, J. in Erjavec, T. (2018). Citiranje jezikoslovnih podatkov v slovenskih znanstvenih objavah: stanje in priporočila. V D. Fišer in A. Pančur (ur.), *Zbornik konference Jezikovne tehnologije in digitalna humanistika 2018* (str. 77–84). Univerza v Ljubljani, Filozofska fakulteta.

Furlan, M. (2018). O govejem lastnem imenu Hrdagata in kletvici (h)ardigata. *Jezikoslovni zapiski, 24*(1), 131–141. doi: 10.3986/JZ.24.1.6938

Haspelmath, M. (2014). The Generic Style Rules for Linguistics. *Zenodo*. doi: 10.5281/zenodo.253501

Hudeček, K. in Mihaljević, M. (2019). Hrvatsko mocijsko nazivlje. *Jezikoslovni zapiski, 25*(1), 107–126. doi: 10.3986/jz.v25i1.7569

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

Jelovšek, A. in Erjavec, T. (2019). A corpus-based study of 16th-century Slovene clitics and clitic-like elements. *Slovene Linguistic Studies, 12*, 3–19. Dostopno prek http://hdl.handle.net/1808/29671

Kačič, Z., Horvat, B., Zögling Markuš, A., Veronik, R., Rojc, M., Žgank, A., Sepesy Maučec, M. in Rotovnik, T. (2002). *SNABI database for continuous*

*speech recognition 1.2*. Slovenian language resource repository CLARIN.SI. Dostopno prek http://hdl.handle.net/11356/1051

Kilgarriff, A. in Renau, I. (2013). esTenTen, a vast webcorpus of Peninsular and American Spanish. *Procedia-Social and Behavioral Sciences, 95,* 12–19. doi: 10.1016/j.sbspro.2013.10.617

Krek, S., Erjavec, T., Dobrovoljc, K., Holz, N., Ledinek, N. in Može, S. (2015). Training corpus ssj500k 1.4 Slovenian language resource repository CLARIN.SI. Dostopno prek http://hdl.handle.net/11356/1052

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. (2019). *Training corpus ssj500k 2.2*. Slovenian language resource repository CLARIN.SI. Dostopno prek http://hdl.handle.net/11356/1210

Krvina, D. (2019). Zaporednost dejanj in njen vpliv na rabo glagolskega vida v slovenščini. *Slovene Linguistic Studies, 12*, 75–83. doi: 10.3986/sjsls.12.1.05

Kulčar, M. (2018). Povezanost vida in vezljivosti pri netvorjenih in predponskoobrazilno tvorjenih glagolih. *Jezikoslovni zapiski, 24*(1), 45–62. doi: 10.3986/JZ.24.1.6932

Ljubešić, N., Fišer, D. in Erjavec, T. (2014). TweetCaT: A tool for building Twitter corpora of smaller languages. V N. Calzolari (ur.), *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)* (str. 2279–2283). Reykjavik, Islandija.

Ljubešić, N. in Klubička, F. (2016). *Croatian web corpus hrWaC 2.1*. Slovenian language resource repository CLARIN.SI. Dostopno prek http://hdl.handle.net/11356/1064

Ljubešić, N., Miličević Petrović, M. in Samardžić, T. (2019). Jezična akomodacija na Twitteru: primjer Srbije. *Slavistična revija, 67*(1), 87–106. Dostopno prek https://srl.si/ojs/srl/article/view/2019-1-1-6

Logar Berginc, N., Grčar, M., Brakus, M., Erjavec, T., Arhar Holdt, Š. in Krek, S. (2012). *Korpusi slovenskega jezika Gigafida, KRES, ccGigafida in ccKRES: gradnja, vsebina, uporaba*. Ljubljana: Trojina, zavod za uporabno slovenistiko, Fakulteta za družbene vede. Dostopno prek https://www.fdv.uni-lj.si/docs/default-source/zalozba/pages-from-logar-et-al---korpusi.pdf?sfvrsn=2

Logar Berginc, N., Erjavec, T., Krek, S., Grčar, M. in Holozan, P. (2013). *Written corpus ccKres 1.0*. Slovenian language resource repository CLARIN. SI. Dostopno prek http://hdl.handle.net/11356/1034

Logar Berginc, N., Gantar, P. in Kosem, I. (2014). Collocations and examples of use: a lexical-semantic approach to terminology. *Slovenščina 2.0, 2*(1), 41–61. doi: 10.4312/slo2.0.2014.1.41-61

Marvin, T., Derganc, J., Beguš, S. in Battelino, S. (2018). Word Selection in the Slovenian Sentence Matrix Test for Speech Audiometry. V D. Fišer in A. Pančur (ur.), *Zbornik konference Jezikovne tehnologije in digitalna humanistika 2018* (str. 181–187). Univerza v Ljubljani, Filozofska fakulteta.

Marvin, T., Battelino, S., Beguš, S. in Derganc, J. (2019). Porazdelitev fonemov v slovenščini in izdelava matričnega testa za govorno avdiometrijo. *Slavistična revija, 67*(4), 537–550. Dostopno prek https://srl.si/ojs/srl/article/view/2019-4-1-1

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

Orel, I. (2019). Ženske dvojinske glagolske oblike v starejšem slovenskem knjižnem jeziku. *Slavistična revija, 67*(2), 273–280. Dostopno prek https://srl.si/ojs/srl/article/view/2019-2-1-15

Petrič, T. (2019). Modal Particles in German Declarative Sentences and their Slovenian Counterparts. *Linguistica, 59*(1), 235–251. doi: 10.4312/linguistica.59.1.235-251

Pisanski Peterlin, A. in Mikolič Južnjič, T. (2018). Subject Personal Pronouns in Slovene: Pragmatic Aspects of a Grammatical Category. *Slovenščina 2.0, 6*(2), 127–153. doi: 10.4312/slo2.0.2018.2.127-153

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

Pori, E. in Kosem, I. (2018). In the Search of Lexicographically Relevant Collocation: The Example of Grammatical Relations Containing Adverbs. *Slovenščina 2.0, 6*(2), 154–185. doi: 10.4312/slo2.0.2018.2.154-185

Rath, A. (2019). Anmerkung zur slowenischen Klitikakette (naslonski niz). *Slovene Linguistic Studies, 12*, 95–112. doi: 10.3986/sjsls.12.1.06

Rozman, T., Stritar Kučuk, M., Kosem, I., Krek, S., Krapš Vodopivec, I., Arhar Holdt, Š. in Stabej, M. (2013). *Learners' corpus Šolar 1.0*. Slovenian language resource repository CLARIN.SI. Dostopno prek http://hdl.handle.net/11356/1036

Rozman, T., Arhar Holdt, Š., Pollak, S. in Kosem, I. (2018). Kolokacije v korpusu Šolar. *Jezik in slovstvo, 63*(2–3), 117–128. Dostopno prek https://www.jezikinslovstvo.com/pdf.php?part=2018|2-3|117-128.

Sharoff, S. (2006). Creating general-purpose corpora using automated search engine queries. *WaCky! Working papers on the Web as Corpus*. Dostopno prek http://wackybook.sslmit.unibo.it/pdfs/sharoff.pdf

Stopar, A. in Ilc, G. (2019). Stilistična (ne)zaznamovanost moških in ženskih poimenovalnih parov za poklice v angleščini in slovenščini. *Slavistična revija, 67*(2), 333–342. Dostopno prek https://srl.si/ojs/srl/article/view/2019-2-1-21

Stramljič Breznik, I. (2018). Ženske ne povedo nič pametnega: jezikovnokorpusna analiza stereotipa. *Jezikoslovni zapiski, 24*(1), 27–44. doi: 10.3986/JZ.24.1.6931

Štebe, J., Bezjak, S. in Vipavc Brvar, I. (2015). *Priprava raziskovalnih podatkov za odprt dostop. Priročnik za raziskovalce*. Ljubljana: Založba FDV. Dostopno prek https://www.dlib.si/details/URN:NBN:SI:DOC-06SLBVXX

Štebe, J., Dolinar, M. in Bezjak, S. (2019). *Smernice za oblikovanje politik znanstvenih založb glede navajanja raziskovalnih podatkov v znanstvenih publikacijah in zagotavljanja dostopa do primarnih podatkov, uporabljenih v člankih* (Verzija 2.3.). Dostopno prek https://www.rd-alliance.org/system/files/documents/Smernice_za_razvoj_politike_zalo%C5%BEb_RDA_Slovenija_V2_3.pdf

Štumberger, S. (2015). Slovaropisna obravnava novejše leksike. *Slovene Linguistic Studies*, *10*, 153–166. Dostopno prek https://ojs.zrc-sazu.si/sjsls/article/view/7365

Trivunović, E. (2019). Diahrono raziskovanje biblijskih in izbiblijskih frazemov. *Jezikoslovni zapiski*, 25(2), 47–61. doi: 10.3986/JZ.25.2.3

Uhlik, M. in Žele, A. (2018). Brezosebne zgradbe v slovenščini: kontrastiva z drugimi južnoslovanskimi jeziki in ruščino. *Jezikoslovni zapiski*, *24*(2), 99–112. doi: 10.3986/jz.v24i2.7112

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

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

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

Vončina, M. (2016). Zaključena znanstvena zbirka podatkov – primeri katalogizacije in Sicris vrednotenja. [Delavnica ADP, 26. 10. 2016.] Dostopno prek https://www.adp.fdv.uni-lj.si/adp_delavnica_okt2016/presentations/2016_Mira-Voncina_Znanstvena_zbirka_podatkov.pdf

Zwitter Vitez, A., Zemljarič Miklavčič, J., Krek, S., Stabej, M. in Erjavec, T. (2013). *Spoken corpus Gos 1.0*. Slovenian language resource repository CLARIN.SI. Dostopno prek http://hdl.handle.net/11356/1040

Zwitter Vitez, A. (2018). Enota analize spontanega govora: interakcija prozodije, pragmatike in skladnje. *Jezik in slovstvo, 63*(2–3), 157–175. Dostopno prek https://www.jezikinslovstvo.com/pdf.php?part=2018|2-3|157-175

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

Žele, A. (2018). O aktualnostnočlenitveni stavi v slovenščini. *Jezik in slovstvo, 63*(2–3), 59–73.

Žitnik, S., Šubelj, L. in Bajec, M. (2014). SkipCor: Skip-mention coreference resolution using linear-chain conditional random fields. *PloS one, 9*(6), e100101. doi: 10.1371/journal.pone.0100101

Žitnik, S., Draskovic, D., Nikolić, B. in Bajec, M. (2017). nutIE—A modern open source natural language processing toolkit. *Proceedings of the 25th Telecommunication Forum (TELFOR)*, 1–4. doi: 10.1109/TELFOR.2017.8249486

Žitnik, S. in Bajec, M. (2018). Coreference Resolution for Slovene on Annotated Data from coref149. *Slovenščina 2.0, 6*(1), 37–67. doi: 10.4312/slo2.0.2018.1.37-67

# 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, mainly (mainly language corpora) citation in six leading Slovene scientific journals (*Jezik in slovstvo*, *Slavistična revija*, *Slovenščina 2.0*, *Linguistica*, *Slovene Linguistic Studies* and *Jezikoslovni zapiski*) and in the proceedings of two scientific conferences focused on linguistics (*Jezikovne tehnologije in digitalna humanistika* and *Obdobja*) for the period of the last seven years, i.e. from 2013 to 2019. We consider 1,074 papers and analyse the results both quantitatively and qualitatively. From the quantitative perspective, we show that, overall, only about a fourth of the papers includes the use of language resources, and that in the later period (2018–2019) the use of language resources is over twice as frequent as it is in the earlier period (2013–2017). We classify the manner of language resource citation into five categories (e.g. *citing the hyperlink in the texts* or *citing the key paper about the resource*) and show that how a resource is cited is, to a large extent, dependent on the instructions for authors of the particular publication. Our qualitative analysis focuses mainly on resources deposited in the repository of the CLARIN.SI research infrastructure, where we show that they are, with few exceptions, incorrectly cited. We summarise the finding using the so-called Austin principles, show what has already been achieved in the scope of the CLARIN.SI infrastructure and propose guidelines for citing linguistic research data and how to implement them.

**Keywords:** Open Science, Research Data Citation, Language Resources, Austin Principles, Slovenian Journals and Conference Proceedings