VREDNOST KORPUSA JANES ZA SLOVENSKO NORMATIVISTIKO

Špela ARHAR HOLDT
Zavod za uporabno slovenistiko Trojina
Filozofska fakulteta Univerze v Ljubljani

Kaja DOBROVOLJC
Zavod za uporabno slovenistiko Trojina

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

Namen pričujočega prispevka je preveriti vrednost korpusa Janes za normativistične raziskave. Korpus Janes namreč prinaša besedila, ki za razliko od gradiva v referenčnih korpusih večinoma niso jezikovno korigirana in zato realneje izkazuje tendence rabe oz. (ne)intuitivnost obstoječih jezikovnih pravil v širši jezikovni skupnosti. Za študijo primera smo izbrali zveze samostalnika z neujemalnim levim prilastkom (*solo petje*, *RTV prispevek*). Analiza razkriva: da se referenčni korpus Kres in korpus Janes glede zapisa teh zvez pomembno razlikujeta; da je raba tovrstnih zvez v korpusu Janes pogostejša in bolj raznolika kot v korpusu Kres; da se v obeh korpusih pojavlja visok delež zvez, ki v rabi izkazujejo variantnost v zapisovanju, tudi na ravni posameznih prilastkov; in – vsaj na prvi pogled – presenetljivo, da je raba v korpusu Janes konsistentnejša, kar nakazuje, da jezikovna regulacija obravnavanega problema povečuje variantnost v jezikovni rabi. Prispevek temelji na konferenčni temi, ki smo jo podatkovno in vsebinsko razširili, vključili smo tudi razpravo o možni nadaljnji obravnavi izbranega jezikovnega problema, širše pa o pomenu in načinu vključitve korpusa Janes v metodologijo slovenske normativistike.

Ključne besede: korpus Janes, korpus Kres, normativistika, intuitivnost jezikovnih pravil, neujemalni levi prilastek

1 LEKTORIRANOST BESEDIL V SLOVENSKIH KORPUSIH

Slovenski referenčni besedilni korpusi[1] vsebujejo po večini besedila iz časopisov in revij ter strokovnih in leposlovnih književnih del. Ker so v našem prostoru naštete besedilne vrste v veliki (čeprav ne enostavno določljivi) meri pred objavo jezikovno lektorirane, lahko privzamemo, da referenčni korpusi kot odslikava jezikovne realnosti prinašajo predvsem jezikovno pregledano in popravljeno gradivo. Za vrsto primerov korpusne uporabe je to dejstvo ustrezno, včasih celo izrazito zaželeno.[2] Kadar pa raziskovalce zanimajo izvorne, nekorigirane tendence jezikovne rabe, se naslonitev na podatke iz referenčnega korpusa izkaže za metodološko nezadostno.[3]

V teh primerih se je treba opreti na jezikovno gradivo, ki je bilo objavljeno brez lektorskih oz. uredniških posegov. Za slovenščino je tudi za te namene od nedavnega na voljo korpus Janes, ki prinaša uporabniško generirane spletne vsebine različnih tipov: tvite, bloge, uporabniške komentarje in zapise z uporabniških forumov (Fišer in dr. 2015). Seveda ni mogoče izhajati iz predpostavke, da prav nobena od objav v korpusu Janes ni bila jezikovno pregledana, kakor tudi ni mogoče predpostavljati, da so lektorirana prav vsa besedila v knjigah, revijah in časopisih, zajetih v referenčni korpus.[4] Na drugi strani zaenkrat ostaja neodgovorjeno vprašanje, na kakšen način v slovenski

[1] Danes sta to Gigafida in Kres (Logar in dr. 2012), v preteklosti so bili v uporabi FIDA, FidaPLUS in Nova beseda.

[2] Tipičen primer uporabnikov, ki pogosto izražajo željo po lektoriranosti jezikovnega gradiva, so učitelji, ki želijo korpusne podatke uporabljati za jezikovnodidaktične namene.

[3] Pri čemer je potrebno poudariti, da so želje, da bi referenčni korpus vseboval in s pomočjo metaoznak tudi razmejeval tako lektorirano kot nelektorirano gradivo – in s tem torej omogočil najširši možni domet raziskav – prisotne že od korpusa FIDA. Vendar pa ločevanje gradiva po tem ključu ni trivialen postopek, kar je bil tudi eden od razlogov za opustitev korpusne metaoznake *lektorirano* oz. *nelektorirano* ob nadgradnji korpusa FidaPLUS v Gigafida (Logar in dr. 2012: 20).

[4] Gigafida in Kres vsebujeta tudi raznovrstna spletna besedila, pri katerih je še težje podajati zanesljive ocene glede morebitne ne/lektoriranosti. V zvezi s tem je mogoče omeniti trenutno potekajoči projekt nadgradnje referenčnega korpusa, ki napoveduje poskus ločevanja besedil, za katere je glede na okoliščine in medij objave mogoče sklepati, da je bil avtorjev namen pisati v standardnem jeziku, od besedil, pri katerih takšno sklepanje ni mogoče (Krek in dr. 2016).

(opomba se nadaljuje na naslednji strani)

računalniško posredovani komunikaciji odsotnost jezikovnega pregleda druge osebe vpliva na količino in naravo avtorjeve samokorekcije: je slednje manj, ker avtorji v spletnih žanrih ne čutijo enake stopnje zavezanosti jezikovni normi kot pri pisanju drugih vrst besedil, ali je samokontrole v strahu pred možnimi jezikovnimi napakami celo več, tudi v smeri jezikovne hiperkorekcije?[5]

Kljub naštetim pomislekom pa je mogoče z dovoljšno gotovostjo predpostaviti, da se glede prisotnosti lektorske ali uredniške jezikovne intervencije gradivo korpusa Janes od referenčnega pomembno razlikuje. Od jezikoslovnih področij, za katera je ta razlika relevantna, nas v pričujočem prispevku zanima predvsem normativistika, in sicer v segmentu, ko se posveča funkcioniranju jezikovnih pravil v jezikovni rabi. Trenutno velja, da se primerjalne korpusne raziskave osredotočajo predvsem na ugotavljanje, v kolikšni meri se v (korigirani in/ali nekorigirani) jezikovni rabi pojavljajo odstopi od obstoječe jezikovne norme (npr. v Jakop 2008; Michelizza 2015: 100–160; Popič, Fišer 2015; Škrjanec in dr. 2015). Točka interesa pri tem pristopu je jezikovni uporabnik ter njegova na spletu drugačna in zato za jezikoslovni pogled zanimiva jezikovna raba. Ideja pričujočega prispevka je ta zorni kot obrniti in primerjalno analizo uporabiti za vrednotenje »uspešnosti« jezikovnih pravil v jezikovni praksi oz. za oceno intuitivnosti določenega jezikovnega pravila za širšo jezikovno skupnost. S tem želimo opozoriti na to, da se je v trenutni situaciji ob vprašanju, kaj (če sploh kaj) lahko slovenska normativistika naredi za računalniško posredovano komunikacijo,[6] nujno vprašati tudi, kaj lahko

[5] Odgovor bi bil najbrž primerljiv zaključku, ki ga ponudi Crystal, ko našteva različne možne razloge za odstope od norme v mrežni govorici: lahko, da uporabnik pravila ne pozna; lahko ga pozna, pa mu je vseeno; lahko želi upoštevati pravila, pa ni dovolj tipkarsko spreten; morda je v naglici površen in ne prebere za sabo; morda je odstop od norme namenski, ker se uporabnik želi prilagoditi načinu komunikacije vrstnikov ali pa želi z odstopom doseči poseben jezikovni učinek. Lahko, da gre za zmes več od naštetih dejavnikov ali kaj povsem drugega. Misliti pa si je mogoče, da so med dejavniki vpliva predvsem starost, spol, izobrazba, kot tudi jezikovni okus oz. preference uporabnika ter njegove značajske lastnosti (Crystal 2011: 59–60).

[6] Nekaj premislekov v zvezi s tem vprašanjem ponuja zapis okrogle mize *Slovenščina Janes: pogovorna, nestandardna, spletna ali spretna?* (Stabej in dr. 2016).

(opomba se nadaljuje na naslednji strani)

računalniško posredovana komunikacija naredi za slovensko normativistiko.

Kot primer jezikovnega problema, ki ga opisani pristop lahko pomaga dodatno osvetliti, smo izbrali v slovenistiki dobro poznano vprašanje zapisovanja zvez samostalnika z neujemalnim levim prilastkom (*alfa samec*, *servo volan*, *USB ključek* oz. *alfasamec*, *servovolan*, *USB-ključek*).[7] Izbira je utemeljena z naslednjimi tremi razlogi: ker je v literaturi že bilo identificirano, da pri zapisovanju tovrstnih zvez narazen/skupaj v rabi (kot tudi v pravilih) vlada precejšnja variantnost; ker gre za problem, pri katerem je pričakovana in v praksi prisotna jezikovna regulacija; in ker je mogoče predvideti, da pri tem vprašanju ni veliko odstopov od norme za dosego posebnega stilskega učinka, ki bi oteževali kvantitativno analizo podatkov.

Prispevek je nastal kot razširitev konferenčne teme (Arhar Holdt, Dobrovoljc 2015), ob čemer smo razpravo vsebinsko poglobili in razširili gradivo z novimi korpusnimi primeri. V nadaljevanju najprej predstavimo izbrani jezikoslovni problem, nato opredelimo metodologijo pridobivanja korpusnih podatkov in rezultate kvantitativne ter kvalitativne korpusne analize. Na podlagi izsledkov preverjamo vrednost korpusa Janes za normativistične raziskave in trenutne možnosti za izvedbo širših, sintetičnih korpusnih raziskav izbrane tematike.

2 PREDSTAVITEV OBRAVNAVANEGA JEZIKOVNEGA PROBLEMA

Razlike v jezikoslovnem razumevanju zvez tipa *alfa samec*, *servo volan*, *RTV prispevek* (oz. kot medponskoobrazilne zloženke zapisano skupaj: *alfasamec*, *servovolan* in kot podredne zloženke s kratično/črkovno/številčno prvo sestavino z vezajem: *RTV-prispevek*) so v slovenskem prostoru prisotne že več kot pol stoletja. V tem času so bile temi posvečene številne razprave, običajno v

[7] V literaturi se pojavljajo različna poimenovanja tovrstnih zvez, kar je povezano z živahno jezikoslovno razpravo, ki spremlja tematiko. Za pričujočo objavo smo se skladno z recenzentskimi priporočili odločili za poimenovanje *zveze samostalnika z neujemalnim levim prilastkom*, s čimer se vsaj v uvodnem delu izognemo vprašanju, kako besednovrstno uvrstiti prvi del zveze in kako poimenovati dejstvo, da se slednji v rabi praviloma ne pregiba. Metodološkim težavam pa se s tem ne izognemo, k čemur se vračamo v razdelku 3.

povezavi s pripravo jezikovnih priročnikov, v katerih je bilo treba podati kategorizacijske rešitve na večji količini avtentičnega jezikovnega gradiva. Polemika je dvotirna: na eni strani se dotika vprašanja, katere od tovrstnih zvez zapisovati skupaj in katere narazen, na drugi strani pa, kako v primeru zapisa narazen besednovrstno uvrščati prvi del besedne zveze. Za te besede je namreč značilno, da so po skladenjski vlogi podobne pridevnikom, po obliki pa samostalnikom – ker pa se pri sklanjanju zveze ne pregibajo, je iz oblikoskladenjskih lastnosti težko določiti optimalno kategorizacijsko rešitev.

V zvezi z besednovrstnim opredeljevanjem so bile predlagane različne možnosti, začetkom debate pa je mogoče slediti v čas priprav na izid Slovarja slovenskega knjižnega jezika. Kot povzema Černelič-Kozlevčar (1988), je poskusni snopič SSKJ za prvi del zvez tipa *avto garaža* uvedel poimenovanje *nesklonljivi pridevnik*, vendar Pogorelec (1965) v svojem odzivu uvrsti te besede med samostalnike, Korošec (1967) pa zagovarja ohlapnejše poimenovanje *nesklonljivi prilastek*, ki najde pot tudi v SSKJ. Na to odločitev se je kritično odzval Toporišič (mdr. 1988), ki izbrano poimenovanje kot poskus izogibanja jasnemu besednovrstnemu uvrščanju označi kot »posebno slabost SSKJ« in strne, »da se I. Černelič-Kozlevčar v teoriji besednih vrst da voditi tudi (ali predvsem) skupinskim interesom slovarnikov in ne more postati (dovolj jasno in dosledno) samo služabnica resnice, pravega spoznanja« (Toporišič 1988: 445). Ta ocena razkriva pravo razsežnost nesoglasja med skupinami jezikoslovcev, v katerem ena stran razlaga v slovarju heterogeno implementirane rešitve (tudi) s stanjem v jezikovni rabi (Rigler 1971), druga pa ponuja sistemski pristop, ki levo od samostalniškega jedra predvideva le pridevnik, problem neujemalnih prilastkov pa rešuje (tudi) z argumentiranjem doslednejšega zapisovanja tovrstnih zvez skupaj in s priporočilom glede pretvorbe problematičnih zvez v drugo skladenjsko obliko (Toporišič 1971). Ker je doprinos na obeh straneh zelo bogat, obenem pa ni namen pričujočega prispevka vstopati v razpravo v jezikovnoteoretičnem smislu, na tem mestu napotujemo bralca k natančnejšim pregledom diskusije, npr. v Logar (2005;

2012). Razen tega je treba omeniti še nekaj novejših razprav, npr. Gložančev (2012), ki kritično ocenjuje rešitve obravnavanega vprašanja v slovarju SP 2001, ter (Kern 2012; Gantar 2015: 117–123), ki predstavljata rešitve v novejših slovarskih virih: Slovarju novejšega besedja slovenskega jezika in Leksikalni bazi za slovenščino.

Če je vprašanje besednozvezne kategorizacije morda naloga predvsem za jezikovni opis, je vprašanje zapisovanja zvez narazen ali skupaj večji izziv za jezikovni predpis. Različne možnosti zapisovanja se pojavijo že v Slovenskem pravopisu iz leta 1950,[8] SP 1962 pa prvi predstavi pravilo ((§ 74), da se besede, ki so v obeh delih tujega izvora in »smo jih sprejeli kot zvarjeno celoto«, zapisujejo skupaj (*avtogaraža*, *kinooperater*), zveze iz domače besede in tuje pa narazen (*alfa žarki*, *avto promet*). V obeh pravopisih se tudi že predlaga kot ustreznejša skladenjska rešitev oblika, v kateri je določujoča beseda postavljena za jedrno (*žarki alfa*) ali pa se prvi del preoblikuje v pridevnik (*avtomobilski promet*). Zanimivo je tudi, da SP 1962 pri zvezah s kratično sestavino rabo vezaja predpisuje samo takrat, ko se simbol piše z malo črko (*h-mol*), sicer ne (*H vitamin*). Kot rečeno, je v sedemdesetih in osemdesetih sledila živahna razprava, v kateri avtorji utemeljujejo razloge za različen zapis tovrstnih zvez narazen ali skupaj (povzeto v Logar 2005).

Trenutni rezultat je tak, da pravopisna pravila pri zloženkah s črkovno ali številčno prvo sestavino dosledno predpisujejo rabo vezaja, npr. *TV-program* (§ 496). Zapis skupaj velja tudi za zloženke s količinskim številom v prvem delu (*petletka*, § 495) in za medponske zloženke tipa *zobozdravnik* (§ 495). Pri ostalih zvezah se tipično dovoljuje zapis skupaj ali narazen (*alfasamec* ali *alfa samec*), skupaj s priporočilom glede preoblikovanja zveze (*samec alfa*) (§ 497–500), vendar so pri različnih skupinah zvez rešitve v jezikovnih priročnikih ter jezikovni rabi zelo raznolike. Dobrovoljc in Jakop (2011: 113–114) ugotavljata,

[8] Npr. pri geslu *kinematograf* (ibid. 276): »**kinematograf** -a *m in* kino -a *m*, - u -o -u -om, kinematografski - a -o: ~ i aparat, ~ a dvorana, ~ o gledališče, ~ a predstava, *v vsakdanji rabi tudi okrajšave o tujih zgledih*: kino dvorana, kino predstava, *toda*: kinooperater, kinostudio, kinoprojektor, kinoaparat«

da se dvojnice glede zapisa v normi ne prekrivajo z dvojnicami v jezikovni rabi, da so obstoječa pravila mestoma nejasna, jezikovna raba pa izrazito neustaljena. Neustaljenost v rabi in neskladje s predpisom so izkazale tudi raziskave N. Logar, ki jih povzema Logar (2012). Dosedanje korpusne analize sicer prinašajo nekatere izsledke, ki so okrepili predvsem argumente za zapis narazen (kar je v luči dosedanje jezikoslovne polemike skladno s pričakovanji). Ob tem pa je za našo razpravo še zlasti pomembno, da avtorji problematizirajo relevantnost uporabljenih korpusnih virov za pristop k tematiki, saj na osnovi lektoriranih besedil ni mogoče realno oceniti obsežnosti in narave obravnavanega problema v jezikovni praksi (Logar 2012: 119; Dobrovoljc, Jakop: 115). Popravki na ravni zapisa narazen/skupaj, kot tudi preoblikovanja tovrstnih zvez v drugo skladenjsko obliko, so namreč med pričakovanimi lektorskimi posegi.[9] Posledično referenčni korpus težko razkrije, kakšna vrsta zapisovanja se zdi jezikovni skupnosti de facto intuitivna – je pa ta podatek izredno dobrodošlo, če ne nujno (obenem pa seveda ne edino) vodilo, ki ga je pri ocenjevanju in nadgradnji jezikovnih pravil potrebno upoštevati.[10]

V tem prispevku s kvantitativno in kvalitativno korpusno analizo primerjamo, kako se glede zapisa obravnavane vrste besed oz. zvez razlikujeta dva korpusna vira: uravnoteženi referenčni korpus Kres (Logar in dr. 2012) in korpus uporabniških vsebin Janes v različici v3 (Fišer in dr. 2015). V prispevku preverjamo naslednje hipoteze: da se korpusa Kres in Janes glede rabe zvez samostalnika z neujemalnim levim prilastkom pomembno razlikujeta; da je raba tovrstnih zvez v korpusu Janes pogostejša kot v korpusu Kres; da se v obeh korpusih pojavlja visok delež zvez, ki v rabi izkazujejo variantnost v zapisovanju; in da v primeru dvojnic v korpusu Kres prevladuje zapis skupaj oz. z vezajem, v korpusu Janes pa zapis narazen.

[9] Čeprav v množici drugih posegov po pogostnosti niso med prvimi (Popič 2014: 229–230).
[10] Dosedanje pristope k standardizaciji slovenščine povzema Dobrovoljc (2013), ki izpostavlja potrebo, da se v metodo vključijo tudi empirični podatki o avtentični sodobni jezikovni rabi. Dobrovoljc (2008) v tem smislu ponuja tudi prvo preverbo določil iz SP 2001 v jezikovni rabi, in sicer prav na primeru zapisovanja skupaj in narazen.

3 METODA LUŠČENJA PODATKOV

V raziskavi smo se osredotočili na zapis zvez samostalnika z neujemalnim levim prilastkom, pri čemer smo se omejili le na tiste prilastke, ki so v korpusih označeni kot samostalniški. S tem kriterijem smo želeli zajeti primere, ki jih Gantar (2015: 321) opredeljuje kot dvodelne samostalniške strukture sbz1 sbz0, npr. *golf igrišče*, *celofan papir*.[11] Že v prvem koraku luščenja pa je postalo jasno, da zaradi že omenjenih težav na ravni besednovrstnega uvrščanja prilastkov v jezikovnih virih podatki v tem smislu niso homogeni. Kljub temu smo delo nadaljevali po izhodiščnem načrtu, saj je pomembno metodološko vprašanje raziskave, do kakšnih rezultatov je mogoče priti z rabo obstoječega korpusnega gradiva, na katerih točkah pa je potrebna nadgradnja, da bo gradivo za tovrstne raziskave sploh uporabno.

Kot potencialne zveze samostalnikov z neujemalnim levim samostalniškim prilastkom smo z orodjem noSketchEngine iz obeh korpusov izluščili nize dveh zaporednih samostalnikov,[12] samostalnika v imenovalniku in samostalnika s poljubnimi oblikoskladenjskimi lastnostmi, pri katerih se dana oblika prvega samostalnika ne glede na velikost črk v celotnem korpusu pojavi pred vsaj tremi različnimi oblikami leme jedrnega samostalnika (npr. *rtv prispevek*, *rtv prispevka*, *rtv prispevkom*). Če je bil ta pogoj izpolnjen, je bil niz oblike prilastka in leme jedra prepoznan kot potencialna zveza samostalnika z neujemalnim levim samostalniškim prilastkom (*rtv prispevek*).

Pregled rezultatov je razkril, da tokenizacijske, besednovrstne in oblikoskladenjske interpretacije, kakršne so bile korpusnim besedilom pripisane v postopku strojnega oblikoskladenjskega označevanja, na podatke vplivajo tako zaradi neenotnosti označevalnikov kot zaradi nedoslednosti v obstoječih jezikovnih virih.

[11] Vidovič Muha (2011: 289–298) pa jih besedotvorno utemeljuje kot zloženke, ki imajo v skladenjski podstavi samostalniški prilastek v različnih skladenjskopomenskih razmerjih z jedrom (npr. *koktajlobleka, basklarinet, matpozicija*).

[12] V korpusu Janes so bili kot nerelevantni že v postopku luščenja izločeni samostalniki, ki se začnejo z znakoma @ ali #.

3.1 Neenotnost označevalnikov

Korpusa sta označena z različnima (statističnima) označevalnikoma: korpus Kres z označevalnikom Obeliks (Grčar in dr. 2012), korpus Janes pa z označevalnikom ToTaLe (Erjavec in dr. 2005). Čeprav oba označevalnika svoj model znanja gradita na istih jezikovnih virih, tj. označevalni shemi JOS (Erjavec, Krek 2008), leksikonu besednih oblik Sloleks (Dobrovoljc in dr. 2015) in učnem korpusu ssj500k (Krek in dr. 2013), med njima zaradi drugačne zasnove lahko prihaja do razlik pri obravnavi nekaterih specifičnih pojavov, povezanih z našo raziskavo.

Na ravni tokenizacije moramo tako upoštevati, da označevalnika uporabljata drugačna pravila za segmentacijo besedil na posamezne besede oz. korpusne pojavnice, zlasti pri obravnavi zvez z ločilom, kot so zloženke z vezajem. Medtem ko označevalnik Obeliks te ne glede na stičnost ločila vedno tokenizira kot niz treh samostojnih pojavnic (*C*, -, *vitamin; prostor*, -, *čas*), označevalnik ToTaLe zveze z obojestičnim ločilom obravnava kot eno samo pojavnico (*C-vitamin; prostor-čas*), zveze z levo-, desno- ali nestičnim vmesnim ločilom pa kot niz treh pojavnic, kar smo pri luščenju podatkov o pogostosti rabe zapisa z vezajem tudi upoštevali.

Prav tako med označevalnikoma prihaja do razlik na ravni besednovrstne kategorizacije krajšav, saj označevalnik Obeliks med okrajšave prišteva zgolj vnaprej določene kratice s končno levostično piko (npr. *dr.*), označevalnik ToTaLe pa tudi izbrane akronime in simbole (npr. *EUR*, *MB*, *CO2*; *km*, *min*, *kg*), ki jih Obeliks umešča med samostalnike. Ker se v korpusu Janes oznake pripisujejo normaliziranim (standardiziranim) pojavnicam in ne neposredno pojavnicam, kakršne se pojavljajo v neoznačenih besedilih, so enakim izhodiščnim pojavnicam zaradi omejene natančnosti postopka avtomatske normalizacije (Ljubešič in dr. 2014) včasih pripisane različne interpretacije (kratice *FB*, *BMW*, *html* so v korpusu Janes denimo označene bodisi kot samostalniki bodisi kot napake označevalnega programa).

Situacija, ki jo opisujemo, sicer ni nespremenljiva: že v času priprave prispevka je bila izdana nova različica korpusa Janes z izboljšavami na ravni oznak, dodaten razvoj pa napoveduje tudi priprava ročno označenega učnega korpusa za označevanje nestandardne slovenščine (Čibej in dr., v tisku). Čeprav je pričakovati, da bodo razlike v delovanju označevalnikov vedno obstajale – v obravnavanem primeru je to delno pogojeno z lastnostmi nestandardnega jezika v primerjavi s standardnim – bi morali v prihodnosti stremeti k čim višji usklajenosti teh sistemov, še zlasti spričo dejstva, da primerjalne korpusne analize v našem prostoru s porastom razpoložljivih virov za slovenščino postajajo vedno pogostejše in kompleksnejše. Na drugi strani pa v prispevku izbrana metoda razkriva pomembnost uporabnikove natančne seznanjenosti s specifikami uporabljenih jezikovnih virov, ki močno presega interpretacijo rezultatov zgolj na osnovi predočenih korpusnih podatkov (Logar in dr. 2015: 468).

3.2 Nedoslednost jezikovnih virov

Druga, vsebinska, omejitev strojnega označevanja je posledica nedosledne obravnave neujemalnih prilastkov v obeh omenjenih jezikovnih virih. V leksikonu besednih oblik Sloleks se neujemalni levi samostalniški prilastki pojavljajo v treh različnih tipih leksikonskih enot. Med prve spadajo samostalniki, ki jim je pripisana paradigma za vsa števila in sklone, a z eno nespremenljivo obliko z ničto končnico, npr. prvi deli večbesednih lastnih imen (*Cape*, *Buenos*, *Las*, *Slovenj* itd.), akronimi (*FBI*, *DNK*, *MoOZ*), simboli (*cal*, *din*, *mio*) oz. druge krajšave (*abc*), nazivi (*doktor*, *profesor*) in samostalnik *foto*. V drugi tip spadajo samostalniki, ki jim je pripisana paradigma z eno samo pregibno obliko v imenovalniku ednine, med katerimi so prav tako najpogostejši različni tipi krajšav (*cm*, *dag*, *Hz*; *MB*, *USD*, *EUR*), kar kaže na nedoslednost pri vnosu enot s podobnimi pomenskimi in skladenjskimi lastnostmi. V tretji tip spadajo samostalniki, ki imajo prepisano zgolj paradigmo s pregibnimi oblikami (npr. *avto*, *solo*, *golf*), torej njihova potencialna ali pogosta raba v pridevniški rabi ni eksplicitno signalizirana z

nesklonljivo ali enooblikovno paradigmo. Dva samostalnika tega tipa (*pat*, *spam*) pa se v leksikonu pojavljata tudi kot nesklonljiva pridevnika.

Podobne nedoslednosti razkriva tudi oblikoskladenjska klasifikacija besedišča v skladenjskem kontekstu (učni korpus ssj500k). Na ravni določanja besednih vrst so nekaterim prilastkom pripisane različne kategorije (npr. pridevnik in prislov *neto*, pridevnik in samostalnik *bruto*, *jumbo*, *super*) ali pa pripisana besedna vrsta ni v skladu z leksikonom (npr. samostalniki *orto*, *ultra*, *instant*). Na ravni določanja oblikoskladenjskih lastnosti pa je samostalnikom namesto priporočene oz. prevladujoče imenovalniške interpretacije ponekod pripisan sklon samostalniškega jedra (npr. *video* tož. *izhod* tož., *v San* mest. *Francisc*u mest., *foto* tož. ed. *krožke* tož. mn.).

Če sklenemo, temeljna jezikovna vira za označevanje slovenskih besedil trenutno izkazujeta določeno mero nedoslednosti tako znotraj posameznega vira kot med seboj. V primeru morebitnih nadgradenj bi bila tako potrebna predvsem natančnejša opredelitev meril za besednovrstno ločevanje med samostalniki in pridevniki, na primer po načelu prevladujoče skladenjske vloge v rabi, kot predlaga Gantar (2015: 117–119). Po besednovrstni kategorizaciji bi nato leksikon Sloleks veljalo posodobiti tako, da se iz samostalnikov izločijo enote, ki se pojavljajo predvsem v pridevniški vlogi (npr. prvi deli lastnih imen, samostalnik *pat*), v leksikonu pa bi bila nepregibna paradigma nato pripisana zgolj tistim samostalnikom, ki se tudi v samostalniški vlogi ne pregibajo (akronimi, simboli in drugi tipi krajšav), s čimer ne bi bilo več potrebe po ohranjanju enooblikovnih paradigem.

Ne glede na to, ali v vlogi levega prilastka nastopa (v splošni jezikovni rabi) pregibni ali nepregibni samostalnik, pa bi moral biti ta v učnem korpusu dosledno označen kot samostalnik v imenovalniku ednine (tj. neujemanje, npr. *avto*im. ed. *sejmov*). Pridevnike, ki se sklanjajo z ničto končnico, pa je v leksikonu smiselno še naprej navajati s celotno paradigmo in jim v učnem korpusu pripisovati lastnosti samostalniškega jedra (tj. ujemanje, npr. *bež* tož. ed. ž. *barvo*).

Izpostavljeni omejitvi z vidika kvantitativnih primerjav v pričujočem prispevku

sicer nista problematični, saj predpostavljamo, da glede na prekrivnost izhodiščnih jezikovnih virov označevalnika obravnavane skladenjske strukture označujeta s podobno natančnostjo, zaradi česar sta delež in nabor nerelevantnih oz. manjkajočih zadetkov v obeh korpusih primerljiva. Kot podrobneje izpostavimo pri opisu kvalitativne kategorizacije izluščenih zvez (razdelek 5), pa bi veljalo ob nadaljnjih analizah posameznih podskupin neujemalnih levih prilastkov označevanje poenotiti in iskanje razširiti tudi na samostalnike v neimenovalniških sklonih oz. na nesamostalniške pojavnice.

4 KVANTITATIVNA PRIMERJAVA REZULTATOV

4.1 Pogostost rabe zvez z neujemalnim levim prilastkom

Kot prikazujejo podatki v tabeli 1, smo iz korpusa Kres z opisano metodo izluščili 3.054, iz korpusa Janes pa 7.840 različnih potencialnih zvez z neujemalnim levim prilastkom (različnic). Primerjava relativne pogostosti pregibnih zvez (števila pojavnic)[13] v obeh korpusih razkriva, da se v korpusu Janes pojavlja 1,7-krat več tovrstnih zvez kot v korpusu Kres (505 proti 846 pojavitvam na milijon pojavnic), kar kaže na pogostejšo rabo tega skladenjskega mehanizma v nelektoriranih uporabniških spletnih vsebinah.

Če pogostost rabe primerjamo še z vidika razmerja med številom pojavitev in številom različnih zvez, vidimo, da je povprečna pogostost posamezne zveze v korpusu Kres sicer višja kot v korpusu Janes (16,5 pojavitev zveze primerjavi z 10,8 pojavitev na 100 milijonov pojavnic), a je nabor zvez v korpusu Janes bistveno daljši (31 proti 61 različnih zvez na milijon pojavnic). Če je torej v korpusu Kres raba levega neujemalnega samostalniškega prilastka omejena na manjše število pogostih zvez, je raba v korpusu Janes razpršena na raznolikejši nabor zvez.

[13] Pri analizi pogostosti zvez namesto pogostosti vseh oblik zvez primerjamo zgolj pogostost oblik v neimenovalniških oblikah, s čimer relativiziramo pojavitve zvez, ki so zelo pogoste, a je njihovo pregibanje z neujemalnim prilastkom redko (npr. *~~janez janša~~, janez janše, janez janši*). V tabeli 1 so sicer navedeni podatki za oba načina štetja.

	Kres		**Janes**		**Skupno**
	Abs.	Rel.	Abs.	Rel.	Abs.
pojavnice (vse)	95.897	987	212.808	1.662	
pojavnice (pregibne)	49.047	505	108.303	846	
različnice – zveze	**3.054**	31	**7.840**	61	**888**
različnica – prilastki	1.432	15	2.851	22	719

Tabela 1: Primerjava pogostosti zvez (npr. *video posnetek*) oz. prilastkov (npr. *video*) v korpusih Kres in Janes.

Med 50 najpogostejšimi lastnoimenskimi zvezami[14] v korpusu Kres so: *new york*, *los angeles*, *špas teater*, *las vegas*, *al kaida*, *buenos aires*, *slovenj gradec*, *new jersey*, *san francisco*, *skb banka*, *new orleans*, *sr slovenija*, *internet explorer*, *york times*, *don kihot*, *union olimpija*, *bmw serija*, *big brother*, *manchester united*, *premier liga*, *ac milan*, *mercator center*, *visual basic*, *lake city*, *lions klub*, *real madrid*, *spring lake*, *rotary klub*, *beverly hills*, *btc city*, *san antonio*, *wall street*, *sankt peterburg*, *new delhi*, *jv evropa*, *don juan*, *cmc publikum*, *washington post*, *roland garros*, *viba film*, *hypo banka*, *sng opera*, *pro plus*, *acroni jesenice*, *san pieter*, *rtv slovenija*, *cafe teater*, *red bull*, *sng drama* in *vw pasat*.

Med 50 najpogostejšimi občnoimenskimi zvezami v korpusu Kres so: *loto številka*, *video posnetek*, *b člen*, *grand hotel*, *bruto plača*, *žiro račun*, *c člen*, *sms sporočilo*, *ects točka*, *rock skupina*, *web stran*, *dos aplikacija*, *uv žarek*, *tv program*, *pop glasba*, *video kamera*, *krep papir*, *feng šu*, *rock glasba*, *moto zveza*, *fitnes center*, *tv oddaja*, *etno glasba*, *wellness center*, *tv postaja*,

[14] Na tem mestu navajamo vse zveze z nesklonljivo prvo sestavino, kakršne smo iz obeh korpusov izluščili z zgoraj opisanim metodološkim postopkom, torej tudi primere tipa *al kaida,* ki ne sodijo v začrtani okvir raziskave. Naknadna (kvalitativna) analiza podatkov sledi v razdelku 5.1. Zveze so razvrščene glede na pogostost pojavljanja vseh oblik (imenovalniške in pregibnih), navajamo pa jih v lematizirani obliki, izhodiščnem zapisu narazen in z malimi črkami (luščenje podatkov je potekalo neobčutljivo na velikost začetnic). Delitev na lastnoimenske in občnoimenske zveze je zgolj ponazoritvena.

tempera barva, *pop kultura*, *tv dnevnik*, *avto magazin*, *kiper prikolica*, *avto šola*, *html datoteka*, *jazz festival*, *bruto znesek*, *video kaseta*, *moto šport*, *banana republika*, *lcd zaslon*, *ip telefonija*, *b liga*, *cd plošča*, *hot dog*, *ph vrednost*, *golf igrišče*, *klub kartica*, *b kategorija*, *hip hop*, *č člen*, *klima naprava* in *ip naslov*.

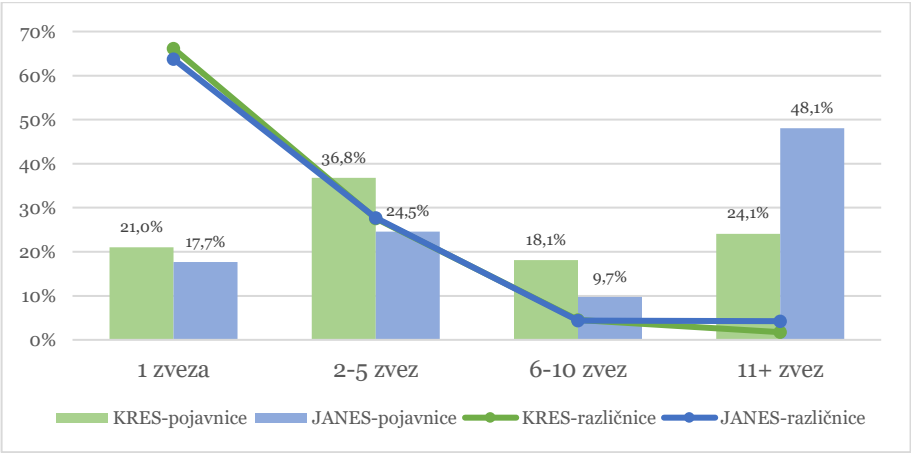
Med 50 najpogostejšimi lastnoimenskimi zvezami v korpusu Janes so: *manchester united*, *coca cola*, *opel victoria*, *factor banka*, *opel astra*, *real madrid*, *premier liga*, *wall street*, *xenon kit*, *janez janša*, *big bang*, *orto bar*, *hong kong*, *van gaal*, *banja luka*, *bin laden*, *bmw forum*, *union olimpija*, *xenon žarnica*, *bmw serija*, *lj obvoznica*, *top gear*, *mercator center*, *opel kadet*, *manchester city*, *xenon luč*, *seat leon*, *ac milan*, *sport klub*, *rock otočec*, *poli salama*, *avto net*, *ford fiesta*, *fiat bravo*, *tv klub*, *a kanal*, *new york*, *pro plus*, *pop tv*, *renault servis*, *sagnac interferometer*, *x factor*, *jugo zastava*, *xenon žaromet*, *red bull*, *chuck norris*, *pika kartica*, *diners kartica*, *renault laguna* in *baby center*.

Med 50 najpogostejšimi občnoimenskimi zvezami v korpusu Janes so: *video posnetek*, *fb stran*, *led dioda*, *led luč*, *buba švab*, *banana republika*, *tv program*, *foto utrinek*, *sd kartica*, *stand-up komedija*, *tv postaja*, *fb profil*, *tv oddaja*, *leasing hiša*, *tv serija*, *fashion vek*, *tv dnevnik*, *park senzor*, *pvc okno*, *big brother*, *tv ekran*, *led bliskavica*, *kasko zavarovanje*, *dpf filter*, *mainstream medij*, *top forma*, *fb prijatelj*, *usb ključek*, *slick guma*, *qr koda*, *lizing hiša*, *servo volan*, *tuš kabina*, *rock glasba*, *chip tuning*, *avto hiša*, *tv kanal*, *video vsebina*, *top model*, *dizel motor*, *alfa samec*, *tv hiša*, *fuzbal tekma*, *trigger točka*, *led lučka*, *pop kultura*, *tv prenos*, *dsg menjalnik*, *sms sporočilo* in *šoping center*.

4.2 Raznolikost samostalnikov v vlogi neujemalnega levega prilastka

Korpus Janes poleg večjega deleža in daljšega seznama zvez vsebuje tudi raznolikejši nabor samostalnikov, ki nastopajo v skladenjski vlogi levega neujemalnega prilastka (15 proti 22 različnih prilastkov na milijon pojavnic).

Kot je razvidno iz grafa na sliki 1, oba korpusa izkazujeta primerljive deleže prilastkov glede na število različnih zvez, v katerih se ti pojavljajo, pri čemer se v obeh korpusih večina prilastkov pojavlja zgolj v eni sami zvezi (66,2 % vseh različnih prilastkov v Kresu oz. 63,8 % vseh različnih prilastkov v Janesu), npr. *union (olimpija)*, *skb (banka)*, *chuck (norris)*, *coca (cola); ects (točka)*, *qr (koda)*, *qwerty (tipkovnica)* in podobno. Dejstvo, da se večina prilastkov pojavlja zgolj v omejenem naboru zvez, potrjuje, da se ta skladenjski mehanizem povezuje predvsem s pregibanjem stalnih besednih zvez, pri katerih je torej celotna zveza tista, ki določa način rabe oz. pregibanja prilastka.



Slika 1: Delež prilastkov glede na število različnih zvez, v katerih se prilastek pojavlja v korpusih Kres in Janes.

V obeh korpusih se le majhen delež prilastkov (1,7 % v Kresu in 4,2 % v Janesu) pojavlja v več kot 10 različnih zvezah, npr. *video* (60 različnih zvez v Kresu, 88 v Janesu), *tv* (58/97), *pop* (33/53), *rock* (26/41), *b* (19, 12), *foto* (18/42), *3d* (14/45), *elektro* (14/16), *solo* (13/11), *jazz* (13/20), *grand* (12/13) itd., a obenem zveze s temi prilastki predstavljajo četrtino vseh pojavitev zvez z neujemalnim levim prilastkom v korpusu Kres oz. skoraj polovico vseh pojavitev v korpusu Janes (četrti par stolpcev na sliki 1).

Seznam tovrstnih pogostih prilastkov v korpusu Janes obenem potrjuje, da gre

za produktivni slovnični mehanizem, ki je povezan tudi oz. predvsem s prevzemanjem novega besedišča, saj na njem prevladujejo časovno oz. področno specifični prilastki, kot dokazujejo primeri tipa *google* (38 različnih zvez), *fb* (37), *eu* (32), *vw* (30), *twitter* (28), *led* (26), *hd* (22), *euro* (21), *gps* (21), *live* (21), *apple* (20), *porno* (20), *online* (19), *usb* (19), *android* (16) itd.

4.3 Prekrivnost zvez z neujemalnim levim prilastkom

888 je zvez, ki se pojavljajo v obeh korpusih, kar predstavlja približno tretjino izluščenih zvez v korpusu Kres, a le desetino izluščenih zvez v korpusu Janes. Nadaljnja analiza zvez, ki se pojavljajo zgolj v korpusu Janes, kaže, da lahko to razliko deloma pripišemo že izpostavljenemu dejstvu, da so v besedilih korpusa Janes tudi sicer pogosteje rabljene prevzete zveze, v katerih se običajno pojavljajo neujemalni levi prilastki, npr. *stand-up* (*komedija*, *scena*), *kickstarter* (*projekt*, *kampanja*), *live* (*stream*, *prenos*), *led* (*luč*, *bliskavica*). Poleg tega se v korpusu Janes kot neujemalni prilastki pojavljajo samostalniki, ki so razmeroma pogosti tudi v korpusu Kres, a v njem redko nastopajo v tej skladenjski vlogi. Med njimi izstopajo zveze s stvarnimi in osebnimi imeni, npr. *Harry Potter*, *Alan Ford*, *Fiat Panda*, v katerih se sklanjajo samo priimki oz. drugi deli imen (npr. *Harry Potterja*), pa tudi zveze z nekaterimi splošneje rabljenimi leksikalnimi enotami, npr. *privat* (*firma*, *sporočilo*), *kasko* (*zavarovanje*, *kritje*), *placebo* (*efekt*, *tabletka*), *rally* (*voznik*, *avto*), *LJ* (*obvoznica*, *tablica*), *SDS* (*poslanec*, *volilec*).

Med pogostimi zvezami, ki se pojavljajo izključno v korpusu Kres, po drugi strani ni izstopajočih skupin. Poleg zvez, ki iz korpusa Janes niso bile izluščene zaradi razlik v označevanju (*žiro* PRID/SAM *račun*) ali premajhne variabilnosti pregibnih oblik (*national geographic*), prevladujejo časovno ali področno specifični primeri (*c člen*, *juan cruz*, *cimos koper*, *geoplin slovan*) in nerelevatni rezultati (*predlog zakon*, *župan občina*).

4.4 Zapisovanje zvez z neujemalnim levim prilastkom

V tretjem koraku kvantitativne primerjave rabe zvez z neujemalnim levim prilastkom nas je zanimalo, v kolikšni meri pri prepoznanih zvezah z levim

prilastkom v korpusih prihaja do variantnosti pri njihovem zapisovanju.[15] Rezultati v tabeli 2 kažejo, da je delež zvez z zapisovalnimi dvojnicami oz. trojnicami (zvez, ki se poleg zapisa narazen v korpusu vsaj enkrat pojavijo tudi v zapisu skupaj in/ali z vezajem) v obeh korpusih približno enak, a presenetljivo nekoliko pogostejši v besedilih korpusa Kres (29 % v korpusu Kres in 25 % v korpusu Janes).

Medtem ko se v korpusu Kres kaže predvsem preklapljanje med zapisoma narazen in z vezajem oz. narazen in skupaj, je variantnost zapisovanja v korpusu Janes enakomerneje porazdeljena med vse tri tipe variantnosti, vključno z variantnostjo vseh treh načinov zapisa.

Načini zapisa zveze	**Kres**	**Janes**
samo zapis narazen (npr. *loto številka)*	71 %	75 %
zapis narazen in z vezajem (npr. *tv film*, *tv-film*)	13 %	8 %
zapis narazen in skupaj (npr. *špas teater*, *špasteater)*	11 %	9 %
zapis narazen, z vezajem in skupaj (npr. *new york*, *newyork*, *new-york*)	5 %	7 %

Tabela 2: Primerjava variantnosti zapisovanja zvez z neujemalnim levim prilastkom.

Po drugi strani je razmerje v deležu variantnosti v zapisovanju nekoliko

[15] Glede na izbrani metodološki pristop v analizo gradiva že v izhodišču zajemamo zgolj zveze, ki se v rabi pojavljajo kot niz dveh zaporednih besed, ločenih s presledkom, ne pa tudi zvez, ki se zapisujejo zgolj z vezajem in/ali skupaj, nikoli pa v zapisu narazen. Dodatna analiza sicer pokaže, da je delež zvez z neujemalnim prilastkom, ki se pri pregibanju pojavljajo v zapisu z vezajem, a nikoli narazen, zanemarljivo majhen: v korpusu Kres so to predvsem zloženke tipa *e-pošta, e-račun, e-poslovanje*, nekatere druge tvorjenke (*t-majica, t-celica, post-socializem, voki-toki*) in nekatera lastna imena (*Browne-Smith, C-Max, Saint-Simon*), v korpusu Janes pa takih zvez ni. Zveze, ki se v rabi pojavljajo samo kot enobesedne pojavnice (tvorjenke), nikoli pa v zapisu narazen ali z vezajem, je mogoče izluščiti šele na podlagi vnaprej znanega nabora prilastkov, ki lahko nastopajo kot prvi del besednozvezne podstave. Teh primerov sicer ni veliko, npr. v korpusu Kres se med 24 samostalniškimi zloženkami s prvo sestavino *solo-* izključno v zapisu skupaj pojavljajo zgolj (zelo redke) pojavnice *soločelist, solosnemanje, solopihalo, soloklarinetist, soloigranje, soloigra, soloharfistka* in *solobalerina*.

drugačno, če pod drobnogled vzamemo samo prilastke, saj se jih v korpusu Kres 68,5 % vedno pojavlja samo v zvezah v zapisu narazen, v korpusu Janes pa je takih 62,1 % vseh prilastkov. Med tistimi zapisovalno nevariabilnimi prilastki, ki se pojavljajo v vsaj treh različnih zvezah, se poleg lastnih imen (*android*, *erasmus*, *microsoft; iphone*, *youtube*, *skype*), nekaterih občnih imen (*bowling*, *metal*, *wellness; kasko*, *dizel*, *press*), nazivov (*don*) in pridevniških prilastkov (*ornk*, *fejk*) v izključnem zapisu narazen v obeh korpusih pojavljajo tudi nekatere kratice v podrednih zloženkah (*dos*, *gif; iq*, *uv*), za katere trenutna jezikovna pravila (SP 2001, § 496) sicer predvidevajo pisanje z vezajem.

5 KVALITATIVNA PRIMERJAVA REZULTATOV

5.1 Kategorizacija zvez, ki se pojavljajo v obeh korpusih

Da bi lahko natančneje določili vsebino izluščenih podatkov, smo 888 zvez, ki se pojavljajo v obeh korpusih, razvrstili v pet robustnih kategorij, kakor prikazuje slika 2.[16]

[1] Nerelevantni rezultati (150 primerov; 17 %): raznovrstne kombinacije, ki so ustrezale pogojem luščenja, vendar niso relevantne za raziskavo (*york city*, *pearl jama*, *družba človek*).

[2] Lastna imena (205 primerov; 23 %), tako domača (*butan plin*, *ford fiesta*) kot tuja (*financial times*, *indiana jones*).

[3] Citatna oz. polcitatna poimenovanja (37 primerov; 4 %), npr. *after party*, *bad boy*, *fair play*, *press center*, *team building*.

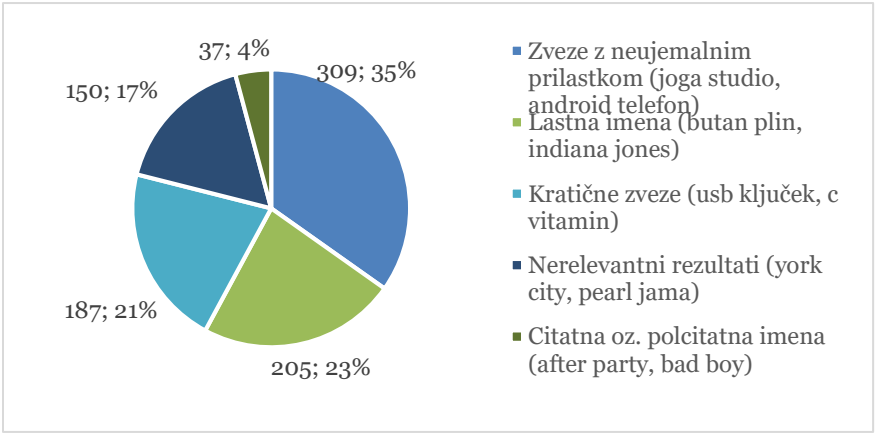
[4] Kratične zveze (187 primerov; 21 %), npr. *rtv prispevek*, *usb ključek*, *c vitamin*, *lcd zaslon*, tudi *zf film*, *fb stran*.

Občna imena z nekratičnim prilastkom, tako z neujemalnim samostalniškim prilastkom, ki je bodisi lastno *(android telefon*) bodisi občno ime *(joga studio*), kot tudi zveze z okrajšano prvo sestavino *(info točka*) ali neujemalnim

[16] Zglede navajamo v zapisu z malimi črkami. Prav tako v zgledih ne navajamo vseh variant zapisa skupaj / narazen / z vezajem: privzeta oblika zapisa pri zgledih je narazen, izjeme od tega načela pa so v besedilu posebej napovedane.

pridevniškim prilastkom *(mikro podjetje)*.

Za nadaljevanje raziskave so zanimive predvsem zveze s kratico v prvem delu in zveze z nekratičnim neujemalnim prilastkom. Skupno ti dve skupini predstavljata 56 % vseh prekrivnih zvez. Ta podatek je relevanten, ker nakazuje, v kolikšni meri obravnavani skladenjski vzorec v rabi presega vrsto zvez, ki se jim jezikoslovne razprave običajno posvečajo (tj. za zapis narazen/skupaj težavne zveze, kakršne obravnavamo tudi v nadaljevanju). Izvorni načrt raziskave je bil še natančneje razvrstiti zveze z neujemalnim prilastkom, in sicer skladno z Dobrovoljc in Jakop (2011: 114) na zveze z (I) okrajšano prvo sestavino (*eko šola*), (II) nesklonljivimi pridevniki (*mini krilo*) in (III) samostalniki v pridevniški rabi (*golf igrišče*). Vendar so se v praksi izkazale težave z enoznačnim razmejevanjem podatkov v naštete skupine, zato v nadaljevanju raznovrstne zveze obravnavamo skupaj.



Slika 2: Kategorije 888 zvez, ki se pojavljajo v korpusih Janes in Kres.

5.2 Zapisovanje zvez, ki se pojavljajo v obeh korpusih

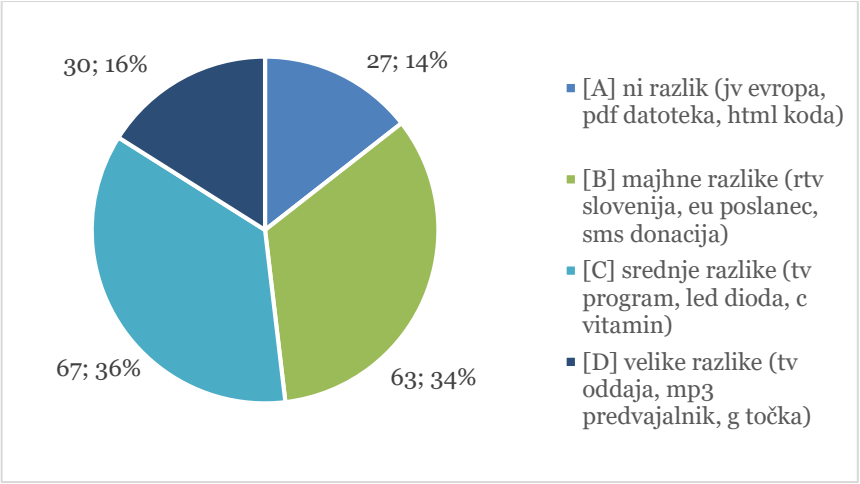
V naslednjem koraku raziskave nas je zanimalo, v kolikšni meri se obravnavana korpusa razlikujeta glede trendov v zapisu besednih zvez tipa *USB ključek/USB-ključek* in *joga studio/jogastudio*. Za vse ustrezajoče podatke so bila izračunana razmerja, v kolikšnem deležu se posamezna zveza pojavlja zapisana narazen, skupaj ali z vezajem. Nato smo deleže primerjali med obema korpusoma in zveze razvrstili v štiri skupine:[17]

[1] Zveze, pri katerih **ne prihaja do razlik**, npr. *loto številka*, *tempera barva*, *pat pozicija*, ki se v obeh korpusih pišejo izključno narazen.
[2] Zveze, pri katerih se posamezni deleži **razlikujejo do 25 odstotnih točk**, npr. *pop pevka* se v Janesu zapisuje narazen v 99,3 %, v Kresu pa v 90,2 % primerov.
[3] Zveze, pri katerih je **razhajanje med 25 in 50 odstotnimi točkami**, npr. *solo petje* se v korpusu Janes zapisuje narazen v 71,7 %, v Kresu v 45,1 % primerov.
[4] Zveze, pri katerih so **razhajanja večja od 50 odstotnih točk**, npr. *lcd zaslon* je v korpusu Janes zapisan narazen v 97,7 %, v Kresu pa v 47 % primerov.

Čeprav pri redko rabljenih zvezah nekoliko manj zanesljive, so se na tovrsten način opredeljene razlike izkazale za ustrezno izhodišče ugotavljanja smiselnosti uporabe korpusa Janes kot komplementarni vir ob korpusu Kres, omogočile pa so tudi osnovno identifikacijo trendov jezikovne rabe, ki se v korpusu Janes kažejo drugače kot v korpusu Kres.

Slika 3 prikazuje razlike v zapisu zvez, pri katerih je na prvem mestu kratica. Teh je med analiziranimi podatki 187 in, kot je bilo izpostavljeno v razdelku 2, naj bi se glede na pravopisna pravila (§ 496) dosledno zapisovale z vezajem.

[17] V raziskavi podatke razvrščamo v vnaprej določene velikostne razrede. Razlog je heterogenost podatkov, o kateri pišemo v nadaljevanju, predvsem velike razlike v variantnosti zapisovanja zvez tipa *led dioda* in *rock legenda* (Sliki 3 in 4), ki jih je lažje primerjati z uporabo univerzalnih (čeprav intuitivno razmejenih) razredov. V nadaljevanju bi bilo smiselno poskusiti ključne odstotkovne meje določiti tudi na osnovi obravnavanega gradiva.

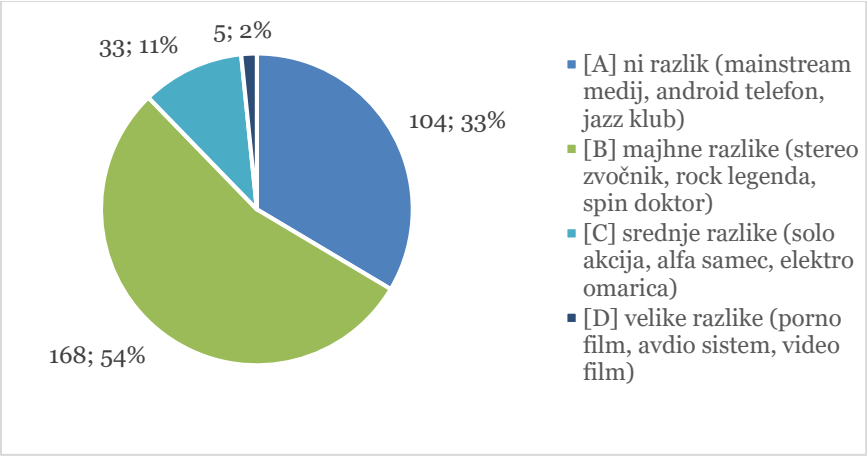


Slika 3: Razlike zapisa kratičnih zvez v korpusih Kres in Janes.

Pri zvezah, ki jih najdemo v skupinah [A] in [B], je v obeh korpusih najti tendenco k zapisovanju zvez narazen. Pri zvezah, ki jih najdemo v skupinah [C] in [D], v Janesu po večini prevladuje zapis narazen, v korpusu Kres pa se zapis narazen giblje med 30 in 75 % v skupini [C] oz. med 12 in 49 % v skupini [D]. Razlike so, po pričakovanjih, na račun zapisa z vezajem. Korpus Janes kaže nekoliko velikodušnejšo rabo vezaja pri zvezah, kjer je na prvem mestu posamezna črka, vendar tudi pri slednjih ne dosledno: več kot 50-odstotno pojavitev z vezajem v korpusu Janes izkazujeta samo primera *e-naslov* (v 93,1 %) in *b-vitamin* (v 61,5 % primerov).[18]

Slika 4 prikazuje razlike v zapisovanju zvez z nekratičnim neujemalnim levim prilastkom. Teh je med analiziranimi podatki 309, trenutna jezikovna pravila za te zveze predvidevajo zapis skupaj ali narazen, kot je razloženo v razdelku 2 (oz. kot strnjeno predstavljata Dobrovoljc in Jakop 2011: 113–122).

[18] V primerjavi s 43 tovrstnimi primeri v korpusu Kres (razlog, da jih ni več, gre iskati tudi v specifikah izbrane metodologije).



Slika 4: Razlike zapisa občnih imen z nekratičnim prilastkom.

Če pri kratičnih zvezah v kategorijah [C] in [D] najdemo 52 % zvez, je na sliki 4 ta delež le 12-%. Splošno gledano sta korpusa torej pri zapisovanju občnih imen z nekratičnim prilastkom skladnejša in po večini gre za skladnost v zapisu narazen. Vendar pa zapis narazen ni dominanten pri prav vseh posameznih primerih; v Janesu več kot 50-odstotno pojavitev zapisa skupaj izkazuje 18 primerov: *avtocesta*, *videoposnetek*, *fotogalerija*, *videospot*, *avtošola*, *avtohiša*, *kinodvorana*, *elektromotor*, *motošport*, *fotozgodba*, *videokaseta*, *turbomotor*, *betablokator*, *avtosalon*, *fotodelavnica*, *elektroinženir*, *narkokartel* in *videokonferenca*. V korpusu Kres je takih primerov 40: *avtocesta*, *elektromotor*, *videospot*, *fotogalerija*, *avtohiša*, *nacionalsocialist*, *kinodvorana*, *videozaslon*, *avtosalon*, *videokaseta*, *elektroinženir*, *fotozgodba*, *videokonferenca*, *videonadzor*, *avtošola*, *evroobmočje*, *pornofilm*, *elektromaterial*, *betakaroten*, *narkokartel*, *videofilm*, *videokamera*, *videoigrica*, *videoigra*, *avtoklub*, *elektroomarica*, *videoposnetek*, *avdiosistem*, *baskitarist*, *fotoutrinek*, *videosporočilo*, *fotoalbum*, *solopetje*, *videoprodukcija*, *fotonatečaj*, *videoprispevek*, *avdiokaseta*, *fotostudio*, *evrokovanec*, *videoprojekcija*.

5.3 Podatki na ravni posameznega prilastka

Značilno za analizirane podatke je, da se raba posameznega prilastka v različnih zvezah razlikuje. Če si ogledamo skupine zvez, ki vsebujejo po vsaj tri različne primere z enakim prilastkom, dobimo naslednje rezultate:

[1] V obeh korpusih se dosledno izrisuje trend pisanja narazen pri skupini zvez, kjer je prilastek lastno ime (*android*, *erasmus*, *linux*). Podobno se kaže pri zvezah s prilastki *fitnes*, *golf*, *reli*, *welness*, *vikend*, *house*, *jazz*, *latino* in *metal*.

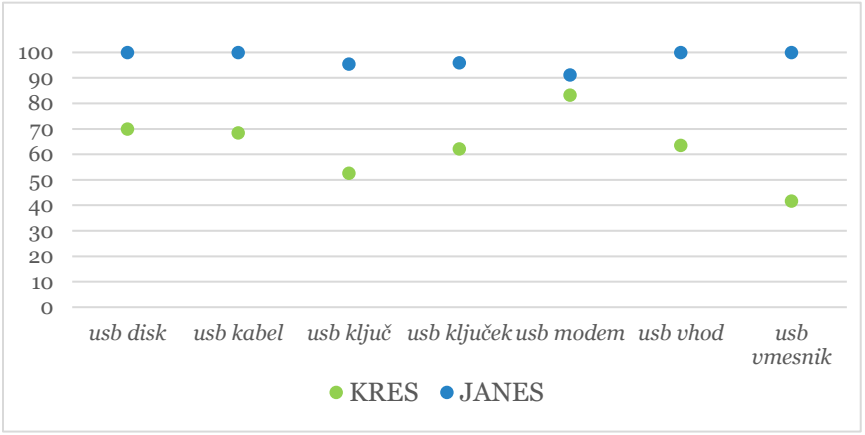
[2] Nekoliko manj skladen je zapis zvez s prilastki *avto*, *bas*, *beta*, *foto*, *kino*, *seks*, *pop* in *rock*. Tu se pri posameznih zvezah s temi prilastki pojavljajo odstopanja od splošnega trenda glede zapisovanja narazen/skupaj, vendar so te razlike relativno skladne v obeh korpusih. Na primer, pri zvezah s prilastkom *rock* je po večini opaziti težnjo k zapisovanju narazen. V korpusu Janes je ta tendenca jasneje izražena, z raznolikimi odstopanji (nihanja so med 100 in 84,6 %) pa prevladuje tudi v korpusu Kres: *rock* [*izvajalec*, *klasika*, *kultura*, *scena*, *festival*, *legenda*, *glasba*, *zasedba*] itd.

[3] Nekatere skupine pa prinašajo zapise, ki se razlikujejo tako od zveze do zveze kot tudi med korpusoma. Gre torej za podatke, ki so zajeti v skupinah [C] in [D] na sliki 4. Sem spadajo npr. zveze s prilastki *avdio*, *elektro*, *evro*, *makro*, *moto*, *solo* in *video*.[19] Razlika je praviloma na račun zvez, ki se v Janesu pišejo narazen, v Kresu pa skupaj; vendar tudi pri slednjih ni mogoče generalizirano trditi, da v korpusu Kres zapis skupaj jasno in prepričljivo prevladuje.

Za boljše razumevanje prikazujemo del analiziranih podatkov na slikah 5 in 6,

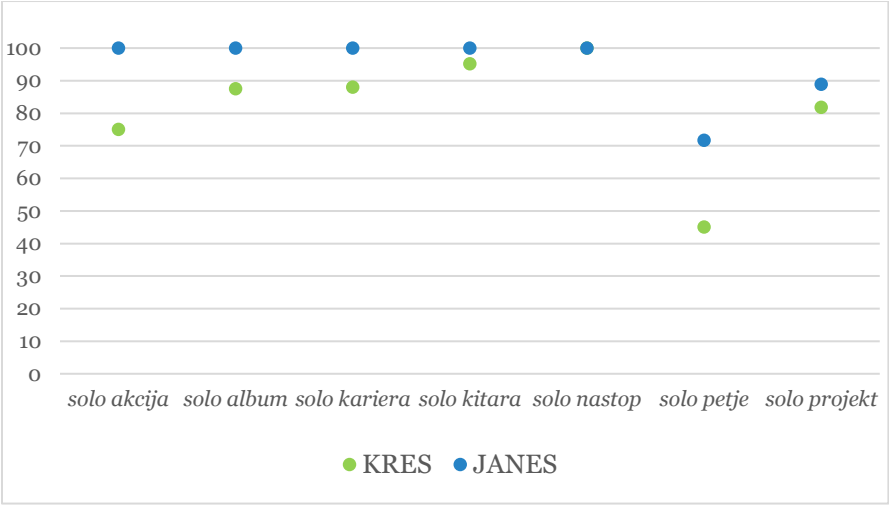
[19] Kot je razvidno, so torej v rabi najbolj heterogeni primeri, ki se končujejo na samoglasnik, za katere bi sicer intuitivno predvidevali, da se tudi v nekorigirani jezikovni rabi pogosteje zapisujejo skupaj (po vzoru medponskoobrazilnih zloženk tipa *zobozdravnik)*. Kot rečeno, se tak trend v rezultatih ne potrdi.
(opomba se nadaljuje na naslednji strani)

ki prikazujeta delež narazen pisanih zvez s prilastkoma *usb*[20] ter *solo*.[21] Morda najbolj presenetljivi rezultat analize je, da so nedoslednosti v rabi v korpusu Kres precej višje kot v korpusu Janes, kar odpira zanimivo vprašanje o dejanskem rezultatu jezikovne korekcije obravnavanega problema, k čemur se vračamo v naslednjem poglavju. Dodati je mogoče, da se v SSKJ in slovarju SP 2001 (www.fran.si, dostop 30. 8. 2015) od navedenih primerov pojavita v zapisu skupaj dve iztočnici, *soloakcija* in *solopetje* (slednja z omembo narazen pisane dvojnice), kar sovpada z nižjim deležem zapisov narazen v Kresu, vendar bi bilo za ugotavljanje neposrednih povezav med referenčnimi priročniki in rabo treba pregledati več gradiva.



Slika 5: Delež zvez s prilastkom *usb*, ki so v korpusih Kres in Janes napisane narazen.

[20] Število pojavitev v korpusu Kres (prva številka v oklepaju) in Janes (druga številka v oklepaju): *usb disk* (7; 33), *usb kabel* (13; 120), *usb ključ* (30; 187), *usb ključek* (28; 243), *usb modem* (25; 83), *usb vhod* (14; 74), *usb vmesnik* (5; 10).
[21] Število pojavitev v korpusu Kres (prva številka v oklepaju) in Janes (druga številka v oklepaju): *solo akcija* (6; 73), *solo album* (21; 50), *solo kariera* (22; 46), *solo kitara* (20; 20), *solo nastop* (10; 17), *solo petje* (46; 43), *solo projekt* (9; 16).



Slika 6: Delež zvez s prilastkom *solo*, ki so v korpusih Kres in Janes napisane narazen.

6 UGOTOVITVE RAZISKAVE IN SKLEP

V prispevku predstavljena analiza je pokazala, da se referenčni korpus Kres in korpus uporabniško generiranih spletnih besedil Janes glede rabe zvez samostalnika z neujemalnim levim prilastkom pomembno razlikujeta. Raba levih neujemalnih prilastkov je v korpusu Janes bistveno pogostejša in bolj raznolika kot v korpusu Kres (razdelek 4.1 in 4.2). Razlike se pojavljajo tudi v načinu zapisovanja zvez: gradivo korpusa Janes izkazuje prepričljive tendence k zapisu narazen, medtem ko v korpusu Kres raba nekoliko bolj variira med zapisom narazen in skupaj/z vezajem (razdelek 4.4). Raziskovanje obravnavanega pojava na gradivu referenčnega korpusa torej podaja bistveno drugačne izsledke kot raziskovanje gradiva, ki ni bilo jezikovno pregledano s strani lektorja oz. urednika; ključna je npr. razlika med ugotovitvijo Dobrovoljc in Jakop (2011: 114), da pravila o rabi vezaja v zvezah tipa *USB-ključek* ne upošteva več kot tretjina uporabnikov, ter ugotovitvijo, da se v gradivu korpusa

Janes to pravilo upošteva zgolj izjemoma.[22] V tem smislu raziskava jasno potrjuje v literaturi že omenjeno predpostavko, da pri določenih jezikovnih pojavih opiranje na jezikovno pregledano gradivo ne zadostuje in korpus Janes se tako kaže kot nujno dopolnilo raziskav, ki se s tovrstnimi vprašanji ukvarjajo.

Pomembne so tudi ugotovitve, ki izvirajo iz metodologije predstavljene raziskave (razdelek 3). Pripravljavcem slovenskih jezikovnih virov je namenjeno opozorilo, da neskladno označevanje korpusov otežuje jezikoslovne analize, ki jih slovenistični prostor v tem trenutku brez dvoma zelo potrebuje. Na drugi strani je treba poudariti nujo, da se korpusne raziskave (tudi na področju normativistike) izvajajo sistematično in ciljno, na podlagi naprednejših podatkovnih luščenj, ki omogočajo celovite povzemalne zaključke, in ob dobrem poznavanju uporabljenih virov ter njihovih označevalnih specifik.

Gradivo nakazuje, da je obravnavani skladenjski vzorec povezan s prevzemanjem stalnih besednih zvez in pregibanjem večbesednih lastnih imen, in potrjuje ugotovitve, da dandanes v slovenščino vstopa predvsem iz angleščine (Gložančev 2012: 125).[23] V domeni normativistike je, da se opredeli do vprašanja, kako ta vstop sprejeti in usmerjati (če se odločimo, da je usmerjanje zaželeno), vendar je ob tem treba izhajati iz dejstva, da je skladenjski vzorec v rabi pogost in produktiven ter da se bo skupaj s prihajajočimi besednimi zvezami njegova prisotnost najverjetneje še nadalje krepila. To izhodišče poudarjajo tudi druge sodobnejše razprave (npr. Gložančev 2012; Kern 2012). V zvezi s to temo je potrebno ponovno spomniti

[22] Čeprav primerjava ni povsem enoznačna, saj Dobrovoljc in Jakop vključujeta tudi zloženke s številko, ki jih v pričujočem prispevku ne analiziramo. Redko raba vezaja v tovrstnih zvezah na spletnih forumih ugotavlja tudi Jakop (2008), sicer na manjši količini gradiva.

[23] Časovni pogled, ki bi razkril, kako točno proces prevzemanja poteka, v raziskavo ni vključen, bi bil pa zelo dobrodošel. Zlasti bi bilo koristno vedeti, v kolikšni meri so zveze z neujemalnimi prilastki zgolj vmesni korak do skladenjske prilagoditve zveze (*rock band* > *rock skupina* > *rokovska skupina*) – in koliko (lahko) jezikovni popravki vplivajo na hitrost teh prilagoditev; pa tudi, katere vrste zvez se ne prilagodijo, bodisi zato, ker jih uporabniki zaradi določenega razloga preferirajo v citatni obliki ali ker skladenjska prilagoditev ni mogoča.

na argument, da »nepridevniški prilastek stoji v slovenskem jeziku *po definiciji* desno od odnosnice« (Toporišič 1974: 35, poud. avt.), če stoji na levi, pa mora biti obravnavan bodisi kot besedotvorna sestavina zloženke, preoblikovan v sklonljiv pridevnik ali premeščen na desno od jedra. S tem argumentom se namreč želja po varovanju slovenske skladnje pred tujimi vplivi preslika oz. prenese na vprašanje besednovrstnega opredeljevanja prilastkov. Ravno iz te problematične povezave, kjer se v vprašanje slovnične kategorizacije vpletajo ideološka načela in kjer se predpisovanje pojavlja na mestu, kjer bi moral stati opis, izvirajo težave tako za prihodnji jezikovni opis, ki mora prilastke obravnavati ob upoštevanju določenih kategorizacijskih omejitev, kot tudi za normativistiko, ki že desetletja išče pot med zapisovanjem zvez z neujemalnimi prilastki narazen in skupaj. Rezultat je že večkrat omenjena nedoslednost v pravilih, kot tudi v rabi.[24]

Variantnost v rabi v obeh korpusih odslikavajo tudi rezultati raziskave (razdelek 5.2). Korpusa se razlikujeta predvsem v zapisovanju kratičnih zvez, kjer so bistvene razlike prisotne kar pri 52 % analiziranega gradiva in pretežno enoznačne: v korpusu Janes prevladuje zapis brez vezaja, v korpusu Kres pa je rabe vezaja več, vendar slednja ne prevladuje dosledno. Pri zapisovanju občnih imen z nekratičnim prilastkom sta korpusa skladnejša, bistveno se razlikujeta v 12 % analiziranih podatkov. Kljub temu je mogoče tudi pri teh podatkih zaključiti, da korpus Janes izkazuje višji delež zapisovanja narazen kot korpus Kres, Kres pa sorazmerno višji, vendar v splošnem še vedno ne prevladujoč delež zapisovanja skupaj. Če smo pri obravnavanem jezikovnem problemu pričakovali, da bo korpus Janes prepričljivejši v zapisu obravnavanih zvez narazen, je nedoslednost v zapisovanju skupaj v korpusu Kres nekoliko presenetila, sploh pri kratičnih zvezah, kjer so jezikovna pravila povsem

[24] Izhodišče, da lahko že besednovrstna kategorizacija nakazuje status normativne (ne)sprejemljivosti jezikovnih elementov, se v sodobnosti pojavlja npr. pri Kern (2012: 143). Mogoče je argumentirati, da je (ne)sprejemljivost skladenjskih vzorcev (tudi v slovarskih priročnikih, če so normativnega tipa) smiselno uporabniku sugerirati drugače kot s pripisom besedne vrste; tudi zato, ker slednji besednovrstnih oznak v tem smislu najbrž sploh ne interpretirajo (o slovarskih uporabnikih npr. Arhar Holdt in dr. 2015).

enoznačna. Razloge gre deloma iskati v sestavi korpusa Kres, ki vsebuje tudi določen segment nelektoriranih besedil, kot tudi v specifikah izbrane metodologije. Vendar je tudi z naštetimi zadržki mogoče zaključiti, da je nelektorirana jezikovna produkcija v rabi doslednejša oz. da lektorska jezikovna regulacija pri obravnavanem problemu pravzaprav dodatno krepi variantnost v jezikovni rabi. Morda je ravno ta ugotovitev ključna motivacija za normativistične raziskave na jezikovno nepregledanem gradivu, v povezavi s tem pa je nujna ustrezna posodobitev aktualnih priročnikov standardnega jezika, po katerih se lektorska regulacija ravna. Primerov, ki bi bili tako očitni, kot je obravnavani v tem prispevku, v slovenščini morda ni veliko, vsekakor pa se jih zdi nujno identificirati in zanje ponuditi rešitve, ki bodo v praksi delovale.

Pri tem je treba razumeti tudi, da podatki izkazujejo predvsem variantnost na ravni prilastka: posamezen prilastek lahko v različnih zvezah kaže različne zapisovalne tendence (razdelek 5.3). Na ravni posamezne zveze pa so trendi zapisovanja v večini primerov jasni: frekvenčno uravnoteženih dvojnic v zapisu ni veliko. Ob tem morda ne bo odveč ponoviti poudarek, da je malo tudi primerov, ki v rabi prevladujejo v zapisu skupaj (npr. *avtocesta*, *elektromotor*, *videospot*, *fotogalerija*): med zvezami z neujemalnimi prilastki predstavljajo v korpusu Janes 5,8 % zvez, kar v celoti 888 raznovrstnih zvez, ki smo jih dobili z luščenjem obravnavanega skladenjskega vzorca, pomeni zgolj 2 % podatkov.[25]

Kakšne sklepe je torej mogoče zapisati na podlagi predstavljenih izsledkov? V prvi vrsti je tu dejstvo, da delo še ni končano. Za dokončno opredelitev stanja bi bilo metodo luščenja treba razširiti, da bi sistematično zajela podatke, ki se vedno zapisujejo z vezajem oz. skupaj, saj je bilo v pričujočem prispevku slednje upoštevano v ločenem koraku. Raziskati bi bilo treba tudi vrsto in pogostost besednozveznih prilagoditev (*tenis igrišče* > *teniško igrišče*, *igrišče za tenis*) v obeh korpusih. Izbrano metodologijo bi bilo mogoče nadgraditi, da bi pri luščenju ločevali podatke glede na način zapisa z velikimi in malimi črkami, saj

[25] Da je primerov, ki prevladujejo v zapisu skupaj, malo, izpostavljajo tudi druge razprave, npr. Gložančev (2012: 125) in Kern (2012: 143).

pri kratičnih prilastkih tipa *tv, sms* zapis z malimi črkami lahko nakazuje razvoj v smer besednega zapisa (*teve, esemes*). V kontekstu predhodnih raziskav bi bilo v nadaljevanju zanimivo primerjati tudi razmerje med skupaj in narazen pisanimi zvezami/zloženkami pri tistih primerih, kjer je v prvem delu že zloženka (*stand-up komedija*, *kickstarter projekt*), na drugi strani pa gradivo osvetliti tudi s časovnega vidika in primerjati trende pri zapisu novih besed oz. zvez s tistimi, ki so v jeziku prisotne že dlje časa. Pri slednjem bi bilo relevantno upoštevati tudi stopnjo podomačenosti (npr. odkriti morebitne razlike v zapisu zvez z *reli – rally* ali *avdio – audio*).

Čeprav iskanje enoznačnega odgovora glede prihodnjega reševanja zapisovanja in kategorizacije zvez z neujemalnim levim prilastkom ni v namenu raziskave, analiza razkriva nekaj uporabnih dejstev, ki jih je pri nadaljnjem delu mogoče upoštevati. Rezultati nedvoumno nakazujejo, da je v jezikovno nekorigirani rabi zapis narazen prevladujoča in privzeta možnost, tako pri zvezah, kjer je na prvem mestu kratica, kot pri ostalih obravnavanih primerih. Če bi ob prihajajoči reviziji jezikovnih pravil želeli slediti izkazani intuiciji jezikovne skupnosti, bi se uporaba vezaja v zvezah tipa *USB-ključek* opustila, pri tipu *solo petje* pa bi prednostno predlagali zapis narazen, razen seveda v primerih, kjer se v rabi izkazuje dejansko variantnost ali ustaljenost zapisa skupaj (tip *avtocesta*). Na drugi strani bi kazalo v prihodnosti vprašanje besednovrstne kategorizacije neujemalnih prilastkov obravnavati osvobojeno od normativnega poslanstva in s primarnim ciljem zagotoviti dosledno obravnavo, ki jo bo mogoče iz jezikovnih priročnikov aplicirati tudi na označevalnike. Trenutno se namreč nedosledna in nejasna kategorizacija prenaša na označevanje korpusnih podatkov in s tem na rezultate luščenj, ki naj bi bili podlaga za nov jezikovni opis – kar je vsekakor veriga, ki bi jo bilo v luči pripravljajočih se jezikovnih priročnikov potrebno prekiniti.

Kar se tiče glavnega vprašanja raziskave, je mogoče skleniti, da je za analizo nekorigirane jezikovne rabe in s tem posredno za oceno intuitivnosti obstoječih jezikovnih pravil za jezikovno skupnost korpus Janes kot gradivni vir izjemnega

pomena. V tem smislu si želimo, da bi čim hitreje našel pot v metodologijo slovenske normativistike in pripomogel k rešitvam, ki bodo ustrezale tako metodološkim kriterijem stroke kot tudi potrebam jezikovne skupnosti po kvalitetnem, sodobnem in razumljivo predstavljenem jezikovnem predpisu.

ZAHVALA

Raziskava, opisana v prispevku, je bila podprta v okviru nacionalnega temeljnega projekta »Viri, orodja in metode za raziskovanje nestandardne spletne slovenščine« (ARRS J6-6842, 2014–2017). Anonimnima recenzentoma se zahvaljujeva za natančno branje ter strokovno utemeljene in konstruktivne predloge.

LITERATURA

Arhar Holdt, Š., Čibej, J. in Zwitter Vitez, A. (2015): S pomočjo uporabniških jezikovnih vprašanj in mnenj do boljšega slovarja. Gorjanc in dr. (ur.): *Slovar sodobne slovenščine: problemi in rešitve*: 196–214. Ljubljana: Znanstvena založba Filozofske fakultete UL.

Arhar Holdt, Š. in Dobrovoljc, K. (2015): Zveze samostalnika z nesklonljivim levim prilastkom v korpusih Janes in Kres. V D. Fišer (ur.): *Zbornik konference Slovenščina na spletu in v novih medijih*: 4–9. Ljubljana: Znanstvena založba Filozofske fakultete.

Crystal, D. (2011): *Internet Linguistics: A Student Guide*. London, New York: Routledge.

Černelič-Kozlevčar, I. (1988): Reševanje besednovrstnih vprašanj v Slovarju slovenskega knjižnega jezika. *Sodobni slovenski jezik, književnost in kultura (Obdobja 8)*: 289–300. Ljubljana: Univerza Edvarda Kardelja, Znanstveni inštitut Filozofske fakultete.

Čibej, J., Fišer, D. in Erjavec, T. (2016): Normalisation, Tokenisation and Sentence Segmentation of Slovene Tweets. *Proceedings of the 10th Language Resources and Evaluation Conference*, v tisku. Portorož:

ELRA.

Dobrovoljc, H. (2008): Vpliv variantnega predpisa na jezikovno rabo (Šest let po izidu Slovenskega pravopisa 2001). V M. Jesenšek (ur.): *Od Megiserja do elektronske izdaje Pleteršnikovega slovarja:* 84–109. Maribor: Filozofska fakulteta.

Dobrovoljc, H. (2013): Smernice jezikovne standardizacije v teoriji, izročilu in praksi. V A. Žele (ur.): *Družbena funkcijskost jezika: vidiki, merila, opredelitve (Obdobja 32)*: 93–99. Ljubljana: Znanstvena založba Filozofske fakultete.

Dobrovoljc, H. in Jakop, N. (2011): *Sodobni pravopisni priročnik med normo in predpisom*. Ljubljana: Založba ZRC.

Dobrovoljc, K., Krek, S., Holozan, P., Erjavec, T., in Romih, M. (2015): *Morphological lexicon Sloleks 1.2*. Dostopno prek: http://hdl.handle.net/11356/1039.

Erjavec, T., Ignat, C., Pouliquen, B., in Steinberger, R. (2005): Massive multi-lingual corpus compilation: Acquis Communautaire and totale. *Proceedings of the 2nd Language & Technology Conference*: 32–36. Poznan, Poland.

Erjavec, T., in Krek, S. (2008): Oblikoskladenjska priporočila in označeni korpusi JOS. *Zbornik Šeste konference Jezikovne tehnologije*: 49–53. Ljubljana: Institut »Jožef Stefan«

Fišer, D., Erjavec, T., Čibej, J. in Ljubešić, N. (2015): Gradnja in analiza korpusa spletne slovenščine JANES. V M. Smolej (ur.): *Slovnica in slovar - aktualni jezikovni opis*: 217-223. Ljubljana: Znanstvena založba Filozofske fakultete.

Fran, slovarji Inštituta za slovenski jezik Frana Ramovša ZRC SAZU, 2014–, različica 3.0. Dostopno prek: www.fran.si (15. 4. 2016).

Gantar, P. (2015): *Leksikografski opis slovenščine v digitalnem okolju*. Ljubljana: Znanstvena založba Filozofske fakultete.

Gantar, P., Krek, S., Kosem, I., Šorli, M., Grabnar, K., Pobirk, O., Zaranšek, P. in Drstvenšek, N. (2012)*: Leksikalna baza za slovenščino*. [Ljubljana]: Ministrstvo za izobraževanje, znanost, kulturo in šport.

Gložančev, A. (2012): Novejša slovenska leksika v luči obravnavae samostalniških zloženk v Slovenskem pravopisu 2001. V H. Dobrovoljc in N. Jakop (ur.): *Pravopisna stikanja: Razprave o pravopisnih vprašanjih*: 125–39. Ljubljana: Založba ZRC.

Grčar, M., Krek, S., in Dobrovoljc, K. (2012): Obeliks: statistični oblikoskladenjski označevalnik in lematizator za slovenski jezik. *Zbornik Osme konference Jezikovne tehnologije*: 89–94. Ljubljana: Institut »Jožef Stefan«.

Jakop, N. (2008): Pravopis in spletni forumi - kva dogaja? V M. Košuta (ur.): *Slovenščina med kulturami*: 315–27. Celovec: Slavistično društvo Slovenije.

Kern, B. (2012). Pisanje skupaj in narazen v Slovarju novejšega besedja slovenska jezika. V H. Dobrovoljc in N. Jakop (ur.): *Pravopisna stikanja: Razprave o pravopisnih vprašanjih*: 141–49. Ljubljana: Založba ZRC.

Korošec, T. (1967): O novejši tvorbi sklopov v slovenščini. *Gospodarski vestnik*, 12. 5. 1967: 180–87.

Krek, S., Gantar, P., Arhar Holdt, Š. in Gorjanc, V. (2016): Nadgradnja korpusov Gigafida, Kres, ccGigafida in ccKres. *Konferenca Jezikovne tehnologije in digitalna humanistika 2016*, v pripravi.

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.

Ljubešić, N., Erjavec, T., in Fišer, D. (2014): Standardizing tweets with character-level machine translation. *CICLing: 15th International Conference on Intelligent Text Processing and Computational Linguistics*, Lecture notes in computer science: 164–175. Kathmandu, Nepal.

Logar, 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.

Logar, N. (2005): Filter vrečka ali filtervrečka, foto posnetek ali fotoposnetek, ISDN paket ali ISDN-paket? V M. Jesenšek (ur.): *Knjižno in narečno besedoslovje slovenskega jezika:* 222–49. Maribor: Slavistično društvo.

Logar, N. (2012): Razmejitev med besednimi zvezami in zloženkami v sodobnem jezikovnem gradivu. V H. Dobrovoljc in N. Jakop (ur.): *Pravopisna stikanja: Razprave o pravopisnih vprašanjih*: 113–23. Ljubljana: Založba ZRC.

Logar, N., Dobrovoljc, K. in Arhar Holdt, Š. (2015): Gigafida: Interpretacija korpusnih podatkov. V M. Smolej (ur.): *Slovnica in slovar - aktualni jezikovni opis*: 467–77. Ljubljana: Znanstvena založba Filozofske fakultete.

Michelizza, M. (2015): *Spletna besedila in jezik na spletu*. Ljubljana: Založba ZRC, ZRC SAZU.

Pogorelec, B. (1965): Ob Poskusnem snopiču slovarja slovenskega knjižnega jezika. *JiS*, 9 (7-8): 232–42.

Popič, D. (2014): *Korpusnojezikoslovna analiza vplivov na slovenska prevodna besedila* [doktorska disertacija]. Filozofska fakulteta UL.

Popič, D. in Fišer, D. (2015): Vejica je mrtva, živela vejica. V M. Smolej (ur.): *Slovnica in slovar - aktualni jezikovni opis*: 609–18. Ljubljana:

Znanstvena založba Filozofske fakultete.

Rigler, J. (1971): H kritikam pravopisa, pravorečja in oblikoslovja v SSKJ. *Slavistična revija*, 19 (4): 433–62.

Slovar novejšega besedja (spletna različica na portalu Fran: 2014). Ljubljana: ZRC SAZU. Dostopno prek: http://www.fran.si/131/snb-slovar-novejsega-besedja (15. 4. 2016).

Slovar slovenskega knjižnega jezika (1970-1991/spletna različica na portalu Fran: 2014). Ljubljana: ZRC SAZU. Dostopno prek: http://www.fran.si/130/sskj-slovar-slovenskega-knjiznega-jezika (15. 4. 2016).

Slovenski pravopis, *elektronska izdaja* (1989 in 2001/spletna različica na potalu Fran: 2014). Ljubljana: Inštitut za slovenski jezik Frana Ramovša ZRC SAZU. Dostopno prek: http://www.fran.si/134/slovenski-pravopis (15. 4. 2016).

Slovenski pravopis (1950). Ljubljana: SAZU. Digitalizirana različica dostopna prek: http://pravopisi.trojina.si/ebooks/pravopis1950/pravopis1950.html (15. 4. 2016).

Slovenski pravopis (1962). Ljubljana: SAZU. Digitalizirana različica dostopna prek: http://pravopisi.trojina.si/ebooks/pravopis1962/pravopis1962.html (15. 4. 2016).

Stabej, M., Dobrovoljc, H., Krek, S., Gantar, P., Popič, D., Arhar Holdt, Š., Fišer, D. in Robnik Šikonja, M. (2016): Slovenščina Janes: Pogovorna, nestandardna, spletna ali spretna? *Slovenščina 2.0*, 4 (2): 101–127.

Škrjanec, I., Popič, D in Fišer, D. (2015): Arheologija začetnice pri stvarnih lastnih imenih. V D. Fišer (ur.): *Zbornik konference Slovenščina na spletu in v novih medijih:* 80–86. Ljubljana: Znanstvena založba Filozofske fakultete.

Toporišič, J. (1971): Pravopis, pravorečje in oblikoslovje v SSKJ I. *Slavistična*

revija, 19 (1): 55–75.

Toporišič, J. (1974): Besednovrstna vprašanja slovenskega knjižnega jezika. *Jezik in slovstvo*, 20 (2–3): 33–39.

Toporišič, J. (1988): Jezikoslovje s Simpozija Obdobja 8. *Slavistična revija*, 36 (4): 437–49.

Vidovič Muha, A. (2011): *Slovensko skladenjsko besedotvorje* [druga, razširjena izdaja]. Ljubljana: Znanstvena založba Filozofske fakultete.

THE VALUE OF THE JANES CORPUS FOR SLOVENIAN LANGUAGE STANDARDIZATION

The main objective of this article is to assess the value of the Janes corpus for research in the field of language standardization. Unlike the existing reference corpora of written Slovenian, the newly available Janes corpus of user-generated content mostly consists of texts that have not been modified by a proofreading expert; it therefore offers a more realistic insight into the trends of language use, as well as the intuitiveness of existing language rules, within a wider language community. We illustrate this methodological potential in a case study of nominal phrases with nonagreeing premodifiers, such as *solo petje* and *RTV prispevek*, by comparing their usage in Janes and the reference Kres corpus. The results reveal: this type of phrases is used more often in Janes and includes a longer list of candidates than in Kres; both corpora include a large number of phrases with variant spelling as either one or two words, irrespective of the premodifier in question; and, somewhat surprising, Janes displays a more consistent language use, suggesting that prescriptive regulation actually increases the level of inconsistency in language use. The article, a revised and enhanced extension of a prior conference paper, concludes with a discussion on possible future approaches to this linguistic issue and advocates for inclusion of Janes into Slovenian language standardisation methodology.

Keywords: Janes corpus, Kres corpus, language standardisation, intuitiveness of language rules, nonagreeing premodifier