Zveze samostalnika z nesklonljivim levim prilastkom v korpusih Janes in Kres

Špela Arhar Holdt,*♦ Kaja Dobrovoljc*

* Zavod za uporabno slovenistiko Trojina, Dunajska 116, 1000 Ljubljana
♦ Filozofska fakulteta Univerze v Ljubljani, Aškerčeva 2, 1000 Ljubljana
spela.arhar@trojina.si, kaja.dobrovoljc@trojina.si

Povzetek

Korpus Janes prinaša uporabniško generirane spletne vsebine (tvite, forume, bloge, komentarje), ki so za razliko od gradiva v ostalih slovenskih korpusih večinoma nekorigirane s strani druge osebe, npr. lektorja ali urednika. Prispevek preverja vrednost novega korpusnega gradiva za normativistične raziskave, in sicer z analizo pogostosti in zapisovanja zvez samostalnika z nesklonljivim levim prilastkom (*solo petje, RTV prispevek*) v korpusih Janes in Kres. Gradivo korpusa Janes razkrije jasnejše trende zapisovanja tovrstnih zvez narazen, v primerjavi s korpusom Kres, kjer je število zvez znatno nižje, trendi v zapisu pa so bolj heterogeni. Rezultati podatkovnega luščenja so v analizi vsebinsko kategorizirani in natančneje preučeni. Na drugi strani izbrana metodologija razgali težave, ki jih pri luščenju podatkov povzroči neenotnost avtomatskega označevanja, in s tem pokaže na zadrege s kategorizacijo nesklonljivih levih prilastkov, ki se ob samostalniku pojavljajo.

Noun phrases with uninflected premodifiers in the Janes and Kres corpora

Unlike other existing Slovenian corpora, the Janes corpus of user-generated content (tweets, forums, blogs, comments) mostly consists of texts that have not been modified by any third party, such as a proofreading expert or an editor. The aim of this paper is to explore the potential of this newly available corpus data for normative language research in a case study of usage frequency and orthography of nominal phrases with uninflected premodifiers, such as *solo petje* and *RTV prispevek*, in the Janes and Kres corpus. In comparison with the Kres reference corpus, which contains significantly less phrases of this type and a more heterogeneous orthography, language data in Janes reveal clearer tendencies towards writing such nominal phrases as two separate words. A subset of the extracted data is further categorized and analysed in more detail, while the methodology itself reveals inconsistencies in automatic POS tagging due to the challenging task of linguistic categorisation of uninflected premodifiers in general.

1 Uvod

Besedilni korpusi kot vzorčene zbirke napredno označenega jezikovnega gradiva predstavljajo izhodišče za raziskovanje avtentične jezikovne rabe, s tem pa nepogrešljivo orodje za izvedbo vseh vrst jezikoslovnih raziskav. Za raziskovanje jezikovnih prvin sodobne pisne slovenščine se (poleg specializiranih virov) uporabljajo referenčni korpusi, danes predvsem Gigafida in Kres, v preteklosti so bili v podobni vlogi korpusi FIDA, FidaPLUS in Nova beseda.

Za naštete vire je značilno, da vsebujejo velik (in težko natančno določljiv) delež lektoriranih besedil. Lektorski posegi v besedila, namenjena javni objavi, so del slovenske jezikovne prakse in njihov obstoj v korpusih ustrezna odslikava realnega stanja. Uporaba lektoriranega gradiva pa je lahko problematična oz. nezadostna v primerih, ko raziskovalca zanimajo primarne, nekorigirane tendence jezikovne rabe, na področju normativistike denimo pri ocenjevanju intuitivnosti določenega jezikovnega pravila za jezikovno skupnost.

Novonastali korpus Janes, ki prinaša uporabniško generirane spletne vsebine (tvite, uporabniške komentarje, bloge in zapise z uporabniških forumov), ponuja možnost za vpogled v jezikovno produkcijo brez lektorskih oz. uredniških posegov, seveda ob upoštevanju specifik v korpus zajetih besedilnih vrst.[1] Kot primer raziskovalnega vprašanja, ki mu tovrstni podatki lahko koristijo, smo v prispevku izbrali rabo zvez samostalnika z nesklonljivim (samostalniškim) levim prilastkom.[2]

2 Predstavitev problema

Vprašanje zapisovanja, sočasno pa tudi jezikovnosistemskega uvrščanja zvez, kot so *alfa samec, servo volan, RTV prispevek* (oz. kot medponskoobrazilne zloženke zapisano skupaj: *alfasamec, servovolan* oz. z vezajem *RTV-prispevek*), je v slovenskem jezikoslovnem prostoru prisotno že desetletja in v tem času so bile temi posvečene – običajno v povezavi s pripravo oz. izidom jezikovnih priročnikov – številne razprave.[3] Ker na tem mestu ni prostora za izčrpen povzetek argumentov, napotujemo bralca k obstoječim pregledom diskusije, npr. v Logar (2005).

Pojav besedilnih korpusov je raziskovalcem ponudil možnost obsežnejših in hitrejših podatkovnih analiz, ne pa tudi enoznačnega odgovora na zgoraj opredeljeno vprašanje. Dobrovoljc in Jakop (2011: 113–114) tako ugotavljata, 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

[1] V prispevku puščamo ob strani sicer zanimivo vprašanje vpliva prisotnosti oz. odsotnosti jezikovnega pregleda s strani druge osebe na količino in naravo avtorjevih samokorekcij.

[2] Za opazovano skupino zvez, za katero so v sorodni literaturi glede na kategorizacijska izhodišča predlagana različna poimenovanja, v prispevku uporabljamo krovni izraz *zveze samostalnika z nesklonljivim levim prilastkom*. Pri tem ne želimo sugerirati absolutne nesklonljivosti prilastkov v tovrstnih zvezah, kakršna bi zahtevala dodatne korpusne analize, temveč predvsem njihovo statistično izstopajočo paradigmatsko fiksiranost v danih zvezah. Nadaljnja zamejitev na *samostalniške* nesklonljive leve prilastke pa opozarja na metodološka izhodišča raziskave, ki so podrobneje razložena v razdelku 3.

[3] Referenčni prispevki k tematiki so mdr. (Rigler, 1971; Toporišič 1971; Vidovič Muha, 1988) v povezavi s slovarjem SSKJ, (Gložančev, 2012) v povezavi s SP 2001, rešitve v novejših slovarskih virih pa predstavljata npr. (Gantar, 2015; Kern, 2012).

(2012). Dosedanje korpusne analize so sicer okrepile argumente za zapis narazen, vendar se ob podatkih izpostavlja vprašanje relevantnosti uporabljenih korpusnih virov, saj na osnovi lektoriranih besedil ni mogoče realno oceniti obsežnosti in narave obravnavanega problema.[4]

V tem prispevku raziskujemo, kako se na ravni obravnavane vrste besed oz. zvez razlikujeta dva za raziskave prosto dostopna korpusna vira: uravnoteženi referenčni korpus Kres (Logar et al., 2012) in korpus uporabniških vsebin Janes (Fišer et al., 2015). S tem preverjamo vrednost korpusa Janes za normativistične raziskave in trenutne možnosti za izvedbo širših, sintetičnih korpusnih raziskav izbrane tematike.

3 Luščenje korpusnih podatkov

Kot potencialne zveze samostalnikov z nesklonljivim levim samostalniškim prilastkom smo iz obeh korpusov izluščili tiste nize dveh zaporednih samostalnikov, pri katerih se dana oblika prvega samostalnika[5] 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 nesklonljivim levim samostalniškim prilastkom (npr. *RTV prispevek*).

Ker se izbrana metoda luščenja deloma opira na besednovrstne oznake, kakršne so bile pojavnicam pripisane v postopku strojnega oblikoskladenjskega označevanja, je pri načrtovanju in interpretaciji izluščenih podatkov tako potrebno upoštevati dve metodološki omejitvi.

Prva izhaja iz dejstva, da sta korpusa označena z različnima (statističnima) označevalnikoma: korpus Kres z označevalnikom Obeliks (Grčar et al., 2012) in korpus Janes z označevalnikom ToTaLe (Erjavec et al., 2005). Čeprav oba označevalnika svoj model znanja gradita na istih jezikovnih virih, leksikonu besednih oblik Sloleks (Dobrovoljc et al., 2015) in učnem korpusu ssj500k (Krek et al., 2013), med njima lahko prihaja do razlik pri tokenizaciji[6] besedila ali obravnavi nekaterih specifičnih jezikovnih sredstev.[7]

Druga, vsebinska, omejitev strojnega označevanja je posledica nedosledne obravnave nesklonljivih prilastkov v obeh omenjenih jezikovnih virih (z izjemo kratic in lastnih imen), zlasti pri vprašanjih besednovrstne kategorizacije (kako ločujemo med pridevniki in samostalniki) in njihove obravnave v besedilnem kontekstu (kako slovnične lastnosti jedra vplivajo na označevanje spola, sklona in števila nesklonljivih pridevnikov oz. sklona nesklonljivih samostalnikov v vlogi pridevnika).[8]

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 nesklonljivih levih prilastkov označevanje poenotiti in iskanje razširiti tudi na pojavnice z nesamostalniškimi oznakami.

4 Kvantitativna primerjava rezultatov

4.1 Pogostost zvez z nesklonljivim 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 nesklonljivim levim prilastkom. Primerjava njihove relativne pogostosti v obeh korpusih razkriva, da se v korpusu Janes pojavlja skoraj enkrat več tovrstnih zvez kot v korpusu Kres, kar kaže na izrazito pogostejšo rabo tega skladenjskega mehanizma v nelektoriranih uporabniških spletnih vsebinah.

	Kres različnice		Kres pojavnice		Janes različnice		Janes pojavnice		Prekrivne različnice
Pogostost	Abs.	Rel.	Abs.	Rel.	Abs.	Rel.	Abs.	Rel.	Abs.
zveze (npr. *alfa samec*)	3.054	**31**	95.897	987	7.840	**61**	212.808	1.662	**888**
prilastki (npr. *alfa*)	1.432	**15**	95.897	987	2.851	**22**	212.808	1.662	**719**

Tabela 1: Pogostost zvez in prilastkov v obeh korpusih.

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 dejstvu, da je v besedilih korpusa Janes tudi sicer pogosteje rabljeno prevzeto besedišče, ki običajno nastopa v vlogi levega nesklonljivega prilastka, npr. *stand-up* (*komedija, scena*), *kickstarter* (*projekt, kampanja*), *live* (*stream, prenos*). Drugo dejstvo je, da se v korpusu Janes kot nesklonljivi prilastki pojavljajo samostalniki, ki so razmeroma pogosti tudi v korpusu Kres, a v njem redko nastopajo v tej

[4] Med običajnimi lektorskimi posegi so popravki na ravni zapisa narazen/skupaj, kot tudi preoblikovanja tovrstnih zvez v obliko, ki je v slovenski skladnji pogostejša.

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

[6] Zloženke z vezajem (npr. *C-vitamin*) označevalnik Obeliks denimo tokenizira kot niz treh pojavnic (C, -, vitamin), označevalnik ToTaLe pa kot eno samo pojavnico (C-vitamin).

[7] Krajšava *html* je denimo v korpusu Kres v vseh pojavitvah označena kot samostalnik, v korpusu Janes pa tudi kot napaka tokenizacije (oznaka Np).

[8] O označevalnih težavah besed, kot so *latino, afro, mini,* piše (Gantar, 2015: 118–120) in izpostavlja, da v jezikovni rabi pri tovrstnih lemah običajno izstopa bodisi samostalniška bodisi pridevniška vloga, kar je mogoče oz. je treba v slovarskem opisu (in posledično pri označevanju) ustrezno upoštevati.

skladenjski vlogi. Med njimi izstopajo zveze s stvarnimi in osebnimi imeni, npr. *Fiat Panda*, *Harry Potter*, v katerih se torej sklanjajo samo priimki, pa tudi zveze z nekaterimi časovno in funkcijsko manj zaznamovanimi prevzetimi občnimi imeni, npr. *privat* (*firma, sporočilo*), *kasko* (*zavarovanje, kritje*), *placebo* (*efekt, tabletka*), *rally* (*voznik, avto*).

4.2 Zapisovanje zvez z nesklonljivim levim prilastkom

V drugem koraku kvantitativne primerjave rabe zvez z nesklonljivim levim prilastkom nas je zanimalo, v kolikšni meri pri prepoznanih zvezah z levim prilastkom v korpusih prihaja do variantnosti pri njihovem zapisovanju. Rezultati 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. Ob rezultatih, ki jih predstavlja Tabela 2, pa je treba upoštevati specifike luščenja podatkov, ki trenutno ne zajema zvez, ki se v korpusih pojavljajo zgolj v zapisu skupaj in/ali z vezajem.[9]

Zapis	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 nesklonljivim levim prilastkom.

5 Kategorizacija prekrivnih zvez

Da bi lahko natančneje določili vsebino izluščenih podatkov, smo 888 zvez, ki se pojavljajo v obeh korpusih, razvrstili v pet robustnih kategorij.[10]

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

[2] **Lastna imena** (zemljepisna, stvarna), tako domača (*butan plin, ford fiesta*) kot tuja (*financial times*), v podatkih pa se pojavljajo tudi osebna imena (npr. *indiana jones, chuck norris*).

[3] **Citatna oz. polcitatna poimenovanja,** npr. *after party, bad boy, fair play, press center, team building*.

[4] **Kratične zveze,** npr. *rtv prispevek, usb ključek, c vitamin, lcd zaslon,* tudi *zf film, fb stran.*

[5] **Občna imena z nekratičnim prilastkom,** tako z nesklonljivim samostalniškim prilastkom, ki je bodisi lastno (*android telefon*) bodisi občno ime (*joga studio*), kot tudi zveze z okrajšano prvo sestavino (*info točka*) ali nesklonljivim pridevniškim prilastkom (*mikro podjetje*).

Rezultati luščenja v točki [5] so v jezikoslovnem smislu precej heterogeni. Medtem ko so zveze tipa *android telefon* in *joga studio* glede na metodologijo pričakovane (in med seboj tudi jasno ločljive), so se zveze tipa *info točka* in *mikro podjetje* med podatki znašli zaradi specifik obravnave v označevalnih virih (gl. pogl. 3). Slednja odslikava težave pri enoznačnem ločevanju med zvezami z okrajšano prvo sestavino (*eko šola*), nesklonljivimi pridevniki (*mini krilo*) in samostalniki v pridevniški rabi (*golf igrišče*).[11] Težave z razmejevanjem, kot tudi želja ugotoviti morebitne tendence v rabi, ki bi razmejevanje lahko utemeljile, so razlog, da v nadaljevanju raziskave raznovrstne zveze obravnavamo skupaj. Ker pa za različne od naštetih skupin veljajo različne normativne smernice glede zapisa, je pri razumevanju in posploševanju podatkov potrebna dodatna previdnost. Rezultate kategorizacije prikazuje Tabela 3.

Kategorija	Število zvez
Nerelevantni rezultati	150
Lastna imena	205
Citatna oz. polcitatna imena	37
Kratične zveze	187
Zveze z nesklonljivim prilastkom	309

Tabela 3: Kategorije prekrivnih zvez.

V nadaljevanju prispevka se od predstavitve rezultatov pogostosti rabe različnih vrst nesklonljivih prilastkov v obeh korpusih premikamo k natančnejšemu pregledu zapisovanja zvez dveh izbranih podskupin: kratičnih zvez [4] in občnih imen z nekratičnim prilastkom [5].

6 Zapisovanje zvez: Janes vs. Kres

V zadnjem 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:

[9] Pilotni poskus luščenja zvez z vezajem, ki se v korpusu Janes nikoli ne pojavijo v zapisu narazen, sicer kaže, da med tistimi s pogostostjo nad 100 pojavitev kot zveze z nesklonljivim prilastkom pojavljajo samo zveze s krajšavo *e-* (npr. *e-volitve*).

[10] Luščenje podatkov je potekalo neobčutljivo na velike začetnice (skupaj obravnavamo *fb stran, FB stran* in *Fb stran*). Pri navajanju zgledov v poglavjih 5 in 6 ne zapisujemo vseh evidentiranih oblik, ampak navajamo vse zapise 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.

[11] Kategorije in primeri po (Dobrovoljc in Jakop, 2011: 114).

[A] Zveze, pri katerih **ne prihaja do razlik,** npr. *loto številka, tempera barva, pat pozicija*, ki se v obeh korpusih pišejo izključno narazen.
[B] 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.
[C] 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.
[D] 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. V nadaljevanju razlike med korpusoma predstavljamo ločeno glede na tip zveze.

6.1 Kratične zveze

Zvez, pri katerih je na prvem mestu kratica, je med podatki 187. Glede na Pravopis (§ 496) naj bi se tovrstne zapisovale z vezajem. Tabela 4 prikazuje, kolikšne so razlike v deležu narazen zapisanih zvez v obeh korpusih.

Rang	Delež	Primeri iz korpusa Janes
[A] ni razlik	14 % zvez	*jv evropa, pdf datoteka, html koda, uefa liga, sv vojna*
[B] majhne razlike	34 % zvez	*rtv slovenija, eu poslanec, dsg menjalnik, nba liga, sms donacija*
[C] srednje razlike	36 % zvez	*tv program, rtv prispevek, led dioda, usb ključek, c vitamin*
[D] velike razlike	16 % zvez	*tv oddaja, mp3 predvajalnik, iq test, 3d model, g točka*

Tabela 4: Razlike zapisa kratičnih zvez.

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).[12]

Različne zveze torej prinašajo različna razmerja v zapisu, pri čemer so nedoslednosti v rabi precej višje v korpusu Kres, kar prikazuje Tabela 5, v kateri so prikazana razmerja za zveze s kratico *USB*.

Zveza	Zapis narazen Kres	Zapis narazen Janes
usb disk	70,0 %	100,0 %
usb kabel	68,4 %	100,0 %
usb ključ	52,6 %	95,4 %
usb ključek	62,2 %	96,0 %
usb modem	83,3 %	91,2 %
usb vhod	63,6 %	100,0 %
usb vmesnik	41,7 %	100,0 %

Tabela 5: Narazen zapisane zveze s kratico *USB*.[13]

6.2 Občna imena z nekratičnim prilastkom

Raznovrstnih zvez z nesklonljivim levim prilastkom je med podatki 309. Trenutna jezikovna pravila za te zveze predvidevajo zapis skupaj ali narazen, kar predstavljata (Dobrovoljc in Jakop, 2011: 113–122). Tabela 6 prikazuje razlike v deležu narazen zapisanih zvez v obeh korpusih.

Rang	Delež	Primeri iz korpusa Janes
[A] ni razlik	34 % zvez	*mainstream medij, android telefon, diesel motor, jazz klub, beta verzija*
[B] majhne razlike	54 % zvez	*stereo zvočnik, rock legenda, pleksi steklo, kino spored, spin doktor*
[C] srednje razlike	11 % zvez	*solo akcija, video predvajalnik, alfa samec, tapas bar, elektro omarica*
[D] velike razlike	1 % zvez	*porno film, avdio sistem, video zaslon, video film*

Tabela 6: Razlike zapisa občnih imen z nekratičnim prilastkom.

Če pri kratičnih zvezah v kategorijah [C] in [D] najdemo 52 % zvez, je v Tabeli 6 ta delež le 12 %. Splošno gledano sta torej v zapisovanju občnih imen z nekratičnim prilastkom korpusa skladnejša in po večini gre za skladnost v zapisu narazen, ki povprečno gledano v podatkih močno prevladuje. 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.

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

[1] V obeh korpusih se dokaj dosledno zapisuje narazen skupina zvez, kjer je prilastek lastno ime (*android, erasmus, linux*). Podobno velja za zveze s prilastki *fitnes, golf, reli, welness, vikend, house, jazz, latino* in *metal*.

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

[13] 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).

[2] V nekaterih skupinah zvez se najdejo pri posameznih primerih glede zapisovanja odstopanja od splošnega trenda, vendar so te razlike relativno skladne v obeh korpusih, npr. pri zvezah s prilastki *avto, bas, beta, foto, kino, seks, pop* in *rock.*

[3] Nekatere skupine pa prinašajo heterogene zapise, ki se tudi med korpusoma razlikujejo (skupini [C] in [D] v Tabeli 6), npr. zveze s prilastki *avdio, elektro, evro, makro, moto, solo* in *video*. Razlika je praviloma na račun zvez, ki se v Kresu pišejo skupaj, v Janesu pa narazen. Zveze s prilastkom *solo* prikazuje Tabela 7.[14]

Zveza	Zapis narazen Kres	Zapis narazen Janes
solo akcija	75,0 %	100,0 %
solo album	87,5 %	100,0 %
solo kariera	88,0 %	100,0 %
solo kitara	95,2 %	100,0 %
solo nastop	100,0 %	100,0 %
solo petje	45,1 %	71,7 %
solo projekt	81,8 %	88,9 %

Tabela 7: Narazen zapisane zveze s prilastkom *solo*.[15]

7 Sklep

V prispevku predstavljena analiza je potrdila tezo, da se referenčni korpus Kres in korpus uporabniško generiranih spletnih besedil Janes glede rabe zvez samostalnika z nesklonljivim levim prilastkom pomembno razlikujeta. Kvantitativni del analize je potrdil hipotezo, da je raba tovrstnih zvez v korpusu Janes bistveno pogostejša kot v korpusu Kres in da se v obeh korpusih pojavlja visok delež zvez, ki v rabi izkazujejo variantnost v zapisovanju (narazen ali skupaj oz. z vezajem).

Natančnejši pregled zvez, ki se pojavljajo v obeh korpusih, je pokazal, da so izluščeni podatki različnih vrst: lastna imena, občna citatna oz. polcitatna poimenovanja, kratične zveze in občna imena z nekratičnim prilastkom. Zadnja skupina je v jezikoslovnem smislu heterogena, saj vsebuje tako primere z nesklonljivim samostalniškim prilastkom, kot tudi primere z nesklonljivim pridevnikom oz. okrajšano prvo sestavino. Raznorodnost rezultatov opozarja na potrebo po nadaljnjih gradivno utemeljenih raziskavah tematike, s korenitim premislekom jezikovnosistemskega razvrščanja nesklonljivih prilastkov, ki trenutno vpliva tudi na priklic podatkov v oblikoskladenjsko označenem gradivu.

Drugi del raziskave je natančneje osvetlil razlike v zapisovanju kratičnih imen in občnih imen z nekratičnim prilastkom v korpusih Janes in Kres. 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 lektorsko reguliranem 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. Poseben izziv za normativistiko, kakor tudi za jezikovni opis, je dejstvo, da se v rabi variantnost pogosto izkazuje že na ravni posameznega prilastka, ki v različnih zvezah kaže različne zapisovalne tendence. Te so, kot smo pokazali v prispevku, v določenih primerih med korpusoma skladne, v določenih primerih različne, vsaj na osnovi obravnavanih podatkov pa se ne zdijo povezane z obstoječo jezikoslovno tipologizacijo prilastkov.

Za dokončno opredelitev stanja bi bilo metodo luščenja treba razširiti, da bi zajela tudi podatke, ki se vedno zapisujejo z vezajem oz. skupaj, po možnosti pa tudi del tipičnih lektorskih besednozveznih parafraz (*tenis igrišče* > *teniško igrišče, igrišče za tenis*), ob tem pa seveda ustrezno zaobiti v prispevku izpostavljene označevalne zadrege. 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 (*stan-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.

Nekoliko posplošena ugotovitev pričujočega prispevka, da je nelektorirana jezikovna produkcija v rabi doslednejša (četudi gre pri tem mestoma za odstop od obstoječega jezikovnega predpisa), da torej lektorska jezikovna regulacija pravzaprav krepi variantnost v jezikovni rabi, vsekakor predstavlja pomemben argument v razpravi o bodočih normativističnih odločitvah glede obravnavane tematike.

8 Zahvala

Raziskava, opisana v prispevku, je bila opravljena v okviru nacionalnega temeljnega projekta »Viri, orodja in metode za raziskovanje nestandardne spletne slovenščine« (J6-6842, 2014–2017), ki ga financira ARRS.

9 Literatura

Helena Dobrovoljc in Nataša Jakop. 2011. *Sodobni pravopisni priročnik med normo in predpisom*. Ljubljana: Založba ZRC.

Kaja Dobrovoljc, Simon Krek, Peter Holozan, Tomaž Erjavec in Miro Romih. 2015. Morphological lexicon Sloleks 1.2, *Slovenian language resource repository CLARIN.SI*, http://hdl.handle.net/11356/1039.

Tomaž Erjavec, Camelia Ignat, Bruno Pouliquen, Ralf Steinberger. 2005. Massive multi-lingual corpus compilation: Acquis Communautaire and totale. V *Proceedings of the 2nd Language & Technology Conference*, str. 32–36. Poznan, Poland.

Darja Fišer, Tomaž Erjavec, Jaka Čibej in Nikola Ljubešić, 2015. Gradnja in analiza korpusa spletne slovenščine JANES. V: *Slovnica in slovar - aktualni jezikovni opis (Obdobja 34)*.

[14] Dodati je mogoče, da se v SSKJ in slovarju Pravopisa (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.

[15] 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).

Polona Gantar. 2015. *Leksikografski opis slovenščine v digitalnem okolju*. Ljubljana: Filozofska fakulteta. V tisku.

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

Miha Grčar, Simon Krek in Kaja Dobrovoljc. 2012. Obeliks: statistični oblikoskladenjski označevalnik in lematizator za slovenski jezik. V: *Zbornik Osme konference Jezikovne tehnologije*, str. 89–94. Ljubljana: Institut Jožef Stefan.

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

Simon Krek, Tomaž Erjavec, Kaja Dobrovoljc, Sara Može, Nina Ledinek in Nanika Holz. 2013. Training corpus ssj500k 1.3, *Slovenian language resource repository CLARIN.SI*, http://hdl.handle.net/11356/1029.

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

Nataša Logar, Miha Grčar, Marko Brakus, Tomaž Erjavec, Špela Arhar Holdt in Simon Krek. 2012. *Korpusi slovenskega jezika Gigafida, KRES, ccGigafida in ccKRES : gradnja, vsebina, uporaba*. Ljubljana: Trojina, zavod za uporabno slovenistiko.

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

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

Jože Toporišič. 1971. Pravopis, pravorečje in oblikoslovje v SSKJ I. *Slavistična revija*, 19(1): 55–75.

Ada Vidovič Muha. 1988. *Slovensko skladenjsko besedotvorje ob primerih zloženk*. Ljubljana: Partizanska knjiga.