UDK 81‘367.4:81‘373:81‘322
Polona Gantar
Filozofska fakulteta Univerze v Ljubljani
apolonija.gantar@guest.arnes.si

Špela Arhar Holdt
Fakulteta za računalništvo in informatiko Univerze v Ljubljani
Filozofska fakulteta Univerze v Ljubljani
spela.arharholdt@ff.uni-lj.si

Senja Pollak
Inštitut Jožef Stefan, Ljubljana
senja.pollak@ijs.si

LEKSIKALNE NOVOSTI V BESEDILIH RAČUNALNIŠKO POSREDOVANE KOMUNIKACIJE

V prispevku opišemo leksikalno analizo izluščenih podatkov za določen kolokacijski okvir iz korpusov Janes in Kres ter predstavimo rezultate, ki so zanimivi za spremljanje leksikalnih novosti v slovenski leksiki in za njeno posodobitev v slovarjih. Izluščene podatke smo analizirali primerjalno glede na aktualne slovarje za slovenščino z vidika še neregistriranega besedišča, z vidika vstopanja v tipične kolokacije in stalne zveze ter z vidika pomenskih sprememb. Jezikoslovna analiza izluščenih kolokacij je med drugim pokazala, da je mogoče s primerjalno analizo prepoznati glavne značilnosti in trende leksikalnih novosti ter zaznati problematične točke, kjer leksikalne novosti zlasti pod vplivom tujejezičnih elementov v slovenščino vnašajo tudi spremembe v zapisu in skladenjski vlogi.

Ključne besede: kolokacije, leksikologija, nestandardna slovenščina, slovarska baza, korpusna analiza

This article presents a lexical analysis of data extracted for a specific collocation window from the Janes and Kres corpora of Slovene. The results of the analysis are shown to be of interest in the monitoring of lexical innovations in Slovene vocabulary and for updating dictionaries. The extracted data were compared with existing dictionaries of Slovene in terms of new vocabulary, typical collocations, and set phrases, as well as semantic shifts. The linguistic analysis of the extracted collocations shows, among other things, that a contrastive comparison can be used to idenitfy the main characteristics and trends regarding lexical innovations, as well as to highlight their problematic aspects–e.g., when lexical innovations–particularly when under the influence of foreign language elements–also introduce changes in spelling and syntactic roles.

Keywords: collocations, lexicology, nonstandard Slovene, dictionary database, corpus analysis

1 Uvod: Predstavitev problema in teoretično ozadje

Digitalna doba prinaša na področje slovaropisja številne novosti, med katerimi so zlasti pomembne optimizacije slovaropisnih delotokov. Računalniško podprta metodologija omogoča časovno učinkovito obdelavo velike količine jezikovnega gradiva in urejanje rezultatov v podatkovne baze, ki so že v izhodišču zasnovane karseda širokonamensko. Premik slovaropisnega interesa od priprave posameznega slovarja do gradnje slovarskih podatkovnih baz prinaša potrebo po širšem spremljanju jezikovne produkcije in beleženju njenih značilnosti ne glede na to, ali bodo podatki objavljeni v določenem slovarskem viru ali ne. Če je bila v preteklosti glavna naloga zbiranje in selekcija jezikovnih podatkov za vključitev v slovar, je danes ključno urejanje podatkov z namenom omogočiti njihovo optimalno izrabo, vključno s prikazom, ki mora biti za uporabnika informativen in smotrn. Prvič v zgodovini slovaropisja se tako srečujemo s podatkovnim obiljem, ki na eni strani prinaša možnost za širok, celosten oz. nepaberkovalen pristop h gradivu, na drugi pa potrebo po dobrem poznavanju uporabljenih postopkov in izvornih jezikovnih virov, saj v nasprotnem primeru rezultatov v podatkovnih bazah ni mogoče ustrezno umestiti in opredeliti.

Med jezikovnimi viri, ki zaslužijo podrobno razumevanje z vidika vrednosti za slovensko slovaropisje, je korpus računalniško posredovane komunikacije (RPK) Janes (Fišer 2018). Komunikacija, kot poteka prek družbenih omrežij, spletnih forumov in blogov, odraža jezikovne lastnosti, ki so značilne za digitalni medij in posebnosti njegove rabe. Korpus Janes tako prinaša besedišče, ki ne odraža le novosti v družbeni realnosti, ampak tudi novosti, ki nastajajo iz značilnosti sodobnega spletnega komuniciranja (Crystal 2001: 17 uvede pojem *netspeak*). Na leksikalni ravni se študije novosti osredotočajo na primerjavo besedišča in stilističnih posebnosti znotraj različnih žanrov (Tagliamonte 2016) ter na zaznavanje neologizmov (Grieve idr. 2017). Na ravni leksikalne semantike se vse bolj uveljavlja prepoznavanje pomenskih sprememb, bodisi z diahronega vidika (Mitra idr. 2015, Hamilton idr. 2016) bodisi z vidika primerjave ciljnega korpusa z referenčnim korpusom (Cook idr. 2013). Za slovenščino je bil pri zaznavanju pomenskih premikov uporabljen pristop distribucijskega modeliranja (Fišer in Ljubešić 2018), rezultati pa so se pokazali kot dragocen aplikativni vir za sodobno slovaropisje.

V pričujočem prispevku h gradivu korpusa Janes pristopamo z avtomatskim luščenjem kolokacij. Kolokacije kot tipične sopojavitve besed predstavljajo že od Firthovih leksikalnih študij, ki jih dobro povzema izjava »You shall know a word by the company it keeps« (1957: 179), enega ključnih modelov za prepoznavanje leksikalnih enot in razločevanje besednih pomenov. Postopek avtomatskega luščenja se osredotoča na kolokacije s samostalniki, tipičnimi za korpus Janes, npr. *sledilec, aplikacija, profil*, ter na kolokacije, ki vsebujejo samostalnike, ki se tipično pojavljajo tako v korpusu Janes kot v korpusu Kres, npr. *slika, stranka, klub*. Metodologija je bila preverjena v predhodni raziskavi, ki se je osredotočala na oceno postopkov luščenja in njihov domet (Pollak idr. 2018). Predhodna raziskava je nakazala, da so kolokacije dober pokazatelj leksikalnih novosti, tako na ravni vstopanja nove leksike in pomenskih sprememb v leksikalni fond slovenščine, kakor tudi na ravni aktualne jezikovne rabe, ki odseva družbeno in

predmetno realnost. Izhajajoč iz prvih rezultatov smo naredili v pričujoči raziskavi korak naprej in gradivo kategorizirali z vidika aplikativne vrednosti za slovensko slovaropisje.

2 Gradivo in opis metodoloških postopkov

Podatke za leksikalno analizo smo pridobili z metodo avtomatskega luščenja kolokacij iz korpusa Janes. Korpus vključuje štiri tipe računalniško posredovanih vsebin: tvite, forumske objave, novičarske komentarje in bloge (Fišer idr. 2018). Različica korpusa v0.3, iz katere smo pridobili gradivo za raziskavo, vsebuje 128.078,158 besed. Ob tem je treba povedati, da v fazi luščenja kolokacijskih kandidatov bistveno izboljšana različica korpusa Janes, tj. v0.4 (Fišer idr. 2016), še ni bila na voljo. Ena glavnih možnosti izboljšave metodološkega pristopa je tako v uporabi nove različice, ki je obsežnejša, časovno posodobljena, dodani so številni metapodatki in popravljene napake v kodiranju.

Kolokacijski podatki, ki smo jih analizirali, zajemajo kolokacije s t. i. tviterskimi lemami in kolokacije z lemami splošnega besedišča. Prve so bile izbrane na podlagi geslovnika, ki je bil izdelan za pripravo slovarja tviterščine (Gantar idr. 2016). Splošne leme pa vključujejo najpogostejše samostalnike v korpusu Janes, ki so hkrati tudi del najpogostejšega besedišča v referenčnem korpusu Kres (Logar in Krek 2012). Luščenje kolokacij je potekalo prek funkcije COLLOCATION orodja Sketch Engine (Kilgarriff idr. 2004). Na podlagi vhodnega seznama samostalniških lem smo izluščili pridevniške, samostalniške in glagolske kolokatorje, ki se nahajajo v neposrednem besedilnem okolju, tj. eno mesto pred in za izbrano lemo. Postopki avtomatskega luščenja kolokacijskih kandidatov iz korpusov, ki so bili že izdelani pri zasnovi Leksikalne baze za slovenščino (Kosem idr. 2013) in pri izdelavi Baze kolokacijskega slovarja (Krek idr. 2016: 102), podrobneje določajo besedilni okvir kolokacij. Na podlagi obeh raziskav je ugotovljeno, da med kolokacijsko najbolj produktivne sodijo strukture s samostalniškim ali glagolskim jedrom: samostalnik + samostalnik v rodilniku, glagol + samostalnik v tožilniku in pridevnik + samostalnik, ki se ujemata v celotni paradigmi. Sledijo strukture z glagolom + samostalnikom v rodilniku. Besedilno okolje, ki zajema 1 besedo pred in 1 besedo za iskanim samostalnikom v naši raziskavi, tako v celoti pokriva najproduktivnejše skladenjske strukture. Poleg tega smo določili, da se morajo kolokatorji za samostalniške leme v korpusu pojaviti najmanj 10-krat in imeti statistično vrednost logDice (Rychlý 2008) najmanj 3. Za splošne leme smo postopek ponovili na uravnoteženem referenčnem korpusu Kres in obdržali le tiste kolokacije, ki se ne pojavljajo na seznamu referenčnega korpusa, torej kolokacije, značilne predvsem za računalniško posredovano komunikacijo.

V nadaljevanju smo najprej avtomatsko nato pa še ročno izločili nerelevantne zveze. Za avtomatsko filtriranje smo razvili postopek za izločitev (a) kolokatorjev, ki vsebujejo URL-naslove, (b) podvojitev na podlagi opuščanja diakritikov v korpusu (npr. *odlicen/odličen film*); (c) podvojitev z različno pripisano besedno vrsto istemu kolokatorju (npr. *film_top-p* in *šit_top-s*); ter (č) kolokatorjev, ki imajo pripisano različno lemo v posameznem korpusu (npr. *edin-p, edini-p*). Pri ročnem pregledu smo izločili še primere z napačnim pripisom leme, kar je pogosto posledica nestandardnega zapisa ali slabega delovanja označevalnikov pri lastnih, zlasti tujih imenih (npr. **arest banka* za *Erste*

Bank). Iz nadaljnje obravnave smo izločili tudi primere, kjer je kombinacija leme in kolokatorja nastopala kot del širše zveze, npr. **piti red* za *piti red bull*, ali pa zveza ni ustrezala podstavni skladenjski strukturi, zlasti ob opustitvi ločil, npr. **jutro prijatelj* za *dobro jutro, prijatelj*. Med izločenimi zvezami je bilo mogoče opaziti tudi primere, ki bi bili ob pravilni lematizaciji zanimivi tudi za nadaljnjo analizo leksikalnih novosti, npr. **gama hrana → GMO hrana*, **vaga vozilo → VAG vozilo, *ozji družina → ožja družina*.

Temu je sledila leksikalna analiza, ki je temeljila na primerjavi izluščenih podatkov s stanjem v aktualnih slovarjih slovenskega jezika, dostopnih prek portala Fran:[1] zlasti s Slovarjem slovenskega knjižnega jezika prve in druge izdaje (SSKJ1 in SSKJ2) ter s Slovarjem novejšega besedja slovenskega jezika (SNB). Analizirane podatke smo razdelili v tri večje skupine:[2] (1) prepoznavanje novih besed, (2) prepoznavanje novih pomenov besed in (3) prepoznavanje aktualne rabe, kot se kaže na ravni kolokacij in stalnih zvez. Kvantitativni rezultati leksikalne analize v posameznih fazah luščenja so prikazani v Tabeli 1.

Tabela 1: Zastopanost kolokacij glede na tip leksikalne novosti v posameznih fazah analize.

		Tviterske leme		**Splošne leme**		**Skupaj**	**%**
		Kolok. levo	**Kolok. desno**	**Kolok. levo**	**Kolok. desno**		
Število vseh lem za leksikalno analizo[3]		1645	867	1105	903	4520	100
Nerelevantne leme		**247**	**141**	**364**	**345**	**1097**	**24,3**
Relevantne leme	**Primer kolokacije**	**1398**	**726**	**741**	**558**	**3423**	**75,7**
a) Novo besedišče	*tribute* [*skupina*]	298	225	52	12	587	17
b) Kolokacijske novosti	*projektna vlada*	834	315	545	220	1914	56
c) Pomenske novosti	*gasilska* [*fotka*] (‚skupinska‘)	59	25	84	31	199	6
d) Drugo	*nikome ništa*	207	161	60	295	723	21

[1] Fran: https://fran.si/.

[2] Glede na primerjalno izhodišče s SSKJ tu predlagana razdelitev v grobem sovpada s tipologijo novejše leksike, ki je bila izdelana z namenom posodobitve in dopolnitve slovarjev Inštituta za slovenski jezik Frana Ramovša ZRC SAZU (Gložančev 2009: 13–15 in Gložančev idr. 2009).

[3] Število lem po avtomatskem filtriranju.

4 Analiza leksikalnih novosti

Kot je prikazano v Tabeli 1, je bilo po ročnem pregledu v nadaljnjo analizo sprejetih skupno 3423 kolokacij tako za splošne kot tviterske leme ali kar 75,7 % avtomatsko pripravljenega gradiva. Na podlagi nadaljnje analize, kot prikazuje spodnji del Tabele 1, je bilo največ leksikalnih novosti zaznanih znotraj kolokacij in stalnih besednih zvez (56 %). Dobrih 20 % kolokacij smo opredelili z oznako »drugo«, kamor smo vključili lastna imena, tujejezične in citatne zveze ter leksikalno manj zanimive tipične sopojavitve. Ta del vključuje tudi frazeološke enote in citatne izraze, ki potrebujejo samostojno jezikoslovno analizo. Sledi na novo registrirano besedišče (17 %) in zaznane pomenske spremembe (6 %).

4.1 Zaznavanje novega besedišča

Med novimi besedami, tj. takimi, ki v obstoječih slovarjih v času naše raziskave niso bile registrirane,[4] predstavljajo opazen delež tujejezični elementi, ki tvorijo kolokacije z v slovenskem leksikalnem fondu že ustaljenimi besedami. Ti elementi postajajo prek različnih stopenj podomačevanja opazen del slovenskega nestandardnega besedišča in navadno (še) nimajo ustreznic v slovenskem standardnem jeziku, posledično je tujemu zapisu prilagojena tudi skladnja, npr. *aftermarket zadeva, statistika trolanja, standup komedija, styling vozilo, tunning [klub, vozilo], rimejk filma, tribute skupina, coworking prostor, [naslov, oddaja] posta, lajkati fotko, velik like, touchscreen zaslon, default nastavitev, narediti backup, solar sistem, online trgovina, [pink, viola] barva, yang oblika, cabaret [večer, comedy], agility tekma,* npr. namesto standardiziranega *kabaretni večer*, *tekma v agilitiju*. Zlasti tipični so tujejezični elementi v obliki kratic in okrajšav,[5] kot npr. *nba [tekma, klub], nhl igralec, lpg sistem, bbq omaka, lgbt [pravice, skupnost, populacija], [tv, mobile, mobilen, web, desktop …] app, [brezplačen, odprt, dostopen] wifi, wifi [hotspot, povezava, omrežje, signal, geslo, dostop, točka …], sms [trženje, marketing, klub]*. Te zveze predstavljajo dobro izhodišče za preučevanje standardizacijskih trendov in stopenj podomačevanja v slovenščini, zlasti ko poskušamo iskati slovenske ustreznike, ki se naravno vključujejo tudi v standardizirani slovenski zapis. Kolokacije in stalne zveze, ki so v besedilih RPK tipično zapisane citatno (ali v različnih stopnjah podomačevanja), razkrivajo vrsto zvez, ki v slovenščini še nimajo ustreznikov, npr. *grammar naci, spoiler alert, control freak, big data, selfie stick*, *road trip*, zato bi bila pri slovarskih posodobitvah in iskanju slovenskih ustreznikov dobrodošla njihova čim bolj avtomatska detekcija in sistematična analiza.

Opazen delež besed, ki vstopajo v tipične kolokacije RPK, je v izhodišču lastnoimenskih. Na tem mestu jih izpostavljamo zaradi trendov, ki so zanimivi za posodobitev pravopisnih pravil, hkrati pa je smiselno razmisliti tudi o njihovi vključitvi v slovarje,

[4] Natančnejšo opredelitev povzemamo po Gložančev 2009: 12: »Novejše besedje […] je tisto besedje […], ki se je v slovenščini pojavilo oz. uveljavilo v obdobju približno zadnjih dvajsetih let in še ni obravnavano v SSKJ«.

[5] V nadaljevanju navajamo vse primere z malimi črkami oz. malimi začetnicami. V rabi, kot jo izkazuje korpus RPK, prihaja do dvojničnih zapisov.

saj je mogoče predvidevati, da se bodo v dvomih uporabniki konzultirali s priročniki. Tipično gre za imena računalniških programov in aplikacij, operacijskih sistemov ipd., kjer se v rabi kaže kolebanje med veliko in malo začetnico: *apple [računalnik, telefon, mobilnik], microsoft excel, google chrome, adobe photoshop, photoshop [mojster, predelava, program], mozilla firefox, linux okolje, samsung tab, amazon kindle, android [sistem, app], uporabljati [android, chrome], ipad tablica, instagram [uporabnik, profil, fotka], facebook [uporabnik, fan], kickstarter [kampanja, projekt], slovenska wikipedija* ipd.

Določen segment v aktualnih slovarjih neregistriranih besed se iz nestandardnih besedil v splošno besedišče vključuje po že znanih besedotvornih postopkih, npr. *profilka* – ‚slika na uporabnikovem FB profilu': zamenjati profilko, nova profilka, *bizarka* – ‚kar je bizarno': petkova bizarka, *naslovka* – ‚naslovna stran': naslovka revije, *bolniška* – ‚bolniški stalež': mesec bolniške, *armaturka* – ‚armaturna plošča': del armaturke. Tudi sicer so nove besede značilne za nestandardne žanre: *[čista, vnebovpijoča] nebuloza, [kazati, pokazati] fakiča, [dobiti, fasati] popizditis, politično prepucavanje, topšit koncert*, hkrati pa je mogoče detektirati tudi nove besede, ki prihajajo v splošno besedišče s področja računalništva, npr. *[zanimiva, interaktivna, dnevnikova] infografika, prednaložena aplikacija, podariti všeček* ipd.

Z vidika preučevanja leksikalnih trendov je treba izpostaviti kolokacije in stalne zveze, ki vključujejo nove besede, ki niso nujno tipične le za nestandardne žanre RPK. Na naš seznam so prišle prek forumskih besedil, vezanih na konkretno strokovno ali terminološko področje, in v večini primerov zahtevajo samostojen leksikalni opis: *samovozeč avtomobil, delavnovarstvena pravica, nenagradno vprašanje, novonabavna vrednost, libertarna stranka, intrigantna zgodba, duetska pesem* itd. Med slovarsko še neregistriranimi besedami je mogoče prepoznati tudi besedne oblike, ki v obstoječih priročnikih niso obravnavane na ravni iztočnic, čeprav so v določenih kontekstih dobile tudi nov pomen, npr. *nasedli projekt, odbita barva, nabite cene*, ali pa gre za premik skladenjske vloge: *alumni član, bitcoin valuta, privat [namen, lajf], komplet avto, top [hrana, cena], horror film, neon barva, [iskati, insajderski] info, nov intro, biznis model* ipd.

Posebno skupino znotraj sklopa novih besed predstavljajo izrazi, ki v slovenskem leksikalnem fondu dejansko niso novi, vendar pa se v obstoječe slovarske priročnike večinoma niso prebili zaradi stopnje nestandardnosti in/ali prevzetosti. Medtem ko je koncept SSKJ1 dosledno sledil načelu kultiviranja slovenskega knjižnega jezika oz. njegovih uporabnikov in je npr. slengovsko besedišče vključeval na podlagi izkazanosti v splošni rabi, torej kot posebej kvalificirani del knjižnega jezika, se je SSKJ2 pri vključevanju slengizmov odločal bolj ali manj na podlagi principa pogostnosti v slovenskih besedilnih korpusih (Ahlin idr. 2014: 122). Vendar pa je težnja po sprejetju večjega dela nestandardnega besedišča izpeljana nesistematično, tipično gre za prevzete besede, ki običajno niso produkt spletnih besedil, ampak so v slovenskem besedišču prisotne že dlje časa. Med kolokacijami, ki jih najdemo v korpusu RPK, aktualni slovarji pa jih ne vključujejo, so: *[poštimati, porihtati, sprobati] biznis, [klima, mašina] laufati, svež*

luft, [prilimati, limati] fotko, zmanjkati štroma, cuker pasti, vključeni primeri, vendar ne v tipičnih kolokacijah ali stalnih zvezah, pa so denimo: *[pocukrana, cukrana] voda* – ‚voda z okusom', *[lokalna, spletna] štacuna, pleh muzika*, *šlep služba* – ‚vlečna služba', *šlepati se na račun (koga)*. Podrobnejša analiza tovrstne nestandardne leksike je med drugim pokazala, da so pomenski premiki na ravni neformalne komunikacije zelo živi, posledično pa je, kot podrobneje prikažemo v naslednjem poglavju, ta pomemben pokazatelj leksikalnega razvoja in njegovih trendov.

4.2 Zaznavanje novih pomenov in pomenskih premikov

Kot rečeno, govorimo v pričujočem prispevku o novih pomenih v odnosu do stanja v aktualnih splošnih slovarjih in v povezavi s kolokacijskim izhodiščem raziskave, kjer pomenske spremembe besed preučujemo v povezavi s konkretnimi kolokacijami ali stalnimi zvezami, v katerih se pojavljajo.[6] O pridobivanju novih pomenov ali pomenskih odtenkov govorimo v primerih, kjer dobi beseda ob obstoječem v SSKJ že registriranem pomenu dodatni pomen ali več pomenov. Ti pomeni so vezani na specializirane segmente jezikovne rabe, v našem primeru zlasti na (a) nestandardno besedišče RPK, (b) področje računalništva ter (c) področje spletne komunikacije na družbenih omrežjih in forumih.

V prvem sklopu, tj. pretežno znotraj nestandardnega besedišča, lahko prepoznamo besede, ki pomensko specifičnost izražajo v leksikalno precej omejenih kombinacijah, npr. *goveja muzika* – ‚narodnozabavna', *gasilska fotka* – ‚skupinska', in v kolokacijah z obsežnejšim nizom kolokatorjev,[7] npr. *hud [igralec, barva], nor [avto], zakon [avto, telefon, film]* – ‚izredno dober, lep', *bolana [fora, šala, ideja]* – ‚skrajno neprimeren ali nesprejemljiv', *trd [pornič, pornografija]* – ‚ekstremen, nazoren', *doza cukra* – ‚sladke jedi, sladkarije', *konkretna doza* – ‚velika, močna', *kontra [smer, efekt]* – ‚nasprotna, ravno obratna', *[avto, bencinar] piti* – ‚porabljati gorivo', *furati [politiko, kampanjo, državo; imidž, stil]* – ‚voditi, upravljati; izražati', *cuzati [denar, državo]* – ‚izvabljati, finančno izčrpavati'.

V drugi sklop sodi leksika, ki je pridobila pomene, vezane na področje računalništva, npr. *zgodovina brskanja, beležiti zgodovino* – ‚seznam predhodnih poizvedovanj na spletu', *nadgradnja [mobilnika, omrežja]* – ‚posodobitev delovanja računalniškega sistema ali njegove vsebine'; *odprt podatek* – ‚vsakomur dostopen za uporabo in ponovno uporabo brez omejitev avtorskih pravic, kopiranja ali objavljanja', *loviti wifi* – ‚biti

[6] Na tem mestu je treba omeniti opredelitev pomenskih premikov, ki je bila izdelana s primerjavo semantičnih profilov besed v referenčnem korpusu Gigafida in podkorpusu tvitov (Fišer in Ljubešič 2018: 204–12). Prekrivne so zlasti kategorije t. i. večjih pomenskih premikov, kamor sodijo premiki, vezani na dnevne dogodke (*vztrajnik*), na razlike v registru (*penzion* – ‚upokojitev oz. pokojnina') in na razlike v mediju (*sledilec* – ‚uporabnik, ki spremlja objave drugih uporabnikov na družbenih omrežjih'). Med t. i. manjšimi pomenskimi premiki pa oženje pomena (*posodobiti* – ‚nadgraditi delovanje računalniškega sistema ali njegove vsebine').

[7] Nekateri od navedenih novih pomenov so predstavljeni v Slovarju tviteršine (http://lexonomy.cjvt.si/slovar-tviterscine/) ali pa so bili registrirani pri prepoznavanju novejše slovenske leksike (prim. Gložančev idr. 2009). Tu navajamo le nekaj takih, ki v obravnavanih slovarjih (še) niso registrirani.

v območju dosega signala‘, *prenesti [aplikacijo, datoteko]* – ‚naložiti z interneta na računalnik ali telefon‘, *odpreti [aplikacijo, link]* – ‚zagnati‘.

Take pomenske specifičnosti so tipične tudi za besedila družbenih omrežij in forumov, kot npr. *[smetiti, zakleniti, zapreti] temo – tema*: ‚zaključena vsebina, ki se obravnava na spletu‘, *smetiti*: ‚dodajati neprimerne ali neustrezne komentarje na spletu‘, *zakleniti*: ‚narediti temo na spletu nedostopno za nadaljnje komentarje‘, *odpreti [temo, debato]* – ‚začeti‘, *[število, milijon] sledilcev* – ‚kdor avtomatsko spremlja objave uporabnikov‘, *deliti [fotko, objavo]* – ‚posredovati na družbenih omrežjih‘, *[zaklenjen, lažen, uradni, tviter] profil* – ‚javna podoba koga na družbenih omrežjih‘, *profilna [fotka, fotografija]* – ‚na uporabnikovem FB ali tviter profilu‘, *[tviter, google, uporabniški] račun* – ‚dogovor ali pogodba s podjetjem za določeno storitev ali izdelek na spletu‘.

Kot izkazuje naše gradivo, prihaja do pomenskih premikov tudi s področja splošnega jezika na bolj specializirana področja, kot je npr. šport: *obračun skupine;* avtomobilizem: *prijava vozila*, *optika vozila, [android, motor] teče;* in telekomunikacije: *[siolova, telekomova] optika*, *[imeti, potegniti] optiko* – ‚optično omrežje‘. Redkeje, a še vedno opazno so zastopane pomenske širitve s specializiranih področij v splošni jezik, npr. *finalist nagrade* – ne samo v športnem smislu, *donacija stranki* – ne samo v dobrodelnem smislu.

Navedeni primeri kažejo, da so besedila RPK glede na svoj specifični neformalni jezikovni položaj dober pokazatelj tako korenitih kot tudi subtilnih pomenskih premikov v sodobni slovenski leksiki. Čeprav gre tipično, a ne izključno, za prepoznavne registre, kot sta npr. sleng in neformalna raba, je detekcija takih sprememb nujna za celovit leksikalni opis v slovarskih in drugih jezikovnih priročnikih, nenazadnje pa tudi za preučevanje trendov leksikalnega razvoja in oblikovanje jezikovne norme.

4.3 Zaznavanje kolokacijskih novosti

Največji delež izluščenih primerov smo opredelili kot kolokacijske novosti. Gre za tri skupine tipičnih besednih sopojavitev: (a) stalne zveze, ki vsebujejo v slovenski leksiki že ustaljene besede, vendar imajo kot celota nov samostojni pomen (b) kolokacije, ki odražajo za določen čas aktualno družbenopolitično situacijo in (c) kolokacije, ki vnašajo v jezik spremembe na ravni jezikovne norme.

V prvi skupini smo prepoznali na novo nastale stalne zveze in kolokacije, značilne za določeno terminološko ali strokovno področje, ki postajajo legitimen del splošnega besedišča (preverjeno so prisotne tudi v referenčnem korpusu Gigafida),[8] in navadno zahtevajo svoj pomenski opis, npr. na področju zdravstva: *rizičen [kontakt, stik, odnos; dogodek], rizična skupina, nadomestno materinstvo, nadomestno zdravljenje, nadomestni zdravnik, narcistična [motnja, oseba, starši, družina];* avtomobilizma: *nadomestno vozilo, kontrolirano vozilo, nižanje glave;* bančništva: *prenesena vloga;*

[8] Korpus Gigafida: http://www.gigafida.net/.

prava: *kazenska točka, varna hiša;* računalništva: *odprta koda, optični kabel, internet stvari, [pokončni, ležeči] način* (pri uporabi pametnega telefona), *verižni mail* itd. Te zveze so zaradi pomenske samostojnosti in prehajanja v splošni jezik sestavni del sodobne leksike, zato bi jih bilo smiselno čim bolj sistematično detektirati in vključevati v proces posodabljanja slovarskih priročnikov.

V drugo skupino sodijo kolokacije, ki odražajo za določeno obdobje aktualno sociolingvistično pogojeno jezikovno rabo. Zanje je značilno, da je njihov nastanek spodbujen s konkretnimi političnimi, gospodarskimi, kulturnimi, športnimi, naravnimi ipd. dogodki, tako na lokalni kot globalni ravni (prim. Fišer in Ljubešić 2018: 205), npr. *[vstajniška, instant, etablirana] stranka, stranka [levega spektra, leve sredine, bloka, tranzicije], infrastrukturni minister, projektna vlada, evro [komisar, cona], komisarska zgodba, iskanje ministra, zamrznjen predsednik, obsojen predsednik; obdavčitev cerkve, rešitev krize, prezadolženo podjetje, sanirati banke; posledica [žleda, žledoloma, ledene ujme], podpis podpore, vzporedni rezultat, povolilna oddaja, pravice migrantov, odvetnik otroka; babji večer, moški čvek* itd. Opazen segment kolokacij in stalnih zvez tega tipa vsebuje lastna imena,[9] npr. *Alenkina vlada, Bratuškova vlada, Cerarjeva [vlada, stranka, program, minister], Jankovičeva stranka, Janševa družina, Kramarjev zakon, Putinova igra, Schumacherjevo stanje*, ali pa gre preprosto za odraz nove predmetnosti in načina življenja v določenem časovnem obdobju na najrazličnejših segmentih človekovega delovanja, npr. *sprememba himne, prestavljanje ure, zaposlitveni klub, izdelek tedna, podjetniški dogodek, odpiranje podjetij, nočni šoping, zavržena hrana, sirski otrok, feminizacija moških, privatizirati vodo* itd. Načeloma gre za tip leksike, ki zaradi svoje (hipotetično) kratke življenjske dobe navadno ni predmet slovarskih opisov, vsaj ne, ko gre za knjižne izdaje. Digitalno slovaropisje ima nasprotno veliko več možnosti, da tak tip novega besedja vključi v jezikovni opis, kjer se podatki nenehno dopolnjujejo in prerazporejajo glede na relevantnost in aktualnost rabe. Na ta način je mogoče uporabniku ponuditi opis leksikalnih novosti in okoliščin njihovega nastanka, ko so te še žive ne glede na njihovo potencialno kratko življenjsko dobo.

Med izluščenimi podatki je treba izpostaviti še kolokacije, ki same po sebi ne izkazujejo leksikalnih novosti in so v tem smislu manj zanimive za posodobitev slovarskih priročnikov, so pa v leksikalnem fondu zelo pogoste in navadno pod vplivom tujejezičnega izvora v slovenščino vnašajo tudi način zapisa, ki ni v skladu z aktualno normo. Izpostaviti je mogoče imena ustanov in organizacij pa tudi dogodkov in prireditev, ki so zaradi splošnosti poimenovanj izrazito na prehodu v generične izraze. Tu je zelo opazen tudi variantni zapis z veliko začetnico, npr. *zaposlitveni klub, gospodarski klub, hiša rešitve, elektro energija, festival družin, tehnološki večer, menjalnica knjig*, ter zapis kratičnih delov brez vezaja, npr. *ekg srca, rtg slika*. Dejstvo je, da so navedeni primeri prepoznani v tipih besedil, ki v izhodišču ne predvidevajo upoštevanja norme, vendar to ne bi smela biti ovira pri prepoznavanju kritičnih točk, kjer najočitneje prihaja do odstopanj, in ugotovitve vključiti v posodobitev normativnih priročnikov za slovenščino.

[9] V navedenih primerih ohranjamo zapis z veliko začetnico, ker se ta v večini primerov kaže tudi v korpusu.

Za konec je treba omeniti, da je najobsežnejši del izluščenih besed, ki tvorijo kolokacije znotraj te skupine, v obstoječe slovarje že vključen. Po eni strani gre za splošno besedišče v tipičnih kolokacijah, kot npr. *peči piškote, reklamni spot, strogi post* ipd., po drugi strani pa zlasti SSKJ2 vključuje opazen del aktualnega nestandardnega besedišča, čeprav ne vedno v tipičnih kolokacijah. Tabela 2 prikazuje primerjavo izluščenih kolokacij v naši raziskavi s podatki v SSKJ2 za lemi *fotka in aplikacija*.

Tabela 2: Podatki, izluščeni iz korpusa Janes, primerjalno s SSKJ2 *(fotka, aplikacija)*.

Podatki iz SSK2 (samo relevantni del gesla)	**Podatki iz korpusa Janes (oblike so ročno urejene glede na ustrezajoče skladenjske vzorce)**
fótka -e ž (ọ́) pog. *fotografija:* uvodna fotka v reviji; napis pod fotko	[profi, spodnja, uradna, panoramska, jutranja, super, družinska, fantastična, vroča, posneta, srednja, fenomenalna, kvalitetna, nora, predzadnja, umetniška, gasilska, profilna, cover, predstavitvena**] fotka** **[**urejati, lajkati, objaviti, najti, izbrati, čakati, pokazati, dati, prilepiti, prilagati, deliti, tvitniti, slediti, naložiti] **fotko** **[**kup, serija, obdelava] **fotk(e)** **fotka** [notranjosti, hrane, snega]
aplikácija -e ž (á) *računalniški program:* razvijati poslovne, spletne aplikacije; uporaba brezplačnih aplikacij / programska aplikacija; računalniške igre in aplikacije	[odprta, kul, nova, spletna, odlična, priljubljena, prednaložena, brezplačna, posebna, foto, internetna, mobilna, urbana] **aplikacija** [uporabljati, naložiti, prenesti, odpreti, ustvariti, razviti] **aplikacijo** **[**različica, razvoj, testiranje, iphon, posodobitev, twitter] **aplikacije**

Prikazana primerjava izpostavlja obseg in pomembnost kolokacijskih podatkov za prikaz aktualne in hkrati naravne jezikovne rabe, posledično pa tudi potrebo po njihovi vključitvi ne samo v kolokacijske ampak tudi splošne slovarje. Ob predpostavki, da je prihodnost slovarskih priročnikov predvsem na spletu, količina vključenih podatkov v slovar ni več problem, kot tudi ne možnost dinamičnega prehajanja med slovarsko bazo in slovarjem.

4.4 Drugo

Kot je razvidno iz Tabele 1, je večji del sicer leksikalno relevantnih izluščenih kandidatov z vidika posodabljanja slovarskih priročnikov manj zanimiv. V kategorijo »drugo« smo tako vključili besedne zveze, ki jih glede na vključenost v upoštevane slovarje nismo preverjali, ker je šlo bodisi za lastna imena, tujejezične oz. citatne zveze ali leksikalno manj zanimive tipične sopojavitve, npr. *trgovina pivoljub, zakonca login,*

climate control, shipping fee, google search. Med drugim smo v to skupino vključevali besedne zveze, ki so ali tvorijo frazeološke enote. V večini primerov gre za nestandardne in prevzete izraze, npr. *spokati kufre, bravo majster, keš pička, pleh pička, kanon futer, držati štango, trgati gate, za mišji kurac, hvala kurcu, zapreti štacuno, nikome ništa,* ki zahtevajo samostojno jezikoslovno analizo, tudi z vidika načina vključevanja citatnih izrazov v slovenska besedila.

5 Sklep in bodoče delo

Jezikoslovna analiza izluščenih kolokacij je pokazala, da je izbrani postopek neposredno uporaben pri zaznavanju različnih leksikalnih novosti, ki vstopajo v besedišče slovenskega jezika, primarno v nestandardnih besedilih spletne komunikacije, od koder pa opazno prehajajo tudi v splošni jezik.

Ugotoviti je mogoče, da nova leksika prihaja v jezik predvsem prek tujejezičnih in lastnoimenskih elementov, npr. z vpeljevanjem velikega števila kratic in krajšav. Pri tem prehaja različne stopnje podomačevanja (*bitcoin, bitkoin, bitkojn*) pogosto pa se iz tujega jezika prenesejo tudi besednozvezne lastnosti, npr. *bitcoin valuta* namesto *valuta bitcoin* ali *bitcoinska valuta*. Tovrstne novosti na eni strani povzročajo nastanek dvojnic, po drugi pa so motivacija za tvorbo po prevzetem vzorcu: *alumni klub, neon barva, privat zabava* ipd. Sistematično luščenje in analiza tovrstnih podatkov omogočata njihovo vključevanje v jezikovni opis, hkrati pa lahko na področju normativistike predstavljata podstat za revizijo jezikovnih pravil (prim. Arhar Holdt in Dobrovoljc 2016).

Posebno pozornost z vidika nadaljnje slovaropisne obravnave zasluži v slovenskem leksikalnem fondu ustaljena nestandardna leksika, npr. *[zadeva, klima, mašina] laufati*, ki jo obstoječi jezikovni opis bodisi izključuje ali jo navaja nesistematično. Celosten pristop k tovrstnemu gradivu omogoča identifikacijo pomenskih premikov na ravni nestandardne leksike, ki je glede na rezultate raziskave pomemben pokazatelj leksikalnega razvoja. V zvezi s tem je pomembno dejstvo, da nova predmetnost pogosto vpliva tudi na pomenske premike pri ustaljenih (standardnih) besedah s širokim pomenskim obsegom, kot npr. *naložiti, teči, prenesti, loviti*, ki pridejo do izraza prav v tipičnih besednih sopojavitvah, kot so kolokacije. Nenazadnje opazen delež izluščenih podatkov na ravni kolokacij predstavlja odraz aktualne jezikovne rabe in subtilnejših pomenskih sprememb, vezanih na aktualno družbenopolitično situacijo, npr. *[vstajniška, instant] stranka*. S sprotno identifikacijo in vključevanjem tovrstnega gradiva je mogoče uporabniku ponuditi opis leksikalnih novosti in okoliščin njihovega nastanka, ko so te še žive, ne glede na njihovo potencialno kratko življenjsko dobo.

Prispevek je pokazal vrednost kolokacijskega gradiva za dopolnjevanje slovarskih podatkovnih baz in utemeljil, da mora biti leksikalni opis jezika celovit, da vloge vse bolj razširjene računalniško posredovane komunikacije in nestandardnega jezika ni mogoče zanemariti in da nadaljnje delo, tudi v smislu jezikovne intervencije in norme, ne more potekati, ne da bi spremembe prepoznali in jih opisali. V nadaljevanju zato sledi predvsem natančnejša tehnična opredelitev, kako pridobljene podatke vključevati

v slovarsko bazo ter kako jih označiti ter medsebojno povezati za optimalno nadaljnjo izrabo. Končni cilj je izvedba obsežnega, celovitega luščenja kot izhodišča za ročni pregled rezultatov, ki avtomatizaciji navkljub ostaja neobhoden osrednji korak slovaropisnega dela.

6 Zahvala

Prispevek izhaja iz štirih temeljnih raziskovalnih projektov: Kolokacije kot temelj jezikovnega opisa: semantični in časovni vidiki (J6-8255), Nova slovnica sodobne standardne slovenščine: viri in metode (J6-8256), Terminologija in sheme znanja v medjezikovnem prostoru (J6-9372) ter Viri, orodja in metode za raziskovanje nestandardne spletne slovenščine (J6-6842), ki jih financira Agencija za raziskovalno dejavnost Republike Slovenije.

Viri in literatura

Martin Ahlin, Branka Lazar, Zvonka Praznik in Jerica Snoj, 2014: *Slovar slovenskega knjižnega jezika.* Druga, dopolnjena in deloma prenovljena izdaja. Izdali Slovenska akademija znanosti in umetnosti, Znanstvenoraziskovalni center Slovenske akademije znanosti in umetnosti, Inštitut za slovenski jezik Frana Ramovša. Ljubljana: Cankarjeva založba, 2014. 1. knjiga 1152 str., 2. knjiga 1150 str. *Jezik in slovstvo* 59/4. 121–27.

Špela Arhar Holdt in Kaja Dobrovoljc, 2016: Vrednost korpusa Janes za slovensko normativistiko. Ur. D. Fišer. *Slovenščina 2.0: Računalniško posredovana komunikacija*, letnik 4 (2). Ljubljana: Trojina, zavod za uporabno slovenistiko. 1–37. Na spletu.

Paul Cook, Jey Han Lau, Michael Rundell, Diana McCarthy in Timothy Baldwin, 2013: A lexicographic appraisal of an automatic approach for detecting new word senses. *Electronic lexicography in the 21st century: Thinking outside the paper: Proceedings of the eLex 2013 conference*. Tallinn, Estonia. 49–69.

David Crystal 2001: Language and the Internet. Cambridge, New York: Cambridge University Press.

John Rupert Firth, 1957: A Synopsis of Linguistic Theory 1930-55. Studies in Linguistic Analysis: Special Volume of the Philological Society. Ur. F. R. Palmer. Selected Papers of J. R. Firth 1952-59. Bloomington in London: Indiana University Press. 168–205.

Darja Fišer, Nikola Ljubešić in Tomaž Erjavec, 2018: *The Janes project: Language resources and tools for Slovene user generated content*. Language Resources and Evaluation. Na spletu.

Darja Fišer, Tomaž Erjavec in Nikola Ljubešić, 2016: JANES v0.4: korpus slovenskih spletnih uporabniških vsebin. Darja Fišer (ur.): *Slovenščina 2.0, 4 (2).* Ljubljana: Trojina, zavod za uporabno slovenistiko. 67–99.

Darja Fišer in Nikola Ljubešić, 2018: Tviti kot leksikografski vir za analizo pomenskih premikov v slovenščini. Ur. D. Fišer, *Viri, orodja in metode za analizo spletne slovenščine*. Ljubljana: ZIFF. 198–226.

Polona Gantar, Iza Škrjanec, Darja Fišer in Tomaž Erjavec, 2016: Slovar tviterščine. Ur. T. Erjavec in D. Fišer. *Zbornik konference Jezikovne tehnologije in digitalna humanistika 2016*. Ljubljana: FF. 71–76. Na spletu.

Alenka Gložančev, 2009: Analitična osvetlitev novejše slovenske leksike: Uvodni razpravni elaborat k leksikalni zbirki Novejša slovenska leksika (v povezavi s spletnimi jezikovnimi viri (NSLSJV). Ur. A. Žele: *Novejša slovenska leksika (v povezavi s spletnimi jezikovnimi viri).* Ljubljana: Založba ZRC, ZRC SAZU.

Alenka Gložančev, Primož Jakopin, Mija Michelizza, Lučka Uršič in Andreja Žele, 2009: *Novejša slovenska leksika (v povezavi z spletnimi jezikovnimi viri).* Ur. A. Žele. Ljubljana: Založba ZRC, ZRC SAZU.

Jack Grieve, Andrea Nini in Diansheng Gio, 2017: Analyzing Lexical Emergence in Modern American English Online. *English Language and Linguistics* 21/1. 99–127.

William L. Hamilton, Jure Leskovec in Dan Jurafsky, 2016: Diachronic word embeddings reveal statistical laws of semantic change. Ur. K. Erk in N. A. Smith. *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*. Berlin. 1489–501.

Adam Kilgarriff, Pavel Rychlý, Pavel Smrz in David Tugwell, 2004: The Sketch Engine. *Proceedings of the Eleventh EURALEX International Congress*. Lorient: Université De Bretagne Sud. 105–16.

Iztok Kosem, Polona Gantar in Simon Krek, 2013: Avtomatizacija leksikografskih postopkov. Tomaž Erjavec, Jerneja Žganec Gros (ur.): *Jezikovne tehnologije: Slovenščina 2.0, 1 (2).* Ljubljana: Trojina, zavod za uporabno slovenistiko. 139–64.

Simon Krek, Polona Gantar, Iztok Kosem, Vojko Gorjanc in Cyprian Laskowski, 2016: Baza kolokacijskega slovarja slovenskega jezika. Tomaž Erjavec, Darja Fišer (ur.). *Zbornik konference Jezikovne tehnologije in digitalna humanistika, 29. september– 1. oktober 2016.* Ljubljana: ZIFF. 101–05.

Nataša Logar in Simon Krek, 2012: New Slovene corpora within the Communication in Slovene project. Prace Filologiczne 63. 197–207.

Sunny Mitra, Ritwik Mitra, Suman Kalyan Mayti, Martin Riedl, Chris Biemann, Pawan Goyal in Animesh Muhherjee, 2015: An automatic approach to identify word sense changes in text media across timescales. *Natural Language Engineeering* 21/5. 773–98.

Senja Pollak Špela Arhar Holdt in Polona Gantar,, 2018: What‘s New on the Internetz? Extraction and Lexical Categorisation of Collocations in Computer-Mediated Slovene. *International Journal of Lexicography* (v tisku).

Pavel Rychlý, 2008: A Lexicographer-Friendly Association Score. Ur. P. Sojka in A. Horák. *Proceedings of the 2nd Workshop on Recent Advances in Slavonic Natural Language Processing, RASLAN 2008, Karlova Studánka*. Brno: Masaryk University. 6–9.

Slovar novejšega besedja slovenskega jezika [SNB], 2012: Ljubljana: Založba ZRC, ZRC SAZU. Na spletu.

Slovar slovenskega knjižnega jezika [SSKJ1]. 1970–1991, 1994, 1997, 1998, 2000, 2008, 2010, 2014. Ljubljana: Založba ZRC, ZRC SAZU. Na spletu.
Slovar slovenskega knjižnega jezika [SSKJ2], druga dopolnjena in deloma prenovljena izdaja. 2014: Ljubljana: Založba ZRC, ZRC SAZU. Na spletu.
Sali A. Tagliamonte, 2016. So Sick or so Cool? The Language of Youth on the Internet. *Language in Society* 45/1. 1–32. Na spletu.

Summary

Computer-assisted lexicographical methodology provides for efficient processing of large quantities of language materials. The extracted results can be organized in databases that are designed from the start to be useful for a wide variety of purposes. On the one hand, the abundance of data allows for an integral, all-encompassing approach to the processed material. On the other hand, it also requires that the database user is familiar with the manner in which the database was compiled, as well as the original language resources used. Without this knowledge, it is impossible adequately to interpret the results obtained. Among the language resources that warrant a more in-depth understanding in terms of their value for Slovene lexicography is the Janes corpus of Slovene computer-mediated communication. This article analyzes the corpus through automatic extraction of collocations, which are a good indicator of lexical innovations, both in terms of the appearance of new lexical elements and semantic shifts in Slovene vocabulary, as well as in terms of contemporary language use, which reflects new the social and objective realities of a language.

The process of automatic extraction focuses on noun collocations typical for the Janes corpus, as well as collocations containing nouns that typically occur both in the Janes and Kres corpora. The material analyzed is automatically obtained, and then automatically and manually filtered. This is followed by a lexical analysis based on the comparison of the extracted data with their representation in contemporary dictionaries of Slovene. The range and adequacy of the methodology have been tested in previous studies. In this article, we analyze the material in terms of its applied value for Slovene lexicography. A large portion of the discussion is focused on the question of the non-standardness of the material used: from new lexical elements and phrase patterns that are borrowed from foreign languages to elements not included in existing language resources because of past lexicographical guidelines. The comparison between existing dictionary entries and the material obtained with the proposed method shows a number of possibilities for obtaining and enriching not only dictionary content, but for modifying basic dictionary design as well. We argue that the starting point for dictionary design should include the entirety of lexical elements, including non-standard and trending material, because this is conducive not only to the identification of problematic aspects and developmental trends, but also to a flexible and dynamic relationship between a dictionary database and its corresponding (digital) dictionary.