LUŠČENJE KOLOKACIJ IZ KORPUSA UPORABNIŠKIH SPLETNIH VSEBIN

Senja Pollak
Institut »Jožef Stefan«, Ljubljana

UDK 811.163.6'373.74'373.43:004.738.5

Prispevek obravnava kolokacije v slovenščini uporabniških spletnih vsebin, natančneje v tvitih, forumih in blogih. Za luščenje kolokacij novega besedišča uporabimo orodje za izdelavo besednih skic, za luščenje za splet specifičnih kolokacij ustaljenega besedišča pa metodo za primerjanje kolokacij med dvema korpusoma. Kolokacijske kandidate analiziramo, obenem pa preučimo razloge za težavnost luščenja.

kolokacije, uporabniške spletne vsebine, besedne skice, neologizmi, jezikovne tehnologije

This paper presents a study of collocations in Slovene user generated content (UGC): in tweets, forums and blog posts. For extracting collocations of newly coined words word sketches are used, while UGC-specific collocations of general vocabulary are extracted using a method for comparing collocations of two corpora. In addition to analyzing collocations the key obstacles in the extraction process are identified.

collocations, user-generated content, word sketches, neologisms, language technologies

1 Uvod

Jezik je živa tvorba, ki se prilagaja novim vsebinam in tehnološkemu razvoju. Široka uporaba svetovnega spleta in možnosti avtomatskega zajema vsebin za sestavo korpusov omogočajo preučevanje jezika uporabnikov spleta. Termin *jezik uporabniških spletnih vsebin* (angl. *user generated content*) ali poenostavljeno kar *spletni jezik* uporabljamo za jezik neformalne komunikacije na forumih, blogih, družabnih omrežjih, kot je Twitter ipd. Za jezik neformalne pisne spletne komunikacije je značilna pogosta raba jezikovnih oblik, ki se odmika od predpisanega standardnega jezika na ravni ortografije (fonetični zapis besed) in skladnje, medtem ko se inovativna raba jezika, značilna za uporabniške vsebine, najočitneje kaže na leksikalni ravni. Poleg neologizmov, ki so skovani za opisovanje novih pojmov spletnih vsebin (npr. *všeček*) ali za opis aktualnih dogodkov (npr. *virantovanje*), so značilne pogostejše rabe pomenov, ki niso značilni za knjižni jezik (npr. *rabiti* v pomenu *potrebovati*) in omogočajo identifikacijo pomenskih premikov (npr. *hud – dober*).

V prispevku nas zanimajo predvsem nove tipične kombinacije besed oz. kolokacije, v katerih nastopajo besede, ki so značilne za spletne vsebine (poglavje 4), prav

tako pa želimo s pomočjo kolokacij identificirati nove rabe ustaljenega besedišča slovenskega jezika (poglavje 5).

2 Motivacija za raziskavo in sorodna dela

Za sodoben in celovit popis slovenščine na leksikalni ravni je treba opraviti analizo kolokacij tudi v uporabniških spletnih vsebinah. Tako želimo identificirati tipična besedilna okolja novih besed (npr. beseda *hipster* se tipično sopojavlja z besedo *pravi*, ki je torej njen kolokator in skupaj tvorita kolokacijo *pravi hipster*) kot tudi nove kolokatorje obstoječega besedišča, ki so lahko tudi pokazatelji pomenskih premikov (npr. *stisniti (roko)* → *stisniti (datoteko)*).

Pristopi k preučevanju kolokaciji se v grobem delijo na frekvenčne in frazeološke (Nesselhauf 2005). Prvi definirajo kolokacije kot statistične sopojavitve besed znotraj določenega okna (Sinclair 1966; Halliday 1966; Firth 1957). Na drugi strani pa pristopi poudarjajo tudi slovnične (Sinclair 1991; Kjellmer 1987) in pomenske lastnosti kolokacij (Cowie 1981, 1994; Benson 1989; Benson idr. 1997). V slednjih so kolokacije definirane kot besedne zveze, ki se nahajajo med prostimi besednimi kombinacijami in idiomi kot dvema ekstremoma pomenske razstavljivosti. Pri luščenju uporabljamo dve metodi; prva je izključno statistična, druga pa upošteva tudi skladenjske relacije; pri analizi in izboru primerov upoštevamo tudi pomenske vidike.

Preučevanje kolokacij s pomočjo korpusov so v slovenščini z vidika leksikografije obravnavali predvsem P. Gantar in Krek (2011), P. Gantar idr. (2009) ter Kosem idr. (2013). Omenjeni avtorji so uporabo besednih skic (Kilgarriff idr. 2004) prilagodili za avtomatsko luščenje leksikalnih podatkov (ALLP) za slovenščino. S terminološkega vidika so bile kolokacije obravnavane v Š. Vintar (2010) ter N. Logar idr. (2014). Poleg leksikografskih vidikov pa na uporabnost preučevanja kolokacij v okviru drugih aplikacij kažejo številne mednarodne raziskave na področjih poučevanja tujega jezika (Orenha-Ottaiano 2012), luščenja informacij (Lin 1998), strojnega prevajanja (Gerber, Yang 1997), razpoznave sentimenta (Yu 2014) itn.

V predstavljeni raziskavi luščimo in analiziramo kolokacije, ki so specifične za korpus uporabniških spletnih vsebin Janes (»spletni jezik«) in jih primerjamo s »splošnim jezikom«, ki ga predstavlja besedilnovrstno uravnoteženi korpus Kres. Že ustaljeno *metodo besednih skic* dopolnimo s *primerjalno metodo* za primerjanje kolokacij poljubnih lem v dveh različnih korpusih, ki smo jo razvili za primerjavo različnih korpusnih parov (glej pilotno študijo v S. Pollak in Š. Arhar Holdt 2015).

Predstavljeni pristop k luščenju kolokacij uporabniških spletnih vsebinah je uporaben tako za leksikografski opis sodobne slovenščine (za dopolnjevanje obstoječih ali pri izdelavi novih virov) kot tudi za razvoj orodij za strojno obdelavo slovenskega jezika.

3 Predstavitev korpusov

Za raziskavo uporabljamo korpus uporabniških vsebin Janes v0.3 (Fišer idr. 2014), ki vsebuje 161 milijonov pojavnic iz tvitov, forumskih sporočil, blogovskih zapisov in komentarjev spletnih novic, objavljenih v obdobju 2001–2015.

Za primerjavo spletne slovenščine s splošno uporabimo korpus Kres (Logar idr. 2012). Kres je iz korpusa Gigafida (prav tam) vzorčeni uravnoteženi podkorpus in ga uporabljamo kot referenčni korpus. Ima 121 milijonov pojavnic in vsebuje stvarna besedila, leposlovje, časopisje, revije in internetne vsebine.

Treba se je zavedati, da izbor korpusov vpliva na rezultate luščenja. Če bi za referenčni korpus uporabili drug korpus, bi bili tudi rezultati drugačni, zato zaključkov ne gre posploševati. Poleg tega se poraja vprašanje odnosa med govorjeno in spletno slovenščino, vendar je podrobnejša raziskava izven okvira tega prispevka. V posameznih primerih preverimo odnos med spletnim in govorjenim jezikom na podlagi korpusa GOS (Verdonik idr. 2013).

4 Luščenje kolokacij novega besedišča z besednimi skicami

Za leme, ki se v korpusu Kres ne pojavljajo, za analizo kolokacij uporabimo *besedne skice* orodja SketchEngine (Kilgariff idr. 2004), ki kolokatorje razvrščajo po slovničnih relacijah. Za hitrejšo obravnavo uporabimo ALLP[1] API (Kosem idr. 2013), ki za določen seznam besed izvozi kolokacije in korpusne zglede (Kosem idr. 2011).

Metodo preizkusimo na novem besedišču korpusa Janes. S frekvenčnega seznama besed v korpusu Janes izberemo 12 samostalnikov, ki se v njem pojavijo vsaj 100-krat, ne pojavljajo pa se v korpusu Kres (*tviteraš*, *profilka*, *všeček*, *bizarka*, *bitcoin*, *vsegliharstvo*, *virantovanje*, *hešteg*, *radkapa*, *tajmlajn*, *oma*, *plinaš*).[2] Ti samostalniki se z izjemo samostalnika *oma* tudi ne pojavljajo v govornem korpusu GOS, slovar SSKJ 2 pa od naštetih vsebuje iztočnici *tviteraš* in *oma.* Seveda sama odsotnost iz (razmeroma majhnega) korpusa GOS še ne pomeni, da so izrazi izključno odraz slovenščine na spletu.

Pri iskanju kolokacij za naštete leme smo zaradi omejitve na vsaj pet pojavnic za omenjeno kolokacijo izluščili kolokacije le za osem od naštetih lem (omejili smo se na štiri relacije). Za izbrane leme smo izluščili kolokacije, kot so *najljubši/znani/pravi tviteraš*, *nova/huda/lepa/dobra profilka*, *meja/število všečkov*, *rudarjenje bitcoinov.* Zanimivi so tudi neologizmi za splošne pojme in njihove kolokacije, npr. *prava/velika bizarka* (»*Včasih se mi zdi, da je cela država ena velika bizarka.*«), in neologizmi, vezani na sodobne politične vsebine, npr. *rezultat/posledica/cena virantovanja*.

5 Luščenje korpusnospecifičnih kolokacij splošnega besedišča s primerjalno metodo

Za splošno besedišče je naša naloga identificirati tiste kolokacije oz. kolokatorje, ki so specifični za korpus uporabniških vsebin in tako omogočajo preučevanje novih rab. Izdelana primerjalna metoda temelji na funkciji razvrščanja rezultatov *SketchDiff* v orodju Sketch Engine, ki smo jo prilagodili za primerjavo kolokacij različnih korpusov (Pollak, Arhar Holdt 2015). Za vsako lemo izvozimo kolokatorje[3] s funkcijo

1 V času te raziskave ALLP uporabi za korpus Janes še ni v celoti prilagojen, zato smo se omejili le na nekatere relacije. V primeru težav smo uporabili tudi izvoz skic osnovnega orodja.

2 Za izbor lem se zahvaljujem Darji Fišer.

collocations, neodvisno od slovničnih relacij in razvrstitvijo po vrednosti *logDice* (Rychly 2008). Kolokatorje izbrane leme primerjamo z mero, ki jo imenujemo *CorpDiff* in predstavlja razliko med vrednostima v specifičnem in referenčnem korpusu. Pozitivne vrednosti pomenijo kolokacije, specifične za korpus Janes, negativne pa za Kres. Osredotočimo se na kolokacije z vrednostjo *CorpDiff* nad 2. Specifičnost tako definiramo z razliko v pogostosti oz. ključnosti, in ne kot izključne pojavitve v enem od dveh korpusov, česar se je pri interpretaciji treba zavedati.

Za primerjavo rabe splošnega besedišča iz korpusov Kres in Janes izluščimo frekvenčne sezname besed po posameznih besednih vrstah. Iz dvajset najbolj pogostih samostalnikov vsakega korpusa izdelamo skupni seznam presečnih občnih samostalnikov, za katere izluščimo kolokatorje po pravkar opisani metodi. Za vsako lemo ohranimo do 50 kolokacijskih kandidatov (z najvišjo CorpDiff vrednostjo) in tako za 15 lem dobimo skupaj 290 kolokacijskih kandidatov. Kljub temu da je ta metoda neodvisna od oblikoskladenjskih oznak, se za analizo te raziskave omejimo na kolokatorje, ki jim je pri samodejnem označevanju korpusa najpogosteje pripisana besedna vrsta *samostalnik*, *glagol* ali *pridevnik*. Za končno obravnavo tako ohranimo 179 kolokacijskih kandidatov, ki jih prek korpusnih konkordanc podrobneje pregledamo.

5.1 Nerelevantni kandidati

Jezik spletnih uporabniških vsebin je za avtomatsko označevanje izjemno zahteven. V trenutnem stanju orodij je kar 36 odstotkov (65 parov <kolokator, lema>) izluščenih kandidatov nerelevantnih in bi jih bilo z izboljšanjem orodij možno izločiti že v fazi luščenja.

Več kot pol nerelevantnih kandidatov je povezanih z **napačno lematizacijo**. Lahko gre za napake lematizacije kolokatorja (npr. v paru <lip, dan> je *lip* napačna lematizacija okrajšave *lep pozdrav*), napačna lematizacija osnovne leme (<delo, norec> izhaja iz napačne lematizacije glagola *delati* v frazemu *delati se norca*), v nekaterih primerih pa gre za lematizacijo na določno namesto nedoločne oblike pridevnika (npr. <levi, del>), kar je razlog za identifikacijo kolokacije kot specifične za korpus Janes, saj je v korpusu Kres lematizirana kot <lev, del> (razlog je morda tudi v tem, da sta korpusa označena z različnima orodjema).

Druga skupina je povezana z **izpuščanjem strešic** v jeziku spletnih uporabniških vsebin. Npr. kolokacijski par <cez, dan> je prav zaradi izpusta strešic napačno razumljen kot specifičen za uporabniške vsebine, čeprav je kolokacija *čez dan* pogosta tudi v referenčnem korpusu. Z rediakritizacijo bi bilo mogoče tovrstne napake luščenja odpraviti.

Pogoste napake predstavljajo kolokacije, izluščene iz besedil **posameznega uporabnika** (oz. domene). Npr. stavek *Pediater je odvetnik otroka* je v korpusu Janes prisoten kar 300-krat, vendar vedno v istem forumu na koncu sporočil istega uporab-

3 Nastavitve: okno 3, min. frekvenca 10 za kolokator in 5 za kolokacijski niz; za samo analizo smo se osredotočili le na kolokacije s frekvenco nad 1 na milijon. Kot listo praznih besed smo dodali znake za emotikone, za potrebe zastavljene raziskave pa tudi izvzeli predloge.

nika. Tovrstne nerelevantne kandidate bi lahko izločili z izboljšanjem metodologije, natančneje z dodajanjem pogoja razpršenosti kolokacije po korpusu.

5.2 Analizirani kandidati

Preostalih 115 kolokacijskih kandidatov (64 odstotkov), ki ne pripadajo zgornji kategoriji na predprocesiranje vezanih nerelevantnih kandidatov, smo podrobneje analizirali. Prva kategorija zajema kolokacijske kandidate s **specifičnimi kolokatorji.** Ti kolokatorji se pojavljajo pri različnih lemah in povejo več o kolokatorjih samih kot o kolokacijah. Metoda je torej primerna za identifikacijo neformalnih pomenov obstoječega besedišča (izluščeni kandidati tako vsebujejo glagol *rabiti* (*rabiti delo/čas*) v pomenu *potrebovati*, prislov *par* v pomenu *nekaj* (*par dni/ur*), značilen kolokator je tudi pridevniško rabljen *super* (*super stran*) ali pa npr. *top* (*top dogodek/video leta*) v pomenu *najvišje uvrščen*. Izluščili smo tudi nekatere okrajšave (uporabljene brez ločila), npr. *roj dan* (*rojstni dan*). Odmik od norme se vidi v rabi besede z bližnjim pomenom (*delavni* vs. *delovni*) v zvezi *delavno mesto.*

Del izluščenih kolokacij je vezan na **spletne vsebine**, ki se kažejo v za korpus Janes značilnih kolokacijah, kot so *link (do) strani*, (*prva/glavna/spletna/desna stran bloga*). Na drugi strani pa je pri izluščenih kolokacijskih kandidatih opaziti žanrsko specifične diskurzivne elemente, saj forume in tvite zaznamuje velika uporaba naslavljanj in pozdravov. Zanimiva kolokacija je npr. *nasmejan dan*, pogosto pa so izluščeni tudi nespecifični kolokatorji, npr. *hvala*, ki se v povezavi z besedo *dan* pogosto pojavlja na koncu sporočil (*Hvala in lep dan*).

Velik del izluščenih kandidatov predstavljajo **tematsko karakterizirani izrazi.** V večini primerov gre za splošne kolokacije, ki pa so bile izluščene zaradi sestave korpusa. Prevladujejo namreč teme s področij avtomobilizma (npr. *voznikova stran*) in zdravja (*ledveni del hrbtenice*), saj so bili forumi teh vsebin ključni vir za sestavo podkorpusa forumov. V večini primerov smo tako izluščili kandidate, ki niso nujno specifični za spletni jezik. Kljub temu lahko najdemo tudi izraze, ki kažejo na neformalno rabo (npr. *gas do konca*).

Kandidati, ki bi bili zanimivi za posebno obravnavo, so **frazeološke enote**, kot so *konec debate*, *ukrasti državo*, *sveta preproščina*, *krasni novi svet*, *prestaviti uro*, *sveta krava* (*»V politiki, če je količkaj verodostojna, javna in ljudska, svetih krav ne bi smelo biti.«*). Metoda je uporabna tudi za luščenje terminov za novejše koncepte (*pametna ura*). Posebej smo označili tudi kolokacije, v katerih je uporabljeno **lastno ime** (*Islamska država*, *Facebook stran*).

Tako specifični kolokatorji kot kolokacije odpirajo vprašanje odnosa med govorjenim jezikom in jezikom uporabniških vsebin. V primeru specifičnih kolokatorjev (*rabiti*, *par*, *super*, *top*) smo preverili in ugotovili, da se ti pojavljajo tudi v korpusu govorjene slovenščine GOS, v SSKJ 2 pa jim je pripisan kvalifikator *pogovorno.* Tovrstni elementi so skupni neformalni komunikaciji, tako v govorjeni slovenščini kot v spletnih vsebinah. Za izbrane zveze *gas do konca*, *konec debate*, *krasni novi svet*, *sveta krava* pa v korpusu GOS nismo našli pojavitev, pri čemer zaradi majhnega

obsega korpusa še ne moremo sklepati, da to niso izrazi govorjenega jezika. (V SSKJ 2 je s kvalifikatorjem *ekspresivno* vključen izraz *konec debate.*)

6 Zaključki in nadaljnje delo

Prispevek obravnava kolokacije v uporabniških spletnih vsebinah. Za luščenje kolokacij smo uporabili funkcijo besednih skic orodja Sketch Engine in primerjalni pristop za luščenje kolokacij, ki korpus uporabniških vsebin primerja z referenčnim korpusom. Prednost metode besednih skic je hiter izvoz iz korpusa in ureditev besed po slovničnih relacijah. Vendar pa je metoda odvisna od kakovosti oblikoskladenjskega označevanja, predvsem pa ne omogoča primerjave med različnimi korpusi (razen v primeru podkorpusov). Primerjalni pristop pa je namenjen prav identifikaciji novih rab besed, vendar je zaradi velike količine izvoženih kolokatorjev bolj nepregleden in časovno zahtevnejši. Več kot 35 odstotkov kandidatov je nerelevantnih zaradi težav predprocesiranja besedil (predvsem lematizacija). V nadaljevanju imamo namen vključiti primerjalno mero *CorpDiff* v metodo luščenja z besednimi skicami. Alternativna rešitev bi bila, da se oba korpusa v Sketch Engine naloži kot podkorpusa in tako uporabi obstoječo funkcijo SketchDiff za primerjavo kolokacij.

Zahvala

Za pomoč pri uporabi orodij se zahvaljujem Kaji Dobrovoljc, Simonu Kreku in Iztoku Kosmu, za vsebinsko pomoč pa Darji Fišer in Poloni Gantar. Raziskava je bila opravljena v okviru projekta Viri, orodja in metode za raziskovanje nestandardne spletne slovenščine (J6-6842, 2014–2017), ki ga financira ARRS.

Literatura

BENSON, Morton, 1989: The structure of collocational dictionary. *The International Journal of Lexicography* 2. 1–14.

BENSON, Morton, BENSON, Evelyn, ILSON, Robert, 1997: *The BBI Dictionary of English Word Combinations.* Revised edition. Benjamins.

COWIE, Anthony P., 1981: The treatment of collocations and idioms in learners' dictionaries. *Applied linguistics* 2. 223–235.

COWIE, Anthony. P., 1994: Phraseology. *Encyclopedia of Language and Linguistics* 6. Oxford, New York. 3168–3171.

FIRTH, John R., 1957: Modes of Meaning. Frank R. Palmer (ur.): *Papers in Linguistics* 1934–1951. London: Oxford University Press. 190–215.

FIŠER, Darja, ERJAVEC, Tomaž, ZWITTER VITEZ, Ana, LJUBEŠIĆ, Nikola, 2014: JANES se predstavi: metode, orodja in viri za nestandardno pisno spletno slovenščino. Tomaž Erjavec, Jerneja Žganec Gros (ur.): *Zbornik 9. konference Jezikovne tehnologije*. 56–61. Ljubljana: Institut »Jožef Stefan«.

GANTAR, Polona, GRABNAR, Katja, KOCJANČIČ, Polona, KREK, Simon, POBIRK, Olga, REJC, Rok, ŠORLI, Mojca, ŠUSTER, Simon, ZARANŠEK, Petra, 2009: Projekt »Sporazumevanje v slovenskem jeziku« ESS in MŠŠ.

GANTAR, Polona, KREK, Simon, 2011: Slovene Lexical Database. Daniela Majchráková, Radovan Garabík (ur.): *Zbornik 6. konference Natural language processing, multilinguality.* Bratislava: Slovenská akadémia vied, Jazykovedný ústav Ludovíta Štúra. 72–80.

GERBER, Laurie, YANG, Jin, 1997: Systran MT dictionary development. *Past, Present, and Future: Machine Translation Summit* 6. 211–218.

HALLIDAY, M. A. K. (1966). Lexis as a Linguistic Level. *In Memory of F. R. Firth*. London: Longman.

KILGARRIFF, Adam, RYCHLY, Pavel, SMRZ, Pavel, TUGWELL, David, 2004: The Sketch Engine. *EURALEX* 2004. 105–116.

KJELLMER, Göran, 1987: Aspects of English Collocations. *Corpus linguistics and Beyond.* Amsterdam, Atlanta: Rodopi.

KOSEM, Iztok, HUSAK, Miloš, MCCARTHY, Diana, 2011: Iztok Kosem, Karmen Kosem (ur.): *Electronic lexicography in the 21st century: new applications for new users*: *eLex* 2011. Ljubljana: Trojina. 150–159.

KOSEM, Iztok, GANTAR, Polona, KREK, Simon, 2013: Avtomatizacija leksikografskih postopkov. *Slovenščina 2.0* 1/2. 139–164.

LIN, Dekang, 1998: Using collocation statistics in information extraction. *Zbornik 7. konference Message Understanding Conference* (MUC-7).

LOGAR, Nataša, GRČAR, Miha, BRAKUS, Marko, ERJAVEC, Tomaž, ARHAR HOLDT, Špela, KREK, Simon, 2012: *Korpusi slovenskega jezika Gigafida, KRES, ccGigafida in ccKRES: gradnja, vsebina, uporaba*. Ljubljana: Trojina, Fakulteta za družbene vede.

LOGAR, Nataša, GANTAR, Polona, KOSEM, Iztok, 2014: Collocations and examples of use: a lexical-semantic approach to terminology. *Slovenščina 2.0.* 41–61.

NESSELHAUF, Nadja, 2005: *Collocations in a Learner Corpus.* Amsterdam: John Benjamins Publishing.

ORENHA-OTTAIANO, Adriane, 2012: English collocations extracted from a corpus of university learners and its contribution to a language teaching pedagogy. *Language and Culture* 34 (2). 241–251.

POLLAK, Senja, ARHAR HOLDT, Špela, 2015: Identifying corpus-specific collocations: The case of spoken Slovene. *NLP, Corpus Linguistics, Lexicography.* Zbornik 8. konference SLOVKO.

RYCHLY, Pavel, 2008: A Lexicographer-Friendly Association Score. Zbornik konference *Recent Advances in Slavonic Natural Language Processing*. Brno: Masaryk University. 6–9.

SINCLAIR, John, 1966: Beginning the Study of Lexis. *In Memory of F. R. Firth*. London: Longman.

SINCLAIR, John, 1991: *Corpus Concordance Collocation*. Oxford University Press.

SSKJ 2: *Slovar slovenskega knjižnega jezika*, druga, dopolnjena in deloma prenovljena izdaja,.

VERDONIK, Darinka, KOSEM, Iztok, ZWITTER VITEZ, Ana, KREK, Simon, STABEJ, Marko, 2013: Compilation, transcription and usage of a reference speech corpus: the case of the Slovene corpus GOS. *Language resources and evaluation* 47/4. 1031–1048.

VINTAR, Špela, 2010: Bilingual term recognition revisited: the bag-of-equivalents term alignment approach and its evaluation. *Terminology* 16/2. 141–158

YU, Ning, 2014: Sentiment analysis in UGC. Marie-Francine Moens, Juanzi Li, Tat-Seng Chua (ur.): *Mining User Generated Content*. CRC Press. Taylor and Francis book. 43–66.