Polona Gantar
Univerza v Ljubljani
Filozofska fakulteta in
Fakulteta za računalništvo in informatiko

Špela Arhar Holdt
Univerza v Ljubljani
Filozofska fakulteta in
Fakulteta za računalništvo in informatiko

Iztok Kosem
Univerza v Ljubljani
Filozofska fakulteta in
Fakulteta za računalništvo in informatiko in
Institut »Jožef Stefan«
Laboratorij za umetno inteligenco

Simon Krek
Institut »Jožef Stefan«
Laboratorij za umetno inteligenco in
Univerza v Ljubljani
Fakulteta za računalništvo in informatiko in
Filozofska fakulteta

DOI: 10.4312/jis.68.4.157-175
1.01

SOPOMENKE 2.0 IN KOLOKACIJE 2.0: NOVI KORAKI ZA SLOVENSKE ODZIVNE SLOVARJE

V prispevku opišemo nadgradnjo dveh slovarjev, *Slovarja sopomenk sodobne slovenščine* in *Kolokacijskega slovarja sodobne slovenščine*, ki sta s svojo prvo izdajo v leksikografski prostor vnesla koncept odzivnega slovarja. Za oba slovarja velja, da sta digitalno zasnovana, v izhodišču strojno pripravljena in postopno izboljšana z uvajanjem novih metodoloških postopkov. Ključna lastnost obeh slovarjev je vključevanje uporabnikov v leksikografski proces z možnostjo dodajanja lastnih predlogov in vrednotenja slovarskih podatkov.

V prispevku opišemo nadgrajeni, drugi različici slovarjev, in sicer leksikografsko obdelavo podatkov, zlasti dodajanje pomenskih informacij in informacij o jezikovni rabi pri vrednotenjsko občutljivem besedišču; vključevanje novih vsebin, kot so protipomenke in dodajanje slovničnih informacij pri kolokacijah; implementiranje izsledkov uporabniških raziskav tako pri vsebini slovarskih podatkov kot pri nadgradnji slovarskih vmesnikov; prednosti novega metodološkega postopka pri luščenju podatkov iz korpusa in način vključevanja podatkov v Digitalno slovarsko bazo.

Slovarja predstavljata primer dobre prakse tudi za druge jezikovne skupnosti, saj temeljita na čim večji izrabi jezikovnotehnoloških postopkov pri izdelavi slovarja, hkrati pa uspešno združujeta leksikografski proces in številne možnosti uporabniškega sodelovanja.

Ključne besede: *Slovar sopomenk sodobne slovenščine*, *Kolokacijski slovar sodobne slovenščine*, odzivni slovar, strojno pridobivanje leksikografskih podatkov, uporabniško vključevanje

Thesaurus 2.0 and Collocations 2.0: New steps for Slovene responsive dictionaries

In this paper, we describe the upgrading of two dictionaries, the *Thesaurus of Modern Slovene* and the *Collocation Dictionary of Modern Slovene*, which introduced the concept of a responsive dictionary into the lexicography with their first edition in 2018. Both dictionaries are considered to be digitally born, automatically created at the outset and gradually improved by the introduction of new methodological procedures. A key feature of both dictionaries is the involvement of users in the lexicographic process, with the possibility of adding their suggestions and evaluating the dictionary data.

In this paper, we describe the upgraded, second versions of the dictionaries, namely the lexicographic data processing, and in particular the addition of semantic and linguistic usage information for negative vocabulary; the inclusion of new content, such as antonyms and grammatical information for collocations; the implementation of user research findings both in the content of the dictionary data and in the upgrading of the dictionary interfaces; the advantages of the new methodological procedure for extracting data from the corpus; and the integration of the data into the Digital Dictionary Database.

The two dictionaries also represent an example of good practice for other linguistic communities, as they are based on maximising the use of language technologies in dictionary production while successfully combining the lexicographic process and the many possibilities for user participation.

Keywords: *Thesaurus of Modern Slovene*, *Collocations Dictionary of Modern Slovene*, responsive dictionary, machine extraction of lexicographic data, user involvement

1 Uvod

V prispevku predstavimo nadgrajeni različici *Slovarja sopomenk sodobne slovenščine*[1] in *Kolokacijskega slovarja sodobne slovenščine*,[2] ki smo ju pripravili v okviru projekta »SoKol« (Sopomenke in Kolokacije 2.0). Nadgradnja obeh virov, ki sta bila v svoji prvi različici objavljena leta 2018, je rezultat novih metodoloških postopkov pri strojnem pridobivanju podatkov, jezikoslovnih analiz in ocen ter usmerjenih uporabniških raziskav. Z vključitvijo obeh slovarskih virov v Digitalno slovarsko bazo smo se spoprijeli tudi z izzivi organizacije podatkov v digitalni bazi, njihovo povezljivostjo, shranjevanjem in načini posodabljanja.

Vsebinska in metodološka nadgradnja je bila tudi priložnost za izboljšavo slovarskih vmesnikov. Najprej v poenotenju celostne vizualne podobe virov Centra za jezikovne vire in tehnologije (CJVT), znotraj tega pa še glede na izsledke uporabniških študij, izdelanih za vrednotenje slovarske vsebine in prikaza podatkov v slovarskem vmesniku.

[1] Povezava: https://viri.cjvt.si/sopomenke/slv/ (dostop 30. 11. 2023).

[2] Povezava: https://viri.cjvt.si/kolokacije/slv/ (dostop 30. 11. 2023).

V nadaljevanju prispevka najprej predstavimo koncept, razvoj in nadgradnjo *Slovarja sopomenk sodobne slovenščine*, nato pa se osredotočimo na izboljšave, ki vključujejo pripravo sopomenske podatkovne baze, izdelavo smernic za pomensko členjenje in razvrščanje sopomenk pod pomene, izdelavo smernic za prepoznavanje negativno zaznamovanega besedišča in označitev s slovarskimi oznakami, pripravo protokolov za leksikografsko obravnavo uporabniško dodanih sopomenk, vključitev protipomenk v slovarski prikaz in nove možnosti uporabniškega sodelovanja pri širjenju vira in izboljšavi njegove kakovosti.

Nato podrobno predstavimo izhodišča za nadgradnjo *Kolokacijskega slovarja sodobne slovenščine*, zlasti izzive pri vključevanju kolokacij v digitalne slovarske vire, rezultate uporabniških raziskav, nov postopek avtomatskega luščenja podatkov in določanje mehanizmov za prikaz relevantnega izbora kolokacij v slovarju. Na koncu predstavimo še izboljšave kolokacijskega vmesnika, zlasti funkcijo kolokacijskih primerjav pri sopomenkah, ki podatkovno povezuje oba nadgrajena slovarja.

Prispevek zaključimo z izzivi in načrti, ki jih imamo za prihodnje nadgradnje obeh slovarskih virov.

2 *Slovar sopomenk sodobne slovenščine 2.0*

Slovar sopomenk sodobne slovenščine 1.0 (SSSS 1.0; Arhar Holdt idr. 2018) je v slovenski pa tudi mednarodni slovarski prostor uvedel koncept odzivnega slovarja, ki je digitalno zasnovan in v prvi fazi izključno strojno pripravljen jezikovni vir, ki omogoča hiter dostop do odprtih podatkov o sodobni jezikovni rabi. Slovar postopoma izboljšujemo na podlagi leksikografskega urejanja podatkov in z vključevanjem uporabnikov tako v proces širjenja slovarja kot tudi v ocenjevanje strojno in uporabniško pripravljenih podatkov (Arhar Holdt idr. 2018: 404).

V prvi različici je SSSS vseboval 105.473 iztočnic in 368.117 sopomenk. Ustvarjen je bil strojno na podlagi že obstoječih virov: Oxford®-DZS *Veliki angleško-slovenski slovar* (Krek idr. 2017) in referenčnega korpusa pisne slovenščine *Gigafida* (Logar Berginc idr. 2012). Podatki, objavljeni v SSSS 1.0, niso bili leksikografsko obdelani. Iztočnice in sopomenski kandidati so bili predstavljeni v osnovni obliki brez besednovrstnih oznak ali drugih metapodatkov, ki bi pomagali razločevati med oblikami. Pomenske opise so nadomeščale strojno generirane pomenske gruče, podatki pa so bili brez slovarskih oznak, razen iz Oxfordovega slovarja podedovanega nabora področnih oznak, kot so npr. biologija, kemija, pravo, šport ipd.

2.1 Čiščenje podatkov in uvoz v Digitalno slovarsko bazo

Prvi korak pri pripravi druge različice SSSS se je nanašal na uvoz podatkov iz baze slovarja sopomenk v Digitalno slovarsko bazo (DSB; Kosem idr. 2021b).

Ta postopek je omogočil povezovanje podatkov z drugimi slovarskimi bazami, olajšal urejanje podatkov in optimiziral njihovo ponovno uporabo. Pred uvozom v DSB smo iz sopomenske baze izluščili sopomenske pare, ki so vsebovali področne oznake, ter jih pregledali, popravili ali prilagodili označevalnemu sistemu v digitalni bazi. Nato smo na podlagi predhodnih množičenjskih kampanj, namenjenih odstranjevanju šuma (Čibej in Arhar Holdt 2019), izločili 8.878 problematičnih iztočnic, kot so npr. redundantne večbesedne enote, pri katerih gre za ponovitev vsebovanega dela besedne zveze (npr. *število spermijev – število spermijev v ejakulatu*). Nadalje smo iztočnice, ki v prvi različici slovarja niso vsebovale oznake za besedno vrsto, zaradi česar so bile sopomenke enakopisnih iztočnic prikazane združeno, razdvoumili in sopomenke ustrezno razvrstili (skupno 4.560 enot), npr. *svet* (sam.): *področje*, *ozemlje*, *sfera* itd. vs. *svet* (prid.): *sakrosankten*, *nedotakljiv*, *posvečen*, *sakralen* ipd.

2.2 Pomenska analiza in razčlenitev pomenov

Za pomensko analizo smo izbrali 2.000 iztočnic na podlagi presečne množice treh virov: *Slovarja sopomenk sodobne slovenščine* 1.0 (Krek idr. 2018), *Kolokacijskega slovarja sodobne slovenščine* 1.0 (Kosem idr. 2019) in *Velikega slovensko-madžarskega slovarja* (Kosem idr. 2021a). Pri selekciji smo upoštevali besednovrstno zastopanost (samostalniki, glagoli, pridevniki in prislovi), eno- oziroma večpomenskost, slovarske oznake in potencialno sovražno besedišče. Pomene smo opremili s pomenskimi indikatorji (Gantar 2015: 164), primarno na podlagi *Velikega slovensko-madžarskega slovarja* oziroma s pomočjo na novo izdelanih pomenskih analiz. Pod posamezne pomene pomensko razčlenjenih iztočnic smo nato razvrstili sopomenke. Delo je zaradi lažje kategorizacije, preglednosti in obvladovanja leksikografskega procesa potekalo v tabelarični obliki (Google preglednice), kamor smo podatke najprej uvozili, potem pa pregledali, dopolnili in uredili, nato pa uvozili nazaj v podatkovno bazo. V procesu leksikografske analize smo oblikovali tudi smernice, ki opredeljujejo postopek razvrščanja sopomenk pod pomene, načine preverjanja sopomenskosti in notranjih sopomenskih povezav.

Da bi zagotovili natančnost pomenske klasifikacije, smo preverjali rabo v referenčnem korpusu pisne slovenščine *Gigafida* 2.0 (Krek idr. 2020), spremljevalnem korpusu slovenskega jezika *Trendi* (Kosem idr. 2022), korpusu spletne slovenščine *JANES* 1.0 (Fišer idr. 2020) in korpusu akademske slovenščine *KAS* 2.0 (Žagar idr. 2021). V težavnejših primerih smo preverjali pomensko vrednost sopomenk z vidika tvorjenja prekrivnih kolokacij, kar nam je pomagalo pri pomenskem razvrščanju. Za to nalogo smo uporabili *Kolokacijski slovar sodobne slovenščine* 1.0, *Slovar sopomenk sodobne slovenščine* 1.0, korpus *Gigafida* 2.0 in funkcijo Sketch Diff orodja Sketch Engine (Kilgarriff idr. 2014). Primer gesla z razvrščenimi sopomenkami je predstavljen v tabeli 1.

Iztočnica	Pomen	Sopomenke
sova (samostalnik)	[1: ptica]	skovik
	[2: kdor je pozno pokonci]	ponočnjak, nočni ptič
	[3: zoprna ženska; izraža negativen odnos]	babura, babnica

Tabela 1: Razvrstitev sopomenskih kandidatov glede na pomensko razčlenjeno iztočnico *sova*

Če sopomenke iz prve različice SSSS korpusna raba ni potrdila, je pri pomensko členjenih geslih nismo upoštevali. V procesu pomenske analize smo naleteli na številne mejne primere, kjer smo se raje odločali za vključevanje kot izključevanje iz dveh razlogov. Prvi je načrtovanje čiščenja in strukturiranja podatkov v naslednjih različicah slovarja, drugi pa dejstvo, da želimo s konceptom odzivnega slovarja zagotoviti uporabnikom čim več podatkov, kar vključuje tudi široko razumevanje pojma sopomenskosti, in jim tako omogočiti čim večjo izbiro sopomenskih kandidatov znotraj konkretne uporabniške izkušnje. Primeri mestoma vključujejo tako pomensko bližino, npr. *lasulja – perika*, *lasni vložek*, *tuf* kot tudi razmerja nad- oziroma podpomenskosti, npr. *čoln – plovilo*, *dingi*, *barka*; *jazbina – brlog* ipd.

2.3 Slovarske oznake

V prvi različici SSSS iztočnice in njihove sopomenke niso vsebovale eksplicitnih informacij o rabi ter slogovni in pragmatični vrednosti, razen omejenega nabora področnih oznak. V zvezi s tem je raziskava, v kateri je sodelovalo 671 uporabnikov, pokazala, da jih več kot tretjina (37 %) odsotnost slovarskih oznak ocenjuje kot problematično (Arhar Holdt 2020: 472). Prav zaradi tega je bila vključitev označevalnega sistema pri tem tipu besedišča ena od prednostnih nalog slovarske nadgradnje, logični nasledek pa tudi posodobitev uporabniškega vmesnika, ki bi uporabnikom omogočil sistematično označevanje predlaganih sopomenk.

Označevalni sistem za besedišče z izrazito negativnim vrednotenjem smo zasnovali na podlagi leksikografskih smernic, ki jih uporabljamo pri izdelavi slovarskih virov na CJVT UL, npr. pri izdelavi *Velikega slovensko-madžarskega slovarja* (Kosem idr. 2021a). Za označevanje smo predvideli tri oznake, in sicer za prepoznavanje elementov sovražnega govora (oznaka *sovražno*), elementov nevljudnosti ali žaljivosti (*grobo*) ter elementov negativnega vrednotenja ali konotacije (*izraža negativen odnos*). Vsaka oznaka je v vmesniku prikazana z ikono, ki ji je dodana razlaga možnega vpliva uporabe označene besede:

> *sovražno*: Z uporabo besede lahko izražamo sovražen, nestrpen odnos do posameznika ali družbene skupine.
> *grobo*: Zaradi družbenih in moralnih norm se marsikateremu uporabniku jezika beseda lahko zdi groba ali neprimerna. Uporaba lahko povzroči nelagodje, razburi ali užali.

izraža negativen odnos: Beseda lahko ni nevtralna.[3] Z uporabo besede se lahko posmehujemo, izražamo neodobravanje ali kritiko do nekaterih lastnosti posameznikov, predmetov ali dejanj.

Oznake smo iztočnicam in njihovim sopomenkam pripisovali ročno v procesu leksikografske analize in pomenske členitve. Za podrobnejši opis gl. Arhar Holdt idr. 2023a.

2.4 Samodejno luščenje in izbor protipomenk

Ena od izboljšav SSSS 2.0 je tudi izdelava metodologije za avtomatsko pridobivanje protipomenk, ki temelji na strojnem učenju s pomočjo besednih vložitev in vnaprej pripravljenih velikih jezikovnih modelov, ter vključitev prikaza protipomenk v slovarski vmesnik. Metodologijo, ki je podrobneje opisana v Arhar Holdt idr. (2023b), bomo v naslednjih iteracijah skušali še izboljšati, trenutno pa je bilo na tej podlagi v slovar vključenih 2.544 protipomenskih parov različnih besednih vrst, ki so bili strojno izluščeni in ročno pregledani, npr. *aktivnost – neaktivnost*, *nedejavnost*, *inercija*, *mirovanje* (*botanika*, *zoologija*); *čist – nečist*, *umazan*, *zamazan* ali *malokdaj – pogosto*, *redno*, *velikokrat*, *često*, *navadno*, *običajno*. Od različice 2.0 naprej lahko uporabniki v slovar dodajajo tako sopomenke kot protipomenke, za protipomenske podatke pa so po vzoru sopomenskih na voljo tudi vse druge informacije in povezave, npr. primerjava rabe prek kolokacij in povezava z referenčnim korpusom. Kot velja za sopomenskost, bo pri slovaropisnem presojanju tudi protipomenskost dopuščena široko. Prve analize in priporočila za pripravo smernic so bila pripravljena v Mozetič idr. (2022), končna metodologija pa bo izbrana po prvi celostni analizi uporabniško dodanega gradiva.

2.5 Primerjava sopomenk prek kolokacij

Nov način luščenja kolokacij (gl. 3.2) je pokazal tudi potrebo po vključevanju dodatnih struktur v kolokacijsko primerjavo. Poleg tega so bila na podlagi raziskave o relevantnosti kolokacij v izbranih skladenjskih strukturah v prvi različici slovarja (Arhar Holdt 2021b) izdelana priporočila glede vključevanja oziroma izključevanja kolokacijskih struktur iz primerjalne funkcije. Najpomembnejši spremembi se nanašata na (a) samostalniške strukture, kjer smo strukturo samostalnik + predlog + samostalnik, npr. *misel izpred let*, nadomestili s produktivnejšo strukturo samostalnik + samostalnik v rodilniku, npr. *razpredanje misli*, in (b) pridevniške

[3] Celoten označevalni sistem, ki ga oblikujemo pri izdelavi slovarskih virov CJVT, predvideva tudi oznaki »lahko izraža negativen odnos«, s katero želimo opozoriti na primere, kjer se v določenih kontekstih sicer nezaznamovana ali pozitivno konotirana beseda lahko rabi nevtralno ali celo negativno (npr. *škrat*, *čarovnica*) in oznako »lahko izraža pozitiven odnos« za besede, ki so v izhodišču negativno zaznamovane in imajo v nekaterih kontekstih tudi pozitivno konotacijo (npr. *mrha*, *blazen*). Treba je še omeniti, da oznake pripisujemo samo pomenom, ki kažejo prepoznavno splošno in ne individualno rabo.

strukture, kjer smo nadomestili strukturo pridevnik + predlog + samostalnik, npr. *drag za davkoplačevalce*, s priredno strukturo pridevnik + in + pridevnik, npr. *drag in potraten*.

Pri pridobivanju primerjalnih kolokacij smo pri izdelavi druge različice osredotočili na njihovo tipičnost glede na (a) pojavljanje z iztočnico in (b) glede na pojavljanja s sopomenko. Na ta način smo iz primerjave izločili manj tipične kolokacije, ki so v prvi različici med jedrni prikaz uvrščale manj relevantne in pogosto tudi manj pogoste kolokacijske primerjave.

2.6 Uredniški protokol pri vrednotenju uporabniško dodanih sopomenk

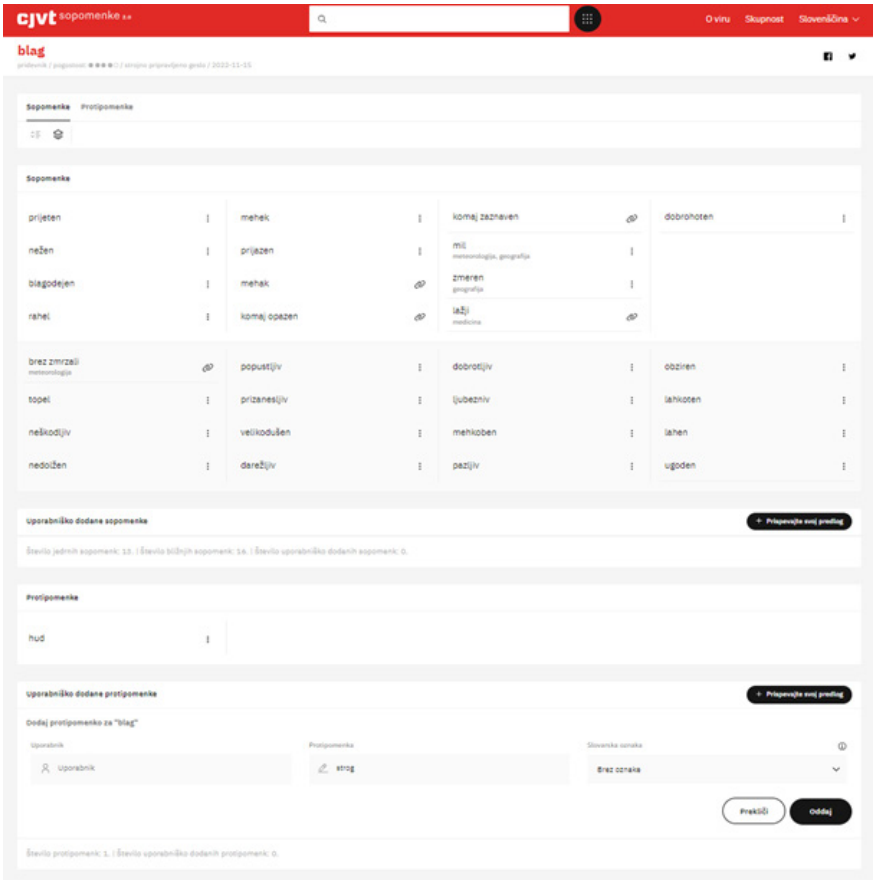
Z namenom ohraniti konsistentnost podatkov in kakovost podatkovne zbirke, hkrati pa v skladu s široko konceptualno zasnovo slovarja upoštevati predloge uporabnikov v čim večji meri, smo v procesu nadgradnje zasnovali smernice za vrednotenje uporabniško dodanih sopomenk. V ta namen je 6 leksikografov analiziralo 972 uporabniško dodanih sopomenk ter jih kategoriziralo kot »primerne«, »neprimerne« in »pogojno primerne« za vključitev v slovarsko bazo. Analiza je bila dopolnjena z uporabniško študijo, ki je zajela širši nabor uporabniških skupin (učitelje, prevajalce in lektorje) in dala vpogled v preference uporabnikov pri prepoznavanju sopomenskosti glede na leksikografe in v razlike pri izboru kriterijev za prepoznavanje sopomenskosti med posameznimi uporabniškimi skupinami (Gapsa in Arhar Holdt 2023).

Ključni parametri protokola za vrednotenje uporabniško dodanih sopomenk vključujejo preverjanje (a) obstoja sopomenskega predloga v avtentični jezikovni rabi na podlagi korpusov, (b) uvrstitve sopomenskega predloga pod ustrezno iztočnico (glede na pomen in slovnične kategorije) ter (c) ustreznosti potencialno pripisanih slovarskih oznak. Ob tem upoštevamo še povratne informacije, ki jih uporabniki slovarja lahko dodajajo z glasovanjem o ustreznosti ali neustreznosti predlaganih sopomenk. Uporabniško dodane sopomenke bodo na podlagi izdelanih smernic obravnavane v naslednji različici slovarja, v vmesnem času pa bomo analizirali uporabniško pripisane slovarske oznake in ugotovili, katere oznake uporabniki še pogrešajo in ali konsistentno uporabljajo predlagane.

2.7 Nadgradnja slovarskega vmesnika

Poleg vsebinskih izboljšav smo pri pripravi druge različice veliko pozornosti namenili nadgradnji slovarskega vmesnika. V sodelovanju z oblikovalsko ekipo smo zasnovali elemente vmesnika, ki omogočajo skladno vizualno podobo vseh elementov v slovarskih virih CJVT, ter logično strukturo, vključno z barvami, ikonami, tipografijo in oblikovanjem elementov, kot so iskanje, preklapljanje, deljenje in vključevanje uporabnikov. Na novo smo zasnovali tudi glavo in nogo slovarja,

razdelek *O slovarju* ter funkcije, ki jih v različici 1.0 ni bilo, vključno s protipomenkami, razvrščenimi po pomenih, ikonami za negativno zaznamovano besedišče ter uporabniškimi mehanizmi za dodajanje oznak in razvrščanjem predlaganih sopomenk in protipomenk pod ustrezni pomen.



Slika 1: Strojno pripravljeno geslo v prenovljenem vmesniku SSSS 2.0

SSSS 2.0 ponuja dve različni postavitvi slovarskega gesla: strojno pripravljeno geslo (slika 1) in pomensko členjeno geslo z ročno razporejenimi sopomenkami. Obe postavitvi vključujeta metapodatke, kot so besednovrstna opredelitev, kazalnik pogostosti iztočnice v referenčnem korpusu *Gigafida* 2.0 in podatek o metodologiji priprave gesla (strojno, ročno) ter zavihek za protipomenke.

Postavitev strojno pripravljenega gesla prikazuje sopomenske kandidate v dveh delih: jedrne sopomenke na belem ozadju, bližnje sopomenke z ohlapnejšo pomensko relacijo pa na sivem. Postavitev vključuje tudi razdelek z uporabniško dodanimi sopomenkami in z možnostjo dodajanja novih predlogov.

Postavitev pomensko členjenega gesla vsebuje poleg metapodatkov vrstico s pomenskim menijem, ki deluje kot filter za posamezni pomen, številčno razvrstitev pomenov in pomenski opis v obliki pomenskega indikatorja. Sopomenke so, vključno z vrednotenjsko oznako pri pomenu ali ikono pri sopomenki, razporejene pod posamezne pomene na podlagi leksikografske analize (gl. 2.2). S klikom na ikono se odpre krajše pojasnilo o vrednotenjskem potencialu, ki ga ima označena beseda v rabi (gl. 2.3).

Pri pomensko členjenih geslih lahko uporabniki dodajajo predloge sopomenk pri posameznih pomenih, pri čemer je privzeta možnost *brez oznake*, druge oznake, tj. *sovražno*, *grobo* in *izraža negativen odnos*, pa je mogoče izbrati v spustnem meniju. V SSSS 2.0 je na voljo tudi vnosno polje, v katerega lahko uporabniki vpišejo katero drugo oznako po lastni presoji (tudi denimo za pozitivno zaznamovanost, ki se ji na projektu nismo posvečali). Predlogi uporabnikov bodo dragoceni ne le za dopolnjevanje prosto dostopne slovarske zbirke, temveč tudi za analize dojemanja sistema označevanja (npr. v primerjavi z leksikografskim) in njegovo morebitno izboljšanje.

V nadgrajenem slovarskem vmesniku SSSS 2.0 je, kot omenjeno v razdelku 2.5., dodana možnost primerjave prekrivnosti kolokatorjev pri posamezni sopomenki, kjer so najprej prikazane kolokacije, ki se pojavljajo z obema sopomenkama, nato pa kolokacije, ki se običajno pojavljajo le z eno od sopomenk. Na primer za sopomenski par *sodoben – današnji* sta tipični kolokaciji *sodoben čas* in *današnji čas*, medtem ko pridevnik *sodoben* običajneje kolocira z *umetnost*, *tehnologija*, *družba*; pridevnik *današnji* pa z *dan*, *seja*, *tekma*.

3 *Kolokacijski slovar sodobne slovenščine 2.0*

Pomembnost kolokacij je v slovarskem prostoru že dolgo prisotna, njihovo vključevanje v slovarske vire pa je deležno številnih analiz, še posebej od pojavnosti korpusov naprej. Za razvoj in nadgradnjo prve različice *Kolokacijskega slovarja sodobne slovenščine* (KSSS 1.0) smo na podlagi izkušenj vrste sorodnih projektov (prim. Kallas idr. 2015; Colman in Tiberius 2018; Hudeček in Mihaljević 2020) in na podlagi jezikoslovnih in uporabniških raziskav izpostavili nekaj ključnih izhodišč, ki naj bi jih obsegala nadgrajena različica slovarja.

Eno od pomembnejših vprašanj je količina dobrih kot tudi slabih strojno pridobljenih korpusnih podatkov. Med pregledovanjem obsežnih seznamov kolokacijskih kandidatov morajo namreč leksikografi prepoznati dobre kolokacije, zavreči slabe in pogosto opraviti ožji izbor tudi med dobrimi. To še zdaleč ni enostavno; medtem ko nekatere slabe kolokacijske kandidate lahko prepoznamo takoj, lahko druge potrdimo kot take šele z analizo korpusnih primerov. Ključno pri tem je, da so merila kolokacijskosti jasna in da so razmerja do drugih večbesednih enot opredeljena. Izhodišča za določanje kolokacijskosti, ki smo jih upoštevali pri izdelavi KSSJ 2.0, so opisana v Gantar idr. 2021.

Z obilico korpusnih podatkov je povezano tudi vprašanje žanrske specifičnosti korpusnih besedil in kakovosti korpusne označenosti. Izvor slabih kolokacijskih kandidatov lahko pogosto pripišemo problematični vsebini korpusa, npr. strojno prevedenim besedilom s spleta (Koppel idr. 2019) ali napakam pri lematizaciji, oblikoskladenjskem označevanju ali skladenjskem razčlenjevanju (Koppel idr. 2019; Pori in Kosem 2021).

Drug izziv je uporaba podatkovnega modela, ki definira način shranjevanja in vrsto v bazi shranjenih podatkov, tj. katere leksikografske odločitve shraniti (samo dobre ali tudi slabe kandidate), kako spremljati najnovejše spremembe v jeziku in kako jih vključevati v obstoječe podatke.

3.1 Uporabniške raziskave

Na razvoj druge različice KSSS je najbolj vplivala raziskava, v kateri smo preučevali odnos različnih skupin uporabnikov: učiteljev slovenščine kot L1, učiteljev slovenščine kot L2, lektorjev in prevajalcev ter leksikografov, do KSSS in način njegove uporabe (Pori idr. 2020; Pori idr. 2021). V evalvacijskem intervjuju, ki je temeljil na metodi vodenega glasnega razmišljanja, smo 40 uporabnikov prosili, naj izvajajo naključna iskanja po lastni izbiri, izvajajo vnaprej določena iskanja ter komentirajo splošno uporabnost slovarja in podobo njegovega vmesnika. Najpomembnejše ugotovitve, ki jih lahko povzamemo iz raziskave, so, da je odnos uporabnikov do vključevanja strojno pridobljenih kolokacij večinoma pozitiven, da pa mora biti opozorilo o naravi teh podatkov in stopnji njihove izdelave jasno izpostavljeno. Uporabniki so slovarski vmesnik ocenili kot dober, vse funkcije so prepoznali kot uporabne in enostavne za uporabo. Možnost dostopa do vseh kolokacij znotraj posamezne strukture prek klika se je uporabnikom zdela koristna, so se pa pojavili dvomi, ali večina uporabnikov sploh pride do dodatne vsebine, kot so npr. korpusni zgledi, pri čemer so uporabniki povezave do korpusa ocenili kot zelo pomembne, celo ključne. Nekateri uporabniki so izrazili potrebo po dodatnih informacijah o kolokacijah, na primer o pogostosti ali statistični jakosti. Zlasti lektorji in prevajalci so menili, da je del o množičnem zbiranju podatkov koristen, čeprav običajno nimajo časa prispevati. Po drugi strani so učitelji izrazili pomisleke glede uporabnosti funkcije, če bi jo uporabljali manj izkušeni uporabniki jezika.

Druga pomembna raziskava, ki smo jo upoštevali, je preučila želje in pričakovanja 415 uporabnikov KSSS glede vrstnega reda kolokacij v slovarskem vmesniku (Arhar Holdt 2021a). Ugotovitve so pokazale, da se pričakovanja uporabnikov glede skladenjskih struktur bolj ali manj ujemajo s predlaganim vrstnim redom v slovarju. Po drugi strani pa so pričakovali razvrstitev kolokacij po pogostosti in ne po statistični jakosti, kot so bile kolokacije razvrščene v različici KSSS. Zanimivo je, da tudi pri tujih slovarjih ta pristop ni enoten: estonski

kolokacijski slovar[4] ureja kolokacije znotraj slovarskih gesel po pogostosti, nizozemski *Woordcombinaties* (Colman in Tiberius 2018), *Mrežnik* (Hudeček in Mihaljević 2020) in *Macmillan Collocations Dictionary* (Rundell 2010) pa po abecednem redu.

Z vidika množičnega zbiranja podatkov za KSSS je pomembna tudi raziskava, v kateri je 6 jezikoslovcev na podlagi izluščenih kolokacij in naključno izbranih korpusnih zgledov glasovalo o njeni ustreznosti (Pori in Kosem 2021). Možni odgovori na vprašanje, ali je kandidat kolokacija, so bili: »da«, »ne« in »ne vem«. Čeprav je bil glavni cilj oceniti zanesljivost metode strojnega pridobivanja kolokacij, je raziskava pokazala, da je za odločanje o njeni primernosti ključna jasna opredelitev, kaj (ustrezna) kolokacija dejansko je. Poleg tega so uporabniki v pilotni študiji poudarili, da so nekatere pomanjkljivosti, kot je denimo uvrstitev kolokacije pod napačno strukturo razvidne le zahtevnejšim uporabnikom, ki razumejo skladenjske lastnosti kolokacij. Druga pomembna ugotovitev je, da je za ugotovitev ustreznosti kolokacije navadno potrebno pogledati več kot le en primer.

Dragocene izkušnje glede uporabniškega dodajanja kolokacij smo pridobili pri razvoju *Igre besed* (Arhar Holdt idr. 2021). V zvezi s tem smo izvedli eksperiment, v katerem smo skupino študentov prosili, naj zglede, ki vsebujejo kolokacije, pripišejo ustreznim pomenom izbranih iztočnic. Ugotovitve so pokazale, da je taka naloga izjemno zanesljiva (v več kot 80 % primerov je bilo soglasje označevalcev 100-odstotno) in primerna za različne namene: ugotavljanje razumljivosti pomenske delitve in pomenskih indikatorjev, ugotavljanje kakovosti/primernosti korpusnih zgledov, posredno pa tudi potrjevanje ustreznosti kolokacij.

Ugotovitve omenjenih raziskav smo uporabili kot izhodišče za načrtovanje druge različice KSSS.

3.2 Strojno pridobivanje podatkov – nova metodologija

Ena od pomembnih metodoloških izboljšav v primerjavi s prvo različico KSSS je postopek pridobivanja kolokacij in zgledov iz korpusa. V nasprotju z luščenjem na podlagi zgolj oblikoskladenjskih podatkov, so bile kolokacije za drugo različico izluščene povsem na novo na podlagi skladenjsko razčlenjenih korpusnih podatkov (Krek idr. 2022; Krek idr. 2021). Novi formalizem opredeljuje odvisnostno skladenjsko razmerje znotraj kolokacije ter omogoča omejitve na kateri koli ravni označenega korpusa, od morfologije do skladenjskih razmerij in drugih označenih korpusnih ravni (Krek idr. 2022: 241). Označene podatke na posamezni korpusni ravni smo uporabili tudi za določanje realizacijske oblike kolokacijskih komponent, kar je pomembno tako za shranjevanje kolokacij v podatkovni zbirki in za njihovo predstavitev uporabnikom.

[4] *Estonian Collocatin Dictionary* je del spletnega portala Sõnaveeb Inštituta za estonski jezik, povezava: https://sonaveeb.ee/?lang=en (dostop 30. 11. 2023).

V avtomatsko luščenje smo za pripravo druge različice slovarja vključili vseh 82 struktur, ker se je izkazalo, da so bile nekatere pomensko relevantne pri posameznih iztočnicah v prvi različici izključene, npr. struktura: samostalnik + glagol (*veter* + *nagaja*, *odnese*, *odpihne*; *izvedenec*, *društvo*, *Cerkev* + *se angažira*). Hkrati smo omejili število kolokacij na strukturo, da bi se izognili obremenitvi uporabnikov s preveliko obilico podatkov. Privzeto največje število prikazanih kolokacij za določeno skladenjsko strukturo v različici 2.0 je 10, hkrati pa za strukture, ki so se v raziskavah izkazale kot bolj produktivne (npr. glagol + samostalnik v tožilniku, pridevnik + samostalnik, samostalnik + samostalnik v rodilniku), prikažemo do 25 kolokacij.

Poleg tega so bila na podlagi raziskave o relevantnosti kolokacij v izbranih skladenjskih strukturah v prvi različici slovarja (Arhar Holdt 2021a) izdelana priporočila glede vključevanja oziroma izključevanja kolokacijskih struktur iz primerjalne funkcije. Kar se tiče izbora iztočnic, so bile kolokacije izluščene za vse samostalnike (brez lastnih imen), pridevnike, prislove in glagole, ki so vključeni v Digitalno slovarsko bazo. Edini dodatni uporabljeni parameter je bila najmanjša pogostnost 4 za kolokacije. Od 138.032 kandidatnih iztočnic jih je 81.445 izpolnjevalo ta pogoj; večina iztočnic je enobesednih, le 128 jih je besednozveznih. Pri strojnem luščenju smo upoštevali zgoraj navedene omejitve na skladenjsko strukturo, razen za 1.608 gesel, ki so bila izbrana za popolno ročno preverjanje (gl. 3.3).

Nov pristop smo uporabili tudi pri avtomatskem luščenju korpusnih zgledov. V različici 1.0 smo konfiguracije funkcije GDEX optimizirali za ekstrakcijo zgledov na ravni kolokacije. Čeprav je ta pristop dal dobre rezultate, je bila GDEX klasifikacija korpusnega stavka odvisna od iztočnice in ne od stavka kot celote. Zato smo se odločili, da bomo izdelali eno konfiguracijo GDEX za celoten korpus *Gigafida* 2.0 in vsakemu stavku v korpusu dodelili oceno GDEX. Na ta način smo avtomatsko pridobili tudi nabor identifikatorjev korpusnih stavkov, v katerih se pojavi posamezna kolokacija. S tem pristopom smo za KSSS 2.0 pridobili do štiri korpusne zglede z najvišjo GDEX oceno za vsako kolokacijo.

3.3 Shranjevanje kolokacijskih podatkov: Digitalna slovarska baza in podatkovno skladišče

Kolokacije so skupaj z drugimi vrstami leksikalnih podatkov shranjene v Digitalni slovarski bazi (Kosem idr. 2021b), ki je zasnovana kot enotna podatkovna baza za slovenščino in je namenjena sestavljanju jezikovnih virov in strojni obdelavi naravnega jezika. Zaradi nadgrajene metodologije strojnega pridobivanja smo morali novo izluščene kolokacije ponovno uvoziti v Digitalno slovarsko bazo in ustrezno nadomestiti prvotno strojno izluščene, ki so že bile v bazi.

Uvoz podatkov je potekal prek t. i. podatkovnega skladišča, v katerem hranimo vse kolokacijske kandidate, pridobljene iz korpusa (skupaj več kot 63 milijonov). Skladišče vsebuje ID-je korpusnih stavkov, v katerih je kolokacija najdena, pomene, pod katere kolokacija sodi, ustreznost kolokacije za slovar glede na njene sestavne komponente itd. Uporaba podatkovnega skladišča olajša analizo podatkov, statistiko, pridobivanje podatkov in vzdrževanje povezave s korpusnimi metapodatki. Evidenca ne le dobrih, temveč tudi slabih kolokacijskih kandidatov, je ključnega pomena za preprečevanje podvajanja dela v prihodnosti.

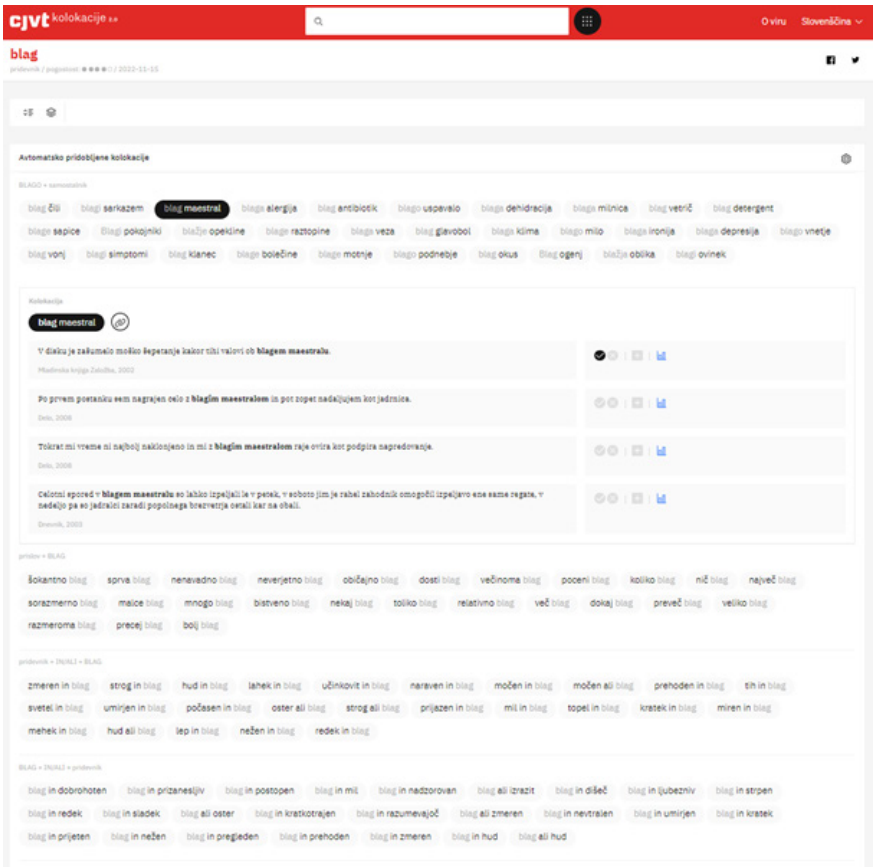
Za izbranih 1.608 iztočnic smo izluščene kolokacije v celoti ročno pregledali in ovrednotili, pri čemer smo kot edino omejitev pri luščenju vzeli pogostnost 4 in več. Pri odločanju smo imeli na voljo tri odločitve: izluščen kolokacijski kandidat je (a) kolokacija; (b) kolokacija, vendar ni relevantna za kolokacijski slovar, in (c) ni kolokacija. Kolokacije iz prve skupine smo vključili v slovar, kolokacije iz druge skupine pa smo ohranili v Digitalni slovarski bazi.

Ročna analiza izbranih kolokacij za 1.608 iztočnic je služila tudi za oceno kakovosti strojno pridobljenih podatkov za vsako skladenjsko strukturo. Rezultati so pokazali nabor struktur, za katere so bili v veliki meri izluščeni kolokacijski kandidati ocenjeni kot dobri. Na primer, za strukturi pridevnik + predlog + pridevnik in pridevnik + samostalnik, je bilo takih več kot 90 % izluščenih kolokacij. Med strukturami, za katere je bil značilen največji odstotek neustreznih kolokacijskih kandidatov (več kot 80 % izluščenih kolokacij) pa lahko navedemo: pridevnik + in/ali + pridevnik, samostalnik + zanikan glagol, samostalnik + samostalnik v dajalniku in samostalnik v imenovalniku + glagol v 3. osebi.

3.4 Nadgradnja slovarskega vmesnika

Vmesnik KSSS 1.0 je doživel nekaj bistvenih sprememb zaradi uskladitve z vmesniki drugih jezikovnih virov CJVT, predvsem pa zaradi izsledkov uporabniških raziskav. Ključne spremembe se nanašajo na razširitev postavitve strani, spremembo pisave in premik polja s pomenskim menijem in filtrom za strukture s položaja v levem stolpcu v zgornjo vrstico nad vsebino (gl. sliko 2). V drugi različici KSSS smo prevzeli tudi prikaz gesla iz drugih slovarjev CJVT UL, v katerem so zdaj kolokacije v glavnem oknu razporejene po pomenih.

Na podlagi uporabniške študije so bile uvedene še druge spremembe, kot je denimo zmanjšanje število klikov za dostop do dodatnih podatkov ter različne možnosti upravljanja s podatki, npr. omejitev pogleda na izbor najpogostejših kolokacij; urejanje kolokacij po pogostosti, abecednem redu, obrnjenem abecednem redu in dolžini; prikaz ali skrivanje iztočnice v kolokaciji ipd. (slika 2).



Slika 2: Strojno pripravljeno geslo v prenovljenem vmesniku KSSS 2.0

Veliko pozornosti smo namenili izboljšanju jasnosti predstavitve kolokacij v vmesniku. Piramido za označevanje stopnje izdelanosti gesla smo nadomestili s tremi tipi gesel: postavitev pomensko razčlenjenega gesla z ročno pregledanimi kolokacijami, postavitev pomensko razčlenjenega gesla z ročno razvrščenimi kolokacijami pod pomene in s strojno pridobljenimi kolokacijami, ki niso pomensko razvrščene, ter postavitev s samo strojno pridobljenimi kolokacijami. V novi postavitvi smo izpostavili tudi skladenjske strukture, ki so bile v prvi različici vidne le ob prehodu z miško, ohranili pa smo način prikazovanja zgledov, do katerih uporabniki dostopajo s klikom na kolokacijo, še vedno pa je na voljo tudi povezava do korpusa, ki prikazuje vse korpusne zglede za določeno kolokacijo.

Druga pomembnejša nadgradnja vmesnika je povezana z uporabniško izkušnjo pri vključevanju v množičenjske naloge. V prvi različici je bila ta možnost omejena na označevanje dobrih ali slabih kolokacij z glasovanjem. V drugi različici smo uporabniški doprinos prestavili na raven zgleda; uporabniki lahko potrdijo ustreznost kolokacije v vsakem prikazanem zgledu, hkrati pa ga lahko razporedijo

v ustrezen pomen, če je pomenska členitev že izdelana (slika 2). V primeru, da uporabniki iskanega pomena ne najdejo, lahko dodajo zgled pod nov oziroma drug pomen, ki tako lahko vključuje več novih pomenov skupaj.

4 Zaključek in nadaljnje delo

V posodobljenih različicah *Slovarja sopomenk sodobne slovenščine* in *Kolokacijskega slovarja sodobne slovenščine* so bile odpravljene nekatere najpomembnejše pomanjkljivosti prejšnjih različic. Čeprav je bil projekt nadgradnje obeh slovarjev omejen v obsegu in ni bilo mogoče ročno urejati celotne zbirke podatkov, smo z nadgradnjo, ki vključuje nove metodološke pristope pri strojnem pridobivanju podatkov in upošteva izsledke usmerjenih uporabniških raziskav začrtali jasno smer prihodnjega razvoja. Temeljna prednost obeh slovarjev je, da sta del Digitalne slovarske baze, kar omogoča lažje upravljanje ter medsebojno povezljivost podatkov, ki so rezultat različnih slovarskih projektov.

Pri nadgradnji ***Slovarja sopomenk sodobne slovenščine*** dajejo smernice, ki smo jih pripravili za pomensko členitev in razvrščanje sopomenk pod pomene, označevanje negativno zaznamovanega besedišča in za ocenjevanje uporabniških predlogov, trdno podlago za nadaljnjo širitev in izboljšavo slovarske zbirke. V različici 3.0 bo naša pozornost še naprej usmerjena v strojno pridobivanje protipomenk, ki je dalo obetavne rezultate, zahteva pa nadaljnje vrednotenje in izvedbo na obsežnejši količini podatkov. Poleg tega načrtujemo metodološke izboljšave in posodobitve pri luščenju, izboru in vizualizaciji sopomenk ter pri vključevanju kolokacijskih podatkov, ki se izboljšujejo tudi z razvojem *Kolokacijskega slovarja sodobne slovenščine*.

Druga različica ***Kolokacijskega slovarja sodobne slovenščine*** prinaša številne spremembe tako na ravni vsebine kolokacijskih podatkov, kot tudi pri načinu njihove predstavitve uporabniku. Pri spremembah smo upoštevali nove dosežke na področju strojnega pridobivanja kolokacij iz korpusov in ugotovitve različnih uporabniških študij. Dolgoročno nameravamo dodati še druge možnosti prikazovanja kolokacij, na primer po vzoru nemškega *Elexiko* (Storjohann 2005) in hrvaškega *Mrežnika* (Hudeček in Mihaljević 2020) na podlagi tipičnih vprašanj in/ali pomenskih lastnosti, npr. z uporabo t. i. semantičnih tipov (Kosem in Pori 2021). V načrtu je tudi izvedba nadaljnjih uporabniških študij za nove izboljšave vmesnika, na podlagi ocene 1.608 ročno pregledanih iztočnic pa bo izboljšana tudi metoda strojnega luščenja.

Zahvala

Projekt Nadgradnja temeljnih slovarskih virov in podatkovnih baz CJVT UL je v letih 2021–2022 financiralo Ministrstvo za kulturo Republike Slovenije. Raziskovalna programa št. P6-0411 (Jezikovni viri in tehnologije za slovenski jezik) in št. P6-0215 (Slovenski jezik – bazične, kontrastivne in aplikativne raziskave) sofinancira Javna agencija za znanstvenoraziskovalno in inovacijsko dejavnost Republike Slovenije iz državnega proračuna.

Literatura

Arhar Holdt, Špela, 2020: How users responded to a responsive dictionary: the case of the Thesaurus of Modern Slovene. *Rasprave: Časopis Instituta za hrvatski jezik i jezikoslovlje* 46/2. 465–48. DOI: https://doi.org/10.31724/rihjj.46.2.1.

Arhar Holdt, Špela, 2021a: Razvrstitev kolokacij v slovarskem vmesniku: uporabniške prioritete. Kosem, Iztok (ur.): *Kolokacije v slovenščini*. Ljubljana: Znanstvena založba Filozofske fakultete. 125–157. DOI: https://doi.org/10.4312/9789610605379.

Arhar Holdt, Špela, 2021b: Kolokacije v Slovarju sopomenk sodobne slovenščine: evalvacija podatkov in predlog za izboljšavo. Kosem, Iztok (ur.): *Kolokacije v slovenščini*. Ljubljana: Znanstvena založba Filozofske fakultete. 269–296. DOI: https://doi.org/10.4312/9789610605379.

Arhar Holdt, Špela, Čibej, Jaka, Dobrovoljc, Kaja, Gantar, Polona, Gorjanc, Vojko, Klemenc, Bojan, Kosem, Iztok, Krek, Simon, Laskowski, Cyprian in Robnik-Šikonja, Marko, 2018: Thesaurus of Modern Slovene: By the Community for the Community. Krek, Simon idr. (ur.): *Proceedings of the XVIII EURALEX International Congress: Lexicography in global contexts*. Ljubljana: Znanstvena založba Filozofske fakultete. 401–410. DOI: https://doi.org/10.4312/9789610600961.

Arhar Holdt, Špela, Gantar, Polona, Kosem, Iztok, Pori, Eva, Robnik Šikonja, Marko in Krek, Simon, 2023b: Thesaurus of Modern Slovene 2.0. Medved, Marek idr. (ur.): *eLex 2023: electronic lexicography in the 21st century (eLex 2023): proceedings of the eLex 2023 conference: [Brno], 27–29 June 2023.* Brno: Lexical Computing CZ. 366–381. https://elex.link/elex2023/wp-content/uploads/82.pdf. (Dostop 30. 11. 2023.)

Arhar Holdt, Špela, Kosem, Iztok, Pori, Eva, Gorjanc, Vojko, Krek, Simon in Gantar, Polona, 2023a: Negativno zaznamovano besedišče v Slovarju sopomenk sodobne slovenščine 2.0. *Slovenščina 2.0* 11/1, 8–32. DOI: https://doi.org/10.4312/slo2.0.2023.1.8-32.

Arhar Holdt, Špela, Logar, Nataša, Pori, Eva in Kosem, Iztok, 2021: Game of words: play the game, clean the database. Gavriilidou, Zoe idr. (ur.): *Lexicography for inclusion: EURALEX XIX: Congress of the European Association for Lexicography: 7-9 September 2021, Vol. 2*. Komotini: Democritus University of Thrace. 41–49. https://euralex.org/publications/game-of-words-play-the-game-clean-the-database/. (Dostop 30. 11. 2023.)

Colman, Lut in Tiberius, Carole, 2018: A good match: a Dutch collocation, idiom and pattern dictionary combined. Krek, Simon idr. (ur.): *Proceedings of the XVIII EURALEX International Congress: Lexicography in global contexts*. Ljubljana: Znanstvena založba Filozofske fakultete. 233–246. DOI: https://doi.org/10.4312/9789610600961.

Čibej, Jaka in Arhar Holdt, Špela, 2019: Repel the syntruders! A crowdsourcing cleanup of the Thesaurus of modern Slovene. Kosem, Iztok idr. (ur.): *Proceedings of the eLex 2019 conference, Electronic lexicography in the 21st century: Smart lexicography. Sintra, Portugal*. Brno: Lexical Computing. 338–356. https://elex.link/elex2019/wp-content/uploads/2019/09/eLex_2019_19.pdf. (Dostop 30. 11. 2023.)

Fišer, Darja, Ljubešić, Nikola in Erjavec, Tomaž, 2020: The Janes project: language resources and tools for Slovene user generated content. *Language Resources and Evaluation* 54/1. 223–246. DOI: https://doi.org/10.1007/s10579-018-9425-z.

Gantar, Polona, 2015: *Leksikografski opis slovenščine v digitalnem okolju*. Ljubljana: Znanstvena založba Filozofske fakultete. DOI: https://doi.org/10.4312/9789612377922.

Gantar, Polona, Krek, Simon in Kosem, Iztok, 2021: Opredelitev kolokacij v digitalnih slovarskih virih za slovenščino. Kosem, Iztok (ur.): *Kolokacije v slovenščini*. Ljubljana: Znanstvena založba Filozofske fakultete. 15–41. DOI: https://doi.org/10.4312/9789610605379.

Gapsa, Magdalena in Arhar Holdt, Špela, 2023: How Lexicographers Evaluate User Contributions in The Thesaurus of Modern Slovene in Comparison to Dictionary Users. Medved, Marek idr. (ur.): *eLex 2023: electronic lexicography in the 21st century (eLex 2023): proceedings of the eLex 2023 conference: [Brno], 27–29 June 2023.* Brno: Lexical Computing CZ. Brno: Lexical Computing CZ. 178–200. https://elex.link/elex2023/wp-content/uploads/47.pdf. (Dostop 30. 11. 2023.)

Hudeček, Lana in Mihaljević, Milica, 2020: Collocations in the Croatian Web Dictionary – Mrežnik. *Slovenščina 2.0* 8/2. 78–111. DOI: https://doi.org/10.4312/slo2.0.2020.2.78-111.

Kallas, Jelena, Kilgarriff, Adam, Koppel, Kristina, Kudritski, Elgar, Langemets, Margit, Michelfeit, Jan, Tuulik, Maria in Viks, Ülle, 2015: Automatic generation of the Estonian Collocations Dictionary database. Kosem, Iztok idr. (ur.): *Electronic lexicography in the 21st century: linking lexical data in the digital age. Proceedings of the eLex 2015 conference, 11–13 August 2015, Herstmonceux Castle, United Kingdom.* Ljubljana: Trojina; Brighton: Lexical Computing. 1−20. https://elex.link/elex2015/proceedings/eLex_2015_01_Kallas+etal.pdf. (Dostop 30. 11. 2023.)

Kilgarriff, Adam, Baisa, Vít, Bušta, Jan, Jakubíček, Miloš, Kovář, Vojtěch, Michelfeit, Jan, Rychlý, Pavel in Suchomel, Vít, 2014: The Sketch Engine: ten years on. *Lexicography* 1/1. 7–36. DOI: https://doi.org/10.1007/s40607-014-0009-9.

Koppel, Kristina, Kallas, Jelena, Khokhlova, Maria, Suchomel, Vít, Baisa, Vít in Michelfeit, Jan, 2019: SkELL corpora as a part of the language portal Sõnaveeb: problems and perspectives. Kosem, Iztok idr. (ur.): *Proceedings of the eLex 2019 conference, Electronic lexicography in the 21st century: Smart lexicography. Sintra, Portugal*. Brno: Lexical Computing. 763–782. https://elex.link/elex2019/wp-content/uploads/2019/09/eLex_2019_43.pdf. (Dostop 30. 11. 2023.)

Kosem, Iztok in Pori, Eva, 2021: Slovenske ontologije semantičnih tipov: samostalniki. Kosem, Iztok (ur.): *Kolokacije v slovenščini*. Ljubljana: Znanstvena založba Filozofske fakultete. 159–202. DOI: https://doi.org/10.4312/9789610605379.

Kosem, Iztok, Bálint Čeh, Júlia, Ponikvar, Primož, Zaranšek, Petra, Kamenšek, Urška, Koša, Peter, Gróf, Annamária, Böröcz, Nándor, Harmat Császár, Jolanda, Szíjártó, Imre, Šantak, Borut, Gantar, Polona, Krek, Simon, Roblek, Rebeka, Zgaga, Karolina, Logar, Urban, Pori, Eva, Arhar Holdt, Špela in Gorjanc, Vojko, 2021a: *Comprehensive Slovenian-Hungarian Dictionary 1.0.* Repozitorij CLARIN.SI. http://hdl.handle.net/11356/1453. (Dostop 30. 11. 2023.)

Kosem, Iztok, Čibej, Jaka, Dobrovoljc, Kaja, Erjavec, Tomaž, Ljubešić, Nikola, Ponikvar, Primož, Šinkec, Mihael in Krek, Simon, 2022: *Monitor corpus of Slovene Trendi 2022-10*. Repozitorij CLARIN.SI. http://hdl.handle.net/11356/1681. (Dostop 30. 11. 2023.)

Kosem, Iztok, Gantar, Polona, Krek, Simon, Arhar Holdt, Špela, Čibej, Jaka, Laskowski, Cyprian, Pori, Eva, Klemenc, Bojan, Dobrovoljc, Kaja, Gorjanc, Vojko in Ljubešić, Nikola, 2019: *Collocations Dictionary of Modern Slovene KSSS 1.0*. Repozitorij CLARIN.SI. http://hdl.handle.net/11356/1250. (Dostop 30. 11. 2023.)

Kosem, Iztok, Krek, Simon in Gantar, Polona, 2021b: Semantic data should no longer exist in isolation: the Digital Dictionary Database of Slovenian. Gavriilidou, Zoe idr. (ur.): *Lexicography for inclusion: EURALEX XIX: Congress of the European Association for Lexicography: 7-9 September 2021*. Komotini: Democritus University of Thrace. 81–83. https://elex.is/wp-content/uploads/2021/09/Semantic-Data-should-no-longer-exist-in-isolation-the-Digital-Dictionary-Database-of-Slovenian_Kosem-Krek-Gantar_EURALEX2020.pdf. (Dostop 30. 11. 2023.)

Krek, Simon, Arhar Holdt, Špela, Erjavec, Tomaž, Čibej, Jaka, Repar, Andraž, Gantar, Polona, Ljubešić, Nikola, Kosem, Iztok in Dobrovoljc, Kaja, 2020: Gigafida 2.0: the reference corpus of written standard Slovene. Calzolari, Nicoletta idr. (ur.): *LREC 2020: Twelfth International Conference on Language Resources and Evaluation: May 11-16, 2020, Palais du Pharo, Marseille, France.* Paris: ELRA – European Language Resources Association. 3340–3345. http://www.lrec-conf.org/proceedings/lrec2020/LREC-2020.pdf. (Dostop 30. 11. 2023.)

Krek, Simon, Gantar, Polona in Kosem, Iztok, 2022: Extraction of collocations from the Gigafida 2.1 corpus of Slovene. Klosa-Kückelhaus, Annette idr. (ur.): *EURALEX 2022, Proceedings of the XX EURALEX International Congress, 12-16 July 2022*. Mannheim: IDS-Verlag. 240–252. https://euralex.org/wp-content/themes/euralex/proceedings/Euralex%202022/EURALEX2022_Pr_p240-252_Krek-Gantar-Kosem.pdf. (Dostop 30. 11. 2023.)

Krek, Simon, Gantar, Polona, Kosem, Iztok in Dobrovoljc, Kaja, 2021: Opis modela za pridobivanje in strukturiranje kolokacijskih podatkov iz korpusa. Arhar Holdt, Špela (ur.): *Nova slovnica sodobne standardne slovenščine.* Ljubljana: Znanstvena založba Filozofske fakultete. 160–194. DOI: https://doi.org/10.4312/9789610605478.

Krek, Simon, Laskowski, Cyprian in Robnik Šikonja, Marko, 2017: From translation equivalents to synonyms: creation of a Slovene thesaurus using word co-occurrence network analysis. Kosem, Iztok idr. (ur.): *Proceedings of eLex 2017: Lexicography from Scratch*. Leiden: Dutch Language Institut; Brno: Lexical Computing; Ljubljana: Trojina. 93–109. https://elex.link/elex2017/wp-content/uploads/2017/09/paper05.pdf. (Dostop 30. 11. 2023.)

Krek, Simon, Laskowski, Cyprian, Robnik Šikonja, Marko, Kosem, Iztok, Arhar Holdt, Špela, Gantar, Polona, Čibej, Jaka, Gorjanc, Vojko, Klemenc, Bojan in Dobrovoljc, Kaja, 2018*: Thesaurus of Modern Slovene 1.0.* Repozitorij CLARIN.SI. http://hdl.handle.net/11356/1166. (Dostop 30. 11. 2023.)

Logar Berginc, Nataša, Grčar, Miha, Brakus, Marko, Erjavec, Tomaž, Arhar Holdt, Špela in Krek, Simon, 2012: *Korpusi slovenskega jezika Gigafida, KRES, ccGigafida in ccKRES: gradnja, vsebina, uporaba*. Ljubljana: Trojina; Fakulteta za družbene vede. DOI: https://doi.org/10.4312/9789610603542.

Mozetič, Tina, Sever, Miha, Justin, Martin in Pegan, Jasmina, 2022: Evalvacijska kategorizacija strojno izluščenih protipomenskih parov. Fišer, Darja in Erjavec, Tomaž (ur.): *Zbornik konference jezikovne tehnologije in digitalna humanistika*. Ljubljana: Inštitut za novejšo zgodovino. 331–338. https://nl.ijs.si/jtdh22/pdf/

JTDH2022_Mozetic-et-al_Evalvacijska-kategorizacija-strojno-izluscenih-protipomenskih-parov.pdf. (Dostop 30. 11. 2023.)

Pori, Eva in Kosem, Iztok, 2021: Evalvacija avtomatskega luščenja kolokacijskih podatkov iz besednih skic v orodju Sketch Engine. Kosem, Iztok (ur.): *Kolokacije v slovenščini*. Ljubljana: Znanstvena založba Filozofske fakultete. 43–77. DOI: https://doi.org/10.4312/9789610605379.

Pori, Eva, Čibej, Jaka, Kosem, Iztok in Arhar Holdt, Špela, 2020: The attitude of dictionary users towards automatically extracted collocation data: a user study. *Slovenščina 2.0* 8/2. 168–201. DOI: https://doi.org/10.4312/slo2.0.2020.2.168-201.

Pori, Eva, Kosem, Iztok, Čibej, Jaka in Arhar Holdt, Špela, 2021: Evalvacija uporabniškega vmesnika Kolokacijskega slovarja sodobne slovenščine. Kosem, Iztok (ur.): *Kolokacije v slovenščini*. Ljubljana: Znanstvena založba Filozofske fakultete. 235–268. DOI: https://doi.org/10.4312/9789610605379.

Rundell, Michael idr. (ur.), 2010: *Macmillan Collocations Dictionary for Learners of English*. Oxford: Macmillan Education.

Storjohann, Petra, 2005: Elexiko: A Corpus-Based Monolingual German Dictionary. *Hermes, Journal of Linguistics* 34. 55–82. https://ids-pub.bsz-bw.de/frontdoor/deliver/index/docId/5005/file/Storjohann_elexiko_A_Corpus_Based_Monolingual_German_Dictionary_2005.pdf. (Dostop 30. 11. 2023.)

Žagar, Aleš, Kavaš, Matic in Robnik Šikonja, Marko, 2021: Corpus KAS 2.0: Cleaner and with New Datasets. Luštrek, Mitja idr. (ur.): *Informacijska družba – IS 2021: Zbornik 24. mednarodne multikonference: 4.-8. oktober 2021, Ljubljana, Slovenia*. Ljubljana: Institut „Jožef Stefan“. https://doi.org/10.5281/zenodo.5562228. (Dostop 30. 11. 2023.)