Kolokacijski slovar sodobne slovenščine

Iztok Kosem,*† Simon Krek,† Polona Gantar,* Špela Arhar Holdt,*‡
Jaka Čibej,*†‡ Cyprian Laskowski*

* Oddelek za prevajalstvo, Filozofska fakulteta, Univerza v Ljubljani
Aškerčeva 2, 1000 Ljubljana
iztok.kosem@ff.uni-lj.si
polona.gantar@guest.arnes.si
spela.arharholdt@ff.uni-lj.si
jaka.cibej@ff.uni-lj.si
cyprianadam.laskowski@ff.uni-lj.si
† Odsek za tehnologije znanja, Institut »Jožef Stefan«
Jamova cesta 39, 1000 Ljubljana
simon.krek@guest.arnes.si
‡ Fakulteta za računalništvo in informatiko, Univerza v Ljubljani
Večna pot 113, 1000 Ljubljana

Povzetek

V prispevku predstavljamo Kolokacijski slovar sodobne slovenščine, nov leksikalni vir za slovenščino. Vir temelji na uporabi sodobnih leksikografskih metod, ki vključujejo avtomatsko luščenje leksikalnih podatkov iz korpusov, množičenje in hitro odzivnost na spremembe v jeziku. Med pomembnejšimi lastnostmi korpusa je prikazovanje gesel v različnih fazah izdelave, kar je novost v slovenskem in tudi mednarodnem prostoru in nadgrajuje idejo rastočega slovarja, pri čemer je eden glavnih razlogov za vpeljavo tega pristopa upoštevanje potreb uporabnikov. Poleg predstavitve vira in metodologije njegove izdelave se prispevek osredotoča na vmesnik, ki uvaja številne novosti prikaza kolokacijskih podatkov, pa tudi slovarskih podatkov nasploh. Prispevek zaključuje predstavitev načrtov za nadaljnje delo.

Collocations Dictionary of Modern Slovene

The paper presents a new lexical resource for Slovene, namely the Collocations Dictionary of Modern Slovene. The resource is being compiled using state-of-the-art lexicographic methods such as automatic extraction of lexical data from corpora, crowdsourcing, and quick responsiveness to language change. An important aspect of the dictionary compilation is that all entries (whether automatically generated, post-processed, finalized by lexicographers, etc.) are immediately published, while dictionary users are provided with the information on their status, i.e. the stage in the compilation process. After the presentation of the Collocations Dictionary and the methodology of its compilation, the paper focuses on the interface, which introduces several innovations in the presentation of collocational information. The paper concludes with an overview of future plans.

1. Uvod

Na mednarodni ravni se v zadnjih letih kaže porast zanimanja za izdelavo kolokacijskih virov. Slovarji kolokacij so nastali oz. nastajajo npr. za estonski (Estonian Collocations Dictionary; Kallas et al., 2015), nemški (German Collocations Dictionary; Roth, 2013; Häcki Buhofer et al., 2014) in španski jezik (DiCE; Vincze et al., 2011; Vincze in Alonso Ramos, 2013). Kolokacijski slovarji v tujini so bili vsaj do zdaj skoraj vedno izdelani za tuje govorce določenega jezika, praksa pa je pokazala, da so kolokacijski podatki zelo koristni tudi za materne govorce, kar navsezadnje potrjuje tudi vse večja tendenca splošnih enojezičnih slovarjev po vključevanju kolokacijskih podatkov. Vseeno pa omenjeni kolokacijski slovarji in podobni viri še vedno ne izkoriščajo vseh prednosti, ki jih ponujajo digitalni mediji; ravno nasprotno, nekateri avtorji, npr. nemškega kolokacijskega slovarja, so pri zasnovi v (pre)veliki meri upoštevali omejitve tiskane različice.

V prispevku predstavljamo Kolokacijski slovar sodobne slovenščine (KSSS), pri čemer največ pozornosti posvečamo metodologiji in postopkom priprave podatkov ter vmesniku, preko katerega bo slovar na voljo uporabnikom. Slovar je rezultat avtomatskih postopkov luščenja kolokacijskih podatkov iz korpusov, ki so bili za slovenščino v zadnjih letih razviti in nenehno izboljševani (npr. Kosem et al., 2013a, 2013b; Gantar et al., 2016). Glavni namen je nasloviti potrebo slovenskih govorcev po jezikovnih virih, usmerjenih v izboljševanje jezikovne produkcije, hkrati pa jezikovnotehnološki skupnosti in ostalim zainteresiranim deležnikom ponuditi obsežne računalniško procesljive podatke o sodobni slovenščini. Poleg tega smo želeli storiti pomemben korak naprej na področju prikazovanja kolokacijskih podatkov in izkoristiti čim več prednosti digitalnih medijev. Tako je bil eden od izzivov izdelati vmesnik, ki bi zadovoljil potrebe različnih uporabnikov, tako maternih kot tujih govorcev slovenščine.

2. Vzorčna baza kolokacij sodobne slovenščine

Vzorčna baza kolokacij je nastala na podlagi poskusnega projekta Baze kolokacijskega slovarja slovenskega jezika (Krek et al. 2016) in vsebuje avtomatsko izluščene kolokacijske podatke (kolokacije z zgledi) za 2.500 gesel, razdeljene po skladenjskih relacijah. Poskusna gesla so na voljo prek posebnega vmesnika (http://bkssj.cjvt.si), o katerem je bila leta 2016 opravljena evalvacijska študija med uporabniki, ki je ponudila tudi informacije o tem, kakšen odnos imajo uporabniki do avtomatsko izluščenih kolokacijskih podatkov. Rezultati študije so pokazali, da nekatere uporabnike neizčiščeni

rezultati oz. neustrezni kolokacijski kandidati ter nestrukturiranost in (pre)velika količina podatkov mestoma motijo, vendar pa se jim scela takšen vir zdi koristen in uporaben. Dejansko je nekatere vprašane celo bolj kot prevelika količina podatkov zmotila premajhna količina gesel.

3. Kolokacijski slovar sodobne slovenščine

Na podlagi strokovnih analiz, pa tudi uporabniških komentarjev vmesnika BKSSJ, smo se lotili avtomatskega izvoza podatkov za veliko večji nabor iztočnic in izdelavo kolokacijskega slovarja (Kosem et al. 2018), in sicer prek Sketch Engine API (Gantar et al., 2015; 2016). Podatki so bili izluščeni iz referenčnega korpusa Gigafida (Logar et al. 2014). Prvotno smo ocenjevali, da bo izvoz zajel približno 50.000 iztočnic, vendar pa se je številka po čiščenju šuma v frekvenčnem seznamu, izločitvi lastnoimenskih iztočnic in iztočnic s prenizko frekvenco in posledično pomanjkanjem koristnih kolokacijskih podatkov skrčila na 35.989 iztočnic, ki vsebujejo skoraj 8 milijon kolokacij in malo manj kot 37 milijonov pripadajočih korpusnih zgledov. Pri izvozu smo uporabili enake nastavitve kot za 2.500 iztočnic BKSSJ (gl. Krek et al., 2016), nekoliko smo izboljšali le konfiguracijo GDEX[1] za slovenščino (Kosem, 2015), npr. kaznovali smo stavke, ki se končajo s podpičjem ipd. Izvožene podatke smo v postopku postprocesiranja dodatno prečistili (deduplikacija zgledov, odstranjevanje kolokacij z vsemi enakimi zgledi ipd.) in prilagodili (pripis iztočnice v ustrezni obliki, zapis kolokatorja v ustrezni obliki glede na podatke v oblikoslovnem leksikonu Sloleks ipd.).

Dandanes v praksi najdemo dva prevladujoča načina objave slovarjev. Prvi, ki ostaja zvest tradicionalnim metodam, je objava slovarja, ko so vsa gesla dokončana. Drugi način, ki je postal standard za spletne slovarje, pa je objava novih slovarskih gesel v rednih intervalih (ponavadi enkrat letno, mogoče celo pogosteje) – temu načinu Klosa (2013) pravi rastoči slovar. Za naše namene noben od omenjenih načinov objave ni bil ustrezen. Tudi pri rastočem slovarju lahko traja več let, preden količina gesel doseže dejansko uporabno vrednost za uporabnike.[2] Posledično smo se odločili za pristop, ki smo ga v slovenskem prostoru prvi predlagali v Krek et al. (2013) in pri katerem je čimveč jezikovnih podatkov odprto ponujenih uporabnikom takoj, ko je z jezikoslovnega vidika ocenjeno, da uporabna vrednost za jezikovno skupnost ustrezno odtehta podatkovni šum; tako pripravljeni podatki morajo vsebovati jasno informacijo o stopnji jezikoslovne pregledanosti.

V KSSS smo se odločili za uporabo naslednjih petih stopenj oz. faz slovarskih gesel:

(1) Avtomatsko izluščeni podatki, ki so postprocesirani (deduplikacija zgledov ipd.), dodano je tudi avtomatsko gručenje kolokatorjev glede na semantične lastnosti, tj. glede na semantični tip, npr. ustanove, predmeti ipd.

(2) Podatki po implementaciji leksikalnogramatičnih oz. statističnih »filtrov«. Na primer, pri vseh strukturah smo odstranili kolokator *biti*, saj se je glagol v veliki večini analiziranih primerov pojavljal v neustreznih kolokacijah oz. je bil pomensko izpraznjen. V bodoče načrtujemo izdelavo obsežnejših seznamov kandidatov za izločanje (angl. stoplist). Kot drugo smo iz avtomatsko izluščenih podatkov izločili vse predložne strukture, ki jih ne potrjuje Slovenski pravopis (Toporišič ur., 2001), saj se je v večini primerov izkazalo, da gre za napačno prepoznane strukture zaradi napak v označevanju.[3]

(3) Podatki z zgolj potrjenimi kolokacijami, ki pa še niso razporejene po pomenih. Poudariti velja, da na tej stopnji ne izločamo samo nekolokacij, temveč tudi statistično šibkejše oz. semantično manj relevantne kolokacije.[4] Upoštevati je namreč treba razliko med statističnimi kolokacijami, tj. statistično relevantnimi sopojavitvami dveh (ali več) besed, in semantičnimi kolokacijami, ki opravljajo določeno semantično funkcijo in so posledično relevantne za kolokacijski slovar. Kot primer lahko navedemo kolokacije tipa *bolnišnica + v +* samostalnik v mestniku (npr. *bolnišnica v Ljubljani, bolnišnica v Izoli*), ki so sicer prepoznane kot statistično relevantne, a za kolokacijski slovar niso zanimive. Vseeno so naši kriteriji za vključitev gradiva, ki se bodo sproti dopolnjevali tudi na podlagi ugotovitev temeljnega raziskovalnega projekta Kolokacije kot temelj jezikovnega opisa: semantični in časovni vidiki (KOLOS; J6-8255), manj strogi kot kriteriji nekaterih znanih kolokacijskih slovarjev – npr. avtorji kolokacijskega slovarja založbe Macmillan so izločali že iztočnice, kot so *hiša*, *kupiti* in *dober*, ker naj ne bi imele statistično zelo relevantnih kolokatorjev. Vendar pa že pri *hiša* najdemo kolokacije, kot so *stanovanjska hiša, medijska hiša, gradnja hiše*, katerih jakost je zelo visoka.

(4) Pomensko členjena gesla: kolokacije in pripadajoči zgledi so razporejene po pomenih (več o tem v nadaljevanju).

(5) Dokončno pregledano in z morebitnimi manjkajočimi podatki (npr. oznake) opremljeno geslo.

Precej razmisleka je bilo vloženega v snovanje pomenskih opisov (Kosem et al., 2017). Kot prvo smo se odločili, da bomo uporabili samo pomene, ne pa tudi podpomenov, saj je ta rešitev zaradi zelo majhnih razlik med posameznimi podpomeni bolj smiselna in posledično prijazna uporabnikom. Poleg tega podpomene pogosto še bolj učinkovito kot razlage ponazarjajo kolokacijski nizi.

Pri pomenskih opisih smo se odločili za uporabo kratkih indikatorjev namesto daljših razlag, saj predvidevamo, da uporabniki pomene besed, ki jih iščejo, bodisi poznajo ali pa ne potrebujejo natančnih razlag, temveč le osnovne pomenske namige za prepoznavo ustreznega pomena. Indikatorjem podobne mehanizme že dolgo uporabljajo slovarji za tuje govorce, predvsem angleščine (npr. Longman Dictionary of Contemporary English), zadnje

[1] GDEX (Good Dictionary EXamples; Kilgarriff et al., 2008) je del korpusnega orodja Sketch Engine, s katerim rangiramo kandidate za dobre slovarske zglede.

[2] Dober primer omenjene problematike je eSSKJ (https://fran.si/201/esskj-slovar-slovenskega-knjiznega-jezika, pri katerem je bilo v prvih dveh letih priprave objavljenih 611 gesel (v prvem letu celo manj kot 100 gesel).

[3] Med izločenimi strukturami so sicer mogoče tudi takšne, ki bi bile lahko legitimne, vendar pa moramo pred njihovo ponovno vključitvijo opraviti podrobnejšo analizo podatkov.

[4] Potrjene statistične kolokacije, ki jih ne vključimo v KSSS, sicer ostajajo v interni bazi, saj so relevantne za leksikografske (npr. izdelavo splošnih enojezičnih slovarjev) in jezikovnotehnološke namene.

čase pa tudi splošni enojezični slovarji (npr. Veliki slovar poljskega jezika).[5]

V KSSS uporabljamo več različnih tipov indikatorjev, npr. sinonime iztočnic, nadpomenke kolokatorjev določenega pomena, najbolj tipični kolokator pomena, področje rabe ipd. Težimo k jedrnatosti, tj. indikatorji so praviloma eno- ali dvobesedni.

Ker večino pomenske informacije prinašajo kolokacije same in njihovi zgledi, je ključna vloga indikatorjev predvsem v vzpostavljanju jasnih razlik med pomeni. Razločevalnost ima pri oblikovanju indikatorjev prednost med sistematičnostjo, kar pomeni, da lahko pri posamezni iztočnici za različne pomene izberemo različne tipe indikatorjev, npr.:

prevajati *(glagol)*

1. jezike
2. energijo
3. dražljaje
4. v računalništvu
5. v drugačno obliko

briljanten *(pridevnik)*

1. o občudovanju
2. iz briljantov
3. bleščeč

Zaradi že prej omenjene stopenjskosti gesel bo imel KSSS lastnosti odzivnega slovarja (Krek et al., 2017; Arhar Holdt et al., 2018), saj se bo odzival na spremembe v jeziku tako, da bodo na podlagi analiz novih podatkov, npr. nove verzije referenčnega korpusa slovenskega jezika, posodabljana tudi že objavljena gesla.

Izziv, ki ga prinaša odzivnost, pa je kratek čas za pripravo podatkov, saj sodobni uporabniki pričakujejo, da so jim slovarske informacije na voljo zelo hitro (Müller-Spitzer, 2014). To potrebo v prvi meri pokrivamo z vključitvijo podatkov v različnih fazah obdelave, pri čemer je končni cilj seveda ponuditi leksikografsko pregledane in redno ažurirane slovarske informacije. Ker je količina kolokacijskih podatkov za pregledovanje zelo velika, poleg tega pa smo tudi kadrovsko in finančno omejeni, smo za pomoč pri čim hitrejši izdelavi gesel KSSS v postopek vpeljali tudi metode množičenja, ki smo jih zasnovali in preizkusili že pri pripravi Predloga za izdelavo Slovarja sodobnega slovenskega jezika (Krek et al., 2013). Odločitev za vpeljavo množičenja se zdi še toliko bolj samoumevna, saj digitalni svet zdaj omogoča tovrstno podporo leksikografskih delotokov.

3.1. Množičenje podatkov za KSSS

Glavni namen vključitve neleksikografov v proces izdelave slovarja je razbremeniti leksikografe rutinskih nalog in njihovo znanje in energijo usmeriti v zahtevnejše leksikografske naloge, kot sta npr. pomenska členitev in pri pripravi indikatorjev. Ena od nalog, ki se nam je zdela primerna za množičenje, je uvrščanje zgledov pod pomene. Zgled, ki v našem primeru ponazarja konkretno kolokacijo, mora množičnik uvrstiti v enega od ponujenih pomenov.

Takšno nalogo smo izvedli na 3.295 kolokacijah iz 88 gesel KSSS, pri čemer smo ponudili po dva zgleda na kolokacijo (skupaj 6.590 mikronalog).[6] Nalogo smo pripravili v lokalni inštalaciji platforme Pybossa (Slika 1).[7] Za nalogo smo uporabili štiri označevalce, študente jezikoslovja, za vsako mikronalogo smo želeli dobiti tri odgovore. Poleg zgleda so označevalci imeli na voljo informacijo o kolokaciji, ki jo je zgled ponazarjal, ter pomene iztočnice, katere kolokacijo so označevali. Poleg pomenov so označevalci lahko izbrali tudi odgovor "Nič od naštetega", s čimer naj bi opozorili, da gre za pomen, ki ga ni na seznamu ponujenih, in "Ne vem", če niso vedeli, katerega od ponujenih pomenov izbrati.

Slika 1: Naloga uvrščanja zgledov pod pomene v Pybossi.

Ujemanje označevalcev je bilo precej visoko, strinjali so se v 79-86 % kolokacij (v povprečju v 83 % kolokacij, povprečna Cohenova kapa je bila 0,83). V 65 % kolokacij oz. 4.258 kolokacijah so se v odgovoru strinjali vsi trije označevalci. V 1.387 primerih (21 %) sta se strinjala po dva označevalca, le 147 primerov (2 %) pa je bilo povsem brez ujemanja. Večina primerov brez ujemanja (106 primerov) je bila označena z "Ne vem" ali z "Nič od naštetega" (54 primerov).[8] Na podlagi teh preizkusnih rezultatov lahko zaključimo, da je raba množičenja vsaj za takšno vrsto slovarsko vezane naloge, precej koristna.

Rezultati so pokazali še dodatno korist naloge, in sicer pridobivanje povratnih informacij o ustreznosti ubeseditve indikatorjev in pomenske členitve, pa tudi o morebitnih manjkajočih pomenih. Tako smo recimo pri glagolu *prihraniti* prvotno imeli ločena pomena za "manj porabiti" (npr. *prihraniti pri stroških*, *prihraniti denar*) in "varčevati" (npr. na banki), analiza odgovorov množičenja pa je pokazala, da bi bilo treba bodisi spremeniti ubeseditev enega ali celo obeh indikatorjev ali pa pomena združiti v en sam pomen. S takšno množičenjsko nalogo tako že dobivamo uporabniške povratne informacije, ki jih ponavadi raziskovalci oz. založniki pridobivajo, če sploh, šele v študijah po objavi slovarja oz. slovarskih gesel.

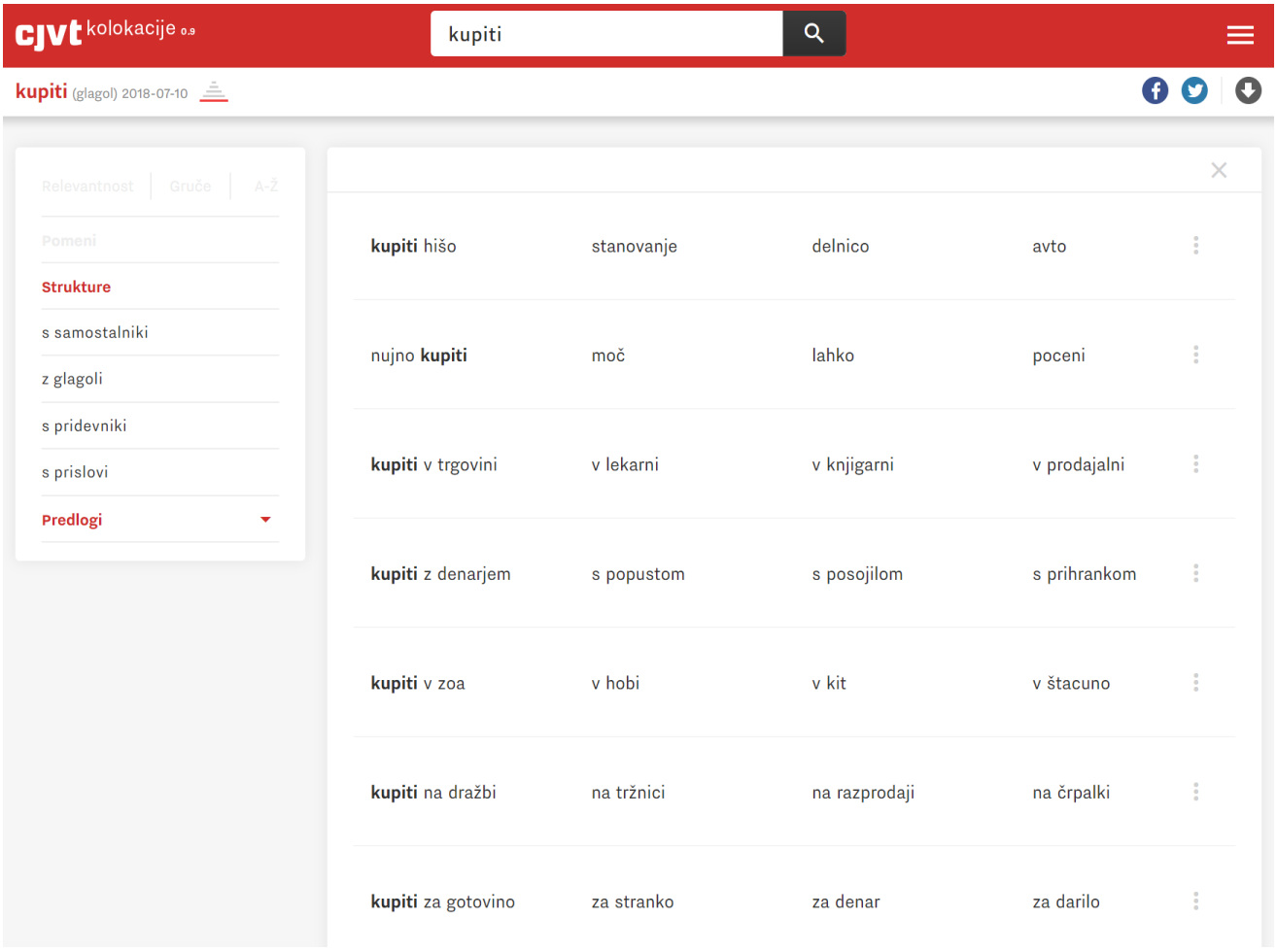
[5] http://wsjp.pl
[6] Število vseh kolokacij v 88 geslih je sicer še večje, a smo izločili (potrjene) kolokacije, ki smo jih našli v Leksikalni bazi za slovenščino (Gantar in Krek, 2011; Gantar et al., 2012).
[7] https://pybossa.com
[8] 26 primerov je bilo označenih tako z "Ne vem" kot z "Nič od naštetega".

4. Vmesnik KSSS

Veliko pozornosti smo posvetili tudi zasnovi vmesnika KSSS.[9] Pri pripravi funkcionalnosti vmesnika smo izhajali predvsem iz informacij, pridobljenih pri uporabniški evalvaciji vmesnika poskusnih 2.500 gesel (Krek et al., 2016). Prva ključna odločitev je bila, da se vmesnik v jedrnem delu posveča kolokacijam, medtem ko so ostali tipi informacij, npr. pomeni, skladenjske relacije ipd., podani kot filtri. Na ta način se odmikamo od pomensko temelječega podajanja kolokacij, ki ga uporabljajo predvsem tiskani (kolokacijski slovar založbe Oxford University Press), pa tudi digitalno zasnovani kolokacijski slovarji (npr. kolokacijski slovar estonskega jezika).

Daleč največji izziv je bil, kako uporabnikom na jasen in nevsiljiv način posredovati informacijo o različnih stopnjah izdelanosti gesel. Čeprav je ta implicitno razvidna iz razpoložljivih funkcionalnosti vmesnika, npr. odsotnost filtra Pomeni nakazuje, da pomenska analiza na podatkih še ni bila opravljena, smo hoteli posamezne stopnje izdelave gesla nakazati tudi eksplicitno. Po daljšem razmisleku in diskusijah smo se odločili za uporabo ikone v obliki petstopenjske piramide, saj najbolje ponazarja postopek izdelave gesla: na začetku je podatkovno bogata, a vseeno že precej zanesljiva avtomatsko izluščena osnova, ki jo z vsakim korakom (proti vrhu) čistimo oz. pilimo. Poleg tega je sestavni del vsakega gesla tudi informacija o datumu zadnje posodobitve.

Pod vrstico, ki vsebuje informacije o iztočnici, tj. besedno vrsto, datum zadnje posodobitve in fazo izdelave gesla, se vmesnik deli na dva dela: na desni je osrednje okno s kolokacijami oz. kolokatorji, na levi pa (ožji) stolpec s filtri in funkcijo razvrščanja.[10] Ob odprtju gesla se uporabniku prikaže neke vrste kolokacijski profil iztočnice, saj mu ponudimo po nekaj kolokacij na skladenjsko strukturo (Slika 2). Praviloma je vsaki strukturi namenjena ena vrstica v vmesniku, če pa določena struktura močno prevladuje oz. vsebuje izredno velik delež relevantnih kolokacij iztočnice, lahko obsega več kot eno vrstico in posledično več kolokacij. V tem uvodnem prikazu uporabnik lahko izbere posamezno strukturo in si ogleda vse kolokacije v njej ali pa že izbere konkretno kolokacijo in si ogleda korpusne zglede.



Slika 2: Uvodna stran gesla (primer glagola *kupiti*).

Namen levega stolpca je uporabniku omogočiti, da čim hitreje pride do želenih podatkov. Na vrhu stolpca je možnost razvrščanja kolokacij, ki je na voljo šele, ko uporabnik odpre posamezno strukturo. Privzeto so kolokacije razvrščene po relevantnosti oz. statistični jakosti, ostali možnosti sta Gruče (razvrščanje kolokatorjev v skupine glede na semantično podobnost) in A-Ž (po abecednem vrstnem redu).

[9] http://viri.cjvt.si/kolokacije/slv/. Povezava bo aktivna od sredine oktobra 2018, ko bo slovar uradno objavljen.

[10] V različici za mobilne telefone so filtri in razvrščanje na voljo prek menija na priklic na vrhu zaslona.

Možnostim razvrščanja sledijo trije tipi filtrov. Prvi filter, Pomeni, prikazuje pomene iztočnice. Pomembna lastnost tega filtra je, da so vsi pomeni iztočnice ves čas prikazani uporabniku – pomeni, v katerih se izbrana kolokacija ali skladenjska struktura ne pojavlja, so namreč zgolj osiveni, ne pa odstranjeni.

Pod Pomeni sledi filter Strukture, ki omogoča filtriranje kolokacij glede na besedno vrsto (prva raven) ali glede na relevantne oblikoskladenjske lastnosti kolokatorja, kot so npr. sklon, stopnja, število ipd.

Zadnji filter v levem stolpcu vmesnika je Predlogi in omogoča filtriranje predložnih struktur. Ker predlogi niso omejeni na določeno besedno vrsto, smo filter podali ločeno in tako uporabnikom omogočili kombinirano uporabo teh dveh filtrov.

Pri filtrih Strukture in Predlogi ima uporabnik lahko vedno izbrano samo eno možnost, ne more npr. hkrati gledati struktur iztočnice s samostalniki in glagoli. Za tako rešitev smo se odločili, ker je glavni namen filtrov omejiti količino podatkov v desnem oknu.

Pomembna lastnost filtrov Pomeni, Strukture in Predlogi je, da se prilagajajo tudi takrat, ko uporabnik manipulira s podatki v osrednjem oknu, npr. ko izbere posamezno strukturo ali kolokacijo. Na ta način filtri opravljajo informativno vlogo o konkretni kolokaciji ali strukturi. Tak način filtriranja slovenski uporabniki že poznajo, saj je bil uporabljen že v vmesnikih korpusov Gigafida, Kres in Gos.[11]

Poleg filtrov v levem stolpcu so na voljo tudi filtri v osrednjem oknu. Stalno aktivni filter je Pogostost, pri katerem uporabnik vleče drsnik proti "redko" oz. "pogosto". Filter je ponujen na vrhu desnega, glavnega okna, ker se nanaša na pogostost oz. redkost kolokatorjev oz. besed v celotnem korpusu, ne pa na pogostost kolokacij. Glavni namen filtra, ki ga bomo (kot ostale funkcionalnosti) še testirali med uporabniki, je omogočiti določenim skupinam uporabnikov dodatno izločanje nerelevantnih kolokatorjev, npr. učiteljem slovenščine kot tujega jezika izločanje redkejših kolokatorjev oz. besed v jeziku.

Dodatni filtri v osrednjem oknu so ponujeni glede na lastnosti iztočnice (ali iztočnice in kolokatorja) zgolj na ravni posamezne strukture, pa še to samo takrat, ko je njihova uporaba glede na lastnosti iztočnice smiselna. Tako je npr. pri pridevniških iztočnicah na voljo filter za moški, ženski in srednji spol.

Po izboru posamezne kolokacije znotraj strukture se uporabniku prikažejo tudi navigacijski gumbi, ki omogočajo enostavno premikanje med sosednjimi kolokacijami. Na ta način se odpravlja potreba po nenehnem drsenju navzdol in klikanju na naslednje kolokacije, ki si jih uporabnik želi ogledati.

Pomembno vlogo v vmesniku ima tudi iskalno okno, ki omogoča iskanje po iztočnicah, kmalu pa je predvidena tudi možnosti iskanja po kolokacijah. Tako bo v primeru, ko bo uporabnik poiskal konkretno kolokacijo, ponujen prikaz, ki bo nekoliko drugačen od prikaza iste kolokacije znotraj posamezne iztočnice – podane bodo namreč informacije o pomenih (v kolikor bodo na voljo) in povezave na kolokacije znotraj iste strukture, in sicer o obeh iztočnicah, ki sestavljata kolokacijo.

KSSS je del portala virov Centra za jezikovne vire in tehnologije Univerze v Ljubljani, s katerim je kljub samostojnemu vmesniku nenehno ohranjena povezava, saj so podatki o morebitnih zadetkih iskanja, ki ga je uporabnik prvotno izvedel v KSSS, v ostalih virih na portalu na voljo prek klika na gumb ob iskalnem oknu.

Vmesnik KSSS je zasnovan za različne digitalne medije, tj. računalnike, tablice in mobilne telefone, z ustreznimi prilagoditvami, kot je npr. omejitev funkcionalnosti pri mobilnih telefonih na račun večje uporabniške prijaznosti.

5. Zaključek in nadaljnje delo

KSSS prinaša v slovenski prostor pomembno novost, in sicer novo različico odzivnega slovarja, katerega značilnost so podatki na različni stopnji izdelanosti, tj. od avtomatsko izluščenih do leksikografsko pregledanih. Na ta način KSSS sledi metodologiji, zastavljeni v Krek et al. (2013) in Gorjanc et al. (2015).

V prihodnjih letih načrtujemo razvoj tako na metodološki in vsebinski kot na predstavitveni ravni. Z vidika metodologije bomo v okviru projekta KOLOS raziskali morebitne izboljšave pri luščenju kolokacij, kot je npr. uporaba metod distribucijske semantike, preizkusili pa bomo tudi luščenje na skladenjsko razčlenjenih korpusih. Načrtujemo tudi vpeljavo novih metod množičenja, predvsem prek igrifikacije. Poleg dodajanja novih gesel KSSS bomo podatke vsebinsko posodabljali z novim korpusnim gradivom, v prvi vrsti iz korpusa Gigafida 2.0, ki bo objavljen konec leta 2018.

Na predstavitveni ravni se bomo posvetili predvsem testiranju vmesnika z različnimi tipi uporabnikov. Tako je v okviru projekta KOLOS že v teku raziskava, v kateri kombiniramo vprašalnike z intervjuji, preverili pa bomo predvsem, katere informacije bi uporabniki želeli oz. potrebovali na prvi strani gesla. Izsledki bodo pokazali, kako izboljšati vmesniško izkušnjo, npr. odkrili morebitne manjkajoče (ali odvečne) dele uporabniškega vmesnika.

Za jezikovnotehnološki razvoj bodo kolokacijski podatki iz KSSS na voljo kot baza podatkov v repozitoriju CLARIN.SI pod licenco Creative Commons 4.0 CC-BY.

6. Zahvala

Prispevek izhaja iz dveh temeljnih raziskovalnih projektov: Kolokacije kot temelj jezikovnega opisa: semantični in časovni vidiki (J6-8255) in Nova slovnica sodobne standardne slovenščine: viri in metode (J6-8256), ki ju financira Agencija za raziskovalno dejavnost Republike Slovenije.

Avtorji se tudi zahvaljujemo podpori infrastrukturnih programov ARRS, in sicer Centru za jezikovne vire in tehnologije Univerze v Ljubljani in Centru za uporabno jezikoslovje pri zavodu Trojina (I0-0051), ter mednarodnemu projektu ELEXIS (European Lexicographic Infrastructure), ki ga finančno podpira evropski program za raziskave in inovacije Obzorje 2020.

Vmesnik predstavljenih virov je razvil Studio Kruh v sodelovanju z Leonom Noetom Jovanom.

[11] http://www.gigafida.net, http://www.korpus-kres.net, http://korpus-gos.net

7. Literatura

eSSKJ: Slovar slovenskega knjižnega jezika 2016–2017, www.fran.si, dostop 28. 08. 2018.

Annelies Häcki Buhofer, Marcel Dräger, Stefanie Meier in Tobias Roth. 2014. *Feste Wortverbindungen des Deutschen. Kollokationenwörterbuch für den Alltag*. Tübingen: Francke.

Špela Arhar Holdt, Jaka Čibej, Kaja Dobrovoljc, Polona Gantar, Vojko Gorjanc, Bojan Klemenc, Iztok Kosem, Simon Krek, Cyprian Laskowski, Marko Robnik-Šikonja. 2018. Thesaurus of Modern Slovene: By the Community for the Community. V: J. Čibej, V. Gorjanc, I. Kosem in S. Krek, ur., *Proceedings of the XVIII EURALEX International Congress: Lexicography in Global Contexts*, str. 401-410. Ljubljana, Ljubljana University Press, Faculty of Arts. https://e-knjige.ff.uni-lj.si/znanstvena-zalozba/catalog/view/118/211/2991-1.pdf (dostop 25. 8. 2018).

Polona Gantar in Simon Krek. 2011. Slovene lexical database. V: D. Majchráková in R. Garabík, ur., *Natural language processing, multilinguality*, str. 72–80. Brno, Tribun EU.

Polona Gantar, Simon Krek, Iztok Kosem, Mojca Šorli, Katja Grabnar, Olga Pobirk, Petra Zaranšek in Nina Drstvenšek. 2012. *Leksikalna baza za slovenščino*. Ljubljana, Ministrstvo za izobraževanje, znanost, kulturo in šport. http://www.slovenscina.eu/spletni-slovar/leksikalna-baza, http://hdl.handle.net/11356/1030 (dostop 8. 4. 2018).

Polona Gantar, Vojko Gorjanc, Iztok Kosem in Simon Krek. 2015. Going semi-automatic and crowdsourced: collocation dictionary of Slovene. V: I. Kosem, ur., *Electronic lexicography in the 21st century: linking lexical data in the digital age. eLex 2015, knjiga povzetkov*, str. 37. Ljubljana, Trojina, Institute for Applied Slovene Studies; Brighton, Lexical Computing.

Polona Gantar, Iztok Kosem in Simon Krek. 2015. Leksikografski proces pri izdelavi spletnega slovarja sodobnega slovenskega jezika. V: V. Gorjanc, P. Gantar, I. Kosem in S. Krek, ur., *Slovar sodobne slovenščine: problemi in rešitve*, str. 280–97. Ljubljana, Univerza v Ljubljani, Filozofska fakulteta.

Polona Gantar, Iztok Kosem in Simon Krek. 2016. Discovering Automated Lexicography: The Case of the Slovene Lexical Database. *International Journal of Lexicography*, 29(2):200–225.

Vojko Gorjanc, Polona Gantar, Iztok Kosem in Simon Krek, ur. 2015. *Slovar sodobne slovenščine: problemi in rešitve.* Ljubljana, Univerza v Ljubljani, Filozofska fakulteta.

Annette Klosa. 2013. The lexicographical process (with special focus on online dictionaries). V: R. H. Gouws, U. Heid, W. Schweickard in H. E. Wiegand, ur., *Dictionaries. An international Encyclopedia of Lexicography. Supplement Volume: Recent Developments with Focus on Electronic and Computational Lexicography,* str. 517–24. Berlin in Boston, de Gruyter.

Adam Kilgarriff, Miloš Husák, Katy McAdam, Michael Rundell in Pavel Rychly. 2008. GDEX: Automatically Finding Good Dictionary Examples in a Corpus. V E. Bernal in J. DeCesaris, ur., *Proceedings of the Thirteenth EURALEX International Congress*, str. 425–32. Barcelona, Spain, Institut Universitari de Linguistica Aplicada, Universitat Pompeu Fabra.

Adam Kilgarriff, Miloš Husak in Miloš Jakubíček. 2013. Automatic collocation dictionaries. Predstavitev na konferenci eLex 2013, Tallinn, Estonija. Dostopno na: https://youtu.be/b3KyhPBeoLU.

Iztok Kosem, Polona Gantar in Simon Krek. 2013a. Automation of lexicographic work: an opportunity for both lexicographers and crowd-sourcing. V: I. Kosem, J. Kallas, P. Gantar, S. Krek, M. Langemets in M. Tuulik, ur., *Electronic lexicography in the 21st century: thinking outside the paper. Proceedings of eLex 2013 Conference, 17-19 October 2013, Tallinn, Estonia*, str. 32–48. Ljubljana, Trojina, Institute for Applied Slovene Studies; Tallinn, Eesti Keele Instituut.

Iztok Kosem, Polona Gantar in Simon Krek. 2013b. Avtomatizacija leksikografskih postopkov. *Slovenščina 2.0: empirične, aplikativne in interdisciplinarne raziskave,* 1(2):139–164. http://www.trojina.org/slovenscina2.0/arhiv/2013/2/Slo2.0_2013_2_07.pdf (dostop 8. 4. 2018).

Iztok Kosem. 2015. Slovarski zgledi. V: V. Gorjanc, P. Gantar, I. Kosem in S. Krek, ur., *Slovar sodobne slovenščine: problemi in rešitve*, str. 320–38. Ljubljana, Znanstvena založba Filozofske fakultete UL.

Iztok Kosem, Polona Gantar in Simon Krek. 2017. Sense menus in collocations dictionary of Slovene. V: *Electronic lexicography in the 21st century: lexicography from scratch,* str. 43. Leiden, Dutch Language Institut; Brno, Lexical Computing; Ljubljana, Trojina Institute for Applied Slovene Studies.

Iztok Kosem, Simon Krek, Polona Gantar, Špela Arhar Holdt, Jaka Čibej, Cyprian Laskowski. 2018. Collocations Dictionary of Modern Slovene. V: J. Čibej, V. Gorjanc, I. Kosem in S. Krek, ur., *Proceedings of the XVIII EURALEX International Congress: Lexicography in Global Contexts*, str. 989-97. Ljubljana, Ljubljana University Press, Faculty of Arts. https://e-knjige.ff.uni-lj.si/znanstvena-zalozba/catalog/view/118/211/2939-1.pdf (dostop 25. 8. 2018).

Simon Krek, Polona Gantar, Iztok Kosem, Vojko Gorjanc in Cyprian Laskowski. 2016. Baza kolokacijskega slovarja slovenskega jezika. V: T. Erjavec in D. Fišer, ur., *Zbornik konference Jezikovne tehnologije in digitalna humanistika, 29. september - 1. oktober 2016*, Filozofska fakulteta, Univerza v Ljubljani, Ljubljana, Slovenija = *Proceedings of the Conference on Language Technologies & Digital Humanities, September 29th - October 1st*, 2016 Faculty of Arts, University of Ljubljana, Ljubljana, Slovenia, str. 101–105 Ljubljana, Znanstvena založba Filozofske fakultete: = Ljubljana University Press, Faculty of Arts.

Simon Krek, Cyprian Laskowski, Marko Robnik Šikonja, Iztok Kosem, Špela Arhar Holdt, Polona Gantar, Jaka Čibej, Vojko Gorjanc, Bojan Klemenc in Kaja Dobrovoljc. 2017. *Sopomenke 1.0: Slovar sopomenk sodobne slovenščine.* Znanstvena založba Filozofske fakultete Univerze v Ljubljani. Dostopno na: viri.cjvt.si/sopomenke (dostop 13. 04. 2018).

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

Carolin Müller-Spitzer, ur. 2014. *Using Online Dictionaries*. Berlin in Boston, de Gruyter.

Tobias Roth. 2013. Going Online with a German Collocations Dictionary. V: I. Kosem, J. Kallas, P. Gantar, S. Krek, M. Langemets in M. Tuulik, ur., *Electronic lexicography in the 21st century: thinking outside the paper. Proceedings of the eLex 2013 conference, 17-19 October 2013, Tallinn, Estonia*, str. 152–63. Ljubljana/Tallinn, Trojina, Institute for Applied Slovene Studies/Eesti Keele Instituut.

Jože Toporišič, ur. 2001. Slovenski pravopis. Ljubljana: Založba ZRC, ZRC SAZU.

Orsolya Vincze, Estela Mosqueira in Margarita Alonso Ramos. 2011. An online collocation dictionary of Spanish. V: I. Boguslavsky in L. Wanner, ur., *Proceedings of the 5th International Conference on Meaning-Text Theory*, str. 275–86. Barcelona.

Orsolya Vincze in Margarita Alonso Ramos. 2013. Testing an electronic collocation dictionary interface: Diccionario de Colocaciones del Espanol. V: I. Kosem, J. Kallas, P. Gantar, S. Krek, M. Langemets in M. Tuulik, ur., *Electronic lexicography in the 21st century: thinking outside the paper. Proceedings of eLex 2013 Conference, 17-19 October 2013, Tallinn, Estonia*, str. 328–37. Ljubljana, Trojina, Institute for Applied Slovene Studies; Tallinn, Eesti Keele Instituut.