Nina Ledinek
ZRC SAZU
Inštitut za slovenski jezik Frana Ramovša

UDK 811.163.6’367:81’322

SKLADENJSKA ANALIZA SLOVENŠČINE IN SLOVENSKI JEZIKOSLOVNO OZNAČENI KORPUSI

V prispevku opozarjamo na možnosti izrabe jezikoslovno označenih korpusov slovenščine za skladenjsko analizo jezika, pri čemer ugotavljamo, da je zaradi slabše razvite slovenske jezikovne infrastrukture – uporabnikom je sicer na voljo vsaj 8 skladenjsko označenih korpusov slovenščine, žal pa zaradi svoje neobsežnosti omogočajo skladenjske analize le v omejenem obsegu – sistematičnih in celostnih na korpusnih podatkih temelječih raziskav slovenske skladnje malo, večinoma pa se opirajo na analizo oblikoskladenjsko označenih korpusov slovenščine.

Ključne besede: skladnja, korpusno jezikoslovje, skladenjsko označevanje korpusov, oblikoskladenjsko označevanje korpusov, drevesnica

1 Uvod

Za večino korpusnojezikoslovnih in na korpusnih podatkih temelječih raziskav jezika je tipično, da se pri njih osredotočamo na identifikacijo, empirično analizo in statistično korelacijo pogostih ponavljajočih se vzorcev soizbir jezikovnih elementov, tako slovničnih kot leksikalnih, čemur sledi jezikoslovna interpretacija podatkov in oblikovanje predpostavk o delovanju jezika. Glede na tipično naravo na korpusu temelječih jezikoslovnih raziskav se torej zdi, da so te idealne za raziskovanje skladnje. Žal pa je tudi zaradi razmeroma pomanjkljive slovenske jezikovne infrastrukture poglobljenih, celostnih in sistematičnih na korpusnih podatkih temelječih analiz slovenske skladnje malo. Raziskovalcem je na voljo premalo (dovolj) univerzalnih in zmogljivih prosto dostopnih orodij za analizo in vizualizacijo skladenjskih podatkov v (obliko)skladenjsko označenih korpusih, predvsem pa so prosto dostopni skladenjsko označeni korpusi slovenščine premalo obsežni. Kljub temu so iz korpusa pridobljeni

interpretirani podatki o slovenski skladnji ključnega pomena za mnoge jezikoslovne raziskave in pri oblikovanju temeljnih jezikovnih priročnikov za slovenščino, tudi nastajajoče tretje izdaje *Slovarja slovenskega knjižnega jezika* (Gliha Komac et al. 2017, Gliha Komac et al. 2018).

2 Dva pogleda na skladenjsko analizo korpusnih podatkov: oblikovanje jezikoslovno označenih korpusov in jezikoslovna analiza označenih korpusov

O skladenjski analizi je v povezavi z jezikoslovno označenimi korpusi slovenščine smiselno razmišljati z dveh vidikov, in sicer z vidika statistično podprte analize skladenjskih pojavov v jeziku, kot se kažejo v (obliko)skladenjsko označenih korpusih, hkrati pa tudi z vidika raziskovanja skladnje za potrebe samega oblikovanja označevalnih modelov za jezikoslovno označene korpuse. Na pomembne prednosti korpusnojezikoslovne oz. na korpusnih podatkih temelječe in strojno podprte analize jezika v primerjavi z drugimi raziskovalnimi paradigmami in metodologijami opozarjata že Čermák in Teubert, ki se v spodnjih navedkih posredno dotakneta obeh omenjenih vidikov: izpostavljata pomembnost jezikoslovnega označevanja korpusov z vidika testiranja obstoječih jezikovnoteoretičnih opisov in formalizmov ter dejstvo, da številni jezikovni pojavi, kot se kažejo v korpusu, še niso celostno opisani, hkrati pa opozarjata tudi na možnosti, ki jih je v jezikoslovje vnesla računalniška analiza jezika, zlasti na možnost opazovanja pojavov v novih korelacijah.

> Danes je očitno, da samo s korpusom lahko pristopamo k popisom jezika. Pri tem gre (1) za dejstvo, da tradicionalni popisi marsikaj izpuščajo – prvič v zgodovini gre za možnost relativno popolnega popisa jezika; (2) za preciziranje oz. za premeščanje mej in osnov mnogim tradicionalnim jezikoslovnim kategorijam in pojavom (npr. za preizkušanje dosedanjih slovnic); (3) za prvi popis pojavov, za katere do sedaj še ni bilo zbranih in urejenih dovolj podatkov; (4) za realno možnost odkriti pojave v povsem novih odvisnostih. (Čermák 1995 v Gorjanc in Krek 2005: 155.)

> Korpusno jezikoslovje širi naše jezikovno znanje, s tem da kombinira tri postopke: (proceduralno) identifikacijo jezikovnih podatkov v korpusu na podlagi določitve kategorij, korelacijo jezikovnih podatkov s pomočjo statističnih metod in na koncu (intelektualno) interpretacijo rezultatov. Prva dva koraka naj bi bila izvedena kolikor je mogoče avtomatsko; tretji korak predpostavlja namernost. Vsaka interpretacija je dejanje in je ravno zato ni mogoče algoritmizirati. (Teubert 1999 v Gorjanc in Krek 2005: 108.)

2.1 Označevanje slovenskih korpusov z oznakami, relevantnimi za skladenjsko analizo

Jezikoslovno označevanje korpusov je postopek, s katerim jezikovnim elementom v korpusu dodajamo interpretativne jezikovnoanalitične oznake, po katerih uporabniki lahko iščejo. Označevanje lahko poteka na različnih ravneh, za raziskave slovenske skladnje pa so relevantni zlasti trije nivoji označevanja, in sicer lematizacija,

oblikoskladenjsko in (večnivojsko) skladenjsko označevanje. Raziskave slovenske skladnje so sicer možne tako na označenih kot na neoznačenih korpusih slovenščine, vendar je za morfološko bogate jezike, kot je slovenščina, jezikoslovno označevanje korpusov ključnega pomena. Korpusne oznake sicer predstavljajo interpretativni poseg v jezikovno realnost, vendar pa omogočajo uspešnejše luščenje podatkov in njihovo večnamensko izrabo, predvsem pa zagotavljajo, da je kvalitetna analiza jezikovnih podatkov časovno obvladljiva (prim. Arhar Holdt 2011: 21–22).

Kot ugotavljamo že v Ledinek 2014 (16–17), so sistemi korpusnega označevanja oblikovani za ciljno rabo in vzpostavljajo uporabnostne rešitve, zato jih je smiselno dojemati zgolj kot pripomoček, ki olajšuje jezikoslovno analizo podatkov, in ne kot končni rezultat jezikoslovne analize. Zaradi zahtev avtomatske analize jezika pri vzpostavljanju sistema označevanja namreč (lahko) prihaja do jezikovnoteoretičnih poenostavitev. Ustrezna dokumentiranost označevalnih sistemov je zato izjemno pomembna. Nujno je tudi, da se uporabniki označenih korpusov s sistemi njihovega označevanja podrobno seznanijo ter z vidika vsakokratne analize prepoznajo in uzavestijo njihove prednosti in slabosti, podatke pa v skladu z njimi ustrezno interpretirajo. Graditelji označenih korpusov morajo označevalne modele kontinuirano evalvirati in stremeti k njihovemu rednemu nadgrajevanju.

Specializirani jezikovni viri za raziskovanje skladnje jezika so skladenjsko označeni korpusi. Ker je slovenska jezikovna infrastruktura v tem pogledu razmeroma skromna, raziskovalci za analizo skladenjskih pojavov v slovenščini pogosto uporabljajo oblikoskladenjsko označene ter lematizirane korpuse, pri čemer izkoriščajo dejstvo, da dajejo oblikoskladenjske oznake razmeroma dobre informacije o potencialni skladenjski strukturi povedi. Ena od pomembnejših raziskav o skladenjskih pojavih v slovenščini, ki poteka na osnovi analize podatkov v oblikoskladenjsko označenem korpusa Gigafida, je predstavljena v razdelku 2.2.1, v nadaljevanju pa prikazujemo, kakšno je v Sloveniji aktualno stanje na področju skladenjskega označevanja korpusov, in sicer z uporabniškega vidika.

2.1.1 Skladenjsko označevanje korpusov

V skladenjsko označenih korpusih ali drevesnicah so predpostavljena skladenjska razmerja s pomočjo skladenjskih analitičnih oznak eksplicirana na velikem vzorcu besedil dejanske jezikovne rabe, kar omogoča statistični pregled vzorcev distribucije skladenjskih struktur in poglobljeno jezikoslovno analizo. Modelov, po katerih gradimo skladenjsko označene korpuse, je veliko, med seboj se precej razlikujejo, zato bomo v nadaljevanju shematično orisali samo temeljne tipološke razlike med njimi, in sicer s 4 vidikov, ki so relevantni tudi glede na tipološke značilnosti slovenskih drevesnic (prim. Ledinek 2014: 18–24).

Glede na kompleksnost označevalnega sistema, tj. glede na število predpostavljenih jezikovnoanalitičnih oznak in kompleksnost njihovega pripisovanja pojavnicam,

običajno ločujemo med skeletnim oz. plitkim skladenjskim označevanjem nasproti popolnemu skladenjskemu označevanju. Popolno skladenjsko označevanje opredeljuje zelo natančna in podrobna analiza skladenjskih razmerij med vsemi pojavnicami v povedi, pri skeletnem oz. plitkem označevanju pa se osredotočamo na analizo predvsem temeljnih skladenjskih razmerij, skladenjska razmerja pa predstavimo bolj shematično in s skromnejšim naborom analitičnih oznak (McEnery in Wison 1996: 44–45; McEnery et al. 2006: 37; Mitkov 2003: 234).

Skladenjsko označevanje pogosto poteka na dveh ravneh, na površinskoskladenjski ter na pomenskoskladenjski ravni. Prva raven izkazuje strukturnoskladenjska ali funkcijskoskladenjska razmerja med pojavnicami povedi, na pomenskoskladenjski ravni pa so običajno označena pomenska razmerja med glagolom in njegovimi argumenti oz. okoliščinami (udeleženske vloge), lahko pa tudi tematsko-rematska struktura povedi, koreferenčna razmerja ipd.

Označevanje korpusov lahko poteka ročno, polavtomatsko ali avtomatsko. Ročno označevanje je danes v rabi zlasti pri pripravi učnih korpusov, tj. za potrebe oblikovanja učne množice, na podlagi katere izvajamo statistično strojno označevanje korpusov, za polavtomatske postopke označevanja pa se pogosto odločamo v primeru, ko rezultate izhodiščnega strojnega označevanja preverimo in napake odpravimo ročno. Zaradi hitre(jše)ga doseganja dobrih rezultatov in manjšega finančnega vložka se v zadnjih letih najpogosteje odločamo za strojno označevanje korpusov. To lahko poteka po pravilih vnaprej pripravljene slovnice ali s pomočjo statističnih metod. Obetavnejše rezultate trenutno dajejo razčlenjevalniki, ki delujejo po načelih statističnega učenja (McEnery in Wilson 1996: 132–133; Ledinek 2014: 20), vedno pogosteje pa raziskovalci ponovno razmišljajo o uporabi hibridnih metod.

Pri pripravi označevalnih modelov za skladenjsko označene korpuse se praviloma opiramo na enega od dveh jezikovnoteoretičnih formalizmov – gradimo ali odvisnostne modele ali modele, ki temeljijo na predpostavkah frazne gramatike. Pri odvisnostnih modelih nas zanima predvsem binarno asimetrično razmerje podrednost : nadrednost, ti modeli pa se v glavnem povezujejo s funkcijskoskladenjsko analizo jezika. V sestavniški strukturi, kot jo predvidevajo modeli, ki temeljijo na frazni gramatiki, je temeljno razmerje del : celota, frazni modeli pa prinašajo podatke o strukturnoskladenjskih razmerjih med pojavnicami v povedi (prim. Abeille 2003: xvi–xvii).

2.1.2 Skladenjsko označeni korpusi slovenščine

Za raziskovanje slovenske skladnje, razvoj skladenjskih razčlenjevalnikov in druge analize je raziskovalcem trenutno na voljo vsaj 8 skladenjsko označenih korpusov slovenščine, ki so označeni po 4 označevalnih modelih. Vsi korpusi so primerki odvisnostnih drevesnic, v temeljnih obrisih jih predstavljamo v nadaljevanju. Poudariti velja, da gre največkrat za razmeroma neobsežne, mnogokrat učne

korpuse, ki so večinoma prosto oz. odprto dostopni, vendar pa zaradi svoje majhnosti omogočajo skladenjske analize jezika le v omejenem obsegu. Ena temeljnih težav z vidika jezikoslovnega raziskovanja slovenske skladnje torej je, da obsežen prosto dostopen skladenjsko označen korpus slovenščine uporabnikom zaenkrat ni na voljo, orodja za raziskovanje obstoječih virov pa so, vsaj z vidika njihovega povprečnega uporabnika, premalo univerzalna.[1]

2.1.2.1 *Slovenska odvisnostna drevesnica*

Prvi skladenjsko označen korpus slovenščine je bil *Slovenska odvisnostna drevesnica*[2] (SDT) (Džeroski et al. 2006; Erjavec in Ledinek 2006). Gre za pisni korpus slovenščine v obsegu približno 30.000 besed oz. 2000 povedi, ki je označen na površinskoskladenjski ravni, in sicer po modelu češkega korpusa *Prague Dependency Treebank*[3] (v nadaljevanju: PDT) (Bemova et al. 1999; Bejček et al. 2013). Korpus sestavlja del vzporednega korpusa Multext-East (Erjavec 2010), tj. prva tretjina romana *1984* Georgea Orwella.

Z enakim označevalnim modelom je bil skladenjsko označen tudi del vzporednega slovensko-angleškega korpusa *SVEZ-IJS* v obsegu 15.000 besed oz. 800 povedi, ki vključuje pravni red Evropske unije (Erjavec 2006).

Slovenska odvisnostna drevesnica ima zelo kompleksen označevalni model, saj predpostavljeni nabor jezikovnoanalitičnih oznak presega 100 enot, pri čemer je sistem njihovega pripisovanja pojavnicam zelo kompleksen. Ker se je nadgradnja korpusa po omenjenem modelu izkazala za kadrovsko in finančno preveč zahtevno, poleg tega pa rabo korpusa za različne namene omejuje njegova prevodnost in besedilna homogenost, je bil v okviru projektov Jezikoslovno označevanje slovenskega jezika[4] (v nadaljevanju: JOS) in Sporazumevanje v slovenskem jeziku[5] (v nadaljevanju: SSJ) oblikovan bolj robusten jezikovnospecifičen model skladenjskega označevanja.

2.1.2.2 Učni korpus *ssj500k* in *Janes-Syn*

Učni korpus *ssj500k*, ki je nastal v okviru projektov JOS in SSJ, predstavlja najobsežnejši slovenski prosto oz. odprto dostopen skladenjsko označen

[1] Za slovenščino sicer obstaja namensko razvit skladenjski razčlenjevalnik (Dobrovoljc et al. 2012; <http://www.slovenscina.eu/tehnologije/razclenjevalnik> (dostop 20. 6. 2018)), s pomočjo katerega je mogoče graditi skladenjsko označene korpuse slovenščine, vendar pa večina jezikoslovcev, razumljivo, nima potrebnega specializiranega znanja za tovrstne postopke. Predstavljene možnosti zato v največji meri ostajajo neizkoriščene.

[2] <http://nl.ijs.si/sdt/>. (Dostop 20. 6. 2018.)

[3] <https://ufal.mff.cuni.cz/pdt3.0>. (Dostop 20. 6. 2018.)

[4] <http://nl.ijs.si/jos/>. (Dostop 20. 6. 2018.)

[5] <http://www.slovenscina.eu/>. (Dostop 20. 6. 2018.)

korpus slovenščine (Ledinek in Erjavec 2009; Ledinek 2014; Krek et al. 2018). Površinskoskladenjsko označen del korpusa vključuje približno 235.000 pojavnic. Robusten jezikovnospecifičen označevalni model za površinskoskladenjsko označevanje korpusa predvideva 10 jezikovnoanalitičnih oznak, ki jih glede na njihovo vlogo delimo v tri skupine. 5 oznak prvega nivoja je namenjenih označevanju zlasti znotrajbesednozveznih skladenjskih razmerij, 4 oznake drugega nivoja približno sovpadajo s površinskoskladenjskimi kategorijami, z njimi torej večinoma označujemo strukture, ki bi jim pri stavčnočlenski analizi pripisali vlogo osebka, predmeta ter lastnostnega in nelastnostnega prislovnega določila, ena sama oznaka tretjega nivoja pa se uporablja zlasti za označevanje (nad)stavčnih in skladenjsko manj predvidljivih struktur, npr. pastavkov, dostavkov, pristavkov, členkov, eliptičnih struktur, stavčnih priredij ipd. Glede na robustnost označevalnega modela je korpus *ssj500k* namenjen raziskovanju zlasti jedrnih skladenjskih pojavov, predvsem tistih, ki jih določa glagolska vezljivost.

Po nekoliko modificiranem modelu odvisnostnega označevanja JOS-SSJ je označen tudi korpus *Janes-Syn*, skladenjsko označen korpus računalniško posredovane komunikacije v obsegu približno 4400 pojavnic (Arhar Holdt et al. 2016; Arhar Holdt et al. 2017).

Po modelu JOS-SSJ je skladenjsko označen tudi uravnoteženi korpus *KRES*, ki pa ni prosto dostopen (Krek in Dobrovoljc 2014).

2.1.2.3 Pomenskoskladenjsko označen učni korpus slovenščine

Po zgledu korpusa PDT je z oznakami udeleženskih vlog, torej na pomenskoskladenjski ravni, označena tudi približno četrtina skladenjsko označenega korpusa *ssj500k* (Krek et al. 2018). Nastali učni korpus je prvi pomenskoskladenjsko označen korpus slovenščine. Nabor analitičnih oznak je glede na nabor korpusa PDT nekoliko modificiran. Trenutno je predvidenih 24 analitičnih oznak, kaže pa se potreba po vključitvi potencialnih oznak za dodatne udeleženske vloge. Označevalni model (zaenkrat) predvideva pomenskoskladenjsko obveznost zgolj delovalniških udeleženskih vlog, okoliščinske so razumljene kot neobvezne (Dobrovoljc et al. 2016a).

2.1.2.4 *Univerzalna odvisnostna drevesnica za slovenščino* in *Univerzalna odvisnostna drevesnica govorjene slovenščine*

Po modelu *Universal Dependencies*[6] (v nadaljevanju: UD) je označen pisni skladenjsko označen korpus slovenščine *Univerzalna odvisnostna drevesnica za slovenščino* (Dobrovoljc et al. 2016b). Pri UD gre za pobudo za medjezično

[6] <http://universaldependencies.org/>. (Dostop 20. 6. 2018.)

usklajeno skladenjsko označevanje korpusov za namen primerjalnih evalvacij, za možnost kontrastivnih jeziko(slo)vnih analiz, za razvoj večjezičnih skladenjskih razčlenjevalnikov, za spodbude na področju medjezičnega učenja jezikovnih modelov ipd. (Nivre et al., 2016). Poleg slovenske drevesnice je po modelu UD trenutno označenih več kot 70 drevesnic 50 različnih svetovnih jezikov.

Kot osnova za pripravo *Univerzalne odvisnostne drevesnice za slovenščino* je služil skladenjsko označen del korpusa *ssj500k*, jezikovnoanalitične oznake modela UD pa so bile pojavnicam pripisane avtomatsko. Spremembe na ravni pripisovanja jezikovnoanalitičnih oznak se pojavljajo na oblikoskladenjski in skladenjski ravni, pri čemer je bil zaradi številnih razlik med označevalnima modeloma na obeh analitičnih ravneh, zlasti pa pri skladenjskem opisu, eden od ključnih korakov oblikovanje zahtevnega sistema pretvorbenih pravil. V nadaljevanju opredeljujemo temeljne razlike v modelih skladenjskega označevanja korpusov *ssj500k* in *Univerzalne odvisnosten drevesnice za slovenščino*.

Tako model JOS-SSJ kot model UD sodita med odvisnostne modele skladenjskega označevanja korpusov, vendar pa je med njima precej razlik. Kot smo omenili že zgoraj, je bil model JOS-SSJ oblikovan za potrebe analize zlasti temeljnih skladenjskih razmerij v stavku (oz. povedi), kot jih opredeljuje predvsem vezljivost glagola, v določeni meri tudi za analizo strukture besednih zvez, medtem ko model UD več pozornosti namenja označevanju tudi drugih skladenjskih razmerij, zanimivih tudi z vidika interakcije, npr. dostavkom, pastavkom, nagovorom, medmetom in členkom kot modifikacijskim elementom. Tudi sicer je model JOS-SSJ s predvidenimi 10 jezikovnoanalitičnimi površinskoskladenjskimi oznakami bolj robusten kot model UD. Ta namreč predpostavlja 40 skladenjskih oznak, zato je tudi analiza skladenjskih razmerij v okviru besednih zvez bolj podrobna. Model JOS-SSJ v okviru svojih oznak do določene mere upošteva tudi pomenska razmerja, saj so prislovna določila, ki bi jih lahko opredelili kot lastnostna v širšem smislu (in tudi nekatere druge strukture, ki bi jih opredelili kot modifikacijske in ne kot elemente propozicije) označena z drugo oznako kot prislovna določila propozicijskega tipa. Model UD tovrstnih razlik v označevanju ne predvideva, saj razlikuje le med udeleženci in vsemi drugimi strukturami stavka, in sicer neodvisno od njihove pomenskoskladenjske oz. površinskoizrazne obveznosti. Hkrati model UD opozarja na stavčno oz. besednozvezno realizacijo stavčnočlenskih oz. pomenskoskladenjskih razmerij (Dobrovoljc et al. 2016b).

Zaradi razlik v robustnosti označevalnih modelov obeh korpusov vseh skladenjsko označenih povedi iz korpusa *ssj500k* ni bilo mogoče avtomatsko pretvoriti v strukture, označene po modelu UD, zato je po modelu UD trenutno označen le del skladenjsko označenega korpusa *ssj500k* v obsegu približno 8.000 povedi oz. 140.000 pojavnic.

Po modelu UD je označena tudi *Univerzalna odvisnostna drevesnica govorjene slovenščine*, prvi skladenjsko označen korpus govorjene slovenščine, hkrati pa tudi prvi govorni korpus, skladenjsko označen po modelu UD (Dobrovoljc in Nivre 2016). Korpus obsega približno 30.000 pojavnic in je vzorčen iz govornega korpusa *Gos*. *Univerzalna odvisnostna drevesnica za slovenščino* in *Univerzalna odvisnostna drevesnica govorjene slovenščine* sta objavljeni kot dela večje zbirke drevesnic *Universal Dependencies 2.0* (Nivre et al. 2017), po kateri je mogoče brskati s spletnima orodjema PML Tree Query[7] in SETS treebank search.[8]

2.2 Skladenjska analiza jezikoslovno označenih korpusov slovenščine

Zaradi slabše opremljenosti slovenščine z obsežnejšimi skladenjsko označenimi viri trenutno večji del skladenjskih analiz jezika, ki so celovito podprte s korpusnim gradivom, nastaja na osnovi podatkov v oblikoskladenjsko označenih korpusih slovenščine. Med pomembnejše raziskave tega tipa se umeščajo raziskave o tipičnih skladenjskih vzorcih, ki jih sooblikujejo konkretni leksemi (prim. Gantar 2015; Ledinek 2014).

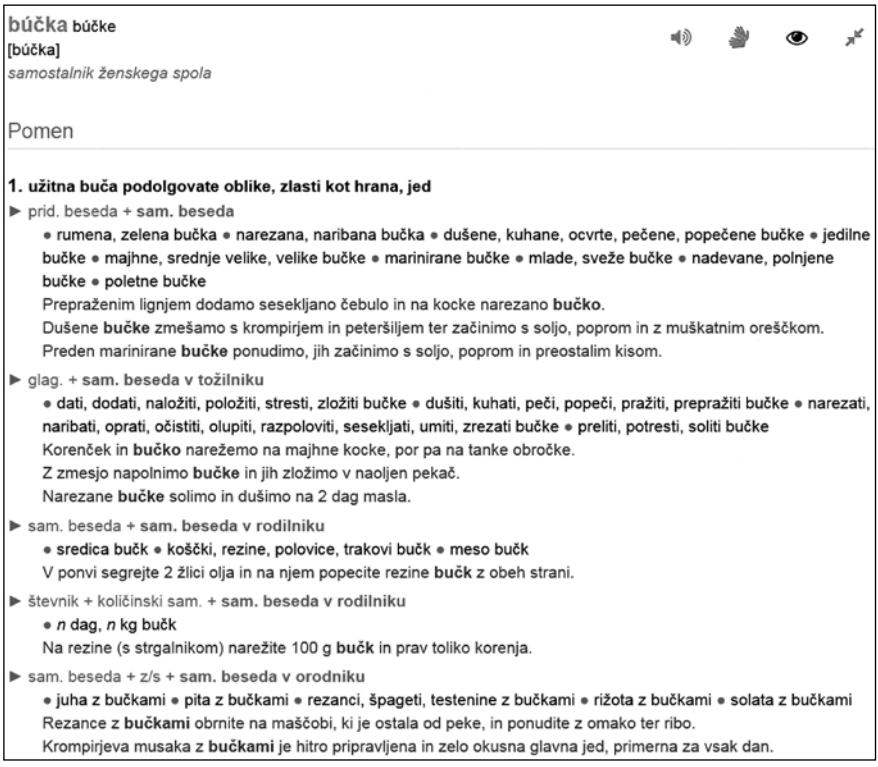
2.2.1 Skladenjski vzorci leksikalnih enot v *eSSKJ*, tretji izdaji *Slovarja slovenskega knjižnega jezika*

Ena pomembnejših aktualnih raziskav slovenske skladnje, tudi z vidika njene celovitosti in obsežnosti, poteka v okviru priprave *eSSKJ*, tretje izdaje *Slovarja slovenskega knjižnega jezika*, ki nastaja na Inštitutu za slovenski jezik Frana Ramovša ZRC SAZU (prim. Gliha Komac et al. 2015), saj eno od redakcijskih faz predstavlja analiza relevantnega besedilnega gradiva z vidika identifikacije tipičnih skladenjskih vzorcev, katerih del je iztočnica oz. leksikalna enota, ki vključuje iztočnico, v konkretnem pomenu ali podpomenu. Ekspliciten prikaz tipičnih skladenjskih vzorcev leksikalnih enot skupaj s prepoznanimi tipičnimi kolokatorji, ki zasedajo mesta v teh vzorcih, je namreč običajen del sodobnih slovarskih priročnikov (prim. Sliko 1).

Pri skladenjski analizi gradiva za slovar tipične skladenjske vzorce, katerih del je leksikalna enota, razložena v *eSSKJ*, tretji izdaji *Slovarja slovenskega knjižnega jezika*, izhodiščno prepoznavamo na podlagi analize in jezikoslovne interpretacije statističnih podatkov o tipičnem kolokacijsko-koligacijskem okolju leme, kot jih izkazuje modul Besedne skice orodja Sketch Engine (Kilgarriff et al. 2004), in sicer na podlagi slovnice besednih skic, ki je bila za analizo korpusa *Gigafida* oblikovana v okviru priprave *eSSKJ*, tretje izdaje *Slovarja slovenskega knjižnega jezika*. Slika 2 predstavlja del podatkov o tipičnem sobesedilnem okolju besede *bučka*, kot ga predstavlja modul.

[7] <http://lindat.mff.cuni.cz/services/pmltq/#!/home>. (Dostop 20. 6. 2018.)

[8] <http://bionlp-www.utu.fi/dep_search/>. (Dostop 20. 6. 2018.)



búčka búčke
[búčka]
samostalnik ženskega spola

Pomen

1. užitna buča podolgovate oblike, zlasti kot hrana, jed

▶ prid. beseda + **sam. beseda**
• rumena, zelena bučka • narezana, naribana bučka • dušene, kuhane, ocvrte, pečene, popečene bučke • jedilne bučke • majhne, srednje velike, velike bučke • marinirane bučke • mlade, sveže bučke • nadevane, polnjene bučke • poletne bučke
Prepraženim lignjem dodamo sesekljano čebulo in na kocke narezano **bučko**.
Dušene **bučke** zmešamo s krompirjem in peteršiljem ter začinimo s soljo, poprom in z muškatnim oreščkom.
Preden marinirane **bučke** ponudimo, jih začinimo s soljo, poprom in preostalim kisom.

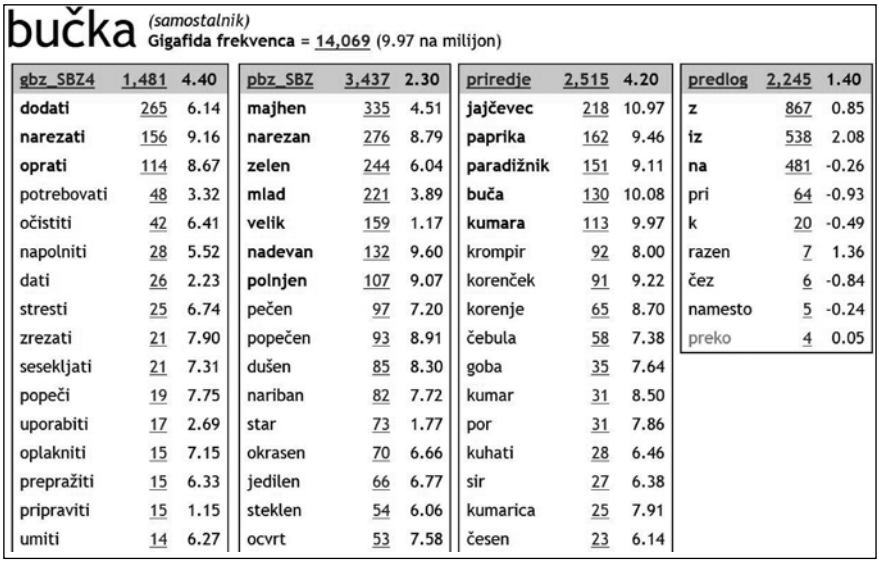
▶ glag. + **sam. beseda v tožilniku**
• dati, dodati, naložiti, položiti, stresti, zložiti bučke • dušiti, kuhati, peči, popeči, pražiti, prepražiti bučke • narezati, naribati, oprati, očistiti, olupiti, razpoloviti, sesekljati, umiti, zrezati bučke • preliti, potresti, soliti bučke
Korenček in **bučko** narežemo na majhne kocke, por pa na tanke obročke.
Z zmesjo napolnimo **bučke** in jih zložimo v naoljen pekač.
Narezane **bučke** solimo in dušimo na 2 dag masla.

▶ sam. beseda + **sam. beseda v rodilniku**
• **sredica bučk** • koščki, rezine, polovice, trakovi bučk • meso bučk
V ponvi segrejte 2 žlici olja in na njem popecite rezine **bučk** z obeh strani.

▶ števnik + količinski sam. + **sam. beseda v rodilniku**
• *n* dag, *n* kg bučk
Na rezine (s strgalnikom) narežite 100 g **bučk** in prav toliko korenja.

▶ sam. beseda + z/s + **sam. beseda v orodniku**
• juha z bučkami • pita z bučkami • rezanci, špageti, testenine z bučkami • rižota z bučkami • solata z bučkami
Rezance z **bučkami** obrnite na maščobi, ki je ostala od peke, in ponudite z omako ter ribo.
Krompirjeva musaka z **bučkami** je hitro pripravljena in zelo okusna glavna jed, primerna za vsak dan.

Slika 1: Prikaz dela tipičnih skladenjskih vzorcev, ki jih v enem od svojih pomenov sooblikuje iztočnica *bučka* (*eSSKJ*, tretja izdaja *Slovarja slovenskega knjižnega jezika*).



bučka *(samostalnik)*
Gigafida frekvenca = 14,069 (9.97 na milijon)

gbz_SBZ4	1,481	4.40
dodati	265	6.14
narezati	156	9.16
oprati	114	8.67
potrebovati	48	3.32
očistiti	42	6.41
napolniti	28	5.52
dati	26	2.23
stresti	25	6.74
zrezati	21	7.90
sesekljati	21	7.31
popeči	19	7.75
uporabiti	17	2.69
oplakniti	15	7.15
prepražiti	15	6.33
pripraviti	15	1.15
umiti	14	6.27

pbz_SBZ	3,437	2.30
majhen	335	4.51
narezan	276	8.79
zelen	244	6.04
mlad	221	3.89
velik	159	1.17
nadevan	132	9.60
polnjen	107	9.07
pečen	97	7.20
popečen	93	8.91
dušen	85	8.30
nariban	82	7.72
star	73	1.77
okrasen	70	6.66
jedilen	66	6.77
steklen	54	6.06
ocvrt	53	7.58

priredje	2,515	4.20
jajčevec	218	10.97
paprika	162	9.46
paradižnik	151	9.11
buča	130	10.08
kumara	113	9.97
krompir	92	8.00
korenček	91	9.22
korenje	65	8.70
čebula	58	7.38
goba	35	7.64
kumar	31	8.50
por	31	7.86
kuhati	28	6.46
sir	27	6.38
kumarica	25	7.91
česen	23	6.14

predlog	2,245	1.40
z	867	0.85
iz	538	2.08
na	481	-0.26
pri	64	-0.93
k	20	-0.49
razen	7	1.36
čez	6	-0.84
namesto	5	-0.24
preko	4	0.05

Slika 2: Vzorčni prikaz dela podatkov o kolokacijsko-koligacijskem okolju leme *bučka*, kot ga prikazuje modul Besedne skice orodja Sketch Engine (korpus Gigafida).

Modul Besedne skice uporablja formalni mehanizem poizvedovanja po atributih, ki so pojavnici pripisani v okviru postopkov jezikoslovnega označevanja, v primeru analize podatkov za potrebe priprave *eSSKJ*, tretje izdaje *Slovarja slovenskega knjižnega jezika* – oblikoskladenjskega označevanja in lematizacije. Za uporabo modula potrebujemo poleg jezikoslovno označenega korpusa, zapisanega v ustrezni obliki, tudi slovnico besednih skic. Ta določa, katere metajezikovne informacije naj algoritem pri izdelavi besednih skic upošteva, ko s pomočjo regularnih izrazov lušči podatke, zahtevane na osnovi nabora vnaprej določenih slovničnih relacij, t. i. gramrelov. Modul temelji na nekoliko razširjeni različici jezika *Corpus Query Language*, ki je bil razvit v devetdesetih letih prejšnjega stoletja (Christ 1994). Slika 3 prikazuje eno od (možnih) slovničnih relacij slovnice besednih skic, ki za raziskovanje skladenjskih pojavov izkorišča oblikoskladenjske podatke – predstavljena skica je namenjena luščenju podatkov o tipičnih pridevnikih, ki razvijajo konkreten samostalnik, in obratno.

```
*DUAL
=kakšen?/kdo-kaj?
2:[tag="P.*"] [tag="[PRZKVL].*"|word=",“|word="se“|word="si“]{0,5}
1:[tag="S.*"]
```

Slika 3: Ena od slovničnih relacij slovnice besednih skic, ki je bila na korpusu FidaPLUS uporabljena za potrebe oblikovanja *Leksikalne baze za slovenščino*.

Da bi zagotovili vsebinsko zanesljivost podatkov v slovarski bazi *eSSKJ*, je relevantnost vseh avtomatsko izluščenih podatkov o tipičnih skladenjskih vzorcih ter tipičnih kolokatorjih, ki zasedajo mesta v njih, ročno preverjena, podatki pa ustrezno jezikoslovno interpretirani. Skladenjski vzorci, identificirani v prvi fazi raziskave, so po potrebi dopolnjeni tudi z vzorci, ki jih prepoznamo ob ročni analizi naključnega nabora približno 300 konkordanc, katerih jedro konkordančnega niza je iztočnica oz. večbesedna leksikalna enota, ki jo sestavlja tudi iztočnica. Tovrstna dopolnila so potrebna zlasti v primerih, ko zaradi izrazitih razlik v frekvenci rabe posameznih pomenov ali podpomenov leksikalnih enot podatki o skladenjskih vzorcih, izluščeni s pomočjo statističnih metod, za vse pomene niso na voljo, ali v primerih, ko je frekvenca rabe določene leme v korpusu Gigafida prenizka, da bi bil modul Besedne skice pri izkazovanju njenega tipičnega sobesedilnega okolja lahko uspešen (prim. Ledinek 2014: 104).

V okviru slovarskih sestavkov *eSSKJ*, tretje izdaje *Slovarja slovenskega knjižnega jezika*, objavljenih do konca leta 2017, je navedenih približno 260 različnih tipičnih skladenjskih vzorcev (npr. *PBZ*[9] *sbz* (ZELEN pulover); *gbz SBZ4* (poslušati GLASBO); *pbz SBZ2* (vajen ČISTOČE), *rbz GBZ* (GOVORITI žaljivo); gbz Inf-GBZ (začeti BLOGATI), *gbz z SBZ6* (živeti z BABICO) ipd.). Skladenjski vzorci so strojno berljivi in so v podatkovni bazi zapisani v transparentnem (opiranje

[9] Enoti *PBZ sbz* in *pbz SBZ* predstavljata isti skladenjski vzorec, velike črke kažejo na to, glede na kateri element vzorec opazujemo – z vidika pridevnika ali z vidika samostalnika.

na površinskoskladenjske kategorije) in dovolj fleksibilnem formalnem jeziku, ki ni vezan na specifične jezikovnoteoretične formalizme, hkrati pa je kompatibilen z oblikoskladenjskimi oznakami JOS, ki v slovenskem prostoru trenutno veljajo za nekakšen standardni nabor oblikoskladenjskih oznak (Erjavec et al. 2010). Za opis segmentov skladenjskih vzorcev v slovarski bazi uporabljamo kategorije tipa *sbz4* (samostalniška beseda v tožilniku), *Inf-gbz* (nedoločnik), *pbz5* (pridevniška beseda v mestniku), *rbz* (prislov) ipd., pri čemer je oznaka neodvisna od notranje strukturiranosti besedne zveze (kot *sbz1* se torej v gradivu lahko uresničujejo enote tipa *hrček*, *dejstvo, da je predsednik priznal napako*, *prijetno dekle*, *moj najnovejši uspeh v deskanju na snegu* ipd.). Za sistemsko označevanje struktur smo se odločili, ker je tovrstno opredeljevanje z vidika raziskovanja skladenjskih razmerij najbolj osnovno, hkrati pa je podatke o tipični notranji strukturiranosti besednih zvez mogoče pridobiti kasneje, s pomočjo luščenja podatkov iz skladenjsko označenih korpusov.

Raziskava, ki jo izvajamo v okviru priprave *eSSKJ*, tretje izdaje *Slovarja slovenskega knjižnega jezika*, ni pomembna samo z vidika identifikacije konkretnih tipičnih skladenjskih vzorcev, ki jih sooblikujejo v slovarju obravnavane leksikalne enote, ampak je relevantna predvsem zato, ker omogoča analizo rabe in primerjavo distribucije rabe tipičnih skladenjskih vzorcev na zelo obširnem vzorcu jezikovnih enot, ki so celostno pomensko in slovnično analizirane. Podatki o tipičnih skladenjskih vzorcih in kolokatorjih namreč predstavljajo le razmeroma majhen del podatkov, navedenih v obširni in preudarno strukturirani strojno berljivi slovarski podatkovni bazi, zapisani v označevanem jeziku XML. Struktura baze raziskovalcem omogoča iskanje po skladenjskih vzorcih v povezavi s tipično pomenskostjo leksikalnih enot, s tipičnimi oblikoslovnimi lastnostmi enot, ki v vzorce vstopajo, oziroma v povezavi ali v razmerju do vseh drugih leksikalnih, slovničnih, pragmatičnih, normativnih in drugih podatkov, ki jih prinaša temeljni enojezični razlagalni slovar slovenščine, pri čemer velja, da so vsi navedeni podatki ročno preverjeni in ustrezno jezikoslovno interpretirani. Ker načrtujemo, da bodo v slovarju po enotni metodologiji analizirane leksikalne enote, ki se pojavljajo v okviru približno 100.000 enobesednih slovarskih iztočnic,[10] bo nabor podatkov v nastajajoči večnamenski podatkovni bazi, ki jih bo mogoče upoštevati pri analizi skladenjskih pojavov v slovenskem jeziku, zelo velik. Podatke je oz. bo mogoče uporabiti ne le za oblikovanje slovarjev in za jezikoslovne raziskave, npr. za potrebe oblikovanja slovnice in drugih jezikovnih opisov, ampak tudi pri (nad)gradnji najrazličnejših naprednih jezikovnih tehnologij.

3 Zaključek

Kratek in pregleden opis stanja na področju (možnosti) računalniško podprtih skladenjskih raziskav na osnovi podatkov v jezikoslovno označenih korpusih slovenščine zaključujemo z mislijo, da so jezikoslovno označeni korpusi in jezikovne

[10] Skladenjskih vzorcev (zaenkrat) ne navajamo pri členkih in medmetih, saj je njihovo skladenjsko okolje nekoliko manj predvidljivo, ne identificiramo pa jih niti pri veznikih in predlogih, katerih vloga je zlasti vzpostavljanje slovničnih razmerij.

tehnologije v širšem smislu pomembno prispevali k razvoju slovenskega jezikoslovja, žal pa je napredek na področju skladenjskega opisa manjši, kot bi si želeli. Deloma je to posledica nekoliko slabše razvite slovenske jezikovne infrastrukture, kljub temu pa velja opozoriti na dejstvo, da se mnogo računalniško podprtih raziskav jezika osredotoča na luščenje podatkov na podlagi predpostavljenih (jezikoslovnih) kategorij, korelacijo jezikovnih podatkov s pomočjo statističnih metod ipd., pogosto pa manjka temeljna faza, tj. jezikoslovna interpretacija jezikovnih podatkov. Razlog za to je verjetno dejstvo, da je ta faza raziskovalno najzahtevnejša, najbolj dolgotrajna in je, kot ugotavlja že Teubert, ni mogoče avtomatizirati in algoritmizirati. K spremembi stanja bi gotovo pripomogla bolj učinkovita slovenska jezikovna in raziskovalna politika, ki bi spodbujala izvajanje temeljnih in dolgotrajnejših bazičnih jezikoslovnih raziskav (prim. Ahačič et al. 2017; Ahačič et al. 2018).

Literatura

Abeillé, Anne (ur.), 2003: *Treebanks: Building and Using Syntactically Annotated Corpora*. Dordrecht: Kluwer Academic Publishers.

Ahačič, Kozma, et al., 2017: *Ciljni raziskovalni projekt Jezikovna politika Republike Slovenije in potrebe uporabnikov: raziskovalno poročilo.* Ljubljana. ZRC SAZU, <https://isjfr.zrc-sazu.si/sites/default/files/raziskovalno_porocilo_28_11_2017.pdf.>. (Dostop 20. 6. 2018.)

Ahačič, Kozma, et al., 2018: *Pravna ureditev in programski dokumenti o jezikovni rabi in praksah jezikovnih uporabnikov v Republiki Sloveniji in uporabnikov slovenskega jezika v sosednjih državah in po svetu.* Ljubljana: ZRC SAZU, Inštitut za slovenski jezik, Založba ZRC. Ur. Nataša Gliha Komac in Polonca Kovač.

Arhar Holdt, Špela, 2011: *Luščenje besednih zvez iz besedilnega korpusa z uporabo dvodelnih in tridelnih oblikoskladenjskih vzorcev*. Ljubljana: Trojina, zavod za uporabno slovenistiko.

Arhar Holdt, Špela, et al., 2016: Syntactic annotation of Slovene CMC: first steps. Fišer, Darja, in Beißwenger, Michael (ur.): *Proceedings of the 4th Conference on CMC and Social Media Corpora for the Humanities*. Ljubljana: Znanstvena založba Filozofske fakultete UL. 3–6.

Arhar Holdt, Špela, et. al., 2017: *CMC training corpus Janes-Syn 1.0*. Slovenian language resource repository CLARIN.SI, <http://hdl.handle.net/11356/1086>. (Dostop 20. 6. 2018.)

Bejček, Eduard, et al., 2013: *Prague Dependency Treebank 3.0*. LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics, Charles University, <http://hdl.handle.net/11858/00-097C-0000-0023-1AAF-3>. (Dostop 20. 6. 2018.)

Bémová, Alla, et al., 1999: *Annotations at Analytical Level: Instructions for Annotators*. Praga: UK MFF UFAL.

Christ, Oliver, 1994: A modular and flexible architecture for an integrated corpus query system. *Proceedings of COMPLEX ,94*. 23–32.

Čermák, František, 2005: Jezikovni korpus: sredstvo in vir spoznanj. [Prevod članka iz revije *Slovo a slovesnost* 56, 1995, 119–140; prev. Andreja Žele]. Gorjanc, Vojko, in Krek, Simon (ur.): *Študije o korpusnem jezikoslovju*. Ljubljana: Krtina. 137–171.

Dobrovoljc, Kaja, et al., 2012: Skladenjski razčlenjevalnik za slovenščino. Erjavec, Tomaž, in Žganec Gros, Jerneja (ur.): *Zbornik Osme konference Jezikovne tehnologije*. Ljubljana: Institut Jožef Stefan. 42–47.

Dobrovoljc, Kaja, et al., 2016a: Označevanje udeleženskih vlog v učnem korpusu za slovenščino. Erjavec, Tomaž, in Fišer, Darja (ur.): *Jezikovne tehnologije in digitalna humanistika*. Ljubljana: Znanstvena založba Filozofske fakultete. 106–110.

Dobrovoljc, Kaja, et al., 2016b: Pretvorba korpusa ssj500k v Univerzalno odvisnostno drevesnico za slovenščino. Erjavec, Tomaž, in Fišer, Darja (ur.): *Jezikovne tehnologije in digitalna humanistika*. Ljubljana: Znanstvena založba Filozofske fakultete. 190–192.

Dobrovoljc, Kaja, in Nivre, Joakim, 2016: The Universal Dependencies Treebank of Spoken Slovenian. *Proceedings of the Tenth International Conference on Language Resources and Evaluation, LREC'16.* ELRA. 1566–1573.

Džeroski, Sašo, et al., 2006: Towards a Slovene Dependency Treebank. *Proceedings of the Fifth International Conference on Language Resources and Evaluation, LREC'06*. ELRA. 1388–1391.

Erjavec, Tomaž, 2006: The English-Slovene ACQUIS corpus. *Proceedings of the Fifth InternationalConference on Language Resources and Evaluation, LREC'06*. ELRA. 2138–2141.

Erjavec, Tomaž, 2010: MULTEXT-East Version 4: Multilingual Morphosyntactic Specifications, Lexicons and Corpora. *Proceedings of the Seventh International Conference on Language Resources and Evaluation, LREC'10*. ELRA.

Erjavec, Tomaž, et al., 2010: The JOS linguistically tagged corpus of Slovene. *Proceedings of the Seventh International Conference on Language Resources and Evaluation, LREC'10.* ELRA.

Erjavec, Tomaž, in Ledinek, Nina, 2006: Slovenska odvisnostna drevesnica: prvi rezultati. Erjavec, Tomaž, in Žganec Gros, Jerneja (ur.): *Zbornik 5. slovenske in 1. mednarodne konference Jezikovne tehnologije IS-LTC*. Ljubljana: Institut Jožef Stefan. 162–167.

Erjavec, Tomaž, in Ledinek, Nina, 2009: Odvisnostno površinskoskladenjsko označevanje slovenščine: specifikacije in označeni korpusi. Stabej, Marko (ur.): *Infrastruktura slovenščine in slovenistike*. (Obdobja 28). Ljubljana: Znanstvena založba Filozofske fakultete. 219–224.

Gantar, Polona, 2015: *Leksikografski opis slovenščine v digitalnem okolju*. Ljubljana: Znanstvena založba Filozofske fakultete.

Gliha Komac, Nataša, et al., 2015: *Koncept novega razlagalnega slovarja slovenskega knjižnega jezika.* Ljubljana: Inštitut za slovenski jezik Frana Ramovša ZRC SAZU.

Gliha Komac, Nataša, et al., 2017: *Slovar slovenskega knjižnega jezika 2016*. Ljubljana: Založba ZRC, ZRC SAZU.

Gliha Komac, Nataša, et al., 2018: *Slovar slovenskega knjižnega jezika 2017*. Ljubljana: Založba ZRC, ZRC SAZU.

Kilgarriff, Adam, et al., 2004: The Sketch Engine. Williams, Geoffrey, in Vessier, Sandra (ur.): *Euralex*. Lorient: Faculte des Lettres et des Sciences Humaines, Universite de Bretagne Sud. 105–115.

Krek, Simon, et al., 2018: *Training corpus ssj500k 2.1*. Slovenian language resource repository CLARIN.SI, <http://hdl.handle.net/11356/1181>. (Dostop 20. 6. 2018.)

Krek, Simon, in Dobrovoljc, Kaja, 2014: Sketch grammar: RegEx-over-POS or dependency parser? A comparison of two MWE extraction methods. *PARSEME 2nd General Meeting.*

Ledinek, Nina, 2014: *Slovenska skladnja oblikoskladenjsko in skladenjsko označenih korpusih slovenščine*. Ljubljana: Založba ZRC, ZRC SAZU.

McEnery, Tony, et al., 2006: *Corpus-Based Language Studies. An Advanced Resource Book*. London: Routledge.

McEnery, Tony, in Wilson, Andrew, 1996: *Corpus Linguistics*. Edinburgh: Edinburgh University Press.

Mitkov, Ruslan (ur.), 2003: *The Oxford Handbook of Computational Linguistics*. Oxford: Oxford University Press.

Nivre, Joakim, et al., 2016: Universal Dependencies v1: A Multilingual Treebank Collection. *Proceedings of the Tenth International Conference on Language Resources and Evaluation, LREC'16.* ELRA. 1659–1666.

Nivre, Joakim, et al., 2017: *Universal Dependencies 2.0*. LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics, Charles University, <http://hdl.handle.net/11234/1-1983>. (Dostop 20. 6. 2018.)

Teubert, Wolfgang, 2005: Korpusno jezikoslovje in leksikografija. [Prevod dela Korpuslinguistik und Lexikographie, *Deutsche Sprache* 4/99, 1999; prev. Mojca Savski]. Gorjanc, Vojko, in Krek, Simon (ur.): *Študije o korpusnem jezikoslovju*. Ljubljana: Krtina. 103–136.