Strojno berljiv Vezljivostni leksikon slovenskih glagolov

Polona GANTAR
Filozofska fakulteta Univerze v Ljubljani, apolonija.gantar@ff.uni-lj.si

Abstract

In this paper, we briefly describe selected models for displaying valency data, and the transfer of existing good practices to the creation of a machine-readable Valency Lexicon of Slovene Verbs, which was produced by automatic extraction of valency patterns from the morphologically, syntactically and semantically annotated Gigafida 2.1 corpus. First, we describe the numerical and statistical representation of the data included in the Lexicon as well as the structure and type of data in it. We then linguistically evaluate the automatically extracted data through a comparative analysis of the selected verb in both the existing Dictionary of Slovenian Transitive Verbs and the newly created Valency Lexicon. We conclude by highlighting possible improvements of the Lexicon, especially in terms of linking data into a single data model – the Slovene Digital Dictionary Database.

Ključne besede: vezljivostni leksikon, vezljivostni vzorci, udeleženske vloge, strojno luščenje vezljivostnih podatkov, korpus Gigafida

Keywords: valency lexicon, valency patterns, semantic roles, automatic extraction of valency data, corpus Gigafida

1 Uvod

Eden od najpogosteje omenjanih izzivov semantičnega spleta v času digitalne transformacije in intenzivnega razvoja umetne inteligence je spreminjanje človeku razumljivih informacij v strojno berljive podatke. Cilj teh prizadevanj je razviti metode, ki so sposobne pretvoriti stavke v obliko, ki omogoča računalniško obdelavo, ter s tem omogočiti strojem, da razumejo človeški jezik. Naloga jezikoslovja v teh prizadevanjih ni trivialna, saj mora zagotoviti, da so podatki, namenjeni strojnemu procesiranju, realni, da ustrezajo specifikam konkretnega jezika in da so hkrati na voljo v čim večjem obsegu. Če torej semantične tehnologije uporabljajo formalno semantiko, da bi opomenile raznolike in neobdelane podatke, ki nas obkrožajo, potem je za jezikoslovje ključno, da izdela jezikovne vire, ki vsebujejo formalizirane jezikoslovne podatke na različnih nivojih. Da bi bilo te vire mogoče izdelati, je potreben jezikoslovni premislek o tem, kaj opredeljuje določeno jezikoslovno kategorijo in kako formalizirati določen jezikoslovni opis, da bo strojno berljiv in hkrati čim bolj univerzalen, da ga bo mogoče vključiti v večjezične modele. Znotraj teh prizadevanj imajo informacije o vezljivostnih lastnostih glagolov, ki tradicionalno veljajo za središče stavka, ključno vlogo pri številnih na pravilih temelječih nalogah računalniškega procesiranja naravnih jezikov, kot so strojno prevajanje, iskanje informacij, povzemanje besedil, odgovarjanje na vprašanja itd. (Kettnerová et al. 2012).

Izhajajoč iz omenjenih potreb in z namenom zagotoviti strojno procesljive podatke, ki bi omogočili izdelavo novega slovničnega opisa slovenskega jezika, ki bo izhajal iz jezikovne realnosti, je bil v okviru projekta Nova slovnica sodobne standardne slovenščine: viri in metode samostojen sklop prizadevanj namenjen izdelavi metodologije za avtomatsko luščenje vezljivostnih vzorcev iz korpusa ter izdelavi strojno procesljivega vezljivostnega leksikona, ki bo vključeval pomensko-skladenjske podatke, uporabne tako za analizo in sintezo besedil kot tudi uporabo v drugih aplikativnih nalogah strojnega procesiranja naravnega jezika. Poleg omenjenih ciljev je treba izpostaviti pomen izdelanega Leksikona tudi z vidika novih za slovenščino

še neizdelanih jezikoslovnih analiz, ki v slovenski prostor prinašajo nova teoretična spoznanja na področju glagolske vezljivosti, kot tudi spoznanja, ki bodo koristna za uporabnike jezikovnih priročnikov in nadaljnje slovnične analize.

V prispevku najprej opišemo nekatere najširše uporabljane vezljivostne leksikone za tuje jezike, ki so nastali na podlagi korpusnih podatkov, formalizacijo in vrsto pomensko-skladenjskih podatkov v njih ter možnosti njihovega prikaza v spletnih vmesnikih. Kratkemu opisu tujejezičnih virov pridružujemo opis vezljivostnih vzorcev v obstoječih slovenskih virih in izpostavljamo dobre prakse, ki smo jih upoštevali pri izdelavi strojno berljivega Vezljivostnega leksikona. V jedru prispevka opišemo njegovo zgradbo in vsebino, in sicer pripravo geslovnika, nabor uporabljenih udeleženskih vlog ter formaliziran zapis vezljivostnih vzorcev. Sledita številčna in jezikoslovna analiza avtomatsko izluščenih podatkov – zadnja temelji na primerjalni študiji glagola *brskati* v ročno izdelanem Vezljivostnem slovarju slovenskih glagolov A. Žele (VSSG) in novem avtomatsko izdelanem Vezljivostnem leksikonu (VL). Prispevek zaključimo z ugotovitvami, ki jih prinaša primerjalna študija, ter možnostmi, ki bi jih bilo smiselno upoštevati pri nadaljnjem razvoju Leksikona.

2 Modeli za prikaz informacij v strojno berljivih vezljivostnih leksikonih

Med izbranimi tujejezičnimi vezljivostnimi leksikoni,[1] ki jih na kratko opišemo v nadaljevanju, se osredotočamo na modele, ki so bili najširše uporabljeni pri prenosu – tipično iz angleščine – na posamezne jezike. Teoretično gledano, temeljijo FrameNet, Vallex in Pattern Dictionary of English Verbs na semantičnem izhodišču, ki se udejanja na posameznem jeziku lastnih slovničnih in skladenjskih pravilih. Metodološko je v obravnavanih modelih v ospredju korpusna analiza rabe besed v realnem sobesedilu s čim večjim deležem

1 Med modeli, ki vsebujejo vezljivostne vzorce in druge z njimi povezane jezikovne ter enciklopedične podatke z možnostjo medjezičnega povezovanja, je treba omeniti vsaj še: Mosaic Knowledge Graph (https://mosaickg.apps.allenai.org/), ConceptNet (https://conceptnet.io/), Babelnet (https://babelnet.org/) in Verbatlas (http://verbatlas.org/).

avtomatizacije postopkov pridobivanja korpusnih podatkov in ohranjanjem ročne analize, ko gre za identifikacijo pomenskih vrednosti. Skupna točka obravnavanih leksikonov je tudi možnost strojnega procesiranja in uporabnost v človeku namenjenih slovarskih aplikacijah. Izbrane modele smo vzeli v obzir tudi zaradi različnih možnosti prikazovanja vezljivostnih podatkov v spletnem okolju z možnostjo prenosa dobrih praks tudi na slovenske podatke.

Med slovenskimi modeli izpostavljamo edini trenutno najobsežnejši vir vezljivostnih podatkov za slovenščino: Vezljivostni slovar slovenskih glagolov (Žele 2008), ki je na voljo tudi v spletnem okolju portala Fran. Temu pridružujemo prikaz vezljivostnih vzorcev v spletni aplikaciji, ki je bila izdelana na podlagi avtomatsko pridobljenih podatkov iz korpusov Kres in ssj500k v pilotni študiji projekta Nova slovnica sodobnega slovenskega jezika: viri in metode, ter prikaz vezljivostnih vzorcev v Leksikalni bazi za slovenščino, ki predstavlja združitev vezljivostnih vzorcev in stavčno oblikovanih pomenskih definicij.

2.1 FrameNet

Med najbolj znane in široko uporabljane[2] strojno berljive semantične leksikone, ki vključujejo vezljivostne podatke, sodi angleški FrameNet,[3] ki temelji na teoriji shemske semantike (angl. *frame semantics*; Fillmore 1976, Fillmore et al. 2003). V FrameNetu se skladenjske realizacije elementov pomenske sheme ali okvirja (ang. *frame*) imenujejo valenčni vzorci in so predstavljeni v obliki relacij ali t. i. tripletov ('FE.PT.GF'), ki opredeljujejo shemski element (*frame element*; 'FE'), tip besedne zveze (*phrase type*; 'PT') in slovnično funkcijo (*grammatical function*; 'GF'). Valenčni vzorci opisujejo celoten nabor vezljivostnih možnosti za vsako leksikalno enoto oz. njen pomen. Tako je denimo za glagol *aktivirati* v pomenu 'narediti (kaj) aktivno ali delujoče' navedenih 12 različnih vezljivostnih vzorcev (Slika 1), ki jih sestavljajo shemski elementi: *Agent* ('vršilec'), *Cause*

2 Pregled Framenetov za posamezne jezike je na: https://framenet.icsi.berkeley.edu/fndrupal/framenets_in_other_languages.

3 Vir: https://framenet.icsi.berkeley.edu/fndrupal/.

(‘vzrok’), *Device* (‘sredstvo’), *Manner* (‘način’), *Place* (‘kraj’), *Purpuse* (‘namen’) in *Time* (‘čas’).

Frame Elements and Their Syntactic Realizations

The Frame Elements for this word sense are (with realizations):

Frame Element	Number Annotated	Realization(s)
Agent	(11)	CNI.-- (3) NP.Ext (8)
Cause	(30)	CNI.-- (1) DNI.-- (2) INI.-- (5) N.Dep (1) NP.Ext (10) PP[by].Dep (10) PP[in].Dep (1)
Device	(42)	NP.Ext (19) NP.Obj (21) N.Head (2)
Manner	(5)	AVP.Dep (5)
Place	(3)	PP[in].Dep (2) PP[inside].Dep (1)
Purpose	(4)	PP[for].Dep (1) Sub.Dep (1) VPto.Dep (1) Sfin.Dep (1)
Time	(1)	PP[as].Dep (1)

Slika 1: Vezljivostni vzorci in oblikoskladenjske realizacije shemskih elementov za pomen glagola *activate* (‘aktivirati’) v FrameNetu.[4]

Kot prikazuje Slika 1, je za vsak shemski element naveden opis oblikoskladenjskih realizacij. Shemski element *Agent* se denimo lahko realizira kot zunanja samostalniška zveza (‘NP.Ext’) ali kot odsotni element, npr. v pasivnih stavkih (angl. *constructional null instantiation*; ‘CNI’). Za slovenščino je bila framenetovska metodologija preizkušena s kontrastivnega vidika na pomenski skupini glagolov premikanja (Može 2013) in pri oblikovanju pomenskih shem in stavčnih vzorcev v Leksikalni bazi za slovenščino (Gantar 2015).

2.2 Pattern Dictionary of English Verbs

Med pristopi, ki združujejo leksikalni in gramatični opis glagolskih pomenov, je pomemben tudi projekt Corpus Pattern Analysis (CPA; Hanks 2004, 2008, Hanks in Pustejovsky 2005), katerega rezultat je Pattern Dictionary of English Verbs.[5] Projekt temelji na projiciranju

4 Vir: https://framenet.icsi.berkeley.edu/fndrupal/frameIndex.
5 Vir: http://pdev.org.uk/#about_cpa.

pomena iz sobesedila na posamezno besedo in izhaja iz teorije jezikovnih konvencij ter možnosti njihove izrabe (angl. *Theory of Norms and Exploitations*; Hanks 1994, 2013, Hanks in Pustojevsky 2004, 2005). Posamezni pomeni glagolov so v slovarju po sistemu CPA povezani s prototipičnimi konteksti, v katerih se glagol pojavlja v realnih besedilih. Vzorce sestavlja osnovna argumentna zgradba glagola, v kateri so stavčni udeleženci opisani s pomočjo semantičnih vrednosti, imenovanih implikature, ki določajo pomenski opis glagola v vzorcu. Kot prikazuje Slika 2, je na primer pomen glagola *aktivirati* 'cause to start to function' ('povzročiti, da začne (kaj) delovati') opredeljen z dvema argumentoma: vršilcem [Anything] in sredstvom, ki se realizira s pomenskimi implikaturami [Device | Body_Part = Cell / Organ] ('sredstvo | del telesa = celica | organ'), medtem ko drugi pomeni tega glagola lahko zahtevajo drugačno število argumentov in drugačne pomenske implikature.

activate

1	[[Anything]] **activate** [[Device | Body_Part = *Cell* | *Organ*]] [[Anything]] causes [[Device | {Body_Part = Cell | Organ}]] to start to function
2	[[Stuff 1 = *Chemical*]] **activate** [[Stuff 2 = *Chemical*]] The presence of [[{Stuff 1 = Chemical}]] causes [[Stuff 2 = Chemical]] to convert to a reactive form
3	[[Human | Institution | Eventuality 1]] **activate** [[Eventuality 2]] [[Human | Institution | Eventuality 1]] causes [[Eventuality 2]] to happen or begin
4	[[Human | Institution]] **activate** [[Rule]] [[Human | Institution]] causes [[Rule]] to come into effect
5	[[Human 1 | Eventuality]] **activate** [[Human 2 | Human_Group]] [[Human 1 | Eventuality]] causes [[Human 2 | Human_Group]] to become involved in a particular situation or process

Slika 2: Glagolski vzorci za pomene glagola *activate* ('aktivirati') v Vezljivostnem slovarju angleških glagolov.[6]

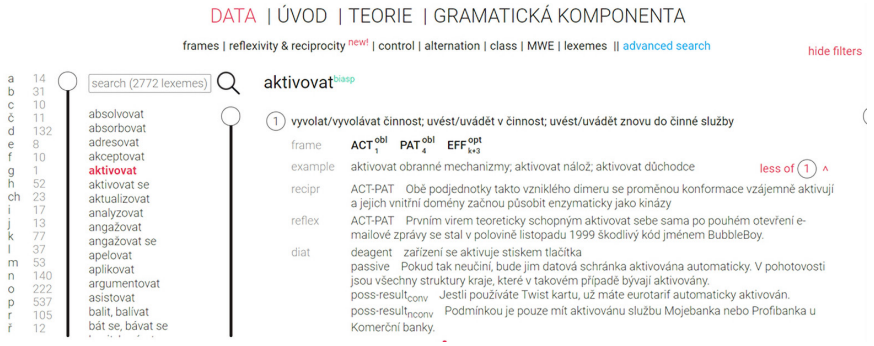
Vsak glagolski vzorec je povezan z naključno izbranimi ročno označenimi korpusnimi primeri v angleškem nacionalnem korpusu (British National Corpus), ki ponazarjajo rabo glagola v danem pomenu v realnem besedilnem kontekstu, označeni argumenti pa omogočajo povezavo z najpogostejšimi kolokacijami, ki se v korpusu pojavljajo na teh mestih.

6 Vir: https://pdev.sketchengine.eu/#about_cpa.

Analiza korpusih vzorcev po vzoru CPA se v veliki meri spogleduje s projektom FrameNet, vendar pa se za razliko od FrameNeta, kjer so v ospredju možnosti združevanja posameznih glagolov v sheme na podlagi njihovega podobnega skladenjskega in pomenskega obnašanja v sobesedilu, CPA osredotoča na sistematično analizo tipičnih pomenskih vzorcev posameznega glagola, manj pa je za ta model zanimiva možnost njihovega medsebojnega povezovanja.

2.3 Vallex

Za morfološko bogate jezike je vzoren primer strojno procesljivega vezljivostnega leksikona češki valenčni leksikon Vallex (Lopatková 2003, Lopatková et al. 2016, Kettnerová 2012).[7] Vallex 4.0 je trenutno zadnja različica Vezljivostnega leksikona pogostejših češkich glagolov, ki ga izdelujejo na Inštitutu za formalno in aplikativno jezikoslovje na Fakulteti za matematiko in fiziko Karlove univerze v Pragi. Gre za elektronsko podatkovno zbirko jezikoslovno označenih in dokumentiranih čeških glagolov (Slika 3), katerih opis temelji na Češkem nacionalnem korpusu in Praški odvisnostni drevesnici (PDT), ki smo jo kot izhodišče uporabili tudi pri določanju udeleženskih vlog za slovenščino (Gantar et al. 2018, Krek et al. 2016).



Slika 3: Prikaz vezljivostnega vzorca za pomen glagola *aktivovat* ('aktivirati') v češkem vezljivostnem leksikonu Vallex 4.0.[8]

7 Vir: https://ufal.mff.cuni.cz/vallex/3.0/.
8 Vir: https://ufal.mff.cuni.cz/vallex/3.0/#/lexeme/aktiv1/0.

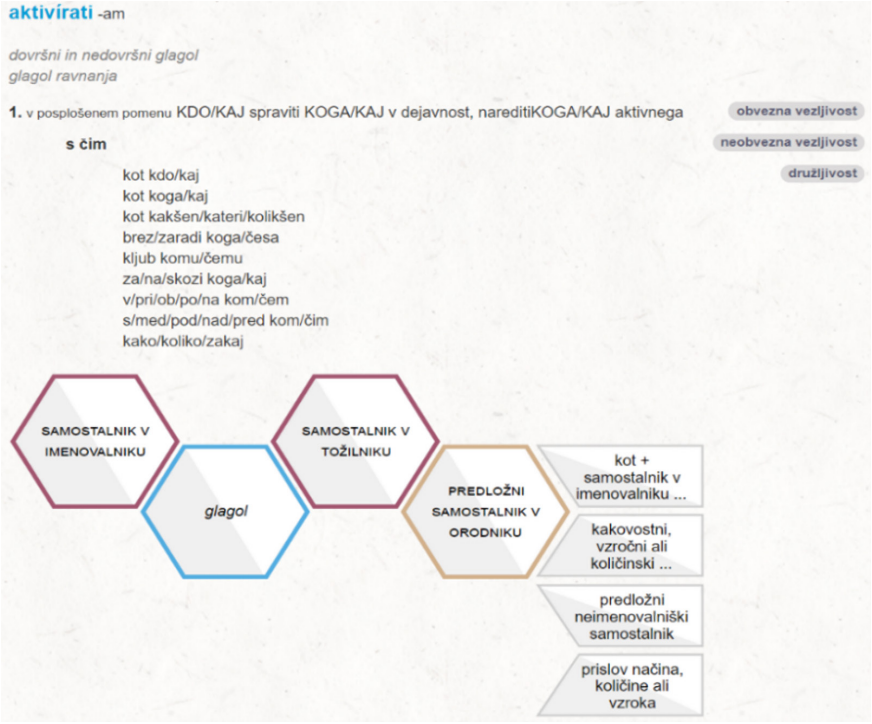
Kot prikazuje Slika 3, je v Vallexu za vsak glagol oz. za vsak glagolski pomen podana informacija o glagolskem vidu, kratka sinonimna razlaga ter vezljivostni vzorec, znotraj katerega so opredeljene udeleženske vloge (t. i. funktorji), morfološka oblika (oblikoslovne realizacije udeležencev) in njihova obligatornost (obvezen, tipičen, opcijski). Poleg tega je za vsak pomen glagola in njegov vezljivostni vzorec navedena povezava na korpus z označenimi stavki in t. i. diateze, ki vključujejo posebnosti pri realizaciji deagentnih in pasivnih zgradb, deležniške oblike in druge skladenjske realizacije.

2.4 Vezljivostni slovar slovenskih glagolov

Za slovenščino so vezljivostni vzorci najbolj podrobno predstavljeni v Vezljivostnem slovarju slovenskih glagolov (Žele 2018), ki je nadgrajena spletna različica tiskanega slovarja (Žele 2008). Spletna različica, ki je na voljo na portalu Fran,[9] vključuje grafični prikaz vezljivostnih vzorcev (Slika 4) s pomočjo t. i. vezljivostnih shem, ki ponazarjajo skladenjsko vezljivost z vidika obveznosti skladenjskih položajev (obvezna in neobvezna vezljivost ter družljivost) in glede na oblikoskladenjske realizacije udeležencev (besedna vrsta, sklon).

Slovar poleg grafične predstavitve oblikoskladenjskih vzorcev prinaša tudi zapis vezljivostnih vzorcev s pomočjo vprašalnic v obliki t. i. besedilne razlage, npr. za pomen glagola *aktivirati*: 'KDO/KAJ spraviti koga/KAJ v dejavnost'. Na mestu glagola v razlagi nastopa sinonim ali abstraktni zastopnik pomena glagola v iztočnici (in ne glagol, za katerega se določa vezljivostni vzorec). Vezljivostni vzorci so razvrščeni glede na obveznost argumentov in vključujejo niz predložnih in prislovnih variant, ki temeljijo na jezikovnosistemski predvidljivosti. Poleg formalizacije vezljivostnih mest na podlagi ročne analize je v VSSG najpomembnejša pomenska razčlenitev glagolov, ki je ključen podatek za ustrezno razvrstitev glagolskih vzorcev pod pomene. Glede na to, da se slovar v pomenski členitvi in pri pomenskem opisu zanaša na stanje v SSKJ, pogostokrat ne odraža realnega jezikovnega stanja: odsotnost realnih leksikalnih podatkov,

9 Vir: https://fran.si/iskanje?FilteredDictionaryIds=218&View=1&Query=%2A.



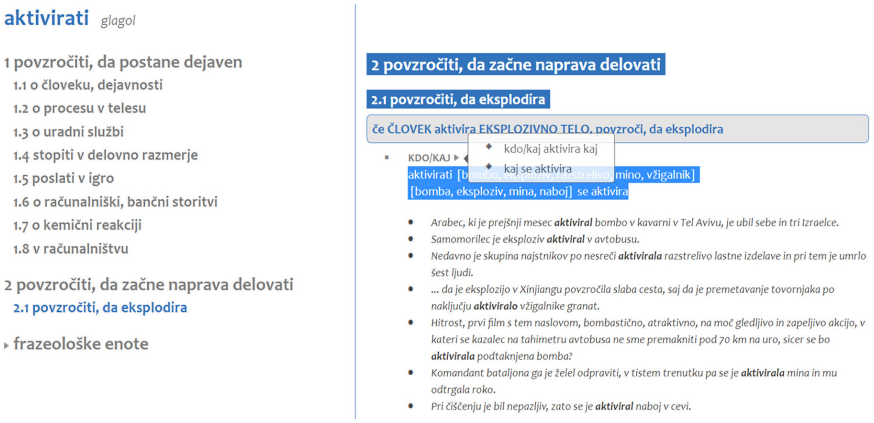
Slika 4: Prikaz vezljivostnih vzorcev za pomen glagola *aktivirati* v spletni različici Vezljivostnega slovarja slovenskih glagolov.[10]

kot bomo prikazali v primerjalni analizi v nadaljevanju, se kaže bodisi v neprepoznavanju pomena bodisi v neprepoznavanju tipičnih leksikalnih zapolnitev udeležencev v vzorcu.

2.5 Leksikalna baza za slovenščino

Poskus slovarske opredelitve vezljivostnih vzorcev za posamezne pomene glagolov je mogoče najti tudi v Leksikalni bazi za slovenščino (Gantar 2015). Prikaz vezljivostnih vzorcev v spletnem vmesniku (Slika 5) vsebuje navedbo možnih stavčnih realizacij v obliki vprašalnic, kot smo videli tudi v VSSG, pri čemer so v Leksikalni bazi navedeni samo vzorci, za katere je mogoče najti potrditve v realnih korpusnih zgledih.

10 Vir: https://fran.si/209/vezljivostni-slovar/4410086/aktivirati?FilteredDictionaryIds=218&View=1&Query=*aktivirati.



Slika 5: Prikaz vezljivostnih vzorcev za pomen glagola *aktivirati* v spletni različici Leksikalne baze za slovenščino.[11]

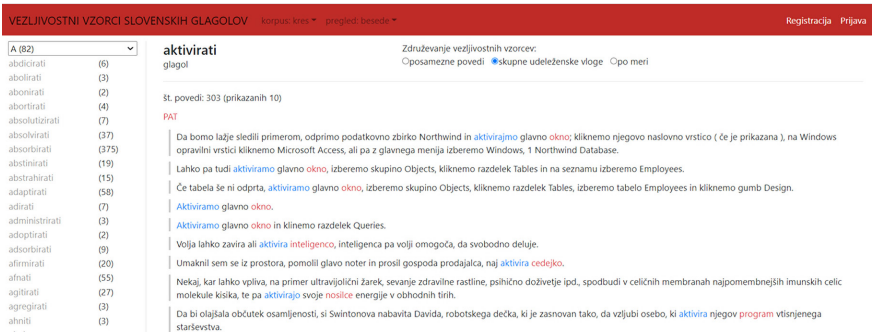
Uporabnost takega prikaza je v možnosti povezovanja udeleženskih mest, opredeljenih s semantičnimi tipi v t. i. pomenski shemi s strukturo *če*-stavka, s kolokacijami in korpusnimi zgledi. Na primer za pomen glagola *aktivirati* ‘povzročiti, da postane dejaven’, se semantični tipi (velike črke) na udeleženskih mestih v pomenski shemi: »če ČLOVEK aktivira EKSPLOZIVNO TELO, povzroči, da eksplodira«, povezujejo s kolokacijami: *aktivirati [bombo, eksploziv, razstrelivo, mino, vžigalnik]*; *[bomba, eksploziv, mina, naboj] se aktivira*. Ta možnost, kot jo predvideva tudi Pattern Dictionary of English Verbs, temelji na povezovanju predhodno definiranih skladenjskih struktur na ravni besedne zveze v na novo izdelanem Frekvenčnem seznamu kolokacij na podlagi korpusa Gigafida 2.1 (Krek et al. 2021a, Krek et al. 2021b) z udeleženskimi mesti znotraj stavčnega vzorca.

2.6 Spletni prikaz avtomatsko izluščenih vezljivostnih vzorcev iz korpusov ssj500k in Kres

Pred izdelavo Vezljivostnega leksikona, ki je predmet tega prispevka, so bili stavčni vzorci poskusno strojno izluščeni na podlagi ročno

11 Vir: http://ssj.slovenscina.eu/spletni-slovar?dictId=79&entryId=836433&key=A.

označenih udeleženskih vlog v učnem korpusu ssj500k (Krek et al. 2020b) in na podlagi uravnoteženega korpusa Kres.[12] Za prikaz vezljivostnih vzorcev v obeh korpusih je bil v okviru projekta Nova slovnica sodobne standardne slovenščine: viri in metode razvit tudi spletni vmesnik (Voje 2018), ki omogoča različne prikaze vezljivostnih vzorcev in nanje vezanih informacij, kot jih je mogoče pridobiti iz oblikoskladenjsko, skladenjsko in semantično označenega korpusa. Slika 6 prikazuje vmesnik, ki omogoča pregled vzorcev v seznamu glagolov, vezanih na posamezni korpus.



Slika 6: Spletni prikaz vezljivostnih vzorcev za glagol *aktivirati* na podlagi korpusa Kres.[12]

V prikazu *pregled besede* (index: besede) je mogoče dobiti seznam vseh glagolov z navedbo števila pojavitev v korpusu ter za vsak glagol seznam vezljivostnih vzorcev, zapisanih s pomočjo udeleženskih vlog, in z navedbo korpusnih zgledov, ki ta vzorec potrjujejo. Glagol je v vzorcu obarvan modro, prečenje miške prek udeleženske vloge pa rdeče obarva realizacijo te vloge v stavku. Vezljivostne vzorce pri posameznem glagolu je mogoče prikazovati za vsak stavek posebej ali pa združeno vse stavke, ki ustrezajo določenemu vzorcu (prikaz *skupne udeleženske vloge*). Vmesnik omogoča tudi pregledovanje vzorcev z vidika zastopanosti posamezne udeleženske vloge (index: udeleženske vloge). V tem prikazu je izhodišče seznam udeleženskih vlog skupaj s številom stavkov, ki to vlogo vsebujejo. Prikaz posamezne udeleženske vloge je enako kot v prejšnjem prikazu

12 Vir: http://www.korpus-kres.net/Support/About.
13 Vir: https://vezljivostni.cjvt.si/home/words/aktivirati#.

mogoče filtrirati glede na posamezne povedi in glede na povedi, ki pripadajo istemu vezljivostnemu vzorcu.

Vmesnik predstavlja testno verzijo spletnega prikaza, ki vključuje tudi možnost vključevanja jezikovne skupnosti pri nadgradnjah leksikona po vzoru odzivnih slovarjev (prim. Arhar Holdt et al. 2018), kot je npr. možnost pomenske razčlenitve glagola in razvrščanje vzorcev skupaj s korpusnimi zgledi pod posamezne glagolske pomene.

3 Vezljivostni Leksikon slovenskih glagolov

V okviru projekta Nova slovnica sodobne standardne slovenščine: viri in metode je bil samostojni sklop aktivnosti namenjen izdelavi računalniško berljivega Vezljivostnega leksikona slovenskih glagolov na podlagi korpusa Gigafida 2.1. Vezljivostni leksikon je zasnovan kot samostojna podatkovna baza, ki bo prek enotnega podatkovnega modela vključena v t. i. Digitalno slovarsko bazo, ta pa bo predstavljala podatkovno izhodišče za izdelavo spletnega Slovarja sodobnega slovenskega jezika (Gorjanc et al. 2015). Ključna značilnost celostne Digitalne slovarske baze je povezanost leksikalnih enot, tako eno- kot večbesednih, na podlagi njihovih skupnih in individualnih lastnosti, kot so pomen, zgradba (npr. prek enotnega sistema skladenjskih struktur), oblikoskladenjske lastnosti in korpusne reference. Trenutno so v Digitalno slovarsko bazo vključeni kolokacijski in oblikoskladenjski podatki, ki so prek samostojnih slovarskih vmesnikov (Kolokacijski slovar sodobne slovenščine;[14] Slovenski oblikoslovni leksikon Sloleks 2.0[15]) na voljo tudi uporabnikom.

Leksikon, ki vključuje vezljivostne vzorce z osnovnimi pomensko-skladenjskimi značilnostmi za najpogostejše slovenske glagole, je pod licenco CC BY-SA 4.0 dostopen na slovenskem repozitoriju CLARIN.SI (Krek et al. 2021c). Postopek luščenja vezljivostnih vzorcev je potekal avtomatsko na podlagi predhodno definiranih udeleženskih vlog za slovenščino (Gantar et al. 2018, Krek et al. 2016) in odvisnostnih skladenjskih povezav po sistemu JOS (Erjavec et al.

14 Dostopno na: https://viri.cjvt.si/kolokacije/slv/.
15 Dostopno na: https://viri.cjvt.si/sloleks/slv/.

2010a, Erjavec et al. 2010b), ki jih vključuje korpus pisne standardne slovenščine Gigafida (Krek et al. 2020a) v različici 2.1.

V nadaljevanju opišemo pripravo geslovnika, nabor udeleženskih vlog za slovenščino, formalni zapis vzorcev v Leksikonu ter avtomatsko izluščene podatke, ki jih Leksikon vsebuje v svoji prvi različici.

3.1 Priprava geslovnika

Izhodišče za pripravo geslovnika predstavlja seznam glagolov iz korpusa Gigafida 2.1 s frekvenco najmanj 3 (pribl. 22.000 lem). Ker je seznam vseboval tudi neglagolske leme, smo odstranili vse oblike, ki se ne končajo na *-ti* ali *-či*, kar je seznam zmanjšalo za približno 2.000 lem. Ta seznam smo prekrižali s seznamom glagolov, ki so del Slovenskega oblikoslovnega leksikona Sloleks 2.0 (Dobrovoljc et al. 2019), ter upoštevali presečno množico. Tej množici smo dodali glagole, ki jih vsebuje Vezljivostni slovar slovenskih glagolov (Žele 2018) ter ročno pregledan[16] ter prečiščen seznam glagolov s pojavitvijo nad 10 v Gigafidi, ki jih ni v Sloleksu ali VSSG, ter seznam glagolov s pojavitvijo med 3 in 10 v Gigafidi, ki jih ni v Sloleksu ali VSSG. S seznama smo nato izločili glagole, ki v trenutni različici leksikona Sloleks 2.0 predstavljajo šum ali odklon od standardiziranega zapisa (npr. *ščistiti, vskladiti, zavžiti* ipd.). Končni geslovnik za strojno luščenje glagolskih vzorcev vsebuje seznam 14.595 glagolov z minimalno frekvenco 3 pojavitev v korpusu Gigafida 2.1.

3.2 Nabor udeleženskih vlog

Nabor udeleženskih vlog za slovenščino (Gantar et al. 2018, Krek et al. 2016), kot prikazuje Tabela 1, temelji na naboru oznak Praške odvisnostne drevesnice PDT 2.0 (Lopatková 2003), ki je bil uporabljen pri izdelavi češkega vezljivostnega leksikona Vallex. Odločitev za češki sistem semantičnih oznak je podprta z možnostjo medjezičnega povezovanja, saj je omenjeni sistem uporabljen tudi za druge

16 Ročni pregled glagolskih lem so na podlagi navodil izdelali študenti jezikoslovnih smeri višjih letnikov ali podiplomskega študija.

Tabela 1: Seznam udeleženskih vlog v Vezljivostnem leksikonu.

Oznaka	Udel. vloga	Opis
ACT	vršilec	delujoči udeleženec, povzročitelj ali nosilec dejanja
PAT	prizadeto	prizadeti predmet dejanja
REC	prejemnik	prejemnik, posredni udeleženec dejanja; nedelovalniški udeleženec, ki mu je dejanje v škodo ali v prid; lastnosti imetnika predmeta; komunikacijska funkcija
ORIG	izvor	izhodišče, izvor/vir/povod dejanja; oseba (skupina), po kateri nekdo nekaj podeduje, posvoji, dobi
RESLT	učinek	učinek, rezultat, cilj dejanja
LOC	kraj	konkretna lokacija, kraj, mesto dejanja; smer v prostoru
SOURCE	izhodišče	začetna točka v prostoru
GOAL	cilj	končna točka v prostoru
EVENT	dogodek	časovno-prostorsko določen dogodek
TIME	čas	konkretni trenutek ali interval dejanja; trenutek ali interval, ki izvira iz dejanja; trenutek ali interval, ki sledi dejanju
DUR	trajanje	trajanje stanja; trajanje dejanja; trajanje dejanja; konkretni trenutek začetka; konkretni trenutek konca
FREQ	pogostnost	frekvenca dejanja
AIM	namen	namen dejanja; namen gibanja, premikanja
CAUSE	vzrok	vzrok dejanja
CONTR	protivnost	posledičnost ali protivnost dejanja
COND	pogojnost	pogoj za obstoj dejanja ali dogodka
REG	ozir	glede na; ključno merilo (pravilo) za ovrednotenje dejanja; primerjava
ACMP	spremstvo	predmet, oseba ali dogodek, ki spremlja dejanje ali druge udeležence
RESTR	omejitev	izjema, omejitev
MANN	način	načinovna lastnost dejanja; rezultat ob koncu dejanja
MEANS	sredstvo	sredstvo ali orodje za izvedbo dejanja
QUANT	količina	kakovostna razlika med dogodki, stanji, predmeti, mera, razpon ali intenziteta dejanja ali okoliščine
MWPRED	večbesedni predikat	zveze z nedoločniki; fazni in nemodalni glagoli
MODAL	modalna zveza	zveze modalnega glagola in nedoločnika; zveze *biti* + modalni prislov
PHRAS	frazeološka enota	odvisni del glagolske frazeološke enote

jezike,[17] prilagoditev za slovenščino pa temelji na ročni analizi glagolskih argumentov v učnem korpusu ssj500k (Krek et al. 2020b). Pri končnem naboru udeleženskih vlog smo želeli ohraniti čim večjo robustnost v številu oznak, ki ne bi predstavljala prepodrobnega pomenskega drobljenja in bi hkrati omogočala konsistentnost označevanja. Končni sistem vključuje 5 delovalniških, 17 udeleženskih in 3 oznake za udeležence znotraj glagolske zveze. Postopek avtomatskega označevanja korpusa Gigafida 2.1 z udeleženskimi vlogami, kot tudi kvantitativna evalvacija rezultatov je podrobneje opisana v Gantar et al. (2018).

Princip vezljivostnega vzorca temelji na pripisu udeleženskih vlog stavčnim udeležencem, in sicer delovalnikom (osebek, predmet) in okoliščinam (prislovna določila), ki jih določa prepozicija za dani pomen glagola. Konkretno to pomeni, da je npr. v stavku *Dodatno moč so nam dali naši navijači,* glagolu *dati* v danem pomenu mogoče pripisati 3 udeležence: tistega, ki je zavestni povzročitelj dejanja (ACT: *navijači*), tistega, ki je prizadeti predmet dejanja (PAT: *moč*), in tistega, ki je prejemnik dejanja oz. mu je dejanje v prid (REC: *nam*).

3.3 Formalni zapis vezljivostnih vzorcev

Vezljivostni vzorci so v Vezljivostnem leksikonu podani za vsak glagol v samostojni datoteki v formatu XML, ki je v iztočnici zastopan kot lema (Primer 1). Poleg identifikacijske številke, ki je pripisana vsaki lemi in je podedovana iz leksikona Sloleks (Dobrovoljc et al. 2019), sta glagolu pripisana še glagolski vid ter absolutna pogostost v korpusu Gigafida ter učnem korpusu ssj500k, če se glagol v njem pojavlja.

17 Prim. EngVallex (Cinkova et al. 2014), CzEngVallex (Urešová et al. 2015) in Crovallex (Preradović et al. 2009).

```
<head>
  <headword>
    <lemma>brskati</lemma>
  </headword>
  <lexicalUnit id="544" type="single">
    <lexeme lexical_unit_lexeme_id="544">brskati</lexeme>
  </lexicalUnit>
  <grammar>
    <category>glagol</category>
    <grammarFeature name="vid">nedovršni</grammarFeature>
  </grammar>
  <measureList>
    <measure source="Gigafida 2.0" type="frequency">9042</measure>
  </measureList>
</head>
```

Primer 1: Podatki v glavi geselskega članka za leksikonsko enoto *brskati* v Vezljivostnem leksikonu.

3.3.1 Podatki o udeleženskih vlogah

Vsakemu glagolu je v Leksikonu pripisan seznam vseh udeleženskih vlog, ki se pojavljajo v vezljivostnih vzorcih, ki jim glagol pripada. Relevantnost vsake udeleženske vloge za konkretni glagol je, kot kaže Primer 2, posredno ovrednotena z dvema frekvenčnima podatkoma: »valency_pattern_ratio« označuje odstotek vezljivostnih vzorcev glagola, v katerih je prisotna posamezna udeleženska vloga, »valency_sentence_ratio« pa označuje odstotek vseh korpusnih stavkov, ki vsebujejo konkretni glagol in udeleženske vlogo.

```
<statisticsContainer>
  <semanticRole>LOC</semanticRole>
  <measureList>
    <measure source="Gigafida 2.0" type="valency_pattern_ratio">0.5238
    </measure>
    <measure source="Gigafida 2.0" type="valency_sentence_ratio">0.8445
    </measure>
  </measureList>
</statisticsContainer>
```

Primer 2: Zapis statističnih vrednosti za pojavljanje udeleženske vloge LOC v vseh vzorcih in vseh korpusnih stavkih, ki vsebujejo glagol *brskati* v Vezljivostnem leksikonu.

Na podlagi teh podatkov je denimo za glagol *brskati* mogoče ugotoviti, da se pojavlja v vzorcih z vsemi udeleženskimi vlogami

(razen RESTR in EVENT), pri čemer se, kot kaže Tabela 2, udeleženske vloge LOC, ACT, MANN, TIME in PAT pojavljajo v največ vzorcih, ki jim ta glagol pripada, v nekoliko drugačnem zaporedju: LOC, ACT, MANN, TIME, PAT pa se te vloge pojavljajo glede na zastopanost v vseh korpusnih stavkih z glagolom *brskati*.[18] Stolpca z znakom * prikazujeta vrsti red glede na pogostnost po obeh parametrih.

Tabela 2: Statistične vrednosti za 5 najrelevantnejših udeleženskih vlog glagola *brskati* v Vezljivostnem leksikonu glede na zastopanost v vseh vezljivostnih vzorcih in vseh korpusnih stavkih, ki ta glagol vsebujejo.

Udel. vloga	valency_pattern_ratio	*	valency_sentence_ratio	*
LOC	0,5238	1	0,8445	1
TIME	0,3333	2	0,197	4
MANN	0,3258	3	0,1984	3
ACT	0,3208	4	0,2293	2
PAT	0,2531	5	0,1028	5

3.3.2 Podatki o vezljivostnih vzorcih

Podatki, ki se v Leksikonu vežejo na posamezni vezljivostni vzorec, so: identifikacijska številka vezljivostnega vzorca in število korpusnih stavkov, v katerih se glagol v določenem vezljivostnem vzorcu pojavlja. Zgradba Leksikona predvideva tudi razvrstitev vezljivostnih vzorcev z vsemi pripadajočimi podatki pod vsak potencialni pomen obravnavanega glagola, vendar v trenutni različici glagoli (še) niso pomensko razčlenjeni in posamezni pomeni niso definirani, kar je ena od prioritetnih nalog pri njegovi nadgradnji.

Za vsako udeleženso vlogo znotraj prepoznanega vezljivostnega vzorca je predviden tudi podatek o skladenjski strukturi,[19] v kateri se udejanja posamezna udeleženska vloga v vzorcu. Podatek o strukturi je primarno namenjen prepoznavanju konkretnih leksikalnih

18 V trenutni različici Leksikona vezljivostni vzorci niso razdeljeni med potencialne glagolske pomene, je pa kljub temu mogoče sklepati, da frekvenčno najpogostejši vzorci bodisi pripadajo tudi najpogostejšim pomenom oz. da frekvenčno najpogostejši vzorci tvorijo pomensko-skladenjsko okolje več glagolskim pomenom.

19 Seznam struktur v formatu XML je skupaj z Vezljivostnim leksikon dostopen na repozitoriju CLARIN.SI ter podrobneje opisan v Krek et al. (2021b).

zapolnitev, ki so za udeleženske vlogo v vzorcu značilne. Na primer za glagol *brskati* v vzorcu, ki ga tvorijo udeleženske vloge 'ACT–LOC–DUR', je značilno, da se udeleženska vloga LOC realizira s predlogi: *v, na, po, pred, pod* in *za*, kar omogoča tudi njihovo izpostavitev v korpusnem zgledu, hkrati z drugimi leksikalnimi zapolnitvami na mestu identificiranih udeleženskih vlog, kot prikazujeta primera 3 in 4.

```
<semanticRole>LOC</semanticRole>
  <syntacticStructureList>
     <syntacticStructure id="15">
        <component num="2">
        <lexeme sloleks="261">v</lexeme>
        </component>
        <component num="2">
        <lexeme sloleks="216">na</lexeme>
        </component>
        <component num="2">
        <lexeme sloleks="234">po</lexeme>
        </component>
     </syntacticStructure>
     <syntacticStructure id="16">
        <component num="2">
        <lexeme sloleks="242">pred</lexeme>
        </component>
        <component num="2">
        <lexeme sloleks="236">pod</lexeme>
        </component>
        <component num="2">
        <lexeme sloleks="276">za</lexeme>
        </component>
     </syntacticStructure>
  </syntacticStructureList>
```

Primer 3: Realizacija udeleženske vloge LOC znotraj predefiniranih skladenjskih struktur v vezljivostnem vzorcu Vezljivostnega leksikona.

```
<corpusExample corpusName="Gigafida 2.0" exampleId="GF5834751.2435.2">
<tree role="ACT">Kar ni tako neverjetno</tree>,
<tree role="ACT"><comp num="1" structure_id="70">moški</comp></tree>
<tree role="DUR"><comp num="1" structure_id="43">vedno</comp></tree>
<comp role="headword">brskajo</comp>
<tree role="LOC"><comp num="2" structure_id="15">po</comp>
<comp num="3" structure_id="15">torbicah</comp> svojih soprog</tree>.
</corpusExample>
```

Primer 4: Korpusni zgled z označenimi leksikalnimi realizacijami za posamezno udeleženske vlogo znotraj stavka v Vezljivostnem leksikonu.

Vezljivostni leksikon predvideva tudi zapis vezljivostnega vzorca v uporabniku razumljivi obliki s pomočjo vprašalnic, ki ustrezajo posamezni udeleženski vlogi v predvidenem zaporedju, kot prikazuje Tabela 3.

Tabela 3: Reprezentacijski zapis udeleženske vloge v vezljivostnem vzorcu v Vezljivostnem leksikonu.

Zaporedje v vzorcu	Udeležena vloga	Zapis v vzorcu	Zaporedje v vzorcu	Udeležena vloga	Zapis v vzorcu
1	**ACT**	KDO/KAJ	14	**ORIG**	IZVOR
2	**PAT**	KOGA/KAJ	15	**FREQ**	KOLIKOKRAT
3	**RESLT**	REZULTAT	16	**SOURCE**	OD KOD
4	**REC**	KOMU/ČEMU	17	**AIM**	S KAKŠNIM NAMENOM
5	**TIME**	KDAJ	18	**QUANT**	ŠTEVILO
6	**MANN**	KAKO	19	**EVENT**	NA DOGODKU
7	**LOC**	KJE	20	**CONTR**	KLJUB ČEMU
8	**MEANS**	S ČIM	21	**ACMP**	S KOM/ČIM
9	**GOAL**	ČEMU	22	**RESTR**	Z OMEJITVIJO
10	**REG**	GLEDE NA KOGA/KAJ	23	**MWPRED**	ne prevajamo
11	**DUR**	KOLIKO ČASA	24	**MODAL**	ne prevajamo
12	**CAUSE**	ZAKAJ	25	**PHRAS**	ne prevajamo
13	**COND**	POD KATERIM POGOJEM			

Na podlagi predvidenih vprašalnic se npr. vezljivostni vzorec 'ACT_LOC_DUR_COND' za glagol *brskati* v reprezentacijskem zapisu glasi: 'KDO/KAJ brska KJE KOLIKO ČASA POD KATERIM POGOJEM', kar se v izluščenem korpusnem zgledu uresničuje kot: *Nekateri bralci*-ACT *najbrž ne bodo nikoli*-DUR *(samo) brskali po internetu*-LOC*, ker preprosto radi kupujejo v živo*-COND.

Nekaterih udeleženskih vlog, npr. EVENT, ORIG, ni mogoče »prevesti« v ustrezno vprašalnico ali pa vloga zastopa več različnih možnih realizacij, odvisno od sobesedila – npr. udeleženska vloga RESLT zastopa tako rezultate dejanja kot tudi povedkova določila, zato v reprezentacijskem zapisu ohranjamo obliko opisa semantične vloge. Pravilo reprezentacijskega zapisa vezljivostnega vzorca tudi predvideva, da se glagol (podčrtano), če se v vzorcu mesto vršilca

realizira, ne izpiše v nedoločniku, ampak v 3. osebi ednine: 'KDO--brska-KJE-S ČIM': *uporabnik*-ACT *brska po podatkovni bazi*-LOC *s pomočjo gesel*-MEANS; 'brskati-KJE': *brskala sem po biografiji*-LOC.

4 Številčna analiza strojno izluščenih podatkov

Prva različica Vezljivostnega leksikona vsebuje vezljivostne vzorce za 14.595 glagolov, ki se pojavljajo v 25.025 različnih vezljivostnih vzorcih, ki jih tvori 25 udeleženskih vlog, vključno s potrditvami v 1.918,766 zgledih korpusov Gigafida in ssj500k.

4.1 Glagoli

Med 14.595 glagoli, ki so zastopani v Vezljivostnem leksikonu, so: *imeti, ostati, priti, dobiti, dati, igrati, predstaviti, delati, prevajati* in *videti* zastopani z največjim številom vezljivostnih vzorcev (glej Prilogo 1). Glede na zastopanost glagolov v številu korpusnih stavkov pa si med prvimi desetimi sledijo: *imeti, morati, iti, začeti, priti, dobiti, povedati, želeti, moči* in *vedeti* (glej Prilogo 2). Približno polovica glagolov ima v Leksikonu 22 ali manj oz. več različnih vzorcev, medtem ko ima 728 glagolov (npr. *zamrznjevati, tonificirati, sotrpeti, zakostenevati, včrtavati, vrtičkariti, zatogotiti, zasedlati, zbranati, zihrati, zlosati, zatrmariti, vsekniti, zarotovati, zaraskati*) naveden en sam vezljivostni vzorec.

4.2 Udeleženske vloge

Od 25 udeleženskih vlog, ki tvorijo vezljivostne vzorce, se najpogosteje glede na vse vezljivostne vzorce v korpusu pojavljajo udeleženske vloge: PAT, ACT, MANN, TIME in LOC (Tabela 4). Glede na število stavkov v korpusu Gigafida in ssj500k pa je vrstni red prvih petih nekoliko drugačen: PAT, ACT, GOAL, TIME, MANN. Stolpca z znakom * prikazujeta vrsti red glede na pogostnost po obeh parametrih. Za udeleženske vloge MWPRED, MODAL in PHRAS nimamo podatka glede na zastopanost v vseh korpusnih stavkih, zato jih pri vrstnem redu, ki sledi pogostnosti, nismo upoštevali.

Tabela 4: Zastopanost posamezne udeleženske vloge v Vezljivostnem leksikonu glede na število glagolov, pri katerih se pojavlja v vseh vzorcih in v vseh korpusnih stavkih.

Oznaka	Udel. vloga/vzorci	*	Udel. vloga/stavki	*
PAT	13.764	1	849.063	1
ACT	13.540	2	835.526	2
MANN	12.866	3	586.857	5
TIME	12.565	4	596.835	4
LOC	11.705	5	463.486	6
GOAL	10.057	6	649.804	3
MEANS	9.339	7	224.321	11
REC	9.311	8	290.240	7
CAUSE	9.305	9	263.247	8
COND	9.119	10	259.909	9
RESLT	9.075	11	198.439	13
DUR	8.501	12	249.875	10
FREQ	8.381	13	213.543	12
REG	7.422	14	156.280	14
SOURCE	7.333	15	122.132	16
ORIG	6.165	16	75.894	17
MWPRED	5.442	17	--	
AIM	4.960	18	67.725	19
MODAL	4.943	19	--	
CONTR	4.851	20	62.392	20
QUANT	4.571	21	73.470	18
ACMP	3.250	22	27.512	22
PHRAS	2.792	23	--	
EVENT	2.544	24	30.519	21
RESTR	3	25	128.816	15

Med udeleženskimi vlogami, ki izstopajo po pogostnosti pojavljanja v korpusnih stavkih, nekoliko nepričakovano izstopa GOAL. Na podlagi obstoječih smernic smo vlogo GOAL pripisovali udeležencem, ki odražajo cilj prizadevanja/dejanja, ki pa ga je mogoče razumeti tudi lokacijsko, zaradi česar predvidevamo, da obstaja nekonsistentnost že na ravni ročnega označevanja. Udeleženska vloga z vrednostjo GOAL je tudi slabše prepoznana pri natančnosti

avtomatskega pripisa udeleženskih vlog (Tabela 6). Predvidevamo tudi, da prihaja zaradi sorodnih skladenjskih realizacij do neustreznega prepoznavanja pomensko različnih udeleženskih vlog, npr. MEANS (sredstvo) in ACMP (spremstvo), kar kažejo tudi manj zanesljivi podatki pri avtomatskem označevanju (Tabela 6).

4.3 Vezljivostni vzorci

Vezljivostni leksikon vsebuje 25.025 različnih vezljivostnih vzorcev. Tabela 5 prikazuje 15 najpogostejših glede na vse vzorce v korpusu in glede na zastopanost v vseh korpusnih stavkih.

Tabela 5: 15 najpogostejših vzorcev v Vezljivostnem leksikonu glede na pogostnost vzorca in glede na število stavkov v korpusu, ki ta vzorec vsebujejo.

Vezljivostni vzorec	Pogostost vzorci	Vezljivostni vzorec	Pogostost stavki
PAT	11.803	**PAT**	14.415.799
ACT_PAT	9.886	**ACT_PAT**	9.277.371
ACT	9.690	**ACT**	5.722.425
PAT_MANN	9.386	**MODAL**	4.376.040
PAT_TIME	9.133	**PAT_TIME**	3.513.077
MANN	8.652	**PAT_MANN**	3.004.601
PAT_LOC	8.132	**ACT_PAT_TIME**	2.843.834
ACT_MANN	8.000	**RESULT**	2.798.512
ACT_PAT_MANN	7.979	**PAT_LOC**	2.245.115
ACT_PAT_TIME	7.900	**ACT_RESLT**	1.931.392
LOC	7.784	**ACT_TIME**	1.823.471
ACT_TIME	7.574	**ACT_PAT_MANN**	1.807.735
TIME	7.529	**ACT_LOC**	1.692.840
ACT_LOC	7.524	**LOC**	1.524.588
PAT_TIME_MANN	7.129	**ACT_MANN**	1.384.355

5 Jezikoslovna analiza strojno izluščenih podatkov na primeru glagola *brskati*

Ovrednotenje avtomatsko izluščenih vezljivostnih vzorcev temelji na primerjalni analizi obstoječega Vezljivostnega slovarja

slovenskih glagolov (VSSG) in na novo izdelanega Vezljivostnega leksikona (VL). V analizi puščamo ob strani različna obsega obeh virov[20] in se osredotočamo na vrsto vezljivostnih podatkov, ki ju prinašata, kot tudi na način njihovega prikaza v spletni različici VSSG oz. v formalnem zapisu VL. Za ustrezno razumevanje vrednotenjske analize je potrebno izpostaviti konceptualne razlike v zasnovi obeh virov.

VSSG temelji na jezikovnosistemski predvidljivosti glagolske vezljivosti, ki v ospredje postavlja skladenjsko izhodišče s teoretično naslonitvijo na t. i. strukturnoskladenjsko vezljivost in normativno vrednost prikazane glagolske vezljivosti, npr. na to, s katerim sklonom je ustreznejše vezati uporabljeni glagolski pomen (Žele 2008: 7). Izhodiščna teoretično-metodološka osnova za VSSG sta delo F. Daneša in sodelavcev Větné vzorce v češtině (1987) ter monografija A. Žele Vezljivost v slovenskem knjižnem jeziku (2001). Pomenska členitev glagolov pa temelji na Slovarju slovenskega knjižnega jezika. V VSSG je vezljivost prepoznana kot del jezikovnega sistema oz. kot pomensko- in strukturnoskladenjski pojav, ki vzročno-posledično povezuje pomensko, skladenjskofunkcijsko in izrazno ravnino in je v besedilu uresničevana predvsem kot vezava ali primik. V VVSG ostajata vezava kot osnovni način izražanja (desne) vezljivosti in primik kot osnovni način izražanja družljivosti tudi temeljna vidika razvrščanja vezljivostnih vzorcev, čeprav avtorica v teoretičnem modelu navaja tako »vezavnodružljive« kot »primičnovezljive« izjeme (Žele 2008: 9). Omenjeno izhodišče – ob odsotnosti semantičnih opredelitev udeležencev v VSSG – ostaja v jedru konceptualnega razlikovanja med obema primerjanima viroma. VL namreč pri razvrščanju vezljivostnih vzorcev ne izhaja iz razlikovanja med obvezno in neobvezno vezljivostjo ter družljivostjo na obliko- in funkcijskoskladenjski ravni, ampak na t. i. tektogramatični oz. pomenski ravni: na eni strani je torej mogoče govoriti o izraženosti oz. neizraženosti udeležencev, na drugi pa o pomenski obveznosti oz. neobveznosti udeleženskih

20 VSSG vezljivostno analizira 2.591 glagolov kot slovarskih gesel (2.061 glavnih izhodiščnih gesel in 530 kazalčnih gesel); VL vključuje 14.595 glagolov in skupno 25.025 različnih vezljivostnih vzorcev.

mest: tako je denimo v vezljivostnih vzorcih VL obveznost neizraženega udeleženskega mesta mogoče razbrati iz njegove »zunanje« prisotnosti, npr. pri vršilcih z osebno glagolsko obliko glagola, splošnih vršilcih, povratnosvojilnih strukturah ipd. Tako izhodišče sledi teoretični podstavi, uporabljeni v češkem Vallexu, ki udeležence razvršča glede na obveznost, opcijskost ali fakultativnost.

Druga, že omenjena razlika med obema viroma je opredelitev udeležencev z naborom semantičnih oznak v VL. Te poleg semantičnih lastnosti udeležencev kažejo tudi na omejene skladenjske možnosti izražanja različnih semantičnih vlog: konkretno se npr. izražanje načina (MANN) in izražanje sredstva (MEANS) lahko skladenjsko izraža z istim skladenjskim inventarjem, npr. s predložno zvezo: *brskati z radovednostjo* (MANN) : *brskati s palico* (MEANS).

Za jezikoslovno evalvacijo strojno izluščenih vezljivostnih podatkov smo izbrali glagol *brskati*, ki ima v korpusu Gigafida 9.042 pojavitev, predvideva več udeleženskih mest in je vključen tudi v Vezljivostni slovar slovenskih glagolov.[21] Za ustrezno vrednotenje primerjalne analize je treba na strani VL upoštevati še stopnjo natančnosti avtomatskega označevanja korpusa Gigafida z udeleženskimi vlogami (Tabela 6), na strani VSSG pa dejstvo, da je izdelan ročno in da vključuje tako pomensko razčlenitev kot tudi pomenske definicije, kar omogoča razvrstitev vezljivostnih vzorcev pod glagolske pomene. Primerjalna analiza je bila izvedena ročno, in sicer so bili v VSSG upoštevani le v slovarju navedeni zgledi, pri VL pa smo upoštevali večje število korpusnih realizacij, tj. stavkov, ki so bili za posamezni vzorec dejansko izluščeni iz korpusa, vendar jih zaradi preobsežnosti v Leksikon nismo vključili.[22]

Ocena natančnosti avtomatskega označevanja udeležencev v korpusnih stavkih s pomočjo orodja mate-tool (Björkelund et al. 2009) je bila s kvantitativnega vidika opravljena na korpusu ssj500k (Gantar et al. 2018). Avtomatsko označeni podatki za posamezno

21 VSSG vključuje tudi glagole (npr. *babiti se, beleti, laizirati se* ipd.), ki v korpusu Gigafida nimajo pojavitve.

22 Kot omenjeno, smo v korpus vključili le po en primer izluščenih stavkov iz korpusov ssj500k in Gigafida za vsak vezljivostni vzorec.

udeleženskо vlogo so bili nato primerjani z metriko F1,[23] kot prikazuje Tabela 6.

Tabela 6: Natančnost avtomatskega pripisa udeleženske vloge v korpusu ssj500k (povzeto po Gantar et al. 2018).

Udel. vloga	F1	Udel. vloga	F1	Udel. vloga	F1	Udel. vloga	F1	Udel. vloga	F1
ACT	0,94	**MANN**	0,76	**LOC**	0,59	**SOURCE**	0,37	**ORIG**	0,24
MWPRED	0,91	**REC**	0,74	**FREQ**	0,59	**CAUSE**	0,35	**AIM**	0,2
MODAL	0,9	**MEANS**	0,64	**GOAL**	0,53	**REG**	0,34	**CONTR**	0,14
PAT	0,88	**TIME**	0,62	**DUR**	0,5	**PHRAS**	0,31	**ACMP**	0,08
RESLT	0,8	**QUANT**	0,62	**COND**	0,46	**EVENT**	0,29	**REST**	0

Po pričakovanju je avtomatski pripis udeleženske vloge natančnejši pri udeleženskih vlogah, ki se v korpusu pojavljajo najpogosteje (F1 = <0,5): ACT, MWPRED, MODAL, PAT, RESLT, MANN, REC, MEANS, TIME, QUANT, LOC, FREQ, GOAL in DUR, manj natančen je avtomatski pripis pri manj pogostih udeleženskih vlogah, kot so COND, SOURCE, CAUSE, REG, PHRAS, EVENT, ORIG, AIM, CONTR, ACMP in REST, kar je treba upoštevati tudi pri analizi jezikoslovne ustreznosti izluščenih podatkov glede na ročno obdelavo vezljivostnih vzorcev v VSSG.

V **Vezljivostnem slovarju slovenskih glagolov** se glagol *brskati* pojavlja v dveh pomenih, in sicer: 1. ‘razkopavati’ in 2. ‘stikati’. Za oba pomena so navedeni enaki vezljivostni vzorci, z izjemo dodatnega vezljivostnega vzorca ‘KDO/KAJ brska za ČIM’ pri 2. pomenu, razporejeni glede na obvezno oz. neobvezno vezljivost ter glede na družljivost.

Obvezna vezljivost predvideva zastopanost vršilca dejanja ter lokacijo, ki se izraža s predložnimi samostalniki v mestniku ali s prislovi kraja:

- KDO/KAJ brska v KOM/ČEM KJE/KOD
- KDO/KAJ brska pri KOM/ČEM KJE/KOD
- KDO/KAJ brska po KOM/ČEM KJE/KOD
- KDO/KAJ brska ob KOM/ČEM KJE/KOD

23 Mera F1 se uporablja za ocenjevanje klasifikacijske točnosti na podlagi harmonične sredine preciznosti in priklica.

Neobvezna vezljivost predvideva prisotnost sredstva, ki se uresničuje s predložnim samostalnikom v orodniku:

- KDO/KAJ brska s ČIM
- KDO/kaj brska za ČIM

Družljivost pa predvideva izražanje lokacije, sredstva in načina s predložnimi samostalniki v tožilniku ali orodniku, s stavčno povedjo ali prislovom načina:

- KDO/KAJ brska na KAJ
- KDO/KAJ brska s ČIM
- KDO/KAJ brska KAKO

Različne realizacijske možnosti – pri čemer zgledi ne potrjujejo vseh izpostavljenih predlogov v vzorcu – so ponazorjene s tremi zgledi za vsak pomen. Neobvezna vezljivost in družljivost sta, predvidevamo, v zgledih nakazani s poševnico oz. oklepaji:

1. razkopavati
    - *Kokoši /s kremplji/ razkopavajo*[24] *po gnoju.*
    - *Otrok (s palico) brska po pesku.*
    - */S prstom/ je brskal po nosu.*
2. stikati
    - *(Za pomembnimi listinami) je brskala po tujih predalih.*
    - *(Za določenimi besedami) je brskal po slovarjih.*
    - preneseno *Rada je /z vsiljivo radovednostjo/ brskala po tujih življenjih.*
    - čustvenostno *Vse življenje brska po knjigah.*

V **Vezljivostnem leksikonu** se glagol *brskati* pojavlja v 399 različnih vzorcih, pri čemer se 336 vzorcev v korpusu pojavi manj kot 10-krat, kar 179 vzorcev pa se v korpusu pojavi zgolj enkrat. Za analizo vezljivosti tega glagola so tako zanimivi predvsem vzorci, ki se v korpusu pojavljajo več kot 100-krat. Razvrstitev vzorcev po pogostnosti je skupaj z reprezentacijskim zapisom in korpusnim zgledom prikazana v Tabeli 7.

24 Iz zgleda ni jasno, ali je uporaba sinonimnega glagola (*razkopavati*) namesto obravnavanega *brskati* namenska ali napaka.

Tabela 7: Najpogostejši vezljivostni vzorci za glagol *brskati* v Vezljivostnem leksikonu.

Vezljivostni vzorec	Pogostost/ korpus	Reprezentacijski zapis	Realizacija
LOC	3.122	brskati KJE	brskati po biografiji
ACT_LOC	751	KDO/KAJ brska KJE	ljudje brskajo po smetnjakih
MANN_LOC	654	brskati KAKO KJE	rad brska po vrtu
TIME_LOC	568	brskati KDAJ KJE	medtem je brskal po telefonu
ACT_MANN_LOC	223	KDO/KAJ brska KAKO KJE	mediji mrzlično brskajo po preteklosti
ACT_TIME_LOC	207	KDO/KAJ brska KDAJ KJE	medtem najstnice vneto brskajo po trgovinah
PAT_LOC	179	brskati KOGA/KAJ KJE	brskati po spominu za številom
PAT	158	brskati KOGA/KAJ	brskati za podrobnostmi
TIME	146	brskati KDAJ	brskati dalje
LOC_DUR	145	brskati KJE KOLIKO ČASA	od sedaj naprej brskati
TIME_MANN_LOC	117	brskati KDAJ KAKO KJE	vedno rad brska po internetu

Podatek o zastopanosti posamezne udeleženske vloge v vezljivostnih vzorcih glagola *brskati* skupaj z najbolj tipičnimi vezljivostnimi vzorci v Tabeli 8 podaja posredno tudi informacijo o obveznosti udeleženskih vlog. Pri razumevanju obveznosti je treba udeleženske vlogo vršilca (ACT) v vzorcih, kot so predstavljeni v VL, upoštevati tudi njegovo izraženost oz. neizraženost: za glagol *brskati* je tako semantična prisotnost vršilca pomensko obvezna, vendar ne nujno tudi izražena.

Iz predstavljenih podatkov v tabelah 7 in 8 je mogoče zaključiti, da so relevantni vzorci za glagol *brskati*, (tj. tisti, ki sodijo med najpogostejše), kot jih prinaša VL, deloma prekrivni z vzorci v VSSG. V Tabeli 9 so upoštevani vsi vzorci v VSSG in samo tisti v VL, ki se pojavljajo za konkretni glagol več kot 90-krat. Vzorce smo v obeh virih razdelili na predvidene pomenske sklope, kar nam je omogočilo primerjavo. Ker se lahko različne udeleženske vloge realizirajo z enakimi skladenjskimi možnostmi, smo pod »lokacijske« realizacije

v VSSG uvrstili vse izpostavljene predložne možnosti (*v, na, pri, po* in *ob*) in realizacije s krajevnimi prislovi (*kje, kod*). Kot pomenske vloge »sredstva« smo upoštevali realizacije s predlogom *s*, ki smo jih ponovili tudi pri pomenski vlogi »način«.

Tabela 8: Razvrstitev udeleženskih vlog glagola *brskati* v Vezljivostnem leksikonu glede na zastopanost v vzorcih in vseh korpusnih stavkih.

Udel. vloga	Udel. vloga/vzorec	Udel. vloga/korpus
LOC	0,5238	0,8445
TIME	0,3333	0,2293
MANN	0,3258	0,1984
ACT	0,3208	0,1970
PAT	0,2531	0,1028
DUR	0,1855	0,0570
REC	0,1579	0,0287
COND	0,1554	0.0277
GOAL	0,1253	0,0248
FREQ	0,1078	0,0246
CAUSE	0,1003	0,0181
MEANS	0,0827	0,0175
REG	0,0827	0,0096
MWPRED	0,0501	0,0059
SOURCE	0,0501	0,0056
AIM	0,0301	0,0055
QUANT	0,0276	0,0023
MODAL	0,0251	0,0022
ACMP	0,0201	0,0022
CONTR	0,0175	0,0013
ORIG	0,015	0,0012
PHRAS	0,0025	0,0008
RESLT	0,0022	0,0001

Tabela 9: Primerjava vezljivostnih vzorcev v VSSG in ustreznih najpogostejših v VL.

Pomenska vloga udeležencev	VSSG	VL
LOKACIJA	KDO/KAJ brska v KOM/ČEM KJE/KOD	brskati KJE
	KDO/KAJ brska na KOM/ČEM KJE/KOD	KDO/KAJ brska KJE
	KDO/KAJ brska pri KOM/ČEM KJE/KOD	brskati KAKO KJE
	KDO/KAJ brska po KOM/ČEM KJE/KOD	brskati KDAJ KJE
	KDO/KAJ brska ob KOM/ČEM KJE/KOD	KDO/KAJ brska KAKO KJE
	KDO/KAJ brska na KAJ	KDO/KAJ brska KDAJ KJE
		brskati KOGA/KAJ KJE
		brskati KJE KOLIKO ČASA
		brskati KDAJ KAKO KJE
		brskati KJE S ČIM
SREDSTVO	brskati s ČIM	brskati KJE S ČIM[25]
NAČIN	brskati na KAJ	brskati KAKO KJE
	brskati s ČIM	KDO/KAJ brska KAKO KJE
	brskati KAKO	brskati KDAJ KAKO KJE
PRIZADETO	brskati za KOM/ČIM	brskati KOGA/KAJ KJE
		brskati KOGA/KAJ
ČAS/TRAJANJE		brskati KDAJ KJE
		KDO/KAJ brska KDAJ KJE
		brskati KDAJ
		brskati KDAJ KAKO KJE
		brskati KJE KOLIKO ČASA

Udeleženske vloge, ki se v zvezi z glagolom *brskati* najpogosteje pojavljajo v vezljivostnih vzorcih VL, so LOC, TIME, MANN, PAT, ACT in DUR v navedenem zaporedju (Tabela 10). Če jih prepišemo v pomensko-skladenjske realizacije, kot jih izkazuje VSSG, in jih primerjalno ovrednotimo glede na obveznost, lahko vidimo, da sta znotraj obvezne vezljivosti prepoznana predvsem vršilec in lokacija, ki sta na prvem mestu tudi v VL, pri čemer je iz VL še razvidno, da se vršilec v stavku pogosto ne realizira. Udeleženska vloga PAT je prepoznana

25 Vezljivostni vzorci, ki vsebujejo udeleženske vlogo MEANS (sredstvo) oz. se realizirajo s predložnim samostalnikom v orodniku, so izkazani tudi v VL, vendar se ne uvrščajo med 11 najpogostejših vzorcev za ta glagol.

kot neobvezna vezljivost, MANN (način), ki se v VL uvršča visoko na seznamu najrelevantnejših udeleženskih vlog, pa je v VSSG prepoznan le v okviru družljivosti.

Tabela 10: Udeleženske vloge za glagol brskati v VL glede na pripisano obveznost v VSSG.

Udeleženska vloga VL	Obveznost VSSG
LOC	obvezna vezljivost
TIME	ni izpričana
MANN	družljivost/neobvezna vezljivost
ACT	obvezna vezljivost
PAT	neobvezna vezljivost
DUR	ni izpričana
(MEANS)[26]	neobvezna vezljivost

Razlika med obema viroma se kaže predvsem v umanjkanju vezljivostnih vzorcev v VSSG z udeleženso vlogo TIME (čas) in DUR (trajanje), ki v VSSG niso izkazani niti v okviru družljivosti, in v izpostavljeni neobvezni vezljivosti v VSSG, ki predvideva bodisi sredstvo (MEANS), ki se v VL ne izkazuje med 5 najpogostejšimi vzorci, tega glagola (glej tudi Tabelo 9), bodisi MANN (način), ki je sicer nakazan tudi z realizacijo v prislovu (*kako*). Distribucija vzorcev z izraženim sredstvom med drugim vzbuja pomisleke o ustrezni razdelitvi vezljivostnih vzorcev v VSSG pod oba pomena, saj je ob pregledu zgledov, ki smo jih za vzorce z udeleženso vlogo MEANS (sredstvo) izluščili iz korpusa Gigafida, mogoče ugotoviti, da so vezani predvsem na pomen 'razkopavati': *Kokoši /s kremplji/ razkopavajo po gnoju*, ne pa tudi na pomen 'stikati',[27] ki je v VSSG ponazorjen s kvalifikatorjem preneseno: *Rada je /z vsiljivo radovednostjo/ brskala po tujih življenjih*. Iz zgleda je tudi razvidno, da predložni predmet v orodniku ne pokriva vloge sredstva, kot smo na podlagi vprašalnice predvidevali, pač pa način, ki se, kot izpostavljata oba vira, pojavlja kot tipična udeleženska vloga v vezljivostnih vzorcih tega glagola.

26 Ni med najpogostejšimi 5 udeleženskimi vlogami.

27 Pregled izluščenih zgledov kaže, da bi bila pomenska opredelitev 'iskati' ali 'pridobivati podatke, informacije' ustreznejša, saj pomen vključuje zelo pogoste realizacije v sodobni pisni slovenščini, kot npr.: *brskati po spletu/internetu, brskati po preteklosti/spominu/arhivu, brskati za informacijami, brskati med knjigami, brskati v službi* ipd.

Primerjava v načinu prikaza in vrsti vezljivostnih podatkov v obeh virih je potrdila predvsem razliko v njuni konceptualni zasnovi. Vezljivostni vzorci se v VSSG osredotočajo na izražanje skladenjsko-pomenske obveznosti udeležencev, ki niso pomensko opredeljeni z udeleženskimi vlogami, ampak jih določajo oblikoslovne kategorije, kot sta besedna vrsta in sklon. Osredotočanje VSSG na opredeljevanje vezljivostne obveznosti z ločevanjem med obvezno in neobvezno vezljivostjo na eni in družljivostjo na drugi strani se zdi z vidika uporabnosti vzorcev za strojno procesiranje in za namene semantičnih analiz manj pomembna informacija, vsekakor pa bi bilo njeno vrednost smiselno preveriti tudi pri slovarskih uporabnikih. VL na drugi strani v ospredje postavlja prepoznavanje tipičnosti glagolskega vzorca in udeleženske vloge v njem s prikazom realnih in hkrati tipičnih leksikalnih realizacij na udeleženskih mestih, ki so opredeljena tudi z mednarodno uporabljanimi semantičnimi oznakami. Zadnje je pomembno tudi z vidika opisa realnega jezikovnega stanja. Če na eni strani VSSG izpostavlja sistemske možnosti skladenjskih realizacij, za katere zgledi ne kažejo nujno tudi realne potrditve, je v VL v ospredju semantična opredelitev udeležencev, iz izluščenih zgledov in z izpostavitvijo najbolj produktivnih realizacij v njem pa je na voljo tudi podatek o tipičnih skladenjskih uresničitvah pomenskih lastnosti udeležencev. Ob tem je treba poudariti, da vključitev zgolj dveh stavkov za vsak vzorec iz korpusa Gigafida v Leksikon tega podatka uporabnikom leksikona ne ponuja neposredno, zato je VL v svoji prvi različici namenjen predvsem izboljšavi strojnega luščenja vezljivostnih vzorcev, neposredna slovarska uporabnost Leksikona pa bo mogoča šele z vključitvijo v Digitalno slovarsko bazo z možnostjo povezovanja kolokacijskih podatkov in identificiranih udeleženskih mest. Pomanjkljivost VL ostaja v nenatančnosti avtomatskega prepoznavanja zlasti manj pogostih udeleženskih vlog. Tako denimo že omenjena predložna zveza »s kom/čim« predstavlja problem za ustrezno ločevanje med sredstvom in načinom tudi za avtomatski model. Podrobnejša analiza izluščenih stavkov s to udeleženso vlogo razkriva predvsem napačne pripise udeleženske vloge sredstva (podčrtano), npr. *Slovenski smučarji*-ACT *najbrž te dni*-TIME *<u>z zavistjo</u>*-MEANS *brskajo po*

spletnih straneh švicarske zveze-LOC, natančnejši pa je pripis udeleženske vloge načina (podčrtano): *Domačini*-ACT *kljub temu*-CONTR *vztrajno*-MANN *brskajo med naplavljenimi predmeti*-LOC. Poleg tega avtomatski sistem ne uspe vedno ustrezno prepoznati prisotnosti udeleženske vloge, npr. v primeru *Medtem ko zgodovinarji in politiki*-ACT *brskajo po arhivih*-LOC, *da bi dokazovali kod bi postavili mejne kamne na slovensko-hrvaški meji*-AIM /.../, ni prepoznana udeleženska vloga TIME (podčrtano), kar posledično vpliva tudi na število in zapis vezljivostnih vzorcev v Leksikonu. Omenjeni slabosti bo, predvidevamo, mogoče izboljšati s povečanjem učne množice ročno označenih vezljivostnih vzorcev v prihodnjih nadgradnjah.

6 Zaključek in smernice za nadaljnje delo

V prispevku opisani avtomatsko izdelani Vezljivostni leksikon predstavlja tako v količini vključenih podatkov kot v njihovi relevantnosti dobro izhodišče za sodoben opis vezljivosti slovenskih glagolov. Uporabnost formaliziranega opisa je predvsem v njegovi strojni berljivosti, s čimer so omogočene nadaljnje jezikoslovne raziskave in uporaba v jezikovnotehnoloških nalogah, uporaba Leksikona za slovarske namene pa potrebuje nadaljnje izboljšave. Uporabnost Leksikona glede na obstoječe vezljivostne vire se kaže med drugim tudi v naslonitvi na splošno upoštevane dobre prakse v smislu medjezikovne kompatibilnosti uporabljenih oznak in skladenjskih razmerij med udeleženci, ki se vedno potrjujejo tudi z realnimi korpusnimi zgledi. Ugotovitve, ki jih je prinesla jezikoslovna ocena izluščenih podatkov tudi v primerjavi z obstoječim ročno izdelanim Vezljivostnim slovarjem, bo mogoče uporabiti za izboljšanje metodologije pri nadaljnjih luščenjih. Tu imamo v mislih izboljšavo mehanizma za prepoznavanje kolokacijskih in drugih tipičnih zapolnitev udeleženskih mest, kjer ostajajo odprte tudi možnosti opredeljevanja leksikalnih realizacij udeleženskih položajev s semantičnimi tipi na podlagi katere od že omenjenih semantičnih ontologij.

Med nalogami, ki jih v prihodnje predvideva nadgradnja Vezljivostnega leksikona, je izboljšanje ročnega označevanja na podlagi

novih smernic, ki bodo upoštevale ugotovitve dosedanjih evalvacij (prim. Gantar v tisku), ter povečanje obsega ročno označenih stavkov v učnem korpusu, kot je predvideno v okviru projekta Razvoj slovenščine v digitalnem okolju.[28] Nadalje ostaja ena od prioritet pomenska razčlenitev in pomenski opis glagolov na podlagi sodobnih korpusnih podatkov ter izdelava spletnega vmesnika za pregledovanje vezljivostnih vzorcev po vzoru odzivnih slovarjev z možnostjo vključevanja jezikovne skupnosti.

Zahvala

Prispevek je nastal v okviru raziskovalnega projekta Nova slovnica sodobne standardne slovenščine: viri in metode (J6-8256) ter v okviru programske skupine Slovenski jezik – bazične, kontrastivne in aplikativne raziskave (P6-0215), ki ju financira Agencija za raziskovalno dejavnost Republike Slovenije.

Reference

Arhar Holdt, Š., Čibej, J., Dobrovoljc, K., Gantar, A., Gorjanc, V., Klemenc, B., Kosem, I., Krek, S., Laskowski, C. in Robnik Šikonja, M. (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. 401–411). Ljubljana: Znanstvena založba Filozofske fakultete. Dostopno prek: https://e-knjige.ff.uni-lj.si/znanstvena-zalozba/catalog/view/118/211/3000-1.

Björkelund, A., Hafdell, L. in Nugues, P. (2009). Multilingual semantic role labeling. V J. Hajič (ur.), *Proceedings of the Thirteenth Conference on Computational Natural Language Learning (CoNLL): Shared Task* (str. 43–48). Association for Computational Linguistics. Dostopno prek: https://aclanthology.org/W09-1206.pdf.

Cinková, S., Fučíková, E., Šindlerová, J. in Hajič, J. (2014). EngVallex: English Valency Lexicon, LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL). Faculty of Mathematics and Physics, Charles University. http://hdl.handle.net/11858/00-097C-0000-0023-4337-2.

28 Spletna stran projekta: https://www.slovenscina.eu/.

Daneš, F. in Hlavsa. Z. (1987). *Větné vzorce v češtině*. Praga: Academia.
Dobrovoljc, K., Krek, S., Holozan, P., Erjavec, T., Romih, M., Arhar Holdt, Š., Čibej, J., Krsnik, L. in Robnik-Šikonja, M. (2019). Morphological lexicon Sloleks 2.0, Slovenian language resource repository CLARIN.SI. http://hdl.handle.net/11356/1230.
Erjavec, T., Fišer, D., Krek, S. in Ledinek, N. (2010a). The JOS Linguistically Tagged Corpus of Slovene. V N. Calzolari, K. Choukri, B. Maegaard, J. Mariani, J. Odijk, S. Piperidis, M. Rosner in D. Tapias (ur.), *LREC 2010: Proceedings of the Seventh International Conference on Language Resources and Evaluation* (str. 1806–1809). European Language Resources Association. Dostopno prek: http://www.lrec-conf.org/proceedings/lrec2010/pdf/139_Paper.pdf.
Erjavec, T., Krek, S., Arhar, Š., Fišer, D., Ledinek, N., Saksida, A., Sivec, B. in Trebar, B. (2010b). Oblikoskladenjske specifikacije JOS V1. Dostopno prek: http://nl.ijs.si/jos/msd/html-sl/index.html.
Fillmore, C. J. (1976). Frame semantics and the nature of language. V S. R. Harnad, H. D. Steklis in J. Lancaster (ur.), Origin and Development of Language and Speech. *Annals of the New York Academy of Sciences*, 280 (1), 20–32. New York: New York Academy of Sciences. https://doi.org/10.1111/j.1749-6632.1976.tb25467.x.
Fillmore, C. J., Johnson, R. J., Petruck in M. R. L. (2003). Background to Framenet. *International Journal of Lexicography*, 16 (3), 235–250. https://doi.org/10.1093/ijl/16.3.235.
Gantar, P. (2015). *Leksikografski opis slovenščine v digitalnem okolju*. Ljubljana: Znanstvena založba Filozofske fakultete. Dostopno prek: https://e-knjige.ff.uni-lj.si/znanstvena-zalozba/catalog/view/62/138/2602-1.
Gantar, P. (v tisku). Analiza udeleženskih vlog s skladenjskega, pomenskega in leksikalnega vidika. V M. Smolej in M. Schlambergar (ur.), *Zbornik prispevkov s Simpozija o skladnji*. Ljubljana: Znanstvena založba Filozofske fakultete.
Gantar, P., Štrkalj Despot, K., Krek, S. in Ljubešić, N. (2018). Towards semantic role labeling in Slovene and Croatian. V D. Fišer in A. Pančur (ur.), *Zbornik konference Jezikovne tehnologije in digitalna humanistika* (str. 93–98). Ljubljana: Znanstvena založba Filozofske fakultete. Dostopno prek: http://nl.ijs.si/jtdh18/JTDH-2018-Proceedings.pdf.
Gorjanc, V., Gantar, P., Kosem, I. in Krek, S. (ur.) (2015). *Slovar sodobne slovenščine: problemi in rešitve*. Ljubljana: Znanstvena založba Filozofske

fakultete. E-izdaja (2017). Ljubljana: Znanstvena založba Filozofske fakultete. https://doi.org/10.4312/9789612379759.

Kettnerová, V., Lopatková, M. in Bejček, E. (2012). The Syntax-Semantics Interface of Czech Verbs in the Valency Lexicon. V R. Vatvedt Fjeld in J. M. Torjusen (ur.), *Proceedings of the 15th EURALEX International Congress* (str. 434– 443). Department of Linguistics and Scandinavian Studies, University of Oslo. Dostopno prek: https://euralex.org/wp-content/themes/euralex/proceedings/Euralex%202012/pp434-443%20Kettnerova,%20Lopatkova%20and%20Bejcek.pdf.

Hanks, P. (1994). Linguistic Norms and Pragmatic Exploitations or Why Lexicographers Need Prototype Theory and Vice Versa. V F. Keifer, G. Kiss in J. Pajzs (ur.), *Papers in Computational Lexicography. Complex '94*, 89–113.

Hanks, P. (2004). Corpus Pattern Analysis. V G. Williams in S. Vessier (ur.), *Proceedings of the 11th EURALEX International Congress* (str. 87–97). Faculté des lettres et des sciences humaines, Université de Bretagne-Sud. Dostopno prek: https://euralex.org/wp-content/themes/euralex/proceedings/Euralex%202004/009_2004_V1_Patrick%20HANKS_Corpus%20pattern%20analysis.pdf.

Hanks, P. (2008, Marec 15). *Mapping meaning onto use: a Pattern Dictionary of English Verbs* [predstavitev na konferenci]. AACL 2008: American Association for Corpus Linguistics, Provo, Utah, ZDA. Dostopno prek: https://nlp.fi.muni.cz/projects/cpa/.

Hanks, P. (2013). *Lexical Analysis: Norms and Exploitations*. Cambridge: MIT Press.

Hanks, P. in Pustejovsky, J. (2004). Common Sense About Word Meaning: Sense in Context. V P. Sojka, I. Kopeček in K. Pala (ur.), *Text, Speech and Dialogue: proceedings* (Lecture Notes in Computer Science, vol. 3206) (str. 15–17). Berlin; Heidelberg: Springer. https://doi.org/10.1007/978-3-540-30120-2_2.

Hanks, P. in Pustejovsky, J. (2005). A Pattern Dictionary for Natural Language Processing. *Revue Française de Linguistique Appliquée*, 10 (2), 63–82. https://doi.org/10.3917/rfla.102.82.

Krek, S., Gantar, P., Dobrovoljc, K. in Škrjanec, I. (2016). Označevanje udeleženskih vlog v učnem korpusu za slovenščino. V T. Erjavec in D. Fišer (ur.), Zbornik konference Jezikovne tehnologije in digitalna humanistika (str. 106–110). Ljubljana: Znanstvena založba Filozofske fakultete.

Dostopno prek: http://www.sdjt.si/wp/wp-content/uploads/2016/09/JTDH-2016_Krek-et-al_Oznacevanje-udelezenskih-vlog-v-ucnem--korpusu-za-slovenscino.pdf.

Krek, S., Arhar Holdt, Š., Erjavec, T., Čibej, J., Repar, A., Gantar, P., Ljubešić, N., Kosem, I. in Dobrovoljc, K. (2020a). Gigafida 2.0: the reference corpus of written standard Slovene. V N. Calzolari (ur.), *LREC 2020: Twelfth International Conference on Language Resources and Evaluation*: *conference proceedings* (str. 3340–3345). Pariz: European Language Resources Association. Dostopno prek: http://www.lrec-conf.org/proceedings/lrec2020/LREC-2020.pdf.

Krek, S., Erjavec, T., Dobrovoljc, K., Gantar, P., Arhar Holdt, Š., Čibej, J. in Brank, J. (2020b). The ssj500k Training Corpus for Slovene Language Processing. V D. Fišer in T. Erjavec (ur.), *Jezikovne tehnologije in digitalna humanistika: zbornik konference* (str. 24–33). Ljubljana: Inštitut za novejšo zgodovino. Dostopno prek: http://nl.ijs.si/jtdh20/pdf/JT-DH_2020_Krek-et-al_The-ssj500k-Training-Corpus-for-Slovene--Language-Processing.pdf.

Krek, S., Gantar, P., Kosem, I., Dobrovoljc, K., Arhar Holdt, Š., Čibej, J., Laskowski, C. A., Klemenc, B. in Krsnik, L. (2021a). Frequency lists of collocations from the Gigafida 2.1 corpus, Slovenian language resource repository CLARIN.SI. http://hdl.handle.net/11356/1415.

Krek, S., Gantar, P., Kosem, I. in Dobrovoljc, K. (2021b). Opis modela za pridobivanje in strukturiranje kolokacijskih podatkov iz korpusa. V Š. Arhar Holdt (ur.), *Nova slovnica sodobne standardne slovenščine: viri in metode* (str. 160–197). Ljubljana: Znanstvena založba Filozofske fakultete.

Krek, S., Gantar, P., Krsnik, L., Laskowski, C., Dobrovoljc, K., Arhar Holdt, Š., Čibej, J., Kosem, I., Klemenc, B., Robnik-Šikonja, M. in Gorjanc, V. (2021c). Valency lexicon extracted from the Gigafida 2.1 corpus, Slovenian language resource repository CLARIN.SI. http://hdl.handle.net/11356/1418.

Lopatková, M. (2003). Valency in the Prague Dependency Treebank: Building the Valency Lexicon. *The Prague Bulletin of Mathematical Linguistics,* 79–80, 37–60. Dostopno prek: http://ufal.mff.cuni.cz/pbml/79-80/lopatkova.pdf.

Lopatková, M., Kettnerová, V., Bejček, E., Vernerová, A. in Žabokrtský, Z. (2016). *Valenční slovník českých sloves VALLEX*. Praga: Karolinum.

Mikelić Preradović, N., Boras, D. in Kisicek, S. (2009). CROVALLEX: Croatian verb valence lexicon. V V. Luzar-Stiffler, I. Jarec in Z. Bekic (ur.), *Proceedings of the ITI 2009 31st International Conference on information technology interfaces* (str. 533–538). https://doi.org/10.1109/ITI.2009.5196142.

Može, S. (2009). Semantično označevanje korpusa slovenščine po modelu FrameNet. V M. Stabej (ur.), *Infrastruktura slovenščine in slovenistike, Obdobja 28* (str. 265–269). Ljubljana: Znanstvena založba Filozofske fakultete in Center za slovenščino kot drugi/tuji jezik. Dostopno prek: https://centerslo.si/wp-content/uploads/2015/10/28-Moze.pdf.

Urešová, Z. Fučíková, E., Hajič, J. in Šindlerová, J. (2015). CzEngVallex: LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL). Faculty of Mathematics and Physics, Charles University. http://hdl.handle.net/11234/1-1512.

Vezljivostni slovar slovenskih glagolov, druga, dopolnjena spletna izdaja. Dostopno prek: www.fran.si.

Voje, K. (2018). *Avtomatska izdelava vezljivostnih vzorcev za slovenske glagole*. Diplomsko delo. Univerza v Ljubljani, Fakulteta za računalništvo in informatiko. Dostopno prek: https://repozitorij.uni-lj.si/IzpisGradiva.php?id=106000&lang=slv.

Žele, A. (2001). *Vezljivost v slovenskem knjižnem jeziku: S poudarkom na glagolu*. Ljubljana: Založba ZRC.

Žele, A. (2008). *Vezljivostni slovar slovenskih glagolov*. Ljubljana: Založba ZRC.

Priloga 1: Seznam 50 glagolov z največjim številom vzorcev v Vezljivostnem leksikonu.

imeti	4859
ostati	4400
priti	4083
dobiti	3783
dati	3285
igrati	2989
predstaviti	2941
delati	2732
prevajati	2625
videti	2619
vzeti	2606
narediti	2572
sodelovati	2540
iti	2524
pripeljati	2494
izgubiti	2443
pomagati	2426
govoriti	2426
nastopiti	2420
pokazati	2387
doseči	2348
vrniti	2327
pasti	2210
voziti	2200
stati	2185
prihajati	2166
poslati	2096
postati	2034
povečati	2013
dvigniti	2009
plačati	1999
spraviti	1997
zmagati	1990
peljati	1987
znižati	1982
obrniti	1980
zbrati	1968
pustiti	1876
odpraviti	1848
uvrstiti	1847
povedati	1847
plačevati	1841
spremeniti	1829
prejeti	1809
nameniti	1805
postaviti	1791
hoditi	1785
pripraviti	1767
delovati	1763
voditi	1759
stopiti	1751

Priloga 2: Seznam glagolov z največjo frekvenco stavkov, ki se pojavljajo v vezljivostnih vzorcih v Vezljivostnem leksikonu.

1	imeti	3054841
2	morati	1978844
3	iti	1361145
4	začeti	1115895
5	priti	1017046
6	dobiti	959830
7	povedati	919724
8	želeti	870894
9	moči	787696
10	vedeti	741256
11	videti	638272
12	postati	637465
13	praviti	606482
14	pomeniti	567388
15	ostati	529493
16	dejati	509807
17	najti	503982
18	odločiti	501947
19	doseči	499666
20	dati	474899
21	igrati	474480
22	narediti	457930
23	reči	449018
24	hoteti	429719
25	govoriti	426294
26	pričakovati	409648
27	uspeti	408480
28	pokazati	406575
29	voditi	405752
30	pomagati	402565
31	delati	397779
32	veljati	394632
33	pripraviti	390593
34	kazati	376309
35	zgoditi	374008
36	zdeti	367321
37	sprejeti	366503
38	živeti	361229
39	potrebovati	357082
40	misliti	350867
41	potekati	349638
42	predstaviti	348925
43	meniti	346598
44	čakati	345443
45	zahtevati	342642
46	dodati	337253
47	končati	332375
48	sodelovati	330390
49	ugotoviti	324324
50	delovati	316336