Glagolske večbesedne enote v učnem korpusu ssj500k 2.1

Polona Gantar,* Špela Arhar Holdt,† Jaka Čibej,‡ Taja Kuzman,♣ Teja Kavčič♦

* Oddelek za prevajalstvo, Filozofska fakulteta, Univerza v Ljubljani
Aškerčeva 12, 1000 Ljubljana
apolonija.gantar@guest.arnes.si
† CJVT (Fakulteta za računalništvo in informatiko, Filozofska fakulteta), Univerza v Ljubljani
Večna pot 113, 1000 Ljubljana
spela.arhar@cjvt.si
‡ Laboratorij za umetno inteligenco, Institut »Jožef Stefan«
Jamova cesta 39, 1000 Ljubljana
jaka.cibej@ijs.si
♣kuzman.taja@gmail.com
♦teya10teja@gmail.com

Povzetek

V prispevku predstavljamo kategorije glagolskih večbesednih enot, kot so bile oblikovane v okviru mednarodne COST akcije PARSEME Shared Task 1.1. za 26 različnih jezikov, in izdelavo učnega korpusa glagolskih večbesednih enot za slovenščino. Osnovni namen prispevka je opisati prve kvantitativne in kvalitativne analize, ki bodo predstavljale izhodišča za izdelavo modela za strojno luščenje večbesednih enot iz korpusnih besedil in za izdelavo leksikona večbesednih enot za slovenščino. V prvem delu prispevka predstavimo postopek označevanja ter označevalne smernice s prilagoditvami za slovenščino, korpusno gradivo ter označevalni program. V drugem delu prispevka natančneje predstavimo 3.364 večbesednih glagolskih enot (iz skupno 13.511 ročno pregledanih povedi) po pripisanih kategorijah: inherentno povratni glagoli, zveze z glagoli v pomensko oslabljeni rabi, predložnomorfemski glagoli in glagolski idiomi. Prispevek sklenemo z razpravo in načrti za prihodnje delo

Verbal Multi-Word Expressions in the Slovene Training Corpus SSJ500k 2.1

The paper presents the categories of verbal multi-word expressions (VMWEs) as developed within the international PARSEME COST Action Shared Task 1.1 for 26 different languages, and the annotation of a Slovene training corpus of VMWEs. The main goal of the paper is to describe the first quantitative and qualitative analyses of VMWEs that will serve as a basis for building a model for the automatic extraction of VMWEs from corpus texts, as well as for the compilation of a lexicon of Slovene MWEs. We begin the paper by presenting the annotation process, the annotation guidelines adapted to Slovene, the corpus, and the annotation tool used. This is followed by a detailed analysis of 3,364 VMWEs (from a total of 13,511 manually annotated sentences) divided into four categories: inherently reflexive verbs, light-verb constructions, inherently adpositional verbs, and verbal idioms. We conclude the paper with a discussion and the description of our plans for future work.

1. Uvod

Večbesedne enote (VE) so prepoznane kot obsežen del mentalnega leksikona govorcev določenega jezika, zato so pomembne tako za jezikoslovne raziskave kot za izgradnjo računalniško procesljivih jezikovnih virov, ki omogočajo izdelavo elektronskih leksikonov VE in razvoj orodij za njihovo procesiranje.

Obstaja več definicij VE, ki se razlikujejo glede na metodološko-teoretična izhodišča in raziskovalne cilje. Jezikoslovni, ali natančneje slovarski vidik, postavlja v ospredje semantične lastnosti VE in jih opredeljuje kot različne tipe zvez, ki izkazujejo določeno stopnjo idiomatičnega pomena (Atkins in Rundell, 2008: 166) ali z drugimi besedami, kot zveze, katerih celostni pomen ni vsota pomenov posameznih sestavin. Definicija VE, oblikovana za namene strojnega procesiranja, na drugi strani izpostavlja (ne)zmožnost njihove razstavljivosti na samostojne lekseme ob ohranitvi pomenskih lastnosti in skladenjske funkcije ter izražanje t. i. leksikalne, skladenjske, pomenske, pragmatične in statistične idiomatičnosti (Baldwin in Kim, 2010: 3). Čeprav ne obstaja splošno sprejeta definicija VE, se tako jezikoslovna kot NLP skupnost strinjata, da je osnovna lastnost, ki ločuje VE od prostih zvez, specifično razmerje, ki obstaja med elementi VE. To razmerje se navadno obravnava v okviru konceptov, kot so kolokabilnost (ali statistična idiomatičnost), idiomatičnost (ali semantična nerazstavljivost), sintaktična (ne)fleksibilnost, ki vključuje tudi možnost notranje modifikacije zveze in nezaporednost leksikaliziranih elementov, ter leksikalna variantnost. Zaradi naštetega predstavljajo VE problem ne samo pri jezikovnih analizah, ampak tudi pri strojnem procesiranju in avtomatski prepoznavi v besedilu.

Eden od načinov za izboljšanje jezikovnotehnoloških nalog, ki vključujejo obravnavo VE, je poznavanje njihovih temeljnih jezikovnih lastnosti ter – na tej osnovi – razvoj metode in standardov za prepoznavanje različnih tipov VE v tekočem besedilu. Če želimo omogočiti, da bo čim večji nabor novorazvitih postopkov dobro deloval tudi za slovenščino, je treba izdelati jezikoslovne analize, ki upoštevajo specifike slovenščine in so hkrati kompatibilne na medjezikovni ravni.

Okvir analize, katere rezultate opisujemo v prispevku, določa sodelovanje v okviru COST akcije PARSEME Shared Task 1.1, rezultati pa bodo uporabni pri izdelavi elektronskega leksikona VE za slovenščino, ki je ena od aktivnosti projekta *Nova slovnica slovenskega jezika: viri in metode*, posredno pa tudi za izdelavo jezikovnih priročnikov, kot sta npr. Slovar sodobnega slovenskega jezika (Gorjanc et al., 2015) in na korpusu temelječa znanstvena slovnica.

V prispevku najprej opišemo postopek prepoznavanja potencialnih glagolskih večbesednih enot (GVE) in posamezne kategorije, kot so definirane v smernicah PARSEME Shared Task 1.1. Nato opišemo postopek označevanja GVE v učnem korpusu ter orodje za označevanje. V nadaljevanju opišemo prve rezultate analiz ročno označenih primerov, in sicer tako s kvantitativnega kot kvalitativnega vidika. V prvem primeru nas je zanimala zastopanost posamezne kategorije v korpusu, frekventnejši posamezni primeri ter zastopanost posameznih elementov znotraj zveze. Jedro prispevka je namenjeno opisu strukturnih, skladenjskih in pomenskih lastnosti prepoznanih GVE.

2. Kategorije glagolskih večbesednih enot in postopek prepoznavanja

Kategorizacija GVE temelji na smernicah, izdelanih v okviru PARSEME Shared task 1.1 (Bathia et al., 2017), definicija posamezne kategorije pa se opira na pomenske in skladenjske lastnosti glagolske zveze, ki so opisane v obliki odločitvenih drevesnic. Prepoznavanje in kategorizacija sta potekala v treh korakih. V prvem koraku smo identificirali zveze glagola z vsaj še eno besedo, ki predstavljajo potencialne GVE. V drugem koraku smo prepoznavali leksikalizirane elemente zveze, tj. elemente brez katerih GVE ne more obstajati, v tretjem koraku pa smo se na podlagi podrobnih jezikovnih testov v obliki generičnih in specifičnih jezikovnih meril odločali, v katero kategorijo sodi prepoznana GVE.

Na podlagi smernic so prepoznane GVE razdeljene na kategorije znotraj dveh razredov, določenih glede na to, ali je kategorijo mogoče aplicirati na večino jezikov, vključenih v raziskavo, ali pa je značilna samo za posamezne jezike. Univerzalne kategorije vključujejo *zveze z glagoli v pomensko oslabljeni rabi* (ang. LVC: Light Verb Constructions), ki so nadalje ločne na celostne (ang. LVC.full) in na kavzativne oz. vzročnostne (ang. LVC.cause), ter na *glagolske idiome* (ang. VID: Verbal Idioms). T. i. kvaziuniverzalne kategorije, ki so vezane na posamezne jezikovne skupine, vključujejo *inherentno povratne glagole* (ang. IRV: Inherently Reflexive Verbs), značilne za večino slovanskih jezikov, ter zveze glagola z izpredložnim morfemom (ang. VPC: Verb-Particle Constructions), značilne predvsem za germanske jezike. Zadnja kategorija je bila v drugi verziji Smernic dopolnjena s tipom predložnih glagolskih zvez (ang. IAV: Inherently Adpositional Verbs), ki predvidevajo odprto skladenjsko mesto in so značilne tudi za slovenščino in nekatere druge slovanske jezike. V prispevku jih imenujemo *predložnomorfemski glagoli* z leksikaliziranim predložnim morfemom.

Za slovenščino je bilo mogoče registrirati GVE za vse predvidene kategorije, razen za VPC, pri čemer obstajajo za posamezne kategorije posebnosti, povezane bodisi s skladenjskimi in morfološkimi lastnostmi slovenščine bodisi s slovničnimi kategorijami, ki so splošno uveljavljene v jeziku in se deloma razlikujejo od drugih jezikov. Na slovenske posebnosti bomo v nadaljevanju opozorili ob posameznih tipih GVE.

3. Korpus in označevalnik

Za označevanje GVE smo uporabili učni korpus ssj500k 2.0 (Krek et al., 2017), ki vključuje približno 500,000 pojavnic in nekaj manj kot 28.000 stavkov iz vzorčenih odstavkov korpusa FidaPLUS (Arhar Holdt in Gorjanc, 2007). Korpus je v celoti označen na oblikoskladenjski ravni (Grčar et al., 2012), v posameznih deležih pa še na ravni lastnoimenskih entitet in skladenjskih razčlemb (Dobrovoljc et al., 2012). Naslednja različica korpusa ssj500k 2.1 vključuje tudi semantične oznake v obsegu 5.500 stavkov (Krek et al., 2018). V prvi fazi označevanja je bilo s kategorijami GVE, kot jih je določala prva verzija Smernic (Candito et al., 2016), označenih 11.411 stavkov s strani dveh označevalcev, nestrinjanja v odločitvah pa so bila prediskutirana in ustrezno popravljena. V drugi fazi so bile kategorije avtomatsko preoznačene na podlagi druge verzije Smernic ter ročno pregledane. S posodobljenimi kategorijami je bilo v drugi fazi dodatno označenih še 2.100 stavkov s strani enega označevalca, pri čemer so bili problematični primeri prav tako prediskutirani in ustrezno popravljeni. Celotni izplen vseh pojavitev GVE v učnem korpusu, kot je razvidno iz Tabele 1, je 3.364 enot.

Za označevanje smo v prvi fazi uporabili orodje SentenceMarkup System, ki je bilo primarno razvito za skladenjsko označevanje slovenščine (Dobrovoljc et al., 2012). Orodje smo prilagodili za namene označevanja GVE tako, da smo mu dodali neodvisen in hkrati medsebojno povezljiv nivo (prim. Gantar et al., 2017). V drugi fazi je označevanje potekalo v spletni anotacijski platformi FLAT (FoLiA Linguistic Annotation Tool), ki je bila prilagojena za namene PARSEME Shared Task in preizkušena na 13 sodelujočih jezikih (Slika 1).

Slika 1: Označevalnik FLAT.

Platforma FLAT omogoča označevanje nizov besedila z vnaprej določenimi kategorijami in dodeljevanje datotek različnim označevalcem. Pri uvozu podpira formata XML in TSV, izvoz končnih datotek pa je v formatu XML. Vnesene oznake se med označevanjem shranjujejo samodejno. Vmesnik omogoča tudi iskanje po besedilih s pomočjo iskalnih pogojev v jeziku CQL.

4. Kvantitativna analiza

Označene GVE so bile po koncu označevanja uvožene v učni korpus ssj500k 2.1 (Krek et al., 2018). Od 13.511 stavkov, pregledanih med prvo in drugo fazo označevanja, jih vsaj eno GVE vsebuje 2.920, kar znaša približno 22 %. Vsak od teh stavkov v povprečju vsebuje 1,15 GVE, glede na celoten pregledani nabor stavkov pa je količina GVE na stavek približno 0,25, kar pomeni, da na GVE v povprečju naletimo v vsakem četrtem stavku.

Tabela 1 prikazuje razpored označenih GVE glede na kategorije. Vseh različnih GVE (brez večkratnih pojavitev ene same enote) je bilo slabih 1.100. Po absolutni frekvenci

največji delež zajema kategorija IRV (48 %), najmanj enot pa je v kategoriji LVC.cause (2 %). Po visokem številu različnih GVE izstopata kategoriji VID in IAV, najmanj raznoliki pa sta kategoriji LVC.full in LVC.cause.

Kategorija	Vse GVE	Delež	Različne GVE
IRV	1.627	48 %	345
IAV	710	21 %	154
VID	724	22 %	457
LVC.cause	64	2 %	27
LVC.full	239	7 %	103
Skupaj	3.364	100 %	1.086

Tabela 1: Razporeditev označenih GVE glede na kategorije.

Pregledani stavki so večinoma vzeti iz besedil s pisnim prenosnikom (13.277 stavkov oz. 98 %), iz govornega prenosnika pa je le 234 stavkov (2 %). Glede na besedilno zvrst je največ stavkov (9.017 oz. 67 %) iz periodičnih publikacij (časopisi in revije), 3.968 stavkov (29 %) je iz knjižnih besedil, preostali 4 % pa so uvrščeni pod drugo. Glede na čas objave besedila pregledani stavki zajemajo obdobje med letoma 1991 in 2006: 3.616 stavkov (27 %) je bilo objavljenih pred letom 2000, 9.375 (69 %) pa med letoma 2000 in 2006. Pri 520 stavkih (4 %) čas objave ni znan. Večina stavkov (10.859 oz. 80 %) je vzeta iz lektoriranih besedil. Pri 2.459 stavkih (18 %) metapodatek o lektoriranosti ni na voljo, pri 193 stavkih (2 %) pa besedilo ni bilo lektorirano.

Tabela 2 prikazuje najpogostejše strukture GVE glede na besedno vrsto sestavine (G – glagol, S – samostalnik, P – pridevnik, R – prislov, D – predlog, Z – zaimek). Strukture, ki so se v korpusu pojavile manj kot 10-krat, so združene v kategorijo Drugo. Najpogostejše strukture so G + Z, G + D, G + S in G + D + S, ki skupaj zajemajo kar 85 % vseh označenih GVE.

Struktura	Primer	Frekvenca	Delež
G + Z	*bati se*	1.663	49 %
G + D	*priti do*	535	16 %
G + S	*imeti odnos*	372	11 %
G + D + S	*biti pod vtisom*	303	9 %
G + Z + P	*biti si edini*	146	4 %
G + R	*biti res*	136	4 %
G + Z + D + S	*ujeti se v past*	24	1 %
G + P	*biti jasno*	20	1 %
G + P + S	*imeti glavno besedo*	19	1 %
S + G + D + S	*biti na robu propada*	12	<1 %
G + Z + S	*vzeti si čas*	11	<1 %
Drugo	-	123	4 %
Skupaj	-	3.364	100%

Tabela 2: Razporeditev označenih GVE glede na besednovrstno strukturo.

5. Kvalitativne analize

Kvalitativna analiza označenih primerov GVE v učnem korpusu zajema njihove strukturne in pomenske lastnosti. Hkrati so bile na podlagi Smernic, ki definirajo posamezno kategorijo znotraj PARSEME Shared Task, prepoznane specifike slovenščine, ki se kažejo na ravni strukturnih in pomenskih testov. Pri tem nas je zanimala vzorčenost strukture znotraj posamezne kategorije, skladenjsko okolje zveze kot celote ter leksikalne zapolnitve na predvidenih udeleženskih mestih. Na podlagi korpusnih primerov smo želeli prepoznati tudi kazalce pomenske celovitosti, ki so uporabni pri avtomatskem prepoznavanju v besedilu.

5.1. Inherentno povratni glagoli (IRV)

Smernice PARSEME Shared Task 1.1 kot samostojne GVE obravnavajo glagole s prostim morfemom *se/si*, imenovali jih bomo inherentno povratni glagoli. Gre za jezikovnospecifično kategorijo, kjer so kot IRV prepoznane samo zveze, kjer glagol brez *se/si* bodisi ne obstaja (*zdeti se*) bodisi prisotnost *se/si* glagolu spreminja pomen in/ali funkcijo (*dati se* – 'moči'). Za preizkus leksikalne trdnosti zveze kot celote in za ločevanje od vseh drugih zvez glagola + *se/si*,[1] je mogoče uporabiti več pomensko-skladenjskih testov, ki preverjajo obnašanje glagola z vidika odpiranja skladenjskih položajev zveze kot celote, pri čemer so določene spremembe v vzorcu in vlogi udeležencev, ki jih taka glagolska zveza predvideva, lahko tudi znanilec pomenskih sprememb.

V učnem korpusu predstavljajo IRV največji delež znotraj obravnavanih kategorij (gl. Tabelo 1). Med 1.627 označenimi primeri so bili štirje primeri napačno kategorizirani, v dveh primerih pa elementi zveze niso bili ustrezno označeni. Ti primeri so bili iz nadaljnje obravnave izločeni. Med pravilno označenimi primeri (1.621) je bilo mogoče identificirati 339 različnih IRV, med katerimi se v korpusu zveze *bati se*, *dati se*, *dogajati se*, *izkazati se*, *lotiti se, odločiti se, počutiti se, pogovarjati se*, *pojaviti se*, *spominjati se, spomniti se, strinjati se, udeležiti se, vrniti se*, *zavedati se*, *zdeti se* in *zgoditi se* pojavijo več kot 20-krat. Variantnost morfema *se/si* se kaže pri manjšem deležu primerov (*premisliti se/si*, *prizadevati se/si*, *upati se/si*, *zapomniti se/si*), v drugih primerih je morfem ustaljen, npr. *bati se, zdeti se; zamišljati si, zaželeti si*.

Najbolj prepoznavna lastnost IRV, za katere je značilno, da brez *se/si* ne obstajajo, je, da glagolska zveza kot celota ne prenese neposrednega predmetnega določila, ki nastopa v udeleženski vlogi prizadetega. Znotraj te skupine je mogoče ločiti dve tipični situaciji: (a) glagolska zveza ne predvideva predmetnega določila v svojem širšem stavčnem vzorcu, npr. *dreti se* : **dreti (se) koga*, *drstiti se* : **drstiti (se) koga*, lahko pa je (b) predmetno določilo del stavčnega vzorca ob prisotnosti morfema *se/si*, npr. *bati se koga* : **bati koga*, *izogibati se koga/komu* : **izogibati koga*, zlasti pogosto s predložnim ali dajalniškim predmetnim določilom, npr. *strinjati se s kom*, *pogovarjati se s kom*, *odzvati se na kaj, odpovedati se komu/čemu* ipd. Lahko bi rekli, da gre v prvem primeru za neprehodne IRV, kamor poleg naštetih sodijo tudi *zvečeriti se, mračiti se* ipd., ter prehodnimi IRV, kjer je (zlasti predložno) predmetno določilo pričakovani del širšega stavčnega vzorca, ki ga napoveduje glagolska zveza.

[1] Upoštevajoč večfunkcijskost *se/si* so iz obravnave IRV izločeni primeri, kjer gre bodisi za pasivne zgradbe (npr. *kazati se*), povratnost (*umivati se, zlomiti si (roko)*) ali vzajemnost (*poljubiti se*) (Gantar et al., 2017).

Pri glagolih, ki lahko obstajajo tudi brez morfema *se/si*, so kot IRV določene samo tiste zveze, kjer morfem glagolu spreminja pomen. Tudi pomensko celovitost zveze glagola in morfema je mogoče prepoznati na podlagi skladenjsko-pomenskih lastnosti, ki jih posamezni pomen zveze definira v svojem stavčnem vzorcu. V prvi skupini nastopajo (prehodni) glagoli, pri katerih prisotnost oz. odsotnost *se/si* povzroči očitno pomensko spremembo, ta pa je pogosto vezana na pomenske lastnosti osebka, ki je v primeru IRV navadno človeško+, hkrati pa glagoli v taki zvezi pogosto nastopajo v svojem prenesenem pomenu, ki ima tudi prepoznavno dobesedno ustreznico, npr. *delati se* – 'pretvarjati se' : *delati kaj* – 'početi, izdelovati'; *pobrati se* – 'opomoči si' : *pobrati kaj* – 'dvigniti s tal'.

V drugi skupini prehodnih glagolov (tj. glagolov, ki dovoljujejo neposredno predmetno določilo, *si/se*, če se pojavlja ob njih, pa lahko izraža pravo povratnost, zaradi česar zveza ni prepoznana kot IRV, npr. *umivati se* – *umivati koga*) je za ločevanje povratnih zvez od IRV treba upoštevati predvsem primere, kjer prisotnost *se/si*, kljub temu da vrača dejanje/stanje na osebek, zagotavlja zadosten pomenski prenos, da je zvezo mogoče obravnavati kot celoto, npr. *dokazovati se* : *dokazovati kaj*, *gristi se* : *gristi kaj*, *naslikati se* : *naslikati koga/kaj*, in sicer tudi v primerih, kjer osebek ni konkretiziran in izraža splošnost, npr. *izplačati se* : *izplačati koga/kaj*, *vleči se* : *vleči koga/kaj*, ali vzajemnost dejanja, npr. *ljubiti se* : *ljubiti koga/kaj*.

Pri prepoznavanju zveze glagola s prostim morfemom *se/si* kot IRV je treba upoštevati tudi številne primere, kjer *se/si* izraža pasiv, npr. *ponavljati kaj – kaj se ponavlja*, *zagotavljati kaj – kaj se zagotavlja*. V takih primerih *se/si* ni del glagola, pač pa samo ena od skladenjskih možnosti umikanja osebka iz stavčnega vzorca.

Leksikalizirane zveze glagola in morfema *se/si* v slovenistični literaturi niso bile obravnavane (izključno) z vidika leksikalne celovitosti, npr. kot samostojna kategorija stalnih besednih zvez, pač pa predvsem z vidika funkcije morfema oz. povratnega zaimka (Toporišič, 2000: 503, 579; Žele, 2012). Pri tem se ugotavlja vloga *se/si* z vidika izražanja različnih stopenj vršilskosti oz. osebkove (ne)udeleženosti, kot npr. v primeru needninskega (*bratiti se* ali *zbrati se*) ali splošnega vršilca dejanja (*tiskati se*) (Žele, 2012: 44, Toporišič, 1982: 244). Njihova ustrezna prepoznava v besedilu z vidika pomenskoskladenjske celovitosti, kot jo opisujemo v prispevku, je pomembna predvsem za strojno prepoznavanje večbesednih enot in njihovo razlikovanje od »prostih« zvez tega tipa. Posledično gre v primeru IRV za enote leksikona, ki jih je kot take smiselno obravnavati v slovarju, bodisi kot samostojne iztočnice bodisi v okviru večpomenskosti.

5.2. Zveze z glagoli v pomensko oslabljeni rabi (LVC)

Da je znotraj sistema Parseme večbesedna enota označena kot LVC, mora ustrezati naslednjim pogojem: sestavljena mora biti iz glagola in samostalnika oz. samostalniške besedne zveze, ki je lahko v obliki predložne zveze, npr. *imeti mnenje, biti v dvomih*, in odpirati mora lastna vezljivostna mesta (npr. *kdo ima predavanje za koga*). Pomensko mora biti povezana z dogajanjem (*imeti predavanje*) ali stanjem (*biti v dvomih*). Glagolski del je lahko dveh tipov: (a) če je glagol pomensko oslabljen oz. k pomenu prispeva pretežno na kategorialni ravni, zvezo uvrstimo v podkategorijo LVC.full, npr. *biti v pomoč*; (b) če lahko osebek razumemo kot vzrok ali vir izraženega dejanja/stanja, zvezo uvrstimo v podkategorijo LVC.cause, npr. *imeti učinek, spraviti v smeh*. V algoritmu za presojanje, ali je določena zveza kandidatka za označevanje ali ne, se upošteva še abstraktnost samostalnika (tip *imeti avto* se ne uvršča med večbesedne enote, idiomatične zveze tipa *imeti mačka* v pomenu 'slabo počutje po uživanju alkohola' pa se uvrščajo v kategorijo VID) in pri uvrščanju v LVC.full zmožnost pretvorbe z izpustom glagola *Janez ima predavanje* → *Janezovo predavanje* (glede slednjih gl. tudi 5.4).

Različne možnosti opredeljevanja zvez s pomensko oslabljenimi glagoli v slovenskem in širšem prostoru pregledno predstavi Soršak (2013), po kateri povzemamo tudi poimenovanje *oslabljenopomenski glagol* (nasproti *polnopomenskemu*), skupaj z opozorilom, da je najustrezneje govoriti o glagolih v pomensko oslabljeni rabi. Kot glavno razliko sistema Parseme v primerjavi z dosedanjimi slovenskimi opredelitvami je mogoče izpostaviti delitev na LVC.full in LVC.cause ter dejstvo, da je pretvorljivost oz. zamenljivosti s polnopomenskim glagolom omenjena le med dodatnimi pogoji v odločevalnih drevesnicah, ni pa med osnovnimi določevalnimi parametri.

Med skupno 303 primeri, označenimi kot LVC (en primer je bil označen napačno), jih je v kategoriji LVC.full 238 (78,8 %) in v LVC.cause 64 (21,2 %). V podatkih se pojavljata dve vrsti struktur: (a) kombinacije glagola in samostalnika (263 oz. 87,1 %) in (b) zveza glagola in predložne samostalniške zveze (39 oz. 12,9 %). Močno prevladujejo zveze z glagolom *imeti* (65,6 %), nekoliko pogosteje se pojavljata tudi *biti* (13,6 %) in *da(ja)ti* (skupaj 9,6 %). Drugi glagoli (*narediti, postaviti, postavljati, ostati, voditi, namenjati, delati, storiti, vzbujati, zbujati, dobiti, zastaviti, spraviti, doseči* in *nositi*) se pojavljajo redkeje in so pogosto vezani na eno samo identificirano zvezo (npr. *ostati v spominu*, *namenjati pozornost*, *(v)zbujati vtis*).

Zveze glagola in predložne zveze so glede na razmerja nekoliko više zastopane v kategoriji LVC.cause. V podatkih se pojavljajo izključno zveze s predlogoma *v* (33 oz. 84,6 %) in *na* (6 oz. 15,4 %). Velika večina podatkov (24 oz. 61,5 %) vsebuje *biti v* (*biti v pomoč, biti v podporo, biti v navadi, biti v interesu, biti v dvomih, biti v korist, biti v prednosti, biti v težavah, biti v užitek, biti v sporu, biti v skrbeh*). S 6 pojavitvami sledijo zveze z *imeti v* (*imeti v lasti, imeti v načrtu, imeti v spominu*), nato s po 3 pojavitvami *ostati v* (*ostati v spominu*) ter *biti na* (*biti na voljo*), samo po 1 pojavitev pa imajo *dati na* (*dati na voljo*), *imeti na* (*imeti na izbiro*) in *spraviti v* (*spraviti v smeh*).

V označenih večbesednih enotah nastopa relativno omejen nabor samostalnikov, skupno jih je 97. Najpogostejša sta *težava* (21) in *pravica* (20), sledijo *možnost, mnenje, učinek, vloga, vpliv, vtis, pomoč, občutek, prednost, sreča, korist, vprašanje, volja*, *posledica*. Po pričakovanjih se nekateri od teh samostalnikov pojavljajo izključno v zvezah LVC.full (*pravica, možnost, mnenje, vloga*), drugi v LVC.cause (*učinek, vpliv, vtis, pomoč*), ponekod pa je pripis kategorije vezan na pomen glagola, npr. *dati prednost* → LVC.cause ter *imeti prednost* → LVC.full.

Pri večini primerov (79 oz. 81,4 %) se samostalnik v podatkih pojavlja z enim samim glagolom, npr. *imeti pravico, biti v pomoč, dati predlog*. Dodatni 3 primeri se pojavljajo z vidskimi pari, npr. *dati/dajati soglasje*. Ločeno skupino predstavlja 5 samostalnikov (*težava, mnenje,*

korist, interes, vzrok), ki se pojavljajo z glagoloma *biti* in *imeti*, npr. *biti v težavah* vs. *imeti težave* (prim. Vidovič Muha, 1998: 307–308, ki utemeljuje povezavo med glagoloma prek izražanja prostorske umeščenosti). Preostalih 10 samostalnikov se pojavlja z različnimi glagoli: *dati, dajati, doseči, imeti učinek*; *dajati, narediti, vzbujati, zbujati vtis*; *postaviti, postavljati, zastaviti vprašanje; biti na voljo, dati na voljo, imeti na voljo*; *imeti v spominu, ostati v spominu*; *dati ime, nositi ime*; *biti v podporo, dati podporo, imeti podporo*; *biti v skrbeh, delati skrbi, imeti skrbi*.

Kot omenjeno (Soršak, 2013), pomenska oslabljenost glagolov ne pomeni pomenske praznosti, kar potrjujejo tudi označeni podatki. Tako v skupini LVC.full kot LVC.cause se pojavljajo glagoli, ki so v rabi prisotni tudi s polnim pomenom, pomensko oslabljenost pa v zvezah LVC dopolnjuje samostalniški del (npr. *imeti* v pomenu 'posedovati' nasproti *imeti posledice* v pomenu 'sprožiti, povzročiti, voditi v posledice'). V pomenskem smislu skupine samostalnikov, ki se pojavljajo v LVC.cause po pričakovanjih opisujejo rezultat določenega dejanja, naj bo to opredelitev vrste rezultata (*učinek, vpliv, vtis*) oz. pozitivna (*korist, užitek*) ali negativna posledica (*muka, preglavica*). Pomensko oslabljeni glagol veže rezultat na stavčni osebek (*nekdo* oz. *nekaj daje, naredi, vzbuja vtis*, je torej povzročitelj dejanja). V določenih primerih so zveze LVC pretvorljive v polnopomenski glagol s sorodno morfološko podobo (npr. *imeti, dajati, dosegati učinek – učinkovati*; *imeti vpliv – vplivati*), ne pa vedno (npr. *vzbujati vtis*; *imeti posledice*).

Na drugi strani je skupina samostalnikov pri LVC.full pomensko bolj heterogena. Poskus delitve v pomenske skupine razkrije, da bi kot stično točko številnih zvez morda lahko izpostavili načrtovanje in ocenjevanje uspeha. Pojavljajo se npr. zveze s samostalniki, ki so vezani na (a) komunikacijo (*mnenje, predlog, vprašanje, izjava, soglasje*), opisujejo (b) potencial za uspeh (*možnost, prednost, priložnost, naskok*), (c) začetne korake (*obljuba, napoved, načrt, pobuda*) ali (č) potencialne razloge za neuspeh (*napaka, pomanjkljivost*). Pojavljajo se tudi skupine, ki opredeljujejo (d) negativno stanje (*težava, strah, dvom, zamera*), (e) pozitivne lastnosti (*moč, pogum, potrpljenje*), (f) dosežene rezultate (*izobrazba, status, posel, mir*) in (g) odnos do še nerealiziranih ciljev (*želja, ambicija, vizija, interes*). Tudi pri tej skupini zvez velja, da so pretvorljive v polnopomenski glagol (npr. *imeti mnenje – meniti*; *dati soglasje – soglašati*) ali ne (*imeti ambicije*; *dati priložnost*).

Kot je razvidno iz navedenih primerov, pokriva kategorija LVC pomensko različne večbesedne enote, ki ponujajo možnost za nadaljnje premisleke in popravke označevanja. Z vidika natančnosti je mogoče premisliti in natančneje opredeliti mejo med zvezami LVC in kolokacijami, pri čemer je lahko vodilo pojavljanje z več glagoli (npr. *postaviti, postavljati, zastaviti vprašanje*). Z vidika priklica pa je treba preveriti, ali so bile v korpusu v resnici označene vse relevantne enote, v prvem koraku morda s pomočjo preverbe besednih skic za identificirane glagole in samostalnike. S slovenističnega vidika bi kazalo na podatkih preveriti še ugotovitve (Žele, 2012: 227–28), da je abstraktni samostalnik navadno v obliki za ednino in v tožilniku.

5.3. Predložnomorfemski glagoli (IAV)

Predložnomorfemski glagoli, imenovani tudi glagoli z leksikaliziranim predložnim morfemom (prim. Žele, 2002), so bili v drugi fazi označevanja vključeni kot neobvezna poskusna kategorija. V Smernicah so kot IAV definirani glagoli, ki brez predložnega morfema ne obstajajo, npr. *simpatizirati z, sprevreči se v, sklicevati se na, apelirati na*, in glagoli, ki jim predložni morfem občutno spremeni pomen, npr. *biti za* – 'strinjati se', *priti do* – 'zgoditi se', *hoditi v/na* – 'obiskovati'. Pri tem je pomembno upoštevati, da udeleženci, ki jih predvideva glagolska zveza kot celota, niso del glagolske večbesedne enote, za razliko od npr. *stati na + trdnih tleh*, so pa bodisi skladenjsko obvezni ali neobvezni (*gre za : *kdo/kaj gre za, vendar: gre za koga/kaj*) in omejeni s slovničnimi, npr. sklon (*simpatizirati s (kom)*), in pomenskimi kategorijami, npr. *priti do (nesreče)* : *priti do (cilja)*.

Na obravnavo predlogov kot prostih glagolskih morfemov naletimo že v Metelkovi slovnici (1825: 247-256, cit. po Žele, 2002: 99), podrobneje jih obravnava tudi Breznik (1916: 250; 1934: 225, cit. po ibid.), izraz »prosti predložni glagolski morfem« pa se v slovenščini ustali v šestdesetih letih (Toporišič, 1967: 111). Podrobneje sta glagole z leksikaliziranim predložnim morfemom v slovenski literaturi obravnavali Žele (2002) in Kržišnik (1994). Prva z vidika stopnje leksikaliziranosti predloga (t. i. leksikalizirani, neleksikalizirani in vezavnodružljivi morfemi), druga pa z vidika frazne trdnosti, tj. bodisi kot frazeološke enote, kjer gre zgolj za strukturno ustaljenost, npr. *biti ob (čem)* – 'nahajati se (ob čem)', ali kot frazeme, kjer gre za leksikalno ustaljenost, npr. *biti ob (kaj)* – 'izgubiti; ne imeti več'.

V učnem korpusu predstavljajo IAV približno petino označenih primerov GVE (gl. Tabelo 1). Med 710 primeri vseh pojavitev, je bilo mogoče identificirati 154 različnih IAV, med katerimi se v korpusu vsaj dvajsetkrat pojavijo zveze *iti za* (vedno z glagolom v tretji osebi ednine – *gre za*), *priti do, vplivati na, skrbeti za, temeljiti na, naleteti na, veljati za* in *biti proti*. V skladu s smernicami smo kot IAV označevali tudi glagolske zveze, sestavljene iz inherentno povratnega glagola (gl. pogl. 5.1) in leksikaliziranega predložnega morfema, kot npr. *ukvarjati se z, nanašati se na, zavzemati se za* ipd.

Pri IAV leksikalizirani predložni morfem običajno sledi glagolu, kar potrjuje 86 % označenih primerov, in sicer se v veliki večini primerov nahaja neposredno za glagolom oz. v njegovi neposredni bližini (+ 3 besede). Izjemo predstavlja *gre za*, kjer je vrivanje služi referiranju na predhodno ubeseditev, npr. *gre (v tem primeru) za*.

Primeri, kjer se predložni morfem z vidika izbire besednega reda nahaja pred glagolom, so v učnem korpusu veliko redkejši. V teh primerih glagol nikoli ne sledi neposredno predložnemu morfemu, razdalja med njima pa je občutno večja, in sicer v petini primerov znaša tri besede ali več. Ta tendenca se zdi zanimiva za strojno prepoznavanje IAV, kjer besednoredna distribucija predloga pred glagolom predvideva upoštevanje razmeroma široke okolice glagola.

Glagole z leksikaliziranim predložnim morfemom je mogoče prepoznati tudi glede nekaterih skupnih pomenskih lastnosti, npr. za izražanje (a) funkcije ali lastnosti, *veljati*

za (favorita, človeka),[2] *imenovati za (direktorja), šteti za, (uspeh), označiti za (laž), narediti za (politika), razglasiti za (svetnika), spoznati za (nevarnega), smatrati za (sovražnika)*, b) (ne)strinjanje, npr. *biti za (globalizacijo), govoriti za (združevanje), biti proti (vojni), imeti (kaj) proti*, c) izhajanja, upoštevanja, npr. *temeljiti na (dejstvu), graditi na (zaupanju), nanašati se na (podatke), nasloniti se na (tradicijo), navezovati/navezati se na (besede), opirati se na (izkušnje), izhajati iz (predpostavke)*, č) začetek ali spremembo dejanja/stanja, npr. *pasti v (komo), spuščati se v (polemiko), pahniti v (obup), priti v (formo), priti/prihajati do (spremembe), pasti pod (vpliv), pripeljati do (spoznanja), prerasti v (ljubezen), sprevreči se v (nasprotje)*, d) spremembo lastnosti, oblike, npr. *pretvoriti v (energijo)*, e) preživljanje, prestajanje, npr. *iti skozi (proces), (morati) dati skozi, (ceneje) priti skozi*, f) aktivno delovanje, npr. *ukvarjati se z, baviti se z, ubadati se z, skrbeti za, poskrbeti za, zavzemati se za, potegovati se za, prizadevati si za* itd.

Z vidika obnašanja v širšem stavčnem vzorcu je za IAV značilno, da prisotnost predložnega morfema pogosto spremeni vezljivostne lastnosti glagola, npr. (a) ko prvotno neprehodni glagol postane prehodni, tipično s predložnim določilom, kot v primerih *živeti od koga/česa, gre za koga/kaj*, (b) ko pride do spremembe sklona predložnega določila, *obrniti se na koga : obrniti se h komu, spoznati se na kaj : spoznati se s kom, klicati po kom/čem : klicati koga/kaj*. Prepoznati je bilo mogoče tudi številne primere glagolov premikanja, ki kot IAV spremenijo pomen v neprostorsko vrednotenje stanja, na primer *priti skozi* – 'preživeti', *hoditi v* – 'obiskovati', *pahniti v* – 'povzročiti, da začne kdo doživljati kaj neprijetnega', *priti/pripeljati/prihajati do* – 'zgoditi se'. Glagolom s širokim pomenskim obsegom predložni morfem tipično zoži pomen, kot v primerih *biti za, govoriti za, imeti proti* ipd. Treba pa je omeniti tudi glagole, ki jim v pomenu znotraj IAV obvezno sledi abstraktni predmet, kot npr. *pasti v (nemilost, depresijo, vrtinec nizkotnosti), dišati po (prevari), pokati od (od veselja), postreči z (zanimivostmi)*.

Prepoznavanje predložnomorfemskih glagolov predstavlja izziv tako za označevalce kot za strojno učenje, saj se med leksikalizirani morfem in glagol lahko vrivajo druge besede, poleg tega pa številne zveze glagola s predlogom niso leksikalizirane, npr. *pasti v luknjo/na tla/pod vlak/čez previs*, lahko izkazujejo dobesedni pomen ob tem da ohranjajo nespremenjen tudi sklon predmetnega določila, npr. *stati za (vrati)* – 'nahajati se' : *stati za (dejanji)* – 'podpirati', in so hkrati lahko tudi večpomenske, npr. *priti do (spremembe)* – 'zgoditi se' in *priti do (denarja)* – 'dobiti'.

Analiza predstavlja izhodišča za strojno prepoznavanje tovrstnih enot, hkrati pa ponuja možnosti za bolj poglobljene raziskave, zlasti na ravni vezljivosti, prepoznavanja stavčnih vzorcev in pomenskih lastnosti udeležencev.

5.4. Glagolski idiomi (VID)

Smernice opredeljujejo glagolske idiome[3] (VID) kot zvezo dveh leksikaliziranih sestavin, pri katerih glagol predstavlja skladenjsko jedro, ki predvideva vsaj enega udeleženca znotraj stavčnega vzorca. Udeleženci imajo lahko različne skladenjske vloge, npr. neposrednega ali predložnega predmetnega določila, *plačati ceno*, *zravnati z zemljo*, osebka, *stara zgodba se ponavlja*, prislovnega določila, *spati kot ubit*, odvisnega stavka, *vedeti, koliko je ura*, itd. Poleg omenjenega, mora taka zveza izkazovati tudi samostojen pomen, kar pomeni, da mora ob določenih spremembah skladenjskih in pomenskih funkcij ohranjati svoj pomen. Kot nabor takih sprememb, ki zvezi ohranjajo pomen, Smernice navajajo možnost pojavljanja sestavin v predvidenih paradigmah (sklanjatveni in spregatveni), tvorjenje časov, tvorjenje aktivnih in pasivnih zgradb, leksikalno variantnost itd.

Definicija znotraj Smernic Parseme se od slovenske razlikuje v tem, da obravnava GVE kot glagolsko jedro stavka, ki predvideva leksikalizirane elemente znotraj svojega stavčnega vzorca – pomenskoskladenjski pristop, medtem ko se v slovenski literaturi izpostavlja predvsem zmožnost opravljanja povedkove funkcije zveze kot celote (Toporišič, 1973/74; Kržišnik, 1994) – funkcijsko-skladenjski pristop. S tega vidika so v slovenščini problematične zveze, ki sicer vključujejo glagol kot ustaljeni del, vendar kot celota ne nastopajo nujno le v vlogi povedka, pač pa tudi v vlogi predmetnega določila, *(ne spodobi se) voditi za nos*, ali v vlogi stavka (*srce se trga (komu)*).

V učnem korpusu je bilo kot GID označenih 724 enot, kar predstavlja 22 % vseh GVE (gl. Tabelo 1). GID z več kot 10 pojavitvami po pričakovanju vključujejo glagol *biti* (tudi *imeti*), določilo pa je glede na besednovrstno opredelitev izmuzljivo, saj se lahko hkrati pojavlja v prislovni in členkovni, pridevniški ali samostalniški funkciji, npr. *biti jasno, biti si na jasnem; biti žal, biti stvar (koga/česa).* Z več kot 5 pojavitvami najdemo še *biti kos, biti prav; priti prav*, *igrati vlogo, pustiti pri miru, priskočiti na pomoč* in *imeti opravka s/z* ter t. i. ustaljene diskurzne označevalce (prim. Dobrovoljc, 2017): *kot se pravi, se pravi, kdo ve*. Glagoli, ki tipično tvorijo različne VID, so poleg *biti* in *imeti* še *vzeti, postaviti, priti, dati* in *iti* (v 10 ali več različnih VID). Med samostalniškimi sestavinami z več kot 10 pojavitvami izstopata *roka* in *glava* ter *beseda, nič, stran* in *vrata* z vsaj 5 pojavitvami v različnih VID.

Kot omenjeno, po frekventnost strukture izstopajo zveze glagola *biti* in prislova/pridevnika/samostalnika, ki jih je glede na strukturno ustaljenost in pomensko izpraznjenost glagola smiselno obravnavati kot ustaljene glagolske zveze oz. leksikonske enote (*biti všeč/res/mar/prida/prav/kos, biti jasno/žal/narobe/stvar/moč*), manj pa se zdi smiselno na podlagi njihove distribucijske omejenosti na pomožnik odpirati samostojno besedno vrsto – povedkovnik (Toporišič, 2000; Žele, 2011). V to skupino sodijo tudi zveze s pomensko širokim *imeti*: *imeti prav/rad, ne imeti pojma/smisla, imeti smisel za* ipd.

Druga struktura, ki je opazno zastopana v učnem korpusu, je zveza glagola in samostalnika oz. samostalniške besedne zveze. Med glagoli izstopata *delati* (*delati družbo/gužvo/izjeme/preglavice/razlike/sceno/škodo*) in *dati* (*dati košarico, dati polet, dati pečat, dajati videz* ipd.), ki strukturno sovpadajo z LVC, vendar ne prenesejo

[2] Ob IAV navajamo še tipične kolokatorje na podlagi korpusa Gigafida za lažje ustrezno pomensko razdvoumljanje.

[3] S tem izrazom se oddaljujemo od slovenske tradicije, ki bi na tem mestu uporabila izraz *glagolski frazemi*. Razlogi so predvsem v drugačnem dojemanju skladenjske vloge glagolske sestavine, kot je pojasnjeno v prispevku.

določenih pretvorb, ki jim sicer podlegajo LVC, npr. izražanje svojine, ki se pri LVC ohranja: *Miha ima predavanje → Mihovo predavanje*, pri VID pa taka pretvorba ob ohranitvi pomena ni mogoča: *Miha dela družbo/gužvo* ipd. → **Mihova družba/preglavice*. Tipično se samostalniška zveza razširja s pridevnikom, ki je bodisi leksikaliziran, *imeti polne roke, zadati smrtni udarec*, varianten: *ubrati drugo/drugačno pot, preteči veliko/dosti vode*, ali zgolj tipičen vrivek v sicer ustaljeno glagolsko zvezo: *služiti si (vsakdanji, nogometni* ipd.) *kruh*. Največji delež predstavljajo v učnem korpusu VID s strukturo glagola in predložne zveze, kjer spet izstopa *biti*, npr. *biti na dosegu roke, biti v konfliktu, biti na preizkušnji, biti na razpolago/voljo, biti na tleh, biti na udaru, biti pod pritiskom, biti pri srcu, biti pri stvari* ipd., sicer pa so zastopani tudi drugi glagoli, npr. *postaviti ob bok, potegniti na dan, priti na dan, dati na izbiro, dati na led, priti do izraza, priti na misel, voditi/vleči za nos, trkati na vrata, stati ob strani, pasti v oči, požirati z očmi, zavijati z očmi*, *postaviti/postavljati na stranski tir, škrtati z zobmi* ipd., sem pa smo šteli tudi primere kot *vzeti (kaj) nase, postaviti se zase, obdržati (kaj) zase* ipd. Zlasti za zveze glagola s samostalniško zvezo in s predložno samostalniško zvezo je z vidika ustaljenosti treba opozoriti na obvezno ali tipično zanikanje, npr. *ne moči*[4] *(komu) do živega, ne moči si kaj, (kaj) ni po godu (komu), (kaj) ne gre v račun (komu), ni ne duha ne sluha o (kom/čem), ni para (komu)*, *ne gre iz glave (komu)* ipd.

V manjšem deležu so v učnem korpusu zastopane tudi druge strukture, npr. stavčne: *solze stopijo v oči (komu), oči so večje od želodca, noge nesejo (koga), stara zgodba se ponavlja, kamen se odvali od srca (komu)*, *časi se spreminjajo, vrabci že čivkajo*, tudi v obliki pregovorov, npr. *bolje preprečiti kot zdraviti, samo osel gre dvakrat na led*, in primerjav: *igrati se (s kom/čim) kot mačka z mišjo, delati (s kom/čim) kot svinja z mehom, steči kot namazano* ipd., zveze glagola in prislova, npr. *priti skupaj, daleč priti, iti predaleč, narediti svoje, ustreliti mimo, imeti zadosti, dobro iti*, ter zveze glagola in zaimenskega morfema, *zagosti jo (komu), ubrati jo, mahniti jo* ipd.

VID se v stavčni vzorec vključujejo na različne načine. Glagolske zveze odpirajo predvidljiva skladenjska mesta, ki jih zapolnjujejo udeleženci s svojimi tipičnimi pomenskimi vlogami, kot smo nakazali pri posameznih primerih zgoraj. Že ob hitrem pregledu primerov v korpusu je mogoče zaznati tudi ustaljenost ali večjo pogostnost nekaterih glagolskih oblik (npr. 3. oseba, zanikanje) pa tudi predvidljivost zapolnitev udeleženskih mest na leksikalni ravni.

V naši raziskavi stavčni vzorci, ki jih narekujejo VID (in druge kategorije GVE), niso bili sistematično raziskani, je pa v ta namen mogoče uporabiti podatke, ki jih vsebuje učni korpus na skladenjski in semantični ravni. V prvem primeru s formaliziranimi skladenjskimi povezavami, v drugem pa s pripisom semantičnih vlog udeležencem na teh mestih. Na ta način bi bilo mogoče identificirati širše stavčne vzorce za posamezni tip GVE in jih uporabiti pri nadaljnjem strojnem luščenju.

6. Razprava in zaključek

Kategorizacija glagolskih večbesednih enot na podlagi Smernic Parseme 1.1 ima dva osnovna namena: (a) določiti merila za prepoznavanje glagolskih večbesednih enot, ki jih je smiselno pri jezikovnem opisu (slovar, slovnica) in strojnem luščenju obravnavati kot celote, tj. elemente leksikona, ter (b) formalizirati opis v skladu z večjezikovno primerljivimi merili.

Na podlagi označenih glagolskih večbesednih enot v učnem korpusu je bilo mogoče izdelati prve kvantitativne in kvalitativne analize za posamezno kategorijo in na njihovi podlagi prepoznati določene načine vzorčenja na skladenjski in pomenski ravni. Ti vzorci predstavljajo dobro izhodišče za izdelavo pravil pri strojnem luščenju GVE in za nadaljnje jezikoslovne opise. Metodološko gledano gre tudi za preusmeritev fokusa s funkcijskoskladenjskega vidika v opis medsebojno povezanih lastnosti na oblikoskladenjski, skladenjski, pomenski in leksikalni ravni.

Glagoli, ki tipično tvorijo GVE, so po pričakovanju glagoli z zelo širokim pomenskim obsegom, npr. *biti, dati, imeti*, zaradi česar izgubljajo svoje leksikalne, ohranjajo pa morfološke lastnosti, skladenjsko funkcijo in pozicijo v stavčnem vzorcu. Pomenska udeleženost glagola v odnosu do posameznih sestavin v zvezi kot celoti je pogosto težko določljiva zaradi velike pogostnosti glagolskih zvez, med katerimi številne ne izkazujejo idiomatičnega branja, prim. zveze z glagoli v pomensko oslabljeni rabi, inherentno povratne glagole in glagole z leksikaliziranim predložnim morfemom. Zaradi tega jih je v tekočem besedilu težko ločiti od prostih zvez pa tudi od kolokacij, ki so frekventne pomensko smiselne in strukturno pravilne besedne povezave. Seznam GVE, ki smo jih prepoznali v korpusu, tako že predstavlja nabor leksikonskih enot, ki jih je kot take mogoče uporabiti pri avtomatskem prepoznavanju v besedilu in nadaljnjem strojnem učenju mehanizmov.

Na drugi strani so prve strukturne in pomenske analize pokazale, da (a) posamezni tipi GVE tvorijo prepoznavne strukturne vzorce, npr. glagol + samostalniška zveza, zlasti predložna, da (b) leksikalizacija elementov vpliva na spremembe v udeleženskih mestih in njihovih semantičnih vlogah, npr. *vreči se po kom – vreči se v kaj – ven se vreči – vreči koga ven* ipd., (c) da je npr. variantnost glagolov predvidljiva z vidika dovršnih in nedovršnih parov, *plačati/plačevati ceno, dati/dajati si opravka s čim*, (č) da zaporedje glagolskih sestavin v posamezni GVE navadno ni ustaljeno, se pa (d) kažejo določene tendence v besednem redu ter (e) številu in zastopanosti vrinjenih elementov, kot tudi, da je (f) na določene leksikalne zapolnitve mogoče sklepati iz frekvenčnih podatkov in elementov besedilnega okolja, ter da je (g) za lažje strojno prepoznavanje GVE smiselno v formaliziran opis vključiti informacije na vseh ravneh korpusne označenosti. V nadaljevanju raziskav, usmerjenih v prepoznavanje večbesednih enot z glagolskim jedrom, bomo zato upoštevali vse vrste podatkov, ki so v zvezi s posameznimi besedami in zvezami v korpusu že na voljo, tj. oblikoskladenjsko označenost, skladenjsko razčlenjenost in pripis pomenskih vlog stavčnim udeležencem.

Za ustrezno identifikacijo različnih VE v jeziku bomo v nadaljevanju izdelali tudi tipologijo večbesednih enot, ki ne

[4] V primeru, da je realizacija izključno vezana na 3. osebo, je to razvidno tudi iz osnovne oblike VID. Nedoločnik v osnovni obliki ohranjamo takrat, ko glagol predvideva tudi prvo- in drugoosebne osebke.

predvidevajo glagolskega jedra, kot npr. ustaljene samostalniške zveze tipa *žlahtna kapljica, kaplja v morje, domača tla*, kjer predstavlja poseben izziv ugotavljanje trdnosti povezave z glagolom, kot npr. *(kaj) je kaplja v morje*, (*zmagati, odigrati) na domačih tleh, (finale) na domačih tleh*. Poleg tega predstavlja v nadaljevanju izziv tudi prepoznavanje VE s samostojnim, vendar ne metaforičnim pomenom, npr. *formula ena, velika začetnica, druga svetovna vojna*, ki se na eni strani približujejo terminološkim na drugi pa lastnoimenskim enotam.

Pri identifikaciji VE bo pozornost treba nameniti tudi ustaljenim skladenjskim strukturam, ki sicer niso pomensko samostojne, imajo pa predvidljivo zgradbo in opravljajo samostojno skladenjsko vlogo, npr. *v času od do*, kamor sodijo tudi številne ustaljene predložne zveze kot npr. *med drugim, v celoti, v skladu z/s, po besedah* ipd. ter t. i. besedilni povezovalci, ki so opazna sestavina tako pisne kot govorne komunikacije (Dobrovoljc, 2017).

7. Zahvala

Raziskava je potekala v okviru projekta ARRS J6-8256 *Nova slovnica sodobne standardne slovenščine: viri in metode* ter sodelovanja v okviru IC1207 PARSEME COST Action[5] in IS1305 ENeL COST Action.[6]

8. Literatura

Špela Arhar Holdt in Vojko Gorjanc. 2007. Korpus FidaPLUS: nova generacija slovenskega referenčnega korpusa. *Jezik in slovstvo* 52(2): 95–110.

Sue B. T. Atkins, in Michael Rundell. 2008. *The Oxford Guide to Practical Lexicography*. New York, Oxford University Press.

Timothy Baldwin in Su Nam Kim. 2010. 'Multiword Expressions' V *Handbook of Natural Language Processing*, Second Edition, str. 267–292, CRC Press, Boca Raton, USA.

Archna Bhatia, Claire Bonial, Marie Candito, Fabienne Cap, Silvio Cordeiro, Vassiliki Foufi, Polona Gantar, Voula Giouli, Carlos Herrero, Uxoa Iñurrieta, Mihaela Ionescu, Alfredo Maldonado, Verginica Mititelu, Johanna Monti, Joakim Nivre, Mihaela Onofrei, Viola Ow, Carla Parra Escartín, Manfred Sailer, Carlos Ramisch, Renata Ramisch, Monica-Mihaela Rizea, Agata Savary, Nathan Schneider, Ivelina Stonayova, Sara Stymne, Ashwini Vaidya, Veronika Vincze in Abigail Walsh. 2017. *PARSEME shared task 1.1 annotation guidelines* (last updated on November 30, 2017). http://parsemefr.lif.univ-mrs.fr/parseme-st-guidelines/1.1/.

Marie Candito, Fabienne Cap, Silvio Cordeiro, Vassiliki Foufi, Polona Gantar, Voula Giouli, Carlos Herrero, Mihaela Ionescu, Verginica Mititelu, Johanna Monti, Joakim Nivre, Mihaela Onofrei, Carla Parra Escartín, Manfred Sailer, Carlos Ramisch, Monica-Mihaela Rizea, Agata Savary, Ivelina Stonayova, Sara Stymne in Veronika Vincze. 2016. *PARSEME shared task 1.0 annotation guidelines - version 1.6b* (last updated on November 26, 2016). http://parsemefr.lif.univ-mrs.fr/parseme-st-guidelines/1.0/.

Kaja Dobrovoljc. 2017. Multi-word discourse markers and their corpus-driven identification: the case of MWDM extraction from the reference corpus of spoken Slovene. *International journal of corpus linguistics*, 22(4), 551–582.

Kaja Dobrovoljc, Simon Krek in Jan Rupnik. 2012. Skladenjski razčlenjevalnik za slovenščino. V *Zbornik Osme konference Jezikovne tehnologije*, str. 42–47, Ljubljana, Institut Jožef Stefan.

Polona Gantar, Simon Krek in Taja Kuzman. 2017. Verbal multiword expressions in Slovene. *Europhras 2017*, str. 247–259. Springer.

Lara Godec Soršak. 2013. Glagoli z oslabljenim pomenom v Slovarju slovenskega knjižnega jezika. *Slavistična revija*: 61(3): 507–522.

Vojko Gorjanc, Polona Gantar, Iztok Kosem in Simon Krek (ur.). 2015. *Slovar sodobne slovenščine: problemi in rešitve*. Ljubljana, Znanstvena založba Filozofske fakultete.

Miha Grčar, Simon Krek in Kaja Dobrovoljc. 2012. Obeliks: statistični oblikoskladenjski označevalnik in lematizator za slovenski jezik. V *Zbornik Osme konference Jezikovne tehnologije.* Ljubljana, Institut Jožef Stefan.

Simon Krek, Kaja Dobrovoljc, Tomaž Erjavec, Sara Može, Nina Ledinek, Nanika Holz, Katja Zupan, Polona Gantar in Taja Kuzman. 2017. Training corpus ssj500k 2.0, *Slovenian language resource repository CLARIN.SI*, http://hdl.handle.net/11356/1165.

Simon Krek, Kaja Dobrovoljc, Tomaž Erjavec, Sara Može, Nina Ledinek, Nanika Holz, Katja Zupan, Polona Gantar, Taja Kuzman, Jaka Čibej, Špela Arhar Holdt, Teja Kavčič, Iza Škrjanec, Dafne Marko, Lucija Jezeršek in Anja Zajc. 2018. Training corpus ssj500k 2.1, *Slovenian language resource repository CLARIN.SI*, http://hdl.handle.net/11356/1181.

Erika Kržišnik. 1994. *Slovenski glagolski frazemi (ob primeru glagolov govorjenja). Doktorska disertacija*. Filozofska fakulteta, Univerza v Ljubljani.

Jože Toporišič. 2000. *Slovenska slovnica*. Maribor, Založba Obzorja.

Jože Toporišič. 1982. *Nova slovenska skladnja*. Ljubljana, Državna Založba Slovenije.

Jože Toporišič. 1976. *Slovenska slovnica*, Maribor, Obzorja.

Jože Toporišič. 1973/74. K izrazju in tipologiji slovenske frazeologije. *Jezik in slovstvo* (8): 273–279.

Ada Vidovič-Muha. 1998. Pomenski preplet glagolov imeti in biti – njuna jezikovnosistemska stilistika. *Slavistična revija* [na spletu], 46(4): 293–323.

Andreja Žele. 2002. Prostomorfemski glagoli kot slovarska gesla. *Jezikoslovni zapiski* 8(1), 95–108.

Andreja Žele. 2012. *Pomensko-skladenjske lastnosti slovenskega glagola*, (Zbirka Linguistica et philologica, 27). Ljubljana, Založba ZRC, ZRC SAZU.

Andreja Žele. 2011. Povedkovnik kot skladenjska in slovarska kategorija. *Jezikoslovni zapiski*, 17(1): 27–34.

[5] http://www.parseme.eu.

[6] http://www.elexicography.eu.