Zapis kanonične oblike frazeoloških enot v Leksikonu večbesednih enot za slovenščino

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

Abstract

This paper discusses the rules for recording the canonical form of phraseological units (PhUs) as an independent type of multiword units (MWUs) in the newly created Multiword Expressions lexicon, which is an integral part of the Slovene Digital Dictionary Database intended for creating the online Dictionary of Modern Slovene. First, we briefly describe different types of MWUs and how they were included in general dictionaries of the Slovene language, and then establish the relationship between the terms: dictionary form, basic form, lemma and canonical form. The latter represents the record of the basic unit in the machine-readable Multiword Expressions lexicon, which is defined in terms of the number and sequence of components, syntactic relations between components and their morphological properties. Based on the extracted data for a pre-selected list of PhUs, we create a system of semantically interconnected variant and transformational PhUs and present concrete solutions on selected examples.

Ključne besede: leksikon večbesednih enot, kanonična oblika frazeološke enote, Digitalna slovarska baza

Keywords: multiword expressions lexicon, canonical form of phraseological units, Digital Dictionary Database

1 Uvod

Večbesedne enote (VE) predstavljajo obsežen del slovarjev, saj so tako kot posamezne besede nosilke pomena v najširšem smislu – ne samo kot enote z leksikalnim pomenom, ampak tudi kot enote, ki vsebujejo kulturološke posebnosti in imajo lahko specializirane komunikacijske vloge. Po nekaterih podatkih predstavljajo VE enako količino besedišča določenega jezika kot enobesedne (Jackendoff 1997: 156), hkrati pa so produktivne tudi pri nastajanju nove leksike in pri prevzemanju iz drugih jezikov (Gantar et al. 2018a).

Zaradi večbesednosti in semantičnih lastnosti, ki jih imajo kot celota, so VE vse bolj pomembne tudi za računalniško procesiranje naravnega jezika in njihovo avtomatsko prepoznavanje v besedilu. Ta pomembnost izhaja iz dejstva, da večbesednost omogoča več fleksibilnosti posameznih komponent in enote kot take. Izziv tako za jezikoslovni kot računalniški del predstavlja dejstvo, da VE za razliko od besed vzpostavljajo tudi skladenjsko razmerje med sestavinami, zahtevajo prilagajanje sestavin znotraj zveze morfološkim pravilom in lahko posamezne sestavine zamenjujejo ali mednje vrivajo druge besede. Z vidika avtomatskega luščenja predstavljajo VE problem tudi zato, ker lahko oblikovno sovpadajo s prostimi besednimi zvezami, ki ne izkazujejo celostnega pomena, npr. *čakati na zeleno luč*, *prebiti led* ipd. Vse te lastnosti delajo večbesedne enote težje prepoznavne v besedilu, ko govorimo o njihovem avtomatskem procesiranju, in težje ulovljive v abstraktni zapis, ko govorimo o njihovem prikazovanju v slovarju.

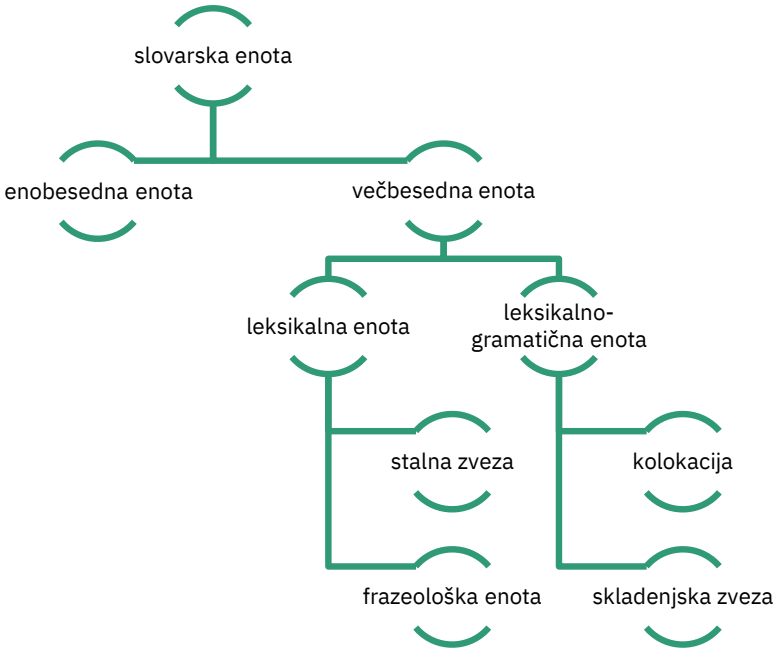
V prispevku najprej predstavimo različne tipe VE, ki smo jih identificirali kot potencialne enote za vključitev v Leksikon, ki bo predstavljal integralni del Digitalne slovarske baze, namenjene izdelavi spletnega Slovarja sodobnega slovenskega jezika (Gorjanc et al. 2015). Nato opišemo načine vključevanja različnih VE v nekatere splošne slovarje za slovenščino ter njihov zapis. Osrednji del prispevka namenimo obravnavi zapisa FE v kanonični obliki v Leksikonu VE. Najprej opredelimo izraz kanonična oblika glede na osnovno obliko ali lemo in glede na slovarsko obliko FE. V nadaljevanju

opišemo postopek izdelave Leksikona VE, in sicer njegovo zgradbo, pripravo izhodiščne liste FE, postopek avtomatskega luščenja iz korpusa in analizo izluščenih primerov, katere cilj je izdelati sistem medsebojnega povezovanja variantno in pretvorbeno povezanih FE v slovarski bazi. Rešitve prikažemo na posameznih primerih, ki smo jih upoštevali pri izdelavi Leksikona. Prispevek zaključimo s temeljnimi ugotovitvami in smernicami za nadaljnje delo.

2 Tipologija večbesednih enot

Pri zasnovi pravil za oblikovanje zapisa kanonične oblike VE v Leksikonu smo izhajali iz tipologije, kot smo jo oblikovali pri izdelavi Leksikalne baze za slovenščino (Gantar 2015) in uporabili pri izdelavi digitalnih slovarskih virov za slovenščino (Gantar et al. 2021). Pri izgradnji slovarske baze smo slovarske enote, tj. enote, ki predvidevajo določene slovarske informacije (Slika 1), z vidika zgradbe opredelili glede na eno- in večbesedne, zadnje pa še glede na to, ali predvidevajo opis pomena ali ne. V prvi skupini so **večbesedne leksikalne enote**, katerih pomen je več kot vsota pomenov njihovih sestavin (Rundell 2008: 168), zaradi česar potrebujejo razlago v slovarju, v drugi pa **večbesedne leksikalno-gramatične enote**, ki v slovarju niso nujno predmet pomenskega opisa, lahko pa predvidevajo kake druge slovarsko relevantne informacije, npr. opis skladenjske ali besedilne vloge, npr. *ne glede na* – veznik; *kot rečeno* – besedilni povezovalec. Kot leksikalno-gramatične enote obravnavamo tudi kolokacije, katerih vloga je v slovarju dvojna: prikazati tipično sobesedilno rabo, ki je značilna za naravni govor maternih govorcev, in razdvoumljati pomene večpomenskih besed, npr. *[češka, norveška, danska …] krona : [trnova, briljantna …] krona : [zobna] krona.*

Večbesedne leksikalne enote smo nadalje razdelili glede na to, ali je njihov celostni pomen poimenovalen ali pa ima primarno ekspresivno oz. vrednotenjsko vrednost. Prvo skupino sestavljajo t. i. **stalne zveze** tipa *topla greda, varnostni trikotnik, črna luknja* ipd., ki navadno sodijo na določeno strokovno področje, zlasti na prehajanju v splošni jezik (prim. Krek et al. 2021b), ko govorimo o splošnem



Slika 1: Delitev slovarskih enot glede na zgradbena in pomenska merila.

referenčnem korpusu standardnega jezika Gigafida 2.0 (Krek et al. 2020a). Stalnih zvez ni mogoče vedno nedvoumno ločevati od kolokacij, zlasti v primeru relativne pomenske transparentnosti, npr. *solatni bife, letni dopust, tuji jezik*. Osnovno merilo za ločevanje stalnih zvez od kolokacij zato ostaja leksikografova presoja, ali zveza potrebuje razlago (stalna zveza) ali ne (kolokacija).

Drugi tip večbesednih leksikalnih enot predstavljajo **frazeološke enote** (FE), ki imajo poleg celostnega pomena tudi ekspresivno vlogo, največkrat doseženo po metaforični ali metonimični poti. Kot take predstavljajo FE tisti segment leksike, ki služi za slikovito, ne-nevtralno izražanje. Z drugimi besedami, FE so vedno rezultat govorčeve intence povedati kaj drugače, bolj opazno. V pričujoči razpravi se bomo ukvarjali z zapisom kanonične oblike samo pri tem tipu VE.

Posebej je treba omeniti heterogeni tip t. i. leksikalno-gramatičnih enot, katerih skupni imenovalec je poleg večbesednosti tudi

smiselnost njihovega vključevanja v slovar zaradi tipičnih vlog, ki jih opravljajo v besedilu (povezovanje, izražanje okoliščin, stopnje ipd.). Poleg **kolokacij** in razširjenih kolokacij lahko tu izpostavimo še **zveze s pomensko oslabljenimi glagoli**, npr. *imeti pogum, dati na razpolago*, t. i. **skladenjske zveze** tipa, *pod okriljem (koga/česa), v nasprotju z/s (kom/čim), za razliko od (koga/česa)*, **predložne glagole**, npr. *gre za (koga/kaj), priti do (koga/česa)*, in **inherentno povratne glagole**, kot so: *zdeti se, delati se* itd. Za zadnji dve skupini je značilno, da jih lahko prepoznavamo tudi kot enote s samostojnim leksikalnim pomenom (prim. Gantar et al. 2021, Gantar et al. 2019b).

3 Obravnava večbesednih enot v splošnih slovarjih za slovenščino

Slovarji vključujejo različne tipe VE, v različnih obsegih, na različnih mestih geselske zgradbe in z različnimi slovarskimi informacijami. Poleg pomenskih opisov, ki predstavljajo pri frazeoloških enotah samostojen izziv, zlasti v smislu pomenske razpršenosti in vključevanja pragmatičnih informacij, se slovarji pri obravnavi VE soočajo predvsem s tremi vprašanji: katere tipe VE vključiti v slovar, kako oz. kam VE vključiti v slovarsko makro- oz. mikrostrukturo in v kakšni obliki jih navesti kot slovarske enote.

3.1 Tipi večbesednih enot v splošnih slovarjih

Ključno merilo za vključitev določene VE v slovar je pomen. Na splošno je mogoče reči, da slovarji vključujejo predvsem tiste VE, katerih pomen je več kot vsota pomenov posameznih sestavin.[1] Izhajajoč iz naše tipologije gre predvsem za stalne zveze, frazeološke in paremiološke enote ter pragmatične izraze tipa *kapo dol, saj nisem na glavo padel* ipd. Čeprav je prepoznavanje pomenske samostojnosti zveze kot celote leksikografsko gledano relativen kriterij, ki je v prvi vrsti odvisen od

1 Načeloma se upošteva dejstvo, da pomena celote ni mogoče razbrati iz pomenov posameznih sestavin, pri čemer je pomensko razmerje med sestavinami VE glede na njen celostni pomen lahko različno interpretirano (prim. Atkins in Rundell 2008: 168, Gantar et al. 2019a: 144).

lastnosti in namena slovarja ter vsakokratne leksikografove presoje, je, kot pravita Atkins in Rundell (2008: 167) potreba po razlagi še vedno najbolj uporabno merilo za odločanje glede tega, katere večbesedne enote vključiti v slovar in kam jih znotraj slovarja umestiti.

Vključenost VE, ki niso prepoznane kot enote z leksikalnim pomenom, je v splošnih slovarjih različna. Nekateri slovarji vključujejo kolokacije kot poseben tip primerov rabe (v SSKJ t. i. iztržki), ali pa so skladenjske zveze, če so v slovar vključene, prikazane znotraj tipičnih zgledov z opozorili kot »v zvezi«, »s predlogom« ipd., kot prikazuje obravnava zvez *pod okrilje (koga/česa)* in *pod okriljem (koga/česa)* v SSKJ2 kot dela slovarskega zgleda (podčrtano):

okrílje -a s, rod. mn. okrílij in okrílj (ȋ) s predlogom
1. knjiž. *varstvo, zaščita:* iti iz mesta pod okriljem vojaške enote; biti pod okriljem zidov / zateči se pod okrilje močnejšega
// *pokroviteljstvo:* vzeti mladega pesnika pod svoje okrilje; sklicati posvetovanje pod okriljem Unesca

Primer 1: Obravnava zvez *pod okrilje (koga/česa)* in *pod okriljem (koga/česa)* v SSKJ2.[2]

3.2 Umestitev večbesednih enot v slovarsko makrostrukturo

Umestitev VE v slovarsko makrostrukturo je tesno povezana z organizacijo slovarske baze in z načinom prikazovanja oz. dostopanja slovarskih uporabnikov do večbesednih enot v slovarju. Odločitve v zvezi z obravnavanjem VE v slovarski bazi zahtevajo teoretično-metodološki premislek na jezikoslovni strani, ki mora biti usklajen s tehničnimi rešitvami v Digitalni slovarski bazi ter z iskalnimi možnostmi in strategijami, ki jih uporabljajo uporabniki pri iskanju VE prek slovarskih vmesnikov.

Splošni slovarji vključujejo večbesedne enote v slovarsko makrostrukturo predvsem glede na strukturna in pomenska merila. Strukturno gledano, so večbesedne enote vedno zveze dveh ali več besed, pri čemer so te besede, zlasti ko govorimo o splošnih

2 Vir: www.fran.si, dostop 15. 11. 2021.

slovarjih, v slovarjih navadno že obravnavane kot iztočnice. Znotraj iztočnic so VE obravnavane tipično v samostojnih razdelkih, ki so namenjeni določenemu tipu večbesedne enote, npr. frazeološko gnezdo, terminološko gnezdo, grafična ločitev ipd. Drugi tipi večbesednih enot, kot so denimo ustaljene zveze s pomensko izpraznjenimi glagoli, prislovne in predložne zveze, po navadi v slovarjih ne nastopajo kot slovarske enote (glej zgoraj primer za *okrilje* v SSKJ2).

Tak način povezanosti sestavin večbesedne enote z večbesedno enoto kot celoto kot tudi način medsebojne povezanosti posameznih VE, ki temelji na hierarhiji in je zasnovana na logiki tiskanega medija, zahteva v relacijski podatkovni bazi drugačen pristop. V e-slovarjih, tako splošnih kot specializiranih, ki temeljijo na strukturiranih digitalnih bazah z vključenimi različnimi slovarskimi in drugimi jezikovnimi podatki, obstaja trend obravnavanja večbesednih enot kot samostojnih slovarskih enot oz. iztočnic z različnim statusom, pri čemer je ključno prepoznavanje enot s pomenom (prim. Tavast et al. 2018) ne glede na njihovo eno- ali večbesednost. V slovarski bazi je zato pomembna predvsem njihova prepoznavnost v smislu pomenske enote, saj to omogoča tudi povratno pridobivanje iz korpusa in povezljivost pomenov na katerikoli ravni: na ravni sestavin VE, oblik, skladenjske zgradbe, in semantičnega tipa.

4 Osnovna oblika večbesedne enote v korpusu, slovarju in slovarski bazi (leksikonu)

Izraz *kanonična oblika*, kot ga uporabljamo v prispevku in nam pomeni zapis (večbesedne) enote v leksikonu, ki je določen s formalno (tj. strojno berljivo) opredelitvijo sestavin ter razmerij med njimi, moramo opredeliti glede na izraz *osnovna oblika* ali *lema*, ki je določena v korpusu na podlagi oblikoskladenjskih kategorij, ter glede na izraz *slovarska oblika*, ki je oblika iztočnice v slovarju in ne sledi nujno korpusni lemi. Izraza slovarska in kanonična oblika imata po svoji definiciji podobno vlogo, saj opredeljujeta zapis VE v slovarskih virih, razliko, ki jo vzpostavljamo med njima, pa upravičujemo z dejstvom, da je leksikonska baza poleg slovarske namenjena tudi strojni rabi in

tem, da pravila ki opredeljujejo zapis slovarske oblike VE v obstoječih splošnih slovarjih, ne upoštevajo skladenjskih razmerij med sestavinami VE in njihovih oblikoskladenjskih lastnosti. Ko v prispevku govorimo o kanonični obliki, nimamo v mislih podrejanja različnih variant in pretvorb nadrejeni obliki, kot je to značilno za obravnavane slovarje, pač pa obravnavamo vse leksikonske enote na istem nivoju, pri čemer mora njihov zapis slediti pravilom, ki jih podrobneje opišemo v nadaljevanju.

Slovarska oblika[3] je torej tista oblika, v kateri beseda ali zveza nastopa v slovarski iztočnici. Pri pregibnih besednih vrstah veljajo splošna pravila glede nabora slovničnih kategorij, ki so zastopane v slovarski obliki. Pri samostalnikih je to navadno imenovalnik ednine, pri glagolih nedoločnik in pri pridevnikih moški spol ednine ter navadno nedoločna oblika (prim. SSKJ2 Uvod). V korpusnem jezikoslovju se za osnovno obliko besede, ki naj bi zastopala različne morfološke oblike pregibnih besed, uporablja izraz lema, ki pa se uporablja kot termin tudi v leksikografskem procesu. Vendar pa je – izhajajoč iz različnih jezikovnih posebnosti – lahko interpretacija leme v korpusih posameznih jezikov različna[4] kot tudi ni nujno, da je korpusna lema prekrivna z obliko, ki jo ima beseda v slovarski iztočnici. Odločitve o tem, katere oblike združiti pod krovno lemo, so tako dogovorne, posledično pa vplivajo tudi na luščenje VE iz korpusa na podlagi oblikoskladenjskih oznak in skladenjskih razmerij, določenih v korpusu.

Hkrati je definiranje slovarske oblike pri VE bolj zapleteno kot pri besedah iz več razlogov. Pri VE imamo opraviti z več kot eno besedo, osnovna oblika VE pa ne more biti vsota osnovnih oblik posameznih sestavin, npr. **iti kakor po maslo* za *gre kakor po maslu*, saj besedna zveza v morfološko bogatih jezikih zahteva morfološko prilagajanje

3 Za obravnavo FE v slovarjih je tudi v slovenski literaturi (Kržišnik 1996, 2004, Gantar 2007, Perdih in Ledinek 2019, Meterc 2019) opravljenih več raziskav, ki obravnavajo problem frazeološke variantnosti in oblik, v katerih se FE pojavljajo v besedilih, v odnosu do slovarske oblike, vključno s potrebo po ločevanju tipičnosti na eni strani in individualnosti na drugi, ki navadno ni predmet slovarske obravnave.

4 Lema lahko vključuje tudi povezane oblike znotraj več besednih vrst, npr. *igrati – igra* (npr. v angleškem jeziku ista oblika dve različni besedni vrsti), ali celo izpeljane oblike tipa *igrati – igralec*, tj. različni obliki znotraj različnih besednih vrst.

sestavin znotraj zveze. VE se kot besedne zveze prilagajajo besedilu tudi navzven, s tem ko vstopajo v različne skladenjske vloge: *zdrava pamet – po zdravi pameti, biti zdrave pameti*, predvidevajo »prosta« skladenjska mesta: *zlesti (komu) pod kožo* in prevzemajo različne upovedovalne možnosti, kot je npr. zanikanje, velelnost, prehajanje v stavčno obliko ipd.

S slovarskega vidika se zdi torej nujno vzpostaviti razmerje med slovarsko obliko, »ki jo tvorijo zaporedje in vrsta sestavin, minimalno število sestavin in razmerja med njimi« (Kržišnik 1996: 134 po Filipec in Čermák 1985: 184), in oblikami rabe (t. i. frazeološkimi oblikami; Toporišič 1973/74: 273), s katerimi se večbesedne enote prilagajajo sobesedilu. Take oblikoslovne prilagoditve v slovarjih naj ne bi bile zastopane (Kržišnik 1996: 134). Na drugi strani je oblike rabe, za katere je mogoče presoditi, da so v jeziku ustaljene in hkrati zastopajo pomensko enakovredne bodisi stilno zaznamovane pomene, mogoče obravnavati kot normirane različice izhodiščne oblike (Kržišnik ibid.), in jih obravnavati tudi v slovarju. Za razliko od normiranih frazeoloških variant, Kržišnik (ibid.) loči tudi t. i. modificirane rabe, ki so lahko bodisi ustvarjalne (t. i. prenovitve) bodisi napaka. Zadnje sproža – zlasti v povezavi z obravnavo VE v obstoječih slovarjih za slovenščino – vprašanje, kako prepoznati slovarsko nerelevantno modificirano rabo ali celo napako, še posebej, če je ta razmeroma pogosta. Z vidika avtomatskega luščenja frazeoloških enot na podlagi korpusa se zdi zato ključno upoštevati vse variante in oblike rabe določene FE in šele na podlagi leksikografske analize prepoznati samostojne FE in njihovo medsebojno pomensko povezanost, kot bomo pokazali v nadaljevanju.

4.1 Pravila za zapis večbesedne enote v kanonični obliki

Čeprav je kanonično obliko VE z lastnostjo FE, kot bomo pokazali v nadaljevanju, mogoče določiti šele na podlagi kontekstualne analize pomensko povezanih variant in pretvorb z vsaj delno prekrivnimi sestavinami, je treba za ustrezno identifikacijo skladenjskih struktur prepoznati vzorce, v katerih se FE pojavljajo v besedilnih realizacijah.

Iz teh vzorcev je mogoče izluščiti najtipičnejše in jih zapisati kot leksikonske enote, pri čemer smo sledili načelu, da mora zapis števila, zaporedja in oblike sestavin slediti čim bolj enotnim pravilom, ki se odražajo v kanonični obliki. Enotni vrstni red elementov v kanonični obliki leksikonske enote, kot prikazuje Tabela 1, smo določili z abstraktno stavčno strukturo, v kateri si sestavine sledijo na podlagi predvidljivega zaporedja znotraj glagolskega stavka oz. podredne besedne zveze.

Tabela 1: Vzorčni seznam leksikonskih enot v kanonični obliki v Leksikonu VE.

Samostalnik v osebku	Glagol	Brezpredl. predmet-1	Brezpredl. predmet-2	Predl. predmet	Prislovno določilo
	barvati	(kaj)		s črnimi barvami	
	naložiti	križ	(komu)		
	naložiti	križ			na (čigavo) ramo
	naložiti	križ			na (čigava) ramena
	naložiti	križ	(komu)		na pleča
	nositi	težak križ			
	naložiti	težak križ	(komu)		
(kaj)	ne da	miru	(komu)		
	ne dati	miru			
	ne dati	miru	(komu)		
	ne moči			iz lastne kože	
	ne moči			iz svoje kože	
	ne moči			iz (kakšne) kože	
	ne moči				mimo (česa)
	ne priplavati				po juhi
	priplavati				po juhi
	ne priplavati				po kisli juhi
	priplavati				po kisli juhi
	ne imeti iskati	kaj			(kje)
	ne migniti			niti s prstom	
	niti migniti			s prstom	
(kaj)	pade		(komu)	v naročje	kot zrela hruška

Ob leksikaliziranih sestavinah so v leksikonski enoti posebej označena (z zaimki v oklepaju) predvidena vezljivostna in druga »odprta« mesta. Ta mesta so v leksikonski enoti zapisana, če njihova

prisotnost/odsotnost ali pomenske lastnosti (zajete tudi v slovničnih kategorijah, kot sta npr. živo+/-) vplivajo na pomensko interpretacijo FE. V zaporedju si sledijo po enakih pravilih kot leksikalizirane oz. variantne sestavine. Podrobnejša pravila za zapis posameznih sestavin v kanonični obliki VE navajamo v nadaljevanju.

Samostalnik/samostalniška zveza v osebku

Leksikalne sestavine na osebkovem mestu so tipično samostalniške besede oz. samostalniške zveze v imenovalniku ednine: *čas zaceli rane*; *(kaj) je vrh ledene gore*. Nedoločni zaimek *kaj* na osebkovem mestu je v kanonični obliki leksikonske enote izražen le, če samostalnik na tem mestu ne odraža kategorije živosti: *(kaj) je bob ob steno; (kaj) je na čigavem zelniku zraslo*. Zaimek v kanoničnem zapisu leksikonske enote ni izražen, če na osebkovem mestu lahko nastopajo samostalniki, ki niso omejeni s kategorijo človeško+. V tem primeru sugerira ustrezne realizacije glagol v nedoločniku: *gledati se kot pes in mačka, govoriti steni* vs. **(kaj) govori steni*. Posebnost je zapis glagola *moči*, ki v svoji nedoločniški obliki sugerira delovalnike z lastnostjo živo, a ga kljub temu v kanonično obliki FE navajamo v 3. osebi ednine, ker tak zapis odraža tipično glagolsko obliko in se zdi zaradi tega tudi bolj intuitiven: *(kdo) ne more iz svoje kože* vs. *ne moči iz svoje kože*.

Glagol ali glagolska zveza

Glagolske sestavine v kanonični obliki leksikonske enote tipično navajamo v nedoločniku: *dati možgane na pašo*. Glagolska oblika se prilagodi osebku, kadar je ta v kanonični obliki izražen, njegova prisotnost pa vpliva na pomen FE: *(kaj) ne da miru (komu)* – 'kaj vznemirja koga' vs. *ne dati miru (komu)* – 'kdo nadleguje koga', ali na možnost dobesedne rabe: *(kaj) drži (koga) pokonci* – 'kaj daje komu psihično in moralno podporo' vs. *držati koga pokonci* – 'kdo fizično podpira koga'.

Glagol kot leksikalizirana sestavina FE pa zahteva še druge odločitve glede kanoničnega zapisa v leksikonski enoti, ki izhajajo iz njegovih slovničnih lastnosti. V nekaterih primerih se tako zastavlja

vprašanje uporabe nevtralnega sedanjika nasproti (v nekaterih primerih) tipičnega preteklika ali prihodnjika (*vrag odnese šalo : vrag je odnesel šalo; iz te moke ni/ne bo/ni bilo kruha; za las manjkati – za las je manjkalo*). Zlasti številne pragmatične oz. t. i. besedilne FE potrebujejo načelne odločitve glede zapisa ustrezne oblike glagolske sestavine, npr. *trikrat lahko ugibate/ugibaš; da dol padeš/padete – da padeš dol*, *daj/dajte no mir*, kot tudi glede zaporedja sestavin, npr. *afne guncati – guncati afne; prodajati bučke – bučke prodajati; suhe žemlje ribati – ribati suhe žemlje.* V takih primerih se je sicer mogoče zanašati na najfrekventnejše realizacije, vendar pa včasih najfrekventnejša oblika ni hkrati tudi najbolj povedna za ustrezno uporabo v besedilu, kar je zlasti pomembno pri učencih slovenščine kot tujega ali drugega jezika, zato bi bilo tovrstne primere s tega vidika smiselno preveriti neposredno pri uporabnikih.[5]

Neposredni in posredni predmet

V tej vlogi tipično nastopajo samostalniki ali samostalniške zveze v neimenovalniških sklonih. V zaporedju sestavin dajemo prednost brezpredložnemu predmetu s tipično realizacijo v tožilniku, ki mu praviloma sledi predmet v dajalniku ali predložni predmet v neimenovalniških sklonih. Pravilo smo upoštevali tako pri navajanju leksikaliziranih sestavin FE, kot pri zapolnljivih vezljivostnih mestih, npr. *dati brco (komu) v rit*; *položiti prst (komu) na usta*; *položiti (kaj) (komu) na jezik*.

Okoliščine

Prislovne in samostalniške predložne zveze (podčrtano), ki nastopajo v vlogi prislovnih določil, si v kanoničnem zapisu sledijo za predložnimi določili: *ustreliti <u>v prazno;</u> pustiti (koga/kaj) <u>pri miru</u>; slediti (komu) tesno <u>za petami</u>*. Ta mesta so lahko v leksikonski enoti tudi samo predvidena in izražena z ustreznim zaimkom: *ne imeti (<u>kje</u>) kaj iskati*. Kot je razvidno iz zadnjega primera, smo pri zaporedju

5 Nekatere probleme v zapisu kanonične oblike pri FE je z uporabniškega vidika analizirala Zala Vidic (2021) v svoji magistrski nalogi.

sestavin v primeru prevladujočih realizacij na podlagi korpusa, temu prilagodili tudi zapis.

Modifikatorji

Pridevniške, prislovne in členkovne modifikatorje, npr. *(lasten, svoj … ne, niti, le*), ki so leksikalizirane sestavine FE, v zapisu navajamo pred elementi, ki jih modificirajo (podčrtano), npr. *ne počutiti se dobro v svoji koži, le/samo s prstom migniti, niti s prstom ne migniti*. Predvidene modifikatorje z različnimi leksikalnimi zapolnitvami navajamo z nedoločnim zaimkom v ustreznem sklonu v oklepaju (podčrtano): *postaviti se v (čigavo) kožo; igrati po (čigavih) notah, zaplavati v (kakšne) vode; zaplavati v (katere) vode.*

5 Izdelava Leksikona večbesednih enot

Strojno berljivi leksikoni VE,[6] ki so namenjeni pripravi elektronskih leksikografskih virov in izdelavi naprednih semantično orientiranih jezikovnotehnoloških aplikacij, obstajajo za različne jezike (prim. Ljubešić et al. 2014 za hrvaščino, Bejček in Straňák 2010 za češčino, Tanabe et al. 2014 za japonščino, Fotopoulou et al. 2014, Markantonatou et al. 2019 za grščino, Odijk 2013, Grégoire 2010 za nizozemščino, Ahlén 2013 za švedščino, Smørdal Losnegaard 2019 za norveščino). Pri izdelavi Leksikona VE za slovenščino (Krek et al. 2021a) smo sledili dvema ciljema, izdelati metodologijo za prepoznavanje znanih VE v korpusu ter izdelati model leksikona, v katerem bodo strukturirane v korpusu identificirane VE skupaj z vsemi relevantnimi jezikovnimi podatki. V našem podatkovnem modelu Leksikon VE predstavlja samostojno t. i. satelitsko digitalno podatkovno bazo, ki vsebuje vse specifične podatke o VE in je hkrati integrirana v celostno slovarsko bazo. Leksikon večbesednih enot je na repozitoriju CLARIN.SI[7] dostopen pod licenco CC BY-SA 4.0.

6 Pregled strojno procesljivih virov za posamezne jezike, ki vsebujejo različne tipe VE, je mogoče najti na: https://sites.google.com/site/mwesurveytest/home.

7 Vir: http://hdl.handle.net/11356/1421.

5.1 Zgradba Leksikona

Prva različica Leksikona vsebuje 5.241 večbesednih enot z lastnostjo frazeološke enote (glej pripravo izhodiščne liste FE v nadaljevanju). Za vsako enoto je definiran zapis v kanonični obliki po pravilih, ki smo jih opisali v razdelku 4.1, in sicer število, vrsta in zaporedje sestavin:

```
<headword>
    <lemma>kaj ne da miru komu</lemma>
</headword>
```

Primer 2: Zapis kanonične oblike FE v zgradbi Leksikona večbesednih enot.

Vsaka leksikonska enota je definirana s skladenjsko strukturo,[8] ki jo opredeljuje identifikacijska številka:

```
<lexicalUnit type="MWE" structure_id="122">
```

Primer 3: Opredelitev skladenjske strukture v zgradbi Leksikona večbesednih enot.

V konkretnem primeru identifikacijska številka id=«122« zastopa strukturo: »z-l-gg-s2-z«, ki jo v danem zaporedju določajo zaimek, členek, glagol, samostalnik v rodilniku in zaimek. Vseh struktur, ki določajo večbesedne enote (syntactic_structure type=«other« in »collocation«), je v Leksikonu 1.480.

Formalni zapis skladenjske strukture v Leksikonu vsebuje tudi podatek o zaporedju sestavin, ki je lahko ustaljen (‘fixed’) ali spremenljiv (‘variable’) ter o skladenjskem razmerju med sestavinami FE, ki temelji na sistemu JOS (Erjavec et al. 2010a, Erjavec et al. 2010b). Vsaki sestavini FE so pripisane še oblikoslovne omejitve na ravni besedne vrste in drugih slovničnih kategorij, npr. števila, sklona in glagolske osebe:

8 Seznam vseh struktur, upoštevanih v Leksikonu večbesednih enot, je v formatih XML in XSD dodan Leksikonu na slovenskem repozitoriju CLARIN.SI.

```
<syntactic_structure type="other" label="z-l-gg-s2-z" id="122">
    <!-- example: kaj ne da miru komu-->
    <system type="JOS">
      <components order="fixed">
        <component cid="1" type="core" label="z"/>
        <component cid="2" type="core" label="l"/>
        <component cid="3" type="core" label="gg"/>
        <component cid="4" type="core" label="s2"/>
        <component cid="5" type="core" label="z"/>
      </components>
      <dependencies>
        <dependency from="3" label="ena" to="1"/>
        <dependency from="3" label="del" to="2"/>
        <dependency from="#" label="modra" to="3"/>
        <dependency from="3" label="dve" to="4"/>
        <dependency from="3" label="dve" to="5"/>
      </dependencies>
      <definition>
        <component cid="1">
          <restriction type="morphology">
            <feature POS="pronoun"/>
          </restriction>
        </component>
        <component cid="2">
          <restriction type="morphology">
            <feature POS="particle"/>
          </restriction>
        </component>
        <component cid="3">
          <restriction type="morphology">
            <feature POS="verb"/>
            <feature type="main"/>
          </restriction>
        </component>
        <component cid="4">
          <restriction type="morphology">
            <feature POS="noun"/>
            <feature case="genitive"/>
          </restriction>
        </component>
        <component cid="5">
          <restriction type="morphology">
            <feature POS="pronoun"/>
          </restriction>
        </component>
      </definition>
    </system>
  </syntactic_structure>
```

Primer 4: Opredelitev zaporedja sestavin, skladenjskih razmerij in oblikoslovnih omejitev v zgradbi Leksikona večbesednih enot.

Znotraj leksikona je vsaka sestavina skladenjske strukture zapolnjena s konkretno leksikalno realizacijo, kot je bila izluščena iz

korpusa: predvideno oblikoskladenjsko definirano pozicijo znotraj strukture torej zaseda konkretna beseda v svoji osnovni in realizacijski obliki:

```
<lexicalUnit type="MWE" structure_id="122">
        <component num="1">
          <lexeme lemma="kaj" msd="Zv-sei">kaj</lexeme>
        </component>
        <component num="2">
          <lexeme lemma="ne" msd="L">ne</lexeme>
        </component>
        <component num="3">
          <lexeme lemma="dati" msd="Ggdste">da</lexeme>
        </component>
        <component num="4">
          <lexeme lemma="mir" msd="Somer">miru</lexeme>
        </component>
        <component num="5">
          <lexeme lemma="kdo" msd="Zv-med">komu</lexeme>
        </component>
</lexicalUnit>
```

Primer 5: Zapis konkretnih leksikalnih realizacij v zgradbi Leksikona večbesednih enot.

Sledi zapis pomenskih informacij. Vsak pomen FE ima svojo identifikacijsko številko in seznam pomenov, s katerimi je povezan na podlagi svoje definicije:

```
<senseList>
    <sense key="s.24">
      <relatedSenseList>
        <relatedSense senseKey="s.26"/>
      </relatedSenseList>
      <definitionList>
        <definition>kaj vznemirja koga; vzbuja zanimanje pri kom
        </definition>
      </definitionList>
```

Primer 6: Zapis pomenskih informacij v zgradbi Leksikona večbesednih enot.

Konkretno v primeru zgoraj, je pomen <sense key=«s.24«> pri FE *kaj ne da miru komu* z definicijo ‘kaj vznemirja koga; vzbuja zanimanje pri kom’[9] povezan s pomenom <relatedSense senseKey=«s.26«/>, ki ga v Leksikonu najdemo pri FE *žilica ne da miru komu*.

9 Definicije za 94 FE v Leksikonu so izdelane na podlagi korpusne analize.

Sledi razdelek s korpusnimi zgledi, v katerih so pri posameznem pomenu FE označene tudi sestavine s pomočjo identifikacijskih številk:

```
<exampleContainerList>
<exampleContainer>
<corpusExample exampleId="GF9913201.308.2">Hedonistična
<comp num="1">plat</comp> vaše osebnosti
<comp num="5">vam</comp>
<comp num="2">ne</comp> bo
<comp num="3">dala</comp>
<comp num="4">miru</comp>, dokler ji ne boste zares prisluhnile.
</corpusExample>
</exampleContainer>
</exampleContainerList>
```

Primer 7: Zapis korpusnih zgledov v zgradbi Leksikona večbesednih enot.

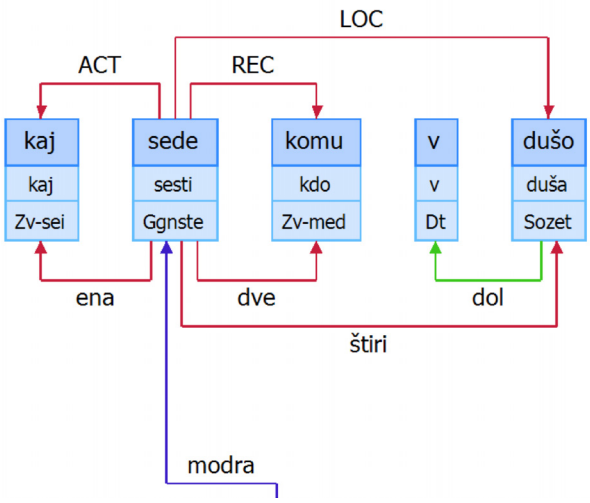
5.2 Luščenje FE iz korpusa

Pri izdelavi leksikonov večbesednih enot gre, metodološko gledano, za dva postopka, ki sta nujno medsebojno povezana (Bejček et al. 2013). Prvi postopek zadeva prepoznavanje VE v tekočem besedilu na podlagi liste VE, ki temelji na obstoječih leksikonih in slovarjih ali ročno označenih korpusih (Savary et al. 2019). Rezultat luščenja na tej podlagi je nabor izhodiščnih VE, kot so zastopane v tekočem besedilu, potencialno v vseh možnih, zanesljivo pa v vseh tipičnih skladenjskih in semantičnih realizacijah (tj. korpusnih stavkih). Drugi postopek se nanaša na odkrivanje VE v besedilih ne glede na obstoječe VE. Ta postopek je z vidika izgradnje leksikona, ki želi kontinuirano in neodvisno od obstoječih virov spremljati pojavljanje VE v besedilih, sicer bolj relevanten, a hkrati pri kompleksnih tipih VE tudi manj natančen.[10] V naši raziskavi smo uporabili prvi pristop, ki na podlagi čim več vstopnih podatkov prepoznava tako pričakovane skladenjske strukture in njihove leksikalne zapolnitve kot tudi še neregistrirane besedne kombinacije, ki so potencialno slovarsko relevantne.

10 Za metodologijo odkrivanja še neznanih večbesednih enot za slovenščino na podlagi korpusa glej Škvorc et al. 2021.

5.2.1 *Priprava podatkov*

Da bi podatke, ki smo jih predvideli v leksikonu, lahko avtomatsko izluščili iz korpusa, smo potrebovali izhodiščni nabor FE. Za izdelavo liste FE smo uporabili leksikalne vire, ki so prosto dostopni in vključujejo VE, ki ustrezajo lastnostim FE, kot smo jih opredelili v tipologiji, in sicer iz Leksikalne baze za slovenščino (Gantar et al. 2013), Slovarja slovenskih frazemov (Keber 2011) in učnega korpusa ssj500k 2.0 (Krek et al. 2020b), v katerem so označeni glagolski frazemi na podlagi smernic, določenih v okviru COST akcije PARSEME.[11] Da bi na podlagi seznama izhodiščnih FE lahko iz korpusa izluščili zahtevane podatke, smo potrebovali oblikoskladenjsko in skladenjsko označen korpus. V ta namen smo uporabili korpus Gigafida 2.0 (Krek et al. 2020a),[12] ki v različici 2.1 vključuje tudi dodatne nivoje označevanja, in sicer na skladenjski ravni po sistemu JOS (Erjavec et al. 2010a, Erjavec et al. 2010b)[13] in UD (Dobrovoljc et al. 2017),[14] lastnoimenske entitete in udeleženske vloge (Gantar et. al. 2018b). Zaporedje



Slika 2: Skladenjsko razčlenjena FE v kanonični obliki v orodju Q-Cat.

11 Vir: https://parsemefr.lis-lab.fr/parseme-st-guidelines/1.1/.
12 Vir: https://viri.cjvt.si/gigafida/System/About.
13 Vir: http://nl.ijs.si/jos/index-en.html.
14 Vir: https://universaldependencies.org/.

sestavin v kanonični obliki leksikonske enote pod seboj združuje vse nivoje informacij, ki jih vsebuje korpus (lema, MSD, skladnja), kar prikazuje skladenjsko razčlenjena FE v orodju Q-Cat (Brank 2021) na Sliki 2, ki smo ji ročno dodali še nivo udeleženskih vlog.

5.2.2 Postopek luščenja

Za postopek avtomatskega luščenja so bile vse izhodiščne FE skladenjsko razčlenjene in pretvorjene v skladenjske strukture, ki so predstavljale osnovo za luščenje primerov rabe posamezne FE iz korpusa. Z namenom, da bi zajeli tudi variantnost in potencialne nove FE, smo upoštevali možnost zapolnjevanja posameznih sestavnih elementov FE s katero koli drugo besedo, kot prikazuje Tabela 2.

Tabela 2: Seznam izluščenih frazeoloških kandidatov za FE *barvati kaj s črnimi barvami* iz korpusa Gigafida 2.0.

0	**barvati**	**kaj**	**s**	**črnimi**	**barvami**	**FE**
	x	**A**	**C**	**x**	**x**	
1	slikati	dneve	s	črnimi	barvami	DA
	a	a	C	C	C	
2	slikati	nevarnosti	s	črnimi	barvami	DA
	a	a	C	C	C	
3	barvati	obrazke	s	pisanimi	barvami	NE
	C	a	C	a	C	
4	barvati	jajčka	s	posebnimi	barvami	NE
	C	a	C	a	C	
5	barvati	dogajanja	s	črnimi	odtenki	DA
	C	a	C	C	a	
6	barvati	kozarčke	s	posebnimi	barvami	NE
	C	a	C	a	C	

Kot prikazuje Tabela 2, smo pri luščenju primerov rabe iz korpusa za vsako izhodiščno FE (0) poiskali vse ustrezne realizacije v korpusu (1–6). Določili smo sestavine FE, ki se lahko spreminjajo (x), konstantne sestavine glede na izhodiščno FE (C) ter predvidena vezljivostna mesta (A). V korpusnih realizacijah (1–6) smo na

spremenljivih mestih (x) zabeležili konkretne leksikalne realizacije (a). V zadnjem stolpcu smo označili, ali zveza nastopa v frazeološkem pomenu ali ne. Ti seznami so nam nato služili za analizo pojavnih oblik FE v realni pisni rabi, kot je izkazana v pisnem korpusu standardne slovenščine, za določanje pravil za zapisovanje kanoničnih oblik FE v Leksikonu ter za razmejevanje variant in pretvorb pomensko povezanih FE od drugih samostojnih FE.

5.3 Analiza izluščenih podatkov

V prvi fazi smo izluščene primere, ki so vsebovali izhodiščne FE, analizirali na podlagi kontekstualnih podatkov in izključili nefrazeološke rabe (glej primere 3, 4 in 6 v Tabeli 2). Nato smo na podlagi tipičnih realizacij beležili variantne in pretvorbene oblike, v katerih se FE pojavljajo, kar je predstavljalo izhodišče za oblikovanje kanoničnega zapisa leksikonskih enot.

5.3.1 Variantnost

Najbolj očitna lastnost, ki jo je pokazala analiza pojavnih oblik, je variantnost, ki je kljub definicijski ustaljenosti ena najbolj prepoznavnih lastnosti FE, zlasti v korpusnih pristopih (Moon 1998, Gantar 2007). Pri analizi pojavnih oblik FE v izluščenih korpusnih primerih smo se srečali z različnimi tipi variantnosti, ki so večinoma prepoznani tudi na slovenskem gradivu (Kržišnik 2004, Meterc 2019).

V izhodišču smo variantnost opredelili kot možnost zamenjevanja posameznih sestavine FE ob ohranitvi njenega osnovnega pomena, npr. *luč na koncu tunela/predora*. Najbolj očitne in v korpusnem pristopu najlažje prepoznavne so leksikalne variante, kot kaže zgornji primer. Variantnost pa v FE ni omejena samo na leksikalno raven, pač ločimo tudi oblikoslovne variante, ki zajemajo variantnost na ravni slovničnih kategorij, vezanih na posamezno sestavino FE, npr. število: *izplačati na roko/roke;* sklona: *prilivati olja/olje na ogenj;* določnosti*: začaran/začarani krog*. Variantnost je lahko vezana tudi na prosta mesta, ki jih odpira FE, npr.: *črni oblaki se zgrinjajo nad kom/čim* / *nad koga/kaj*. Kot variante pa je mogoče obravnavati tudi

potencialne modifikacije posameznih sestavin z dodatnimi elementi, npr. *priplavati po (kisli, prežgani, slani, neslani) juhi*, ter obstoj daljše oblike FE s t. i. fakultativnim delom, npr. *(kaj) pade (komu) v naročje* in *(kaj) pade (komu) v naročje kot zrela hruška*, tudi kadar gre za vezljivostna mesta: *znajti se v začaranem krogu – znajti se v začaranem krogu (česa).*

V primerih, kjer so variante posamezne sestavine FE zelo številne, kot npr. v primeru izhodiščne FE *začaran krog* v Tabeli 3,[15] pa je na mestu premislek, katere sestavine še obravnavati kot variante in kdaj je že mogoče govoriti o elementih besedilnega okolja. Podobno kot pri kolokacijah, katerih ključna opredelitev so statistične vrednosti, nam tudi v tem primeru pri odločitvah pomagajo številčni podatki. Ker mehanizem za luščenje upošteva skladenjska razmerja med sestavinami FE ter njihove morfološke lastnosti, je mogoče določiti frekvenčni prag tako za leksikalne izbire na variantnih mestih kot za različna skladenjska razmerja, ki so definirana z naborom skladenjskih struktur.

Tabela 3: Variantnost glagolske sestavine za izhodiščno FE *začaran krog* (5.671 pojavitev v korpusu Gigafida 2.0) v različnih skladenjskih strukturah glede na statistične vrednosti.

	gg-p4-s4				gg-zp-d-p4-s4	gg-d-p4-s4			gg-zp-d-p5-s5	gg-d-p2-s2				gg-zp-d-p2-s2
	pojavitve v okolici	MI³	LL	logDice		pojavitve v okolici	MI³	LL	logDice		pojavitve v okolici	MI³	LL	logDice
prekiniti	177	24,135	1,910	6,461	**vrteti (se)**	697	30.483	9,908	8,13	**izstopiti**	90	22,673	1,030	6,746
pretrgati	57	22,262	0,724	7,162	**znajti se**	395	26,184	4,131	6,273	**rešiti (se)**	70	18,760	0,493	3,840
presekati	54	22,376	0,706	7,271	**ujeti (se)**	116	21,649	5,240	1,047	**stopiti**	39	16,376	0,237	3,139
skleniti	49	17,191	0,308	3,302	**voditi**	99	18,845	0,574	2,957	**izvleči (se)**	24	16,917	0,210	4,812
razkleniti	15	20,098	0,225	6,291	**pasti**	50	16,931	0,293	2,994	**izviti (se)**	21	18,649	0,245	6,079

15 Za prikaz problema smo uporabili osnovni konkordančnik korpusa Gigafida 2.0 in iskanje po okolici zveze *začaran krog* v razponu +/- 3 besede. Dobljeni seznam smo filtrirali glede na besedno vrsto elementov v sobesedilu in zabeležili število pojavitev v definirani okolici in statistične vrednosti MI3, LL in logDice. Statistične vrednosti so bile izbrane glede na ugotovitve v Kosem et al. (2021).

5.3.2 Pretvorbenost

Poleg variantnosti izkazuje večina FE tudi različne pretvorbene možnosti, kamor štejemo prilagajanja celotne FE sobesedilu v smislu spremembe skladenjske vloge, npr. posamostaljenje: *priplavati po kisli juhi – kisla juha*, prehoda v stavčno oz. pregovorno obliko: *vrteti se v začaranem krogu – krog je začaran*; *začaran krog se sklene; igrati se z ognjem – kdor se igra z ognjem, se opeče*; zanikanja: *priplavati po kisli juhi – ne priplavati po kisli juhi*; možnosti trpne rabe: *obračati denar – denar se obrača*, spremembe v kategoriji živosti pri udeležencih, npr. *(kaj) ne pusti koga pri miru – ne pustiti (koga) pri miru* in spremembe v vezljivostnem vzorcu FE, npr. *(kaj) je na (čigavih) ramenih – (kaj) je na ramenih (koga)*. Med samostojne pretvorbe je mogoče šteti še prehod v t. i. besedilne ali pragmatične FE, npr. *(kdo) ni padel na* glavo in *saj nisem na glavo padel.*

Med pretvorbeno povezanimi FE je kot samostojne leksikonske enote smiselno navajati predvsem osamosvojene samostalniške zveze, ki sicer nastopajo ob variantnih glagolih in imajo kot osamosvojene zveze tudi potrditve v korpusnih primerih, npr. *rešiti se, izviti se … iz začaranega kroga – začarani krog*; *ugrizniti, zagristi v kislo jabolko – kislo jabolko; dobiti, imeti debelo kožo – debela koža*; *zaklati kokoš, ki nese zlata jajca – zlato jajce*. V ta sklop sodijo tudi posamostaljenja tipa: *prepirati se za oslovo senco – prepiranje za oslovo senco – prepir za oslovo senco*. Med pretvorbe, ki jih v Leksikonu VE navajamo kot povezane leksikonske enote, sodijo tudi primeri z izkazanimi prostimi vezljivostnimi mesti, npr. *imeti kurjo polt – (kaj) naredi kurjo polt (komu)*.

Tudi o pretvorbah, vezanih na določeno FE, je mogoče govoriti samo v povezavi s pomenom. Pretvorbeno povezane so samo tiste FE, ki ob svoji pretvorbi ohranjajo pomen. V vseh drugih primerih moramo FE obravnavati kot samostojno – pretvorbeno nepovezano leksikonsko enoto, kot prikažemo v Tabelah 4 in 5.

5.3.3 Povezanost variantnih in pretvorbenih oblik FE

Kompleksnost problematike pri določanju leksikonske enote FE, ki jo med drugim povzročata variantnost in možnost pretvorb, ponazarjamo na primeru izhodiščne FE, ki vsebuje predložno zvezo *s prstom* in glagolom *migniti*. Postopek luščenja je predvidel možnosti različnih realizacij na mestu obeh sestavin, na podlagi izluščenih primerov pa je bilo mogoče evidentirati še druge sestavine, ki se pojavljajo v besedilnem okolju stavčnega vzorca (Tabela 4).

Tabela 4: Seznam pomensko povezanih leksikonskih enot za izhodiščno FE *s prstom migniti*.

	Leksikonska enota																					Pomen
1											samo	s	prstom					migniti				*'biti vpliven; imeti moč'*
2											samo	s	prstom					migniti	pa			
3											samo	z	mezincem					migniti				
4											le	s	prstom					migniti				
5		ne	da	bi								s	prstom					mignil				*'ne da bi se bilo treba truditi'*
6		ne	da	bi	bilo	treba	komu					s	prstom					migniti				
7		ne	da	bi			kdo					s	prstom					mignil				
8		ne	da	bi		moral	kdo					s	prstom					migniti				
9	kdo	ne		bi							niti	s	prstom					mignil				
10											niti	s	prstom					migniti				
11											niti	z	mezincem					migniti				
12												s	prstom				ne	migniti				
13												z	mezincem				ne	migniti				
14	nihče										niti	s	prstom				ne	migne				*'nič ne narediti; ne ukrepati'*
15											niti	s	prstom				ne	migniti				
16											niti	s	prstom				ne	migniti	da			
17											niti	z	mezincem				ne	migniti				
18											še	s	prstom				ne	migniti				
19												s	prstom	niti			ne	migniti				
20											niti	s	prstom				ne	migniti	za	koga/kaj		
21												s	prstom				ne	migniti	za	koga/kaj		
22												s	prstom				ne	migniti	pri	čem		

Pri združevanju vzorcev v leksikonske enote smo variantne sestavine pri posameznih sestavinah FE šteli kot samostojne enote. Primere realizacij v pretekliku in prihodnjiku smo združili v leksikonsko enoto z nevtralno sedanjiško obliko glagola: *ne bo niti s prstom mignil zate → niti s prstom ne migniti za (koga)*. Prav tako smo v eno leksikonsko enoto združili primere z izraženim osebkom in primere, ki so osebek predvidevali, čeprav v stavku ni bil eksplicitno izražen, npr. *(kdo) niti s prstom ne migne, da → niti s prstom ne migniti, da.* Členke smo v kanonični obliki razporedili glede na to, katero sestavino modificirajo: *niti s prstom; niti migniti*.

Kompleksna slika variant in pretvorb ne razkriva le tipičnosti vzorca pri določeni FE, ampak kaže posledice tudi za njen pomen. Kot lahko vidimo iz Tabele 4, se pomensko osamosvojijo vzorci z glagolom v trdilni obliki (*migniti*) in členkoma *samo/le* (primeri 1–4). V nasprotju s členkom *niti*, ki se pojavlja pri FE s pomenom 'nič ne narediti; ne ukrepati', členka *samo* in *le* sugerirata FE s samostojnim pomenom: 'biti vpliven; imeti moč', ki je v korpusu sicer redkeje zastopan. Drugo pomensko samostojno skupino sestavljajo leksikonske enote s trdilnim glagolom (*migniti*) in zanikano pogojniško zvezo *ne da bi* (primeri 5–8), navadno še v kombinaciji z modalnim *morati* ali *treba*. Ta kombinacija je ključna za izražanje pomena 'ne da bi se bilo treba truditi'. Prav tako je za ta pomen potrebna izraženost delovalnika, ki pa ga v leksikonski enoti ni mogoče navajati na prvem mestu, kot sicer določa naše pravilo o enotnem zaporedju sestavin v kanonični obliki FE, še posebej, ker se delovalnik lahko pojavlja tudi v neimenovalniškem sklonu (podčrtano): *ne da bi moral (kdo) s prstom migniti – ne da bi bilo treba <u>(komu)</u> s prstom migniti*. Najpogosteje je kombinacija besed *s prstom* + *migniti* vezana na pomen 'nič ne narediti; ne ukrepati' (primeri 9–22). Kot kažejo primeri, se pomen realizira tako s trdilnimi kot nikalnimi oblikami glagola (*migniti, ne migniti*). V primeru trdilne glagolske oblike je prisotnost nikalnega členka *niti* obvezna, pri nikalnih oblikah pa imamo lahko tako enojno kot dvojno zanikanje: *<u>niti</u> s prstom migniti* in *<u>niti</u> s prstom <u>ne migniti</u>*, pri čemer je pri dvojnem zanikanju členek *niti* vezan na sestavino *prst* (*niti s prstom*), pri drugem tipu zanikanja pa na

glagol (*ne migniti*). Zanimiva, vendar v realni rabi zelo redko izkazana možnost, je dvojno zanikanje, vezano neposredno na glagol (primer 19 v Tabeli 4): *s prstom niti ne migniti*, ki je poleg tega, da sugerira drugačno pomensko interpretacijo, v korpusu tudi redko izkazana, zato je nismo navajali kot samostojne leksikonske enote. Pri skupini FE za pomen 'nič ne narediti; ne ukrepati' je treba izpostaviti tudi fakultativno odpiranje vezljivostnega mesta (*za koga/kaj*, *pri čem*) ter prisotnost odvisnega stavka, ki ga nakazuje veznik da*: niti s prstom ne migniti, da …*

Z vidika povezovanja variantnih in pretvorbeno povezanih FE v Leksikonu je pomembno prepoznavati pomensko vrednost FE, saj nam to omogoča povezljivost FE ne samo na ravni leksikonskih enot, pač pa tudi med posameznimi pomeni. Sistem povezovanja variantno in pretvorbeno povezanih FE na ravni leksikonskih enot in posameznih pomenov prikazuje Tabela 5.

Tabela 5: Sistem variantno in pretvorbeno povezanih FE na ravni pomena.

kaj **ne da miru** komu	*kaj vzbuja zanimanje pri kom*							
kaj **ne pusti** koga **pri miru**	*kaj vzbuja zanimanje pri kom*							
dajte no mir		*izraža nestrinjanje*						
dati mir			*biti miren; ne razgrajati*	*ne nadlegovati*				
dati mir komu				*ne nadlegovati*				
pustiti koga **pri miru**				*ne nadlegovati*				
ne dati miru					*biti aktiven*	*razgrajati*	*nadlegovati*	
ne dati miru komu							*nadlegovati*	
ne pustiti koga **pri miru**							*nadlegovati*	
pustiti kaj **pri miru**								*ne ukvarjati se s čim*

Sobesedilna analiza korpusnih primerov, ki vsebujejo FE, navedene v Tabeli 5, je pokazala, da sta FE *(kaj) ne dati miru (komu)* in *(kaj) ne pusti (koga) pri miru* povezani v pomenu 'kaj vznemirja koga; kaj vzbuja zanimanje pri kom'. FE *dajte no mir*, ki se rabi tudi v obliki: *daj no mir*, izraža nestrinjanje, dvom in zato variantno in pretvorbeno ni povezana s katero od navedenih leksikonskih enot. FE *dati mir* je glede na korpusne primere mogoče prepoznati v dveh pomenih: 1. 'biti miren; ne razgrajati' in 2. kot opozorilo, prošnja 'prenehati nadlegovati, vznemirjati', s FE *dati mir (komu)* in *pustiti (koga) pri miru* pa jo je mogoče povezati le v 2. pomenu. Za FE *ne dati miru* smo registrirali tri pomene: 1. 'vztrajati, biti aktiven', 2. 'razgrajati' in 3. 'nadlegovati, vznemirjati', vendar se s FE *ne dati miru (komu)* in *ne pustiti koga pri miru* povezuje le v tretjem pomenu. FE *pustiti (kaj) pri miru* in *dajte no mir* se kljub prekrivnim sestavinam zaradi pomena 'ne se ukvarjati s čim' in 'izraža nestrinjanje' ne povezujeta z nobeno od navedenih leksikonskih enot.

6 Zaključek in nadaljnje delo

Naš namen je bil izdelati jezikovni vir, ki bo uporaben pri izdelavi Slovarja sodobnega slovenskega jezika in za številne jezikovnotehnološke naloge. Leksikon VE predstavlja tako sestavni del celostne Digitalne slovarske baze, ki omogoča strukturiranje različnih tipov jezikovnih podatkov, od morfologije ter eno- in večbesednih leksikalnih enot do stavčnih vzorcev in pomenskih informacij.

VE enote so z vidika vključevanja in zapisa v digitalnih virih lahko problematične z več vidikov. Njihova pojavnost v besedilu je razpršena, saj imajo kot večbesedne enote veliko različnih možnosti prilagajanja besedilu, posamezne besede se lahko na podlagi oblikovnih in pomenskih možnosti znotraj zveze zamenjujejo in prevzemajo različne oblike. Med sestavine VE se lahko vrivajo druge sestavine in nekatere VE lahko nastopajo tudi kot proste zveze, torej brez leksikalnega pomena. Njihov zapis v digitalni bazi pa kljub temu zahteva ustrezno formalizacijo, ki omogoča tudi povratno luščenje iz korpusa. VE je zato, podobno kot besedne enote, treba pri vključevanju v

leksikon obravnavati kot leme ter ločevati različne oblike od njihovih pojavnic. Ob vključitvi pojava variantnosti in pretvorbenih možnosti je osnovna naloga pri določanju zapisa VE kot leksikonske enote določitev možnega obsega variacije oz. ugotovitev, na kateri točki odstopanje od kanonične oblike krši medsebojno odvisnost med obliko in pomenom VE, kar je pogoj za prepoznavnost nove VE.

Postopek luščenja FE na podlagi predhodno definiranih FE nam omogoča prepoznavanje relativno ustaljenih variant in pretvorbenih možnosti, pri čemer je frekvenčni prag variantnosti in pretvorbenosti mogoče prilagoditi glede na frekventnost celotne FE in glede na druge parametre, ki jih omogoča korpus.

Analiza realne rabe FE na podlagi izluščenih podatkov iz korpusa nas napeljuje na nekatere sklepe, ki jih je smiselno upoštevati pri oblikovanju kanoničnih oblik leksikonskih enot v digitalno zasnovanih jezikovnih virih. VE je v slovarjih smiselno obravnavati na enak način kot enobesedne iztočnice. To je pomembno predvsem za identifikacijo zveze kot celote – ne le prek posameznih njenih sestavin – in zaradi možnosti vzpostavitve pomenskih in drugih povezav med posameznimi FE. Hkrati ima to posledice tudi za možnost iskanja FE po celotni zvezi in ne le po kateri od njenih leksikalnih sestavin, kot je sicer praksa v tradicionalnih tiskanih slovarjih. Digitalni slovarski medij namreč ne samo da ni problematičen z vidika vključevanja velike količine podatkov zaradi prostorske neomejenosti, ampak – glede na to da temelji na digitalno organiziranih podatkih v bazi – omogoča tudi različne prikaze VE: bodisi kot samostojnih leksikalnih enot bodisi v povezavi s posamezno sestavino kot iztočnico. Analiza korpusnih primerov daje na prvi pogled nepregledno število možnih realizacij v smislu zaporedja sestavin, variant, skladenjskih pretvorb in načinov vklapljanja FE v sobesedilo. Vendar pa je bojazen, da bi v takih primerih število leksikonskih enot v slovarju preveč naraslo, odveč, saj digitalna baza nima prostorskih omejitev, hkrati pa se je pri naboru povezanih leksikonskih enot mogoče zanašati na frekvenčne podatke o zastopanosti posamezne variante, pretvorbe ipd. ter zanemariti redke in enkratne pojavitve. Poleg pomembnega spoznanja, da je kanoničnih oblik FE za razliko od enobesednih

lahko več in ne ena sama, je pomembno tudi povezovanje med leksikonskimi enotami znotraj istega pomenskega polja. To dejstvo, ki narekuje organizacijo podatkov v digitalni bazi je pomembno tudi z uporabniškega vidika. Uporabniki lahko prek povezanih variant in pretvorb ugotovijo, katere rabe so za posamezno FE možne oz. sprejemljive in katere ne. Pomensko povezane FE uporabniku pokažejo način umeščanja v sobesedilo, kar je zlasti pomembno za učenje slovenščine kot tujega jezika in besedilno produkcijo na sploh.

Princip organizacije podatkov v digitalni slovarski bazi, kjer predstavljajo variantno in pretvorbeno povezane VE znotraj posameznega pomena samostojne leksikonske enote, izpostavlja tudi vprašanje oblikovanja njene kanonične oblike. V Leksikonu VE smo v ta namen določili pravila, ki določajo enotno zaporedje, obliko in vrsto sestavin, ki temeljijo na abstraktnem zaporedju znotraj prostega stavka. Ob tem je treba poudariti, da abstraktno enotno zaporedje sestavin ni idealna rešitev pri vseh tipih FE. Posebnost v tem smislu so npr. besedilne FE tipa *dajte no mir, pojdi se solit*, kjer pravila za določanje kanonične oblike, npr. *dati mir, iti se solit,* kot smo jih prikazali v prispevku, uporabniku ne dajejo realne slike o rabi FE v kontekstu. V ta namen nameravamo v prihodnje več pozornosti nameniti uporabniškim raziskavam, kjer bomo s pomočjo različnih kanoničnih oblik preverjali prepoznavnost FE in ali lahko uporabniki na podlagi kanonične oblike ustrezno uporabijo FE v sobesedilu.

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

Ahlén, K. (2013). *Building a MWE Lexicon for Swedish (SweMWELex)*. Neobjavljen rokopis. Dostopno prek: https://cl.lingfil.uu.se/~nivre/master/karin1.pdf, Univerza v Uppsali.

Atkins, B. T. S. in Rundell, M. (2008). *The Oxford Guide to Practical Lexicography*. New York: Oxford University Press.

Bejček, E. in Straňák, P. (2010). Annotation of Multiword Expressions in the Prague Dependency Treebank. *Language Resources and Evaluation*, 44 (1), 7–21. https://doi.org/10.1007/s10579-009-9093-0.

Bejček, E., Straňák, P. in Pecina, P. (2013). Syntactic Identification of Occurrences of Multiword Expressions in Text using a Lexicon with Dependency Structures. V V. Kordoni, C. Ramisch in A. Villavicencio (ur.), *Proceedings of the 9th Workshop on Multiword Expressions* (str. 106–115). Association for Computational Linguistics. Dostopno prek: https://aclanthology.org/W13-1016.pdf.

Brank, J. (2021). Q-CAT Corpus Annotation Tool 1.2, Slovenian language resource repository CLARIN.SI. http://hdl.handle.net/11356/1442.

Dobrovoljc, K., Erjavec, T. in Krek, S. (2017). The Universal Dependencies Treebank for Slovenian. V T. Erjavec, J. Piskorski, L. Pivovarova, J. Šnajder, J. Steinberger in R. Yangarber (ur.), *Proceedings of the 6th Workshop on Balto-Slavic Natural Language Processing* (str. 33–38). The Association for Computational Linguistics. Dostopno prek: https://aclanthology.org/W17-1406.pdf.

Erjavec, T., Krek, S., Arhar, Š., Fišer, D., Ledinek, N., Saksida, A., Sivec, B. in Trebar, B. (2010a). Oblikoskladenjske specifikacije JOS V1.1. Dostopno prek: http://nl.ijs.si/jos/msd/html-sl/index.html.

Erjavec, T., Fišer, D., Krek, S. in Ledinek, N. (2010b). The JOS Linguistically Tagged Corpus of Slovene. V N. Calzolari, K. Choukri, B. Maegaard, J. Mariani, J. Odijk, S. Piperidis, M. Rosner, 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.

Filipec, J. in Čermák, F. (1985). *Česka lexikologie*. Praga: Academia.

Fotopoulou, A., Markantonatou, S. in Giouli, V. (2014). Encoding MWEs in a Conceptual Lexicon. V V. Kordoni, M. Egg, A. Savary, E. Wehrli in S. Evert (ur.), *Proceedings of the 10th Workshop on Multiword Expressions*

(MWE) (str. 43–47). Association for Computational Linguistics. Dostopno prek: https://aclanthology.org/W14-0807.pdf.

Gantar, P. (2007). *Stalne besedne zveze v slovenščini: korpusni pristop*. Ljubljana: Založba ZRC. https://doi.org/10.3986/9789612540364.

Gantar, P. (2015). *Leksikografski opis slovenščine v digitalnem okolju*. Ljubljana: Trojina, zavod za uporabno slovenistiko. E-izdaja (2018). Ljubljana: Znanstvena založba Filozofske fakultete. Dostopno prek: https://e-knjige.ff.uni-lj.si/znanstvena-zalozba/catalog/view/62/138/2602-1.

Gantar, P., Krek, S., Kosem, I., Šorli, M., Kocjančič, P., Grabnar, K., Yerošina, O., Zaranšek, P. in Drstvenšek, N. (2013). Slovene lexical database 1.0, Slovenian language resource repository CLARIN.SI. http://hdl.handle.net/11356/1030.

Gantar, P., Arhar Holdt, Š. in Pollak, S. (2018a). Leksikalne novosti v besedilih računalniško posredovane komunikacije. *Slavistična revija*, 66 (4), 459–472. Dostopno prek: https://srl.si/ojs/srl/article/view/2018-4-1-4.

Gantar, P., Štrkalj Despot, K., Krek, S. in Ljubešić, N. (2018b). 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.

Gantar, P., Colman, L., Parra Escartín, C. in Martínez Alonso, H. (2019a). Multiword expressions: between lexicography and NLP. *International Journal of Lexicography*, 32 (2), 138–162. https://doi.org/10.1093/ijl/ecy012.

Gantar, P., Arhar Holdt, Š., Čibej, J. in Kuzman, T. (2019b). Structural and Semantic Classification of Verbal Multi-Word Expressions in Slovene. *Prispevki za novejšo zgodovino,* 59 (1), 99–119. Dostopno prek: http://www.dlib.si/stream/URN:NBN:SI:DOC-BKIGNYPE/4d72dc31-9f1a-4ccf-86de-da3690ac7f54/PDF.

Gantar, P., Krek, S. in Kosem, I. (2021). Opredelitev kolokacij v digitalnih slovarskih virih za slovenščino. V I. Kosem (ur.), Kolokacije v slovenščini (str. 15–41). Ljubljana: Znanstvena založba Filozofske fakultete.

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.

Grégoire, N. (2010). DuELME: a Dutch electronic lexicon of multiword expressions. *Language Resources and Evaluation*, 44 (1/2), 23–39. https://doi.org/10.1007/s10579-009-9094-z.

Jackendoff, R. (1997). Twistin' the Night Away. *Language*, 73, 534–559.

Keber, J. (2011). Dictionary of Slovenian Phrasemes, Slovenian language resource repository CLARIN.SI. http://hdl.handle.net/11356/1129.

Kosem, I., Krek, S. in Gantar, P. (2020). Defining collocation for Slovenian lexical resources. *Slovenščina 2.0: empirične, aplikativne in interdisciplinarne raziskave*, 8 (2), 1–27. https://doi.org/10.4312/slo2.0.2020.2.1-27.

Kosem, I., Logar, N., Dobrovoljc, K. in Ljubešić, N. (2021). Razvrščanje in relevantnost kolokatorjev v slovenščini: novi pristopi. V I. Kosem (ur.), *Kolokacije v slovenščini* (str. 79–124). Ljubljana: Znanstvena založba Filozofske fakultete.

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, A., Laskowski, C., Krsnik, L., Kosem, I., Brank, J., Dobrovoljc, K., Arhar Holdt, Š., Čibej, J., Robnik-Šikonja, M., Klemenc, B. in Gorjanc, V. (2021a). Multiword Expressions lexicon extracted from the Gigafida 2.1 corpus, Slovenian language resource repository CLARIN.SI. http://hdl.handle.net/11356/1421.

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.

Kržišnik, E. (1996). Norma v frazeologiji in odstopi od nje v besedilih. *Slavistična revija*, 44 (2), 133–154. Dostopno prek: https://srl.si/ojs/srl/article/view/COBISS_ID-2620770.

Kržišnik, E. (2004). Poskusni zvezek slovenskega frazeološkega slovarja. *Slavistična revija*, 52 (2), 199–208. Dostopno prek: https://srl.si/ojs/srl/article/view/COBISS_ID-26054754.

Ljubešić, N., Dobrovoljc, K., Krek, S., Peršurić Antonić, M. in Fišer, D. (2014). hrMWELex: a MWE lexicon of Croatian extracted from a parsed gigacorpus. V T. Erjavec in J. Žganec Gros (ur.), *9. konferenca jezikovne tehnologije Informacijska družba IS 2014* (str. 25–32). Dostopno prek: http://library.ijs.si/Stacks/Proceedings/InformationSociety/2014/2014_IS_CP_Volume-G_%28LT%29.pdf.

Markantonatou, S., Zakis, G., Moutzouri, V. in Chantou, M. (2019). IDION: A database for Modern Greek multiword expressions. V A. Savary, C. Parra Escartín, F. Bond, J. Mitrović in V. Barbu Mititelu (ur.), *Proceedings of the Joint Workshop on Multiword Expressions and WordNet (MWE-WN 2019)* (str. 130–134). Association for Computational Linguistics. Dostopno prek: https://aclanthology.org/W19-5115.pdf.

Meterc, M. (2019). Analiza frazeološke variantnosti za slovarski prikaz v eSSKJ-ju in SPP-ju. *Jezikoslovni zapiski: zbornik Inštituta za slovenski jezik Frana Ramovša*, 25 (2), 33–45. https://doi.org/10.3986/JZ.25.2.2.

Moon R. (1998). *Fixed Expressions and Idioms in English: A Corpus-Based Approach*. Oxford: Clarendon Press.

Odijk, J. (2013). Identification and Lexical Representation of Multiword Expressions. V P. Spyns, J. Odijk (ur.), *Essential Speech and Language Technology for Dutch* (str. 201–217). Berlin; Heidelberg: Springer. https://doi.org/10.1007/978-3-642-30910-6_12.

Perdih, A. in Ledinek, N. (2019). Multi-word Lexical Units in General Monolingual Explanatory Dictionaries of Slavic languages. *Slovene Linguistic Studies/Slovenski Jezik*, 12 (22), 113–134. https://doi.org/10.3986/sjsls.12.1.07.

Savary, A., Cordeiro, S. R. in Ramisch, C. (2019). Without lexicons, multiword expression identification will never fly: A position statement. V A. Savary, C. Parra Escartín, F. Bond, J. Mitrović in V. Barbu Mititelu (ur.), *Proceedings of the Joint Workshop on Multiword Expressions and WordNet (MWE-WN 2019)* (str. 79–91). Association for Computational Linguistics. Dostopno prek: https://aclanthology.org/W19-5110.pdf.

Shudo, K., Kurahone, A. in Tanabe, T. (2011). A Comprehensive Dictionary of Multiword Expressions. V D. Lin, Y. Matsumoto in R. Mihalcea (ur.), *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies* (str. 161–170). Association for Computational Linguistics. Dostopno prek: https://aclanthology.org/P11-1017.pdf.

Slovar slovenskega knjižnega jezika, druga, dopolnjena in deloma prenovljena izdaja. Dostopno prek: www.fran.si.

Smørdal Losnegaard, G. (2019). Predicting The Unpredictable: Developing a lexicon model for Norwegian MWEs. *CLARIN2019 Book of Astracts*. https://www.clarin.eu/sites/default/files/clarin2019_phdposter_10_losnegaard.pdf.

Škvorc, T., Gantar, P. in Robnik-Šikonja, M. (2021). Strojno prepoznavanje idiomov z globokimi nevronskimi mrežami. V Š. Arhar Holdt (ur.), *Nova slovnica sodobne standardne slovenščine: viri in metode* (str. 231–258). Ljubljana: Znanstvena založba Filozofske fakultete.

Tanabe, T., Takahashi, M. in Shudo, K. (2014). A lexicon of multiword expressions for linguistically precise, wide-coverage natural language processing. *Computer Speech and Language*, 28 (6), 1317–1339. https://doi.org/10.1016/j.csl.2013.09.001.

Tavast, A., Langemets, M., Kallas, J. in Koppel, K. (2018). Unified Data Modelling for Presenting Lexical Data: The Case of EKILEX. V S. Krek, J. Čibej, V. Gorjanc in I. Kosem, *Proceedings of the 18th EURALEX International Congress: Lexicography in Global Context* (str. 749–761). Ljubljana: Znanstvena založba Filozofske fakultete. Dostopno prek: https://e-knjige.ff.uni-lj.si/znanstvena-zalozba/catalog/download/118/211/2920-1?inline=1.

Toporišič, J. (1973/1974). K izrazju in tipologiji slovenske frazeologije. *Jezik in slovstvo*, 19 (8), 273–279.

Vidic, Z. (2021). *Oblikovanje kanoničnih oblik pri frazeoloških enotah v strojno berljivem Leksikonu večbesednih enot – uporabniški vidik*. Magistrsko delo. Univerza v Ljubljani, Filozofska fakulteta.