Analiza udeleženskih vlog s skladenjskega, pomenskega in leksikalnega vidika

*Polona Gantar**

Izvleček

V prispevku je prikazana možnost uporabe učnega korpusa za slovenščino ssj500k pri jezikoslovnih analizah ter metodologija pri pripravi in analizi jezikovnih podatkov. V raziskavi, ki se osredotoča na združevanje skladenjskega in pomenskega nivoja jezikovnega opisa, smo analizirali razmerje med skladenjskimi osebki in udeleženskо vlogo vršilca dejanja. Rezultate analiz bo mogoče uporabiti pri izboljšavi učnih množic za strojno učenje ter pri slovničnih in pomenskih opisih v slovnicah in slovarjih za slovenščino.

Ključne besede: učni korpus ssj500k, odvisnostna razmerja, semantične vloge, semantični tipi

Abstract – An analysis of semantic roles from a syntactic, semantic and lexical point of view

The paper presents the possibilities of using the learning corpus ssj500k for Slovene in linguistic analyses and the methodology of preparing and analysing corpus linguistic data. In a study focusing on combining the syntactic and semantic levels of linguistic description, we analyse the relationship between syntactic subject and the semantic role ACT (actor). The results of the analyses will be used to improve training sets for machine learning and for grammatical and semantic descriptions in grammars and dictionaries for Slovene.

Keywords: training corpus ssj500k, syntactic dependencies, semantic roles, semantic types

* Univerza v Ljubljani, Filozofska fakulteta, apolonija.gantar@ff.uni-lj.si.

1 Uvod

Strojno procesljivi podatki o jeziku na oblikoslovni, skladenjski in vse bolj tudi na pomenski ravni postajajo ključni tako za raziskovanje jezika kot za uporabo v jezikovnotehnološke namene, izdelavo jezikovnih aplikacij in jezikovnih priročnikov, kot so leksikoni, slovarji in slovnice. Z razvojem korpusnega jezikoslovja postajajo strojno procesljive zbirke vse večje, kategorije označenih jezikovnih podatkov številnejše, možnosti njihove uporabe pa vse kompleksnejše, zlasti ko gre za medsebojno povezovanje na različnih jezikovnih ravninah. Pri jezikoslovnih analizah, pri čemer slovenščina ni izjema, se dostikrat kaže problem izolirane ali ravninske obravnave jezikovnih pojavov, čeprav se ugotavlja njihova soodvisnost tako rekoč na vseh ravninah kot tudi v odnosu do jezikovne rabe.

Za slovenščino je prepoznava potrebe po strojno procesljivih učnih množicah, ki omogočajo jezikoslovne analize in nadaljnji razvoj jezikovnih virov, prisotna že od časa izdelave prvega splošnega korpusa, ko je bilo treba premisliti načine jezikoslovnega označevanja, od segmentacije in tokenizacije do oblikoskladenjskih, skladenjskih in nenazadnje tudi pomenskih oznak (Jakopin in Bizjak 1997; Krek 2010). Osnovni premisleki so zahtevali na eni strani upoštevanje specifik slovenskega jezika, zlasti morfološko bogatost, besednovrstno konverznost in prostost besednega reda, po drugi pa tudi razmislek o ločevanju besed od besednih zvez, npr. pri zapisu z vezajem, že na ravni tokenizacije. Poleg omenjenega je bil pri izbiri jezikoslovnih oznak potreben tudi premislek o mednarodni prenosljivosti z uporabo standardiziranega nabora oznak za več jezikov, npr. Universal dependencies[1] na ravni morfologije in skladnje, ki se uporablja za več evropskih jezikov – poleg sistema JOS (Erjavec et al. 2011) – tudi za slovenščino (Dobrovoljc et al. 2017).

Za nadaljnji razvoj korpusov je izrednega pomena razvoj učnih množic ali t. i. učnih korpusov,[2] ki vsebujejo ročno pripisane oznake na različnih ravneh. V večnivojsko označenem korpusu je mogoče jezikovne pojave opazovati linearno, tj. znotraj leksikalnega in skladenjskega konteksta, kot tudi vertikalno oz. medravninsko z upoštevanjem morfološke podstave in pomenskih lastnosti besed. Korpusno izhodišče hkrati predstavlja analizo realnih, najpogosteje sodobnih jezikovnih podatkov, pri čemer se je treba zavedati specifik korpusa in pri njegovi izdelavi sprejetih jezikoslovnih odločitev ter v zvezi z njimi interpretirati jezikovne podatke.

1 https://universaldependencies.org/.

2 Za pregled nastajanja učnih korpusov za slovenščino gl. Krek (2010).

V pričujočem prispevku najprej opišemo korpus in orodje, ki smo ju uporabili pri analizi povezav med skladenjsko vlogo osebka in njegovimi semantičnimi vlogami. V ta namen najprej povzamemo jezikoslovne odločitve, ki so bile sprejete pri skladenjskem in pomenskem označevanju učnega korpusa ter določajo teoretično izhodišče raziskave. Jedrni del prispevka je namenjen analizi udeleženske vloge »aktanta« glede na skladenjsko vlogo in glede na semantične tipe besed, ki se pojavljajo v tej skladenjski in pomenski vlogi. Prispevek zaključimo z ugotovitvami, za katere menimo, da jih je smiselno upoštevati pri nadaljnjem razvoju učnega korpusa ter pri razumevanju skladenjske in pomenske soodvisnosti jezikovnih pojavov in njihovem jezikovnem opisu.

2 Učni korpus ssj500k

V raziskavi smo uporabili učni korpus ssj500k 2.2[3] (Krek et al. 2020), ki je označen na 7 ravneh, njegova zasnova pa temelji na sistemu JOS (Erjavec et al. 2011), razvitem v istoimenskem projektu,[4] vključno z nadgradnjami, ki so potekale v okviru projekta Sporazumevanje v slovenskem jeziku.[5] Korpus ssj500k je uravnotežen nabor vzorcev besedil iz korpusa sodobne pisne slovenščine FidaPLUS (Arhar in Gorjanc 2007), ki zajemajo obdobje od leta 1990 do leta 2000. Celotni korpus vsebuje 586.248 besed, ki pripadajo 27.829 stavkom. Kot prikazuje tabela 1, je celoten korpus ročno segmentiran in lematiziran z oblikoskladenjskimi oznakami, razvitimi v projektu JOS, ter po sistemu UD za morfološko označevanje (Dobrovoljc et al., 2019).

Tabela 1: Obseg posameznih označenih nivojev v korpusu ssj500k

Nivo označevanja	Pojavnice	Stavki	Dokumenti	%
segmentacija	586.248	27.829	1.655	100
lematizacija	586.248	27.829	1.655	100
oblikoskladnja JOS	586.248	27.829	1.655	100
oblikoskladnja UD	586.248	27.829	1.655	100
skladnja JOS	235.864	11.411	617	40

3 http://hdl.handle.net/11356/1210.

4 JOS: http://nl.ijs.si/jos/

5 SSJ: http://www.slovenscina.eu/; http://www.slovenscina.eu/Vsebine/Sl/Kazalniki/K2.aspx.

Nivo označevanja	Pojavnice	Stavki	Dokumenti	%
skladnja UD	140.670	8.000	581	24
semantične vloge	112.048	5.501	228	19
imenske entitete	194.637	9.488	498	33
večbesedne enote	280.522	13.511	754	48

Drugi označeni nivoji obsegajo le del celotnega korpusa, in sicer raven večbesednih enot pribl. polovico korpusa, pomenska raven (udel. vloge) pribl. petino korpusa, raven imenskih entitet pa tretjino korpusa. Korpus ssj500k tako predstavlja največji, najbolj bogato označen ter najširše uporabljan učni in testni korpus za slovenščino.

3 Orodje za analizo korpusa Q-CAT

Poleg učnih korpusov so zlasti za jezikoslovne analize pomembna tudi orodja, ki omogočajo analize na različnih jezikoslovnih ravneh. Ko govorimo o slovenščini, je splošna poizvedovanja po korpusih mogoče opraviti že na ravni konkordančnika (prim. Arhar Holdt et al. 2019), za kompleksnejše korpusnojezikoslovne analize pa so na voljo specializirana orodja in vključitev korpusov v prostodostopne konkordančnike, ki omogočajo tudi zahtevnejše analize, npr. Sketch Engine[6] in Kontext.[7]

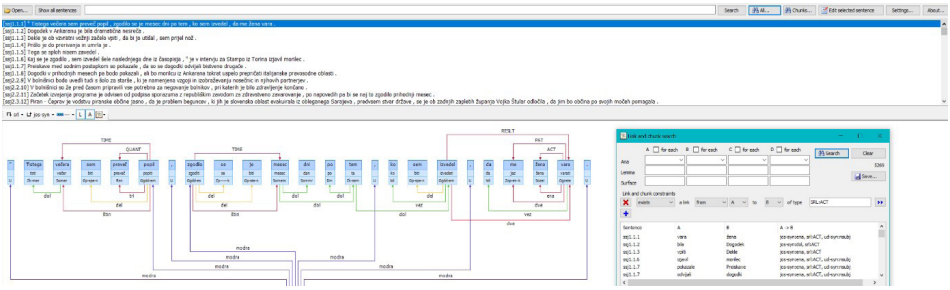
Za našo analizo smo uporabili orodje Q-CAT, ki je bilo razvito v okviru projekta Nova slovenska slovnica (Brank 2019; Dobrovoljc 2019) kot nadgradnja programa SentenceMarkup, ki je bil za ročno skladenjsko razčlenjevanje in označevanje imenskih entitet razvit že pri projektu Sporazumevanje v slovenskem jeziku. V nadgrajeni različici orodje omogoča jezikoslovno označevanje in analizo korpusnih besedil na oblikoslovni in višjih označevalnih ravneh, dodajanje poljubnega števila označevalnih ravni različnih tipov, dinamično prilagajanje nastavitev in kompleksnejša iskanja.

Uporabnik lahko s posebnim vmesnikom (slika 1) oblikuje različne iskalne pogoje, na podlagi katerih program poišče, prešteje in izpiše vse povedi v korpusu, ki temu iskalnemu pogoju ustrezajo. Orodje omogoča tudi dodajanje novih ravni označevanja, kjer določimo ime nivoja, seznam uporabljenih oznak

6 https://www.clarin.si/noske/.

7 https://www.clarin.si/kontext/corpora/corplist.

ter vrsto označevalne ravni, tj. ali gre za označevanje povezav ali nizov – in v tem primeru še možnost označevanja prekinjenih oz. neprekinjenih nizov. Za našo raziskavo smo uporabili možnost kombiniranega iskanja na skladenjskem in pomenskem nivoju.



Slika 1: Orodje za označevanje in analizo korpusa Q-CAT.

4 Raziskava

Osnovni namen naše raziskave je pokazati možnosti jezikoslovnih analiz na podlagi večnivojsko označenega korpusa, cilj pa ugotoviti, kje so težavnejša mesta na ravni skladenjsko-pomenskih razmerij z vidika sodobnega slovničnega opisa ter kje so možnosti za izboljšavo jezikoslovnih smernic pri avtomatskem označevanju in izgradnji jezikovnih virov za slovenščino v prihodnje.

4.1 Jezikoslovna izhodišča

Jezikoslovna izhodišča naše raziskave temeljijo na jezikoslovnih odločitvah, ki so bile sprejete pri izdelavi sistemov za jezikoslovno označevanje pri izgradnji učnega korpusa ssj500k. Za namene naše raziskave – v kateri nas zanima pomensko-skladenjsko razmerje med udeleženci z lastnostjo aktanta in s tipično vlogo osebka – sta zato zanimiva zlasti skladenjski in pomenski nivo označevanja. Prvi temelji na sistemu JOS, drugi pa na naboru udeleženskih vlog za slovenščino po sistemu Praške odvisnostne drevesnice (Hajič et al. 2017; Hajič et al. 2003). S pripisom semantičnih tipov udeležencem z lastnostjo aktanta ali s tipično skladenjsko vlogo osebka smo skladenjsko-pomenskemu izhodišču dodali še leksikalni vidik po sistemu, ki je bil razvit v okviru projekta Kolos (Kosem in Pori 2021).

4.1.1 Skladenjski nivo

Označevanje učnega korpusa na skladenjskem nivoju sledi smernicam, ki so opisane v sistemu Jezikoslovno označevanje slovenščine JOS (Ledinek in Erjavec 2009) in so bile nadgrajene pri projektu Sporazumevanje v slovenskem jeziku (Dobrovoljc et al. 2012). V uporabljenem odvisnostnem sistemu so skladenjski odnosi znotraj vsake povedi predstavljeni z drevesno strukturo, pri čemer vsaka povezava pripada enemu od desetih tipov na treh nivojih. Povezava tretjega nivoja (tj. modra povezava na korenski element) se uporablja za povezovanje hierarhično najvišjih pojavnic znotraj zaključenih skladenjskih drevesnic.

Povezave prvega nivoja označujejo razmerja znotraj besednih zvez, in sicer »dol« – jedro in določilo (tudi povedkovo) besedne zveze: *tistega*(dol) ← *večera*;[8] »del« – dele zloženega povedka: *sem*(del) ← popil; »prir« – jedra v prirednih zvezah znotraj stavka, »vez« – besede ali ločila v vezniški vlogi: *vzgoja*(prir) ← *in*(vez) ← i*zobraževanje*, ter »skup« – nepolnopomenske besede z močno tendenco sopojavljanja: *tako*(skup) → *da*. Za namene naše raziskave so najbolj pomembne povezave drugega nivoja, ki označujejo stavčne člene, in sicer: »ena« – osebek stavka: *žena*(ena) ← *vara*; »dve« – predmet stavka: *me*(dve) ← *vara*; »tri« – prislovno določilo lastnosti in »štiri« – prislovno določilo, kraja, časa, načina in vzroka: *tistega večera*(štiri) ← *popil*; *preveč*(tri) ← *popil*.

Z oznako »ena«, ki je v jedru našega zanimanja, so označeni tudi zanikani osebki in partitativni rodilniki, kar je v skladu s pripisom udeleženske vloge ACT na pomenski ravni. V nasprotju s pomensko ravnjo pa na skladenjski ravni t. i. logični osebki, ki so pomensko prepoznani kot vršilci, niso cilj povezave »ena«, pač pa so označeni kot predmeti stavka in so torej cilj povezave »dve«. Samostalniške zveze v vlogi povedkovega določila ob glagolu *biti* so na skladenjski ravni cilj povezave »dol«, medtem ko so na pomenski ravni tipično prepoznane kot izhodišče ali pobuda glagolskega dejanja, zato so označene kot »vršilec« (ACT), cilj glagolskega dejanja pa kot »prizadeto« (PAT): *Dogodek v Ankaranu*(dol-ACT) *je bila dramatična nesreča*(ena-PAT).

4.1.2 Pomenski nivo

Pomenska raven obsega pripis udeleženskih vlog stavčnim udeležencem. Na podlagi razmeroma široko mednarodno sprejetega sistema, ki izhaja iz Praške odvisnostne drevesnice in je med drugim uporabljen pri izdelavi češkega

8 Puščica označuje izhodišče, tj. jedro besedne zveze, in cilj označevanja.

vezljivostnega leksikona Vallex, predstavlja izhodišče za določanje udeleženskih vlog t. i. propozicija ali pomenska podstava povedi, ki ima dve temeljni sestavini:[9] povedje (ali predikat) in udeležence oz. participante. Udeleženci se nadalje delijo na delovalnike (argumente) in okoliščine (cirkumstante). Semantične oznake so torej glede na elemente propozicije pripisane sestavnim elementom zloženega povedka (povedkovo določilo in povedkov prilastek), delovalnikom (osebek, predmet) ter okoliščinam (prislovna določila). Tudi nabor udeleženskih vlog v tabeli 2, ki so bile uporabljene za označevanje korpusa (Gantar et al. 2018), je nastal na podlagi združevanja nabora oznak praške odvisnostne drevesnice PDT 2.0, in sicer vključuje 5 delovalniških, 17 udeleženskih in 3 oznake za elemente znotraj glagolske zveze.

Tabela 2: Udeleženske vloge v korpusu ssj500k

Udel. vloga	**SLO**	**OPIS**
ACT	vršilec, aktant	delujoči udeleženec, povzročitelj ali nosilec dejanja
PAT	prizadeto	prizadeti predmet dejanja
REC	prejemnik/koristnik	prejemnik, posredni udeleženec dejanja; nedelovalniški udeležnec, ki mu je dejanje v škodo ali v prid
ORIG	izvor/podedovanost	izhodišče, izvor/vir/povod dejanja; oseba, skupina, po kateri nekdo kaj podeduje, posvoji, dobi
RESLT	učinek	učinek, rezultat, cilj dejanja
LOC	kraj/smer	konkretna lokacija, kraj, mesto dejanja; smer v prostoru
SOURCE	začetna lokacija	začetna točka v prostoru
GOAL	končna lokacija	končna točka v prostoru
EVENT	dogodek	časovno-prostorsko določen dogodek
TIME	obdobje/sočasnost/začetek/konec	trenutek ali interval dejanja; trenutek ali interval, ki izvira iz dejanja; trenutek ali interval, ki sledi dejanju
DUR	trajanje stanja/dejanja	trajanje stanja, dejanja; konkretni trenutek začetka, konca
FREQ	pogostnost	frekvenca dejanja

9 Za razliko od slovenske slovnice, ki uporablja izraz *aktant* za vse delovalnike (Toporišič 2000: 492), uporabljamo v raziskavi izraz *aktant* po vzoru PDT za t. i. prvega delovalnika oz. vršilca dejanja.

Udel. vloga	SLO	OPIS
AIM	namen/namera	namen dejanja; namen gibanja, premikanja
CAUSE	Vzrok	vzrok dejanja
CONTR	dopustnost/protivnost	nepričakovana posledičnost dejanja
COND	pogojnost	pogoj za obstoj dejanja ali dogodka
REG	ozir/merilo/primerjava	ključno merilo, pravilo za ovrednotenje dejanja; primerjava
ACMP	spremstvo	predmet, oseba ali dogodek, ki spremlja dejanje ali udeležence
RESTR	omejitev	izjema, omejitev
MANN	način/rezultat	načinovna lastnost dejanja; rezultat ob koncu dejanja
MEANS	sredstvo	sredstvo ali orodje za izvedbo dejanja
QUANT	količina/razlika/ stopnja	kakovostna razlika med dogodki, stanji, predmeti; mera, razpon ali intenziteta dejanja ali okoliščine
MWPRED	večbesedni predikat	zveze z nedoločniki; fazni in nemodalni glagoli
MODAL	modalna zveza	zveze modalnega glagola in nedoločnika; zveze biti + modalni prislov
PHRAS	frazeološka enota	odvisni del glagolske frazeološke enote

V izhodišču našega zanimanja je udeleženska vloga aktant, ki v splošnem zajema vršilce, pobudnike dejanja. Pri pripisovanju te udeleženske vloge pa smo upoštevali še dodatna pravila. Samostalniška povedkova določila ob glagolu *biti* smo označevali kot »prizadeto« (PAT), samostalnike v imenovalniku, ki nastopajo kot osebki glagola *biti*, pa kot »vršilce« (ACT): *območje medenice*(ACT) *je središče telesa*(PAT); *problem beguncev*(ACT) *je stvar države*(PAT); *naprava BICOM*(ACT) *je zbirka impulzov*(PAT). Glede na omenjena izhodišča že na ravni označevanja korpusa predvidevamo tu največja odstopanja med skladenjskim in pomenskim nivojem, predvsem zaradi težav pri odločanju o izhodišču in določilu stavka na pomenski ravni in o polno- oz. nepolnopomenski vlogi glagola *biti*, ki odloča med osebkovo in povedkovodoločilno vlogo na skladenjski ravni.

4.1.3 Leksikalni nivo

Za pripis semantičnih tipov argumentom, ki jim je v korpusu pripisana vloga aktanta ali pa so cilj skladenjske povezave ena, smo uporabili semantično ontologijo za samostalnike, ki je bila oblikovana v okviru projekta Kolos (Kosem in Pori 2021). Ontologija vsebuje 21 krovnih skupin s podkategorijami, npr. ČLOVEK; ČLOVEK-aktivnost, ČLOVEK-lastnost, ki so v posameznih podkategorijah lahko še podrobneje razčlenjene, npr. ČLOVEK-lastnost- pripadnost (tabela 3). Za namene konkretne raziskave smo uporabili oznake prvega in drugega nivoja ter jih ročno pripisali 2.455 argumentom, ki smo jih nato upoštevali pri leksikalni analizi. Argumentom, kjer je cilj povezave (tj. potencialni osebek) glagol, semantičnega tipa nismo pripisovali. Sem se uvrščajo zveze z nedoločniki, npr. je najti, je poudariti, ni videti ipd. Semantični tipi so bili tako pripisani le samostalniškim argumentom glede na neposredni kontekst v stavku, npr. trgovina z orožjem (AKTIVNOST-DEJAVNOST--gospodarska-) : na policah trgovin (ARTEFAKT-zgradba-storitve-).

Tabela 3: Semantični tipi, določeni v okviru projekta Kolos (Kosem in Pori 2021)

01	ČLOVEK---	ČLOVEK-aktivnost--	ČLOVEK-aktivnost-nosilec-
			ČLOVEK-aktivnost-funkcija-
		ČLOVEK-lastnost--	
		ČLOVEK-naziv--	
02	TELO---		
03	ŽIVAL---		
04	RASTLINA---		
05	MIKROORGANIZEM---		
06	GLIVA---		
07	HRANA---		
08	SNOV---		
09	ARTEFAKT---		
10	PROSTOR---		
11	OBLIKA---		

12	POJAV---		
13	PROCES---		
14	MERA---		
15	ČAS---		
16	ČUSTVO---		
17	LASTNOST---		
18	STANJE---		
19	KOGNICIJA---		
20	AKTIVNOST---		
21	SKUPINSKO---		

4.2 Metodologija

Za namene pričujoče raziskave smo iz korpusa izluščili dva seta stavkov po naslednjih merilih: prvi set vsebuje stavke ali dele stavka, ki vsebujejo na semantični ravni oznako ACT in katerokoli povezavo na skladenjski ravni. Takih primerov je 5.238. Drugi set vsebuje stavke ali dele stavka, ki vsebujejo na skladenjski ravni povezavo »ena«, na semantični ravni pa udeleženec ni označen kot ACT, ampak mu je pripisana katerakoli druga udeleženska vloga. Takih primerov je v učnem korpusu 6.812. Med temi smo izločili tiste, kjer udeleženska vloga argumentom ni pripisana, saj v celotnem korpusu semantično označeni del vsebuje le četrtino korpusa, medtem ko je skladenjsko označenega 40 % korpusa. Takih stavkov je v korpusu 6.179, kar je naš skupni nabor analiziranih primerov zmanjšalo na 5.871.

V nadaljevanju smo 2.455 udeležencem, kar predstavlja približno 40 % vseh upoštevanih primerov, ročno pripisali še semantični tip na podlagi predstavljene ontologije. Seznami so nam nato služili za kvantitativno in kvalitativno analizo, ki nam je pomagala odgovoriti na vprašanja:

- kako se udeleženska vloga aktanta obnaša glede na tip skladenjske povezave;
- kateri delovalniki in okoliščine so cilj skladenjske povezave ena;
- kateri so semantični tipi udeležencev z lastnostjo ACT in udeležencev, ki so cilj skladenjske povezave ena.

4.3 Analiza in ugotovitve

4.3.1 Udeleženska vloga aktanta glede na tip skladenjske povezave

Kot prikazuje tabela 4, je v učnem korpusu ssj500k udeleženska vloga ACT pripisana 5.238 udeležencem. Kar v 93 % je tej udeleženski vlogi na skladenjski ravni pripisan osebek (povezava »ena«), kar je glede na merila skladenjskega označevanja pričakovano. V preostalih 7 % pa udeleženski vlogi ACT na skladenjski ravni ni pripisana vloga osebka, ampak predmeta (povezava »dve«; 4 %), prislovnega določila (povezava »tri«; 0,1 %; povezava »štiri« 1,5 %) ali povedkovega določila (povezava »dol« 1,3 %).

Tabela 4: Udeleženska vloga ACT glede na tip skladenjske povezave

jos-syn:*	**srl: ACT**	**5.238**	**100 %**	**primer**
jos-syn:ena	srl: ACT	4.878	93 %	sem izvedel, da me **žena** vara
jos-syn:dve	srl: ACT	201	4 %	**vodstvu** piranske občine je jasno
jos-syn:štiri	srl: ACT	79	1,5 %	na agronomskem **oddelku** so vpeljali študij živinoreje
jos-syn:dol	srl: ACT	71	1,3 %	**Dogodek** v Ankaranu je bila dramatična nesreča.
jos-syn:tri	srl: ACT	6	0,1 %	taka postanem tudi **sama**
jos-syn:del	srl: ACT	1		se **mi** zdi nespametno
jos-syn:vez	srl: ACT	2		**kaj** mu sledi; **kdo** in kako bo živel

ACT = ena (osebek stavka)

Udeleženci z lastnostjo ACT po pričakovanju v večini primerov ustrezajo cilju skladenjske povezave »ena«, ki predvideva osebke glagolskega dejanja: *Bilo je tistega dne, ko sem izvedel, da me žena*(ena-ACT) *vara*. V našem podatkovnem setu se to potrjuje v 93 %. Z oblikoskladenjskega vidika pričakujemo na osebkovem mestu predvsem samostalnike v imenovalniku (gl. primer zgoraj), v rodilniku pa, ko gre za pomen količine (t. i. partitivni osebek), npr. *predstavilo se bo sto razstavljavcev*(ena-ACT); *ljubiteljev konjeniškega športa*(ena-ACT) *je veliko*, ali za osebke zanikanega glagola, npr. *brez slovenščine ni Evrope*(ena-ACT); *zapletov*(ena-ACT) *še ni konec*. Kadar so na mestu osebka pridevniške besede, gre za posamostaljene pridevnike, npr. *prisotni*(ena-ACT) *so se odločili*; *tu se največ zadržujejo mladi*(ena-ACT), ali za izražanje določnosti ob izpustu

samostalniškega jedra zveze, npr. *menijo, da so nove mize bistveno preširoke, stare*(ena-ACT) *so bile boljše*. Zlasti stavčnočlenska vloga je kot sredstvo za izražanje besedilne določnosti v slovenščini še precej neraziskana.

ACT = dve (predmet stavka)

Kadar so aktanti na skladenjski ravni označeni kot predmet stavka, imamo, kot kažejo primeri, opravka z logičnimi osebki[10] v neimenovalniških sklonih, kjer po pričakovanju prihaja do diskrepanc na skladenjski in pomenski ravni zaradi odločitve, da logični osebki na skladenjski ravni niso prepoznani kot osebki stavka:

da vam(dve-ACT) *bo toplo*

nekaterim(dve-ACT) *ni všeč*

Sloveniji(dve-ACT) *ni mar*

V primeru, da sta v stavku izkazani dve udeleženski mesti, je logično pričakovati tudi diskrepanco med osebkovim mestom (ena) na skladenjski ravni in prizadetim udeležencem (PAT) na pomenski:

razvoj(ena-PAT) je(dve-ACT) *ne zanima*

cena(ena-PAT) se jim(dve-ACT) *zdi previsoka*

ne skrbi jih(dve-ACT) *usoda*(ena-PAT)

astronome(dve-ACT) *zanima nekaj drugega*(ena-PAT)

ženskam(dve-ACT) *se je težje odločiti*(ena-MWPRED)

Kot kažejo primeri, skladnja povsod, kjer obstaja možnost različne interpretacije osebkove in predmetne skladenjske vloge, favorizira imenovalniško obliko za pripis osebkove vloge in neimenovalniško za pripis vloge predmeta, in sicer ne glede na to, ali gre za vršilce dejanja ali nosilce stanja. Na pomenski ravni pa se ponekod koleba predvsem med vlogo aktanta (ACT) in prejemnika (REC), npr. *mu*(dve-REC) *je zmanjkalo denarja*(dve-ACT), kar kaže potrebo po natančnejši razmejitvi med pomenskima vlogama aktanta (ACT) in prejemnika (REC) pri izboljšanju učne množice v nadgradnji korpusa.

10 Toporišič 2000: 698: »O logičnem osebku govorimo tedaj, če vršilec dejanja itd. prav tako ni v imenovalniku, v drugem sklonu pa je zaradi zanikanega povedka ali zaradi pridevniškega količinskega prilastka, /.../ npr. *Jožku se je sanjalo*.«

ACT = tri, štiri (prislovno določilo)

Aktanti, ki so na skladenjski ravni opredeljeni kot cilj prislovnodoločilnih povezav, so prislovna določila kraja in časa (povezava »štiri«), redkeje načina (povezava »tri«). V prvem primeru gre tipično za prostorske realizacije osebka, kar je skladno z odločitvami na pomenski ravni označevanja: *V bolnišnici*(štiri-ACT) *bodo uvedli šolo za starše*; *Pri Fujifilmu*(štiri-ACT) *so objavili*; *V Banki Slovenije*(štiri-ACT) *menijo*.

Prislovno določilo načina (povezava »tri«) je na skladenjski ravni pripisano aktantom z leksikalno zapolnitvijo »sam«, kadar semantično nadomešča osebek: *taka postanem tudi sama*(tri-ACT), *sami*(tri-ACT) *smo si krivi*, vendar to razmerje med skladenjsko in pomensko ravnjo ni vedno dosledno, saj se pogosto beseda »sam« tako pomensko (način-MANN) kot skladenjsko (»tri« – prisl. določilo načina) razume kot načinovni prislov: *kot so sami*(tri-MANN) *povedali*; *vzgojili smo jih sami*(tri-MANN); *težko boste karkoli naredili sami*(tri-MANN). Prav v zvezi z zadnjimi primeri bi bilo pri nadaljnjem razvoju morfosintaktičnega označevanja smiselno poenotiti obe ravni v razmerju ena-ACT.

ACT = dol (določilo zloženega povedka)

Tretji sklop prestavljajo stavki, kjer je aktantom na skladenjski ravni pripisana vloga določila, vloga osebka pa je pripisana drugemu delu zloženega stavka. Gre za t. i. kopulativne stavke,[11] ki odkrivajo problem določila v zloženem povedku oz. problem osebkovega/predmetnega odvisnika v večstavčni povedi:

Dogodek v Ankaranu(dol-ACT) *je bila dramatična nesreča*(ena-PAT).

Gostja večera(dol-ACT) *bo Desa Muck*(ena-PAT).

Pravilen odgovor(dol-ACT) *je bil Grad Podčetrtek*(ena-PAT).

Večina potnikov(dol-ACT) *so bile ženske*(ena-PAT).

Prikazani primeri potrjujejo predvsem odločitve, ki so bile sprejete na skladenjski ravni, medtem ko bi bilo na semantični ravni smiselno dosledno opredeljevati samostalnike v vlogi povedkovega določila kot rezultat dejanja (RESLT). Uskladitev skladenjskega in pomenskega nivoja pri t. i. kopulativnih stavkih, bi torej izgledala takole:

Dogodek v Ankaranu(ena-ACT) *je bila dramatična nesreča*(dol-RESLT) – Kdo ali kaj je bila dramatična nesreča?

11 Po Toporišiču (2000: 557) imajo kopulativni stavki v vlogi povedka vez ali kopulo ter obvezno prisotno povedkovo določilo, ki ga definira kot nesamostojni stavčni člen, ki ga uporabljamo pri pomensko nepopolnih glagolih stanja in poteka – na primer pri glagolih *biti, postati* in *ostati.*

Gostja večera(dol-RESLT) *bo Desa Muck*(ena-ACT) – Kdo ali kaj bo gostja večera?

Pravilen odgovor(dol-RESLT) *je bil Grad Podčetrtek*(ena-ACT) – Kdo ali kaj je bil pravilen odgovor?

Večina potnikov(dol-RESLT) so bile ženske(ena-ACT) – *Kdo ali kaj je bila večina potnikov*?

4.3.2 Skladenjska vloga osebka glede na tip semantične vloge

Kot prikazuje tabela 5, je skoraj v 70 % skladenjskemu mestu osebka (ena) na pomenski ravni pripisana udeleženska vloga prizadeto (PAT). V približno četrtini primerov je nedoločniški glagolski osebek opredeljen kot del glagolske zveze (MWPRED – večbesedni predikat), v slabih 3 % je cilj skladenjske povezave »ena« REZULTAT dejanja, kar ustreza označevalnim smernicam (gl. tudi primere pri pogl. ACT = dol). Drugih udeleženskih vlog, ki so pripisane potencialnim osebkom, je zanemarljivo malo in predstavljajo posebnosti ali napake.

Tabela 5: Skladenjska povezava »ena« glede na udeleženske vloge (ki niso ACT)

srl: ni ACT	**jos-syn: ena**	**633**	**100 %**	**primer**
PAT	jos-syn: ena	435	69 %	**Stvar** je malce bolj zapletena.
MWPRED	jos-syn: ena	159	25 %	Obenem je treba **poudariti**.
RESLT	jos-syn: ena	18	2,8 %	Zdi se mi, **da ni opazila nikogar**. **Strah** me je tega.
QUANT	jos-syn: ena	8	1,2 %	Cena prevoza je 700 **tolarjev**.
REC	jos-syn: ena	7	1,1 %	Bi **mi** lahko tisto prinesli jutri zvečer?
PHRAS	jos-syn: ena	2	0,3 %	Spet mi je **beseda** za hip obstala v grlu. Samo osel gre dvakrat na led.
COND	jos-syn: ena	1	0,1 %	Kakšna svoboda za manjšino bi bila, **če bi se morali izogibati popisu**?
DUR	jos-syn: ena	1	0,1 %	Ne sprašujta, **koliko let je trajalo**.
REG	jos-syn: ena	1	0,1 %	Povsem očetov **sin** sem svojega skril.
TIME	jos-syn: ena	1	0,1 %	Se je to **stoletje** gospodarsko ugodno razvijalo.

ena = PAT (prizadeto)

Prvo skupino sestavljajo stavki, ki predstavljajo različne odločitve pri označevanju samostalniške besede ali zveze v povedkovem določilu na skladenjski in semantični ravni, kar smo že obravnavali pri problemu ACT = dve (predmet stavka), npr. *sem namreč zdravljeni alkoholik*(ena-PAT). Na semantični ravni se pri takih primerih interpretira neizraženost, vendar ne tudi semantična odsotnost vršilca, torej: *kdo*(ACT) *je zdravljeni alkoholik*, pri čemer smo že ugotovili nedoslednosti pri semantičnem označevanju povedkovodoločilnih elementov, kjer opažamo nedoslednosti med PAT in RESLT.

V drugo skupino sodijo stavki, kjer je skladenjskim osebkom (ena) na pomenski ravni pripisana vloga prizadeto (PAT). Gre predvsem za izražanje trpnika, kjer je osebek sicer semantično prisoten, vendar ne nujno izražen, npr.

stvar(ena-PAT) *je zapletena*(dol-RESLT)

amandma(ena-PAT) *je umaknjen*(dol-RESLT)

cilji(ena-PAT) *so se stalno spreminjali*(dol-RESLT)

Prvi dan v mesecu(ena-PAT) *je bil posvečen*(dol-RESLT) *soncu*

Temu sledi tudi različnost interpretacije na skladenjski in pomenski ravni.

ena = MWPRED (večbesedni predikat)

Gre za zveze z glagolom biti in nedoločnikom, kjer je nedoločnik na skladenjski ravni opredeljen kot osebek oz. cilj skladenjske povezave »ena«, na semantični ravni pa je v skladu s smernicami zveza označena kot MEPRED, torej kot zloženi predikat. Kot kažejo primeri, lahko iz korpusa izluščimo zlasti zveze z nedoločniki in modalnimi elementi (*treba, mogoče, lahko*): *bilo je čutiti*(ena-MWPRED), mogoče je delovati(ena-MWPRED), *težko je določiti*(ena-MWPRED), *treba je paziti*(ena-MWPRED).

ena = RESLT (rezultat dejanja)

Kadar je cilj skladenjske povezave »ena« – torej potencialni osebek – na pomenski ravni označen kot rezultat dejanja (RESLT), gre za problem različne interpretacije osebkovega in povedkovodoločilnega dela, ki smo ga izpostavili že pri kopulativnih stavkih (gl. ACT=dol), tu pa gre predvsem za interpretacijo osebkovega odvisnika, ki je na pomenski ravni lahko prepoznan kot rezultat dejanja (RESLT), na skladenjski pa kot osebkov odvisnik (ena):

zdi se, da ...(ena-RESLT)

od nekdaj ga je zanimalo, kako ...(ena-RESLT)

Razliko v označevanju na obeh nivojih je mogoče pojasniti z več parametri. Na pomenski ravni je odločitev za oznako RESLT pogojena v večini primerov s prisotnostjo vršilca ali nosilca dejanja v neimenovalniškem sklonu, npr. *od nekdaj ga*(dve-ACT) je zanimalo, čeprav je mogoče ugotoviti, da tako označevanje ni dosledno upoštevano, saj so povedkova določila lahko označena tudi kot »prizadeto« (PAT): *Desa Muck*(PAT) *bo gostja večera*(ACT). Na drugi strani je na skladenjski ravni odločitev za osebkov odvisnik pogojena s prepoznavanjem konkretnega glagola kot polnopomenskega in ne pomožnega, saj bi v nasprotnem primeru pričakovali oznako »dol« tudi pri osebkovih odvisnikih – na enak način, kot je ta skladenjska oznaka uporabljena za povedkova določila: *Desa Muck*(ena) *bo gostja večera*(dol).

4.3.3 Semantični tipi udeležencev na mestu osebka oz. s pomensko vlogo vršilca dejanja

V okviru raziskave smo želeli tudi ugotoviti, katerim semantičnim tipom, ki smo jih določili na podlagi slovenske semantične ontologije za samostalnike (Kosem in Pori 2021), tipično pripadajo skladenjski osebki (1.220 udeležencev) oz. vršilci dejanja (1.235 udeležencev), kot prikazuje tabela 6.

Tabela 6: Zastopanost semantičnih tipov na mestu osebka oz. v vlogi aktanta

Semantični tip	**Osebek**	**ACT**	**Udeleženci**	**%**
	1.220	1.235	2.455	100
ČLOVEK	**623**	**530**	**1.153**	**47**
SKUPINSKO	**81**	**194**	**275**	**11**
ARTEFAKT	**115**	**117**	**232**	**9**
A K T I V N O S T _ DEJANJE	**101**	**116**	**217**	**9**
ČAS	58	35	93	4
TELO	30	50	80	3
KOGNICIJA	23	65	88	3
PROSTOR	30	15	45	2
MERA	29	21	50	2
ŽIVAL	16	24	40	2
STANJE	16	22	38	2
RASTLINA	21	5	26	1

Semantični tip	Osebek	ACT	Udeleženci	%
ČUSTVO	19	0	19	1
SNOV	14	12	26	1
HRANA	11	5	16	1
LASTNOST	11	3	14	1
POJAV	8	13	21	1
PROCES	7	5	12	0
MIKROORGANIZEM	4	0	4	0
OBLIKA	1	3	4	0

V slabi polovici (47 %) so udeleženci na mestu osebka ali z lastnostjo aktant prepoznani v okviru krovnega semantičnega tipa ČLOVEK. Opazneje so zastopani še semantični tipi SKUPINSKO, ARTEFAKT in AKTIVNOST_DEJANJE (v tabeli 6 poudarjeno). Preostali semantični tipi so zastopani v manj kot 10 % označenih udeležencev.

Leksikalne enote, ki jih določa semantični tip ČLOVEK, je mogoče še podrobneje opredeliti s podkategorijami ČLOVEK-aktivnost, ČLOVEK--lastnost in ČLOVEK-naziv-ime. V prvi skupini lahko prepoznamo zlasti nosilce aktivnosti (avtor, kupec, kolesar, kmet), nosilce funkcije (*član, članica, dijak, minister*) in nosilce poklica (novinar, arhitekt, igralka). V drugi skupini prepoznamo nosilce lastnosti (*alkoholik, bolnica, lepotica*), ki je lahko geografska (*domačini, Američan, državljan, Celjan*), sorodstvena (*žena, otrok, oče*) ali nesorodstvena (*deklica, fant, moški, človek, ljudje*), v tretji pa nosilce naziva (*gospod, gospodična*) in lastna imena.

Semantični tip SKUPINSKO je mogoče pripisati 11 % udeležencev, zastopani pa so s samostalniki, ki označujejo skupine: *ljudje, ansambel, družina, ekipa, osebje, množica*, ter organizacije in ustanove: *agencija, banka, družba, klub, podjetje*.

Samostalniki v vlogi semantičnega tipa ARTEFAKT so pripisani udeležencem v 9 % ter označujejo dokumente (*besedilo, časopis, določba, knjiga*), zgradbe (*cerkev, hiša, kavarna*), naprave (*brskalnik, gonilnik, ključ, mobilnik*) in vozila (*avtomobil, helikopter*).

Semantični tip AKTIVNOST_DEJANJE je pripisan udeležencem v 9 % in vključuje samostalnike, kot so *akcija, boj, napad*, izstopajo še dogodki (*dogodek, festival*) in komunikacija (*argument, beseda, diskurz, govor, izjava*).

Na mestu aktantov in osebkov smo sicer identificirali samostalnike, ki pripadajo večini vseh predvidenih semantičnih kategorij (20 od 21), pri čemer se v 2 % ali več pojavljajo še samostalniki semantičnih tipov ČAS (*čas, dan, doba*), TELO (*oči, noga, kri*), KOGNICIJA (*država, politika, zakon*), PROSTOR (*območje, Slovenija, Evropa*), MERA (*cena, stopnja*), ŽIVAL (*čebela, maček*) in STANJE (*nevarnost, problem, bolečina*).

5 Zaključek in nadaljnje delo

V prispevku smo prikazali eno od možnosti uporabe učnega korpusa za slovenščino pri jezikoslovnih analizah ter metodologijo pri pripravi in analizi jezikovnih podatkov. Na podlagi izhodiščnega vprašanja: Kdo ali kaj je aktant v slovenščini, smo želeli ugotoviti, kako se udeleženska vloga aktanta obnaša glede na tip skladenjske povezave; kateri delovalniki in okoliščine so cilj skladenjske povezave »ena« ter kateri so semantični tipi udeležencev s to skladenjsko vlogo in lastnostjo vršilca.

Naše ugotovitve so na eni strani pokazale pričakovano ujemanje na skladenjski in pomenski ravni v razmerju osebek in vršilec dejanja, kjer odločitve sledijo skladenjskemu ali semantičnemu izhodišču pri označevanju podatkov. Na drugi strani so se pokazale diskrepance med skladenjsko in semantično ravnjo, ki podpirajo izhodiščne jezikoslovne odločitve oziroma gre za napake ali nedoslednosti, ki jih je smiselno odpraviti pri predvidenih nadgradnjah učnega korpusa.

Ena od možnosti izboljšave je zlasti premislek o usklajenem pomenskoskladenjskem označevanju logičnih osebkov, tj. s povezavo »ena« (osebek stavka) na skladenjski ravni (trenutno označeno z »dve« – predmet stavka) in z udeleženske vlogo ACT (aktant) na pomenski ravni, npr. *ne skrbi jih*(ena-ACT) *usoda*(dve-PAT). Na pomenski ravni se pri takih primerih pojavljajo nekonsistentnosti pri razmejevanju med udeleženskima vlogama aktanta (ACT) in prejemnika (REC), npr. *mu*(REC) *je zmanjkalo denarja*(ACT), kar kaže potrebo po natančnejši opredelitvi logičnih osebkov tudi s semantičnega vidika.

Po pričakovanju obstajajo najbolj očitne nekonsistentnosti pri označevanju na pomenski in skladenjski ravni pri interpretaciji jedra in določila v kopulativnih stavkih, kjer je v vlogi kopule glagol *biti*, v vlogi osebka in povedkovega določila pa samostalniška beseda ali zveza v imenovalniku: *Gostja večera bo Desa Muck*. Na skladenjski ravni velja pravilo, po katerem je povedkovo določilo cilj povezave »dol«, kadar je glagol v povedku (navadno *biti*) nepolnopomenski,

drugemu delu stavka pa je pripisana vloga osebka: *Gostja večera*(dol) *bo Desa Muck*(ena). Na pomenski ravni velja odločitev, da je samostalnik v osebku označen kot aktant (ACT), samostalnik v povedkovem določilu pa kot rezultat dejanja (RESLT), vendar pa se na mestu povedkovega določila pojavlja tudi oznaka prizadeto (PAT): *Gostja večera*(ACT) *bo Desa Muck*(PAT). Pri nadaljnjem označevanju učnega korpusa bi bilo mogoče sistematične nekonsistentnosti, ki so se pokazale na podlagi analiziranih primerov, razmeroma preprosto identificirati in jih strojno odpraviti ali ponovno premisliti jezikoslovna izhodišča.

Področje semantičnih ontologij je pri prepoznavanju konceptov in njihovih povezav na medjezikovni ravni eno od ključnih pri izgradnji t. i. semantičnih mrež oz. baz znanja. V raziskavi smo zato želeli preizkusiti tudi možnost pripisovanja semantičnih tipov samostalniškim udeležencem, ki zasedajo osebkovo mesto ali pa jim je pripisana vloga aktanta, in ugotoviti, ali je mogoče prepoznati določene semantične skupine, katerim ti udeleženci pripadajo. Poleg konkretnih leksikalnih zapolnitev, ki se kažejo na teh mestih, bi bilo mogoče v tem delu raziskavo razširiti tudi na morfološko raven, kjer se ponujajo možnosti prepoznavanja tako besednovrstne zastopanosti teh položajev kot tudi njihova morfološka zgradba. Na ta način bi bilo mogoče sistematično analizirati distribucijo posameznih sklonov na preučevanih udeleženskih mestih, vlogo zaimenskih besed, lastnoimenskih entitet itd.

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.

Literatura

Arhar Holdt, Špela, Dobrovoljc, Kaja, Logar, Nataša, 2019: Simplicity matters: user evaluation of the Slovene reference corpus. Language resources and evaluation 53/1. 173–190.

Arhar Holdt, Špela, Gorjanc, Vojko, 2007: Korpus FidaPLUS: Nova generacija slovenskega referenčnega korpusa. Jezik in slovstvo 52/2. 95–110.

Brank, Janez, 2019: Q-CAT Corpus Annotation Tool 1.1, Slovenian language resource repository CLARIN.SI, http://hdl.handle.net/11356/1282.

Dobrovoljc, Kaja, 2019: Q-CAT: Orodje za ročno označevanje in analizo besedilnih korpusov: priročnik za uporabo. https://slovnica.ijs.si/wp-content/uploads/2019/10/Q-CAT_prirocnik.pdf.

Dobrovoljc, Kaja, Erjavec, Tomaž, Ljubešić, Nikola, 2019: Improving UD processing via satellite resources for morphology. Proceedings of the Third Workshop on Universal Dependencies, UDW SyntaxFest, Paris, France, 26 August 2019. Association for Computational Linguistics. 24–34.

Dobrovoljc, Kaja, Erjavec, Tomaž, Krek, Simon, 2017: The Universal Dependencies Treebank for Slovenian. Proceedings of the EACL workshop. The 6th Workshop on Balto-Slavic Natural Language Processing, April 4, 2017 Valencia, Spain. Stroudsburg: ACL. 33–38.

Dobrovoljc, Kaja, Krek, Simon, Rupnik, Jan, 2012: Skladenjski razčlenjevalnik za slovenščino. Erjavec, T., Žganec Gros J. (ur.): Zbornik Osme konference Jezikovne tehnologije. Ljubljana: Institut Jožef Stefan.

Erjavec, Tomaž, Krek, Simon, Fišer, Darja, Ledinek, Nina, 2011: Projekt JOS: jezikoslovno označevanje slovenskega jezika = Project JOS: linguistic annotation of Slovene. Ljubljana: Institut Jožef Stefan, Odsek za tehnologije znanja. http://nl.ijs.si/jos/.

Erjavec, Tomaž, 2012: MULTEXT-East: Morphosyntactic Resources for Central and Eastern European Languages. Language Resources and Evaluation 46/1. 131–142.

Gantar, Polona, Štrkalj Despot, Kristina, Krek, Simon in Ljubešić, Nikola: 2018: Towards semantic role labeling in Slovene and Croatian. Proceedings of the conference on Language Technologies and Digital Humanities, Ljubljana, Slovenia: Faculty of Arts.

Hajič J., Hajičová, E., Mírovský, J., 2017: Prague Dependency Treebank. Ide N., Pustejovsky J. (eds): Handbook of Linguistic Annotation. Springer, Dordrecht.

Hajič, Jan, Panevová, Jarmila, Urešová, Zdeňka, Bémová, Alevtina, Kolářová, Veronika in Pajas, Petr, 2003: PDTVALLEX: Creating a Large-coverage Valency Lexicon for Treebank Annotation. E. Hinrichs, J. Nivre (ed): Proceedings of The Second Workshop on Treebanks and Linguistic Theories, volume 9 of Mathematical Modeling in Physics, Engineering and Cognitive Sciences, Vaxjo, Sweden: Vaxjo University Press. 57–68.

Jakopin, Primož, Bizjak Končar, Aleksandra, 1997: O strojno podprtem oblikoslovnem označevanju slovenskega besedila. Slavistična revija 45/3–4. 513–532.

Kosem, Iztok in Pori, Eva, 2021: Slovenske ontologije semantičnih tipov: samostalniki. I. Kosem (ur.): Kolokacije v slovenščini. Ljubljana: Znanstvena založba Filozofske fakultete Univerze v Ljubljani (v tisku).

Krek, Simon, Erjavec, Tomaž, Dobrovoljc, Kaja, Gantar, Polona, Arhar Holdt, Špela, Čibej, Jaka, Brank, Janez, 2020: The ssj500k training corpus for Slovene language processing. Fišer, Darja, Erjavec, Tomaž (ur.): Jezikovne tehnologije in digitalna humanistika: zbornik konference: 24.–25. september 2020, Ljubljana: Inštitut za novejšo zgodovino. 24–33.

Krek, Simon, 2010: Pridobivanje jezikovnih podatkov iz besedilnih korpusov za namen izdelave enojezičnih slovarjev in slovnic: doktorska disertacija. Ljubljana: FF.

Ledinek, Nina, Erjavec, Tomaž, 2009: Odvisnostno površinskoskladenjsko označevanje slovenščine: specifikacije in označeni korpusi. Zbornik Simpozija Obdobja: Infrastruktura slovenščine in slovenistike. Ljubljana.

Toporišič, Jože, 2000: Slovenska slovnica. Maribor: založba Obzorja.