

ZAHVALA

Zahvala ob oddaji doktorske raziskave gre v prvi vrsti mentorju dr. **Vojku Gorjancu**: za potrpežljivo čakanje na rezultate, hitro branje in dragocene komentarje, še posebej pa za nalezljivo navdušenje nad odkrivanjem novih idej, ljudi in dežel. Enako pomembna zahvala gre **Simonu Kreku** za nedeljske pogovore, nove izzive in vizijo prihodnosti. Hvaležna sem za vse dane priložnosti, še bolj pa za izkazano zaupanje, ki mi je omogočilo priložnosti tudi izkoristiti.

Zahvala gre celotni ekipi podjetja Amebis, d. o. o., Kamnik, ki me je kot mlado raziskovalko sprejela v svoje vrste in mi s podatki, nasveti ter programsko podporo pomagala pri pripravi doktorske naloge. Posebno zahvalo dolgujem direktorju **Miru Romihu** za razumevanje dela doma in v tujini.

Številni drugi vzorniki in podporniki bodo na tem mestu ostali neimenovani. Prav tako moji najpomembnejši bližnji, hvala vsem, ker ste, kar ste, vztrajajte z mano še naprej. Posebna zahvala za vso podporo gre družini, ki ji posvečam to delo.

*T'ain't what you do it's the way that you do it
T'ain't what you do it's the time that you do it
T'ain't what you do it's the place that you do it
And that's what gets results*

Sy Oliver & Trummy Young

KAZALO VSEBINE

0 UVOD	I
I RAZISKOVALNO IZHODIŠČE.....	2
1 KORPUSNI PRISTOP V JEZIKOSLOVJU.....	2
2 RAČUNALNIŠKO JEZIKOSLOVJE IN OBDELAVA NARAVNEGA JEZIKA.....	4
3 KORPUSI IN LEKSIKOGRAFIJA	6
II OZNAČEVANJE KORPUSNIH BESEDIL TER LUŠČENJE LEKSIKALNIH PODATKOV	8
1 OZNAČEVANJE KORPUSNIH BESEDIL.....	8
1.1 Popolni korpusni pristop.....	8
1.2 Oznake korpusa FidaPLUS	9
1.2.1 Leme.....	9
1.2.2 Oblikoskladenjske oznake	10
1.2.3 Nabor oblikoskladenjskih oznak JOS	11
1.2.4 Uspešnost oblikoskladenjskega označevanja slovenščine	12
2 LUŠČENJE LEKSIKALNIH PODATKOV IZ KORPUSA	13
2.1 Luščenje terminoloških besednih zvez	14
2.2 Luščenje večdelnih besednih nizov	15
III LEKSIKALNA PODATKOVNA ZBIRKA ZA SLOVENŠČINO	16
1 OPREDELITEV	16
1.1 Slovarska leksikalna zbirka za slovenščino	17
1.2 Slovenski wordnet	18
2 LEKSIKALNA ZBIRKA ASES	19
2.1 Organizacija leksikalne zbirke	20
2.1.1 Beseda (oblika)	21
2.1.2 Pomen	23
2.2 Nabor informacij v leksikalni zbirki	25
2.2.1 Primer 1 – pajek	26

2.2.2	Primer 2 – moder	28
2.2.3	Primer 3 – izdati	31
3	NADGRADNJA ZBIRKE NA OSNOVI KORPUSNIH PODATKOV	35
3.1	Izhajanje iz korpusnih jezikovnih podatkov.....	35
3.2	Upoštevanje enopomenskosti leksikalne enote.....	37
3.3	Odločitev za polavtomatsko metodo nadgradnje.....	37
IV	RAZISKOVALNI CILJI IN POSTOPKI.....	39
1	NAMEN RAZISKAVE	39
1.1	Raziskovalno izhodišče.....	39
1.2	Ključni pojmi.....	39
1.3	Raziskovalni cilji in vprašanja.....	41
2	POTEK RAZISKAVE	42
2.1	Priprava podatkovnih virov.....	42
2.1.1	Izdelava podkorpusa	42
2.1.2	Pretvorba podkorpusa za pripravo seznama vzorcev	45
2.1.2.1	Preimenovanje oznak za obravnavane leme.....	46
2.1.2.2	Odstranitev oznak iz besedila.....	46
2.1.2.3	Menjava ločil s posebnimi oznakami.....	47
2.1.2.4	Dodajanje oznake za konec besedilnega dela	47
2.1.2.5	Končno stanje pretvorjenega besedila	47
2.2	Priprava seznama vzorcev.....	48
2.2.1	Izdelava seznamov vzorcev s programom Oxford Wordsmith Tools	48
2.2.2	Prva obdelava seznama vzorcev.....	49
2.3	Analiza najpogostejših vzorcev	50
2.3.1	Priprava podatkov	50
2.3.2	Razvrščanje vzorcev v skupine za nadaljnjo analizo.....	52
2.3.3	Združevanje vzorcev v vzorčne tipe	52
2.4	Luščenje besednih nizov in evalvacija izluščenih podatkov	53
2.4.1	Napake avtomatskega označevanja.....	53
2.4.2	Kategorije oblikoskladenjskih oznak	53
2.4.3	Organizacija vzorčnih tipov glede na relevantnost za luščenje.....	53
3	PROGRAMI	54
3.1	Priprava in obdelava podkorpusa	54
3.1.1	Štetje lem v podkorpusu	54

3.1.2	Preimenovanje obravnavanih lem	54
3.1.3	Odstranitev oznak v besedilu podkorpora	55
3.1.4	Pretvorba ločil v oznake	56
3.1.5	Dodajanje oznake za konec besedilnega dela	56
3.2	Priprava vzorčnih tipov	57
3.2.1	Odstranjevanje vzorcev, vsebujočih določene oznake.....	57
3.2.2	Luščenje vzorčnih zapolnitev.....	57
3.2.3	Luščenje vzorcev za posamezni vzorčni tip	58
3.3	Luščenje besednih nizov	59
3.3.1	Osnovno luščenje	59
3.3.2	Dva samostalnika	59
3.3.3	Pridevnik s samostalniškim določilom.....	60
3.3.4	Samostalnik s pridevniškim določilom	61
3.3.5	Neujemalne kombinacije.....	62
3.3.6	Priredne zveze	62
3.3.7	Štetje besednih nizov	63
3.4	Dodatna obravnava samostalniških zvez	63
3.4.1	Preverjanje relevantnosti vzorčnih tipov za luščenje samostalniških zvez	63
V ANALIZA.....		65
1	VZORCI S SAMIMI POLNOPOMENSKIMI BESEDNIMI VRSTAMI.....	65
1.1	Pajek – dvodelni vzorci	66
1.1.1	Luščenje zvez Sam + Sam	67
1.1.2	<i>Pajek</i> + Sam	68
1.1.2.1	Analiza označenosti.....	72
1.1.3	Sam + <i>Pajek</i>	74
1.1.4	Luščenje zvez Prid + Sam ter Sam + Prid	76
1.1.5	<i>Pajek</i> + Prid.....	77
1.1.5.1	Analiza označenosti.....	79
1.1.6	Prid + <i>Pajek</i>	79
1.1.6.1	Analiza označenosti.....	81
1.1.7	<i>Pajek</i> + Prisl	82
1.1.8	Prisl + <i>Pajek</i>	82
1.1.9	Luščenje zvez Sam + Glag ter Glag + Sam.....	83
1.1.10	Glag + <i>pajek</i>	83
1.1.11	<i>Pajek</i> + Glag.....	84
1.2	Pajek – tridelni vzorci.....	85
1.2.1	Prid + Prid + <i>pajek</i>	86
1.2.2	Prid + <i>pajek</i> + Sam	86
1.2.3	Preostali vzorci	87
1.3	Strasten – dvodelni vzorci.....	88
1.3.1	<i>Strasten</i> + Sam.....	89
1.3.1.1	Analiza označenosti.....	90

1.3.2	Sam + <i>strasten</i>	91
1.3.3	<i>Strasten</i> + Prid	92
1.3.3.1	Analiza označenosti	92
1.3.4	Prid + <i>strasten</i>	92
1.3.4.1	Analiza označenosti	93
1.3.5	Prisl + <i>strasten</i>	94
1.3.6	Glag + <i>strasten</i>	95
1.4	Strasten – tridelni vzorci	95
1.4.1	Prisl + <i>strasten</i> + Sam	96
1.4.2	Glag + <i>strasten</i> + Sam	97
1.4.3	<i>Strasten</i> + Prid + Sam	97
1.4.4	<i>Strasten</i> + Sam + Prid	98
1.4.5	<i>Strasten</i> + Sam + Sam	98
1.4.6	Preostali vzorci	99
1.5	Plesati – dvodelni vzorci	99
1.5.1	Prisl + <i>plesati</i>	100
1.5.1.1	Analiza označenosti	102
1.5.2	<i>Plesati</i> + Prisl	103
1.5.2.1	Analiza označenosti	104
1.5.3	<i>Plesati</i> + Prid	105
1.5.4	Prid + <i>plesati</i>	105
1.5.5	Glag + <i>plesati</i>	106
1.5.6	<i>Plesati</i> + Sam	106
1.5.6.1	Analiza označenosti	107
1.5.7	Sam + <i>plesati</i>	108
1.6	Plesati – tridelni vzorci	109
1.6.1	<i>Plesati</i> + Prid + Sam	110
1.6.2	Preostali vzorci	111
1.7	Temeljito – dvodelni vzorci	111
1.7.1	Prisl + <i>temeljito</i>	112
1.7.2	<i>Temeljito</i> + Prisl	113
1.7.2.1	Analiza označenosti	114
1.7.3	Luščenje zvez Prisl + Prisl	114
1.7.4	<i>Temeljito</i> + Glag	115
1.7.5	Glag + <i>temeljito</i>	117
1.7.6	Sam + <i>temeljito</i>	118
1.7.7	<i>Temeljito</i> + Sam	118
1.7.8	<i>Temeljito</i> + Prid	119
1.8	Temeljito – tridelni vzorci	121
1.8.1	Prisl + <i>temeljito</i> + Glag	121
1.8.2	Glag + <i>temeljito</i> + Glag	122
1.8.3	<i>Temeljito</i> + Glag + Sam	123
1.8.4	Sam + <i>temeljito</i> + Glag	124
1.8.5	Prid + Sam + <i>temeljito</i>	125
1.8.6	Preostali vzorci	125

2	VZORCI S PREDLOGOM	126
2.1	Dvodelni vzorci	126
2.2	Pajek – dvodelni vzorci	127
2.2.1	Pred + <i>pajek</i>	127
2.2.1.1	Analiza označenosti	127
2.2.2	<i>Pajek</i> + Pred	128
2.3	Strasten – dvodelni vzorci	128
2.3.1	Pred + <i>strasten</i>	128
2.3.2	<i>Strasten</i> + Pred	129
2.4	Plesati – dvodelni vzorci	129
2.4.1	Pred + <i>plesati</i>	129
2.4.1.1	Analiza označenosti	129
2.4.2	<i>Plesati</i> + Pred	130
2.5	Temeljito – dvodelni vzorci	131
2.5.1	Pred + <i>temeljito</i>	131
2.5.1.1	Analiza označenosti	131
2.5.2	<i>Temeljito</i> + Pred	132
2.6	Kolokacijska analiza predloga	132
2.7	Tridelni vzorci	133
2.8	Pajek – tridelni vzorci	133
2.8.1	<i>Pajek</i> + Pred + Sam	134
2.8.2	Sam + Pred + <i>pajek</i>	134
2.8.3	Glag + Pred + <i>pajek</i>	135
2.8.4	Preostali vzorci	135
2.9	Strasten – tridelni vzorci	135
2.9.1	Pred + <i>strasten</i> + Sam	136
2.9.2	<i>Strasten</i> + Sam + Pred	136
2.9.3	Preostali vzorec	137
2.10	Plesati – tridelni vzorci	137
2.10.1	<i>Plesati</i> + Pred + Sam	137
2.10.2	<i>Plesati</i> + Pred + Prid	138
2.10.3	Prisl + <i>plesati</i> + Pred	139
2.10.4	Pred + Sam + <i>plesati</i>	139
2.10.5	Preostala vzorca	140
2.11	Temeljito – tridelni vzorci	140
2.11.1	Pred + Sam + <i>temeljito</i>	140
2.11.2	<i>Temeljito</i> + Glag + Pred	141
3	VZORCI Z VEZNIKOM	142

3.1	Nerelevantni vzorci	143
3.2	Pajek – tridelni vzorci	143
3.2.1	Simetrična vzorčna tipa	143
3.2.1.1	Analiza označevanja	145
3.2.2	Preostali vzorec	145
3.2	Strasten – tridelni vzorci	143
3.3.1	Simetrična vzorčna tipa	146
3.3.1.1	Analiza označenosti	147
3.3.2	Preostali vzorci	148
3.4	Plesati – tridelni vzorci	148
3.4.1	Simetrična vzorčna tipa	149
3.4.2	Preostali vzorci	150
3.5	Temeljito – tridelni vzorci	151
3.5.1	Simetrična vzorčna tipa	151
3.5.2	Preostala vzorca	153
4	VZORCI Z DRUGIMI BESEDNIMI VRSTAMI	153
4.1	Vzorci z oznako za pomožni glagol	153
4.2	Vzorci z oznako za členek	155
4.3	Vzorci z oznako za okrajšavo	157
4.4	Vzorci z oznako za zaimek	158
4.5	Vzorci z oznako za števniki	158
4.6	Delež vnaprej odstranjenih vzorcev	159
5	VZORCI Z NELEMATIZIRANIMI BESEDAMI	161
5.1	Pajek	161
5.2	Strasten	162
5.3	Plesati	163
5.4	Temeljito	165
6	MANJ POGOSTI VZORCI Z LEMO PAJEK	167
VI	REZULTATI ANALIZE	171
1	IZBOLJŠAVA AVTOMATSKEGA OBLIKOSKLADENJSKEGA OZNAČEVANJA	171

2	DOLOČANJE RELEVANTNOSTI VZORČNIH TIPOV	172
2.1	Delitev vzorčnih tipov glede na rezultate analize	172
2.1.1	Za luščenje relevantni vzorčni tipi	172
2.1.2	Vzorčni tipi, ki vsebujejo glagol	174
2.1.3	Za luščenje nerelevantni vzorčni tipi	175
2.2	Luščenje samostalniških besednih zvez	178
2.2.1	Novi vzorčni tipi	178
2.2.2	Luščenje besednih zvez	179
2.2.3	Analiza izluščenih zvez	181
VII	SKLEP	183
VIII	POVZETEK	186
IX	ABSTRACT	193
X	PRILOGE	201
1	NABOR OBLIKOSKLADENJSKIH OZNAK JOS	201
2	NABOR OBLIKOSKLADENJSKIH OZNAK MULTTEXT-EAST	204
3	LEGENDA ZAPISA VZORČNIH TIPOV	207
4	PATTERN-TYPE FORMAT	208
XI	LITERATURA	209
XII	KAZALA	220
1	KAZALO TABEL	220
2	KAZALO SLIK IN GRAFOV	224
3	AVTORSKO KAZALO	225
4	STVARNO KAZALO	226

UVOD

O

Pričujoče delo združuje jezikoslovni pogled s postopki avtomatske obdelave naravnega jezika – in kot druge raziskave z interdisciplinarnimi izhodišči prinaša vse na slednje vezane prednosti ter slabosti. Za uvod se zdi zato primerno nameniti nekaj besed stanju, ki ga prinaša presek računalništva ter jezikoslovja za slovenščino v danem trenutku.

Kaj lahko jezikoslovcu sodelovanje pri razvoju jezikovnih tehnologij ponudi? Na eni strani pogled v področje pred časom neslutnih možnosti razvoja aplikacij, ki postajajo nepogrešljiv pripomoček tvorcem besedil pri vsakodnevni rabi jezika; gradnja in označevanje jezikovnih virov, razvoj avtomatskega slovnicega pregledovanja, strojnega prevajanja itd. so področja, ki jezikoslovčev strokovni doprinos vsekakor potrebujejo. Na drugi strani iz sodelovanja pri razvoju izdelkov, vedno usmerjenem v reševanje trenutnih problemov in iskanje kompromisov med teorijo in prakso, sledi uvid, da je omenjeni doprinos produktiven le v primeru, da zmore (na ustrezno fleksibilen način) izkazovati določeno stopnjo samoomejevanja. Na vsak način torej velik izziv.

V slovenskem prostoru potekajo računalniškojezikoslovne raziskave v okviru številnih institucij, tako raziskovalnih kot razvojnih. Pomemben doprinos, če jih naštejemo le nekaj, prinašajo raziskovalci Filozofske fakultete, Inštituta Jožef Stefan, Inštituta Frana Ramovša ZRC SAZU, zavoda Trojina, v pričujočem delu večkrat omenjenega podjetja Amebis itd. Ob specializiranih raziskavah manjšega obsega so bili v času od potrditve doktorske teme sprejeti nekateri večji projekti, temelječi na institucionalnem povezovanju in usmerjeni k jasno začrtanim ciljem. Taka sta denimo *Jezikoslovno označevanje slovenščine*, še zlasti pa *Sporazumevanje v slovenskem jeziku*, ki za cilje postavljata izboljšanje kvalitete ter dostopnosti jezikovnih virov za slovenski jezik – tako za obdelavo naravnega jezika kot za opisno jezikoslovje.¹ Nekateri rezultati omenjenih raziskav in projektov so predstavljeni tudi v pričujočem delu.

Trenutno se v Sloveniji veliko pozornosti posveča razvoju korpusnih virov in izboljšavi njihovega avtomatskega označevanja, iz korpusov se z različnimi metodami luščijo raznovrstni leksikalni podatki, pripravlja se leksikalna zbirka, ki bo osnova novim slovarskim priročnikom. Kar se tiče obdelave naravnega jezika, smo bili v zadnjih letih priča porastu statističnih metod, učenja sistemov na osnovi učnih podatkov ter poskusov integracije statistike s starejšimi, na osnovi pravil delujočimi, sistemi.

Kot bo vidno v nadaljevanju, pričujočo raziskavo uvrščamo v področje računalniške leksikografije. Ker se ukvarja predvsem z avtomatskim delom metode, ne pa toliko z (sicer na vseh mestih predvideno) jezikoslovno analizo podatkov, je na prvi pogled bolj računalniška kot leksikografska. Ne gre pa spregledati dejstva, da jezikoslovje predstavlja temelje metode in na vseh mestih – kljub zgoraj omenjeni potrebi po kompromisih – vodi njen razvoj in predloge za nadgradnjo. Ker vsekakor skuša biti v prvi vrsti uporabno, je mogoče doktorsko delo širše uvrstiti v področje uporabnega jezikoslovja; še zlasti, če slednje razumemo – malo za šalo, predvsem pa v razmislek – kot avtor članka o jezikoslovju v slovenski *Wikipediji*: »Uporabno jezikoslovje je del jezikoslovja, ki uporablja jezikoslovne teorije za reševanje problemov v resničnem svetu.«²

Doktorska naloga prinaša v prvih treh poglavjih opredelitev raziskovalnih izhodišč: prvo poglavje predstavlja korpusnojezikoslovne teoretske podlage, drugo označevanje korpusnih besedil ter luščenje leksikalnih podatkov ter tretje pojem leksikalne podatkovne zbirke. Četrto poglavje prinaša opis raziskovalnih ciljev ter postopkov. Peto in šesto poglavje prinašata praktični del naloge: analizo podatkov ter strnjene rezultate analize. Sledi še sklep z evalvacijo raziskave in, kot je v navadi, povzetek v slovenskem ter angleškem jeziku.

¹ Internetni strani projektov – <<http://nl.ijs.si/jos/>> ter <<http://www.slovenscina.eu/>>.

² <<http://sl.wikipedia.org/wiki/Jezikoslovje>>. Vse v doktorski nalogi navedene internetne strani so bile glede dostopa preverjene 31. 3. 2009.

RAZISKOVALNO IZHODIŠČE

Poglavje predstavlja opredelitev raziskovalnega izhodišča pričujočega doktorskega dela, ki prinaša preplet korpusnojezikoslovnih ugotovitev z metodami obdelave naravnega jezika, z natančnejšo umestitvijo v področje računalniške leksikografije. V nadaljevanju so na kratko predstavljene tri teme: (I) korpusni pristop v jezikoslovju, (II) obdelava naravnega jezika v razmerju do računalniškega jezikoslovja ter (III) korpusna ter računalniška leksikografija.

1 Korpusni pristop v jezikoslovju

Ko pride do opredeljevanja mesta korpusnega jezikoslovja znotraj jezikoslovja, je med pogosto navajanimi Leechev prispevek iz leta 1992, *Corpora and theories of linguistic performance*. Geoffrey Leech v tem prispevku priznava, da korpusnega jezikoslovja zaradi neobstoja lastnega raziskovalnega področja ne gre obravnavati na enakem nivoju kot denimo sociolingvistiko ali psiholingvistiko, kljub temu pa se odreka redukciji korpusnega jezikoslovja na zgolj metodološki pristop. Korpusno jezikoslovje mu pomeni nov raziskovalni podvig, obenem pa novo filozofijo pogleda na jezik, ki pred pojavom korpusa kot metodološkega orodja ni bila mogoča.

Kot štiri ključne značilnosti korpusnega jezikoslovja – značilnosti, ki so naravna posledica odločitve za uporabo korpusa pri jezikoslovnem delu – Leech navaja osredotočenost na: (I) jezikovno **performanco**, ne kompetenco, (II) jezikoslovni **opis**, ne na jezikovne univerzalijske, (III) **kvantitativne ter tudi kvalitativne** jezikovne modele, (IV) **empiristični**, ne racionalistični znanstveni pristop (Leech 1992: 127).

Kot številni drugi (korpusni) jezikoslovci Leech korpusno jezikoslovje postavlja predvsem v opozicijo jezikoslovju, ki temelji na ugotovitvah Noama Chomskega in kroga jezikoslovcev s podobnimi jezikoslovnimi izhodišči.³ Na tem mestu vprašanja zgodovine korpusnega jezikoslovja puščamo ob strani⁴, s stališčem, da novosti v obravnavi jezika, ki jih korpusno jezikoslovje brez dvoma prinaša, ne gre iskati apriorno v razmerju do ostalih obstoječih jezikoslovnih paradig, ampak ob osredotočanju na izkušnje korpusnojezikoslovne prakse, kakor o tem piše denimo Teubert:

»Dokler ni bilo mogoče obdelati velikih količin jezikovnih podatkov na sistematičen proceduralen način, ni bilo nobene možnosti, da bi kompleksne povezave sopojavljanja med elementi besedila (tj. besedami) opisali drugače kot s skladenjskimi pravili. Takšna pravila opisujejo vedenje razredov elementov, npr. samostalnikov, v razmerju do vedenja razredov drugih elementov, npr. pridevnikov v vlogi prilastkov. [...] Da smo pripravljeni klasificirati sopojavljanje določenih elementov besedila kot segment besedila, ki spada skupaj [...], moramo segment besedila prepoznati kot ponavljajoč pojav. [...] Korpusno jezikoslovje je metodološko sposobno opisati sopojavljanje kot statistični pojav.« (Teubert 2000 v Gorjanc in Krek 2005: 114)

Resnično novost v jezikoslovju predstavlja na besedilni korpus kot kombinacijo avtentičnega besedilnega gradiva v elektronski obliki ter programskega orodja za obdelavo jezikovnih podatkov vezana metodološka

³ V zvezi s tem so zanimive ugotovitve Jacqueline Léon, ki v svojem prispevku *Claimed and unclaimed sources of corpus linguistics* obravnava korpusnojezikoslovno sklicevanje na večni spor s Chomskim kot enega od poskusov retrospektivne konstrukcije zgodovine korpusnega jezikoslovja, ki naj bi služila vzpostavitvi slednjega kot ločenega raziskovalnega izhodišča. Za temelj argumenta navaja zgoraj navedeni Leechev članek in ga označuje kot poskus vzpostavitve »antičomskijanske« jezikoslovne paradigme (Léon 2005 v Teubert in Krishnamurty 2007: 327).

⁴ Teme se dotikajo številni avtorji, mdr. npr. Gorjanc 2005, McEnery in Wilson 1996, McEnery et al. 2006, Teubert in Krishnamurty (ur.) 2007.

revolucija. Nove metodološke možnosti, ki jih je iznajdba korpusa prinesla s sabo, so predpogoj za nove ugotovitve, ki šele v sekundarni fazi spreminjajo pogled na jezik oz. jezikoslovje do te mere, da je o korpusnem jezikoslovju mogoče govoriti o več kot novi metodi.

Po letih korpusne gradnje, razvoja programske opreme za označevanje ter obdelavo korpusnih besedil, predvsem pa uzaveščanja raznovrstnih tipov vprašanj, na katera korpusi lahko odgovarjajo, zveza *korpusno jezikoslovje* ohranja dvojni značaj. Na eni strani označuje ločeno interdisciplinarno (a vedno iz jezikoslovja izhajajočo) raziskovalno disciplino, na drugi aktivnosti, vezane na jezikoslovno analizo korpusov, pri katerih korpus večinoma predstavlja metodološko dopolnitev ali nadgradnjo že obstoječih raziskovalnih pristopov.

Širše gledano **korpusno jezikoslovje** pomeni vse aktivnosti, povezane s korpusi: od snovanja ter gradnje korpusov, razvoja raznovrstne programske opreme za delo s korpusi, na korpusih temelječih produktivno usmerjenih raziskav (tako na področjih razvoja jezikovnih tehnologij kot opisnega jezikoslovja) do teoretičnojezikoslovnih raziskav vseh jezikovnih ravnin (Gorjanc 2005: 23). Zveza v najširšem smislu torej pomeni zbirni pojem za številne interdisciplinarne aktivnosti v zvezi s korpusi.

Znotraj samega korpusnega jezikoslovja se za opredelitev najbolj z jezikoslovjem povezanih aktivnosti predlaga poimenovanje **korpusni pristop**. Korpusni pristop v jezikoslovju je definiran kot kombinacija kvantitativne ter kvalitativne analize velikega nabora avtentičnega jezikovnega gradiva, ki poteka z uporabo specializiranih računalniških programov.⁵ Kvantitativna analiza pomeni avtomatski (strojni, računalniški) del analize, ki ga – upoštevajoč raziskovalčeva navodila – izvede program. Ko so podatki pripravljeni, nastopi kvalitativni del analize, ki temelji na raziskovalčevi interpretaciji in nadaljnji obdelavi dobljenih rezultatov (Gorjanc 2005: 23–24).

Glede na namen raziskovalčeve uporabe korpusnega pristopa se ločita dve metodi: (I) uporaba korpusa kot **vir preverjanja hipotez** ter (II) uporaba korpusa kot **vir gradnje hipotez** (Gorjanc 2005: 24).

Na eni strani se korpusi in korpusna orodja uporabljajo kot dopolnilo že obstoječih jezikoslovnih metod. Korpus služi kot orodje za preverjanje ali dopolnjevanje obstoječih hipotez o jeziku, kar tipično vključuje združevanje rezultatov korpusne analize s podatki iz slovarjev, slovnice, pravopisnih priročnikov, šolskih učbenikov itd.

Na drugi strani se korpus uporablja kot vir gradnje hipotez o jeziku. Izhodišče jezikoslovne raziskave so iz korpusa pridobljeni podatki, priprava katerih skuša biti v čim večji meri objektivna in neodvisna od morebitnih jezikoslovčevih predpostavk o jeziku. Jezikoslovčeva interpretacija jezikovnih podatkov nastopi v večji meri šele za tem, ko je stanje ugotovljeno na podlagi jezika oz. iz jezika samega. Podatki, temelječi na statistično obdelanem jezikovnem materialu, namreč pogosto prinašajo jezikovnouporabniški intuiciji o tipičnosti nasprotujočo sliko jezika.⁶

Glede na zapisano je možna opredelitev pričujoče raziskave kot korpusnojezikoslovnega dela, ki korpus uporablja kot vir gradnje hipoteze, s poudarkom, da v raziskavi preučevana metoda temelji na oblikoskladenjsko označenem ter lematiziranem korpusu. Vprašanje označevanja korpusnih besedil v odnosu do odpovedi vnaprejšnji interpretaciji jezikovne realnosti je natančneje opredeljeno v poglavju II-1.

⁵ Na drugih mestih v literaturi se zgoraj opisana podskupina korpusnih aktivnosti pogosto poimenuje kar *korpusno jezikoslovje*, tako npr. Ooi definira korpusno jezikoslovje kot raziskovalno področje, ki se ukvarja z raziskovanjem jezika na osnovi pisnih ali govornih korpusov, običajno s pomočjo računalnika, s katerim se podatki shranjujejo, procesirajo ter analizirajo (Ooi 1998: 34).

⁶ Pri čemer je nujno upoštevati dejstvo, da je privzemanje tako korpusnih podatkov kot jezikoslovne intuicije kot edino možno metodološko izhodišče z današnjega vidika nesmiselno (Gorjanc 2003: 20).

2 Računalniško jezikoslovje in obdelava naravnega jezika

Pred definicijo obdelave naravnega jezika (ang. *natural language processing*) je potrebno zapisati nekaj besed o pojmu računalniško jezikoslovje (ang. *computational linguistics*). Podobno kot za korpusno je v literaturi za računalniško jezikoslovje zaslediti več opredelitev.

Računalniško jezikoslovje avtorjem na eni strani pomeni tisti del jezikoslovja, ki se pri analizi na kakršen koli način srečuje z računalniki⁷ – vezano je torej na metodološke pristope znotraj jezikoslovja. Računalniško jezikoslovje je v tem pomenu zbirni pojem za različne jezikoslovne aktivnosti – v celoti zajema npr. tako računalniško leksikografijo, kakor je predstavljena v nadaljevanju, kot tudi korpusno jezikoslovje (ne glede na to, ali slednje razumemo kot metodo ali ločeno raziskovalno področje).

Poskus zamejitve interesov računalniškega jezikoslovja s strani jezikoslovca podaja Karlgren, ko na začetku devetdesetih piše takole:

»Po dvanajstih uglednih mednarodnih konferencah v organizaciji Mednarodnega odbora za računalniško jezikoslovje [International Committee for Computational Linguistics] in po vse večjem številu publikacij in lokalnih srečanj, posvečenih tej temi, bi lahko sklepali, da do zdaj vsi, ki nam je kakor koli do zadeve, vemo, za kaj v računalniškem jezikoslovju gre. Temu ni tako. [...] Brez dvoma gre pri računalniškem jezikoslovju za računanje in jezikoslovje, s poudarkom na 'in'. Temeljna koncepta sta računanje, ne računalnik ter jezikoslovje, ne procesiranje jezika.« (Karlgrén 1990: 97, prev. Š. Arhar).

Na drugi strani služi zveza (prav nasprotno) za poimenovanje raziskovalne smeri, ki se ukvarja predvsem z »računalniškimi sistemi za razumevanje in generiranje naravnih jezikov« (Ooi 1998: 24). V uvodu v *The Oxford Handbook of Computational Linguistics* lahko preberemo, da je izraz nastal v jezikovnotehnoloških krogih, ko je postalo očitno, da se interes financerjev od strojnega prevajanja obrača k splošnejšim k avtomatski obdelavi jezika usmerjenim raziskavam (Mitkov (ur.) 2003: xvii); v nadaljevanju razvoja discipline pa je kot izraz, ki počasi začne zamenjevati računalniško jezikoslovje, opredeljena zveza »(statistična) **obdelava naravnega jezika**«. V številnih virih najdemo poimenovanji kot zamenljivi, npr. v *International Encyclopedia of Linguistics* (Frawley (ur.) 2003).

Na tem mestu bomo s pojmom **računalniško jezikoslovje** poimenovali raziskovalno smer, ki se ukvarja z (I) raziskovanjem jezika s pomočjo statističnih metod obdelave jezikovnega gradiva ter (II) razvojem računalniških orodij, jezikovnih virov ipd., kadar slednje izvira iz jezikoslovnih potreb oz. je usmerjeno k jezikoslovnim ciljem.

Raziskovalno področje, ki izhaja z računalniške strani in jezik ter jezikoslovne ugotovitve uporablja predvsem za avtomatsko ali polavtomatsko pridobivanje različnih tipov podatkov, potrebnih za razvoj računalniških aplikacij, bomo na tem mestu imenovali **obdelava naravnega jezika**.

V pričujoči raziskavi predstavlja obdelava naravnega jezika – predvsem kar se tiče namena raziskave ter uporabljenih metod, dopolnitev korpusnojezikoslovnemu raziskovalnemu izhodišču. Kar obe raziskovalni smeri družijo in vedno bolj zbližuje⁸, je (I) vezanost na razvoj novih tehnoloških možnosti za obdelavo naravnega jezika,

⁷ Slej ko prej raziskovanje jezika brez takšnega ali drugačnega stika z računalnikom s stališča metodološke aktualnosti najbrž ne bo več smotrno, zato takšno poimenovanje na dolgi rok najbrž nima perspektive. Trenutno avtorji še vedno postavljajo »nove« jezikoslovne pristope v nasprotje oz. primerjavo s »tradicionalnimi« po principu računalniško vs. predračunalniško oz. neračunalniško.

⁸ Sicer je največji očitek (korpusnega) jezikoslovja obdelavi naravnih jezikov, da jezika ne raziskuje, ampak zgolj matematično obdeluje (Teubert in Krishnamurty 2007: 5–6), kar posledično (lahko) vodi v neupoštevanje številnih ugotovitev, ki v jezikoslovju že dolgo veljajo za dejstvo (npr. nearbitrarnost niza besed v stavku, razlike med jezikovnimi žanri itd.).

kar pomeni podobno metodo pridobivanja podatkov iz jezikovnih virov ter (II) naravnost k uporabnosti – kot bo vidno v nadaljevanju korpusno jezikoslovje stremi k produkciji nove generacije jezikovnih priročnikov, obdelava naravnega jezika pa k razvoju jezikovnih tehnologij.

Poleg v prejšnjem poglavju omenjenih dveh pristopov znotraj jezikoslovja prinaša obdelava naravnega jezika tretji pogled na korpus. Ti predstavljajo enega od možnih virov pridobivanja informacij o jeziku, kar pomeni razlike predvsem v dveh točkah: (I) prioretiziranje kvantitativne analize – za razliko od korpusnih pristopov v jezikoslovju, kjer sta kvantitativni ter kvalitativni del analize obravnavana kot enakovredno ključna, stremi obdelava naravnega jezika k čim večji avtomatizaciji procesiranja besedilnih podatkov, ter (II) odnos do sestave jezikovnega vira – za razliko od korpusnega jezikoslovja, kjer velik poudarek leži na snovanju in gradnji korpusov, obdelava naravnega jezika uporablja širši spekter virov za pridobivanje podatkov (strojnوبرljivi slovarji, internet, prevodni spomini, ontologije raznega tipa itd.). Načrtna gradnja korpusov znotraj obdelave naravnega jezika je redka – če nam korpus pomeni uravnoteženo, notranje strukturirano besedilno zbirko, ki je grajena kot reprezentativni vzorec določenega jezikovnega žanra.

O vlogi interpretacije pri korpusnojezikoslovni ter računalniškojezikoslovni obravnavi jezika (avtor uporablja izraz računalniško jezikoslovje za to, kar na tem mestu imenujemo obdelava naravnega jezika) ter pomenu sestave jezikovnega vira piše Teubert:

»Korpusno jezikoslovje širi naše jezikovno znanje, s tem da kombinira tri postopke: (proceduralno) identifikacijo jezikovnih podatkov v korpusu na podlagi določitve kategorij, korelacijo jezikovnih podatkov s pomočjo statističnih metod in na koncu (intelektualno) interpretacijo rezultatov. Prva dva koraka naj bi bila izvedena kolikor je mogoče avtomatsko; tretji korak predpostavlja **namernost**. Vsaka interpretacija je dejanje in je ravno zato ni mogoče algoritmizirati. V tem pojmovanju leži bistvena razlika med korpusnim ter računalniškim jezikoslovjem, ki jezik razume **proceduralno**.« (Teubert 2000 v Gorjanc in Krek 2005: 108, poudarki Š. Arhar)

Kot bo razvidno iz nadaljevanja, je središče interesa pričujočega doktorskega dela pridobivanje leksikalnih podatkov iz korpusa, ne pa tudi njihova interpretacija – v kolikor se slednja pojavlja, je usmerjena k evalvaciji uporabljene metode pridobivanja podatkov, ne pa k podatkom samim. S tega stališča je raziskava bliže obdelavi naravnega jezika. Na drugi strani na ravni organizacije leksikalnih podatkov izhajamo iz korpusnojezikoslovnih izhodišč: interes raziskave je podatke pripraviti na način, ki v naslednjem koraku predvideno jezikoslovno interpretacijo omogoča oz. olajšuje. Ker je primarni namen luščenja leksikalnih podatkov dopolnjevanje leksikalnih zbirk za slovenščino, te pa so, kot bo razloženo v poglavju III-1, lahko namenjene tako za razvoj jezikovnih tehnologij kot za pripravo jezikovnoopisnih virov za človeškega uporabnika, je dvojnost raziskovalnega izhodišča toliko bolj očitna.

3 Korpusi in leksikografija

Nov pogled, ki ga obravnava jezika prinaša korpusno jezikoslovje, izvira predvsem iz ugotovitev, ki jih je s sabo prinesel projekt COBUILD (*Collins Birmingham University International Language Database*), tj. gradnja ter analiza korpusa *Bank of English* za potrebe izdelave enojezičnega angleškega slovarja za neangleške govorce (*Collins COBUILD English Language Dictionary*, prva izdaja 1987). S stališča leksikografije danes priznana revolucionaren projekt imamo lahko za temelj uporabne veje korpusnega jezikoslovja (več o projektu denimo v Sinclair 1987).

Statistična obdelava velike količine jezikovnega materiala se odraža v identifikaciji **sopojavitvenih odnosov** med besedami, od katerih sta za pričujočo raziskavo zanimivi predvsem kolokacija ter koligacija⁹ (Sinclair 1998: 12–13).

Kolokacija, kakor jo razumemo na tem mestu¹⁰, predstavlja odnos med dvema besedama, jedrom ter kolokatorjem, za kateri je na osnovi izbrane statistične metode ugotovljena sopojavitvena povezava.¹¹ Za besedi ni nujno, da si v besedilih neposredno sledita, prav tako ni v obravnavo zajet skladenjski nivo. Primer: pridevnik *oranžen* se sopojavlja s kolokatorji *barva, abonma, odtenek, cvet, ton, svetloba, revolucija*.¹²

Koligacija predstavlja kolokaciji primerljiv odnos na ravni dveh slovničnih kategorij (npr. samostalniki se sopojavljajo s predlogom na levi) ali med jedrno besedo ter kategorijo (pridevnik *oranžen* se sopojavlja s samostalniki na desni).¹³

Že leta 1987 Sinclair navaja naslednje ugotovitve, izvirajoče iz leksikografske obravnave korpusnih primerov: (I) v večini primerov obravnave korpusnih pojavitev se v jezikovni strukturi izkazujejo določeni vzorci, (II) izkazuje se, da vzorec ne prinaša samo tipične skladnje, ampak tudi nabor tipičnih leksikalnih kolokatorjev, (III) pomena ne prinašajo posamezne besede, postavljene ena za drugo po določenih pravilih, pomen se tvori s simultano izbiro ter vzpostavitvijo daljših **leksikalnih enot**¹⁴ (Sinclair 1987 v Sinclair 1996: 160).

Predvsem pri obdelavi naravnega jezika je bilo v osemdesetih in tudi še devetdesetih letih prejšnjega stoletja osredotočanje na posamezno besedo kot nosilko pomena oz. pomenov pogosta praksa, saj je določitev besedne meje kot meje leksikalne enote računalniško (relativno) enostaven postopek. Takšna obravnava jezika je v neskladju z dvema dejstvoma: (I) mnogo pomenov zahteva za realizacijo v kontekstu prisotnost več kot ene

⁹ Ostala dva odnosa sta še »pomenska preferenca« (ang. *semantic preference*), ki predstavlja tendenco sopojavljanja s pomensko sorodnimi kolokatorji (glagol *jesti* se npr. sopojavlja s samostalniki, označujočimi vrste hrane), ter »pomenska prozodija« (ang. *semantic prosody*), ki predstavlja opredelitev pomena celote s stališča pozitivnih oz. negativnih konotacij (npr. glagol *povzročiti* se v slovenščini sopojavlja s samostalniki *škoda, zmeda, nesreča, razdejanje, razburjenje, preplah, panika, katastrofa* ...).

¹⁰ Za primerjavo različnih definicij kolokacije glej npr. McEnery et al. 2006: 82. O različnih vrstah kolokacij piše Fontenelle 1997: 16–23.

¹¹ Statistične metode pridobivanja kolokatorjev so predstavljene npr. v Oakes 1998, Manning in Schütze 2003.

¹² Kolokatorji so pridobljeni iz korpusa FidaPLUS z uporabo statistike LL za mesto na desni od jedra.

¹³ Nekaterim avtorjem koligacija (oz. koligacijska preferenca) pomeni tudi tendenco jedrne besede, da se npr. pojavlja na določenih mestih stavka ali se v kateri od oblik pojavlja znatno pogostejše kot v drugih (se pojavlja npr. pretežno v množini itd.). Več o tem v Atkins in Rundell 2008: 304–307.

¹⁴ Primeri leksikalnih enot na tem mestu so: *set fire to, set on fire, set eyes on, set free*.

besede ter (II) vzorci sopojavljanja besed (ki je v jeziku precej močnejše, kot to običajno izkazuje predkorporusni jezikoslovni opis) so neposredno povezani s pomenom (Sinclair 1998).

Navedene ugotovitve tvorijo izhodišče pričujoče raziskave v dveh točkah: (I) na eni strani nas zanima premik od obravnave posamezne besede kot nosilke (potencialnega) pomena k obravnavi večbesednih leksikalnih enot ter na drugi (II) jezikovni vzorec kot kombinacija koligacijskih ter kolokacijskih informacij, saj ponuja združenje tradicionalno ločenega pogleda na slovar ter slovnico. V literaturi najdemo za ta predmet jezikoslovnega interesa poimenovanje **leksikogramatika**.¹⁵

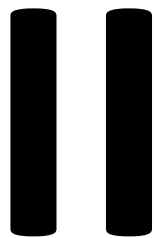
Navedene ugotovitve seveda ne prinašajo sprememb samo na področje leksikografije, ampak tudi na druga področja, predvsem se o njih veliko piše v povezavi s prevodoslovjem, besediloslovjem (predvsem v zvezi s primerjavo različnih jezikovnih žanrov), poučevanjem jezika (v angleškogovorečem prostoru intenzivno na področju poučevanja angleščine kot drugega oz. tujega jezika) itd. V pričujočem delu skušamo izbrane segmente novega pogleda na leksikografijo prenesti še na področje (slovenske) obdelave naravnega jezika, in sicer v segment, ki je z leksikografijo tesno povezan, tj. gradnja leksikalnih podatkovnih zbirk.

Raziskovalno področje, ki na različne načine združuje leksikografijo ter nove tehnologije, avtorji pogosto imenujejo **računalniška leksikografija**. Patrick Hanks denimo opredeljuje kot glavna interesa računalniške leksikografije: (I) izrabo človeku namenjenih slovarjev za računalniške namene ter (II) uporabo računalniških metod za izdelavo novih slovarjev (Hanks 2003: 49). Precej splošni opredelitvi pridajamo še Ooijevo definicijo računalniške leksikografije kot raziskovalnega področja, v grobem definiranega s tremi cilji: (I) razvoj postopkov za avtomatizacijo leksikografskega dela (različne računalniške aplikacije za urejanje, izmenjavo itd. leksikalnih podatkov), (II) izdelava leksikalnih zbirk za strojno rabo ter (III) izdelava leksikalnih zbirk za človeško rabo (Ooi 1998: 30).¹⁶

Kot rečeno, se k razlikam med leksikalnimi zbirkami za strojno ter človeško rabo, kot jih loči Ooi, vračamo v poglavju III-1, pred tem pa sledi poglavje, posvečeno označevanju korpusnih virov ter luščenju podatkov iz njih.

¹⁵ Nerazdružljivost slovarja in slovnice je bila izpostavljena tudi v slovenskem jezikoslovju: »Slovar in slovnica sta se potrdila kot metajezikovna fikcija dejansko neločljive celote, ki ji pravimo jezik.« (Vidovič Muha 2006: 39) Izčrpen pregled teme v praksi ponuja Hallidayjeva funkcijska slovnica (Halliday 2004).

¹⁶ Ooi navaja številne avtorje, ki jim računalniška leksikografija pomeni predvsem avtomatsko pridobivanje leksikalnih podatkov iz strojnoblerljivih slovarjev (ta metoda je bila izredno popularna na prehodu v devetdeseta leta prejšnjega stoletja – glej npr. Boguraev in Briscoe 1989; nekoliko novejši in metodološko izvirnejši primer je denimo Fontenelle 1997), vendar za razliko od Hanksa že navaja tudi ugotovitev, da so za gradnjo leksikalne zbirke, namenjene strojni rabi, podatki iz slovarjev, narejenih za človeka, le delno uporabni (Ooi 1998: 31).



OZNAČEVANJE KORPUSNIH BESEDIL TER LUŠČENJE LEKSIKALNIH PODATKOV

1 Označevanje korpusnih besedil

1.1 Popolni korpusni pristop

Kot je bilo nakazano v poglavju I-1, jezikoslovni pristop, ki postavlja hipoteze na osnovi besedilnega korpusa, jezikoslovčevu intuicijo o jeziku (na mestih, kjer je to mogoče) nadomešča z analizo avtentičnih jezikovnih podatkov. V raziskovalni praksi, imenovani **popolni korpusni pristop**, se ta temeljna premisa odraža v odločitvi za delo z golimi, jezikoslovno neoznačenimi korpusnimi besedili, saj se ima vsakršno pripisovanje kategorij besedam v korpusu za apriorno interpretacijo podatkov.

Korpus pri tem pristopu ni obravnavan le kot vir primerov jezikovne rabe, ampak kot nerazdružljiva celota jezikovnih podatkov. Teoretične jezikoslovne ugotovitve izvirajo neposredno iz korpusa, ki ni na nikakršen način prilagojen jezikoslovnim kategorijam nekorpusnega izvora. Nove kategorije temeljijo na pogostnosti v korpusu obstoječih vzorcev jezika (pri čemer je potencialno relevantna tudi informacija o neobstoju določenega vzorca) (Tognini-Bonelli 2001 v Teubert in Krishnamurty 2007: 74).

Na drugi strani se je s širitvijo korpusnega jezikoslovja na področja morfološko bogatejših jezikov jezikoslovno označevanje korpusov – vsaj npr. lematizacija, tj. pripis osnovne oblike besedam – izkazalo za potrebno. Obenem so nova spoznanja o možnostih obdelave označenih jezikovnih podatkov botrovala razvoju programske opreme t. i. drugega korpusnojezikoslovnega vala: razvijalci programov sledijo specifičnim raziskovalnim potrebam, programi so ciljno usmerjeni v pridobivanje, organizacijo in prikaz samo tiste vrste podatkov, ki jih uporabnik za določeno nalogo potrebuje.¹⁷

František Čermák npr. v razpravi, prvič objavljeni leta 1995, že piše o **zunanjih** ter **notranjih korpusnih podatkih**. O zunanjih podatkih govorimo pred vključitvijo v korpus (gre torej za neoznačena pisna ter transkribirana govorjena besedila), po vključitvi ter obdelavi besedil pa govorimo o notranjih podatkih:

»Tako dostopni in strojno berljivi notranji podatki v samem računalniku so takšne vrste in lastnosti, kakršne tvorci korpusa glede na zamišljeni cilj in uporabo dodajo. Kakorkoli je to tudi mogoče, praktično noben korpus danes ne daje na razpolago samo podatkov v obliki preprostih linearnih besednih verig. [...] Stopenjsko tovrstnega označevanja je lahko toliko, kolikor velika je potreba in kolikor jih je mogoče računalniško (programsko) uspešno vključiti in uveljaviti [...]« (Čermák 1995 v Gorjanc in Krek 2005: 148)

¹⁷ Tipičen predstavnik tovrstnega programa je denimo program *Besedne skice* oz. *Word Sketches* <<http://www.sketchengine.co.uk/>>, ki na podlagi oblikoskladenjsko označenega korpusa ter nabora slovnčnih pravil prikazuje različne vrste kolokacijskih podatkov za leksikografske potrebe (Krek in Kilgariff 2006).

Pripisovanje oznak v korpusu dejansko pomeni interpretativni poseg v jezikovno realnost, vendar po drugi strani omogoča izrabo označenih podatkov na ravni združujočih kategorij.¹⁸ Pričujoča raziskava temelji na stališču, da kvalitetna uporaba korpusa ne pomeni vztrajanja pri obdelavi neoznačenih besedil, ampak uzaveščenje (potencialno) šibkih mest označenosti uporabljenega vira s strani korpusnega uporabnika, s strani korpusnega graditelja pa evalvacija ter stremljenje k izboljšavam metod označevanja – zlasti v primeru avtomatskega označevanja širše uporabljenih korpusnih virov, k čemur se vračamo v II-1.2.4.

1.2 Oznake korpusa FidaPLUS

V pričujočem delu se osredotočamo na oznake trenutno aktualnega referenčnega korpusa FidaPLUS, ki je (kot bo vidno poglavju IV-2.1.1) vir za pripravo v doktorski raziskavi uporabljenega podkorpusa.¹⁹ Ostali projekti označevanja slovenskih korpusnih besedil na tem mestu ostajajo ob strani, tako npr. vprašanja skladišnega ter semantičnega kot vseh drugih oblik označevanja slovenščine.²⁰

1.2.1 Leme

Lematizacija je postopek pripisovanja osnovne oblike besedam v korpusnem besedilu. Lematizator, uporabljen za lematizacijo korpusa FidaPLUS, je razvit na podjetju Amebis, d. o. o., Kamnik.²¹ Program, ki temelji na leksikonu besednih oblik, je za slovenščino dragocen zaradi možnosti tako (I) razdvoumljanja besednih oblik v primeru obstoja več možnih lem kot tudi (II) tvorbe v leksikonu neobstoječih lem na osnovi besednih končnic.²²

V primeru neobstoja v korpusnem besedilu najdene besedne oblike v leksikonu sta predvidena dva koraka: (I) vključitev upoštevanja tipičnih odklonov od knjižne norme v sodobnem pisnem jeziku: oblika *stricom* se lematizira v *stric*, *nevem* v *vedeti* itd., (II) avtomatska tvorba leme na osnovi besedne končnice, npr. oblika *Pomurci* se lematizira v *Pomurec* ter *Goodyearju* v *Goodyear* (o problemih tovrstne lematizacije v Arhar in Gorjanc 2007: 102). V primeru, da leme ni mogoče avtomatsko uganiti, ostane besedna oblika v korpusu nelematizirana (in je kot taka označena), njen obstoj pa je zabeležen v ločeno datoteko, ki je osnova nadgradnje oblikoslovnega leksikona.²³

Pogostejši od primerov neobstoja leme v leksikonu so primeri, ko je za eno obliko možnih več različnih lem. Iskanje prave možnosti poteka v več korakih: (I) izločitev lem, ki so za dano obliko na osnovi slovničnih pravil najmanj verjetne (npr. pri primerih, ki sredi stavka izkazujejo veliko začetnico, so ohranjene le lastnoimenske leme itd.), (II) izločitev manj verjetnih lem glede na rezultate avtomatske stavčnočlenske analize besedila – slednja poteka s stavčnim razčlenjevalnikom, prav tako razvitim na omenjenem podjetju, ter (III) izbor najverjetnejše leme z delnim upoštevanjem skladišnega konteksta.

¹⁸ Sistematičen pregled argumentov za in proti označevanju korpusov prinaša npr. McEnery et al. 2006: 29–32. Odličen vir o označevanju je tudi Garside et al 1997.

¹⁹ Korpus FidaPLUS je na voljo na internetni strani <<http://www.fidaplus.net>>.

²⁰ O skladiščnem označevanju denimo Erjavec in Ledinek 2006, Ledinek 2007. O pripravi na semantično označevanje Fišer 2005, Fišer in Erjavec 2008. O označevanju diskurza npr. Verdonik 2006; 2008, Verdonik et al. 2008. Označevanje govornih zbirk Žganec Gros et al. 2000a; 2000b. Gradnja govornega korpusa Zemljarič Miklavčič 2006. Gradnja korpusa usvajanja slovenščine Stritar 2006a, 2006b. O označevanju korpusa s stališča izbire ustreznega formata zapisa Erjavec 2003. Seznam literature še zdaleč ni popoln.

²¹ <<http://www.amebis.si>>.

²² Drugi poskusi lematizacije slovenskih besedil so opisani npr. v Džeroski in Erjavec 2000, Mladenić 2002.

²³ K temi se vračamo v poglavju V-5, ki prinaša nabor nelematiziranih oblik, nastopajočih v besednih nizih z v raziskavi obravnavanimi primeri.

Primer lematiziranega besedila prinaša Tabela 1 v naslednjem poglavju: nazoren primer razdvoumljanja lem je denimo pridevniška oblika *belim*, za katero je iz začetnega nabora možnih lem *bel*, *beliti*, *Belo* uspešno izbrana ustrezna *bel*.

1.2.2 Oblikoskladenjske oznake

Sistem oblikoskladenjskega²⁴ označevanja, ki je bil uporabljen za označevanje korpusa FidaPLUS, je rezultat projekta Multext-East.²⁵ Prva verzija nabora oznak Multext-East je nastala med letoma 1995 in 1997, trenutno je na voljo v tretji različici, ki prinaša tabele oblikoskladenjskih oznak in pojasnila o sistemu njihove rabe za deset evropskih jezikov, med njimi tudi slovenskega (Erjavec 2004).²⁶

Na osnovnem nivoju sistem predvideva besednovrstno uvrstitev obravnavane besedne oblike, na drugem nivoju pa se označuje njene nadaljnje lastnosti glede na slovanske ter slovnične kategorije, ki jih posamezni besedni vrsti (lahko avtomatsko) pripišemo – pridevniku v opisanem sistemu npr. vrsto, stopnjo, spol, število, sklon, določnost ter živost. V besedilu pripisana oznaka ima obliko črkovnega niza, v katerem vsaka od črk predstavlja vrednost ene od kategorij (oznaka *Soser* npr. označuje samostalnik, občno ime, srednji spol, ednino, roditeljski).

Tudi pripisovanje oblikoskladenjskih oznak je za korpus FidaPLUS v celoti izvedeno z označevalnikom podjetja Amebis; v veliki meri temelji na podatkih, zajetih v leksikalni zbirki ASES (glej poglavje III-2). V besedilu najdena oblika je primerjana s podatki iz leksikalne zbirke, posledično ji je v prvi vrsti pripisana besedna vrsta, na podlagi tega podatka pa še vrsta kategorialnih vrednosti, ki so posamezni besedni vrsti pripisljive. Podobno kot leme so tudi oblikoskladenjske oznake (oz. oznake MSD, kot se v tem kontekstu pogosto imenujejo) v korpusu FidaPLUS razdvoumljene.

Označeni korpusni dokument je v Konkordančniku ASP32, s katerim korpus FidaPLUS uporabljamo na internetu, dostopen s klikom iz konkordančnega niza. Prav tako so v konkordančniku dostopni podatki o sistemu oblikoskladenjskega označevanja, ki omogoča lažje branje oblikoskladenjskih oznak (o uporabi konkordančnika Arhar 2006). Spodnji primer prinaša primer označenosti stavka *Pogosti avstralski pajek z belim zadkom je veljal za povzročitelja gnitja mesa*. Ohranjena je barvna shema, ki v konkordančniku omogoča večjo preglednost zapisa: vrste oznak (lemma, msd itd.) so zapisane z modro barvo, vrednosti posamezne oznake zeleno, oznake formata xml pa nakazuje rdeča barva.

²⁴ Oblikoskladenjska raven označevanja, kot pove ime, prinaša oznake, opredeljujoče slovanske ter slovnične lastnosti, ki jih izpričuje posamezna besedna oblika v besedilu. V določeni meri avtomatsko označevanje glede pripisanih kategorij sledi ugotovitvam jezikoslovja, pogosto pa to v celoti ni mogoče. Označevanje namreč poteka brez upoštevanja pomena, omejeno je tudi upoštevanje skladenjskih informacij konteksta označevane besede. Oblikoskladenjsko označevanje je omejeno na pripisovanje informacij, ki so avtomatsko ugotovljive iz besedne oblike same, čemur je prilagojen nabor označevalnih kategorij. Oznake, navedene v Prilogah 1 in 2, prim. denimo z opredelitvijo prepleta kategorialnih lastnosti besed ter njihove skladenjske vloge v besedilu v Vidovič Muha 2000: 30–38.

²⁵ <<http://nl.ijs.si/ME/V3/>> – na internetni strani je na voljo vsa dokumentacija, vključno s tabelami oznak in povezavami na relevantno literaturo. Tabele oznak so za lažje razumevanje nadaljevanja poglavja navedene tudi kot Priloga 2.

²⁶ V zvezi z oblikoskladenjskim označevanjem slovenščine je potrebno omeniti še: (I) nabor oznak Inštituta za slovenski jezik Frana Ramovša ZRC SAZU, ki je nastal za potrebe označevanja korpusa Beseda (Jakopin in Bizjak 1997; Lönneker in Jakopin 2004; Jakopin in Bizjak 2008), (II) nabor oznak LC-STAR, ki se uporablja pri gradnji jezikovnih virov za simultano prevajanje govora (Verdonik in Rojc 2004; Verdonik, Rojc in Kačič 2004), (III) razvoj oblikoskladenjskega označevalnika s konkordančnikom SLON na Kemijskem inštitutu Ljubljana (Zupan in Čeh 2008) Nabor oznak Multext-East je bil uporabljen tudi za statistično označevanje slovenščine (Erjavec in Sarossy 2006). Pregled teme za angleški prostor ponuja npr. Mitkov (ur.) 2003.

```
<p ID="F0018254.2216"><s ID="F0018254.2216.1">

<w lemma="pogost" msd="Pkomeid" lemmas="pogost" msds="Pkomeid" lemmass="pogostiti
pogost"msdss="Gppstexnxxxxxd, Gpvsdexnxxxxxd, Pkomeid, Pkometdxn, Pkozed, Pkozem, Pkommi,
Pkozdi, Pkosdi, Pkozdt, Pkosdt">Pogosti</w>

<w lemma="avstralski" msd="Pvomeid" lemmas="avstralski" msds="Pvomeid"
lemmass="avstralski"msdss="Pvomeid, Pvometdxn, Pvozed, Pvozem, Pvommi, Pvozdi, Pvosdi, Pvozdt,
Pvosdt">avstralski</w>

<w lemma="pajek" msd="Somei" lemmas="pajek" msds="Somei" lemmass="pajek" msdss="Somei,
Sometxxn">pajek</w>

<w lemma="z" msd="Dpeo" lemmas="z" msds="Dpeo" lemmass="z" msdss="Dpeo, Dper">z</w>

<w lemma="bel" msd="Pkomeo" lemmas="bel" msds="Pkomeo" lemmass="bel beliti Belo" msdss="Pkomeo,
Pkoseo, Pkommd, Pkozmd, Pkosmd, Gppspexnxxxxxn, Slseo">belim</w>

<w lemma="zadek" msd="Someo" lemmas="zadek" msds="Someo" lemmass="zadek" msdss="Someo,
Sommd">zadkom</w>

<w lemma="biti" msd="Gvpstexn" lemmas="biti" msds="Gvpstexn" lemmass="biti jesti on"
msdss="Gvpstexn, Gppstexnxxxxxn, Zotzerxxdxs">je</w>

<w lemma="veljati" msd="Gpdrxemtxxxxxn" lemmas="veljati" msds="Gpdrxemtxxxxxn" lemmass="veljati"
msdss="Gpdrxemtxxxxxn">veljal</w>

<w lemma="za" msd="Dpet" lemmas="za" msds="Dpet" lemmass="za" msdss="Dpet, Dper, Dpeo">za</w>

<w lemma="povzročitelj" msd="Sometxxd" lemmas="povzročitelj" msds="Sometxxd"
lemmass="povzročitelj" msdss="Sometxxd, Somer, Somdi, Somdt">povzročitelja</w>

<w lemma="gnitje" msd="Soser" lemmas="gnitje" msds="Soser" lemmass="gnitje" msdss="Soser, Sosmi,
Sosmt">gnitja</w>

<w lemma="meso" msd="Soser" lemmas="meso" msds="Soser" lemmass="meso"
msdss="Soser">mesa</w>

<c type="PUN">.</c></s></p>
```

Tabela 1: Primer označenega besedila v korpusu FidaPLUS.

Razdvoumljenost oznak se kaže v trostopenjskosti označevanja: oznaki *lemmass* ter *msdss* prinašata celoten nabor možnih oznak, *lemmas* ter *msds* vmesno stopnjo razdvoumljanja, *lemma* ter *msd* pa končno stanje po razdvoumljanju (več v Arhar in Gorjanc 2007: 101–103).

1.2.3 Nabor oblikoskladenjskih oznak JOS

Pomemben korak k izboljšavi kvalitete oblikoskladenjskega označevanja slovenskih besedil je bil izveden v sklopu projektu JOS, Jezikoslovno označevanje slovenščine²⁷, katerega cilj je priprava prosto dostopnega večnivojsko (oblikoskladenjsko, skladenjsko ter pomensko) označenega milijonskega korpusa:

²⁷ <<http://nl.ijs.si/jos/>>.

»Projekt JOS skuša zapolniti vrzel pri jezikovnih virih za slovenščino z izdelavo standardiziranih prosto dostopnih označenih korpusov, skupaj z revidiranim naborom oblikoskladenjskih specifikacij. V članku poročamo o prvih rezultatih: korpusu "jos100k", ki predstavlja zlati standard za označevanje, in korpusu "jos1M", na katerem trenutno poteka delo. Oba korpusa vsebujeta vzorčene odstavke iz korpusa FidaPLUS in sta označena z razdvoumljenimi in ročno preverjenimi lemmami in oblikoskladenjskimi oznakami.« (Erjavec in Krek 2008a: 49)

Za pričujoče delo je zanimiva predvsem omenjena revizija in nadgradnja v sklopu projekta Multext-East pripravljenega sistema oblikoskladenjskih oznak. Revizijo ter nadgradnjo so vodila naslednja načela²⁸: (I) načelo strnjivosti ter berljivosti oznak je vodilo k odpravi prostih mest v označevalnih tabelah, (II) načelo pripisljivosti je vodilo k odstranitvi kategorij oz. vrednosti, za katere se je izkazalo, da avtomatsko niso pripisljive na zadovoljivi ravni (zaradi potrebe po upoštevanju besednega konteksta²⁹, zaradi dvoumnosti itd.), (III) načelo uravnovešenosti pa je vodilo k preoblikovanju označevalne tabele na mestih, kjer so se oznake kazale za preveč ali premalo razpršene oz. preveč ali premalo specifične (Arhar in Ledinek 2008: 55).

Ker je revizija sistema oznak dokumentirana v navedenih dveh prispevkih, celotne oblikoskladenjske specifikacije v izčrpnih oblikah pa so na voljo na internetni strani <<http://nl.ijs.si/jos/josMSD-sl.html>>, se na tem mestu ne posvečamo vprašanju sprememb označevalnih kategorij glede na izvorni sistem Multext-East. Ker pa so v doktorski raziskavi že uporabljene nove oblikoskladenjske oznake (glej poglavje IV-2.1.1), ki jih v nadaljevanju pričujočega dela ne razvezujemo oz. razlagamo, so označevalne tabele sistema oznak JOS za lažje razumevanje na voljo tudi kot Priloga 1.

1.2.4 Uspešnost oblikoskladenjskega označevanja slovenščine

Avtomatsko pripisovanje oznak, zlasti če so slednje kompleksnega tipa, kar za oblikoskladenjske oznake JOS lahko trdimo, nikoli ne more biti v celoti uspešno.³⁰ Kot kaže ocena označevanja slovenščine, ki jo podajata Erjavec in Krek 2008, je bila za označevalnik, ki nas na tem mestu najbolj zanima, tj. Amebisov označevalnik, ugotovljena 85,7 % uspešnost, kar pomeni, da je oblikoskladenjska oznaka, pripisana s strani avtomatskega označevalnika, v 85,7 % enaka ročno pripisani oznaki.³¹

V raziskavo, ki jo avtorja predstavljata v navedenem prispevku, je bila zajeta primerjava treh označevanj istega nabora korpusnih besedil. Korpus jos100k – približno 100.000 besed velik podkorpus korpusa FidaPLUS (o pripravi korpusnega vira Erjavec in Krek 2008: 50) – je bil označen z označevalnikom podjetja Amebis ter s

²⁸ V večji meri povzeta po priporočilih za oblikoskladenjsko označevanje korpusov *EAGLES* <<http://www.ilc.cnr.it/EAGLES/annotate/annotate.html>>.

²⁹ Problemi se kažejo na mestih, kjer je za določitev kategorije oz. vrednosti potrebno (I) upoštevanje besedne okolice na skladijski ravni, ker je določitev v teh primerih vezana na niz pojavnic in ne le na posamezno pojavnico, (II) razdvoumljanje na osnovi semantičnih informacij, ki je mogoče le na ravni ročnega označevanja – ali pa še to ne (za natančnejšo razlago s primeri glej Arhar in Ledinek 2008). K problemu označevanja skladijsko ter semantično odvisnih jezikovnih pojavnosti v sklopu avtomatskega oblikoskladenjskega označevanja se vračamo v poglavju V-4.2 v zvezi z označevanjem besedne vrste členki v slovenščini.

³⁰ Atkins in Rundell trdita, da označevalne napake za potrebe korpusne leksikografije niso relevantne, saj se slednja osredotoča na tipično v jeziku. Jezikovne regularnosti se v dovolj velikem, solidno označenem korpusu zanesljivo kažejo, označevalne napake pa se izgubijo v množici podatkov (Atkins in Rundell 2008: 91–92). Čeprav se z avtorjema načeloma strinjamo, je potrebno pridati, da je v primerjavi z označevanjem angleščine označevanje slovenščine, predvsem na oblikoskladenjskem nivoju, zahtevnejša naloga, katere izvedbo je v vsakem primeru smiselno neprekinjeno preverjati ter nadgrajevati.

³¹ V načrtu je natančna analiza razlik v avtomatsko ter ročno pripisanih oznakah; podatki trenutno še niso na voljo.

statističnim označevalnikom TnT, oznake pa so bile tudi ročno pregledane ter popravljene. Rezultate, ki so seveda najbolj zanimivi na mestih primerjave med ročno ter avtomatsko pripisanimi oznakami, avtorja navajata v obliki tabele (ibid.: 52):

	število besed	razlaga
1	100.003	vse besede v korpusu JOS 100k
2	86.617	oznake, ki jih je pripisal TnT, se ujemajo z ročno pripisanimi
3	85.719	oznake, ki jih je pripisal Amebisov označevalnik, se ujemajo z ročno pripisanimi
4	78.011	oznake obeh označevalnikov se ujemajo z ročno pripisanimi
5	7.708	oznake Amebisovega označevalnika se ujemajo z ročno pripisanimi, oznake TnT-ja ne
6	8.606	oznake TnT-ja se ujemajo z ročno pripisanimi, oznake Amebisovega označevalnika ne
7	3.238	oznake nobenega od označevalnikov se ne ujemajo z ročno pripisanimi, avtomatsko pripisane oznake so enake
8	2.440	oznake nobenega od označevalnikov se ne ujemajo z ročno pripisanimi, avtomatsko pripisane oznake so različne

Tabela 2: Natančnost označevanja korpusa jos100k.

Kot je vidno iz gornje tabele, so izboljšave označevanja mogoče pri točkah 5, 6, 7 ter 8. Prvima dvema od naštetih, tj. primerom, kjer eden od označevalnikov pripiše pravilno oznako, drugi ne – z izhodiščno predpostavko, da informacija o tem, katera od obeh oznak je pravilna, ni na voljo – se posveča prispevek Rupnika in drugih (2008). Avtorji testirajo nabor statističnih učnih algoritmov ter parametrov za razvoj sistema za avtomatsko identifikacijo pravilne oznake v opisanih primerih. V najboljših pogojih poročajo o 79,73 % uspešnosti (Rupnik, Grčar in Erjavec 2008: 114).

Težje rešljiv problem predstavljajo mesta, na katerih sta oba označevalnika neuspešna (7 in 8 v gornji tabeli), še bolj pa v tabeli sicer neizpostavljeni primeri, kjer prihaja do različnega označevanja pri dveh človeških označevalcih.³²

Iz povedanega sledi, da je na poti k izboljšavi kvalitete oblikoskladenjskega označevanja smotrno veliko pozornosti posvetiti poskusom izboljšave uspešnosti (vsakega od posameznih) avtomatskih označevalnikov, ki so za slovenščino na voljo. V doktorski raziskavi se osredotočamo na oblikoskladenjske oznake, pripisane z Amebisovim označevalnikom, in sicer predvsem s stališča vpliva označevalnih napak na kvaliteto luščenja podatkov iz korpusa.

2 Luščenje leksikalnih podatkov iz korpusa

Avtomatsko pridobivanje leksikalnih podatkov iz korpusnih in primerljivih virov je v literaturi imenovano **luščenje leksikalnih podatkov** (ang. *lexical data extraction*). Podatki, ki jih iz korpusov pridobivamo, so lahko

³² Priprava t. i. zlatega standarda (ang. *gold standard*) za določanje uspešnosti avtomatskega označevanja zahteva več kot enega človeškega označevalca. Oznake korpusa jos100k so bile v celoti pregledane s strani dveh, primere, pri katerih sta slednja pripisala različni oznaki, pa je pregledal še tretji označevalec (Erjavec in Krek 2008a: 52).

različnih vrst: v pričujoči raziskavi se osredotočamo na luščenje **večdelnih besednih nizov**, v katerih nastopajo izbrane enobesedne iztočnice, kakor bo natančneje predstavljeno v poglavju II-2.2.

Za luščenje podatkov se uporabljajo različni metodološki pristopi, ki jih lahko v osnovi delimo na (I) statistične, (II) jezikoslovno osnovane ter (III) hibridne. Prvi pristop temelji zgolj na upoštevanju pogostnosti in sopojavljanju podatkov v korpusnem viru, pri drugem se obenem uporabljajo tudi pravila, ki vključujejo jezikoslovne (npr. na skladnjo vezane) pogoje luščenja, tretji pristop pa združuje prva dva (Vintar 2002: 78; Vintar 2008: 100–101).

V pričujoči raziskavi uporabljena metoda je **luščenje podatkov na osnovi oblikoskladenjskih oznak**. Za slovenščino je bila metoda že preizkušena za pridobivanje terminoloških besednih zvez – usmerjena v pridobivanje specifičnega nabora leksikalnih enot je metoda temu ustrezno specializirana (glej npr. Vintar 1999; 2002; 2008; Logar in Vintar 2008; Vintar in Erjavec 2008). V nadaljevanju je na kratko prestavljen primer luščenja terminologije, kakor ga opisujeta Vintar in Erjavec (2008), sledi pa primerjalna opredelitev luščenja podatkov v sklopu pričujoče raziskave.

2.1 Luščenje terminoloških besednih zvez

Namen raziskave, ki jo opisujeta Vintar in Erjavec, je luščenje samostalniških terminoloških besednih zvez s področja informatike. Avtorja luščita besedne zveze iz specializiranega korpusa iKorpus in jih v naslednjem koraku z izbrano statistično formulo glede na primerjavo s podatki referenčnega korpusa uredita po terminološki relevantnosti (več o tem Vintar in Erjavec 2008: 67–68).

Luščenje poteka na osnovi vnaprej pripravljenega nabora skladenjskih vzorcev, ki prinašajo opredelitev sosledja besednih vrst v besedni zvezi ter njenega jedra (v tabeli podčrtano), kakor prikazuje spodnja tabela (ibid.: 67):³³

vzorec	
PRID + <u>SAM</u>	<u>SAM</u> + PRID + SAM
<u>SAM</u> + SAM	<u>SAM</u> + PRED + SAM
<u>SAM</u> + SAM + SAM	<u>SAM</u> + PRED + PRID + SAM
PRID + PRID + <u>SAM</u>	PRID + <u>SAM</u> + PRED + SAM
PRID + <u>SAM</u> + SAM	<u>SAM</u> + PRED + SAM + SAM

Tabela 3: Vzorci za luščenje samostalniških terminoloških besednih zvez.

Skladenjski vzorci torej opredeljujejo besedne zveze, ki se iz korpusa luščijo na osnovi opredelitve besedne vrste, ki jo posamezni besedni obliki v korpusu pripisuje oblikoskladenjska oznaka. Skladenjski vzorci v gornji tabeli so, kot pišeta avtorja, pridobljeni na osnovi podatkov obravnavanega specializiranega korpusa (ibid.: 67).

Po urejanju izluščenih besednih nizov glede na terminološko relevantnost, kar poteka na ravni lematiziranih besednih nizov, skušata avtorja poiskati ustrezne kanonične besednozvezne različice, *plošča CD* namesto *plošča cd* ter *digitalni fotoapararat* namesto *digitalen fotoapararat*. Pri pretvorbi izhajata iz v besedilih izpričane imenovalniške oblike (če se izluščeni lematizirani besedni niz v korpusu nikoli ne pojavi z jedro besedo v imenovalniku, pretvorba v kanonično obliko ni mogoča) (ibid.: 68). Kot bo vidno v naslednjem poglavju, je

³³ Primarni interes avtorjev je pri luščenju terminologije posvečen samostalniškim zvezam, z željo širitve metode tudi na druge polnopomenske besedne vrste (Vintar 2008: 40–41). Seznane skladenjskih vzorcev najdemo denimo v Vintar 1999: 169 ter Vintar 2008: 40.

vprašanje priprave ustrezne končne oblike izluščenih podatkov tudi v središču pozornosti doktorske raziskave: ta je pravkar predstavljeni mestoma sorodna, čeprav ima druge točke interesa in je glede načina luščenja podatkov elementarnejša.

2.2 Luščenje večdelnih besednih nizov

Namen pričujoče raziskave je dopolnjevanje leksikalne zbirke za obdelavo naravnega jezika z večbesednimi leksikalnimi enotami (več o tem sledi v poglavju III-3). Glavni vodili luščenja sta **zaporednost besed** in **pogostnost besednega niza**. Zaporednost besed pomeni, da so izluščeni besedni nizi v izvornem besedilu zaporedne enote, pri čemer tekom samega luščenja v isto skupino zajemamo tako proste kot stalne besedne zveze oz. tako neidiomatične kot idiomatične (frazeme, idiome).

V isto skupino so denimo zajeti primeri tipa *rdeča mravlja* ter tipa *rdeča bluza*, s čimer odpade ločevanje prostih besednih zvez od stalnih (tj. večbesednih leksemov; o tem npr. Vidovič Muha 2000: 26–28). V isto skupino so obenem uvrščeni primeri tipa *rdeča četrť* ter *rdeči petelin*. Z vprašanjem razmerja med navedenimi tipi besednih zvez se natančneje ukvarja Gantar, tudi v razmerju do korpusnih podatkov:

»Relativnost razmerij med eno in večbesednimi leksikalnimi enotami ter med pomensko transparentnimi in pomensko netransparentnimi SBZ se v korpusu zabrisuje, kar seveda ne pomeni, da je na podlagi korpusnih podatkov nemogoče določiti temeljni predmet leksikalne problematike, pač pa predvsem preseči skrajnostno delitev na besede in besedne zveze na eni strani ter na stalne in proste zveze na drugi.« (Gantar 2006: 158)

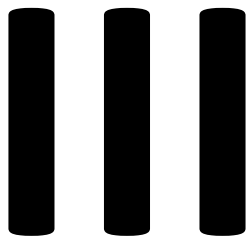
Pomenska analiza izluščenih nizov, kot rečeno v poglavju I-2, ostaja izven interesa pričujoče raziskave, ki se osredotoča predvsem na pripravo ter evalvacijo uporabljene metode; je pa leksikografska analiza podatkov predvidena kot nujni vmesni korak med luščenjem ter nadaljnjo uporabo pridobljenih podatkov.

Pri pridobivanju večdelnih nizov iz korpusa ne uporabljamo nobene od kolokacijskih statistik, ampak se opiramo le na pogostnost izluščenih nizov v korpusu. Glede na specifične potrebe nadaljnjih raziskav je seveda mogoče statistike izračunati naknadno, trenutno pa se zdi smiselno ugotoviti, kakšne podatke je možno pridobiti v primeru ugotavljanja tipičnosti rabe zgolj na osnovi visoke pogostnosti nizov v korpusnih besedilih.

Za razliko od v prejšnjem poglavju opisanega luščenja terminologije, ki izhaja iz skladišnega vzorca, želimo na tem mestu izhajati iz enobesedne iztočnice, tj. besede, ki v skladišne vzorce vstopa. Raziskava se osredotoča na obravnavo vseh skladišnih vzorcev, ki jih za obravnavan nabor iztočnic v korpusu najdemo. Na ta način želimo skladišne vzorce obravnavati v prvi vrsti kot koligacijske informacije o izbrani besedi, na osnovi vzorcev izluščeni besedni nizi pa podajajo kolokacijske informacije o obravnavani besedi (glej poglavje I-3). Skladišni vzorci so v raziskavi obravnavani ne le s stališča tipičnosti ter relevantnosti, temveč tudi obratno – interes je posvečen tudi vzorcem, ki za obravnavane iztočnice ne prinašajo za luščenje zanimivih rezultatov (pojmovanje relevantnosti izluščenih podatkov je opredeljeno v IV-1.2).

Kot rečeno je v raziskavi veliko pozornosti posvečene pripravi končne oz. prikazne oblike izluščenih podatkov, tj. oblike, v kateri je izluščeni besedni niz poslan v ročno analizo. Ker sta ročna analiza ter urejanje podatkov dolgotrajno opravilo, stremimo k temu, da so podatki pripravljeni na način, ki omogoča čim hitrejšo branje in obdelavo večbesednih enot. Raziskava se osredotoča tudi na oblikoskladišne oznake, izpričane pri obravnavanih izluščenih primerih. Dotikamo se torej še enega problema, ki je v literaturi že bil identificiran: »Žal se pri vzorcih razkrije kakovost označevanja, saj se morebitne napake pokažejo bodisi v obliki nepravilnih zvez bodisi v obliki nižjega priklica.« (Vintar 2002: 81)

Z željo omogočanja primerjave v doktorski raziskavi uporabljenih skladišnih vzorcev z naborom, predstavljenim v Tabeli 3, je večja pozornost predvidena na ravni luščenja podatkov za samostalniške iztočnice. Obenem je predvideno tudi luščenje besednih nizov s pridevniškim, prislovnim ter glagolskim izhodiščem.



LEKSIKALNA PODATKOVNA ZBIRKA ZA SLOVENŠČINO

Poglavje prinaša opredelitev leksikalne podatkovne zbirke s poudarkom na razlikah med zbirkami, namenjenimi obdelavi naravnega jezika, ter zbirkami, namenjenimi pripravi jezikovnih virov za človeškega uporabnika. Poglavje prinaša predstavitev treh primerov gradnje leksikalne zbirke za slovenščino in je sklenjeno z izpostavitvijo možnosti nadgradnje jezikovnotehnološke zbirke ASES s korpusnimi podatki.

1 Opredelitev

Zvezo **podatkovna zbirka** (ang. *database, data base*) v tem delu uporabljamo v pomenu, ki izvira iz računalniške terminologije. V svoji izvorni disciplini predstavlja natančno zamejen pojem, ki se v literaturi pogosto ločuje od podatkovne baze, podatkovnega sistema, podatkovnega skladišča ipd. Na dejstvo, da so postale računalniške podatkovne zbirke različnih vrst del raziskovalnih metod tudi drugih strok, je vezana delna determinologizacija, po kateri podatkovna zbirka (oz. pogosto sopomensko uporabljana **podatkovna baza**) pomeni v grobem: interaktivno elektronsko zbirko izbrane vrste podatkov, ki so organizirani in med seboj povezani na določen način; ali če se po definicijo zatečemo v izvirno računalništvo: »mehanizirana, večuporabniška, formalno definirana in centralno nadzorovana zbirka podatkov« (Mohorič 1995: 14).³⁴

Iz navedenega sledi, da **leksikalna podatkovna zbirka** prinaša na izbrani način formalizirane ter organizirane podatke leksikalnega tipa. Glede na vrsto podatkov, ki jih zbirka zajema, način organiziranosti, namen gradnje itd. lahko leksikalne zbirke delimo na različne načine. Na tem mestu je primarna delitev, kot je že bilo nakazano v poglavju I-3, glede na **namen uporabe** podatkov. S tega stališča ločujemo: (I) zbirke, namenjene obdelavi naravnega jezika in posledično razvoju oz. delovanju ene ali več specifičnih jezikovnih tehnologij – v nadaljevanju poglavja jih imenujemo **jezikovnotehnološke leksikalne zbirke**, ter (II) zbirke, namenjene pripravi jezikovnoopisnih virov za človeškega uporabnika – v nadaljevanju jih imenujemo **slovarske leksikalne zbirke**.³⁵

Tako ena kakor druga oblika leksikalnih podatkovnih zbirk sta si sorodni v temeljih zasnove: prinašata besedišče jezika z deloma prekrivnim naborom podatkov o slednjem. Druži ju tudi dejstvo, da sami po sebi nista končni produkt, ampak vmesna stopnja na poti do določene jezikovne tehnologije oz. jezikovnega vira slovarskega tipa.

Ker pa je končni cilj snovanja ene in druge zbirke drugačen, je bistvena razlika med njima v prvi vrsti tip podatkov, ki jih prinašata. Slovarska leksikalna zbirka se osredotoča na leksiko v tradicionalnem leksikografskem smislu: stremi k za človeka uporabnemu jezikoslovnemu opisu besedišča. Jezikovnotehnološka se na drugi strani osredotoča na za računalniško obdelavo uporabne leksikalne podatke, glavni namen je zagotoviti tehnologiji informacije, ki jih uporabnik slovarja kot govorec naravnega jezika poseduje oz. jih je sposoben inferirati.³⁶

³⁴ Omeniti je potrebno, da nekateri avtorji s pojmom podatkovna zbirka označujejo tudi besedilne korpuse, slovarje, tezavre itd. (npr. Kruyt 2003; Varantola 2003). V pričujočem delu slednje imenujemo širše **jezikovni viri**.

³⁵ Predlagana poimenovanja temeljijo na produktu, h kateremu vodi priprava vsake od vrst zbirk (jezikovne tehnologije nasproti slovarjem), pri čemer se zavedamo dejstva, da določeni tipi leksikalnih zbirk lahko ostanejo zunaj predlagane kategorizacije (v primeru, da je gradnja denimo usmerjena k obema ciljema).

³⁶ S tega stališča je relevantnost obstoja besede v zbirki enaka relevantnosti katerega koli tipa metainformacije o tej besedi.

V nadaljevanju poglavja sta najprej na kratko predstavljena dva projekta gradnje leksikalnih zbirk za slovenščino: (I) snovanje in gradnja slovarske leksikalne zbirke ter (II) gradnja slovenskega wordneta. Namen na tem mestu ni opis celotne situacije na področju gradnje leksikalnih zbirk za slovenščino, temveč zgolj natančnejša opredelitev pomenskega polja leksikalne zbirke. Izbrana primera obenem ponujata izhodišče za primerjavo z jezikovnotehnoško leksikalno zbirko ASES, natančnemu opisu katere je spričo središčnosti za pričujočo raziskavo posvečena večja pozornost.

1.3 Slovarska leksikalna zbirka za slovenščino

V našem prostoru prvi elaborirani primer snovanja slovarske leksikalne podatkovne zbirke, temelječe na podatkih iz (referenčnega) korpusa, je prispevek Vojka Gorjanca, Simona Kreka ter Polone Gantar iz leta 2005. Avtorji identificirajo potrebo po gradnji jezikovnega vira za izdelavo različnih vrst slovarskih priročnikov, ki naj bi uporabniku ponudili metodološko aktualen jezikoslovni opis sodobnega slovenskega jezika, kakršen se kaže v rabi:

»Namen leksikalne podatkovne zbirke je tako predvsem pridobiti podatke o realnem jeziku, torej aktualnem leksikalnem naboru v slovenščini, o pomenih leksikalnih enot in njihovem tipičnemu ubeseditelju. [...] Leksikalna podatkovna zbirka ima torej namen popisati stanje v slovenskem jeziku izključno na podlagi podatkov iz referenčnega korpusa slovenskega jezika. Gradnja take zbirke je neodvisna od morebitnih kasnejših slovarskih realizacij, kjer je potrebno upoštevati še npr. tip slovarja, uporabnika, velikost itd.« (Gorjanc, Krek in Gantar 2005: 5)

Leksikalna zbirka je zasnovana dosledno korpusno, kar avtorji članka prikazujejo tako s predstavitvijo metodoloških izhodišč gradnje kot tudi z opisom izvedene poskusne analize 783-ih najpogostnejših lem na črko b, kakor se pojavljajo v takrat aktualnem referenčnem korpusu FIDA:

»Izhodiščno vodilo je v leksikalni zbirki prikazati aktualno stanje slovenščine na leksikalni ravni: obstoj leksikalnih enot, njihovo dejansko obliko in pomen ter tipično ubeseditelje. Poseben poudarek velja registraciji različnih vrst besedne povezovalnosti: kolokacije, skladenjski vzorci, pomensko netransparentne zveze in idiomi.« (ibid.: 17)

V prispevku so predvideni problemi, nastajajoči ob avtomatskem razdvoumljanju besednih oznak ali zaradi korpusnega šuma, nakazane so smernice glede prikaza in obravnave večbesednih leksemov, smernice glede ugotavljanja in prikaza pomena iztočnice, obravnave zgledov rabe, razložena je metoda za pridobivanje kolokatorjev iz korpusa, izbran je tudi najustreznejši format leksikalne zbirke.

Realizacija v prispevku načrtane leksikalne zbirke je predvidena v sklopu projekta *Sporazumevanje v slovenskem jeziku*.³⁷ Leksikalna zbirka kot eden izmed ciljev projekta je opredeljena takole:

»Cilj aktivnosti [leksikalna zbirka] je baza podatkov v skladenjskih, pomenskih, frazeoloških in drugih lastnostih besedišča slovenskega jezika. Določanje standardnih postopkov za analizo korpusa s pomočjo specializirane programske opreme in standardov za izdelavo posamezne leksikalne enote v leksikalni podatkovni bazi poteka od junija 2008 do decembra 2008. Izdelava končne verzije standarda z vzorčnimi primeri za vse besedne vrste bo potekala od januarja 2009 do junija 2009. Izdelava leksikalne baze bo potekala od julija 2009 do julija 2012 po sklopih: A – K: do julija 2010, L – P: do julija 2011, R – Ž: do julija 2012. Predvidena sta dva glavna namena uporabe leksikalne baze: za potrebe leksikografije/leksikologije, za

³⁷ Projekt, katerega upravičenec je Amebis, d. o. o., Kamnik (s konzorcijskimi partnerji: Inštitut Jožef Stefan, Univerza v Ljubljani, Znanstvenoraziskovalni center SAZU ter zavod za uporabno slovenistiko Trojina), delno financira Evropska unija iz Evropskega socialnega sklada ter Ministrstvo za šolstvo in šport. Projekt se izvaja v okviru Operativnega programa razvoja človeških virov za obdobje 2007–2013. Internetna stran projekta: <<http://www.slovenscina.eu>>.

potrebe procesiranja naravnih jezikov.« (<http://www.slovenscina.eu/Vsebine/SI/Aktivnosti/LeksikalnaBaza.aspx>, dostop 6. 3. 2009).

Pri gradnji jezikovnih virov za manjše jezikovne skupnosti, kakršna je tudi slovenska, je pred začetkom projektov smiselno premisliti o načinih omogočanja širše uporabnosti vira, predvsem v smislu dostopnosti zainteresiranim raziskovalnim skupnostim ter posledično o standardiziranem formatu zapisa podatkov, ki bi omogočal njihovo enostavno izmenljivost ter uporabo. Kot je razvidno iz gornjega navedka, je leksikalna zbirka, predvidena v sklopu projekta Sporazumevanje v slovenskem jeziku, zasnovana na način, ki bo omogočal njeno čim širšo izrabo, tudi s stališča dvojnosti namena gradnje (za leksikografske potrebe ter za obdelavo naravnega jezika).

1.4 Slovenski wordnet

Poseben pristop h gradnji leksikalnih zbirk predstavljajo **ontologije**, leksikalne zbirke, katerih cilj je identifikacija ter formalizacija pojmovnih mrež. Večinoma temeljijo na organizaciji leksike glede na pomenske odnose med besedami (nad- ter podpomenskost, sopomenskost, protipomenskost, pomenska vsebovanost itd.), lahko pa podatke dopolnjujejo tudi z védenjem o svetu (ang. *world knowledge*).

Primer ontološkega jezikovnega vira v slovenskem prostoru predstavlja slovenski wordnet – sloWNet³⁸, o gradnji ter možnostih izrabe katerega piše Darja Fišer: wordnet postavlja v množico leksikalnih zbirk³⁹, ki si na različne načine »vse prizadevajo predstaviti hierarhijo jezikovno neodvisnih pojmov in uporabljajo podobne relacije za ustvarjanje povezav med posameznimi pojmi« (Fišer 2005: 18).

Temeljno organizacijsko načelo wordneta je razvrščanje leksike v sopomenske nize (t. i. sinsete), pri čemer je ključnega pomena ugotavljanje medleksemskih pomenskih odnosov, poleg sopomenskosti predvsem nad- ter podpomenskosti. Zaradi proste dostopnosti je wordnet izredno priljubljena osnova razvijanja raznovrstnih postopkov za obdelavo naravnega jezika, ima pa nekaj pomanjkljivosti, ki iz izbrane metode gradnje izvirajo.

Jezikovna neodvisnost pojmov predvideva lahko prenosljivost sistema relacij iz enega jezika (običajno je to angleščina, za katero je wordnet najbolj razvit) v drugega. V zvezi s prenosom so sicer predvideni tako problemi z leksikalnimi vrzelmi kot tudi denotacijskimi razlikami, kar se oboje izkazuje tudi pri poskusu prenosa ene od besednih mrež iz angleščine v slovenščino (ibid.: 20–23). Drugi problem, na katerega opozarja avtorica, je nekonsistentnost organiziranosti leksike v izvorni, angleški besedni mreži (ibid.: 26).

Avtorica se skuša omenjenim problemom izogniti s pridobivanjem leksikalnih ter relacijskih podatkov iz razširjenega nabora slovenskih jezikovnih virov; poleg angleškega wordneta uporablja korpus FIDA, podatke iz enojezičnega ter dvojezičnih slovarjev ter izbranih strokovnih priročnikov. Problem, ki se ob takšnem kombiniranju pojavi (poleg odmika od želene avtomatiziranosti gradnje zbirke), so razlike v jezikovnih podatkih iz posameznih virov, pa tudi razkorak med stremljenjem po formalizaciji naravnega jezika ter sočasnim upoštevanjem dejanske jezikovne rabe⁴⁰ (ibid.: 26–29).

³⁸ <<http://lojze.lugos.si/~darja/sloWnet.html>> – internetna stran prinaša opis projekta gradnje slovenskega wordneta s povezavami na relevantno literaturo. SloWNet je prosto dostopen za raziskovalne namene.

³⁹ Skupaj z informacijami, kje najti več podatkov o projektih, avtorica navaja v tej skupini še *Cyc*, *EuroWordNet*, *FrameNet* ter *HowNet*.

⁴⁰ Poskusi formalizacije naravnega jezika so problematični v smislu, da temeljijo na ideji urejenosti ter predvidljivosti jezika. Naravni jeziki so »anarhični, ne sledijo pravilom, neprestano se spreminjajo, so nepredvidljivi« (Teubert in Krishnamurty (ur.) 2007: 6). S tega stališča je obdelava naravnega jezika postavljena pred nehvaležno nalogo.

Novejši opis stanja slovenskega wordneta je na voljo denimo v Fišer in Erjavec (2008). Poleg posodobljenih podatkov o obsegu zbirke, ki so navedeni spodaj, prinaša prispevek tudi predstavitev preizkusa novih metod avtomatskega dopolnjevanja zbirke:

»Slovenski wordnet trenutno vsebuje skoraj 20.000 različnih literalov [pomensko opredeljenih iztočnic], ki pokrivajo večino osnovnih konceptov ter kar precej specifičnih, ki so bili večinoma pridobljeni iz Wikivirov in so s področja biologije. V wordnetu so zaenkrat predvsem samostalniki, poleg enobesednih literalov je precej tudi večbesednih. Najpogostejša relacija med sinseti je hipernimija, ki predstavlja 46 % vseh relacij v wordnetu.« (Fišer in Erjavec 2008: 41).

2 Leksikalna zbirka ASES

Med podjetji, ki se ukvarjajo z razvojem končnih jezikovnotehnoloških izdelkov za slovenščino, je trenutno vodilno podjetje Amebis, d. o. o., Kamnik. Jezikovne tehnologije, ki v sklopu podjetja nastajajo, so npr. strojni prevajalnik Presis (za jezikovni par slovenščina-angleščina), črkovalnik μ BesAna, slovnični pregledovalnik BesAna, sintetizator govora Govorec, programirani sogovornik Klepec itd.⁴¹

Razvoj naštetih izdelkov je osnovan na leksikalni zbirki ASES (Amebisov skupni elektronski slovar). Začetek nastajanja zbirke sega v obdobje, ko korpusni viri (večjega obsega, uravnoteženi) za slovenščino še niso bili na voljo. Kar se tiče nabora leksike, se zadnja leta zbirka posodablja predvsem iz referenčnih korpusov, ni pa osnovana na korpusnojezikoslovnih načelih.⁴² V prispevku iz leta 2002 najdemo takle opis zbirke:

»Osnovne enote sistema ASES so med seboj povezani pojmi, preko katerih se slovenske besede povezujejo z besedami v drugih jezikih. Poleg teh povezav se pojmi med seboj lahko povezujejo tudi v različne druge skupine ter v enega ali več delnih tezavrov. Pojmi poleg nekaterih pomenskih in drugih statističnih informacij vsebujejo še povezave na ustrezne besede oz. besedne zveze, sinonimne in antonimne povezave itd. Same besede vsebujejo osnovne morfološke informacije ter podatke o zlogovanju in izgovorjavi.« (Romih in Holozan 2002a: 166)

Ker je sistem ASES trenutno najrelevantnejša leksikalna zbirka za avtomatsko obdelavo slovenskega jezika in ker v literaturi še ni bil podrobneje predstavljen, je v pričujočem delu predstavitvi namenjenega več prostora. Sistem ASES v nadaljevanju poglavja obravnavamo predstavitveno, z namenom osnovnega očrta sistema oz. organizacijske strukture, ki jo prinaša; pri tem si pomagamo z analizo treh iztočnic: glagolske *izdati*, samostalniške *pajek* ter pridevniške *moder*.⁴³ Evalvacija zbirke v smislu izpostavljanja močnih ali šibkih mest oz. potencialov za izboljšavo v predstavitvenem delu ni predvidena, sledi pa delno v sklopu predloga nadgradnje zbirke v poglavju III-3.

⁴¹ <<http://www.amebis.si>>. O naštetih izdelkih več npr. v Romih in Holozan 2002a; 2000b; 2000c; Arhar in Romih 2006.

⁴² Glede na metodološke smernice gradnje lahko leksikalne zbirke delimo na (I) **predkorpusne leksikalne zbirke** – ki so starejšega datuma, njihova gradnja je bolj ali manj ciljno usmerjena, pogosto vezana na specifične probleme obdelave naravnega jezika; glede nastanka so metodološko različne, združujejo ročni vnos, nadgrajen z metodami ekstrakcije podatkov iz strojnoberljivih slovarjev, korpusov ipd., ter (II) **korpusne leksikalne zbirke** – izhodišče za gradnjo zbirke je nabor korpusnih leksikalnih informacij, tipično pogostnostna lista besed izbranega korpusa; korpus kot vir podatkov ostaja v središču metodološkega interesa, informacije so leksikogramatičnega tipa, frekventnost jezikovnih pojavov je bistvena; korpusni podatki se lahko dopolnjujejo z informacijami iz drugih virov.

⁴³ Na tem mestu gre zahvala Petru Holozanu za podatke o zbirki ter slike vmesnika.

2.1 Organizacija leksikalne zbirke

Sistem ASES prinaša nabor besed ter besednih zvez (število vseh iztočnic marca 2009 je blizu 872.500), in sicer skupaj za slovenščino in druge jezike, aktualne za razvoj avtomatskega prevajanja.⁴⁴ Povezava med jeziki je vzpostavljena s pomočjo elementov, imenovanih *pomeni*, ki so formalizacija semantičnega dela posameznega (enobesednega) jezikovnega znaka.

Pomeni so definirani predvsem jezikovnokontrastivno, torej glede na potrebe strojnega prevajanja, in sicer: (I) **pomenskoločevalno** – kadar ima ena beseda več možnih prevodov, se pomeni ločujejo glede na te možnosti (glej primer *moder* v III-2.2.2 – pomen *moder* (*barva*) se prevaja v angleški *blue*, pomen *moder* (*pameten*) pa v angleški *wise*) ter (II) **pomenskozdrževalno** – kadar se več besed prevaja na enak način, so združene pod enim samim pomenom (ponovno primer *moder* v III-2.2.2 – pomen *moder* (*barva*) je povezan tako z obliko *moder* kot tudi z obliko *sinji*; tako prva kot druga se prevajata v angleški *blue*). Na ta način ASES prinaša tudi določen nabor podatkov o sopomenskosti v slovenščini.

Jedrni del zbirke predstavlja abecedno urejen seznam iztočnic različnih vrst (vrsto iztočnice v zbirki predstavljajo enočrkovne oznake, ki so v nadaljevanju navedene v oglatih oklepajih). Za poimenovanje različnih tipov iztočnic se v leksikalni zbirki ASES uporablja besedišče, ki je v jezikoslovju običajno definirano drugače kot v zbirki (tu je npr. privzet **dvojnostni odnos med besedo ter pomenom**, ne pa *obliko* in *pomenom*). V nadaljevanju poglavja (samo znotraj III-2) uporabljamo poimenovanja, kot so razumljena v leksikalni zbirki, na tem mestu pa jih v izogib terminološki zmedi na kratko definiramo:

(I) osnovna formalna enota zbirke je **beseda** [B], ki predstavlja oblikovni del jezikovnega znaka in kot taka prinaša informacije o možnih oblikah besede (*oblike*), skupaj z oblikam ustrezajočimi oblikoskladenjskimi oznakami; beseda je povezana s *pomeni* ter *uporabami* (običajno je to nabor besednih zvez, v katerih se izhodiščna beseda pojavlja);

(II) **pomen** [P] je formalizacija pomenskega dela jezikovnega znaka. Povezan je na *besede* različnih jezikov, ki so v tem kontekstu imenovane *izvori*. Urejanje pomenov prinaša tudi možnost povezovanja pomenov v t. i. *odnose* (npr. nadpomenskost, protipomenskost, manjšalnica, prebivalec, lastnost, ...), mogoče pa jih je tudi kvalificirati glede na rabo (npr. napačno, pogovorno, strokovno, redko, množinsko, britansko, ameriško, ...);

(III) **zveza** [Z] prinaša besedne zveze, ob katerih je informacija o njihovi sestavljenosti;

(IV) **skupina** [S] prinaša nabor besed, ki jih je smiselno združiti zaradi potreb obdelave naravnega jezika, npr. nedovršno ter dovršno varianto glagola, samostalni in iz njega izpeljani svojilni pridevnik, britansko ter ameriško različico zapisa besede itd.;

(V) **glagolska predloga** [G] prinaša informacije o glagolski vezljivosti.⁴⁵

V nadaljevanju sledi nekaj slik za boljšo predstavo organizacije leksikalne zbirke in raznovrstnih iztočnic, ki jih sistem prinaša. Na primeru obravnave para iztočnic *pajek* (*pajek* – *živo* nasproti *pajek* – *neživo*) sta predstavljena osnovni vrsti iztočnic, tj. *beseda* ter *pomen*. Iztočnice vrste *zveza* ter *skupina* niso nadalje obravnavane, saj zgolj organizirajo leksikalno zbirko na sekundarnem nivoju (združujejo enote prvega reda). *Glagolske predloge* so kot vmesna stopnja med povezavo glagolskih besednih oblik ter pomenov bolj komplicirane in zato podrobneje razložene ob primeru *izdati* v poglavju III-2.2.3.

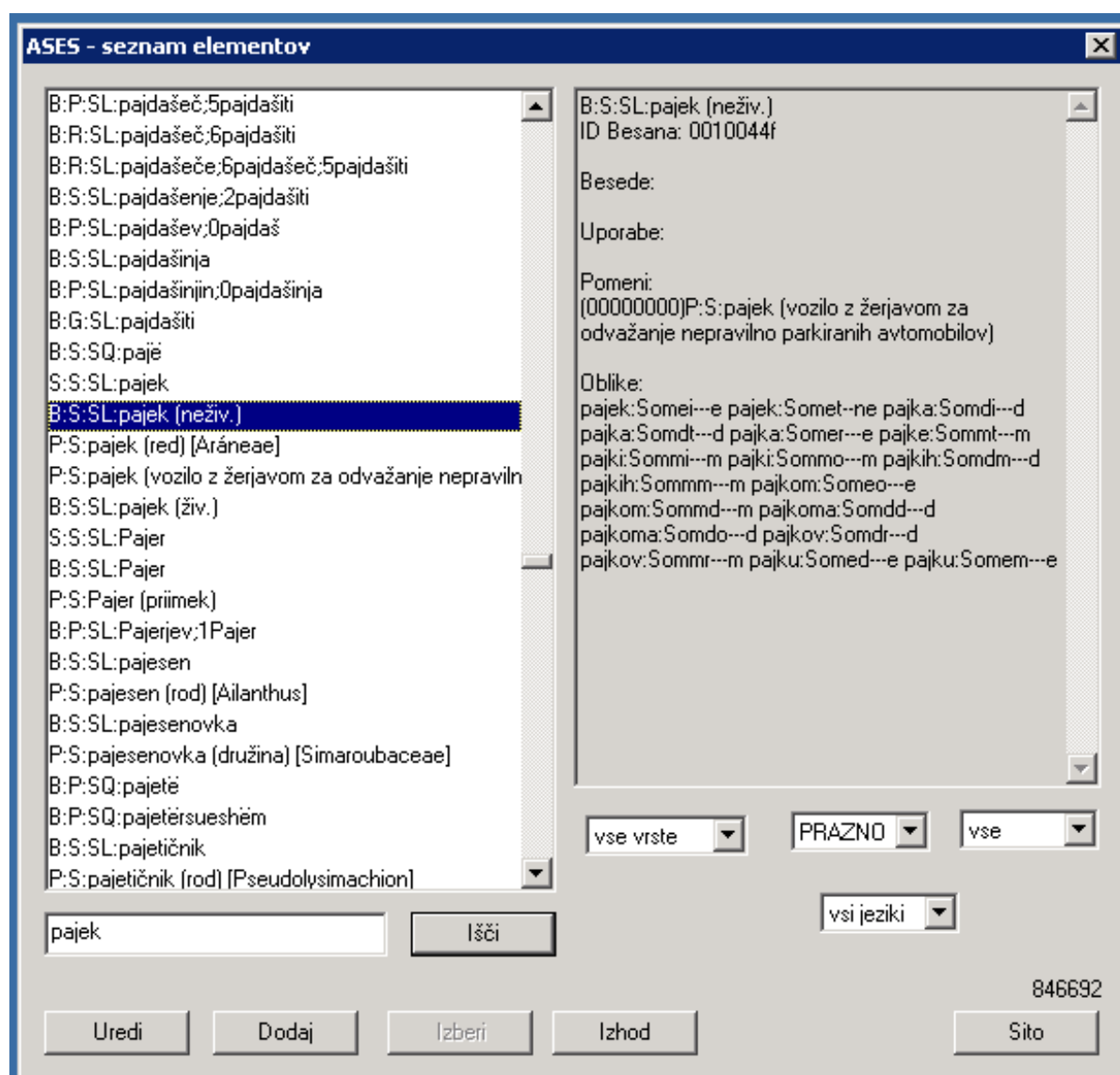
⁴⁴ Trenutno je na tržišču le prevajalnik za par slovenščina-angleščina, zato v sistemu najdemo največ angleških iztočnic, poleg njih pa tudi nemške, francoske ter albanske.

⁴⁵ Pojem glagolska vezljivost na tem mestu uporabljamo splošno; kot bo vidno v nadaljevanju, so podatki v zbirki le delno primerljivi z jezikoslovnim sistemom opisa glagolske vezljivosti, kot ga najdemo npr. v Žele 2001 ali Žele 2008.

2.1.1 Beseda (oblika)

Spodnja slika⁴⁶ prikazuje obravnavo oblikovnega dela jezikovnega znaka *pajek*. Iztočnica se imenuje *B:S:SL:pajek (neživ.)* Oznaka pred iztočnico pomeni, da gre za *besedo, samostalni, slovenski jezik*. Iztočnica v imenu prinaša še podatek, namenjen ločevanju med enakopisnima oblikama *pajek – neživo* in *pajek – živo*.

Kot je razvidno iz primera, vmesnik leksikalne zbirke prinaša v levem oknu abecedno urejeni seznam iztočnic, v desnem oknu pa so navedeni na izbrano iztočnico vezani podatki: (I) povezava s pomenom *P:S:pajek (vozilo z žerjavom za odvažanje nepravilno parkiranih avtomobilov)*; pomen je opredeljen z definicijo slovarskega tipa⁴⁷, kar sicer v sistemu ASES ni pravilo – pomenske opredelitve iztočnic so namenjene le večji preglednosti leksikalne zbirke, zato so običajno asociacijskega tipa (kot npr. razlikovanje med dvema pomenoma pridevnika *moder: moder (barva)* ter *moder (pameten)*), (II) v kategoriji *Oblike* so navedene sklonne oblike samostalnika z ustrežajočimi oblikoskladenjskimi oznakami.⁴⁸ Preglednejša predstavitev podatkov sledi v poglavju III-2.2.1.



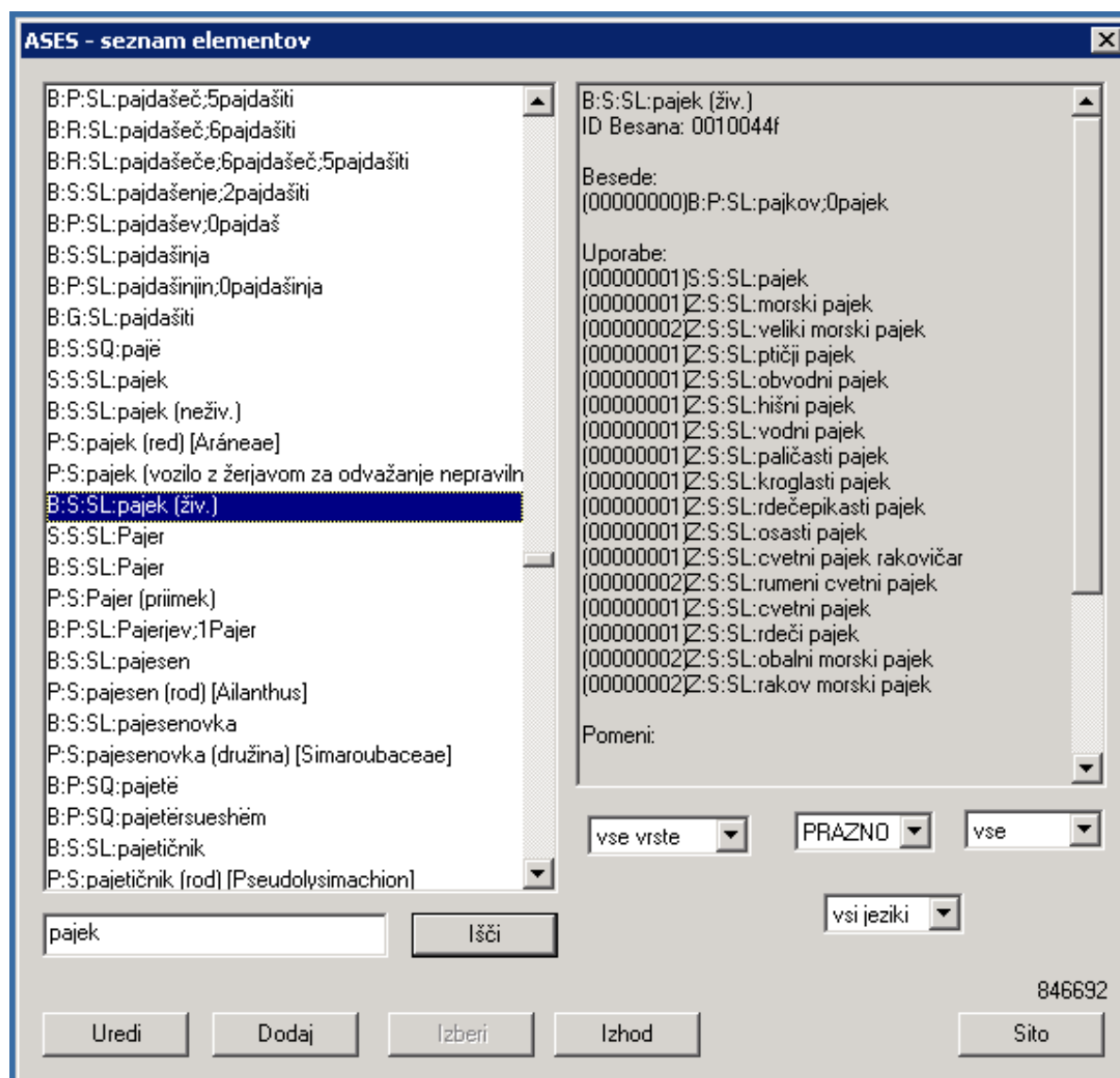
Slika 1: Vmesnik sistema ASES – B:S:SL:pajek (neživ.).

⁴⁶ V času priprave doktorske naloge je na podjetju Amebis potekala posodobitev vmesnika leksikalne zbirke. Na tem mestu so navedene slike iz časa pred posodobitvijo.

⁴⁷ V obravnavanem primeru je vir definicije Slovar slovenskega knjižnega jezika.

⁴⁸ Slednje ustrezajo označevalnemu sistemu Multext-East (glej poglavje II-1.2.2).

Naslednja slika prikazuje iztočnico *B:S:SL:pajek (živ.)*. Pri tej so v desnem oknu navedene tudi t. i. *uporabe*: pri obravnavanem primeru gre za nabor bioloških terminoloških besednih zvez z obravnavano iztočnico.

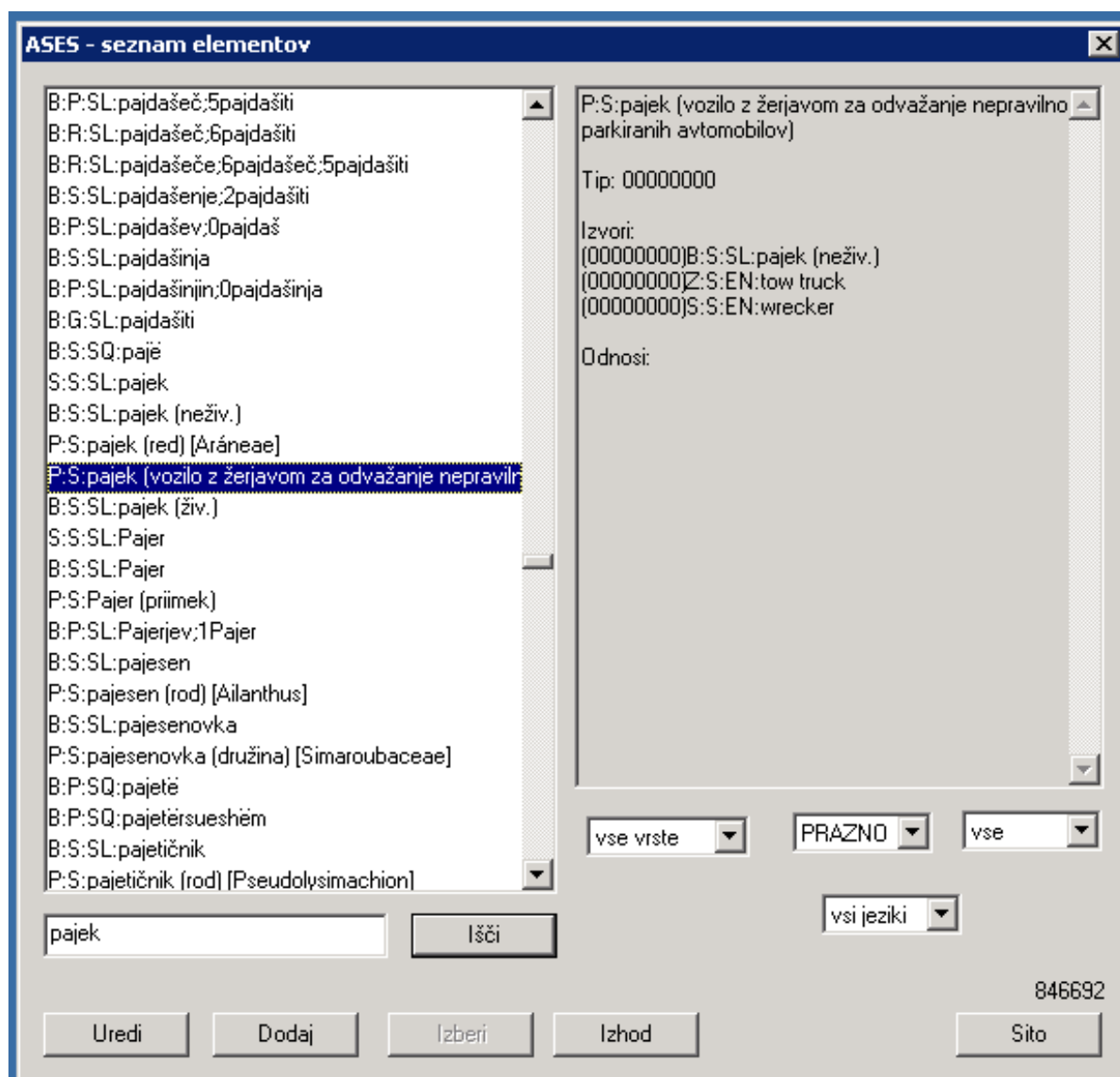


Slika 2: Vmesnik sistema ASES – B:S:SL:pajek (živ.).

2.1.2 Pomen

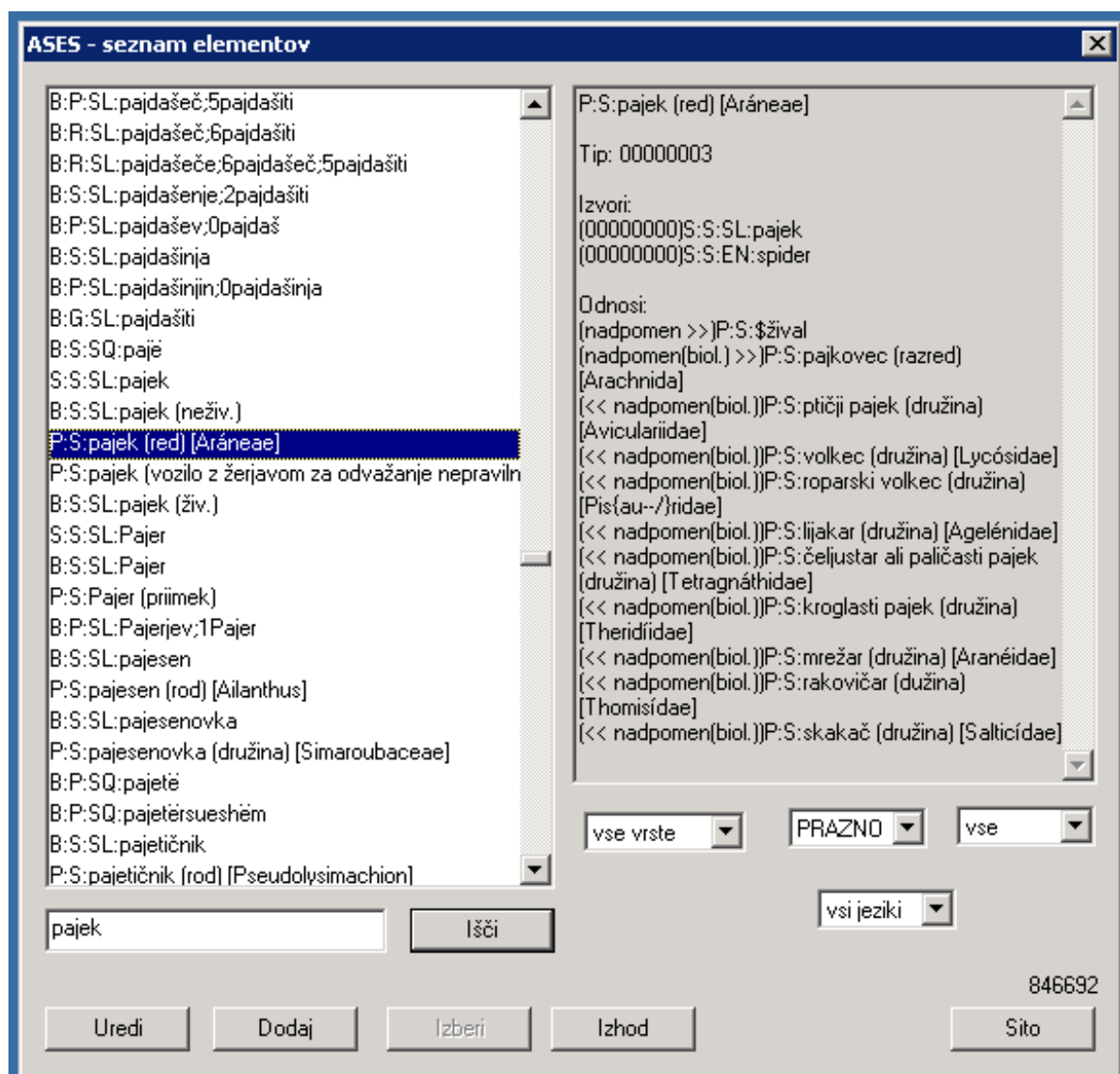
Ker sta *pajek* – živo ter *pajek* – neživo ločeni iztočnici, ASES prinaša tudi dve iztočnici za *pomen*, in sicer *P:S:pajek* (vozilo z žerjavom za odvažanje nepravilno parkiranih avtomobilov) ter *P:S:pajek* (red *Araneae*). Oznaka *P:S* pomeni, da gre za *pomen*, *samostalniški*. Pomen je običajno opredeljen v slovenščini, nima pa jezikovne oznake, ker se veže z *besedami* različnih jezikov.

Na naslednji sliki je prikaz obravnave pomena *P:S:pajek* (vozilo z žerjavom za odvažanje nepravilno parkiranih avtomobilov); v desnem oknu so prikazani *izvori*, tj. besedne oblike, s katerimi je obravnavani pomen povezan. Prva na seznamu je slovenska *B:S:SL:pajek* (neživ.), sledita pa dve angleški, zveza *Z:S:EN: tow truck* ter skupina *S:S:EN:wrecker*.



Slika 3: Vmesnik sistema ASES – P:S: pajek (vozilo z žerjavom za odvoz nepravilno parkiranih avtomobilov).

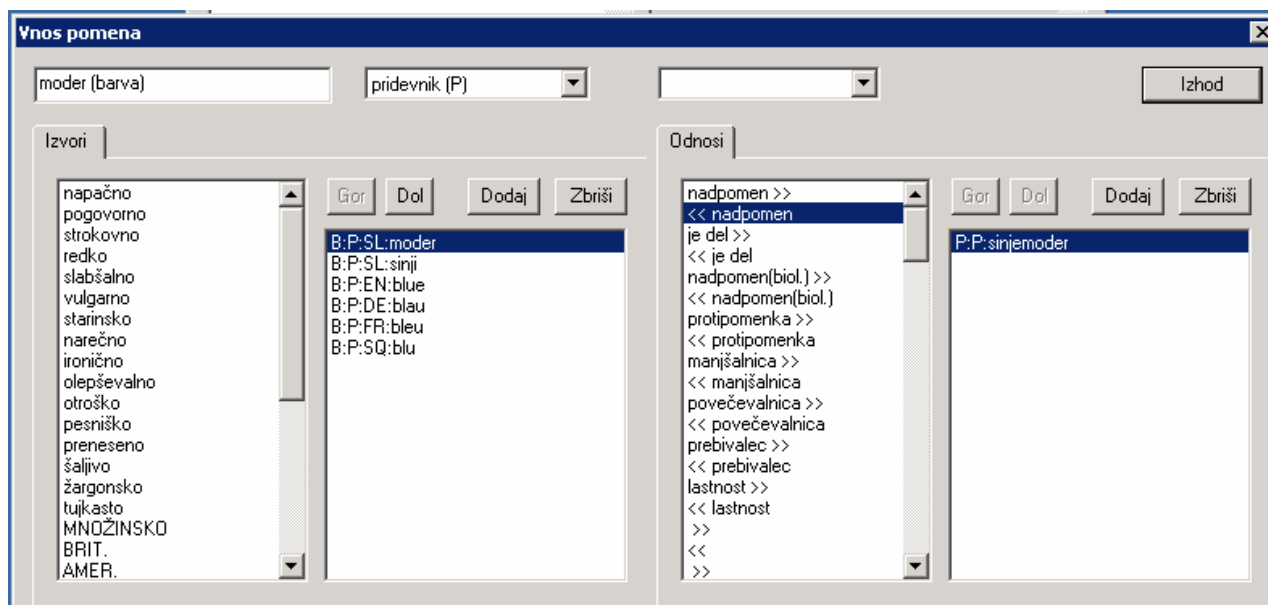
Slika 4 prinaša informacije o pomenski iztočnici *P:S:pajek (red) [Araneae]*. Okno na desni od niza iztočnic prinaša nabor *izvorov*, slovenskega *B:S:SL:pajek (živ.)* ter angleškega *S:S:EN:spider*. V rubriki *odnosi* so našteje informacije o nadpomenskih ter podpomenskih odnosih, pri obravnavanem primeru predvsem tiste, ki temeljijo na biološki klasifikaciji. Natančneje so podatki predstavljeni v poglavju III-2.2.1.



Slika 4: Vmesnik sistema ASES – P:S: pajek (red) [Araneae].

Spodnja slika prikazuje še možnosti urejanja pomenov. Obravnavana iztočnica je pridevniška, in sicer *P:P:moder (barva)*. V drugem od štirih stolpcev so naštet *izvori* (torej besedne oblike različnih jezikov, s katerimi je obravnavani pomen povezan: poleg dveh slovenskih so na seznamu še angleška, nemška, francoska ter albanska ustreznica). Prvi od stolpcev prinaša niz kvalifikatorjev, ki jih lahko za posamezen primer izberemo: za na sliki označeni izvor *B:P:SL:moder* ni predvidena nobena omejitev rabe, zato ni v povezavi z njo izbran noben kvalifikator.

Zadnja dva stolpca prinašata možnost povezovanja pomenov v pomenske odnose⁴⁹ – pomen *P:P:sinjemoder* je označen kot podpomenski obravnavanemu *P:P:moder (barva)*. Preglednejša predstavitev podatkov sledi v III-2.2.2.



Slika 5: Vmesnik sistema ASES – urejanje pomena *P:S: moder (barva)*.

2.2 Nabor informacij v leksikalni zbirki

Iz opisa organizacije sistema ASES je razvidno, da gre za kompleksno zasnovano leksikalno zbirko, ki združuje oz. v različne vrste odnosov postavlja različne vrste iztočnic. V nadaljevanju bodo za izbrane primere (samostalnik *pajek*, glagol *izdati* ter pridevnik *moder*) informacije iz leksikalne zbirke predstavljene v shematizirani obliki, ki omogoča strnjen pregled nad njimi.

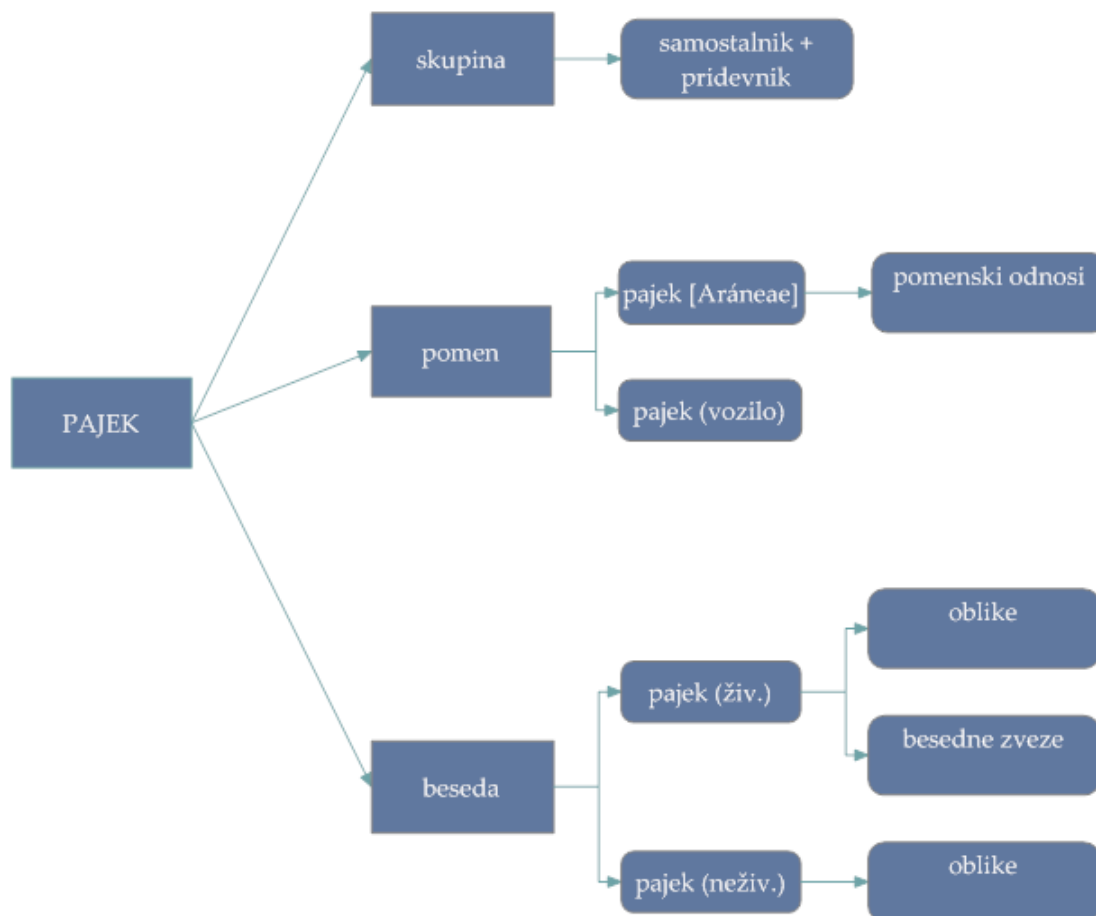
Primeri so poskus prikaza tipične obravnave posamezne od izbranih treh besednih vrst, ne pa tudi predstavitev vseh možnih povezav, ki jih leksikalna zbirka za slednje omogoča; z drugimi besedami, analiza izbranih primerov je namenjena prikazu trenutnega stanja leksikalne zbirke, ne njenega celovitega potenciala. Prav tako, kot rečeno, ni veliko prostora namenjenega interpretaciji ter evalvaciji prikazanih podatkov, saj to dvoje presega namen pričujočega dela.

⁴⁹ Kot je razvidno s seznama, ne gre le za pomenske odnose (nadpomensko, podpomensko, protipomensko), ampak tudi besedotvorne odnose (manjšalnica, povečevalnica) ali odnose zunajjezikovne realnosti (lastnost, prebivalec). Nabor odnosov v sistemu ASES ni končen, novi tipi odnosov se dodajajo glede na trenutne razvojne potrebe.

2.2.1 Primer 1 – pajek

Za iskalni pogoj *pajek* leksikalna zbirka ASES prinaša tri vrste iztočnic, in sicer: (I) skupino, v kateri so združeni samostalniki z izpeljanimi pridevniki, (II) dva pomena, pri enem od njiju najdemo tudi podatke o pomenskih odnosih, (III) dve besedi, obe prinašata nabor besednih oblik, ena pa še nabor besednih zvez.

Opisano členitev iztočnic prikazuje spodnja slika, odnosi med iztočnicami bodo predstavljeni v nadaljevanju poglavja. Besedne oblike, besedne zveze ter pomenski odnosi so navedeni v tabelah pod sliko.



Slika 6: ASES – pajek – členitev iztočnic.

Oblike besed ter ustrezajoče oblikoskladenjske oznake

pajek (živ.)	pajek (neživ.)
pajek:Somei---e	pajek:Somei---e
pajka:Somdi---d	pajek:Somet--ne
pajka:Somdt---d	pajka:Somdi---d
pajka:Somer---e	pajka:Somdt---d
pajka:Somet--de	pajka:Somer---e
pajke:Sommt---m	pajke:Sommt---m
pajki:Sommi---m	pajki:Sommi---m
pajki:Sommo---m	pajki:Sommo---m
pajkih:Somdm---d	pajkih:Somdm---d
pajkih:Sommm---m	pajkih:Sommm---m
pajkom:Someo---e	pajkom:Someo---e
pajkom:Sommd---m	pajkom:Sommd---m

pajkoma:Somdd---d	pajkoma:Somdd---d
pajkoma:Somdo---d	pajkoma:Somdo---d
pajkov:Somdr---d	pajkov:Somdr---d
pajkov:Sommr---m	pajkov:Sommr---m
pajku:Somed---e	pajku:Somed---e
pajku:Somem---e	pajku:Somem---e

Tabela 4: ASES – pajek: Oblike besed ter ustrezajoče oblikoskladenjske oznake.

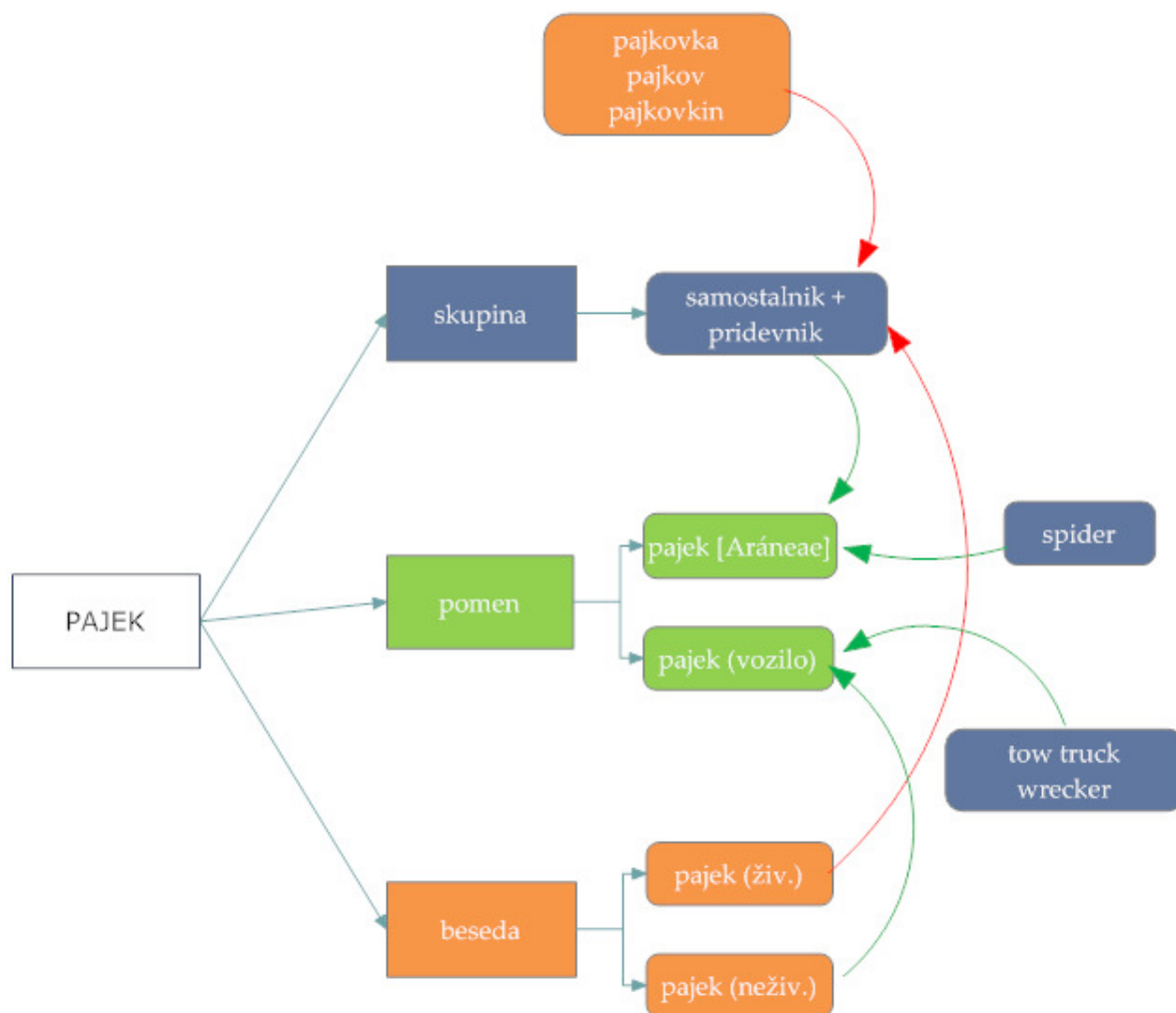
besedne zveze	pomenski odnosi
pajek (živ.)	pajek (živ.)
morski pajek	pajek (red) [Aráneae]
veliki morski pajek	
ptičji pajek	nadpomenka
obvodni pajek	žival
hišni pajek	
vodni pajek	nadpomenka (biol.)
paličasti pajek	pajkovec (razred) [Arachnida]
kroglasti pajek	
rdečepikasti pajek	podpomenka (biol.)
osasti pajek	ptičji pajek (družina) [Aviculariidae]
cvetni pajek rakovičar	volkec (družina) [Lycósidae]
rumeni cvetni pajek	roparski volkec (družina) [Pisridae]
cvetni pajek	lijakar (družina) [Agelénidae]
rdeči pajek	čeljstar ali paličasti pajek (družina) [Tetragnáthidae]
obalni morski pajek	kroglasti pajek (družina) [Theridíidae]
rakov morski pajek	mrežar (družina) [Aranéidae]
	rakovičar (družina) [Thomisidae]
	skakač (družina) [Salticida]

Tabela 5: ASES – pajek: Besedne zveze in pomenski odnosi.

Kot je razvidno iz Tabele 5, so tako besedne zveze kot pomenski odnosi, ki jih v leksikalni zbirki najdemo pri obliki oz. pomenu *pajek (živ.)*, specializiranega tipa – na eni strani terminološke besedne zveze, ki poimenujejo vrste pajkov, npr. [*morski, ptičji, hišni, rdeči ...*] *pajek*, na drugi uvrstitev pajka v ustrezno vejo biološkega debla glede na klasifikacijo živih bitij.⁵⁰

V nadaljevanju je predstavljena še ena slikovna ponazoritev, na kateri so v ospredju odnosi med različnimi vrstami iztočnic. Zaradi želje po večji preglednosti je vsaka vrsta iztočnice prikazana z drugo barvo: *skupine* so modre, *pomeni* zelene ter *besede* oranžne barve. Povezave so prikazane s puščicami dveh barv, rdeče puščice prikazujejo uvrščanje besed v skupino, zelene pa povezave med pomeni ter njihovimi izvori.

⁵⁰ Zanimiva je dvojnost uvrstitve v podpomensko-nadpomenski odnos s splošno nadpomenko *žival* na eni strani ter »biološko« nadpomenko (nadpomensko glede na biološko klasifikacijo) *pajkovec*.



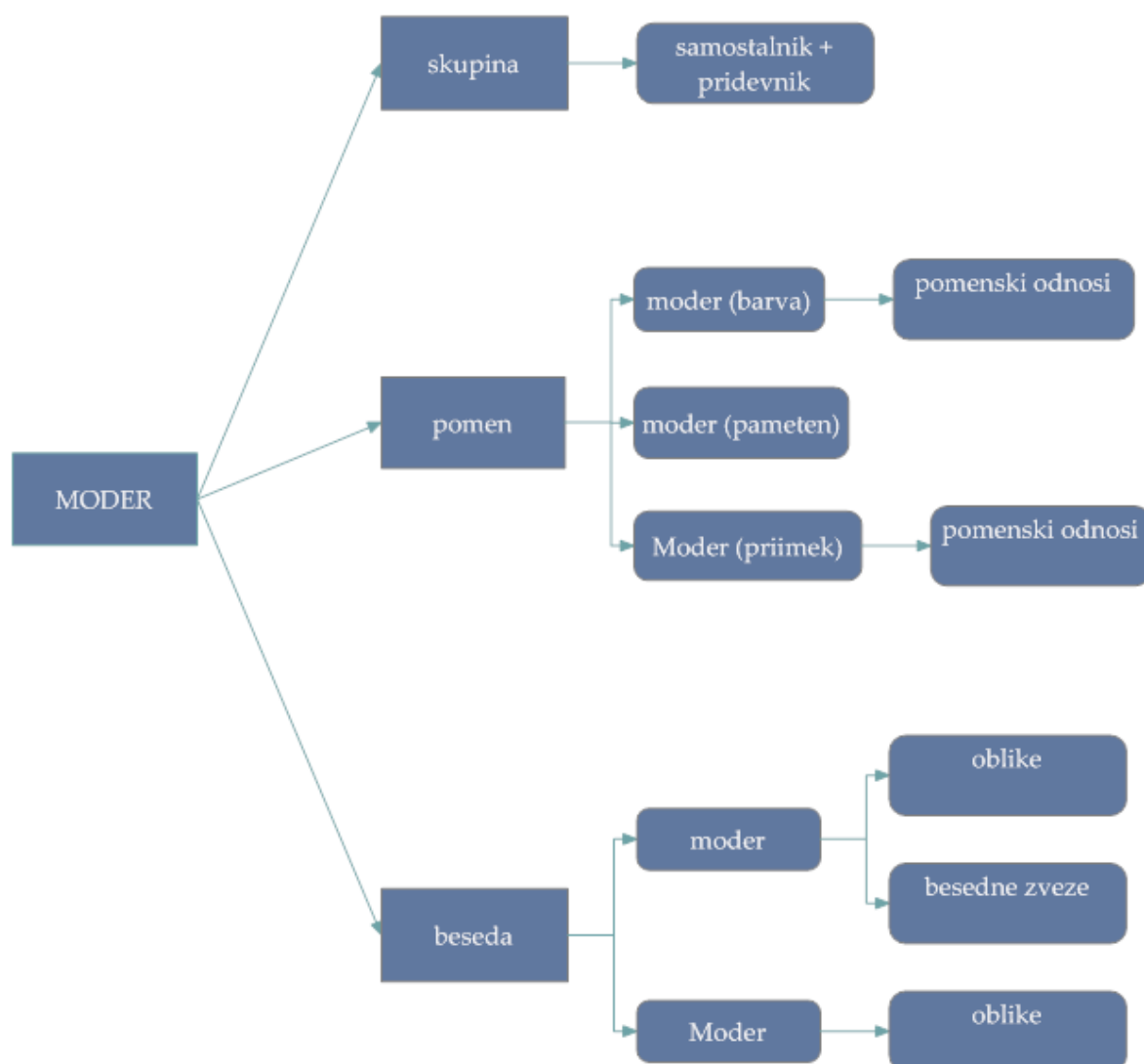
Slika 7: ASES – pajek – odnosi med iztočnicami.

Skupino *pajek* poleg besede *pajek (živ.)* v leksikalni zbirki sestavljajo še samostalnik *pajkovka* ter izpeljana svojilna pridevnika *pajkov* ter *pajkovkin*. Namen skupin je združevanje iztočnic v enote višjega reda – povezave, ki jih vzpostavlja skupina kot celota, veljajo za vse člane skupine. Pomen *pajek (red)* [Aráneae] je povezan s skupino *pajek* ter angleško skupino *spider*, pomen *pajek (vozilo)* pa je povezan z besedo *pajek (neživ.)* ter angleškima skupinama *tow truck* in *wrecker*.

2.2.2 Primer 2 – moder

Tudi primer *moder* prinaša tri vrste iztočnic: (I) skupino, ki je kot pri prejšnjem primeru nabor samostalnikov ter svojilnih pridevnikov, (II) tri pomeni, dva od pomenov prinašata še nadaljnje informacije o pomenskih odnosih, (III) dve besedi, obe z informacijami o besednih oblikah, ena pa prinaša še nabor besednih zvez.

Nabor besednih iztočnic pri tem primeru prinaša pridevniško *moder* ter samostalniško *Moder* – gre za lastno ime oz. priimek – pomeni pa so trije, *moder (barva)*, *moder (pameten)* ter *Moder (priimek)*. V nadaljevanju poglavja sledi še ponazoritev odnosov med iztočnicami, besedne oblike, besedne zveze ter pomenski odnosi pa so tudi tokrat navedeni v tabelah pod sliko.



Slika 8: ASES – moder – členitev iztočnic.

Oblike besed ter ustrezajoče oblikoskladenjske oznake

moder		Moder
moder:Pkomein	modrejšem:Pkpmem	Moder:Smmei---e
moder:Pkometn-n	modrejšem:Pkpsem	Modra:Smmdi---d
modra:Pkomdi	modrejšemu:Pkpmed	Modra:Smmdt---d
modra:Pkomdt	modrejšemu:Pkpsed	Modra:Smmer---e
modra:Pkosmi	modrejši:Pkpmeid	Modra:Smmet--de
modra:Pkosmt	modrejši:Pkpmetd-n	Modre:Smmt---m
modra:Pkozei	modrejši:Pkpmmi	Modri:Smmmi---m
modre:Pkommt	modrejši:Pkpsdi	Modri:Smmmo---m
modre:Pkozer	modrejši:Pkpsdt	Modrih:Smmdm---d
modre:Pkozmi	modrejši:Pkpzdi	Modrih:Smmmm---m
modre:Pkozmt	modrejši:Pkpzdt	Modrom:Smmeo---e
modrega:Pkomer	modrejši:Pkpzed	Modrom:Smmdm---m
modrega:Pkomet--d	modrejši:Pkpzem	Modroma:Smdd---d
modrega:Pkoser	modrejših:Pkpmdm	Modroma:Smmdo---d
modrejša:Pkpmdi	modrejših:Pkpmdr	Modrov:Smmdr---d

modrejša:Pkpmtdt	modrejših:Pkpmmm	Modrov:Smmmr---m
modrejša:Pkpsmi	modrejših:Pkpmmr	Modru:Smmed---e
modrejša:Pkpsmt	modrejših:Pkpsdm	Modru:Smmem---e
modrejša:Pkpzei	modrejših:Pkpsdr	
modrejše:Pkpmmt	modrejših:Pkps	
modrejše:Pkpsei		
modrejše:Pkpset		
modrejše:Pkpzer		
modrejše:Pkpzmi		
modrejše:Pkpzmt		
modrejšega:Pkpmer		
modrejšega:Pkpmet--d		
modrejšega:Pkpser		

Tabela 6: ASES – moder: Oblike besed ter ustrezajoče oblikoskladenjske oznake.

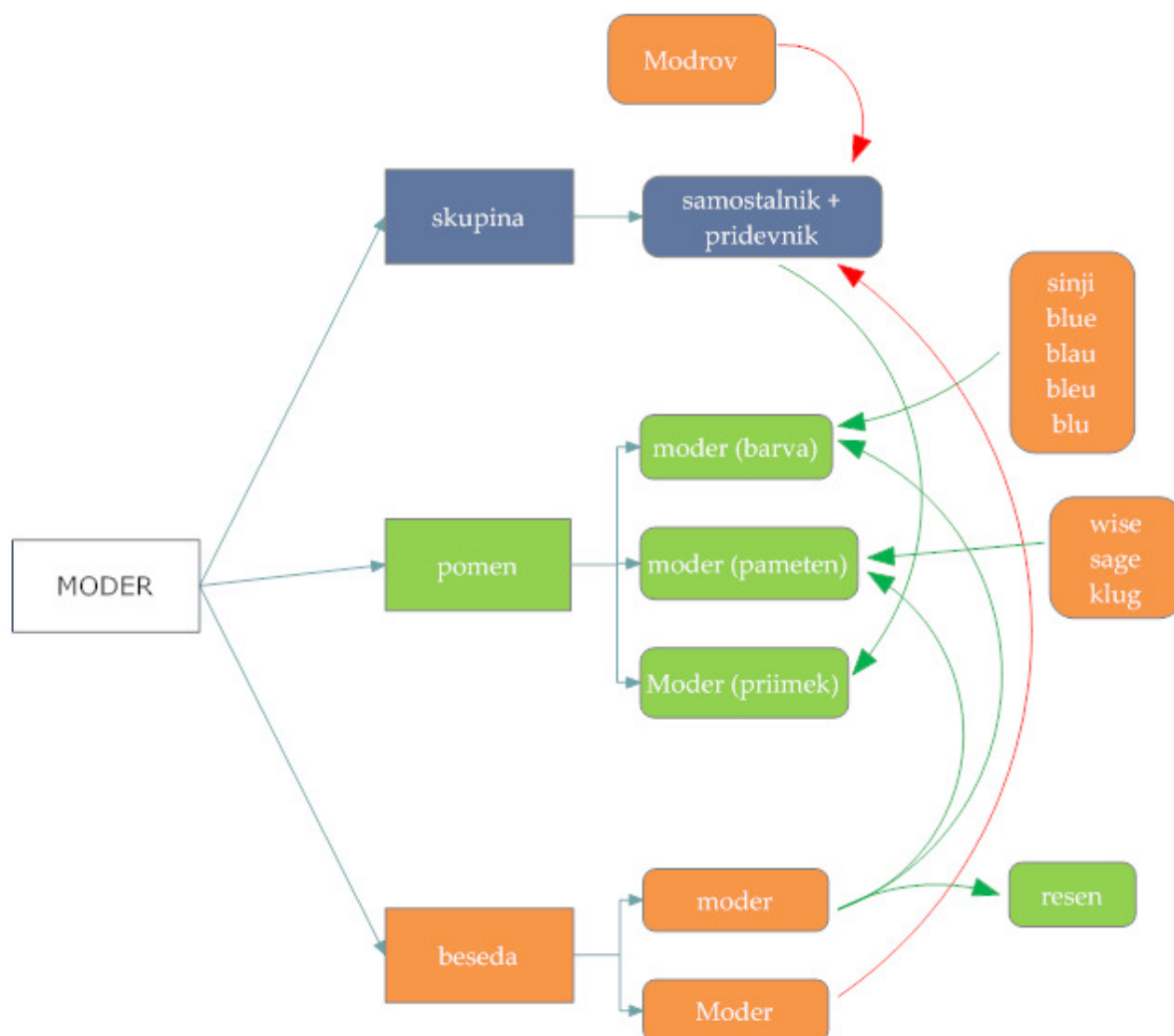
besedne zveze	pomenski odnosi	
moder	moder (barva)	Moder (priimek)
modra galica	podpomenka	nadpomenka
modri angel	sinjemoder	priimek
svetlo moder		
modra duglazija		
modri egiptovski lotos		
modra kurja češnjica		
modri volovski jezik		
modro milje		
modri jetičnik		
modro kosteničevje		
modri glavinec		
modri meček		
bermudski modri meček		
jesenska modra čebulica		
dvolistna modra čebulica		
travniška modra čebulica		
modra stožka		
modra zamorska mačka		
modri gnu		
modra ovca		

Tabela 7: ASES – moder: Besedne zveze in pomenski odnosi.

Besedne zveze so vezane na pridevnik, ponovno gre za nabor (po večini bioloških) stalnih besednih zvez, tj. poimenovanj rastlin ter živali, npr. *modri jetičnik*, *modri glavinec*, *jesenska modra čebulica*; *modra zamorska mačka*, *modri gnu* itd. Da prinaša zbirka – kar se tiče stalnosti ter idiomatičnosti – raznovrstne besedne zveze, priča denimo skupna uvrstitev primerov *modra galica*, *svetlo moder*, *modri angel*.

Kot podpomenka je v zbirki k pomenu *moder (barva)* uvrščen pomen *sinjemoder*, pomen *Moder (priimek)* pa je uvrščen kot podpomenka pomenu *priimek*.

Sledi ponazoritev odnosov med iztočnicami, barvna legenda ostaja enaka kot pri prejšnjem primeru: *skupine* so modre, *pomeni* zelene ter *besede* oranžne barve, rdeče puščice povezujejo besede v skupine, zelene pa povezujejo pomene z oblikami.



Slika 9: ASES – moder – odnosi med iztočnicami.

V skupino je poleg samostalnika *Moder* uvrščen še izpeljani svojilni pridevnik *Modrov*. Skupina je povezana s pomenom *Moder (priimek)*. Beseda *moder* je povezana s tremi pomeni, *moder (barva)*, *moder (pameten)* ter *resen*. Pomen *moder (barva)* je povezan še z besedami *sinji*, angleško *blue*, nemško *blau*, francosko *bleu* ter albansko *blu*. Pomen *moder (pameten)* pa je povezan še z angleškima *wise* in *sage* ter nemško *klug*.

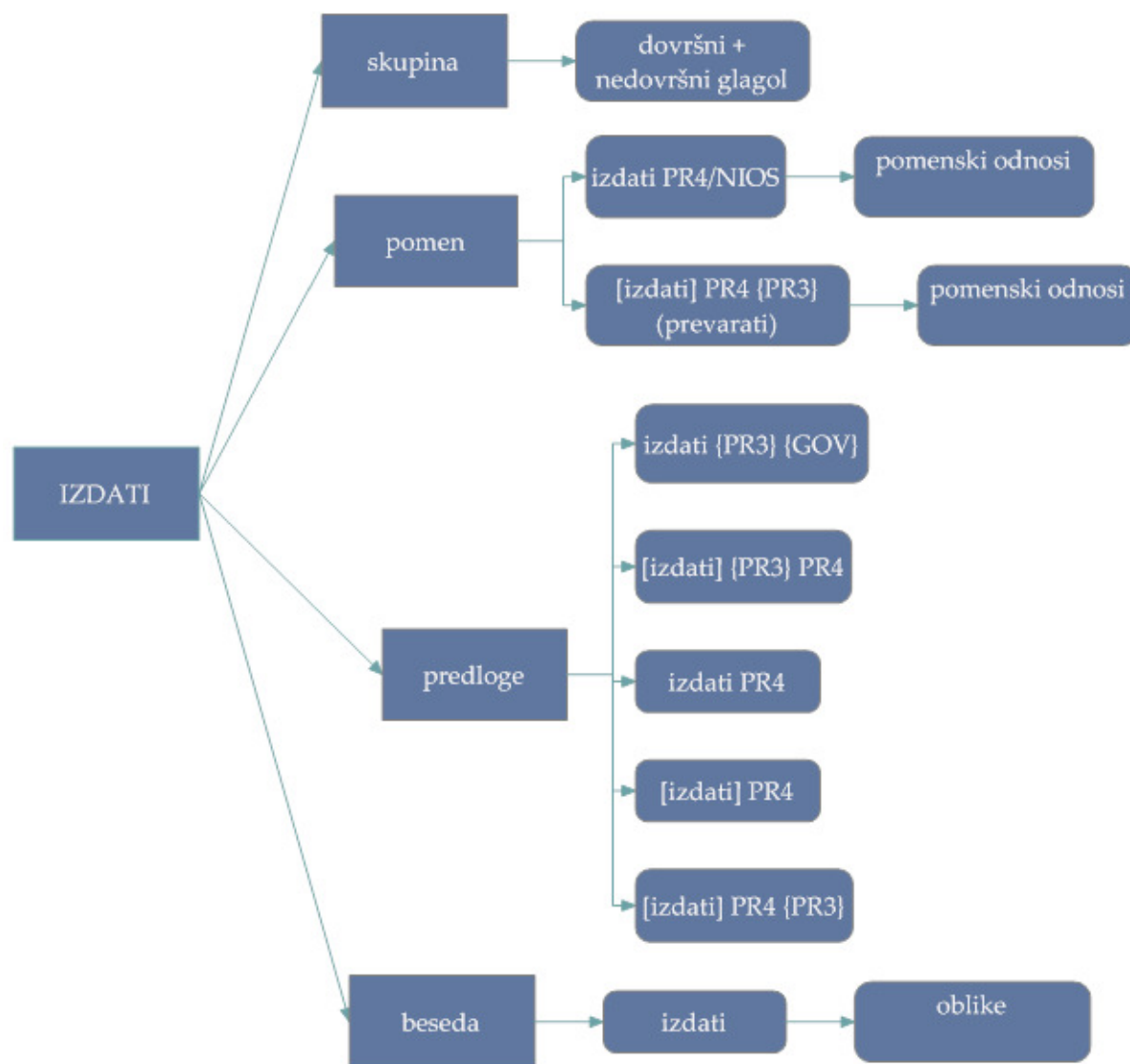
2.2.3 Primer 3 – izdati

Tretji primer, *izdati*, prinaša štiri vrste iztočnic: (I) skupina tokrat prinaša kombinacijo dovršnega ter nedovršnega glagola, (II) beseda je ena sama, prinaša informacijo o besednih oblikah, (III) pomena sta dva, oba prinašata informacije o pomenskih odnosih, poleg tega v leksikalni zbirki najdemo še (IV) pet glagolskih predlog, ki podajajo informacije skladenjskega nivoja.

Za zapis skladenjskih informacij v glagolskih predlogah se uporablja poseben format, ki ga je treba pred nadaljevanjem na kratko razložiti. Spodnja tabela prinaša seznam simbolov, ki jih srečujemo v nadaljevanju tega poglavja (poleg teh za opredeljevanje skladenjskih odnosov v sistemu ASES obstajajo še drugi), skupaj z razlago vsakega simbola:

simbol	razlaga
PR3	predmet v tretjem sklonu
PR4	predmet v četrtem sklonu
{}	element znotraj zavutih oklepajev je neobvezen
[]	glagol znotraj oglatih oklepajev stoji namesto skupine dovršnega ter nedovršnega glagola
OS ter NIOS	predhodni element tipično je (OS) ali ni (NIOS) oseba
GOV	nadomešča oznako PR4 na mestih, kjer tipično nastopa premi ali odvisni govor
PRD	<i>direct object</i> , uporablja se pri angleških primerih
\$	določa poseben tip pomena, ki sam ni povezan z oblikami, so pa nanj vezani drugi pomeni oz. določena pravila avtomatskega prevajalnika
\$\$	določa poseben tip pomena, ki služi formalizaciji pravil v zvezi s kategorijama oseba ter spol pri glagolih

Tabela 8: Razlaga simbolov za ponazarjanje skladenjskih odnosov v sistemu ASES.



Slika 10: ASES – izdati – členitev iztočnic.

Za primer s pomočjo Tabele 8 interpretiramo zapis obeh pomenov na Sliki 10: (I) **izdati PR4/NIOS** – kombinacija zgolj dovršne oblike glagola izdati s predmetom v tožilniku, na mestu katerega tipično nastopa beseda ali besedna zveza, ki ne izraža podspola človeškosti, (II) **[izdati] PR4 {PR3} (prevarati)** – kombinacija dovršne ali nedovršne oblike glagola izdati z obveznim predmetom v tožilniku ter neobveznim predmetom v dajalniku. V normalnih oklepajih sledi za lažje razumevanje še pomenska opredelitev glagola v tej rabi, tj. prevarati.

Oblike besed ter ustrezajoče oblikoskladenjske oznake

izdati

izda:Gppste--n-----d	izdan:Gpds-emr-----d
izdaš:Gppsde--n-----d	izdana:Gpds-dmr-----d
izdaj:Gpvsde-----d	izdana:Gpds-ezr-----d
izdajmo:Gpvspm-----d	izdana:Gpds-msr-----d
izdajo:Gppstm--n-----d	izdane:Gpds-mzr-----d
izdajta:Gpvsdd-----d	izdani:Gpds-dsr-----d
izdajte:Gpvsdm-----d	izdani:Gpds-dzr-----d
izdajva:Gpvspd-----d	izdani:Gpds-mmrr-----d
izdal:Gpdr-emt-----d	izdano:Gpds-esr-----d
izdala:Gpdr-dmt-----d	izdasta:Gppsdd--n-----d
izdala:Gpdr-ezt-----d	izdasta:Gppstd--n-----d
izdala:Gpdr-mst-----d	izdaste:Gppsdm--n-----d
izdale:Gpdr-mzt-----d	izdat:Gpm-----d
izdali:Gpdr-dst-----d	izdata:Gppsdd--n-----d
izdali:Gpdr-dzt-----d	izdata:Gppstd--n-----d
izdali:Gpdr-mmt-----d	izdate:Gppsdm--n-----d
izdalo:Gpdr-est-----d	izdati:Gpn-----d
izdam:Gppspe--n-----d	izdava:Gppsdp--n-----d
izdamo:Gppspm--n-----d	

Tabela 9: ASES – izdati: Oblike besed ter ustrezajoče oblikoskladenjske oznake.

pomenski odnosi	
izdati PR4/NIOS	[izdati] PR4 {PR3} (prevarati)
GP2+ \$ni oseba	GP2+ domovina država narod \$oseba

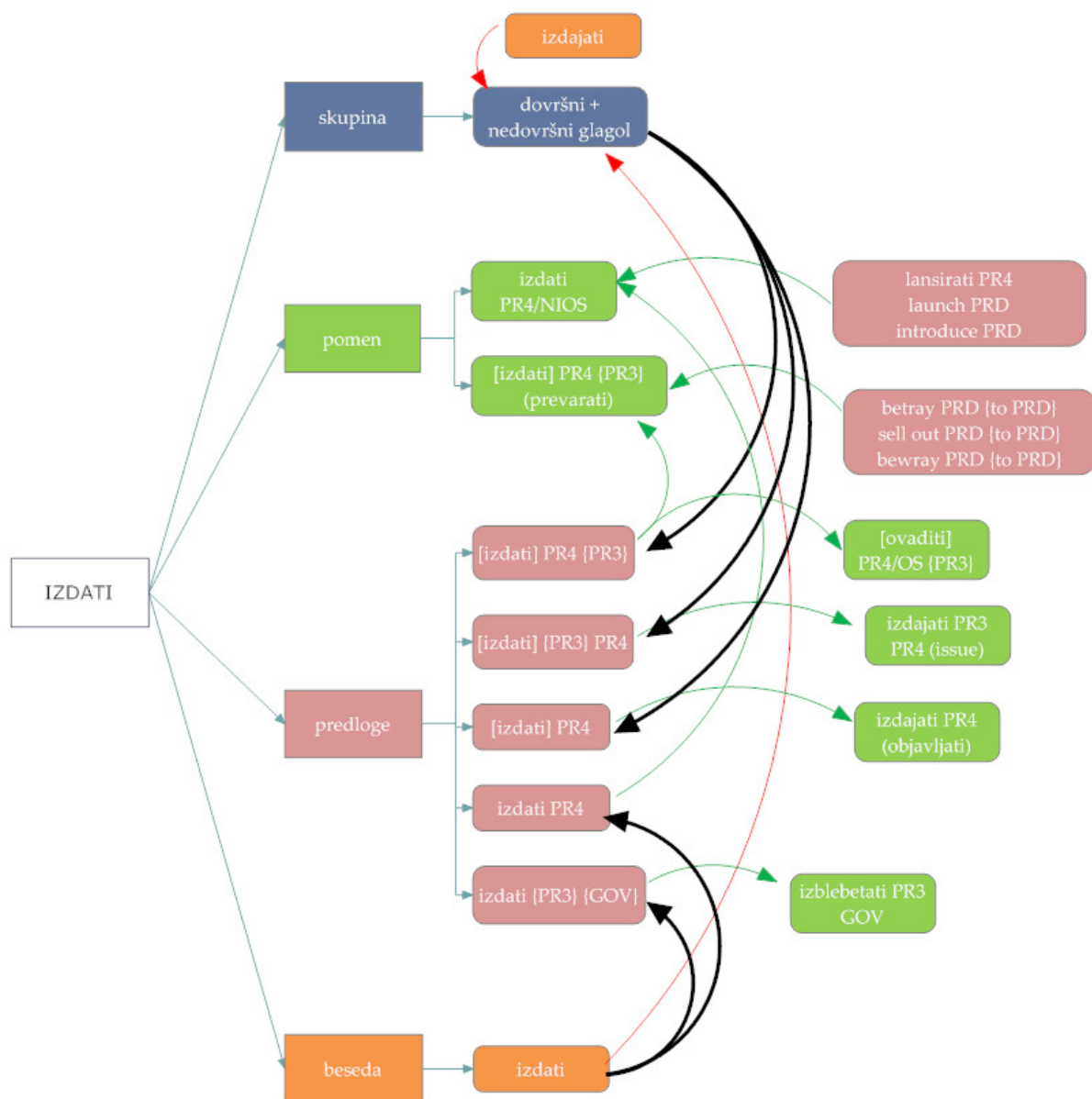
Tabela 10: ASES – izdati: Pomenski odnosi.

Tabela 10 prinaša za oba pomena nabor odnosov vrste GP2+. S to oznako so označeni odnosi med glagoli ter samostalniki, ki se tipično pojavljajo za obravnavano glagolsko iztočnico.⁵¹ Za pomen *izdati PR4/NIOS* je zabeleženo, da se povezuje s samostalnikom, ki ne izkazuje podspola človeškosti. Pri pomenu *[izdati] PR4 {PR3} (prevarati)* pa je zabeležena povezava s samostalniki, ki izkazujejo človeškost, oz. s samostalniki *domovina*, *država* ter *narod*.

⁵¹ Na tem mestu se podatki v leksikalni zbirki ASES približujejo koligacijsko-kolokacijskim, vendar le s posameznimi primeri.

Spodnja slika prinaša ponazoritev odnosov med iztočnicami. Poleg modre za *skupine*, oranžne za *besede* ter zelene za *pomene* slika prinaša rožnato barvo za *glagolske predloge*. Rdeča barva puščic ostaja za označevanje uvrščanja besed v skupine, zelena za povezovanje pomenov z oblikami oz. v tem primeru s predlogami, ki so vmesni člen pri povezovanju pomenov ter oblik pri glagolih. Nova je črna barva puščic, s katero so označene povezave med besedami oz. skupinami ter predlogami. Besedne oblike so torej prek predloge (črne puščice) povezane s pomeni (zelene puščice).

V skupino je kot rečeno poleg dovršnega *izdati* uvrščena še nedovršna varianta *izdajati*. Beseda *izdati* je povezana z dvema predlogama, *izdati PR4* (ta je nadalje vezana na pomen *izdati PR4/NIOS*) ter *izdati {PR3} {GOV}* (ki je vezana na pomen *izblebetati PR3 GOV*). Skupina je povezana s preostalimi tremi predlogami: od teh je *[izdati] PR4* nadalje vezana na pomen *izdajati PR (objavljati)*, predloga *[izdati] {PR3} [PR4]* je vezana na pomen *izdajati PR3 PR4 (issue)*, predloga *[izdati] PR4 {PR3}* pa je vezana na dva pomena, *[ovaditi] PR4/OS {PR3}* ter *izdajati PR3 PR4 (issue)* ter *[izdati] PR4 {PR3} (prevarati)*. Ta pomen je povezan še z angleškimi predlogami *betray PRD {to PRD}*, *sell out PRD {to PRD}* ter *bewray PRD {to PRD}*. Pomen *izdati PR4/NIOS* pa je povezan še s predlogami *lansirati PR4* ter angleškima *launch PRD* in *introduce PRD*.



Slika 11: ASES – izdati – odnosi med iztočnicami.

3 Nadgradnja zbirke na osnovi korpusnih podatkov

Kljub temu da je leksikalna zbirka ASES primarno zasnovana na enobesednih iztočnicah, se v njeni organiziranosti na različnih mestih kaže želja po vključevanju kompleksnejših leksikalnih informacij: od nabora stalnih besednih zvez, v katerih se iztočnica pojavlja, do informacij o glagolski vezljivosti, različnih vrstah odnosov med besedami (pomenskih, besedotvornih, izvirajočih iz védenja o svetu), o tipičnem ter netipičnem sopojavljanju besed itd. Obenem je iz primerov, predstavljenih v prejšnjem poglavju, razvidno, da so podatki naštetih vrst na različnih mestih zbirke različno zastopani, kar se lahko odraža v neuravnoteženi vsebini leksikalne zbirke.⁵²

Glede na zapisano je smiselni razmislek o posodobitvi leksikalne zbirke na način, ki upošteva v poglavju I-3 izpostavljene korpusnojezikoslovne ugotovitve, obenem pa seveda potrebe obdelave naravnega jezika. V nadaljevanju poglavja so predstavljena izhodišča nadgradnje, pri čemer je pozornost posvečena naslednjim vprašanjem: (I) izhajanje iz korpusnih podatkov, (II) upoštevanje enopomenskosti leksikalne enote ter (III) odločitev za polavtomatsko metodo nadgradnje.

Doktorska raziskava se v empiričnem delu posveča enemu od segmentov nadgradnje, tj. luščenju večdelnih besednih nizov iz besedilnega korpusa, z željo dopolnitve enobesednih iztočnic zbirke z ustreznimi kolokacijsko-koligacijskimi informacijami (za opredelitev metode luščenja podatkov glej II-2.2, za opredelitev ciljev in opis poteka raziskave pa poglavje IV).

5.1 Izhajanje iz korpusnih jezikovnih podatkov

Za osnovo leksikalne zbirke, ki naj bi služila večnamenski avtomatski obdelavi slovenščine, je najbolj smiselno uporabiti **referenčni korpus** slovenskega jezika, ki kot uravnoteženi vzorec jezika vsakdanje rabe prinaša podatke o splošni jezikovni rabi. Predvideno je, da se ti podatki v nadaljevanju gradnje dopolnjujejo oz. nadgrajujejo s podatki iz specializiranih korpusov ter drugih uporabnih virov⁵³: z zajemom čim širšega nabora podatkov o rabi leksike je omogočena kvalitetnejša reprezentacija posamezne iztočnice. Pri snovanju zbirke je ključno načrtovanje sopostavljanja ter posledično **sovplivanja jezikovnih podatkov** iz različnih virov: podatkov, pridobljenih iz vzorca jezika vsakdanje rabe ter vzorca jezika v specializirani rabi, ne gre razmejevati, pač pa je potrebno informacije iz različnih virov ustrezno integrirati v zbirko na način, da skupaj prinašajo popolnejšo sliko jezikovne rabe.

S tega stališča je bistvenega pomena **ohranjanje informacije o jezikovnem viru**, od koder je bil podatek pridobljen, v leksikalni zbirki. Na ta način je obenem omogočeno ločeno urejanje in **prioretiziranje podatkov** za potrebe razvoja specifične jezikovne tehnologije. Pri razvoju avtomatskega prevajanja jezika je npr. mogoča predpostavka, da je na vhodu programa jezikovno nespecializirano besedilo, kar pomeni pripis višje prioritete leksikalnih podatkov iz referenčnega korpusa (slednje se odraža npr. v prilagojenem vrstnem redu prevodnih različic, ki jih prevajalnik po obdelavi besedila ponudi). Na drugi strani razvoj npr. slovnicega pregledovalnika predvideva večji poudarek na leksikalnih podatkih, pridobljenih iz jezikovnonormativnih priročnikov ipd.

Prvi korak pri nadgradnji zbirke predstavlja dopolnitev obstoječih iztočnic⁵⁴ s **podatki o pogostnosti** v referenčnem korpusu FidaPLUS in glede na to nadaljnja obdelava iztočnic, ki se v korpusu izkazujejo za

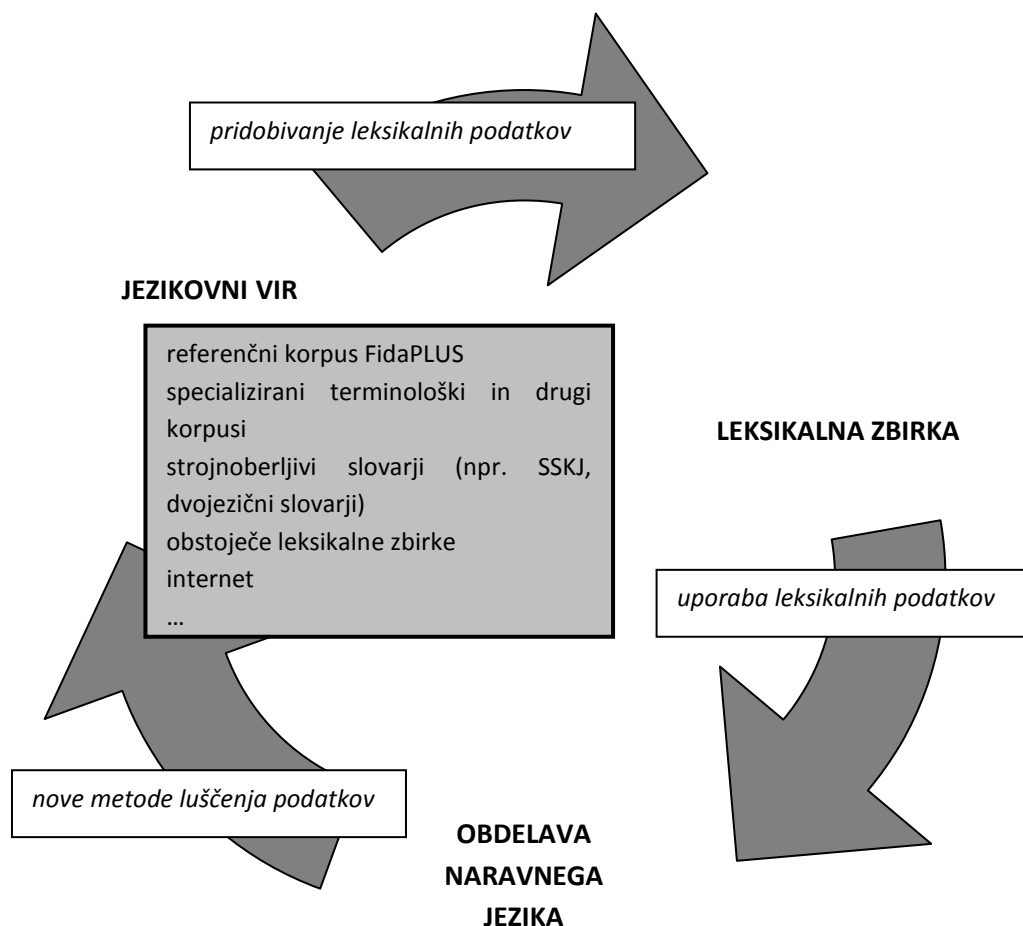
⁵² Za predkorpusne leksikalne zbirke je značilno, da so namensko dopolnjevale glede na specifične razvojne cilje, kar je glede na čas, ki ga zahteva ročni vnos podatkov, seveda smiselno.

⁵³ Za razvoj specializiranih računalniških aplikacij je običajno najprimerneje uporabljati posebej za razvojno nalogo grajene korpusne vire, ni pa to vedno v praksi izvedljivo.

⁵⁴ Pogostnost je iztočnicam vrste *beseda* (glej III-2.1.1) lahko avtomatsko pripisana na osnovi pogostnostnega seznama besed v referenčnem korpusu. Pogostnostno opredeljevanje drugih vrst iztočnic zbirke ASES je seveda bistveno zahtevnejša naloga.

najpogostejše.⁵⁵ V prvih fazah nadgradnje je tako večina pozornosti usmerjene k sodobni slovenščini vsakdanje rabe, v nadaljevanju pa je predvideno dopolnjevanje podatkov glede na trenutne razvojne potrebe, pri čemer pridejo v poštev specializirani korpusi in tudi nekorpusni jezikovni viri.⁵⁶

Ključnega pomena je, da ostane nadgrajevanje zbirke v središču interesa jezikovnotehnološke razvojne skupnosti in je kot tako ena izmed razvojnih priorit. Razvojni krog, ki je predviden, vključuje tri elemente: (I) jezikovni vir, iz katerega so pridobljeni leksikalni podatki, ti so vključeni v (II) leksikalno zbirko, kar omogoča (III) razvoj obdelave naravnega jezika, ki z novimi metodami za ekstrakcijo podatkov pripomore k izrabi jezikovnih virov na vedno višjih nivojih.⁵⁷



Slika 12: Razvojni krog nadgrajevanja leksikalne zbirke.

⁵⁵ Za urejevalce zbirke koristna bi bila nadgradnja vmesnika zbirke, da bi omogočal enostaven dostop do najpogostnejših iztočnic – npr. možnost urejanja iztočnic glede na pogostnost v izbranem izhodiščnem jezikovnem viru, obenem pa ni odveč razmislek o načinu grafičnega opredeljevanja iztočnic glede na pogostnost.

⁵⁶ Marginalizacija netipičnega v jeziku je pri obdelavi naravnega jezika lahko velika prednost v izhodiščnih korakih dopolnjevanja leksikalne zbirke, ne more pa ostati edino vodilo gradnje.

⁵⁷ V literaturi pogosto zasledimo Leechevo različico razvojnega kroga, ki predstavlja odnos med leksikalno zbirko ter korpusom (glej npr. Ooi 1998: 68). Ooi predstavi tudi svojo, nekoliko kompleksnejšo različico (ibid.: 72). Različica, predstavljena na tem mestu, je prilagojena za slovensko situacijo.

5.2 Upoštevanje enopomenskosti leksikalne enote

Vključevanje semantičnih podatkov v obdelavo naravnega jezika je problem, ki se ga raziskovalci lotevajo na tak ali drugačen način, kljub številnim poskusom pa celostna avtomatska prepoznavna pomena še vedno ostaja nedosežen cilj. Kljub temu da se v jezikoslovju že desetletja pojavljajo ideje o premiku pozornosti od enobesedne k večbesednim enotam⁵⁸, ostaja obdelava naravnega jezika še vedno v veliki meri vezana na enobesedno izhodišče, ki je temelj obravnave večbesednih ter znotraj tega frazeoloških enot, avtomatskemu razdvoumljanju besednega pomena⁵⁹ itd.

V poglavju I-3 je bilo govora o enopomenskosti jezikovnega vzorca, ki združuje jedrno besedo s podatki o njenem tipičnem ubesediljenju, kakor se izkazuje v izbranem korpusnem viru ob uporabi izbrane statistične metode: koligacijski podatki povedo, katere strukture se v povezavi z jedrno besedo statistično relevantno pojavljajo, kolokacijski podatki pa, katere so statistično relevantne leksikalne zapolnitve teh skladenjskih struktur. Kontekstualne informacije, ki jih prinaša jezikovni vzorec, so po ugotovitvah korpusne leksikografije v veliki večini dovoljšen indikator za razdvoumljanje pomena, kakor ga razume obdelava naravnega jezika – ali če pogledamo na problem z drugega zornega kota: problem »dvoumnosti besed« je posledica napačnega prepričanja, da je samostojna beseda običajno nosilka pomena. Nekateri besede so pomenske enote, mnoge niso – pri čemer je pomenska enota beseda skupaj z vsemi besedami besedilnega konteksta, ki so potrebne za razdvoumljanje te besede (Teubert v Halliday et al. 2004: 83).

Temeljni za leksikalno zbirko, kakršno predstavljamo na tem mestu, so **koligacijski ter kolokacijski podatki o besedah**, kar pomeni celovit premik pozornosti od enobesedne k večbesednim leksikalnim enotam. Z organizacijskega vidika leksikalne zbirke se kaže za smotrno ohranjanje enobesedne iztočnice kot izhodiščne enote, vendar le na ravni forme, ne pa tudi pomena. V praksi to pomeni, da posamezne besede niso kvalificirane, povezane s prevodnimi ustreznici, pomensko členjene ipd., pač pa je besedna oblika zgolj vodilka do kompleksnejših podatkov, ki so ustrezno metajezikovno obdelani.

Vprašanju načina vključitve koligacijsko-kolokacijskih podatkov v zbirko se na tem mestu ne posvečamo, mogoče so raznovrstne rešitve, končna odločitev pa temelji predvsem na predvideni uporabi podatkov za razvoj posameznih produktov. Kot je razvidno iz povedanega, je ključen razmislek o vključevanju tako skladenjskih podatkov, denimo skladenjskih vzorcev, v katerih se iztočnica pojavlja, kot tudi nabora kolokatorjev, ki z iztočnico v vzorcih nastopajo; na vseh mestih, kjer je to mogoče, pa je smiselno vključiti tudi podatek o pogostnosti v korpusu (torej tako na ravni vzorcev kot kolokatorjev).

5.3 Odločitev za polavtomatsko metodo nadgradnje

Na področju obdelave naravnega jezika je odločitev za avtomatske metode pridobivanja ter obdelave podatkov prva izbira, saj je ročna obdelava podatkov draga ter zamudna. Človeški doprinos h gradnji jezikovnih virov je tipično predviden le na mestih, kjer avtomatske metode ne dosegajo potrebne kvalitete oz. je poseg

⁵⁸ Pri nas o večbesednih enotah kot nosilkah pomena piše npr. Vidovič Muha pri uvrščanju stalnih besednih zvez med lekseme: Leksem kot poimenovalna enota jezika je širši od besede, saj zajema tudi stalne besedne zveze, npr. vrstne (generične) pojme, kot so *beli medved*, *rdeči bor*, *češko pivo* ipd.« (Vidovič Muha 2000: 17). Korpusna leksikografija gre s kolokacijskostjo še korak naprej, kakor piše denimo Gantar: »S tem ko je korpus omogočil prepoznavno sopojavljanje posameznih besed, je hkrati tudi prevrednotil pojme kot so leksikalna in skladenjska enota, kar je v prenosu na leksikografsko prakso z upoštevanjem dejstva, da je težišče leksikografije na pripisovanju pomena jezikovnemu znaku, ki se manifestira vedno in samo v besedilu, pomenilo enakovredno obravnavo besedne in besednozvezne problematike ter vključitev SBZ in očitnih skladenjskih vzorcev tudi v splošne in ne zgolj specializirane slovarje.« (Gantar 2006: 158)

⁵⁹ Prim. denimo Sinclairjevo premiso, da je dvoumnost v besedilu posledica metode opazovanja in ne same besedilne strukture (Sinclair 1998: 7).

strokovnjaka neizbežen (kot je bilo predstavljeno v poglavju II-1.2.4 denimo pregledovanje ter popravljanje avtomatsko obdelanih podatkov za določitev uspešnosti avtomatskega sistema). Na drugi strani je bilo v poglavju I-2 izpostavljeno stališče korpusnega jezikoslovja, ki zahteva po kvantitativnih metodah procesiranja podatkov tudi kvalitativne – avtomatizacija metode torej ni vrednota sama na sebi, ampak le predpriprava jezikoslovni analizi.

Pri gradnji leksikalnih zbirk je smotrno razmišljati o **polavtomatskih metodah gradnje**. O ključnosti sodelovanja jezikoslovcev pri razvoju jezikovnih tehnologij ni dvoma: jezikoslovno védenje je nepogrešljivo v fazi snovanja ter izbire metod gradnje, tekom gradnje pa ne gre brez sodelovanja usposobljenih leksikografov ter drugih profilov strokovnjakov, glede na konkretne razvojne potrebe (npr. prevajalcev v povezavi z razvojem strojnega prevajanja ipd.). Samo pridobivanje leksikalnih podatkov iz korpusov, njihova priprava, organizacija itd. preden so izročeni v obdelavo graditelju zbirke, pa mora biti v čim večji meri avtomatizirana. S stališča zmanjševanja časa ter stroškov človeške interpretacije ter organizacije jezikovnih podatkov je smotrno razmišljati o razvoju **uporabniku prijaznega programskega orodja**, ki bi omogočalo enostavno obdelavo avtomatsko pridobljenih podatkov (selekcioniranje, sortiranje, klasificiranje, organiziranje v medsebojne odnose itd.) ter njihov uvoz v leksikalno zbirko.

Na drugi strani je smiselno poskrbeti za pregledno ter za človeškega uporabnika **logično strukturirano reprezentacijo podatkov** v zbirki sami, saj slednje omogoča boljše raziskovalno oz. razvojno izhodišče za izrabo leksikalnih podatkov. Gradnja jezikovnotehnološke leksikalne zbirke je sicer usmerjena v računalniško rabo, kar pomeni, da na vseh nivojih gradnje vključujemo le informacije, ki so relevantne za avtomatsko obdelavo jezika, človeškega uporabnika pa upoštevamo le v smislu graditelja zbirke oz. razvijalca metod ali produktov, ki na zbirki temeljijo.⁶⁰

⁶⁰ Primer je denimo vključevanje pomenskih opredelitev oz. definicij ali zgledov rabe le na mestih, kjer je to potrebno za graditelja zbirke – pri čemer je seveda jasno, da so informacije tega tipa toliko bolj relevantne za gradnjo slovarskega tipa leksikalnih zbirk (glej III-1).

IV

RAZISKOVALNI CILJI IN POSTOPKI

Poglavje prinaša opis doktorske raziskave, začenši z opredelitvijo namena raziskave, ki prinaša raziskovalno izhodišče, razlago ključnih pojmov ter shemo raziskovalnih ciljev in vprašanj. Drugi del poglavja je posvečen opisu poteka, predvsem pripravi in obdelavi podkorpusa ter pripravi in prvi obdelavi seznamov vzorcev. Naslednji koraki raziskave bodo natančneje predstavljeni v poglavju V in deloma VI, zato so na tem mestu opisani le na kratko. Pričujoče poglavje zaključuje predstavitev programskih skript, ki so bile znotraj doktorske raziskave pripravljene za izvedbo posameznih raziskovalnih nalog.

1 Namen raziskave

1.1 Raziskovalno izhodišče

Raziskovalno izhodišče doktorske naloge predstavlja korpusnojezikoslovna opredelitev **sopojavitvenih odnosov** kolokacije ter koligacije kot odnosa med jedrno, izhodiščno enobesedno enoto in besedami ter slovničnimi kategorijami, ki glede na izbrano statistično metodo opredeljujejo tipični kontekst jedrne besede v rabi (I-3). V povezavi z omenjenim je ključna opredelitev **pomenske enote** kot besede z vsemi elementi konteksta, ki so potrebni za razdvajanje te besede (III-3.2).

Premik od obravnave enobesedne iztočnice k večbesednim leksikalnim enotam skuša doktorska naloga omogočiti v povezavi z **nadgradnjo leksikalne zbirke ASES** na osnovi korpusnih podatkov (III-2). Trenutno najrelevantnejša slovenska jezikovnotehnološka leksikalna zbirka že omogoča vključevanje zahtevnejših vrst leksikalnih podatkov (besednozveznih, vezljivostnih, pomenskoodnosnih itd.), vendar slednje spričo dolgotrajnosti ročnega vnosa v zbirki niso uravnoteženo zastopane (III-3).

Doktorska naloga skuša zapolniti vrzel na področju vključevanja iz korpusa pridobljenih kolokacijsko-koligacijskih podatkov v leksikalno zbirko, pri čemer je predvidena polavtomatska metoda (III-3.3): avtomatsko luščenje večdelnih besednih nizov za obravnavano iztočnico, čemur sledi ustrezna interpretacija ter ročna organizacija podatkov. Doktorska naloga se osredotoča na prvi, avtomatski del metode, v zvezi s čimer je preizkušena metoda **luščenja leksikalnih podatkov iz lematiziranega ter oblikoskladenjsko označenega korpusa** (II-2.2). V raziskavi je uporabljen podkorpus referenčnega korpusa FidaPLUS, označen z naborom oznak JOS (II-1.2).

Luščenje večdelnih besednih nizov, preizkušeno v pričujočem delu, temelji na **zaporednosti besed** ter **pogostnosti niza** v korpusu, pri čemer je pozornost posvečena predvsem vzorcem, ki jih z upoštevanjem oblikoskladenjskih oznak iz korpusnih besedil lahko pridobimo: tako s stališča razvrščanja primerov, ki jih posamezen vzorec prinaša, glede na tip, kot tudi s stališča ustreznosti primerom pripisanih oblikoskladenjskih oznak (II-2.2).

1.2 Ključni pojmi

Središče interesa doktorske raziskave predstavlja **vzorec**: iz besedilnega korpusa pridobljena kombinacija izbrane leme (avtomatsko pripisane osnovne oblike besede) in obkrožajočih oblikoskladenjskih oznak: dvodelni vzorec (bigram) prinaša lemo ter oblikoskladenjsko oznako na levi oz. desni od leme, tridelni (trigram) lemo v kombinaciji z dvema oznakama. Vzorec dopolnjuje informacija o pogostnosti v korpusnem viru: *KAG PAJEK (88)*

denimo pomeni, da se lema *pajek*, pred katero se pojavlja oblikoskladenjska oznaka za glavni števniki arabskega zapisa, v korpusu pojavlja 88-krat.

Vzorec kot kombinacija avtomatsko pripisanih oznak ne predstavlja realne podobe jezikovnih podatkov v korpusu, ampak le **približek**, ki je lahko zanesljiv le toliko, kolikor je zanesljivo avtomatsko označevanje (glej II-1.2.4). Povedano drugače: namesto primerom realne jezikovne rabe, ki jih prinaša besedilni korpus, je pri opisani metodi luščenja interes posvečen abstrahirani podobi jezika, kakor se izraža na ravni v korpusu pripisanih oblikoskladenjskih oznak.

Del besedila, ki ga z uporabo vzorca iz korpusa avtomatsko izluščimo, imenujemo v pričujočem delu **zapolnitev**. Kot pri samem vzorcu je tudi na ravni zapolnitve ključna informacija o pogostnosti. Za zgoraj navedeni vzorec *KAG PAJEK* denimo izluščimo iz korpusa seznam 39-ih različnih zapolnitev, od katerih je najpogostnejša *14 pajkov (11)*. Poimenovanje »zapolnitev« nakazuje primarnost vzorca napram zapolnitvi, kar je na prvi pogled v nasprotju z očitnim dejstvom, da so v procesu gradnje jezikovnega vira jezikovni podatki kronološko vedno pred oznakami. Potreben je poudarek, da podatkov, ki jih iz korpusa pridobivamo z luščenjem, ne gre povsem enačiti s podatki v izhodiščnih besedilih: med obdelavo so delci korpusnega besedila iztrgani iz besedilnega konteksta in nadalje obdelani ter organizirani na način, odvisen od uporabljene metode.⁶¹

Ker vzorci prinašajo oblikoskladenjske oznake, ki so glede prinašajočih kategorij in vrednosti precej bogate (glej Priloga 1), je v izogib razpršenosti izluščenih podatkov predvideno združevanje vzorcev v **vzorčne tipe**. Združevanje vzorcev in posledično uporaba posplošenih kategorij za luščenje se odraža tudi na ravni kompleksnosti obdelave podatkov, kjer postane ključno vprašanje, katere kategorije in vrednosti oblikoskladenjskih oznak upoštevati pri luščenju ter končnem prikazu podatkov. Na tej ravni namesto o zapolnitvah govorimo o **izluščenih besednih nizih**, v primeru, da izluščeni besedni niz ustreza jezikoslovni definiciji, pa tudi o **izluščenih besednih zvezah**.

Besedni niz v pričujočem delu pomeni katero koli (zaporedno) sopojavitev dveh ali več besed, pri čemer so – kot piše Gantar – v jeziku najpogostejše kombinacije med slovničnimi ter slovničnimi in predmetnopomenskimi besedami, v slovenščini npr. *se je, da je, ki je, ki ga je, ki se je, pa se je, za okolje in prostor, ne glede na to* itd. (Gantar 2007: 41). Besedne zveze razumemo kot podmnožico besednega niza, pri čemer je ponovno ključna zaporednost v besedilu⁶² in obenem izražanje skladenjske ter pomenske povezanosti besednozveznega jedra (ki je predmetnopomenska beseda) ter določil, npr. *pajek za seno, pajek v mreži, pajek tarantela, odvoz s pajkom* itd. Kolokacija kot tretji primerljivi pojem vključuje statistično merjenje povezanosti med jedrno besedo ter kolokatorjem, ki sam na sebi na jedro ni skladenjsko vezan (glej I-3), kolokatorji za *pajek* so denimo [*SIP, vreteno, lisice, odvoz, kosilnica, zgrabljajnik, prodati*] itd.

Rezultate luščenja, s tem pa tudi izhodiščne vzorce oz. vzorčne tipe, v raziskavi opredeljujemo s stališča **relevantnosti za nadaljnjo uporabo**, tj. uvrstitev v leksikalno zbirko. Pri določanju relevantnosti se upošteva: (I) pogostnost v korpusu, (II) kvaliteta označenosti besed ter (III) zadovoljiv priklic⁶³ izluščenih enot, na ravni izluščenih večbesednih nizov pa predvsem (IV) skladenjska in pomenska povezanost elementov niza (glej II-2.2). Naštete točke so namenjene opredeljevanju relevantnosti v pričujoči raziskavi, končna odločitev glede uporabnosti določene vrste besednega niza pa je prepuščena graditeljem leksikalne zbirke.

⁶¹ S tega stališča je pojem zapolnitve še najbolj podoben »zadetku« pri uporabi korpusnega konkordančnika: podobnost je na ravni pridobivanja oz. prikaza korpusnih podatkov, ki ustrezajo določenemu pogoju, pri čemer pa je že opravljen določen del interpretacije ter organizacije podatkov.

⁶² Dva glavna problema, vezana na luščenje zaporednih besednih nizov, sta **gnezdenje ter prekinjenost zvez** (Sinclair 1998: 10; Vintar 2008: 102).

⁶³ **Priklic** izluščenih enot pomeni delež **pridobljenih** podatkov glede na **vse** relevantne podatke, ki so v viru na voljo. Vprašanje priklica je vezano predvsem na luščenje glagolskih besednih zvez, saj so slednje, kot se potrjuje v nadaljevanju raziskave, težje zajemljive v sklopu dvo- ter tridelnih vzorcev.

1.3 Raziskovalni cilji in vprašanja

V nadaljevanju so glede na namen raziskave opredeljeni glavni raziskovalni cilji in v njihovem okviru temeljna raziskovalna vprašanja.

1. Priprava podatkovnega vira

1.1 Priprava podkorpusa referenčnega korpusa FidaPLUS, ki ustreza naslednjima pogojema:

- 1.1.1 kvantitativna obvladljivost z razpoložljivo strojno in programsko opremo;
- 1.1.2 oblikoskladenjska označenost z naborom oznak JOS.

1.2 Priprava različice podkorpusa za izdelavo pogostnostih seznamov vzorcev.

2. Analiza vzorcev/vzorčnih tipov in njihovih zapolnitev/izluščenih besednih nizov za izbrane leme

2.1 Izdelava pogostnostih seznamov vzorcev za izbrane leme.

2.2 Luščenje zapolnitev najpogostnejših vzorcev za izbrane leme in analiza podatkov, pri čemer so v ospredju naslednja raziskovalna vprašanja:

- 2.2.1 kateri od vzorcev prinašajo za luščenje relevantne podatke in kateri ne;
- 2.2.2 ali so možne posplošitve na ravni ne/relevantnosti glede na vsebnost ali mesto določene oblikoskladenjske oznake znotraj vzorcev;
- 2.2.3 na kakšen način je glede na vsebnost oz. mesto oblikoskladenjskih oznak znotraj vzorcev možno združevanje vzorcev v skupine (vzorčne tipe).

2.3 Luščenje zapolnitev vzorčnih tipov za izbrane leme in analiza podatkov, pri čemer so v ospredju naslednja raziskovalna vprašanja:

- 2.3.1 kateri od vzorčnih tipov prinašajo za luščenje relevantne besedne zveze in kateri ne;
- 2.3.2 ali so možne posplošitve na ravni ne/relevantnosti glede na vsebnost ali mesto določene oblikoskladenjske oznake znotraj vzorčnih tipov;
- 2.3.3 katera mesta oblikoskladenjskih oznak prinašajo informacije, ki jih je potrebno upoštevati pri luščenju podatkov, in katera ne;
- 2.3.4 katera mesta oblikoskladenjskih oznak prinašajo informacije, ki jih je potrebno upoštevati pri prikazu podatkov, in katera ne.

2.4 Organizacija vzorčnih tipov glede na izpričano ne/relevantnost za nadaljnjo uporabo.

2.5 Prenos izbranega dela ugotovitev na nove primere (leme), pri čemer so v ospredju naslednja raziskovalna vprašanja:

- 2.5.1 ali se pri obravnavi vzorcev z novo lemo pojavljajo v raziskavi še neidentificirani vzorci/vzorčni tipi in če da, na kakšen način jih je mogoče uvrstiti med obstoječe podatke;
- 2.5.2 ali izluščeni podatki za novo lemo potrjujejo opredelitev relevantnosti in če ne, do kakšnih

sprememb prihaja.
<p>3. Analiza avtomatskega označevanja</p> <p>3.1 V primeru, da obravnavani podatki na ravni oblikoskladenjskih oznak izkazujejo označevalne napake, analiza označevanja, pri čemer so v ospredju naslednja raziskovalna vprašanja:</p> <p>3.1.1 ali je mogoče označevalne napake natančneje opredeliti glede na raven, na kateri se pojavljajo (na besednovrstni ravni, na ravni posamezne kategorije);</p> <p>3.1.2 ali je mogoče označevalnim napakam s pomočjo obravnavanih podatkov (izluščeni besedni nizi, korpusne konkordance) najti vzrok;</p> <p>3.1.3 ali je mogoče najti način preprečevanja pojavljanja označevalnih napake v bodoče, in katere informacije so za to potrebne.</p>
<p>4. Priprava programskih skript za avtomatizacijo raziskovalnih postopkov</p> <p>4.1 Priprava programskih skript za pripravo podatkovnega vira.</p> <p>4.2 Priprava programskih skript za luščenje ter urejanje vzorcev, vzorčnih tipov ter njihovih zapolnitev/ besednih nizov.</p>
<p>5. Evalvacija raziskovalnih postopkov in orodij</p> <p>5.1 Evalvacija v raziskavi uporabljenih postopkov in orodij z izpostavitvijo možnosti nadaljnjih izboljšav.</p>

Tabela 11: Raziskovalni cilji ter vprašanja.

Nadaljevanje poglavja prinaša opis poteka doktorske raziskave.

2 Potek raziskave

2.1 Priprava podatkovnih virov

2.1.1 Izdelava podkorpusa

Korpus, uporabljen v raziskavi, je podkorpus trenutnega referenčnega korpusa za slovenščino, tj. korpusa FidaPLUS (oz. njegove različice korpusa Fida+X, o čemer več v nadaljevanju).⁶⁴ Priprava podkorpusa temelji na potrebi po kvantitativni omejitvi uporabljenega jezikovnega vira⁶⁵, ki je vezana na zmogljivost razpoložljive programske ter strojne opreme, predvidene za obdelavo korpusnih podatkov. Izdelavi podkorpusa botruje tudi želja po delu s korpusnimi besedili, posodobljenimi na ravni oblikoskladenjskih oznak, tj. označenimi z naborom oznak JOS (glej II-1.2.3).

V podkorpus so bili uvrščeni vsi odstavki korpusa FidaPLUS, ki prinašajo (vsaj) eno od 15-ih za analizo izbranih besed (več o tem v nadaljevanju). Kot rečeno v II-2.2, se v raziskavi osredotočamo na luščenje podatkov za eno samo izhodiščno besedo (oz. besedni obliki pripisano lemo) naenkrat – kar pomeni odločitev za pripravo korpusnega vira, ki prinaša čim večjo količino besedilnega konteksta za obravnavano iztočnico. Ker na opisani

⁶⁴ Podkorpus je pripravil dr. Tomaž Erjavec, za kar mu gre na tem mestu iskrena zahvala.

⁶⁵ Referenčni korpus FidaPLUS prinaša okrog 621.150.000 besed.

način pripravljen podkorporus na ravni obravnave izbranih besed ohranja reprezentativnost izvirnega korpusa, odpade vprašanje uravnoteževanja podkorporusa s stališča zajetih besedilnih zvrsti oz. virov.

Pred pripravo podkorporusa je torej potrebna izbira nabora besed, ki bodo v nadaljevanju raziskave izhodišče za analizo. Nabor izbranih primerov je naključen in ne skuša biti na nikakršen način reprezentativen, kar s stališča raziskave ni problematično, saj nas nabor zanima predvsem s stališča leksikalnih podatkov, ki jih s testirano metodo lahko iz podkorporusa izluščimo. Posledično je glavni kriterij pri izbiri iztočnic kvantitativna obvladljivost nabora pojavitev, tj. število korpusnih zadetkov do 100.000, najbolje nekje med 5.000 ter 25.000.

Drugi pogoj je zastopanost vsake od štirih polnopomenskih besednih vrst z vsaj enim primerom, saj je v raziskavi predvidena analiza po enega primera za vsako od teh besednovrstnih skupin. Kot omenjeno v II-2.2, je predvidena večja pozornost na ravni luščenja podatkov za samostalniške iztočnice, zato je seznam samostalnikov v gornji tabeli nekoliko daljši od ostalih. Ker se na ravni oblikoskladenjskega označevanja samostalniki mdr. ločujejo tudi z oznakama *lastno* oz. *občno ime*, je bil v nabor ter analizo zajet tudi lastnoimenski samostalnik.

SAMOSTALNIK	GLAGOL	PRIDEVNIK	PRISLOV
pajek	plesati	strasten	temeljito
Slovenka	izdati	moder	
oven	izpeljati		
debata	sklanjati		
vonj	izsiliti		
matica			
lipicanec			

Tabela 12: Izhodiščne besede za gradnjo podkorporusa.

Kot bo vidno v nadaljevanju, so v analizo v prvem koraku zajete besede *pajek*, *plesati*, *strasten* ter *temeljito*, v drugem koraku pa še *Slovenka*, *oven* ter *debata*.⁶⁶

Za izdelavo podkorporusa je uporabljena različica korpusa FidaPLUS v formatu xml, t. i. Fida+X, kar pomeni določene razlike v primerjavi z različico, ki je na internetu na voljo v Konkordančniku ASP32. O pripravi korpusa Fida+X, ki je uporabljen tudi za pripravo korpusov v sklopu projekta JOS (glej II-1.2.3), pišeta Erjavec in Krek:

»Prvi korak na poti od korpusa FidaPLUS do korpusa JOS je bila pretvorba v format XML po priporočilih TEI P4 [...], da bi s tem ohranili standardni format in omogočili uporabo orodij za delo s formatom XML, predvsem XSLT. Format TEI P4 je sicer povratno združljiv s formatom TEI P3 korpusa FIDA in XML je podmnžica formata SGML, vendar končni korpus FidaPLUS kot podatkovna zbirka ni v celoti skladen niti s formatom SGML niti s specifikacijami MULTEXT-East, zato je bil proces pretvorbe zahtevnejši, kot je bilo pričakovati. Procesiranje je bilo izvedeno s pomočjo niza skript v programskem jeziku Perl, končni pretvorjeni korpus FidaPLUS pa imenujemo Fida+X. Ta je za malenkost manjši kot izvirni korpus, ker smo izpustili besedila, pri katerih hevrstični postopki s pomočjo Perl skript niso zadostovali za njihovo kompatibilnost s standardom TEI P4.« (Erjavec in Krek 2008a: 50)

Prvi korak priprave podkorporusa je zajemal luščenje ustreznih odstavkov iz korpusa Fida+X, drugi pa preoznačevanje besedil z oblikoskladenjskimi oznakami, posodobljenimi v okviru že večkrat omenjenega projekta JOS.⁶⁷

⁶⁶ Torej ne vse besede s seznama: pri gradnji podkorporusa je bilo za primer zajetih več besed, če bi se med potekom raziskave pokazala potreba po menjavi primera ali analizi dodatnih primerov.

Rezultat priprave podkorpusa je 22.563 datotek formata xml v skupni velikosti 1,21 GB. Datoteke prinašajo izluščene odstavke korpusnih besedil, skupaj z informacijo o besedilnem viru v obliki glav korpusnih dokumentov. Glave so navedene ena za drugo na začetku dokumenta, odstavki si ustrezno sledijo v nadaljevanju datoteke. Odstavki prinašajo le informacijo o razdvoumljenem stanju korpusnih oznak (le oznake *lemma* ter *msd* – prim. tristopenjsko označeno korpusno besedilo v Tabeli 1). Sledi primer odstavka korpusnega besedila:

```
<p id="F0000001.244">
<s id="F0000001.244.1">
<w lemma="Benko" msd="Slmetd">Benka</w>
<w lemma="Pulko" msd="Slmei">Pulko</w>
<c>,</c>
<w lemma="prvi" msd="Kbvzei">prva</w>
<w lemma="Slovenka" msd="Slzei">Slovenka</w>
<c>,</c>
<w lemma="ki" msd="Vd">ki</w>
<w lemma="biti" msd="Gp-ste-n">je</w>
<w lemma="v" msd="Dm">v</w>
<w lemma="petinpolleten" msd="Ppnsem">petinpolletnem</w>
<w lemma="potovanje" msd="Sosem">potovanju</w>
<w lemma="z" msd="Do">z</w>
<w lemma="motor" msd="Someo">motorjem</w>
<w lemma="obkrožiti" msd="Ggdd-ez">obkrožila</w>
<w lemma="svet" msd="Sometn">svet</w>
<c>,</c>
<w lemma="biti" msd="Gp-pte-n">bo</w>
<w lemma="tudi" msd="L">tudi</w>
<w lemma="februarja" msd="Rnn">februarja</w>
<w lemma="nadaljevati" msd="Ggnd-ez">nadaljevala</w>
<w lemma="multimedijski" msd="Ppnzmt">multimedijske</w>
<w lemma="predstavitev" msd="Sozmt">predstavitve</w>
<w lemma="svoj" msd="Zp-zer">svoje</w>
<w lemma="pot" msd="Sozer">poti</w>
<c>.</c>
</s>
</p>
```

Tabela 13: Primer označenega besedila v podkorpusu.

Pripravljeni podkorpus zajema skupno 26.799.466 pojavnic (štete so oznake za leme), kar predstavlja glede na oceno števila besed referenčnega korpusa okrog 4,3 % izvirnega korpusnega obsega.

Kot razloženo v gornjem navedku se različici korpusa FidaPLUS ter Fida+X glede kvantitete zajetih podatkov razlikujeta⁶⁸, zato tabela v nadaljevanju prinaša primerjavo pogostnosti posamezne od 15-ih izbranih lem v

⁶⁷ Kot omenjeno v II-1.2.2, referenčni korpus, ki je na voljo na internetu, v času poteka doktorske raziskave še ni preoznačen z novim naborom oznak, ampak prinaša nabor oznak Multext-East.

⁶⁸ Pri pretvorbi korpusa so bile nekatere oznake dopolnjene oz. popravljene, nekateri fragmenti besedila pa zaradi nedosledne označenosti izpuščeni; možni razlog za precej visoke razlike so tudi morebitni zapleti na ravni štetja lem v obeh korpusih.

korpusu FidaPLUS – leme so štete s pomočjo Konkordančnika ASP32 v internetni različici korpusa – ter podkorpusa Fida+X – leme so štete s pomočjo pripravljenega programa (glej IV-3.1.1):

	FidaPLUS	podkorpus Fida+X	razlika
vonj	23.279	23.414	+135
pajek	9.122	9.222	+100
matica	6.307	6.410	+103
Slovenka	14.378**	14.526	+148
debata	9.142	9.258	+116
lipicanec	4.402	4.442	+40
oven	6.415*	6.480	+65
plesati	15.443	15.662	+219
izdati	82.142	75.907	-6.235
izpeljati	31.529	27.829	-3.700
sklanjati	1.359	1.380	+21
izsiliti	5.338	5.015	-323
strasten	6.263	6.337	+74
moder	52.169*	52.680	+511
temeljito	22.701	22.877	+176

* – upoštevano je število po izločitvi zadetkov, ki so označeni kot lastno ime

** – upoštevano je število po izločitvi zadetkov, ki so označeni kot občno ime

Tabela 14: Razlike v številu lem v korpusu FidaPLUS ter podkorpusu Fida+X.

Razlike v podatkih so za nadaljevanje raziskave pomembne predvsem v sklopu analize avtomatskega označevanja, kjer so za interpretacijo podkorpusnih podatkov uporabljeni konkordančni nizi Konkordančnika ASP32. Z upoštevanjem dejstva, da so podatki v obeh korpusih različni tudi na ravni nabora pripisanih oblikoskladenjskih oznak, odločitev predstavlja šibko mesto v metodi. V raziskavi je bila sprejeta zaradi zmogljivosti in praktičnosti Konkordančnika ASP32 za zahtevnejše oblike iskanja po korpusu⁶⁹; v prid odločitvi priča tudi dejstvo, da pripravljeni podkorpus z novimi oznakami ni bil ponovno označen, ampak so bile oznake le avtomatsko pretvorjene, kar pomeni, da morebitne napake na ravni prvotnega pripisa v korpusu FidaPLUS ostajajo tudi na ravni podkorpusa Fida+X.

2.1.2 Pretvorba podkorpusa za pripravo seznama vzorcev

Podkorpus, opisan v prejšnjem poglavju, se v raziskavi uporablja kot vir za luščenje vzorčnih zapolnitev oz. besednih nizov. Izključno za pripravo seznamov vzorcev, ki so predpogoj za prvi korak analitičnega dela raziskave, pa se uporablja na poseben način preoblikovana različica podkorpusa.

Vzorec je, kot rečeno, kombinacija obravnavane leme ter oblikoskladenjskih oznak, ki se ob njej pojavljajo. Izhodišče za pripravo seznamov vzorcev – slednji so urejeni glede na pogostnost, ki je za analizo ključen podatek – je torej besedilo, v katerem namesto obravnavane besede nastopa lema, namesto ostalih besednih oblik pa oblikoskladenjske oznake.

Takšen zapis je na strukturni ravni primerljiv z izhodiščnim besedilom: ohranjena sta linearnost zapisa ter vrstni red oznak, ohranjena so stavčna ločila ter oznake formata xml, ki določajo meje povedi ter odstavkov. Glavna ideja na tem mestu je, da je mogoče za pridobivanje podatkov iz tako pripravljenega besedila uporabiti orodja,

⁶⁹ Možna alternativa na tem mestu bi bila uporaba programskega orodja Oxford WordSmith Tools, ki v novejših različicah podpira delo z označenimi korpusnimi podatki. Več o programu v nadaljevanju.

ki so primarno namenjena obdelavi korpusnih besedil. Za pripravo seznamov vzorcev je bil v raziskavi uporabljen program Oxford WordSmith Tools (različica 4.0.0.387), čemur je posvečeno poglavje IV-2.2.1.

Ker je priprava seznamov vzorcev vezana na zmogljivosti izbranega programskega orodja, je potrebna pretvorba izhodiščnih besedil podkorpusa v obliko, ki omogoča čim enostavnejše nadaljnje delo. Vsi v nadaljevanju opisani koraki pretvorbe besedil so izvedeni z uporabo v raziskavi pripravljenih programov (za opis programov glej poglavje IV-3).

Metoda pretvorbe besedil poteka po naslednjem postopku: (I) preimenovanje oznak *lemma* pri obravnavanih 15 lemah, (II) odstranitev neželenih oznak iz besedila, (III) menjava ločil s posebnimi oznakami ter (IV) dodajanje oznak za konec povedi. V nadaljevanju so naštet koraki na kratko opisani ter predstavljeni s primeri.

2.1.2.1 Preimenovanje oznak za obravnavane leme

Ker je v nadaljevanju predvidena posebna obravnava primerov, ki izkazujejo katero od 15 obravnavanih lem (ostale želimo iz besedila odstraniti, skupaj z oznako *lemma*), je oznaka *lemma* pri obravnavanih zamenjana z oznako *mojalemma*. Kratek opis programa prinaša poglavje IV-3.1.2.

```
<p id="F0000001.244">
<s id="F0000001.244.1">
<w lemma="Benko" msd="Slmetd">Benka</w>
<w lemma="Pulko" msd="Slmei">Pulko</w>
<c>,</c>
<w lemma="prvi" msd="Kbvzei">prva</w>
<w mojalemma="Slovenka" msd="Slzei">Slovenka</w>
<c>,</c>
<w lemma="ki" msd="Vd">ki</w>
<w lemma="biti" msd="Gp-ste-n">je</w>
<w lemma="v" msd="Dm">v</w>
<w lemma="petinpolleten" msd="Ppnsem">petinpolletnem</w>
<w lemma="potovanje" msd="Sosem">potovanju</w>
<w lemma="z" msd="Do">z</w>
<w lemma="motor" msd="Someo">motorjem</w>
<w lemma="obkrožiti" msd="Ggdd-ez">obkrožila</w>
<w lemma="svet" msd="Sometn">svet</w>
<c>,</c>
```

Tabela 15: Preimenovanje oznak za obravnavane leme.

2.1.2.2 Odstranitev oznak iz besedila

V tem koraku so iz besedila odstranjene oznake, vezane na format xml: oznake za začetek in konec besede in ločila ter za napoved lem (*lemma*) in oblikoskladenjskih oznak (*msd*). Odstranjene so vse leme, razen označenih z *mojalemma*. Pri slednjih je nato odstranjena oblikoskladenjska oznaka. Na koncu so torej ohranjene le gole oblikoskladenjske oznake ter obravnavane leme:

```
<p id="F0000001.244">
<s id="F0000001.244.1"> Slmetd Slmei , Kbvzei Slovenka , Vd Gpxstexn Dm Ppnsem Sosem Do Someo
Ggddxez Sometn , Gpxptexn L Rnn Ggndxez Ppnzmt Sozmt Zpxzer Sozer . </s>
</p>
```

Tabela 16: Odstranitev oznak iz besedila.

Kot je razvidno iz gornjega primera, sta v tem koraku izvedena še dva postopka: (I) vezaji znotraj oblikoskladenjskih oznak so zamenjani z znakom x v izogib razbijanju oznak na več delov⁷⁰ ter (II) odstranjeni so prehodi v nove vrstice, razen na koncu povedi. Na tak način je zagotovljena oblika tekočega besedila, potrebna za nadaljnjo obdelavo podatkov. Kratek opis uporabljenih programov prinaša poglavje IV-3.1.3.

2.1.2.3 Menjava ločil s posebnimi oznakami

Ker je pri luščenju besednih nizov ločila smiselno obravnavati kot mejnike (več o tem v IV-2.2.2), jih je treba pri pretvorbi besedil zamenjati z besednimi oznakami, ki jih je v nadaljevanju enostavno poiskati, in vzorce, ki oznake vsebujejo, odstraniti iz nadaljnje obravnave. Znotrajpovedna ločila so v besedilih zamenjana z oznako VMES, končna pa z oznako KONC. Kratek opis programa prinaša poglavje IV-3.1.4.

```
<p id="F0000001.244">
<s id="F0000001.244.1"> Slmetd Slmei VMES Kbvzei Slovenka VMES Vd Gpxstexn Dm Ppnsem Sosem Do
Someo Ggddxez Sometn VMES Gpxptexn L Rnn Ggndxez Ppnzmt Sozmt Zpxzer Sozer KONC </s>
</p>
```

Tabela 17: Menjava ločil z oznakami.

2.1.2.4 Dodajanje oznake za konec besedilnega dela

V korpusnih besedilih so pogosti primeri, kjer konci povedi oz. besedilnih delov ne prinašajo končnega ločila (npr. pri naslovih, tabelarnih podatkih, alinejskem naštevanju itd.). Meja povedi je, kot rečeno, za pripravo seznamov vzorcev ključnega pomena in mora biti pri obdelavi podatkov jasno nakazana. Problem je rešen s pripisom posebne oznake NIKONC na omenjena mesta besedil. Kratek opis programa prinaša poglavje IV-3.1.5.

Sledi primer stavka iz korpusnega besedila, v katerem na koncu besedilnega dela ni končnega ločila. Spodaj je navedeno pretvorjeno besedilo z vstavljenjo oznako NIKONC.

```
<s id="F0000001.57.1">
<w lemma="1550" msd="Kag">1550</w>
<w lemma="protestant" msd="Somei">protestant</w>
<w lemma="Primož" msd="Slmei">Primož</w>
<w lemma="Trubar" msd="Slmei">Trubar</w>
<w lemma="izdati" msd="Ggdste">izda</w>
<w lemma="prvi" msd="Kbvzdt">prvi</w>
<w lemma="slovenski" msd="Ppnzdt">slovenski</w>
<w lemma="knjiga" msd="Sozdt">knjigi</w>
</s>

<s id="F0000001.57.1"> Kag Somei Slmei Slmei izdati Kbvzdt Ppnzdt Sozdt NIKONC </s>
```

Tabela 18: Dodajanje oznake za konec povedi.

2.1.2.5 Končno stanje pretvorjenega besedila

Končno stanje pretvorjenega besedila prikazuje primer v nadaljevanju. Kot je razvidno, v celotnem odstavku ostaja izpisana obravnavana lema *moder*, ki je obkrožena z oblikoskladenjskimi oznakami. Ločila so zabeležena v obliki oznak VMES ter KONC, vstavljen je znak za manjkajoče končno ločilo NIKONC.

⁷⁰ Obdelava korpusnih besedil pogosto predvideva tokenizacijo (razbijanje besedila na manjše enote glede na vnaprej definirane besedne meje) na mestu vezaja.

Ostale oznake, vezane na format xml – za označevanje delov besedila, tj. stavkov in odstavkov, skupaj z identifikacijskimi oznakami – ostajajo, ker ne vplivajo na nadaljnjo obdelavo podatkov.

```
<p id="F0000001.84">
<s id="F0000001.84.1"> Zkxzei Sozei Ggvste L Ppnsmt Sosmt Ppnmmr Sommr Vp Sozer KONC </s>
<s id="F0000001.84.2"> Dm Somem Ggnspmxn Rnn Ppnmmt Sommt VMES Pdnzmi N KONC </s>
<s id="F0000001.84.3"> Dt Kbvmdm Kbgmmm Sommm Ggdste Somei Ggnste Somei VMES Vd Ggvste Kbvmei
N VMES Rnn Vp Zpxxxdxk Ggnstm Ppnmeid VMES Ppnmeid VMES moder Vp Ppnmeid Somei VMES Vp
Kbvmei N KONC </s>
<s id="F0000001.84.4"> Do Ppnmeo Ggdste Ppnmeid Somei VMES Kbvmei Pdnmein VMES Rnn Kbzmeid
Somei VMES O NIKONC </s>
<s id="F0000001.84.5"> L Dt Sozet Ppnmer Somer Ggnste Somei Ggdn Rnn Sozmr Vp Sosmt VMES L Vp Ppnsei
Ggvn Zzxsei Znxsei Slzmr KONC </s>
<s id="F0000001.84.6"> Zkxsei Gpxstexn Sozei Sozer Do Pdnmeo Someo KONC </s>
<s id="F0000001.84.7"> Rnn Somei Ggnste Zsdmete Sometn Vp Ggnste VMES Rnn Gpxpdexn Pdnmein Dt
Ppnzet Sozet KONC </s>
<s id="F0000001.84.8"> Rnn Ggnsde Sometn KONC </s>
<s id="F0000001.84.9"> Zspmeie Somei Gpxstexn L Ppnzei Soser N Sozed VMES Vd Ggnste Dm Slmem Vp
Gpxstexn Somei Ppnmer Somer VMES Kbvmei Pdnmein KONC </s>
</p>
```

Tabela 19: Končno stanje pretvorjenega besedila.

Na predstavljeni način je pretvorjen celotni podkorpus. Po predelavi 22.563 datotek obsega skupno le še 340 MB. Nadaljevanje poglavja prinaša opis postopka izdelave pogostnostnega seznama vzorcev na osnovi pripravljenega podatkovnega vira.

2.2 Priprava seznama vzorcev

2.2.1 Izdelava seznamov vzorcev s programom Oxford Wordsmith Tools

Oxford Wordsmith Tools⁷¹ je nabor programskih orodij, prvotno pripravljenih za leksikografske potrebe založbe Oxford University Press. Zaradi zmogljivosti (programska orodja se redno posodablja) in prijaznosti uporabniku, se je uporaba programa zelo razširila – ker zasnova programa to omogoča, tudi na obdelavo drugih jezikov (v slovenski literaturi je program predstavljen denimo v Vintar 2008: 90–99).

V raziskavi je uporabljena funkcija Clusters znotraj orodja Concord. Clusters (slovenska ustreznica angleškemu *cluster* bi lahko bila *skup*, kakor je predlagano v Gorjanc 2005) glede na izbrane parametre izdela po pogostnosti urejen seznam n-gramov, pri čemer upošteva podatke, ki se pojavljajo v konkordančnem nizu za obravnavano besedo. Konkordančni nizi sami so zaradi oblike podatkovnega seta manj uporabni, kot je razvidno iz spodnjega primera s konkordančnim jedrom *pajek*, zato so v raziskavi puščeni ob strani:

⁷¹ <<http://www.lexically.net/wordsmith/>>.

	N Concordance
15	Sommr KONC Rnn Ggvsde Ppnmmt pajek VMES Vd Ggnstm Rnn Dm
16	Vd Gpxstexn L Dm Rnn Pdnmem pajek Rnn Ppnser Soser VMES Vp Vp
17	Rnn Vd Gpxstexn Gpxdxes Sosei pajek L Sozei Dt Sosei KONC Vd Gpxg
18	Sozmi Do Sozmo Dt Sozdt Vp Ppnmet pajek VMES L Psnmein Somei Do
19	Do Soseo Dm Somem Vp Someo Do pajek VMES Do Zotmmo Vp Vd Gpxg
20	Kbzzem Sozem VMES Vp Gpxptmxn pajek L Rnn Ggndxmm KONC

Tabela 20: Del konkordančnega niza z jedrom *pajek* in oblikoskladenjskimi oznakami.

S programom Clusters so pripravljene sezname vzorcev za dvodelne, tridelne ter štiridelne kombinacije leme z okoliškimi oznakami. Z ustrezno zamejitvijo okna obravnave je doseženo, da je v vzorcu vedno zajeta tudi jedrna beseda (pri tridelnih kombinacijah pridejo denimo v poštev tiste, ki upoštevajo do vključno dve mesti levo in dve mesti desno od jedrne besede).

Rezultat predstavlja spodnji primer, tj. prvih 10 mest seznama tridelnih vzorcev z lemo *pajek*. *N* je zaporedna številka vzorca v pogostnostnem seznamu, *Freq.* pogostnost vzorca, *Length* dolžina vzorca. Pregled vzorcev kaže, da gre večinoma za kombinacije z oznakami za ločila:

N	Cluster	Freq.	Length
1	KAG NIKONC PAJEK	827	3
2	PAJEK SOMEI KAG	374	3
3	PAJEK VMES VD	371	3
4	VMES PAJEK VMES	320	3
5	PAJEK SOZMR VMES	312	3
6	PAJEK VMES KAG	261	3
7	VMES PAJEK DT	251	3
8	NIKONC PAJEK SOMEI	243	3
9	VMES PAJEK SOZMR	239	3
10	KAG VMES PAJEK	233	3

Tabela 21: Del seznama tridelnih vzorcev z lemo *pajek*.

Za vsako od 15 lem so bili torej izdelani po trije sezname glede na dolžino vzorca (seznam dvodelnih, tridelnih ter štiridelnih vzorcev), skupno 45 seznamov, shranjenih v obliki xls, ki omogoča enostavno nadaljnjo obdelavo podatkov.

2.2.2 Prva obdelava seznama vzorcev

Predvideno je, da vzorci, kakršni nas zanimajo na tem mestu, ne presegajo meja povedi, zato je smiselna odstranitev vzorcev, ki vsebujejo oznaki (I) KONC – zamenjuje končno ločilo povedi ter (II) NIKONC – zaznamuje konec povedi, kjer ločilo v besedilu ni prisotno.

Prav tako so iz nadaljnje obravnave izločeni vzorci, vsebujoči oznako VMES, ki v podatkovnem viru zamenjuje znotrajpovedna ločila. Odločitev za izločitev teh vzorcev je bila sprejeta kljub dejstvu, da potencialno prinašajo relevantne besedne zveze, tj. predvsem pristavčno zložene besedne zveze. Kot je evidentirano v Slovenski slovnici, je »[p]ristavčno zložena besedna zveza [...] podobna podredno zloženi, le da je določujoči del prosto pridružen: *Prešeren, naš največji pesnik*; sicer pa isto predmetnost imenuje z dveh stališč.« (Toporišič 2004⁴: 558)

Pristavčno zložene besedne zveze zaradi svoje posebne strukture zahtevajo samostojno raziskavo s prilagojenim postopkom luščenja. V pričujoči raziskavi je za zveze tega tipa predvideno prenizko število pridobljenih primerov glede na čas, namenjen njihovi identifikaciji, zato jih trenutno puščamo ob strani.

Za avtomatsko izločitev vzorcev, ki vsebujejo oznake KONC, NIKONC ali VMES, je bil pripravljen program, ki ga opisuje poglavje IV-3.2.1. Gre za preprost program, ki bere podatke iz tekstovne datoteke vrstico za vrstico in v drugo tekstovno datoteko zapiše samo tiste vrstice, ki ne vsebujejo neželenih oznak.

Po prvi selekciji vzorcev se sezname precej skrajšajo – več kot ima vzorec mest, večji je posledično delež odstranjenih vzorcev, kot prikazuje spodnji primer:

vzorci z lemo pajek	dvodelni	tridelni	štiridelni	skupaj
pred izločitvijo	484	5.331	12.460	18.275
po izločitvi	478	3.870	5.871	10.219
delež izločenih	1,24 %	27,41 %	52,88 %	44,08 %

Tabela 22: Delež pri prvi selekciji izločenih vzorcev z lemo pajek.

Vzorci z oznakami za ločila se sicer pojavljajo predvsem na vrhu pogostnostnega seznama, kar ponazarja spodnji primer: od prvih 30 najpogostnejših vzorcev po opisani selekciji ostanejo le trije za nadaljnjo obravnavo (odstranjeni vzorci so v tabeli obarvani sivo):

mesto	vzorec	pogostnost			
1	KAG NIKONC PAJEK	827	16	NIKONC PAJEK SOZMR	166
2	PAJEK SOMEI KAG	374	17	VMES PPNMEID PAJEK	157
3	PAJEK VMES VD	371	18	PAJEK DT KAG	155
4	VMES PAJEK VMES	320	19	SOMEI VMES PAJEK	153
5	PAJEK SOZMR VMES	312	20	VP PAJEK VMES	142
6	PAJEK VMES KAG	261	21	PAJEK KAG VMES	141
7	VMES PAJEK DT	251	22	VMES PAJEK SOMEI	140
8	NIKONC PAJEK SOMEI	243	23	VMES PAJEK KAG	131
9	VMES PAJEK SOZMR	239	24	NIKONC PAJEK VMES	130
10	KAG VMES PAJEK	233	25	PAJEK DT SOSET	124
11	PPNMEID PAJEK VMES	194	26	PPNMMI PAJEK VMES	124
12	VMES PAJEK VP	192	27	VMES VP PAJEK	122
13	SOMEI PAJEK VMES	188	28	PAJEK VMES SOMEI	119
14	DO PAJEK VMES	175	29	PAJEK VMES VP	117
15	DO PAJEK KONC	170	30	VMES SOMEI PAJEK	111

Tabela 23: Prva selekcija najpogostejših 30 tridelnih vzorcev z lemo pajek.

V nadaljevanju je opisana analiza seznamov vzorcev.

2.3 Analiza najpogostejših vzorcev

2.3.1 Priprava podatkov

Naslednji korak raziskave je analiza vzorčnih zapolnitev za najpogostejše vzorce. Na tem mestu so v obravnavo zajeti sezname za štiri leme različnih besednih vrst, tj. *pajek*, *strasten*, *plesati* in *temeljito*. Glavni namen prvega

pregleda vzorčnih zapolnitev je ugotoviti možnosti (I) nadaljnje avtomatske selekcije vzorcev ter (II) združevanja vzorcev v skupine za nadaljnjo obravnavo.

V analizo je zajetih po (približno) 100 najpogostejših dvo-, tri- in štiridelnih vzorcev za vse štiri naštetе leme: izhodiščna meja za zajem v analizo je bila postavljena na 100. mesto seznama; v primeru, da enako pogostnost kot vzorec na stotem mestu seznama prinašajo tudi sledeči vzorci, pa so bili v analizo zajeti tudi slednji. Tabela 24 prinaša podatke o pogostnosti vsakega od posameznih primerov na stotem mestu seznama ter končno število v analizo zajetih vzorcev za vsako od lem. Skupno število zajetih vzorcev je 1231:

	<i>pajek</i>	<i>strasten</i>	<i>plesati</i>	<i>temeljito</i>	skupno število
dvodelni					
pogostnost na 100. mestu	13	11	32	61	
končni delež zajetih	101	101	100	100	402
trodelni					
pogostnost na 100. mestu	17	18	42	67	
končni delež zajetih	101	103	103	101	408
štiridelni					
pogostnost na 100. mestu	6	8	12	22	
končni delež zajetih	116	103	102	100	421
skupno število	318	307	305	301	1231

Tabela 24: Število v analizo zajetih vzorcev.

Za luščenje in urejanje vzorčnih zapolnitev sta bila pripravljena programa, predstavljena v poglavju IV-3.2.2 – prvi za luščenje zapolnitev vsakega od vzorcev, drugi za razvrščanje dobljenih zapolnitev po pogostnosti ter štetje vseh zapolnitev. Končni rezultat so datoteke s podatki za vseh 1231 vzorcev, kakršne ponazarja spodnji primer za vzorec PAJEK DT SOSET:

pajek	za	seno	35
PAJEK	za	seno	5
PAJEK	ZA	SENO	3
pajka	za	seno	2
pajek	za	obračanje	2
pajkoma	za	prestavljanje	2
PAJKA	za	obračanje	2
pajka	na	leto	2
pajka	za	odstranjevanje	1
pajkov	za	spletanje	1
Pajek	na	leto	1
pajka	na	drstenje	1
PAJKA	za	seno	1
PAJEK	za	obračanje	1
pajka	za	obračanje	1
pajku	za	zmanjšanje	1
PAJKE	za	seno	1
VSEH SKUPAJ			62

Tabela 25: Primer vzorčnih zapolnitev za PAJEK DT SOSET.

Kot je razvidno iz primera, je metoda luščenja na tem mestu enostavno izpisovanje besednih oblik, ki ustrezajo izbranemu vzorcu, v novo datoteko (ta je zaradi preglednosti poimenovana glede na vzorec, ki ga prinaša, npr.

pajek Dt Sozet.txt). Kot omenjeno v poglavju IV-1.2, na ravni pridobivanja zapolnitev izluščeni podatki niso urejeni na zahtevnejši ravni – gre izključno za izpis podatkov ter enostavno štetje; štetje zapolnitev je, kot je razvidno iz primera, občutljivo na velike začetnice, zato v mnogih primerih razpršuje podatke (npr. ločen prikaz zapolnitve *pajek za seno (35)* ter *PAJEK za seno (5)*), kar pa v trenutnem koraku raziskave ni problematično.

2.3.2 Razvrščanje vzorcev v skupine za nadaljnjo analizo

Naslednji korak raziskave predstavlja ročni pregled vzorčnih zapolnitev, pri čemer je pozornost posvečena predvsem vprašanju, kateri od vzorcev prinašajo za luščenje relevantne podatke in kateri ne (poimovanje relevantnosti je opredeljeno v poglavju IV-1.2).

Pregled podatkov izkazuje, da je mogoče relevantnost oz. nerelevantnost vzorca v določeni meri predvideti tako glede na (I) **vsebnost** oblikoskladenjskih oznak kot na (II) njihovo **mesto znotraj vzorca**.

Vzorci, ki se izkazujejo za problematične v smislu prinašanja za obravnavo manj zanimivih zapolnitev, denimo vsebujejo oznako za (I) pomožno glagolsko obliko, (II) zaimek, (III) okrajšavo, (IV) števniki ali (V) členek. Posebno skupino tvorijo tudi vzorci, vsebujoči oznako za (VI) nelematizirano besedo. Za našete skupine se je izkazala smiselna nadaljnja skupna obravnava, ki ji je posvečeno poglavje V-4.

Po opredelitvi gornjih skupin sledi selekcija seznamov vzorcev, iz katerih so odstranjeni problematično oznako vsebujoči vzorci. Kratek opis programa, uporabljenega v tem koraku raziskave, sledi v poglavju IV-3.2.1.

Preostale primere je možno glede na vsebnost vzorca razvrstiti v skupine, ki vsebujejo (I) **veznik**, (II) **predlog** ali (III) **same polnopomenske besede**. Pri določanju relevantnosti se v teh primerih izkazuje za pomemben kriterij mesto določene oznake znotraj vzorca – za nerelevantne se denimo kažejo primeri, ki prinašajo veznik na zadnjem mestu vzorca, oznako za predlog pred oznako za glagol itd. Naštetim trem skupinam je v raziskavi posvečena večina pozornosti (glej V-1, V-2 ter V-3).

Prvotni namen raziskave je bil pregled tako dvodelnih ter tridelnih kot tudi štiridelnih vzorcev, vendar so bili tekom analize zadnji iz obravnave izločeni. Pri pregledu štiridelnih vzorcev s polnopomenskimi besednimi vrstami se je namreč pokazalo, da členjenost oblikoskladenjskih oznak⁷² v teh vzorcih močno razprši podatke, kar v preveliki meri podaljšuje njihovo ročno analizo. Obravnava štiri- in večdelnih vzorcev je predvidena v sklopu nadaljnjih raziskav.

2.3.3 Združevanje vzorcev v vzorčne tipe

Po zgledu skladenjskih vzorcev, kakršni se v slovenščini uporabljajo za luščenje terminologije (glej poglavje II-2.1), je združevanje vzorcev v vzorčne tipe osnovano na redukciji oblikoskladenjske oznake na **opredelitev besedne vrste**: vzorci PAJEK PPNMEID, PAJEK PPNMEIN ter PAJEK PPNZER so denimo združeni v vzorčni tip *pajek + pridevnik (pajek + Prid)*. Potrebno je poudariti, da je takšna združitev le izhodišče za luščenje besednih nizov, analiza katerih je v večji meri usmerjena tudi k obravnavi ostalih v oznakah pripisanih oblikoskladenjskih kategorij (glej IV-2.4.2).

Za združevanje vzorcev v vzorčne tipe je bil pripravljen specializiran program (opis programa v IV-3.2.3).

⁷² S številom elementov v vzorcih se pogostnosti pojavitev posameznega vzorca nižajo. Členjenost oznak to razpršenost še povečuje, pri čemer je treba upoštevati tudi dejstvo, da so oznake za različne besedne vrste različno členjene. Obravnavana metoda torej na ravni večdelnih vzorcev zahteva določene prilagoditve.

2.4 Luščenje besednih nizov in evalvacija izluščenih podatkov

Glavnino raziskovalnega dela prinaša analiza v prejšnjem poglavju opredeljenih skupin vzorcev z namenom razvoja čim kvalitetnejše metode za luščenje dvo- ter tridelnih besednih zvez za slovenščino. Raziskovalna vprašanja, ki vodijo analizo, so na ravni (I) napak avtomatskega označevanja, (II) kategorij oblikoskladenjskih oznak ter (III) organizacije vzorčnih tipov glede na relevantnost za luščenje. Vprašanja so v nadaljevanju poglavja le na kratko opredeljena, saj je tematiki v celoti posvečeno poglavje V, deloma pa tudi poglavje VI.

2.4.1 Napake avtomatskega označevanja

Ker luščenje, kakršno je predstavljeno v pričujoči raziskavi, temelji na v korpusnem besedilu pripisanih oblikoskladenjskih oznakah, je ključnega pomena iskanje **tipičnih napak pripisanih oznak** – predvsem tistih, za katere je predviden negativen vpliv na ravni luščenja podatkov. Med analizo (glej poglavje V) je avtomatskemu označevanju posvečena pozornost na mestih, kjer izluščeni podatki ne ustrezajo oblikoskladenjskim oznakam v vzorcu. Na vsakem takem mestu sledi kratka predstavitev evidentiranega označevalnega problema s primeri ter predlogom rešitve.

Na tem mestu je potreben ponoven poudarek, da se analiza namensko osredotoča na šibka mesta označevanja, ob strani pa pušča vse primere, v katerih je označevanje ustrezno. V vseh primerih je potrebno imeti pred očmi dejstvo, da oblikoskladenjsko označevanje za slovenščino izpričuje relativno visoko mero natančnosti – kot omenjeno v poglavju o II-1.2.4, je za obravnavani označevalnik ugotovljena 85,7 % uspešnost.

2.4.2 Kategorije oblikoskladenjskih oznak

Analiza na ravni kategorij oblikoskladenjskih oznak je namenjena poskusu ugotovitve, katere kategorije je pri luščenju koristno upoštevati kot razločevalne, katere pa je smotno združevati v izogib prevelike razpršenosti izluščenih podatkov. V določenih primerih je tako na ravni obdelave ter končnega prikaza podatkov predvideno upoštevanje osnovne oblike besede, medtem ko je na drugih mestih potrebno upoštevanje določene druge besedne oblike (prim. denimo zvezi *pajek zaklopničar* in *strup pajka*).

Vprašanje relevantnosti kategorij oznak je torej ključno tako na ravni razvrščanja besednih nizov glede na vrste, ki se ob luščenju izkazujejo, ter urejanja slednjih po pogostnosti, kot tudi na ravni končnega prikaza podatkov, ki mora biti pripravljen na način, da je za človeškega uporabnika čim bolj enostavno berljiv: kot rečeno v III-3.3, je smiselno stremeti k podatkom, ki bodo graditelju leksikalne zbirke predstavljeni v pregledni obliki, ki omogoča hitro in enostavno nadaljnjo obravnavo.

2.4.3 Organizacija vzorčnih tipov glede na relevantnost za luščenje

Eden od ciljev raziskave je tudi priprava seznamov relevantnih vzorcev oz. vzorčnih tipov, ki se v nadaljevanju lahko uporabljajo za avtomatsko pridobivanje izbranega tipa leksikalnih podatkov (bodisi na način, predstavljen v doktorski raziskavi, bodisi kot del kake druge oblike luščenja podatkov), obenem pa priprava seznama nerelevantnih vzorcev, ki služi kot izhodišče za avtomatsko izločanje nerelevantnih vzorcev iz seznama vzorcev za obravnavano lemo.

3 Programi

Namesto navajanja celotnih programskih kod, ki so bile pripravljene v doktorski raziskavi, sledi v nadaljevanju kratek opis programskih skript glede na nalogo, za katero so bile uporabljene. Vsi programi so napisani v jeziku Perl v okolju ActiveState Komodo IDE 4.4.⁷³

Programi so med raziskavo nastajali sproti, tj. za reševanje specifičnih raziskovalnih vprašanj, kakor so se pojavljala. To se v določeni meri odraža v razdrobljenosti oz. neintegriranosti posameznih postopkov v celoto – tudi na mestih, kjer bi bilo slednje vsekakor pričakovano (npr. uporaba enega programa za luščenje vzorčnih zapolnitev ter drugega za štetje podatkov ipd.). Na mnogih mestih uporaba pripravljenih programov tudi še zahteva ročno pripravo podatkov na vhodu posameznega programa ali pa sprotno spreminjanje programske kode glede na trenutni raziskovalni cilj (npr. vsakokratni ročni vnos vzorčnega tipa, ki ga želimo uporabiti za luščenje besednih nizov itd.).

Zaključiti je torej možno, da je nabor pripravljenih programov ključnega pomena za uspešno izvedbo pričujoče raziskave, ni pa v trenutni obliki primeren za luščenje podatkov v širše zastavljenih raziskavah. Pred tem korakom je potrebna integracija vseh ključnih stopenj luščenja (izbira relevantnih vzorcev za luščenje, ločeno luščenje podatkov glede na definirane vrste, razvrščanje po pogostnosti, ustrezen prikaz dobljenih podatkov) v smiselno programsko celoto, obenem pa seveda ustrezna prilagoditev programov (oz. v doktorskem delu opisane metode nasploh) uporabljenemu podatkovnemu viru, namena posamezne raziskave itd.

3.1 Priprava in obdelava podkorpusa

3.1.1 Štetje lem v podkorpusu

ime programa	Štetje lem v podkorpusu.pl
namen programa	Program prešteje pojavitve izbranih 15-ih lem v podkorpusu, obenem pa vse leme podkorpusa.
razlog za uporabo programa	Število obravnavanih lem v podkorpusu je potrebno za primerjavo s številom teh lem v korpusu FidaPLUS, kakršen je na voljo na internetu. Podatek o številu vseh lem je potreben za opredelitev obsega podkorpusa. Več informacij v IV-2.1.1.
kratek opis delovanja	<ol style="list-style-type: none"> 1. Program po vrsti odpira, obdeluje ter zapira podkorpuse datoteke; 2. v trenutno odprti datoteki vrstico za vrstico išče in šteje izbrane leme (išče npr. vzorec <i>lemma</i>="pajak" itd.), obenem pa šteje vse leme besedila; 3. po koncu štetja izpiše pogostnost za vsako od obravnavanih lem ter vseh lem v podkorpusu.

3.1.2 Preimenovanje obravnavanih lem

ime programa	Preimenovalnik obravnavanih lem.pl
namen programa	Program preimenuje korpusne oznake <i>lemma</i> v <i>mojalemma</i> za izbranih 15 lem.
razlog za uporabo programa	Preimenovanje oznak v nadaljevanju omogoča ločevanje izbranih lem od vseh ostalih. Več informacij v IV-2.1.2.1.

⁷³ <<http://www.activestate.com/komodo/>>.

kratek opis delovanja	<ol style="list-style-type: none"> 1. Program po vrsti odpira, obdeluje ter zapira podkorpuse datoteke; 2. v trenutno odprti datoteki vrstico za vrstico išče pojavitve izbrane leme; 3. v primeru najdene pojavitve zamenja njeno oznako <i>lemma</i> z <i>mojalemma</i>; 4. zapiše trenutno obravnavano vrstico v novo datoteko (ohrani ime datoteke, zapiše jo v drugo mapo).
-----------------------	--

3.1.3 Odstranitev oznak v besedilu podkorpusa

ime programa	Odstranitev besedilnih oznak 1.pl
namen programa	Program iz besedila odstrani izbrane oznake, postavi prehode besedila v novo vrstico na ustrezna mesta in zamenja vezaje znotraj oblikoskladenjskih oznak z znakom x.
razlog za uporabo programa	Pretvorba besedil podkorpusa omogoča pripravo pogostnostnih seznamov vzorcev. Več informacij v IV-2.1.2.2.
kratek opis delovanja	<ol style="list-style-type: none"> 1. Program po vrsti odpira, obdeluje ter zapira podkorpuse datoteke; 2. vrstico za vrstico odstranjuje oznake, vezane na format xml, skupaj z lemami, besednimi oblikami ter prehodi v nove vrstice, da ostanejo le gole oblikoskladenjske oznake; 3. program iz besedila odstrani prehode v novo vrstico na mestih za oznako za konec stavka; 4. doda presledke pred oznake za konec stavka in doda prehode v novo vrstico za oznakami za konec odstavka; 5. zamenja vezaje znotraj oblikoskladenjskih oznak z znakom x; 6. odstrani na format xml vezane oznake za ločila, da ostanejo le gola ločila; 7. zapiše trenutno obravnavano vrstico v novo datoteko (isto ime datoteke, druga mapa).
razlaga	<p>Izvorni zapis za vsako korpusno pojavnico predstavlja spodnji primer:</p> <pre><w lemma="Trubar" msd="Slmei">Trubar</w></pre> <p>Med obdelavo so iz besedila odstranjeni deli pred oblikoskladenjsko oznako ter za njo:</p> <pre><w lemma="Trubar" msd="Slmei">Trubar</w></pre> <p>Mesta, ki vsebujejo preimenovane oznake <i>mojalemma</i>, med postopkom ostanejo nespremenjena:</p> <pre><s id="F0000001.84.3"> Dt Kbvmdm Kbgmmm Sommm Ggdste Somei Ggnste Somei , Vd Ggvste Kbvmei N , Rnn Vp Zpxxdxk Ggnstm Ppnmeid , Ppnmeid ,<w mojalemma="moder" msd="Ppnmeid Vp Ppnmeid Somei , Vp Kbvmei N . </s></pre>
ime programa	Odstranitev besedilnih oznak 2.pl
namen programa	Program iz besedila odstrani preostanek oznak.
razlog za uporabo	Pretvorba besedil podkorpusa omogoča pripravo pogostnostnih seznamov vzorcev.

programa	Več informacij v IV-2.1.2.2.
kratek opis delovanja	<ol style="list-style-type: none"> 1. Program po vrsti odpira, obdeluje ter zapira podkorpuse datoteke; 2. vrstico za vrstico išče pojavitve katere od 15 lem; 3. v primeru najdene pojavitve odstrani iz besedila oznako <i>mojalemma</i>, skupaj z ustrežajočo oblikoskladenjsko oznako, da ostane le gola lema; 4. zapiše trenutno obravnavano vrstico v novo datoteko (isto ime datoteke, druga mapa).

3.1.4 Pretvorba ločil v oznake

ime programa	Pretvorba ločil v oznake.pl
namen programa	Program v podatkovnem naboru poišče ločila in jih zamenja z izbranimi črkovnimi oznakami.
razlog za uporabo programa	Pretvorba besedil podkorpusa omogoča pripravo pogostnostnih seznamov vzorcev. Več informacij v IV-2.1.2.3.
kratek opis delovanja	<ol style="list-style-type: none"> 1. Program odpre mapo z datotekami podatkovnega vira ter datoteke po vrsti odpira, obdeluje ter zapira; 2. vrstico za vrstico išče ločila, nekončna ločila zamenja z oznako VMES, končna s KONC; 3. zapiše trenutno obravnavano vrstico v novo datoteko (isto ime datoteke, druga mapa).
razlaga	Z oznako ne želimo menjati (I) vezaja, kadar označuje prosto mesto znotraj oblikoskladenjske oznake (npr. <i>Gvdr-emt</i>), zato so bili ti vezaji predhodno pretvorjeni v znak x, ter (II) pike, kadar se nahaja znotraj kode ID (npr. <i>F0000001.818</i>), čemur se izognemo z ohranjanjem presledkov pred stavčnimi ločili pri odstranjevanju besedilnih oznak.

3.1.5 Dodajanje oznake za konec besedilnega dela

ime programa	Dodajanje oznake NIKONC.pl
namen programa	Program v podatkovnem naboru poišče mesta, kjer na koncu besedilnega dela ni končnega ločila, ter na ta mesta vstavi oznako NIKONC.
razlog za uporabo programa	Pretvorba besedil podkorpusa omogoča pripravo pogostnostnih seznamov vzorcev. Več informacij v IV-2.1.2.4.
kratek opis delovanja	<ol style="list-style-type: none"> 1. Program odpre mapo z datotekami podatkovnega vira ter datoteke po vrsti odpira, obdeluje ter zapira; 2. pred vse oznake za konec stavka, pred katerimi še ni oznake KONC, vstavi oznako NIKONC; 3. zapiše trenutno obravnavano vrstico v novo datoteko (isto ime datoteke, druga mapa).

3.2 Priprava vzorčnih tipov

3.2.1 Odstranjevanje vzorcev, vsebujočih določene oznake

ime programa	Odstranjevanje nerelevantnih vzorcev – oznake za ločila.pl
namen programa	Program s pogostnostnega seznama vzorcev odstrani tiste, ki vsebujejo neželene oznake. Več informacij v IV-2.2.2.
razlog za uporabo programa	Program se uporablja za prvo selekcijo seznamov vzorcev, za odstranjevanje vzorcev, ki vsebujejo oznake za ločila oz. konec povedi.
kratek opis delovanja	<ol style="list-style-type: none"> 1. Program odpre seznam vzorcev za obravnavano lemo v tekstovni različici; 2. program bere seznam vrstico za vrstico; 3. razen v primeru, da trenutno obravnavana vrstica vsebuje neželeno oznako VMES, KONC ali NIKONC, program prepíše vrstico v novo datoteko.
ime programa	Odstranjevanje nerelevantnih vzorcev – oblikoskladenjske oznake.pl
namen programa	Program s seznama vzorcev odstrani tiste, ki vsebujejo neželene oblikoskladenjske oznake. Več informacij v IV-2.3.2 ter V-4.6.
razlog za uporabo programa	Program se uporablja za drugo selekcijo seznama vzorcev, za odstranjevanje vzorcev, ki vsebujejo katero od neželenih oblikoskladenjskih oznak (npr. za členek, okrajšavo, števnik, nelematizirano besedo itd.).
kratek opis delovanja	<ol style="list-style-type: none"> 1. Program odpre seznam vzorcev za obravnavano lemo v tekstovni različici; 2. program bere seznam vrstico za vrstico; 3. razen v primeru, da trenutno obravnavana vrstica vsebuje neželeno oblikoskladenjsko oznako, program prepíše vrstico v novo datoteko.

3.2.2 Luščenje vzorčnih zapolnitev

ime programa	Preliminarno luščenje.pl
namen programa	Program iz podkorpusa izlušči vzorčne zapolnitve in jih zapiše v ustrezne datoteke.
razlog za uporabo programa	Program se uporablja za luščenje vzorčnih zapolnitev za analizo najpogostnejših vzorcev za vsako od lem <i>pajek</i> , <i>strasten</i> , <i>plesati</i> , <i>temeljito</i> . Več informacij v IV-2.3.1.
kratek opis delovanja	<ol style="list-style-type: none"> 1. Uporabnik vnese elemente vzorcev v spremenljivke (program omogoča hkratno obdelavo okrog 30-ih vzorcev); 2. program po vrsti odpira, obdeluje ter zapira podkorpusne datoteke; 3. za vsakega od najavljenih vzorcev program preverja obstoj v podkorpusu s pomočjo vrste regularnih izrazov; 4. v primeru, da se besedilo ujema z regularnim izrazom, program odpre datoteko, ki jo pri prvem najdenem zadetku poimenuje glede na elemente vzorca (npr. <i>Sozet pajek.txt</i>); 5. program zapiše najdeni zadetek v datoteko, ki jo po uporabi zapre.
razlaga	Regularni izrazi prinašajo strukturo korpusnega besedila, vključujočo spremenljivke na ustreznih mestih:

	<pre><w lemma="([^\"]*)" msd="\$V2">([^\"]*)</w> <w lemma="\$V3" msd="([^\"]*)">([^\"]*)</w></pre> <p>Regularni izrazi se razlikujejo glede na mesto, ki ga v obravnavanem vzorcu zaseda lema – gornji primer omogoča iskanje dvodelnega vzorca, ki ima na prvem mestu oblikoskladenjsko oznako, na drugem lemo. Vrednost spremenljivk, kot rečeno, na začetku programa določi uporabnik.</p> <p>Opisani program je pripravljen v treh različicah, za dvodelne, tridelne ter štiridelne vzorce.</p>
ime programa	Štetje preliminarnih podatkov.pl
namen programa	Program prešteje vzorčne zapolnitve v tekstovni datoteki in jih prikaže v tabelarni obliki.
razlog za uporabo programa	Program se uporablja za urejanje vzorčnih zapolnitev, izluščenih s programom <i>Preliminarno luščenje</i> , glede na pogostnost. Podatki so zapisani v tabelarni obliki, ki omogoča njihovo nadaljnjo analizo. Več informacij v IV-2.3.1.
kratek opis delovanja	<ol style="list-style-type: none"> 1. Program odpre mapo s tekstovnimi datotekami, vsebujočimi izluščene zapolnitve, ter datoteke po vrsti odpira, obdeluje ter zapira; 2. za vsako od tekstovnih datotek v obdelavi program ustvari enakoimensko datoteko za pisanje v formatu xls; 3. program bere vhodno datoteko vrstico za vrstico in vsako trenutno vrstico shrani v tabelo, v kateri pri vsaki novi pojavitvi ustrezni vzorčni zapolnitvi poveča vrednost za ena; 4. po koncu branja program uredi podatke glede na vrednost, tj. pogostnost, začenši z najvišjo pogostnostjo; 5. program sešteje pogostnosti vseh besednih nizov; 6. program bere urejeni seznam in razbija vzorčne zapolnitve na elemente (besede); 7. program zapiše v izhodno datoteko besede zapolnitve ter pogostnost zapolnitve na način, da je vsaka beseda v svojem stolpcu; 8. na koncu seznama doda vrstico s podatkom o vsoti vseh pogostnosti besednih nizov.
razlaga	Opisani program je pripravljen v treh različicah, za dvodelne, tridelne ter štiridelne vzorce.

3.2.3 Luščenje vzorcev za posamezni vzorčni tip

ime programa	Luščenje vzorcev za vzorčni tip.pl
namen programa	Program iz seznama vzorcev izlušči tiste, ki ustrezajo izbranemu vzorčnemu tipu, in jih zapiše v ločeno tabelo.
razlog za uporabo programa	Program se uporablja za pripravo tabel, ki za vsakega od vzorčnih tipov vsebujejo vse ustrezajoče vzorce, kar omogoča informacijo o zastopanosti vzorčnega tipa v korpusnem viru. Več informacij v IV-2.3.3.
kratek opis delovanja	<ol style="list-style-type: none"> 1. Program odpre seznam vzorcev za obravnavano lemo v tekstovni različici; 2. bere datoteko vrstico za vrstico, razbija vsebino vrstic na mestih presledkov in

	shranjuje na ta način pridobljene elemente v urejeni seznam (ang. <i>array</i>);
	3. v primeru, da se vzorec ujema z regularnim izrazom, ki opredeljuje obravnavani vzorčni tip, se elementi urejenega seznama po vrsti pišejo v tabelo vsak v svoj stolpec.
razlaga	Opisani program je pripravljen v dveh različicah, za dvodelne ter tridelne vzorce.

3.3 Luščenje besednih nizov

3.3.1 Osnovno luščenje

ime programa	Osnovno luščenje.pl
namen programa	Program iz podkorpusa izlušči podatke, ki ustrezajo izbranemu vzorčnemu tipu, in jih zapiše v ustrezno datoteko.
razlog za uporabo	Program se uporablja za osnovno luščenje besednih nizov. Pri vsakem primeru so izluščeni vsi podatki – leme, oznake ter oblike – kar omogoča nadaljnjo analizo na ravni kvalitete označenosti ter oblikoskladenjskih kategorij, ki jih je pri nadaljnjem luščenju potrebno upoštevati.
kratek opis delovanja	<ol style="list-style-type: none"> 1. Program po vrsti odpira, obdeluje ter zapira podkorpusne datoteke; 2. bere datoteko vrstico za vrstico in išče vse zadetke, ki ustrezajo regularnemu izrazu, v katerem so definirane informacije o obravnavani lemi ter oblikoskladenjskih oznakah; 3. v primeru najdenega zadetka v posebno datoteko zapiše izbrani nabor informacij.
razlaga	<p>Regularni izrazi, ki se uporabljajo, prinašajo strukturo korpusnega besedila v formatu xml, definirano na mestih, ki jih predvideva vzorčni tip, in nedefinirano na vseh ostalih mestih. V spodnjem primeru je pri prvi besedi definirana lema (<i>temeljito</i>), pri drugi pa oblikoskladenjska oznaka (N). Ostala mesta strukture so opredeljena z obliko <code>[^<]*</code>, ki omogoča najdenje katere koli zapolnitve (pomeni torej »poljubno«):</p> <pre><w lemma="(temeljito)" msd="([<]*)">([<]*)</w> <w lemma="([<]*)" msd="(N)">([<]*)</w></pre> <p>Oklepaji znotraj regularnega izraza v jeziku Perl omogočajo naknadno sklicevanje na posamezne segmente najdenega podatka. Gornji regularni izraz omogoča nadaljnjo uporabo (zapis v novo datoteko) leme, oblikoskladenjske oznake ter izpričane oblike besede:</p> <pre>temeljito zadihtal Rnn N temeljito zadihtal</pre> <p>Opisani program je pripravljen v dveh različicah, za dvodelne ter tridelne vzorce.</p>

3.3.2 Dva samostalnika

ime programa	Luščenje zvez dveh samostalnikov.pl
namen programa	Program iz podkorpusa izlušči podatke, ki ustrezajo izbranemu vzorčnemu tipu, in jih

	zapiše v ustrezno datoteko.
razlog za uporabo programa	Program se uporablja za luščenje besednih zvez dveh samostalnikov. Je nadgradnja osnovnega luščenja besed, ki upošteva kategorije oblikoskladenjskih oznak na mestih, potrebnih za ustrezno razvrščanje izluščenih podatkov glede na tipe samostalniških zvez s samostalniškim prilastkom. Več informacij v V-1.1.1.
kratek opis delovanja	<ol style="list-style-type: none"> 1. Program po vrsti odpira, obdeluje ter zapira podkorpuse datoteke; 2. bere datoteko vrstico za vrstico in išče vse zadetke, ki ustrezajo regularnemu izrazu, v katerem so definirane informacije o obravnavani lemi ter oblikoskladenjskih oznakah; 3. v primeru najdenega zadetka razbije oblikoskladenjske oznake na posamezne znake, ki jih shrani v urejeni seznam; 4. ustrezne znake oblikoskladenjskih oznak iz urejenega seznama shrani v spremenljivke, ki jih nato primerja med sabo (npr. oznako za spol prve besede ter oznako za spol druge besede); 5. v primeru izpolnjevanja danega pogoja (glej spodaj), piše podatke (leme, oznake ter besedne oblike) v ustrezno datoteko.
razlaga	<p>Regularni izrazi, ki se uporabljajo, se od izrazov osnovnega luščenja razlikujejo v tem, da prinašajo opredelitev oblikoskladenjske oznake za vse obravnavane besede:</p> <pre><w lemma="(pajek)" msd="(S.....?)">([<]*)</w> <w lemma="([<]*)" msd="(S.....?)">([<]*)</w></pre> <p>Ker imajo oblikoskladenjske oznake regularno strukturo, lahko s klicem določenega mesta oznake dobimo informacijo, ki je za določen tip zvez relevantna; največkrat je to oznaka za sklon samostalnika, v enem primeru pa tudi podatek o občnosti/lastnosti samostalnika. Program pri razvrščanju zadetkov v datoteke preverja naslednje pogoje:</p> <ul style="list-style-type: none"> - sklona sta enaka (<i>\$sklon1 eq \$sklon2</i>); - oba sklona sta v rodilniku (<i>(\$sklon1 eq \$sklon2) and (\$sklon2 eq "r")</i>); - oba sklona sta v dajalniku (<i>(\$sklon1 eq \$sklon2) and (\$sklon2 eq "d")</i>); - prvi samostalnik ni v rodilniku, drugi je v rodilniku (<i>(\$sklon1 ne \$sklon2) and (\$sklon2 eq "r")</i>); - prvi samostalnik ni v dajalniku, drugi je v dajalniku (<i>(\$sklon1 ne \$sklon2) and (\$sklon2 eq "d")</i>); - sklona samostalnikov nista enaka, drugi samostalnik je lastno ime (<i>(\$sklon1 ne \$sklon2) and (\$obcnost eq "I")</i>); - sklona samostalnikov nista enaka, drugi samostalnik ni ne v rodilniku ne v dajalniku (<i>(\$sklon1 ne \$sklon2) and (\$sklon2 ne "r") and (\$sklon2 ne "d")</i>).

3.3.3 Pridevnik s samostalniškim določilom

ime programa	Luščenje zvez pridevnika s samostalniškim določilom.pl
namen programa	Program iz podkorpusa izlušči podatke, ki ustrezajo izbranemu vzorčnemu tipu, in jih zapiše v ustrezno datoteko.
razlog za uporabo programa	Program se uporablja za luščenje besednih zvez pridevnika s samostalniškim določilom. Več informacij v V-1.1.6.

kratek opis delovanja	<p>V osnovi je program enak programu za luščenje zvez dveh samostalnikov, z ustrezno prilagoditvijo na ravni regularnega izraza ter zapisa ustreznih mest oblikoskladenjskih oznak v spremenljivke. Program preverja naslednje pogoje:</p> <ul style="list-style-type: none"> - sklona pridevnika in samostalnika sta enaka ($\\$sklon1 eq \\$sklon2$); - oba sklona sta v rodilniku ($(\\$sklon1 eq \\$sklon2) and (\\$sklon2 eq "r")$); - oba sklona sta v dajalniku ($(\\$sklon1 eq \\$sklon2) and (\\$sklon2 eq "d")$); - pridevnik ni v rodilniku, samostalnik je v rodilniku ($(\\$sklon1 ne \\$sklon2) and (\\$sklon2 eq "r")$); - pridevnik ni v dajalniku, samostalnik je v dajalniku ($(\\$sklon1 ne \\$sklon2) and (\\$sklon2 eq "d")$); - sklona nista enaka, samostalnik ni ne v rodilniku ne v dajalniku ($(\\$sklon1 ne \\$sklon2) and (\\$sklon2 ne "r") and (\\$sklon2 ne "d")$).
-----------------------	--

3.3.4 Samostalnik s pridevniškim določilom

ime programa	Luščenje zvez samostalnika s pridevniškim določilom.pl
namen programa	Program iz podkorpusa izlušči podatke, ki ustrezajo izbranemu vzorčnemu tipu, in jih zapiše v ustrezno datoteko.
razlog za uporabo programa	Program se uporablja za luščenje besednih zvez samostalnika z levim pridevniškim določilom, pri čemer je glavni kriterij ujemanje v spolu, sklonu ter številu. Več informacij v V-1.1.4.
kratek opis delovanja	<p>V osnovi je program enak programu za luščenje zvez dveh samostalnikov, z ustrezno prilagoditvijo na ravni regularnega izraza ter zapisa ustreznih mest oblikoskladenjskih oznak v spremenljivke. Program preverja naslednji pogoj:</p> <ul style="list-style-type: none"> - pridevnik in sledeči samostalnik se ujemata v sklonu, spolu in številu ($(\\$sklon1 eq \\$sklon2) and (\\$spol1 eq \\$spol2) and (\\$stevilo1 eq \\$stevilo2)$).

ime programa	Luščenje zvez samostalnika s pridevniškim določilom (inverzija).pl
namen programa	Program iz podkorpusa izlušči podatke, ki ustrezajo izbranemu vzorčnemu tipu, in jih zapiše v ustrezno datoteko.
razlog za uporabo programa	Program se uporablja za luščenje besednih zvez samostalnika z levim pridevniškim določilom, kadar sta besedi v obrnjenem besednem redu; glavni kriterij je ujemanje v spolu, sklonu ter številu. Več informacij v V-1.1.4.
kratek opis delovanja	<p>V osnovi je program enak programu za luščenje zvez dveh samostalnikov, z ustrezno prilagoditvijo na ravni regularnega izraza ter zapisa ustreznih mest oblikoskladenjskih oznak v spremenljivke. Program preverja naslednji pogoj:</p> <ul style="list-style-type: none"> - samostalnik in sledeči pridevnik se ujemata v sklonu, spolu in številu ($(\\$sklon1 eq \\$sklon2) and (\\$spol1 eq \\$spol2) and (\\$stevilo1 eq \\$stevilo2)$).

3.3.5 Neujemalne kombinacije

ime programa	Luščenje neujemalnih kombinacij samostalnika ter pridevnika.pl
namen programa	Program iz podkorporusa izlušči podatke, ki ustrezajo izbranemu vzorčnemu tipu, in jih zapiše v ustrezno datoteko.
razlog za uporabo programa	Program se uporablja za luščenje kombinacij samostalnika s pridevnikom na levi, ki ne izkazuje ujemanja v spolu, sklonu in številu. Podatki so namenjeni preverjanju avtomatskega označevanja. Več informacij v V-1.1.4.
kratek opis delovanja	Program je enak programu <i>Luščenje zvez samostalnika s pridevniškim določilom</i> , le da zapisuje v datoteko primere, ki se <u>ne</u> ujemajo z danim pogojem (se torej ne ujemajo v spolu, sklonu in številu).

ime programa	Luščenje neujemalnih kombinacij dveh pridevnikov.pl
namen programa	Program iz podkorporusa izlušči podatke, ki ustrezajo izbranemu vzorčnemu tipu, in jih zapiše v ustrezno datoteko.
razlog za uporabo programa	Program se uporablja za luščenje kombinacij dveh pridevnikov, ki ne izkazujeta ujemanja v spolu, sklonu in številu. Podatki so namenjeni preverjanju avtomatskega označevanja. Več informacij v V-1.3.3.1.
kratek opis delovanja	Program je enak prejšnjemu opisanemu programu, le da zapisuje v datoteko primere, v katerih se ne ujemata dva pridevnika.

3.3.6 Priredne zveze

ime programa	Luščenje prirednih zvez.pl
namen programa	Program iz podkorporusa izlušči podatke, ki ustrezajo izbranemu vzorčnemu tipu, in jih zapiše v ustrezno datoteko.
razlog za uporabo programa	Program se uporablja za luščenje tridelnih prirednih besednih zvez, ki prinašajo dve (enakovrstni ter ujemajoči se) polnopomenski besedi, povezani s prirednim veznikom. Več informacij v V-3.
kratek opis delovanja	<ol style="list-style-type: none"> 1. Program po vrsti odpira, obdeluje ter zapira podkorporusne datoteke; 2. bere datoteko vrstico za vrstico in išče vse zadetke, ki ustrezajo regularnemu izrazu, v katerem so definirane informacije o obravnavani lemi ter oblikoskladenjskih oznakah; 3. v primeru najdenega zadetka razbije oblikoskladenjske oznake na posamezne znake, ki jih shrani v urejeni seznam; 4. del oznake, ki označuje sklon vsake od besed, shrani v spremenljivki, ki ju nato primerja med sabo; 5. v primeru, da sta oznaki za sklon enaki, piše podatke (leme, oznake ter besedne oblike) v ustrezno datoteko.
razlaga	Pripravljene so štiri različice programa, za vsako od obravnavanih besednih vrst.

Različici za luščenje prirednih zvez dveh glagolov oz. prislovov ne prinašata preverjanja ujemanja v sklonu.

ime programa	Luščenje prirednih zvez (neujemanje).pl
namen programa	Program iz podkorpora izlušči podatke, ki ustrezajo izbranemu vzorčnemu tipu, in jih zapiše v ustrezno datoteko.
razlog za uporabo programa	Program se uporablja za luščenje tridelnih prirednih besednih zvez, ki prinašajo dve (enakovrstni ter neujemajoči se) polnopomenski besedi, povezani s prirednim veznikom. Podatki so namenjeni preverjanju avtomatskega označevanja. Več informacij v V-3.2.1.1.
kratek opis delovanja	Program je enak programu <i>Luščenje koordinacij</i> , le da podatke v datoteko zapisuje v primeru, da se samostalnika oz. pridevnika, povezana s prirednim veznikom, v sklonu <u>ne</u> ujemata.

3.3.7 Štetje besednih nizov

ime programa	Števec vrstic.pl
namen programa	Program bere vrstice tekstovne datoteke, šteje ponovitve vrstic in jih ureja glede na pogostnost. Na koncu zapiše po pogostnosti urejen seznam v novo datoteko.
razlog za uporabo programa	Program se uporablja za urejanje izluščenih besednih nizov v pogostnostne seznane.
kratek opis delovanja	<ol style="list-style-type: none"> 1. Program odpre tekstovno datoteko, vsebujočo besedne nize; 2. bere vrstico za vrstico in vsako trenutno vrstico shrani v tabelo, v kateri pri vsaki novi pojavitvi ustrezni vzorčni zapolnitvi poveča vrednost za ena; 3. po koncu branja program uredi podatke glede na vrednost, tj. pogostnost, začenši z najvišjo pogostnostjo; 4. program zapiše seznam v izhodno datoteko.

3.4 Dodatna obravnava samostalniških zvez

3.4.1 Preverjanje relevantnosti vzorčnih tipov za luščenje samostalniških zvez

ime programa	Preverjanje relevantnosti – samostalniške zveze.pl
namen programa	Program lušči besedne zveze za vse vzorčne tipe, ki so bili v raziskavi definirani kot relevantni za luščenje samostalniških besednih zvez.
razlog za uporabo programa	Program se v raziskavi uporablja za evalvacijo uvrstitve vzorčnih tipov v tabelo relevantnih za luščenje samostalniških besednih zvez. Za luščenje se uporablja nabor novih samostalniških lem, ki jih uporabnik vnese na začetek programske kode. Več informacij v VI-2.2.2.
kratek opis delovanja	Program združuje luščenje različnih tipov samostalniških zvez, kakršno je bilo predstavljeno v prejšnjih poglavjih:

	<ol style="list-style-type: none"> 1. Uporabnik v programsko kodo vnese samostalniško lemo za obdelavo; 2. program po vrsti odpira, obdeluje ter zapira podkorporne datoteke; 3. bere datoteko vrstico za vrstico in išče vse zadetke, ki ustrezajo regularnemu izrazu, v katerem so definirane informacije o obravnavani lemi ter oblikoskladenjskih oznakah; 4. v primeru najdenega zadetka po potrebi (če luščenje za trenutno obravnavani tip besedne zveze to zahteva) razbije oblikoskladenjske oznake na posamezne znake, ki jih shrani v urejeni seznam in 5. ustrezne dele oznak shrani v spremenljivke, ki jih nato primerja med sabo; 6. v primeru izpolnjevanja danih pogojev zapiše najdene podatke (leme, oznake ter besedne oblike) v ustrezno datoteko (vsak tip besedne zveze v ločeno datoteko).
razlaga	<p>Pripravljeni sta dve različici programa, različica za dvodelne vzorce trenutno lušči 22, različica za tridelne pa 26 različnih tipov besednih nizov, kandidatov za besedne zveze.</p>

3.4.2 Odstranjevanje nerelevantnih vzorcev

ime programa	Odstranjevanje nerelevantnih vzorcev – popolna selekcija.pl
namen programa	Program s seznama vzorcev odstrani izbrani nabor vzorcev.
razlog za uporabo programa	Program se uporablja za selekcijo seznama vzorcev za nove obravnavane leme, z namenom identifikacije prej spregledanih vzorcev oz. vzorčnih tipov. Več informacij v V-6 ter VI-2.2.1.
kratek opis delovanja	<ol style="list-style-type: none"> 1. Program odpre seznam vzorcev za obravnavano lemo v tekstovni različici; 2. bere seznam vrstico za vrstico; 3. razen v primeru, da trenutno obravnavana vrstica prinaša vzorec, ki se ujema s katerim od navedenih v regularnem izrazu, program prepíše vrstico v novo datoteko.



ANALIZA

V nadaljevanju je predstavljena analiza najpogostejših vzorcev, ki vsebujejo katero od štirih obravnavanih lem: *pajek*, *strasten*, *plesati* ter *temeljito*. Glede na oblikoskladenjske oznake, ki jih obravnavani vzorci prinašajo, se poglavja ločijo na vzorce (I) s samimi polnopomenskimi besednimi vrstami, (II) predlogi ter (III) z vezniki, za tem pa so predstavljeni še vzorci z oblikoskladenjskimi oznakami za (IV) druge besedne vrste ter (V) nelematizirane besede.⁷⁴

Osnovna delitev poglavij je glede na obravnavano lemo, znotraj tega pa glede na dolžino vzorca (dvodelni ali tridelni vzorci). Glavnino poglavij predstavlja predstavitev nabora izluščenih besednih nizov ter analiza podatkov s stališča (I) njihove relevantnosti za vključitev v leksikalno zbirko, (II) upoštevanja oz. neupoštevanja različnih kategorij oblikoskladenjskih oznak tako pri luščenju kot pri končnem prikazu izluščenih podatkov, v posameznih primerih pa še (III) ustreznosti pripisanih oblikoskladenjskih oznak (glej IV-2.4). Natančnejše jezikoslovne analize izluščenih podatkov pričujoče delo ne prinaša (glej I-2).

Večina poglavja se osredotoča na analizo najpogostejših izpričanih vzorcev oz. vzorčnih tipov za posamezne leme. Dopolnitev s stališča širitve zornega kota raziskave tudi na manj pogoste vzorce prinaša podpoglavje V-6, kjer so obravnavani redkejši vzorci za primer samostalnika *pajek*.

Združene rezultate analize prinaša poglavje VI.

1 Vzorci s samimi polnopomenskimi besednimi vrstami

Med v pregledno analizo zajetimi najpogostejšimi vzorci za vsako od obravnavanih lem (glej IV-2.3) predstavljajo vzorci s samimi polnopomenskimi besedami naslednje deleže (prva številka v tabeli predstavlja število vzorcev s samimi polnopomenskimi besedami, druga vse ročno pregledane vzorce, sledi delež, zaokrožen na eno decimalno mesto):

	dvodelni vzorci	tridelni vzorci
<i>pajek</i>	65 (101) 64,4 %	10 (101) 9,9 %
<i>strasten</i>	70 (101) 69,3 %	24 (103) 23,3 %
<i>plesati</i>	61 (100) 61,0 %	8 (103) 7,8 %
<i>temeljito</i>	73 (100) 73,0 %	31 (101) 30,7 %

Tabela 26: Delež vzorcev s samimi polnopomenskimi besednimi vrstami med vsemi najpogostejšimi.

⁷⁴ Kot rečeno v poglavju IV-2.3.2, so bili iz celotnega nabora najprej izločeni vzorci z nelematiziranimi besedami, okrajšavami, členki, števniki itd., nato vzorci z vezniki oz. predlogi, tako da so na koncu na seznamu ostali le vzorci z oznakami za same polnopomenske besedne vrste. Izdelava vzorčnih skupin je torej potekala od predvidoma manj relevantnih vzorcev do bolj relevantnih. Analiza pa poteka – kot pričujoče poglavje – v obratnem vrstnem redu, od za luščenje predvidoma zanimivejših vzorcev do manj zanimivih.

1.1 Pajek – dvodelni vzorci

Najpogostejših 65 dvodelnih vzorcev z lemo *pajek* in oznako za polnopomensko besedo na levi ali desni lahko glede na besednovrstno opredelitev spremljajoče oblikoskladenjske oznake razvrstimo v osem vzorčnih tipov: lema *pajek* se pojavlja s samostalnikom, pridevnikom, prislovom ali (glavnim) glagolom⁷⁵ bodisi na levi ali desni strani. Število vzorcev, uvrščenih v posamezni vzorčni tip, prikazuje naslednja tabela:

vzorčni tip	število vzorcev
<i>pajek</i> + Sam ⁷⁶	10
Sam + <i>pajek</i>	19
<i>pajek</i> + Prid	3
Prid + <i>pajek</i>	15
<i>pajek</i> + Prisl	1
Prisl + <i>pajek</i>	1
<i>pajek</i> + Glag	9
Glag + <i>pajek</i>	7

Tabela 27: Razvrstitev najpogostejših vzorcev z lemo *pajek* in oblikoskladenjskimi oznakami za polnopomenske besede v vzorčne tipe.

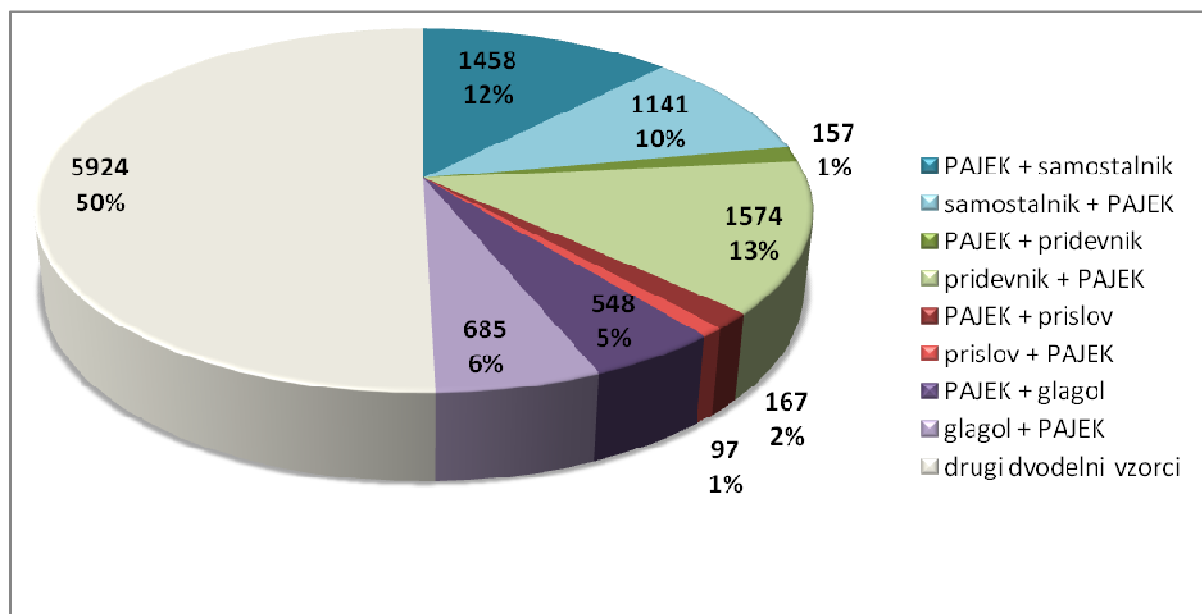
V nadaljevanju nas zanimajo podatki o pojavljanju evidentiranih vzorčnih tipov v celotnem korpusu. Iz narave oblikoskladenjskih oznak, ki prinašajo za različne besedne vrste različne nivoje členjenosti, sledi dejstvo, da je bolj od samega števila različnih vzorcev, ki jih v korpusu najdemo za posamezen vzorčni tip, pomenonosna informacija o številu korpusnih pojavitev vseh vzorcev posameznega tipa.⁷⁷ Spodnji graf zato prinaša informacijo o vsoti pogostnosti vzorcev, zajetih v posamezni vzorčni tip.

V grafično predstavitev so zajeti vsi dvodelni vzorci z lemo *pajek*. Kot je razvidno, predstavljajo vzorci s samimi polnopomenskimi besedami polovico celotnega nabora, kategorija *drugi dvodelni vzorci* pa drugo polovico nabora. Od obravnavanih vzorčnih tipov je v korpusu najpogostnejši Prid + *pajek* (13 %), sledita mu *pajek* + Sam (12 %) ter Sam + *pajek* (10 %). Kombinacije z glagolom so redkejše (6 % Glag + *pajek* oz. 5 % *pajek* + Glag), še manjša pa je pogostnost kombinacij s prislovom (2 % *pajek* + Prisl oz. 1 % Prisl + *pajek*). Samo 1 % predstavlja tudi vzorčni tip *pajek* + Prid.

⁷⁵ Glede na trenutno aktualen sistem oblikoskladenjskega označevanja se glagolske oblike označujejo kot *glavne* ali *pomožne*. Prve so zajete v obravnavo v pričujočem poglavju, druge pa v poglavju V-4.1.

⁷⁶ Vzorčni tipi so opredeljeni z oznakami za besedne vrste: **Sam** – samostalnik, **Prid** – pridevnik, **Glag** – glagol, **Prisl** – prislov. Podpisana številka opredeljuje sklon, npr. **Sam₂** – samostalnik v rodilniku, poševnica pa pomeni negacijo, npr. **Sam₂** – samostalnik v nerodilniku. Celotna legenda je navedena kot Priloga 3.

⁷⁷ Podatek, da se kombinacija leme *pajek* s samostalnikom na desni pojavlja v korpusu 1458-krat, je za pričujočo raziskavo bolj zanimiv od podatka, da se za omenjeni vzorčni tip v korpusu najde 29 različnih vzorcev.



Graf 1: Vsota pogostnosti pojavitev vzorcev za dvodelne vzorčne tipe z lemo *pajek* in oblikoskladenjskimi oznakami za polnopomenske besede.

V nadaljevanju poglavja sledi obravnava vzorcev glede na navedene vzorčne tipe, začenši s predstavitvijo metode luščenja zvez dveh samostalnikov.

1.1.1 Luščenje zvez Sam + Sam

Zveze dveh samostalnikov se v jezikoslovni literaturi ločujejo na več vrst. Po tipologiji *Slovenske slovnice* se izkazujejo (I) zveze samostalnika z **ujemalnim** desnim samostalniškim prilastkom ter (II) zveza samostalnika z **neujemalnim** (imenovalnim, rodilniškim, dajalniškim) desnim samostalniškim prilastkom (Toporišič 2004⁴: 560–561). Literatura pri obravnavi skladenjskih vzorcev za luščenje zvez dveh samostalnikov ločuje različne vrste izluščenih besednih nizov glede na sklon drugega od samostalnikov (glej npr. Vintar 1999: 169, Gantar 2007: 183).

Podobna delitev se izkazuje tudi na ravni korpusnih podatkov: na prvih mestih seznama vzorcev leme *pajek* s samostalnikom na desni prevladujejo kombinacije z drugim samostalnikom v imenovalniku ali rodilniku.⁷⁸ Izluščene vzorčne zapolnitve potrjujejo potrebo po ločevanju primerov glede na sklon oz. ne-/ujemanje samostalnikov, kar zahteva posebno pozornost na ravni izdelave metode avtomatskega luščenja.

Avtomatsko razporejanje zvez dveh samostalnikov v želene skupine je mogoče na osnovi upoštevanja oblikoskladenjskih oznak obeh samostalnikov. Potencialno ujemanost je mogoče identificirati, kadar je oznaka za sklon obeh samostalnikov enaka, pri čemer je potrebna pozornost na mestu, kjer sta oba samostalnika v rodilniku ali dajalniku, saj so te zveze na ravni avtomatske prepoznavne dvoumne (prim. *pajka tarantele* ter *pajka vrste*). Neujemanost je možno identificirati v primerih, da se samostalnika v sklonu ne ujemata, obenem pa je desni samostalnik v rodilniku ali dajalniku.

Ker sistem oblikoskladenjskega označevanja z ločevanjem lastnih imen od občnih to dopušča, je možen tudi poskus identifikacije zvez z neujemalnim imenovalnim desnim prilastkom, pri čemer se dvoumnost izkazuje v primerih, da sta oba samostalnika v imenovalniku (prim. *pajek Franci* ter *pajek Pottinger*).

⁷⁸ Na prvem mestu pogostnostnega seznama je vzorec PAJEK SOMEI (595), na drugem mestu PAJEK SOZMR (560).

Za luščenje in razvrščanje zvez, ustrežajočih skladiškemu vzorcu Sam + Sam, je bil na osnovi predstavljenih izhodišč pripravljen poseben program (za natančnejši opis programa glej IV-3.3.2), ki iz besedil izlušči ter v sedem ločenih datotek zapiše vse kombinacije dveh zaporednih samostalnikov, ki se:

- A) ujemata v sklonu;
- B) ujemata v sklonu, ki je roditelj;
- C) ujemata v sklonu, ki je dajalnik;
- D) ne ujemata v sklonu, ob tem, da je drugi samostalni v roditelju;
- E) ne ujemata v sklonu, ob tem, da je drugi samostalni v dajalniku;
- F) ne ujemata v sklonu, ob tem, da je drugi samostalni v sklonu, ki ni roditelj ali dajalnik;
- G) ne ujemata v sklonu, ob tem, da je drugi samostalni lastno ime.

Za uvrstitev v leksikalno zbirko relevantni besedni nizi so pričakovani kot rezultat luščenja pri kombinacijah A (tip *pajek tarantela*), D (tip *vrsta pajkov*) ter E (tip *podpora Slovenkam*). Kombinaciji B in C sta namenjeni filtriranju primerov, pri katerih iz oznake sklon ni mogoče avtomatsko ugotoviti, za kateri tip besedne zveze gre. Primere, ki se pojavljajo na seznamih B ter C, bi bilo mogoče deloma avtomatsko razporediti k primerom seznama A ter D oz. E glede na nabor besednih nizov, ki se v slednjih seznamih že pojavljajo (npr. zveza *pajka tarantele* se avtomatsko razvrsti na seznam A, ker se tam na relativno visokem mestu že pojavlja zveza *pajek tarantela*). O predlogu razvrščanja je več napisanega v poglavju V-1.1.2.

Kombinacije tipa G prinašajo nabor samostalnikov, ki jim sledi samostalni, označen za lastno ime (npr. *pajek SPIDER*, *pajek Krivograd* itd.), vendar samo v primerih, da se samostalnika v sklonu ne ujemata: na ta način bi bilo teoretično mogoče ločiti zveze z neujemalnim prilastkom tipa *pajek SIP* od zvez z ujemalnim prilastkom tipa *pajek Franci*. V praksi se izkazuje v povezavi s to metodo vsaj dva problema: (I) omejenost priklica zelenih zvez z ravnijo kvalitete označevanja lastnih imen (oz. iz neevidentirane lastnoimenskosti izvirajoče napačne lematizacije) ter (II) že omenjena dvoumnost zvez na ravni imenovalniškega sklonu.

Kombinacije tipa F so namenjene luščenju primerov, ki z neujemalnostjo, ki pa ne izpričuje roditelja ali dajalnika za drugi samostalni, ne padejo v nobeno od predvidenih kategorij; ti primeri zahtevajo predvsem analizo s stališča označenosti.

1.1.2 Pajek + Sam

Pričujoče poglavje prinaša rezultate luščenja kombinacij samostalnika *pajek* ter potencialnega desnega samostalniškega prilastka. Pridobljene zveze⁷⁹ so urejene glede na vrsto luščene kombinacije, kakor so bile slednje opisane v prejšnjem poglavju.

Prvi seznam prinaša kombinacije, pri katerih se samostalnika ujemata v sklonu. V seznamu sta oba samostalnika lematizirana, kar je tudi ustrezna prikazna oblika⁸⁰ za obravnavane podatke:

besedni niz		pogostnost v korpusu	
pajek	sip	370	pajek VO 5
pajek	spider	68	pajek klatež 5
pajek	pottinger	30	pajek spajder 4
pajek	skakač	20	pajek SRO 4
pajek	zaklopničar	12	pajek FAHR 4
pajek	križevac	10	pajek samuraj 3
pajek	Pottinger	9	pajek vrsta 3

⁷⁹ V nadaljevanju so v tabelah besednih nizov vedno navedeni primeri, katerih pogostnost v korpusu je 3 ali več.

⁸⁰ Kot omenjeno v II-2.2, je **prikazna oblika** podatkov oblika, v kateri je izluščeni besedni niz poslan v nadaljnjo ročno obravnavo.

pajek	obračalnik	8	pajek	SPIDER	3
pajek	tarantela	7	pajek	spaider	3
pajek	Franci	7	pajek	tip	3
pajek	Jus	7	pajek	Jernej	3
pajek	Krivograd	7	pajek	FONTANESI	3
pajek	lijakar	6	pajek	TKM	3
pajek	križavec	6			

Tabela 28: Izluščeni podatki *pajek* + *Sam* – ujemalne zveze.

Sledi seznam kombinacij samostalnika *pajek* v nerodilniku (pregled oznak priča, da gre večinoma za imenovalnik, kar potrjuje predvidevanja) ter sledečega samostalnika v rodilniku. Ustrezna prikazna oblika za tovrstne podatke je prikaz prvega samostalnika v lematizirani obliki, drugega pa v ustrezni rodilniški obliki. K problemu napačnega označevanja (npr. *sip*, *kot*, *od* itd.) se vračamo v poglavju V-1.1.3.

besedni niz	pogostnost v korpusu	pajek	krone	6
pajek SIP	453	pajek	vrste	3
pajek sip	82	pajek	DEUTZ	3
pajek širine	23	pajek	vozil	3
pajek SPIDER	9	pajek	POTTINGER	3
pajek Sip	9	pajek	FAHR	3
pajek kot	7	pajek	od	3

Tabela 29: Izluščeni podatki *pajek*₂ + *Sam*₂.

Rezultat luščenja zvez, kjer sta oba samostalnika v rodilniku, prikazuje tabela v nadaljevanju. Podatke je mogoče z upoštevanjem ustrezne prikazne oblike primerjati⁸¹ z nizi v Tabeli 28 oz. Tabeli 29. Barvna shema tabele je razložena v nadaljevanju:

zveza v rodilniku	pogostnost v korpusu		
pajka skakača	3	pajka velikosti	1
pajka Francija	2	PAJKOV CECCATO	1
pajkov križevcev	2	pajkov regije	1
pajkov tarantel	2	pajkov zaklopničarjev	1
pajka tarantele	2	pajkov rakovičarjev	1
pajka vrste	2	pajka volkca	1
pajka od	1	pajkov vrste	1
pajkov baldahinarjev	1	pajka niti	1
pajka lijakarja	1	pajkov skakačev	1

Tabela 30: Izluščeni podatki *pajek*₂ + *Sam*₂.

V gornji tabeli so besedni nizi, ki obenem obstajajo v Tabeli 28, obarvani svetlo sivo, nizi, ki obstajajo v Tabeli 29, pa temno sivo.⁸² Neobarvana sta dva primera, *pajka velikosti* ter *pajkov regije*, ki se v korpusu pojavljata samo enkrat, kar pomeni, da z opisanim postopkom ni mogoče ugotoviti, za kateri tip zveze gre; tovrstni primeri pri obravnavi sicer ne povzročajo velikih težav, saj zaradi nizke pogostnosti ostajajo na obrobju interesa.

⁸¹ Podatki so na tem mestu primerjani ročno. Pred razvojem avtomatskega razvrščanja bi bilo smotno predlagani postopek preveriti na širšem naboru podatkov, da se ugotovi njegova zanesljivost – predvsem za nizkopogostne besedne zveze.

⁸² V Tabelah 28 in 29 so sicer predstavljeni le nizi s pogostnostjo 3 ali več, vendar je smiselno razvrstitve izvajati z upoštevanjem celotnih seznamov nizov.

Ker Tabela 28 vsebuje vse besedne nize, ki glede na oblikoskladenjske oznake izkazujejo ujemanje v sklonu, vsebuje v celoti tudi primere Tabele 30. Prerazvrščanje podatkov zato poteka na sledeči način: (I) v primeru, da se primer iz Tabele 30 pojavlja edino ali relevantno pogostejše v Tabeli 28 (npr. *pajek tarantela* – *pajka tarantele*), je zgolj odstranjen iz Tabele 30, (II) če se primer iz Tabele 30 pojavlja v Tabeli 29 pogostejše kot v Tabeli 28 (npr. *pajek od*), je izbrisan iz Tabele 30, pogostnost niza je odšteteta iz Tabele 29 ter prišteta v Tabeli 30. Enaka je obravnava, če se primer v Tabelah 28 in 29 pojavlja z enako pogostnostjo (npr. *pajek vrste*).

Po razvrstitvi podatkov je stanje v obravnavanih seznamih naslednje (spremembe so označene podčrtano):

besedni niz	pogostnost v korpusu		
pajek sip	370	pajek SRO	4
pajek spider	68	pajek FAHR	4
pajek pottinger	30	pajek samuraj	3
<u>pajek skakač</u>	<u>24</u>	pajek vrsta	3
<u>pajek zaklopničar</u>	<u>13</u>	pajek SPIDER	3
<u>pajek križavec</u>	<u>12</u>	pajek spaider	3
pajek Pottinger	9	pajek tip	3
pajek obračalnik	8	pajek Jernej	3
<u>pajek tarantela</u>	<u>13</u>	pajek FONTANESI	3
<u>pajek Franci</u>	<u>9</u>	pajek TKM	3
pajek Jus	7	<u>pajek baldahinar</u>	<u>3</u>
pajek Krivograd	7	<u>pajek CECCATO</u>	<u>3</u>
<u>pajek lijakar</u>	<u>7</u>	<u>pajek rakovičar</u>	<u>3</u>
pajek križavec	6	<u>pajek volkec</u>	<u>3</u>
pajek VO	5		
pajek klatež	5		
pajek spajder	4		

Tabela 31: Razvrščeni podatki *pajek* + Sam – ujemalne zveze.

besedni niz	pogostnost v korpusu		
pajek SIP	453	pajek krone	6
pajek sip	82	<u>pajek vrste</u>	<u>6</u>
pajek širine	23	pajek DEUTZ	3
pajek SPIDER	9	pajek vozil	3
pajek Sip	9	pajek POTTINGER	3
pajek kot	7	pajek FAHR	3
		<u>pajek od</u>	<u>4</u>

Tabela 32: Razvrščeni podatki *pajek* + Sam₂.

Novi podatki prinašajo izboljšavo na ravni (I) ustrezne uvrstitve besednega niza glede na pogostnost (npr. uvrstitev niza *pajek tarantela* na višje mesto tabele), (II) uvrstitev novih primerov med najpogostejše (zaradi povišanja pogostnosti so v Tabelo 31 uvrščeni primeri *pajek [baldahinar, CECCATO, rakovičar, volkec]*), (III) odstranitev neustreznih primerov s seznama (iz Tabele 31 je denimo odstranjen niz *pajek vrsta*).

Po enakem postopku, kot je prikazano za primer luščenja zvez z neujemalnim rodilniškim prilastkom, je mogoče obravnavati zveze z neujemalnim dajalniškim prilastkom. Primerov tega tipa je sicer za obravnavani primer bistveno manj.⁸³ S pogostnostjo 3 ali več se v korpusu pojavlja le zveza *pajek olivi* (18), ki pa je posledica napačne lematizacije: drugi samostalni v tej zvezi označuje ime stroja *Olivi*, tj. v navedenem primeru ni v dajalniškem sklonu, temveč v imenovalniku. Prav tako je najti en sam primer pri luščenju primerov z obema samostalnika v dajalniku, tj. *pajku dežnikarju* (1).

⁸³ V uvodu poglavja je bila kot primer tega tipa navedena zveza *podpora Slovenkam*, ker za samostalni *pajek* ni najti relevantnega primera.

Za analizo ostanejo še primeri, kjer je drugi samostalnik označen kot lastno ime.⁸⁴ Zaradi problemov na ravni označevanja lastnih imen je predvidena manjša relevantnost pridobljenih tovrstnih podatkov (k problemom označevanja se vračamo v poglavju V-1.1.3). Spodnji seznam je naveden v celoti:

besedni niz	pogostnost v korpusu				
pajek	Francija	14	pajek	VO	1
pajek	SPIDER	9	pajek	DAROS	1
pajek	Šmalc	3	pajek	SPAJDER	1
pajek	POTTINGER	3	pajek	Hans	1
pajek	DEUTZ	3	pajek	MAZZOTTI	1
pajek	FAHR	3	pajek	Krivograd	1
pajek	DVOVRETENSKI	2	pajek	TKM	1
pajek	FONATANESI	1	pajek	Louis	1
pajek	Juh	1	pajek	Dora	1

Tabela 33: Izluščeni podatki *pajek* + Sam_{LI} – neujemalni.

Za primerjavo je v celoti naveden še seznam vseh primerov za vzorčni tip *pajek* + Sam_{LI}.

besedni niz	pogostnost v korpusu				
pajek	Francija	14	pajek	POTRESAČ	1
pajek	SPIDER	12	pajek	MARANGON	1
pajek	Krivograd	8	pajek	FONATANESI	1
pajek	Franci	7	pajek	FCR	1
pajek	Jus	7	pajek	Rus	1
pajek	FAHR	7	pajek	Draga	1
pajek	VO	6	pajek	ITL	1
pajek	DEUTZ	5	pajek	RO	1
pajek	POTTINGER	5	pajek	Maja	1
pajek	SRO	4	pajek	Dora	1
pajek	TKM	4	pajek	Louis	1
pajek	Šmalc	3	pajek	DISDERID	1
pajek	Jernej	3	pajek	Zoran	1
pajek	FONTANESI	3	pajek	Juh	1
pajek	Henry	2	pajek	Mitja	1
pajek	DAROS	2	pajek	Marko	1
pajek	SPAJDER	2	pajek	Hans	1
pajek	Srečko	2	pajek	MAZZOTTI	1
pajek	CECCATO	2	pajek	Bola	1
pajek	DVOVRETENSKI	2			

Tabela 34: Izluščeni podatki *pajek* + Sam_{LI} – vsi.

Kot je razvidno iz obeh tabel, so glede ujemanosti prilastka zveze tako (I) neujemalne, predvsem gre za imena delovnih strojev, npr. *pajek* [*Spider*, *Krivograd*, *Fahr*, *Deutz*, *Pottinger*], kot (II) ujemalne, kjer se beseda *pajek* pojavlja kot priimek, npr. *Pajek* [*Jernej*, *Hans*, *Dora*, *Maja*, *Mitja* ...]. Ker so zveze večinoma redke, je avtomatsko ločevanje med enim ter drugimi težka naloga – ki jo, kot rečeno, otežujejo tudi problemi na ravni lematizacije besednih oblik.⁸⁵

Kot omenjeno v poglavju II-1.2.1, poteka lematizacija besedil primarno na osnovi leksikonskih podatkov, delno pa tudi z ugibanjem leme na osnovi besednih končnic. S stališča označevanja lastnih imen je takšna metoda v

⁸⁴ Za označevanje lastnoimenskega samostalnika pri zapisu vzorčnih tipov uporabljamo oznako Sam_{LI}.

⁸⁵ Zlasti če se te zapisujejo tako raznovrstno, kot izkazuje obravnavani primer – *SIP*, *Sip*, *spider*, *Spider*, *spajder*, *spajder*, *SPIDER* ...

osnovi uspešna toliko, kolikor je leksikon na ravni prinos lastnih imen bogat. Ker lastna imena predstavljajo na ravni leksikona težko opisljivo skupino, obenem pa tvorijo pomemben delež besedil v naravnem jeziku, se je v sklopu obdelave naravnega jezika oblikovalo ločeno raziskovalno področje, ki se osredotoča na prepoznavo lastnih imen (ang. *named entity recognition*, za uvod v temo glej npr. Mitkov (ur.) 2003: 545–550). Pred razvojem sistema za avtomatsko prepoznavo lastnih imen⁸⁶ za slovenski jezik se kaže kot najbolj smiselna (čeprav začasna, zamudna ter dokaj omejena) rešitev vključevanje najpogostnejših (med luščenjem zvez evidentiranih) primerov lastnoimenskih zvez v leksikalno zbirko za vsak obravnavani primer sproti.

1.1.2.1 Analiza označenosti

Že omenjeni problemi avtomatskega označevanja so vezani na poimenovanje znamk delovnih strojev, npr. *pajek* [*SIP*, *Olivi*, *Pottinger*], za katere v leksikalni zbirki pred označevanjem korpusa ni bilo ustreznih vnosov. Lematizacija je posledično v več primerih neustrezna, pogosto temelji na obstoječih občnoimenskih lemah – *SIP* se lematizira v samostalnik *sip* ali *sipa*, *Olivi* v *oliva*. Podobno je v leksikalni zbirki neevidentirana uporaba samostalnika *pajek* kot priimka, kar pomeni, da se vsi primeri pojavitve priimka lematizirajo, kot da gre za občno ime (temu primerno ima beseda *pajek* v izluščenih besednih nizih vedno malo začetnico).

Naslednja tema analize označenosti so besedni nizi, v poglavju V-1.1.1 predstavljeni kot »tip F« – gre za kombinacije samostalnikov, ki glede na oblikoskladenjske oznake ne ustrezajo nobeni od pričakovanih besednozveznih kategorij, zato zahtevajo pozornost s stališča ustreznosti označevanja. Primeri – za luščenje kombinacij samostalnika *pajek* s sledečim samostalnikom jih je 89 – so v nadaljevanju navedeni glede na vzorec, v sklopu katerega se pojavljajo, navedeni pa sta tako izpričani kot tudi lematizirani obliki samostalnikov:

vzorec		lematizirani besedni niz		izpričani besedni niz		pogostnost
Sometd	Slzei	pajek	Francija	pajka	Francija	14
Sometd	Somei	pajek	sip	PAJKA	SIP	6
		pajek	sip	pajka	Sip	3
		pajek	Šmalc	PAJKA	Šmalc	3
Sometd	Slmei	pajek	strup	pajkov	strup	3
Sommr	Somei	pajek	tarantela	pajka	tarantelo	2
Somdi	Sozet	pajek	vozilo	pajek	vozilo	2
Somei	Sozet	pajek	muha	pajek	muho	2
		pajek	mreža	pajek	mrežo	2
Someo	Somei	pajek	kot	pajkom	kot	2
Somer	Somei	pajek	bolnik	pajka	bolnik	2
Sommr	Somei	pajek	kot	pajkov	kot	2
		pajek	kadaver	pajkov	kadaver	2
Somed	Sozet	pajek	plezanje	pajku	plezanje	1
Somed	Sozet	pajek	noga	pajku	nogo	1
Somei	Slmetd	pajek	Juh	pajek	Juha	1
Somei	Slmmt	pajek	Louis	pajek	Louise	1
Somei	Sometn	pajek	avto	Pajek	avto	1
Somei	Sozet	pajek	kosilo	pajek	kosilo	1
		pajek	uničenje	pajek	uničenje	1
		pajek	srečka	Pajek	srečko	1
		pajek	vrnitev	pajek	vrnitev	1
		pajek	ura	Pajek	URO	1
		pajek	sreča	pajek	srečo	1
		pajek	vaba	Pajek	vabo	1
		pajek	ura	pajek	uro	1
		pajek	nitka	Pajek	nitko	1

⁸⁶ Poskus opisujeta denimo Arčan in Vintar 2006.

		pajek	ura	pajek	URO	1
Somei	Sozmt	pajek	oči	pajek	oči	1
Someo	Somei	pajek	velikan	pajkom	velikan	1
Someo	Sommi	pajek	meščan	pajkom	meščani	1
Someo	Sommt	pajek	avtomobil	pajkom	avtomobile	1
Someo	Sosei	pajek	prebivalstvo	pajkom	prebivalstvo	1
Someo	Sozei	pajek	policija	pajkom	policija	1
		pajek	država	pajkom	država	1
Somer	Sometd	pajek	klatež	pajka	klateža	1
Somer	Sozei	pajek	nit	pajka	nit	1
Sometd	Somei	pajek	sip	pajka	SIP	1
Sometd	Sozei	pajek	igra	pajka	igra	1
Sometn	Somei	pajek	sip	pajek	SIP	1
		pajek	spider	pajek	spider	1
Sommi	Sozet	pajek	pajčevina	pajki	pajčevino	1
Sommm	Somei	pajek	obisk	pajkih	obisk	1
Sommo	Sommi	pajek	delavec	pajki	delavci	1
		pajek	skakač	pajki	skakači	1
		pajek	volkec	pajki	volkci	1
Sommr	Slmei	pajek	Hans	pajkov	Hans	1
Sommr	Somei	pajek	gen	pajkov	gen	1
		pajek	ples	Pajkov	ples	1
		pajek	izloček	pajkov	izloček	1
		pajek	svet	Pajkov	svet	1
		pajek	zadek	pajkov	zadek	1
Sommr	Sommi	pajek	samec	pajkov	samci	1
Sommt	Slmei	pajek	Krivograd	pajke	krivograd	1
Sommt	Somei	pajek	sip	PAJKE	Sip	1
		pajek	sip	pajke	SIP	1
Sommt	Sozei	pajek	večina	Pajke	večina	1

Tabela 35: Izluščeni podatki *pajek* + *Sam* – analiza označenosti.

Kot je razvidno iz tabele, so napake oznak različnih vrst. Najvišje v seznamu je napačna lematizacija imena *Franci* v *Francija*, veliko problemov povzročajo že omenjena imena znamk (*SIP*, *spider*), pogosta pa je tudi napačna lematizacija pridevnika *pajkov* v samostalniško lemo, npr. *pajkov* [*strup*, *kadaver*, *izloček*]. Pozornost je potrebna še na ravni napačnega označevanja zvez tipa *pajki* [*skakači*, *volkci*] ter *pajka* [*tarantelo*, *klateža*].

Večini navedenih napak bi se bilo mogoče izogniti na podlagi vnosa tipičnih zvez v leksikalno zbirko ter upoštevanja besednozveznih podatkov pri nadaljnjem označevanju, o čemer bo v nadaljevanju še govora.

Ustrezno lematizirane, vendar glede vnosa v leksikalno zbirko manj zanimive so denimo kombinacije samostalnika *pajek* (največkrat v imenovalniku) ter sledečega samostalnika v tožilniku, npr. *pajek* [*muho*, *oči*, *vozilo*], *pajku* [*nogo*] itd.

Na več mestih v tem poglavju se kaže za problematično tudi lematizacija nekaterih funkcijskih besed v samostalnike (*pajek* [*kot*, *od*]). Tudi ta problem je rešljiv le z upoštevanjem konteksta pri označevanju, o čemer je več napisanega v V-1.3.1.1. Za primer dejanske rabe, ki priča o neustreznosti samostalniških oznak za *kot* v besednih nizih *pajek kot*, so na tem mestu navedeni zadetki iz korpusa FidaPLUS.⁸⁷

⁸⁷ Iskalni pogoj, uporabljen za pridobivanje primerov iz korpusa, je opredeljen v imenu posameznega konkordančnega niza. V iskalnem pogoju uporabljena sintaksa je razložena denimo v Arhar 2007.

in nezanimivo. Povedano drugače, če film o zlobnih **pajkih kot** za šalo preživi kritik, ki se boji celo prijaznih stavek se je glasil, da sem proti lisicam in **pajkom kot** edinemu ukrepu za urejanje prometa. Zato smo pred npr. tropske stonoge, ose stezičarke, ki uporablja **pajka kot** gostitelja svojih ličink in rdečega nosatega medveda. v resnici dejala, da je proti uvedbi lisic in **pajkov kot** edinima sredstvoma. Novi odlok uvaja rumene pentlje, nekoč Himalajo dvignile kontinentalne plošče, bodo preskrbljene z vodnimi **pajki kot** svojo hrano. A ta simbioza si je upal tudi sam poskusiti. Potem pa so **pajki kot** za stavo skakali z ene krste na drugo. in so kljub videzu bolj v sorodu s škorpijoni in **pajki kot** z raki. Pogosto jih imenujejo živi fosili in vrsta tem območju policisti prisotni večkrat kot drugje - tako s **pajki kot** z drugimi sredstvi. "Z redarji smo imeli na past za žužke. Če pa vstopi kaj nevarnega, **pajek kot** po čudežu izgine.

Konkordančni niz FidaPLUS 1: #1pajek_kot#25.

1.1.3 Sam + pajek

Pregled izluščenih vzorcev, ki spadajo v na tem mestu obravnavani vzorčni tip, kaže, da več kot polovica zapolnitev odpade na primere, kjer se samostalnik *pajek* pojavlja za samostalnikom v imenovalniku (SOMEI PAJEK – 427-krat v korpusu). V nadaljevanju so najprej predstavljeni primeri zvez, ujemajočih se v sklonu (Tabela 36), nato pa še zvez s samostalnikom *pajek* v rodilniku (Tabela 37).

Podatki so že prerazvrščeni na način, opisan v V-1.1.2. Kljub temu v spodnji tabeli še ostajajo nekatere napačno uvrščene zveze (označene s sivo barvo):⁸⁸

besedni niz		pogostnost v korpusu			
obračalnik	pajek	191	DVOVRETENSKI	pajek	4
človek	pajek	190	podoba	pajek	4
mož	pajek	18	sip	pajek	4
Lidija	pajek	12	Vladimir	pajek	4
Bogdan	pajek	10	voznik	pajek	4
Samo	pajek	10	skupina	pajek	3
ženska	pajek	7	Janez	pajek	3
Olga	pajek	7	odrasla	pajek	3
Mirko	pajek	6	osat	pajek	3
ŠTIRIVRETENSKI	pajek	5	Maks	pajek	3

Tabela 36: Razvrščeni podatki Sam + pajek – ujemalne zveze.

Sicer v Tabeli 36 glede na pogostnost v korpusu močno prednjačita zvezi *obračalnik pajek* (191)⁸⁹ ter *človek pajek* (190), veliko pa je tudi na tem mestu primerov, iz katerih je razvidno, da se beseda *pajek* v slovenščini pojavlja tudi kot priimek, tj. [*Lidija, Bogdan, Samo, Olga, Mirko, Vladimir, Janez, Maks*] *pajek*.

⁸⁸ Deloma gre to pripisati sovpadu zvez glede na slovnično število – ker so v Tabeli 36 samostalniki lematizirani, sta zvezi *voznik pajka* ter *voznik pajkov* združeni v zvezi *voznik pajek*, s posledično združeno pogostnostjo. V Tabeli 37 samostalnik *pajek* ni lematiziran, kar pomeni, da so nizi z različnim slovničnim številom obravnavani kot različne enote.

⁸⁹ Zveza *obračalnik pajek* je sicer tipičen primer, ki mu je avtomatsko težko določiti vrsto glede na ujemalnost prilastka, saj se v korpusu pojavlja le v imenovalniški ali tožilniški obliki.

besedni niz		pogostnost v korpusu			
vrsta	pajkov	43	odprava	pajkov	4
oblika	pajka	10	gnezdo	pajkov	4
ugriz	pajkov	5	samec	pajka	4
večina	pajkov	5	strup	pajka	4
uporaba	pajka	5	napad	pajka	3
mreža	pajkov	5	pomoč	pajkov	3
uvedba	pajka	5	sorodnik	pajkov	3
ukinitev	pajka	5	zbirka	pajkov	3
življenje	pajkov	5	raziskovalec	pajkov	3
poznavalec	pajkov	5	skupina	pajkov	3
posredovanje	pajka	5	število	pajkov	3
kot	pajka	5	uporaba	pajkov	3
preučevanje	pajkov	4	množica	pajkov	3

Tabela 37: Izluščeni podatki Sam + *pajek*₂.

Na tem mestu se postavlja pomembno vprašanje upoštevanja slovničnega števila kot ločevalnega za ločevanje besednih nizov obravnavanega tipa: iz gornje tabele so na eni strani razvidni primeri, ki upravičujejo ločevanje zvez glede na število samostalnika oz. ohranitev drugega samostalnika v množini – denimo [*množica, vrsta, večina, zbirka*] *pajkov* itd. Na drugi strani vidimo tudi primere, kjer ločevanje ni smotno, npr. *uporaba pajka / uporaba pajkov*. Za slednji tip nizov je zaželen prikazna oblika, ki združuje oba niza pod edninsko obliko.

Možen poskus avtomatskega združevanja temelji na upoštevanju prekrivnosti jedrnega samostalnika v obravnavanih primerih: če se na seznamu pojavljata dva niza z enakim prvim samostalnikom in drugim, ki izpričuje različne oblike glede na število, se oba primera združita in seštejeta pod obliko za ednino.⁹⁰ V spodnji tabeli so prikazani na opisani način dopolnjeni podatki, spremembe so označene podčrtano.

besedni niz		pogostnost v korpusu			
vrsta	<u>pajka</u>	<u>45</u>	napad	pajka	<u>4</u>
oblika	pajka	10	<u>voznik</u>	<u>pajka</u>	<u>4</u>
uporaba	pajka	<u>8</u>	<u>samica</u>	<u>pajka</u>	<u>4</u>
posredovanje	pajka	<u>6</u>	pomoč	<u>pajka</u>	<u>4</u>
samec	pajka	<u>6</u>	sorodnik	pajkov	3
mreža	<u>pajka</u>	<u>6</u>	zbirka	pajkov	3
večina	pajkov	5	raziskovalec	pajkov	3
ugriz	pajkov	5	skupina	pajkov	3
življenje	pajkov	5	število	pajkov	3
poznavalec	pajkov	5	uporaba	pajkov	<u>3</u>
uvedba	pajka	5	množica	pajkov	3
ukinitev	pajka	5	<u>slika</u>	<u>pajka</u>	<u>3</u>
kot	pajka	5	<u>gen</u>	<u>pajka</u>	<u>3</u>
strup	pajka	<u>5</u>			
preučevanje	pajkov	4			
odprava	pajkov	4			
gnezdo	pajkov	4			

Tabela 38: Dopolnjeni podatki Sam + *pajek*₂.

⁹⁰ Postopek še ni preverjen na večji količini gradiva, zato podatki o zanesljivosti oz. dejanski uporabnosti niso na voljo. Na tem mestu so podatki razvrščeni ročno.

Kot je vidno, so spremembe ponovno na mestih (I) uvrstitve (združenih) besednih nizov višje v pogostnostnem seznamu (npr. v primerih [*uporaba, posredovanje, samec*] pajka), (II) uvrstitev novih primerov na seznam (npr. [*slika, gen*] pajka) ter (III) odstranitev prerazvrščenih primerov s seznama (npr. *uporaba* pajkov).

Pomembna sprememba v tabeli je seveda združitev besednih nizov z enakim prvim samostalnikom – samo v edninski obliki se po novem pojavljajo primeri [*vrsta, mreža, voznik, samica, pomoč*] pajka. Obenem pregled nespremenjenih primerov kaže mesta, kjer bi bila sprememba zaželeno, vendar zaradi omejenih podatkov s predlaganim postopkom ne more biti izvedena, npr. [*ugriz, življenje, gnezdo, sorodnik*] pajkov. Za nekatere primere je ohranitev množinske oblike seveda ustrezna, npr. [*večina, zbirka, skupina, število, množica*] pajkov.

Besednih nizov ostalih tipov (s samostalnikom *pajek* v dajalniku ali označenim kot lastno ime) za na tem mestu obravnavani vzorčni tip korpusni podatki ne izpričujejo. Prav tako ne pokaže nič bistveno novega glede na ugotovitve v V-1.1.3 analiza pripisanih oblikoskladenjskih oznak.

1.1.4 Luščenje zvez Prid + Sam ter Sam + Prid

Za vzorčna tipa *pajek* + Prid ter Prid + *pajek* je bila izvedena krajša analiza relevantnosti v oblikoskladenjskih oznakah pridevniku pripisanih kategorij za luščenje.⁹¹ Podatki so v nadaljevanju na kratko predstavljeni, utemeljenost sprejetja določenih odločitev pa se preverja v nadaljevanju predstavitev analize pri vseh primerih besednih nizov, ki vsebujejo pridevnik.

Kar se tiče oznak **vrste** pridevnika⁹²: pri vzorčnem tipu *pajek* + Prid v korpusu prevladujejo kombinacije z besedami, označenimi za splošne pridevnike (135), manj je kombinacij z deležniškimi pridevniki (20) – npr. *pajka polnjene, pajku podaljšana, pajkom odpeljanih* itd. – ter svojilnimi pridevniki (2), tj. *pajek [Čopova, Arnesove]*. Kar se tiče **stopnje** pridevnika⁹³: med korpusnimi pojavitvami prevladujejo oznake nedoločene stopnje (153), redko se pojavljata primernik (3) – *pajek večji, pajki pogostejši, pajki učinkovitejši* – ter presežnik (1), tj. *pajki najpogostejši*.

Tudi pri vzorčnem tipu Prid + *pajek* v korpusu prevladujejo kombinacije z besedami, označenimi za splošne pridevnike (1471), bistveno manj je kombinacij z deležniškimi pridevniki (92) – npr. [*nošeni, podivjani, predramljeni, risani*] *pajek* itd. – ter svojilnimi (11) pridevniki, npr. [*Nigradov, Sipov*] *pajek*. Kar se tiče stopnje pridevnika, po pričakovanju ponovno prevladuje nedoločena stopnja (1537), redka sta primernik (28) – npr. [*starejši, manjši, močnejši*] *pajek* – ter presežnik (9), npr. [*najlepših, najnevarnejših, najmanjših*] *pajkov*. Iz navedenega je razvidno, da kategoriji vrsta in stopnja za avtomatsko luščenje nista posebej relevantni (pri čemer izhajamo iz predpostavke, da v korpusu redkeje zastopane oznake prinašajo za luščenje podatkov manj zanimive zapolnitve).

Kar se tiče označevanja **spola, sklona** ter **števila** se pri vzorčnem tipu *pajek* + Prid poleg pridevnikov moškega spola (105) na mestu za samostalnikom *pajek* pojavljajo tudi pridevniki ženskega (35) ter srednjega (17) spola. Analiza vzorčnih zapolnitev kaže, da besedni nizi s pridevniki, ki se s samostalnikom ne ujemajo v spol, sklonu ter številu, niso zanimivi za luščenje, ker v opisanih primerih pridevnik najverjetneje določa samostalnik na desni, npr. *pajek italijanske (proizvodnje), pajek črne (barve), pajki pralne (naprave)* itd. Podobno moški spol pridevnika prevladuje pri vzorčnem tipu Prid + *pajek*: poleg pridevnikov moškega spola, ki predstavljajo veliko

⁹¹ Označevalni sistem je na voljo kot Priloga 1, za natančnejše podatke o označevanju pridevnika po sistemu JOS glej oblikoskladenjske specifikacije <<http://nl.ijs.si/jos/msd/html-sl/msd.A.html>>.

⁹² Označevalni sistem JOS ločuje med *splošnimi*, *svojilnimi* ter *deležniškimi* pridevniki. Splošni pridevniki združujejo lastnostne ter vrstne, ker so ti zaradi enakopisnosti oblik v besedilu avtomatsko pogosto neločljivi.

⁹³ Označevalni sistem JOS ločuje med *nedoločeno stopnjo*, *primernikom* ter *presežnikom*. Nedoločena stopnja združuje označevanje nestopnjevanih pridevniških oblik s tistimi, ki so stopnjeване opisno (ko stopnje torej ne moremo avtomatsko prepoznati iz označevane besede same, temveč le na osnovi njene neposredne besedilne okolice). Kot primerniki in presežniki so označene oblike, ki stopnjo izražajo obrazilno.

večino (1549), se pred samostalnikom *pajek* pojavljajo tudi pridevniki srednjega (15) ter ženskega (10) spola (ki so, kot bo vidno v nadaljevanju, pogosto rezultat neustreznega označevanja).

Za luščenje so torej v prvi vrsti zanimivi pari samostalnika ter pridevnika, ki se ujemata v spolu, sklonu in številu, obenem pa podatkov ne želimo nadaljnje ločevati glede na oznake vrste ter stopnje pridevnika. Glede na ta izhodišča sta bila pripravljena dva specializirana programa (za opis glej poglavje IV-3.3.4). Za preverjanje ustreznosti označevanja je bil pripravljen še program, ki lušči neujemalne kombinacije (glej IV-3.3.5). V nadaljevanju sledijo rezultati luščenja za vzorčna tipa *pajek* + Prid ter Prid + *pajek*.

1.1.5 *Pajek* + Prid

Rezultati luščenja v spolu, sklonu ter številu ujemajočih se parov samostalnika *pajek* ter pridevnika na desni so naslednji (zaradi majhnega števila so navedeni vsi najdeni primeri):

besedni niz	pogostnost v korpusu	pajek	progast	1
pajek dvovretenski	25	pajek rdeč		1
pajek štirivretenski	12	pajek suh		1
pajek živ	3	pajek lačen		1
pajek nevaren	3	pajek roparski		1
pajek hidravlični	2	pajek plašen		1
pajek velik	2	pajek navaden		1
pajek strupen	2	pajek potreben		1
pajek dolgodlak	2	pajek pravi		1
pajek biološki	1	pajek izenačen		1
pajek Maman	1	pajek poln		1
pajek star	1	pajek neškodljiv		1
pajek kriv	1	pajek zoprni		1
pajek neopazen	1	pajek kosmat		1
pajek pogost	1	pajek zmožen		1

Tabela 39: Izluščeni podatki *pajek* + Prid.

Analiza nabora primerov skupaj z minimalnim besedilnim kontekstom kaže, da na tem mestu obravnavani vzorčni tip prinaša za vključitev v leksikalno zbirko potencialno zanimive podatke dveh tipov:

- (I) kombinacijo samostalnika *pajek* in določujočega ujemalnega pridevnika na desni – tj. levega prilastka na desni (Toporišič 2004⁹⁴: 562). Korpus FidaPLUS kaže, da se ta tip pojavlja predvsem v besedilnem žanru malih oglasov⁹⁴, prim. naslednje primere:

PAJEK dvovretenski italijanski, ugodno prodam. Tel.: 041/

PAJEK dvovretenski, rotacijska kosilnico z gnetilnikom, mlin za koruzo traktorski

PAJEK dvovretenski, z gibljivimi kolesi za manjši traktor, kupim.

PAJEK štirivretenski in rotacijsko kosilnico, prodam. Tel.: 041

PAJEK štirivretenski spider 350, še zapakiran z garancijo, prodam za

PAJEK štirivretenski, silokombajn, odjemalec silaže, trosilec tehnostroj in IMT

Konkordančni niz FidaPLUS 2: #1pajek_dvovretenski, #1pajek_štirivretenski.

⁹⁴ Mali oglasi iz časopisa *Kmečki glas* so v korpusu FidaPLUS izredno pogosti (tudi zaradi podvajanja zadetkov), kar pojasni visoko število besednih nizov obravnavanega tipa.

- (II) kombinacijo samostalnika *pajek* in ujemajočega pridevnika, ki se skladenjsko pojavlja v vlogi povedkovega določila, npr.:

Tako bo otrok razumljivo pomislil, da je *pajek nevaren*. Raje mu razložite, da je pajek neškodljiv,

Odkrili so, da je puščavski *pajek zmožen* pojesti samo proteine iz določene živali, če se je

Kobe pravi, da je *pajek potreben*, saj je že nekajkrat moral posredovati v nujnih primerih

Konkordančni niz FidaPLUS 3: #1pajek_nevaren, #1pajek_zmožen, #1pajek_potreben.

Oba predstavljena tipa besednih nizov sta avtomatsko do določene mere ločljiva od primerov, ko pridevnik določa samostalnik na svoji desni. Pod pogojem, da je pri obdelavi jezika upoštevan skladenjski kontekst, identifikacija samostalnika tako na levi kot na desni od obravnavanega pridevnika vodi v pripis večje verjetnosti povezanosti pridevnika z desnim samostalnikom:

SAMONAKLADALCO SIP 19, *pajek hidravlični* dvig, balirko in zgrabljajnik 380, prodam.

Vsaka zvita vrv postane kača, vsak *pajek strupen* stvor, vsaka čebela zlovešča grožnja smrti.

Fritz Vollrath in svilnati *pajek biološki* reaktor brezhibne ekobalance.

Konkordančni niz FidaPLUS 4: #1pajek_hidravlični, #1pajek_strupen, #1pajek_biološki.

Ker omenjeni postopek trenutno ni na voljo, je odločitev za upoštevanje besednih nizov, izluščenih za obravnavani vzorec, kot relevantnih za zbirko na strani graditelja zbirke. Nekoliko enostavnejša možnost avtomatskega določanja relevantnosti primerov Tabele 39 za vključitev v zbirko je tudi primerjava na tem mestu izluščenih kandidatov s seznamom besednih nizov za Prid + *pajek*. Postopek je osnovan na predvidevanju, da se (I) besedni pari, ki izpričujejo obrnjeni besedni red, v korpusu pojavljajo tudi v nezaznamovani obliki ter (II) da se pridevniki, ki vstopajo v povedkova določila, pojavljajo tudi kot levi prilastki samostalnika.

Spodnja tabela prinaša (ponovno ročno dopolnjene) podatke Tabele 39. Primeri, ki se ne pojavljajo tudi pri luščenju nizov za Prid + *pajek* (glej V-1.1.8), so v spodnji tabeli prečrtani. Pri tistih, ki se pojavljajo, pa je informativno navedena pogostnost pojavitev:

besedni niz	pogostnost v korpusu	pajek	progast	1 (1)
pajek dvovretenski	25 (29)	pajek rdeč		1 (230)
pajek štirivretenski	12 (42)	pajek suh		1
pajek živ	3 (9)	pajek lačen		1
pajek nevaren	3 (15)	pajek roparski		1
pajek hidravlični	2 (67)	pajek plašen		1 (1)
pajek velik	2 (53)	pajek navaden		1 (1)
pajek strupen	2 (88)	pajek potreben		1
pajek dolgodlak	2	pajek pravi		1 (2)
pajek biološki	1	pajek izenačen		1
pajek Maman	1	pajek poln		1
pajek star	1 (4)	pajek neškodljiv		1
pajek kriv	1	pajek zoprn		1
pajek neopazen	1	pajek kosmat		1 (17)
pajek pogost	1 (1)	pajek zmožen		1

Tabela 40: Razvrščeni podatki *pajek* + Prid.

Razvidno je, da so podatki po opravljeni primerjavi le delno primernejši za uvrstitev v leksikalno zbirko. Za ustrezno ločevanje med primeri bi potrebovali korpusne podatke, označene na skladiščnem nivoju. Na tem mestu je zanimivejši od samih izluščenih primerov za vzorčni tip *pajek* + Prid koligacijski potencial samostalnika – pri vnosu besednega niza *dvovretenski pajek* (29) je denimo mogoče v leksikalni zbirki označiti, da se besedi v korpusu pojavljata tudi v obrnjenem vrstnem redu – kar je pri izpričani pogostnosti 25 smiselno, za razliko od redkeje pojavljajočih se primerov, npr. *pajek* [*plašen, navaden, pravi, progast, pogost*].

1.1.5.1 Analiza označenosti

Na tem mestu raziskave so evidentirane predvsem napake označevanja na ravni ločevanja prislovnih oblik od pridevniških, kar je, kot bo izpostavljeno še v nadaljevanju dela (glej V-1.3.4.1), eden od glavnih problemov trenutnega oblikoskladiščnega označevanja. V potrditev trditve je na tem mestu naveden nabor korpusnih zadetkov za primere kombinacij besede *pajek* s pridevnikom srednjega spola na desni.

prometa, po katerem lahko nepravilno parkirana vozila odstranijo s **pajkom samo**, če vozilo neposredno ogroža promet, kakršno nam ne bo dovolilo pohoditi rastoče rastline ali ubiti **pajka samo** zato, ker sta tu. Razumimo torej, da redarji, ampak tatovi. Zato podatke o odvozi s **pajkom Javno** podjetje Parkirišča redno sporoča policistom, zdaj pa so jih se suhe južine, ali strokovno poimenovano opilioni, od **pajkov močno** razlikujejo! Njihovo glavoprsje široko prehaja v zadek,

2. Zakaj je za cvetnega **pajka koristno**, da lahko spreminja svojo barvo?

je v predvolilnem boju obljubljala, da bodo lisice in **pajki skrajno** sredstvo za kaznovanje tistih, ki so napačno parkirali. vsak dan, ko bo s **pajkom odpeljano** vozilo parkirano na Cesti dveh cesarjev, pa 320 tolarjev Ko pa sem jaz zraven, lahko postavim dva ptičja **pajka samo** pet centimetrov vsaksebi, a se ne bosta odzvala napadalno ali predvsem) tistega, ki je po mnenju voznika **pajka neprevilno** parkiran ali pa mu je potekel čas, ki ga so živali videti pri svojem delu vesele, urne. **Pajki mnogo** predejo ali na že naprednih mrežah čakajo. Rega pleza že lansko poletje, so bili lastnikom vklenjenih ali s **pajkom odpeljanih** vozil šele pred kratkim poslani prvi plačilni nalogi, domačega pajka. Skupina Nasinih znanstvenikov je leta 1995 dala **pajkom različna** znana mamila. Pajki, ki so zaužili kofein,

Konkordančni niz FidaPLUS 5: #1pajek_#2P??s*.

Poleg ustrezno označenih pridevniških primerov (*javno, skrajno, odpeljano, različna*) se pojavljajo tudi nepravilno označeni prislovi (*samo, koristno, mnogo*). Na prvi pogled dokaj preprosta, a predvideno zanesljiva metoda pri odločanju, ali pri reševanju dvoumnosti obravnavanega tipa pripisati pridevniško ali prislovno oznako, se zdi preverjanje obstoja ustreznega samostalnika na desni blizu dvoumne oblike. Slednji postopek bi bilo potrebno dodati na ustrezno mesto razdvoumljanja lematizacije (glej II-1.2.1), denimo na mesto prvega odstranjevanja manj verjetnih lem za obravnavano obliko.

1.1.6 Prid + *pajek*

Kombinacija samostalnika in ujemalnega pridevnika na levi (tj. levega ujemalnega pridevniškega prilastka) je glede na podatke o zastopanosti v korpusu (vzorčni tip Prid + *pajek* predstavlja 13 % med vsemi dvodelnimi vzorci) za obravnavano lemo najbolj zanimiva. O tem priča tudi dolžina seznama izluščenih besednih zvez, na katerem se s pogostnostjo 3 ali več pojavlja kar 78 različnih besednih nizov:

besedni niz		pogostnost v korpusu	imenovan	pajek	
rdeč	pajek	229	dolgonog	pajek	5
ptičji	pajek	139	ogromen	pajek	5
strupen	pajek	88	zlat	pajek	5
hidravlični	pajek	67	kovinski	pajek	5
morski	pajek	62	zelen	pajek	5
velik	pajek	53	spremenjen	pajek	4
štirivretenski	pajek	42	nenavaden	pajek	4
vretenski	pajek	39	spreten	pajek	4
majhen	pajek	35	obvoden	pajek	4
dvovretenski	pajek	29	star	pajek	4
črn	pajek	25	gnusen	pajek	4
kosmat	pajek	17	dvoreden	pajek	4
cveten	pajek	16	vrten	pajek	4
ljubljski	pajek	16	evropski	pajek	4
orjaški	pajek	16	istoimenski	pajek	4
nevaren	pajek	15	morbiden	pajek	4
rakovičast	pajek	14	moder	pajek	4
omrežen	pajek	13	svilnat	pajek	4
obračalen	pajek	13	Srebrast	pajek	4
voden	pajek	11	umeten	pajek	4
velikanski	pajek	11	živeč	pajek	3
avstralski	pajek	11	rjav	pajek	3
novomeški	pajek	10	razbeljen	pajek	3
zloglasen	pajek	10	predramljen	pajek	3
spleten	pajek	9	tropski	pajek	3
rabljen	pajek	9	vilinski	pajek	3
živ	pajek	9	strašen	pajek	3
smrtonosen	pajek	9	mesten	pajek	3
mlad	pajek	9	občinski	pajek	3
rdečepikast	pajek	8	tolst	pajek	3
nov	pajek	8	avtomobilski	pajek	3
koprski	pajek	8	komunalen	pajek	3
hišen	pajek	8	italijanski	pajek	3
jamski	pajek	7	trotočkovni	pajek	3
morilski	pajek	7	drag	pajek	3
srebrast	pajek	7	brazilski	pajek	3
iskalen	pajek	7	mali	pajek	3
različen	pajek	6	radioaktiven	pajek	3
plastičen	pajek	5			

Tabela 41: Izluščeni podatki Prid + pajek.

Z izjemo pridevnika *vretenski*, ki pri obdelavi besedila nastane zaradi razbijanja primerov tipa 2-*vretenski* na dve enoti, in nenavadnega dvojnega pojavljanja pridevnika *srebrast*, ki ima v sedmih primerih lemo zapisano z malo začetnico, v štirih pa z veliko, gornji seznam prinaša nabor zvez, kandidatke za vključitev v leksikalno zbirko.

Žal avtomatsko ni mogoče ločevati med zvezami s kakovostnim in vrstnim pridevnikom (ni možna diferenciacija med *rdeč pajek* ter *rdeči pajek* v primeru avtomatskega označevanja v besedilnem kontekstu – razen seveda v redkih primerih, ko je določnost izražena na ravni besedne oblike). Pri predstavljeni metodi luščenja se pri obravnavi pridevnikov moškega spola upošteva lematizirane oblike, kot je prikazano v gornji tabeli. Ločevanje kakovostnosti od vrstnosti je na ta način prepuščena leksikografski obravnavi. Za pridevnike srednjega ter

ženskega spola je potrebno razmisliti o prilagoditvi metode, saj se pridevniške oblike vseh spolov lematizirajo v moško oz. nezaznamovano osnovno obliko (več o tem sledi v V-1.3.1).

Kot je bilo omenjeno v II-2.1, Erjavec in Vintar (2008: 68) pri pretvorbi izluščenih nizov, kandidatov za terminološke besedne zveze, v ustrezno kanonično obliko za vsakega od primerov v korpusu poiščeta izpričano imenovalniško obliko pridevnika moškega spola ter z njo nadomestita lematizirano obliko (npr. pretvorba *hidravličan pajek* v *hidravlični pajek*). Postopek, ki je uporaben predvsem za pogoste primere v korpusu, bi bilo možno uporabiti tudi pri pripravi prikazne oblike podatkov za vnos v leksikalno zbirko. Ker pa se na tem mestu ne osredotočamo le na stalne besedne zveze (glej II-2.2), bi bilo potrebno v opisanem postopku posebej upoštevati primere, kjer besedni nizi v imenovalniku izpričujejo dve pridevniški obliki:

zelo **strupen pajek** s črno-rdečim pikastim zadkom (dve besedi)

režiserja Franceta Štiglica iz leta 1966, 2. velik **strupen pajek** iz južne Italije, 3. švicarski slikar in grafik

NAVPIČNO: 1 **strupeni pajek** črne barve z rdečo liso v zmerno toplih krajih

ugriznejo približno sedem ljudi. Pred dobrim tednom je ta **strupeni pajek**, ki se najraje zadržuje

Funnell previdno odpre vrata - v zarjaveli škatli morda preži **strupen pajek**. Kače, pajki in druga

pomisli, kaj bi se zgodilo, če bi me **strupeni pajek** ugriznil?"

Navpično: 1. **strupen pajek** s črno-rdečim pikastim zadkom (dve besedi)

v dobro režirani kriminalki: ko je avstralskega policista pičil **strupeni pajek**, so imeli zdravniki na voljo

Strupeni pajek tokrat napadel

Na centrali trgovske verige, pri kateri se je pojavil **strupeni pajek**, so za Dnevnik povedali, da se je

Rumeni pajek (*Chericanthum puncturatum*) je naš najbolj **strupeni pajek**. Posebno napadalne so

ki mi je rekel, da me je pičil **strupeni pajek**, ker tedaj še niso poznali lymske bolezni. Čez

Konkordančni niz FidaPLUS 6: #1**strupen**P???ei*_#1**pajek**.

Kljub nekaterim napakam v označevanju (glej naslednje poglavje) luščenje zvez glede na ujemanje v spolu, sklonu in številu daje dobre rezultate. Ne gre pa na tem mestu pozabiti zvez, ki niso ujemanje, so pa za obravnavo vsekakor zanimive, tj. zvez pridevnika s samostalniškim dopolnilom (Toporišič 2004⁴: 327). Pri trenutno obravnavanem vzorčnem tipu bi bile pričakovane zveze pridevnika z rodilniškim ali dajalniškim samostalniškim dopolnilom: **Prid + pajek₂ / pajek₂ + Prid** ter **Prid + pajek₃ / pajek₃ + Prid**. Za luščenje tovrstnih besednih nizov je bil po zgledu luščenja zvez dveh samostalnikov (glej V-1.1.1) pripravljen specializiran program (opis programa v IV-3.3.3). Edini primer besedne zveze pridevnika s samostalniškim dopolnilom je *pajku podoben* (3) oz. *podoben pajku* (6).

1.1.6.1 Analiza označenosti

Luščenje nizov, kjer pred samostalnikom *pajek* nastopa pridevnik ženskega ali srednjega spola, poleg ustrezno označene zveze pridevnika z dopolnilom (*podobna pajku*) prinaša rezultate, navedene v Tabeli 42. Najpogostejša napaka v spodnjem naboru je denimo označenost pridevniške oblike kot imenovalnik dvojine srednjega spola (primeri so podčrtani):

oblikoskladenjski oznaki		besedni obliki		Pdnzmi	Sommt	dehidrirane	pajke
Pdnzmi	Sommt	modificirane	pajke	<u>Pdnzdi</u>	<u>Somei</u>	<u>spremenjeni</u>	<u>pajek</u>
Ppnzdi	Somei	jamski	pajek	<u>Pdnzdi</u>	<u>Somei</u>	<u>Najeti</u>	<u>pajek</u>

<u>Pdnsdi</u>	<u>Somei</u>	<u>Prekanjeni</u>	<u>pajek</u>	<u>Pdnsdi</u>	<u>Sommi</u>	<u>Kuhani</u>	<u>pajki</u>
<u>Pdnsdi</u>	<u>Sommi</u>	<u>Zastrupljeni</u>	<u>pajki</u>	<u>Pdnsdi</u>	<u>Somei</u>	<u>spremenjeni</u>	<u>pajek</u>
<u>Ppnsei</u>	<u>Sommr</u>	<u>polno</u>	<u>pajkov</u>	<u>Pdnsdi</u>	<u>Sometd</u>	<u>ubiti</u>	<u>pajka</u>
<u>Pdnsdi</u>	<u>Sommi</u>	<u>prebujeni</u>	<u>pajki</u>				

Tabela 42: Prid + *pajek* – neujemanje v spolu.

Primeri neustrezno označenih kombinacij samostalnika *pajek* s pridevnikom moškega spola, ki pa se ne ujema v sklonu, so navedeni spodaj. Kot je razvidno iz primerov, do neujemanja lahko prihaja tudi zaradi napačnega označevanja sklonov samostalnika:

oblikoskladenjski oznaki		besedni obliki	
Ppnmer	Sometd	rdečega	pajka
Ppnmmo	Sommi	vodnimi	pajki
Ppnmer	Sometd	morskega	pajka
Ppnmmd	Someo	ptičjim	pajkom
Ppnmmo	Sommi	ptičjimi	pajki
Ppnmdi	Sometd	ptičja	pajka

Tabela 43: Prid + *pajek* – neujemanje v sklonu.

Kljub glede na velikost korpusa izredno redkim označevalnim napakam na tem mestu ne bo odveč poudarek, da je v primerih obravnavanega tipa pri razdvoumljanju oblikoskladenjskih oznak smiselno vedno izbrati oznako, ki je skladna z ujemalnostjo. Tudi na tem mestu gre razmisliti v smer upoštevanja besednozveznih leksikalnih podatkov pri avtomatskem označevanju: dopolnjevanje leksikalne zbirke z informacijo, da so [*rdeči, vodni, morski, ptičji*] *pajek* v jeziku tipične besedne zveze, lahko ob ustrezni prilagoditvi označevanja vodi v pripisovanje ujemajočega se spola, sklonov ter števila obema besednima oblikama v besedilu hkrati.⁹⁵

1.1.7 *Pajek + Prisl*

Korpusnih pojavitev kombinacije samostalnika *pajek* s prislovom na levi je 167, kar je v primerjavi s pojavitvami drugih vzorčnih tipov malo. Pregled besednih nizov kaže, da za vključitev v leksikalno zbirko niso relevantni, kar potrjujejo naslednji primeri:

besedni niz		pogostnost v korpusu	
pajek	lahko	5	pajka več 3
pajek	prede	4	pajki lahko 3
pajke	tako	3	pajke pogosto 3
pajek	očitno	3	pajka lahko 3

Tabela 44: Izluščeni podatki – *pajek + Prisl*.

1.1.8 *Prisl + pajek*

Nizov, označenih kot kombinacija samostalnika *pajek* s prislovom na desni, je v korpusu 97, kar ponovno priča o relativni redkosti obravnavanega vzorčnega tipa v korpusu. Pregled primerov kaže, da velik delež odpade na napačno lematizirane, npr. na zvezi *moč pajka* oz. *moč pajkov*, ki se v korpusu pojavljata kot del televizijskega sporeda in sta zato dokaj pogostni (skupaj 21 primerov):

⁹⁵ S tem rešimo problem le na ravni označevanja tipičnih oz. zelo pogostih zvez, kar pa se zdi ustrezen prvi korak.

besedni niz	pogostnost v korpusu	več	pajkov	
moč pajka	16	kako	pajki	3
lahko pajek	9	veliko	pajkov	3
moč pajkov	5			

Tabela 45: Izluščeni podatki – Prisl + pajek.

Kot vidimo, so med primeri tudi nizi, ki izražajo količino, npr. [veliko, več] pajkov. Slednji bi za nadaljnjo obravnavo sicer lahko bili zanimivi, vendar je zaradi redke zastopanosti ter načeloma za leksikalno zbirko nerelevantnih nizov obravnavani vzorčni tip trenutno opredeljen za manj zanimivega za luščenje podatkov.

1.1.9 Luščenje zvez Sam + Glag ter Glag + Sam

Zveze samostalnika ter glagola niso jedrnega interesa za pričujočo raziskavo, ker je predvideno, da se bodo učinkoviteje luščile iz besedil, označenih na skladijski ravni.⁹⁶ Jasno je, da z luščenjem dvodelnih ter tridelnih vzorcev ni mogoče doseči zadovoljivega priklica zvez med denimo samostalnikom (v vlogi osebka ali predmeta) ter glagolom, saj so omejeni s številom elementov ter njihovo zaporednostjo v vzorcu. Kljub temu je na tem mestu namenjenega nekaj prostora luščenju kombinacij samostalnikov ter glagolov oz. glagola in drugih besednih vrst (glej V-1.3), predvsem v smislu prikaza možnosti oz. omejitev, ki jih ta metoda ponuja.

Oblikoskladijske oznake za glagolske oblike so sicer zelo členjene (glej Priloga 1), kar se odraža v razpršenosti vzorčnih zapolnitev. V pričujoči raziskavi so zato vse glagolske oblike obravnavane na ravni leme. Samostalnike pa se zdi bolje obravnavati na ravni besedne oblike, kar omogoča ločevanje besednih nizov tipa *odpeljati pajek* – pri katerih je mogoče sklepati, da je v izhodiščnem stavku *pajek* v vlogi osebka – od nizov tipa *poklicati pajka*, pri katerih je mogoče sklepati, da je v izhodiščnem stavku *pajek* v vlogi predmeta.

1.1.10 Glag + pajek

Spodnja tabela prinaša nabor primerov za obravnavani vzorčni tip:

besedni niz	pogostnost v korpusu			
odpeljati pajek	148	odpraviti	pajke	4
bati pajkov	28	odvažati	pajek	4
poklicati pajka	17	pičiti	pajek	3
ubiti pajka	7	najti	pajka	3
imeti pajke	6	morati	pajek	3
priiti pajek	6	kupiti	pajka	3
kupiti pajek	6	marati	pajkov	3
imeti pajek	5	dvigniti	pajek	3
videti pajka	5	poslati	pajka	3
zagledati pajka	4	lesti	pajek	3
živeti pajek	4	igrati	pajki	3
preganjati pajke	4	dati	pajkom	3
oboževati pajke	4	opaziti	pajka	3
gnati pajka	4	črniti	pajek	3
imeti pajka	4	dobiti	pajka	3
pobijati pajke	4			

Tabela 46: Izluščeni podatki – Glag + pajek.

⁹⁶ Priprava sistema za skladijsko označevanje slovenskih besedil je eden od ciljev projekta Jezikoslovno označevanje slovenščine.

Rezultati sicer prinašajo zanimivo podobo sopojevjanja samostalnika *pajek* ter glagola na levi, vendar so, kot rečeno, omejeni glede priklica, obenem pa zaradi nedoločniške oblike v kombinaciji z imenovalniškim samostalnikom slabše berljivi. Možna rešitev slednjega problema je avtomatsko nadomeščanje nedoločnika z ustrezno sedanjiško obliko (pretvorba *odpeljati pajek* v *odpelje pajek* itd.), kar pa je lahko zgolj provizorično, saj pomeni precejšen poseg v podatkovni nabor.

Mogoče je tudi nadaljnje razvrščanje besednih nizov glede na sklon samostalnika, pri čemer se oblikujejo vzorčni tipi **Glag + pajek₁ / pajek₁ + Glag // Glag + pajek₂ / pajek₂ + Glag // Glag + pajek₃ / pajek₃ + Glag ter Glag + pajek₄ / pajek₄ + Glag**.

Za primer sledi v nadaljevanju preurejena tabela s podatki, pri čemer je potreben poudarek, da je preurejanje ročno (tako pretvorba glagolske oblike kot tudi razporejanje nizov glede na sklon samostalnika):

Glag + pajek ₁			Glag + pajek ₄		
ODPELJE	pajek	148	poklicati	pajka	17
PRIDE	pajek	6	ubiti	pajka	7
IMA	pajek	5	imeti	pajke	6
ŽIVI	pajek	4	<u>kupiti</u>	<u>pajek</u>	<u>6</u>
ODVAŽA	pajek	4	videti	pajka	5
PIČI	pajek	3	zagledati	pajka	4
MORA	pajek	3	preganjati	pajke	4
DVIGNE	pajek	3	oboževati	pajke	4
LEZE	pajek	3	gnati	pajka	4
ČRNI	pajek	3	imeti	pajka	4
IGRAJO	pajki	3	pobijati	pajke	4
Glag + pajek₂			odpraviti	pajke	4
bati	pajkov	28	najti	pajka	3
marati	pajkov	3	kupiti	pajka	3
Glag + pajek₃			poslati	pajka	3
dati	pajkom	3	opaziti	pajka	3
			dobiti	pajka	3

Tabela 47: Urejeni podatki – Glag + pajek.

Brez upoštevanja konteksta je urejanje podatkov problematično na ravni ločevanja imenovalnika od tožilnika v primeru, da samostalni *pajek* izpričuje podspol neživosti – prim. npr. v gornji tabeli podčrtani primer *kupiti pajek*. Ročno razvrščanje omogoča lažjo identifikacijo tovrstnih primerov, je pa ustrezno zamudnejše. Ker ni bila opravljena natančna analiza uspešnosti avtomatsko pripisanih sklonov v gornjih primerih, vprašanje, na katerem mestu je smotrnejše podatke poslati v ročno obdelavo (pred oz. namesto avtomatskega razvrščanja glede na sklon samostalnika, po razvrščanju), ostaja trenutno neodgovorjeno.

1.1.11 Pajek + Glag

Tako kot v prejšnjem poglavju je tudi na tem mestu analiza namenjena zgolj prikazu besednih nizov, ki so z ekstrakcijo zvez tega tipa pridobljivi. Spodnja tabela prikazuje nabor najpogostejših zvez z glagolom v nedoločniku ter samostalnikom v izpričani obliki; imenovalniška oblika samostalnika, kot rečeno, sugerira osebkovo stavčnočlensko vlogo, ostale oblike pa predmetne:

besedni niz			pogostnost v korpusu		
pajek	odpeljati	49	pajki	delati	4
pajki	napadati	40	pajek	splesti	4
pajkom	odpeljati	15	pajkov	bati	4
pajki	imeti	11	pajki	začeti	4
pajek	imeti	11	pajki	prinašati	4

pajek	spresti	7	pajkov	živeti	3
pajki	plesti	7	pajek	tkati	3
pajek	prodati	6	pajek	začeti	3
pajek	odvažati	6	pajka	zadeti	3
pajki	odvažati	6	pajki	izločati	3
pajek	boriti	5	pajkov	imeti	3
pajki	živeti	5	pajek	prežati	3
pajek	plesti	5	pajka	pripraviti	3
pajek	odvleči	5	pajek	živeti	3
pajek	ugrizniti	5	pajek	prepresti	3
pajek	loviti	5	pajek	morati	3
pajek	potegniti	4	pajki	delovati	3
pajki	voziti	4	pajek	poškodovati	3
pajkom	odvažati	4	pajka	nastaviti	3
pajki	presti	4	pajek	vedeti	3

Tabela 48: Izluščeni podatki – pajek + Glag.

Za razliko od prejšnjega vzorčnega tipa se na tem mestu pojavlja vprašanje obravnave nizov s samostalniki v predložnih sklonih (v mestniku ter orodniku), npr. (*s*) *pajkom odvažati*. Pri luščenju besednih nizov, ki imajo na prvem mestu samostalnik, ki ga želimo obravnavati nelematiziranega, je ključno zajeti v obravnavo tudi potencialne predhodne predloge.⁹⁷ V sklopu daljših vzorcev je smiselno obravnavati tudi nize z večglagolskimi zvezami, npr. *pajek [morati, začeti]*.

Še nekaj besed o ločevanju podatkov glede na slovnično število samostalnika: v gornjih primerih se kaže za nepotrebno – prim. *pajek [odpelje, ima, sprede, proda, odvažajo, plete, odvleče, ugrizne, lovi, potegne ...]* ter *pajki [napadajo, imajo, pletejo, odvažajo, živijo, vozijo ...]*. To pomeni, da bi bilo mogoče oblike po številu združiti, zaradi preglednosti najbolje pod edninsko obliko (podoben primer združevanja je predstavljen v V-1.1.3). Na tem mestu ostajajo tudi opozorila glede posega v podatkovno realnost.

1.2 Pajek – tridelni vzorci

Med najpogostejšimi (101) tridelnimi vzorci z lemo *pajek* je takih, ki poleg obravnavane leme prinašajo dve polnopomenski besedi, le 10 (9,9 %).⁹⁸ Vzorci, ki se pojavijo več kot enkrat, so uvrščeni v ustrezne vzorčne tipe, tisti s pogostnostjo 1 pa so zaenkrat obravnavani le na ravni vzorca ter vzorčnih zapolnitev⁹⁹, predvsem kar se tiče relevantnosti izluščenih podatkov za vključitev v leksikalno zbirko.

vzorčni tip	število vzorcev
Prid + Prid + <i>pajek</i>	2
Prid + <i>pajek</i> + Sam	2
drugi vzorci	6

PAJEK SOMEI SOMEI

⁹⁷ Mogoče se je seveda zanašati na oblikoskladenjske oznake, ki so na ravni označevanja predložnih sklonov predvidoma dokaj zanesljive. V zvezi s tem bi bile potrebne dodatne raziskave.

⁹⁸ Večina visokopogostnih vzorcev odpade na kombinacije s funkcijskimi besednimi vrstami, nabor vzorcev, sestojčih iz polnopomenskih besednih vrst, pa je temu ustrezno krajši.

⁹⁹ Analiza se osredotoča na najbolj tipično, kar pomeni, da pri delu marsikaj ostaja ob strani. Deloma so podatki (za samostalniške besedne zveze) dopolnjeni v poglavju V-6.

PPNMEID SOMEI PAJEK
 SOMEI PAJEK SOMEI
 PAJEK RNN RNN
 PPNMMI PAJEK GGNSTM
 RNN PPNMMR PAJEK

Tabela 49: Nabor najpogostejših tridelnih vzorcev z lemo *pajek* in oblikoskladenjskimi oznakami za polnopomenske besede.

Navedeni vzorčni tipi ter vzorci so v nadaljevanju poglavja na kratko predstavljeni, ne prinaša pa pričujoče poglavje analize oblikoskladenjskih oznak ter vprašanj luščenja in prikaza podatkov, ker bi to v večji meri pomenilo podvajanje informacij poglavja V-1.1.

1.2.1 Prid + Prid + *pajek*

Spodnji seznam potrjuje intuicijo, da gre na tem mestu za vzorčni tip, ki prinaša relevantne besedne zveze. Navedene so vse, ki se pojavljajo s pogostnostjo 2 ali več (predstavljeni so primeri, ki na ravni oblikoskladenjskih oznak izkazujejo ujemanje pridevnikov s samostalnikom):

besedni niz		pog. v korpusu		navaden	vrten	pajek	2
rdečenog	ptičji	pajek	6	dvovretenski	obračalen	pajek	2
vretenski	hidravličen	pajek	6	strupen	ptičji	pajek	2
velik	morski	pajek	6	ohranjen	štirivretenski	pajek	2
velik	kosmat	pajek	6	mali	morski	pajek	2
orjaški	ptičji	pajek	5	vretenski	italijanski	pajek	2
rabljen	štirivretenski	pajek	4	zloglasen	novomeški	pajek	2
velik	strupen	pajek	4	užiten	ptičji	pajek	2
živeč	voden	pajek	3	velik	črn	pajek	2
velik	hišen	pajek	3	rumen	cveten	pajek	2
kosmat	ptičji	pajek	3	debel	črn	pajek	2
ogromen	črn	pajek	2	nov	dvovretenski	pajek	2
odkrit	ptičji	pajek	2	velik	evropski	pajek	2
strupen	evropski	pajek	2	perujski	ptičji	pajek	2
orhidejin	rakovičast	pajek	2	velik	ptičji	pajek	2

Tabela 50: Izluščeni podatki – Prid + Prid + *pajek*.

Več informacij o luščenju ter prikazu tovrstnih podatkov je na voljo v poglavju V-1.1.6.

1.2.2 Prid + *pajek* + Sam

Tudi na tem mestu obravnavani vzorčni tip prinaša za nadaljnjo obravnavo relevantne besedne nize, vendar je nabor nekoliko krajši:

besedni niz		pogostnost v korpusu		štirivretenski	pajek	širina	2
dvovretenski	pajek	sipa	11	vretenski	pajek	širina	2
štirivretenski	pajek	sip	6	vretenski	pajek	pottinger	2
štirivretenski	pajek	sipa	4	jamski	pajek	met	2
obračalen	pajek	spider	3	ptičji	pajek	Henry	2
vretenski	pajek	sip	3	rjav	pajek	samotar	2
brazilski	pajek	klatež	3	dvovretenski	pajek	sip	2

Tabela 51: Izluščeni podatki – Prid + *pajek* + Sam

Vzorčni tip bi bilo mogoče zapisati na način **Prid + (pajek + Sam)**, saj pridevnik določa samostalniško besedno zvezo na desni. V naslednjem koraku so smiselne nadaljnje delitve vzorčnega tipa glede na vrsto samostalniške zveze (glej V-1.1.2) in njihova ločena obravnava. V gornji tabeli delitev še ni izvedena, zato je denimo primer *štirivretenski pajek širina* v neustrezni prikazni obliki. Poleg omenjenega poglavja prinaša za luščenje tovrstnih podatkov relevantne informacije tudi poglavje V-1.1.6.

1.2.3 Preostali vzorci

Spodnja tabela prinaša po nekaj najpogostejših primerov zapolnitev za posameznega od preostalih vzorcev z lemo *pajek* in dvema polnopomenskima besedama:

PAJEK SOMEI SOMEI	pogostnost v korpusu
pajek SIP spider	52
pajek sip spider	15
pajek pottinger HIT	4
pajek spider SIP	3
PPNMEID SOMEI PAJEK	
resnični mož pajek	7
novi človek pajek	7
resnični človek pajek	4
SOMEI PAJEK SOMEI	
obračalnik pajek SIP	17
obračalnik pajek sip	3
PAJEK RNN RNN	
pajek tam gotovo	
pajek tako napačno	2
PPNMMI PAJEK GGNSTM	
ptičji pajki živijo	2
RNN PPNMMR PAJEK	
najbolj strupeni pajki	6

Tabela 52: Preostali tridelni vzorci z lemo *pajek* in dvema polnopomenskima besedama.

Kot je razvidno iz podatkov, so za nadaljnjo obravnavo zanimivi vzorci, v katerih se obravnavani samostalnik *pajek* pojavlja v kombinacijah z (I) drugimi samostalniki (*PAJEK SOMEI SOMEI*, *SOMEI PAJEK SOMEI*) oz. (II) samostalnikom in pridevnikom (*PPNMEID SOMEI PAJEK*), kar seveda ustreza dosedanjim ugotovitvam o slovenski skladnji na ravni samostalniških besednih zvez. V sklopu prvega tipa najdemo zveze samostalniškega jedra z desnoprilastkovno samostalniško besedno zvezo (*pajek SIP spider*, *obračalnik pajek SIP*), v sklopu drugega tipa pa z levim pridevniškim prilastkom določano samostalniško besedno zvezo (*resnični mož pajek*).

Na osnovi podatkov lahko torej tvorimo naslednje vzorčne tipe: ***pajek (Sam + Sam) / Sam (pajek + Sam) / Prid (Sam + pajek)***.

Potencialno zanimivi za uvrstitev v leksikalno zbirko so tudi vzorci, ki prinašajo besedno zvezo samostalnika z levim pridevniškim prilastkom, pridevnik pa določen še s prislovom na levi (*najbolj strupeni pajki*). Potencialno zanimive so tudi kombinacije samostalniške besedne zveze ter glagola, z upoštevanjem opozoril glede nizkega priklica podatkov (*ptičji pajki živijo*).

Nerelevantni za nadaljnjo analizo so besedni nizi, ki na zadnjem mestu prinašajo prislov (*pajek tam gotovo*).

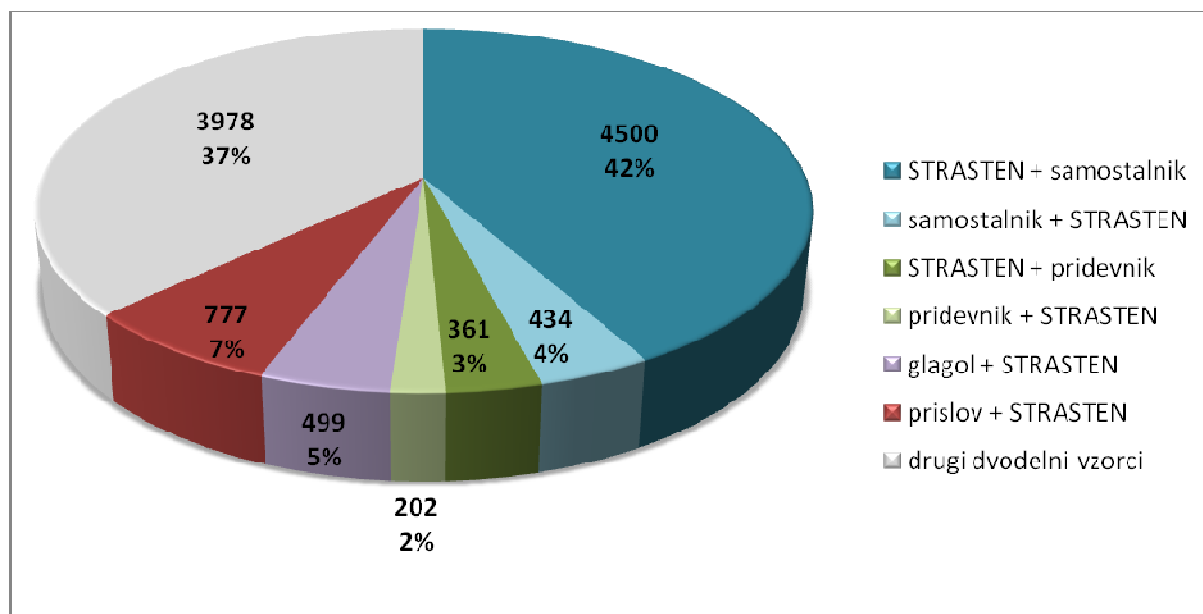
1.3 Strasten – dvodelni vzorci

Najpogostejših 70 dvodelnih vzorcev z lemo *strasten* in oznako za polnopomensko besedo na levi ali desni lahko razvrstimo v šest vzorčnih tipov: lema *strasten* se pojavlja s samostalnikom oz. pridevnikom na levi ali desni ter prislovom ali glagolom na levi. Zastopanost najpogostejših vzorcev v posameznem vzorčnem tipu prikazuje naslednja tabela:

vzorčni tip	število vzorcev
<i>strasten</i> + Sam	35
Sam + <i>strasten</i>	11
<i>strasten</i> + Prid	9
Prid + <i>strasten</i>	4
Prisl + <i>strasten</i>	2
Glag + <i>strasten</i>	9

Tabela 53: Razvrstitev najpogostejših vzorcev z lemo *strasten* in oblikoskladenjskimi oznakami za polnopomenske besede v vzorčne tipe.

Sledi graf z informacijo o vsoti pogostnosti vseh korpusnih pojavitev za vsakega od na tem mestu definiranih vzorčnih tipov, skupaj s številom zapolnitev, ki jih slednji ne pokrivajo (*drugi dvodelni vzorci*).



Graf 2: Vsota pogostnosti pojavitev vzorcev za dvodelne vzorčne tipe z lemo *strasten* in oblikoskladenjskimi oznakami za polnopomenske besede.

Graf kaže, da korpusne pojavitve vzorcev s polnopomenskimi besedami predstavljajo več kot pol vseh dvodelnih. Med obravnavanimi vzorci po pogostnosti prednjači vzorčni tip *strasten* + Sam (42 %), sledijo Prisl + *strasten* (7 %), Glag + *strasten* (5 %) ter Sam + *strasten* (4 %). Najredkeje se pojavljata vzorčna tipa *strasten* + Prid (3 %) ter Prid + *strasten* (2 %).

V nadaljevanju sledi obravnava vzorcev glede na navedene vzorčne tipe.

1.3.1 Strasten + Sam

Kot je bilo omenjeno v V-1.1.6, se po trenutno vzpostavljenem sistemu označevanja slovenskih besedil vsi pridevniki lematizirajo v nezaznamovano obliko (ki oblikovno sovпада z obliko za moški spol) pridevnika. Pri obravnavanem vzorčnem tipu *strasten* + Sam je predvideno luščenje nabora zvez, kjer se pridevnik in samostalnik ujemata v spolu, sklonu ter številu, pri čemer se pri zvezah s samostalniki srednjega ter ženskega spola prikaz z lematizirano obliko obravnavanega pridevnika zdi manj ustrezen (oz. slabše pregleden). Na tem mestu je zato predlagana tristopenjska obravnava tovrstnih primerov:

- (I) v prvi fazi luščenje zvez v tri različne skupine glede na spol pridevniku sledečega samostalnika;
- (II) sledi sprememba pridevniške leme v ustrezno osnovno obliko;
- (III) nato pa šele urejanje besednih zvez glede na pogostnost.¹⁰⁰

Na ta način dobimo tri ločene nabore besednih zvez, ki jih predstavljajo spodnje tabele (zaradi velike pogostnosti zvez tega tipa so navedene le zveze s pogostnostjo 5 ali več).

besedni niz - m. spol		pogostnost v korpusu	
strasten	kadilec	283	strasten motorist 9
strasten	ljubitelj	177	strasten izbruh 9
strasten	poljub	147	strasten večer 9
strasten	zbiralec	115	strasten nasprotnik 9
strasten	ribič	104	strasten pilot 8
strasten	lovec	102	strasten kolesar 8
strasten	igralec	75	strasten ton 7
strasten	ljubimec	69	strasten filatelist 7
strasten	zbiratelj	65	strasten boj 7
strasten	objem	61	strasten dan 7
strasten	navijač	59	strasten borec 7
strasten	seks	54	strasten nogometaš 7
strasten	bralec	41	strasten gurman 7
strasten	gobar	39	strasten mladenič 7
strasten	zagovornik	37	strasten šahist 7
strasten	privrženec	34	strasten kockar 7
strasten	človek	34	strasten pogled 7
strasten	oboževalec	32	strasten zaljubljenec 7
strasten	odnos	27	strasten govor 7
strasten	hazarder	25	strasten glas 7
strasten	športnik	22	strasten zagovor 7
strasten	trenutek	19	strasten raziskovalec 7
strasten	jadrlec	18	strasten kuhar 6
strasten	ples	18	strasten mrk 6
strasten	golfist	17	strasten grof 6
strasten	smučar	17	strasten način 6
strasten	planinec	16	strasten par 6
strasten	kvartopirec	15	strasten nabiralec 6
strasten	popotnik	13	strasten prizor 6
strasten	pivec	13	strasten začetek 5
strasten	fotograf	12	strasten obiskovalec 5
strasten	potapljač	12	strasten pogovor 5
strasten	prijatelj	12	strasten vrtnar 5
strasten	tango	12	strasten potrošnik 5

¹⁰⁰ Opisani postopek trenutno še ni v celoti avtomatiziran, podatki za pričujoče poglavje so bili deloma pripravljeni z uporabo obstoječih programov, deloma z ročnimi pretvorbami pridevniških oblik v tabelah.

strasten	ugankar	12	strasten	nakupovalec	5
strasten	iskalec	12	strasten	zvok	5
strasten	jedec	10	strasten	govorec	5
strasten	plesalec	10	strasten	izliv	5
strasten	moški	10	strasten	temperament	5
strasten	občutek	9	strasten	kritik	5

Tabela 54: Izluščeni podatki – *strasten* + Sam_m.

besedni niz – ž. spol	pogostnost v korpusu	strastna	debata	9
strastna ljubezen	203	strastna igra		8
strastna noč	177	strastna beseda		8
strastna kadilka	47	strastna polemika		8
strastna želja	41	strastna zaljubljenost		8
strastna ženska	33	strastna igralka		7
strastna romanca	32	strastna obsedenost		7
strastna ljubiteljica	24	strastna lepotica		6
strastna avantura	18	strastna urica		6
strastna narava	17	strastna glasba		6
strastna oboževalka	16	strastna samovolja		6
strastna zveza	15	strastna ljubimka		6
strastna predanost	14	strastna privlačnost		6
strastna bralka	14	strastna jahalka		5
strastna zgodba	14	strastna barva		5
strastna razprava	13	strastna pesem		5
strastna afera	12	strastna zagovornica		5
strastna zbirateljica	12	strastna navezanost		5
strastna ljubica	11	strastna oblika		5
strastna privrženost	10	strastna zavzetost		5
strastna gobarka	9	strastna oseba		5
strastna navijačica	9			

Tabela 55: Izluščeni podatki – *strastna* + Sam_ž.

besedni niz – sr. spol	pogostnost v korpusu	strastno	zanimanje	11
strastno razmerje	129	strastno navdušenje		9
strastno čustvo	34	strastno prebujenje		9
strastno poljubljanje	32	strastno poželenje		8
strastno ljubljenje	17	strastno sovražstvo		8
strastno kajenje	16	strastno srečanje		8
strastno življenje	14	strastno zbiranje		8
strastno iskanje	13	strastno ustvarjanje		6
strastno pismo	11	strastno hrepenenje		6

Tabela 56: Izluščeni podatki – *strastno* + Sam_s.

Luščenje zvez pridevnika s samostalniškim dopolnilom za obravnavano lemo *strasten* ne prinaša nobenega relevantnega primera.

1.3.1.1 Analiza označenosti

Pri luščenju zgoraj navedenih besednih nizov je bilo upoštevano ujemanje pridevnika s samostalnikom glede spola, sklona ter števila. Kot je bilo omenjeno v poglavju 1.1.4, je bil za analizo označevanja pripravljen tudi program za luščenje nizov, za katere oblikoskladenjske oznake ne izkazujejo ujemanja. Luščenje za obravnavani

vzorčni tip vrne 31 primerov, ki so spodaj navedeni v izpričani obliki, poleg njih pa navajamo še pripisane oznake:

besedni niz	oblikosklad. oznaki	strastne	vode	Ppnmmt	Sozmt
Strastna razmerja	Ppnzei Soser	strastnih	kot	Ppnmmr	Sozmr
STRASTNA TAYLOR	Ppnzei Slmei	strastnih	laikih	Ppnmmr	Sommm
STRASTNIH IZSESKOV	Ppnmmr Slmei	Strastna	Marko	Ppnzei	Slmei
strastnimi flamenko	Ppnmmo Somei	Strastna	Marko	Ppnzei	Slmei
strastni imaginarij	Ppnmeid Sozmr	strastnega glasu		Ppnmet	Somer
strastno ljubezen	Ppnsei Sozet	strastnega glasu		Ppnmet	Somer
strastni imaginarij	Ppnmeid Sozmr	strastnega žurerja		Ppnmet	Somer
strastne nepristranosti	Ppnzer Sozdi	strastnega bluza		Ppnmer	Sozei
strasten love	Ppnmein Sommt	Strastna Primož		Ppnzei	Slmei
strastna vznesenost	Ppnsmi Sozei	strastni antididaktičnosti		Ppnzem	Sozdt
strastne tango	Ppnzer Somei	strastnih kot		Ppnmmr	Sozmr
strastnimi ritmi	Ppnmmo Sozmo	strastne vprašanosti		Ppnzer	Sozdi
strastni ženski	Ppnmeid Sozed	strastna Petra		Ppnzei	Slmdi
strasten deloalkoholik	Ppnmein Sozmr	strastni Dmitrij		Ppnmeid	Sozmr
strastna bralce	Ppnzei Sommt	STRASTNEGA UGODJA		Ppnmet	Soser

Tabela 57: Analiza označenosti – neujemalne zveze *strasten* + *Sam*.

Sivo so obarvane zveze, ki so označene napačno, se pa obenem pojavljajo med najbolj tipičnimi, kakor so navedene v Tabelah 54–56. Tem napakam bi se bilo mogoče ogniti z upoštevanjem besednozveznih podatkov iz leksikalne zbirke (kot je bilo izpostavljeno v 1.1.6.1).

Težje je zaobiti druge na tem mestu evidentirane napake, ker temeljijo denimo na problemih označevanja (I) lastnih imen (npr. *strastna Taylor*, *strastni Dmitrij*) ali (II) neznanih besed (*strasten love*, *strastni imaginarij*, *strastnega bluza* itd.). Problematični za označevanje so tudi primeri, kjer pridevnik namesto zgolj samostalnika določa priredno besedno zvezo (*strastna Marko*, *strastna Primož*).

Ponovno se na seznamu pojavljajo napake lematizacije veznika *kot*. Možna rešitev tega problema se zdi natančna kolokacijska analiza samostalnikov *kot* in *kota*, ki bi pri nadaljnjem označevanju služila kot izhodišče za razdvoumljanje lem. Na osnovi podatkov bi bilo možno zanesljivejše ločevanje med označevanjem primerov tipa *strastni kot*, kjer je verjetnost, da gre za veznik večja, od primerov tipa *pravi kot*, kjer je dvoumnost težje rešljiva. Možno je seveda tudi upoštevanje minimalne skladijske okolice problematičnega primera, ki je v primerih samostalnika *kot* tipično drugačna kot v primeru veznika. Vse naštetu velja tudi za druge primere funkcijskih besed z enakopisnimi polnopomenskimi pari (npr. problem lematizacije predloga *od*).

1.3.2 Sam + *strasten*

Za razliko od prejšnjega poglavja na tem mestu ni pričakovano veliko število za uvrstitev v leksikalno zbirko relevantnih izluščenih besednih nizov. Intuicijo potrjuje spodnji seznam zvez s pogostnostjo 3 ali več:

besedni niz	pogostnost v korpusu	učenka	strastno	4
kot strastnega	24	letih	strastnega	3
Slovenci strastni	5	politike	strasten	3
vlogo strastne	4	noči	strastnega	3
kot strastnemu	4	srce	strastne	3
vlogo strastnega	4			

Tabela 58: Izluščeni podatki – Sam + *strasten*.

Pregled nizov v kontekstu kaže, da v primeru *politike strasten* ne gre za zvezo pridevnika z dopolnilom. Manj zanimivi so tudi preostali primeri iz tabele.

1.3.3 Strasten + Prid

Kot je razvidno iz nadaljevanja, so nerelevantni za nadaljnjo obravnavo tudi spodnji primeri:

besedni niz	pogostnost v korpusu	strasten	intimen	
strasten ljubezenski	108	strasten filmski		4
strasten nogometen	14	strasten strankarski		3
strasten ljubiteljski	11	strasten glasben		3
strasten seksualen	7	strasten podvoden		3
strasten spolen	6	strasten italijanski		3
strasten mlad	6	strasten slovenski		3
strasten romantičen	6	strasten dvoren		3
strasten športen	5	strasten nov		3
strasten političen	5	strasten telesen		3
strasten surrealističen	4			

Tabela 59: Izluščeni podatki – *strasten* + Prid.

K relevantnejšim podatkom, ki jih prinaša daljši vzorčni tip, se vračamo v poglavju V-1.4.3. Kljub nezanimivosti izluščenih nizov za vključitev v leksikalno zbirko pa je analizo mogoče nadaljevati na ravni identifikacije neustrezno označenih primerov, čemur je posvečeno naslednje poglavje.

1.3.3.1 Analiza označenosti

Za luščenje zvez dveh pridevnikov, ki se ne ujemata v spolu, sklonu ter številu, je bil pripravljen poseben program (za opis glej IV-3.3.5). Luščenje prinaša le nekaj primerov tovrstnih besednih nizov:

besedni niz	oblikoskladenjski oznaki
strastna osebna	Ppnzei Ppnsmi
strastnih holivudskih	Ppnzmr Ppnzdr
strastnega sluteča	Ppnmet Ppnzei
strastne prepovedane	Ppnzer Pdnzmi
strastnega Cameroninega	Ppnmet Psnmer
strastni ognjeni	Ppnmeid Pdnsdi
Strastna Ewan	Ppnzei Ppnmein

Tabela 60: Analiza označenosti – neujemalne zveze *strasten* + Prid.

Označevalne napake so, kot je razvidno, podobne že opisanim v prejšnjih poglavjih (npr. V-1.3.1.1). Kakor bo izpostavljeno tudi pri analizi označenosti naslednjega vzorčnega tipa, bi bilo razdvoumljanje oblikoskladenjskih oznak pri tovrstnih primerih smiselno osnovati na oblikoskladenjski podobnosti med pridevniki znotraj pridevniškega niza posamezne besedne zveze.

1.3.4 Prid + *strasten*

Podobno kot v prejšnjem poglavju tudi tukaj predstavljeni primeri pridevniških parov sami na sebi niso zanimivi za nadaljnjo obravnavo:

besedni niz	pogostnost v korpusu	nov	strasten	
sam strasten	25	neznan	strasten	9
nekdanji strasten	9	dolg	strasten	6
				6

poln	strasten	5	ogrožen	strasten	3
številn	strasten	4	pozabljen	strasten	3
zadnji	strasten	4	vroč	strasten	3
umrl	strasten	3	dolgoleten	strasten	3
podoben	strasten	3			

Tabela 61: Izluščeni podatki – Prid + strasten.

1.3.4.1 Analiza označenosti

Pregled oznak, pripisanih v tem poglavju obravnavanim besednim nizom, pokaže probleme označenosti predvsem na ravni napačne lematizacije prislovov v pridevnike. Nabor primerov za napake tega tipa je možno identificirati s pregledom primerov, označenih za pridevnik v srednjem spolu - kot je razvidno iz spodnjega nabora, ni v nobenem od primerov z oznako za srednji spol označen tudi sledeči pridevnik, kar je dodaten kazalec, da gre za potencialno označevalno napako.

besedni obliki		oblikosklad. oznaki	
prikrito	strastne	Pdnsei	Ppnzer
prijajeno	strastne	Pdnsei	Ppnmmt
pootročeno	strastna	Pdnsei	Ppnzei
Podrejeno	strastnim	Pdnsei	Ppnzmd
odmaknjeno	strastna	Pdnsei	Ppnzei
odmaknjeno	strastna	Pdnsei	Ppnzei
zadržano	strastna	Pdnsei	Ppnzei
enkratno	strastne	Ppnsei	Ppnzer
derviško	strastnih	Ppnsei	Ppnmmt
animalično	strastnega	Ppnsei	Ppnmet
animalično	strastnega	Ppnsei	Ppnmet
polno	strastne	Ppnsei	Ppnzer

Tabela 62: Analiza označenosti Prid + strasten.

Nekoliko drugačen, vendar primerljiv nabor zadetkov dobimo z iskanjem kombinacij pridevnika srednjega spola ter pridevnika *strasten* v korpusu FidaPLUS:

ne vidi razloga za to, da bi Trump o **domnevnem strastnem** srečanju s Karo lagal. "Konec koncev," ljubljeno najstniško hčerko. Na drugi strani pa mu življenje **pestri strastna** in zapeljiva ljubica Oliva. Katera stran V prejšnji številki so bila v fotografski objektiv **ujeta strastna** čustva Steffania do takrat še neznane blondinke. zavoda Svoboda osvobaja, v prostor, ki sicer pozna **mnoga strastna** čustva, velike zamisli in logične premisleke, Mnogi vidijo uspeh judizma v njihovem **nenehnem strastnem** iskanju smisla: v Bogu, stvarjenju, človekovi Je (do)končno zapuščanje teritorija in ustvarjanje **ekskluzivnega strastnega** razmerja s heroinom, tokrat z težišče belogardizma pa je na oglatih, brezobzirnih, **derviško strastnih** laikih. Bistvo klerikalizma je v hladnem triu op. 66 v začetnem in končnem stavku sledimo **skladateljevemu strastnemu** sporočilu, podobna živopisnost rudimentarna zmes kitare, sinkopiranega ploskanja, eventualnih kastanjet in **animalično strastnega** glasu, nekaj denarja in v šali dodal, da je zaradi **Paspaljevega strastnega** kajenja njegovo posteljo večkrat želel

delo. V ljubezni lahko do 16. novembra pričakujete **določena strastna** doživetja. Kljub temu zaradi velike na drugi strani<, kjer smo bili priče spočetju **novega strastnega** ljubezenskega razmerja med naivnima vendar je njegovo poslanstvo to, kar Mrak polaga v **Iškarjotova strastna** vprašanja: "Zakaj preizkušnja? Zakaj temperamentno in hkrati intrigantsko pismo Frana Govekarja Vladimirju Levcu, **polno strastne** vitalnosti in podlaga za sentiment, ki pa mestoma zraste v pravo **tangovsko strastno** odbijanje in približevanje tudi z udarci oddaljenih koncih planeta, ki še ni videlo in slišalo **mojstrovega strastnega** igranja. Zdaj pa kaže, da bo letos v Ali kako ljubijo Mehičani. Evforični avtobiografski zapis pisateljevega **nekajmesečnega strastnega** razmerja s

Konkordančni niz FidaPLUS 7: #2P??s* _#1strasten.

Razvidno je, da se kljub ustreznemu označevanju nekaterih primerov pojavljajo označevalne napake, ki bi bile potencialno rešljive z upoštevanjem samostalniškega jedra besedne zveze pri označevanju določujočih pridevnikov. Sicer je na tem mestu izpostavljeni problem obravnavan tudi na drugih mestih, npr. v V-1.1.5.1.

1.3.5 Prisl + strasten

Nabor zvez, izluščenih za obravnavani vzorčni tip, prinaša potencialno zanimive besedne nize – v primeru, da je na ravni vnosa v leksikalno zbirko relevantna informacija o tem, na kakšen način je obravnavani pridevnik z levim določilom tipično določen. V nasprotnem primeru je smiselno nize obravnavati v sklopu daljših vzorčnih tipov (prim. poglavje V-1.4.1).

besedni niz	pogostnost v korpusu	hkrati	strasten	
najbolj	strasten 120	zasebno	strasten	6
zelo	strasten 86	divje	strasten	5
tako	strasten 82	lahko	strasten	5
bolj	strasten 77	prav	strasten	5
malo	strasten 30	hudo	strasten	5
veliko	strasten 27	toliko	strasten	4
kako	strasten 18	vseeno	strasten	4
nekaj	strasten 16	sem	strasten	4
sicer	strasten 16	posebno	strasten	4
izjemno	strasten 16	dokaj	strasten	4
nekoč	strasten 13	pogosto	strasten	3
nadvse	strasten 11	obenem	strasten	3
vedno	strasten 11	očitno	strasten	3
enako	strasten 10	noro	strasten	3
dovolj	strasten 8	najprej	strasten	3
pretirano	strasten 7	resnično	strasten	3
zdaj	strasten 7	naravnost	strasten	3
precej	strasten 6	verjetno	strasten	3
izredno	strasten 6	nekaj	strasten	3
več	strasten 6			

Tabela 63: Izluščeni podatki – Prisl + strasten.

Intuicija je, da je nabor prislovov, ki se tipično pojavljajo pred pridevniki, ki se lahko stopnjujejo, v določeni meri predvidljiv¹⁰¹: na seznamu je poleg zvez s stopnjevalnima prislovoma *bolj* in *najbolj* pričakovati še vrsto mernih

¹⁰¹ Jasno je, da začetna izbira štirih lem za podrobno obravnavo v veliki meri vpliva na nabor vzorčnih tipov, ki so v doktorskem delu zajeti v analizo. Če bi bil izbran pridevnik (ali prislov; glej Vidovič Muha 2000: 76), ki izkazuje vrstnost namesto kakovostnosti, bi bila torej v ospredju druga vprašanja.

prislovov (npr. *zelo*, *dovolj*, *precej* itd.). Z luščenjem primerov za dovolj velik nabor kakovostnih pridevnikov bi bilo mogoče skupino mernih prislovov natančneje identificirati ter načeloma zamejiti, izdelani seznam pa z nadaljnjim luščenjem zgolj dopolnjevati oz. ga uporabljati kot referenčno listo za iskanje primerov, ki pri posameznem obravnavanem pridevniku odstopajo od pričakovanega. Več o luščenju nizov s prislovi sledi v poglavju V-1.7.3

Pri obravnavi nizov prislova in pridevnika brez samostalniškega jedra analiza gornjih primerov kaže, da ločevanje glede na spol ni potrebno. Glede na nizek priklic prislovov v primerniški oz. presežniški obliki (tudi stopnja pri prislovu je označena samo v primerih, ko je izražena morfološko, opisno stopnjevanje ostaja zunaj dometa oblikoskladenjskega označevanja) se prav tako zdi glede na trenutne podatke manj smiselno ločevanje zvez glede na stopnjo prislova.

1.3.6 Glag + *strasten*

Kot je razvidno iz nadaljevanja, besedni nizi, ki jih prinaša vzorčni tip Glag + *strasten*, sami na sebi niso relevantni za nadaljnjo obravnavo. Analiza nizov v daljših vzorčnih tipih sledi v V-1.4.2:

besedni niz	pogostnost v korpusu	pritegniti	strasten	
preživeti	67	splesti	strasten	4
postati	39	naplesti	strasten	4
razviti	25	preživljati	strasten	4
imeti	19	delati	strasten	4
začeti	11	občutiti	strasten	3
privoščiti	8	prepletati	strasten	3
vneti	8	posvečati	strasten	3
želei	7	pisati	strasten	3
predati	7	predajati	strasten	3
slediti	6	govoriti	strasten	3
pričeti	6	umreti	strasten	3
gojiti	4	skrivati	strasten	3
pričarati	4	pozabiti	strasten	3

Tabela 64: Izluščeni podatki – Glag + *strasten*.

1.4 Strasten – tridelni vzorci

Nabor najpogostejših tridelnih vzorcev, ki prinašajo lemo *strasten* in dve polnopomenski besedi, je nekoliko daljši od nabora za lemo *pajek* – prinaša 24 primerov, ki jih je mogoče razvrstiti v pet vzorčnih tipov:

vzorčni tip	število vzorcev
Prisl + <i>strasten</i> + Sam	4
Glag + <i>strasten</i> + Sam	4
<i>strasten</i> + Prid + Sam	2
<i>strasten</i> + Sam + Prid	3
<i>strasten</i> + Sam + Sam	8
drugi vzorci	3
RNN RNN STRASTEN	
RNN GGDSDD STRASTEN	
SOMEI STRASTEN SOMETD	

Tabela 65: Nabor najpogostejših tridelnih vzorcev z lemo *strasten* in oblikoskladenjskimi oznakami za polnopomenske besede.

1.4.1 Prisl + *strasten* + Sam

Kot je bilo izpostavljeno v V-1.3.1, je pri luščenju zvez samostalnika z levim pridevniškim prilastkom potrebna ločena obravnava primerov glede na spol samostalnika. Rezultate za vse tri skupine predstavljajo besedni nizi s pogostnostjo 2 ali več:

besedni niz – m. spol	pog. v korpusu	tako	strasten	kadilec	2
najbolj strasten kadilec	10	najbolj	strasten	spev	2
najbolj strasten ribič	7	vseeno	strasten	privrženec	2
najbolj strasten navijač	5	najbolj	strasten	zbiratelj	2
malo strasten kadilec	5	vedno	strasten	pivec	2
najbolj strasten zagovornik	4	slepo	strasten	tempo	2
zelo strasten ljubimec	4	zdaj	strasten	biljardist	2
bolj strasten poljub	4	enako	strasten	zavijalec	2
najbolj strasten ljubitelj	4	zelo	strasten	par	2
tako strasten gobar	4	tako	strasten	igralec	2
nekaj strasten poljub	3	tako	strasten	odnos	2
tako strasten ribič	3	najbolj	strasten	znanstvenik	2
izjemno strasten človek	3	zelo	strasten	človek	2
najbolj strasten ljubimec	3	nekdanj	strasten	kadilec	2
dokaj strasten poljub	3	sicer	strasten	kadilec	2
tako strasten ljubitelj	3	najbolj	strasten	vraževerec	2
najbolj strasten trenutek	2				

Tabela 66: Izluščeni podatki – Prisl + *strasten* + Sam.

besedni niz – ž. spol	pog. v korpusu	zelo	strastna	noč	2
zelo strastna ženska	4	zelo	strastna	nрав	2
najbolj strastna noč	3	pretirano	strastna	planinka	2
nekaj strastna noč	3	najbolj	strastna	ljubezen	2
zelo strastna narava	3	tako	strastna	kota	2
bolj strastna nakupovalka	3	zelo	strastna	oseba	2
tako strastna kadilka	2				

Tabela 67: Izluščeni podatki – Prisl + *strastna* + Sam.

besedni niz – s. spol	pog. v korpusu
najbolj strastno cimljenje	2
naravnost strastno zavzemanje	2
dolgo strastno poljubljanje	2

Tabela 68: Izluščeni podatki – Prisl + *strastno* + Sam.

Kot je bilo izpostavljeno v V-1.3.5, se določen nabor prislovov v gornjih zvezah ponavlja, kar potrjuje smiselno izdelavo in uporabo referenčnega seznama mernih prislovov. Na tem mestu je potrebno še opozorilo na probleme avtomatskega ločevanja med primeri tipa *malo strastnih kadilcev* ter *malo strasten kadilec*. Za identifikacijo bi bilo v obravnavo potrebno pritegniti upoštevanje sklona samostalnika oz. samostalniške besedne zveze (tj. ločevanje primerov z rodilniškimi oblikami od ostalih).

1.4.2 Glag + *strasten* + Sam

Pri obravnavi pričujočega vzorčnega tipa je potrebno ponoviti opozorilo o zadržkih oz. omejitvah luščenja besednih nizov, ki posegajo na skladiščno raven. V spodnji tabeli je na ravni samostalnika ter pridevnika ohranjena izpričana oblika, glagol pa je v osnovni obliki:

besedni niz			pog. v korp.				
preživeti	strastno	noč	58	umreti	strastne	urice	3
razviti	strastna	romanca	5	splesti	strasten	kadilec	3
razviti	strastna	ljubezen	5	postati	strastna	zveza	3
razviti	strastna	privlačnost	5	začeti	strasten	inkvizitor	3
vneti	strastna	ljubezen	5	posvečati	strastno	razmerje	3
pričeti	strastno	razmerje	5	preživljati	strastnemu	zbiranju	3
razviti	strastno	razmerje	4	preživljati	strastne	noči	3
naplesti	strastna	ljubezen	4	postati	strasten	lovec	3

Tabela 69: Izluščeni podatki – Glag + *strasten* + Sam.

V primerih, kjer je besedna zveza v imenovalniku oblikovno drugačna od tožilniške, je omogočeno ločevanje nizov glede na predvidene stavčnočlenske odnose izvornih stavkov (prim. besedno zvezo v vlogi predmeta v *preživeti strastno noč* in osebka v *razviti strastna romanca*). Drugače je v primerih, kjer sta obliki prekrivni (npr. *[pričeti, razviti, začeti] strastno razmerje*). Na teh mestih je za razvrstitev potrebna ročna analiza konkordančnega niza.

1.4.3 *Strasten* + Prid + Sam

Glede na sestavljenost pričujočega vzorčnega tipa je pri prikazu podatkov ponovno potrebno upoštevanje ujemalnosti pridevnikov s samostalnikom:

besedni niz – m. spol			pog. v korp.				
strasten	nogometen	navijač	11	strastna	ljubezenska	zveza	6
strasten	ljubezenski	prizor	8	strastna	mlada	umetnica	5
strasten	nogometen	navdušenec	3	strastna	ljubezenska	romanca	5
strasten	športen	ribič	3	strastna	surrealistična	zgodba	4
strasten	italijanski	potapljač	2	strastna	ljubezenska	noč	4
strasten	podvoden	ribič	2	strastna	ljubezenska	dogodivščina	4
strasten	masoven	nacizem	2	strastna	dvorna	gospodična	3
strasten	ljubiteljski	vrtnar	2	strastna	telesna	ljubezen	2
strasten	ljubezenski	krč	2	strastna	nova	generacija	2
strasten	eksperimentalen	fizik	2	strastna	etnična	identifikacija	2
strasten	spolen	odnos	2	strastna	ljubezenska	izkušnja	2
strasten	romantičen	film	2	strastna	ljubezenska	naveza	2
strasten	intimen	odnos	2	strastna	danska	kadilka	2
				besedni niz – s. spol			
besedni niz – ž. spol				strastno	ljubezensko	razmerje	24
strastna	ljubezenska	zgodba	28	strastno	ljubezensko	pismo	3
strastna	ljubiteljska	vrtnarka	6				

Tabela 70: Izluščeni podatki – *strasten* + Prid + Sam, vsi spoli.

Pri obravnavanih primerih je pretvorb pridevniških oblik veliko in ker slednje trenutno potekajo ročno, je priprava nabora besednih nizov dokaj zamudna. Za potrebe izboljšave metode luščenja je potreben premislek o lematizaciji pridevnikov na način, da bi bila spolu ustrezna osnovna pridevniška oblika dostopna v besedni

oznaki.¹⁰² Ob vseh izpostavljenih rešitvah, ki jih prinaša za avtomatsko obdelavo jezika premik od enobesednosti k besednozveznosti, se zdi vztrajanje pri pripisovanju zgolj nezaznamovane oblike vsem pridevnikom nesmiselno, saj v jezikovni rabi pridevniki spol vedno izražajo, četudi je vezan na kontekst – obenem pa je ta izraženost, kot izpričujejo primeri, na ravni besednih zvez drugačnega pomena kot denimo izražanje števila ali sklona.

1.4.4 Strasten + Sam + Prid

Oblikoskladenjska oznaka za pridevnik na zadnjem mestu po do sedaj obravnavanih podatkih pogosto pomeni nedokončanost izluščenih zvez. Spodnji nabor besednih nizov potrjuje intuicijo, da obravnavani vzorčni tip ni relevanten za nadaljnjo obravnavo. Seznam prinaša pridevnik *strasten* ter samostalniki v osnovni obliki, drugi pridevnik pa v izpričani obliki:

besedni niz		pogostnost v korpusu	
strasten	prebujenje	mladostne	8
strasten	grof	Vronskim	6
strasten	zbiratelj	starih	5
strasten	motorist	blage	4
strasten	zbiralec	umetniških	4
strasten	zbiralec	starih	3
strasten	ljubitelj	stare	3

Tabela 71: Izluščeni podatki – *strasten* + Sam + Prid.

1.4.5 Strasten + Sam + Sam

Kot je bilo prikazano v poglavju V-1.1.1, zahteva luščenje besednih nizov za vzorčne tipe, ki vsebujejo Sam + Sam, posebno metodo. Program, primerljiv opisanemu v navedenem poglavju, za tri- ali večdelne vzorce trenutno še ni na voljo, zato so na tem mestu predstavljeni le nizi z drugim samostalnikom v izpričani rodilniški obliki. Za obravnavani pridevnik *strasten* so tovrstne zveze namreč najpogostejše, kar je razvidno iz seznama vzorcev, zajetih v pričujoči vzorčni tip: zveze z rodilnikom zavzemajo kar 84 % vseh (522 od 620). Spodnji podatki so ločeni glede na spol jedrnega samostalnika:

besedni niz – m. spol			pog. v korp.				
strasten	igralec	golfa	16	strasten	zbiratelj	starin	4
strasten	zbiralec	umetnin	11	strasten	ljubitelj	golfa	4
strasten	ljubitelj	glasbe	7	strasten	izliv	ljubezni	3
strasten	ljubitelj	knjig	6	strasten	bralec	poezije	3
strasten	ljubitelj	narave	6	strasten	kadilec	cigaret	3
strasten	zbiralec	orožja	6	strasten	ljubitelj	športa	3
strasten	ljubitelj	iger	5	strasten	zbiralec	znamk	3
strasten	zbiralec	starin	5	strasten	ljubitelj	opere	3
strasten	zbiratelj	umetnin	5	strasten	iskalec	resnice	3
strasten	igralec	lota	5	strasten	navijač	nogometašev	3
strasten	bralec	časopisov	4	strasten	ljubitelj	konj	3
strasten	iskalec	prave	4	strasten	uničevalec	iluzije	3
strasten	ljubitelj	orhidej	4	strasten	kadilec	cigar	3
strasten	zbiralec	plošč	3				

¹⁰² Sicer večji odstop od trenutne označevalne prakse bi bila denimo dvonivojskost leme v smislu pripisovanja (I) tako spolu ustrezne pridevniške oblike kot (II) leme za nezaznamovani spol. S stališča luščenja podatkov, kakršno je opisano na tem mestu, je zaželena vsebovanost ustrezne oblike znotraj oznake. Možno je sicer naknadno pridobivanje podatkov o ustreznih pridevniških oblikah (iz leksikona besednih oblik ali npr. korpusa, kot je predlagano v Erjavec in Vintar 2008: 68), kar pa pomeni dodaten korak v metodi luščenja. Najpreprostejša, a manj prijetna za nadaljnjo obravnavo zvez, pa je seveda ohranitev oblik v nezaznamovani različici, npr. *strasten ljubezenski zgodba*, *strasten seksualen srečanje*.

strasten	ljubitelj	nogometa	3				
strasten	zbiratelj	orožja	3				
besedni niz – ž. spol				besedni niz – s. spol			
strastna	reševalka	križank	3	strastno	odkrivanje	umetnosti	3
strastna	igra	draguljev	3	strastno	zbiranje	instrumentov	3

Tabela 72: Izluščeni podatki – *strasten* + Sam + Sam, vsi spoli.

1.4.6 Preostali vzorci

Vzorci, ki niso bili uvrščeni v nobenega od do sedaj predstavljenih vzorčnih tipov, so trije. Za vsakega od njih prinaša spodnja tabela seznam zapolnitev s pogostnostjo 3 ali več.

RNN RNN STRASTEN			pogostnost v korpusu
prav	tako	strastni	4
prav	tako	strastna	4
eden	najbolj	strastnih	3
precej	bolj	strastna	3
tako	najbolj	strasten	3
kar	nekaj	strastnih	3
res	zelo	strasten	3
RNN GGSDDD STRASTEN			
skupaj	preživita	strastno	12
kmalu	pričneta	strastno	5
SOMEI STRASTEN SOMETD			
kot	strastnega	športnika	12

Tabela 73: Nabor preostalih tridelnih vzorcev z lemo *strasten* in dvema polnopomenskima besedama.

Če nas, kot rečeno v V-1.3.5, s stališča gradnje zbirke zanima določanje pridevnika kot takega, je potencialno uporaben prvi od navedenih vzorcev – oz. podatki, ki jih lahko pridobimo z luščenjem vzorčnega tipa **Prisl + Prisl + *strasten***.

Ostala dva vzorca ne prinašata za uvrstitev v zbirko zanimivih zapolnitev, prvi zaradi nevsebovanja jedra samostalniške besedne zveze (*skupaj preživita strastno*), drugi zaradi neustrezne označenosti veznika *kot* za samostalnik (*kot strastnega športnika*).

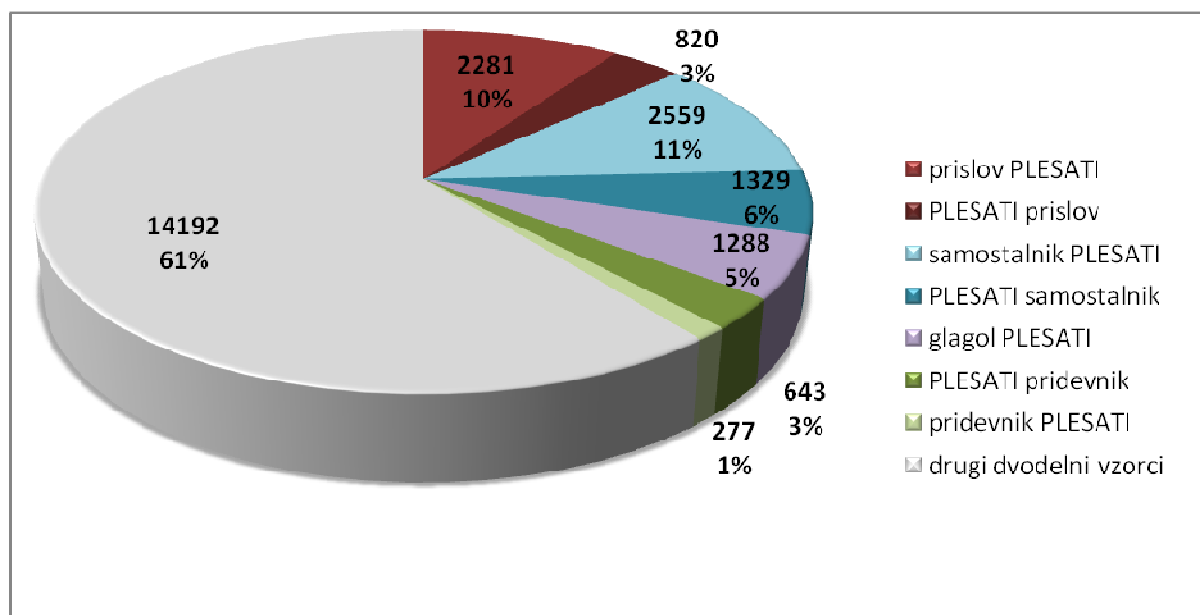
1.5 Plesati – dvodelni vzorci

Najpogostejših 61 dvodelnih vzorcev z lemo *plesati* ter oznako za polnopomensko besedo lahko glede na besednovrstno opredelitev spremljajoče oblikoskladenjske oznake uvrstimo v sedem različnih tipov: lema *plesati* se pojavlja s prislovom, samostalnikom ter pridevnikom na levi oz. desni strani ali z glagolom na levi:

vzorčni tip	število vzorcev
Prisl + <i>plesati</i>	1
<i>plesati</i> + Prisl	2
<i>plesati</i> + Prid	8
Prid + <i>plesati</i>	2
Glag + <i>plesati</i>	14
<i>plesati</i> + Sam	12
Sam + <i>plesati</i>	22

Tabela 74: Razvrstitev najpogostejših vzorcev z lemo *plesati* in oblikoskladenjskimi oznakami za polnopomenske besede v vzorčne tipe.

Graf z informacijo o vsoti pogostnosti vseh korpusnih pojavitev za vsakega od zgoraj definiranih tipov kaže, da so med obravnavanimi vzorci v korpusu najpogostejše kombinacije s samostalnikom (11 %) oz. prislovom (10 %) na levi od obravnavanega glagola. Prav tako so pogoste kombinacije s samostalnikom na desni (6 %) ter glagolom na levi (5 %). Sledijo kombinacije s prislovom (3 %) oz. pridevnikom (3 %) na desni, še redkejše pa so kombinacije s pridevnikom na levi (1 %). Sicer obravnavani vzorci predstavljajo manjši delež, kot se je izkazalo pri obravnavi lem *pajek* ter *strasten* – le 39 % vseh.



Graf 3: Vsota pogostnosti pojavitev vzorcev za dvodelne vzorčne tipe z lemo *plesati* in oblikoskladenjskimi oznakami za polnopomenske besede.

V nadaljevanju so predstavljeni posamezni vzorčni tipi.

1.5.1 Prisl + *plesati*

Obravnavani vzorčni tip prinaša številne besedne nize, zato so v nadaljevanju navedeni le tisti s pogostnostjo 5 ali več. Najvišje v seznamu nizov je primer *rad plesati* (508), ki sicer združuje oblike za različne spole in števila – ločevanje se na tem mestu ni izkazalo za smiselno.¹⁰³

¹⁰³ Vprašanju avtomatskega označevanja povedkovnika se pričujoče delo ne posveča, na tem mestu zato le na kratko: ob upoštevanju dejstva, da se označevalna praksa trudi vsaj v izhodišču ločevati nivoje označevanja (oblikoskladenjsko, skladenjsko, semantično označevanje), se zdi na oblikoskladenjski ravni smiselno vztrajati pri pripisovanju predvsem slovarskega tipa leksikalnih informacij, kar bi v konkretnem primeru pomenilo označevanje nabora potencialnih (spol izražajočih) povedkovnikov za pridevnike. Trenutno so označevani kot prislovi, kar deloma izvira iz specifik oblikoskladenjskih paradig, ki jih izražajo.

besedni niz		pogostnost v korpusu			
rad	plesati	508	navdušeno	plesati	9
lahko	plesati	171	nato	plesati	9
dobro	plesati	96	neutrudno	plesati	9
vedno	plesati	63	kmalu	plesati	8
veliko	plesati	63	neumorno	plesati	8
skupaj	plesati	54	letos	plesati	8
kako	plesati	47	treba	plesati	8
več	plesati	45	takrat	plesati	8
tako	plesati	39	nikoli	plesati	8
najraje	plesati	36	sedaj	plesati	7
veselo	plesati	35	redno	plesati	7
divje	plesati	33	preprosto	plesati	7
lepo	plesati	33	razposajeno	plesati	7
spet	plesati	32	zjutraj	plesati	7
zdaj	plesati	31	zraven	plesati	7
naprej	plesati	25	ponavadi	plesati	7
tam	plesati	25	takole	plesati	7
sam	plesati	23	zares	plesati	7
nekoč	plesati	23	malo	plesati	7
odlično	plesati	21	izvrstno	plesati	7
kar	plesati	20	slabo	plesati	7
danes	plesati	18	izzivalno	plesati	7
prvič	plesati	17	tukaj	plesati	6
trenutno	plesati	16	dobesedno	plesati	6
najprej	plesati	16	znova	plesati	6
včasih	plesati	14	živahno	plesati	6
pogosto	plesati	13	hkrati	plesati	6
nekdaj	plesati	12	dolgo	plesati	6
bolj	plesati	12	nocoj	plesati	6
res	plesati	12	enkrat	plesati	6
potem	plesati	11	težko	plesati	6
sicer	plesati	11	večkrat	plesati	6
kdaj	plesati	11	ponovno	plesati	5
aktivno	plesati	10	tja	plesati	5
raje	plesati	10	uspešno	plesati	5
noro	plesati	10	zapeljivo	plesati	5
zvečer	plesati	9	pridno	plesati	5

Tabela 75: Izluščeni podatki – Prisl + *plesati*.

Vprašanje smiselnosti ločevanja prislovov glede na stopnjo se je odprlo že na drugih mestih (glej V-1.3.5), kjer je bilo tudi zapisano, da takšno ločevanje (tudi zaradi nizkega priklica) ni smiselno; v pričujočem poglavju, kjer je nabor stopnjevanih prislovov večji, so za primer in v premislek ločeno navedeni besedni nizi, kjer se prislovi pojavljajo v primerniški oz. presežniški obliki. Seznam je naveden v celoti, podatki pa za obe stopnji združeni:

besedni niz		pogostnost v korpusu			
najbolje	plesati	10	boljše	plesati	1
bolje	plesati	8	slabše	plesati	1
manj	plesati	5	težje	plesati	1
najlepše	plesati	4	najpogostejše	plesati	1
lažje	plesati	3	pozneje	plesati	1
več	plesati	3	prijetneje	plesati	1
najdlje	plesati	1	globlje	plesati	1
lepše	plesati	1	pogostejše	plesati	1
			laže	plesati	1

hitreje	plesati	1	močnejše	plesati	1
dlje	plesati	1			

Tabela 76: Izluščeni podatki – Prisl (primernik/presežnik) + plesati.

Kot je razvidno iz primerov, ločena obravnava zvez s stopnjevanimi oblikami sicer prinaša določen nabor informacij o obravnavanem glagolu, vendar se na tem mestu potrjuje intuicija, da so te za vključitev v leksikalno zbirko obrobnejšega pomena.

1.5.1.1 Analiza označenosti

V pričujočem poglavju se analiza označenosti osredotoča na en sam problematičen primer: kot bo vidno tudi iz tabele v poglavju V-1.5.4, se primeri tipa *sam plesati* pojavljajo označeni kot kombinacije glagola *plesati* bodisi s prislovom bodisi pridevnikom. Iz konkordančnih nizov v korpusu FidaPLUS za kombinacije s pridevnikom (konkordančni niz 8) ter prislovom (konkordančni niz 9) je razvidno – spodaj je za vsakega od nizov naveden vzorec 10 konkordanc – da ločevanje pri označevanju ni konsistentno:

I) beseda *sam* je označena kot pridevnik

naju ni pustil na kakšno veselico, pa sva velikokrat **sama plesala** v tem taktu. Imam pa ohranjen še en tram delokrog skoraj povsem usmerjen na plesno sceno, kjer še **sama plešem**, koreografiram, raziskujem, tudi sta pripravila. Zamikalo me je, da bi še **sam plesal**."

erotičnost pri njem sploh ne pride v poštev, dekleta **sama plešejo**, fantje sami, je uživanje nad gibanjem samim,

, lahko bi tudi pel, plesal... **Sami plešete** ?

Ker je oboževal Freda Astaireja, je začel tudi **sam plesati** step. Povedal je, da mu je šlo tako

Nežni gibi balerin Billyja popolnoma prevzamejo in kmalu začne tudi **sam plesati**. Wilkinsonova je nad sliši, in pove, da je pred leti tudi **sama plesala**.

je po mnenju glasbenih kritikov takšen, da noge kar **same plešejo**. Fantje, ki so že imeli turnejo po ZDA šanku, cuzajo pivo in buljijo v punčke, ki **same plešejo**. Potem gredo domov, kjer se ata in mama

Konkordančni niz FidaPLUS 8: *sam*#2P*_#1plesati*.

II) beseda *sam* je označena kot prislov

na primer vidi v lokalih v Novi Gorici: Rusinje **samo plešejo**, z Bolgarkami, Ukrajinkami in Romunkami pa se da so vas videli samo kot plesalko. Bi vam bilo **samo plesati** premalo

22.25 Ples na severu: Četrtek Želiva si **samo plesati**, francosko plesni film

studio je poslal helikopterje z reševalci, medtem ko je **sama plesala** rumbo z Mitchumom. Preračunljivost ali

POTEM: "Če **sam plešeš** striptiz, je to seksi, če to počne tujec

. Obred se je bližal koncu. Sredi dvorane je **sama plesala** le še Nemka, zatem ko je z nevoščljivim pogledom telo, mišična masa je popolnoma drugače razporejena. Odkar **sama plešem** pri En knapu, sem pridobivala na

, dekleta prihajajo plesat v ta lokal.<< **Samo plesat** ?<< Odvisno od razpoloženja

je to vsakdanje delo. Dolgo let je tudi **sama plesala**, sedaj pa poskuša znanje prenesti otrokom.

V Butterflyu niso **samo plesale**

Konkordančni niz FidaPLUS 9: *sam*#2R*_#1plesati*.

Očitno je, da je na tem mestu potrebno poenotenje označevanja. V slovarskih priročnikih (SSKJ, SP) je za večino gornjih primerov sicer predlagana besednovrstna uvrstitev besede *sam* med zaimke, za prislov pa gre v primerih tipa *Rusinje samo plešejo*.

1.5.2 Plesati + Prisl

Nabor spodnjih primerov je za uvrstitev v leksikalno zbirko le delno zanimiv. Nerelevantni so denimo nizi tipa *plesati [tako, zelo, kar, res]*. Zanimivi pa so npr. primeri *plesati [skupaj, pozno, tesno, dolgo]*. Avtomatsko ločevanje med primeri trenutno ni mogoče.

Ker se mestoma na levi od glagola pojavljajo enaki prislovi kot na desni (npr. *plesati [skupaj, veliko, več, naprej ...]*), bi bila s primerjavo obeh seznamov mogoča pridobitev dodatne informacije o sopojavljanju različnega tipa prislovov z obravnavanim glagolom.¹⁰⁴

besedni niz		pogostnost v korpusu		
plesati	tako	118	plesati doma	5
plesati	skupaj	75	plesati dolgo	5
plesati	veliko	33	plesati skozi	4
plesati	okoli	31	plesati značilno	4
plesati	kar	28	plesati res	4
plesati	več	24	plesati zgoraj	4
plesati	naprej	24	plesati sem	4
plesati	zelo	23	plesati povsem	4
plesati	sam	21	plesati povsod	4
plesati	pozno	18	plesati drugače	4
plesati	tesno	16	plesati izključno	4
plesati	dobro	16	plesati najprej	4
plesati	okrog	16	plesati vedno	4
plesati	tja	14	plesati avtonomno	4
plesati	lahko	12	plesati daleč	3
plesati	bolj	12	plesati solo	3
plesati	naokrog	12	plesati visoko	3
plesati	rad	11	plesati neprenehoma	3
plesati	tam	10	plesati zares	3
plesati	čisto	8	plesati hkrati	3
plesati	zato	8	plesati popolnoma	3
plesati	prav	6	plesati drugam	3
plesati	nekje	6	plesati dovolj	3
plesati	malo	6	plesati komaj	3
plesati	sicer	5	plesati slabo	3
plesati	zdaj	5	plesati največ	3
plesati	toliko	5		

Tabela 77: Izluščeni podatki – *plesati* + Prisl.

¹⁰⁴ Zanimiv bi bil morda poskus avtomatske identifikacije prislovov, ki so besednoredno bolj vezani kot drugi.

1.5.2.1 Analiza označenosti

Na tem mestu se za problematično izkazuje predvsem vprašanje uspešnosti avtomatskega ločevanja med prislovi ter enakopisnimi predlogi (npr. *okoli*, *naokrog*, *skozi*). Pregled zadetkov korpusa FidaPLUS za oba primera kaže, da je trenutno ločevanje med besednima vrstama manj uspešno, kar izkazujejo spodnji primeri (po 10 zadetkov za vsakega od iskalnih pogojev):

I) beseda *okoli* je označena kot prislov

Kje je pust? Kje so maškare, ki so **plesale okoli** moje postelje, ko sem se jaz davil s slino
vzame Taylor obleko. Pop glasba se zdi glasnejša. **Pleše okoli** Taylor. Njena noga udarja ritem. Rame se ji
nastavljala pred njim! Vsak dan v novih cunjicah je **plesala okoli** Tima, kot bi odkrivala spomenik narodnim herojem! Ko
pa po njeni izrecni želji. Tip, ki je **plesal okoli** Jenny, gleda za njima očitno slabe volje; Jennyjin
hiši so bliski, če se tako izrazimo, dobesedno **plesali okoli** glave in kot po čudežu se ni nikomur skrivil niti
ga skupaj z mrtvim kozlom. Vsa vas je začela **plesati okoli** njega in peti. Nenadoma pa so potegnili odejo z
piše. Je opazno večja od drugih, ki **plešejo okoli** nje, se sučejo in nenavadno gibljejo, kakor da
Morda se zavedate bliskov informacij, ki **plešejo okoli** vas in z vami... nenaden utrinek svetlobe
sem pel, ko sem s hčerkama v naročju **plesal okoli** hiše. Njuni tanki lasi so me žgečkali po vratu
Palico moramo imeti vseskozi pod kontrolo in ne sme **plesati okoli** roke, ko jo spustimo na koncu odriva. Pomembno

Konkordančni niz FidaPLUS 10: #1plesati_okoliR*.

II) beseda *okoli* je označena kot predlog

najboljši, ta bo krasen minister, je menda kar **plesal okoli** njega. In premier je na koncu prikimal. Da
Noge uživajo v hladnih valčkih plitvih jezerc, majhne postrvi **plešejo okoli** gležnjev, težko se je sploh spomniti na žalostno
se svojemu približam brez hlačk, tudi gola sem že **plesala okoli** njega in ga izzivala, a če ima v mislih
Tem pa tudi pojema sapa, in kot mačke **plešejo okoli** vrele kaše, ti iščejo rešitev, kako bi njihov
Leta 1993 je po ulicah Berlina na Love parade **plesalo okoli** 30.000 pretežno zadetih plesalcev, leta 1999 pa
hodili po razmetanih vasicah sredi ničesar in zvečer peli in **plesali okoli** skupnega ognja. Vaščane sem sicer prosil za to in
to, da sem vedno v ženini senci. Vsi **plešejo okoli** nje, vsi jo poznajo, jaz pa sem "
vedno manjšo površino jader), svetleča konica kompasa pa **pleše okoli** severa. Za nama se po južnem nebu vlečejo temni
Mladi **plešejo okoli** policistov

Žal bo morala Bolgarova dama v nadaljevanju igre **plesati okoli** kmeta c3 kot kakšen lačen medved. Star bolgarski pregovor

Konkordančni niz FidaPLUS 11: #1plesati_okoliD*.

Pregled celotnih konkordančnih nizov za oba primera kaže, da beseda *okoli* v obravnavanem besednem nizu *plesati okoli* nastopa bodisi kot predlog, v redkih primerih pa kot članek približnostne mere; prislovna raba v korpusu FidaPLUS sploh ni izpričana. Rešitev označevalnih problemov tega tipa bi bilo že predlagano

upoštevanje skladijske okolice pri razdvoumljanju oznak – pri čemer se dvoumnim oblikam z (v sklonu ujemajočim) se samostalnikom na desni pripiše večja verjetnost, da so besednovrstno predlogi.

1.5.3 Plesati + Prid

Z upoštevanjem izvirnega konteksta spodnjih primerov je mogoče ugotoviti, da so s skladijskega vidika spodnji primeri treh tipov: pridevnik je bodisi v vlogi (I) levega prilastka samostalniške besedne zveze (*plesati [folklorne, klasični, standardne]* – ti primeri sami na sebi niso relevantni, ker zveza ni celotna) bodisi v vlogi (II) povedkovega prilastka (npr. *plesati [bos, sama, naga]*), ali pa konvertirani v stavku nastopajo v vlogi (III) osebka (*plesati [mladi, stari]*).

besedni niz	pogostnost v korpusu	plesati	orientalski	4
plesati samo	21	plesati sami		4
plesati folklorne	14	plesati ritualni		4
plesati klasični	12	plesati folklorna		4
plesati sama	9	plesati prekmurske		4
plesati standardne	8	plesati latinskoameriške		4
plesati belokranjske	7	plesati tradicionalna		3
plesati tradicionalne	7	plesati žensko		3
plesati bos	7	plesati bele		3
plesati dunajski	7	plesati gorenjske		3
plesati sam	6	plesati grške		3
plesati kozjanske	5	plesati same		3
plesati srbsko	5	plesati zmagoviti		3
plesati akrobatski	5	plesati Labodje		3
plesati mladi	5	plesati klasičen		3
plesati glavno	5	plesati nori		3
plesati grški	5	plesati sodobni		3
plesati ljudske	5	plesati latino		3
plesati tradicionalni	5	plesati trebušni		3
plesati naga	5	plesati bananin		3
plesati celo	5	plesati cele		3
plesati trebušne	5	plesati smrtni		3
plesati mlade	5	plesati cela		3
plesati stari	5	plesati veseli		3
plesati divji	5	plesati različne		3
plesati bosa	4	plesati bosi		3
plesati mlada	4	plesati primabalerina		3
plesati srednjeveške	4	plesati stare		3
plesati štajerske	4			

Tabela 78: Izluščeni podatki – plesati + Prid.

Avtomatsko ločevanje med naštetimi tipi je možno le z upoštevanjem skladijskega konteksta, ne pa tudi izključno na ravni oblikoskladijskih oznak. Nekatere od primerov v daljših besednih nizih prinaša poglavje V-1.6.1.

1.5.4 Prid + plesati

Kot je vidno iz spodnje tabele, je na tem mestu ponovno najti primere različnih vrst: na eni strani številne primere kombinacij z (I) napačno označenimi prislovi (npr. [*samo, dobro, razigrano, dolgo*] *plesati*; glej V-1.1.5.1, V-1.3.4.1) oz. skladijskimi (II) osebki (npr. *nastopajoči plesati*) ali (III) povedkovimi prilastki (npr. *pijan plesati*). Kot je bilo rečeno v prejšnjem poglavju, je avtomatsko ločevanje med primeri zgolj na osnovi oblikoskladijskih oznak trenutno nemogoče.

besedni niz	pogostnost v korpusu	sam	plesati	
samo plesati	27	pripravljen	plesati	3
dobro plesati	20	naga	plesati	3
celo plesati	14	dano	plesati	3
sama plesati	7	sposobno	plesati	3
razigrano plesati	7	pijan	plesati	3
dolgo plesati	6	zavzeto	plesati	3
malo plesati	6	nastopajoči	plesati	3
objeta plesati	5	kratkim	plesati	3
sposobni plesati	5	veseli	plesati	3
sproščeno plesati	5	nagi	plesati	3
mogoče plesati	4			

Tabela 79: Izluščeni podatki – Prid + *plesati*.

1.5.5 Glag + *plesati*

Pričujoči vzorčni tip prinaša besedne zveze dveh glagolov. Večina izluščenih primerov prinaša zveze, v katerih glagol *plesati* nastopa v obliki nedoločnika (npr. *[začeti, znati, naučiti] plesati*) ali namenilnika (npr. *[iti, hoditi, priti] plesat*).

besedni niz	pogostnost v korpusu	odpraviti	plesat	
začeti plesati	244	zahoteti	plesati	7
znati plesati	160	ljubiti	plesati	6
iti plesat	113	marati	plesati	6
naučiti plesati	93	nameravati	plesati	6
učiti plesati	82	dovoliti	plesati	5
hoteti plesati	67	pustiti	plesati	4
želeči plesati	59	znati plesat		4
morati plesati	50	začeti plesat		4
nehati plesati	46	odločiti	plesati	4
videti plesati	41	upati	plesati	4
hoditi plesat	37	poskušati	plesati	3
moči plesati	33	dejati	plesat	3
smeti plesati	21	spraviti	plesat	3
prenehati plesati	17	želeči plesat		3
priti plesat	14	oditi	plesat	3
dati plesati	13	siliti	plesati	3
pričeti plesati	11	misлити	plesati	3
povabiti plesat	11	naučiti plesat		3

Tabela 80: Izluščeni podatki – Glag + *plesati*.

Dragocena informacija, razvidna iz primerov, je, da korpusna raba izkazuje variantnost pri uporabi nedoločnika oz. namenilnika v nekaterih od primerov (npr. *začeti plesati/plesat*, *želeči plesati/plesat*), pri čemer je sicer potrebno upoštevati razlike v pogostnosti rabe, ki je tipično v prid nedoločniški obliki. V gornji tabeli so primeri dvojne rabe označeni sivo.

1.5.6 *Plesati* + Sam

Ob upoštevanju zadržkov, izraženih v V-1.1.9, spodnji seznam prinaša za uvrstitev v leksikalno zbirko potencialno zanimive besedne nize, med katerimi prednjačijo na eni strani primeri obravnavanega glagola v kombinaciji s samostalnikom v tožilniku, ki izraža tip plesa (*plesati [tango, kolo, balet, step]*), na drugi strani v imenovalniku, ki izražajo skladenjski osebek (*plesati [otroci, Mojca, člani, Alenka]*).

besedni niz	pogostnost v korpusu	plesati	rokenrol	5
plesati tango	74	plesati ženske		5
plesati kolo	66	plesati leta		5
plesati balet	65	plesati sence		5
plesati step	58	plesati gole		4
plesati kot	37	plesati čardaš		4
plesati valček	35	plesati plesalke		4
plesati striptiz	32	plesati kan		4
plesati salso	29	plesati Hana		4
plesati ples	27	plesati fokstrot		4
plesati polko	26	plesati baleta		4
plesati četvorko	22	plesati račke		4
plesati hip	21	plesati Tomaž		4
plesati sambo	19	plesati balerina		4
plesati jazz	17	plesati dejan		3
plesati flamenko	16	plesati Manca		3
plesati polke	14	plesati čačača		3
plesati moški	11	plesati striptiza		3
plesati breakdance	10	plesati prava		3
plesati plese	10	plesati ljudje		3
plesati otroci	10	plesati solisti		3
plesati Mojca	10	plesati duhovi		3
plesati kankan	10	plesati Andreja		3
plesati gola	10	plesati tvist		3
plesati dekleta	9	plesati flamenco		3
plesati pravljico	8	plesati laboda		3
plesati vlogo	8	plesati Olga		3
plesati twist	8	plesati prav		3
plesati rock	7	plesati Rosana		3
plesati člani	7	plesati pav		3
plesati kazačok	7	plesati miši		3
plesati kola	7	plesati snežinke		3
plesati plesalci	7	plesati maja		3
plesati dan	6	plesati čarovnice		3
plesati tanga	6	plesati džez		3
plesati greh	6	plesati klon		3
plesati skupina	6	plesati swing		3
plesati članice	6	plesati privid		3
plesati Alenka	6	plesati iskrice		3
plesati rumbo	6	plesati mambo		3
plesati Regina	6	plesati maturanti		3
plesati Goran	6	plesati koreografijo		3

Tabela 81: Izluščeni podatki – *plesati* + Sam.

V sklopu obravnave vzorčnih tipov z lemo *plesati* besedni nizi niso nadalje ročno razvrščeni glede na sklon samostalnikov. Za primer postopka glej V-1.1.10.

1.5.6.1 Analiza označenosti

Problem, evidentiran na tem mestu, je označevanje tožilniških samostalniških oblik za imenovalniške. Spodaj je naveden seznam besednih nizov, pri katerih je samostalnik označen kot imenovalnik ednine moškega spola (za druga dva spola je nabor primerljiv navedenemu):

PLESATI SOMEI		pogostnost v korpusu	plesati	odsev	1
plesati	tango	33	plesati	šov	1
plesati	striptiz	21	plesati	hudič	1
plesati	hip	21	plesati	lik	1
plesati	valček	17	plesati	prizor	1
plesati	balet	19	plesati	obup	1
plesati	ples	14	plesati	član	1
plesati	jazz	14	plesati	stari	1
plesati	step	10	plesati	občutek	1
plesati	kot	9	plesati	kraljevič	1
plesati	moški	8	plesati	duet	1
plesati	flamenko	8	plesati	show	1
plesati	greh	6	plesati	labod	1
plesati	rock	6	plesati	klovn	1
plesati	kan	4	plesati	pajacek	1
plesati	twist	4	plesati	pevec	1
plesati	swing	3	plesati	gol	1
plesati	privid	3	plesati	tvist	1
plesati	pav	3	plesati	oče	1
plesati	rokenrol	3	plesati	kantri	1
plesati	kazačok	3	plesati	Demeter	1
plesati	džez	3	plesati	riž	1
plesati	pod	2	plesati	metuljček	1
plesati	los	2	plesati	go	1
plesati	prav	2	plesati	svet	1
plesati	fokstrot	2	plesati	čarlston	1
plesati	kankan	2	plesati	Pinč	1
plesati	solo	2	plesati	bolero	1
plesati	čačača	2	plesati	duhovnik	1
plesati	brat	2	plesati	Ginger	1
plesati	baletnik	2	plesati	mož	1
plesati	Klon	2	plesati	pravi	1
plesati	človek	2	plesati	čas	1
plesati	solist	1			

Tabela 82: Analiza označenosti – PLESATI SOMEI.

Razen nekaterih izjem (npr. *plesati [kot, moški, pod, brat, človek, Demeter, Ginger]* itd.) gre v gornji tabeli za primere, kjer bi moral biti sklon samostalnika označen za tožilnik – ob zavedanju, da je v naboru pričakovati tudi trpniške primere, kjer je označenost samostalnika v imenovalniku ustrezna (primer iz korpusa FidaPLUS: *Ha, tako se pleše tango!*). Avtomatsko ločevanje teh primerov od ostalih trenutno ni mogoče.

1.5.7 Sam + plesati

Pričujoči vzorčni tip ponovno prinaša različne vrste primerov: samostalnik pred glagolom *plesati* je skladenjsko bodisi (I) osebek (*[zvezde, zemlja, ljudje, otroci] plesati*), (II) predmet (*[tango, kolo] plesati*), lahko pa gre za (III) napačno označen pridevnik v vlogi povedkovega prilastka (*gola plesati*) ali (IV) kakšno drugo besedno vrsto (*[raje, kot, tam] plesati*). Pojavljajo se tudi primeri s samostalnikom v mestniku oz. orodniku, ki so kot predložne zveze obravnavane v sklopu poglavja V-2.10.4, npr. *[zraku, skupinah, Zmincem] plesati*.

besedni niz		pogostnost v korpusu	zemlja	plesati	15
miši	plesati	26	noč	plesati	13
zvezde	plesati	25	raje	plesati	13
let	plesati	19	ljudje	plesati	13

otroci	plesati	12	tam	plesati	3
večer	plesati	11	začetku	plesati	3
čas	plesati	11	turneji	plesati	3
dan	plesati	11	Dorica	plesati	3
časa	plesati	10	odru	plesati	3
ženske	plesati	9	ligi	plesati	3
dekleta	plesati	9	gola	plesati	3
ljudi	plesati	8	Slovencev	plesati	3
leto	plesati	8	skupinah	plesati	3
skupini	plesati	8	upokojenci	plesati	3
Slovenija	plesati	8	klubu	plesati	3
leta	plesati	7	željo	plesati	3
plesalke	plesati	7	kolo	plesati	3
predstavi	plesati	7	demoni	plesati	3
kongres	plesati	6	dvorišču	plesati	3
pari	plesati	6	cesti	plesati	3
Slovenci	plesati	6	letih	plesati	3
plesalci	plesati	6	Zmincem	plesati	3
zraku	plesati	6	medo	plesati	3
kot	plesati	6	prijatelji	plesati	3
moški	plesati	6	dvorani	plesati	3
leti	plesati	5	vrsti	plesati	3
hormoni	plesati	5	Tržič	plesati	3
Kamnik	plesati	4	univerza	plesati	3
Sloveniji	plesati	4	času	plesati	3
tedna	plesati	4	šoli	plesati	3
prijateljicami	plesati	4	zgoščenke	plesati	3
plese	plesati	4	dni	plesati	3
plesalka	plesati	4	poroki	plesati	3
konji	plesati	4	vlogi	plesati	3
stoletja	plesati	4	štoru	plesati	3
glasbi	plesati	4	obiskovalci	plesati	3
tango	plesati	4	folkloru	plesati	3
trgu	plesati	4	glasbe	plesati	3
maturanti	plesati	3	jesen	plesati	3
kresnice	plesati	3	glasbo	plesati	3
sloni	plesati	3	Krško	plesati	3

Tabela 83: Izluščeni podatki – Sam + plesati.

Izpostaviti je potrebno predvsem primere tipa *[noč, večer, čas] plesati*, pri katerih kljub navidezni oblikovni ustreznosti ne gre za stavčni osebek, ampak del zveze, ki predstavlja v stavku prislovno določilo (primer iz korpusa FidaPLUS: *Nizozemci so dolgo v noč plesali zmagoviti ples.*). Navedeni poudarek je podkrepitev dejstev, da je (I) predstavljena metoda luščenja podatkov za obravnavo zvez, ki presegajo raven oblikoskladnje, manj ustreznost ter da je (II) ročna analiza izluščenih podatkov na vsak način nujen del predstavljene metode.

1.6 Plesati – tridelni vzorci

Za analizo je na voljo 8 tridelnih vzorcev z lemo *plesati* ter dvema polnopomenskima besedama. Na osnovi vzorcev je bil izdelan en sam vzorčni tip:

vzorčni tip

število vzorcev

plesati + Prid + Sam

2

drugi vzorci

6

RNN RNN PLESATI

RNN PLESATI RNN

PLESATI RNN RNN

PLESATI SLZEI SLMEI

PPNZEM SOZEM PLESATI

SOMMI RNN PLESATI

Tabela 84: Nabor najpogostejših tridelnih vzorcev z lemo *plesati* in oblikoskladenjskimi oznakami za polnopomenske besede.

1.6.1 Plesati + Prid + Sam

Spodnje besedne nize lahko – ponovno z upoštevanjem konteksta – razvrstimo na primere, ki prinašajo zveze skladenjskega povedka z (I) osebkom (*plesati folklorne skupine*), (II) predmetom (*plesati orientalske plese*) ali (III) prislovnim določilom (*plesati celo noč*). Avtomatsko ločevanje med primeri je trenutno neizvedljivo, ker zahteva upoštevanje pomena – problemi so denimo že na ravni razvrščanja primerov glede na sklon samostalniške besedne zveze, kadar gre za dvoumne oblike tipa *plesati klasični balet*.

besedni niz			pog. v korpusu	plesati	celonočne	plese	
plesati	folklorne	skupine	11	plesati	edina	pličanič	2
plesati	klasični	balet	9	plesati	sodobni	ples	2
plesati	dunajski	valček	6	plesati	divja	jaga	2
plesati	ljudske	plese	5	plesati	škotske	plese	2
plesati	belokranjske	plese	5	plesati	orientalske	plese	2
plesati	srbsko	kolo	5	plesati	preprosti	ljudje	2
plesati	glavno	vlogo	5	plesati	domače	plese	2
plesati	tradicionalne	plese	5	plesati	družabne	plese	2
plesati	orientalski	ples	4	plesati	indijanske	plese	2
plesati	celo	noč	4	plesati	mladi	grofje	2
plesati	folklorne	skupina	4	plesati	paško	kolo	2
plesati	trebušne	plese	4	plesati	ljubljski	balet	2
plesati	ritualni	ples	4	plesati	kmečki	fantje	2
plesati	standardne	plese	4	plesati	osvajalski	ples	2
plesati	srednjeveške	plese	4	plesati	izrazni	ples	2
plesati	grške	plese	3	plesati	veseli	ljudje	2
plesati	smrtni	ples	3	plesati	krilati	mladci	2
plesati	folklorne	plese	3	plesati	irske	plese	2
plesati	naga	dekleta	3	plesati	mlada	ženska	2
plesati	labodje	jezero	3	plesati	belokranjske	pesmi	2
plesati	zmagoviti	ples	3	plesati	ritualne	plese	2
plesati	klasičen	balet	3	plesati	vilinsko	kolo	2
plesati	primabalerina	gledališča	3	plesati	ljubke	balerine	2
plesati	žensko	vlogo	3	plesati	veliko	število	2
plesati	trebušni	ples	3	plesati	mlade	ženske	2
plesati	narodne	plese	2	plesati	obredni	ples	2
plesati	balkansko	kolo	2	plesati	metliško	obredje	2
plesati	bele	pike	2	plesati	brezova	metla	2
plesati	počasne	plese	2				

Tabela 85: Izluščeni podatki – *plesati* + Prid + Sam.

1.6.2 Preostali vzorci

Za vsakega od preostalih vzorcev, ki se niso uvrstili v vzorčne tipe, so v nadaljevanju predstavljeni po 3 najpogostnejši primeri zapolnitev.

RNN RNN PLESATI	pogostnost v korpusu
zelo rad plesati	74
zelo dobro plesati	10
tako dobro plesati	8
RNN PLESATI RNN	
lahko plesati same	4
dejansko plesati zato	2
lahko plesati tu	2
PLESATI RNN RNN	
plesati tesno skupaj	6
plesati zelo dobro	5
plesati kar tako	3
PLESATI SLZEI SLMEI	
plesati Mojca Majcen	6
plesati Regina Križaj	6
plesati Hana Cimperman	4
PPNZEM SOZEM PLESATI	
folklorni skupini plesati	5
zadnji turneji plesati	3
belokranjski inačici plesati	2
SOMMI RNN PLESATI	
ljudje radi plesati	5
mladi radi plesati	3
otroci radi plesati	3

Tabela 86: Nabor preostalih tridelnih vzorcev z lemo *plesati* in dvema polnopomenskima besedama.

Razen za peti vzorec v gornji tabeli, ki prinaša primere s samostalniško besedno zvezo, vendar brez predloga (*folklorni skupini plesati*), so izluščeni podatki za uvrstitve v zbirko potencialno relevantni – ob upoštevanju, da prinašajo po večini podatke, ki jih ni smotno obravnavati izključno na oblikoskladenjski ravni. To velja tako za primere, ki na skladenjski ravni izvornih stavkov izražajo povedek ter osebek (*plesati Mojca Majcen; ljudje radi plesati*), kot tudi za primere s prislovnimi določili (*zelo rad plesati; plesati tesno skupaj*).

1.7 Temeljito

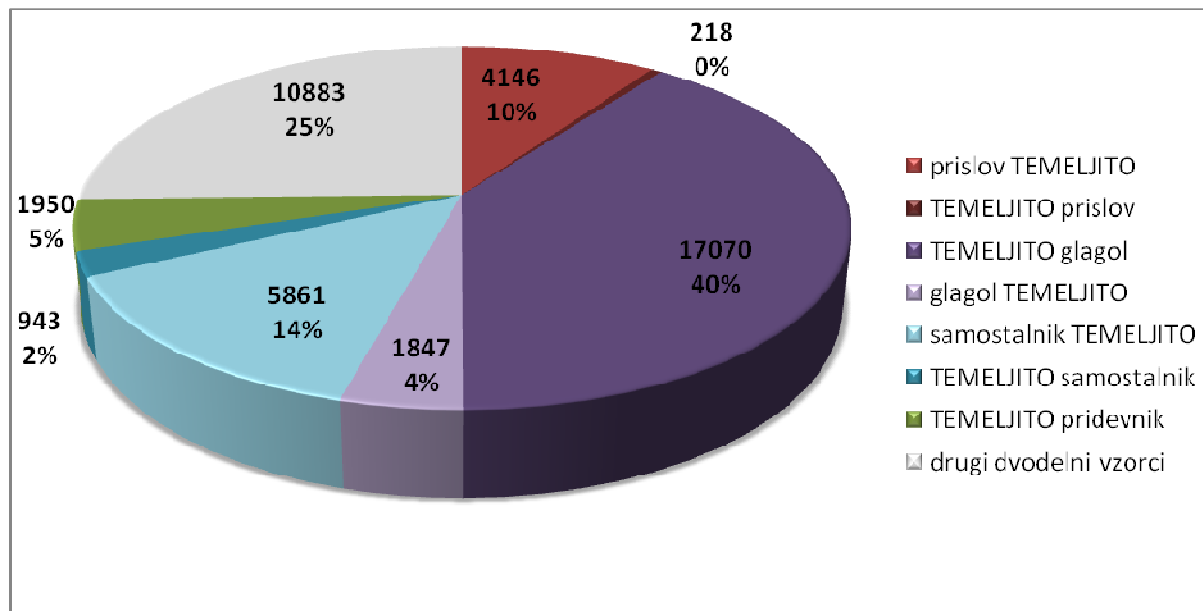
Najpogostejših 73 vzorcev z lemo *temeljito* in polnopomensko besedo lahko razvrstimo v sedem vzorčnih tipov: s prislovom, glagolom ali samostalnikom na levi oz. desni ter pridevnikom na levi od obravnavane leme:¹⁰⁵

vzorčni tip	število vzorcev
Prisl + <i>temeljito</i>	1
<i>temeljito</i> + Prisl	1
<i>temeljito</i> + Glag	24
Glag + <i>temeljito</i>	11
Sam + <i>temeljito</i>	25

¹⁰⁵ Ker so oblikoskladenjske oznake za prislove bistveno manj členjene od ostalih oznak za polnopomenske besede (glej Priloga 1) in ker se kaže za smotno v povezavi s spodaj prikazanim grafom, sta na tem mestu izjemoma vzorčna tipa *temeljito* + Prisl ter Prisl + *temeljito* izdelana na osnovi enega samega vzorca.

<i>temeljito</i> + Sam	3
<i>temeljito</i> + Prid	8

Tabela 87: Razvrstitev najpogostejših vzorcev z lemo *temeljito* in oblikoskladenjskimi oznakami za polnopomenske besede v vzorčne tipe.



Graf 4: Vsota pogostnosti pojavitev vzorcev za dvodelne vzorčne tipe z lemo *temeljito* in oblikoskladenjskimi oznakami za polnopomenske besede.

Kot je razvidno iz grafičnega prikaza, zajemajo v Tabeli 87 navedeni vzorčni tipi kar s 75 % vseh dvodelnih vzorcev z obravnavano lemo. Velika večina odpade na kombinacije prislova *temeljito* z glagolom na desni (40 %). Sledijo kombinacije s samostalnikom (14 %) oz. prislovom (10 %) na levi, kombinacije s pridevnikom na desni (5 %), glagolom na levi (4 %) ter samostalnikom na desni (2 %). Najmanj pogoste so kombinacije s prislovom na desni (manj kot 1 %).

1.7.1 Prisl + *temeljito*

Na prvi pogled seznam kombinacij prislova *temeljito* s prislovom na desni prinaša predvidljive rezultate. Precej visoko pogostnost imajo kombinacije z besedami, ki so na skladenjski ravni del povedka, npr. [*treba, lahko, potrebno*] *temeljito*, visoko na seznamu sta seveda tudi niza s stopnjevalnima prislovoma [*bolj, najbolj*] *temeljito*. Sicer so v tabeli po večini zveze s prislovi načina in mere (npr. [*zelo, dovolj, precej*] *temeljito*) ter kombinacije s prislovi drugih vrst, npr. časovnimi ([*lani, včeraj, oktobra*] *temeljito*). Vprašanje, ali je besedi *temeljito* pripisana prislovna lema ustrezna ali gre morda za napačno označeno pridevniško obliko, je brez upoštevanja konteksta nerešljivo.

Ker je seznam precej dolg, so na tem mestu navedeni le primeri s pogostnostjo 5 ali več.

besedni niz	pogostnost v korpusu	najbolj	temeljito	116	
treba	temeljito	389	potrebno	temeljito	105
tako	temeljito	331	skupaj	temeljito	88
najprej	temeljito	258	dovolj	temeljito	87
lahko	temeljito	233	prej	temeljito	87
bolj	temeljito	216	res	temeljito	67
zelo	temeljito	131	kako	temeljito	55
enkrat	temeljito	119	vedno	temeljito	53

nato	temeljito	48	torej	temeljito	11
zdaj	temeljito	46	verjetno	temeljito	11
letos	temeljito	44	včeraj	temeljito	11
posebej	temeljito	41	nujno	temeljito	11
zares	temeljito	38	poprej	temeljito	11
takoj	temeljito	28	vnaprej	temeljito	11
spet	temeljito	25	zvečer	temeljito	10
prvič	temeljito	23	doma	temeljito	10
lani	temeljito	21	sedaj	temeljito	10
zato	temeljito	21	vsakič	temeljito	10
večkrat	temeljito	20	strokovno	temeljito	10
sicer	temeljito	20	doslej	temeljito	9
očitno	temeljito	20	kmalu	temeljito	9
rad	temeljito	19	nikoli	temeljito	9
ponovno	temeljito	19	zagotovo	temeljito	9
potem	temeljito	19	kasneje	temeljito	8
nekajkrat	temeljito	19	vselej	temeljito	8
predhodno	temeljito	18	takrat	temeljito	8
veliko	temeljito	18	zjutraj	temeljito	8
malo	temeljito	17	nedavno	temeljito	8
dvakrat	temeljito	17	izredno	temeljito	8
dokaj	temeljito	17	pozno	temeljito	7
posebno	temeljito	17	zadnjič	temeljito	7
nekoliko	temeljito	17	izjemno	temeljito	7
sam	temeljito	16	gotovo	temeljito	7
znova	temeljito	16	občasno	temeljito	7
končno	temeljito	16	pogosto	temeljito	6
tokrat	temeljito	16	naprej	temeljito	6
resnično	temeljito	16	raje	temeljito	6
mogoče	temeljito	15	prihodnje	temeljito	5
precej	temeljito	15	kdaj	temeljito	5
kar	temeljito	15	karseda	temeljito	5
čim	temeljito	15	domov	temeljito	5
nadvse	temeljito	14	trikrat	temeljito	5
medtem	temeljito	14	prav	temeljito	5
nazadnje	temeljito	13	čimprej	temeljito	5
čimbolj	temeljito	12	spomladi	temeljito	5
danes	temeljito	11			

Tabela 88: Izluščeni podatki – Prisl + temeljito.

Da bi bili rezultati uporabni za vnos v leksikalno zbirko, bi bilo treba luščenje nadgraditi. S tukaj predstavljeno metodo avtomatsko ločevanje med primeri, ki jih imamo lahko za prislovne besedne zveze (tip *zelo temeljito*), od kombinacij, ki na skladenjski ravni izražajo več prislovnih določil (tip *včeraj temeljito*), ni mogoče. Več o možnostih izboljšave luščenja zvez Prisl + Prisl sledi v poglavju V-1.7.3.

1.7.2 Temeljito + Prisl

V primerjavi s prejšnjim vzorčnim tipom je trenutno obravnavani v korpusu relativno redek, pojavlja se z enim samim vzorcem z 218 zapolnitvami. Kot je razvidno s seznama, so primeri, kakršni so luščeni na tem mestu, oblikovno dvomni – iz besednega niza ni jasno, ali prislov *temeljito* nastopa v zvezi s prislovom na desni ali gre za zveze z napačno označenimi pridevniki:

besedni niz	pogostnost v korpusu	temeljito	javno	4
temeljito strokovno	31	temeljito	tehnično	4
temeljito kadrovske	9	temeljito	globinsko	4
temeljito vsebinsko	7	temeljito	zakonsko	3
temeljito zvočno	6	temeljito	pravno	3
temeljito oblikovno	6	temeljito	lahko	3
temeljito varnostno	5	temeljito	literarno	3
temeljito mehansko	5	temeljito	sistematsko	3
temeljito gospodarsko	5	temeljito	raziskovalno	3
temeljito zdravstveno	5			

Tabela 89: Izluščeni podatki – temeljito + Prisl.

1.7.2.1 Analiza označenosti

Pregled konkordančnega niza za prve tri najpogostejše primere Tabele 89 v korpusu FidaPLUS kaže tako zveze s pridevnikom kot prislovom, kar s stališča avtomatskega označevanja pomeni, da je doseg višje natančnosti mogoč le z upoštevanjem besednozveznega konteksta problematične besedne oblike. Za primer je navedenih nekaj konkordanc za iskalni pogoj *temeljito_strokovno*:

v uradni jezik SES. Sestavljanje vsebine vprašanja seveda zahteva **temeljito strokovno** pripravo in preučitev izluščili skupne predloge in vsak od njih bo šel skozi **temeljito strokovno** presojo in pravno analizo. V skladu s alkohola. Za to delo se je kljub slabemu vidu **temeljito strokovno** izpopolnil v Zagrebu, kjer je opravil v začetku leta 1995. Zato je potrebno predloženi zakon **temeljito strokovno** proučiti. je še precej vprašanj, tako da bo potrebno stvar **temeljito strokovno** in urbanistično preučiti. Čeprav Alojz zakonske predloge še pred vložitvijo v postopek pred državnim zborom **temeljito strokovno** in iz prakse obdelati da bodo firme vstopale v te procese veliko bolj **temeljito strokovno** pripravljene. Najmanj potrebno in

Konkordančni niz FidaPLUS 12: *temeljito_strokovno*.

Dani iskalni pogoj sicer v korpusu FidaPLUS vrne 37 zadetkov, iskanje z omejitvijo glede besednovrstne označenosti (iskanje kombinacij s primeri, kjer je beseda *strokovno* označena za prislov) vrne 26 zadetkov, med katerimi gre za zveze s prislovi le v 14 primerih. Avtomatsko ločevanje med pridevniki ter prislovi se je pokazalo za problematično že na drugih mestih, zato na tem mestu le reference na V-1.1.5.1, V-1.3.4.1, V-1.5.4.

1.7.3 Luščenje zvez Prisl + Prisl

Za razliko od označevanja drugih polnopomenskih besednih vrst, pri katerih se trenutni nivo členjenosti oznak pogosto kaže za preveč specifičnega za luščenje, se pri prislovi nasprotno odpira vprašanje dodatnega označevanja vrste prislova.

Zadržki so na tem mestu jasni. Metoda avtomatskega ločevanja prislovov v ustrezne skupine, temelječa bodisi na podatkih iz leksikalne zbirke bodisi iz skladenjsko označenega korpusa, trenutno še ni na voljo, kar pomeni, da bi bilo kakršno koli kategorizacijo treba izvesti ročno oz. polavtomatsko.¹⁰⁶ Prav tako še ni raziskano, katere

¹⁰⁶ Razmislek gre v smer uporabe nabora prislovov, evidentiranih v slovarskem delu Slovenskega pravopisa (kjer so prislovi kvalificirani glede na vrsto), vendar je slednje izvedljivo le v primeru, da so izluščeni sezname prislovov ročno pregledani, ovrednoteni glede uporabnosti za oblikoskladenjsko označevanje in preverjeni s stališča realne jezikovne rabe. Možno bi bilo izdelati izhodiščni nabor prototipskih prislovov za vsako od

kategorije vrste bi bilo prislovom sploh smiselno pripisovati. Glede na to, da se v povezavi z razvrščanjem prislovov odpira tudi vprašanje ravni avtomatske pripisljivosti, s tem pa natančnosti ter priklica pri luščenju označenih podatkov, se zdi označevanje razvejanega nabora prislovnih vrst, kakršnega denimo prinaša slovnica (Toporišič 2004⁴: 407–409), trenutno težje izvedljivo in manj smiselno. Poskus ločevanja na osnovnejši ravni, ki bi izviralo iz potreb obdelave naravnega jezika, pa se kljub vsemu kaže za zaželenega.¹⁰⁷

Toporišič npr. na najvišji ravni ločuje *okolščinske prislove* – »prislovi, ki dobivajo prilastek za seboj« – od *svojstvenostnih* – »tisti, ki določilo dobivajo pred seboj« (Toporišič 2004⁴: 407–408). Delitev prislovov na ta način se zdi za avtomatsko obdelavo smiselna, zlasti zato, ker temelji na pozicijskem kriteriju. Če sledimo kategorizaciji prislovnih besednih zvez v slovnici, se zdi za luščenje (bolj od določitve jedra v posameznem tipu prislovnih zvez) koristna hipoteza, da je v zvezah z jedrom, ki je okoliščinski prislov, enakega tipa tudi desni prilastek – denimo *včeraj jutraj* ali *tu spodaj*. Na drugi strani so zveze, ki imajo za jedro svojstvenostni ali okoliščinski prislov, za določilo pa levi prilastek svojstvenostnega tipa – denimo *zelo počasi* ali *zelo levo*, pa tudi npr. *gospodarsko neupravičeno* (Toporišič 2004⁴: 568–571).

Obenem se na prvi pogled kaže vredno poskusa ločevanje svojstvenostnih prislovov na skupino *prislovov mere* (ter morda *kratnosti*¹⁰⁸) na eni strani od t. i. *prislovov ozira* (na osnovi česar bi bilo omogočeno avtomatsko ločevanje primerov tipa *zelo neupravičeno* od *gospodarsko neupravičeno*). Razlog leži v intuiciji, da prva skupina predstavlja relativno zamejen del besedišča, ki ga je mogoče avtomatsko izluščiti, nato pa dobljeni seznam po potrebi dopolnjevati, medtem ko za drugo skupino slednje ne velja – obenem predstavljajo prislovi ozira nabor, ki zaradi prekrivnosti s pridevniško obliko na ravni avtomatskega označevanja zahteva posebno pozornost.

Čeprav je pričujoče poglavje osredotočeno na luščenje zvez dveh prislovov, pa predlagana metoda rešuje tudi v prejšnjih poglavjih izpostavljene probleme označevanja ter luščenja zvez določujočega prislova ter pridevnika, za katere je prav tako rečeno, da »imajo pridevniki, kadar gre za lastnostno določitev (in vzročnostno), določila pred seboj, kadar pa gre za t. i. okoliščine prostora, časa – za seboj,« pri čemer se ponovno izkazuje za ločeno skupina mernih prislovov (Toporišič 2004⁴: 565).

1.7.4 Temeljito + Glag

O zvezah glagola z določujočim prislovom je bilo nekaj napisanega že v V-1.5.1 in V-1.5.2. Na tem mestu predstavljeni vzorčni tip se izkazuje za izredno pogostnega (najbolj pogostnega med vsemi, obravnavanimi v pričujoči raziskavi), spodnji seznam prinaša zato le primere s pogostnostjo 20 ali več.

besedni niz	pogostnost v korpusu	
temeljito spremeniti	963	temeljito preoblikovati 47
temeljito pregledati	857	temeljito preštudirati 45
temeljito pripraviti	818	temeljito prezidati 45
temeljito obnoviti	735	temeljito prežvečiti 43
temeljito očistiti	642	temeljito oklestiti 41
temeljito premisliti	553	temeljito poučiti 40
temeljito prenoviti	532	temeljito preučevati 38
temeljito preučiti	388	temeljito zdrgniti 38
temeljito razmisliti	348	temeljito sanirati 37

izbranih kategorij ter v nadaljevanju s pomočjo statističnih metod luščiti ostale kandidate za posamezno skupino.

¹⁰⁷ Koristna bi bila npr. že samo izdelava referenčnega seznama prislovov mere, ki tipično določajo sledeči prislov oz. pridevnik (glej V-1.3.5).

¹⁰⁸ Na tem mestu operiramo s kategorijami, kakršne so predstavljene v slovnici; če se izkaže, da je za avtomatsko obdelavo to smiselno, je seveda možna drugačna razvrstitev v skupine (in obenem poimenovanje skupin).

temeljito oprati	343	temeljito zamajati	37
temeljito premešati	328	temeljito poglobiti	37
temeljito pogovoriti	315	temeljito posušiti	36
temeljito pripravljati	302	temeljito pojasniti	35
temeljito preveriti	217	temeljito pregledovati	35
temeljito lotiti	207	temeljito izboljšati	34
temeljito preiskati	201	temeljito pogledati	33
temeljito proučiti	198	temeljito posvetovati	33
temeljito raziskati	198	temeljito preskusiti	33
temeljito obdelati	196	temeljito skrtačiti	32
temeljito ogledati	180	temeljito razpravljati	32
temeljito odcediti	167	temeljito prepričati	31
temeljito analizirati	166	temeljito umešati	31
temeljito seznaniti	163	temeljito raziskovati	31
temeljito umiti	149	temeljito čistiti	31
temeljito pretresti	143	temeljito pregnesti	31
temeljito sprati	126	temeljito razmišljati	31
temeljito pretehtati	125	temeljito oprhati	30
temeljito zmešati	123	temeljito mešati	30
temeljito posvetiti	121	temeljito preverjati	30
temeljito preurediti	120	temeljito zagreniti	29
temeljito prevetriti	112	temeljito oceniti	29
temeljito prebrati	106	temeljito razložiti	28
temeljito predstaviti	101	temeljito prebrskati	28
temeljito popraviti	99	temeljito zamisliti	27
temeljito predelati	97	temeljito razčleniti	27
temeljito počistiti	94	temeljito obravnavati	27
temeljito izkoristiti	94	temeljito reformirati	27
temeljito izprašati	93	temeljito namočiti	26
temeljito ukvarjati	90	temeljito umivati	26
temeljito opraviti	89	temeljito nadzorovati	26
temeljito spreminjati	83	temeljito zaslišati	26
temeljito spoznati	80	temeljito splakniti	25
temeljito preizkusiti	77	temeljito proučevati	24
temeljito obnavljati	76	temeljito zaznamovati	24
temeljito zaliti	74	temeljito zmanjšati	24
temeljito osušiti	73	temeljito pretakniti	23
temeljito prečesati	72	temeljito pozanimati	23
temeljito odstraniti	70	temeljito zbrisati	23
temeljito posodobiti	69	temeljito spremljati	23
temeljito poznati	69	temeljito posneti	23
temeljito izprati	62	temeljito reorganizirati	23
temeljito prenavljati	61	temeljito obrezati	23
temeljito obrisati	61	temeljito ohladiti	22
temeljito prezračiti	54	temeljito preiskovati	22
temeljito pospraviti	54	temeljito izrabiti	22
temeljito poseči	52	temeljito zaščititi	21
temeljito urediti	48	temeljito razkužiti	20
temeljito poskrbeti	47	temeljito morati	20

Tabela 90: Izluščeni podatki – *temeljito* + Glag.

Kljub omejitvam pri luščenju glagolskih zvez na osnovi oblikoskladenjskih oznak se zdi nabor predstavljenih informacij zanimiv za vključitev v leksikalno zbirko.

1.7.5 Glag + temeljito

Za razliko od prejšnjega vzorčnega tipa se na tem mestu ponovno pojavlja vprašanje ustrezne označenosti dvoumnih oblik za prislovne in ne pridevniške.

besedni niz	pogostnost v korpusu			
morati	temeljito	500	objaviti	temeljito 6
začeti	temeljito	149	pridobiti	temeljito 6
opraviti	temeljito	106	skušati	temeljito 6
lotiti	temeljito	83	ukazati	temeljito 6
zahtevati	temeljito	55	dočakati	temeljito 6
želeti	temeljito	44	pozabiti	temeljito 5
odločiti	temeljito	42	slediti	temeljito 5
nameravati	temeljito	41	izdelati	temeljito 5
pripraviti	temeljito	41	predstaviti	temeljito 5
doživeti	temeljito	36	očistiti	temeljito 5
izvesti	temeljito	24	dobiti	temeljito 5
uspeti	temeljito	24	končati	temeljito 5
dati	temeljito	23	iti	temeljito 5
potrebovati	temeljito	22	vzeti	temeljito 4
odpreti	temeljito	20	imeti	temeljito 4
moči	temeljito	18	pustiti	temeljito 4
veljati	temeljito	17	naročiti	temeljito 4
napovedati	temeljito	16	spremljati	temeljito 4
pripravljeni	temeljito	15	obljubiti	temeljito 4
opravljati	temeljito	15	priporočati	temeljito 4
narediti	temeljito	14	izdati	temeljito 4
znati	temeljito	12	omogočiti	temeljito 3
hoteti	temeljito	12	zagovarjati	temeljito 3
kazati	temeljito	12	misliti	temeljito 3
skleniti	temeljito	11	lotevati	temeljito 3
omogočati	temeljito	11	pokazati	temeljito 3
izpeljati	temeljito	11	izkazati	temeljito 3
pričakovati	temeljito	10	početi	temeljito 3
predlagati	temeljito	9	preživeti	temeljito 3
utegniti	temeljito	9	zagotoviti	temeljito 3
splačati	temeljito	9	deti	temeljito 3
zaslužiti	temeljito	8	postaviti	temeljito 3
poskušati	temeljito	8	sprožiti	temeljito 3
zagotavljati	temeljito	8	obravnavati	temeljito 3
pričeti	temeljito	8	zalivati	temeljito 3
zaliti	temeljito	7	začenjati	temeljito 3
terjati	temeljito	7	poskusiti	temeljito 3
načrtovati	temeljito	7	videti	temeljito 3
izvajati	temeljito	7	odrediti	temeljito 3
čistiti	temeljito	7	delati	temeljito 3

Tabela 91: Izluščeni podatki – Glag + temeljito.

Natančnejša analiza označenosti na tem mestu ne sledi, saj je bil problem označevanja prislovnih oz. pridevniških oblik že večkrat izpostavljen. Primere, kjer je obravnavana beseda *temeljito* nedvomno prislov, prinašajo besedni nizi, predstavljeni v poglavju V-1.8.2.

1.7.6 Sam + temeljito

Kot je razvidno iz nadaljevanja, na tem mestu obravnavani vzorčni tip ne prinaša za vključitev v leksikalno zbirko relevantnih besednih nizov (navedeni so primeri s pogostnostjo 10 ali več).

besedni niz	pogostnost v korpusu	policisti	temeljito	
letih	temeljito 111	grad	temeljito	16
sestavine	temeljito 70	prostore	temeljito	15
zadevo	temeljito 62	strokovnjaki	temeljito	15
kožo	temeljito 46	let	temeljito	14
leto	temeljito 40	hrano	temeljito	14
času	temeljito 34	gobe	temeljito	14
stvari	temeljito 33	krompir	temeljito	14
dan	temeljito 31	telo	temeljito	14
časa	temeljito 28	Sloveniji	temeljito	13
dni	temeljito 28	sistem	temeljito	13
mesecih	temeljito 27	roke	temeljito	13
stavbo	temeljito 27	čas	temeljito	12
stvar	temeljito 26	prst	temeljito	12
hišo	temeljito 24	testo	temeljito	11
teden	temeljito 24	dela	temeljito	11
leti	temeljito 23	časom	temeljito	11
dneh	temeljito 23	zadeve	temeljito	11
res	temeljito 22	mešanico	temeljito	11
leta	temeljito 22	primer	temeljito	11
razmere	temeljito 21	podjetje	temeljito	11
delo	temeljito 20	mesto	temeljito	11
letu	temeljito 19	cerkev	temeljito	10
življenje	temeljito 18	obraz	temeljito	10
stoletju	temeljito 18	Slovenija	temeljito	10
vodo	temeljito 18	nakupom	temeljito	10
vlada	temeljito 17	zadevi	temeljito	10
lase	temeljito 17	otroka	temeljito	10
zmes	temeljito 16	minutah	temeljito	10
uporabo	temeljito 16	dom	temeljito	10

Tabela 92: Izluščeni podatki – Sam + temeljito.

1.7.7 Temeljito + Sam

Skladno z intuicijo, da zveze prislova pred samostalnikom v slovenščini niso tipične, spodnji podatki kažejo nabor besednih nizov, ki prinašajo kombinacijo samostalnikov z (neustrezno lematiziranim) pridevnikom *temeljito*:

besedni niz	pog. v korpusu	temeljito	obravnav	
temeljito prenova	95	temeljito	razprava	11
temeljito obnova	60	temeljito	pomladitev	11
temeljito analiza	58	temeljito	prevetritev	11
temeljito čiščenje	43	temeljito	sanacija	11
temeljito preiskava	41	temeljito	pregled	10
temeljito reforma	35	temeljito	kot	10
temeljito sprememba	28	temeljito	poročilo	10
temeljito poznavanje	24	temeljito	preureditev	9
temeljito priprava	18	temeljito	umivanje	9
temeljito raziskava	16	temeljito	revizija	9

temeljito	obdelava	9	temeljito	popravilo	4
temeljito	posodobitev	8	temeljito	preučitev	4
temeljito	preverjanje	7	temeljito	informacija	4
temeljito	znanje	7	temeljito	ureditev	4
temeljito	kontrola	7	temeljito	kopanje	3
temeljito	študija	6	temeljito	iskanje	3
temeljito	rekonstrukcija	6	temeljito	izobrazba	3
temeljito	predstavitev	6	temeljito	prestrukturiranje	3
temeljito	izdelava	5	temeljito	preoblikovanje	3
temeljito	reorganizacija	5	temeljito	zaščita	3
temeljito	akcija	5	temeljito	predelava	3
temeljito	ogledalo	5	temeljito	bilanca	3
temeljito	spremljanje	5	temeljito	kota	3
temeljito	načrtovanje	5	temeljito	rez	3
temeljito	pranje	5	temeljito	razčlenitev	3
temeljito	čistka	4	temeljito	prenovitev	3
temeljito	preobrazba	4	temeljito	delo	3
temeljito	izobraževanje	4	temeljito	škropljenje	3
temeljito	proučitev	4	temeljito	ocena	3
temeljito	poseg	4	temeljito	odstranjevanje	3
temeljito	obvladovanje	4	temeljito	tortura	3

Tabela 93: Izluščeni podatki – temeljito + Sam.

Ker gornji podatki že sami na sebi sugerirajo napačno označevanje pridevniških oblik kot prislovnih, npr. *temeljita* [prenova, obnova, analiza, preiskava, reforma], je natančnejša analiza, podprta s korpusnimi primeri, na tem mestu izpuščena.

1.7.8 Temeljito + Prid

Spodnja tabela prinaša za vključitev v zbirko relevantne primere pridevnika s prislovom *temeljito*. Kot je razvidno, nabor pridevnikov v veliki meri sestoji iz opisnih deležnikov na -n/-t, npr. *temeljito* [prenovljen, pripravljen, obdelan; umit, zalit, podprt], razen tega pa še npr. zveze z vrstnim pridevnikom na -ski, npr. *temeljito* [kadrovske, vsebinske, gospodarske, spomladanske] itd.

besedni niz	pogostnost v korpusu	temeljito	preverjen	14	
temeljito	prenovljen	162	temeljito	izobražen	14
temeljito	obnovljen	145	temeljito	navlažen	13
temeljito	pripravljen	123	temeljito	notranji	12
temeljito	obdelan	102	temeljito	mejen	11
temeljito	očiščen	54	temeljito	preoblikovan	11
temeljito	spremenjen	47	temeljito	zasnovan	11
temeljito	seznanjen	41	temeljito	kadrovski	10
temeljito	preizkušen	40	temeljito	odcejen	10
temeljito	opran	34	temeljito	predstavljen	10
temeljito	premišljen	34	temeljito	premešan	10
temeljito	predelan	32	temeljito	povezan	10
temeljito	pregledan	26	temeljito	zaščiten	9
temeljito	opravljen	25	temeljito	pretehtan	9
temeljito	raziskan	24	temeljito	poškropljen	9
temeljito	umit	23	temeljito	drugačen	9
temeljito	posodobljen	20	temeljito	razložen	8
temeljito	izdelan	19	temeljito	obveščen	8
temeljito	opremljen	16	temeljito	prevetren	8
temeljito	strokoven	14	temeljito	osvežen	8

temeljito	ohlajen	8	temeljito	dodelan	4
temeljito	vsebinski	7	temeljito	preskušen	4
temeljito	opisan	7	temeljito	prezidan	4
temeljito	preurejen	7	temeljito	pregret	4
temeljito	dokumentiran	7	temeljito	adaptiran	4
temeljito	počiščen	6	temeljito	proučen	3
temeljito	zalit	6	temeljito	razdejan	3
temeljito	izpeljan	6	temeljito	praven	3
temeljito	poučen	6	temeljito	prilagojen	3
temeljito	zastražen	6	temeljito	spomladanski	3
temeljito	urejen	6	temeljito	zaseden	3
temeljito	podkovan	6	temeljito	preiskan	3
temeljito	omočen	6	temeljito	pojasnjen	3
temeljito	prepleten	6	temeljito	kmetijski	3
temeljito	institucionalen	6	temeljito	poškodovan	3
temeljito	založen	5	temeljito	zastavljen	3
temeljito	tehnološki	5	temeljito	znanstven	3
temeljito	kriminalističen	5	temeljito	zloščen	3
temeljito	saniran	5	temeljito	razčlenjen	3
temeljito	načrtovan	5	temeljito	toploten	3
temeljito	nadzorovan	5	temeljito	reformiran	3
temeljito	gospodarski	5	temeljito	zaprt	3
temeljito	finančen	5	temeljito	odebeljen	3
temeljito	hišen	5	temeljito	informiran	3
temeljito	osušen	5	temeljito	pretlačen	3
temeljito	izšolan	5	temeljito	osvetljen	3
temeljito	podprt	5	temeljito	pripovedovan	3
temeljito	razgledan	5	temeljito	študijski	3
temeljito	izboljšan	5	temeljito	obrobljen	3
temeljito	prečiščen	5	temeljito	vzdrževan	3
temeljito	oblikovan	4	temeljito	porušen	3
temeljito	uničen	4	temeljito	računalniški	3
temeljito	javen	4	temeljito	teoretičen	3
temeljito	lepoten	4	temeljito	gradben	3
temeljito	pomit	4	temeljito	pokrit	3
temeljito	skrit	4	temeljito	načet	3
temeljito	domišljen	4	temeljito	nastavljiv	3
temeljito	zavarovan	4	temeljito	izveden	3
temeljito	preučen	4	temeljito	čistilen	3
temeljito	analiziran	4	temeljito	voden	3
temeljito	socialen	4	temeljito	razvit	3
temeljito	dopolnjen	4	temeljito	omajan	3
temeljito	splaknjen	4	temeljito	nov	3

Tabela 94: Izluščeni podatki – temeljito + Prid.

Iz izluščenih podatkov seveda ni razvidno, ali prislov v resnici določa pridevnik (primer iz korpusa FidaPLUS: *V Razkrižju so v soboto odprli temeljito obnovljen in dograjen kulturni dom*) ali gre za napačno lematizacijo pridevnika v prislov (FidaPLUS: *pričakuje temeljito kadrovske in programske prenove stranke*) ali prislova v pridevnik (FidaPLUS: *ker bi jo sicer morali temeljito kadrovske prevetriti*). Vsaj za deležniške oblike pa se – po hitrem pregledu korpusnih zadetkov – zdi, da so označene ustrezno.

1.8 Temeljito – tridelni vzorci

Za analizo je na seznamu najpogostejših vzorcev z lemo *temeljito* ter dvema oblikoskladenjskima oznakama za polnopomensko besedo na voljo 31 vzorcev, ki jih je možno združiti v 5 vzorčnih tipov:

vzorčni tip	število vzorcev
Prisl <i>temeljito</i> Glag	11
Glag <i>temeljito</i> Glag	4
<i>temeljito</i> Glag Sam	3
Sam <i>temeljito</i> Glag	8
Prid Sam <i>temeljito</i>	2
drugi vzorci	3
RNN RNN TEMELJITO	
TEMELJITO PPNZET SOZET	
SOZET RNN TEMELJITO	

Tabela 95: Nabor najpogostejših tridelnih vzorcev z lemo *temeljito* in oblikoskladenjskimi oznakami za polnopomenske besede.

V nadaljevanju poglavja so predstavljeni podatki glede na navedene vzorčne tipe.

1.8.1 Prisl + *temeljito* + Glag

Obravnavani vzorčni tip prinaša naslednje besedne nize (s pogostnostjo 5 ali več):

besedni niz		pogostnost v korpusu		potrebno	temeljito	pregledati	10
skupaj	temeljito	premešati	45	tako	temeljito	pripraviti	10
treba	temeljito	premisliti	40	letos	temeljito	obnoviti	10
treba	temeljito	obnoviti	34	najprej	temeljito	oprati	10
enkrat	temeljito	premisliti	30	lahko	temeljito	spoznati	10
lahko	temeljito	pripraviti	21	treba	temeljito	pretehtati	9
treba	temeljito	pripraviti	21	najprej	temeljito	seznaniti	9
treba	temeljito	razmisliti	21	lahko	temeljito	pregledati	9
najprej	temeljito	preučiti	19	skupaj	temeljito	premisliti	9
najprej	temeljito	pregledati	18	treba	temeljito	pogovoriti	8
tako	temeljito	spremeniti	17	najprej	temeljito	premisliti	8
enkrat	temeljito	pregledati	17	zelo	temeljito	pripraviti	8
treba	temeljito	prenoviti	15	najprej	temeljito	preveriti	8
lahko	temeljito	ogledati	15	lahko	temeljito	prekriti	8
najbolj	temeljito	umivati	15	najprej	temeljito	obnoviti	8
treba	temeljito	očistiti	15	potrebno	temeljito	spremeniti	8
treba	temeljito	preučiti	14	najprej	temeljito	ogledati	8
lahko	temeljito	spremeniti	13	tako	temeljito	očistiti	8
treba	temeljito	spremeniti	13	lahko	temeljito	posvetiti	8
treba	temeljito	pregledati	12	najprej	temeljito	razmisliti	7
najprej	temeljito	očistiti	12	potrebno	temeljito	prenoviti	7
lahko	temeljito	očistiti	11	potrebno	temeljito	očistiti	7
res	temeljito	pripraviti	11	prej	temeljito	pogovoriti	7
treba	temeljito	preveriti	10	treba	temeljito	umiti	7
najprej	temeljito	umiti	10	ponovno	temeljito	preučiti	7
bolj	temeljito	pripraviti	10	potrebno	temeljito	proučiti	7

treba	temeljito	raziskati	7	očitno	temeljito	spremeniti	5
treba	temeljito	prevetrili	7	najprej	temeljito	počistiti	5
enkrat	temeljito	razmisliti	7	tako	temeljito	spreminjati	5
bolj	temeljito	lotiti	7	predhodno	temeljito	posvetovati	5
najprej	temeljito	pripraviti	7	enkrat	temeljito	proučiti	5
zares	temeljito	pripraviti	6	potrebno	temeljito	pripraviti	5
potrebno	temeljito	razmisliti	6	najprej	temeljito	preizkusiti	5
treba	temeljito	popraviti	6	prej	temeljito	premisli	5
treba	temeljito	proučiti	6	medtem	temeljito	spremeniti	5
takoj	temeljito	pregledati	6	takoj	temeljito	umiti	5
najprej	temeljito	pogovoriti	6	spet	temeljito	očistiti	5
enkrat	temeljito	pretehtati	6	dvakrat	temeljito	obnoviti	5
lahko	temeljito	zamajati	6	vnaprej	temeljito	pripraviti	5
bolj	temeljito	pregledati	6	nato	temeljito	oprati	5
potrebno	temeljito	obnoviti	6	nato	temeljito	očistiti	5
tako	temeljito	pripravljati	6	treba	temeljito	analizirati	5
prej	temeljito	očistiti	6	zdaj	temeljito	pripravljati	5
tako	temeljito	obnoviti	6	treba	temeljito	počistiti	5
najprej	temeljito	prebrati	6	treba	temeljito	urediti	5
bolj	temeljito	posvetiti	5	zelo	temeljito	pripravljati	5
najprej	temeljito	prenoviti	5				

Tabela 96: Izluščeni podatki – Prisl + temeljito + Glag.

Tabela prinaša raznovrstne primere: prednjačijo kombinacije s prislovi, ki skladenjsko nastopajo s pomožnim glagolom *biti* kot del povedka (*treba, lahko, potrebno*). Od drugih primerov se pojavljajo besedni nizi s prislovi, ki izražajo stopnjo oz. mero (*bolj, zares*), ter drugimi, npr. časovnimi prislovi (*najprej, medtem, nato*). Na tem mestu se kaže za smotrnega premislek o izločitvi prve omenjene skupine nizov oz. njihovi obravnavi med zvezami, ki na skladenjski ravni izvornih stavkov izražajo modalnost z glagolsko obliko (take so zveze, predstavljene v sledečem poglavju) – v kolikor se slednje sploh izkažejo za relevantne za vključitev v leksikalno zbirko.

1.8.2 Glag + temeljito + Glag

Spodnja tabela prinaša nabor kombinacij dveh glagolov, pomožnega ter polnopomenskega, med njima pa obravnavanega prislova *temeljito* kot določilo slednjega:

besedni niz		pogostnost v korpusu					
morati	temeljito	premisлити	47	morati	temeljito	proučiti	7
morati	temeljito	pripraviti	35	začeti	temeljito	pripravljati	7
morati	temeljito	razmisлити	29	morati	temeljito	analizirati	7
morati	temeljito	spremeniti	28	morati	temeljito	potruditi	6
začeti	temeljito	obnavljati	25	morati	temeljito	preoblikovati	6
morati	temeljito	pogovoriti	23	morati	temeljito	lotiti	6
morati	temeljito	obnoviti	21	odločiti	temeljito	spremeniti	6
morati	temeljito	pregledati	15	morati	temeljito	poznati	6
morati	temeljito	očistiti	13	morati	temeljito	preveriti	5
začeti	temeljito	prenavljati	13	nameravati	temeljito	obnoviti	5
morati	temeljito	prenoviti	11	morati	temeljito	sprati	5
nameravati	temeljito	prenoviti	10	uspeti	temeljito	obnoviti	5
morati	temeljito	preučiti	10	morati	temeljito	preiskati	5
morati	temeljito	prežvečiti	9	odločiti	temeljito	prenoviti	5
začeti	temeljito	spreminjati	8	kazati	temeljito	razmisлити	4
morati	temeljito	pretehtati	7	veljati	temeljito	razmisлити	4
				morati	temeljito	seznaniti	4

morati	temeljito	oprati	4	morati	temeljito	prečesati	3
začeti	temeljito	ukvarjati	4	morati	temeljiteje	pripraviti	3
želei	temeljito	pripraviti	4	želei	temeljiteje	seznaniti	3
morati	temeljito	prevetriti	4	začeti	temeljiteje	nadzorovati	3
utegniti	temeljito	spremeniti	4	morati	temeljito	pretresti	3
morati	temeljito	raziskati	4	morati	temeljito	posvetiti	3
začeti	temeljito	preučevati	3	nameravati	temeljito	posodobiti	3
nameravati	temeljito	preoblikovati	3	nameravati	temeljito	pripraviti	3
morati	temeljito	zamisliti	3	veljati	temeljito	pregledati	3
želei	temeljito	obnoviti	3	hoteti	temeljito	spoznati	3
morati	temeljito	reformirati	3	začeti	temeljiteje	preučevati	3
morati	temeljito	posodobiti	3	morati	temeljito	stlačiti	3
morati	temeljito	vprašati	3	morati	temeljito	razgibati	3
morati	temeljito	preštudirati	3	morati	temeljiteje	premisli	3
odločiti	temeljito	pospraviti	3	hoteti	temeljito	pripraviti	3
pričeti	temeljiteje	pripravljeni	3	začeti	temeljiteje	delati	3
dati	temeljito	pregledati	3	morati	temeljito	izboljšati	3
začeti	temeljito	preurejati	3	odločiti	temeljito	pripraviti	3

Tabela 97: Izluščeni podatki – Glag + temeljito + Glag.

Nabor nizov se kaže za relevantnega za luščenje v smislu doprinosu informacij, izraženih na ravni pomožnega glagola – v primeru, da so na ravni gradnje leksikalne zbirke slednje uporabne. Možna je denimo združena obravnava z dvodelnimi zvezami (glej V-1.7.4) z dopolnjevanjem podatkovnih seznamov (npr. *temeljito spremeniti* – 963, *morati temeljito spremeniti* – 28; *temeljito premisliti* – 557, *morati temeljito premisliti* – 47).¹⁰⁹

1.8.3 Temeljito + Glag + Sam

Nabor spodnjih nizov je – ob upoštevanju problemov priklica – relevanten za luščenje. Navedeni so primeri s pogostnostjo 5 ali več:

besedni niz	pogostnost v korpusu						
temeljito	umiti	roke	40	temeljito	lotiti	prenove	7
temeljito	spremeniti	življenje	30	temeljito	oprati	roke	6
temeljito	spremeniti	podobo	28	temeljito	očistiti	kožo	5
temeljito	zagreniti	življenje	25	temeljito	umivati	roke	5
temeljito	spremeniti	način	16	temeljito	prenoviti	moštvo	5
temeljito	umivati	zobe	16	temeljito	umiti	obraz	5
temeljito	lotiti	dela	14	temeljito	obnoviti	cesto	5
temeljito	pregledati	poslovanje	11	temeljito	očistiti	okolico	5
temeljito	spremeniti	odnos	10	temeljito	spremeniti	razmerje	5
temeljito	spremeniti	razmerja	10	temeljito	pregledati	prostore	5
temeljito	premešati	karte	9	temeljito	lotiti	preiskave	5
temeljito	preučiti	razmere	9	temeljito	spremeniti	strukturo	5
temeljito	analizirati	stanje	8	temeljito	preiskati	okolico	5
temeljito	očistiti	obraz	8	temeljito	pretakniti	prostore	5
temeljito	očistiti	zobe	8	temeljito	pregledati	vsebino	5
temeljito	pripraviti	tla	7	temeljito	spremeniti	zgodovino	5
temeljito	umiti	zobe	7	temeljito	analizirati	vzroke	5
temeljito	prebrati	navodila	7	temeljito	pospraviti	dvorišča	5

¹⁰⁹ Tovrstna obravnava je seveda mogoča na vseh mestih, kjer tridelni vzorec oz. vzorčni tip zajema dvodelnega, ki že sam na sebi prinaša za vključitev v zbirko relevantne nize.

Tabela 98: Izluščeni podatki – temeljito + Glag + Sam.

Iz podatkov je razvidno, da obravnavani vzorčni tip po večini sestavljajo vzorci s samostalniki v tožilniku, ki nastopajo na stavčni ravni v vlogi predmeta, npr. *temeljito [umiti roke, umiti zobe, očistiti obraz]*.

1.8.4 Sam + temeljito + Glag

Z upoštevanjem vseh že izpostavljenih zadržkov so potencialno zanimivi za uvrstitev v zbirko tudi spodnji nizi (pogostnost 5 ali več):

besedni niz	pogostnost v korpusu			leto	temeljito	prenoviti	6
letih	temeljito	spremeniti	32	stavbo	temeljito	predelati	6
sestavine	temeljito	premešati	31	zadevo	temeljito	premisлити	6
kožo	temeljito	očistiti	19	lase	temeljito	sprati	6
zadevo	temeljito	raziskati	18	cedilu	temeljito	odcediti	6
sestavine	temeljito	obdelati	16	leto	temeljito	obnoviti	6
letih	temeljito	prenoviti	16	testo	temeljito	pregnesti	6
razmere	temeljito	spremeniti	15	Sloveniji	temeljito	spremeniti	6
letih	temeljito	obnoviti	14	časom	temeljito	prenoviti	5
življenje	temeljito	spremeniti	13	desetletjih	temeljito	spremeniti	5
sestavine	temeljito	zmešati	11	zadevo	temeljito	proučiti	5
krompir	temeljito	oprati	11	policisti	temeljito	pregledati	5
stvari	temeljito	spremeniti	11	otroka	temeljito	pregledati	5
hrano	temeljito	prežvečiti	10	obraz	temeljito	očistiti	5
stvar	temeljito	premisлити	9	letu	temeljito	prenoviti	5
stavbo	temeljito	prenoviti	9	hišo	temeljito	prenoviti	5
gobe	temeljito	odrgniti	9	leto	temeljito	pregledati	5
delo	temeljito	opraviti	9	mestu	temeljito	razgreti	5
mesecih	temeljito	obnoviti	9	leta	temeljito	obnoviti	5
zadevo	temeljito	preučiti	9	policija	temeljito	pregledati	5
uporabo	temeljito	oprati	8	nakupom	temeljito	premisлити	5
zmes	temeljito	premešati	8	teden	temeljito	očistiti	5
limono	temeljito	oprati	8	leto	temeljito	očistiti	5
grad	temeljito	prezidati	7	desetletju	temeljito	spremeniti	5
času	temeljito	spremeniti	7	letu	temeljito	spremeniti	5
cekina	temeljito	maščevati	7	leti	temeljito	obnoviti	5
kožo	temeljito	sprati	7	strokovnjaki	temeljito	pregledati	5
zadevo	temeljito	preiskati	7	stavbo	temeljito	prezidati	5
minutah	temeljito	sprati	7	vodo	temeljito	skrtačiti	5
hišo	temeljito	preiskati	7	nasade	temeljito	pregledati	5
zadevi	temeljito	pogovoriti	6	odločitvijo	temeljito	premisлити	5

Tabela 99: Izluščeni podatki – Sam + temeljito + Glag.

Tudi v gornji tabeli je najti kombinacije s stavčnočlenskim (I) predmetom (*hrano temeljito prežvečiti, kožo temeljito očistiti*), (II) osebkom (*policisti temeljito pregledati, strokovnjaki temeljito pregledati*), nekaj pa je tudi kombinacij z (nepopolnimi) samostalniškimi zvezami, na prvi pogled večinoma v stavčni vlogi (III) prislovnega določila (*letih temeljito spreminiti, časom temeljito prenoviti*). Kot je že bilo izpostavljeno v V-1.6.1, so problematične za avtomatsko obravnavo predvsem dvoumne oblike, npr. *sestavine temeljito zmešati, grad temeljito prezidati, testo temeljito pregnesti* itd.

1.8.5 Prid + Sam + temeljito

Na tem mestu predstavljeni besedni nizi niso relevantni za nadaljnjo obravnavo:

besedni niz		pog. v korp.					
zadnjih	letih	temeljito	28	dobljene	informacije	temeljito	3
zadnjem	času	temeljito	11	preteklih	letih	temeljito	3
tekočo	vodo	temeljito	10	javnih	del	temeljito	3
zadnjem	letu	temeljito	9	prihodnjem	letu	temeljito	3
prihodnjih	dneh	temeljito	8	celotno	zadevo	temeljito	3
izgubljenega	cekina	temeljito	7	naslednji	dan	temeljito	3
naslednje	leto	temeljito	6	koalicijski	partnerji	temeljito	3
prihodnje	leto	temeljito	6	osnovno	šolo	temeljito	3
najkrajšem	času	temeljito	6	zadnjem	obdobju	temeljito	3
kratkem	času	temeljito	5	zahodni	polobli	temeljito	3
prihodnjih	letih	temeljito	5	zimsko	sezono	temeljito	3
lanskega	leta	temeljito	5	poznejših	letih	temeljito	3
novem	mestu	temeljito	5	dokončno	odločitvijo	temeljito	3
lansko	leto	temeljito	5	minulih	mesecih	temeljito	3
minulih	letih	temeljito	4	zakonskih	določil	temeljito	3
zadnjih	desetletjih	temeljito	4	preteklih	mesecih	temeljito	3
naslednjih	letih	temeljito	4	pravnega	vidika	temeljito	3
zgodovinsko	znamenitost	temeljito	4	debelo	črevo	temeljiteje	3
divji	novinci	temeljito	4	zadnjem	desetletju	temeljito	3
prejšnji	teden	temeljito	4	novi	lastnik	temeljito	3
zemeljskih	delih	temeljito	4	končanem	delu	temeljito	3
govejega	mesa	temeljito	3	prosti	čas	temeljito	3

Tabela 100: Izluščeni podatki – Prid + Sam + temeljito.

1.8.6 Preostali vzorci

Za vsakega od preostalih vzorcev so v nadaljevanju predstavljeni najpogostnejši primeri zapolnitev.

RNN RNN TEMELJITO	pogostnost v korpusu
prav tako temeljito	19
čim bolj temeljito	18
treba najprej temeljito	11
TEMELJITO PPNZET SOZET	
temeljito mejno kontrolo	11
temeljito notranjo preno	5
temeljito javno razpravo	4
SOZET RNN TEMELJITO	
zadevo najprej temeljito	3
zmes zelo temeljito	3
zadevo prej temeljito	2

Tabela 101: Nabor preostalih tridelnih vzorcev z lemo *temeljito* in dvema polnopolnomenkima besedama.

Prvi od navedenih vzorcev prinaša pogojno zanimive podatke, v primeru nadgradnje metode na način, ki omogoča ločevanje primerov, kjer zveza na levi določa zadnjega v nizu prislovov (*prav tako temeljito*) od drugih zvez (*treba najprej temeljito*). Zveze tipa *prav tako*, *čim bolj* itd. bi bilo denimo smiselno luščiti posebej in tekom nadaljnje obravnave obravnavati kot povezane enote. Pogojno zanimive podatke prinaša tudi drugi vzorec, čeprav so navedeni primeri očitno napačno označeni. Zadnji od vzorcev ne prinaša relevantnih besednih zvez.

2 Vzorci s predlogom

Pričujoče poglavje strukturno (in mestoma vsebinsko) sledi prejšnjemu, s to razliko, da je – kot priča naslov – namesto vzorcem s samimi polnopomenskimi besedami posvečeno vzorcem z oblikoskladenjskimi oznakami za predlog. Med najpogostejšimi vzorci za vsako od obravnavanih lem (glej IV-2.3) predstavljajo vzorci s predlogi naslednje deleže (prva številka v tabeli predstavlja število vzorcev s predlogom, druga vse ročno pregledane vzorce, sledi delež, zaokrožen na eno decimalko):

	dvodelni vzorci	tridelni vzorci
<i>pajek</i>	9 (101) 8,9 %	23 (101) 22,8 %
<i>strasten</i>	7 (101) 6,9 %	21 (103) 20,4 %
<i>plesati</i>	5 (100) 5,0 %	31 (103) 30,1 %
<i>temeljito</i>	2 (100) 2,0 %	13 (101) 12,9 %

Tabela 102: Delež vzorcev s predlogom med vsemi najpogostejšimi.

2.1 Dvodelni vzorci

Na prvi pogled kombinacije obravnavanih štirih lem s predlogom na levi ali desni strani niso posebej zanimive za vnos v leksikalno zbirko. V sledečih poglavjih navedeni vzorčni tipi izkazujejo nerelevantnost predvsem za kombinacije predloga s pridevnikom *strasten* (V-2.3.1, V-2.3.2) ter na desni od samostalnika *pajek* (V-2.2.2) oz. prislova *temeljito* (V-2.5.2). Pogojno zanimive so zveze samostalnika s predlogom na levi (V-2.2.1), vendar le toliko, kolikor so zanimivi podatki o pojavljanju obravnavanega samostalnika v različnih sklonskih oblikah (za gradnjo leksikalne zbirke te informacije niso posebej zanimive, ker so pridobljive iz oblikoskladenjske oznake samostalnika).

Med vsemi kombinacijami se za najbolj relevantno kaže kombinacija glagola *plesati* s predlogom na desni (V-2.4.2), ker izkazuje določen nabor informacij o glagolski vezljivosti.¹¹⁰

Vse do sedaj naštetе kombinacije se kažejo za potencialno relevantne v sklopu širših besednih zvez, kar se potrjuje v nadaljevanju poglavja pri obravnavi tridelnih vzorcev.

Precej v tem poglavju predstavljenih vzorcev pa je zanimivih za analizo avtomatskega označevanja. Pozornost je posvečena predvsem primerom, ki na ravni oznak izkazujejo na prvi pogled nepričakovane kombinacije, tj. vzorčni tip s predlogom pred glagolom (V-2.4.1.1) oz. prislovom (V-2.5.1.1) ter pojav primerov neujemanja sklona predloga s sklonom sledečega samostalnika (V-2.2.1.1).

Kljub temu, da se v navedenih podatkih ne izkazujejo, je možno predvideti primere, ki na ravni vključevanja podatkov v leksikalno zbirko ne bi smeli biti spregledani, denimo besedne zveze tipa *po krivici*. Vsaj deloma bi bilo možno takšne primere evidentirati v sklopu obravnave **kolokacijske okolice posameznih predlogov**. Slednji kot predmet analize sicer niso zajeti v pričujočo raziskavo, zato poglavje V-2.6 prinaša zgolj shematičen primer tipa podatkov, ki so pridobljivi s trenutno razpoložljivimi viri.

¹¹⁰ Sicer o predlogih v zvezah glagola s predložnimi zvezami že slovnica piše: »Ti predlogi so prosti morfemi glagolov: spadajo torej h glagolu, ne k samostalniški zvezi, ki jo vežejo.« (Toporišič 2004⁴: 582) O problematiki natančneje denimo Gantar 2007: 86–88.

2.2 Pajek – dvodelni vzorci

2.2.1 Pred + pajek

Spodnji nizi sami na sebi niso relevantni za vključitev v zbirko, njihova obravnava je bolj smiselna znotraj širših zvez (glej npr. V-2.8.2 in V-2.8.3).

besedni niz	pogostnost v korpusu	po	pajkih	6
s pajkom	505	proti	pajkom	6
s pajki	85	med	pajki	6
pred pajki	60	poleg	pajkov	5
o pajkih	40	od	pajka	5
za pajke	27	brez	pajkov	5
za pajka	25	do	pajkov	5
o pajku	18	proti	pajku	4
na pajke	17	v	pajku	3
za pajek	17	na	pajku	3
na pajka	17	na	pajkih	3
pri pajkih	13	za	pajkom	3
zaradi pajka	10	namesto	pajka	3
v pajka	9	pri	pajku	3
pred pajkom	7	nad	pajkom	3
od pajkov	7			

Tabela 103: Izluščeni podatki – Pred + pajek.

2.2.1.1 Analiza označenosti

Spodnja tabela prinaša vse vzorce obravnavanega vzorčnega tipa z namenom pregleda ujemanja sklona predloga ter sledečega samostalnika. Neujemanja so označena s sivo barvo:

vzorec	pogostnost v korpusu	Dt	Sometn	13
Do Someo	520	Do Sommi	6	
Do Sommo	147	Dt Somei	5	
Dm Sommm	63	Dd Someo	4	
Dt Sometd	51	Dd Somed	4	
Dt Sommt	48	Dd Sommd	3	
Dm Somem	27	Dr Sometd	2	
Dr Sommr	25	Do Somdo	1	
Dr Somer	21	Dt Sommr	1	

Tabela 104: Analiza označenosti Pred + pajek.

V primerih, ki na ravni oblikoskladenjskih oznak izkazujejo neujemanje predloga ter sledečega samostalnika v sklonu, je utemeljeno vnaprejšnje sklepanje, da gre za napake v označevanju oz. avtomatskem razdvoumljanju besednih oblik. Možna rešitev problema je že večkrat predlagano označevanje z upoštevanjem besednozveznosti, pri čemer se pri razdvoumljanju oznak predlogu in samostalniku, ki sta glede na skladenjski kontekst z visoko verjetnostjo del predložne zveze, izbereta istosklonski oznaki. V nadaljevanju sledi nekaj primerov označevalnih napak iz korpusa FidaPLUS.

Tone Kuntner pripoveduje o svojih najbližjih. Zaradi sinovega navdušenja **nad pajki** so vsi Kuntnerjevi postali Taranetele so

pravi velikani **med pajki**. Nekatere vrste skupaj z nogami merijo do 30 cm

stoječega prometa, nad katerega se je mestna oblast spravila s pajki in lisicami, se pojavi v primeru, če se okolica hiše, prenočevala pa sva v frčadi<< med pajki in žužki. V tem času sva hišo, ob pajki. Čeprav poglji žrtvujejo za hrano kraljici, se med pajki in poglji splete tesno družabništvo, saj pajki

Konkordančni niz FidaPLUS 13: #2Dpeo_#1pajekS???i*.

TELICO simentalco, brejo 4 mesece, prodam ali zamenjam za pajek ali trosilec hlevskega gnoja.

POD Pajek v mreži 23

MOTOKULTIVATOR gorenje in kosilnico prodam ali zamenjam za pajek . Tel.: 041/354

CITROEN ZX, prodam ali zamenjam za pajek ali diatonično harmoniko. Tel.: 07/81

Konkordančni niz FidaPLUS 14: #2Dpet_#1pajekS???i*.

Prilagodite taktiko. Ogenj in kopja so dobra orožja tako proti pajkom kot proti orkom, ki sčasoma uletijo; gluh včeraj razmišlala, bi ali ne bi gorice poškropila proti pajkom in drugim držovnim škodljivcom. Pa sma

Konkordančni niz FidaPLUS 15: #2Dped_#1pajekS???o*.

Iz korpusnih primerov gre sklepati, da je vzrok pripisovanja neustrezne oblikoskladenjske oznake vsebnost nelematizirane oz. neznane besede v stavku (npr. *pogljij, uletijo, držovnim škodljivcom*). Vsebnost neznane besede vpliva na uspešnost avtomatske stavčne analize, posledično pa so pripisane napačne oznake.

2.2.2 Pajek + Pred

Relevantnejši za vnos v zbirko od spodaj navedenih so besedni nizi, ki vsebujejo še samostalnik na desni – glej poglavje V-2.8.1.

besedni niz	pogostnost v korpusu	pajek	do	10
pajek na	469	pajek	pri	7
pajek v	163	pajek	od	7
pajek za	148	pajek	brez	4
pajek z	62	pajek	med	4
pajek iz	40	pajek	pred	4
pajek po	16	pajek	proti	3

Tabela 105: Izluščeni podatki – pajek + Pred.

2.3 Strasten – dvodelni vzorci

2.3.1 Pred + strasten

Relevantnejši od spodaj navedenih so besedni nizi, ki vsebujejo še samostalnik na desni – glej poglavje V-2.9.1.

besedni niz	pogostnost v korpusu	po	strastni	34
v strastno	109	o	strastni	31
v strastnem	75	s	strastno	31
s strastnim	72	za	strastno	28
za strastne	49	s	strastnimi	24

v	strastni	22	po	strastnih	8
za	strastnega	20	o	strastnih	8
med	strastnim	19	do	strastnega	6
po	strastnem	16	v	strastne	6
v	strasten	16	od	strastne	6
med	strastnimi	15	ob	strastnem	6
pri	strastnih	15	za	strastna	5
na	strastne	13	k	strastnemu	5
v	strastnih	13	pri	strastnem	5
o	strastnem	12	do	strastne	5
za	strasten	11	zaradi	strastnih	4
od	strastnih	11	na	strastnem	3
v	strastnega	11	poleg	strastnih	3
poleg	strastnega	11	na	strastno	3
zaradi	strastne	10	na	strasten	3
med	strastne	10	kljub	strastni	3
iz	strastnega	10	brez	strastnega	3
iz	strastne	10	med	strastno	3
zaradi	strastnega	9	k	strastni	3
od	strastnega	9			

Tabela 106: Izluščeni podatki – Pred + *strasten*.

2.3.2 Strasten + Pred

Navedeni primeri sami na sebi niso zanimivi za vnos v leksikalno zbirko, zaradi nepogostnosti pa v raziskavo tudi niso bili uvrščeni primeri širših zvez (denimo s samostalnikom na desni).

besedni niz	pogostnost v korpusu	strasten	do	4
strasten v	14	strasten	na	3
strasten od	5	strasten	pri	3
strasten glede	5			

Tabela 107: Izluščeni podatki – *strasten* + Pred.

2.4 Plesati – dvodelni vzorci

2.4.1 Pred + plesati

Zveze predloga pred glagolom so, kot rečeno, za slovenščino nepričakovane, zato so na tem mestu navedeni vsi primeri, sledi pa analiza označenosti obravnavanih zvez.

besedni niz	pogostnost v korpusu	krog	plesala	1
pred plesat	6	od	pleše	1
iz Plešiva	4	na	plesala	1
navkljub pleši	1	po	plesati	1
Prek pleše	1			

Tabela 108: Izluščeni podatki – Pred + *plesati*.

2.4.1.1 Analiza označenosti

V nadaljevanju sledi nekaj primerov, označenih kot kombinacija predloga pred glagolom v korpusu FidaPLUS: izdelana sta konkordančna niza za prva dva najpogostejša niza Tabele 108. Iz konkordanc je razvidno, da gre v

prvem primeru v resnici za kombinacijo glagola s prislovom. Zadetki so ponovljeni oz. nerazpršeni, tj. iz istega vira. V drugem primeru je napačno označen samostalnik *Plešivo*, sicer geografsko lastno ime:

Mačica pravi: Reci, da ne greš **pred plesat**, kot da ti da zlate čevlje.<< In
in mačica pravi: Reci, da ne pojdeš **pred plesat**, kot da ti da židano oblačilo.<< Deklica
ji je rekla: Reci, da ne pojdeš **pred plesat**, kot da ti da zlato kočijo in konje.
Mačica pravi: Reci, da ne greš **pred plesat** kot da ti da zlate čevlje.<< In tako
odgovoriti in mačica pravi: Reci, da ne pojdeš **pred plesat**, kot da ti da židano oblačilo. Deklica reče
mačica ji je reklak: REci, da ne pojdeš **pred plesat**, kot da ti da zlato kočijo in konje.

Konkordančni niz FidaPLUS 16: *pred_plesat*.

Beli pri Predvoru (na sliki z vinogradnikom Robijem Torošem **iz Plešiva** v Goriških brdih) je odlična igralka
Reya iz Kozane (sivi pinot) in Andrej Kristančič **iz Plešiva** (sauvignon); dva prihajata s Krasa: iz
Pa zato, ker je Joži Toroš, vinogradnica **iz Plešiva** v Brdih pripovedovala o tem, da je poživljajoča rastlina
Zmagovalci so v skupini suhih sivih pinotov Borut Blažič **iz Plešiva** pri Dobrovem, polsuhih Franci Cvetko iz

Konkordančni niz FidaPLUS 17: *iz_Plešiva*.

Ker je pričakovana verjetnost dejanskega pojavljanja predlogov neposredno pred glagoli nizka (kar potrjujejo tudi navedene konkordance), je analiza primerov, pri katerih oblikoskladenjske oznake prinašajo tovrstno kombinacijo, dobro izhodišče za izboljšavo avtomatskega označevanja.

2.4.2 Plesati + Pred

Glagole s sledečimi predlogi je mogoče obravnavati kot posebno vrsto kolokacijskih enot in jih temu ustrezno organizirati v leksikalno zbirko. S tega stališča so spodnji podatki za luščenje relevantni.

besedni niz		pogostnost v korpusu			
plesati	v	931	plesati	od	44
plesati	z	781	plesati	med	42
plesati	na	649	plesati	brez	22
plesati	po	449	plesati	nad	11
plesati	ob	155	plesati	iz	7
plesati	pred	114	plesati	sredi	6
plesati	do	106	plesati	k	6
plesati	pri	100	plesati	skozi	5
plesati	za	90	plesati	proti	5
plesati	okoli	82	plesati	o	4
plesati	okrog	68	plesati	čez	3
plesati	pod	52	plesati	mimo	3
			plesati	zaradi	3

Tabela 109: Izluščeni podatki – *plesati + Pred*.

Abstraktno sliko kombinacij glagola s predlogom je na naslednjem nivoju reprezentacije podatkov možno dopolniti s konkretnimi zapolnitvami mest na desni (glej poglavje V-2.10.1) oz. ustreznimi (skladenjskimi ali pomenskimi) kategorijami, ki konkretne zapolnitve predstavljajo.¹¹¹

2.5 Temeljito – dvodelni vzorci

2.5.1 Pred + *temeljito*

Besedni nizi, kakršne izpričuje na tem mestu predstavljeni vzorčni tip, niso zanimivi za vključitev v leksikalno zbirko.

besedni niz		pogostnost v korpusu			
za	temeljito	137	na	temeljitejše	7
z	temeljito	73	skozi	temeljito	5
v	temeljito	54	znotraj	temeljito	4
na	temeljito	31	do	temeljitejše	3
za	temeljitejše	19	pod	temeljito	3
po	temeljito	13	pred	temeljito	3
med	temeljito	8			

Tabela 110: Izluščeni podatki – Pred + *temeljito*.

2.5.1.1 Analiza označenosti

Analiza označenosti na tem mestu izvira iz predhodne identifikacije problemov pri avtomatskem ločevanju pridevniških oblik od prislovnih, kar vodi v predvidevanje, da so visoke pogostnosti nekaterih primerov v Tabeli 110 posledica napačnega označevanja. Analiza konkordančnega niza za iskalni pogoj kombinacije predloga *za* s kot prislov označeno besedo *temeljito* sume potrjuje: prevladujejo zadetki z napačno označenim pridevnikom (spodaj je navedenih naključno izbranih 10 zadetkov konkordančnega niza, vseh zadetkov je v korpusu FidaPLUS 132):

Po poti smo pobirali tudi smeti, vendar bi bilo **za temeljito** čiščenje potrebnega veliko več časa," je poudaril takšno dolgoročno najemno pogodbo, po kateri bi najemnik poskrbel **za temeljito** prenovo doma in nakup Sloveniji znani in bi v njih našli kar dovolj razlogov **za temeljito** drugačen pristop, pa se državni zbor ne da: pa lahko pripišemo soncu. Kaj lahko storite? Poskrbite **za temeljito** zaščito pred soncem z novo kolekcijo

Razmere so že tako slabe, da se je **za temeljito** akcijo, ki naj bi pospešila svetovno porabo bakra, drgnjenje zob z zobno ščetko in uporaba ustnih vod dovolj **za temeljito** ustno nego in da so občasne Radenci, Gospodarska zbornica Slovenije in Socius - vendarle odločili **za temeljito** poživitev in prenovo Vodja protestantske stranke ulstrskih unionistov David Trimble se je zavzel **za temeljito** prenovo stranke. Z generacije razreda C so se pri Mercedes-Benzu odločili **za temeljito**, a očem precej skrito prenovo. Novi unije, finski general Gustav Haegg Lund, se je zavzel **za temeljito** prenovo evropsko-ameriške varnostne

Konkordančni niz FidaPLUS 18: *za_temeljito*R*.

¹¹¹ Primer možne reprezentacije podatkov tega tipa za slovenščino prinaša Vezljivostni slovar (Žele 2008), vendar obravnavana lema *plesati* v slovar ni zajeta, zato natančnejša primerjava na tem mestu ni mogoča.

Od desetih navedenih primerov je pravilno označen en sam (*temeljito drugačen pristop*). Avtomatsko pripisovanje besedne vrste v zvezah tipa *za temeljito ustno nego* je problematično, ker brez upoštevanja pomena ni razvidno, ali oblika *temeljito* v besedni zvezi določa pridevnik ali samostalniško besedno zvezo. Možno je reševanje z upoštevanjem oz. primerjavo kolokacijskih podatkov prislova *temeljito* s pridevnikom *temeljiti* – pri čemer je izhodiščna (ne pa še na podatkih preverjena) predpostavka, da bo statistika izkazala potrebne razlike v kolokabilnosti zveze *temeljita nega* ter *temeljito usten*.¹¹²

2.5.2 Temeljito + Pred

Spodaj navedeni nizi sami na sebi niso zanimivi za vključitev v zbirko, zaradi nepogostnosti pa v raziskavo tudi niso bili uvrščeni primeri širših zvez (denimo s samostalnikom na desni).

besedni niz		pogostnost v korpusu			
temeljito	v	17	temeljito	pri	5
temeljito	z	8	temeljiteje	od	5
temeljito	o	8	temeljito	od	3
temeljito	na	8			

Tabela 111: Izluščeni podatki – temeljito + Pred.

2.6 Kolokacijska analiza predloga

V pričujočem delu je funkcijskim besednim vrstam posvečeno malo pozornosti, tako s stališča analize kot posledično priprave ustreznih postopkov za luščenje in organizacijo leksikalnih podatkov. Kot pa je bilo izpostavljeno v uvodu v pričujoče poglavje, se zdi nekatere tipe zvez smiselno skušati zajeti prav v sklopu obravnave funkcijskih besednih vrst. Nadaljevanje poglavja se teme zgolj dotika, večina vprašanj pa ostaja odprtih za nadaljnje raziskave.

Za potrebe izdelave primera seznama kolokatorjev za izbrani predlog je deloma lahko v pomoč konkordančnik korpusa FidaPLUS.¹¹³ Seznam 25 samostalnikov, ki se na mestu za predlogom *po* pojavljajo v mestniku, je glede na (na 500.000 zadetkov omejeni) konkordančni niz naslednji – prvi (sivi) seznam prinaša podatke glede na pogostnost, drugi glede na statistično vrednost LL:¹¹⁴

¹¹² V primerih, kot je opisani, je problem nadgrajevanja avtomatskega označevanja iz že avtomatsko označenih besedil posebej očiten. V pričujoči raziskavi privzemamo, da je označevanje že na dovolj visokem nivoju, da je iz pravilno označenega dela podatkov mogoče dobiti dovolj zanesljive informacije za izboljšavo šibkih mest označevalnika.

¹¹³ Funkcijske besedne vrste so v jeziku seveda izredno pogoste, zato omejitev dolžine konkordančnega niza pri obravnavi predlogov vodi k manjši natančnosti podatkov (seznam kolokacij prinaša vzorec vzorca jezika). Konkordančnik ASP32 v resnici ni bil razvit za obdelavo podatkov, kakršno srečujemo v pričujočem delu, zato ga na tem mestu uporabljamo zgolj za ponazoritev potenciala obravnavanega tipa analize.

¹¹⁴ »Rezultat te statistike prinaša informacijo o razmerju med dejanskim ter pričakovanim stanjem sopojavljanja dveh besed, pri čemer je pričakovano stanje, da sta besedi med seboj popolnoma neodvisni, tj. da se sopojavljata po naključju. Kadar se dejansko ter pričakovano stanje ujemata, je rezultat statistike nič. Višji ko je rezultat, manjša je verjetnost, da se besedi sopojavljata naključno.« (Arhar in Gorjanc 2007: 105)

	SEZNAM POGOSTNOSTI	POGOSTNOST	SEZNAM LL	VREDNOST LL
po	besedah	6810	besedah	60.790.450.481
	svetu	3729	mnenju	27.309.524.062
	mnenju	3718	podatkih	25.613.921.044
	Sloveniji	3274	svetu	21.264.104.484
	podatkih	3018	dogovoru	20.521.322.158
	letu	2909	zakonu	19.956.170.078
	zakonu	2548	zaslugi	19.068.466.591
	vrsti	2410	potrebi	18.355.474.898
	dogovoru	2150	vrsti	17.760.449.869
	koncu	2115	Sloveniji	16.453.794.125
	zaslugi	1962	letu	15.942.454.770
	potrebi	1872	vojni	12.623.079.744
	vojni	1816	telefonu	11.276.519.990
	tekmi	1534	objavi	11.073.425.243
	Evropi	1455	koncu	10.897.575.020
	telefonu	1397	navodilih	10.539.993.037
	številu	1325	številu	10.368.204.062
	objavi	1234	vrnitvi	9.848.656.250
	pošti	1163	pošti	9.738.020.411
	navodilih	1145	naročilu	9.362.027.263
	glavi	1137	tekmi	9.131.972.682
	vrnitvi	1107	želji	8.447.215.909
	cesti	1059	Evropi	8.225.632.888
	zmagi	1037	glavi	8.138.473.601
	želji	995	naključju	8.005.714.730

Tabela 112: Kolokatorji (Sam₅) predloga *po*.

Visoke številke v stolpcu *pogostnost* pričajo o tem, da je visokopogostnih zadetkov oz. kandidatov za kolokatorje veliko število, iz obeh seznamov besed pa je razvidno, da so s semantičnega vidika zveze različnih tipov (prim. denimo *po Sloveniji* in *po mnenju*). Avtomatsko ločevanje enega tipa zvez od drugih se zdi z opisano metodo nedosegljivo, vsekakor pa je mogoče podatke, kakršni so, vključiti v leksikalno zbirko in pri obdelavi naravnega jezika uporabljati na besednozvezni ravni.

Kot rečeno so v zvezi s funkcijskimi besednimi vrstami potrebne nadaljnje analize podatkov, ki bodo omogočile čim širše zajetje relevantnih besednih zvez (prim. denimo zveze tipa *v obliki pajka* v V-2.8.4).

2.7 Tridelni vzorci

Poglavje prinaša nabor za luščenje uporabnih vzorčnih tipov, ki prinašajo npr. predložne samostalniške zveze (V-2.8.1, V-2.8.2) ter glagolske predložne zveze (V-2.8.3, V-2.10.1, V-2.10.4). Pogojno zanimivi so primeri, kjer se glagol pojavlja s predlogom na desni ter določujočim prislovom na levi (V-2.11.2).

Nezanimivi za vključitev v zbirko so nizi, kjer se na zadnjem mestu pojavlja predlog (V-2.9.2, V-2.8.4, V-2.10.5), pridevnik (V-2.9.3, V-2.10.2) ali prislov (V-2.11.1). Manj zanimivi so tudi nizi, kjer predlog nastopa pred samostalniško besedno zvezo (V-2.9.1).

2.8 Pajek – tridelni vzorci

Tridelnih vzorcev z lemo *pajek* ter oblikoskladenjsko oznako za predlog je med vsemi najpogostejšimi 23 (22,8 %). Vzorce je mogoče razvrstiti v 3 vzorčne tipe, ki – kot bo vidno v nadaljevanju – vsi prinašajo za uvrstitev v zbirko relevantne besedne nize:

vzorčni tip	število vzorcev
<i>pajek</i> + Pred + Sam	5
Sam + Pred + <i>pajek</i>	8
Glag + Pred + <i>pajek</i>	4
drugi vzorci	6
DM SOZEM PAJEK	
PAJEK DM PPNZEM	
DO PAJEK RNN	
PAJEK SOZMR DT	
DO PAJEK DT	
PPNMEID PAJEK DT	

Tabela 113: Nabor najpogostejših tridelnih vzorcev z lemo *pajek* in oblikoskladenjsko oznako za predlog.

2.8.1 *Pajek* + Pred + Sam

Spodaj navedeni besedni nizi pričajo o relevantnosti obravnavanega vzorčnega tipa za luščenje podatkov:

besedni niz	pogostnost v korpusu	pajek	v	predsobi	4
pajek za seno	105	pajek	iz	Maribora	3
pajek v mreži	14	pajek	na	koncu	3
pajek na svetu	13	pajek	v	ujetništvu	3
pajek za obračanje	9	pajek	na	vrtnu	3
pajek za odvoz	7	pajek	na	delu	3
pajek v spalnici	6	pajek	na	stropu	3
pajek iz Bistrice	5	pajek	z	oljem	3
pajek na leto	4				

Tabela 114: Izluščeni podatki – *pajek* + Pred + Sam.

2.8.2 Sam + Pred + *pajek*

Tudi nabor na tem mestu navedenih nizov je relevanten za nadaljnjo obravnavo (gre za enak tip kot v prejšnjem poglavju, z razliko v mestu leme *pajek* v vzorčnem tipu):

besedni niz	pogostnost v korpusu	odvoz	s	pajki	9
odvoz s pajkom	206	zgodba	o	pajkih	8
vozilo s pajkom	52	avtomobil	s	pajki	5
strah pred pajki	31	srečanje	s	pajkom	4
Strah pred pajki	22	zveza	s	pajkom	3
avtomobil s pajkom	22	strokovnjak	za	pajke	3
film o pajkih	17	strah	pred	pajkom	3
vozilo s pajki	10	razprava	o	pajku	3

Tabela 115: Izluščeni podatki – Sam + Pred + *pajek*.

Zanimivi sta zvezi na tretjem in četrtem mestu tabele (obarvani sivo), pri čemer druga označuje lastno ime, tj. naslov filma. Razlika je evidentirana, ker je bila v 31 primerih samostalniki *strah* pripisana lema z malo, v 22 primerih pa z veliko začetnico. V primeru, da v leksikalni zbirki, ki je izhodišče za označevanje, lastnoimenska različica ne bi obstajala, bi bila razlika izgubljena (glej primer *plesati z volkovi* v V-2.10.1).

2.8.3 Glag + Pred + *pajek*

Tudi v nadaljevanju predstavljeni rezultati (kljub vsebnosti glagola) izkazujejo visoko stopnjo relevantnosti za vključitev v leksikalno zbirko:

besedni niz	pogostnost v korpusu	odstraniti	s	pajkom	4
odpeljati s pajkom	72	zamenjati	za	pajek	4
voziti s pajkom	24	iti	za	pajka	3
menjati za pajek	10	spominjati	na	pajka	3
spremeniti v pajka	7	naložiti	na	pajka	3
odvažati s pajkom	6	ukvarjati	s	pajki	3
odvažati s pajki	4				

Tabela 116: Izluščeni podatki – Glag + Pred + *pajek*.

2.8.4 Preostali vzorci

Od preostalih vzorcev, ki se pojavljajo dovolj pogosto, da so bili zajeti v pričujočo analizo, je zanimiv le prvi v spodnji tabeli. Zveze tipa v obliki *pajka* ne bi smele ostati ne vključene v leksikalno zbirko, vprašanje je le, na katerem mestu je obravnava najbolj smiselna; glede na to, da je pri primerih tipa *po krivici* predlagana obravnava v sklopu predloga, se ta možnost ponuja tudi na tem mestu, vendar (kot rečeno v poglavju V-2.6) ideja zahteva s podatki podprto raziskavo.

DM SOZEM PAJEK	pogostnost v korpusu
v obliki pajka	13
PAJEK DM PPNZEM	
pajek v občinski	3
pajek na Ptujski	3
DO PAJEK RNN	
s pajkom tako	2
PAJEK SOZMR DT	
pajek SIP na	65
pajek sip na	22
DO PAJEK DT	
s pajkom za	9
s pajkom na	6
PPNMEID PAJEK DT	
hidravlični pajek za	12
hidravlični pajek na	6

Tabela 117: Nabor preostalih tridelnih vzorcev z lemo *pajek* in oznako za predlog.

2.9 Strasten – tridelni vzorci

Med najpogostejšimi tridelnimi vzorci je takih z lemo *strasten* in oblikoskladenjsko oznako za predlog 21 (20,4 %). Razvrstimo jih lahko v 2 vzorčna tipa, ki pa ne prinašata za nadaljnjo obravnavo zanimivih besednih nizov.

vzorčni tip	število vzorcev
Pred + <i>strasten</i> + Sam	16
<i>strasten</i> + Sam + Pred	4
drugi vzorci	1
GGDSTE DT STRASTEN	

Tabela 118: Nabor najpogostejših tridelnih vzorcev z lemo *pajek* in oblikoskladenjsko oznako za predlog.

2.9.1 Pred + *strasten* + Sam

V nadaljevanju navedeni besedni nizi imajo za dopolnjevanje leksikalne zbirke manjšo vrednost. Zveze pridevnika ter samostalnika so že bile predstavljene (glej V-1.3.1), glede na dvodelne pa spodnji nabor ne prinaša dodatnih informacij – razen v primeru, da je v središču interesa kvaliteta označevanja na ravni pripisanih sklonov, čemur pa je bilo posvečene dovolj pozornosti na predhodnih mestih (npr. v V-2.2.1.1).

besedni niz	pogostnost v korpusu						
v	strastno	razmerje	51	s	strastnim	tangom	4
v	strastnem	objemu	30	s	strastno	ljubeznijo	4
po	strastni	noči	19	za	strastne	zbiralce	4
o	strastni	ljubezni	16	s	strastno	vnemo	3
v	strastno	romanco	13	med	strastnim	poljubom	3
s	strastnim	poljubom	11	po	strastni	ljubezni	3
poleg	strastnega	ugankarja	9	za	strastne	igralce	3
pri	strastnih	kadilcih	9	k	strastnemu	odkrivanju	3
v	strastno	noč	6	po	strastnem	poljubljanju	3
v	strastno	afero	6	na	strastne	poljube	3
s	strastnim	grofom	6	v	strastnem	sovraštvu	3
za	strastne	kadilce	5	s	strastno	glasbo	3
s	strastnim	seksom	5	do	strastne	noči	3
s	strastnim	poljubljanjem	5	zaradi	strastne	ljubezni	3
med	strastnim	poljubljanjem	4	za	strasten	seks	3
za	strastno	ljubezen	4	za	strastno	avanturo	3
za	strastne	ribiče	4	za	strastna	čustva	3
iz	strastne	debate	4	v	strastno	ljubezen	3
zaradi	strastnega	razmerja	4	v	strastno	avanturo	3
s	strastnimi	poljubi	4	za	strastne	noči	3
za	strastnega	ribiča	4	v	strasten	poljub	3

Tabela 119: Izluščeni podatki – Pred + Prid + *strasten*.

2.9.2 *Strasten* + Sam + Pred

Iz nabora spodaj navedenih nizov je razvidno, da tudi na tem mestu obravnavani vzorčni tip ni relevanten za luščenje podatkov (podatki, kot je vidno, niso razvrščeni glede na spol samostalnika).

besedni niz	pogostnost v korpusu						
strasten	ljubezen	do	30	strasten	romanca	za	4
strasten	noč	z	29	strasten	srečanje	z	4
strasten	razmerje	z	24	strasten	navdušenje	nad	4
strasten	želja	po	20	strasten	hrepenenje	po	4
strasten	ljubezen	med	14	strasten	ljubezen	z	4
strasten	odnos	do	12	strasten	poljubljanje	z	3
strasten	lovec	na	11	strasten	poljub	z	3
strasten	zanimanje	za	9	strasten	seks	na	3
strasten	objem	z	8	strasten	romanca	z	3
strasten	poljub	na	7	strasten	urica	z	3
strasten	igralec	na	6	strasten	zaljubljenec	v	3
strasten	kadilec	v	5	strasten	boj	za	3
strasten	seks	z	4	strasten	afera	z	3
strasten	sovraštvo	do	4	strasten	kadilec	med	3
strasten	zgodba	o	4	strasten	borec	za	3

Tabela 120: Izluščeni podatki – *strasten* + Sam + Pred.

2.9.3 Preostali vzorec

Tako kot v prejšnjem poglavju tudi na tem mestu navedene vzorčne zapolnitve same na sebi niso zanimive za nadaljnjo obravnavo.

GGDSTE DT STRASTEN	pogostnost v korpusu
zaplete v strastno	32
spusti v strastno	15
spremeni v strastno	5
prelevi v strastnega	3

Tabela 121: Preostali tridelni vzorec z lemo *strasten* in oznako za predlog.

2.10 Plesati – tridelni vzorci

Z lemo *plesati* in oblikoskladenjsko oznako za predlog je med najpogostejšimi tridelnimi vzorci 31 primerov (30,1 %). Vzorce je mogoče razvrstiti v 4 vzorčne tipe, od katerih dva prinašata relevantne besedne nize (*plesati z volkovi*; *z veseljem plesati*) ter dva nerelevantne (*plesati v plesni*; *lahko plesati na*) za nadaljnjo obravnavo. Prav tako sta nerelevantna 2 vzorca, ki sta ostala izven evidentiranih tipov.

vzorčni tip	število vzorcev
<i>plesati</i> + Pred + Sam	18
<i>plesati</i> + Pred + Prid	5
Prisl + <i>plesati</i> + Pred	4
Pred + Sam + <i>plesati</i>	2
drugi vzorci	2
SOMMI PLESATI DM	
PLESATI RNN DT	

Tabela 122: Nabor najpogostejših tridelnih vzorcev z lemo *plesati* in oblikoskladenjsko oznako za predlog.

2.10.1 *Plesati* + Pred + Sam

Besedni nizi, ki jih prinaša obravnavani vzorčni tip, so v korpusu izredno pogoste, zato so v nadaljevanju navedeni le primeri s pogostnostjo 5 ali več. Kot je razvidno iz tabele, prinaša vzorčni tip tako za vključitev v zbirko relevantne primere (npr. *plesati [z volkovi, v krogu, po mizah]*) kot tudi primere, ki sami na sebi niso zaključene celote (npr. *plesati [ob zvokih, na konicah, pod vodstvom]*).

besedni niz		pogostnost v korpusu					
plesati	z	volkovi	42	plesati	v	ritmih	15
plesati	pred	očmi	37	plesati	z	dekletom	13
plesati	ob	zvokih	30	plesati	v	dežju	13
plesati	do	jutra	29	plesati	po	zraku	12
plesati	na	glasbo	29	plesati	v	disku	12
plesati	v	ritmu	28	plesati	po	mizi	11
plesati	v	krogu	26	plesati	ob	glasbi	11
plesati	v	skupini	24	plesati	na	konicah	11
plesati	na	odru	19	plesati	v	diskoteki	11
plesati	po	mizah	19	plesati	od	veselja	10
plesati	po	ulicah	16	plesati	v	parih	10
plesati	na	mizi	15	plesati	pod	vodstvom	9
plesati	v	paru	15	plesati	po	sobi	9
				plesati	v	diskotekah	9

plesati	z	zvezdami	9	plesati	v	kazini	6
plesati	na	ulicah	9	plesati	nad	Škofljico	6
plesati	v	baletu	9	plesati	na	ulici	6
plesati	do	konca	9	plesati	z	Jackom	6
plesati	z	glavo	8	plesati	na	mestu	6
plesati	v	predstavi	8	plesati	s	prijateljicami	6
plesati	na	soncu	8	plesati	po	glavi	5
plesati	na	Dunaju	8	plesati	v	klubih	5
plesati	po	taktih	8	plesati	na	zvoke	5
plesati	po	odru	8	plesati	pri	Kazini	5
plesati	ob	spremljavi	8	plesati	v	čast	5
plesati	pri	skupini	8	plesati	pri	Mojci	5
plesati	na	konici	7	plesati	v	Ljubljani	5
plesati	po	stanovanju	7	plesati	po	prostoru	5
plesati	od	leta	7	plesati	na	grobovih	5
plesati	za	denar	7	plesati	na	robu	5
plesati	po	cestah	6	plesati	v	zboru	5
plesati	pri	folklori	6	plesati	po	taktu	5
plesati	v	predstavah	6	plesati	v	vetru	5
plesati	po	cesti	6	plesati	na	zabavi	5
plesati	do	onemoglosti	6	plesati	v	družbi	5
plesati	v	transu	6	plesati	po	hiši	5

Tabela 123: Izluščeni podatki – *plesati* + Pred + Sam.

Na vrhu tabele je zveza *plesati z volkovi* (obarvana sivo), ki je v korpusu pogosta, ker se pojavlja kot ime filma *Pleše z volkovi*. Zaradi lematizacije se lastnoimenskost ter stalnost oblike zveze izgubita. Na tem mestu je zveza izpostavljena kot protiprimer v poglavju V-2.8.2 evidentirani zvezi *Strah pred pajki*.

2.10.2 *Plesati* + Pred + Prid

Spodaj naštetih nizi sami na sebi niso zanimivi za vključitev v leksikalno zbirko (izjema je sicer napačno lematizirana *plesati s telebajski*).

besedni niz		pogostnost v korpusu					
plesati	v	plesni	26	plesati	v	baletnih	4
plesati	v	folklorni	15	plesati	v	pariškem	4
plesati	v	folklorni	15	plesati	na	ameriško	4
plesati	do	zgodnjih	12	plesati	z	različnimi	4
plesati	z	belim	11	plesati	z	debelo	4
plesati	do	jutranjih	10	plesati	v	veliki	4
plesati	pri	folklorni	9	plesati	v	osnovni	4
plesati	v	plesnem	8	plesati	v	narodni	4
plesati	pri	plesni	6	plesati	v	glasbenih	3
plesati	na	Mozartovo	6	plesati	za	dober	3
plesati	pri	plesnem	5	plesati	v	številnih	3
plesati	za	lastno	5	plesati	za	dobrodelne	3
plesati	v	ljubljski	5	plesati	po	mestnih	3
plesati	na	posneto	5	plesati	v	akademski	3
plesati	v	ljubljskem	5	plesati	z	zaprtimi	3
plesati	v	različnih	5	plesati	v	mirnem	3
plesati	na	plesnih	4	plesati	v	nočnih	3
plesati	v	mariborskem	4	plesati	na	dobro	3
plesati	s	telebajski	4	plesati	po	napačni	3
plesati	s	plesno	4	plesati	v	londonskem	3
plesati	v	baletnem	4	plesati	po	točilnem	3

Tabela 124: Izluščeni podatki – *plesati* + Pred + Prid.

2.10.3 Prisl + *plesati* + Pred

Kot v poglavju V-2.4.2 so tudi na tem mestu navedeni potencialno zanimivi za nadaljnjo obravnavo.

besedni niz		pogostnost v korpusu		honorarno	plesati	v	4
lahko	plesati	na	13	kako	plesati	z	4
lahko	plesati	v	10	vedno	plesati	do	4
lahko	plesati	z	9	veliko	plesati	z	4
rad	plesati	z	9	vedno	plesati	na	4
rad	plesati	v	9	veselo	plesati	v	3
vedno	plesati	z	8	skupaj	plesati	na	3
vedno	plesati	v	7	kako	plesati	na	3
zdaj	plesati	v	7	divje	plesati	med	3
rad	plesati	na	6	prvič	plesati	na	3
naprej	plesati	z	6	lahko	plesati	ob	3
zvečer	plesati	v	6	danes	plesati	v	3
tako	plesati	na	5	danes	plesati	po	3
trenutno	plesati	v	5	lahko	plesati	do	3
vedno	plesati	po	5	res	plesati	z	3
zjutraj	plesati	v	5	vedno	plesati	pred	3
najprej	plesati	v	5	divje	plesati	po	3
kar	plesati	okoli	5	tako	plesati	z	3
prvič	plesati	v	5	brezplačno	plesati	na	3
nekoč	plesati	v	5	lahko	plesati	brez	3
lahko	plesati	po	5	nekajkrat	plesati	na	3
veselo	plesati	na	4	dobro	plesati	v	3

Tabela 125: Izluščeni podatki – Prisl + *plesati* + Pred.

2.10.4 Pred + Sam + *plesati*

Na tem mestu predstavljeni vzorčni tip je zanimiv za luščenje. Zveze so primerljive z naborom v V-2.10.1, so pa v korpusu znatno manj pogoste (tako pojavitve na ravni celotnega vzorčnega tipa kot tudi vsake od evidentiranih zvez posebej).

besedni niz		pogostnost v korpusu		od	veselja	plesati	3
z	veseljem	plesati	9	z	zgoščenke	plesati	3
na	odru	plesati	8	v	zraku	plesati	3
v	noč	plesati	7	v	okolici	plesati	3
na	premieri	plesati	6	v	predstavi	plesati	3
v	Ljubljani	plesati	5	pred	očmi	plesati	3
v	skupini	plesati	5	v	preteklosti	plesati	3
na	dvorišču	plesati	5	na	štoru	plesati	3
v	Sloveniji	plesati	5	v	mladosti	plesati	3
po	zraku	plesati	5	s	prijateljicami	plesati	3
na	primer	plesati	4	do	onemoglosti	plesati	3
od	odraslih	plesati	4	brez	odmora	plesati	3
pred	leti	plesati	4	s	prijatelji	plesati	3
nad	Zmincem	plesati	4	na	cesti	plesati	3
v	resnici	plesati	4	v	krogu	plesati	3
po	malem	plesati	4	od	malega	plesati	3
v	kopalkah	plesati	3	na	Dunaju	plesati	3

od	Slovencev	plesati	3	na	trgu	plesati	3
na	začetku	plesati	3				

Tabela 126: Izluščeni podatki – Pred + Sam + plesati.

2.10.5 Preostala vzorca

Nobeden od spodaj navedenih vzorcev ni zanimiv za nadaljnjo obravnavo, o čemer pričajo primeri najpogostejših vzorčnih zapolnitev.

SOMMI PLESATI DM	pogostnost v korpusu
ljudje plesati po	3
plesalci plesati v	3
PLESATI RNN DT	
plesati pozno v	16
plesati tja med	7
plesati dolgo v	5

Tabela 127: Preostala tridelna vzorca z lemo plesati in oznako za predlog.

2.11 Temeljito – tridelni vzorci

13 najpogostejših tridelnih vzorcev z lemo *temeljito*, ki vsebujejo predlog, med vsemi najpogostejšimi tridelnimi predstavlja 12,9 %. Vzorci so razvrščeni v 2 vzorčna tipa, od katerih prvi ne prinaša relevantnih besednih nizov, drugi pa delno oz. potencialno relevantne (*pred uporabo temeljito* ter *temeljito pripraviti na*).

vzorčni tip	število vzorcev
Pred + Sam + <i>temeljito</i>	5
<i>temeljito</i> + Glag + Pred	8

Tabela 128: Nabor najpogostejših tridelnih vzorcev z lemo temeljito in oblikoskladenjsko oznako za predlog.

2.11.1 Pred + Sam + temeljito

Kot je razvidno iz spodnje tabele, nizi, izluščeni s pomočjo na tem mestu obravnavanega vzorčnega tipa, niso zanimivi za uvrstitev v leksikalno zbirko.

besedni niz	pogostnost v korpusu	v	miru	temeljito	
pred uporabo	temeljito 15	v	celoti	temeljito	5
v Sloveniji	temeljito 13	do	časa	temeljito	4
na teden	temeljito 12	pred	uživanjem	temeljito	4
pred časom	temeljito 10	po	svetu	temeljito	4
pred nakupom	temeljito 9	pred	spanjem	temeljito	4
po kopanju	temeljito 8	pred	vožnjo	temeljito	4
na leto	temeljito 7	pred	zaužitjem	temeljito	4
o zadevi	temeljito 7	v	črevesju	temeljito	4
na dan	temeljito 6	pred	leti	temeljito	4
po uporabi	temeljito 6	pred	dnevi	temeljito	4
na cedilu	temeljito 5	na	ministrstvu	temeljito	4
v Celju	temeljito 5	z	rokami	temeljito	4
pred odhodom	temeljito 5	po	čiščenju	temeljito	4
za preliv	temeljito 5	z	igralkami	temeljito	3
pred sajenjem	temeljito 5	v	knjigi	temeljito	3

v	bazen	temeljito	3	v	Velenju	temeljito	3
na	razpis	temeljito	3	po	umivanju	temeljito	3
o	dohodnini	temeljito	3	pred	odločitvijo	temeljito	3
v	EU	temeljito	3	v	praksi	temeljito	3
v	Italiji	temeljito	3	za	ozimnico	temeljito	3
v	prihodnosti	temeljiteje	3	na	pregled	temeljito	3
v	državi	temeljito	3	pred	podpisom	temeljito	3
o	samomarih	temeljito	3	v	mešalniku	temeljito	3
po	potrebi	temeljito	3	po	vojni	temeljito	3
v	komendi	temeljito	3	z	okolico	temeljito	3
v	klubu	temeljito	3	z	vilicami	temeljito	3
pred	odločanjem	temeljito	3	za	klanje	temeljito	3
pred	setvijo	temeljito	3	do	jeseni	temeljito	3
z	bolnikom	temeljito	3	pred	uvedbo	temeljito	3
na	dvoboj	temeljito	3	v	Laškem	temeljito	3
do	poletja	temeljito	3	z	dalnogledom	temeljito	3

Tabela 129: Izluščeni podatki – Pred + Sam + temeljito.

2.11.2 Temeljito + Glag + Pred

Spodnji nizi so potencialno relevantni v smislu obravnave glagola s sledečim predlogom kot kolokacijske enote (glej poglavje V-2.4.2). Zaradi številčnosti primerov so spodaj navedeni le primeri s pogostnostjo 5 ali več.

besedni niz		pogostnost v korpusu		temeljito	posegati	v	11
temeljito	pripraviti	na	207	temeljito	umešati	v	11
temeljito	pripravljeni	na	125	temeljito	namazati	z	10
temeljito	seznaniti	z	116	temeljito	obdelati	v	10
temeljito	razmisliti	o	113	temeljito	obračunati	z	10
temeljito	premisliti	o	65	temeljito	obnoviti	v	9
temeljito	pogovoriti	z	63	temeljito	izkoristiti	za	9
temeljito	ukvarjati	z	50	temeljito	zamisliti	nad	9
temeljito	pogovoriti	o	46	temeljito	razlikovati	od	9
temeljito	sprati	z	43	temeljito	podučiti	o	9
temeljito	poseči	v	42	temeljito	obrisati	z	9
temeljito	oprati	z	35	temeljito	pripravljeni	za	9
temeljito	poskrbeti	za	34	temeljito	zdrgniti	z	8
temeljito	očistiti	z	32	temeljito	izmiti	z	8
temeljito	poučiti	o	28	temeljito	splakniti	z	8
temeljito	obdelati	z	25	temeljito	odcediti	na	7
temeljito	poglobiti	v	24	temeljito	odstraniti	z	7
temeljito	posvetovati	z	24	temeljito	pozanimati	o	7
temeljito	oprati	pod	22	temeljito	spremeniti	v	7
temeljito	izprašati	o	22	temeljito	poprijeti	za	7
temeljito	zmešati	z	20	temeljito	spoprijeti	z	7
temeljito	premešati	z	19	temeljito	pripraviti	z	7
temeljito	pripraviti	za	19	temeljito	opraviti	z	7
temeljito	odrgniti	z	17	temeljito	obnoviti	pred	7
temeljito	umiti	z	16	temeljito	ukvarjati	v	6
temeljito	vplivati	na	16	temeljito	skrbeti	za	6
temeljito	osušiti	z	15	temeljito	preizkusiti	na	6
temeljito	razmišljati	o	15	temeljito	vzeti	pod	6
temeljito	izprati	z	14	temeljito	pomesti	z	6
temeljito	pobrsinati	po	12	temeljito	počistiti	z	6
temeljito	oprati	v	12	temeljito	razpravljati	o	6

temeljito	vtreti	v	6	temeljito	pregledati	z	5
temeljito	pregledati	v	5	temeljito	obveščati	o	5
temeljito	mešati	z	5	temeljito	počistiti	za	5
temeljito	zaliti	z	5	temeljito	prenoviti	v	5
temeljito	oplakniti	pod	5	temeljito	ločiti	od	5
temeljito	pognojiti	z	5	temeljito	prekriti	z	5
temeljito	natreti	z	5	temeljito	poškropiti	z	5
temeljito	splakniti	pod	5	temeljito	razgledati	po	5

Tabela 130: Izluščeni podatki – temeljito + Glag + Pred.

3 Vzorci z veznikom

Velika večina vzorcev oz. vzorčnih tipov, ki vsebujejo veznik, ni relevantna za luščenje besednih nizov. Med nezanimive za dopolnjevanje leksikalne zbirke spadajo vsi dvodelni nizi, ne glede na to, ali je lema pred veznikom ali za njim, ter vsi tridelni nizi, kjer se veznik pojavlja na začetku ali na koncu vzorca – pri tridelnih nizih so torej zanimivi le tisti, kjer se veznik pojavlja na sredini vzorca. Nekaj primerov za nabor vzorcev, ki ne prinašajo relevantnih informacij, je na voljo v poglavju V-3.1.

Kot bo vidno iz nadaljevanja, med najpogostejšimi vzorci najdemo le primere, ki prinašajo **priredni veznik** – kar je sicer skladno s predpostavko, da so priredni vezniki na besednozvezni ravni pogostejši od podrednih: »[p]riredni vezniki so mogoči v prostem in zloženem stavku, podredni pa le v zloženem (izjema so primerjalni *ko, kakor, kot*)« (Toporišič 2004⁴: 432). Obenem to pomeni, da besedne zveze s primerjalnimi vezniki v pričujočem poglavju ostajajo neanalizirane.

Pri tridelnih vzorcih je predvidljivo, da so za luščenje relevantni tisti, ki prinašajo na obeh straneh prirednega veznika enako besedno vrsto, kar analiza rezultatov potrjuje. Opisani tip vzorcev v nadaljevanju za lažjo obravnavo imenujemo **simetrični vzorci**. Za luščenje slednjih je bil pripravljen specializiran program (za opis programa glej IV-3.3.6). Kadar je med najpogostejšimi poleg simetričnih najti še druge vzorce, so v pričujočem poglavju navedeni le primeri vzorčnih zapolnitev zanje, z željo identifikacije potencialnih problemov na označevalni ravni.

Podatke o zastopanosti vzorcev z oblikoskladenjsko oznako za veznik prinaša spodnja tabela. Prva številka v tabeli predstavlja število vzorcev z veznikom, druga vse ročno pregledane vzorce, sledi delež, zaokrožen na eno decimalno. Druga vrstica tabele prinaša podatke o številu vzorcev, ki so bili prepoznani za nerelevantne glede na pozicijo oznake za veznik znotraj vzorca ter glede na to odstranjeni iz nadaljnje analize (tj. vsi vzorci, ki prinašajo oznako za veznik na začetku ali koncu vzorca, ne glede na dolžino vzorca).

	dvodelni vzorci		tridelni vzorci	
<i>pajek</i>	4 (101)	4,0 %	29 (101)	28,7 %
odstranjenih iz nadaljnje analize	4		12	
<i>strasten</i>	4 (101)	4,0 %	32 (103)	31,1 %
odstranjenih iz nadaljnje analize	4		17	
<i>plesati</i>	4 (100)	4,0 %	31 (103)	30,1 %
odstranjenih iz nadaljnje analize	4		14	
<i>temeljito</i>	4 (100)	4,0 %	23 (101)	22,8 %

odstranjenih iz nadaljnje analize

4

19

Tabela 131: Delež vzorcev z veznikom med vsemi najpogostejšimi.

3.1 Nerelevantni vzorci

Primeri v nadaljevanju so namenjeni zgolj prikazu vrste besednih nizov, ki glede na predpostavko o nerelevantnosti niso vključeni v nadaljnjo obravnavo. Pregledane so bile sicer le zapolnitve najpogostejših vzorcev, ne pa tudi na osnovi vzorčnih tipov luščeni nizi, vendar je na tem mestu utemeljeno predvidevati, da izpuščeni nabor ne prinaša za vključitev v zbirko uporabnih podatkov.

VP PAJEK	pogostnost v korpusu
in pajek	371
in pajki	105
ter pajek	43
PAJEK VP	
pajek in	103
pajkov in	102
pajki in	100
PAJEK SOZMR VP	
pajek SIP in	63
pajek SIP ter	2
VP PAJEK SOMEI	
in pajek SIP	51
ter pajki spider	11
ter pajek SIP	8

Tabela 132: Primeri nerelevantnih vzorcev z lemo *pajek* ter oblikoskladenjsko oznako za veznik.

3.2 Pajek – tridelni vzorci

Poleg simetričnih vzorčnih tipov, ki prinašata za vključitev v zbirko relevantne besedne zveze (*lisice in pajek*; *pajek in lisice*) je na seznamu za analizo še en vzorec, ki pa, kot bo razvidno iz primerov, ni relevanten za luščenje (*pajek in diatonično*).

vzorčni tip	število vzorcev
<i>pajek</i> + Vp + Sam	7
Sam + Vp + <i>pajek</i>	9
drugi vzorci	1
PAJEK VP RNN	

Tabela 133: Nabor najpogostejših tridelnih vzorcev z lemo *pajek* in oblikoskladenjsko oznako za veznik.

3.2.1 Simetrična vzorčna tipa

Spodnji tabeli prinašata primere za simetrična vzorčna tipa, in sicer z lemo *pajek*, oblikoskladenjsko oznako za samostalni ter (na sredini) priredni veznik. Besedni nizi prinašajo oba samostalnika v lematizirani obliki – kar mdr. pomeni, da je beseda *lisice* v spodnji tabeli lematizirana v množinsko obliko, kar je glede na izvorni kontekst primerov ustrezno (na seznamu se sicer pojavlja tudi *lisica*, vendar z nižjo pogostnostjo).

samostalnik + veznik + pajek			pog.	pajek + veznik + samostalnik			pog.
lisice	in	pajek	89	pajek	in	lisice	57
zgrabljajnik	in	pajek	12	pajek	in	škorpijon	20
kača	in	pajek	11	pajek	in	zgrabljajnik	16
plug	in	pajek	11	pajek	in	kača	12
žuželka	in	pajek	10	pajek	in	žuželka	9
dvig	in	pajek	9	pajek	in	traktor	5
škorpijon	in	pajek	8	pajek	in	podgana	5
sip	in	pajek	6	pajek	in	muha	4
kača	ali	pajek	6	pajek	ali	lisice	4
obračalnik	in	pajek	6	pajek	in	pajčevina	4
podgana	in	pajek	5	pajek	in	sadilec	4
grablje	in	pajek	5	pajek	in	vitel	3
redar	in	pajek	5	pajek	ali	kača	3
policist	in	pajek	4	pajek	in	ščurek	3
BCS	in	pajek	4	pajek	in	puhalnik	3
insekt	in	pajek	3	pajek	in	krokodil	3
muha	in	pajek	3	pajek	in	Katarina	3
vijak	in	pajek	3				

Tabela 134: Izluščeni podatki – simetrična priredna vzorčna tipa z lemo *pajek*.

Pogosto se samostalniki, tipično pojavljajoči se v prirednih zvezah z besedo *pajek*, pojavljajo pri obeh obravnavanih vzorčnih tipih – primeri, ki se ponovijo, so zgoraj obarvani sivo. Iz tabele je razvidno tudi, da se od veznikov v najpogostnejših izluščenih zvezah večinoma pojavlja *in*, redko tudi *ali*.

Z upoštevanjem dejstva, da se na predlagani način izgubi morebitna tendenca k stalnosti besednega reda v obravnavanih besednih zvezah, je podatke iz obeh gornjih tabel možno združiti, pri čemer se še jasneje pokažejo razlike v pogostnosti sopoljavljajočih se elementov:

samostalnik, ki stopa v priredje s <i>pajek</i>		pog.		
lisice	154	BCS		4
kača	33	rak		4
škorpijon	30	sadilec		4
zgrabljajnik	29	cisterna		3
žuželka	21	gosenica		3
plug	11	trosilec		3
podgana	10	puhalnik		3
dvig	9	metulj		3
muha	8	pes		3
redar	7	lisica		3
policist	7	ščurek		3
sip	6	krokodil		3
obračalnik	6	vijak		3
insekt	6	plazilec		3
grablje	5	kazen		3
kuščar	5	Katarina		3
traktor	5	čebela		3
pajčevina	4	opica		3
robot	4	vitel		3

Tabela 135: Samostalniki, ki stopajo v priredje s samostalnikom *pajek*.

3.2.1.1 Analiza označevanja

Ob luščenju nizov, predstavljenih v gornjih tabelah, so bili v posebno datoteko zbrani primeri, pri katerih (glede na pripisane oblikoskladenjske oznake) samostalnika ne izkazuje enakega sklonu. Za ta postopek je bil uporabljen specializiran program (opis programa v IV-3.3.6). Morda še bolj kot za analizo označevanja je pregled spodnjih primerov smiselna za identifikacijo nizov, kjer je eden od nastopajočih samostalnikov del širše besedne zveze¹¹⁵, kar sugerira obravnavo teh primerov na drugem mestu (tj. v sklopu luščenja podatkov za daljši vzorčni tip). Spodnja tabela prinaša združene podatke za oba obravnavana vzorčna tipa:

besedni niz		pogostnost v korpusu					
krompirja	in	pajek	17	žuželke	in	pajke	5
kolesa	ter	pajke	13	prikolico	in	pajek	5
gnoja	in	pajek	12	pajek	in	cisterno	4
silaze	in	pajek	7	koruze	in	pajek	3
pajek	in	samonakladalko	7	pajek	in	kosilnico	3
cisterno	in	pajek	6	kosilnico	in	pajek	3
samonakladalko	in	pajek	6	žuželke	in	pajki	3
pajke	in	lisice	5	litrov	in	pajek	3

Tabela 136: Analiza označenosti: Sam + Vp + Sam – neujemanje v sklonu.

Skupina nizov s prvim samostalnikom v rodilniku izkazuje primere, ki so potencialno del širših besednih zvez (npr. [sadirnik krompirja, odjemalnik silaze, nakladalec gnoja] in pajek). Nekaj primerov je neustrezno označenih, ker v leksikalni zbirki ni predvideno, da se beseda *pajek* sklanja po paradigmi za samostalnike, ki izkazujejo podspol neživosti (v primerih tipa *prikolico in pajek* je npr. zadnji samostalnik označen z imenovalniškim sklonom). V primerih *kolesa in pajke* ter *pajke in lisice* pa sta z imenovalniškim sklonom označena samostalnika *kolo* ter *lisice*, *pajek* pa je označen z oblikoskladenjsko oznako za tožilnik.

3.2.2 Preostali vzorec

Za analizo ostaja še en vzorec, ki prinaša kombinacijo leme *pajek* z oblikoskladenjsko oznako za veznik ter prislov. Spodnji primeri kažejo, da vzorec kot tak ni relevanten za luščenje:

PAJEK VP RNN	pogostnost v korpusu
pajek ali diatonično	2
pajek in kmalu	2
pajek ali angleško	2

Tabela 137: Preostali tridelni vzorec z lemo *pajek* in prirednim veznikom.

3.3 Strasten – tridelni vzorci

Poleg obeh simetričnih vzorčnih tipov, ki prinašata relevantne zveze (*romantičen in strasten*; *strasten in čustven*), prinaša spodnja tabela še vzorčni tip, ki vsebuje samostalnik (*uspeh in strasten*) ter dva vzorca, ki vsebujeta oznako za prislov (*svobodno in strastno*; *strastno in viharo*). Nobena od slednjih treh naštetih skupin besednih nizov ni relevantna za nadaljnjo obravnavo.

vzorčni tip	število vzorcev
<i>strasten</i> + Vp + Prid	5

¹¹⁵ Pri avtomatskem iskanju teh primerov se opiramo na sklon samostalnika, kar pomeni, da je postopek uspešen le pri določenem tipu besednih zvez.

Prid + Vp + <i>strasten</i>	5
Sam + Vp + <i>strasten</i>	3
drugi vzorci	2
STRASTEN VP RNN	
RNN VP STRASTEN	

Tabela 138: Nabor najpogostejših tridelnih vzorcev z lemo *strasten* in oblikoskladenjsko oznako za veznik.

3.3.1 Simetrična vzorčna tipa

V nadaljevanju so podatki za oba simetrična vzorčna tipa z lemo *strasten* združeni v skupni tabeli:

besedni niz		pogostnost v korpusu		strasten		in	silovit	4
romantičen	in	strasten	12	strasten	in	predan		4
vroč	in	strasten	11	intimen	in	strasten		4
močen	in	strasten	11	strasten	in	romantičen		4
ljubeč	in	strasten	11	hiter	in	strasten		4
globok	in	strasten	10	strasten	in	strašen		4
lep	in	strasten	10	dinamičen	in	strasten		3
dolg	in	strasten	10	skrivnosten	in	strasten		3
čustven	in	strasten	9	poželjiv	in	strasten		3
divji	in	strasten	8	drzen	in	strasten		3
strasten	in	čustven	8	znan	in	strasten		3
čuten	in	strasten	8	strasten	in	nenasiten		3
nežen	in	strasten	7	goreč	in	strasten		3
strasten	in	dolg	7	strasten	in	tenkočuten		3
strasten	in	ljubeč	7	ponosen	in	strasten		3
strasten	in	temperamenten	6	strasten	in	izkušen		3
strasten	in	vroč	6	navdušen	in	strasten		3
strasten	in	čuten	6	strasten	in	dramatičen		3
pameten	in	strasten	6	strasten	in	ognjevit		3
strasten	in	zahteven	5	strasten	in	eleganten		3
strasten	in	intenziven	5	strasten	in	vročekrven		3
ognjevit	in	strasten	5	strasten	in	brezkompromisen		3
strasten	in	močen	5	posestniški	in	strasten		3
velik	in	strasten	5	zapeljiv	in	strasten		3
strasten	in	zapeljiv	4	svoboden	in	strasten		3
strasten	in	goreč	4	strasten	in	izgubljen		3
strasten	in	divji	4	strasten	in	vdan		3
strasten	in	pustolovski	4	živahen	in	strasten		3
strasten	in	viharen	4					

Tabela 139: Izluščeni podatki – simetrična priredna vzorčna tipa z lemo *strasten*.

V spodnji tabeli pa so z združeno pogostnostjo navedeni še pridevniki, ki se sopoljavljajo s *strasten* bodisi na levi ali desni strani priredne besedne zveze:

pridevnik, ki stopa v priredje s <i>strasten</i>	pog.	čuten	14
romantičen	20	globok	12
ljubeč	18	divji	12
čustven	18	lep	11
vroč	17	nežen	10
dolg	17	ognjevit	8
močen	16	temperamenten	8

goreč	7	poln	3
zapeljiv	7	nenasiten	3
zahteven	7	odločen	3
velik	6	eleganten	3
intenziven	6	tenkočuten	3
viharen	6	brezkompromisen	3
pameten	6	vdan	3
silovit	6	izgubljen	3
predan	5	drzen	3
dramatičen	5	duhovit	3
intimen	4	oster	3
strašen	4	dinamičen	3
svoboden	4	vznemirljiv	3
hiter	4	znan	3
pustolovski	4	nepredvidljiv	3
ponosen	4	navdušen	3
buren	4	odličen	3
posestniški	4	glasen	3
dober	4	iskren	3
živahen	4	očarljiv	3
izkušen	4	vročekrven	3
poželjiv	4	pozoren	3
skrivnosten	4	zanimiv	3
zvest	4	topel	3

Tabela 140: Pridevniki, ki stopajo v priredje s pridevnikom *strasten*.

3.3.1.1 Analiza označenosti

Primerov, kjer glede na oznake prihaja do neujemanja med sklonoma obeh polnopomenskih besed, je manj kot pri zvezah z lemo *pajek*, vendar gre pri vzorčnih tipih, analiziranih na tem mestu, hitreje sklepati na označevalne probleme, saj je v primeru besednozvezne prirednosti¹¹⁶ ujemanje obeh pridevnikov s sledečim (na tem mestu izpuščenim) samostalnikom pričakovati v vseh primerih.

besedni niz		pogostnost v korpusu					
burne	in	strastne	2	neverjetnega	in	strastnega	1
strastne	in	nestrpne	2	strastne	in	silovite	1
strasten	in	pričakujoč	1	strastnemu	in	črno	1
strastni	in	boleči	1	strastnega	in	erotičnega	1
kratki	in	strastni	1	nenadni	in	strastni	1
razmišljajoče	in	strastne	1	strastne	ali	ljubeče	1
nežne	in	strastne	1	nevtralne	ali	strastne	1
strastne	ter	izkušene	1	strastnim	in	zahtevnim	1
zapeljivega	in	strastnega	1	premožne	in	strastne	1
izkušene	in	strastne	1	močne	in	strastne	1
ženstvene	in	strastne	1	strastne	in	šaljive	1
slastne	in	strastne	1	bučni	in	strastni	1
natančne	in	strastne	1				

¹¹⁶ Iz predpostavke so izključeni primeri tipa (primer je izmišljen za potrebe ponazoritve) *Bil je zelo strasten in burnih besednih spopadov se ni nikoli branil*. Ločevanje prvega od drugega na ravni luščenja tridelnih besednih nizov, tj. brez upoštevanja širšega skladenjskega konteksta, sicer ni mogoče, pa vendar je utemeljeno sklepati, da so pri zapolnitvah obravnavanega vzorčnega tipa stavčna priredja redkejša od besednozveznih.

Tabela 141: Analiza označenosti: Prid + Vd + Prid – neujemanje v sklonu.

Ob vnosu prve od zvez kot iskalni pogoj v korpusu FidaPLUS najdemo 3 zadetke:

pravi vihar čustev. Zadovoljni z učinkom, boste preživljali **burne in strastne** trenutke. da bi se osvobojeno in sproščeno
podala tudi v bolj **burne in strastne** glasbene pripovedi novejšega datuma.
sprožili pravi vihar čustev. Zadovoljni z učinkom boste preživljali **burne in strastne** trenutke.

Konkordančni niz FidaPLUS 19: *burne_in_strastne*.

Eden od primerov v gornjem konkordančnem nizu se sicer ponovi, pri čemer je zanimivo, da je označevanje prav v teh dveh primerih različno – pri prvem primeru je ustrezno označeno ujemanje v tožilniškem sklonu, pri drugem ne: glede na oblikoskladenjske oznake naj bi bil prvi pridevnik v rodilniku, drugi v tožilniku.

3.3.2 Preostali vzorci

Med preostalimi vzorci je najti še en vzorčni tip, ki pa sam na sebi ne prinaša za nadaljnjo obravnavo relevantnih izluščenih besednih nizov:

besedni niz		pogostnost v korpusu		dan	in	strasten	
uspeh	in	strasten	7	ljubezen	in	strasten	3
gospodinja	in	strasten	4	seks	in	strasten	3

Tabela 142: Izluščeni podatki – Sam + Vp + *strasten*.

Neuvrščena v vzorčne tipe sta še dva vzorca, pri katerih se lema *strasten* pojavlja v kombinaciji z oznako za prislov. Spodnje vzorčne zapolnitve potrjujejo predvidevanje, da gre v nekaterih primerih za napačno označene pridevniške oblike (ker je bilo o ločevanju pridevniških od prislovnih oblik dosti napisanega že na drugih mestih (denimo V-1.1.5.1, V-1.3.4.1, V-1.7.2.1), za te vzorce v trenutnem poglavju ne sledi natančnejša analiza označevanja).

STRASTEN VP RNN	pogostnost v korpusu
strastno in viharo	4
strastne in obenem	3
strastni in zelo	3
RNN VP STRASTEN	
svobodno in strastno	6
veliko in strastno	3
čutno in strastno	2
hkrati pa strasten	2

Tabela 143: Preostala tridelna vzorca z lemo *strasten* in prirednim veznikom.
3.4 Plesati – tridelni vzorci

Poleg simetričnih vzorčnih tipov, ki prinašata relevantne rezultate (*peti in plesati; plesati in peti*), v spodnji tabeli najdemo še tri v vzorčne tipe neuvrščene vzorce, od katerih nobeden ne prinaša relevantnih besednih nizov (*plesati in tako; zdaj pa plesati; glasbo in plesati*).

vzorčni tip	število vzorcev
<i>plesati</i> + Vp + Glag	7

Glag + Vp + <i>plesati</i>	7
drugi vzorci	3
PLESATI VP RNN	
RNN VP PLESATI	
SOZET VP PLESATI	

Tabela 144: Nabor najpogostejših tridelnih vzorcev z lemo *plesati* in oblikoskladenjsko oznako za veznik.

3.4.1 Simetrična vzorčna tipa

Tudi v pričujočem poglavju so v nadaljevanju podatki za oba simetrična vzorčna tipa z lemo *plesati* predstavljeni v skupni tabeli.

besedni niz	pogostnost v korpusu		rolati	in	plesati	
peti in plesati	388		plesati	in	ploskati	5
plesati in peti	210		plesati	in	poučevati	5
plesati in igrati	71		bobnati	in	plesati	5
prepevati in plesati	55		plesati	in	tekmovati	5
igrati in plesati	47		plesati	in	delati	4
piti in plesati	39		plesati	in	učiti	4
zabavati in plesati	28		plesati	in	mahati	4
plesati in prepevati	25		plesati	ali	igrati	4
plesati in skakati	21		ploskati	in	plesati	4
plesati in uživati	18		plesati	in	vriskati	4
plesati in piti	16		noreti	in	plesati	4
plesati in plesati	14		plesati	in	uganjati	4
plesati in koreografirati	13		plesati	in	improvizirati	3
plesati in nastopati	10		sedeti	ali	plesati	3
plesati in rajati	10		plesati	in	početi	3
poskakovati in plesati	8		plesati	in	jesti	3
plesati in hoditi	8		peti	ter	plesati	3
plesati in veseliti	7		telovaditi	in	plesati	3
plesati in imeti	7		plesati	in	žurati	3
plesati in poslušati	7		plesati	in	trenirati	3
plesati in noreti	7		plesati	in	praznovati	3
jesti in plesati	7		teči	in	plesati	3
skakati in plesati	7		plesati	in	zabijati	3
smejati in plesati	6		priti	in	plesati	3
vriskati in plesati	6		peti	ali	plesati	3
veseliti in plesati	6		plesati	ter	peti	3
plavati in plesati	6		plesati	in	navijati	3
poljubljati in plesati	6		poučevati	in	plesati	3
vteti in plesati	6		plesati	in	slaviti	3
pogovarjati in plesati	6		kuhati	in	plesati	3
plesati in plavati	6		plesati	in	obiskovati	3
plesati in kričati	6		plesati	in	deklamirati	3
rajati in plesati	5		popivati	in	plesati	3
plesati in zabavati	5					

Tabela 145: Izluščeni podatki – simetrična priredna vzorčna tipa z lemo *plesati*.

Sledi še združen prikaz glagolov, ki se v zvezah obravnavanega tipa z glagolom *plesati* najpogosteje pojavljajo:

glagol, ki stopa v priredje s <i>plesati</i>	pog. v korpusu		
peti	608	igrati	123
		prepevati	82

piti	58	slaviti	4
zabavati	35	mahati	4
skakati	30	trenirati	4
uživati	22	uganjati	4
rajati	16	risati	4
koreografirati	14	iti	4
veseliti	13	obiskovati	4
nastopati	13	oboževati	4
plavati	12	početi	4
poskakovati	12	sodelovati	4
noreti	11	stati	3
poslušati	10	improvizirati	3
jesti	10	kolesariti	3
vriskati	10	govoriti	3
ploskati	9	navijati	3
imeti	8	pomagati	3
smejati	8	zabijati	3
hoditi	8	oblačiti	3
poučevati	8	objemati	3
delati	7	recitirati	3
telovaditi	7	seksati	3
popivati	7	večerjati	3
kričati	7	plezati	3
vteti	7	priti	3
pogovarjati	6	loviti	3
poljubljati	6	drsati	3
sedeti	5	znati	3
teči	5	potovati	3
praznovati	5	žurirati	3
deklamirati	5	žvižgati	3
bobnati	5	sprostiti	3
tekmovati	5	žurati	3
rolati	5	kuhati	3
učiti	4	sprehajati	3
gosti	4		

Tabela 146: Glagoli, ki stopajo v priredje z glagolom *plesati*.

3.4.2 Preostali vzorci

Spodaj navedeni trije vzorci ne prinašajo za dopolnitev leksikalne zbirke relevantnih zapolnitev. Iz podatkov je razvidno tudi, da je posebno pozornost pri označevanju potrebno nameniti kombinaciji prislova ter besede *pa*, za katero je (vsaj glede na tukaj pridobljene podatke) redkeje ustrezna oblikoskladenjska oznaka za veznik kot *pa* za členek – primeri so v tabeli podčrtani.

PLESATI VP RNN

plesati in tako	6
plesati in kako	3
plesati in nato	3
plesati in ponavadi	3

RNN VP PLESATI

<u>zdaj pa plesati</u>	7
<u>potem pa plesati</u>	6
<u>gotovo pa plesati</u>	2
<u>najraje pa plesati</u>	2

naokoli in plesati	2
<u>nato pa plesati</u>	2
<u>ponoči pa plesati</u>	2
SOZET VP PLESATI	
glasbo in plesati	10
kitaro in plesati	5
dvorano in plesati	4

Tabela 147: Preostali tridelni vzorci z lemo *plesati* in prirednim veznikom.

3.5 Temeljito – tridelni vzorci

Podatkovna tabela z najpogostejšimi tridelnimi vzorci prinaša 4 vzorce, ki vsebujejo lemo *temeljito* ter priredni veznik. Relevantne podatke prinašata prva dva vzorca (*hitro in temeljito*; *temeljito in strokovno*), druga dva pa nerelevantne (*vodo in temeljito*; *operemo in temeljito*):

vzorci

TEMELJITO VP RNN
RNN VP TEMELJITO
SOZET VP TEMELJITO
GGDSPM VP TEMELJITO

Tabela 148: Nabor najpogostejših tridelnih vzorcev z lemo *temeljito* in oblikoskladenjsko oznako za veznik.

3.5.1 Simetrična vzorčna tipa

Ker so oblikoskladenjske oznake za prislov manj členjene od ostalih oznak za polnopomenske besede in zato vzorci z njimi niso pogosti in ker se je takšna praksa izkazala za uspešno pri prejšnjih obravnavanih primerih, so na tem mestu predstavljeni podatki za vzorčna tipa **Prisl + Vp + temeljito** ter **temeljito + Vp + Prisl**, kljub temu da je med najpogostejšimi najti le vzorca z oznako RNN. Podatki so predstavljeni za oba vzorčna tipa v skupni tabeli:

besedni niz		pogostnost v korpusu		temeljito	in	obsežno	7
hitro	in	temeljito	68	temeljito	in	podrobno	7
počasi	in	temeljito	28	postopoma	in	temeljito	6
resno	in	temeljito	17	temeljito	in	pogosto	6
skrbno	in	temeljito	15	temeljito	in	dokončno	6
temeljito	in	strokovno	15	pravilno	in	temeljito	6
redno	in	temeljito	15	globoko	in	temeljito	6
nato	pa	temeljito	12	vestno	in	temeljito	6
dolgo	in	temeljito	12	nežno	in	temeljito	6
temeljito	in	vsestransko	10	temeljito	in	globoko	5
pravočasno	in	temeljito	10	sistematično	in	temeljito	5
temeljito	in	sistematično	9	pogosto	in	temeljito	5
temeljito	in	celovito	9	takoj	in	temeljito	5
potem	pa	temeljito	9	sproti	in	temeljito	5
dobro	in	temeljito	9	previdno	in	temeljito	5
temeljito	in	pravočasno	8	temeljito	in	sproti	5
temeljito	in	bolj	8	temeljito	in	redno	5
temeljito	in	resno	8	temeljito	in	analitično	5
temeljito	in	hitro	8	večkrat	in	temeljito	4
natančno	in	temeljito	8	konkretno	in	temeljito	4
strokovno	in	temeljito	8	obširno	in	temeljito	4
temeljito	in	natančno	7	preudarno	in	temeljito	4

temeljito	in	obširno	4	temeljito	in	kakovostno	3
prej	in	temeljito	4	temeljito	in	dolgoročno	3
temeljito	in	lepo	4	več	in	temeljito	3
lahko	in	temeljito	4	temeljito	in	čvrsto	3
veliko	in	temeljito	4	načrtno	in	temeljito	3
temeljito	in	dosledno	4	učinkovito	in	temeljito	3
odgovorno	in	temeljito	4	postopno	in	temeljito	3
hkrati	pa	temeljito	3	temeljito	in	odgovorno	3
znova	in	temeljito	3	obsežno	in	temeljito	3
temeljito	in	načrtno	3	temeljito	in	pregledno	3
temeljito	in	zanesljivo	3	temeljito	in	skrbno	3
široko	in	temeljito	3	mirno	in	temeljito	3
temeljito	in	dobro	3	temeljito	in	nežno	3
odločno	in	temeljito	3				

Tabela 149: Izluščeni podatki – simetrična priredna vzorca z lemo *temeljito*.

Kot pri prejšnjih lemah se tudi na tem mestu izkazuje zamenljivost obeh polnoprimerov besed, čeprav za primere na vrhu tabele morda nekoliko manj (*hitro in temeljito* se denimo pojavlja znatno večkrat kot *temeljito in hitro*). Drugo opozorilo velja zvežam z besedo *pa*, ki za prislovom (kot rečeno v V-3.4.2) pogosto ne nastopa v prirednovezniški vlogi (primeri so v gornji tabeli obarvani sivo). Sledi še nabor najpogostejših prislovov, ki se v zvezah obravnavanega tipa pojavljajo s *temeljito*:

prislov, ki stopa v priredje s <i>temeljito</i>	pog.		
hitro	77	analitično	6
počasi	33	veliko	5
resno	25	večkrat	5
strokovno	24	previdno	5
redno	22	lepo	5
skrbno	18	preudarno	5
pravočasno	18	konkretno	5
natančno	15	takoj	5
sistematično	14	učinkovito	5
dobro	13	prej	5
nato	13	dosledno	4
dolgo	13	več	4
globoko	12	zanesljivo	4
sproti	11	dolgoročno	4
celovito	11	pregledno	4
pogosto	11	treba	4
vsestransko	10	postopno	4
obsežno	10	pozorno	3
podrobno	9	kritično	3
potem	9	strogo	3
nežno	9	uspešno	3
bolj	9	odločno	3
vestno	8	obenem	3
obširno	8	mirno	3
pravilno	7	izčrpno	3
odgovorno	7	znova	3
dokončno	7	kakovostno	3
lahko	6	čvrsto	3
načrtno	6	hkrati	3
postopoma	6	široko	3

Tabela 150: Prislovi, ki stopajo v priredje s prislovom *temeljito*.

3.5.2 Preostala vzorca

Preostala dva vzorca sta, kot je vidno iz spodnjih primerov, sama na sebi nerelevantna za nadaljnjo obravnavo.

SOZET VP TEMELJITO	pogostnost v korpusu
vodo in temeljito	19
skledo in temeljito	3
GGDSPM VP TEMELJITO	
operemo in temeljito	16
odcedimo in temeljito	8
očistimo in temeljito	6
zarežemo in temeljito	4

Tabela 151: Preostala tridelna vzorca z lemo *temeljito* in prirednim veznikom.

4 Vzorci z drugimi besednimi vrstami

V pričujočem poglavju so predstavljeni vzorci, ki so bili iz analize, kakršna je prikazana v prejšnjih poglavjih, vnaprej izločeni na podlagi vsebovanja katere od besednih vrst ali oblik, ki je za luščenje predvidoma problematična, npr. členek, števniki, okrajšava itd (glej IV-2.3.2). Problematičnost se izkazuje na ravni (I) predvidenega **nezadostnega priklica** želenih leksikalnih podatkov z opisano metodo (kar naj bi v vseh primerih odpravilo luščenje skladiščno označenih podatkov, ko bo skladiščno označen korpus za slovenščino pripravljen) ali pa (II) **slabe avtomatske pripisljivosti** obstoječih oblikoskladiščnih kategorij korpusnim besedilom – sem sodijo, kot bo vidno v nadaljevanju, npr. problemi označevanja členkov ali okrajšav.

Izhodišče raziskave na tem mestu je torej vnaprej privzeta nerelevantnost nabora določenih skupin vzorcev, pri čemer je za vse predstavljene primere nujen poudarek, da izločitev vzorcev v pričujoči raziskavi predstavlja zgolj začasno metodološko odločitev, ki mora biti v nadaljevanju potrjena ali ovržena na osnovi natančnejših analiz.

Struktura sledečih poglavij je podobna: po kratki predstavitvi problematike sledi tabela s primeri za najpogostnejše tridelne vzorce za vse 4 obravnavane leme, poleg tega pa graf, ki prikazuje število vzorcev obravnavanega tipa glede na število vseh tridelnih vzorcev (podatki za dvodelne vzorce so manj zanimivi, zato so bili za potrebe ponazoritve izbrani le tridelni). Zadnje poglavje prinaša združene podatke o številu na opisani način vnaprej odstranjenih vzorcev, in sicer tako za dvodelne kot za tridelne vzorce.

4.1 Vzorci z oznako za pomožni glagol

Na prvem mestu med vzorci, ki predvidoma ne prinašajo za vključitev v leksikalno zbirko zanimivih rezultatov, je nabor vzorcev, v katerih se pojavlja oblikoskladiščna oznaka za različne oblike pomožnega glagola. Poimenovanje »pomožni glagol« izvirja na tem mestu iz ustroja označevalnega sistema, kjer se na prvi ravni ločijo t. i. »glavne« glagolske oblike od »pomožnih« (glej Priloga 1). Oblike, ki jih slednje oznake pokrivajo, so sicer različne, vendar gre večinoma za nabor oblik glagola *biti*. V izogib poimenovalni zmedi je v nadaljevanju v celoti naveden seznam oznak, o katerih bo v pričujočem poglavju govora, skupaj s primeri oblik ter ustrežajočimi lemmi.

oznaka	primer oblike/leme	Gp-d-ez	bila/bit, bíla/bit, bílá/bit, bílá/bit
Gp-n	bíti/bit, biti, bíti/bit	Gp-d-es	bílo/bit, bíló/bit, bilo/bit
Gp-m	bít/bit, bit/bit	Gp-d-mm	bíli/bit, bili/bit
Gp-d-em	bíl/bit, bil/bit, bíl/bit	Gp-d-mz	bíle/bit, bílé/bit, bile/bit

Gp-d-ms	bila/bití, bíla/bití	Gp-ppe-n	bom/bití
Gp-d-dm	bíla/bití, bílá/bití, bila/bití, bílá/bití	Gp-ppe-d	nebom/bití, nébom/bití
Gp-d-dz	bili/bití	Gp-ppm-n	bomo/bití
Gp-d-ds	bíli/bití, bili/bití	Gp-ppm-d	nebomo/bití
Gp-spe-n	sém/bití, sëm/bití, sem/bití	Gp-ppd	bova/bití
Gp-spe-d	nisem/bití	Gp-ppd-n	bova/bití
Gp-spm-n	smo/bití	Gp-ppd-d	(nebova/bití)
Gp-spm-d	nismo/bití, nesmo/bití	Gp-pde-n	boš/bití
Gp-spd-n	sva/bití, sma/bití	Gp-pde-d	neboš/bití
Gp-spd-d	nisva/bití, nisma/bití, nesva/bití	Gp-pdm-n	boste/bití
Gp-spdzn	sve/bití	Gp-pdd-n	bosta/bití
Gp-sde-n	sí/bití, si/bití, si/bití	Gp-pte-n	bó/bití, bo/bití, bô/bití
Gp-sde-d	nisí/bití, nisi/bití	Gp-pte-d	nebo/bití
Gp-sdm-n	ste/bití	Gp-ptm-n	bojo/bití, bodo/bití
Gp-sdm-d	niste/bití	Gp-ptm-d	nebojo/bití, nebodo/bití
Gp-sdd-n	sta/bití	Gp-ptd-n	bosta/bití
Gp-sdd-d	nista/bití, nesta/bití	Gp-g	bi/bití, bî/bití, bí/bití
Gp-ste-n	jé/bití, jè/bití, je/bití	Gp-g---d	nebi/bití
Gp-ste-d	ní/bití, ni/bití	Gp-vpm	bodimo/bití, bodímo/bití
Gp-stm-n	só/bití, so/bití, sò/bití	Gp-vpd	bodiva/bití
Gp-stm-d	niso/bití, neso/bití, nesó/bití	Gp-vde	bodi/bití, bódi/bití
Gp-std-n	sta/bití	Gp-vdm	bodite/bití
Gp-std-d	nista/bití, neso/bití	Gp-vdd	bodita/bití

Tabela 152: Nabor vseh možnih oblikoskladenjskih oznak za t. i. pomožne glagolske oblike.

Pri metodi, ki jo opisuje pričujoče delo, nizi, v katerih se pojavlja katera od obravnavanih oblik, predvidoma ne izpričujejo relevantnega doprinosa na ravni leksikalnih podatkov. Pogosto so za vnos v leksikalno zbirko zanimivejši krajši besednih nizi, na ravni vzorčnega tipa vsebujoči oznako le za glavni glagol: prim. denimo vzorčno zapolnitev *je odpeljal s pajkom* ter *odpeljal s pajkom*, ki se pri luščenju preoblikuje v *odpeljati s pajkom*.

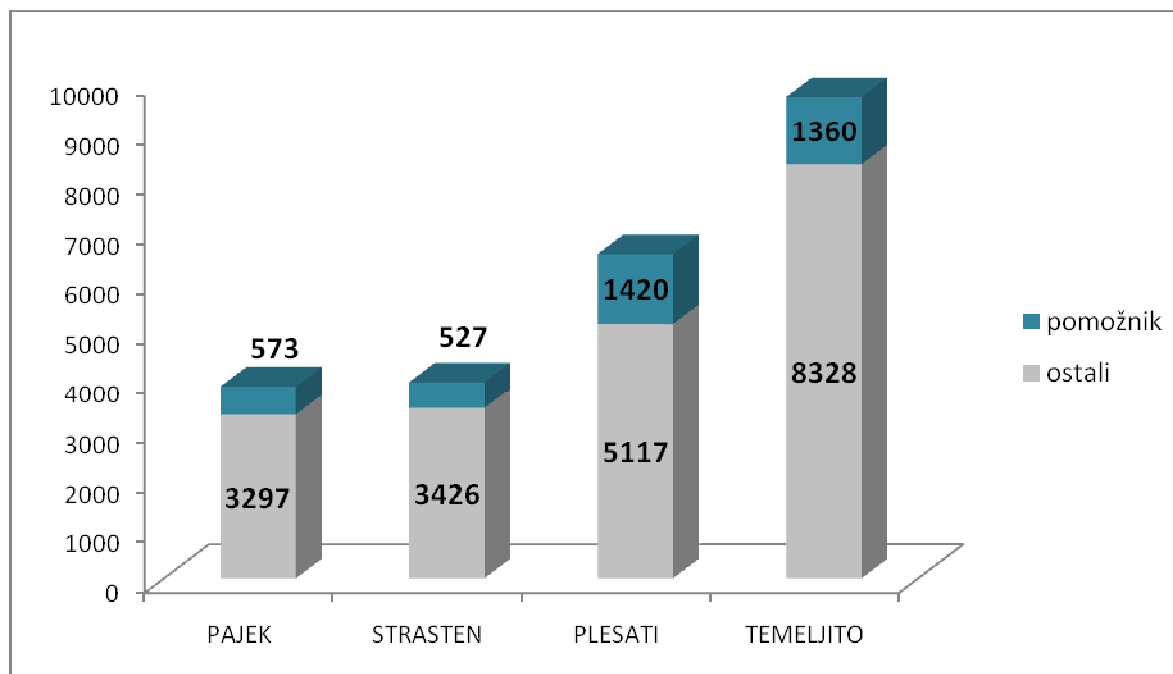
Pred nadaljevanjem so za primer navedene zapolnitve za najpogostejše izpričane tridelne vzorce za vsako od obravnavanih lem.

GPXSTEXN	GGDDXEM	PAJEK	102
je odpeljal pajek			87
GPXSTEXN	STRASTEN	SOMEI	197
je strasten ljubitelj			16
VD	GPXSTEXN	PLESATI	130
ki je plesala			29
GPXSTEXN	RNN	TEMELJITO	408
je treba temeljito			189

Tabela 153: Najpogostejši tridelni vzorci z oblikoskladenjsko oznako za pomožni glagol ter obravnavano lemo.

V gornjih primerih sicer vedno nastopa enaka oblikoskladenjska oznaka, tj. GPXSTEXN, ker je označena oblika *je* v jeziku izredno pogosta. Kljub temu primeri dovolj nazorno kažejo, da z izločitvijo vzorcev s pomožniškimi oblikami – pod pogojem, da je analiza opravljena na drugih mestih – leksikalni podatki niso izgubljeni. Dodaten argument za izpuščanje opisanega tipa vzorcev iz nadaljnje obravnave je že večkrat izpostavljeno dejstvo, da za slovenščino luščenje zaporednih besednih nizov, posebej na ravni obravnave glagola, ne more dajati dovolj zanesljivih rezultatov in je podatke nujno treba kombinirati s pridobljenimi iz skladenjsko označenih besedil.

Z odločitvijo, da se vzorci obravnavanega tipa vnaprej izločijo iz obravnave, se seznam potencialno zanimivih vzorcev precej skrajša. Spodnji graf prikazuje število vzorcev, vsebujočih eno od lem ter vsaj eno oblikoskladenjsko oznako za pomožni glagol (v tridelnih vzorcih se lahko pojavi dvakrat), in sicer v primerjavi s celotnim naborom potencialno zanimivih vzorcev. Kot je razvidno iz spodnjega grafičnega prikaza, podatki potrjujejo predvidevanje, da vzorci z oznako za pomožnik predstavljajo nekoliko večji delež v naboru vzorcev z glagolsko oz. prislovno lemo kot pa pri vzorcih s samostalniško oz. pridevniško lemo.



Graf 5: Število vzorcev, izločenih zaradi vsebovanja oblikoskladenjske oznake za pomožni glagol.

4.2 Vzorci z oznako za členek

Analiza primerov avtomatsko označenih besedil ter izkušnje tako z (I) nadgradnjo sistema za oblikoskladenjsko označevanje kot tudi (II) razvijanjem sistema za skladenjsko označevanje slovenščine¹¹⁷ pričajo o mnogih težavah, vezanih na poskus avtomatskega pripisovanja oznake za členek na oblikoskladenjski označevalni ravni. Problemi se kažejo pri označevanju členka kot besedne vrste, saj je pri določanju členka v besedilu večinoma potrebno upoštevanje skladenjske okolice besedila, prepoznava pa poteka tudi s pomočjo semantičnih kazalcev (prim. kriterija izpustljivosti iz besedila, ter nezmožnosti uporabe vprašalnice). Problem označevanja členkov je torej problem njihovega ločevanja od »enakozvočnih prislovov in veznikov« (Toporišič 2004⁴: 445).

Označevalna praksa se problema loteva na različne načine, pogosto leksikonsko. To pomeni, da so posamezne besede opredeljene za členke v oblikoslovnem leksikonu, ki je osnova za označevanje – kar ponovno predstavlja poskus določitve na ravni besede same. Na ravni leksikona je mogoče uspešno določiti predvsem nabor členkov, ki v rabi nimajo enakopisnih oblik s prislovi ali vezniki (npr. *že, še, le* ipd.). Na tem mestu je potrebno omeniti, da je bila kot rešitev problema v sklopu nadgradnje oblikoskladenjskega označevanja predlagana, vendar zaenkrat v praksi še nepreverjena, ideja združitve členkov ter enakopisnih prislovov na ravni označevanja oblikoskladenjske pod skupno kategorijo, posledično pa njihovo ločevanje šele na ravni označevanja skladnje.

¹¹⁷ Oboje je potekalo v okviru projekta Jezikoslovno označevanje slovenščine. Tematika problema avtomatskega označevanja členkov v znanstveni literaturi trenutno še ni predstavljena, zato na tem mestu trditve niso podprepljene s številčnimi podatki.

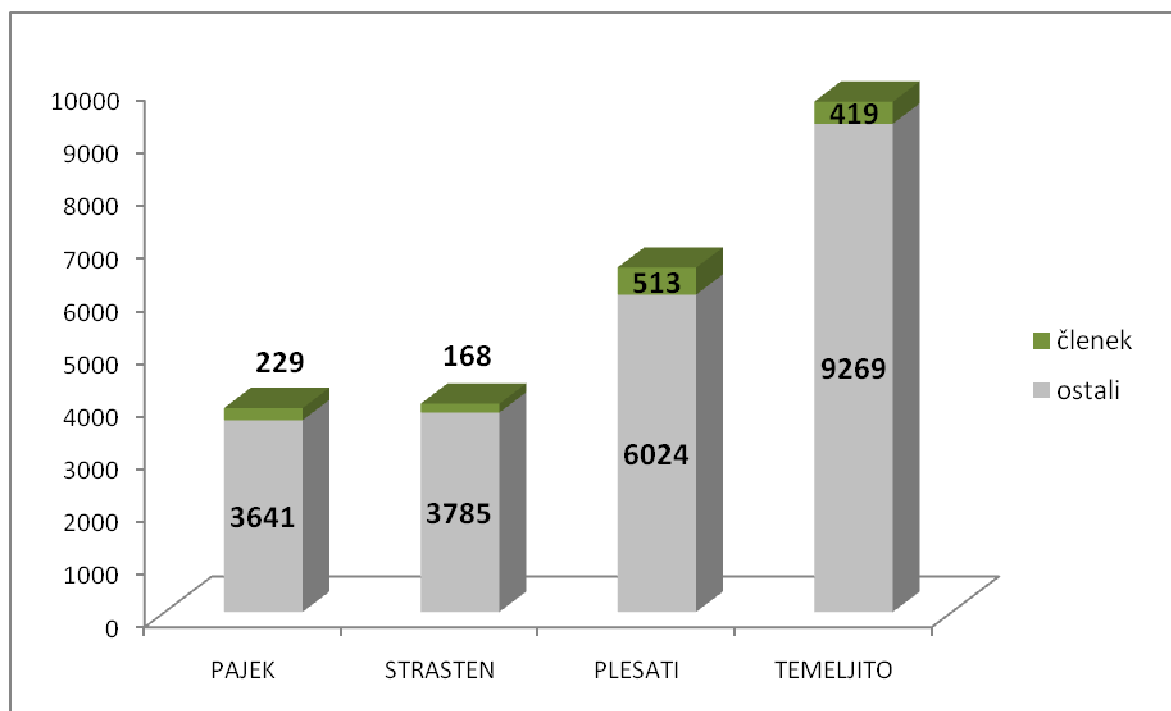
Členki trenutno v korpusu torej niso (in ne morejo biti) označeni na zadovoljivi ravni, kar utemeljuje odločitev za vnaprejšnjo izločitev vzorcev, v katerih se oznaka za članek (oznaka je ena sama, tj. *L*) pojavlja. Dodaten argument je predvidena nizka relevantnost nizov s členki za vključitev v leksikalno zbirko. Spodaj navedeni primeri zvez za najpogostejše tridelne vzorce pričajo v prid navedenim predpostavkam.

PAJEK	L	RNN	25
pajek še nikoli			3
L	STRASTEN	SOMEI	101
tudi strasten lovec			9
L	RNN	PLESATI	147
še vedno plesati			36
L	RNN	TEMELJITO	355
še enkrat temeljito			78

Tabela 154: Najpogostejši tridelni vzorci z oblikoskladenjsko oznako za članek ter obravnavano lemo.

Kljub povedanemu pa (za razliko od pomožnih glagolov v prejšnjem poglavju) besede, trenutno označene za članke, same na sebi zahtevajo natančnejšo kolokacijsko in koligacijsko analizo. Eden od pričakovanih in želenih rezultatov je identifikacija besednih zvez, ki modificirajo stavke oz. stavčne dele in morda poskus (avtomatski obdelavi slovenščine namenjene) kategorizacije slednjih. Pričakovano je, da bo velik doprinos na ravni pridobivanja leksikogramatičnih podatkov o členkih prinesla analiza skladenjsko označenih besedil.

Spodnji graf prinaša podatke o številu vzorcev, ki vsebujejo oblikoskladenjsko oznako za članek glede na število vseh tridelnih vzorcev za vse štiri obravnavane leme. Deleži izločenih vzorcev so pri vseh štirih primerih primerljivi.



Graf 6: Število vzorcev, izločenih zaradi vsebovanja oblikoskladenjske oznake za članek.

4.3 Vzorci z oznako za okrajšavo

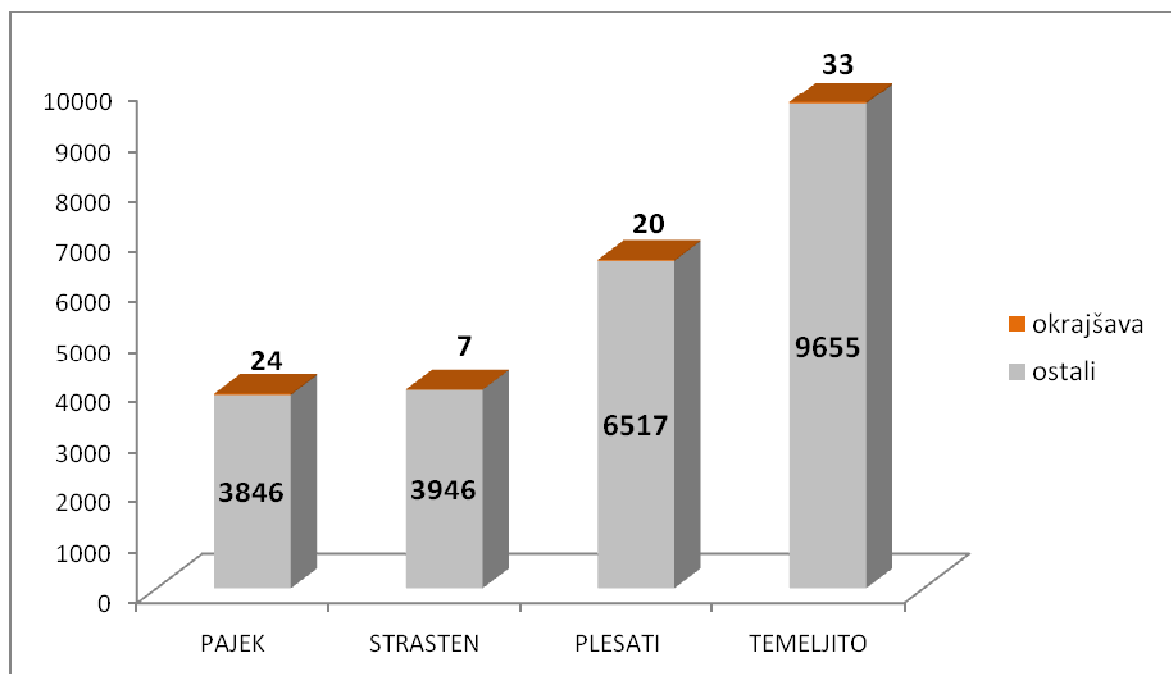
Naslednja skupina vzorcev, ki so bili evidentirani kot nezanimivi za luščenje glede na vsebnost določene oblikoskladenjske oznake, so vzorci z oznako za okrajšavo (oznaka je ena sama, tj. *O*). Graf 7 prikazuje, da je število vzorcev, v katerih se ta oznaka pojavlja, relativno redko; kot je razvidno iz spodnje tabele s primeri zapolnitev, pa so vzorci (razen pri lemi *pajek*) redki tudi glede pogostnosti v korpusu:

PAJEK	KAG	O	86
pajek 5 m			12
O	O	STRASTEN	3
t. i. strasten			1
PLESATI	VD	O	2
plesati kot npr.			2
KAG	O	TEMELJITO	8
100 g temeljito			3

Tabela 155: Najpogostejši tridelni vzorci z oblikoskladenjsko oznako za okrajšavo ter obravnavano lemo.

Pregled označevanja okrajšav priča o možnosti izboljšav, za natančnejšo opredelitev ter izvedbo katerih pa je potrebna specializirana raziskava. Trenutno so v korpusu FidaPLUS z oznako *O* označene nekatere okrajšave oz. kratice (*SIT*, *itd.*, *t. i.* ...), merske enote (*m*, *kg*, *kHz* ...), pa tudi še denimo zapis časa, kadar nastopa s piko (23.15). Pri novejšem označevanju je slednji problem sicer že odpravljen: označevanje poteka po novem z oznako za števnike (glej V-4.5).

Sledi graf s podatki o številu tridelnih vzorcev, ki so odstranjeni s seznama potencialno relevantnih zaradi vsebovanja oznake za okrajšavo.



Graf 7: Število vzorcev, izločenih zaradi vsebovanja oblikoskladenjske oznake za okrajšavo.

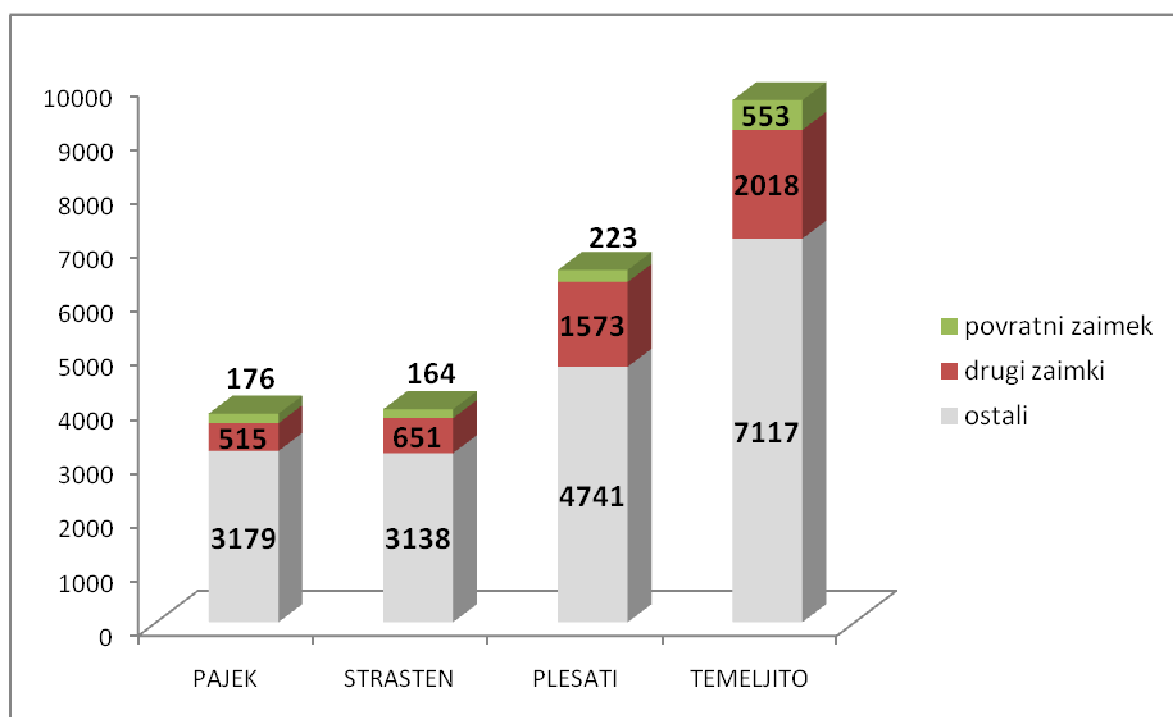
4.4 Vzorci z oznako za zaimek

Oblikoskladenjske oznake za zaimke so po trenutnem označevalnem sistemu z 9 označevalnimi kategorijami (od katerih prva prinaša 9 zaimkovnih skupin) poleg glagolskih najbolj kompleksne (glej Priloga 1). Pregled zaimkovnih skupin in vzorcev z njimi kaže, da so za luščenje leksikalnih podatkov potencialno zanimivi vzorci z oznako za povratni zaimek. Luščenja teh zvez pričujoča raziskava ne prinaša. O relevantnosti preostalih skupin bi bile potrebne dodatne raziskave, na tem mestu je privzeto, da ne prinašajo za nadaljnjo obravnavo relevantnih besednih zvez. Sledi nekaj primerov zvez z zaimkovno oblikoskladenjsko oznako:

PAJEK	ZPXXXXXXK	GPXSTEXN	17
pajek se je			10
ZPXXXXXXK	GGDSTE	STRASTEN	23
se razvije strastna			7
PLESATI	VP	ZPXXXXXXK	222
plesati in se			202
ZPXXXXXXK	GPXSTEXN	TEMELJITO	164
se je temeljito			157

Tabela 156: Najpogostejši tridelni vzorci z oblikoskladenjsko oznako za zaimek ter obravnavano lemo.

Ker vzorci za vse štiri obravnavane leme najpogosteje prinašajo oznako za povratni zaimek (ob oznakah za osebne zaimke, ki so prav tako med najbolj pričakovanimi) in ker je zanjo zaželen natančnejša obravnava v nadaljnjih raziskavah, je ta kategorija v spodnjem grafu prikazana ločeno.



Graf 8: Število vzorcev, izločenih zaradi vsebovanja oblikoskladenjske oznake za zaimek.

4.5 Vzorci z oznako za števniki

Tudi nabor vzorcev, ki vsebujejo oznako za števniki (rimski, arabski ali besedni), je v pričujoči raziskavi privzet za nerelevantnega za luščenje, pri čemer so argumenti za odločitev podobni kot v prejšnjih poglavjih. Obenem je

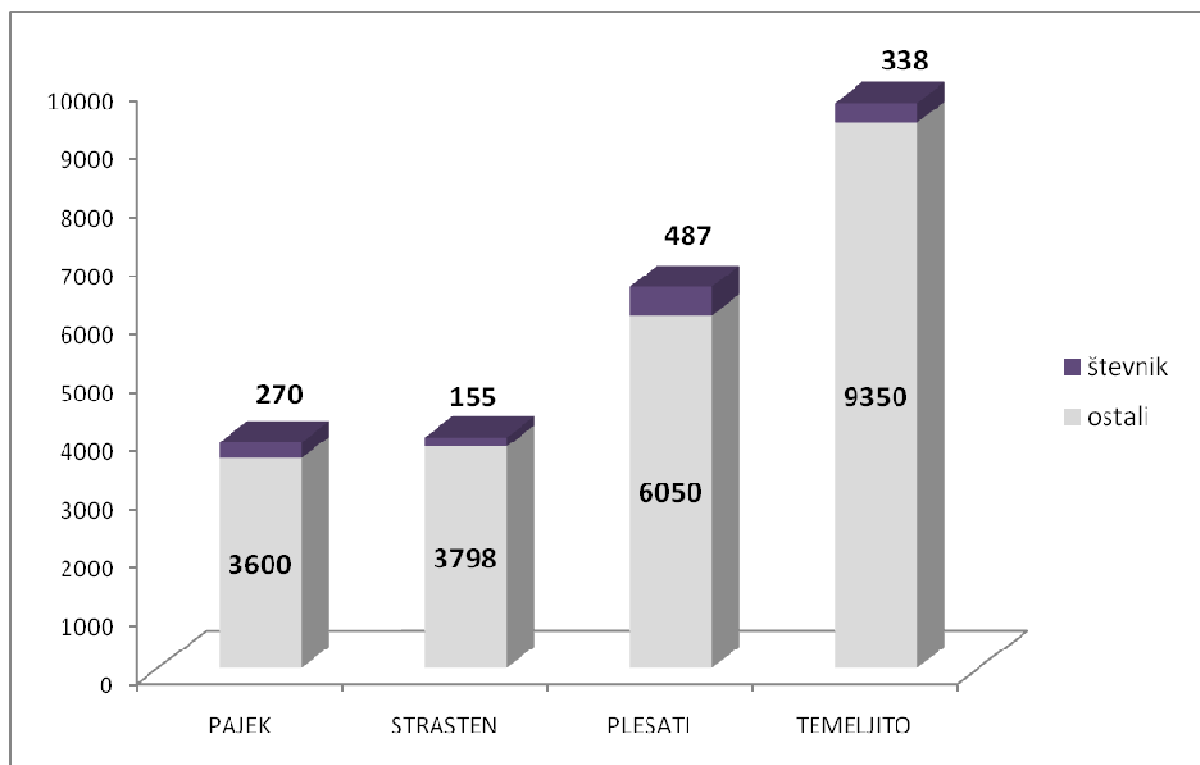
opozorilo, da so v zvezi z obravnavano besedno vrsto zaželeno nadaljnje raziskave, na tem mestu še toliko nujnejše, saj je kvaliteta luščenja besednih nizov s števniki manj vezana na upoštevanje skladišnih informacij oz. izboljšavo kvalitete označevanja kot v prejšnjih primerih. Spodnja tabela prinaša primere zapolnitev za obravnavane leme:

PAJEK	SOMEI	KAG	347
pajak SIP 230			99
KAG	STRASTEN	SOSMI	14
22.30 Strastna razmerja			9
KBVMEI	VP	PLESATI	29
peti in plesati			29
SOSER	KAG	TEMELJITO	42
leta 1994 temeljito			4

Tabela 157: Najpogostejši tridelni vzorci z oblikoskladišjsko oznako za okrajšavo ter obravnavano lemo.

Za razliko od večine prej predstavljenih primerov gornje zapolnitve izkazujejo potencialno relevantnost na ravni tridelnih (*pajak SIP 230*) ali dvodelnih (*leta 1994*) vzorcev (ne pa tudi v primeru označevanja časa oz. napačno lematiziranega *peti in plesati*).

Sledi še graf s podatki o številu tridelnih vzorcev, ki so odstranjeni s seznama potencialno relevantnih zaradi vsebovanja oznake za števniki.

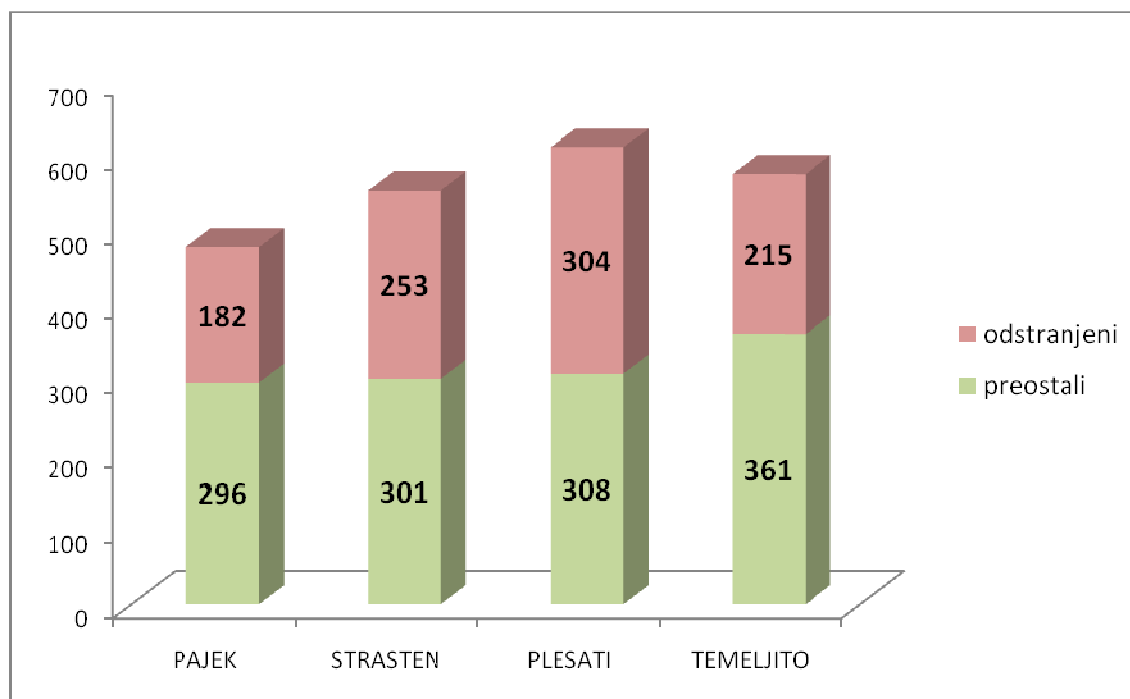


Graf 9: Število vzorcev, izločenih zaradi vsebovanja oblikoskladišjske oznake za števniki.

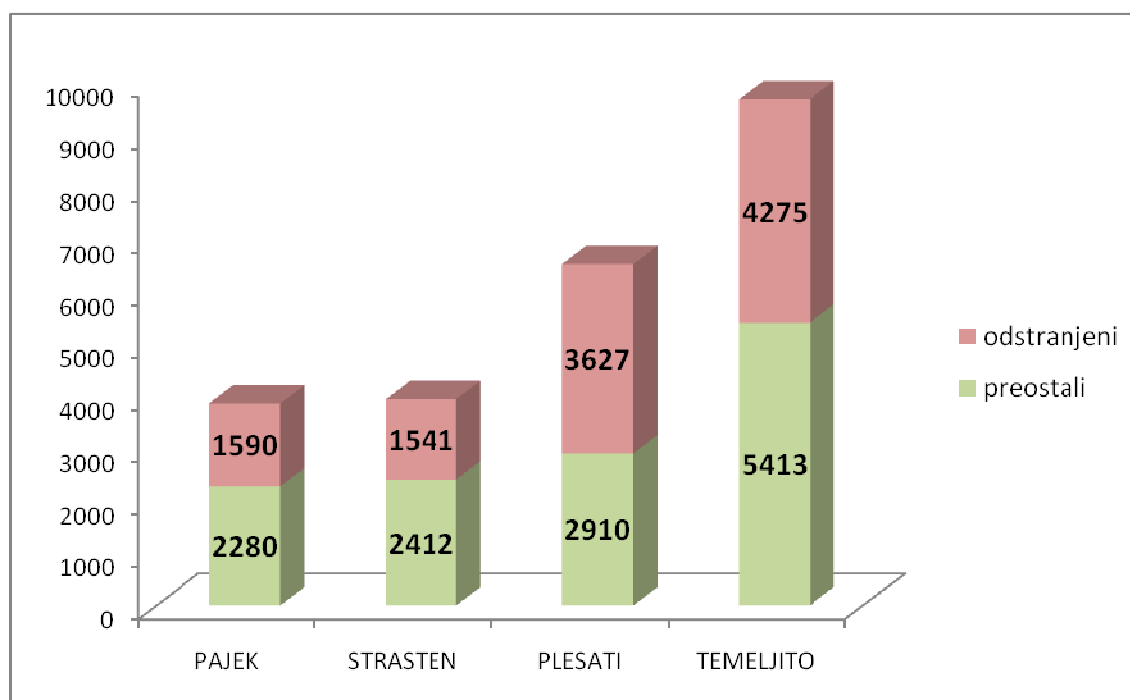
4.6 Delež vnaprej odstranjenih vzorcev

Kot so nakazovali že grafi v predhodnih poglavjih, odločitev za vnaprejšnje odstranjevanje naštetih skupin vzorcev precej skrajša seznam za luščenje potencialno relevantnih vzorcev. Za selekcijo seznama je bil uporabljen specializiran program (opis programa v IV-3.2.1), ki iz celotnega seznama vzorcev v ločeno datoteko

prepiše samo tiste, ki ne spadajo v nobeno od problematičnih skupin. Grafa spodaj prikazujeta podatke po selekciji dvodelnih ter tridelnih vzorcev¹¹⁸.



Graf 10: Število dvodelnih vzorcev, izločenih zaradi vsebovanja katere od problematičnih oblikoskladenjskih oznak.



Graf 11: Število tridelnih vzorcev, izločenih zaradi vsebovanja katere od problematičnih oblikoskladenjskih oznak.

¹¹⁸ Pri tridelnih vzorcih se pogosto dogaja, da en vzorec vsebuje oznaki dveh različnih problematičnih skupin, zato so v grafu vsi izločeni vzorci združeni v eno samo skupino.

Podatki kažejo, da je največji delež odstranjenih vzorcev glede na preostale v naboru z glagolsko lemo (pri tridelnih vzorcih z lemo *plesati* je delež odstranjenih celo višji od deleža preostalih). Grafe je zaradi očitnih razlik med vzorci, ki jih izkazujejo leme različnih besednih vrst, predvsem pa zaradi različne členjenosti ter razpršenosti oblikoskladenjskih oznak, sicer težko primerjati; ključna na tem mestu je zato le primerjava odstranjenega ter preostalega deleža pri vsakem posameznem grafičnem stolpcu.

5 Vzorci z nelematiziranimi besedami

Posebej dragocene podatke za izboljšavo avtomatskega označevanja prinaša nabor vzorcev z oblikoskladenjsko oznako za nelematizirano besedo (oznaka *N*). Trenutna praksa razvoja označevanja nelematiziranim pojavnicam že posveča zadostno pozornost: ob vsakem avtomatskem označevanju se vzporedno izdeluje seznam oblik, ki jim leme ni bilo mogoče pripisati na osnovi podatkov iz leksikalne zbirke.

V tem delu se zato z nelematiziranimi pojavnicami ne ukvarjamo podrobno, v nadaljevanju so bolj ali manj kot zanimivost navedeni primeri zapolnitev za dvodelne vzorce z nelematizirano obliko. Vsi vzorci, ki vsebujejo oznako za nelematizirano besedo, so bili vnaprej izločeni iz nadaljnje analize luščenja, saj zahtevajo v prvi vrsti analizo s stališča razvoja označevalnega sistema in ne s stališča vnosa podatkov v leksikalno zbirko.

Kot bo vidno iz primerov v nadaljevanju, velik delež nelematiziranih besed odpade na lastna imena (*strasten Dionis*, *Clug pleše*), tuje besede (*pajek umbrella*), zatipkane (*strastni bibliofli*, *manjdlje temeljito*) oz. neustrezno tokenizirane (*9.30Mož pajek*) besede, neknjižne besede (*jebeni pajek*, *strasten snifač*, *tamlada plešeta*, *temeljito zluftati*), dosti pa je tudi zvez oz. besed, ki niso označene, ker v leksikalno zbirko do sedaj še niso bile vključene, npr. primeri tipa *osmeronogi pajek*, *strastna golfistka*, *plesati sirtaki*, *temeljito prediskutirati*. Posebna pozornost mora biti seveda posvečena primerom, ki se v korpusu pojavljajo z visoko pogostnostjo, v obravnavanih primerih npr. zveze, ki poimenujejo znamko delovnih strojev – *pajek* [*Deutz*, *Fahr*, *Claas*, *Fella*] ali vrsto plesov – *plesati* [*sirtaki*, *ringaraja*, *hiphop*, *rock'n'roll*].

5.1 Pajek

nelem. + pajek		pog. v korpusu	ov	pajek	
9.30Mož	pajek	10	Predno	pajek	1
10.30Mož	pajek	6	xxxxxxxx	pajek	1
8.00Mož	pajek	5	Čedomir	pajek	1
š	pajek	3	xxxxxxxxxxxx	pajek	1
20.00Zlati	pajek	2	AZlati	pajek	1
xxxxxxxxxxx	pajek	2	jebeni	pajek	1
n	pajek	1	B	pajek	1
osmeronogi	pajek	1	Zlatopredi	pajek	1
trikraki	pajek	1	Najdi.si	pajek	1
9.00Mož	pajek	1	osatega	pajek	1
štiriredni	pajek	1	rdečehrbti	pajek	1
kg12	pajek	1	dvorepi	pajek	1
zlatopredi	pajek	1			

Tabela 158: N + pajek

pajek + nelem.		pog. v korpusu	pajek	daros	
pajek	deutz	72	pajek	kuhn	5
pajek	fahr	69	pajek	marangon	4
pajek	claas	21	pajek	kverneland	3
pajek	fella	12	pajek	disderid	2
pajek	Fahr	7	pajek	pajkljo	2
pajek	stoll	6	pajek	dvo	2

pajek	Spinny	1	pajek	Latrodectus	1
pajek	umbrella	1	pajek	pletetpod	1
pajek	haublitz	1	pajek	Jugoslavu	1
pajek	Argiope	1	pajek	Nephila	1
pajek	Lunajek	1	pajek	Ananseju	1
pajek	Heriberta	1	pajek	tancajo	1
pajek	fontanesi	1	pajek	kg12	1
pajek	Fonzi	1	pajek	Tutaja	1
pajek	Bristowe	1	pajek	zmasakriral	1
pajek	Phoneutrias	1	pajek	Lunajkom	1
pajek	class	1	pajek	kimura	1
pajek	Troglohyphantes	1	pajek	Deutz	1
pajek	Pisaura	1	pajek	profarm	1
pajek	analysis	1	pajek	Claas	1
pajek	zapredala	1	pajek	zapreda	1
pajek	Philaesus	1			

Tabela 159: *pajek* + N

5.2 Strasten

nelem. + strasten		pog. v korpusu	Gerard	strasten	1
Dipendra	strasten	2	Issigonis	strasten	1
Carolus	strasten	1	ovi	strasten	1
Mayfield	strasten	1	Bruggeju	strasten	1
Gaby	strasten	1	detektirajo	strasten	1
razbuhtele	strasten	1	Scarlett	strasten	1
Runfeldt	strasten	1	Born	strasten	1
Howard	strasten	1	Fried	strasten	1
Senneu	strasten	1	Nati	strasten	1
Enrique	strasten	1			

Tabela 160: N + *strasten*

strasten + nelem.		pog. v korpusu	strastne	puhače	1
strastna	žurerka	2	strastnega	sektaša	1
strasten	Dionis	2	strastni	konzumenti	1
strastna	Reggie	2	strastni	Theo	1
strastna	Ilse	2	strastnega	vlekača	1
strastna	Jacinta	2	strastnih	senjorit	1
strastne	Toplečke	2	strastna	golfistka	1
strastna	filmoljubka	2	strastna	rajanka	1
strasten	filmočil	2	strasten	gleduh	1
strasten	tenisač	2	strastnim	snifačem	1
strastnima	seksoma	1	strastna	Carmen	1
strastnega	prdača	1	strastna	Ninočka	1
strastni	smu	1	strastni	filmofili	1
strastna	Manuela	1	strastna	potapljačka	1
strastnih	seksih	1	strastni	antipatikus	1
strasten	espressivo	1	strastna	Nicole	1
strastna	trekerka	1	strastna	Winona	1
strastni	zalizavi	1	strastni	školjkoljubi	1
strastna	modofilka	1	strastni	bibliofli	1
strastna	brezglavka	1	strastni	Claudii	1
strastni	športaš	1			

Tabela 161: *strasten* + N

5.3 Plesati

nelem. + plesati		pog. v korpusu	lkedo	plesal	1
diktirko	plešejo	2	dixielanda	plesalo	1
Guillem	plesati	2	Satchmo	plesali	1
lindžo	plesati	2	Dickie	plesal	1
Break	pleše	2	Jumbo	plesal	1
Petruchia	plesal	2	Breni	plesale	1
Gere	pleše	2	debelajsi	plešejo	1
Clug	pleše	2	alumni	plesali	1
pofukelj	pleše	2	matiasi	plešeta	1
Flip	plešejo	2	hiphop	pleše	1
Jumbo	plešeta	1	Jesseja	plesati	1
Emily	plesala	1	Breakbytu	plesalo	1
mamilaši	plešejo	1	Gerrard	pleše	1
plise	plešejo	1	Duvall	plešeta	1
Bakhanti	plešejo	1	Hopsi	plesali	1
Folkarta	plesale	1	Randu	pleše	1
Nathlia	plesati	1	Nathalie	pleše	1
Ermalai	plesal	1	skret	pleše	1
lambado	plesali	1	singlom	Plešeš	1
Tüdi	plesali	1	Beyoncé	pleše	1
Ares	pleše	1	kombinezionu	plesal	1
Rowland	pleše	1	ognjeviteje	plesalo	1
Lenny	pleše	1	Seat	pleše	1
Swayzejem	plesala	1	Cinecitta	plesala	1
Colette	pleše	1	gumno	plesat	1
kozmičarko	plesal	1	studijih	plesali	1
nauče	plesati	1	Evaldas	pleše	1
Hazel	pleše	1	toplesu	plesale	1
singla	Plešeš	1	Elliot	plesati	1
Nicole	plesala	1	Janček	pleše	1
Nursel	pleše	1	Cluga	plešejo	1
Fergie	pleše	1	bum	pleše	1
Flashdanceu	pleše	1	Kia	pleše	1
Moga	pleše	1	Persefona	pleše	1
Beals	plesala	1	Antonyja	plesati	1
Tejada	pleše	1	Sebastianu	pleše	1
Block	plesal	1	Horsey	plesal	1
Yankeeji	plesali	1	Mačkare	plešejo	1
Cheivo	plesal	1	Maddy	pleše	1
1Je	plesati	1	Traberga	plesale	1
Travolta	plesal	1	Y	pleše	1
Giselle	pleše	1	Smrkljuh	pleše	1
Nialle	pleše	1	Portella	plesala	1
Rosario	pleše	1	Lemmon	plešeta	1
Carabosse	plešejo	1	Paris	plesala	1
pomponi	plesale	1	Maradona	pleše	1
Molly	pleše	1	Coulthard	plesal	1
Lentu	plešejo	1	Gerryjem	plešeta	1
nejevernežem	pleše	1	Rothbarta	plesal	1
Piggy	plesal	1	Neo	plešejo	1

Cityja	plesalo	1	Koora	plešem	1
Madonna	plešeta	1	Blanche	pleše	1
Versaceja	plesal	1	Laibach	pleše	1
Momix	plešejo	1	larfami	plesale	1
Lenčo	plešeta	1	Rabadi	pleše	1
miss	plesale	1	podstati	pleše	1
Yankeeji	plešejo	1	tamlada	plešeta	1
feuerhake	pleše	1	zalugi	pleše	1
Muravidek	pleše	1	Cor	pleše	1
Illmann	plesal	1	Wooja	pleše	1
Bolšoju	plesala	1	Srhoju	plešeta	1
imBodo	plesali	1	Plisecka	pleše	1
Janeiru	plesalo	1	boogie	plesali	1
Adrianna	plešeta	1	housa	plešete	1
Maiykesch	plesali	1	Vadžpaji	pleše	1
asereje	plesali	1	Lacrimasu	plešejo	1
tres	plešejo	1			

Tabela 162: N + *plesati*

plesati + nelem.	pog.	plesati		
plesati break	8	plesati Zorbo		1
plesati sirtaki	8	plesati Rachel		1
plesati sa	7	plesati ballo		1
plesati pogo	7	plesati Mahmud		1
plesati ringaraja	6	plesati butoh		1
plesati hiphop	6	plesati kfersibodi		1
plesati rock'n'roll	4	plesati tanguerosi		1
plesati lala	3	plesati nonstop		1
plesati charleston	3	plesati Elena		1
plesati limbo	3	plesati jive		1
plesati džezbalet	3	plesati Muuji		1
plesati Giselle	3	plesati Sergiu		1
plesati lambado	2	plesati capoeiro		1
plesati boogie	2	plesati Sylvie		1
plesati rap	2	plesati dapke		1
plesati blues	2	plesati papigpolko		1
plesati Georgeta	2	plesati Margot		1
plesati bossa	2	plesati merengue		1
plesati ino	2	plesati kathak		1
plesati polk'n'roll	2	plesati zravn		1
plesati četvorkometka10.00Visoka	1	plesati ROCK'N'ROLL		1
plesati Paquito	1	plesati Mercutia		1
plesati jazzy	1	plesati reels		1
plesati hule	1	plesati ča		1
plesati non	1	plesati Laurent		1
plesati makareno	1	plesati Coria		1
plesati zavzeteje	1	plesati Ans		1
plesati makarene	1	plesati sotiš		1
plesati jazzbalet	1	plesati Shae		1
plesati ans	1	plesati hako		1
plesati lezginko	1	plesati Abigail		1
plesati Valeria	1	plesati Edward		1
plesati komos	1	plesati T1ortolo		1
plesati tai	1	plesati Colasa		1
		plesati čarleston		1

plesati	bogve	1	plesati	hulo	1
plesati	mapuko	1	plesati	Capuero	1
plesati	zoprnn	1	plesati	bharata	1
plesati	kalinko	1	plesati	Gogo	1
plesati	Dabo	1	plesati	To1rtolo	1
plesati	jalgati	1	plesati	Fracki	1
plesati	poškodbeni	1	plesati	sebo	1
plesati	Auroro	1	plesati	lambeth	1
plesati	hod'la	1	plesati	Macareno	1
plesati	Gheghegč	1	plesati	raznobarvi	1

Tabela 163: *plesati* + N

5.4 Temeljito

nelem. + temeljito		pog. v korpusu			
Mannheimu	temeljito	2	Alverstead	temeljito	1
Nairobiu	temeljito	2	podolgem	temeljito	1
rally	temeljito	2	šovbiznisa	temeljito	1
Futaba	temeljito	2	Citroënu	temeljito	1
interiere	temeljito	1	Pro	temeljito	1
Ravenni	temeljito	1	brisačkami	temeljito	1
Hagiju	temeljito	1	Kott	temeljito	1
Winns	temeljito	1	Starr	temeljito	1
Lisac	temeljito	1	okpravi	temeljito	1
motion	temeljito	1	Oise	temeljito	1
sintro	temeljito	1	omogočča	temeljito	1
Eura	temeljito	1	dabi	temeljito	1
Privilege	temeljito	1	kropinom	temeljito	1
Borl	temeljito	1	Stevena	temeljito	1
sprila	temeljito	1	Wrigley	temeljito	1
Derochette	temeljito	1	ScanDisk	temeljito	1
Arrigoni	temeljito	1	međalniku	temeljito	1
Gällivare	temeljito	1	Sai	temeljito	1
Barbie	temeljito	1	corsa	temeljito	1
Alexandro	temeljito	1	stoenko	temeljito	1
maskare	temeljito	1	Lannoy	temeljito	1
Kevom	temeljito	1	Olympiakosa	temeljito	1
Mills	temeljito	1	Bebo	temeljito	1
M3Z	temeljito	1	Colorja	temeljito	1
Cometa	temeljito	1	Samsung	temeljito	1
Brea	temeljito	1	Medex	temeljito	1
kadetta	temeljito	1	exemption	temeljito	1
Viero	temeljito	1	Clemen	temeljito	1
besedoslada	temeljito	1	Dietla	temeljito	1
Gulikersom	temeljito	1	Powerfarm	temeljito	1
vontone	temeljito	1	Perfekthionom	temeljito	1
Larix	temeljito	1	Lotus	temeljito	1
Spohr	temeljito	1	Windows	temeljito	1
Dolfe	temeljito	1	Ošskozi	temeljito	1
Nicholasa	temeljito	1	časo	temeljito	1
solaxu	temeljito	1	Egkh	temeljito	1
Luckyju	temeljito	1	Vasilisa	temeljito	1
Watergate	temeljito	1	priporočivo	temeljito	1
manjdlje	temeljito	1	Alcatel	temeljito	1
			Benecol	temeljito	1

Phare	temeljito	1	Pustoljub	temeljito	1
Oraclu	temeljito	1			

Tabela 164: N + temeljito

temeljito + nelem.		pog. v korpusu			
temeljito	preverbo	6	temeljito	upel	1
temeljito	popudrajte	4	temeljito	spljuval	1
temeljito	prediskutirala	3	temeljito	preliže	1
temeljito	prediskutiral	2	temeljito	preiskali'	1
temeljito	sprevedrilo	2	temeljito	prerajala	1
temeljito	predebatirala	2	temeljito	razsikali	1
temeljito	prediskutirajmo	2	temeljito	posračkali	1
temeljito	prediskutirati	2	temeljito	nakuril	1
temeljito	natrenirati	2	temeljito	predebatirali	1
temeljito	ošvrkal	2	temeljito	preu	1
temeljito	pre	2	temeljito	odljubi	1
temeljito	preanalizirati	2	temeljito	vmasirajte	1
temeljito	prediskutirali	2	temeljito	počedili	1
temeljito	preizprašal	2	temeljito	predebatira	1
temeljito	stestirajo	1	temeljito	premesimo	1
temeljito	zabaviti	1	temeljito	odcede	1
temeljito	preoučiti	1	temeljito	preadaptiral	1
temeljito	prejedrila	1	temeljito	popudrati	1
temeljito	videoanalizo	1	temeljito	oplevite	1
temeljito	spoliramo	1	temeljito	bio	1
temeljito	brifirajo	1	temeljito	preanaliziral	1
temeljito	presprejali	1	temeljito	zamanjali	1
temeljito	prešvicala	1	temeljito	vmasiramo	1
temeljito	spričkamo	1	temeljito	preredčimo	1
temeljito	prizkušena	1	temeljito	zatrpeti	1
temeljito	presesamo	1	temeljito	zadihtal	1
temeljito	stestirali	1	temeljito	odmastimo	1
temeljito	zlufatla	1	temeljito	preškropimo	1
temeljito	premeŭaj	1	temeljito	zmeđajte	1
temeljito	pobiksala	1	temeljito	prostituirali	1
temeljito	porazgovorili	1	temeljito	demistificirala	1
temeljito	prouče	1	temeljito	okepala	1
temeljito	prefotografirali	1	temeljito	vmasirati	1
temeljito	pognetemo	1	temeljito	proučiti	1
temeljito	zrešetal	1	temeljito	zakreditiralo	1
temeljito	premesili	1	temeljito	prečasajo	1
temeljito	preokrenilo	1	temeljito	redefinirati	1
temeljito	zmeš	1	temeljito	preplužila	1
temeljito	preanalizirali	1	temeljito	inštruiral	1
temeljito	ugnetemo	1	temeljito	prebilo	1
temeljito	predebatirati	1	temeljito	zdemokratiziral	1
temeljito	zredigirati	1	temeljito	internalizirali	1
temeljito	obnovljati	1	temeljito	prelopatimo	1
temeljito	zakresala	1	temeljito	redefinirala	1
temeljito	umasirajte	1	temeljito	pretipavali	1
temeljito	prekrenilo	1	temeljito	meliorirala	1
temeljito	premečkate	1	temeljito	demistificirajo	1
temeljito	spreminiti	1	temeljito	premašamo	1
temeljito	raziskaliV	1	temeljito	oribalo	1
			temeljito	osmislitve	1

temeljito	očistiti	1	temeljito	natelovadili	1
temeljito	zrevolucioniral	1	temeljito	oplahniti	1
temeljito	naplavali	1	temeljito	predebatiral	1
temeljito	inštruira	1	temeljito	lotimoi	1
temeljito	podro	1	temeljito	zazro	1
temeljito	uračunal	1	temeljito	sprhala	1
temeljito	obdedalo	1	temeljito	nasankal	1
temeljito	poplevemo	1			

Tabela 165: temeljito + N

6 Manj pogosti vzorci z lemo pajek

Kljub temu da so v pričujoči raziskavi v središču zanimanja predvsem visokopogostni vzorci oz. vzorčni tipi, se je (predvsem tekom analize tridelnih vzorcev) pokazala potreba po natančnejšem pregledu celotnega seznama vzorcev s ciljem identifikacije vzorčnih tipov, ki v analizi do tega trenutka zaradi nižje pogostnosti niso bili obravnavani. V nadaljevanju poglavja je tovrsten postopek predstavljen na primeru nabora vzorcev z lemo *pajek*.

Za prvi korak identifikacije izpuščenih vzorčnih tipov je bil uporabljen program, ki z ustreznega seznama vzorcev odstrani vse do sedaj že evidentirane vzorce (opis programa v IV-3.4.2). Po tem postopku ostane na seznamu tako za dvodelne kot tridelne vzorce nabor primerov, ki jih je možno razvrstiti v tri skupine: (I) kombinacije obravnavane leme z oblikoskladenjskimi oznakami, (II) kombinacije leme s katero od drugih lem, obravnavanih v tej raziskavi, (III) kombinacije leme s kakim drugim znakovnim nizom. Omenjeni trije tipi so razvidni v spodnji tabeli:

vzorec	pogostnost v korpusu		
VD PAJEK	168	26 PAJEK	1
? PAJEK	96	61 PAJEK	1
PAJEK ?	64	ZDR PAJEK	1
PAJEK VD	26	PAJEK ZDR	1
MODER PAJEK	4	PAJEK PLESATI	1
PAJEK PAJEK	4	PAJEK AMP	1
GT PAJEK	2	M PAJEK	1
CDO09 PAJEK	2	PAJEK 05	1
6 PAJEK	2	PAJEK 97228	1
		PAJEK JU	1

Tabela 166: Nabor manj pogostih dvodelnih vzorcev z lemo *pajek*.

Primeri, ki tvorijo prvo navedeno skupino, tj. vzorci, ki so bili izpuščeni iz obravnave zaradi nezadostne pogostnosti, nas v nadaljevanju zanimajo tudi s stališča zapolnitev, vendar na tem mestu zgolj na ravni razvrščanja vzorčnih tipov na relevantne ter nerelevantne. Priprava podatkov je na tem mestu naslednja: (I) novi vzorci so ročno pregledani ter razvrščeni v ustrezne vzorčne tipe, na osnovi teh so (II) izluščeni nabori besednih nizov. Slednji so ponovno (III) ročno pregledani, sledi (IV) razvrstitev vzorčnega tipa na ustrezno mesto v tabeli relevantnosti za luščenje in (V) izbor ustreznega besednega niza za ponazoritev.¹¹⁹

Rezultat obdelave seznamov tako dvodelnih kot tridelnih vzorcev prinaša spodnja tabela¹²⁰, v kateri so vzorčni tipi razvrščeni na relevantne, nerelevantne ter vsebujoče glagol. Slednja skupina je ločena, ker prinaša predvidoma manj zanesljive podatke, ki pa se kljub temu pogosto izkazujejo za zanimive za vključitev v leksikalno zbirko. Vprašanje, ali oz. na kakšen način jih vključiti, na tem mestu torej puščamo graditelju zbirke.

¹¹⁹ Ker so za obravnavane vzorčne tipe izluščene besedne zveze v korpusu manj pogostne (večinoma se vsaka pojavi le enkrat), so primeri izbrani zgolj glede na ponazoritveni potencial, ne pa tudi glede na pogostnost.

¹²⁰ Za legendo zapisa vzorčnih tipov glej Priloga 3.

vzorčni tip	primeri besednega niza
relevantni za nadaljnjo obravnavo	
Sam + Prid + Sam (različni tipi zvez glede na ujemanje v sklonu)	<i>pajek črna vdova; pajek italijanske proizvodnje, razvoj rdečega pajka; zavetje drobnim pajkom; pajku podoben moški;</i>
Sam + Prisl + Prid (ujemalnost)	<i>pajek grozno nevaren</i>
Sam + Sam + Prid (ujemalnost)	<i>SIP pajek dvovretenski, pajek SIP dvovretenski, pajek človeku nevaren</i>
Sam + Prid + Prid (ujemalnost)	<i>pajek dvovretenski italijanski</i>
Pred + Sam + Sam	<i>v obliki pajka</i>
Sam + Vd + Sam	<i>pajek kot sredstvo</i>
Prid + Vd + Sam	<i>strupen kot pajek</i>
Prid + Pred + Sam	<i>nor na pajke</i>
Prisl + Pred + Sam	<i>skupaj s pajki</i>
Prisl + Vd + Sam ₁	<i>urno kot pajek</i>
Prisl + Prisl + Sam	<i>kar nekaj pajkov</i>
z glagolom	
Sam ₁ + Glag + Glag Sam _{/1} + Glag + Glag	<i>pajek začeti stopicati pajke želeti pretihotapiti</i>
Glag + Glag + Sam ₁ Glag + Glag + Sam _{/1}	<i>pričeti mastiti pajek želeti prijeti pajka</i>
Glag + Sam ₁ + Glag Glag + Sam _{/1} + Glag	<i>morati pajek doživeti smeti pajka fotografirati</i>
Sam ₁ + Glag + Sam Sam _{/1} + Glag + Sam	<i>pajek odpeljati avto; dijaki spoznavati pajke pajku odtrgati nogo; avto odpeljati pajek</i>
Sam ₁ + Prid + Glag	<i>pajek neopazen ždeti</i>
Sam + Sam + Glag (različni tipi zvez glede na sklon)	<i>pajek muho pokončati; pajek zaklopničar živeti; pajku nogo odtrgati; število pajkov zmanjšati;</i>
Glag + Sam + Prisl Glag + Prisl + Sam	<i>skočiti pajek naravnost vesti spet pajke</i>
Prisl + Sam ₁ + Glag Prisl + Sam _{/1} + Glag	<i>danes pajek odpeljati najprej pajka omamiti</i>
Glag + Vd + Sam ₁ Glag + Vd + Sam _{/1}	<i>misliti kot pajek narisati kot pajka</i>

nerelevantni za nadaljnjo obravnavo	
Vd + Sam	<i>kot pajek</i>
Sam + Vd	<i>pajek kot</i>
Sam + Pred + Prisl	<i>pajek na najbolj</i>
Sam + Sam + Prisl	<i>pajek mrežo stalno; pajek križevac pogosto, človek pajek nikoli</i>
Sam + Vd + Prisl	<i>**pajek kot nehumano</i>
Prisl + Sam + Prisl	<i>lahko pajek nemoteno</i>
Prisl + Sam + Sam	<i>torej pajek ukrep; sicer uporaba pajka</i>
Prisl + Vd + Sam _{/1}	<i>več kot pajkov</i>
Sam + Prisl + Sam	<i>kislina lahko pajek</i>
Sam + Prid + Prisl	<i>**pajek samo zato</i>
Prid + Prisl + Sam	<i>**spletene mreže pajka</i>
Sam ₁ + Glag + Prisl Sam _{/1} + Glag + Prisl	<i>pajek odvažati nepravilno pajkom izogniti tako</i>
Sam + Prid + Sam (niso vsi primeri nerelevantni)	<i>pajek ujeto žuželko</i>
Sam + Prisl + Sam	<i>pajek dnevno težave</i>
Prisl + Pred + Sam (niso vsi primeri nerelevantni)	<i>nato na pajka</i>
Pred + Sam + Sam (niso vsi primeri nerelevantni)	<i>za otroke pajek</i>
Sam + Vd + Glag	<i>pajek kot pikati</i>
Prid + Glag + Sam	<i>**celo odpeljati pajek</i>
Sam + Prisl + Prid (neujemalnost)	<i>pajek najbolj strupena</i>
Sam + Sam + Prid (neujemalnost)	<i>črni pajek bavarske, pajek ukrep kranjske</i>
Sam + Prid + Prid (neujemalnost)	<i>pajka nepravilno parkiran</i>
Sam + Vd + Prid	<i>pajek kot poskusen</i>
Sam ₁ + Glag + Prid Sam _{/1} + Glag + Prid	<i>pajek imeti poseben pajka predstaviti nemški</i>
Glag + Sam + Prid	<i>prihajati pajek črna</i>
Vd + Sam + Pred	<i>kot pajek na</i>

** – vsi izluščeni primeri izkazujejo napako na ravni oblikoskladenjske označenosti

Tabela 167: Manj pogosti vzorčni tipi z lemo *pajek*.

Kot rečeno so v gornji tabeli predstavljeni vzorčni tipi za obravnavano lemo manj tipični, kljub temu pa je razvidno, da prinašajo uporabne podatke. Podobno dopolnitev bo v nadaljevanju potrebno pripraviti tudi na ravni preostalih treh obravnavanih lem, saj je predvideno, da se bo ob tem nabor vzorčnih tipov še povečal. Na predstavljeni način pripravljena tabela pa seveda zahteva dodatno preverjanje podatkov na osnovi obravnave večjega števila lem za posamezno besedno vrsto. Vprašanju se mestoma posveča naslednje poglavje (VI-2.2).

VI

REZULTATI ANALIZE

Pričujoče poglavje je namenjeno strnjeni predstavitvi rezultatov analize, pri čemer sta v ospredje postavljeni dve glavni raziskovalni vprašanji: (I) izboljšava avtomatskega oblikoskladenjskega označevanja ter (II) določanje relevantnosti vzorčnih tipov za luščenje besednih zvez. Obema od tem ustreza eno od podpoglavij v nadaljevanju.

1 Izboljšava avtomatskega oblikoskladenjskega označevanja

Analiza avtomatske označenosti izluščenih besednih nizov v poglavju V poteka na več mestih. V nadaljevanju so strnjeno predstavljene ugotovitve, ki jih analiza prinaša.

Na ravni pripisovanja **napačne besedne vrste** oblikam v besedilu se pojavlja označevanje (I) svojilnih pridevnikov za samostalnik (*pajkov strup*), (II) funkcijske besede za samostalnik (*pajek kot po čudežu izgine*), med najpogostejšimi so primeri označevanja (III) prislova za pridevnik (*pajki mnogo predejo*) ali (IV) pridevnika za prislov (*temeljito strokovno pripravo*), problem je tudi (V) ločevanje med prislovi ter enakopisnimi predlogi (*plesati okoli roke*). Redkeje se pojavlja npr. (VI) označevanje samostalnika za prislov (*moč pajka*) ali (VII) samostalnika za glagol (*vinogradnica iz Plešiva*). Poleg navedenih problemov je bilo evidentirano še nekonsistentno besednovrstno označevanje besede *sam* (*je začel tudi sam plesati step* – pridevnik; *dolgo let je tudi sama plesala* – prislov).

Pogosti so tudi problemi pripisovanja **napačnih oblikoskladenjskih kategorij** oblikam v besedilu, npr. (I) pripisovanje napačnega spola ali sklona znotraj besednih zvez (*pajka tarantelo* -> Somdi Sozet) oz. (II) na ravni vezave skladenjskega predmeta s povedkom (*plesati tango* -> Somej). Neujemanje v sklonu je še posebej problematično na ravni predloga pred samostalnikom (*zaradi sinovega navdušenja nad pajki* -> Do Sommij).

Našteti problemi se večinoma zdijo rešljivi z nadgradnjo označevanja besednih oblik z upoštevanjem kolokacijskih ter koligacijskih besednozveznih informacij. V primeru, da so v leksikalni zbirki vnesene najbolj tipične besedne zveze, v katerih se besedna oblika pojavlja, je slednjo v besedilu mogoče označiti upoštevajoč omenjene informacije. Na ta način se je možno (vsaj na ravni najbolj tipičnih zvez) izogniti denimo napačnemu označevanju sklona v ujemalnih zvezah samostalnika in pridevnika itd.

K izboljšavi kvalitete označevanja bi veliko prispevalo tudi upoštevanje samih koligacijskih informacij o tipičnih besednih zvezah v slovenščini, npr. dodajanja na koligacijskih podatkih temelječih pravil na mesto razdvoumljanja lem oz. oblikoskladenjskih oznak. V primeru dvoumne prislovne oz. pridevniške oblike je npr. mogoče preverjanje besednovrstne opredelitve neposredne besedilne okolice na desni ob problematične oblike ipd.

Težko rešljivi so problemi, vezani na označevanje **lastnih imen**. Kljub temu se pred razvojem kvalitetnega sistema za avtomatsko prepoznavo lastnih imen v slovenščini kaže možna začasna rešitev vključevanje najpogostnejših lastnoimenskih enot (besed ali besednih zvez) v leksikalno zbirko – pri čemer je privzeto, da so enote kandidatke za uvrstitev avtomatsko identificirane, nato pa ročno pregledane in razvrščene v zbirko. Možno je tudi označevanje tega tipa enot v leksikalni zbirki s posebno oznako, na katero so vezani nadaljnji postopki avtomatske obdelave naravnega jezika (npr. neprevajanje pri strojnem prevajanju ipd.).

Primeri za neustrezno označevanje lastnih imen se v sklopu obravnavanih lem sicer pojavljajo na ravni označevanja tako (I) stvarnih lastnih imen, npr. *pajek [SIP, Sip, Olivi, Pottinger]*, kot (II) osebnih lastnih imen (*strastni Dmitrij*), konkretno tudi denimo na ravni (III) neobstoja besede *pajek* v leksikalni zbirki kot priimka,

npr. *pajek* [Zoran, Maja], ali (IV) napačne lematizacije sicer evidentiranega lastnega imena, npr. *pajka Francija* v *pajek Francija* namesto v *pajek Franci*.

Označevalne napake se, razumljivo, pojavljajo tudi na mestih **označevanja drugih neznanih besed**, npr. (I) tujih besed (*strasten love/hate odnos*) ali (II) v zbirki še neobstoječih slovenskih besed (*plesati s telebajski*). Besede, ki med procesom označevanja ostanejo nelematizirane, se zapišejo v poseben seznam ter naknadno dodajo v leksikalno zbirko, s čimer je seveda smiselno nadaljevati.

Posebno pozornost za nadaljnji razvoj avtomatskega označevanja zahteva **označevanje členkov**, saj gre za kategorijo, ki je na oblikoskladenjski ravni – zaradi enakopisnosti členkov s prislovi ter vezniki – avtomatsko težko pripisljiva. V zvezi s tem je predlagan premislek o označevanju členkov šele na skladenjskem nivoju označevanja. Premislek zahteva tudi **označevanje okrajšav**, kjer je možno ločevati med skupinami, zajemljivimi v leksikon (npr. merske enote), ter nepredvidljivimi, za označevanje katerih bi bilo potrebno razviti specializirane metode avtomatske identifikacije v besedilu.

2 Določanje relevantnosti vzorčnih tipov

Pričujoče poglavje prinaša poskus strnitve ugotovitev, vezanih na relevantnost posameznega vzorčnega tipa za luščenje leksikalnih podatkov. Kot je bilo izpostavljeno v IV-1.2, »relevantnost vzorčnega tipa« v pričujoči raziskavi pomeni dovolj veliko verjetnost, da luščenje besednih nizov za obravnavani vzorčni tip prinaša leksikalne enote, ki so (same na sebi) zanimive za uvoz v leksikalno zbirko. Končna odločitev glede relevantnosti oz. nerelevantnosti določene vrste besednih nizov za nadaljnjo obravnavo je seveda odvisna od namena luščenja znotraj vsake posamezne raziskave – v nadaljevanju predstavljena delitev vzorčnih tipov prinaša torej le eno, nikakor pa ne edino, od možnih izhodiščnih razvrstitev.

Večje spremembe v kategorizaciji vzorčnih tipov so predvidene predvsem kot rezultat analize podatkov, pridobljenih na osnovi obravnave večje količine (besednovrstno različnih) lem. Trenutno so vzorčni tipi pripravljeni na (pre)majhnem številu obravnavanih besed, da bi bilo omogočeno zanesljivo posploševanje. V zvezi s tem so zaželeno nadaljnje raziskave, prvi korak h katerim prinaša poglavje VI-2.2.

2.1 Delitev vzorčnih tipov glede na rezultate analize

Vzorčni tipi, ki so bili natančneje predstavljeni tekom analize v poglavju V, so v nadaljevanju strnjeni v treh tabelah: ločujemo (I) za luščenje relevantne vzorčne tipe od (II) za luščenje nerelevantnih vzorčnih tipov ter (III) vzorčnih tipov, ki vsebujejo glagol. Slednji so predstavljeni ločeno zaradi pričakovane višje stopnje posega na skladenjski nivo jezika, kar posledično pomeni manjšo ustreznost obravnavane metode za pridobivanje teh podatkov.

Spodnja tabela je torej zasnovana podobno kot Tabela 167 v poglavju V-6: legenda zapisa vzorčnih tipov ostaja enaka (glej Priloga 3), dve zvezdici pred primerom besednega niza prinaša tabela na mestu, kjer je bilo za obravnavani vzorčni tip najti le neustrezno označene primere. Z znakom # so označeni vzorčni tipi, ki so bili evidentirani tekom obravnave manj tipičnih vzorčnih tipov za lemo *pajek* (glej poglavje V-6).

2.1.1 Za luščenje relevantni vzorčni tipi

Vzorčni tipi v spodnji tabeli so glede na vrsto besednih zvez, ki jih z luščenjem pridobimo, razvrščeni v 4 skupine: (I) samostalniške besedne zveze, (II) pridevniške besedne zveze, (III) priredne besedne zveze ter (IV) pogojno zanimive zveze.

Nabor samostalniških besednih zvez je po pričakovanjih najboljšežnejši, v osnovi prinaša zveze samostalnika s samostalniškim, pridevniškim prilastkom ali predložno zvezo. Priredne besedne zveze prinašajo besednovrstno

simetrične tridelne vzorčne tipe s prirednim veznikom. Med pogojno zanimive zveze pa se uvrščajo kombinacije pridevnika oz. prislova z določili na levi, vendar brez besednozveznega jedra.

Samostalniške besedne zveze	
Sam + Sam	<i>pajek skakač, človek pajek</i>
Sam + Sam ₂	<i>vrsta pajkov</i>
Sam + Sam ₃	<i>podpora Slovenkam</i>
Sam + Sam _{LI}	<i>pajek Krivograd</i>
Sam + Sam + Sam	<i>pajek SIP spider</i>
Prid + Sam + Sam	<i>dvovretenski pajek sip; strasten igravec golfa, strastna reševalka križank, strastno odkrivanje umetnosti</i>
#Sam + Prid + Sam (različni tipi zvez glede na ujemanje v sklonu)	<i>pajek črna vdova; pajek italijanske proizvodnje, razvoj rdečega pajka; zavetje drobnim pajkom; pajku podoben moški;</i>
Prid + Sam (ujemanje v sklonu, razlikovanje glede na spol)	<i>ptičji pajek; strasten kadilec, strastna ljubezen, strastno razmerje</i>
Sam + Prid (ujemanje)	<i>pajek dvovretenski</i>
#Sam + Prisl + Prid (ujemanje)	<i>pajek grozno nevaren</i>
#Sam + Sam + Prid (ujemanje)	<i>SIP pajek dvovretenski, pajek SIP dvovretenski, pajek človeku nevaren</i>
#Sam + Prid + Prid (ujemanje)	<i>pajek dvovretenski italijanski</i>
Prid + Prid + Sam	<i>rdečenog ptičji pajek; strasten nogometni navijač, strastna ljubzenska zgodba, strastno ljubzensko razmerje</i>
Prisl + Prid + Sam	<i>najbolj strupen pajek; najbolj strasten kadilec, zelo strastna ženska, najbolj strastno zavzemanje</i>
Sam + Pred + Sam	<i>pajek za seno, odvoz s pajkom</i>
#Pred + Sam + Sam	<i>v obliki pajka</i>
#Sam + Vd + Sam	<i>pajek kot sredstvo</i>
Pridevniške besedne zveze	
Prid + Sam ₃	<i>podoben pajku</i>
#Prid + Vd + Sam	<i>strupen kot pajek</i>
#Prid + Pred + Sam	<i>nor na pajke</i>
Priredne besedne zveze	
Sam + Vp + Sam	<i>lisice in pajek</i>
Prid + Vp + Prid	<i>romantičen in strasten</i>
Glag + Vp + Glag	<i>peti in plesati</i>

Prisl + Vp + Prisl	<i>hitro in temeljito</i>
Pogojno zanimive zveze	
Prisl + Prisl	<i>zelo temeljito, temeljito strokovno</i>
Prisl + Prisl + Prisl	<i>prav tako temeljito</i>
# Prisl + Pred + Sam	<i>skupaj s pajki</i>
#Prisl + Vd + Sam ₁	<i>urno kot pajek</i>
Prisl + Prid	<i>najbolj strasten; temeljito prenovljen</i>
Prisl + Prisl + Prid	<i>prav tako strasten</i>
#Prisl + Prisl + Sam	<i>kar nekaj pajkov</i>

Tabela 168: Nabor za luščenje relevantnih vzorčnih tipov.

2.1.2 Vzorčni tipi, ki vsebujejo glagol

V spodnji tabeli so vzorčni tipi razvrščeni v pet skupin, glede na to, kaj se v izluščenem besednem nizu ob glagolu pojavlja: (I) samostalnik oz. samostalniška besedna zveza (primeri so ločeni glede na to, ali se slednje pojavlja v imenovalniku ali neimenovalniku), (II) prislov ali prislovna zveza, (III) glagol, (IV) predlog ali predložna zveza ter (V) pridevnik.

Kombinacije s Sam / SamBZ	
Sam ₁ + Glag Sam _{/1} + Glag	<i>pajek odpeljati, miši plesati pajkov bati; pajka zadeti, kolo plesati</i>
Glag + Sam ₁ Glag + Sam _{/1}	<i>odpeljati pajek, plesati otroci bati pajkov; dati pajkom; poklicati pajka, plesati tango;</i>
Glag + Prid + Sam ₁ Glag + Prid + Sam _{/1}	<i>razviti strastna romanca, plesati folklorne skupine preživeti strastno noč, plesati ljudske plese</i>
Glag + Sam + Sam	<i>plesati Mojca Majcen</i>
Prid + Sam + Glag	<i>ptičji pajek živet</i>
# Sam ₁ + Glag + Glag # Sam _{/1} + Glag + Glag	<i>pajek začeti stopicati pajke želeči pretihotapiti</i>
#Glag + Glag + Sam ₁ #Glag + Glag + Sam _{/1}	<i>pričeti mastiti pajek želeči prijeti pajka</i>
#Glag + Sam ₁ + Glag #Glag + Sam _{/1} + Glag	<i>morati pajek doživeti smeti pajka fotografirati</i>
# Sam ₁ + Glag + Sam # Sam _{/1} + Glag + Sam	<i>pajek odpeljati avto; dijaki spoznavati pajke pajku odtrgati nogo; avto odpeljati pajek</i>
# Sam ₁ + Prid + Glag	<i>pajek neopazen ždeti</i>

<p>#Sam + Sam + Glag (različni tipi zvez glede na sklon)</p> <p>Prisl + Glag + Sam₁ Prisl + Glag + Sam_{/1}</p> <p>Sam₁ + Prisl + Glag Sam_{/1} + Prisl + Glag</p> <p>#Glag + Sam + Prisl</p> <p>#Glag + Prisl + Sam</p> <p>#Prisl + Sam₁ + Glag</p> <p># Sam_{/1} + Glag</p> <p>#Glag + podredni veznik + Sam₁ #Glag + podredni veznik + Sam_{/1}</p>	<p><i>pajek muho pokončati; pajek zaklopničar živeti; pajku nogo odtrgati; število pajkov zmanjšati;</i></p> <p><i>temeljito spremeniti narava</i> <i>temeljito umiti roke</i></p> <p><i>ljudje radi plesati, policisti temeljito pregledati</i> <i>sestavine temeljito premešati</i></p> <p><i>skočiti pajek naravnost</i></p> <p><i>uvesti spet pajke</i></p> <p><i>danes pajek odpeljati</i></p> <p><i>najprej pajka omamiti</i></p> <p><i>misлити kot pajek</i> <i>narisati kot pajka</i></p>
Kombinacije s Prisl / PrislBZ	
<p>Prisl + Glag</p> <p>Prisl + Prisl + Glag</p> <p>Glag + Prisl</p> <p>Glag + Prisl + Prisl</p>	<p><i>dobro plesati; temeljito spremeniti</i></p> <p><i>zelo dobro plesati, res temeljito pripraviti</i></p> <p><i>plesati skupaj, opraviti temeljito</i></p> <p><i>plesati tesno skupaj</i></p>
Kombinacije z Glag	
<p>Glag + Glag</p> <p>Glag + Prisl + Glag</p>	<p><i>začeti plesati</i></p> <p><i>morati temeljito premisliti</i></p>
Kombinacije s Predl	
<p>Glag + Pred</p>	<p><i>plesati v</i></p>
<p>Glag + Pred + Sam</p> <p>Pred + Sam + Glag</p>	<p><i>odpeljati s pajkom, plesati z volkovi</i></p> <p><i>z veseljem plesati</i></p>
Kombinacije s Prid	
<p>Prid + Glag</p> <p>Glag + Prid</p>	<p><i>pijan plesati</i></p> <p><i>plesati bos, plesati folklorne</i></p>

Tabela 169: Nabor vzorčnih tipov, ki vsebujejo glagol.

2.1.3 Za luščenje nerelevantni vzorčni tipi

Nerelevantni vzorčni tipi so združeni v skupine glede na vsebnost »problematičnega elementa« (npr. pridevnika na koncu vzorčnega tipa, kar se odraža v necelosti izluščenih besednih zvez itd.). V določeni meri lahko

predlagana razvrstitev sugerira generalizacijo hevrstik za avtomatsko odstranjevanje nerelevantnih vzorcev, vendar je pred poskusom posploševanja potrebno preveriti ugotovitve na večjem naboru besed.

Trenutno so skupine nerelevantnih vzorcev naslednje: vzorčni tip vsebuje (I) prislov na koncu ali pred samostalnikom, (II) pridevnik na koncu, (III) predlog na začetku ali koncu, (IV) priredni veznik med različnima besednima vrstama ali na začetku oz. koncu vzorca, (V) vnaprej določeno neželjeno oblikoskladenjsko oznako. Predvsem pri slednji skupini so zaželene nadaljnje raziskave, saj so bili vzorci iz nabora odstranjeni pred začetkom analize.

vzorčni tip	primer besednega niza
Prisl na koncu vzorca ali pred Sam	
Sam + Prisl	<i>pajek lahko; sestavine temeljito</i>
Sam + Prisl + Prisl	<i>pajek tam gotovo, zmes zelo temeljito</i>
Prid + Sam + Prisl	<i>naslednje leto temeljito</i>
Pred + Sam + Prisl	<i>s pajkom tako, pred uporabo temeljito</i>
#Sam + Pred + Prisl	<i>pajek na najbolj</i>
#Sam + Sam + Prisl	<i>pajek mrežo stalno; pajek križavec pogosto, človek pajek nikoli</i>
#Sam + Vd + Prisl	<i>**pajek kot nehumano</i>
#Prisl + Sam + Prisl	<i>lahko pajek nemoteno</i>
#Sam + Prid + Prisl	<i>**pajek samo zato</i>
#Prid + Prisl + Sam	<i>**spletene mreže pajka</i>
Glag + Prisl (niso vsi primeri nerelevantni)	<i>plesati tako</i>
Prisl + Glag + Prisl	<i>dejansko plesati zato</i>
#Sam ₁ + Glag + Prisl	<i>pajek odvažati nepravilno</i>
#Sam ₁ + Glag + Prisl	<i>pajkom izogniti tako</i>
Prisl + Sam	<i>lahko pajek; temeljito prenova</i>
#Prisl + Sam + Sam	<i>torej pajek ukrep; sicer uporaba pajka</i>
#Sam + Prisl + Sam	<i>kislina lahko pajek</i>
#Sam + Prisl + Sam	<i>pajek dnevno težave</i>
Prid na koncu vzorca	
Sam + Prid (neujemanje)	<i>vlogo strastne</i>
Prid + Sam + Prid	<i>strastno prebujenje mladostne</i>
Sam + Pred + Prid	<i>pajek v občinski</i>
#Sam + Prisl + Prid (neujemanje)	<i>pajek najbolj strupena</i>

#Sam + Sam + Prid (neujemanje)	<i>črni pajek bavarske, pajek ukrep kranjske</i>
#Sam + Prid + Prid (neujemanje)	<i>pajka nepravilno parkiran</i>
#Sam + Vd + Prid	<i>pajek kot poskusen</i>
Prid + Prid	<i>strasten ljubezenski, nekdanji strasten;</i>
Glag + Prid (niso vsi primeri nerelevantni)	<i>plesati folklorne, preživeti strasten</i>
Prisl + Glag + Prid	<i>skupaj preživeti strastno</i>
Glag + Pred + Prid	<i>zaplesti v strastno, plesati v plesni</i>
#Sam ₁ + Glag + Prid	<i>pajek imeti poseben</i>
#Sam _{1/1} + Glag + Prid	<i>pajka predstaviti nemški</i>
#Glag + Sam + Prid	<i>prihajati pajek črna</i>
Pred na začetku ali koncu vzorca	
Pred + Sam/Prid/Glag/Prisl	<i>s pajkom, v strastno, **iz Plešiva, za temeljito</i>
Pred + Prid + Sam	<i>v strastno razmerje</i>
#Pred + Sam + Sam (niso vsi primeri nerelevantni)	<i>za otroke pajek</i>
Sam/Prid/Prisl + Pred	<i>pajek na, strasten v, temeljito v</i>
Sam + Sam + Pred	<i>pajek SIP na</i>
Prid + Sam + Pred	<i>hidravlični pajek za, strasten ljubezen z</i>
Pred + Sam + Pred	<i>s pajkom za</i>
Prisl + Glag + Pred	<i>lahko plesati na, temeljito pripraviti na</i>
Glag + Prisl + Pred	<i>plesati pozno v</i>
Sam + Glag + Pred	<i>ljudje plesati po</i>
#Vd + Sam + Pred	<i>kot pajek na</i>
Vp med različnima besednima vrstama ali na začetku ali koncu vzorca	
Sam + Vp + Prisl	<i>pajek in kmalu, vodo in temeljito</i>
Sam + Vp + Prid	<i>uspeh in strasten</i>
Prid + Vp + Prisl	<i>strasten in obenem</i>
Prisl + Vp + Prid	<i>**svobodno in strastno</i>
Glag + Vp + Prisl	<i>plesati in tako, operemo in temeljito</i>
Prisl + Vp + Glag	<i>naokoli in plesati</i>
Vp na začetku ali koncu vzorca	<i>in pajek, in pajek SIP</i>

Vsebnost »problematične« oblikoskladenjske oznake	
vzorci, ki vsebujejo oznako za pomožni glagol	<i>je odpeljal pajek</i>
vzorci, ki vsebujejo oznako za članek	<i>tudi strasten lovec</i>
vzorci, ki vsebujejo oznako za okrajšavo	<i>pajek 5 m</i>
vzorci, ki vsebujejo oznako za zaimek	<i>se je temeljito</i>
vzorci, ki vsebujejo oznako za števniki	<i>leta 1994 temeljito</i>
vzorci, ki vsebujejo oznako za nelematizirano besedo	<i>jebeni pajek</i>
Drugi vzorčni tipi	
#Prisl + Vd + Sam _{/1}	<i>več kot pajkov</i>
#Sam + Prid + Sam (niso vsi primeri nerelevantni)	<i>pajek ujeto žuželko</i>
#Prisl + Pred + Sam (niso vsi primeri nerelevantni)	<i>nato na pajka</i>
#Sam + Vd + Glag	<i>pajek kot pikati</i>
#Prid + Glag + Sam	<i>**celo odpeljati pajek</i>

Tabela 170: Nabor za luščenje nerelevantnih vzorčnih tipov.

2.2 Luščenje samostalniških besednih zvez

V prejšnjem poglavju predstavljene tabele imamo lahko za izhodišče nadaljnjih raziskav, pri čemer je ključno predvsem (I) preverjanje predlagane razvrstitve vzorčnih tipov glede na relevantnost pridobljenih besednih nizov, (II) dodajanje novih vzorčnih tipov v ustrezna mesta tabel, (III) evalvacija in izboljšava predlaganih metod luščenja glede na novopridobljene podatke.

Ker so bile za slovenščino do sedaj luščene predvsem samostalniške besedne zveze (glej II-2.1), je z željo omogočanja primerjave rezultatov v nadaljevanju predstavljena delna evalvacija nabora vzorčnih tipov za luščenje samostalniških besednih zvez. Evalvacija poteka na osnovi novega nabora samostalniških lem, tj. *Slovenka*, *oven* in *debata*. Preverjanje podatkov poteka na dveh ravneh: (I) iskanje potencialnih novih vzorčnih tipov za dvodelne ter tridelne vzorce z novimi lemami, (II) luščenje in analiza nizov za nabor vzorčnih tipov, ki so v tabelo uvrščeni kot relevantni za nadaljnjo obravnavo.

2.2.1 Novi vzorčni tipi

Nove vzorčne tipe je možno relativno preprosto poiskati s pomočjo programa, ki s seznama vzorcev za novo obravnavano lemo odstrani vzorce, ustrezajoče vsem že evidentiranim vzorčnim tipom (opis programa v IV-3.4.2). Sledi ročni pregled preostalih vzorcev, v primeru najdenega novega vzorčnega tipa beleženje le-tega ter luščenje ustrežajočih besednih nizov, na podlagi katerih temelji opredelitev relevantnosti novega vzorčnega tipa.

Z opisanim postopkom pridobljene podatke prikazuje Tabela 171:

lema	vzorčni tip	besedna zveza	uvrstitev v tabelo
Slovenka	Prisl + Sam + Prid	<i>lani Slovenke svetovne (1)</i>	nerelevantni
	Sam + Pred + Glag	<i>*Slovenka nasproti stati (1)</i>	nerelevantni
oven	Sam + Prid + Pred	<i>oven obešen na (1)</i>	nerelevantni
	Sam + Prisl + Pred	<i>oven kmalu po (1)</i>	nerelevantni
	Sam + Pred + Pred	<i>oven s po (1)</i>	nerelevantni
	Sam + V + Pred	<i>oven čeprav za (1)</i>	nerelevantni
debata	Glag + Sam + Pred	<i>teči debata o (26)</i>	nerelevantni

Tabela 171: Novi vzorčni tipi za leme *Slovenka*, *oven*, *debata*.

Vsi od gornjih vzorčnih tipov so uvrščeni v tabele k ostalim nerelevantnim.

2.2.2 Luščenje besednih zvez

Pričujoče poglavje prinaša s specializiranim programom (za opis programa glej IV-3.4.1) izluščene samostalniške besedne zveze za nove obravnavane leme. Primeri za vsakega od vzorčnih tipov so izbrani primarno glede na pogostnost, mestoma (pri manj tipičnih vzorčnih tipih) pa tudi glede na ponazoritveni potencial.

Vzorčni tipi v nadaljevanju so razvrščeni v tri glavne skupine: (I) zveze samostalnika oz. samostalniške besedne zveze s samostalnikom oz. samostalniško besedno zvezo, (II) zveze samostalnika s pridevniškimi prilastki ter (III) zveze s predlogom ali veznikom.

vzorčni tip	primer besedne zveze s pogostnostjo	opombe
Sam/SBZ + Sam/SBZ		
Sam + Sam	<i>mati Slovenka (26), Slovenka Tina (89)</i> <i>Maribor oven (43), oven vodnik (13)</i>	<i>Maribor oven (43)</i> – glej razlago spodaj.
Sam + Sam ₂	<i>večina Slovenk (30), Slovenka leta (560)</i> <i>znamenje ovna (88)</i> <i>stvar debate (52), debata svetnikov (3)</i>	<i>TV debata (4), internet debata (1)</i> – glej razlago spodaj
Sam + Sam ₃	<i>podpora Slovenkam (1)</i> <i>Šepetanje ovnom (2)</i> <i>sledenje debatam (2)</i>	
Sam + Sam (drugi Sam se ne sklanja)	<i>revija Slovenka (2)</i> <i>šifra oven (9)</i> <i>kategorija Debate (1)</i>	<i>Oven se pojavlja tudi kot priimek.</i>
Sam + Sam + Sam (različne vrste zvez glede na sklon)	<i>Slovenka Katarina Srebotnik (59), naslov Slovenka leta (39), ekipa trgovine Slovenka (2)</i> <i>oven JSR pasme (20), šifra oven Marija (7), nagrada Železnega ovna (6)</i> <i>Debate Računalništvo Ljubezen (32), TV debate kandidatov (2), izziv vodenja debate (2)</i>	<i>Debate Računalništvo Ljubezen (32)</i> – glej razlago spodaj

Sam + Prid + Sam (različne vrste zvez glede na sklon)	<i>Slovenka srednjih let (4), podoba znanih Slovenk (48)</i> <i>oven solčavske pasme (4), zid nabijajoč oven (4)</i> <i>debata predsedniških kandidatov (5), uro trajajoča debata (7)</i>	<i>zid nabijajoč oven (4) – glej razlago spodaj</i> <i>uro trajajoča debata (7) – glej razlago spodaj</i>
Sam/SBZ s pridevniškimi prilastki		
Prid + Sam	<i>*znan Slovenka (410)</i> <i>plemenski oven (82)</i> <i>*javen debata (267)</i>	problematična lematizacija pridevniških oblik – v tabeli so navedeni nizi z lematiziranimi pridevniškimi oblikami (primeri so označeni z zvezdico)
Prid + Prid + Sam	<i>*številčen znan Slovenka (14)</i> <i>mlad plemenski oven (4)</i> <i>*razen težek debata (10)</i>	
Prisl + Prid + Sam	<i>*visoko uvrščen Slovenka (60)</i> <i>sicer krotek oven (1)</i> <i>*bolj vroč debata (5)</i>	
Prid + Sam + Sam (različne vrste zvez)	<i>*Janin Slovenka leta (21), moden slog Slovenk (11)</i> <i>plemenski oven SJR (1), astrološko znamenje ovna (2)</i> <i>*ekonomski debata leta (2), dneven red debate (4)</i>	
Inverzivni vzorci (zamenjava besednega reda)		
Sam + Prid Sam + Sam + Prid Sam + Prid + Prid	<i>oven plemenski (2)</i>	<i>z lemmama Slovenka in debata ni relevantnih primerov za pričujoči vzorčni tip, samo kombinacije tipa Slovenke uspešne (5)</i>
Zveze s Pred ali Vp oz. Vd		
Sam + Pred + Sam Pred + Sam + Sam	<i>kandidatka za Slovenko (52), Slovenka v finalu (13)</i> <i>luna v ovnu (118), oven pri delu (7) debata o DVB (92), prostor za debato (7)</i> <i>po rodu Slovenka (21), za Slovenko leta (60)</i> <i>v znamenju ovna (91), z ovni pasme (3)</i> <i>na debati omizja (1), pred začetkom debate (5)</i>	<i>kandidatka za Slovenko (52) – glej razlago spodaj</i> <i>debata o DVB (92) – glej razlago spodaj</i>
Sam + Vd + Sam ₁ Sam + Vp + Sam	<i>Slovenka kot voznica (2)</i> <i>Rupel kot oven (1)</i> <i>debata kot znanost (1)</i> <i>Slovenec in Slovenka (635), Slovenka in Slovenec (1114)</i> <i>ovca in oven (39), oven in kozel (17)</i> <i>debata in razmislek (31), predavanje in debata (16)</i>	<i>Slovenec in Slovenka (635) – glej razlago spodaj</i>

Tabela 172: Samostalniške besedne zveze za leme *Slovenka*, *oven*, *debata*.

Na tem mestu je potreben ponoven poudarek, da je predstavljeno luščenje zvez z novimi lemmami le del evalvacije, predvidene za nadaljevanje razvoja opisovane metode. Podobno analizo je potrebno opraviti tudi na ravni preostalih vzorčnih tipov (glagolskih, nerelevantnih) ter seveda za zveze, kjer je jedro nesamostalniška besedna vrsta. Tudi na ravni samostalniških besednih zvez je predvideno dopolnjevanje podatkov za še večji nabor lem.

V poglavju II-2.1 je bil predstavljen nabor vzorcev, ki so bili za slovenščino preizkušeni pri luščenju terminoloških besednih zvez. Kot omenjeno, gre v pričujoči raziskavi za luščenje z drugim namenom, kar je glavni razlog za razlike, ki se pojavljajo ob primerjavi vzorcev ene ter druge raziskave. Pri prvi (luščenje terminologije, v spodnji tabeli siv stolpec) se kaže večja osredotočenost na tipične besedne zveze, v obravnavo pa so zajeti tudi štiridelni vzorci. Doktorska raziskava (bel stolpec) prinaša nekoliko več dvo- oz. tridelnih vzorcev, vendar večinoma na račun pritegnitve v obravnavo manj tipičnega v jeziku.

Vzorca, ki bi za luščenje lahko bila širše uporabna, sta v spodnji tabeli podčrtana. Gre za primera, ki prinašata (I) priredne besedne zveze oz. (II) samostalniške besedne zveze, ki jih določa prislov.

luščenje terminologije	luščenje za leksikalno zbirko	novi vzorci
PRID + SAM	✓	SAM + PRID
SAM + SAM	✓	SAM + SAM + PRID
SAM + SAM + SAM	✓	SAM + PRID + PRID
PRID + PRID + SAM	✓	<u>PRISL + PRID + SAM</u>
PRID + SAM + SAM	✓	PRED + SAM + SAM
SAM + PRID + SAM	✓	SAM + VD + SAM
SAM + PRED + SAM	✓	<u>SAM + VP + SAM</u>
SAM + PRED + PRID + SAM	✗	
PRID + SAM + PRED + SAM	✗	
SAM + PRED + SAM + SAM	✗	

Tabela 173: Primerjava nabora vzorcev za luščenje samostalniških besednih zvez.

2.2.3 Analiza izluščenih zvez

Prva skupina besednih zvez, zveze **s samostalniškim prilastkom**, v celoti prinaša leksikografsko relevantne rezultate. Na ravni dvodelnih vzorcev se kaže za bolj tipični kombinacija dveh v sklonu ujemajočih se samostalnikov (*mati Slovenka-26*) ter zveza z drugim samostalnikom v rodilniku (*Slovenka leta-560*). Pojavitve samostalniških zvez z drugim samostalnikom v dajalniku so redke (*podpora Slovenkam-1*), običajno jih je s seznama kandidatov za besedne zveze potrebno poiskati ročno.¹²¹ Prav tako so redki primeri kombinacij z desnim neujemalnim imenovalnim prilastkom (*revija Slovenka-2*), za katere se je prav tako izkazalo, da jih je avtomatsko težje identificirati: ker je v imenovalniku pogosto celotna besedna zveza, so ti primeri primarno razvrščeni na seznam »ujemalnih« zvez. V primeru, da jih želimo ločevati, je torej potrebna ročna identifikacija.

Med novoizluščenimi primeri se pojavlja tudi tip zveze, ki do sedaj v raziskavi ni bil evidentiran, in sicer s primeroma **TV debata**, **internet debata**. Pri tovrstnih primerih je pričakovana ničta sklanjatev prvega dela zveze, drugi samostalnik pa se sklanja. Zanesljivo avtomatsko ločevanje se zdi težje izvedljivo, ker je primerov malo, niso pa ločljivi v primeru, da je v imenovalniku celotna besedna zveza.

¹²¹ Zaradi nizke pogostnosti se na seznamu izluščenih nizov relevantne zveze izgubijo med nerelevantnimi.

Tudi luščenje zvez **samostalnika in samostalniške besedne zveze** prinaša relevantne besedne zveze za vse nove samostalniške leme. Na ravni luščenja treh samostalnikov se problem izkazuje pri zvezi *Debate Računalništvo Ljubezen*, za katero konkordančni niz korpusa FidaPLUS prinaša nabor ponovljenih konkordanc internetnega izvora (samostalniki najbrž označujejo imena povezav z večjega števila sorodnih internetnih strani). Drugi primer iz gornje tabele, ki ima visoko frekvenco zaradi istega razloga, je *debata o DVB-92*.¹²² Tudi sicer so korpusni šumi te vrste v korpusu FidaPLUS pogosti, kar priča o potrebi po večji pozornosti pri vključevanju internetnih besedil v referenčni korpus.

Primerljivo dobre rezultate daje luščenje kombinacije samostalnika in samostalniške besedne zveze z levim pridevniškim prilastkom (*debata predsedniških kandidatov*). Problemi se kažejo na ravni lematizacije prvega samostalnika, primer *zid nabijajoč oven* denimo izvira iz zveze *iz zidu nabijajoč oven*. Pri zvezah tega tipa je torej potrebna dodatna pozornost.

Luščenje zvez **samostalnika z levim pridevniškim prilastkom** za vse tri leme prinaša nabor visokopogostnih besednih zvez (*plemenski oven-82*). Tridelne zveze so v skladu s pričakovanji redkejšje, vendar nabor primerov kljub temu kaže na dobre rezultate – kar priča o tem, da je luščenje tega tipa zvez z upoštevanjem ujemalnosti uspešna metoda. Problemi se pojavljajo le na ravni prikaza izluščenih podatkov, ki prinaša pridevnike, lematizirane v nezaznamovano pridevniško obliko, npr. *razen težek debata, Janin Slovenka leta* itd.

Nasprotno se na ravni luščenja besednih zvez s **pridevniškim prilastkom na desni** kažejo slabi rezultati. Izluščeni besedni nizi so redki, večinoma pa prinašajo zveze tipa *Slovenke uspešne*, tj. kombinacijo, ki na skladenjski ravni izraža osebek ter povedkovo določilo. Pri luščenju treh tipov zvez za tri leme je bil najden le en relevanten primer, tj. *oven plemenski (2)*.

Luščenje **zvez s predlogom** daje mešane rezultate. Na eni strani omogoča pridobitev zvez, ki so za nadaljnjo obravnavo zanimive (*po rodu Slovenka, luna v ovnu, v znamenju ovna, prostor za debato*), na drugi strani na seznamu najdemo nekončane zveze (*kandidatka za Slovenko, z ovni pasme*).

Redke in manj zanimive so **zveze s podrednim veznikom** (*debata kot znanost-1*), na drugi strani pa so po pričakovanjih izredno pogoste zveze dveh samostalnikov s **prirednim veznikom** (*Slovenec in Slovenka-635*).

Pri analizi izluščenih besednih zvez se na več mestih kaže potreba po **upoštevanju kategorije števila** samostalnika. Nekaj primerov iz gornje tabele: *Maribor oven, Slovenka kot voznica, ovca in oven, Slovenec in Slovenka* itd. Naštete zveze bi želeli izluščiti v obliki, ki prinaša ustrezno obliko za število: *Maribor Ovni, Slovenke kot voznice, ovce in ovni, Slovenci in Slovenke*. V primerih, kjer so glede na število možne različne zveze, je želeno slednje ločevati, npr. *uro trajajoča debata* od *ure trajajoča debata*.

Podobno se kaže potreba po ponovnem razmisleku o **upoštevanju kategorije stopnje** pri prislovih ter pridevniki: primer izluščene zveze *visoko uvrščen Slovenka* se v besedilih v veliki večini pojavlja v obliki *najvišje uvrščena Slovenka*, podoben je tudi primer (ki v gornjo tabelo sicer ni bil uvrščen) *uraden lep Slovenka*, ki se v besedilih pojavlja v obliki *uradno najlepša Slovenka*.

¹²² DVB je kratica za *Digital Video Broadcasting*, sistem digitalne televizije.

VII SKLEP

V raziskavi je bila uporabljena in s tem preverjena metoda luščenja leksikalnih podatkov na osnovi dvodelnih ter tridelnih kombinacij obravnavane leme z oblikoskladenjskimi oznakami neposredne besedilne okolice. Za potrebe luščenja je bilo pripravljenih več specializiranih programov, katerih učinkovitost je bila preverjena na osnovi analize izluščenih besednih nizov v poglavju V. Na več mestih omenjenega kakor tudi pričujočega poglavja so bile izpostavljene možnosti izboljšave metode luščenja oz. opozorila glede predvidenega dodatnega razvoja postopka. Na tem mestu so ugotovitve za vse stopnje obdelave podatkov strnjeno predstavljene.

Na ravni priprave korpusnih besedil ter formiranja začetnega seznama vzorcev se kot problem kaže predvsem **dolgotrajnost obdelave**, vezana na omejeno procesorsko moč uporabljenega računalnika ter seveda številne ročne postopke, s katerimi je bilo dopolnjeno avtomatsko luščenje podatkov. Dolgotrajnost se tekom celotne raziskave pojavlja tudi zaradi parcialne narave izdelanih programskih skript (vsaka od skript se uporablja za eno samo specifično nalogo). Vsekakor dolgotrajna je tudi analiza podatkov, ki pa v nadaljevanju v takšnem obsegu ne bo več potrebna oz. bo potekala na ravni, nakazani v VI-2.2.

Za luščenje podatkov iz celotnega referenčnega korpusa je predvidena izdelava celovitega avtomatiziranega postopka, ki na čim bolj enostaven način združuje vse v pričujoči raziskavi opisane elemente (iskanje relevantnih besednih zvez, razvrščanje in štetje zvez, ustrezen prikaz besednih zvez), kar pomeni, da zgoraj opisane težave za nadaljevanje niso pričakovane.

Glede same **kvalitete izluščenih podatkov** se potrjuje predvidevanje, da so rezultati relevantnejši na mestih, kjer slovenski jezik na skladenjski ravni v okvirih dvodelnih ali tridelnih besednih nizov izkazuje dovoljšno stopnjo regularnosti – iz tega sledi zaključek, da je metoda zanesljivejša na ravni luščenja besednih zvez s samostalniškimi, pridevniškimi ali prislovnimi jedrom. Na ravni kombinacij, ki vsebujejo glagolsko lemo oz. oblikoskladenjsko oznako, je zanesljivost z opisano metodo izluščenih podatkov nižja, predvsem na ravneh, kjer se v izluščenih besednih nizih izkazujejo kombinacije, ki na skladenjski ravni izvornih stavkov prinašajo različne stavčne člene. Za luščenje zvez z glagolom so torej potrebne dodatne raziskave, ki jih bo predvidoma omogočila analiza skladenjsko označenih besedil. Slednjo je sicer več kot smiselno opraviti tudi za izboljšavo luščenja ostalih tipov zvez.

V raziskavi se uporablja **luščenje na osnovi vzorčnih tipov**, ki so večinoma posplošitev nabora sorodnih vzorcev na raven najvišje kategorije oblikoskladenjskih oznak (tj. besedne vrste), obenem pa se na več mestih tekom dejanskega luščenja ter prikazovanja dobljenih podatkov kaže za ključno upoštevanje še katere od označevalnih podkategorij. Povedano drugače: izkazuje se, da upoštevanje vseh podkategorij pri luščenju zvez preveč razpršuje podatke, upoštevanje zgolj besedne vrste pa v enaki meri vodi v izgubo zvez zaradi neustreznega združevanja. Tekom pričujoče raziskave je zato pri posameznem tipu zvez določeno, katere podkategorije je smiselno pri luščenju upoštevati in katerih ne.

V pričujoči raziskavi so najpodrobneje raziskane **možnosti luščenja samostalniških besednih zvez**. Pri luščenju zvez (I) s samostalniškimi prilastki je ključno upoštevanje razlik v sklonu samostalnikov, kakor se pojavljajo v izpričanih besednih zvezah (prim. *oven vodnik* vs. *znamenje ovna* itd.). Glede na sklonske kombinacije lahko med luščenjem zveze razvrščamo, po razvrščanju pa je zaradi dvoumnosti določenih sklonskih kombinacij potrebna dodatna obdelava podatkovnih tabel. Na tej točki so zaželeno nadaljnje izboljšave metode. Metoda luščenja zvez (II) s **pridevniškimi prilastki**, pri čemer je nujno upoštevati ujemanje pridevnika oz. pridevnikov s

samostalnikom v spolu, sklonu in številu, daje dobre rezultate in je ni treba dograjevati – razen na ravni prikaza pridevniških lem v ustreznih oblikah glede na spol jedrnega samostalnika (prim. *javen debata*).

Ob vsem omenjenem se je izkazalo za ustrezno tudi **luščenje predložnih zvez** z besednovrstno različnimi besednozveznimi jedri (samostalnik za predlogom mora biti seveda prikazan v ustrezni sklonski obliki, prim. *pajek za seno*), prav tako tudi **luščenje zvez s podrednim veznikom**, kjer pridejo v poštev kombinacije obravnavane polnopomenske besede s samostalnikom v imenovalniku za veznikom (prim. *strupen kot pajek*). **Luščenje prirednih zvez** prinaša ustrezne rezultate na ravni simetričnih vzorčnih tipov, tj. v primeru, da se levo in desno od veznika pojavljata besednovrstno isti besedi (prim. *romantičen in strasten*).

Za druge vzorčne tipe, trenutno evidentirane za **luščenje pridevniških ter prislovnih zvez**, predstavljena metoda daje dobre rezultate, čeprav je tudi na tem mestu pri manj tipičnih vzorcih potrebno zveze ročno iskati po seznamu. Na ravni luščenja zvez pridevnika s samostalniškim dopolnilom je ključno upoštevanje sklona samostalnika (npr. *podoben pajku*)

Dodaten razmislek je zaželen glede **označevanja vrste prislova**. Na eni strani se zdi smiselno iz celotnega nabora ločevanje prislovov mere, ki predstavljajo bolj ali manj zaključeno skupino, znotraj besednih zvez pa so pomensko manj zanimivi (*zelo strasten kadilec*); enako velja za prislove, ki so na skladenjski ravni del povedka (tip *treba temeljito prenoviti*). Upoštevajoč kategorije v slovnici se zdi glede na sestavljenost prislovnih besednih zvez izvedljivo ločevanje prislovov, ki tipično dobijo prilastek na desni, od tistih, ki dobijo prilastek na levi. Če izmed slednjih izločimo prislove mere, ostanejo t. i. prislovi ozira (*gospodarsko neupravičen*), ki so za avtomatsko obdelavo jezika najbolj zanimivi (zaradi prekrivnosti s pridevniško obliko).

Na ravni zvez z glagolom, kolikor jih imamo za zanimive za vključitev v leksikalno zbirko, je ključno predvsem ohranjanje sklona samostalnika oz. samostalniških besednih zvez, saj slednje sugerira stavčnočlensko vlogo v izvornem stavku (bistveno je torej ločevanje imenovalniških – osebkovnih – enot od neimenovalniških). V razmislek se ponuja nadgradnja prikaza zvez imenovalniškega leksema z glagolom v obliki za tretjo osebo ednine, vendar zgolj z namenom olajšane branja (*odpelje pajek* vs. *odpeljati pajka*).

Dodatne raziskave so potrebne v zvezi z **ugotavljanjem relevantnosti kategorij števila ter stopnje**, saj se na določenih mestih razvrščanja ter prikaza besednih zvez izkazuje za razločevalni, na drugih ne (prim. za relevantnost: *Slovenke in Slovenci* ter *uradno najlepša Slovenka*). Na prvi pogled se zdi smiselna nadgraditev metode na način, ki bi na evidentirano problematičnih mestih upošteval razpršenost različnih oblikoskladenjskih oznak na ravni posamezne kategorije (če se določena besedna zveza npr. pojavlja pretežno v enem samem številu, se prikaže v izpričani obliki, če se pojavlja v ustreznem razmerju vseh treh števil, se prikaže v nezaznamovani edninski obliki).

Načeloma se kaže, da luščenje zvez daje dobro sliko **tipičnega v jeziku**, slabše pa je na ravni manj pogostnih vzorčnih tipov. Zelo tipične zveze so dovolj pogostne, da se pri luščenju podatkov za ustrezni vzorčni tip uvrstijo dovolj visoko v seznamu pogostnosti ne glede na morebitne napake na ravni pripisanih oznak ali izgubljanje podatkov zaradi pomanjkljivosti same metode. Na ravni manj pogostnih podatkov pa je relevantnih zvez manj oz. so od neustrezno označenih oz. uvrščenih ločljive le ročno. Ročna analiza avtomatsko izluščenih podatkov je sicer v vsakem primeru nujna, ker pa je predvidoma zamudna, je smiselno podatke vnaprej selekcionirati. Na kakšen način in do katere mere selekcija poteka, je predmet nadaljnjih odločitev, vezanih na namen posamezne raziskave.¹²³

¹²³ Pri zelo tipičnih vzorcih je predvidljivo, da bodo tipične zveze zasedle ustrezno visoka mesta v pogostnostnem seznamu, pri manj tipičnih vzorcih pa je nizka tudi pogostnost posameznih zvez. Ali posledično iz nadaljnje obravnave izpustiti celoten vzorčni tip, je najbrž vprašanje, vezano na namen posameznega luščenja.

Na tem mestu je možno skleniti, da luščenje besednozveznih podatkov iz oblikoskladenjsko označenega korpusa predstavlja dobro izhodišče za nadgradnjo jezikovnotehnoške leksikalne zbirke za slovenščino. Doktorsko delo predstavlja temelje metode in v zvezi s tem izpostavlja problematična mesta, ki potrebujejo nadaljnjo analizo. Kljub nekaterim v pričujočem poglavju izpostavljenim pomanjkljivostim opisane metode so rezultati – predvsem na ravni luščenja samostalniških besednih zvez, ki jim je bilo v raziskavi posvečene več pozornosti – spodbudni in pričajo o visokem potencialu uporabnosti tovrstnega luščenja podatkov za različne namene.

Doktorsko delo zaključujemo z navedkom Wolfganga Teuberta in Ramesha Krishnamurtyja iz uvoda v njun obsežni zbornik prispevkov na temo korpusnega jezikoslovja. Avtorja utemeljujeta, zakaj obdelava naravnega jezika ne more biti uspešna, dokler se trudi naravne jezike formalizirati, obenem pa opredeljujeta, kako in do katere mere lahko problem opisa nepredvidljivega naravnega jezika reši korpusno jezikoslovje. To vprašanje je, konec koncev, središčno tudi v pričujoči raziskavi.

»Formalni jeziki [...] so urejeni. Jasno ločujejo med tem, kaj je slovnično pravilen izraz in kaj ne. So rigidni. Naravni jeziki so pravo nasprotje. So anarhični, ne sledijo pravilom, stalno se spreminjajo, so nepredvidljivi. Vsak poskus izdelave sistema od zgoraj navzdol, ki bo spravil pod streho nekaj, kar se izogiba redu in je polno posebnosti, je obsojen na propad.

Korpusno jezikoslovje je jezikoslovje od spodaj navzgor, je jezikoslovje *parole*. Izhodiščna točka je vedno korpus, realni jezikovni podatki. Ali bo analiza vnesla red v anarhijo diskurza, je odprto vprašanje. Statistična analiza obsežnih korpusov bo našla ponavljajoče se vzorce in druge vrste verjetnosti. Lahko izmerimo statistično signifikantnost sopojavitev. Lahko opazujemo trende. Lahko odkrijemo regularnosti. Ampak opis tega, kar najdemo, ne bo nikoli proizvedel jezikovnega modela, ki bi bil preprostejši od kompleksnosti realnih jezikovnih podatkov.« (Teubert in Krishnamurty (ur.) 2007: 6–7, prev. Š. Arhar)

VIII

POVZETEK

I Doktorska raziskava združuje korpusnojezikoslovna izhodišča z metodami obdelave naravnega jezika, pri čemer se glede na cilj – avtomatsko pridobivanje leksikalnih podatkov za nadgradnjo jezikovnotehnoške leksikalne zbirke za slovenščino – uvršča na področje računalniške leksikografije.

I-1 Korpusno jezikoslovje je bolj kot z opozicijskega stališča do drugih jezikoslovnih tokov smiselno opredeljevati ob osredotočanju na izkušnje korpusnojezikoslovne prakse. Besedilni korpus, kombinacija avtentičnega besedilnega gradiva v elektronski obliki ter programskega orodja za obdelavo jezikovnih podatkov, je predpogoj za statistično obdelavo jezika, pri kateri sta v središču interesa sopojavljanje jezikovnih elementov ter njihova pogostnost. Glede na nove jezikoslovne ugotovitve, temelječe na korpusnih podatkih, imamo pojav korpusov v jezikoslovju lahko za resnično metodološko revolucijo, ne glede na to, ali korpus pomeni dopolnilo že obstoječih jezikoslovnih metod ali zaključeni jezikovni vir, ki je (edini lahko) osnova za postavitve novih hipotez o jeziku.

I-2 Obdelava naravnega jezika (ONJ) je na računalništvo vezana raziskovalna smer, ki jezik in jezikoslovne ugotovitve uporablja predvsem za (pol)avtomatsko pridobivanje raznovrstnih podatkov, potrebnih za razvoj računalniških aplikacij, ki so z jezikom povezane (jezikovnih tehnologij). ONJ je s korpusnim jezikoslovjem deloma primerljiva, saj se obe smeri poslužujeta podobnih metod pridobivanja podatkov iz jezikovnih virov, obe sta tudi naravnani k uporabnosti. Ločuje pa ju drugačen odnos do korpusa kot jezikovnega vira, saj za razliko od korpusnega jezikoslovja ONJ jezik predvsem obdeluje, ne pa tudi analizira ali interpretira. S tega stališča je korpus v ONJ enakovreden kateremu koli drugemu jezikovnemu viru.

I-3 Posebej velike spremembe v pogledu na jezik prinaša korpus na področje **leksikografije**. Statistična obdelava velike količine jezikovnega materiala se odraža v identifikaciji sopojavitvenih odnosov med besedami, med katerimi sta v središču interesa predvsem kolokacija ter koligacija: prva pomeni odnos med dvema besedama, za kateri je na osnovi izbrane statistične metode ugotovljena sopojavitvena povezava, druga poimenuje primerljiv odnos na ravni dveh slovničnih kategorij oz. jedrne besede ter slovnične kategorije.

Ugotovljeno je bilo tudi, da so vzorci sopojavljanja besed (slednji torej prinašajo tako tipično skladnjo kot nabor tipičnih leksikalnih kolokatorjev) neposredno povezani s pomenom, kar pomeni premik raziskovalnega interesa od obravnave pretežno enobesednih iztočnic ter nabora njihovih abstrahiranih pomenov k obravnavi daljših leksikalnih enot, ki so praviloma enopomenske. Pristop združuje slovar ter slovnico v enotno področje interesa, imenovano **leksikogramatika**.

Računalniška leksikografija združuje leksikografijo z novimi tehnologijami, pri čemer se osredotoča predvsem na razvoj postopkov za avtomatizacijo leksikografskega dela ter gradnjo leksikalnih zbirk za strojno na eni ter človeško rabo na drugi strani.

II Označevanje korpusa (tako lematizacija kot pripisovanje drugih vrst oznak) sicer pomeni interpretativni poseg v jezikovno realnost, ki je toliko večji, če so oznake pripisane avtomatsko, vendar na drugi strani omogoča izrabo označenih podatkov na ravni abstraktnejših kategorij. Doktorska raziskava raziskuje možnosti luščenja leksikalnih podatkov iz lematiziranega ter oblikoskladenjsko označenega besedilnega korpusa.

II-1.1 Vztrajanje pri obdelavi surovih, neoznačenih korpusnih podatkov, ki je značilno za zagovornike popolnega korpusnega pristopa, postane s širitvijo korpusnega jezikoslovja na področja morfološko bogatih jezikov, obenem pa z razvojem zmogljive programske opreme za delo z označenimi korpusi, manj smotno. Danes

korpusno jezikoslovje ločuje med **zunanji** (pred vključitvijo v korpus) ter **notranji** (po vključitvi ter obdelavi besedil) **korpusnimi podatki**.

II-1.2 Trenutno aktualni referenčni korpus za slovenščino, **korpus FidaPLUS**, je v celoti lematiziran ter oblikoskladenjsko označen, oba postopka sta bila izvedena na podjetju Amebis, d. o. o., Kamnik. Korpusne oznake so razdvoumljene na osnovi skladišne analize obravnavanega stavka, v korpusu pa so oznake dostopne tako v nerazdvoumljeni kot razdvoumljeni različici. Za označevanje korpusa FidaPLUS je bil uporabljen sistem oblikoskladišnega označevanja, ki je nastal v okviru projekta Multext-East. Kasneje, v sklopu projekta Jezikoslovno označevanje slovenščine (JOS), je bil ta sistem revidiran ter nadgrajen; prenovljeni sistem imenujemo nabor oblikoskladišnih oznak JOS.

V sklopu projekta JOS je potekla tudi evalvacija avtomatskega oblikoskladišnega označevanja slovenščine, pri čemer je bila za Amebisov označevalnik ugotovljena 85,7 % natančnost. Z vključitvijo statističnega oblikoskladišnega označevalnika TnT ter statistično primerjavo primerov označevanja enega ter drugega označevalnika, je bila uspešnost označevanja nekoliko izboljšana. Izboljšave so mogoče tudi z osredotočanjem na napake, ki se pri označevanju tipično pojavljajo, čemur se posveča tudi pričujoča raziskava.

II-2 Avtomatsko pridobivanje leksikalnih podatkov iz jezikovnega vira imenujemo **luščenje leksikalnih podatkov**. V pričujoči raziskavi uporabljena metoda je luščenje podatkov na osnovi oblikoskladišnih oznak.

II-2.1 Omenjena metoda je v slovenščini že preizkušena za luščenje samostalniških terminoloških besednih zvez. Luščenje poteka na osnovi vnaprej pripravljenih skladišnih vzorcev, ki prinašajo opredelitev sosledja besednih vrst v besedni zvezi ter njenega jedra. Podatki, izluščeni iz specializiranega korpusa, so primerjani s podatki referenčnega korpusa in glede na to z izbrano statistično formulo urejeni glede na terminološko relevantnost. V končnem koraku sledi pretvorba izluščenih nizov v končno besednozvezno obliko na osnovi korpusnih podatkov.

II-2.2 Luščenje v pričujoči raziskavi je deloma primerljivo z opisanim. Luščijo se dvodelni ter tridelni besedni nizi, pri čemer je glavno vodilo zaporednost besed ter pogostnost niza v izvornem korpusu. Pomenska analiza izluščenih podatkov ostaja zunaj interesa raziskave, zato se pri luščenju v isto skupino zajemajo tako proste kot stalne besedne zveze oz. tako neidiomatične kot idiomatične. Izhodišče za luščenje je enobesedna iztočnica (samostalniška, pridevniška, glagolska ali prislovna), za katero so v prvem koraku identificirane koligacije, nato pa na osnovi vzorcev, s katerimi koligacije opredeljujemo, izluščeni besedni nizi, v katerih iztočnica nastopa. Besedni nizi so urejeni glede na pogostnost, ki nam pomeni glavni indikator tipičnosti jezikovnih podatkov.

III Luščenje leksikalnih podatkov v pričujoči raziskavi je namenjeno posodobitvi leksikalne zbirke ASES z večbesednimi leksikalnimi enotami.

III-1 Leksikalna podatkovna zbirka prinaša na izbrani način formalizirane ter organizirane leksikalne podatke različnih vrst. Glede na namen podatkov za razvoj končnega produkta ločujemo jezikovnotehnološke leksikalne zbirke od slovarskih. Prve so namenjene računalniški obdelavi jezika, druge pa stremijo k za človeškega uporabnika zanimivemu jezikoslovnemu opisu. Leksikalna zbirka sama na sebi še ni končni produkt, ampak le podatkovni vir za izdelavo specifične jezikovne tehnologije oz. jezikovnega vira slovarskega tipa.

V slovenskem prostoru je poleg jezikovnotehnološke zbirke ASES, s katero se ukvarja pričujoče delo, potrebno omeniti **III-1.1** slovarsko leksikalno zbirko, ki nastaja v sklopu projekta Sporazumevanje v slovenskem jeziku, ter **III-1.2** gradnjo zbirke ontološkega tipa, tj. slovenskega wordnet.

III-2 Leksikalna zbirka **ASES** nastaja na podjetju Amebis za potrebe razvoja jezikovnih tehnologij za slovenščino (strojno prevajanje, slovnično pregledovanje, sinteza govora, programirani sogovornik itd.).

III-2.1 ASES prinaša okrog 846.700 **iztočnic petih različnih vrst**, in sicer skupaj za slovenščino ter ostale (za strojno prevajanje relevantne) jezike. Osnovni enoti zbirke sta *beseda* (prinaša podatke o oblikovnem delu jezikovnega znaka – nabor besednih oblik z oblikoskladenjskimi oznakami) ter *pomen* (prinaša podatke o pomenskem delu jezikovnega znaka, omogoča povezovanje med jeziki, opredeljevanje pomenskih odnosov med besedami ter kvalifikacijo iztočnice glede na rabo). Organizacijsko vlogo imajo iztočnice vrste *zveza* (prinašajo besedne zveze z obravnavano besedo) ter *skupina* (združujejo posamezne iztočnice, za katere je predvidena enotna obravnava). Iztočnice vrste *glagolska predloga* prinašajo podatke o glagolski vezljivosti.

III-2.2 Analiza nabora vnosov v zbirki ASES (*pajek, izdati, moder*) se osredotoča na trenutno stanje zbirke, pri čemer je na številnih mestih razvidna želja po vključevanju kompleksnejših leksikalnih podatkov, npr. stalnih besednih zvez, informacij o glagolski vezljivosti, podatkov o tipičnem in netipičnem pojavljanju besed ter različnih vrstah odnosov med besedami (pomenskih, besedotvornih, izvirajočih iz vedenja o svetu) itd. Ker je gradnja zbirke usmerjena v sprotno reševanje problemov razvoja posameznih jezikovnih tehnologij (ročno vnašanje podatkov je izredno zamudno), prinašajo določena mesta zbirke popolnejše informacije kot druga.

III-3 Za nadgradnjo zbirke ASES je v pričujočem delu predlagana metoda, osnovana na izrabi korpusnih podatkov.

III-3.1 Osnova nadgradnje zbirke je **referenčni korpus** slovenskega jezika, v nadaljevanju (ključno je neprekinjeno nadgrajevanje zbirke) pa je predvidena tudi uporaba drugih jezikovnih virov, npr. specializiranih korpusov, strojnوبرljivih slovarjev, interneta itd. S tega stališča je ključno ohranjanje informacije o jezikovnem viru v zbirki, kar omogoča ločeno urejanje ter prioretiziranje podatkov za potrebe razvoja specifične jezikovne tehnologije. Prvi korak pri nadgradnji tako predstavlja dopolnitev obstoječih iztočnic s podatki o pogostnosti v referenčnem korpusu in glede na to nadaljnja obdelava iztočnic, ki se v korpusu izkazujejo za najbolj tipične.

III-3.2 Temeljni za vključitev v leksikalno zbirko so koligacijski ter kolokacijski podatki o besedah, kar pomeni premik pozornosti od enobesedne k **večbesednim leksikalnim enotam**.

III-3.3 Podatke za vključitev v zbirko želimo iz korpusa pridobiti avtomatsko, nato pa jih prav tako avtomatsko organizirati v – kolikor je mogoče – berljivo in pregledno obliko za nadaljnjo ročno obdelavo.

IV-1 Središče interesa doktorske raziskave predstavlja **vzorec**: iz besedilnega korpusa pridobljena kombinacija izbrane leme in obkrožajočih oblikoskladenjskih oznak, skupaj s podatkom o pogostnosti pojavljanja vzorca v izvornem korpusnem viru. Del besedila, ki ga z uporabo vzorca izluščimo, imenujemo **vzorčna zapolnitev**.

Ker vzorci vsebujejo oblikoskladenjske oznake, ki so glede kategorij ter vrednosti, ki jih opredeljujejo, precej bogate, je v izogib razpršenosti izluščenih podatkov predlagano združevanje vzorcev v **vzorčne tipe**. Združevanje vzorcev in posledično uporaba posplošenih kategorij za luščenje se odraža tudi na ravni kompleksnosti obdelave podatkov, kjer postane ključno vprašanje, katere kategorije in vrednosti oznak upoštevati pri luščenju ter končnem prikazu podatkov in katerih ne. Deloma obdelane in organizirane podatke imenujemo **izluščeni besedni nizi**.

V raziskavi so besedni nizi natančneje analizirani, pri čemer so temeljna raziskovalna vprašanja vezana na relevantnost posamezne vrste izluščenih podatkov za vključitev v leksikalno zbirko. Obenem je pozornost posvečena tudi besednim nizom pripisanim oblikoskladenjskim oznakam, vendar le na mestih, kjer se slednje izkazujejo za neustrezne.

IV-2 V raziskavi je uporabljen **podkorpus** korpusa Fida+X, ki je različica korpusa FidaPLUS. Podkorpus, ki obsega okrog 26.800.000 lem, je lematiziran ter označen z naborom oblikoskladenjskih oznak JOS, sestavlja pa ga nabor odstavkov iz referenčnega korpusa, ki prinašajo vsaj eno s seznama za analizo izbranih besed. Izhodiščni

seznam prinaša 15 besed, od katerih jih je v raziskavi natančno analiziranih 7: *pajek*, *strasten*, *plesati*, *temeljito* ter *Slovenka*, *oven* in *debata*.

Opisani podkorporus se uporablja kot vir luščenja podatkov, izključno za pripravo **izhodiščnih seznamov vzorcev** pa se uporablja preoblikovana različica podkorporusa, ki vsebuje le izbrani nabor lem, oblikoskladenjske oznake ter oznake za ločila izvirnega korpusnega besedila. Podatki tako pripravljenega podatkovnega vira so urejeni s pomočjo programa Oxford WordSmith Tools. S funkcijo Clusters so glede na izbrane parametre izdelani po pogostnosti urejeni sezname skupov, vsebujočih eno od obravnavanih lem ter oblikoskladenjske oznake, ki se ob lemi pojavljajo. Primeri, znotraj katerih se pojavlja oznaka za ločilo, so odstranjeni s seznamov.

Naslednji korak raziskave je ročni pregled vzorčnih zapolnitev za **najpogostejše** (okrog 100 za vsak seznam) vzorce s štirimi lemmi, *pajek*, *strasten*, *plesati* ter *temeljito*. Pregled podatkov izkazuje, da je mogoče relevantnost vzorca za luščenje v določeni meri predvideti tako glede na vsebnost določene vrste oblikoskladenjskih oznak kot tudi glede na njihovo mesto znotraj vzorca. Posledično so iz nadaljnje analize izločeni vzorci, ki vsebujejo oznake za: pomožno glagolsko obliko, zaimek, okrajšavo, števniki ali členek. V analizo so zajeti primeri, ki vsebujejo oznake za: veznik, predlog ali polnopomenske besede.

Med nadaljnjo analizo so vzorci združeni v vzorčne tipe na osnovi redukcije oblikoskladenjske oznake na opredelitev besedne vrste. Vzorčni tipi so izhodišče za luščenje besednih nizov, analiza katerih se vrača k obravnavi ostalih v oznakah pripisanih oblikoskladenjskih kategorij. Glavni **cilji analize** so: (I) iskanje tipičnih napak besednim nizom pripisanih oznak, (II) poskus ugotovitve, katere kategorije oznak je pri luščenju ter prikazu podatkov smotrno upoštevati kot razločevalne in katere je smotrno združevati v izogib razprševanju podatkov, ter (III) klasifikacija vzorčnih tipov glede na (ne)relevantnost za luščenje.

IV-3 Med raziskavo je bilo v programskem jeziku Perl za obdelavo podatkov pripravljenih mnogo specializiranih **programskih skript** (23 osnovnih različic). Sprotno in problemsko usmerjeno nastajanje programov se odraža v njihovi neintegriteti, kar je pred nadaljevanjem luščenja podatkov znotraj širše zasnovanih raziskav potrebno odpraviti.

V-1 Najbolj zanimive rezultate prinaša luščenje besednih nizov, ki vsebujejo **same polnopomenske besede**.

V-1.1 Na ravni luščenja zvez **dveh samostalnikov** je smiselno ločevati med primeri z ujemalnimi ter neujemalnimi samostalniškimi prilastki, kar je avtomatsko mogoče z upoštevanjem oznake za sklon pri obeh samostalnikih, pri čemer so v primerih dvoumnosti na oblikovni ravni potrebni nadaljnji postopki urejanja podatkov. Analizirani vzorčni tipi so: *pajek* + Sam (*pajek skakač*), *pajek* + Sam₂ (*pajek širine*), *pajek* + Sam₃ (*pajek olivi*), *pajek* + Sam_{LI} (*pajek Pottinger*) ter Sam + *pajek* (*obračalnik pajek*), Sam + *pajek*₂ (*vrsta pajka*).

Na ravni luščenja zvez **samostalnika ter pridevnika** so zanimive samostalniške besedne zveze s pridevnikom, ki se ujema v spolu, sklonu ter številu, ter zveze pridevnika s samostalniškim dopolnilom. Analizirani vzorčni tipi so: *pajek* + Prid (*pajek dvovretenski*), Prid + *pajek* (*rdeč pajek*), *pajek*₃ + Prid (*pajku podoben*) ter Prid + *pajek*₃ (*podoben pajku*). Poleg naštetih sta analizirana še vzorčna tipa: *pajek* + Prisl (*pajek lahko*), Prisl + *pajek* (*lahko pajek*).

Kljub zavedanju da z luščenjem dvo- ter tridelnih vzorcev ni mogoče doseči zadovoljivega priklica zvez, ki posegajo na skladenjsko raven jezika (npr. med samostalnikom v vlogi osebkata ter glagolom v vlogi povedka), so vzorci, ki vsebujejo oznako za glagol, v raziskavi analizirani. Na ravni luščenja zvez **samostalnika ter glagola** se izkazuje potreba po urejanju podatkov glede na sklon samostalnika, za primer Glag + *pajek* npr. Glag + *pajek*₁ (*odpeljati pajek*), Glag + *pajek*₂ (*bati pajkov*), Glag + *pajek*₃ (*dati pajkom*), Glag + *pajek*₄ (*poklicati pajka*). Analiziran je tudi vzorčni tip *pajek* + Glag (*pajek odpeljati*).

V-1.2 Od primerov z besedo *pajek* ter dvema polnopomenskima besedama sta analizirana vzorčna tipa: Prid + Prid + *pajek* (*orjaški ptičji pajek*), Prid + *pajek* + Sam (*štirivretenski pajek sip*), relevantne podatke pa prinašajo tudi: *pajek* + Sam + Sam (*pajek Pottinger HIT*), Sam + *pajek* + Sam (*obračalnik pajek SIP*) ter Prid + Sam + *pajek* (*resnični mož pajek*).

V-1.3 Pri luščenju samostalniških besednih zvez s pridevnikom je potrebno ločevanje primerov glede na spol samostalnika: *strasten* + Sam_m (*strasten kadilec*), *strastna* + Sam_z (*strastna noč*), *strastno* + Sam_s (*strastno razmerje*). Drugi analizirani vzorčni tipi so: Sam + *strasten* (*Slovenci strastni*), *strasten* + Prid (*strasten ljubezenski*), Prid + *strasten* (*nov strasten*), Prisl + *strasten* (*najbolj strasten*) ter Glag + *strasten* (*preživeti strasten*).

V-1.4 Od primerov z besedo *strasten* ter dvema polnopomenskima besedama so analizirani vzorčni tipi: Prisl + *strasten* + Sam (*najbolj strasten kadilec*), Glag + *strasten* + Sam (*preživeti strastno noč*), *strasten* + Prid + Sam (*strasten ljubezenski prizor*), *strasten* + Sam + Prid (*strasten zbiratelj starih*) ter *strasten* + Sam + Sam (*strasten igralec golfa*).

V-1.5 Od primerov kombinacij besede *plesati* s še eno polnopomensko besedo so analizirani vzorčni tipi: Prisl + *plesati* (*dobro plesati*), *plesati* + Prisl (*plesati skupaj*), *plesati* + Prid (*plesati folklorne*), Prid + *plesati* (*sam plesati*), Glag + *plesati* (*začeti plesati*), *plesati* + Sam (*plesati tango*) ter Sam + *plesati* (*miši plesati*).

V-1.6 Primer vzorčnega tipa s še dvema polnopomenskima besedama pa je: *plesati* + Prid + Sam (*plesati klasični balet*).

V-1.7 Pri luščenju **zvez s prislovi** se kaže za smiselne premislek o dodatnem označevanju vrste prislova. Od primerov kombinacij besede *temeljito* s še eno polnopomensko besedo so analizirani vzorčni tipi: Prisl + *temeljito* (*tako temeljito*), *temeljito* + Prisl (*temeljito strokovno*), *temeljito* + Glag (*temeljito spremeniti*), Glag + *temeljito* (*opraviti temeljito*), Sam + *temeljito* (*sestavine temeljito*), *temeljito* + Sam (*temeljito čiščenje*) ter *temeljito* + Prid (*temeljito prenovljen*).

V-1.8 Analizirani tridelni vzorčni tipi pa so: Prisl + *temeljito* + Glag (*skupaj temeljito premešati*), Glag + *temeljito* + Glag (*morati temeljito premisliti*), *temeljito* + Glag + Sam (*temeljito umiti roke*), Sam + *temeljito* + Glag (*kožo temeljito očistiti*) ter Prid + Sam + *temeljito* (*prihodnje leto temeljito*).

V-2 Med dvodelnimi zvezami, ki vsebujejo **predlog**, so najbolj zanimive kombinacije z glagolom na levi, npr. *plesati* + Pred (*plesati v*), določen nabor relevantnih zvez pa je mogoče dobiti tudi s kolokacijsko analizo predloga, npr. *po* + Sam_s: (*po besedah*). Pri obravnavi tridelnih vzorčnih tipov so analizirani: *pajek* + Pred + Sam (*pajek za seno*), Sam + Pred + *pajek* (*odvoz s pajkom*), Glag + Pred + *pajek* (*odpeljati s pajkom*), Pred + *strasten* + Sam (*v strastno razmerje*), *strasten* + Sam + Pred (*strasten odnos do*), *plesati* + Pred + Sam (*plesati z volkovi*), *plesati* + Pred + Prid (*plesati do zgodnjih*), Prisl + *plesati* + Pred (*vedno plesati z*), Pred + Sam + *plesati* (*z veseljem plesati*), Pred + Sam + *temeljito* (*pred uporabo temeljito*) ter *temeljito* + Glag + Pred (*temeljito pripraviti na*).

V-3 V zvezi z luščenjem zvez, ki vsebujejo **veznik**, so najbolj zanimivi primeri, ki prinašajo dve istovrstni besedni vrsti, povezani s prirednim veznikom: Sam + Vp + *pajek* (*kača in pajek*), *pajek* + Vp + Sam (*pajek in škorpion*), Prid + Vp + *strasten* (*romantičen in strasten*), *strasten* + Vp + Prid (*strasten in čustven*), Glag + Vp + *plesati* (*peti in plesati*), *plesati* + Vp + Glag (*plesati in igrati*), *temeljito* + Vp + Prisl (*temeljito in strokovno*), Prisl + Vp + *temeljito* (*hitro in temeljito*).

V-4 Vzorci, ki prinašajo oblikoskladenjsko oznako za **pomožni glagol**, **členek**, **okrajšavo**, **zaimek** oz. **števnik**, se izkazujejo za problematične na ravni nezadostnega priklica želenih leksikalnih podatkov z opisano metodo ali na

ravni slabe avtomatske pripisljivosti obstoječih oblikoskladenjskih kategorij korpusnim besedilom. Našteti vzorci (predstavljajo dobrih 40 % vseh) ostajajo v pričujoči raziskavi ob strani.

V-5 Prav tako niso natančneje analizirani vzorci, vsebujoči oznako za **nelematizirano besedo**, saj v prvi vrsti zahtevajo analizo s stališča razvoja označevalnega sistema in ne vnosa podatkov v leksikalno zbirko. Velik delež nelematiziranih besed odpade na lastna imena (*Clug pleše*), tuje besede (*pajek umbrella*), neustrezno tokenizirane (*9.30Mož pajek*) oz. zatipkane besede (*manjdlje temeljito*), neknjižne besede (*strasten snifač*), dosti pa je tudi zvez oz. besed, ki niso označene, ker v leksikalno zbirko do sedaj še niso bile vključene (*plesati sirtaki*).

V-6 Za samostalniki *pajek* so nadalje analizirani še vzorci, ki zaradi nižje pogostnosti na prejšnjih mestih niso prišli v obravnavo. Za vključitev v leksikalno zbirko relevantne rezultate prinašajo naslednji skladenjski vzorci: Sam + Prid + Sam (različni tipi zvez glede na ujemanje v sklonu, npr. *pajek črna vdova*, *razvoj rdečega pajka*, *pajku podoben moški*), Sam + Prisl + Prid (*pajek grozno nevaren*), Sam + Sam + Prid (*pajek SIP dvovretenski*), Sam + Prid + Prid (*pajek dvovretenski italijanski*), Pred + Sam + Sam (*v obliki pajka*), Prid + Pred + Sam (*nor na pajke*), Prisl + Pred + Sam (*skupaj s pajki*), Prisl + Prisl + Sam (*kar nekaj pajkov*) ter primeri, vsebujoči podredni veznik: Sam + Vd + Sam (*pajek kot sredstvo*), Prid + Vd + Sam (*strupen kot pajek*) ter Prisl + Vd + Sam₁ (*urno kot pajek*).

VI-1 Označevalne napake se v analiziranih besednih nizih pojavljajo na ravni pripisovanja napačne besedne vrste, najpogosteje označevanje prislova za pridevnik (*pajki mnogo predejo*) ali pridevnika za prislov (*temeljito strokovno pripravo*) ter označevanje funkcijskih besed za samostalnike (*pajek kot po čudežu izgine*), problem je tudi ločevanje med prislovi in enakopisnimi predlogi (*plesati okoli roke*). Obenem so pogosti problemi pripisovanja napačnih oblikoskladenjskih kategorij, npr. pripisovanje napačnega spola ali sklona znotraj besednih zvez itd. Našteti problemi se zdijo rešljivi z nadgradnjo avtomatskega označevanja z vključitvijo kolokacijskih ter koligacijskih besednozveznih informacij, k izboljšavi kvalitete označevanja bi veliko prispevalo že dodajanje na koligacijskih podatkih temelječih pravil na mesto razdvoumljanja lem oz. oblikoskladenjskih oznak.

Težko rešljivi so problemi, vezani na označevanje lastnih imen (*pajek Pottinger*, *strastni Dmitrij*), tujih besed (*strasten love/hate odnos*) ali v zbirki še neobstoječih slovenskih besed (*plesati s telebajski*). Besede, ki med procesom označevanja ostanejo nelematizirane, se zapišejo v poseben seznam ter naknadno dodajo v leksikalno zbirko. Pred razvojem kvalitetnega sistema za avtomatsko prepoznavo lastnih imen v slovenščini se kaže kot možna začasna rešitev vključevanje v zbirko tudi najpogostnejših lastnoimenskih besed oz. zvez.

Posebno pozornost za nadaljnji razvoj avtomatskega označevanja zahteva **označevanje členkov**, saj gre za kategorijo, ki je na oblikoskladenjski ravni – zaradi enakopisnosti členkov s prislovi ter vezniki – avtomatsko težko pripisljiva. V zvezi s tem je predlagan premislek o označevanju členkov šele na skladenjskem nivoju označevanja. Premislek zahteva tudi **označevanje okrajšav**, kjer je možno ločevati med skupinami, zajemljivimi v leksikon (npr. merske enote), ter nepredvidljivimi, za označevanje katerih bi bilo potrebno razviti specializirane metode avtomatske identifikacije v besedilu.

VI-2.1 V končni obliki so vzorčni tipi s primeri besednih nizov strnjeni v treh tabelah. **Za luščenje relevantnih vzorčnih tipov** je 31, glede na izluščene podatke so razvrščeni na: samostalniške besedne zveze (17), pridevniške besedne zveze (3), priredne besedne zveze (4) ter pogojno zanimive zveze (7).

Za luščenje nerelevantne vzorčne tipe delimo glede na sestavo na vzorčne tipe: s prislovom na koncu vzorca ali pred samostalnikom, s pridevnikom na koncu vzorca, s predlogom na začetku ali koncu vzorca, s prirednim veznikom med različnima besednima vrstama ali na začetku oz. koncu vzorca ter drugo. Glede na vsebnost oznake so v tabelo uvrščeni tudi vzorci, ki vsebujejo oznako za pomožni glagol, členek, okrajšavo, zaimek, števniki ali nelematizirano besedo.

Vzorčni tipi, ki vsebujejo glagol, so predstavljeni ločeno zaradi pričakovane višje stopnje posega na skladenjski nivo jezika, kar posledično pomeni manjšo ustreznost obravnavane metode za pridobivanje teh podatkov. Delijo se glede na to, kaj se v izluščenem besednem nizu ob glagolu pojavlja: samostalni¹ oz. samostalniška besedna zveza, prislov oz. prislovna zveza, glagol, predlog oz. predložna zveza ali pridevnik.

VI-2.2 Luščenje samostalniških besednih zvez je preverjeno še za leme *Slovenka*, *oven* in *debata*. V celoti relevantne rezultate prinaša luščenje zvez s **samostalniškim prilastkom**. Na ravni dvodelnih vzorcev se kažeta za bolj tipični kombinacija dveh v sklonu ujemajočih se samostalnikov (*mati Slovenka-26*) ter zveza z drugim samostalnikom v rodilniku (*Slovenka leta-560*). Redke so pojavitve samostalniških zvez z drugim samostalnikom v dajalniku (*podpora Slovenkam-1*) ter primeri kombinacij z desnim neujemalnim imenovalniškim prilastkom (*revija Slovenka-2*). Tudi luščenje zvez samostalnika in samostalniške besedne zveze prinaša relevantne besedne zveze za vse nove samostalniške leme (*naslov Slovenka leta-39*).

Enako uspešno je luščenje zvez samostalnika z **levim pridevniškim prilastkom** (*plemensi oven-82*). Tridelne zveze so v skladu s pričakovanji redkejšje, vendar nabor primerov kljub temu kaže na dobre rezultate. Problemi se pojavljajo na ravni prikaza izluščenih podatkov, ki prinaša pridevnike, lematizirane v nezaznamovano pridevniško obliko (*razen težek debata*). Nasprotno pa luščenje besednih zvez s **pridevniškim prilastkom na desni** prinaša redke in manj zanimive rezultate (*oven plemensi-2*).

Luščenje zvez s **predlogom** na eni strani omogoča pridobitev zvez, ki so za nadaljnjo obravnavo zanimive (*po rodu Slovenka*, *v znamenju ovna*), na drugi strani na seznamu najdemo nekončane zveze (*z ovni pasme*). Redke so zveze s **podrednim veznikom** (*debata kot znanost-1*), nasprotno pa izredno pogoste zveze s **prirednim veznikom** (*Slovenec in Slovenka-635*).

VII V raziskavi je bila uporabljena in s tem preverjana metoda luščenja leksikalnih podatkov na osnovi dvodelnih ter tridelnih kombinacij obravnavane leme z oblikoskladenjskimi oznakami neposredne besedilne okolice. Največji problem metode je dolgotrajnost obdelave podatkov. Za luščenje podatkov iz celotnega referenčnega korpusa je predvidena izdelava celovitega avtomatiziranega postopka, ki na čim bolj enostaven način združuje vse v pričujoči raziskavi opisane elemente (iskanje relevantnih besednih zvez, razvrščanje in štetje zvez, ustrezen prikaz besednih zvez).

Glede same kvalitete izluščenih podatkov se potrjuje predvidevanje, da so rezultati relevantnejši na mestih, kjer slovenski jezik na skladenjski ravni v okvirih dvodelnih ali tridelnih besednih nizov izkazuje dovoljšno stopnjo regularnosti – metoda je zanesljivejša na ravni luščenja besednih zvez s samostalniškim, pridevniškim ali prislovnim jedrom, na ravni vzorcev, ki vsebujejo glagolsko lemo oz. oblikoskladenjsko oznako, pa je zanesljivost z opisano metodo izluščenih podatkov nižja. Druga načelna ugotovitev je, da luščenje zvez daje dobro sliko tipičnega v jeziku, slabše pa je na ravni manj pogostnih vzorčnih tipov.

Z upoštevanjem dejstva, da so za nadaljnji razvoj metode potrebne raziskave na večjem naboru gradiva, je možno skleniti, da luščenje besednozveznih podatkov iz oblikoskladenjsko označenega korpusa predstavlja dobro izhodišče za nadgradnjo jezikovnotehnološke leksikalne zbirke za slovenščino. Rezultati – predvsem na ravni luščenja samostalniških besednih zvez, ki jim je bilo posvečene v raziskavi več pozornosti – so spodbudni, in pričajo o visokem potencialu uporabnosti tovrstnih izluščenih podatkov za različne namene.

IX ABSTRACT

I The doctoral research combines a corpus-linguistic point of departure with methods of natural language processing. In view of the goal – the automatic acquisition of lexical data for upgrading the lexical database of Slovene for language technology applications – the research positions itself in the field of computer lexicography.

I-1 Rather than adopting a position of opposition towards other linguistic streams, it is more meaningful to define **corpus linguistics** as focusing on the experience of corpus-linguistic practice. The corpus, a combination of authentic text material in electronic form along with the software tools for processing language data, is a precondition for the statistical processing of language in which the focus of interest is the co-occurrence of language elements and their frequency. In view of new linguistic findings based on corpus data, the phenomenon of corpora in linguistics can be understood as a genuine methodological revolution, irrespective of whether the corpus is seen as a supplement to existing linguistic methods or an integrated language source which is (or is the only thing that can be) the basis for the establishment of new hypotheses about language.

I-2 Natural Language Processing (NLP) is a computer science related research area that uses language and linguistic findings primarily for the (semi)automatic acquisition of the diverse data necessary for the development of computer applications linked with language (language technologies). To some extent, NLP is comparable with corpus linguistics, as both areas make use of similar methods of gathering data from language sources, and both share an orientation towards utility. They are distinguished from each other, however, by a different relationship to the corpus as a language source, as, in contrast to corpus linguistics, NLP primarily processes language rather than analysing and interpreting it. From this perspective, in NLP the corpus is of equal value to any other language source.

I-3 The corpus brings particularly significant changes in the view of language to the area of **lexicography**. The statistical processing of large quantities of language material is reflected in the identification of co-occurrence relationships between words, amongst which the focus of interest is primarily on collocation and colligation: the first referring to the relationship between two words for which a co-occurrence connection is determined on the basis of selected statistical methods, while the second refers to a comparable relationship between two grammatical categories or a word and a grammatical category.

It has also been determined that patterns of the co-occurrence of words (providing both typical syntax and a set of typical lexical collocates) are directly linked with meaning, which represents a shift in research interest from the treatment of predominantly single-word headwords and a set of their abstracted meanings to the treatment of longer lexical units, which are typically monosemous. This approach combines lexicon and grammar in a unified field of interest, called **lexicogrammar**.

Computer lexicography combines lexicography with new technologies, focusing primarily on the development of procedures for the automation of lexicographical work and the construction of lexical databases for machine use, on the one hand, and human use, on the other.

II Annotating the corpus (both lemmatisation and the attributing of other types of annotation) represents an interpretive intervention into language reality, which is all the greater if the annotations are assigned automatically, although, on the other hand, this enables the use of the annotated data on the level of more abstract categories. The present doctoral research investigates the possibilities of extracting lexical data from a lemmatised and morphosyntactically tagged corpus.

II-1.1 With the spread of corpus linguistics to the areas of morphologically rich languages, accompanied by the development of powerful software for working with annotated corpora, insisting on the treatment of raw, unmarked corpus data, which is characteristic of advocates of an entirely corpus-driven approach, becomes less reasonable. Today, corpus linguistics distinguishes between **external** (prior to inclusion in the corpus) and **internal** (subsequent to the inclusion and processing of the texts) **corpus data**.

II-1.2 The current reference corpus for Slovene, **the FidaPLUS corpus**, is entirely lemmatised and morphosyntactically tagged, both procedures having been executed by the Amebis, d. o. o., Kamnik company. The corpus annotations are disambiguated based on the syntactic analysis of the processed sentence, while in the corpus tags are accessible both in non-disambiguated and disambiguated versions. For annotating the FidaPLUS corpus a system of morphosyntactic tags was used that came about within the framework of the project Multext-East. Later, in conjunction with the project Linguistic Annotation of Slovene (Slo.: Jezikoslovno označevanje slovenščine: JOS), the system was revised and upgraded; the renovated system is called the JOS morphosyntactic tagset.

In conjunction with the JOS project an evaluation of the automatic morphosyntactic tagging of Slovene was also undertaken, in which a level of accuracy of 85.7 % was determined for the Amebis tagger. With the inclusion of the TnT statistical morphosyntactic tagger and a statistical comparison of examples annotated by each of the taggers the tagging accuracy was somewhat improved. Further improvements are also possible by focusing on the errors that typically appear in tagging, something to which attention is also devoted in the present research.

II-2 The automatic gathering of lexical data from a language source is called **the extraction of lexical data**. In the present research the method employed is the extraction of data on the basis of morphosyntactic tags.

II-2.1 This method has already been tested in Slovene for extracting terminological noun phrases. Extraction proceeds on the basis of pre-prepared syntactic patterns, which provide a definition of the sequence of the parts of speech in the phrases and of the heads of the phrases. Data extracted from a specialized corpus are compared with data from the reference corpus and on this basis are ordered with a selected statistical formula in terms of terminological relevance. In the final step there follows the transformation of the extracted word sequence into the final form of the phrase on the basis of corpus data.

II-2.2 In the present research extraction is partially comparable with the described method. Bigrams and trigrams of words are extracted, the main criteria being the succession of the words and the frequency of the sequence in the corpus. An analysis of the meaning of the extracted data remains outside the interest of the research, thus the extraction captures both free and set phrases, and both non-idiomatic and idiomatic phrases, in the same group. The point of departure for extraction is a single-word headword (noun, adjective, verb or adverb) for which colligations have been identified in the first step, and then, on the basis of the patterns with which the colligations are defined, word sequences in which the headword appears are extracted. The word sequences are arranged with regard to their frequency, which represents the main indicator of the typicality of the language data.

III The extraction of lexical data in the present research is for the purpose of upgrading the ASES lexical database with multi-word lexical units.

III-1 A **lexical database** consists of lexical data of various kinds, formalised and organised on the basis of the selected method. With regard to the purpose of the data for the development of the final product we distinguish language-technological lexical databases (LT lexical databases) from dictionary lexical databases. The former are aimed at the computer processing of language, while the latter strive for a linguistic description of interest to the human user. The lexical database is not in itself a finished product, but simply a data source for the development of a specific language technology or a dictionary-type language resource.

In the Slovene sphere, in addition to the ASES LT database, which is treated in the present work, it is also necessary to mention **III-1.1** the dictionary lexical database designed within the project Communication in Slovene, and **III-1.2** the database of the ontological type, the Slovene Wordnet.

III-2 The **ASES** lexical database was devised by the company Amebis for the development of language technologies for Slovene (machine translation, grammar checking, speech synthesis, chatbot, etc.).

III-2.1 ASES has around 846,700 **headwords of five different types**, both for Slovene and other languages (relevant for machine translation). The basic units of the database are the *word* (providing data about the formal part of the language sign – a set of word forms with morphosyntactic tags) and *meaning* (providing data about the semantic part of the language sign, enabling connection between languages, defining the semantic relationships between words and qualifying headwords regarding their use). An organisational role is performed by headwords of the type *phrase* (providing phrases with the treated word) and *group* (combining individual headwords for which a unified further treatment is foreseen).

III-2.2 The analysis of a set of entries in the ASES database (*spider, to betray, wise* – Slo.: *pajek, izdati, moder*) focuses on the current state of the database, in numerous places of which there is an evident desire to include more complex lexical data, e.g., set phrases, information about verb subcategorization, data about the typical and atypical co-occurrence of words and various kinds of relationships between words (semantic, word-forming, world knowledge based), etc. Due to the fact that the construction of the database is oriented towards the ongoing solution of the problems of the development of individual language technologies (entering data manually is extremely time consuming), some parts of the database provide more complete information than others.

III-3 In the present work, the suggested method for upgrading the ASES database is based on the use of corpus data.

III-3.1 The basis for upgrading the database is the **reference corpus** of Slovene language. In the future (the constant upgrading of the database is crucial), the use of other language resources is also foreseen, e.g., specialised corpora, machine-readable dictionaries, the Internet, etc. From this perspective it is critical to retain the information about the language source in the database, as it enables the separate ordering and prioritisation of data for the needs of the development of specific language technologies. The first step in upgrading is, therefore, supplementing the existing headwords with the reference-corpus frequencies, followed by the further processing of those headwords that prove to be the most typical in the corpus.

III-3.2 The most important information for inclusion in the lexical database is colligational and collocational data about words, which represents a shift from focusing on single-word lexical units to **multi-word lexical units**.

III-3.3 We seek to acquire the data for inclusion in the database automatically, and then to automatically organise these data – as far as possible – into a readable and clear form for future manual processing.

IV-1 The focus of interest of the present doctoral research is the **pattern**: the combination of the selected lemma and the surrounding morphosyntactic tags acquired from the corpus, together with data about the frequency of appearance of the pattern in the corpus source. The part of the text extracted by using the pattern we call the **pattern content**.

Due to the fact that patterns contain morphosyntactic tags, which are rather rich in terms of the categories and values they define, the combining of patterns into **pattern types** is suggested in order to avoid the dispersion of the extracted data. The combining of patterns and the consequent use of generalised categories for extraction is also reflected on the level of the complexity of the processing of the data, where the question of which

categories and values of tags are to be taken into account (and which are to be disregarded) during the extraction and final presentation of the data becomes critical. We call the partially treated and organised data **extracted word sequences**.

In the research word sequences are more precisely analysed, with the fundamental research questions being related to the relevance of the individual type of extracted data for inclusion in the lexical database. At the same time, attention is also devoted to the morphosyntactic tags attributed to the word sequences, but only in cases where these tags prove to be unsuitable.

IV-2 The **subcorpus** of the Fida+X corpus is used in the research, the latter being a variation of the FidaPLUS corpus. The subcorpus, which has a scope of around 26,800,000 lemmata, is lemmatised and annotated with the JOS morphosyntactic tags, and is made up of a set of paragraphs from the reference corpus that provide at least one of the words from the word list selected for analysis. The initial list provides 15 words, of which the research undertook a precise analysis of 7: *spider*, *passionate*, *to dance*, *thoroughly*, as well as *Slovene* (female), *ram* and *debate* – Slo.: *pajek*, *strasten*, *plesati*, *temeljito*, *Slovenka*, *oven*, *debata*.

The subcorpus described is used as a source for the data extraction, but exclusively for the preparation of the **initial list of patterns** a transformed version of the subcorpus is subsequently used, containing only the selected list of lemmata, the morphosyntactic tags and the punctuation tags of the original corpus text. The data from the data source prepared in this way are organised with the help of the Oxford WordSmith Tools program. The frequency-ordered lists of clusters containing one of the treated lemmata and the morphosyntactic tags that appear along with the lemma are created with regard to the selected parameters using the 'Clusters' function of the program. Cases including punctuation tags are excluded from further analysis.

The next step of the research is a manual review of the pattern content for the **most frequent** patterns (around 100 for each list) with four lemmata: *spider*, *passionate*, *to dance* and *thoroughly* (Slo.: *pajek*, *strasten*, *plesati* and *temeljito*). A review of the data shows that to a certain extent it is possible to predict the relevance of the pattern for extraction with regard to the presence of particular kinds of morphosyntactic tags, as well as to their position within the pattern. Consequently, patterns containing the following tags are excluded from further analysis: auxiliary verb form, pronoun, abbreviation, numeral or particle. Analysis therefore focuses on the patterns containing: conjunctions, prepositions and content words.

In the continuation of the analysis patterns are grouped into pattern types on the basis of a reduction of the morphosyntactic tags to part-of-speech definitions. Pattern types are a starting point for the extraction of word sequences, the analysis of which returns to the treatment of the other ascribed morphosyntactic categories. The main **goals of the analysis** are: (I) finding typical errors in the annotation of the word sequences, (II) attempting to determine which categories of annotation can be reasonably counted as differentiating in the extraction and presentation of data and which should be counted as unifying in terms of avoiding the dispersion of data and (III) classification of pattern types regarding their (ir)relevancy for data extraction.

IV-3 During the research many specialised short **programs** (23 basic versions) for processing data were prepared using the Perl programming language. The ongoing and problem-oriented emergence of the programs is reflected in their lack of integration, which is something that needs to be resolved before continuing the extraction of data within more broadly conceived research.

V-1 The most interesting results were provided by the extraction of word sequences containing **only content words**.

V-1.1 On the level of the extraction of **two-noun phrases** it is meaningful to differentiate between examples with accordant and non-accordant noun attributes, which can be done automatically by taking into account the

tags for the case of both nouns. In cases of formal ambiguities further procedures for ordering the data are necessary. The analysed pattern types are¹²⁴: *spider* + N (*spider jumper*), *spider* + N₂ (*spider [of] width*), *spider* + N₃ (*spider [to the] olive*), *spider* + N_p (*spider Pottinger*) and N + *spider* (*tedder spider*), N + *spider*₂ (*type [of] spider*).

Of interest on the level of the extraction of phrases of **noun and adjective** are noun phrases with an adjective that agrees in gender, case and number, as well as adjectival phrases with a noun complement. The analysed pattern types are: *spider* + Adj (*spider two-spindled*), Adj + *spider* (*red spider*), *spider*₃ + Adj ([to a] *spider similar*) and Adj + *spider*₃ (*similar [to a] spider*). In addition to these, the following pattern types were also analysed: *spider* + Adv (*spider lightly*), Adv + *spider* (*lightly spider*).

In spite of being aware that with the extraction of two-part and three-part patterns it is not possible to achieve a satisfactory retrieval of phrases that address the syntactic level of language (e.g., combinations of a noun in the role of the subject and a verb in the role of the predicate), patterns that contain a verb tag are nonetheless analysed in the research. On the level of the extraction of phrases of **noun and verb**, the need arose to order the data with regard to the case of the noun, for example V + *spider*: V + *spider*₁ (*to tow away spider*), V + *spider*₂ (*to fear spiders*), V + *spider*₃ (*to give [to] spiders*), V + *spider*₄ (*to call a spider*). In addition to these, the following pattern type is also analysed: *spider* + V (*spider to tow away*).

V-1.2 From the examples with the word *spider* and two content words the following pattern types are analysed: Adj + Adj + *spider* (*giant bird spider*), Adj + *spider* + N (*four-spindled spider SIP*). Relevant data is also provided by: *spider* + N + N (*spider Pottinger HIT*), N + *spider* + N (*tedder spider SIP*) and Adj + N + *spider* (*genuine man spider*).

V-1.3 In the extraction of noun phrases with an adjective it is necessary to differentiate between examples with regard to the gender of the noun: *passionate* + N_m (*passionate smoker*), *passionate* + N_f (*passionate night*), *passionate* + N_c (*passionate affair*). Other analysed pattern types are: N + *passionate* (*Slovenes passionate*), *passionate* + Adj (*passionate love*), Adj + *passionate* (*new passionate*), Adv + *passionate* (*most passionate*) and V + *passionate* (*to live passionate*).

V-1.4 From the examples with the adjective *passionate* and two content words the following pattern types are analysed: Adv + *passionate* + N (*most passionate smoker*), V + *passionate* + N (*to spend a passionate night*), *passionate* + Adj + N (*passionate love scene*), *passionate* + N + Adj (*a passionate collector [of] old*) and *passionate* + N + N (*a passionate player [of] golf*).

V-1.5 From the examples of combinations of the verb *to dance* with one other content word the following pattern types are analysed: Adv + *to dance* (*well dance*), *to dance* + Adv (*to dance together*), *to dance* + Adj (*to dance folkloric*), Adj + *to dance* (*alone to dance*), V + *to dance* (*to begin to dance*), *to dance* + N (*to dance the tango*) and N + *to dance* (*mice dance*).

V-1.6 An example of a pattern type with two other content words is: *to dance* + Adj + N (*to dance classical ballet*).

V-1.7 In the extraction of **phrases with adverbs** consideration of the additional annotation of the adverbial types would seem to be sensible. From the examples of combinations of the word *thoroughly* with one other content word the following pattern types are analysed: Adv + *thoroughly* (*so thoroughly*), *thoroughly* + Adv

¹²⁴ For an explanation of the pattern-type format see Appendix 4. Examples are translated from Slovene as literally as possible to retain the structure, even though it is clearly impossible to grasp the actual phrase meaning in such a way (especially due to the polysemy of some of the chosen example words, e.g., *pajek* – spider/tow truck/tedder etc.).

(*thoroughly professionally*), *thoroughly* + V (*to thoroughly change*), V + *thoroughly* (*to settle thoroughly*), N + *thoroughly* (*components thoroughly*), *thoroughly* + N (*thoroughly cleansing*) and *thoroughly* + Adj (*thoroughly renovated*).

V-1.8 The analysed three-part pattern types are: Adv + *thoroughly* + V (*to together thoroughly mix*), V + *thoroughly* + V (*to have to thoroughly consider*), *thoroughly* + V + N (*to thoroughly wash hands*), N + *thoroughly* + V (*skin to thoroughly clean*) and Adj + N + *thoroughly* (*next year thoroughly*).

V-2 Amongst the two-part phrases that contain a **preposition** the combinations with the verb on the left are the most interesting, e.g., *to dance* + P (*to dance in*). It is also possible to obtain a particular set of relevant phrases with a collocation analysis of the preposition, e.g., *after* + N₅: (*after words*). In treating three-part pattern types the following are analysed: *spider* + P + N (*spider for hay*), N + P + *spider* (*removal with a spider*), V + P + *spider* (*to tow away with a spider*), P + *passionate* + N (*in a passionate relationship*), *passionate* + N + P (*a passionate attitude towards*), *to dance* + P + N (*to dance with wolves*), *to dance* + P + Adj (*to dance until early*), Adv + *to dance* + P (*to always dance with*), P + N + *to dance* (*with pleasure to dance*), P + N + *thoroughly* (*before use thoroughly*) and *thoroughly* + V + P (*to thoroughly prepare for*).

V-3 In analysing extracted phrases that contain a **conjunction**, of most interest are examples that provide two words of the same word class connected with the coordinated conjunction: N + Cc + *spider* (*snake and spider*), *spider* + Cc + N (*spider and scorpion*), Adj + Cc + *passionate* (*romantic and passionate*), *passionate* + Cc + Adj (*passionate and emotional*), V + Cc + *to dance* (*to sing and dance*), *to dance* + Cc + V (*to dance and play*), *thoroughly* + Cc + Adv (*thoroughly and professionally*), Adv + Cc + *thoroughly* (*quickly and thoroughly*).

V-4 Patterns that provide morphosyntactic tags with **an auxiliary verb, a particle, an abbreviation, a pronoun or a numeral** prove to be problematic on the level of the insufficient retrieval of the desired lexical data with the described method, or on the level of a low ascribability of the existing morphosyntactic categories to the corpus texts. These patterns (representing more than 40 % of the total pattern count) are put aside in the present research.

V-5 Similarly, patterns containing the tag for **a non-lemmatised word** are also not analysed in the present research, as in the first place they demand an analysis from the point of view of the development of the annotation system and not from that of the entry of data in the lexical database. A large proportion of the non-lemmatised words are proper names (*Clug dances*), foreign words (*spider *umbrella**), unsuitably tokenized words (*9.30Man spider*) or mistyped words (*almonsd thoroughly*), non-literary words (*passionate sniffa*), while a certain number of examples are not annotated because they have not been included in the database so far (*to dance the sirtaki*).

V-6 For the noun *spider* there is also a further analysis of patterns that were not treated previously due to their low frequency. The following syntactic patterns provided relevant results for inclusion in the lexical database: N + Adj + N (various types of phrases with regard to agreement in case, e.g., *spider black widow*, *development [of the] red spider*, *[to a] spider similar man*), N + Adv + Adj (*spider terribly dangerous*), N + N + Adj (*spider SIP two-spindled*), N + Adj + Adj (*spider two-spindled Italian*), P + N + N (*in the form [of a] spider*), Adj + P + N (*crazy about spiders*), Adv + P + N (*together with spiders*), Adv + Adv + N (*quite some spiders*) and examples containing a subordinate conjunction: N + Cs + N (*spider as a means*), Adj + Cs + N (*poisonous as a spider*) and Adv + Cs + N₁ (*quickly as a spider*).

VI-1 In the analysed word sequences **annotation errors** appear on the level of the attribution to the wrong part-of-speech, most frequently the annotation of an adjective as an adverb or an adverb as an adjective and the annotation of nouns as function words. Differentiating between adverbs and conjunctions with the same spelling is also a problem. At the same time, there are frequent errors on the level of the attribution of the wrong morphosyntactic categories, e.g., the attribution of the wrong gender or case within phrases, etc. These

problems would appear to be solvable with the upgrading of the automatic annotation to include collocational and colligational information of the phrase level. An important contribution to the improvement of the quality of tagging would already be the addition of colligation-based rules in the place of the disambiguation of lemmata or morphosyntactic tags.

It is difficult to solve problems connected with the annotation of proper names (*spider Pottinger, passionate Dmitrij*), foreign words (*a passionate *love/hate* relationship*) or Slovene words that do not yet exist in the database (*to dance with Teletubbies*). Words that remain non-lemmatised in the process of annotation are recorded in a special file and later added to the lexical database. Prior to the development of a quality named entity recognition system for Slovene a possible temporary solution would be to include the most frequent proper names (words or phrases) in the database.

The **annotation of particles** demands particular attention in the development of tagging, as it is a category that is difficult to attribute automatically on the morphosyntactic level, due to homography of particles and adverbs or conjunctions. In this regard, it is suggested that consideration be given to the annotation of particles not before the syntactic level of tagging. The **annotation of abbreviations** also demands consideration. Here it is possible to differentiate between groups that can be captured in the lexicon (e.g., units of measurement) and unpredictable examples, for the annotation of which it would be necessary to develop specialised methods for automatic identification in the text.

VI-2.1 In the final form, pattern types with examples of word sequences are compiled in three tables. There are 31 **pattern types relevant to extraction**, categorised with regard to the extracted data into: noun phrases (17) adjective phrases (3), coordinated phrases (4) and conditionally interesting phrases (7).

Pattern types irrelevant to the extraction are classified with regard to their composition into pattern types: with an adverb at the end of the pattern or before the noun, with an adjective at the end of the pattern, with a preposition at the beginning or the end of the pattern, with a coordinated conjunction between different word classes or at the beginning or end of the pattern, and others. With regard to the tags, patterns that contain auxiliary verbs, particles, abbreviations, pronouns, numerals or non-lemmatised words are also placed in the table.

Pattern types that contain a verb are presented separately due to the foreseen higher level of intervention on the syntactic level of language, which consequently means that the treated method for acquiring this data is less suitable. These patterns are divided with regard to what appears in the extracted word sequence alongside the verb: noun or noun phrase, adverb or adverbial phrase, verb, preposition or prepositional phrase, or adjective.

VI-2.2 The extraction of noun phrases is also verified for the lemmata *Slovene* (female), *ram* and *debate* (Slo.: *Slovenka*, *oven* and *debata*). The extraction of two-noun phrases provides entirely relevant results. More typical on the level of bigrams appear to be combinations of two nouns in the same case (*mother Slovene-26*) and a phrase with the second noun in the genitive case (*Slovene [of the] year-560*). There are only rare appearances of noun phrases with the second noun in the dative case (*support [to] Slovenes-1*) and examples of combinations with a right non-agreeing nominative noun qualifier (*journal Slovene-2*). The extraction of three-noun phrases also provides relevant examples for all of the new lemmata (*the title Slovene [of the] year-39*).

Equally successful is the extraction of noun phrases with **accordant adjectives on the left** (*breeding ram-82*). In line with expectations, three-part phrases are less frequent; however, the list of examples still indicates good results. Problems arise on the level of the presentation of the extracted data, since adjectives are lemmatised as genderly unmarked adjectival forms. In contrast, on the level of the extraction of noun phrases with the **accordant adjectives on the right**, fewer and less interesting results are provided (*ram breeding-2*).

The extraction of phrases **with a preposition** enables, on the one hand, the acquisition of phrases that are interesting for further processing (*by birth Slovene, in the sign [of a] ram*), while, on the other hand, incomplete phrases are found (*with rams [of the] breed*). Phrases with a **subordinate conjunction** are rather scarce (*debate as a science-1*), while in contrast phrases with a **coordinated conjunction** are extremely frequent (*a Slovene[m] and a Slovene[fj]-635*).

VII In the research a method of lexical-data extraction on the basis of two-part and three-part combinations of the treated lemma with the morphosyntactic tags of its minimal textual context was used, and thus verified. For the extraction of data from the entire reference corpus the construction of an entirely automated procedure that will combine all of the elements described in the present research in the simplest way (the search of relevant phrases, the categorisation and counting of phrases, a suitable final formalization of the phrases) is foreseen.

With regard to the quality of the extracted data the predicted outcome is confirmed, namely that the results are most relevant where the Slovene language demonstrates a sufficient level of regularity on the syntactic level within the framework of two-part or three-part word sequences – the method is more reliable on the level of the extraction of phrases with a noun, an adjective or an adverb as a phrasal head, while on the level of patterns that contain a verb lemma or morphosyntactic tag the level of reliability of the data extracted with the described method is lower. The other finding in terms of principle is that the extraction of phrases provides a good image of that which is typical in a language but does not perform as well on the level of less frequent pattern types.

Taking into account the fact that for the future development of the method research on a larger set of data will be necessary, it is possible to conclude that the extraction of phrases from a morphosyntactically tagged corpus represents a good point of departure for the upgrading of the LT lexical database in Slovene. The results – particularly on the level of the extraction of noun phrases, to which more attention was devoted in the research – are encouraging and bear witness to the great potential for the use of this kind of extracted data for various purposes.

PRILOGA 1

Nabor oblikoskladenjskih oznak JOS

SAMOSTALNIK				PRIDEVNIK			
1	vrsta	občno ime	o	1	vrsta	splošni	p
		lastno ime	l			svojilni	s
2	spol	moški	m	2	stopnja	deležniški	d
		ženski	z			nedoločeno	n
		srednji	s			primernik	p
3	število	ednina	e	3	spol	presežnik	s
		dvojina	d			moški	m
		množina	m			ženski	z
4	sklon	imenovalnik	i	4	število	srednji	s
		rodilnik	r			ednina	e
		dajalnik	d			dvojina	d
		tožilnik	t			množina	m
		mestnik	m	5	sklon	imenovalnik	i
5	živost	ne	n			rodilnik	r
		da	d			dajalnik	d
						tožilnik	t
						mestnik	m
						orodnik	o
				6	določnost	nedoločni	n
						določni	d
PRISLOV				PREDLOG			
1	stopnja	nedoločeno	n	1	sklon	imenovalnik	i
		primernik	r			rodilnik	r
		presežnik	s			dajalnik	d
2	deležje	ne	n			tožilnik	t
		da	d			mestnik	m
						orodnik	o

GLAGOL

1	vrsta	glavni	g
		pomožni	p
2	vid	dovršni	d
		nedovršni	n
		dvovidski	v
3	oblika	nedoločnik	n
		namenilnik	m
		deležnik	d
		sedanjik	s
		prihodnjik	p
		pogojnik	g
		velelnik	v
4	oseba	prva	p
		druga	d
		tretja	t
5	število	ednina	e
		dvojina	d
		množina	t
6	spol	moški	m
		ženski	z
		srednji	s
7	nikalnost	nezanikani	n
		zanikani	d

VEZNIK

1	vrsta	priredni	p
		podredni	d

ZAIMEK

1	vrsta	osebni	o
		kazalni	k
		nedoločni	n
		svojilni	s
		vprašalni	v
		oziralni	z
		povratni	p
		nikalni	l
		celostni	c
2	oseba	prva	p
		druga	d
		tretja	t
3	spol	moški	m
		ženski	z
		srednji	s
4	število	ednina	e
		dvojina	d
		množina	m
5	sklon	imenovalnik	i
		rodilnik	r
		dajalnik	d
		tožilnik	t
		mestnik	m
		orodnik	o
6	število	ednina	e
	svojine	dvojina	d
		množina	m
7	spol	moški	m
	svojine	ženski	z
		srednji	s
8	oblika	klitični	k
		navezni	z

NEUVRŠČENI

1	vrsta	tujejezični	j
		tipkarska	t
		program	p

ŠTEVNIK			
1	zapis	arabski	a
		rimski	b
		besedni	r
2	vrsta	glavni	g
		vrstilni	v
		zaimkovni	z
		drugi	d
3	spol	moški	m
		ženski	z
		srednji	s
4	število	ednina	m
		dvojina	z
		množina	s
5	sklon	imenovalnik	i
		rodilnik	r
		dajalnik	d
		tožilnik	t
		mestnik	m
6	določnost	orodnik	o
		nedoločni	n
		določni	d

Kategorije **členek** - L, **medmet** - M ter **okrajšava** - O niso nadalje členjene.

PRILOGA 2

Nabor oblikoskladenjskih oznak Multext-East

SAMOSTALNIK				PRIDEVNIK			
1	vrsta	občno_ime	o	1	vrsta	kakovostni	k
		lastno_ime	l			svojilni	s
2	spol	moški	m			vrstni	v
		ženski	z	2	stopnja	osnovnik	o
		srednji	s			primernik	p
3	število	ednina	e			presežnik	s
		dvojina	d			elativ	e
		množina	m	3	spol	moški	m
4	sklon	imenovalnik	i			ženski	z
		rodilnik	r			srednji	s
		dajalnik	d	4	število	ednina	e
		tožilnik	t			dvojina	d
		mestnik	m			množina	m
		orodnik	o	5	sklon	imenovalnik	i
5	/					rodilnik	r
6	/					dajalnik	d
7	živost	ne	n			tožilnik	t
		da	d			mestnik	m
						orodnik	o
				6	določnost	ne	n
						da	d
				7	/		
				8	živost	ne	n
						da	d
PRISLOV				PREDLOG			
1	vrsta	splošni	s	1	vrsta	/	p
2	stopnja	osnovnik	o	2	sestavljeno	enostaven	e
		primernik	p			prironski	p
		presežnik	r	3	sklon, ki	rodilnik	r
		elativ	e		ga zahteva	dajalnik	d
						tožilnik	t
						mestnik	m
						orodnik	o

GLAGOL				ZAIMEK			
1	vrsta	polnopomenski	p	1	vrsta	osebni	o
		naklonski	n			kazalni	k
		vezni	v			nedoločnostni	n
2	oblika	povednik	p			svojilni	s
		velelnik	v			vprašalni	v
		pogojnik	g			oziralnostni	z
		nedoločnik	n			povratni	p
		deležnik	d			nikalni	l
		namenilnik	m			celostni	c
3	čas	sedanjik	s	2	oseba	prva	p
		prihodnjik	p			druga	d
		nesedanjik	r			tretja	t
4	oseba	prva	p	3	spol	moški	m
		druga	d			ženski	z
		tretja	t			srednji	s
5	število	ednina	e	4	število	ednina	e
		dvojina	d			dvojina	d
		množina	m			množina	m
6	spol	moški	m	5	sklon	imenovalnik	i
		ženski	z			rodilnik	r
		srednji	s			dajalnik	d
7	način	tvornik	t			tožilnik	t
		trpni deležnik	r			mestnik	m
8	nikalnost	nezanikani	n			orodnik	o
		zanikani	z	6	število	ednina	e
9	/				svojine	dvojina	d
10	/					množina	m
11	/			7	spol	moški	m
12	/				svojine	ženski	z
13	/					srednji	s
14	vid	nedovršni	n	8	naslonka	ne	n
		dovršni	d			da	d
						navezna	z
				9	nanašanje	osebni	o
						svojilni	s
				10	skladenjska vloga	samostalniški	s
						pridevniški	p
						prislovni	r
				11	/		
				12	živost	ne	n
						da	d
VEZNIK							
1	vrsta	priredni	p				
		podredni	d				
2	oblika	enobesedni	e				
		večbesedni	v				

ŠTEVNIK			
1	vrsta	glavni	g
		vrstilni	v
		množilni	m
		drugi	d
2	spol	moški	m
		ženski	z
		srednji	s
3	število	ednina	e
		dvojina	d
		množina	m
4	sklon	imenovalnik	i
		rodilnik	r
		dajalnik	d
		tožilnik	t
		mestnik	m
		orodnik	o
5	zapis	arabski	a
		rimski	r
		besedni	b
6	določnost	ne	n
		da	d
7	/		
8	/		
9	živost	ne	n
		da	d

Kategorije **členek** - L, **medmet** - M ter **okrajšava** - O niso nadalje členjene.

PRILOGA 3

Legenda zapisa vzorčnih tipov

OSNOVNE OZNAKE

Sam	samostalnik
Prid	pridevnik
Glag	glagol
Prisl	prislov
Pred	predlog
Vp	priredni veznik
Vd	podredni veznik

PODPISANE OZNAKE

1-6	sklon
m, ž, s	spol
LI, OI	lastno ime, občno ime
/	negacija

primer

Sam₂ = samostalnik v rodilniku
 Sam_m = samostalnik moškega spola
 Sam_{LI} = samostalnik lastno ime
 Sam_{/1} = samostalnik v neimenovalniku

Appendix 4

Pattern-type format

basic symbols

N	noun
Adj	adjective
V	verb
Adv	adverb
P	preposition
Cc	coordinating conjunction
Cs	subordinating conjunction

subscribed symbols

1-6	case
m, f, n	gender
p, c	proper name, common name
/	negation

example

N_2 = noun, genitive case
 N_m = noun, masculine gender
 N_p = noun, proper name
 $N_{/1}$ = noun, non-nominative

LITERATURA

Aarts, Bas (2007). Corpus linguistics, Chomsky and fuzzy tree fragments. V Teubert, Wolfgang in Krishnamurty, Ramesh (ur.), *Corpus Linguistics: Critical Concepts in Linguistics*. London, New York: Routledge. 173–181.

Aarts, Jan (2007). Does corpus linguistics exist? Some old and new issues. V Teubert, Wolfgang in Krishnamurty, Ramesh (ur.), *Corpus Linguistics: Critical Concepts in Linguistics*. London, New York: Routledge. 58–73.

Agirre, Eneko in Edmonds, Philip (ur.) (2007). *Word sense disambiguation: Algorithms and applications*. Dordrecht: Springer.

Aijmer, Karin in Altenberg, Bengt (ur.) (1991). *English Corpus Linguistics*. London in New York: Longman.

Arčan, Mihael in Vintar, Špela (2006). Avtomatično prepoznavanje lastnih imen. V Erjavec, Tomaž in Žganec Gros, Jerneja (ur.), *Jezikovne tehnologije*. Ljubljana: IJS. 150–155.

Arhar, Špela in Ledinek, Nina (2008). Oblikoskladenjske oznake JOS: revizija in nadgradnja nabora oznak za avtomatsko skladiščno označevanje slovenščine. V Erjavec, Tomaž in Žganec Gros, Jerneja (ur.), *Jezikovne tehnologije*. Ljubljana: IJS. 54–59.

Arhar, Špela (2007). *Kaj početi z referenčnim korpusom FidaPLUS*. Ljubljana: Univerza v Ljubljani, Filozofska fakulteta.

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

Arhar, Špela in Romih, Miro (2006). Klepec: slovenski programirani sogovornik. V Erjavec, Tomaž in Žganec Gros, Jerneja (ur.), *Jezikovne tehnologije*. Ljubljana: IJS.

Atkins, Sue in Rundell, Michael (2008). *The Oxford Guide to Practical Lexicography*. Oxford: Oxford University Press.

Atkins, Sue in Zampolli, Antonio (ur.) (1994). *Computational approaches to the lexicon*. Oxford: Oxford University Press.

Baker, Mona, Francis, Gill in Tognini-Bonelli, Elena (ur.) (1993). *Text and Technology*. Amsterdam: John Benjamins Publishing Company.

Barnbrook, Geoff (1996). *Language and Computers: A practical Introduction to the Computer Analysis of Language*. Edinburgh: Edinburgh University Press.

Baroni, Marco, Kilgarrieff, Adam, Pomikálek, Jan in Rychlý, Pavel (2006). WebBootCaT: a web tool for instant corpora. V Corino, Elisa et al. (ur.), *Euralex*. Torino, Italia: European Association for Lexicography. 123–131.

Beaugrande, Robert de (2007). "Corporate bridges" twixt text and language. V Teubert, Wolfgang in Krishnamurty, Ramesh (ur.), *Corpus Linguistics: Critical Concepts in Linguistics*. London, New York: Routledge. 93–118.

Bentivogli, Luisa in Pianta, Emanuele (2002). Detecting Hidden Multiwords in Bilingual Dictionaries. V Braasch, Anna in Povlsen, Claus (ur.), *Euralex*. Copenhagen, Denmark: Center for Sprogteknologi. 785–793.

Bentivogli, Luisa in Pianta, Emanuele (2000). Looking for lexical gaps. V Heid, Ulrich et al. (ur.), *Euralex*. Stuttgart, Germany. 663–669.

Biber, Douglas, Conrad, Susan in Reppen, Randi (1998). *Corpus Linguistics: Investigating Language Structure and Use*. Cambridge: Cambridge University Press.

Boguraev, Bran in Briscoe, Ted (ur.) (1989). *Computational Lexicography for Natural Language Processing*. London in New York: Longman.

Braasch, Anna (2006). Exploitation of syntactic patterns for sense group identification. V Corino, Elisa et al. (ur.), *Euralex*. Torino, Italia: European Association for Lexicography. 133–139.

Braasch, Anna in Olsen, Sussi (2004). STO: A Danish Lexicon Resource – Ready for Applications. *International Conference on Language Resources and Evaluation*. Lizbona: ELRI. 1079–1083.

Braasch, Anna in Pedersen, Bolette S. (2002). Recent Work in the Danish Computational Lexicon Project "STO". V Braasch, Anna in Povlsen, Claus (ur.), *Euralex*. Copenhagen, Denmark: Center for Sprogteknologi. 301–314.

Braasch, Anna in Olsen, Sussi (2000). Formalised Representation of Collocations in a Danish Computational Lexicon. V Heid, Ulrich et al. (ur.), *Euralex*. Stuttgart, Germany. 475–488.

Chafe, Wallace (2007). The importance of corpus linguistics to understanding the nature of language. V Teubert, Wolfgang in Krishnamurty, Ramesh (ur.), *Corpus Linguistics: Critical Concepts in Linguistics*. London, New York: Routledge. 41–57.

Christiansen, Tom in Torkington, Nathan (1998). *Perl Cookbook*. Sebastopol, CA: O'Reilly.

Citron, Sabine in Widmann, Thomas (2006). A Bilingual Corpus for Lexicographers. V Corino, Elisa et al. (ur.), *Euralex*. Torino, Italia: European Association for Lexicography. 251–255.

Correard, Marie-Helene (ur.) (2002). *Lexicography and Natural Language Processing: A Festschrift in Honour of B.T. Sue Atkins*. Grenoble: Euralex.

Cozens, Simon in Wainwright, Peter (2000). *Beginning Perl*. Birmingham: Wrox Press.

Čermák, František (2006). Collocations, Collocability and Dictionary. V Corino, Elisa et al. (ur.), *Euralex*. Torino, Italia: European Association for Lexicography. 929–937.

Čermák, František (2000). Combination, Collocation and Multi-Word Units. V Heid, Ulrich et al. (ur.), *Euralex*. Stuttgart, Germany. 489–495.

De Cock, Sylvie in Granger, Sylviane (2004). High Frequency Words: the Bête Noire of Lexicographers and Learners Alike. V Williams, Geoffrey in Vessier, Sandra (ur.), *Euralex*. Lorient: Faculté des Lettres et des Sciences Humaines, Université de Bretagne Sud. 233–243.

De Mauro, Tulio (2006). On Lexicon and Grammar. V Corino, Elisa et al. (ur.), *Euralex*. Torino, Italia: European Association for Lexicography. 19–29.

Duncker, Dorte (2002). Collecting Collocations. V Braasch, Anna in Povlsen, Claus (ur.), *Euralex*. Copenhagen, Denmark: Center for Sprogteknologi. 521–531.

Džeroski, Sašo in Erjavec, Tomaž (2000). Strojno učenje lematizacije neznanjih slovenskih besed. V Erjavec, Tomaž in Žganec Gros, Jerneja (ur.), *Jezikovne tehnologije*. Ljubljana: IJS. 14–19.

- Erjavec, Tomaž in Krek, Simon (2008a). Oblikoskladenjske specifikacije in označeni korpusi JOS. V Erjavec, Tomaž in Žganec Gros, Jerneja (ur.), *Jezikovne tehnologije*. Ljubljana: IJS. 49–53.
- Erjavec, Tomaž in Krek, Simon (2008b). The JOS Morphosyntactically Tagged Corpus of Slovene. *International Conference on Language Resources and Evaluation*. Marakeš: ELRA.
- Erjavec, Tomaž in Sarossy, Bence (2006). Oblikoslovno označevanje slovenskega jezika: primer korpusa SVEZ–IJS. V Erjavec, Tomaž in Žganec Gros, Jerneja (ur.), *Jezikovne tehnologije*. Ljubljana: IJS. 168–173.
- Erjavec, Tomaž in Ledinek, Nina (2006). Slovenska odvisnostna drevesnica: prvi rezultati. V Erjavec, Tomaž in Žganec Gros, Jerneja (ur.), *Jezikovne tehnologije*. Ljubljana: IJS. 162–167.
- Erjavec, Tomaž (2004). MULTTEXT–East Version 3: Multilingual Morphosyntactic Specifications, Lexicons and Corpora. V Lino, Maria Teresa in Xavier, Maria Francisca (ur.), *International Conference on Language Resources and Evaluation, LREC'04*. Lizbona: ELRA. 1535–1538.
- Erjavec, Tomaž (2003). Označevanje korpusov. *Jezik in slovstvo*, 48 (3–4), 61–76.
- Erjavec, Tomaž, Džeroski, Sašo in Zavrel, Jakob (2000). Morphosyntactic Tagging of Slovene: Evaluating PoS Taggers and Tagsets. *International Conference on Language Resources and Evaluation*. Atene: ELRA.
- Erjavec, Tomaž, Gorjanc, Vojko in Stabej, Marko (1998). Korpus FIDA. V Erjavec, Tomaž in Žganec Gros, Jerneja (ur.), *Jezikovne tehnologije*. Ljubljana: IJS.
- Erjavec, Tomaž (1998). Standardizacija zapisa jezikovnih podatkov. V Erjavec, Tomaž in Žganec Gros, Jerneja (ur.), *Jezikovne tehnologije*. Ljubljana: IJS. 119–123.
- Evert, Stefan, Heid, Ulrich, Sauberlich, Bettina, Debus-Gregor, Esther in Scholze-Stubenrecht, Werner (2004). Supporting corpus-based dictionary updating. V Williams, Geoffrey in Vessier, Sandra (ur.), *Euralex*. Lorient: Faculté des Lettres et des Sciences Humaines, Université de Bretagne Sud. 255–264.
- Fillmore, Charles (2007). "Corpus linguistics" or "computer-aided armchair linguistics". V Teubert, Wolfgang in Krishnamurty, Ramesh (ur.), *Corpus Linguistics: Critical Concepts in Linguistics*. London, New York: Routledge. 197–220.
- Fišer, Darja in Erjavec, Tomaž (2008). Predstavitev in analiza slovenskega wordneta. V Erjavec, Tomaž in Žganec Gros, Jerneja (ur.), *Jezikovne tehnologije*. Ljubljana: IJS. 37–42.
- Fišer, Darja (2005). Pristopi k izdelavi leksikalnih podatkovnih zbirk. *Jezik in slovstvo*, 50 (6), 17–32.
- Fliedl, Günther, Homa, Andreas, Maurer-Stroh, Philippa in Weber, Georg (2004). ANCR – The Adjective-Noun Collocation Retriever (A Tool for Generating a Bilingual Dictionary from a German-English Parallel Corpus). V Williams, Geoffrey in Vessier, Sandra (ur.), *Euralex*. Lorient: Faculté des Lettres et des Sciences Humaines, Université de Bretagne Sud. 885–890.
- Fontenelle, Thierry (2006). Developing a Lexicon for an New French Spell-checker. V Corino, Elisa et al. (ur.), *Euralex*. Torino, Italia: European Association for Lexicography. 151–158.
- Fontenelle, Thierry (2004). Lexicalization for Proofing Tools. V Williams, Geoffrey in Vessier, Sandra (ur.), *Euralex*. Lorient: Faculté des Lettres et des Sciences Humaines, Université de Bretagne Sud. 79–86.
- Fontenelle, Thierry (2000). Extracting Phraseology for Content Analysis and Document Retrieval. V Heid, Ulrich et al. (ur.), *Euralex*. Stuttgart, Germany. 351–358.

- Fontenelle, Thierry (1997). *Turning a Bilingual Dictionary into a Lexical-Semantic Database*. Tübingen: Max Niemeyer Verlag.
- Frawley, William (ur.) (2003). *International Encyclopedia of Linguistics*. Oxford University Press.
- Gantar, Polona (2008). (Slovenska) leksika med leksikonom in slovnico. *Jezik in slovstvo*, 53 (5), 19–36.
- Gantar, Polona (2007). *Stalne besedne zveze v slovenščini*. Ljubljana: Založba ZRC, ZRC SAZU.
- Gantar, Polona (2006). Korpusni pristop k frazeologiji in slovarske aplikacije. *Slavistična revija*, 54 (Posebna številka), 151–163.
- Gantar, Polona (2005). Korpusni pristop k prepoznavanju in analizi stalnih besednih zvez v slovenščini. V Kržišnik, Erika in Eismann, Wolfgang (ur.), *Frazeologija v jezikoslovju in drugih vedah*. Ljubljana: UL, FF, Oddelek za slovenistiko Filozofske fakultete. 79–88.
- Garside, Roger, Leech, Geoffrey in McEnery, Tony (ur.) (1997). *Corpus Annotation: Linguistic Information from Computer Text Corpora*. London in New York: Longman.
- Gorjanc, Vojko in Vintar, Špela (2007). Korpusna analiza vloge označevalcev medleksemskih razmerij v organizaciji besedila. *Jezik in slovstvo*, 52 (3–4), 117–130.
- Gorjanc, Vojko, Krek, Simon in Gantar, Polona (2005). Slovenska leksikalna podatkovna zbirka. *Jezik in slovstvo*, 50 (2), 3–19.
- Gorjanc, Vojko in Krek, Simon (ur.) (2005). *Študije o korpusnem jezikoslovju*. Ljubljana: Krtina.
- Gorjanc, Vojko (2005). *Uvod v korpusno jezikoslovje*. Domžale: Izolit.
- Gorjanc, Vojko in Jurko, Primož (2004). Kolokacije in učenje tujega jezika. *Jezik in slovstvo*, 49 (3–4), 49–62.
- Gorjanc, Vojko (2003). Korpusi in jezikoslovje. *Jezik in slovstvo*, 48 (3–4), 19–27.
- Gorjanc, Vojko in Žele, Andreja (2002). Compound Dictionary Entries (The Case of Slovene Noun Phrases). V Braasch, Anna in Povlsen, Claus (ur.), *Euralex*. Copenhagen, Denmark: Center for Sprogteknologi. 607–614.
- Halliday, Michael A. K. (2004). *An introduction to functional grammar*. Oxford: Oxford University Press.
- Halliday, Michael A. K., Teubert, Wolfgang, Yallop, Colin in Čermakova, Anna (2004). *Lexicology and Corpus Linguistics: An introduction*. London: Continuum.
- Hanks, Patrick (ur.) (2008). *Lexicology*. New York: Routledge.
- Hanks, Patrick (2004). Corpus Pattern Analysis. V Williams, Geoffrey in Vessier, Sandra (ur.), *Euralex*. Lorient: Faculté des Lettres et des Sciences Humaines, Université de Bretagne Sud. 87–97.
- Hanks, Patrick (2000). Contributions of Lexicography and Corpus Linguistics to a Theory of Language Performance. V Heid, Ulrich et al. (ur.), *Euralex*. Stuttgart, Germany. 3–13.
- Heid, Ulrich in Gouws, Rufus H. (2006). A Model for a Multifunctional Dictionary of Collocations. V Corino, Elisa et al. (ur.), *Euralex*. Torino, Italia: European Association for Lexicography. 979–988.

- Heid, Ulrich (1994). Relating Lexicon and Corpus: Computational Support for Corpus-Based Lexicon Building in DELIS. V W. Martin, W. Meijs, M. Moerland, E. ten Pas, P. van Sterkenburg, P. Vossen (ur.), *Euralex*. Amsterdam. 459 – 471.
- Holozan, Peter (2004). Uporaba glagolskih predlog pri strojnem prevajanju. V Erjavec, Tomaž in Žganec Gros, Jerneja (ur.), *Jezikovne tehnologije*. Ljubljana: IJS. 128.
- Hunston, Susan in Francis, Gill (2000). *Pattern Grammar: a corpus-driven approach to the lexical grammar of English*. Amsterdam: John Benjamins.
- Jackendoff, Ray (2002). *Foundations of Language: Brain, Meaning, Grammar, Evolution*. Oxford: Oxford University Press.
- Jakopin, Primož in Bizjak Končar, Aleksandra (2008). Part-of-Speech Tagging of Slovenian, 12 years after. V Erjavec, Tomaž in Žganec Gros, Jerneja (ur.), *Jezikovne tehnologije*. Ljubljana: IJS. 104–109.
- Jakopin, Primož in Loenneker, Birte (2004). Query-driven Dictionary Enhancement. V Williams, Geoffrey in Vessier, Sandra (ur.), *Euralex*. Lorient: Faculté des Lettres et des Sciences Humaines, Université de Bretagne Sud. 273–284.
- Jakopin, Primož in Bizjak, Aleksandra (1997). O strojno podprtem oblikoslovnem označevanju slovenskega besedila. *Slavistična revija*, 45 (3–4), 513–532.
- Janssen, Maarten (2006). Orthographic Variation in Lexical Databases. V Corino, Elisa et al. (ur.), *Euralex*. Torino, Italia: European Association for Lexicography. 167–172.
- Karlgren, Hans (1990). Computational Linguistics in 1990. *13th International Conference on Computational Linguistics* Helsinki, Finska. 97–99.
- Kennedy, Graeme (1998). *An Introduction to Corpus Linguistics*. London in New York: Longman.
- Kilgariff, Adam (2006). Collocationality (and how to measure it). V Corino, Elisa et al. (ur.), *Euralex*. Torino, Italia: European Association for Lexicography. 997–1004.
- Kilgariff, Adam, Rychly, Pavel, Smrz, Pavel in Tugwell, David (2004). The Sketch Engine. V Williams, Geoffrey in Vessier, Sandra (ur.), *Euralex*. Lorient: Faculté des Lettres et des Sciences Humaines, Université de Bretagne Sud. 105–115.
- Kilgariff, Adam in Rundell, Michael (2002). Lexical Profiling Software and its Lexicographic Applications – a Case Study. V Braasch, Anna in Povlsen, Claus (ur.), *Euralex*. Copenhagen, Denmark: Center for Sprogteknologi. 807–818.
- Kilgariff, Adam (1997). I don't believe in word senses. *Computers and the Humanities*, 31 (2), 91–113.
- Klímová, Jana (2002). Computational Processing of Czech Derived Words. V Braasch, Anna in Povlsen, Claus (ur.), *Euralex*. Copenhagen, Denmark: Center for Sprogteknologi. 137–144.
- Knowles, Gerry (2007). Corpora, databases and the organisation of linguistic data. V Teubert, Wolfgang in Krishnamurty, Ramesh (ur.), *Corpus Linguistics: Critical Concepts in Linguistics*. London, New York: Routledge. 119–133.
- Krek, Simon in Kilgariff, Adam (2006). Slovene Word Sketches. V Erjavec, Tomaž in Gros, Jerneja (ur.), *Jezikovne tehnologije*. Ljubljana: IJS. 62–65.

- Krek, Simon (2004). Slovarji serije COBUILD in formalizacija definicijskega jezika. *Jezik in slovstvo*, 49 (2), 3–16.
- Krenn, Brigitte (2000). Empirical Implications on Lexical Association Measures. V Heid, Ulrich et al. (ur.), *Euralex*. Stuttgart, Germany. 359–371.
- Kruyt, Truus (2003). Multifunctional linguistic databases: Their multiple use. V van Sterkenburg, Piet (ur.), *A Practical Guide to Lexicography*. Amsterdam: John Benjamins Publishing Company. 194–203.
- Kustova, G. I. in Paducheva, E. V. (1994). Semantic Dictionary as a Lexical Database. V W. Martin, W. Meijs, M. Moerland, E. ten Pas, P. van Sterkenburg, P. Vossen (ur.), *Euralex*. Amsterdam. 479 – 485.
- Laffling, John (1994). An Analogical Dictionary for Machine Translation. V W. Martin, W. Meijs, M. Moerland, E. ten Pas, P. van Sterkenburg, P. Vossen (ur.), *Euralex*. Amsterdam. 486 – 493.
- Ledinek, Nina (2007). Slovenska odvisnostna drevesnica v raziskavah o induktivnem odvisnostnem označevanju. *Jezik in slovstvo*, 52 (1), 3–16.
- Leech, Geoffrey (2007). The value of a corpus in English language research: A reappraisal. V Teubert, Wolfgang in Krishnamurty, Ramesh (ur.), *Corpus Linguistics: Critical Concepts in Linguistics*. London, New York: Routledge. 315–325.
- Leech, Geoffrey (1992). Corpora and theories of linguistic performance. V Svartvik, Jan (ur.), *Directions in corpus linguistics: Proceedings of Nobel symposium 82*. Berlin, New York: Mouton de Gruyter. 125–148.
- Léon, Jacqueline (2005). Claimed and unclaimed sources of corpus linguistics. V Teubert, Wolfgang in Krishnamurty, Ramesh (ur.), *Corpus Linguistics: Critical Concepts in Linguistics*. London in New York: Routledge. 326–341.
- Lezius, Wolfgang, Dipper, Stefanie in Fitschen, Arne (2000). IMSLex – Representing Morphological and Syntactic Information in a Relational Database. V Heid, Ulrich et al. (ur.), *Euralex*. Stuttgart, Germany. 133–139.
- Loenneker, Birte in Jakopin, Primož (2004). Checking POSBesa, a Part-of-Speech Tagged Slovenian Corpus. V Erjavec, Tomaž in Žganec Gros, Jerneja (ur.), *Jezikovne tehnologije*. Ljubljana: IJS. 48–55.
- Logar, Nataša in Vintar, Špela (2008). Korpusni pristop k izdelavi terminoloških slovarjev: od besednih seznamov in konkordanc do samodejnega luščenja izrazja. *Jezik in slovstvo*, 53 (5), 3–18.
- Logar, Nataša (2007). *Korpusni pristop k pridobivanju in predstavitvi jezikovnih podatkov v terminoloških slovarjih in terminoloških podatkovnih zbirkah [doktorska naloga]*. Ljubljana: Univerza v Ljubljani, Filozofska fakulteta.
- MacLeod, Catherine in Grishman, Ralph (2000). The Influence of Corpora on Lexicons: Corpora Use in the Creation of COMLEX Syntax and NOMLEX. V Heid, Ulrich et al. (ur.), *Euralex*. Stuttgart, Germany. 141–148.
- Manning, Christopher D. in Schütze, Hinrich (2003). *Foundations of Statistical Natural Language Processing*. Cambridge, Massachusetts: The MIT Press.
- McEnergy, Tony, Xiao, Richard in Tono, Yukio (2006). *Corpus-Based Language Studies*. London: Routledge.
- McEnergy, Tony in Wilson, Andrew (1996). *Corpus Linguistics*. Edinburgh: Edinburgh University Press.
- Mitkov, Ruslan (ur.) (2003). *The Oxford Handbook of Computational Linguistics*. Oxford Oxford University Press.

Mladenić, Dunja (2002). Automatic word lemmatization. V Erjavec, Tomaž in Žganec Gros, Jerneja (ur.), *Jezikovne tehnologije*. Ljubljana: IJS. 153–159.

Mohorič, Tomaž (1995). *Uvod v podatkovne baze*. Ljubljana: BI–TIM d. o. o.

Moirón, M. Begoña Villada in Bouma, Gosse (2002). A corpus-based approach to the acquisition of collocational prepositional phrases. V Braasch, Anna in Povlsen, Claus (ur.), *Euralex*. Copenhagen, Denmark: Center for Sprogteknologi. 153–158.

Oakes, Michael P. (1998). *Statistics for Corpus Linguistics*. Edinburgh: Edinburgh University Press.

Olsen, Sussi (2002). Some Aspects of the Syntactic Encoding of Nouns in a Computational Lexicon – the STO Project. V Braasch, Anna in Povlsen, Claus (ur.), *Euralex*. Copenhagen, Denmark: Center for Sprogteknologi. 159–168.

Ooi, Vincent (1998). *Computer Corpus Lexicography*. Edinburgh: Edinburgh University Press.

Orešnik, Janez (1989). Računalniško prepoznavanje slovenske skladnje (jezikoslovčev predlog). *Slovenski jezik v znanosti* 2. Ljubljana: Znanstveni inštitut Filozofske fakultete. 129–143.

Pereira, Luísa Alice Santos in Mendes, Amália (2002). An Electronic Dictionary of Collocations for European Portuguese: Methodology, Results and Applications. V Braasch, Anna in Povlsen, Claus (ur.), *Euralex*. Copenhagen, Denmark: Center for Sprogteknologi. 841–849.

Peters, Carol, Federici, Stefano, Montemagni, Simonetta in Calzolari, Nicoletta (1994). From Machine Readable Dictionaries to Lexicons for NLP: the Cobuild Dictionaries – a Different Approach. V W. Martin, W. Meijs, M. Moerland, E. ten Pas, P. van Sterkenburg, P. Vossen (ur.), *Euralex*. Amsterdam. 147 – 157.

Rodger, Liam (2004). On "not going there" in the Bilingual Dictionary: the Case of Grammatical-Word Idioms. V Williams, Geoffrey in Vessier, Sandra (ur.), *Euralex*. Lorient: Faculté des Lettres et des Sciences Humaines, Université de Bretagne Sud. 511–518.

Romih, Miro in Holozan, Peter (2002a). Infrastruktura za razvoj jezikovnih tehnologij – korpus FIDA in sistem ASES. V Erjavec, Tomaž in Žganec Gros, Jerneja (ur.), *Jezikovne tehnologije*. Ljubljana: IJS. 166.

Romih, Miro in Holozan, Peter (2002b). Slovensko-angleški prevajalni sistem. V Erjavec, Tomaž in Žganec Gros, Jerneja (ur.), *Jezikovne tehnologije*. Ljubljana: IJS. 167.

Romih, Miro in Holozan, Peter (2002c). Sporazumevanje z računalnikom v naravnem jeziku. V Erjavec, Tomaž in Žganec Gros, Jerneja (ur.), *Jezikovne tehnologije*. Ljubljana: IJS. 168.

Romih, Miro (1998). Amebis in jezikovne tehnologije. V Erjavec, Tomaž in Žganec Gros, Jerneja (ur.), *Jezikovne tehnologije*. Ljubljana: IJS. 29–33.

Rupnik, Jan, Grčar, Miha in Erjavec, Tomaž (2008). Improving morphosyntactic tagging of Slovene by tagger combination. V Erjavec, Tomaž in Žganec Gros, Jerneja (ur.), *Jezikovne tehnologije*. Ljubljana: IJS. 110–115.

Ruppenhofer, Josef, Baker, Collin F. in Fillmore, Charles J. (2002a). Collocational Information in the FrameNet Database. V Braasch, Anna in Povlsen, Claus (ur.), *Euralex*. Copenhagen, Denmark: Center for Sprogteknologi. 359–369.

Ruppenhofer, Josef, Baker, Collin F. in Fillmore, Charles J. (2002b). The FrameNet Database and Software Tools. V Braasch, Anna in Povlsen, Claus (ur.), *Euralex*. Copenhagen, Denmark: Center for Sprogteknologi. 371–375.

Saint-Dizier, Patrick in Viegas, Evelyne (ur.) (1995). *Computational lexical semantics*. New York: Cambridge University Press.

Schulte im Walde, Sabine (2002). Evaluating Verb Subcategorisation Frames learned by a German Statistical Grammar against Manual Definitions in the *Duden* Dictionary. V Braasch, Anna in Povlsen, Claus (ur.), *Euralex*. Copenhagen, Denmark: Center for Sprogteknologi. 187–197.

Schutz, Rik (2004). Structured Data + Automated Selection and Sorting = Dictionary. V Williams, Geoffrey in Vessier, Sandra (ur.), *Euralex*. Lorient: Faculté des Lettres et des Sciences Humaines, Université de Bretagne Sud. 303–310.

Schwartz, Randal L. in Phoenix, Tom (2001). *Learning Perl*. Sebastopol, CA: O'Reilly.

Seretan, Violeta, Nerima, Luka in Wehrli, Eric (2004). A Tool for Multi-Word Collocation Extraction and Visualization in Multilingual Corpora. V Williams, Geoffrey in Vessier, Sandra (ur.), *Euralex*. Lorient: Faculté des Lettres et des Sciences Humaines, Université de Bretagne Sud. 755–766.

Sinclair, John (2007). Meaning in the framework of corpus linguistics. V Teubert, Wolfgang in Krishnamurty, Ramesh (ur.), *Corpus Linguistics: Critical Concepts in Linguistics*. London, New York: Routledge. 182–196.

Sinclair, John (2005). To complement the dictionary. *Jazyky a jazykověda: Sborník k 65. narozeninám prof. PhDr. Františka Čermáka*. Praga: DrSc. Filozofická fakulta Univerzity Karlovy – Ústav Českého národního korpusu.

Sinclair, John (2004a). In praise of the dictionary. V Williams, Geoffrey in Vessier, Sandra (ur.), *Euralex*. Lorient: Faculté des Lettres et des Sciences Humaines, Université de Bretagne Sud. 1–11.

Sinclair, John (2004b). *Trust the Text: Language, corpus and discourse*. London in New York: Routledge.

Sinclair, John (1998). The Lexical Item. V Weigand, Edda (ur.), *Contrastive Lexical Semantics*. Amsterdam: John Benjamins. 1–24.

Sinclair, John (1996a). *Lexis & Lexicography*. Singapore: UniPress.

Sinclair, John (1996b). The Empty Lexicon. *International Journal of Corpus Linguistics*, 1 (1), 99–119.

Sinclair, John, Hoelter, Martin in Peters, Carol (ur.) (1994). *The language of definition: The formalization of dictionary definitions for NLP*. Luxembourg: Office for official publications of the European Commission.

Sinclair, John (ur.) (1987). *Looking up: An account of the COBUILD Project in lexical computing*. London: Collins.

Smolej, Mojca (2004). Členki kot besedilni povezovalci. *Jezik in slovstvo*, 49 (5), 45–57.

Stabej, Marko (ur.) (2003). *Jezik in slovstvo: Jezikovne tehnologije za slovenščino – tematska številka*, 48 (3–4).

Storjohann, Petra (2006). New Lexicographic Approaches to the Description of Sense Relations. V Corino, Elisa et al. (ur.), *Euralex*. Torino, Italia: European Association for Lexicography. 1201–1212.

Stritar, Mojca (2006a). Merila za oblikovanje korpusov usvajanja tujega jezika. *Jezik in slovstvo*, 51 (5), 59–74.

Stritar, Mojca (2006b). Oblikovanje korpusa usvajanja slovenščine kot tujega jezika. V Erjavec, Tomaž in Žganec Gros, Jerneja (ur.), *Jezikovne tehnologije*. Ljubljana: IJS. 134–139.

Stubbs, Michael (2002). *Words and phrases: corpus studies of lexical semantics*. Oxford: Blackwell Publishing.

- Stubbs, Michael (1993). British traditions in text analysis. V Baker, Mona et al. (ur.), *Text and Technology*. Amsterdam: John Benjamins. 1–33.
- Summers, Della (1996). Computer Lexicography: the importance of representativeness in relation to frequency. V Thomas, Jenny in Short, Mick (ur.), *Using Corpora for Language Research*. New York: Longman Publishing. 260–266.
- Teubert, Wolfgang in Krishnamurty, Ramesh (ur.) (2007). *Corpus Linguistics: Critical Concepts in Linguistics*. London, New York: Routledge.
- Teubert, Wolfgang (2007). Writing, hermeneutics, and corpus linguistics. V Teubert, Wolfgang in Krishnamurty, Ramesh (ur.), *Corpus Linguistics: Critical Concepts in Linguistics*. London, New York: Routledge. 134–159.
- Teubert, Wolfgang (2000). Korpuslinguistik und Lexicographie. *Deutsche Sprache*, 4(99), 292–313.
- Tognini-Bonelli, Elena (2007). The corpus-driven approach. V Teubert, Wolfgang in Krishnamurty, Ramesh (ur.), *Corpus Linguistics: Critical Concepts in Linguistics*. London, New York: Routledge. 74–92.
- Tognini-Bonelli, Elena (2001). *Corpus Linguistics at Work*. Amsterdam: John Benjamins.
- Toporišič, Jože (2004). *Slovenska slovnica: četrta izdaja*. Maribor: Obzorja.
- Van der Meer, Geart (2000). Core, subsense and the *New Oxford Dictionary of English* (NODE). On how meanings hang together, and not separately. V Heid, Ulrich et al. (ur.), *Euralex*. Stuttgart, Germany. 419–431.
- Van Eynde, Frank in Gibbon, Dafydd (ur.) (2000). *Lexicon development for speech and language processing*. Dordrecht: Kluwer Academic Publishers.
- Varantola, Krista (2003). Linguistic corpora (databases) and the compilation of dictionaries. V van Sterkenburg, Piet (ur.), *A Practical Guide to Lexicography*. Amsterdam: John Benjamins Publishing Company. 228–240.
- Verdonik, Darinka (2008). Označevanje vrste diskurzivnih označevalcev. V Erjavec, Tomaž in Žganec Gros, Jerneja (ur.), *Jezikovne tehnologije*. Ljubljana: IJS. 25–28.
- Verdonik, Darinka, Žgank, Andrej in Pisanski Peterlin, Agnes (2008). Validacija označevanja diskurzivnih označevalcev v korpusih Turdis-2 in BNSlint. V Erjavec, Tomaž in Žganec Gros, Jerneja (ur.), *Jezikovne tehnologije*. Ljubljana: IJS. 29–32.
- Verdonik, Darinka (2006). Pragmatically annotated corpora in speech-to-speech translation. V Erjavec, Tomaž in Žganec Gros, Jerneja (ur.), *Jezikovne tehnologije*. Ljubljana: IJS. 50–55.
- Verdonik, Darinka, Rojc, Matej in Kačič, Zdravko (2004). Creating Slovenian Language Resources for Development of Speech-to-Speech Translation Components. *International Conference on Language Resources and Evaluation*. Lizbona: ELRA. 1399–1402.
- Verdonik, Darinka in Rojc, Matej (2004). Jezikovni viri projekta LC-STAR. V Erjavec, Tomaž in Žganec Gros, Jerneja (ur.), *Jezikovne tehnologije*. Ljubljana: IJS. 42–47.
- Vidovič Muha, Ada (2006). Kategorialnost leksemov med slovarjem in slovnico. *Slavistična revija*, 54 (Posebna številka), 23–42.
- Vidovič Muha, Ada (2000). *Slovensko leksikalno pomenoslovje: Govorica slovarja*. Ljubljana: Znanstveni inštitut Filozofske fakultete.

Vidovič Muha, Ada (1993). Glagolske sestavljenke – njihova skladenjska podstava in vezljivostne lastnosti. *Slavistična revija*, 41 (1), 161–192.

Vintar, Špela (2008). *Terminografija: terminološka veda in računalniško podprta terminografija*. Ljubljana: Univerza v Ljubljani, Filozofska fakulteta.

Vintar, Špela in Erjavec, Tomaž (2008). iKorpus in luščenje izrazja za Islovar. V Erjavec, Tomaž in Žganec Gros, Jerneja (ur.), *Jezikovne tehnologije*. Ljubljana: IJS. 65–69.

Vintar, Špela (2002). Avtomatsko luščenje izrazja iz slovensko-angleških vzporednih besedil. V Erjavec, Tomaž in Žganec Gros, Jerneja (ur.), *Jezikovne tehnologije*. Ljubljana: IJS. 78–85.

Vintar, Špela (1999). Računalniško podprto iskanje terminologije v slovensko-angleškem vzporednem korpusu. *Uporabno jezikoslovje*, (7/8), 156–169.

Wanner, Leo, Bohnet, Bernd in Giereth, Mark (2006). What is Beyond Collocations? Insights from Machine Learning Experiments. V Corino, Elisa et al. (ur.), *Euralex*. Torino, Italia: European Association for Lexicography. 1071–1087.

Zemljarič Miklavčič, Jana (2006). Korpus govorne slovenščine. V Erjavec, Tomaž in Žganec Gros, Jerneja (ur.), *Jezikovne tehnologije*. Ljubljana: IJS. 124–127.

Zupan, Jure in Čeh, Blaž (2008). *Navodila za uporabo računalniškega programa SLON-13*. Ljubljana: Poročilo projekta DP–KI 2448.

Zupan, Jure (1999). Problemi in nekaj rešitev računalniških obdelav slovenskih besedil. *Slavistična revija*, 47 (3), 278–296.

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

Žele, Andreja (2006). Vezljivost v slovenskem knjižnem jeziku (s poudarkom na glagolu). *Slavistična revija*, 54 (Posebna številka), 43–55.

Žele, Andreja (2003). Slovarska obravnava povedkovnika. *Jezik in slovstvo*, 48 (2), 3–15.

Žele, Andreja (2001). *Vezljivost v slovenskem jeziku (s poudarkom na glagolu)*. Ljubljana: Inštitut za slovenski jezik Frana Ramovša ZRC SAZU.

Žganec Gros, Jerneja, Mihelič, France in Dobrišek, Simon (2000). Govorne tehnologije: pridobivanje in pregled govornih zbirk za slovenski jezik. *Jezik in slovstvo*, 48 (3–4), 47–59.

Žganec Gros, Jerneja, Mihelič, France, Dobrišek, Simon, Erjavec, Tomaž in Žganec, Mario (2000). A Phonetically and Prosodically Annotated Slovene Speech Corpus. V Erjavec, Tomaž in Žganec Gros, Jerneja (ur.), *Jezikovne tehnologije*. Ljubljana: IJS. 27–30.

SLOVARJI

SSKJ: *Slovar slovenskega knjižnega jezika na CD-romu z Odzadnjim slovarjem slovenskega jezika in Besediščem slovenskega jezika z oblikoslovnimi podatki*. Elektronska izdaja (ver. 1.1). SAZU in ZRC SAZU, Inštitut za slovenski jezik Frana Ramovša, DZS.

SP: *Slovenski pravopis*. Elektronska izdaja (ver. 1.0). SAZU in ZRC SAZU, Inštitut za slovenski jezik Frana Ramovša.

INTERNETNE STRANI

Projekti

Jezikoslovno označevanje slovenščine – <<http://nl.ijs.si/jos/>>.

Sporazumevanje v slovenskem jeziku – <<http://www.slovenscina.eu/>>.

Multext-East – <<http://nl.ijs.si/ME/V3/>>.

SloWNet – <http://lojze.lugos.si/~darja/slownet.html>

Programska oprema

Word Sketches (Besedne skice) – <<http://www.sketchengine.co.uk/>>.

WordSmith Tools – <<http://www.lexically.net/wordsmith/>>.

Komodo IDE – <<http://www.activestate.com/komodo/>>.

Korpusi

FidaPLUS – <<http://www.fidaplus.net>>.

Podjetja

Amebis, d. o. o, Kamnik – <<http://www.amebis.si>>.

Trojina, zavod za uporabno slovenistiko – <<http://www.trojina.si/Vsebine/SI/Domov/Domov.aspx>>.

Priporočila Eagles - <<http://www.ilc.cnr.it/EAGLES/annotate/annotate.html>>.

Jezikoslovje v Wikipediji – <<http://sl.wikipedia.org/wiki/Jezikoslovje>>.

KAZALO TABEL

Tabela 1: Primer označenega besedila v korpusu FidaPLUS.....	11
Tabela 2: Natančnost označevanja korpusa jos100k.....	13
Tabela 3: Vzorci za luščenje samostalniških terminoloških besednih zvez.....	14
Tabela 4: ASES – pajek: <i>Oblike besed ter ustrezajoče oblikoskladenjske oznake</i>	27
Tabela 5: ASES – pajek: <i>Besedne zveze in pomenski odnosi</i>	27
Tabela 6: ASES – moder: <i>Oblike besed ter ustrezajoče oblikoskladenjske oznake</i>	30
Tabela 7: ASES – moder: <i>Besedne zveze in pomenski odnosi</i>	30
Tabela 8: Razlaga simbolov za ponazarjanje skladenjskih odnosov v sistemu ASES.....	32
Tabela 9: ASES – izdati: <i>Oblike besed ter ustrezajoče oblikoskladenjske oznake</i>	33
Tabela 10: ASES – izdati: <i>Pomenski odnosi</i>	33
Tabela 11: Raziskovalni cilji ter vprašanja.....	42
Tabela 12: Izhodiščne besede za gradnjo podkorpusa.....	43
Tabela 13: Primer označenega besedila v podkorpusu.....	44
Tabela 14: Razlike v številu lem v korpusu FidaPLUS ter podkorpusu Fida+X.....	45
Tabela 15: Preimenovanje oznak za obravnavane leme.....	46
Tabela 16: Odstranitev oznak iz besedila.....	46
Tabela 17: Menjava ločil z oznakami.....	47
Tabela 18: Dodajanje oznake za konec povedi.....	47
Tabela 19: Končno stanje pretvorjenega besedila.....	48
Tabela 20: Del konkordančnega niza z jedrom <i>pajek</i> in oblikoskladenjskimi oznakami.....	49
Tabela 21: Del seznama tridelnih vzorcev z lemo <i>pajek</i>	49
Tabela 22: Delež pri prvi selekciji izločenih vzorcev z lemo <i>pajek</i>	50
Tabela 23: Prva selekcija najpogostejših 30 tridelnih vzorcev z lemo <i>pajek</i>	50
Tabela 24: Število v analizo zajetih vzorcev.....	51
Tabela 25: Primer vzorčnih zapolnitev za PAJEK DT SOSET.....	51
Tabela 26: Delež vzorcev s samimi polnopomenskimi besednimi vrstami med vsemi najpogostejšimi.....	65
Tabela 27: Razvrstitev najpogostejših vzorcev z lemo <i>pajek</i> in oblikoskladenjskimi oznakami za polnopomenske besede v vzorčne tipe.....	66
Tabela 28: Izluščeni podatki <i>pajek</i> + Sam – ujemalne zveze.....	69
Tabela 29: Izluščeni podatki <i>pajek</i> ₂ + Sam ₂	69
Tabela 30: Izluščeni podatki <i>pajek</i> ₂ + Sam ₂	69
Tabela 31: Razvrščeni podatki <i>pajek</i> + Sam – ujemalne zveze.....	70
Tabela 32: Razvrščeni podatki <i>pajek</i> + Sam ₂	70
Tabela 33: Izluščeni podatki <i>pajek</i> + Sam _U – neujemalni.....	71
Tabela 34: Izluščeni podatki <i>pajek</i> + Sam _U – vsi.....	71
Tabela 35: Izluščeni podatki <i>pajek</i> + Sam – analiza označenosti.....	73
Tabela 36: Razvrščeni podatki Sam + <i>pajek</i> – ujemalne zveze.....	74
Tabela 37: Izluščeni podatki Sam + <i>pajek</i> ₂	75
Tabela 38: Dopolnjeni podatki Sam + <i>pajek</i> ₂	75
Tabela 39: Izluščeni podatki <i>pajek</i> + Prid.....	77
Tabela 40: Razvrščeni podatki <i>pajek</i> + Prid.....	78
Tabela 41: Izluščeni podatki <i>pridevnik</i> + PAJEK.....	80
Tabela 42: Prid + <i>pajek</i> – neujemanje v spolu.....	82
Tabela 43: Prid + <i>pajek</i> – neujemanje v sklonu.....	82
Tabela 44: Izluščeni podatki – <i>pajek</i> + Prisl.....	82
Tabela 45: Izluščeni podatki – Prisl + <i>pajek</i>	83
Tabela 46: Izluščeni podatki – Glag + <i>pajek</i>	83
Tabela 47: Urejeni podatki – Glag + <i>pajek</i>	84
Tabela 48: Izluščeni podatki – <i>pajek</i> + Glag.....	85

Tabela 49: Nabor najpogostejših tridelnih vzorcev z lemo <i>pajek</i> in oblikoskladenjskimi oznakami za polnopomenske besede.	86
Tabela 50: Izluščeni podatki – Prid + Prid + <i>pajek</i> .	86
Tabela 51: Izluščeni podatki – Prid + <i>pajek</i> + Sam.	86
Tabela 52: Preostali tridelni vzorci z lemo <i>pajek</i> in dvema polnopomenskima besedama.	87
Tabela 53: Razvrstitev najpogostejših vzorcev z lemo <i>strasten</i> in oblikoskladenjskimi oznakami za polnopomenske besede v vzorčne tipe.	88
Tabela 54: Izluščeni podatki – <i>strasten</i> + Sam _m .	90
Tabela 55: Izluščeni podatki – <i>strastna</i> + Sam _z .	90
Tabela 56: Izluščeni podatki – <i>strastno</i> + Sam _s .	90
Tabela 57: Analiza označenosti – neujemalne zveze <i>strasten</i> + Sam.	91
Tabela 58: Izluščeni podatki – <i>samostalnik</i> + STRASTEN.	91
Tabela 59: Izluščeni podatki – <i>strasten</i> + Prid.	92
Tabela 60: Analiza označenosti – neujemalne zveze <i>strasten</i> + Prid.	92
Tabela 61: Izluščeni podatki – Prid + <i>strasten</i> .	93
Tabela 62: Analiza označenosti Prid + <i>strasten</i> .	93
Tabela 63: Izluščeni podatki – Prisl + <i>strasten</i> .	94
Tabela 64: Izluščeni podatki – Glag + <i>strasten</i> .	95
Tabela 65: Nabor najpogostejših tridelnih vzorcev z lemo <i>strasten</i> in oblikoskladenjskimi oznakami za polnopomenske besede.	96
Tabela 66: Izluščeni podatki – Prisl + <i>strasten</i> + Sam.	96
Tabela 67: Izluščeni podatki – Prisl + <i>strastna</i> + Sam.	96
Tabela 68: Izluščeni podatki – Prisl + <i>strastno</i> + Sam.	96
Tabela 69: Izluščeni podatki – Glag + <i>strasten</i> + Sam.	97
Tabela 70: Izluščeni podatki – <i>strasten</i> + Prid + Sam, vsi spoli.	97
Tabela 71: Izluščeni podatki – <i>strasten</i> + Sam + Prid.	98
Tabela 72: Izluščeni podatki – <i>strasten</i> + Sam + Sam, vsi spoli.	99
Tabela 73: Nabor preostalih tridelnih vzorcev z lemo <i>strasten</i> in dvema polnopomenskima besedama.	99
Tabela 74: Razvrstitev najpogostejših vzorcev z lemo <i>plesati</i> in oblikoskladenjskimi oznakami za polnopomenske besede v vzorčne tipe.	100
Tabela 75: Izluščeni podatki – Prisl + <i>plesati</i> .	101
Tabela 76: Izluščeni podatki – Prisl (primernik/presežnik) + <i>plesati</i> .	102
Tabela 77: Izluščeni podatki – <i>plesati</i> + Prisl.	103
Tabela 78: Izluščeni podatki – <i>plesati</i> + Prid.	105
Tabela 79: Izluščeni podatki – Prid + <i>plesati</i> .	106
Tabela 80: Izluščeni podatki – Glag + <i>plesati</i> .	106
Tabela 81: Izluščeni podatki – <i>plesati</i> + Sam.	107
Tabela 82: Analiza označenosti – PLESATI SOMEI.	108
Tabela 83: Izluščeni podatki – Sam + <i>plesati</i> .	109
Tabela 84: Nabor najpogostejših tridelnih vzorcev z lemo <i>plesati</i> in oblikoskladenjskimi oznakami za polnopomenske besede.	110
Tabela 85: Izluščeni podatki – <i>plesati</i> + Prid + Sam.	110
Tabela 86: Nabor preostalih tridelnih vzorcev z lemo <i>plesati</i> in dvema polnopomenskima besedama.	111
Tabela 87: Razvrstitev najpogostejših vzorcev z lemo <i>temeljito</i> in oblikoskladenjskimi oznakami za polnopomenske besede v vzorčne tipe.	112
Tabela 88: Izluščeni podatki – Prisl + <i>temeljito</i> .	113
Tabela 89: Izluščeni podatki – <i>temeljito</i> + Prisl.	114
Tabela 90: Izluščeni podatki – <i>temeljito</i> + Glag.	116
Tabela 91: Izluščeni podatki – Glag + <i>temeljito</i> .	117
Tabela 92: Izluščeni podatki – Sam + <i>temeljito</i> .	118
Tabela 93: Izluščeni podatki – <i>temeljito</i> + Sam.	119
Tabela 94: Izluščeni podatki – <i>temeljito</i> + Prid.	120
Tabela 95: Nabor najpogostejših tridelnih vzorcev z lemo <i>temeljito</i> in oblikoskladenjskimi oznakami za polnopomenske besede.	121

Tabela 96: Izluščeni podatki – Prisl + <i>temeljito</i> + Glag.	122
Tabela 97: Izluščeni podatki – Glag + <i>temeljito</i> + Glag.	123
Tabela 98: Izluščeni podatki – <i>temeljito</i> + Glag + Sam.	124
Tabela 99: Izluščeni podatki – Sam + <i>temeljito</i> + Glag.	124
Tabela 100: Izluščeni podatki – Prid + Sam + <i>temeljito</i> .	125
Tabela 101: Nabor preostalih tridelnih vzorcev z lemo <i>temeljito</i> in dvema polnopomenskima besedama.	125
Tabela 102: Delež vzorcev s predlogom med vsemi najpogostejšimi.	126
Tabela 103: Izluščeni podatki – Pred + <i>pajek</i> .	127
Tabela 104: Analiza označenosti Pred + <i>pajek</i> .	127
Tabela 105: Izluščeni podatki – <i>pajek</i> + Pred.	128
Tabela 106: Izluščeni podatki – Pred + <i>strasten</i> .	129
Tabela 107: Izluščeni podatki – <i>strasten</i> + Pred.	129
Tabela 108: Izluščeni podatki – Pred + <i>plesati</i> .	129
Tabela 109: Izluščeni podatki – <i>plesati</i> + Pred.	130
Tabela 110: Izluščeni podatki – Pred + <i>temeljito</i> .	131
Tabela 111: Izluščeni podatki – <i>temeljito</i> + Pred.	132
Tabela 112: Kolokatorji (Sam ₅) predloga <i>po</i> .	133
Tabela 113: Nabor najpogostejših tridelnih vzorcev z lemo <i>pajek</i> in oblikoskladenjsko oznako za predlog.	134
Tabela 114: Izluščeni podatki – <i>pajek</i> + Pred + Sam.	134
Tabela 115: Izluščeni podatki – Sam + Pred + <i>pajek</i> .	134
Tabela 116: Izluščeni podatki – Glag + Pred + <i>pajek</i> .	135
Tabela 117: Nabor preostalih tridelnih vzorcev z lemo <i>pajek</i> in oznako za predlog.	135
Tabela 118: Nabor najpogostejših tridelnih vzorcev z lemo <i>pajek</i> in oblikoskladenjsko oznako za predlog.	135
Tabela 119: Izluščeni podatki – Pred + Prid + <i>strasten</i> .	136
Tabela 120: Izluščeni podatki – <i>strasten</i> + Sam + Pred.	136
Tabela 121: Preostali tridelni vzorec z lemo <i>strasten</i> in oznako za predlog.	137
Tabela 122: Nabor najpogostejših tridelnih vzorcev z lemo <i>plesati</i> in oblikoskladenjsko oznako za predlog.	137
Tabela 123: Izluščeni podatki – <i>plesati</i> + Pred + Sam.	138
Tabela 124: Izluščeni podatki – <i>plesati</i> + Pred + Prid.	139
Tabela 125: Izluščeni podatki – Prisl + <i>plesati</i> + Pred.	139
Tabela 126: Izluščeni podatki – Pred + Sam + <i>plesati</i> .	140
Tabela 127: Preostala tridelna vzorca z lemo <i>plesati</i> in oznako za predlog.	140
Tabela 128: Nabor najpogostejših tridelnih vzorcev z lemo <i>temeljito</i> in oblikoskladenjsko oznako za predlog.	140
Tabela 129: Izluščeni podatki – Pred + Sam + <i>temeljito</i> .	141
Tabela 130: Izluščeni podatki – <i>temeljito</i> + Glag + Pred.	142
Tabela 131: Delež vzorcev z veznikom med vsemi najpogostejšimi.	143
Tabela 132: Primeri nerelevantnih vzorcev z lemo <i>pajek</i> ter oblikoskladenjsko oznako za veznik.	143
Tabela 133: Nabor najpogostejših tridelnih vzorcev z lemo <i>pajek</i> in oblikoskladenjsko oznako za veznik.	143
Tabela 134: Izluščeni podatki – simetrična priredna vzorčna tipa z lemo <i>pajek</i> .	144
Tabela 135: Samostalniki, ki stopajo v priredje s samostalnikom <i>pajek</i> .	144
Tabela 136: Analiza označenosti: Sam + Vp + Sam – neujemanje v sklonu.	145
Tabela 137: Preostali tridelni vzorec z lemo <i>pajek</i> in prirednim veznikom.	145
Tabela 138: Nabor najpogostejših tridelnih vzorcev z lemo <i>strasten</i> in oblikoskladenjsko oznako za veznik.	146
Tabela 139: Izluščeni podatki – simetrična priredna vzorčna tipa z lemo <i>strasten</i> .	146
Tabela 140: Pridevniki, ki stopajo v priredje s pridevnikom <i>strasten</i> .	147
Tabela 141: Analiza označenosti: Prid + Vd + Prid – neujemanje v sklonu.	148
Tabela 142: Izluščeni podatki – Sam + Vp + <i>strasten</i> .	148
Tabela 143: Preostala tridelna vzorca z lemo <i>strasten</i> in prirednim veznikom.	148
Tabela 144: Nabor najpogostejših tridelnih vzorcev z lemo <i>plesati</i> in oblikoskladenjsko oznako za veznik.	149
Tabela 145: Izluščeni podatki – simetrična priredna vzorčna tipa z lemo <i>plesati</i> .	149
Tabela 146: Glagoli, ki stopajo v priredje z glagolom <i>plesati</i> .	150
Tabela 147: Preostali tridelni vzorci z lemo <i>plesati</i> in prirednim veznikom.	151
Tabela 148: Nabor najpogostejših tridelnih vzorcev z lemo <i>temeljito</i> in oblikoskladenjsko oznako za veznik.	151
Tabela 149: Izluščeni podatki – simetrična priredna vzorca z lemo <i>temeljito</i> .	152

Tabela 150: Prislovi, ki stopajo v priredje s prislovom <i>temeljito</i> .	153
Tabela 151: Preostala tridelna vzorca z lemo <i>temeljito</i> in prirednim veznikom.	153
Tabela 152: Nabor vseh možnih oblikoskladenjskih oznak za t. i. pomožne glagolske oblike.	154
Tabela 153: Najpogostejši tridelni vzorci z oblikoskladenjsko oznako za pomožni glagol ter obravnavano lemo.	154
Tabela 154: Najpogostejši tridelni vzorci z oblikoskladenjsko oznako za členek ter obravnavano lemo.	156
Tabela 155: Najpogostejši tridelni vzorci z oblikoskladenjsko oznako za okrajšavo ter obravnavano lemo.	157
Tabela 156: Najpogostejši tridelni vzorci z oblikoskladenjsko oznako za zaimek ter obravnavano lemo.	158
Tabela 157: Najpogostejši tridelni vzorci z oblikoskladenjsko oznako za okrajšavo ter obravnavano lemo.	159
Tabela 158: N + <i>pajek</i>	161
Tabela 159: <i>pajek</i> + N	162
Tabela 160: N + <i>strasten</i>	162
Tabela 161: <i>strasten</i> + N	163
Tabela 162: N + <i>plesati</i>	164
Tabela 163: <i>plesati</i> + N	165
Tabela 164: N + <i>temeljito</i>	166
Tabela 165: <i>temeljito</i> + N	167
Tabela 166: Nabor manj pogostih dvodelnih vzorcev z lemo <i>pajek</i> .	167
Tabela 167: Manj pogosti vzorčni tipi z lemo <i>pajek</i> .	169
Tabela 168: Nabor za luščenje relevantnih vzorčnih tipov.	174
Tabela 169: Nabor vzorčnih tipov, ki vsebujejo glagol.	175
Tabela 170: Nabor za luščenje nerelevantnih vzorčnih tipov.	178
Tabela 171: Novi vzorčni tipi za leme <i>Slovenka</i> , <i>oven</i> , <i>debata</i> .	179
Tabela 172: Samostalniške besedne zveze za leme <i>Slovenka</i> , <i>oven</i> , <i>debata</i> .	180
Tabela 173: Primerjava nabora vzorcev za luščenje samostalniških besednih zvez.	181

KAZALO SLIK IN GRAFOV

Slika 1: Vmesnik sistema ASES – B:S:SL:pajek (neživ.).	21
Slika 2: Vmesnik sistema ASES – B:S:SL:pajek (živ.).	22
Slika 3: Vmesnik sistema ASES – P:S: pajek (vozilo z žerjavom za odvoz nepravilno parkiranih avtomobilov).	23
Slika 4: Vmesnik sistema ASES – P:S: pajek (red) [Araneae].	24
Slika 5: Vmesnik sistema ASES – urejanje pomena P:S: moder (barva).	25
Slika 6: ASES – pajek – členitev iztočnic.	26
Slika 7: ASES – pajek – odnosi med iztočnicami.	28
Slika 8: ASES – moder – členitev iztočnic.	29
Slika 9: ASES – moder – odnosi med iztočnicami.	31
Slika 10: ASES – izdati – členitev iztočnic.	32
Slika 11: ASES – izdati – odnosi med iztočnicami.	34
Slika 12: Razvojni krog nadgrajevanja leksikalne zbirke.	36

Graf 1: Vsota pogostnosti pojavitev vzorcev za dvodelne vzorčne tipe z lemo <i>pajek</i> in oblikoskladenjskimi oznakami za polnopomenske besede.	36
Graf 2: Vsota pogostnosti pojavitev vzorcev za dvodelne vzorčne tipe z lemo <i>strasten</i> in oblikoskladenjskimi oznakami za polnopomenske besede.	88
Graf 3: Vsota pogostnosti pojavitev vzorcev za dvodelne vzorčne tipe z lemo <i>plesati</i> in oblikoskladenjskimi oznakami za polnopomenske besede.	100
Graf 4: Vsota pogostnosti pojavitev vzorcev za dvodelne vzorčne tipe z lemo <i>temeljito</i> in oblikoskladenjskimi oznakami za polnopomenske besede.	112
Graf 5: Število vzorcev, izločenih zaradi vsebovanja oblikoskladenjske oznake za pomožni glagol.	155
Graf 6: Število vzorcev, izločenih zaradi vsebovanja oblikoskladenjske oznake za členek.	156
Graf 7: Število vzorcev, izločenih zaradi vsebovanja oblikoskladenjske oznake za okrajšavo.	157
Graf 8: Število vzorcev, izločenih zaradi vsebovanja oblikoskladenjske oznake za zaimek.	158
Graf 9: Število vzorcev, izločenih zaradi vsebovanja oblikoskladenjske oznake za števnik.	159
Graf 10: Število dvodelnih vzorcev, izločenih zaradi vsebovanja katere od problematičnih oblikoskladenjskih oznak.	160
Graf 11: Število tridelnih vzorcev, izločenih zaradi vsebovanja katere od problematičnih oblikoskladenjskih oznak.	160

AVTORSKO KAZALO

Arčan, Mihael	72	Ledinek, Nina	9, 12
Arhar, Špela	9, 10, 11, 12, 19, 73, 132	Leech, Geoffrey	2, 36
Atkins, Sue	6, 12	Léon, Jacqueline	2
		Logar, Nataša	14
Bizjak, Aleksandra	10	Lönneker, Birte	10
Boguraev, Bran	7		
Briscoe, Ted	7	Manning, Christopher	6
		McEnery, Tony	2, 6, 8
Chomsky, Noam	2	Mitkov, Ruslan	4, 72
		Mladenić, Dunja	9
Čermák, František	8	Mohorič, Tomaž	16
Džeroski, Sašo	9	Oakes, Michael	6
		Ooi, Vincent	3, 4, 7, 36
Erjavec, Tomaž	9, 10, 12, 13, 14, 19, 43, 81, 98		
		Rojc, Matej	10
Fišer, Darja	9, 18, 19	Romih, Miro	19
Fontenelle, Thierry	6, 7	Rundell, Michael	6, 12
Frawley, William	4	Rupnik, Jan	13
Gantar, Polona	15, 17, 37, 40, 67, 126	Sarossy, Bence	10
Garside, Roger	8	Sinclair, John	6, 7, 37, 40
Gorjanc, Vojko	2, 3, 5, 8, 9, 11, 17, 48, 132	Sritar, Mojca	9
Grčar, Miha	13		
		Teubert, Wolfgang	2, 4, 5, 8, 18, 37, 185
Halliday, Michael	7, 37	Toporišič, Jože	49, 67, 77, 81, 115, 126, 142, 155
Hanks, Patrick	7	Varantola, Krista	16
Holozan, Peter	19	Verdonik, Darinka	9, 10
		Vidovič Muha, Ada	7, 10, 15, 37, 94
Jakopin, Primož	10	Vintar, Špela	14, 15, 40, 48, 67, 72, 81, 98
Kačič, Zdravko	10	Wilson, Andrew	2
Karlgren, Hans	4		
Kilgariff, Adam	8	Zemljarič Miklavčič, Jana	9
Krek, Simon	2, 5, 8, 12, 13, 17, 43	Zupan, Jure	10
Krishnamurty, Ramesh	2, 4, 8, 18, 185		
Kruyt, Truus	16	Žele, Andreja	20, 131
		Žganec Gros, Jerneja	9

STVARNO KAZALO

A

ActiveState Komodo.....	54
Amebis.....	1, 9, 17, 19
antičomskijanstvo	2
ASES.....	10, 16, 17, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 39
vmesnik.....	21, 22, 23, 24, 25, 36
avtomatska prepoznavna pomena	37
avtomatska pripisljivost.....	10, 12, 115, 153, 172
avtomatsko slovnično pregledovanje.....	1

B

Bank of English	6
BesAna.....	19
beseda	
funkcijska	73, 91, 171
neknjižna	161
polnopomenska.....	52, 62, 63, 65, 66, 67, 85, 86, 87, 88, 95, 96, 99, 100, 109, 110, 111, 112, 121, 125, 126, 147, 151, 152, 184
tuja	161, 172
zatipkana.....	161
besediloslovje	7
besedna vrsta	
funkcijska	85, 132, 133
napačno pripisana.....	171
polnopomenska	14, 43, 52, 65, 85, 114
besedna zveza	40
dveh glagolov	85, 106
dveh samostalnikov	59, 60, 61, 67, 81
glagolska	116
idiomatična	15, 17
lastnoimenska.....	72, 138, 171
modifikacijska	156
neidiomatična	15
neujemalna	91, 92, 221
pomensko netransparentna.....	15, 17
pomensko transparentna.....	15
predložna	108, 127, 133, 174, 182, 184
pridevnika s samostalniškim dopolnilom	81, 92, 184
pridevniška.....	172, 173, 184
priredna	91, 144, 172, 173, 181, 182, 184
prislovna	113, 115, 174, 184
pristavčno zložena	49
prosta.....	15
s podrednim veznikom.....	142
samostalniška	60, 63, 68, 85, 87, 94, 96, 99, 105, 110, 111, 132, 133, 172, 173, 174, 178, 179, 181, 182, 183

stalna	15, 30, 37, 81, 138
terminološka.....	ii, 14, 27, 81, 181
tipična	82, 91, 171, 181
v zbirki ASES.....	19, 20, 22, 26, 27, 28, 30, 35
Besedne skice.....	8
besedotvorni odnos	25, 35, 188
bigram	glej vzorec, dvodelni
biološka klasifikacija.....	24, 27

C

COBUILD.....	6
Cyc.....	18

Č

členek	52, 57, 65, 150, 153, 155, 156, 178
približnostne mere.....	104
členjenost oblikoskladenjskih oznak	52, 66, 83, 111, 114, 151, 161
črkovalnik	19

D

denotacijska razlika	18
dvoumnost izluščenih podatkov	67, 68, 79, 97, 105, 110, 113, 117, 124, 171
dvoumnost pomena	37

E

EAGLES	12
empiristični pristop	2
enakopisnost.....	21, 76, 91, 104, 155, 171, 172
enopomenskost jezikovnega vzorca	37
EuroWordNet.....	18
evalvacija	
luščenja podatkov.....	5, 15, 42, 178, 181
označevanja	9
razvrščanja podatkov.....	63
zbirke ASES	19, 25

F

Filozofska fakulteta	1
formalizacija naravnega jezika	18, 185
format xml.....	10, 43, 44, 45, 46, 48, 55, 59
FrameNet	18

G

glagolska predloga	31, 34
--------------------------	--------

glagolska vezljivost	20, 35, 39, 126
gnezdenje zvez	40
gold standard	<i>glej</i> zlati standard
Govorec	19

H

HowNet	18
--------------	----

I

inferiranje	16
Inštitut Frana Ramovša ZRC SAZU	1, 10
Inštitut Jožef Stefan	1, 17
interdisciplinarnost	1, 3

J

jezikoslovje	1
korpusno	2, 3, 4, 5, 6, 8, 38, 185
od spodaj navzgor	185
opisno	1
računalniško	1, 4, 5
uporabno	1
jezikoslovna interpretacija	3, 5
jezikoslovna intuicija	3, 8
jezikoslovni opis	2, 7, 185
Jezikoslovno označevanje slovenščine (JOS) ..	1, 11, 83, 155
jezikovna kompetenca	2
jezikovna performanca	2
jezikovna raba	8, 18, 35, 40, 98, 114
jezikovna tehnologija	1, 3, 5, 16, 19, 35, 38
jezikovni vir...1, 4, 5, 10, 12, 16, 17, 18, 35, 36, 37, 40, 42	
jezikovni znak	20, 21, 37

K

kanonična oblika podatkov <i>glej</i> prikazna oblika podatkov	
Klepec	19
koligacija	6, 39
koligacijska	
analiza	156
informacija	7, 15, 33, 171
preferenca	6
koligacijski podatki	37, 171
kolokabilnost	132
kolokacija	6, 17, 39, 40, 132
kolokacijska	
analiza	91, 132, 156
enota	130, 141
informacija	7, 15, 171
okolica	126
statistika	15

kolokacijska informacija	33
kolokacijski podatki	8, 37, 132
kolokacijsko-koligacijski podatki	33, 35, 37, 39
kolokacijskost	37
kolokator	6, 17, 37, 40, 132, 133
končni prikaz podatkov	40, 53, 65
Konkordančnik ASP32	10, 43, 45, 132
korpus	
Beseda	10
FIDA	17, 18, 43
Fida+X	42, 43, 44, 45
FidaPLUS 6, 9, 10, 11, 12, 35, 39, 41, 42, 43, 44, 45,	
54, 73, 74, 77, 78, 79, 81, 93, 94, 102, 103, 104,	
108, 109, 114, 120, 127, 128, 129, 130, 131, 132,	
148, 157	
iKorpus	14
jos100k	12, 13
jos1M	12
kot metodološko orodje	2, 3
kot reprezentativni vzorec jezika	5
kot vir gradnje hipotez	3
kot vir gradnje hipoteze	8
kot vir preverjanja hipotez	3
referenčni	9, 14, 17, 35, 39, 41, 42, 44, 182, 183
specializirani	35, 36
v raziskavi uporabljeni podkorpus 9, 12, 39, 41, 42,	
43, 44, 45, 46, 48, 54, 55, 56, 57, 59, 60, 61, 62, 63	
korpusni podatki	
notranji	8
zunanji	8
korpusni pristop	2, 3, 5
popolni	8
korpusni šum	17, 182
korpusno besedilo	
neoznačeno	8, 9
kratica	157
kvalificiranje iztočnic	37
glede na rabo	20, 25
kvalitativna analiza	3, 5, 38
kvalitativni jezikovni model	2
kvantitativna analiza	3, 5, 38
kvantitativni jezikovni model	2

L

lastno ime	28, 45, 60, 67, 68, 71, 72, 76, 91, 130, 134, 161, 171
osebno	171
stvarno	171
lekiskalna baza	<i>glej</i> leksikalna zbirka
leksem	37
večbesedni	15, 17
leksikalna enota	6, 37
enobesedna	37, 39

večbesedna	7, 15, 37, 39
leksikalna vrzel	18
leksikalna zbirka ...	16, 17, 19, 20, 25, 27, 28, 31, 33, 35, 72, 79, 134, 145, 171
ASES	<i>glej</i> ASES
gradnja zbirke	7, 18, 38
jezikovnotehnološka	16, 39, 185
korpusna	17, 19, 35
kot del razvojnega kroga	36
predkorpusna	19, 35
slovarska	1, 16, 17, 38
za človeško rabo	7, 17
za strojno rabo	7, 18
leksikografija	6
korpusna	12, 37
računalniška	1, 4, 7
leksikografska aplikacija	7, 8, 48
leksikogramatične informacije	19, 156
leksikogramatika	7
leksikon besednih oblik	9, 155
lematizacija	8, 9, 71, 72, 73, 79, 91, 138, 180, 182
lematizator	9
literal	19
ločilo	45, 46, 47, 49, 50, 55, 56, 57
logaritem verjetnosti (LL)	6, 132
luščenje terminologije	14, 52, 181

M

manjšalnica	20, 25
merska enota	157, 172
modalnost	122
morfem	
prosti glagolski	126
Multext-East	10, 12, 21, 43
načela nadgradnje sistema oznak	12

N

nadpomenskost	18, 19, 20, 24, 25, 27
nametilnik	106
namernost	5
napačna lematizacija	68, 70, 91, 93, 118, 120, 159, 172
nedoločnik	84, 106
nelematizirana beseda	52, 57, 65, 128, 161, 172, 178
neračunalniški pristop	4
nerazpršenost podatkov	130
nezaznamovana oblika pridevnika	81, 89, 98, 182

O

občno ime	10, 43, 45, 67, 72
obdelava naravnega jezika	1, 2, 4, 5, 6, 7, 15, 16, 18, 19, 20, 35, 36, 37, 72, 115, 133, 171, 185

okrajšava	52, 57, 65, 157, 159, 178
ontologija	5, 18
osebik	83, 97, 105, 106, 108, 109, 110, 111, 124, 182
Oxford WordSmith Tools	45, 46, 48
označevalna kategorija	9, 10, 12, 40, 42, 52, 53, 65, 76, 115, 153, 158, 183, 184
napačno pripisana	171
označevalne napake	12, 13, 15, 42, 53, 69, 73, 77, 79, 81, 82, 91, 92, 93, 94, 105, 108, 112, 113, 119, 125, 127, 130, 131, 148, 169, 171, 172, 184
označevalnik	
Amebisov	10, 12, 13
SLON	10
statistični	13
TnT	13
označevanje členkov	153, 155, 172
označevanje lastnih imen	68, 71, 91
označevanje okrajšav	172
označevanje slovenščine	
oblikoskladenjsko	10, 11, 12, 155
označevanje diskurza	9
označevanje govornih zbirk	9
semantično	9, 11
skladenjsko	9, 11, 155
statistično	10
označevanje vrste prislova	114, 184
oznake	
JOS	11, 12, 39, 41, 42, 43, 76
korpusa FidaPLUS	9, 10
LC-STAR	10
Multext-East	10, 12, 21, 44

P

podatkovna baza	<i>glej</i> podatkovna zbirka
podatkovna zbirka	16
podpomenka	27, 30
podpomenskost	18, 24, 25, 27
podspol človeškosti	33
podspol živosti	84, 145
pojmovna mreža	18
pomen	
izlučenih podatkov	15
iztočnice v leksikalni zbirki	17, 20
leksikalne enote	6
posamezne besede	6, 7, 37
pomenska enota	37, 39
pomenska preferenca	6
pomenska prozodija	6
pomenska vsebovanost	18
pomenski odnos	18, 25, 26, 27, 28, 30, 31, 33, 35
pomožni glagol	52, 122, 123, 153, 154, 155, 178
poučevanje jezika	7
povedek	110, 111, 112, 122, 171, 184

povedkov prilastek	105, 108
povedkovnik	100
povedkovo določilo	78, 182
predlog	52, 65, 104, 105, 126, 130, 132, 133, 134, 135, 137, 140, 171, 174, 176
predmet.....	32, 33, 83, 97, 108, 110, 124, 171
predračunalniški pristop.....	4
prekinjenost zvez.....	40
prepoznavna lastnih imen	72, 171
Presis	19
prevodoslovje	7
pridevnik	
deležniški	76, 119, 120
kakovostni	80, 95
svojilni	28, 171
vrstni	80, 119
pridevniški niz.....	92
pridevniški prilastek	2, 183
levi.....	79, 87, 96, 105, 182
na desni.....	77, 182
pridobivanje podatkov	
avtomatsko	4, 5, 7, 37, 53, 183
polavtomatsko	4, 38, 39
prikazna oblika podatkov	14, 15, 68, 69, 75, 81, 87
priklic izluščenih enot.....	15, 40, 68, 83, 84, 87, 101, 115, 123, 153
pripisovanje oznak	
avtomatsko	1, 12, 13
kot poseg v jezikovno realnost.....	9
ročno	12, 13
priporočila TEI	43
priročnik	
jezikovni	5
jezikovnonormativni	35
pravopisni	3
slovarski	1, 17, 103
strokovni	18
prislov	
časovni	122
kratnostni.....	115
merni.....	95, 96, 112, 115, 122, 184
okoliščinski.....	115
ozira	115, 184
stopnje	122
stopnjevalni.....	94, 112
svojstvenostni	115
prislovni prilastek	
desni.....	115, 184
levi.....	115
prislovno določilo	109, 110, 111, 113, 124
proceduralnost obravnave jezika	5
programirani sogovornik	19
programska koda.....	54
programska skripta.....	39, 42, 54, 183

programski jezik Perl	43, 54, 59
protipomenskost.....	18, 19, 20, 25
psiholingvistika.....	2

R

racionalistični pristop.....	2
računalniška aplikacija	1, 4, 7, 35
računalniški pristop	4
računalništvo.....	1, 16
razdvoumljanje	
besednega pomena	37, 39
oznak	9, 10, 11, 12, 17, 79, 91, 105, 127, 171
razpršenost podatkov	12, 40, 52, 53, 83, 161, 183, 184
ročni pregled oznak.....	13, 38
ročni vnos podatkov.....	19, 35, 39, 54

S

samostalniški prilastek	60, 181, 183
desni imenovalni.....	67, 181
desni neujemalni	67, 68, 71, 181
desni ujemalni	67, 68, 71, 181
seznam vzorcev	39, 41, 45, 46, 47, 48, 49, 50, 52, 53, 55, 56, 57, 58, 64, 67, 98, 159, 167, 178, 183
simultano prevajanje govora	10
sinset.....	18
sintetizator govora.....	19
skladenjska enota.....	37
skladenjski odnos	
v zbirki ASES.....	31
skladenjski vzorec	14, 15, 17, 37, 52, 67, 68
skup.....	48
slovnični pregledovalnik.....	19, 35
slovnično število.....	74, 75, 85, 98, 182, 184
sloWNet	18
sociolingvistika	2
sopojavitev	40, 185
sopojavitveni odnos	6, 39
sopojavljanje	2, 14, 35, 37, 103
sopomenskost	18, 19, 20
Sporazumevanje v slovenskem jeziku (SSJ)	1, 17, 18
statistična formula	14
statistična metoda	1, 4, 5, 6, 14, 37, 39, 115
statistična obdelava podatkov	6
statistični pojav	2
statistični učni algoritem	13
stavčni razčlenjevalnik	9
stopnja	182, 184
pridevnika	76, 77
prislova	95, 101
strojni prevajalnik	19, 20, 35
strojno prevajanje	1, 4, 20, 35, 38, 171
strojnoberljivi slovar	5, 7, 19

Š

števnik 40, 52, 57, 65, 153, 158, 159, 178

T

tokenizacija 47, 161
 tradicionalni pristop 4
 trigram *glej* vzorec, tridelni
 Trojina, zavod za uporabno slovenistiko 1, 17
 trpnik 108

U

Univerza v Ljubljani 17
 uspešnost oblikoskladenjskega označevanja 12, 40, 53

V

védenje o svetu 18, 35
 veznik 52, 65, 91, 142, 143, 144, 145, 146, 149, 150,
 151, 155, 172, 174, 176

podredni 142, 182, 184
 priredni 142, 143, 151, 173

W

Wikipedija 1
 Word Sketches *glej* Besedne skice
 wordnet 17, 18, 19

Z

začetnica
 mala 72, 80
 velika 9, 52, 134
 zaimek vii, 52, 158, 178
 osebni 158
 povratni 158
 zlati standard 13
 ZRC SAZU 17