NUI Galway
OÉ Gaillimh

ARAN - Access to Research at NUI Galway

Provided by the author(s) and NUI Galway in accordance with publisher policies. Please cite the published version when available.

Title	Avtomatsko pridobivanje besednih zvez iz korpusa z uporabo leksikona SSJ
Author(s)	Arhar Holdt, Špela; Arcan, Mihael
Publication Date	2011-11-17
Publication Information	Arhar Holdt, Špela, & Aran, Mihael. (2011). Avtomatsko pridobivanje besednih zvez iz korpusa z uporabo leksikona SSJ. Paper presented at the OBDOBJA 30 Meddisciplinarnost v slovenistiki / Interdisciplinarity in Slovene Studies, Ljubljana, Slovenia, 17–19 November .
Publisher	Centre for Slovene as a Second and Foreign Language, Univerity of Ljubljana
Link to publisher's version	https://centerslo.si/simpozij-obdobja/zborniki/obdobja-30/
Item record	http://hdl.handle.net/10379/14956

Downloaded 2021-01-21T22:21:12Z

Some rights reserved. For more information, please see the item record link above.

AVTOMATSKO PRIDOBIVANJE BESEDNIH ZVEZ IZ KORPUSA Z UPORABO LEKSIKONA SSJ

Špela Arhar Holdt
Trojina, zavod za uporabno slovenistiko, Ljubljana

Mihael Arčan
Unit for Natural Language Processing, Galway

UDK 811.163.6'322'373.74'374

Računalniška leksikografija je meddisciplinarno področje, ki se osredotoča na avtomatizacijo leksikografskih postopkov in pripravo leksikalnih podatkovnih zbirk različnih vrst. V prispevku predstavljava postopek avtomatskega pridobivanja besednih zvez samostalnika z ujemalnim pridevniškim prilastkom iz besedilnega korpusa in avtomatsko pripravo izluščenih podatkov v ustrezni besednozvezni obliki z uporabo leksikona besednih oblik SSJ.

računalniška leksikografija, besedilni korpus, luščenje besednih zvez, leksikon besednih oblik SSJ

The field of computational lexicography is an interdisciplinary field, primarily focusing on the automatisation of lexicographic procedures and the building of lexical databases of various kinds. In this paper we describe the automatic extraction of word phrases from a text corpus (phrases that contain adjectives that agree in gender, case, and number with the following noun) and the transformation of extracted lexical data to a syntactically suitable final form by the means of the SSJ morphological lexicon.

computational lexicography, text corpus, word phrase acquisition, SSJ morphological lexicon

1 Računalniška leksikografija

Definicije računalniške leksikografije (RL)[1] v literaturi niso povsem enoznačne, vendar jih po večini združuje dejstvo, da avtorjem »računalniško« v enaki meri, kot jim pomeni »izvedeno *s pomočjo* računalnika«, pomeni tudi »izvedeno *za* računalnik«. Pogled na računalnik, ki je obenem sredstvo in cilj obdelave leksikalnih podatkov, omogoča preplet jezikoslovja in računalništva v novo, meddisciplinarno raziskovalno področje, ki ga na izbranem primeru za slovenščino predstavljava v pričujočem prispevku.

Tipično dvosmerna je denimo Hanksova opredelitev interesov RL: prvič, prestrukturiranje in izraba človeku namenjenih slovarjev za računalniške potrebe, in drugič, uporaba računalniških postopkov za izdelavo novih slovarjev (Mitkov 2003: 49). V literaturi najdemo številne avtorje, ki se osredotočajo na izvedbo prve točke, tj. pridobivanje leksikalnih podatkov iz strojnoberljivih slovarjev, tipično za potrebe avtomatske obdelave naravnega jezika.[2] Ta raziskovalna tema je bila na prehodu v devetdeseta leta prejšnjega stoletja izredno priljubljena, vendar se je kmalu izkazalo, da so za pridobivanje

1 V angleški literaturi *Computational Lexicography*.

2 Tipična primera sta Boguraev in Briscoe 1989 ali Fontenelle 1997. Več o tovrstnem pojmovanju RL je mogoče prebrati v Ooi 1998.

leksikalnih podatkov, namenjenih strojni rabi, slovarji le delno uporabni, saj predvidevajo uporabnikovo inferiranje na mestih, kjer strojna obdelava potrebuje eksplicitne podatke (Ooi 1998: 31).

Ooi (1998: 30) zato ponudi nekoliko drugačno definicijo: prvi glavni cilj RL je razvoj računalniških postopkov za avtomatizacijo leksikografskega dela, drugi pa priprava leksikalnih podatkovnih zbirk (LPZ), namenjenih bodisi strojni bodisi človeški rabi. Osredotočenost na LPZ namesto na slovar predstavlja pomembno razliko. Leksikalna zbirka predstavlja podstat za pripravo jezikovnih priročnikov ali jezikovnih tehnologij, ni pa že sama končni izdelek. LPZ so po vsebini in obliki izredno različne, nekatere so relativno enostavne, druge zelo blizu tega, kar si predstavljamo kot elektronski slovar.[3] Ne glede na vsebino LPZ se RL osredotoča na avtomatske, kvantitativne postopke obdelave jezikovnega gradiva, kvalitativna analiza jezikovnih podatkov pa ostaja v domeni leksikografije – računalniško podprte, vendar v jedru jezikoslovne discipline.

2 Avtomatsko pridobivanje besednih zvez iz besedilnega korpusa

Če si v času priprave slovarja SSKJ leksikografskega dela ni bilo mogoče predstavljati brez ročnega izpisovanja gradiva, si danes težko predstavljamo leksikografijo brez besedilnega korpusa. Avtomatsko pridobivanje leksikalnih podatkov iz korpusnih in primerljivih virov je v literaturi imenovano luščenje leksikalnih podatkov. Podatki, ki jih pridobivamo iz korpusov avtomatsko, so lahko različnih vrst, kar je deloma odvisno od možnosti, ki jih prinaša korpusni vir, in seveda namena vsakega posameznega luščenja.

V prispevku predstavljava postopek luščenja besednih zvez na osnovi korpusnim pojavnicam pripisanih lem in oblikoskladenjskih oznak. Za slovenščino je bila ta metoda že večkrat preizkušena, npr. za luščenje terminoloških besednih zvez.[4]

Kot se je izkazalo pri dosedanjih izvedbah, se metoda sooča z dvema glavnima izzivoma. Prvič, uporaba korpusnih oznak pomeni odvisnost postopka od zmožnosti in kvalitete označevanja. Vemo pa, da avtomatsko pripisovanje oznak v besedila naravnega jezika (zlasti v primeru kompleksnih označevalnih sistemov) nikoli ne more biti povsem natančno niti povsem skladno z obstoječimi jezikoslovnimi ugotovitvami.

Drugič, združevanje jezikovnih podatkov glede na pripisane oznake lahko rezultira v neželeni abstrakciji, ki otežuje nadaljnjo ročno analizo gradiva. Luščenje podatkov vedno pomeni izgubo besedilnega konteksta, zato je ključnega pomena, da kvalitativni del analize pritegne kontekst nazaj v obravnavo (besedilni korpus to seveda omogoča). Obenem je nujno že pri samem luščenju stremeti k pripravi leksikalnih podatkov v ustrezni prikazni obliki, tj. obliki, ki je za leksikografa pregledna in enostavno berljiva.

Ker dobro izkazuje omenjene probleme, obenem pa prinaša za slovenščino zelo tipične besednozvezne podatke, sva v tem prispevku za primer luščenja izbrala samostalniške zveze z levim ujemalnim pridevniškim prilastkom (npr. *državni zbor*, *osnovna šola*, *delovno mesto*).

3 Opis uporabljenih jezikovnih virov

V prispevku uporabljava podatke iz korpusa FidaPLUS 60M (okrog 60 milijonov pojavnic obsegajoč podkorpus referenčnega korpusa FidaPLUS). Podkorpus predstavlja približno 10 % izvornega korpusa; pripravljen je bil zato, ker manjša količina korpusnega gradiva omogoča enostavnejše testiranje programskih postopkov. Podkorpus je pripravljen v formatu XML, je lematiziran in

3 V času pisanja prispevka nastaja za slovenščino LPZ, namenjena tako pripravi virov za človeškega uporabnika kot strojni uporabi (Gantar 2009, 2010).

4 Pregled te teme je v Vintar 2008, 2009.

označen z oblikoskladenjskimi oznakami JOS.[5]

Leksikon besednih oblik SSJ je eden od rezultatov projekta Sporazumevanje v slovenskem jeziku (www.slovenscina.eu). Zajema 100.743 enot, pripravljenih v skladu s standardom Lexical Markup Framework.[6] V času pisanja prispevka se pripravlja vključitev leksikona v terminološki portal Termania, kjer bo prosto dostopen za uporabo.[7]

4 Luščenje podatkov

4.1 Opis postopka

V prvem koraku so bili iz korpusa FidaPLUS 60M izluščeni vsi primeri, kjer se glede na pripisani oblikoskladenjski oznaki skupaj pojavljata pridevnik in samostalnik. Ker med takšnimi rezultati pričakujemo tudi pridevniške zveze z neujemalnim samostalniškim prilastkom (npr. *vreden ogleda*, *dostopen javnosti*, *skrb zbujajoč*), je bil v postopek vključen pogoj, da se pripisani oznaki ujemata v spolu in sklonu.[8]

Zaradi potrebe po združevanju podatkov so bile tako pri pridevnikih kot pri samostalnikih izluščene leme. Med luščenjem je bila s pomočjo izpisa ustrezajočega dela oblikoskladenjske oznake vsakemu primeru pripisana tudi oznaka spola.[9] Rezultati so bili nato glede na pogostnost v korpusu razvrščeni od najbolj do najmanj pogostih. Tabela 1 prikazuje 25 najpogostejših rezultatov luščenja.

Tabela 1: 25 najpogostejših rezultatov luščenja

	izluščeni lemi		pogostost v korpusu	oznaka spola
1	državen	zbor	12.143	M
2	evropski	unija	6310	Z
3	nov	mesto	6256	S
4	osnoven	šola	5790	Z
5	kazniv	dejanje	5523	S
6	svetoven	vojna	5256	Z
7	ustaven	sodišče	4917	S
8	zunanji	minister	4879	M
9	prihodnji	leto	4769	S
10	človekov	pravica	4750	Z
11	nadzoren	svet	4198	M
12	vrsten	red	4165	M
13	deloven	mesto	4096	S
14	New	York	4092	M
15	tiskoven	konferenca	3936	Z
16	nov	Gorica	3785	Z
17	zadnji	leto	3773	S
18	zadnji	čas	3706	M
19	državen	sekretar	3702	M
20	občinski	svet	3667	M
21	mesten	občina	3480	Z
22	poslanski	skupina	3420	Z
23	velik	del	3295	M
24	velik	Britanija	3241	Z
25	notranji	zadeva	3172	Z

4.2 Analiza rezultatov

Prikaz podatkov z lemami ustreza le delno; s tega vidika so problematične predvsem pridevniške leme. Med lematizacijo se namreč vsem pridevnikom pripiše osnovna oblika, ki sovpada z obliko za moški spol, in sicer z nedoločno obliko, kadar obstaja izbira. Sprememba oblike pridevnika je torej potrebna pri ženskem (npr. *evropski unija*, *osnoven šola*, *svetoven vojna*) in srednjem spolu (npr. *nov mesto*, *kazniv dejanje*, *ustaven sodišče*). Izluščeni rezultati pa potrebujejo nadaljnjo avtomatsko obravnavo tudi na ravni

5 Specifikacije označevalnega sistema JOS so na voljo na strani http://nl.ijs.si/jos/josMSD-sl.html in korpus FidaPLUS je na strani www.fidaplus.net.

6 Več o standardu na www.lexicalmarkupframework.org. Več o zasnovi leksikona v Arhar 2009.

7 Stran terminološkega portala Termania: www.termania.net. Leksikon bo mogoče kot podatkovno bazo prenesti na lasten računalnik pod licenco Creative Commons (http://creativecommons.si/licence).

8 Ker se z vsako dodatno specifikacijo pogoja poveča možnost vpliva označevalnih napak na rezultate, je bilo preverjanje ujemanja v številu izpuščeno.

9 Oznaka spola je potrebna za nadaljnjo obdelavo podatkov (glej poglavje 5.2).

kategorije določnosti in zapisa z veliko oz. malo začetnico.

4.2.1 Avtomatsko označevanje določnih in nedoločnih oblik pridevnika

V *Slovenski slovnici* (Toporišič 2004: 320) lahko preberemo, da se v slovenščini pojavlja skupina pridevnikov, pri katerih se »v imenovalniku in enako se glasečem tožilniku ednine oblike za moški spol« izraža razlika med nedoločno in določno obliko.[10] Glede na avtorjeve ugotovitve so ti pridevniki sami na sebi nedoločni, izbira določne oz. nedoločne oblike v jezikovni rabi pa je odvisna od skladenjskih in pomenskih značilnosti besedilnega konteksta (prav tam: 328).

Pri avtomatskem označevanju besedil se to odraža tako, da so tovrstni pridevniki vedno lematizirani v nedoločno obliko, obenem pa je v sklonu, ki razliko lahko izraža, pojavnicam pripisana informacija v oblikoskladenjski oznaki, ali gre za določno ali nedoločno obliko pridevnika. Samo na sebi je takšno označevanje podatkov nesporno, saj na ravni posamezne besedne oblike prinaša dovolj visoko natančnost.

Problem nastane pri luščenju kompleksnejših (v našem primeru npr. besednozveznih) jezikovnih podatkov, saj lahko pridevnik v določenih besednih zvezah nastopa primarno v določni obliki, kar pa se v primeru luščenja lem izgubi: npr. v Tabeli 1 pri *državen [zbor, sekretar]*, *nadzoren svet*, *vrsten red*. Zato je treba pri luščenju besednih zvez iz korpusa razviti postopke na tak način, da se (kjer je to potrebno) v podatkih ohrani oz. naknadno poišče določna oblika pridevnika.

4.2.2 Zapis lastnih imen

Zapis, ki ga na ravni velike ali male začetnice prinaša lema, se v tem prispevku obravnava kot ustrezen, čeprav ni vedno tako (npr. v Tabeli 1 pri *evropski unija*, *nov mesto*, *nov Gorica*, *velik Britanija*). Lastna imena predstavljajo posebej zahtevno poglavje avtomatskega označevanja besedil v naravnem jeziku, zato zahtevajo ločeno obravnavo, ki presega meje pričujočega prispevka.

5 Priprava podatkov v ustrezni prikazni obliki

Podatke je mogoče avtomatsko pretvoriti v ustrezno prikazno obliko, in sicer na osnovi dodatnih luščenj iz korpusa ali z uporabo že pripravljenih podatkovnih virov, kakršen je denimo leksikon besednih oblik. V prispevku predstavljena metoda uporablja oba pristopa, ki ju v nadaljevanju na kratko opredeljujeva.

5.1 Uporaba korpusnih podatkov

Eden od primerov avtomatskega generiranja prikazne oblike na osnovi korpusnih podatkov je predstavljen v prispevku Erjavca in Š. Vintar (2008). Avtorja izluščene besedne nize, kandidate za terminološke besedne zveze, pretvorita v ustrezno prikazno obliko tako, da za vsakega od primerov v korpusu poiščeta imenovalniško obliko, jo izluščita, nato pa z izluščeno pojavnico nadomestita izhodiščno lemo. S tem postopkom rešujeta tako probleme na ravni prikaza spola kot določnosti pridevniške oblike (*spleten stran* → *spletna stran*, *digitalen fotoaparat* → *digitalni fotoaparat*).

Iskanje oblik po korpusu je metoda z relativno omejenim dometom; v ustrezni prikazni obliki je mogoče zapisati le primere, ki se v korpusu pojavijo v imenovalniški obliki. Poleg tega luščenje korpusnih pojavnic prinaša variantne zapise z malimi oz. velikimi črkami (npr. *evropska*, *Evropska*, *EVROPSKA*), kar zahteva dodatno pozornost pri generiranju prikazne oblike. Opisana uporaba korpusa je bila dosedanja praksa pri luščenju besednih zvez, kar gre bolj kot preferenci raziskovalcev pripisati dejstvu, da do nedavnega ni bil na voljo dovolj obsežen prosto

10 Poleg omenjene skupine poznamo tudi pridevnike, ki tovrstne dvojnosti na ravni oblik ne izkazujejo in so posledično s stališča avtomatskega pripisovanja osnovne oblike in naknadnega luščenja podatkov manj problematični.

dostopen leksikon besednih oblik (Vintar 2009: 348). To vrzel na področju jezikovnih virov za slovenščino odpravlja leksikon besednih oblik SSJ.

5.2 Uporaba leksikona

5.2.1 Struktura leksikona SSJ

Leksikon besednih oblik SSJ prinaša za vsako od obravnavanih leksikalnih enot nabor vseh njenih besednih oblik, skupaj z opredelitvijo oblikoskladenjskih lastnosti in pogostnostjo oblike oz. ustrezajoče oblikoskladenjske oznake v korpusu FidaPLUS.

Za primer navajava strukturo opisa ene od oblik pridevnika *državen* (v formatu XML).

Primer 1: Struktura opisa pridevniške oblike v leksikonu SSJ

```
<WordForm>
      <feat att="stopnja" val="nedoločeno" />
      <feat att="spol" val="moški" />
      <feat att="število" val="ednina" />
      <feat att="sklon" val="imenovalnik" />
      <feat att="določnost" val="da" />
      <FormRepresentation>
        <feat att="zapis_oblike" val="državni" />
        <feat att="msd" val="Ppnmeid" />
        <feat att="pogostnost" val="69032" />
      </FormRepresentation>
</WordForm>
```

Če podatke preberemo po vrsti, so pridevniški obliki pripisane oblikoskladenjske lastnosti (stopnja, spol, število, sklon, določnost), opredeljen je zapis oblike (*državni*), oblikoskladenjska oznaka, ki označuje obliko v korpusu (*Ppnmeid*), in podatek, kolikokrat se tako označena oblika pojavlja v korpusu (*69032*).

5.2.2 Spol pridevnika

Za pripravo besednih zvez v ustrezni obliki glede na spol pridevnika je bil pripravljen program, ki (I) primer za primerom bere seznam izluščenih podatkov, (II) pri primerih z oznako spola Z ali S na osnovi pridevniške leme poišče ustrezno obliko v leksikonu SSJ in (III) z njo nato dopolni izvorne podatke. V primeru, da v leksikonu ni iskanega pridevniškega gesla, se primer beleži v ločeno datoteko.

5.2.3 Določnost pridevnika

Priprava pridevnikov v ustrezni obliki glede na določnost je nekoliko bolj zapletena. Najprej je treba ugotoviti, ali je pridevnik z obravnavanega stališča sploh problematičen (glej poglavje 4.2.1). V prvem koraku zato program (I) primer za primerom bere seznam izluščenih podatkov in (II) pri primerih z oznako spola M na osnovi pridevniške leme preveri v leksikonu SSJ, ali obravnavani pridevnik izraža razliko med določno in nedoločno obliko. V primeru variantnosti je potrebna nadaljnja avtomatska obravnava (npr. *državen/državni*), v nasprotnem primeru pa je obravnava zaključena (npr. *zunanji*).

Leksikon SSJ že prinaša podatek o tem, kolikokrat se posamezna pridevniška oblika pojavlja v korpusu FidaPLUS. Npr. v geslu *državen* je na voljo podatek, da se določna oblika *državni* pojavi 69.032-krat, nedoločna oblika *državen* pa samo 101-krat. Ta podatek priča o tipičnem pojavljanju pridevnika in bi lahko sam na sebi razrešil problem prikaza nekaterih besednih zvez. Ker pa je na ravni posameznega primera za izbiro ustrezne oblike bolj dragocena informacija o tem, kako se pridevnik pojavlja znotraj določene besedne zveze (glej poglavje 4.2.1), je treba postopek dopolniti z dodatnim luščenjem korpusnih podatkov.[11]

Za to nalogo je bil pripravljen program, ki (I) v korpusu FidaPLUS 60M poišče vse pojavitve obravnavane besedne zveze v imenovalniku in tožilniku ednine ter nato (II) ločeno prešteje primere, v katerih se znotraj zveze pojavlja pridevnik v določni in nedoločni obliki. Za potrebe preverjanja rezultatov so bile v štetje vključene tudi zveze z morebitno drugačno obliko pridevnika.[12] Sledi

11 Takšna dopolnitev je ključna zlasti za ustrezen prikaz enakopisnih pridevnikov različnega pomena, saj so ti na ravni leksikona zajeti (in prešteti) kot ena leksikalna enota. Npr. v podatku o pogostnosti pojavljanja oblike *bučni* so v leksikonu sešteti tako pridevniki iz zvez tipa *bučni zavitek* kot tudi *bučni aplavz*.

(III) primerjava pogostnosti oblik in (IV) prikaz rezultatov: besedna zveza je prikazana z najbolj pogosto pridevniško obliko; ostale oblike so skupaj s podatkom o pogostnosti navedene ob strani, kar omogoča hitrejšo kvalitativno analizo izluščenih podatkov. V primeru, da v leksikonu ni iskanega pridevniškega gesla, se primer beleži v ločeno datoteko.

Za boljšo predstavo navajava nekaj rezultatov (Tabela 2).

Tabela 2 kaže, da je bila v primeru izluščenega niza lem *oseben avtomobil* kot predlagana oblika izbrana zveza, v kateri se pridevnik pojavlja v določni obliki (*osebni avtomobil*), saj podatki v korpusu izkazujejo, da takšna raba močno prevladuje (1.020 zadetkov v primerjavi z 0 zadetki za zvezo *oseben avtomobil*). V primeru, da kvalitativna analiza pokaže, da je katera od predlaganih besednih zvez neustrezna (ko je npr. zveza v korpusu preredka, da bi bilo mogoče zanesljivo sklepanje o tipični rabi), jo je seveda treba ročno spremeniti. Zadnjo besedo pri pripravi podatkov ima vedno leksikograf; naloga kvalitativne analize je, da podatke pripravi na način, da ima leksikograf z njimi čim manj rutinskega, časovno potratnega dela.

6 Uspešnost predstavljenega postopka

Tabela 3 predstavlja izluščene leksikalne podatke po uporabi vseh v prispevku opisanih programskih postopkov, in sicer za 25 najpogostejših primerov.

Tabela 3: 25 najpogostejših rezultatov z avtomatsko generirano besednozvezno obliko

	izluščeni lemi		pogostost v korpusu		končna oblika
1	državen	zbor	12.143	→	*državni zbor*
2	evropski	unija	6310	→	*evropska unija*
3	nov	mesto	6256	→	*novo mesto*
4	osnoven	šola	5790	→	*osnovna šola*
5	kazniv	dejanje	5523	→	*kaznivo dejanje*
6	svetoven	vojna	5256	→	*svetovna vojna*
7	ustaven	sodišče	4917	→	*ustavno sodišče*
8	zunanji	minister	4879	→	*zunanji minister*
9	prihodnji	leto	4769	→	*prihodnje leto*
10	človekov	pravica	4750	→	*človekova pravica*
11	nadzoren	svet	4198	→	*nadzorni svet*
12	vrsten	red	4165	→	*vrstni red*
13	deloven	mesto	4096	→	*delovno mesto*
14	New	York	4092	→	*New York*
15	tiskoven	konferenca	3936	→	*tiskovna konferenca*
16	nov	Gorica	3785	→	*nova Gorica*
17	zadnji	leto	3773	→	*zadnje leto*
18	zadnji	čas	3706	→	*zadnji čas*
19	državen	sekretar	3702	→	*državni sekretar*
20	občinski	svet	3667	→	*občinski svet*
21	mesten	občina	3480	→	*mestna občina*
22	poslanski	skupina	3420	→	*poslanska skupina*
23	velik	del	3295	→	*velik del*
24	velik	Britanija	3241	→	*velika Britanija*
25	notranji	zadeva	3172	→	*notranja zadeva*

V raziskavi, ki jo opisujeva, je bil leksikon SSJ uporabljen pri obravnavi 979.502 besednih zvez. Izven obravnave je ostalo

Tabela 2: Primer izbiranja ustrezne oblike glede na določnost pridevnika

izluščeni lemi		**Pogostost v korpusu**	**predlagana oblika**	**Pogostost im./tož.**	**alternativna oblika**	**Pogostost im./tož.**
oseben	avtomobil	1719	*osebni avtomobil*	1020	oseben avtomobil	0
javen	razpis	1511	*javni razpis*	1032	javen razpis	2
velik	uspeh	1197	*velik uspeh*	604	veliki uspeh	29
kliničen	center	1169	*klinični center*	1116	kliničen center	0
mlad	človek	629	*mlad človek*	124	mladi človek	20

12 Te najdemo pri besednih zvezah, ki izražajo živost, pri katerih tožilniška oblika ni enaka imenovalniški.

69.297 primerov, ki skupaj prinašajo 37.434 različnih pridevnikov.[13] V grobem bi torej lahko zapisali, da je pokritost podatkov z leksikonom 93,4-odstotna. Glede na to, da je bil leksikon SSJ osnovan na referenčnem korpusu FidaPLUS, podatki, ki jih testirava v pričujočem prispevku, pa prihajajo iz istega vira, je visoka pokritost seveda pričakovana, tovrstna ocena postopka pa pravzaprav nima prave evalvacijske vrednosti. Relevantnejše ocene uspešnosti s tega vidika bodo podane, ko bo postopek uporabljen na zadostni količini gradiva, tudi jezikovno specializiranega, iz različnih korpusnih virov.

Za raziskovalce, ki bodo v prihodnosti izvajali podobna luščenja podatkov, je morda uporabna kratka analiza, v kolikšni meri je bila (oz. ni bila) v rezultatih prikazana ustrezna pridevniška oblika. Za pridobitev ocene je bilo ročno pregledanih 250 izluščenih besednih zvez, in sicer v petih sklopih s po 50 primeri zvez, ki se v korpusu pojavljajo s pogostnostjo 100, 50, 10, 5 ali 1.

Pet primerov je bilo rešenih neustrezno zaradi napak na ravni lematizacije (npr. *levi desna* → *leva desna*) in pet jih je ostalo nerešenih zaradi neobstoja pridevnika v leksikonu (npr. *kuloarski zgodba*, *rdečeoranžen lučka*, *kabinetski razmišljanje*). Poleg tega je mogoče trditi, da pripisana pridevniška oblika ne ustreza še pri dodatnih 25 primerih, pri čemer je bila neustreznost pripisana tudi primerom *novi vinogradnik*, *revni jug*, *perspektivni trg*, ki se v korpusu sicer pojavljajo pogosteje s pridevnikom v določni obliki. Sicer pa pride pri večini primerov do napake zaradi pomanjkanja podatkov v korpusu, saj program v primeru, da se zveza nikoli ne pojavi v imenovalniku oz. tožilniku, izbere eno od pridevniških oblik naključno. Glede na opisano analizo – ki je sicer zgolj okvirna in preliminarna – sledi, da (po najbolj strogih ocenah, vendar brez upoštevanja zapisa z veliko ali malo začetnico) postopek dosega 86-odstotno natančnost.

7 Sklep

Čeprav podrobne analize uspešnosti zaradi že omenjenih razlogov še niso bile opravljene, je mogoče podati oceno, da postopek pretvorbe pridevniških oblik z uporabo leksikona SSJ daje dobre rezultate, zlasti na ravni pripisa ustrezne oblike glede na spol. Izbira oblike glede na določnost dobro deluje pri primerih z relativno visoko pogostnostjo, pri primerih, kjer se imenovalniške oblike v korpusu ne pojavljajo, pa je po pričakovanjih nezanesljiva. Na tem mestu je zaželena izboljšava programa, po kateri bo mogoče na teh mestih upoštevati dodatne informacije, npr. podatke iz leksikona SSJ o pojavljanju pridevniških oblik (glej poglavje 5.2.3).

Preverjanje ustreznosti oblike izluščenih besednih zvez sicer še vedno ostaja kvalitativni analizi, vendar za leksikografijo, ki je v zadnjih desetletjih doživela razvoj, po katerem se pridobivanje tovrstnih jezikovnih podatkov namesto v letih meri v dnevih, takšno stanje pomeni (zgolj še en) korak naprej.

Dostopnost rezultatov

Vse v prispevku uporabljene programske skripte (v programskem jeziku Perl) bodo skupaj z leksikonom na voljo na internetni strani projekta SSJ (www.slovenscina.eu).

Literatura

ARHAR, Špela, 2009: Učni korpus SSJ in leksikon besednih oblik za slovenščino. *Jezik in slovstvo* 54/3–4. 43–56.

BOGURAEV, Bran, BRISCOE, Ted (ur.), 1989: *Computational Lexicography for Natural Language Processing*. London, New York: Longman.

FONTENELLE, Thierry, 1997: *Turning a Bilingual Dictionary into a Lexical–Semantic Database*. Tübingen: Max Niemeyer Verlag.

13 Šteto brez razlikovanja med malimi in velikimi črkami.

GANTAR, Polona, 2009: Leksikalna baza – vse, kar ste vedno želeli vedeti o jeziku. *Jezik in slovstvo* 54/3–4. 69–94.

GANTAR, Polona, 2010: K uporabniku usmerjeni slovnično-leksikalni opisi slovenskega jezika. Vojko Gorjanc, Andreja Žele (ur.): *Izzivi sodobnega jezikoslovja.* Ljubljana: Znanstvena založba Filozofske fakultete. 35–52.

MITKOV, Ruslan (ur.), 2003: *The Oxford Handbook of Computational Linguistics*. Oxford: Oxford University Press.

OOI, Vincent, 1998: *Computer Corpus Lexicography.* Edinburgh: Edinburgh University Press.

TOPORIŠIČ, Jože, [4]2004: *Slovenska slovnica*. Maribor: Obzorja.

VINTAR, Špela, ERJAVEC, Tomaž, 2008: iKorpus in luščenje izrazja za Islovar. Tomaž Erjavec, Jerneja Žganec Gros (ur.): *Zbornik Šeste konference Jezikovne tehnologije, 16. do 17. oktober 2008*. Ljubljana: Institut Jožef Stefan. 65–69.

VINTAR, Špela, 2008: *Terminografija: terminološka veda in računalniško podprta terminografija*. Ljubljana: Znanstvena založba Filozofske fakultete.

VINTAR, Špela, 2009: Samodejno luščenje terminologije – izkušnje in perspektive. Nina Ledinek idr. (ur.): *Terminologija in sodobna terminografija.* Ljubljana: Založba ZRC, ZRC SAZU. 345–356.

Spletne strani

Creative Commons: http://creativecommons.si/licence

Korpus slovenskega jezika FidaPLUS: www.fidaplus.net

Lexical Markup Framework: www.lexicalmarkupframework.org

Oblikoskladenjske oznake JOS: http://nl.ijs.si/jos/josMSD-sl.html

Sporazumevanje v slovenskem jeziku: www.slovenscina.eu