Obeliks:
statistični oblikoskladenjski označevalnik in lematizator za slovenski jezik

Miha Grčar,[1] Simon Krek,[2] Kaja Dobrovoljc[3]

[1, 2] Institut Jožef Stefan
Jamova cesta 39, 1000 Ljubljana
miha.grcar@ijs.si, simon.krek@ijs.si
[3] Trojina, zavod za uporabno slovenistiko
Partizanska cesta 5, 4220 Škofja Loka
kaja.dobrovoljc@trojina.si

Povzetek

V prispevku opisujemo označevalnik za slovenščino Obeliks, ki je bil izdelan v okviru projekta Sporazumevanje v slovenskem jeziku. Označevalnik je sestavljen iz treh komponent: tokenizacijskega modula, ki za stavčno segmentacijo in tokenizacijo uporablja pravila, oblikoskladenjskega označevalnika ter različice lematizatorja LemmaGen, ki je prilagojena za delovanje v kombinaciji z označevalnikom. Učno množico pri opisani različici označevalnika predstavlja učni korpus ssj500k, ki je označen po tabeli oznak JOS. Pri sistemu JOS s 1.903 možnimi oznakami je natančnost označevalnika 91,34 % za celotno oznako in 98,30 % za vrhnjo kategorijo (POS). Natančnost lematizacije je 97,88 % ob upoštevanju velike začetnice ter 98,55 % na ravni črkovnega niza. V prispevku predstavljamo zasnovo označevalnika ter analizo natančnosti označevanja. Označevalnik je prosto dostopen na spletu.

Obeliks: Statistical Morphosyntactic Tagger and Lemmatizer for Slovene

The paper describes Obeliks, a new statistical tagger for Slovene developed within the "Communication in Slovene" project. The new tool consists of three modules: a rule-based sentence splitter and tokenizer, a morphosyntactic tagger, and a version of the LemmaGen lemmatizer which works in combination with the tagger. Obeliks is trained on the ssj500k corpus tagged according to the JOS tagset. In the JOS system which includes 1,903 possible tags, the tagger achieved 91.34% accuracy for all tags and 98.30% for POS only. Lemmatization accuracy is 97.88% with capitalization included and 98.55% for all-lowercase letters. The paper presents the design of the tagger and the analysis of the tagging accuracy. Obeliks is freely available for download on the Web.

1. Uvod

V prispevku opisujemo statistični označevalnik, ki je nastal v okviru projekta Sporazumevanje v slovenskem jeziku.[1] Označevalnik implementira programska oprema Obeliks in zajema (1) segmentacijo in tokenizacijo, (2) oblikoskladenjsko označevanje in (3) lematizacijo.

Segmentacija je proces označevanja začetkov in koncev delov besedil po izbranem kriteriju. V skladu s tradicijo, ki se opira na tipično organizacijo delov besedil pri jezikih z alfabetsko pisavo, ki za označevanje smiselno zaključenih delov besedil uporabljajo ločila (npr. latinična in cirilična), se kot najpomembnejši segment pojavlja stavek, tipično ločen s piko, klicajem ali vprašajem, v nekaterih primerih tudi s tropičjem, pomišljajem ali na drug, manj običajen način.

Tokenizacija je postopek prepoznavanja in določanja posamezih korpusnih pojavnic, ki se na splošno delijo v dve večji skupini. Prva skupina so ločila in simboli, v drugo skupino pa spadajo elementi, ki so zanimivi za kasnejše jezikoslovno označevanje in jih pogosto opredeljujemo z oznako "beseda", čeprav mednje spadajo tudi števila, spletni naslovi ali kombinacije ločil, števk in črk (npr. Micro$oft, U2, AC/DC itd.), ki jih na prvi ravni morda niti ne bi prepoznali pod tem imenom.

Pripisovanje oblikoskladenjskih oznak oz. oblikoskladenjsko označevanje (*POS-tagging*, *part-of-speech tagging*, *word-class tagging*) je ena od najstarejših in najpogostejših oblik dodajanja interpretativnih informacij jezikoslovne narave besedilom, pri čemer posamezni pojavnici v korpusu pripišemo, v kateri osnovni besednovrstni razred spada v specifičnem jezku ter lastnosti, ki jih izkazuje znotraj razreda.

Pri oblikoskladenjskem označevanju je pomembno, kateri model označevanja izberemo; ta je navadno opredeljen s tabelo oznak (*tag set*), ki vsebujejo različno število možnih oznak. Za slovenščino obstaja več tabel oznak; za učenje označevalnika Obeliks je bila izbrana tabela oznak JOS (Erjavec et al., 2010), ki vsebuje 1.903 možne oznake z dvanajstimi vrhnjimi kategorijami.[2] To število je zelo veliko in zato predstavlja zahteven problem za statistične označevalnike.

Zadnja faza v procesu označevanja je lematizacija, ki je jezikovnotehnološki proces pripisovanja osnovne oblike korpusnim pojavnicam pri tistih besednih vrstah, ki so pregibne in tvorijo oblikoslovno paradigmo.

Označevalnik Obeliks procese segmentacije, tokenizacije, oblikoskladenjskega označevanja in lematizacije združuje v enoten proces, ki ga podrobneje opisujemo v naslednjem poglavju.

2. Postopek označevanja

V naslednjih podpoglavjih opisujemo osnovne postopke, tj. segmentacijo in tokenizacijo, oblikoskladenjsko označevanje ter lematizacijo, s poudarkom na oblikoskladenjskem označevanju, ki tvori jedro označevalnika in je osrednja tema tega prispevka.

2.1. Segmentacija in tokenizacija

Segmentacija in tokenizacija v označevalniku Obeliks temeljita na pravilih, ki so bila opredeljena v samostojnem tokenizacijskem programu, opisanem v (Krek, 2010), in so nekoliko prilagojena za konsistentno uporabo v sklopu tokenizacija-oblikoskladenjsko označevanje-lematizacija.

[1] http://www.slovenscina.eu

[2] http://nl.ijs.si/jos/msd/html-sl/index.html.

Izhodišče segmentacijsko-tokenizacijskih načel je slediti čim bolj intuitivnim in razumljivim pravilom, da se izognemo presenečenjem in nerazumevanju pri kasnejših obdelavah tokeniziranih besedil, kar velja predvsem za človeške uporabnike označenih besedil. Kot daleč najbolj problematični ločili se pri tem procesu izkazujeta vezaj in pika.

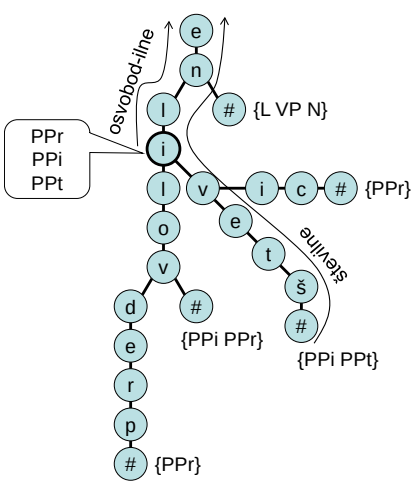
2.2. Oblikoskladenjsko označevanje

Delovanje označevalnika – kot sestavne komponente Obeliksa – temelji na principu nadzorovanega strojnega učenja (angl. *supervised machine learning*) (Mitchell, 1997), kar pomeni, da se označevanja nauči na ročno označenih primerih iz učnega korpusa. Označevalnik v fazi učenja zgradi model znanja, ki mu nato omogoča označevanje novih (še neoznačenih) besed. Osnovne komponente označevalnika so: (i) drevo končnic (*suffix trie*), (ii) algoritem za tvorjenje vektorjev značilk (*feature vectors*), (iii) algoritem za učenje in (iv) algoritem za označevanje. Te koncepte in postopke podrobneje opisujemo v naslednjih podpoglavjih.

2.2.1. Drevo končnic

Drevo končnic je ena poglavitnih komponent označevalnika. Zgrajeno je na podlagi učnega korpusa in leksikona ter "zna" za vsako besedo določiti nabor potencialnih oznak. Glavna prednost drevesa končnic pred običajnim leksikonom je, da lahko deluje tudi pri besedah, ki jih v drevesu ni – v takem primeru algoritem tvori nabor potencialnih oznak na podlagi končnice.

Kot primer navajamo drevo končnic (slika 1), ki vsebuje besede *predvolilne*, *volilne*, *ne*, *številne* in *civilne* (zapisane so od zadnje do prve črke in zaključene z znakom #, od korena drevesa proti listom). Vozlišče z zaključnim znakom vsebuje nabor oznak, ki jih je mogoče pripisati besedi (npr. besedi *ne* lahko pripišemo eno od oznak iz množice {L, VP, N}). Četudi besede ni v drevesu (npr. *osvobodilne*), lahko na podlagi končnice (v tem primeru *-ilne*) oz. pripadajočega vozlišča (označeno na sliki) sklepamo na oblikoslovne oznake, tako da izračunamo unijo vseh množic oznak v poddrevesu s korenom v vozlišču.



Slika 1: Preprost primer drevesa končnic.

kontekst (besede)	tri predhodne besede (w_{-3}, w_{-2}, w_{-1}), trenutna beseda (w_0) in tri sledeče besede (w_{+1}, w_{+2}, w_{+3})
kontekst (oblikoslovne oznake in množice oznak iz drevesa končnic)	tri predhodne oblikoslovne oznake (p_{-3}, p_{-2}, p_{-1}), množice oznak iz drevesa končnic za trenutno besedo in sledeče besede (a_0, a_{+1}, a_{+2}, a_{+3}), posamezne oznake iz množic a (m_0, m_{+1}, m_{+2}, m_{+3})
predpone in končnice	predpone in končnice dolžine od 1 do 4 za tri predhodne besede ($p_{pos,len}$, $s_{pos,len}$, $-3 \le pos \le -1$, $1 \le len \le 4$), trenutno besedo ($p_{0,len}$, $s_{0,len}$, $1 \le len \le 4$) in tri sledeče besede ($p_{pos,len}$, $s_{pos,len}$, $1 \le pos \le 3$, $1 \le len \le 4$)
črkovne in ostale značilke	- Ali beseda vsebuje znak, ki ni črka ali številka? - Ali beseda vsebuje številko? - Ali beseda vsebuje veliko črko? - Ali se beseda začenja z veliko začetnico [in ni prva beseda v stavku]?

Tabela 1: Značilke za oblikoslovno označevanje z uporabo strojnega učenja.

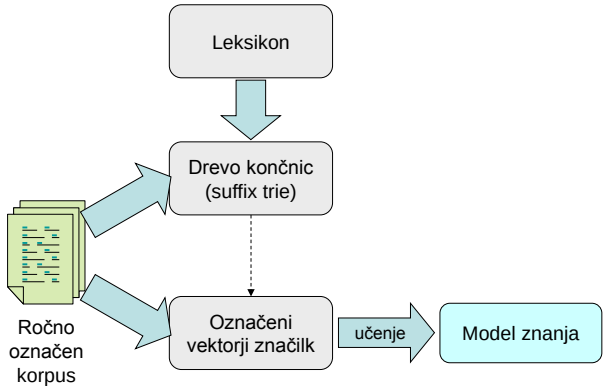
2.2.2. Vektorji značilk

Vsak primer (tj. beseda) je, kot je to navada v strojnem učenju, opisan z vektorjem značilk. Značilke in njihove vrednosti lahko dojemamo kot lastnosti besede, ki pa morajo biti takega tipa, da samostojno ali v kombinaciji z drugimi lastnostmi določajo neko specifiko oblikoslovne oznake, ki se pritiče besedi. Obeliks za značilke uporablja (i) kontekst besede (tj. predhodne in sledeče besede), (ii) predhodne oblikoslovne oznake, (iii) oznake iz drevesa končnic za trenutno in sledeče besede, (iv) predpone in končnice ter (vi) črkovne in ostale značilke, kot je to razvidno iz tabele 1. Značilke so povzete po (Giménez & Márquez, 2004), kjer so tudi bolj podrobno obrazložene. Primer tvorjenja vektorja značilk podaja tabela 2.

L D P ? Še v najboljših časih je redko delovalo.	
w_{-3}=še	$p_{-1,3}$=naj
w_{-2}=v	...
w_{-1}=najboljših	$p_{0,1}$=č
…	$p_{0,2}$=ča
w_{+3}=delovalo	…
p_{-3}=L	$p_{+1,1}$=j
p_{-2}=D	$p_{+1,2}$=je
p_{-1}=P	…
a_0={S}	$p_{+3,1}$=d
a_{+1}={G, Z}	$p_{+3,2}$=de
a_{+2}={R}	…
a_{+3}={P, S, G, …}	$s_{-3,1}$=e
m_0=S	$s_{-3,2}$=še
m_{+1}=G	$s_{-2,1}$=v
m_{+1}=Z	$s_{-1,1}$=h
…	$s_{-1,2}$=ih
m_{+3}=P	…

m_{+3}=S	$s_{0,1}$=h
m_{+3}=G	$s_{0,2}$=ih
…	…
$p_{-3,1}$=š	$s_{+1,1}$=e
$p_{-3,2}$=še	$s_{+1,2}$=je
$p_{-2,1}$=v	…
$p_{-1,1}$=n	$s_{+3,1}$=o
$p_{-1,2}$=na	$s_{+3,2}$=lo
vsebuje znak, ki ni črka ali številka = ne	
vsebuje številko = ne	
vsebuje veliko črko = ne	
se začenja z veliko začetnico = ne	

Tabela 2: Vrednosti značilk za besedo *časih* v stavku *Še v najboljših časih je redko delovalo*.



Slika 2: Učenje označevalnika: (1) iz ročno označenega korpusa in leksikona zgradimo drevo končnic, (2) za vsako besedo tvorimo vektor značilk (pri tem postopku uporabljamo tudi drevo končnic), (3) na podlagi označenih vektorjev značilk tvorimo model znanja.

2.2.3. Algoritem za učenje

Za učenje potrebujemo ročno označeni korpus in leksikon, iz katerih najprej zgradimo drevo končnic. Nato za vsako besedo iz učnega korpusa tvorimo vektor značilk (pri tem za določanje vrednosti značilk *a* in *m* (glej tabeli 1 in 2) uporabimo drevo končnic). Ti vektorji značilk so označeni z oblikoslovnimi oznakami pripadajočih besed in zato lahko uporabimo algoritem za nadzorovano učenje, ki na podlagi označenih vektorjev tvori model znanja.

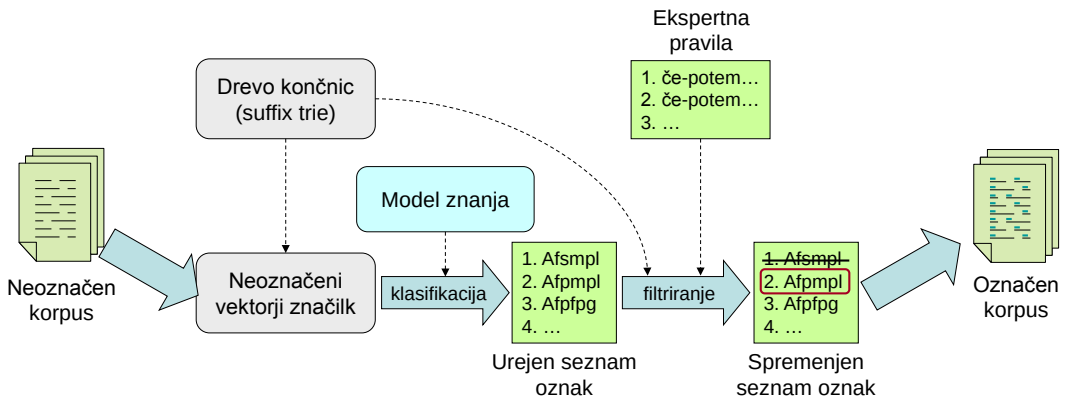
Model vsebuje informacijo o tem, katere lastnosti (tj. pari značilka=vrednost) združujejo določene oblikoslovne kategorije in jih hkrati ločujejo od ostalih. Algoritem, ki ga Obeliks uporablja za učenje, temelji na principu maksimalne entropije (*maximum entropy*) in se pogosto uporablja za označevanje sekvenc (Ratnaparkhi, 1996; Nigam et al. 1999). Postopek učenja prikazuje slika 2.

2.2.4. Algoritem za označevanje

V fazi označevanja algoritem besedam iz neoznačenega korpusa pripiše oblikoslovne oznake (tvori označen korpus). Za vsako besedo najprej tvori vektor značilk na popolnoma enak način kot v fazi učenja. Besede oz. pripadajoči vektorji značilk zaporedno vstopajo v klasifikacijski algoritem, ki na podlagi zgrajenega modela znanja pripiše besedi neko oblikoslovno oznako. Pripisovanje oznake poteka tako, da klasifikacijski algoritem najprej uredi vse oznake padajoče po verjetnosti, da pripadajo dani besedi. Nato s tega seznama nekatere oznake odstranimo. V primeru, da je beseda (v celoti) v drevesu končnic, kar pomeni, da je vsebovana tudi v učnem korpusu in/ali leksikonu, potem so s seznama najprej odstranjene vse oznake, ki v drevesu končnic ne pripadajo besedi. Naknadno seznam spremenimo (ne glede na to, ali drevo končnic vsebuje besedo ali ne) z uporabo ekspertnih pravil, ki jih povzemamo v prilogi. Prvo oznako z vrha tako dobljenega seznama na koncu pripišemo besedi. Postopek označevanja je prikazan na sliki 3.

2.3. Lematizacija

Obeliks za lematizacijo besed uporablja LemmaGen[3] (Juršič, Mozetič, Erjavec, & Lavrač, 2010), implementacijo algoritma za lematizacijo na osnovi pravil tipa RDR (*Ripple Down Rules*). Za vsako kategorijo oblikoslovne oznake (samostalnik, pridevnik, glagol itd.) Obeliks iz učnega korpusa in leksikona zgradi ločeno lematizacijsko drevo. Lematizacija se izvede po označevanju in uporabi oblikoslovno oznako besede za izbiro lematizacijskega drevesa. Če npr., oblikoslovna oznaka predstavlja samostalnik, potem se za lematizacijo pripadajoče besede uporabi "samostalniško" lematizacijsko drevo. Na ta način Obeliks tvori bolj smiselne pare oznaka-lema, saj npr., glagolu *pozná* ne pripiše pridevniške leme *pozen*, temveč glagolsko lemo *poznati*, in obratno.



Slika 3: Označevanje besedila: (1) za vsako besedo tvorimo vektor značilk, (2) klasifikacijski algoritem na podlagi modela znanja, drevesa končnic in ekspertnih pravil pripiše besedi oblikoslovno oznako.

[3] http://lemmatise.ijs.si/

3. Učni korpus ssj500k[4]

Učni korpus ssj500k je bil tako kot označevalnik Obeliks izdelan v okviru projekta Sporazumevanje v slovenskem jeziku in temelji na obeh učnih korpusih, izdelanih v okviru projekta JOS. Sestavljata ga celotni korpus jos100k ter dodatnih 400.000 besed iz enomilijonskega korpusa jos1M. Vsi jezikoslovni metapodatki (oznake, leme, tokenizacija) so bili v korpusu ssj500k še enkrat ročno pregledani, povečana je bila množica skladenjsko označenih in ročno pregledanih povedi. V delu, ki ga zajema korpus jos100k, so bile dodane informacije o lastnih imenih za potrebe strojnih prepoznavalnikov imenskih entitet. Za razliko od korpusov jos100k in jos1M je bila v korpusu ssj500k v celoti ročno pregledana in popravljena tudi stavčna segmentacija in tokenizacija, kar omogoča tudi preverjanje uspešnosti algoritmov pri teh dveh postopkih. Številčni podatki o elementih v korpusu ssj500k so v tabeli 3.

Oznaka	Opis	ssj500k
	besedilo	1.677
<p>	odstavek	8.137
<s>	stavek oz. poved	27.829
<w>	beseda	500.295
<c>	ločilo/simbol	85.953
<w> + <c>	pojavnica	586.248
<links>	element s skladenjskimi povezavami	11.411
<link>	skladenjska povezava	235.865
<chunks>	element s povezavami na imenske entitete	2.178
<chunk>	imenska entiteta	4.398

Tabela 3: Število elementov v učnem korpusu ssj500k

Učni korpus ssj500k je prosto dostopen na spletnih straneh projekta SSJ[5] pod licenco Creative Commons Priznanje avtorstva-Nekomercialno 3.0.[6]

4. Analiza označevanja

Na korpusu ssj500k je bilo izvedeno desetkratno prečno preverjanje natančnosti označevalnika, ki daje naslednje rezultate:

Kategorija	Dod. pogoj	%
natančnost na znanih besedah		93,09
natančnost na neznanih besedah		54,03
skupna natančnost		92,49
natančnost na znanih besedah	(kat.)	98,72
natančnost na neznanih besedah	(kat.)	87,24
skupna natančnost	(kat.)	98,55
natančnost na znanih besedah	(brez ločil)	92,04
natančnost na neznanih besedah	(brez ločil)	53,99
skupna natančnost	(brez ločil)	91,34
natančnost na znanih besedah	(kat., brez ločil)	98,50
natančnost na neznanih besedah	(kat, brez ločil)	87,22
skupna natančnost	(kat., brez ločil)	98,30
natančnost lematizacije	(brez ločil)	97,88
natančnost lematizacije	(male črke, brez ločil)	98,55

Tabela 4: Natančnost označevalnika.

10-kratno prečno preverjanje je metoda, s katero lahko ocenimo natančnost označevalnika na neznanem besedilu (ob predpogoju zadostne heterogenosti označenega korpusa, ki pa je izpolnjen v primeru korpusa ssj500k).[7] Naj na tem mestu še razložimo, da so znane besede tiste besede, ki so vsebovane bodisi v leksikonu bodisi v učnem korpusu, neznane pa tiste, ki se pojavijo izključno v korpusu, ki ga označujemo.

Rezultati kažejo, da ima Obeliks doslej najboljši rezultat pri statističnih označevalnikih, ki so bili uporabljeni za slovenščino (prim. Džeroski et al., 2000). Natančnost označevanja ob upoštevanju zgolj pojavnic, ki so označene kot beseda tj. <w>, brez pojavnic, ki jih je tokenizator opredelil kot ločilo, tj. <c>, je 91,34 %. Ob upoštevanju vrhnje kategorije, ki jih je v tabeli oznak JOS dvanajst, je označevalnik uspešen v 98,30 % primerov. Natančnost lematizatorja je 97,88 % ob upoštevanju velike začetnice pri lemi ter 98,55 % v primeru, da upoštevamo zgolj črkovni niz brez razlikovanja med velikimi in malimi črkami.

Podrobnejše analize napak pri prečnem preverjanju kažejo natančnejšo sliko. Če pri lematizaciji najprej prikažemo napake lematizatorja, pri čemer ne upoštevamo razlike med velikimi in malimi črkami, se napake pojavljajo pri naslednjih oblikoslovnih oznakah (tabela 5):

Oznaka	Opis	Št. Napak
S (Sl+So)	samostalnik (lastno ime + občni)	2.914 (1.790+1.124)
P	pridevnik	1.991
G	glagol	749
R	prislov	577
Z	zaimek	530
K	števnik	127
N	neuvrščeno	116
L	členek	55
V	veznik	30
D	predlog	10
M	medmet	2
O	okrajšava	1
Skupaj		7.102

Tabela 5: Napake lematizacije (ne razlikujemo med velikimi in malimi črkami).

[4] http://www.slovenscina.eu/Vsebine/Sl/Kazalniki/K10.aspx
[5] http://www.slovenscina.eu/tehnologije/ucni-korpus
[6] http://creativecommons.org/licenses/by-nc/3.0/deed.sl
[7] http://en.wikipedia.org/wiki/Cross-validation_%28statistics%29

To tabelo lahko primerjamo s tabelo, v kateri so upoštevane razlike med velikimi in malimi črkami. Rezultate navajamo po besednih vrstah v tabeli 6.

Oznaka	Opis	Št. napak
S (Sl+So)	samostalnik (lastno ime + občni)	5.696 (4.308+1.388)
P	pridevnik	2.082
G	glagol	749
R	prislov	578
Z	zaimek	530
N	številnik	356
K	neuvrščeno	136
L	členek	56
V	veznik	30
D	predlog	12
M	medmet	4
O	okrajšava	2
Skupaj		10.231

Tabela 6: Napake lematizacije (upoštevane male in velike črke).

Pričakovano (prim. Krek, 2010) se na prvem mestu pojavljajo lastna imena, ki so nepredvidljiva in v večini primerov rešljiva zgolj s širjenjem leksikona. Pri pridevnikih, ki so na drugem mestu za samostalniki, gre v treh četrtinah primerov (1.621) za zamenjavo s prislovi. Kategorija lastnih imen in parov pridevnik-prislov torej prispevata približno polovico vseh napak pri lematizaciji.

V tabeli 7 nadalje navajamo napake označevalnika v primerih, kjer je bila oblikoskladenjska kategorija pripisane oznake pravilna, celotna oblikoskladenjska oznaka pa napačna.

Oznaka	Opis	Št. napak
S	samostalnik	17.402
P	pridevnik	8.519
Z	zaimek	5.338
K	številnik	1.428
G	glagol	1.117
D	predlog	818
R	prislov	6
Skupaj		34.628

Tabela 7: Napake označevanja (kategorija pravilna).

Zanimiva je tudi podrobna analiza po posameznih oznakah. V tabeli 8 navajamo prvih petnajst oznak, pri katerih se označevalnik najpogosteje moti.

Zap. št.	Oznaka	Št. napak
1	Sometn	1.657
2	Sozmt	1.405
3	Somei	1.340
4	Slmei	1.243
5	Sozmi	986
6	Slzei	939
7	Soset	702
8	Sosei	660
9	Ppnzmt	611
10	Ppnmetd	528
11	Slmer	474
12	Sometd	459
13	Ppnzmi	443
14	Sozer	431
15	Slmetd	430

Tabela 8: 15 najbolj težavnih oznak (kategorija pravilna)

Analiza kaže, da je denimo pri samostalniku težavna kategorija par imenovalnik-tožilnik moškega spola (oznaki *Sometn*-*Somei*) ali podoben par pri ženskem spolu množine samostalnika (oznake *Sozmt*-*Sozmi*). Predpostavimo lahko, da bo prišlo do težav tam, kjer so oblike v oblikoslovni paradigmi identične, razdalje do okoliških besed, ki bi lahko razdvoumile pravilno oznako, pa tipično daljše od tistih, ki jih upošteva označevalnik.

V tabeli 9 prikazujemo podatke o primerih, kjer se je označevalnik zmotil že pri oblikoskladenjski kategoriji.

Oznaka	Opis	Št. napak
P	pridevnik	2.063
S	samostalnik	1.474
R	prislov	1.189
Z	zaimek	1.028
N	neuvrščeno	762
V	veznik	686
G	glagol	681
L	členek	242
K	številnik	133
D	predlog	128
M	medmet	27
O	okrajšava	4
Skupaj		8.417

Tabela 9: Napake označevanja (kategorija napačna).

Pri tej vrsti napak je najpogostejša zamenjava prislovov s pridevniki in obratno, ki tudi gledano v celoti predstavlja velik izziv za označevalnik (in posledično torej tudi za lematizator). Kot zanimivost v tabeli 10 navajamo še tri najtežje primere glede na posamezno obliko oz. pojavnico.

Oblika	Št.	Oznaka1 / Lema1	Oznaka2 / Lema2
jih	618	Zotzmt--k / *on*	Zotmmt--k / *on*
vse	379	Zc-set / *ves*	Rsn / *vse*
kaj	298	Zv-set / *kaj*	Rsn / *kaj*

Tabela 10: Najtežje oblike oz. pojavnice. Drugi stolpec predstavlja št. napak pri označevanju oblike, tretji (pravilna oznaka) in četrti stolpec (pripisana napačna oznaka) pa podajata najpogostejšo napako pri označevanju oblike.

Prva je oblika *jih* zaimkovne leme *on* v ženskem ali moškem spolu (tožilnika). V tem primeru gre za

razreševanje stavčne ali celo medstavčne koreference, kar je zahtevna naloga za označevalnik. Drugi je zaimek *ves* v srednjem spolu ednine imenovalnika (npr. *vse kaže, da...*) ali prislov z lemo *vse* (npr. *vse večji sum...*). Tretja je oblika *kaj* kot vprašalni zaimek (npr. *kaj je to*) ali prislov (npr. *nič kaj prida*).

Pomembno je torej, da se pri interpretaciji besedil, strojno obdelanih z lematizatorji in označevalniki, zavedamo tistih točk, kjer do napak prihaja pogosteje, in se ne zanašamo zgolj na skupno oceno natančnosti, kot jo kaže desetkratno prečno preverjanje.

5. Zaključek

V prispevku smo opisali novo orodje za označevanje besedil v slovenskem jeziku Obeliks, ki vsebuje module za segmentacijo in tokenizacijo, lematizacijo in oblikoskladenjsko označevanje ter kombinira statistične modele z ekspertnimi pravili za slovenski jezik. Obeliks je dostopen v dveh oblikah: (1) kot programska oprema na portalu SourceForge[8] in (2) kot spletni servis na straneh projekta "Sporazumevanje v slovenskem jeziku".[9] Projekt Obeliks smo zastavili kot dolgoročni projekt, ki bo v prihodnje doživel nadgradnjo, ki bo izhajala iz predstavljene analize napak pri prečnem preverjanju na učnem korpusu. Nadgradnja bo usmerjena predvsem v izboljšanje ekspertnih pravil in povečanje leksikona.

Literatura

Džeroski S., Erjavec T., Zavrel J. (2000). Morphosyntactic Tagging of Slovene: Evaluating PoS Taggers and Tagsets. *Second International Conference on Language Resources and Evaluation, LREC'00*, pp. 1099-1104.

Erjavec, T., Fišer, D., Krek, S., Ledinek, N. (2010). The JOS Linguistically Tagged Corpus of Slovene. V *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC)* (str. 1806-1809). Pariz: ELRA.

Giménez, J., & Márquez, L. (2004). SVMTool: A general POS tagger generator based on Support Vector Machines. *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC'04).* Lisbon, Portugal.

Juršič, M., Mozetič, I., Erjavec, T., & Lavrač, N. (2010). LemmaGen : multilingual lemmatisation with induced Ripple-Down rules. *J. univers. comput. sci.*

Krek, S. (2010). *Pridobivanje jezikovnih podatkov iz besedilnih korpusov za namen izdelave enojezičnih slovarjev in slovnic*: doktorska disertacija. Univ. v Ljubljani, Filozofska fakulteta, Oddelek za slovenistiko.

Mitchell, T. M. (1997). *Machine Learning.* McGraw-Hill.

Nigam, K., Lafferty, J., & McCallum, A. (1999). Using Maximum Entropy for Text Classification. *Proceedings of IJCAI-99 Workshop on Machine Learning for Information Filtering*, (pp. 61-67).

Ratnaparkhi, A. (1996). A Maximum Entropy Model for Part-Of-Speech Tagging. *Proceedings of Empirical Methods in Natural Language Processing*, (pp. 133-142).

Priloga (ekspertna pravila)[10]

w – beseda (kot je zapisana v besedilu)
w_L – beseda w zapisana z malimi črkami
L – seznam dovoljenih oznak
(vhodna spremenljivka in hkrati rezultat aplikacije pravil)
T={ } – množica za filtriranje seznama L (sprva prazna)

(1) če je w ena sama črka
    v T dodaj oznako Some
    če S_{Dm1} vsebuje w_L, v T dodaj oznako Dm
    če S_{Dt1} vsebuje w_L, v T dodaj oznako Dt
    če S_{DoDr1} vsebuje w_L, v T dodaj oznaki Do, Dr
    če S_{Dd1} vsebuje w_L, v T dodaj oznako Dd
    če S_{Krg1} vsebuje w_L, v T dodaj oznako Krg
    če S_{N1} vsebuje w_L, v T dodaj oznako N
    če w_L="a", v T dodaj oznaki Vp, Rsn
  sicer: če je w oblike *številka-končnica*, pri čemer S_{KaKo} vsebuje končnico
    v T dodaj oznake Ka* (“*” pomeni poljubne vrednosti preostalih atributov oznake)
  sicer: če w vsebuje vsaj eno črko in vsaj eno števko ter vsebuje izključno črke in števke
    v T dodaj oznake S*, N, Kag
  sicer: če w vsebuje vsaj eno števko in nobene črke
    če se w konča s piko, v T dodaj oznako Kav
    sicer: v T dodaj oznako Kag
  sicer: v T dodaj vse oznake iz L

(2) če S_{Di} ne vsebuje w_L, potem iz T odstrani oznako Di
(3) če S_{Dr} ne vsebuje w_L, potem iz T odstrani oznako Dr
(4) če S_{Dd} ne vsebuje w_L, potem iz T odstrani oznako Dd
(5) če S_{Dt} ne vsebuje w_L, potem iz T odstrani oznako Dt
(6) če S_{Dm} ne vsebuje w_L, potem iz T odstrani oznako Dm
(7) če S_{Do} ne vsebuje w_L, potem iz T odstrani oznako Do
(8) če S_{Vp} ne vsebuje w_L, potem iz T odstrani oznako Vp
(9) če S_{Vd} ne vsebuje w_L, potem iz T odstrani oznako Vd
(10) če S_L ne vsebuje w_L, potem iz T odstrani oznako L
(11) če S_Z ne vsebuje w_L, potem iz T odstrani oznake Z*
(12) če w ne vsebuje samih črk, potem iz T odstrani oznake M, G*
(13) če w ni kombinacija črk in pomišljajev, potem iz T odstrani oznake P*, R*
(14) če w ni kombinacija črk, števk in pomišljajev, potem iz T odstrani oznake oznake S*
(15) če S_O ne vsebuje w_L in hkrati w ni ena sama črka, ki ji sledi pika, potem iz T odstrani oznako O
(16) če w ni zaporedje rimskih števk, ki mu ali mu ne sledi pika, potem iz T odstrani oznake Kr*
(17) če se w ne začenja z neko predpono iz S_{KbPr}, potem iz T odstrani oznake Kb*
(18) če se w ne začenja s števko, potem iz T odstrani oznake Ka*
(19) če S_{Gp} ne vsebuje w_L, potem iz T odstrani oznake Gp*
(20) nazadnje iz L odstrani vse oznake, ki jih ni v T (pri tem L ohrani vrstni red oznak, ki jih vsebuje)

[8] http://sourceforge.net/projects/obeliks/ (navodila za uporabo se nahajajo na pripadajoči wiki-strani)
[9] http://oznacevalnik.slovenscina.eu/

[10] Seznami besed, ki so del definicije ekspertnih pravil (S_{Dm1}, S_{Dt1}, S_{DoDr1}…), so dostopni na spletni strani https://sourceforge.net/p/obeliks/wiki/SeznamiBesed/