Zbirka primerov rabe vejice Vejica 1.3

Peter Holozan

Amebis, d. o. o., Kamnik
Bakovnik 3, 1241 Kamnik
peter.holozan@amebis.si

Povzetek

Pripravljena je bila nova verzija zbirke primerov rabe vejice Vejica, v kateri so bile popravljene najdene napake iz prejšnje verzije 1.0, dodan pa je bil del z nestandardnimi tviti iz korpusa Janes-Vejica 1.0. Nova verzija je izrazito izboljšala rezultat popravljanja vejic slovničnega pregledovalnika Besana pri delu iz korpusa Lektor. Slovnični pregledovalnik Besana je bil dopolnjen za pregledovanje nestandardnih besedil.

Corpus of comma usage Vejica 1.3

New version of corpus if comma usage Vejica was prepared, correcting the found problems from previous version 1.0. New part wa added containing samples of non-standard twits from corpus Janes-Vejica 1.0. The new version of Vejica improved results for Besana grammar checker comma correcting significantly in the part from corpus Lektor. Besana grammar checker was improved for better handling of non-standard texts.

1. Uvod

Jezikovnotehnološko raziskovanje kot osnovo potrebuje ustrezno označene korpuse oz. primere, kar nam omogoča potem bodisi uporabo strojnega učenja bodisi preizkušanje različnih metod.

Za področje postavljanje vejic je bilo že pripravljenih nekaj zbirk podatkov, vendar imajo obstoječe zbirke nekatere slabosti, zato je smiselno to področje še dopolniti z novo zbirko primerov rabe vejice, ki bo omogočala nadaljnje delo na tem področju.

2. Pregled dosedanjih zbirk primerov

Obstaja že kar nekaj prosto dostopnih korpusov oz. zbirk primerov rabe vejice v slovenščini, ki imajo označene napačno postavljene (torej manjkajoče in odvečne) vejice.

2.1. Korpus Šolar

V korpusu šolskih pisnih izdelkov Šolar so besedila, ki so jih učenci v slovenskih osnovnih in srednjih šolah samostojno tvorili pri pouku. Zajeta so besedila, pri katerih je slovenščina materni jezik avtorjev in ki niso bila napisana posebej za projekt, ampak kot šolska produkcija, jezikovni popravki so pa taki, kot so jih naredili učitelji. (Rozman et al., 2010)

Vključena so besedila od 6. do 9. razreda osnovnih šol, od 1. do 5. letnika srednjih šol in besedila, ki so bila napisana na maturitetnih tečajih. Del korpusa so tudi popravki, ki so jih naredili učitelji.

Prvič je korpus Šolar kot zbirko primerov rabe vejice uporabil Holozan (2012), vendar so bile uporabljene le povedi, ki so imele označene kakšno napako pri vejicah (torej bodisi manjkajočo bodisi odvečno vejico), kar pa ni primerno za ugotavljanje natančnosti. Nekoliko dopolnjena verzija (popravljene se bile nekatere opažene napake pri popravkih vejic) je bila uporabljena še pri preizkusu strojnega učenja za postavljanje vejic (Holozan, 2013).

Korpus je dostopen na naslovu https://www.korpus-solar.net/ in v repozitoriju CLARIN.SI s povezavo http://hdl.handle.net/11356/1036.

2.2. Korpus Lektor

Korpus lektorskih popravkov Lektor je nastal v okviru doktorske naloge (Popič, 2014), vsebuje približno milijon besed. Besedila v korpusu so napisali profesionalni pisci, dobra polovica so prevodi. Baza je v formatu XML in označeni so vsi lektorski popravki.

Korpus je dostopen na naslovu http://www.korpus-lektor.net.

Korpus je bil za analizo vejic uporabljen v Piškur (2015).

2.3. Vejica 1.0

Zbirka primerov rabe vejice Vejica 1.0 je bila pripravljena v okviru doktorske naloge (Holozan, 2016). Sestavljena je iz štirih delov, dva dela imata še poddele (kar je podrobneje opisano v nadaljevanju).

Zbirka primerov vsebuje 113.308 povedi, pri čemer je označenih 17.768 (11,36 %) manjkajočih vejic, 4.608 (3,22 %) vejic pa je označenih za odvečne. (Holozan, 2016)

Zbirka primerov je dostopna v repozitoriju CLARIN.SI s povezavo http://hdl.handle.net/11356/1055 pod licenco CC BY-NC-SA 4.0.

2.3.1. Korpus KUST

Korpus usvajanja slovenščine kot tujega jezika (KUST) je zbirka besedil, ki so jih napisali govorci drugih jezikov, ki se šele učijo slovensko. Korpus je bil predlagan v Stritar (2006) in bil narejen v okviru projekta ESS Uspešno vključevanje otrok, učencev in dijakov migrantov v vzgojo in izobraževanje. Projekt je izvajal Center za slovenščino kot drugi/tuji jezik Filozofske fakultete Univerze v Ljubljani (Rozman et al., 2010). Vključen je bil le manjši del korpusa KUST, in sicer je bilo izbranih 388 povedi, v katerih je bila vsaj ena odvečna ali manjkajoča vejica, pravilne povedi niso vključene (Holozan, 2016).

2.3.2. Korpus Šolar

V poddelu iz korpusa Šolar so bila uporabljena besedila iz korpusa Šolar (bolj podrobno opisanega v točki 2.1)., ki so vsebovala vsaj en učiteljski popravek (ker vsa besedila

niso vsebovala popravkov). Ker pa se je pokazalo, da je veliko napak pri vejicah neoznačenih (za kar je možno več razlogov: če je besedilo vsebovalo preveč napak, je učitelj lahko obupal in nehal popravljati, učitelj je lahko popravil poved drugače in vejica potem ni bila več potrebna, kakšne vejice pa so učitelji najbrž tudi spregledali). Zato so bili vsi ti primeri ročno pregledani in popravljeni, skupaj skoraj 50.000 povedi. (Holozan, 2016)

Primere je popravljala ena oseba (s konzultacijami s strokovnjaki pri problematičnih primerih), kar pomeni, da so vejice čim bolj enotno postavljene, kar lahko pomeni prednost pri računalniški obdelavi, predvsem pri strojnem učenju, težave zaradi neenotne uporabe vejic v uporabljenem korpusu so npr. imeli pri strojnem učenju postavljanja vejic v baskovščini (Alegría et al., 2006).

Primeri iz tega dela imajo označen tudi poddel, ki pove, v katerem razredu oziroma letniku je bil napisan posamezen primer.

Ta poddel je uporabljen v Krajnc (2015) in Krajnc in Robnik-Šikonja (2015).

2.3.3. Korpus Lektor

Primeri iz korpusa Lektor (podrobneje opisanega v točki 2.2) so bili pretvorjeni tako, da so bili označeni le lektorski popravki pri vejicah, vsi drugi lektorski popravki pa so bili izpuščeni. Izpuščene so bile tudi povedi, ki so v celoti napisane v tujih jezikih, vendar to ni bilo narejeno čisto dosledno.

Celoten del je ena enota, posamična besedila iz korpusa Lektor niso bila razporejena kot morebitni poddeli.

2.3.4. Wikipedija

Zadnji del predstavljajo članki iz Wikipedije, pri čemer so bili izbrani članki, ki niso preveč lektorirani, za kar je bila uporabljena kategorija »Članki, ki so potrebni čiščenja«. Vse manjkajoče in odvečne vejice so bile ročno označene. Uporabljenih je bilo 9 člankov, ki skupaj vsebujejo 870 povedi. (Holozan, 2016)

Namen tega dela je bil dobiti domeno, ki bi bila med šolarji iz korpusa Šolar in profesionalnimi pisci iz korpusa Lektor.

2.4. Janes–Vejica 1.0

Ta zbirka primerov je bila narejena iz 500 tvitov (čivkov) v nestandardni slovenščini (izbrani so bili le tviti, ki jim je bila pripisana jezikovna nestandardnost), zbranih v okviru projekta Janes. V njih so bili ročno označeni vsi primeri nestandardne rabe vejic, vsa ta mesta in tudi vse obstoječe vejice pa so bile označene z razlogom za vejico. Primere sta označevala dva označevalca, na koncu pa je v primeru neskladja kuratorka določila končno oznako. (Popič et al., 2016).

Zaradi označenih razlogov za vejice je ta zbirka primerov zelo zanimiva za nadaljnje analize, žal pa ni velika. Zbirka primerov je dostopna v repozitoriju CLARIN.SI pod licenco CC BY-SA 4.0 na naslovu http://hdl.handle.net/11356/1088.

3. Dopolnjevanje zbirke rabe vejice

Po izdelavi je bila zbirka primerov Vejica 1.0 uporabljena za preizkušanje slovničnega pregledovalnika Besana. Pri tem se je pokazalo, da je v Vejica 1.0 kar nekaj težav, ki jih je bilo treba rešiti. Vmes sta bili narejeni zbirki Vejica 1.1 in Vejica 1.2, ki sicer nista bili objavljeni, je pa Vejica 1.2 bila omenjena pri zagovoru doktorata Holozan (2016), po tem pa je bilo narejeno še kar nekaj novih popravkov, zato sem se odločil, da bo nova izdana zbirka označena kot Vejica 1.3.

3.1. Izločanje tujih povedi

Kot prva težava se je pokazalo, da je v delu iz korpusa Lektor še nekaj deset povedi, ki so v celoti v tujem jeziku.

Harlem est né a Lausanne.
Omar est né a Evian.
Les gens nés a Lausanne sont généralement sujets helvétiques.

Slika 1: Primeri izločenih tujih povedi.

V prvih dveh primerih je analizator opozarjal na vejico pred »a«, ker ni zaznal, da gre za tuji jezik, ker so besede Harlem, ne, a, Omar in Evian tudi v slovenskem slovarju. Tu bi se sicer morda dalo še kaj nadgraditi s tem, da bi lastna imena izvzeli iz ugotavljanja jezika, vendar je pri kratkih povedih ugotavljanje jezika vseeno lahko problematično, še najbolj zanesljiva rešitev bi bila uporaba slovarjev in analizatorjev za vse te druge jezike, kar pa je v praksi sitno tako zaradi počasnosti kot potrebnega prostora za tuje slovarje (in npr. v Amebisu bi lahko tako dobro zaznavali le angleščino in nemščino, delno tudi francoščino in albanščino, za hrvaščino imamo le zelo majhen testni slovar, drugih jezikov pa niti nismo pokrivali).

V realni uporabi je pri povedih v tujem jeziku tudi sicer smiselno, da se pravilno nastavi jezik v urejevalniku besedil (že zaradi črkovalnika), kar pomeni, da slovenski slovnični pregledovalnik potem ne bo pregledoval teh povedi.

Ostali so še primeri, v katerih je v tujem jeziku le del, največkrat naslovi del ali pa citirani primeri. Ti primeri sicer delajo težave analizatorju, vendar so to čisto realna uporaba, ki jo bo moral slovenski slovnični pregledovalnik nekoč rešiti.

3.2. Pravilno upoštevanje lektorskih popravkov

Pregled problematičnih primerov je pokazal, da se je pri uvozu iz korpusa Lektor v nekaterih primerih (predvsem pri gnezdenih napakah, torej ko sta bili napaki ena v drugi) zgodilo, da so bili uvoženi tudi drugi lektorski popravki, in sicer tako, da je bilo napisano tako originalno besedilo kot popravek.

Kaj vi mislite menite in zakaj?
Razdelite izročite vsakemu po en flomaster/barvno pisalo.
kupujmo Kupujemo izdelke, narejene iz lokalnega lesa s FSC oznako

Slika 2: Nekaj primerov težav zaradi napačnega upoštevanja popravkov.

Še posebej v primerih, ko je bil popravljen glagol, je to analizatorju naredilo težave, ker je zaradi dvojnega glagola domneval, da nekje manjka vejica, kar je potem poslabševalo natančnost iskanja manjkajočih vejic.

3.3. Popravljene vejice

Pokazalo se je, da nekateri primeri manjkajočih oz. odvečnih vejic niso bili označeni v zbirki primerov. Največ takih primerov je bilo v delu iz korpusa Lektor, potem v delu iz Wikipedije, najmanj (po deležu) napak je bilo v delu

iz korpusa Šolar, ki je bil ob izdelavi zbirke primerov Vejica 1.0 najbolj sistematično ročno preverjen (Holozan, 2016).

Izračuna se kot povprečje indeksiranih vrednosti BDP-ja na prebivalca, pričakovane življenske dobe§, in povprečnega pričakovanega obdobja obiskovanja izobraževalnih ustanov v državi. Poleg zmanjšanja potrebne količine surovine za dano uporabo§, izboljšana učinkovitost zmanjšuje relativno ceno uporabe določene surovine, kar poveča povpraševanje po surovini, to pa lahko izniči prihranke, ki jih je prinesla povečana učinkovitost. Določite§ kateri korporaciji pripadajo naslednje znamke: Da bi si lažje zapomnili§ pišite na tablo: Različna gledišča: Kdo?

Slika 3: Primeri popravljenih vejic, označeni z znakom §.

Sistematično so bili za Vejico 1.3 preverjeni le primeri, pri katerih se z oznakami v zbirki primerov niso ujemali rezultati slovničnega pregledovalnika Besana. Zato se ob dopolnitvah Besane redno najdejo še novi taki primeri napak v zbirki primerov, kar kaže, da bi bilo za še natančnejše popravke treba še enkrat ročno preveriti celotno zbirko primerov. Po drugi strani pa glede na dosedanje število odkritih napak in delež napak, ki jih že popravi Besana, ocenjujem, da število še neodkritih najbrž ne presega enega odstotka sedanjega števila najdenih napak.

3.4. Izboljšanje zaradi popravkov zbirke primerov

Popravljanje zbirke primerov je zelo izboljšalo rezultate popravljanja vejic, dosežene s programom Besana, in sicer predvsem v delu iz korpusa Lektor, opazna izboljšava pa je bila tudi v delih iz korpusa KUST in Wikipedije, kjer so bile predvsem napake pri vejicah, ker sta bila ta dva dela označena ročno in nista bila potem prej še enkrat preverjena.

Za del iz korpusa Lektor se je priklic pri iskanju manjkajočih vejic izboljšal iz 40,5 % na 50,5 %, natančnost pa kar iz 18,8 % na 35,6 %. Pri iskanju odvečnih vejic se je pri delu iz korpusa Lektor priklic popravil iz 26,8 % na 40,5 %, natančnost pa iz 29,7 % na 58,4 %. Izboljšanje je torej res veliko. To izboljšanje kaže, da bo morda smiselno ponoviti poskus s strojnim učenjem postavljanja vejic, pri katerem je Holozan (2016) pokazal, da deluje veliko slabše na delu iz korpusa Lektor kot na delu iz korpusa Šolar. Po drugi strani pa je tam šlo za postavljanje vseh vejic, pri čemer vpliv popravkov najbrž ne bo tako velik, kot je za popravljanje vejic, kajti v delu iz korpusa Lektor je delež napak pri postavljanju vejic majhen – 1,2 % vejic manjka, odvečnih je pa 0,8 % (Holozan, 2016).

3.5. Priključitev zbirke Janes–Vejica 1.0

Odločil sem se, da v zbirko primerov vključim še primere iz korpusa Janes–Vejica 1.0, s čimer bo zbirka dopolnjena še s spletno slovenščino.

Raziskava Popič in Fišer (2018), ki je bila izvedena na tem korpusu, je pokazala, da so v večini primerov vejice postavljanje v skladu s standardom, nestandardne vejice pa so predvsem manjkajoče, odvečnih je malo.

Primeri so bili pretvorjeni tako, da so označeni enako kot drugi deli zbirke primerov (torej z oznakami za manjkajoče in odvečne vejice), odstranjeni so bili sicer zanimivi podatki o tipih vejic. Poenotenje teh primerov v format preostale zbirke primerov Vejica je koristno zato, da lahko nekdo na enak način uporabi zelo raznolike vire pri preizkušanju popravljanja vejic oz postavljanja vseh vejic besedilo.

Vsega skupaj je v tem delu 1369 povedi.

4. Sestava nove zbirke rabe vejice

Vejica 1.3 je sestavljena iz petih delov, KUST in Šolar pa sta še naprej razdeljena na poddele. Število primerov je prikazano v spodnji tabeli.

Vejica 1,3	oznaka	povedi	vejic	manjkajočih	odvečnih	delež manjkajočih	delež odvečnih
KUST		**388**	**221**	**432**	**101**	**66,16 %**	**31,37 %**
KUST de	KUST.de.	8	2	11	0	84,62 %	0,00 %
KUST en	KUST.en.	98	32	71	50	68,93 %	60,98 %
KUST es	KUST.es.	110	100	135	32	57,45 %	24,24 %
KUST it	KUST.it.	61	49	75	8	60,48 %	14,04 %
KUST sh	KUST.sh.	111	38	140	11	78,65 %	22,45 %
Šolar		**49438**	**49125**	**16341**	**3870**	**24,96 %**	**7,30 %**
Šolar OŠ 6	Solar.OS6.	678	432	239	29	35,62 %	6,29 %
Šolar OŠ 7	Solar.OS7.	2457	1288	725	117	36,02 %	8,33 %
Šolar OŠ 8	Solar.OS8.	4003	2503	1434	221	36,42 %	8,11 %
Šolar OŠ 9	Solar.OS9.	3398	2532	700	173	21,66 %	6,40 %
Šolar PŠ 1	Solar.PS1.	1137	1164	601	73	34,05 %	5,90 %
Šolar PŠ 2	Solar.PS2.	619	625	324	51	34,14 %	7,54 %
Šolar PŠ 3	Solar.PS3.	966	819	520	81	38,83 %	9,00 %
Šolar PŠ 5	Solar.PS5.	472	435	292	42	40,17 %	8,81 %
Šolar SŠ 1	Solar.SS1.	4431	3570	1880	348	34,50 %	8,88 %
Šolar SŠ 2	Solar.SS2.	3533	3812	1399	308	26,85 %	7,48 %
Šolar SŠ 3	Solar.SS3.	3703	3209	1508	277	31,97 %	7,95 %

Šolar SŠ 4	Solar.SS4.	2940	3174	1234	246	27,99 %	7,19 %
Šolar G 1	Solar.G1.	7240	8475	2150	775	20,24 %	8,38 %
Šolar G 2	Solar.G2.	3134	3918	816	221	17,24 %	5,34 %
Šolar G 3	Solar.G3.	3613	4391	841	304	16,07 %	6,47 %
Šolar G 4	Solar.G4.	6892	8593	1565	590	15,41 %	6,42 %
Šolar MT	Solar.MT.	222	185	113	14	37,92 %	7,04 %
Lektor	Lektor.Lektor.	**52121**	**71204**	**1133**	**717**	**1,57 %**	**1,00 %**
Wikipedija	Wiki.Wiki.	**869**	**929**	**124**	**71**	**11,78 %**	**7,10 %**
Janes	Janes.Janes.	**1368**	**646**	**387**	**20**	**37,46 %**	**3,00 %**
skupaj		104184	122125	18417	4779	13,10 %	3,77 %

Tabela 1: Sestava zbirke primerov Vejica 1.3.

Tabela 1 prikazuje zgradbo zbirke primerov Vejica 1.3 skupaj z dodatno razdelitvijo delov KUST in Šolar.

Enako kot v Holozan (2016) je delež manjkajočih vejic izračunan tako, da delimo število manjkajočih vejic z vsoto napisanih vejic in števila manjkajočih vejic (niso upoštevanje odvečne vejice), delež odvečnih vejic pa je izračunan kot kvocient števila odvečnih vejic z vsoto števila napisanih vejic in števila odvečnih vejic (niso upoštevanje manjkajoče vejice).

4.1. Format nove zbirke rabe vejice

Zbirka je zgrajena enako, kot je bil zgrajena zbirka primerov rabe vejice Vejica 1.0. Vsaka poved je v svoji vrstici, v vrstici je najprej oznaka povedi, ki je sestavljena iz oznake dela, oznake poddela (uporabljene oznake so naštete v Tabeli 1) in zaporedne številke v poddelu, deli oznake so med sabo ločeni s pikami. Potem sledi poved, ločena s tabulatorjem, mesta, na katerih vejice manjkajo, so označene z znakom »¤«, odvečne vejice pa so nadomeščene z znakom »÷«.

KUST.de.4 Ko sva prišli do skupine¤ so vsi vprašali¤ kaj smo kupili in midve tudi.
KUST.de.5 Danes¤ če gremo ven¤ bom oblekla mojo novo obleko, se že veselim.
Wiki.Wiki.487 Pogosto se v spastične mišice spodnjih okončin dajejo injekcije botulina÷ z namenom¤ da se zmanjša spastično povečan mišični tonus, ki je lahko zelo boleč.
Solar.G1.643 Na leto imamo približno 120 snežnih dni÷ ter 220 kondicijskih enot.
Solar.PS2.170 Odlomek govori o tem¤ kaku je David Goldstein tekel pred smogom¤ voznik avta¤ ki÷ je šel mimo¤ pa mu ni hotel ustaviti.
Lektor.Lektor.266 V drugih državah pa delovanje sindikatov režim zatira÷ ali pa dovoljuje zasebnim podjetjem¤ da ga zatirajo, tudi če v ta namen uporabljajo silo.
Janes.Janes.1182 je dost¤ da je topla, za čez šmorn glih kul.

Slika 4: Nekaj primerov iz zbirke.

Datoteka je zapisana v formatu UTF-8.

4.2. Dostopnost nove zbirke rabe vejice

Zbirka primerov rabe vejice Vejica 1.3 je objavljena v repozitoriju CLARIN.SI pod imenom Vejica 1.3 pod licenco CC BY-NC-SA 4.0 na naslovu http://hdl.handle.net/11356/1185.

5. Izboljševanje popravljanja vejic

Ker so podatki v Holozan (2016) že zastareli, je hkrati z novo zbirko primerov rabe vejice smiselno objaviti še kratko poročilo delu pri računalniškem popravljanju vejic s programom Besana in najnovejše rezultate, ki lahko služijo kot referenčne vrednosti za nadaljnje raziskave na tem področju.

5.1. Izboljšave slovničnega pregledovalnika Besana

Rezultati, objavljeni v Holozan (2016), so bili izboljšani že do zagovora konec leta 2016. Tako se je za problem iskanja manjkajočih vejic skupen priklic popravil iz 57,5 % na 63,5 %, natančnost pa iz 74,9 % na 79,8 %. Do tega trenutka se je Besana še popravila na priklic 75,3 % in natančnost 80,6 %. Vsi ti podatki so na zbirki primerov Vejica 1.0, kot pa kažejo ugotovitve v točki 3.4, bodo rezultati v novi verziji primerov še boljši.

Po eni strani je za izboljšanje Besane pomagalo povečanje slovarja (nove verzije prepoznajo veliko več lastnih imen), še bolj pa dopolnitve stavčnega analizatorja (oz. analizatorja povedi). Ena od pomembnejših dopolnitev za področje vejic je bila obravnava veznika »kot« v različnih skladenjskih vlogah. Vse izboljšave analizatorja so na primer označevanje jos100k izboljšale do te mere, da so zdaj leme pravilno označene v 99,31 %, oblikoskladenjske oznake pa v 96,87 %.

Izboljšano je bilo tudi opozarjanje na postavljanje vejic v primerih, pri kateri sta dve možnosti, prava pa je odvisna od pomena. Tak primer sta npr. »tako da« in »zato ker«, pri katerih je vejica lahko spredaj ali pa vmes, odvisno od poudarka. Prej je Besana vedno postavila vejico na začetku, kar je v tistih primerih, v katerih bi morala vejica biti vmes, pomenilo, da najprej ni ugotovila manjkajoče vejice, dodatno pa je še postavila vejico tja, kjer je ne bi smelo biti. Zato v teh primerih zdaj Besana na začetku opozori, da nekje v stavku manjka vejica, in prepusti uporabniku, da sam izbere, kje bi vejica morala biti.

5.2. Izboljšanje pri nestandardni slovenščini

Na prvi pogled se morda zdi, da dopolnjevanje stavčnega analizatorja za Besano z zelo nestandardno slovenščino ni preveč smiselno, saj takih besedil ne bo nihče popravljal z Besano. Vendar je kar nekaj razlogov, zaradi katerih je to smiselno.

Take nestandardne oblike se ljudem prikradejo tudi, kadar želijo pisati v knjižni slovenščini (ker marsikdo bere

predvsem spletno slovenščino in vedno manj besedil v knjižni slovenščini), zato je smiselno, da jih Besana popravi.

Še pomembnejši razlog pa je, da je stavčni analizator uporabljen tudi drugje. Ena taka uporaba je na primer označevalnik korpusov in v korpusu Gigafida je kar precej nestandardnih besedil, ki so v tem trenutku nenatančno označena, ker je označevalnik naučen predvsem na besedila v knjižni slovenščini (je pa bil pri tem označevanju dosežen velik napredek v okviru projekta Janes, in sicer z vključitvijo nestandardnih učnih podatkov in z vključitvijo Brownovih gruč, pridobljenih iz velikih zbirk surovih nestandardnih podatkov (Ljubešić et al., 2018)). Tak primer je npr. beseda »jas« kot nestandardna oblika osebnega zaimka za prvo osebo ednine. Beseda »jas« ima v Gigafidi 1947 pojavitev, med prvimi 20 zadetki je 16 takih, kjer je mišljen osebni zaimek, vendar so vse označene kot oblika samostalnika »jasa«.

Druga uporaba je v sistemu za virtualne asistente SecondEgo, ki tako bolje prepoznavajo nestandardno slovenščino (pregled uporabe agentov v SecondEgu kaže, da je predvsem pogosta neuporaba strešic pri čšž in velikih začetnic).

Tretja pa je strojni prevajalnik Presis, ki pa je že lahko čisto realna uporaba tudi pri tvitih, še posebej zato, ker ima Google Prevajalnik precej težav z nestandardno slovenščino, kar je logično, ker paralelni korpusi, iz katerih se uči, vsebujejo večinoma le standardno slovenščino.

5.2.1. Nestandardno besedišče

Kako pogosto je nestandardno besedišče že v korpusu Gigafida, kaže npr. primer zapisa »nč« namesto »nič«, za kar je v Gigafidi kar 6410 konkordanc. V nekaj malega primerih gre sicer tudi za kratico »NČ« v pomenu »nedoločen čas«, še vedno pa je praktično 6000 stavkov, v katerih o analizator bolje deloval, ker bo to besedo pravilno prepoznal.

Primer neoznačene manjkajoče vejice v korpusu Šolar, ki ga je našla Besana po tem, ko je bila dodana beseda »kokr« kot nestandardni zapis besede »kakor«, je naslednji:

Pozno zvečer so uzeli najnujnejše stvari¤ pobrali§ kokr so hitro lahko¤ saj se jim je počas že istekal na barko.

Slika 5: Naslov slike.

Besana je dodatno opozorila, da manjka vejica na mestu z oznako §.

Primer nestandardno zapisane besede, ki dela težave pri postavljanju vejic, če je analizator ne prepozna, je npr. »poj« v pomenu »potem«, ki se prekriva z velelnikom glagola »peti«, zaradi česar je Besana pri »Poj bom pa vsakmu pokazal napis na seb, pa naj se mal zamisljo.« prej pred »bom« pričakovala vejico.

Primer nestandardne oblike, ki pa ni bila dodana, ker bi verjetno povzročila več težav, kot pa bi jih rešila, pa je »ja« kot zapis osebnega zaimka »jaz«, ker se prekriva z zelo pogostim medmetom (oz. členkom) »ja«, ki je zelo pogost ravno v nestandardnih besedilih. Če bi bil dodan kot osebni zaimek, Besana v veliko primerih ne bi dodala manjkajoče vejice za »ja« na začetku povedi, ker bi menila, da gre za osebni zaimek, ki je del nadaljevanja. Po drugi strani pa se je kot koristno izkazalo to, da je bil dodan nestandardni osebni zaimek »jas«, ki se prekriva z obliko samostalnika »jasa«.

5.2.2. Tipične redukcije pri pisanju

V nestandardni slovenščini obstajajo nekatere tipične redukcije pri pisanju, ki jih je smiselno upoštevati, kadar besede ni v slovarju.

mogoče bo kdo clo zastopu kdaj kk je brez šihta haha zlobirite interrail pa pridte :) Men prej pr stran grejo, k pa zadi.

Slika 6: Primeri redukcij pri pisanju.

V analizator je bilo dodano pravilo, da »u« na koncu besede tipično nadomešča »il«, »al« ali »el«, če gre za deležnik na -l. Podobno je pri »i« treba preveriti, ali je mogoče namesto »aj«.

Razlika med temi tipičnimi redukcijami in nestandardnim besediščem je v tem, da je nestandardno besedišče dodano v slovar, redukcije pa se upoštevajo sistemsko (je pa treba v slovar dodati oblike, ki se prekrivajo z drugimi standardnimi oblikami).

5.2.3. Neuporaba čšž

V tvitih (in tudi nasploh pri nestandardnem pisanju) je pogosto, da so izpuščene strešice na čšž in so torej namesto teh uporabljene črke csz. Amebisov analizator ima nastavitev, ki pri vseh csz upošteva tudi možnost, da gre za čšž. Težava pa je, ker pri tem lahko nastanejo tudi dvoumnosti, ki lahko otežijo analizo povedi in postavljanje vejic.

@SanjaModric kaj to pomeni za nase ce zvecer zmagajo ...

Slika 7: Primer tvita brez strešic.

Pri tipičnih veznikih (npr. »če«) je bilo treba dodati tudi obliko »ce« pri preverjanju mest, na katerih pogosto manjka vejica. Dodatno se je pri »ce« zapletlo še v kombinaciji z neuporabo velikih začetnic, ker se je pokazala možnost, da je »CE« še oznaka kraja Celje, kar je dodatno zapletlo delo analizatorja povedi.

5.2.4. Neuporaba velikih začetnic

Podobno kot strešice so v nestandardnih besedilih pogosto izpuščene tudi velike začetnice. Zato ima Amebisov analizator tudi nastavitev, da se ne zanaša, da so pravilno uporabljene velike začetnice (ta nastavitev je na primer vključena v virtualnih asistentih v sistemu SecondEgo). Seveda pa ta nastavitev zaplete razdvoumljanje.

ja želim ti čim več dobrih ocen..

Slika 8: Primer neuporabe velikih začetnic.

Besana običajno pri vejicah za medmeti na začetki povedi preveri, da mora biti medmet napisan z veliko začetnico, ta omejitev je bila ob nastavitvi, da začetnice niso pomembne, odstranjena, tako da Besana postavi vejico za »ja« v primeru zgoraj.

5.2.5. Vejice pri začetnih medmetih

Na tvite lahko gledamo kot na premi govor. Ena od značilnosti je veliko število uporabljenih medmetov, ki morajo na začetku povedi biti z vejico ločeni od nadaljevanja.

Bravo Risi! Mah ne grem se vec!! Aja hvala za sveče, grem dans na britof zastojn ajde falil so do 13h ampak pol se je pa zacel #FB http://t.co/pnGkF3N46T

Slika 9: Primeri, pri katerih manjka vejica za začetnim medmetom.

Po eni strani je bilo treba dodati nekatere medmete v slovar (npr. »mah«), po drugi pa je bila prej tudi omejitev, da je pogoj za vejico velika začetnica.

5.2.6. Začetni pozivi pri tvitih

Tviti se zelo pogosto začnejo s seznamom naslovnikov.

@leaathenatabako Evo poznavalke. @jureflux A si vedu, da je to na Islandiji? @JJansaSDS @vladaRS dej ze enkrat tih bod no, kaj si pa ti naredu?

Slika 10: Primeri z začetnimi pozivi.

Po eni strani bi se na te primere dalo gledati, kot da gre za začetne zvalnike in bi jim potem sledila vejica. Vendar so ti pozivi običajno dodani samodejno in zato pisci tvitov nanje ne gledajo kot del besedila (kar kažejo tudi velike začetnice za njimi). Analizator je bil dopolnjen tako, da v primeru, da je nastavljen na obravnavo tvitov, preskoči vsa imena na začetku.

5.3. Trenutni rezultati za slovnični pregledovalnik Besana

Besana trenutno pri popravljanju vejic v zbirki primerov rabe vejice Vejica 1.3 doseže rezultate, prikazane v tabeli 2.

del	priklic	natančnost
KUST	87,04 %	94,35 %
Šolar	77,62 %	90,76 %
Lektor	50,45 %	38,06 %
Wikipedija	79,84 %	86,73 %
Janes	49,48 %	79,26 %
SKUPAJ	75,59 %	85,78 %

Tabela 2: Popravljanje manjkajočih vejic.

Pri popravljanju manjkajočih vejic so rezultati zelo raznoliki, najboljši priklic je pri delih KUST in Wikipedija, najslabši pa pri delih Janes in Lektor. Slabši rezultat pri Janesu kaže na to, da bo treba še dopolniti analizo nestandardne slovenščine, po drugi strani pa je slabši rezultat v Lektorju posledica tega, da tam v precejšnji meri manjkajo vejice, ki jih je težko avtomatsko postaviti (torej zanje ne zadošča, da pogledamo le besede v neposredni okolici), kar se npr. vidi iz tega, da je beseda, pred katero v Lektorju manjka največ vejic, »in«, na katerega odpade 18 % vseh manjkajočih vejic (Holozan, 2015).

Pri natančnosti je najboljši rezultat pri delih KUST in Šolar, najslabši pa pri delih Lektor in Janes. Pri delu Lektor je natančnost izrazito slabša (več kot pol slabša od vseh drugih delov), razlog za to najbrž to, da je delež manjkajočih vejic v tem delu (1,57 %) veliko manjši kot v drugih delih (na drugem mesti je del Wikipedija, ki vsebuje 11,78 % manjkajočih vejic, torej več kot sedemkrat več, pri drugih delih pa je razlika še večja).

del	priklic	natančnost
KUST	27,72 %	96,55 %
Šolar	37,42 %	93,72 %
Lektor	40,45 %	58,47 %
Wikipedija	45,07 %	96,97 %
Janes	30,00 %	40,00 %
SKUPAJ	37,75 %	85,17 %

Tabela 3: Popravljanje odvečnih vejic.

Priklic pri popravljanju odvečnih vejic je slabši kot pri popravljanju manjkajočih, natančnost pa je primerljiva.

Zanimivo je, da je najslabši priklic pri delu KUST, pri katerem je pri manjkajočih vejicah priklic najboljši, drugi najslabši pa je pri Janesu, vendar je pri Janesu le malo primerov odvečnih vejic in tudi sicer delež odvečnih vejic ni velik. Delež odvečnih vejic pri delu KUST je tako več kot desetkrat večji od Janesa in najbrž so vse te odvečne vejice preveč nepredvidljive za Besano (bi se pa bilo temu smiselno bolj posvetiti, če bi želeli narediti verzijo Besane, ki bi bila bolj uporabna za osebe, ki se učijo slovenščino kot drugi jezik).

Natančnost je spet najslabša pri Lektorju, tudi tukaj pa velja, da je delež manjkajočih vejic v Lektorju daleč najmanjši od vseh delov (1 %).

6. Možnosti za nadaljnje delo

Ena smer je še nadaljnje preverjanje pravilnosti označenih napak pri vejicah, vendar takih napak najbrž ni več veliko.

Bolj zanimivo bi bilo razširiti zbirko primerov še z dodatnimi deli. Ena možnost bi bila uporaba SSJ500k, kot je predlagal že Holozan (2016). Ta del sicer najbrž ni tako zanimiv za napake pri vejicah, ker teh verjetno ni tako veliko, je pa zanimiv, ker je ročno označen (približno polovica je tudi skladenjsko razčlenjena), kar omogoča, da se preizkusijo različne metode postavljanja vseh vejic in se ugotovi, kakšen rezultat bi bilo mogoče doseči, če bi imeli res dobro oblikoskladenjsko označevanje in stavčno razčlenjevanje. Bi pa bilo za ta del treba ročno pregledati vejice, kar zahteva precejšen časoven vložek (po drugi strani pa je res, da če bi nas zanimalo postavljanje vseh vejic, in ne popravljanje napačnih vejic, bi lahko uporabili SSJ500k tudi brez ročnega označevanja napačnih vejic, ker delež napačnih vejic najbrž ni prevelik, če pa bi dodatno za iskanje napačnih vejic uporabili kar Besano, bi lahko polovico napak odkrili z dovolj malo dela, kar bi za postavljanje vseh vejic bilo že čisto uporabno).

Glede na to, da imamo že vključena besedila šolarjev, dijakov in profesionalnih piscev, bi bilo zanimivo izdelati še zbirko primerov študentskih besedil. Morda bi se dalo za pripravi take zbirke primerov izkoristiti dejstvo, da morajo na veliko fakultetah biti diplomska in magistrska dela obvezno lektorirana. Ta besedila so zbrana v Korpusu akademske slovenščine (KAS) (Erjavec at al., 2016). Zanimiva bi bila primerjava napak pri postavljanju vejic med besedili s fakultet, ki zahtevajo lekturo, v primerjavi s

tistimi, ki lekture ne zahtevajo. Bi pa bilo treba vejice spet ročno preveriti, kar zahteva kar nekaj dela.

Glede izboljševanja slovničnega pregledovalnika Besana pri popravljanju vejic je še veliko rezerve pri popravljanju nestandardnih besedil, po drugi strani pa to ni ravno tipična uporaba. Zato bo bolj smiselno izboljšati rezultat pri delu iz korpusa Lektor, še posebej natančnost (Besana ima kar nekaj težav pri strokovnih besedilih, in sicer analizator še ne zna pravilno upoštevati citiranj, težave pa mu delajo tudi seznami literature, pri katerih pa je težava, da obstajajo različni formati in da vsebujejo veliko tujih besed).

7. Sklep

Izboljšana verzija zbirka primerov rabe vejice Vejica 1.3 omogoča še boljše preizkušanje programov za popravljanje vejic. Dodatek dela s tviti v spletni slovenščini omogoča še preizkušanje tovrstne nestandardne slovenščine.

Analizator povedi in s tem Besana sta bila dopolnjena, da bolj učinkovito obdelujeta nestandardna besedila, rezultati pa kažejo, da se to da še izboljšati, saj so ti rezultati v delu Janes slabši kot pri drugih delih.

Rezultati popravljanja vejic so sicer občutno boljši od tistih, ki so bili objavljeni v Holozan (2016), po eni strani zaradi izboljševanja Besane, po drugi strani pa zaradi čiščenja primerov. V delu Šolar takoj Besana najde 77,62 % manjkajočih vejic (pri natančnosti 90,76 %), medtem ko je leta 2016 našla 64,95 % manjkajočih vejic (z natančnostjo 89,48 %), še večja pa je razlika pri delu Lektor (priklic na 50,45 % iz 33,84 %, natančnost pa na 38,06 % iz 18,19 %).

Nasploh rezultati kažejo, kako pomembna je izbira primeru pri preizkušanju in je zato za primerjavo nujno, da se uporabijo isti primeri, kar je omogočeno z licenco CC in objavo zbirke primerov rabe vejice v repozitoriju CLARIN.SI.

8. Literatura

Iñaki Alegria, Bertol Arrieta, Arantza Diaz de Ilarraza, Eli Izagirre, Montse Maritxalar. 2006. Using Machine Learning Techniques to Build a Comma Checker for Basque. V *Proceedings of the COLING/ACL 2006 Main Conference Poster Sessions*, str. 1–8. Association for Computer Linguistics. 1–8.

Tomaž Erjavec, Darja Fišer, Nikola Ljubešić, Nataša Logar, Milan Ojsteršek. 2016. Slovenska akademska besedila: prototipni korpus in načrt analiz. V: *Zbornik konference Jezikovne tehnologije in digitalna humanistika 2016*, str. 58–64. Znanstvena založba Filozofske fakultete v Ljubljani.

Peter Holozan. 2012. Kako dobro programi popravljajo vejice v slovenščini. V: Jezikovne tehnologije: *Zbornik C 15. mednarodne multikonference Informacijska družba IS 2012, 8. do 12. oktober 2012*, str. 101–106. Institut Jožef Stefan. 101–106.

Peter Holozan. 2013. Uporaba strojnega učenja za postavljanje vejic v slovenščini. *Uporabna informatika*, XXI/3: 196–209.

Peter Holozan. 2015. Možnosti uporabe jezikovnih tehnologij za določanje težav pri rabi vejice. V: Helena Dobrovoljc in Tina Lengar Verovnik, ur., *Pravopisna razpotja*, str. 77–92. Založba ZRC, Ljubljana.

Peter Holozan. 2016. *Računalniško postavljanje vejic v slovenščini*. Doktorsko delo, Filozofska fakulteta, Univerza v Ljubljani.

Anja Krajnc. 2015. *Postavljanje vejic v slovenščini s pomočjo strojnega učenja.* Diplomsko delo, Fakulteta za računalništvo in informatiko, Univerza v Ljubljani.

Anja Krajnc, Marko Robnik-Šikonja. 2015. Postavljanje vejic v slovenščini s pomočjo strojnega učenja in izboljšanega korpusa Šolar. V *Zbornik konference Slovenščina na spletu in v novih medijih*, str. 38–43. Znanstvena založba Filozofske fakultete v Ljubljani.

Nikola Ljubešić, Tomaž Erjavec, Darja Fišer. 2018. Orodja za procesiranje nestandardne slovenščine. V: Darja Fišer, ur., *Viri, orodja in metoda za analizo spletne slovenščine*, str. 74–99. Znanstvena založba Filozofske fakultete Univerze v Ljubljani, Ljubljana.

Karin Piškur. 2015. *Postavljanje vejic v slovenskih besedilih z orodjem LanguageTool.* Diplomsko delo, Fakulteta za računalništvo in informatiko, Univerza v Ljubljani.

Damjan Popič. 2014. *Korpusnojezikovna analiza vplivov na slovenska prevodna besedila.* Doktorsko delo, Filozofska fakulteta, Univerza v Ljubljani.

Damjan Popič, Darja Fišer, Katja Zupan, Polona Logar. 2016. Raba vejice v uporabniških spletnih vsebinah. V: *Zbornik konference Jezikovne tehnologije in digitalna humanistika 2016*, str. 149–153. Znanstvena založba Filozofske fakultete v Ljubljani.

Damjan Popič in Darja Fišer. 2018. (Ne)normativnost računalniško posredovane komunikacije v slovenščini: merilo vejice. V: Darja Fišer, ur., *Viri, orodja in metoda za analizo spletne slovenščine*, str. 140–159. Znanstvena založba Filozofske fakultete Univerze v Ljubljani, Ljubljana.

Tadeja Rozman, Mojca Stritar, Irena Krapš Vodopivec, Iztok Kosem, Simon Krek. 2010. *Nova didaktika poučevanja slovenskega jezika : sporazumevanje v slovenskem jeziku*. Ministrstvo za šolstvo in šport: Amebis. http://www.slovenscina.eu/Media/Kazalniki/Kazalnik15/Nova_didaktika_Sporazumevanje.pdf.

Mojca Stritar. 2006. Oblikovanje korpusa usvajanja slovenščine kot tujega jezika. V: *Zbornik 5. slovenske in 1. mednarodne konference Jezikovne tehnologije*. Institut "Jožef Stefan".

Peter Holozan. 2016. *Corpus of comma placement Vejica 1.0*, Slovenian language resource repository CLARIN.SI, http://hdl.handle.net/11356/1055.

Damjan Popič; et al. 2017. *Tweet comma corpus Janes-Vejica 1.0*, Slovenian language resource repository CLARIN.SI, http://hdl.handle.net/11356/1088.

Tadeja Rozman; et al. 2013. *Learners' corpus Šolar 1.0*, Slovenian language resource repository CLARIN.SI, http://hdl.handle.net/11356/1036.