Kako dobro programi popravljajo vejice v slovenščini

Peter Holozan

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

Povzetek

Za slovenščino obstaja dva slovnična pregledovalnika, ki med drugim popravljata tudi napake pri vejicah v besedilu, to sta Besana in LanguageTool. S pomočjo zbrane baze primerov napačne rabe vejic (kjer je večina primerov pridobljena iz korpusa Šolar) je bil izračunan priklic za iskanje manjkajočih in odvečnih vejic za oba programa. Pokazalo se je, da oba programa kar dobro odkrivata manjkajoče vejice, pri čemer je Besana uspešnejša, ker zna opozoriti, da vejica manjka, tudi v primerih, ko sicer ne zna točno postaviti vejice. Pri odvečnih vejicah je Besana manj uspešna, LanguageTool pa jih sploh ne odkriva.

How well programs correct commas in Slovenian

There are two grammar checkers for Slovenian, Besana and Language Tool, and both are capable of correcting incorrect positions of commas in text. Thanks to the collection of examples of wrong usage of commas (the majority of the examples came from Šolar) recall for missing and superfluous commas was calculated for both programs. The results showed that both programs successfully detect missing commas, but Besana got better results because it is able to warn us when a comma is missing also in cases where it is not able to place a comma correctly. In superfluous commas Besana is less successful, but LanguageTool does not even detect them.

1 Uvod

Amebis že več kot 20 let razvija slovnični pregledovalnik Besana, ki med drugim popravlja tudi napake pri postavljanju vejic v slovenskih besedilih.

Takoj se zastavi vprašanje, kolikšen delež napak pri vejicah Besana odkrije. Vendar je na to vprašanje težko odgovoriti brez dovolj velikega korpusa besedil, v katerem bi bile označene napake pri vejicah.

S korpusom Šolar (Rozman et al., 2010), v katerem so napake označene, se je pokazala možnost za preizkus Besane in hkrati še za primerjavo s slovenskimi pravili za LanguageTool, odprtokodni slovnični pregledovalnik, ki je vključen tudi v LibreOffice.

Najprej bosta na kratko opisana Besana in LanguageTool.

Sledil bo opis korpusov Šolar (v katerem so zbrani šolski pisni izdelki in učiteljski popravki) in Kust (kjer besedila tujcev, ki se učijo slovenščino) (Rozman et al., 2010) in metode dela.

Na koncu bodo predstavljeni rezultati preizkušanja in možnosti za izboljšanje preizkušanja.

2 Besana

Besana[1] (kar je okrajšava za BESedna ANAliza) je slovnični pregledovalnik, ki ga razvija podjetje Amebis. Namenjen je iskanju napak v besedilih, in sicer predvsem slovničnih, to je takih, ki jih črkovalnik ne more odkriti.

Odkriva npr. neujemanje med pridevniki in samostalniki v sklonu spolu in številu, napačne sklone za predlogi, napačne variante predlogov s/z oz. k/h, napačne predloge pri krajevnih imenih, zanikanje s tožilnikom, napačno tvorbo trpnika, nekatere tipične neknjižne uporabe, napačno dvojino, napačne velike/male začetnice, presledke pri ločilih ipd. Pomemben del pa je tudi opozarjanje na napake pri vejicah, in sicer tako na manjkajoče kot odvečne vejice.

Besana lahko deluje kot samostojen program (Besana Mini) ali pa je vključena kot preverjanje slovnice v Microsoft Word ali LibreOffice. Dodatek pri Besani je še pregibnik, tj. program za pregibanje (spreganje, sklanjanje) besed.

Besana odkriva morebitne napake na dva načina: osnovni način je stavčni analizator, ki ima vgrajene tudi tipične napake (opis nekaterih tipičnih napak, ki jih sprejema analizator, in težav, ki lahko zaradi tega nastanejo, je v Holozan (2006)) oz. so tipične napake skupaj s podatki o vrstnih oznakah vgrajene že v leksikalno podatkovno zbirko ASES (Arhar, Holozan, 2009). Kadar pa analizatorju analiza ne uspe, uporabi Besana pomožna pravila, ki pa so vgrajena neposredno v kodo programa in jih uporabniki ne morejo prilagajati (lahko pa pri vseh vrstah napak nastavijo, ali želijo, da jih Besana opozarja nanje).

Prva verzija Besane (za okolje DOS) je bila napisana v jeziku C, zdaj pa je napisana v jeziku C++.

3 LanguageTool

LanguageTool[2] je odprtokodni program za preverjanje sloga in slovnice. Podpira angleščino, francoščino, nemščino, poljščino, nizozemščino romunščino in še množico drugih jezikov, med katerimi je tudi slovenščina. Odkriva napake, ki jih črkovalniki ne morejo. Program deluje s pravili, ki so narejena za vsak jezik posebej. Osnovna pravila so zapisana v formatu XML, podpira pa tudi kompleksna pravila, napisana v jeziku Java.

Prvotni avtor programa je Daniel Naber, ki je program razvil v okviru svojega diplomskega dela, zdaj pa pri razvoju sodelujejo tudi drugi razvijalci, še posebno pri pisanju jezikovno odvisnih pravil. Prva verzija je bila napisana v programskem jeziku Python, zdaj pa je napisana v jeziku Java.

Je prosto dostopen pod licenco LGPL. Glavni vzdrževalec za slovenska pravila je Martin Srebotnjak, 30. 6. 2012 je bilo za slovenščino 85 pravil (od teh 41 za manjkajoče vejice).

[1] http://besana.amebis.si

[2] http://www.languagetool.org/

```
<rulegroup name="Manjkajoča vejica pred 'zato'" id="ZATO_BREZ_VEJICE">
<rule>
<pattern mark_from="1" mark_to="-1" case_sensitive="yes">
<token regexp="yes" negate="yes">[,\(;:\–]|[Ii]n|ter|[Aa]li|[Ss]amo|[Ll]e|[Zz]golj|[Pp]redvsem|[Ll]ahko|[Aa]mpak|glavnem|[Mm]orda</token>
<token>zato</token>
<token negate="yes">,</token>
</pattern>
<message>Ponavadi je pred 'zato' vejica: <suggestion>, zato</suggestion>!</message>
<short>Najbrž manjka vejica pred 'zato'</short>
<example type="correct">Bilo je vroče<marker>, zato</marker> sva se slekli.</example>
<example type="incorrect">Bilo je vroče<marker> zato</marker> sva se slekli.</example>
</rule>
</rulegroup>
```

Slika 1. Primer pravila iz LanguageTool za odkrivanje manjkajočih vejic pred *zato*.

Na sliki 1 je primer pravila za manjkajoče vejice pred veznikom *zato*. Pravilo pravi, da če najde v besedilu besedo *zato*, pred katero ni ločil ,(;:- in ne besed *in, ter, ali, samo, le, zgolj, predvsem, lahko, ampak, glavnem, morda*, potem opozori, da najbrž manjka pred tem *zato* vejica.

LanguageTool lahko deluje kot samostojni program ali pa je vključen v LibreOffice oz. OpenOffice.

4 Priprava testnih podatkov

Večina primerov za napake pri postavljanju vejic je bila zbrana iz korpusa Šolar, del pa tudi iz korpusa KUST in drugih virov.

Osnovni namen zbirke napak pri vejicah je ugotavljanje priklica, ker vsebuje le povedi z napačno postavljenimi vejicami, ne pa tudi pravilnih povedi, kar bi bilo potrebno za izračun prave natančnosti. Natančnost se zato izračuna le nad primeri z napakami, kot nadomestek pa se lahko izračuna še nad temi primeri s popravljenimi napakami.

Za uporabnike slovničnih pregledovalnikov, ki sami poznajo slovnična pravila, natančnost ni tako zelo pomembna (seveda do neke mere, preveč lažnih napak vseeno postane moteče), saj sami hitro ocenijo, ali gre res za napačno postavljeno vejico in jim program predvsem pomaga pri primerih, ko spregledajo napačno rabo vejico; po drugi strani pa uporabniki, ki sami slovničnih pravil ne poznajo, lahko preveč zaupajo slovničnim pregledovalnikom in zanje bo pomembna tudi visoka natančnost.

4.1 Korpus Šolar

Korpus šolskih pisnih izdelkov Šolar je korpus besedil, ki so jih učenci slovenskih osnovnih in srednjih šol samostojno tvorili pri pouku. Zajeta so besedila, kjer je slovenščina materni jezik avtorjev, niso bila napisana posebej za projekt, ampak so del šolske produkcije, in jezikovni popravki so realni, kakršne so naredili učitelji (Rozman et al., 2010).

V korpusu so besedila učencev od 6. do 9. razreda osnovnih šol in od 1. do 5. letnika srednjih šol.

šola	število/delež besedil	število/delež besed
osnovna šola	505 / 18,7 %	133423 / 13,8 %
poklicno izobraževanje	183 / 6,8 %	52422 / 5,4 %
strokovno izobraževanje	843 / 31,2 %	261496 / 27 %
gimnazija	1172 / 43,3 %	520136 / 53,8 %

Tabela 1: Besedila v korpusu Šolar po šolah.

Tabela 1 prikazuje deleže po vrstah šol, ki so vključene v korpus.

82,3 % besedil je bilo zajeto pri predmetu slovenščina, druga pa pri drugih predmetih (npr. psihologija, sociologija, zgodovina), večina besedil je iz let 2009 in 2010. Pomemben kriterij pri gradnji korpusa je bila tudi regijska uravnoteženost (Rozman et al., 2010).

V korpus so vključeni tudi popravki napak, in sicer tako popravki, ki so jih naredili učitelji (razmerje med besedili, ki so imela učiteljske poprave, in nepopravljenimi besedili je 1459:1292), kot popravki, ki so jih vnesli sestavljavci korpusa (vendar je bilo to narejeno le nad delom korpusa).

4.1.1 Priprava primerov

Za pripravo podatkov so bili uporabljeni popravki iz korpusa, in sicer je bil narejen postopek, ki je vpisal manjkajoče vejice (z znakom ¤) in označil odvečne (z znakom ÷). Korpus je bil nato razdeljen na povedi in izločene so bile vse povedi, ki niso vsebovale odvečnih ali manjkajočih vejic.

Vendar se je kasneje izkazalo, da postopek ni bil popolnoma natančen. Tako je npr. kot odvečne označil vejice, ki sicer v resnici niso odvečne, vendar so bile v popravkih nadomeščene s pomišljajem ali s piko. Tudi pri odvečnih vejicah so bile težave, ker so bile v popravkih kdaj izbrisane posamezne besede ali pa je bilo kaj dodanega. Te primere je bilo treba ročno popraviti, delno si je dalo pomagati s tem, da je bil v takih primerih velikokrat stranski učinek odvečen presledek pred znakom za manjkajočo ali odvečno vejico. Del napak je najbrž ostal, vendar ročno preverjanje rezultatov ni pokazalo, da bi bilo tega zelo veliko, tako da to ne bi smelo bistveno vplivati na rezultate.

Na nekatere od teh napak (in verjetno tudi na napake v samem Šolarju, kjer so popravki za tisto, kar so opazili učitelji, kar pa niso nujno vse napake) pa je pokazal tudi pregled primerov, kjer Besana ni bila uspešna. Te napake so bile popravljene, tako da je zdaj baza primerov napak pri vejicah natančnejša, kot je sam Šolar, bi pa bilo smiselno nekoč v prihodnosti nazaj združiti to bazo primerov, pridobljenih iz Šolarja, s samim korpusom Šolar in tako po eni strani dobiti še seznam pravilnih primerov, da bo mogoče povedati pravo natančnost, po drugi strani pa se bodo pokazale še napake v korpusu Šolar.

4.2 Korpus KUST

Korpus KUST (korpus usvajanja slovenščine kot tujega jezika) je zbirka besedil, ki so jih napisali govorci drugih jezikov, ki se učijo slovenščine. Tak korpus je bil predlagan v (Stritar, 2006), besedila so bila zbrana v

okviru projekta ESS Uspešno vključevanje otrok, učencev in dijakov migrantov v vzgojo in izobraževanje, ki ga je izvajal Center za slovenščino kot drugi/tuji jezik Filozofske fakultete Univerze v Ljubljani (Rozman et al., 2010). Besedila so bila napisana na roko in pretipkana v okviru projekta, dodani so bili podatki o tipu dokumenta, vrsti in stopnji tečaja, na katerem je besedilo nastalo, starosti avtorja besedila, državi izvora, kraju bivanja in njegovem prvem jeziku. V besedilih so bili prekriti podatki, ki razkrivajo identiteto avtorja (če je besedilu npr. omenjeno ime Janez, je to nadomeščeno z XImeX, in to ne glede na sklon). V korpusu je skupaj 32.117 besed v 306 besedilih (Rozman et al., 2010).

jezik	delež
španščina	31,7 %
italijanščina	29,9 %
italijanščina dvojezično s slovenščino	1,8 %
angleščina	11,2 %
srbščina	8,4 %
nemščina	4,3 %
nemščina dvojezično s slovenščino	1,5 %
bosanščina	2,7 %
hrvaščina	0,8 %
makedonščina	3,7 %

Tabela 2: Deleži (glede na število besed) glede na prvi jezik avtorja v korpusu KUST.

Tabela 2 prikazuje deleže glede na prvi jezik avtorja, drugi jeziki imajo največ 1,8 % (Rozman et al., 2010). Starost avtorjev je bila med 13 in 21 let, večina jih je imelo med 16 in 19 let (Rozman et al., 2010).

Primeri iz korpusa KUST so bili vključeni zaradi domneve, da je tukaj več napak kot v besedilih rojenih govorcev in da bi lahko slovnični pregledovalnik po potrebi prilagodili, da bi bil še bolj uporaben za tujce, ki se učijo slovensko.

4.2.1 Priprava primerov

Za pripravo primerov je bilo treba izhajati iz nepopravljenih besedil, saj napake v korpusu KUST še niso označene. Besedilo je bilo razrezano na povedi in razporejeno glede na prvi jezik učenca, pri čemer sem združil hrvaščino, srbščino in bosanščino v isto skupino. Potem je bilo treba ročno poiskati manjkajoče in odvečne vejice, kar je vzelo precej časa, občasno pa je bilo tudi zapleteno razvozlati, kaj je pisec pravzaprav mislil. Na mesta, kjer manjkajo vejice, je bil dopisan znak ¤, odvečne vejice pa so bile nadomeščene z znakom ÷. Jeziki, za katere je bilo manj besedil v korpusu (nemščina, angleščina in srbščina/hrvaščina/bosanščina), so bili obdelani v celoti, kjer je bilo besedil veliko (italijanščina in španščina), je bila obdelana prva tretjina korpusa (ker bi bilo ročno popravljanje vejic preveč dolgotrajno in še več primerov najbrž ne bi bistveno vplivalo na rezultate), vsi drugi jeziki, kjer je bilo le malo besedil oz. besed v njih, pa so bili izpuščeni.

4.3 Drugi viri

Primeri iz korpusov Kust in Šolar so bili dodani že obstoječi bazi primerov za napake pri vejicah. Primeri v njej so bili zbrani iz različnih jezikovnih priročnikov, ki obravnavajo vejice, vaj, člankov na temo vejic (npr. (Šek Mertük, 2011)), diplomske naloge (Žibert, 2006) in drugih primerov za vejice, ki so bili zbrani med razvojem programa Besana v podjetju Amebis.

Vendar pa ti primeri ne odražajo pravega razmerja napak, ker je to tipično le izbor, v priročnikih je običajno več primerov za vejice, ki jih je teže postaviti, v besedilih se pa taki primeri ne pojavljajo tako pogosto. Zato niso tako uporabni za določanje uporabnosti slovničnih pregledovalnikov, seveda so pa uporabni za razvoj pregledovalnikov, saj je ravno pri težjih vejicah pomembno, da bi jih prav postavili.

Iz dela teh virov in iz primerov tretjega letnika gimnazije iz Šolarja je narejen krajši testni nabor, ki se ne uporablja pri izboljševanju Besane. Rezultati tega nabora služijo za preverjanje, da izboljšave niso preveč prilagojene na točno določene primere.

4.4 Priprava baze primerov

Baza je izvedena kot seznam vrstic s tremi stolpci, pri čemer je v prvem stolpcu oznaka kategorije (ki npr. pove, ali gre za primere, ki se bodo uporabljali za kontrolno skupino pri dodajanju pravil za vejico ali za osnovne podatke, ločuje pa tudi podatke iz korpusov Kust in Šolar od drugih podatkov).

V drugem stolpcu je številčni[3] podatek o viru. Primeri iz korpusa Kust so razdeljeni glede na prvi jezik učenca (nemščina, angleščina, španščina, italijanščina, srbščina/hrvaščina/bosanščina), primeri iz korpusa Šolar pa glede na razred oz. letnik in vrsto šole (6. do 9. razred osnovne šole, 1. do 3. in 5. letnik poklicne šole, 1. do 4. letni srednje strokovne šole in 1. do 4. letnik gimnazije). Možno bi bilo tudi združiti podatke iz korpusa Šolar v le štiri kategorije (osnovna šola, poklicna šola, strokovna šola in gimnazija), vendar me je zanimalo tudi, koliko se rezultati znotraj teh skupin ujemajo po letnikih, in ne le povprečki. Slabost te odločitve pa so bolj nepregledne tabele (seveda pa je možno podatke potem po potrebi združevati).

V tretjem stolpcu so primeri stavkov. Manjkajoče vejice so označene z znakom ¤, odvečne pa nadomeščene z znakom ÷.

Baza je izvedena kot preglednica v programu, da pa jo programi potem laže uporabljajo, se naredi izvoz v besedilno obliko, kjer so vrstice ločene s kodo za novo vrstico, stolpci pa s kodo za tabulator.

```
5	25	Ta hčerka še ni pri kruhu¤ saj se šola kot frizerka v Mariboru.
5	25	Veliko časa tudi preživim pred televizijo÷ ter me zanimajo znanstvene odajo.
5	25	Nato ju je policija izsledila v kampu¤ vendar predrzni najstnici sta ušle¤ zatekli sta se zapuščenim otrokom.
5	25	Če ju opazite¤ se javite na policiski postaji.
```

Slika 2. Izsek iz baze primerov (8. razred OŠ iz korpusa Šolar).

[3] Številčne oznake so bile izbrane, da lahko program neposredno zapisuje rezultate v tabelo. Dodana je manjša baza z legendo oznak virov.

Na sliki 2 je delček primerov, zbranih iz korpusa Šolar. Opazno je, da je v besedilu tudi veliko drugih napak, torej mora biti postopek za odkrivanje napak pri vejicah čim bolj odporen na druge napake.

vir	število manjkajočih vejic	število odvečnih vejic
testni nabor	795	203
KUST	401	104
Šolar - osnovna šola	2332	415
Šolar - poklicno izobraževanje	1285	189
Šolar - strokovno izobraževanje	4310	985
Šolar - gimnazija	3362	1155

Tabela 3: Število zbranih primerov.

V tabeli 3 je zbrano število primerov napak pri vejicah po kategorijah. Primeri iz korpusa Šolar močno prevladujejo.

4.4.1 Opcijske vejice

V nekaterih primerih je slovnično možno oboje: da vejica bodisi je bodisi je ni. Take vejice največkrat vplivajo na pomen besedila oz. vsaj na poudarek.

Pravilno postavljanje takih vejic za zdaj presega zmožnosti računalniškega preverjanja (možno bi bilo kvečjemu opozarjanje na morebitno dvoumnost v takih primerih), zato taki primeri niso bili uvrščeni v bazo primerov.

Posebni primer so vejice pri datumih v primerih: *v nedeljo, 9. maja (bo)*, kjer Slovenski pravopis 2001 dopušča tudi pisanje brez vejice. Besana v tem primeru opozarja, če vejice ni, je pa to ločena kategorija napake, tako da je mogoče opozarjanje na vejice v takih primerih preprosto izključiti, če je uporabnik ne želi uporabljati.

5 Postopek preizkušanja

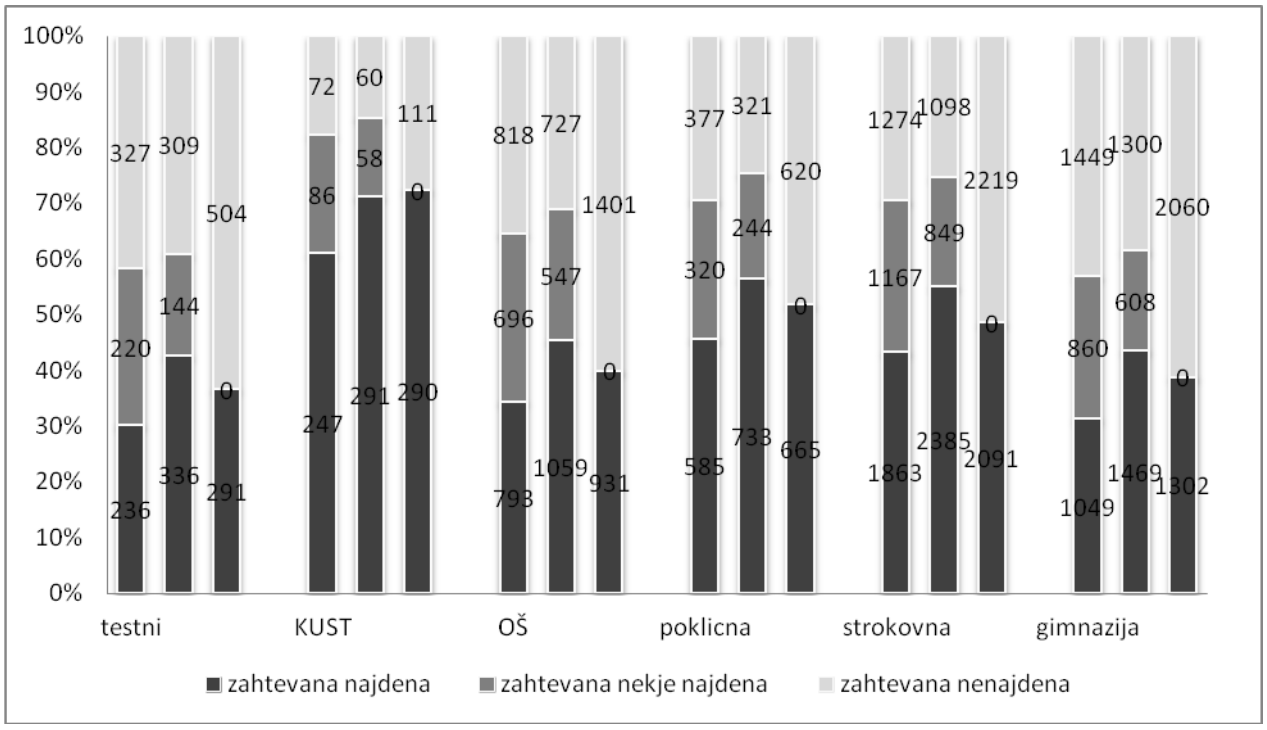
Za preizkušanje je bil uporabljen program Vejicar, ki je sicer namenjen preizkušanju zanesljivosti Besane. Program uporabi bazo primerov, kjer so označene manjkajoče in odvečne vejice (z znakoma ¤ in ÷). Za vsak primer najprej nadomesti znak za odvečno vejico (÷) z vejico in izbriše znak za manjkajočo vejico (¤). Stavek spusti skozi Besano in ugotovi, ali je Besana pravilno opozorila na manjkajoče in odvečne vejice. Rezultate ob skupnem rezultatu izračunava še za posamezne vire in kategorije.

Program Vejicar je bil dopolnjen še za uporabo rezultatov LanguageTool, in sicer tako, da so bila pravila v LanguageTool popravljena tako, da dodajajo pri manjkajoči vejici znak ¤ namesto vejice, datoteka, ki je vsebovala primere (vsakega v svoji vrstici), je bila pognana skozi program (pri čemer so bila vključena le pravila za vejice) in rezultati so bili dodani kot dodatni stolpec v vhodu programa Vejicar.

6 Rezultati

Program Vejicar izpiše rezultate v datoteko, in sicer v obliki tabele, v kateri so elementi ločeni s tabulatorjem, tako da je rezultate preprosto prek odložišča prenesti v preglednico v Excelu, kjer se potem naredijo dodatne obdelave.

Ker je bilo v zadnjem letu iskanje napak pri vejicah v Besani precej izpopolnjeno (tudi po zaslugi zbrane baze primerov), sta bila zabeležena dva rezultata Besane, ki sta narazen 9 mesecev, v tretjem stolpcu pa je rezultat LanguageTool. Korpus Šolar je razdeljen glede na vrsto šole.



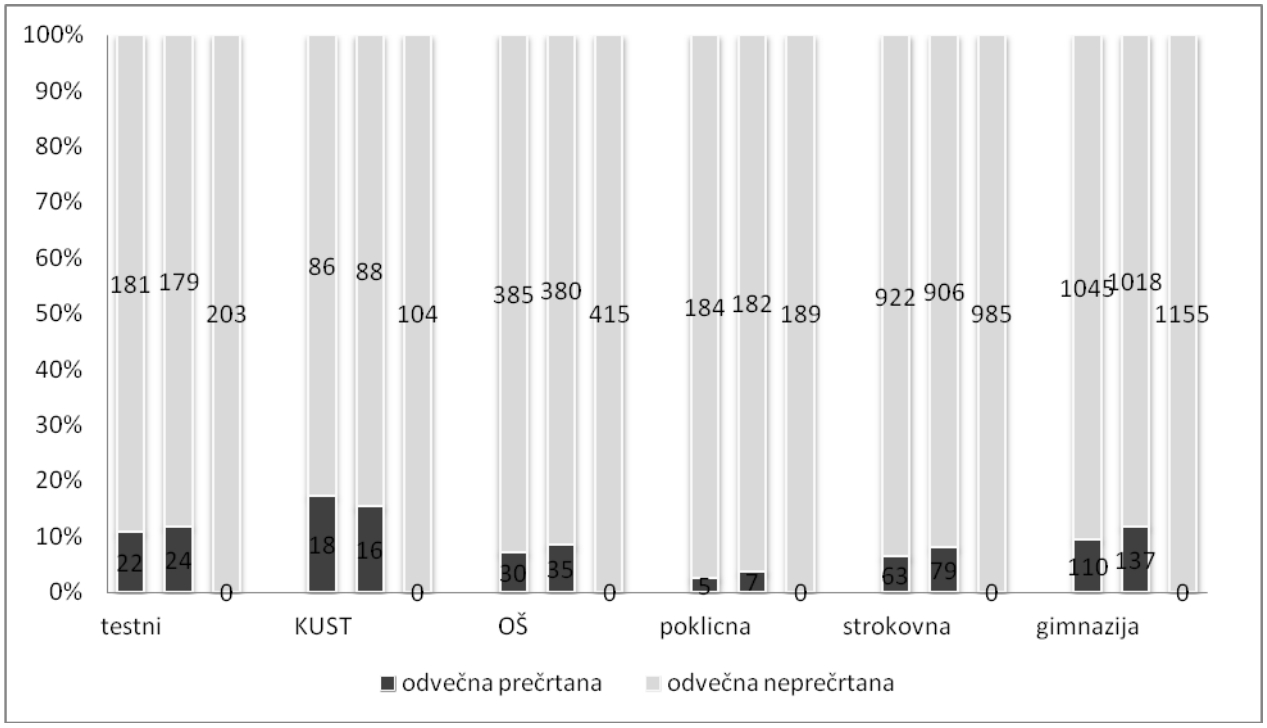
Slika 3: Priklic manjkajočih vejic (po stolpcih: Besana septembra 2011, Besana julija 2012, LanguageTool maja 2012)

Besana opozarja na manjkajoče vejice na dva načina. Pri prvem točno določi mesto, kje manjka vejice, največkrat je to na veznikih, ki zahtevajo vejico. Pri drugem načinu pa le opozori, da vejica manjka nekje v stavku, ne zna pa ugotoviti, kje to je. Do takih opozoril pride npr. v primerih, ko najde v stavku dve osebni glagolski obliki brez vmesnega veznika ali vejice, kar je jasen znak, da bi morala nekje vmes biti vejica.

Rezultat kaže, da ravno ta opozorila, kjer vejica ni natančno najdena, naredijo večino razlike med Besano in LanguageTool. Rezultati obeh programov pa so precej odvisni od piscev, oba bolje delujeta v primeru, ko znajo pisci slabše slovenščino in pozabljajo na vejice pred tipičnimi vezniki.

Natančnost opozarjanja na manjkajoče vejice (le nad primeri z napakami pri vejicah, tako odvečnimi kot manjkajočimi) je pri Besani 87,8 %, pri LanguageTool pa 81,5 %. Za pravo natančnost bi bila potrebna testna baza, ki bi vsebovala tudi povedi s pravilnimi vejicami.

Uporaba korpusa Šolar pa je pokazala, da niso težava le manjkajoče vejice, ampak tudi odvečne. V primerih iz korpusa Šolar skupaj manjka 11340 vejic, odvečnih pa je 2744, kar pomeni, da je v skoraj 20 % primerov problematičnih vejic težava to, da je vejica odveč. Če dodamo še to, da Besana popravi 60 do 70 % manjkajočih vejic, je potem odvečnih vejic skoraj toliko kot manjkajočih, kajti rezultati programov pri odvečnih vejicah so precej slabši kot pri manjkajočih.



Slika 4: Priklic odvečnih vejic (po stolpcih: Besana septembra 2011, Besana julija 2012, LanguageTool maja 2012)

LanguageTool sploh ne išče odvečnih vejic in tudi Besana jih najde le okoli 10 %. To kaže, da je pri odvečnih vejicah še veliko prostora za izboljšave slovničnih pregledovalnikov.

Zanimivo pa je, da ima Besana pri odvečnih vejicah boljši rezultat pri gimnaziji kot pri poklicni oz. strokovni šoli, kar je ravno obratno kot pri manjkajočih vejicah.

7 Kako naprej

Treba bi bilo zbrati še bazo primerov, kjer bi bile vključene tudi pravilne povedi, da bi bilo mogoče izračunati še natančnost popravljanja vejic. Pri izdelavi take baze je treba biti previden, saj kakovost lekture besedil zelo vpliva na rezultat natančnosti, saj v primeru, da za preizkušanje uporabimo besedila, kjer skoraj ni napak, dobimo zelo slab rezultat za natančnost (Helfrich, Music, 2000). Zato je treba zbrati realna besedila, ki še niso bila lektorirana, po možnosti taka, kjer obstaja tudi lektorirana verzija, da ni treba ročno iskati napak. Idealno morajo biti pisci čim bolj raznoliki, z različnim jezikovnim znanjem.

V bazo primerov bi bilo smiselno dodati tudi posebej označene primere za opcijske vejice, torej vejice, ki lahko so ali ne, pri čemer se bolj ali manj spremeni pomen. Zbirka teh primerov bi omogočila preizkušanje tudi takih primerov in morda razvoj postopkov za opozarjanje na morebitne dvoumne primere.

Naslednji korak je podrobnejša analiza, na katerih mestih tipično manjkajo oz. so odveč vejice, s pomočjo česar bo mogoče dopolniti pravila za iskanje manjkajočih oz. odvečnih vejic.

V prihodnosti bo treba v slovničnih pregledovalnikih posvetiti večjo pozornost odkrivanju odvečnih vejic, saj kaže, da je tudi to pogosta težava pri pisanju. Zanimivo pri odvečnih vejicah je, da so v 30 % do 50 % pred vezniki, ki sicer tipično zahtevajo vejico (npr. *ko*, *ki*, *da*), a je v teh primerih ne sme biti, npr. zaradi tega, ker je spredaj kakšen drug veznik ali pa členek (ta podatek je iz pripravljalnega poskusa podrobnejše analize).

Po drugi strani pa bilo smiselno uporabiti korpus Šolar tudi za celovit preizkus slovničnih pregledovalnikov, ki se ne bi omejil le na napake pri vejicah, ampak bi zajel vse napake. Bo pa v tem primeru, posebno pri slogovnih popravkih, treba paziti na to, da popravki niso nujno absolutni, ampak bi tudi dva lektorja različno popravila isto besedilo, kar pomeni, da bo težko narediti popolnoma samodejno preizkušanje, ampak bo treba rezultate ocenjevati ročno, kar bo žal omejilo velikost vzorca, ki ga bo mogoče preizkusiti.

8 Sklep

Rezultat nad testnim naborom je naslednji:

kategorija	Besana	LanguageTool
priklic manjkajočih vejic	60,8 %	36,6 %
natančnost pri manjkajočih vejicah	87,8 %	79,0 %
priklic odvečnih vejic	11,8 %	/
natančnost pri odvečnih vejicah	100 %	/

Tabela 4: Rezultati nad testnim naborom.

Rezultati nad testnim naborom so nekaj slabši, kot so sicer rezultati nad korpusom Šolar, še posebej pa nad korpusom KUST, kar kaže na to, da je v testnem naboru večji delež napak, ki jih je težko odkriti.

Tako Besana kot LanguageTool kar uspešno odkrivata manjkajoče vejice v slovenskih besedilih, Besana jih odkrije dobrih 20 % več, in to predvsem na račun opozoril, kjer ne zna točno postaviti vejice, ve pa, da nekje manjka. Težava pri teh opozorilih pa je, da mora uporabnik dovolj dobro poznati pravila za postavljanje vejic, da lahko sam postavi vejico na pravo mesto. Prav tako pa niso uporabna za popolnoma samodejno postavljanje vejic (npr. pri razpoznavi govora), zato bi bilo dobro v čim več primerih določiti pravo mesto, kje manjka vejica. Spremembe Besane v zadnjem letu so bile izrazito v tej smeri, kar kažejo tudi rezultati na sliki 3 (pred temi popravki je imel LanguageTool boljši rezultat kot Besana, če bi upoštevali le natančno določene manjkajoče vejice).

Besana je tudi bolj natančna pri opozarjanju na manjkajoče vejice, brez nove baze primerov pa ni mogoče reči, kakšna je v resnici natančnost obeh programov.

Odvečne vejice delajo težave obema programoma, LanguageTool jih sploh ne odkriva, Besana pa jih odkrije le okoli 10 %.

Zanimivo bi bilo rezultate za slovenščino primerjati z rezultati za druge jezike, vendar se tu pojavi težava, da večinoma preizkušajo programe za preverjanje slovnice v celoti in pri rezultatih niso posebej navedeni rezultati za vejice. Primer, kjer so rezultati za vejice navedeni ločeno, je za latvijščino v Deksne, Skadiņš (2011), kjer so med tipi najdenih napak navedeni štirje tipi, ki zadevajo vejice. Priklic je med 14 % in 56,3 %, natančnost pa med 70,4 % in 91,3 %. Žal pa ni navedeno, koliko je primerov za posamezen tip, zato ni mogoče izračunati skupnega rezultata za vejice. Ni pa tudi jasno, ali gre le za manjkajoče ali tudi za odvečne vejice, vprašanje je, ali navedeni štirje tipi pokrivajo vse problematične vejice, zaradi različnih pravil o postavljanju vejic pa so rezultati tudi težko primerljivi.

9 Literatura

Arhar, Š., Holozan, P., 2009. ASES – leksikalna podatkovna zbirka za razvoj slovenskih jezikovnih tehnologij. V V. Mikolič (ur.), *Jezikovni korpusi v medkulturni komunikaciji*. Koper: Založba Annales.

Deksne, D., Skadiņš, R. 2011. CFG Based Grammar Checker for Latvian. V *Proceedings of the 18th Nordic Conference of Computational Linguistics NODALIDA 2011*. Riga, Latvia.

Helfrich, A., Music, B. 2000. Design and evaluation of grammar checkers in multiple languages. V *Proceedings COLING '00 Proceedings of the 18th conference on Computational linguistics - Volume 2*. Stroudsburg, PA, ZDA: Association for Computational Linguistics.

Holozan, P., 2006. Dodatne dvoumnosti zaradi popustljivosti analizatorja pri analizi slovenskih stavkov. V T. Erjavec in J. Žganec Gros (ur.), *Zbornik 5. slovenske in 1. mednarodne konference JEZIKOVNE TEHNOLOGIJE 2006*. Ljubljana: IJS.

Rozman, T., Stritar, M., Krapš Vodopivec, I., Kosem, I., Krek, S., 2010. *Nova didaktika poučevanja slovenskega jezika : sporazumevanje v slovenskem jeziku*. Ljubljana: Ministrstvo za šolstvo in šport: Amebis. http://www.slovenscina.eu/Media/Kazalniki/Kazalnik15/Nova_didaktika_Sporazumevanje.pdf.

Stritar, M., 2006: Oblikovanje korpusa usvajanja slovenščine kot tujega jezika, V T. Erjavec, J. Žganec Gros (ur.), *Zbornik 5. slovenske in 1. mednarodne konference Jezikovne tehnologije 2006*. Ljubljana: IJS.

Šek Mertük, P., 2011: Vejica premalo ali preveč pri študentih razrednega pouka. *Revija za elementarno izobraževanje*. Letnik 4, št. 1–2..

Žibert, Ž., 2006. *Slovenska vejica: balast ali skladenjska nujnost slovenskega knjižnega jezika?*. Diplomska naloga. Ljubljana: Fakulteta za družbene vede, Univerza v Ljubljani.