

Kako Besana popravlja vejice?

Peter Holozan



Napovednik

- ▶ polovica doktorske disertacije Računalniško postavljanje vejic v slovenščini
 - zbirka primerov rabe vejice
 - ~~◦ postavljanje vejic s strojnim učenjem~~
 - ~~◦ izboljševanje LanguageToola~~
 - izboljševanje Besane
 - + analizator povedi
- ▶ doktorska disertacija je dostopna na naslovu <http://peter.amebis.si/drvejice.pdf>

Motiv

- ▶ Slovenci imamo pravilno postavljanje vejic za znak višje pismenosti
 - postavitve vejic vpliva na naše dojemanje vsebine
 - veliko piscev ima težave s postavljanjem vejic
 - pravila zahtevajo slovnično znanje
- ▶ možnost za pomoč piscem s pomočjo programov za preverjanje vejic
 - programi že obstajajo, hipoteza, da se jih da izboljšati
- ▶ postavljanje vseh vejic, kadar teh ni, npr. pri razpoznavi govora

Zbirka primerov rabe vejice

- ▶ velika zbirka primerov z označenimi napakami pri postavljanju vejic nujnost za ocenjevanje uspešnosti popravljanja vejic
- ▶ uporabljena korpusa Šolar in Lektor
 - ni bilo mogoče neposredno uporabiti Šolarja, ker je preveč napak (učitelj popravil druge stvari, obupal nad popravljanjem ali pa preprosto spregledal)
 - zato ročni pregled primerov iz korpusa Šolar (3,95 milijona znakov oz. skoraj 2200 lektorskih strani)
- ▶ dodani še primeri iz korpusa Kust
- ▶ dodano manjše število člankov iz Wikipedije

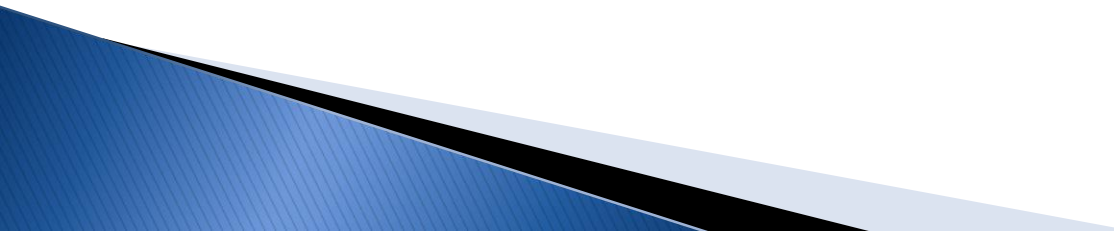
Nekatere dileme pri stavi vejic

- ▶ “tako, da” proti “, tako da”
- ▶ “prosim” kot členek
 - Pridite $\varnothing$ prosim $\varnothing$ jutri.
- ▶ oziralni zaimki za vezniki
 - Družini bi morali prej dojeti, da kar delata $\varnothing$ prinaša samo zlo.
- ▶ trije zaporedni vezniki
 - Ljudje pravijo, da v peklu živijo hudiči $\varnothing$ in da če prideš v pekel $\varnothing$ te tam ne čaka srečno življenje.
- ▶ tipkarske napake
 - Opravičil se je učiteljici in odšel $\varnothing$ da svoje klopi.

Vejica 1,0

- ▶ skupna velikost 113.308 povedi, 17.768 manjkajočih vejic in 4608 odvečnih vejic
- ▶ velika večina iz korpusov Šolar in Lektor
- ▶ zbirka primerov rabe vejice Vejica 1,0 je objavljena v repozitoriju CLARIN.SI pod licenco CC BY–NC–SA 4.0
- ▶ to omogoča drugim raziskovalcem delo na tem področju in primerjavo rezultatov
- ▶ alternativne možnosti pri vejicah
 - velikokrat v parih!
 - kontekst (Klavdij, kralj^x in stražarji so prihiteli.)

Besana

- ▶ program razvija podjetje Amebis
 - ▶ uporablja analizator povedi in potem iz določene strukture povedi ugotavlja, ali so vejice pravilno postavljene
 - ▶ pomožne metode: tipični vezniki, štetje osebnih glagolskih oblik
 - ▶ v nekaterih primerih Besana samo opozori, da nekje manjka vejica, ne ve pa točnega mesta
- 

ASES

- ▶ Amebisova glavna jezikovna baza podatkov
- ▶ večjezična
- ▶ različni med seboj povezani elementi
 - besede (SL: 310.000 lem, 8,3 milijona oblik)
 - zveze (SL: 37.000)
 - skupine (SL: 55.000)
 - predloge (SL: 8200)
 - pomeni (135.000)

Glagolske predloge

- ▶ G:G:SL:[zaljubiti] se {v PR4}=S:G:SL:zaljubiti!1!SE#OS!P1!!#PRZ!P2!B:D:SL:v (t)!NEOB
- ▶ G:G:SL:zatreskati se {v PR4}=B:G:SL:zatreskati!1!SE#OS!P1!!#PRZ!P2!B:D:SL:v (t)!NEOB
- ▶ G:G:SL:zacopati se {v PR4}=B:G:SL:zacopati!1!SE#OS!P1!!#PRZ!P2!B:D:SL:v (t)!NEOB
- ▶ G:G:SL:zagledati se v PR4=B:G:SL:zagledati!1!SE#OS!P1!!#PRZ!P2!B:D:SL:v (t)!
- ▶ G:G:SL:zatrpati se {v PR4}=B:G:SL:zatrpati!1!SE#OS!P1!!#PRZ!P2!B:D:SL:v (t)!NEOB
- ▶ G:G:EN:fall in love {with PRD}=B:G:EN:fall!1!#OS!P1!!#PRT!!Z:R:EN:in love!#PRZ!P2!B:D:EN:with!NEOB
- ▶ G:G:EN:fall for PRD=B:G:EN:fall!1!#OS!P1!!#PRZ!P2!B:D:EN:for!
- ▶ G:G:EN:become infatuated {with PRD}=B:G:EN:become!1!#OS!P1!!#POD!!B:P:EN:infatuated+Pdo!#PRZ!P2!B:D:EN:with!NEOB
- ▶ G:G:DE:verlieben sich {in PR4}=B:G:DE:verlieben!1!SE#OS!P1!!#PRZ!P2!B:D:DE:in (t)!NEOB
- ▶ G:G:DE:verknallen sich {in PR4}=B:G:DE:verknallen!1!SE#OS!P1!!#PRZ!P2!B:D:DE:in (t)!NEOB
- ▶ P:G:zaljubiti se {v PR4}=0!G:G:SL:[zaljubiti] se {v PR4}#0!G:G:SL:zagledati se v PR4#2!G:G:SL:zatreskati se {v PR4}#2!G:G:SL:zacopati se {v PR4}#2!G:G:SL:zatrpati se {v PR4}#0!G:G:EN:fall in love {with PRD}#2!G:G:EN:fall for PRD#2!G:G:EN:become infatuated {with PRD}#0!G:G:DE:verlieben sich {in PR4}#2!G:G:DE:verknallen sich {in PR4}
- ▶ *:G1 +#P:G:zaljubiti se {v PR4}#P:S:\$ \$je oseba

Pridevniške predloge

- ▶ G:P:SL:dolg POC=B:P:SL:dolg!1!#POC!P1!!
 - ▶ G:P:EN:long PRD=B:P:EN:long!1!#PRD!P1!!
 - ▶ P:P:dolg POC=0!G:P:SL:dolg
POC#0!G:P:EN:long PRD
 - ▶ *:G1 + #P:P:dolg POC#P:S:\$ \$omejeno
 - ▶ *:G1 > #P:P:dolg POC#P:S:\$ enota dolžine
-
- ▶ pet kilometrov dolga kolona
 - ▶ kolona, dolga pet kilometrov
 - ▶ Kolona je postala dolga pet kilometrov.

Analiza povedi

- ▶ kot del stavka:
 - izglagolski pridevniki, prislovi in **samostalniki**
 - prilastkovi odvisniki
 - prislovni odvisniki
 - osebkovi/predmetni odvisniki
- ▶ priredja
- ▶ soredja (medmeti, zvalniki)
- ▶ dodatno še upoštevanje izpustov, za zdaj predvsem osebka

Amebisov vmesni jezik

- ▶ To sem prebral v Janezu poslanem pismu.
- ▶

```
*(-POV:(-STAg-npppdvt-----:(2PR4:(-SFR:(-DSF:(-JED:(-SAZkes:{261c3f;4bf16f3}[0]<2d14>))))),(*PVD:(-GPO:[1])),(0PVD:(-GGL:{7a94f9;de62}[2]<23f0>)),(-PDOj:(-PZV:(-PRE:{f99;4354,83d6,7122,6f02e,356ce,10d37,ff104}[3]<f50>)),(-SFR:(-DSF:(-PFR:(-DPF:(-IGPab2b:(DPR3:(-SFR:(-DSF:(-JED:(-SAME:{473562;25fd262}[4]<14c1c>))))), (APVD:(-IGG:{77f8c7;205f2d}[5]<169e0>)), (COSB:(-SFR:(-DSF:(-JED:(-NOZnes-:[5])))))))), (-JED:(-SAME:{34ca9;8aeb14,3463b4,3462d6,3da7b4d}[6]<5f50>))))), (-LOCKp:[7]), (1OSB:(-SFR:(-DSF:(-JED:(-NOZnemp:[2]))))))))
```

Potencialne meje stavkov

- ▶ največ težav je pri levosmernih vejicah za odvisniki
- ▶ osebne glagolske oblike (je, mora, imam)
- ▶ naslonske oblike osebnih zaimkov (jo, ga, jih)
- ▶ nikalnice (ne, ni), ni čisto zanesljivo mesto
 - Če bo deževalo jutri ne pridi.
 - Če bo deževalo pridi jutri.
- ▶ meja se upošteva le, če je uspešna analiza

Analizator povedi

- ▶ trenutni rezultati za jos100k
 - analizator je preveč naučen na ta primer!
 - popravljenih okoli 20 napak v jos100k
 - pravilne leme: 99,07 %
 - pravilne oblikoskladenjske oznake: 96,09 %
 - ne nujno pravilni rezultati:
 - prepoznane besede: 99,33 %
 - analizirani stavki: 90,41 %
 - analizirane povedi: 70,79 %

Metoda izboljševanja Besane

- ▶ najprej po besedah kot pri LanguageToolu
 - tipični vezniki
 - kandidati za pogojne meje
- ▶ pomembnejša pa izdelava programa, ki vejicam pripiše razlog, zakaj vejica tam je
 - to je omogočilo, da se da po posameznih vrstah vejic preveriti, koliko težav imajo pisci (in Besana) z njimi
- ▶ ob tem še izboljšave analizatorja povedi

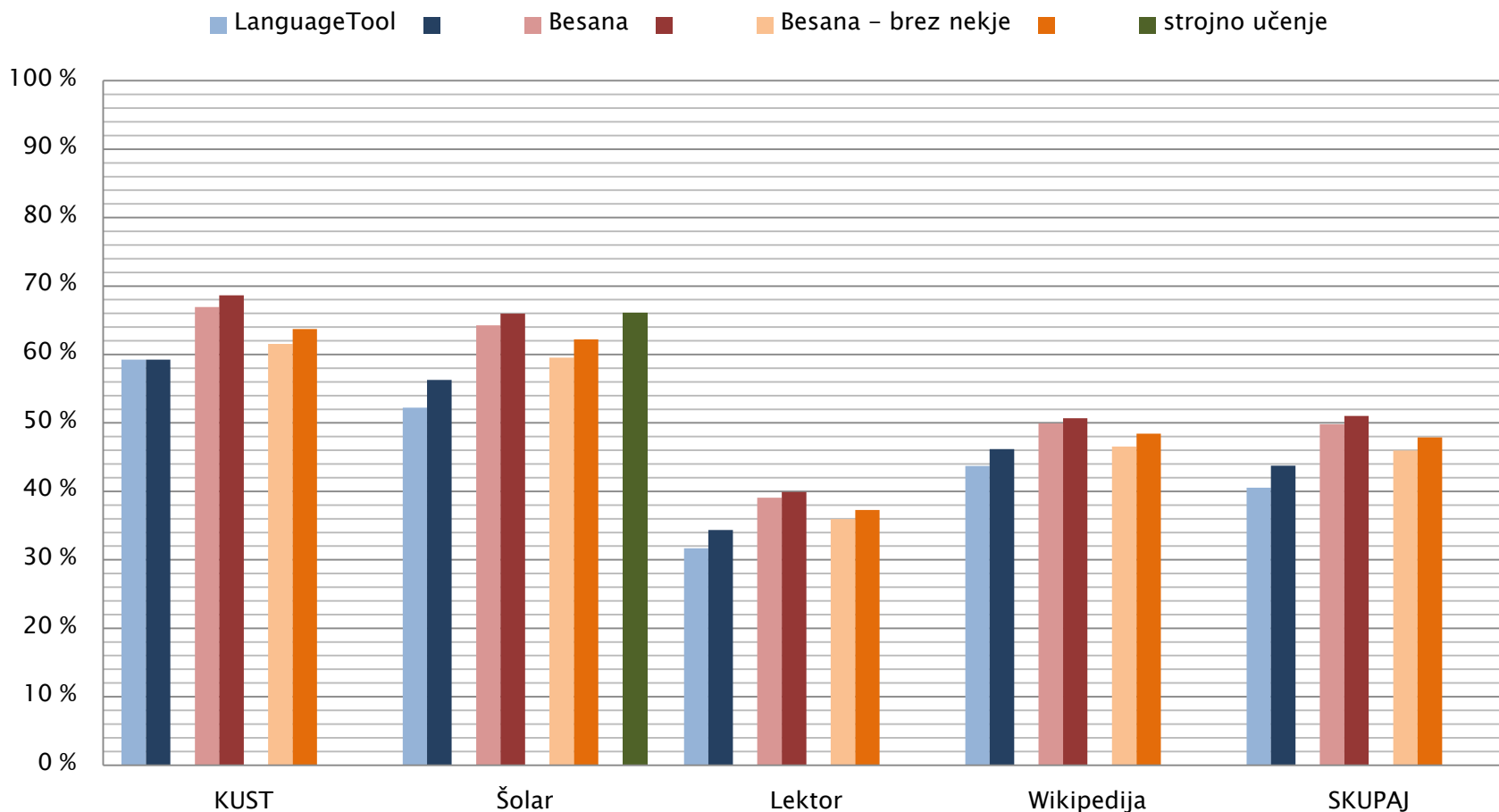
Določanje razlogov za vejice

- ▶ Denar je zelo hitro zapravljen za nesmisle $C1 + E2$, in $E2 - C1$ ko se je zavedel $E2' E2$, kaj je storil $E2'$, se je odločil za vrnitev k očetu $E7$, da ga prosi odpuščanja.
- ▶ Dana nam je možnost odpuščanja $C1 + E8$, in $E8 - C1$ če se nekaj hudega zgodi $E8'$, se potrudimo to popraviti $E8$, kolikor se da $E8' C1$, in za ostalo prosimo odpuščanja.
- ▶ Vera nam ponudi resnico in človek jo sprejme $C1 + E2$, in $E2 - C1$ ko jo sprejme $E2'$, človek vidi $E7$, da je zadovoljin tudi njegov razum.

Problematične vrste vejic

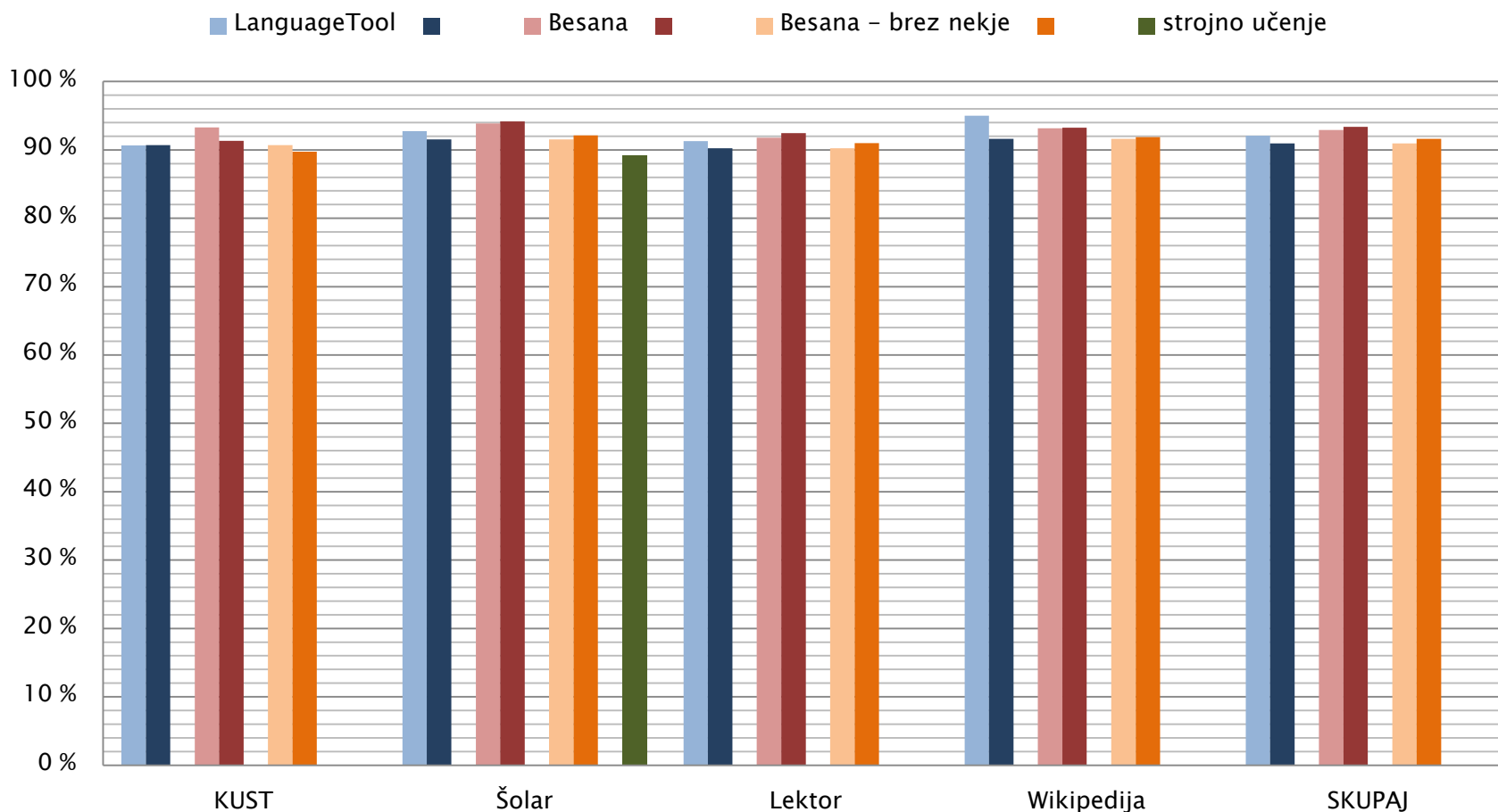
- ▶ levosmerne vejice za odvisniki
 - *Ker ga je punca **prevaralaš** je prosil Meursaulta, da ji napiše pismo in jo povabi k njemu na kavo.*
- ▶ začetki odvisnikov
 - *Tonček je študent (to v današnjem času pomeni gimnazijec) **š** ki so mu vseč vse ženske.*
- ▶ vezalno priredje za vrinjenim odvisnikom
 - *Dano besedilo je epika, ker **pripoveduješ in** je roman, ker je daljše besedilo.*
- ▶ vrinjeni odvisniki
 - *Sam je nezvest **možš in ko** ugotovi, da se njegova žena srečava z neznanim gospodom, je še ljubosumen.*

Priklic pri iskanju vseh vejic



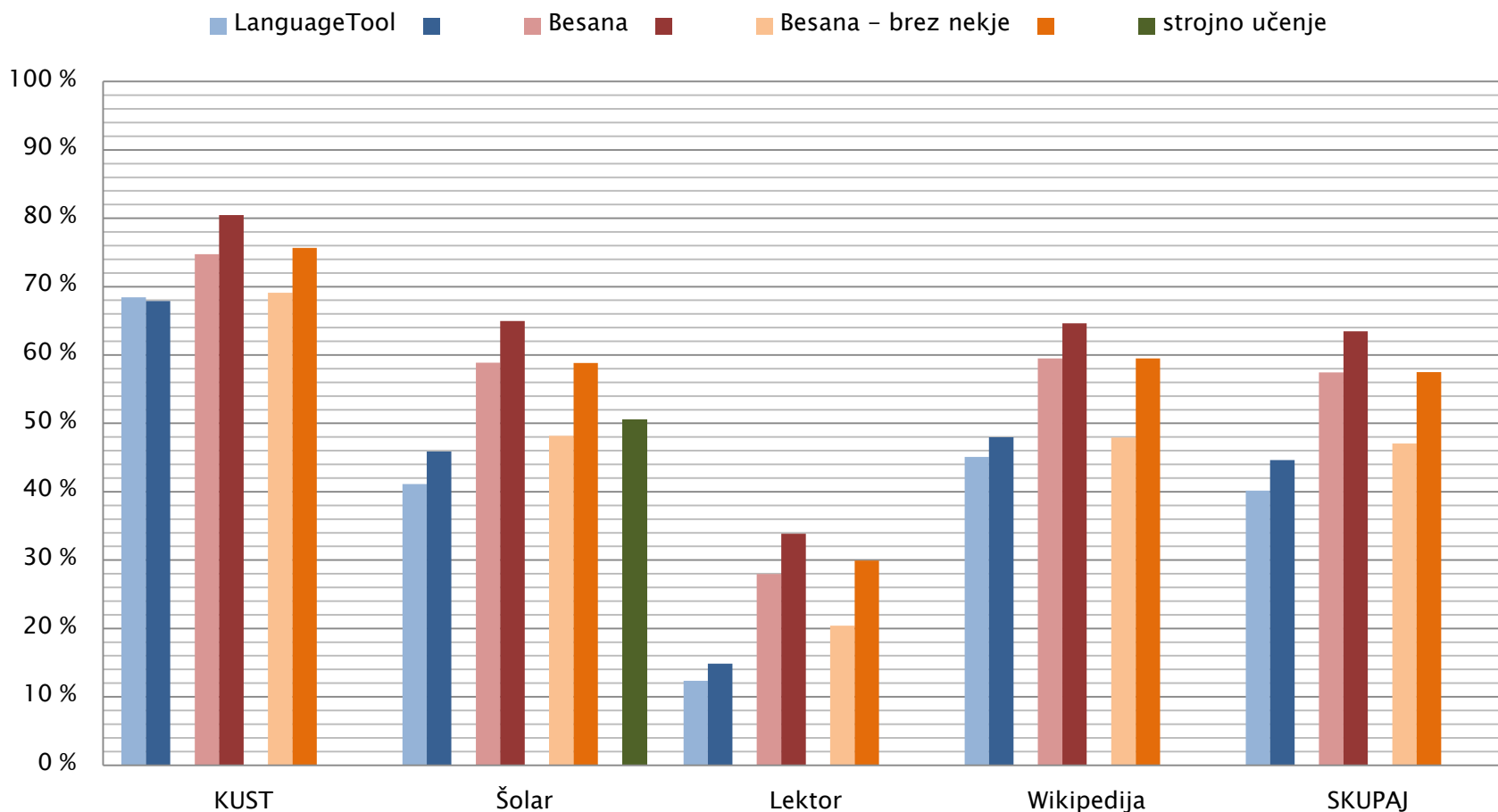
Rezultati:
Šolar 62–66 %
Lektor 37 %

Natančnost pri iskanju vseh vejic



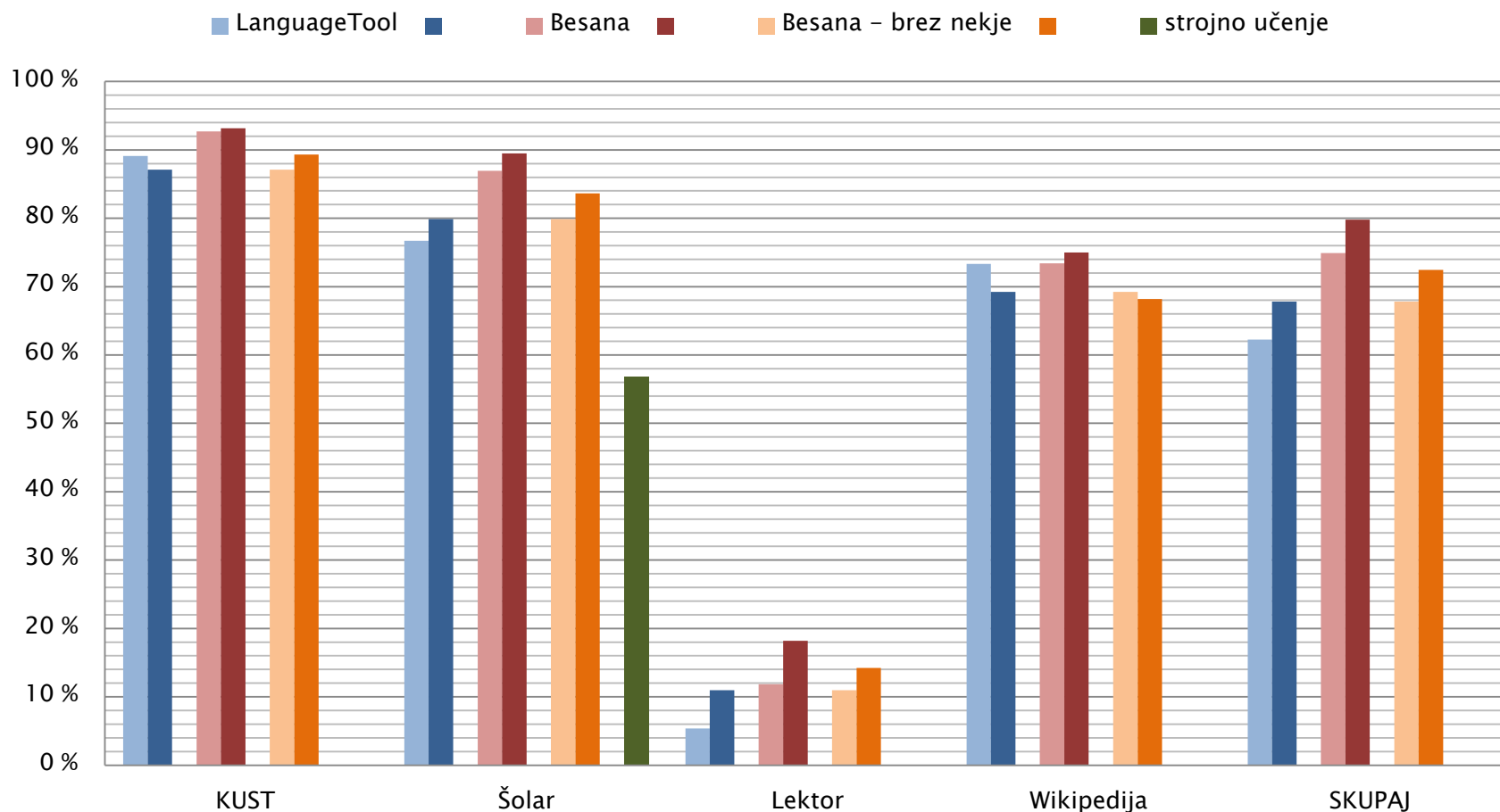
Rezultati:
povsod okoli 90 %

Priklic pri manjkajočih vejicah



Rezultati:
Besana 63 %
LanguageTool 45 %

Natančnost pri manjkajočih vejicah

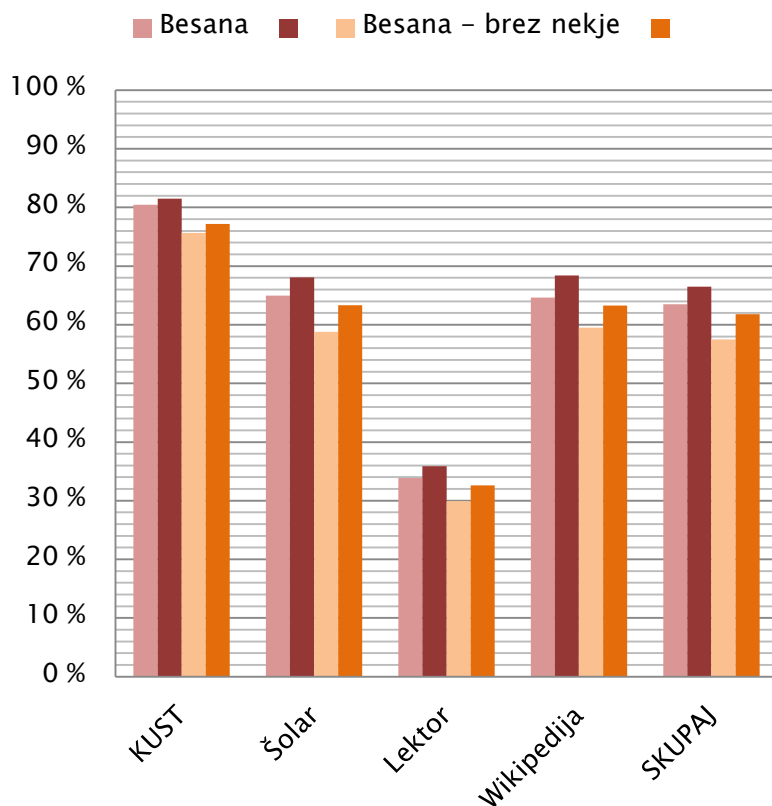


Rezultati:
Besana 80 %
LanguageTool 68 %

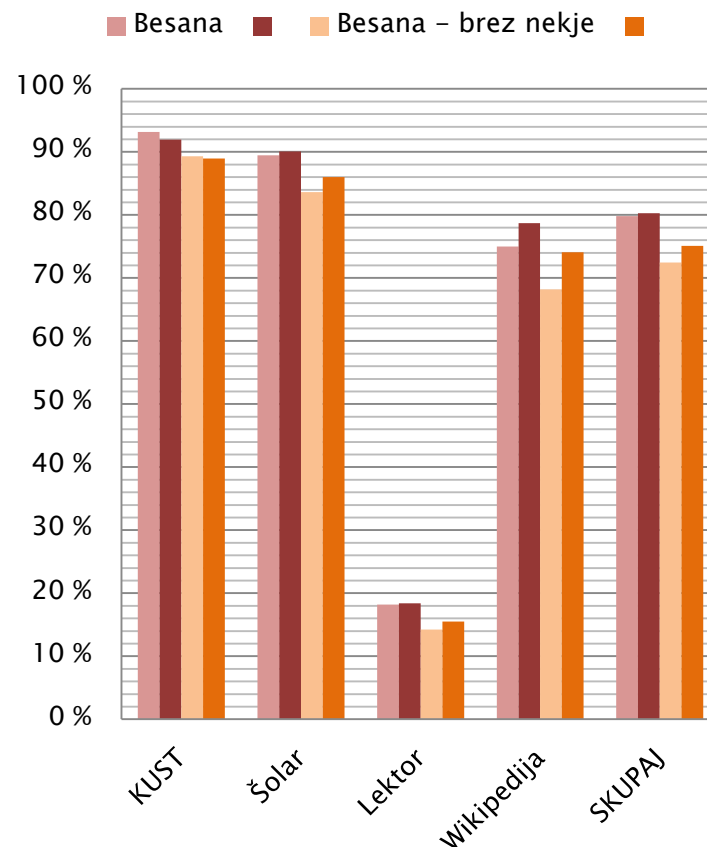
Naknadne izboljšave Besane

Natančnost pri manjkajočih vejicah

Priklic manjkajočih vejic



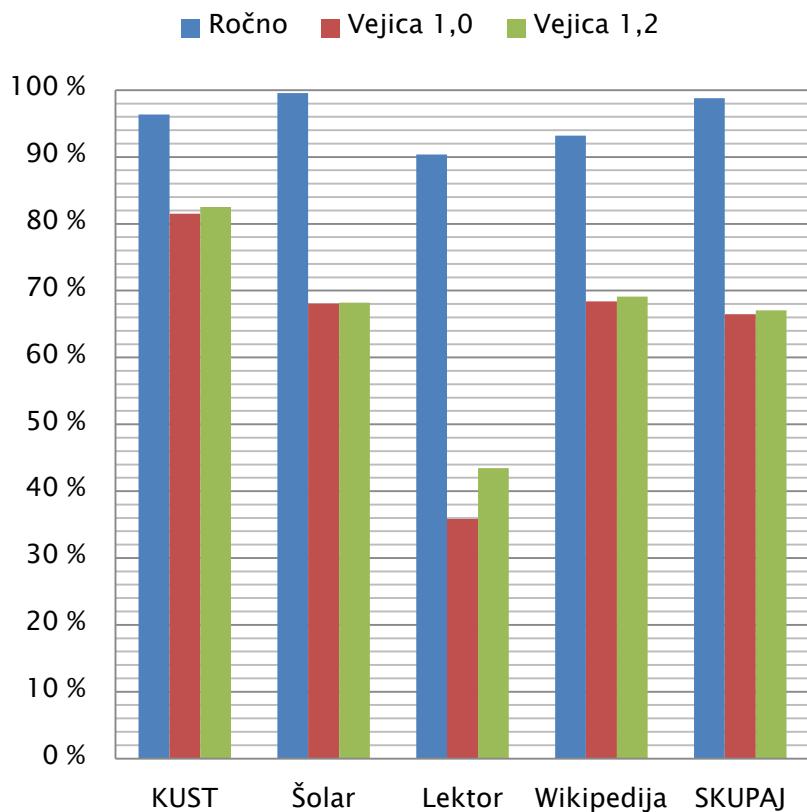
Izboljšava:
63,5 % na 66,5 %
57,5 % na 61,8 %



Izboljšava:
79,8 % na 80,3 %
72,5 % na 75,1 %

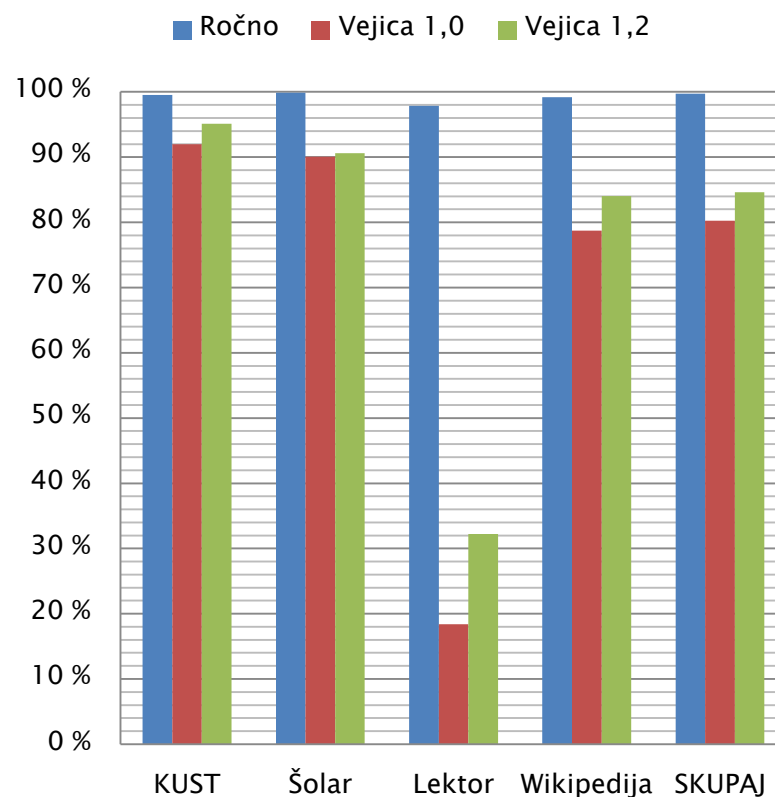
Popravljanje zbirke primerov

Priklic manjkajočih vejic



Izboljšava Lektor:
35,9 % na 43,4 %

Natančnost pri manjkajočih vejicah



Izboljšava Lektor:
18,4 % na 32,2 %

Preostale napake

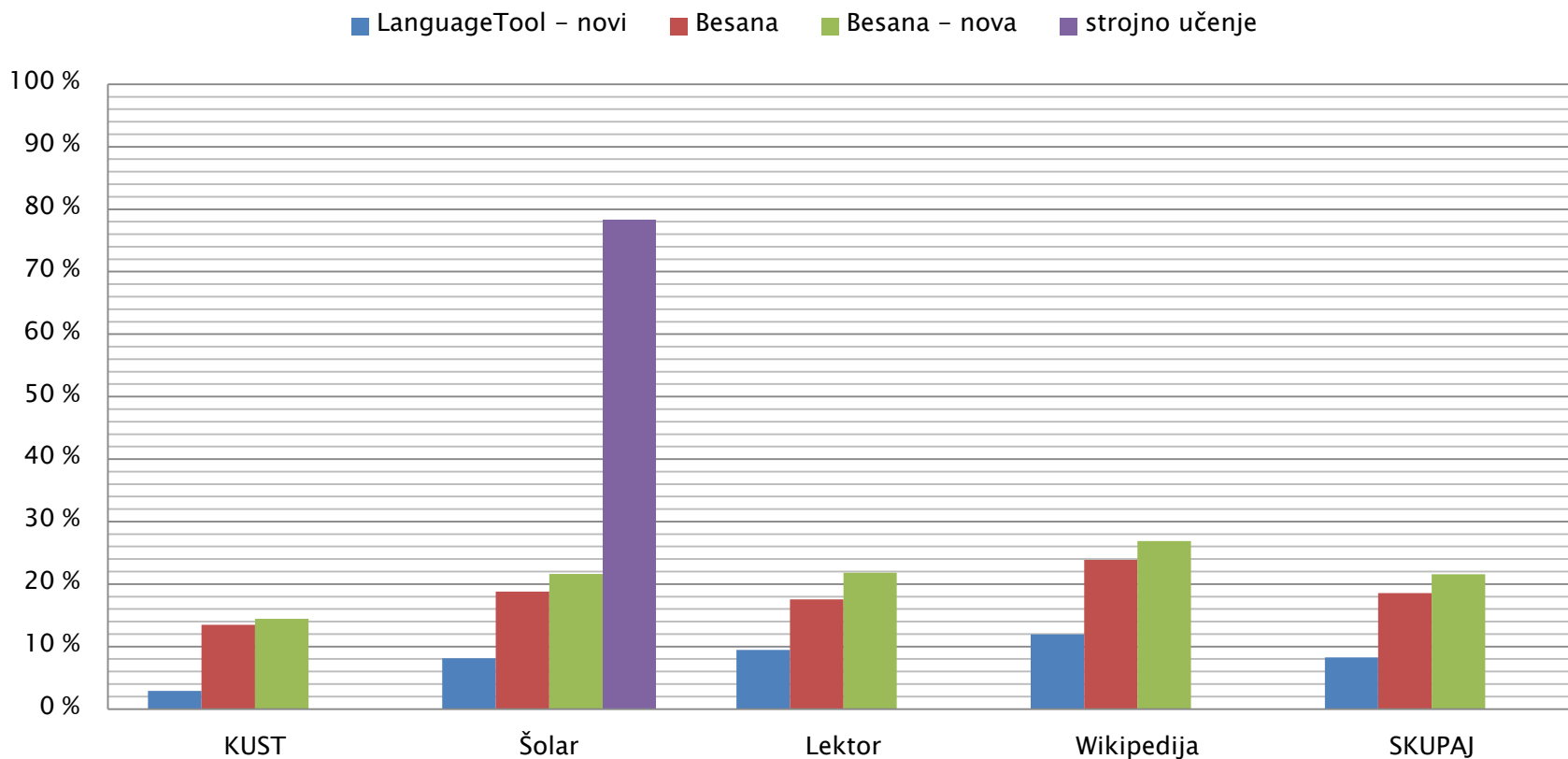
- ▶ težave zaradi drugih napak v besedilu
 - Tu gre za družbeno raznolikost, naprimer revščino **in bogastvom**.
 - Mlada dekleta **so** že v rosnih letih spopadajo z mnogimi ovirami, ki jih morajo premagati.
- ▶ vejice pred in
 - Zgodba se prične odvijati doma, in te že na začetku privede do revščine glavnega junaka.
- ▶ težave analizatorja
 - Večje **kot** so bile te spremembe, več je bilo mešanja in novosti v jezikih.

Možnosti nadaljnjega dela

- ▶ zbirka primerov
 - poprava najdenih napak, občasne objave
 - mogoče bi bilo smiselno dodati še ssj500k
- ▶ Besana
 - izboljševanje analizatorja povedi
 - izpusti (elipse; *Ena rasa ima dve nogi, štiri race pa osem*)
 - besedni red
 - “senčni” stavčni členi
 - Razen Janezu sem pošto poslal tudi Mihi.
 - hitrostna optimizacija — večnitnost
 - dodatna pravila za postavljanje vseh vejic
 - izboljševanje natančnosti

Rezultati – odvečne vejice

Priklic odvečnih vejic



Rezultati – odvečne vejice

Natančnost odvečnih vejic

