Dodatne dvoumnosti zaradi popustljivosti analizatorja pri analizi slovenskih stavkov

Peter Holozan

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

Povzetek

Zaradi neknjižnih oblik, ki se pojavljajo v besedilih, je smiselno, da analizator pri stavčni analizi predvidi, da bi lahko bila v stavku tudi kakšna od pogostih neknjižnih oblik. Ta popustljivost analizatorja pa lahko pripelje do dvoumnosti, kjer je potem možno, da v stavku je neknjižna oblika ali pa je ni in gre za drug pomen. V takih primerih je treba določiti, katera od možnosti je bolj verjetna.

Additional ambiguities caused by the Slovenian analyser's permissiveness

Due to non-standard forms occurring in texts, it makes sense for an analyser to anticipate that a sentence could contain a common non-standard form. However, the analyser's permissiveness can lead to ambiguity where it is possible to analyse a sentence either as containing a non-standard form and having one meaning or as not containing a non-standard form and having another meaning. In such cases, it is necessary to choose which is the likelier possibility.

1 Uvod

Pri stavčni analizi se pojavi problem, da se v besedilu, ki ga je treba analizirati, lahko pojavijo neknjižne oblike. Pri slovničnem pregledovalniku je zelo pomembno, da te oblike čim bolje ugotovimo, da lahko pregledovalnik nanje opozori, pa tudi pri strojnem prevajanju je zaželeno, da je čim odpornejše na neknjižne oblike in napake.

Analizator, ki ga uporabljamo v podjetju Amebis, tako v slovarju vsebuje najpogostejše neknjižne oblike (*življenski*, *nebo* namesto *ne bo*, *jest* namesto *jaz*), pri manj pogostih pa jih označevalnik ugiba sproti, če je beseda neznana (npr. vezava *ne* s povednim sedanjikom glagola). Označevalnik tudi dodaja tipično zamenjane sklone (dajalnik in mestnik) in števila (dvojina in množina).

Težava se pojavi, kadar je nek stavek mogoče analizirati z uporabo popustljivosti ali pa brez tega, ker se mora potem analizator odločiti, katere možnost je verjetnejša.

2 Namen članka

V članku bodo najprej opisani nekateri tipični neknjižni pojavi, ki se pojavljajo v slovenskih besedilih, in načini, kako jih analizator lahko upošteva.

Sledil bo opis iskanja stavkov, pri katerih pride zaradi upoštevanja neknjižnih oblik do dvoumnosti.

Te stavke bom razvrstil po značilnih skupinah in poskusil najti metode, kako izbrati analizo, ki je bolj verjetna.

Na koncu bo predstavljeno, kako pri teh primerih delujeta prevajalni sistem Presis in slovnični pregledovalnik BesAna.

3 Nekatere tipične neknjižne oblike v slovenskih besedilih

3.1 Neknjižne besede

Te besede (npr. *življenski*, *cuker*, *jest*) so rešene tako, da so vnesene v slovar s posebno oznako in povezane na ustrezne pomene. Ta način ima dodatno prednost, ker zmanjšuje verjetnost, da bi se neknjižna beseda po pomoti vnesla v slovar kot knjižna, saj je tam že vnesena, le da ima posebno oznako.

3.2 Neknjižno sklanjanje in spreganje

Najpogostejše take oblike (*zadanemo*, *otroci* za orodnik množine) so vnesene že v slovar. Dodatno označevalnik pri označevanju išče še nekatere oblike, kot so na primer pogovorne oblike opisnih deležnikov na -l (*gledu* namesto *gledal*) in podaljševanje osnove pri sklanjanju imen (*Mirota*, *Lukata*).

3.3 Zanikanje povednega sedanjika

Pogosto postaja tudi, da se povedni sedanjik glagola zanika tako, da se predpona ne- prilepi k glagolu, podobno kot je to pri pridevnikih in samostalnikih (*nevem*). To je zelo pogosto pri oblikah za prihodnjik in pogojnik glagola *biti*: *nebom*, *nebo*, *nebi*. Te oblike so za glagol *biti* vnesene že v slovar, za druge glagole pa jih pri neznanih besedah poskusi najti označevalnik.

3.4 Neknjižna uporaba povratnih svojilnih zaimkov

V knjižni slovenščini je treba v primeru, ko se svojina nanaša na osebek (izjema je splošno lastništvo (Herrity, 2000)), uporabiti povratni svojilni zaimek, kar pa se pogosto opušča (*Popravljam moj članek.*).

V teh primerih sam analizator nima težav, dodatno le označi, kadar se nepovratni svojilni zaimek ujema z osebkom, da lahko potem na podlagi tega slovnični pregledovalnik opozori uporabnika na morebitno neknjižno uporabo.

3.5 Napol vikanje

Pri napol vikanju je pomožni glagol v množini kot pri vikanju, opisni deležnik je pa v ednini kot pri tikanju (*Kdaj boste prišla?*, *Kdaj boste prišel?*).

3.6 Zanikanje s tožilnikom

Predmet v tožilniku se pri zanikanju stavka spremeni v rodilnik (pri glagolu *biti* v pomenu *obstajati* preide v

rodilnik iz imenovalnika tudi osebek (*Sosede ni doma.*)). To se pogosto opušča.

Amebisov analizator pri zanikanih stavkih sprejema tako predmete v tožilniku kot v rodilniku, vendar prve posebej označi v analizi, da lahko slovnični pregledovalnik potem ugotovi, da gre za možno neknjižno uporabo.

3.7 Neuporaba dvojine

Do tega lahko pride že pri samem številniku (*pred dvemi leti*), te neknjižne oblike so vnesene v slovar. Druga možnost pa je, da je v neknjižnem številu samostalnik oz. pridevnik (*dve ure*). Da bi bil analizator v takih primerih lahko uspešen, označevalnik vse oblike za množino podvoji, da lahko pomenijo tudi neknjižno dvojino.

3.8 Zamenjevanje mestnika z dajalnikom

Do tega prihaja pri moškem spolu: *na velikemu vrtu*. Označevalnik dajalniku ednine moškega spola pripiše še možnost, da gre za neknjižni mestnik, s čimer potem analizator uspešno analizira take primere.

3.9 Zamenjevanje nedoločnika in namenilnika

V knjižni slovenščini mora biti ob glagolih premikanja namenilnik, ob drugih glagolih pa nedoločnik. V pogovornem jeziku se nedoločnik pogosto zamenja z namenilnikom (*Moram delat.*), včasih pa zaradi hiperkorektnosti (ker se posebej trudi, da ne bi pozabil na nedoločnike) kdo uporabi tudi nedoločnik namesto namenilnika ob glagolih premikanja (*Grem pisati.*).

4 Iskanje stavkov, pri katerih pride do dvoumnosti

Vhodni korpus, ki smo ga uporabili pri iskanju stavkov, je seznam primerov, ki so jih prevajali uporabniki spletne različice slovensko-angleškega strojnega prevajalnika Presis. Ker so ti primeri popolnoma nelektorirani (in so vnašalci pogosto tudi namerno uporabljali pogovorni jezik), je med njimi zelo veliko neknjižnih oblik. Dopolnjen je še z različnimi primeri, na katere smo naleteli pri ročnem preverjanju delovanja strojnega prevajanja in iskanja slovničnih napak.

Za te primere smo se odločili, ker smo želeli v besedilih čim več neknjižnih oblik, ki se pojavljajo pri pisanju, zato smo se izognili uporabi korpusa Fida, ki sicer vsebuje tudi precej nelektoriranih besedil, vendar so tudi pri teh besedilih pisci večinoma izobraženi in se trudijo pisati čim bolj knjižno, kar zelo zmanjša število neknjižnih oblik.

Pokazalo se je, da je zaradi same zasnove analizatorja relativno zapleteno najti take pare analiz, kjer pride do dodatne dvoumnosti zaradi upoštevanja neknjižnih oblik. Najenostavnejša rešitev bi bila, da bi za vsak stavek izbrali vsakič po eno možno analizo in potem uporabili slovnični pregledovalnik. Dvoumni stavki bi bili tisti, pri katerih bi se pojavila vsaj po ena analiza z in brez pripomb pregledovalnika. Žal se je pokazalo, da je že število analiz včasih zelo veliko, zato jih analizator potem poreže. Pri zaključku analize se izgubijo tudi nekateri vmesni rezultati, zato bi bilo najlaže ponoviti kar celotno analizo. Vendar je analizator časovno precej zahteven in bi tak način dela bil zelo počasen.

Zato je bil s pomočjo pogojnega prevajanja raje dodan v analizator del kode, ki sproti ugotavlja, ali je bila v analizi uporabljena kakšna označena neknjižna oblika ali ne. Če se pri analizi stavka pojavijo analize obeh vrst, je stavek dvoumen.

Pri napol vikanju in svojilnih zaimkih na možno dvoumnost opozarja že slovnični pregledovalnik, zato smo tam uporabili kar njegov rezultat.

5 Analiza primerov

Seznam možnih dvoumnosti je bil ročno pregledan in dvoumnosti razvrščene po tipičnih skupinah.

5.1 Možna neuporaba povratnega svojilnega zaimka

Pri prvi in drugi osebi je sklepanje o manjkajočem povratnem svojilnem zaimku precej zanesljivo (*Poklical sem mojo prijateljico.*). Problematične dvoumnosti se pojavijo le pri pogojnih stavkih z izpuščenim osebkom, kjer ne moremo biti prepričani o osebi.

Rad bi opravičil mojega brata.
Rad vam bi predstavil mojega soseda Ivana.
Rad bi vam predstavil moj hobi.
V spomin na mojega deda bi ga želel obnoviti.
Na koncu bi se rad opravičil zaradi moje angleščine.

Primer 1. Primeri pogojnih stavkov z možno neuporabo povratnega svojilnega zaimka

Če bi se zanesli na to, da v besedilu ni neknjižnih uporab, bi lahko v teh stavkih sklepali na to, da osebek ni v prvi osebi. Kot pa kaže primer 1, je največkrat bolj verjetno, da manjka povratni svojilni zaimek.

Nekoliko drugačen je položaj pri tretji osebi.

Janez je ljubil njegovo ženo.
Okrog njenega grebena je krožila cela jata vodnih ptičev.
Irski raziskovalci so raziskali pomembnost zajtrka pri njihovih študentih
Vstopila je v njene hlačke.
Uporabnik lahko vnaprej definira opozorilni klic na njegovo telefonsko številko.
lagal ji je glede njegovega premoženja
da bo tako dober kot njegov oče

Primer 2. Primeri svojilnega zaimka in osebka v tretji osebi

V primeru 2 se vidi, da je pri tretji osebi veliko več možnosti, da je stavek res napisan knjižno (kar pa bi bilo največkrat mogoče ugotoviti šele iz sobesedila). To je potem treba upoštevati tudi pri slovničnem pregledovalniku, ki mora v teh primerih veliko manj opozarjati kot pri prvi in drugi osebi.

5.2 Napol vikanje

Pri napol vikanju za ženski spol pride do dvoumnosti zaradi tega, ker se oblika prekrije z drugo osebo dvojine za srednji spol. Na srečo pa je (razen morda v kakšni znanstveni fantastiki) zelo malo verjetno, da bi kdo

ogovarjal skupino bitij srednjega spola, tako da se da zelo zanesljivo sklepati, da gre vedno za napol vikanje.

Kam ste šla potem?
ki ste mi ga sporočila
S katero pošiljko ste poslala sete.
Nič mi niste pisala.
Ali ste že kdaj obiskala mladinski hotel?
Boste ponudbo potrdila?
vi ste dolgočasna

Primer 3. Primeri napol vikanja

Vendar pa tudi tukaj lahko pride do dodatnih dvoumnosti:

Vi ste ženska.

Primer 4. Primer napačno ugotovljenega napol vikanja

V primeru 4 pride do dvoumnosti zaradi tega, ker besedo *ženska* prepozna tudi kot pridevnik, kar bi pomenilo napol vikanje. V teh primerih mora analizator zato dati prednost analizi s samostalnikom.

5.3 Analize z odvečno dvojino

Zaradi dodatne dvojine v označevalniku se analize lahko povečajo še za (neknjižno) dvojino. Pri imenovalniku do tega ne pride, ker se mora ujemati še z glagolom, se pa to zgodi pri predložnih zvezah.

Pred leti me je povozil traktor.

Primer 5. Nepotrebno dodana dvojina

V primeru 5 tako analizator najde tudi možnost, da bi se to zgodilo pred dvema letoma.

Te možnosti skoraj vedno odvečne, kadar pa niso, je to nemogoče ugotoviti brez sobesedila. Na srečo pri prevajanju v angleščino potem to dvoumnost izgubimo, slovnični pregledovalnik pa tudi ne opozarja nanjo.

5.4 Dvoumnost zaradi zanikanja s tožilnikom

Popustljivost analizatorja pri zanikanju s tožilnikom prinese težave pri ženskem spolu.

Ne gledam slike.

Primer 6. Zanikanje s tožilnikom ali ne

Pri primeru 6 je tako število slik odvisno od tega, ali je predmet v rodilniku ali v tožilniku. Tega brez sobesedila ni mogoče ugotoviti, Amebisov analizator zato da prednost knjižni obliki.

Možna rešitev te težave bi bila, da bi program pogledal celotno besedilo in preštel vse nedvoumno zanikane predmete in iz tega sklepal, kaj pisec večkrat uporablja.

5.5 Nebo

Lepljenje nikalnice *ne* ob obliko glagola *biti* za prihodnjik je vedno pogostejši pojav. V večini primerov jo lahko opazi že črkovalnik, zaplete pa se pri tretji osebi ednine, kjer se neknjižna oblika prekrije z relativno pogostim samostalnikom *nebo*.

To nebo poceni.
Oblačno nebo.
lepo nebo v odboju
oljnato nebo
Nad nami je nebo.
nebo v bolečini

Primer 7. Dvoumnosti z besedo *nebo*

V večini primerov je pravi pomen nebo, nekatere primere pa je mogoče razrešiti tudi z dodatnimi pravili. Tako ima glagol *poceniti* označeno, da je vezava s predmetom *nebo* malo verjetna, tako da pride v tem primeru na prvo mesto glagol *biti*.

Zamenjava pri uporabnikih nebo problem.

Primer 8. Zanimiva dvoumnost

V primeru 8 je stavek, kjer na prvi pogled ni bilo videti, da bi analiza lahko bila dvoumna. Pokaže pa se, da je mogoče stavek razumeti kot: *Nebo zamenjava problem pri uporabnikih.*, kjer *zamenjava* oblika glagola *zamenjavati* (drugi možnosti sta še glagol *zamenjati* in samostalnik *zamenjava*). Možna rešitev za ta stavek bi bila, da bi se zmanjšale verjetnosti za analize, ki vsebujejo glagol *zamenjavati*, ki je relativno redek.

5.6 Jest

V pogovornem jeziku se *jest* pogosto pojavlja kot nadomestek za *jaz*. Dvoumnosti povzroča dejstvo, da je beseda tudi namenilnik glagola *jesti*, kar povzroči, da zelo pogosto postanejo dvoumni stavki, ki vsebujejo glagole premikanja.

Jest grem.
Jest grem na ples.
jest grem domov
grem samo jest
grem malo jest
jest grem pa v kvalifikacije
šel sem jest domov

Primer 9. Dvoumnosti z besedo *jest*

Te dvoumnosti je žal brez sobesedila zelo težko razrešiti. Tudi tukaj bi bil verjetno pravi način to, da bi analizator iz okoliškega besedila ocenil, koliko pisec uporablja pogovorni jezik.

5.7 Dvoumnosti zaradi domnevnih namenilnikov namesto nedoločnikov

Kar nekaj pogostih besed se prekriva z možnimi namenilniki (*spet*, *izpit*, *past*, *rit*, *pet*, …). V kombinaciji z glagoli, ki se lahko vežejo tudi nedoločniki (ob modalnih glagolih, kjer je dvoumnosti manj, ker se običajno ne vežejo s čim drugim, sta taka primera pogosto *imeti* (*Imam pisati članek.*) in *biti* (*Delati je naporno.*, *Živeti je pesem.*)).

Danes imamo spet računalništvo.
Jutri moram spet v službo
Jutri imam izpit.
Daj pet sto.
Ima veliko rit.
moja prijateljica ima božansko rit
spet imam moje zobe
tu je past domovine

Primer 10. Možni neknjižni namenilniki

V veliki večini primerov je prava analiza tista brez namenilnika (na splošno se *imeti* in *biti* ne vežeta pogosto z nedoločnikom), tako da je znižanje verjetnosti analizam z njim dovolj zanesljiva metoda za ugotavljanje prave analize.

Ta statut je veljavno sprejet.
Dobro razvit je turizem.

Primer 11. Prekrivanje namenilnika in deležnika na -t

Včasih se z namenilnikom prekrijejo tudi deležniki na -t. Tudi v teh primerih je verjetnost, da gre res za neknjižno obliko, po navadi majhna.

5.8 Mam in mamo

V pogovornem jeziku se pojavlja krajšanje glagola *imeti* v *meti*. Do dvoumnosti lahko pri povednem sedanjiku pride zaradi prekrivanja z oblikami samostalnika *mama*.

poškodovan prst mam
štiri ure mam časa
in mamo Ivano

Primer 12. Pogovorna oblika *imeti*

Ugotavljanje pravega pomena je tukaj precej zoprno. Možna rešitev je, da se prepove, da je *mama* desni prilastek samostalnika *ura*, kar pa razreši le delček primerov.

5.9 Radio in radij

V pogovornem jeziku se samostalnik *radio* pogosto piše kot *radijo*. Na to tudi ne (z izjemo imenovalnika in tožilnika ednine) opozarja črkovalnik, ker se oblike prekrivajo z oblikami samostalnika *radij*.

popravilo radijev
Težava je pri rezanju radija.
Pri krivljenju kabla je upoštevati minimalne radije krivljenja.
Oglašujemo preko radija in časopisov.
veliki radiji
nekotirani notranji radiji so taki
kar se trenutno vrti na radiju
Vsi nekotirani radiji.
dobre dogovore smo sklenili tudi z radiji in časopisnimi hišami
Zanima me, če si lahko na svojo spletno stran dodam povezavo do vašega radija?
Že cel mesec oglašuje po radiju.
slišal sem na radiju

Primer 13. Prekrivanje oblik radia in radija

Te dvoumnosti je zelo težko razrešiti, edini možen način je z uporabo pomenov. Tako se na primer da dodati pravilo, da se *popravilo* ne veže s pomenoma *radij (element)* in *polmer*.

5.10 Vso

Relativno pogosta je neknjižna uporaba oblike *vso* namesto *vse* za tožilnik ednine srednjega spola pridevniškega zaimka *ves*.

V prostorih je vso pohištvo
, da ponudi vso bogastvo vonjav.
ruši vso dosedanje delo

Primer 14. Dvoumnosti z besedo *vso*

Težavo povzroči to, da je *vso* lahko tudi posamostaljen tožilnik ednine ženskega spola. Stavke iz primera 14 analizator zato razume tudi tako, da je *vso* v njih predmet v tožilniku, resnični predmet pa osebek.

Vsi trije primeri, ki so bili najdeni v vzorcu, so taki, da bi bilo bolje dati prednost analizi, kjer je v besedilu neknjižna oblika. Delno bi se dalo to reševati po posameznih primerih glagolov (ni posebno verjetno, da bi pohištvo kaj pojedlo), smiselno bi bilo pa preizkusiti tudi splošnejše pravilo, da v primerih, kjer besedi *vso* takoj sledi osebek v ednini srednjega spola, zmanjšamo verjetnost tej analizi. Nadaljnje preizkušanje na daljših primerih bo pokazalo, če je tako pravilo dovolj zanesljivo za uporabo.

6 Primeri uporabe

Analizator v prevajalnem sistemu Presis že zna razrešiti nekatere od zgoraj naštetih dvoumnosti.

Vhod	Izhod
To nebo poceni.	This won't be cheap.
Oblačno nebo.	Cloudy sky.
Jest grem.	I go to eat.
Jest gledam televizijo.	I watch television.

Tabela 1: Primeri strojnih prevodov prevajalnika Presis

Tudi slovnični pregledovalnik BesAna uporablja isti analizator. Zato tudi ta v prvem primeru takoj svetuje, da bi bilo bolj knjižno napisati *To ne bo poceni.*

7 Literatura

Herrity, Peter, 2000. *Slovene: A Comprehensive Grammar*. Routledge.

Žagar, France, 1987. *Pouk slovenske slovnice in pravopisa v višjih razredih osnovne šole.* Založba Obzorja Maribor, prvi natis.

Žagar, France, 1991. *Slovenska slovnica in jezikovna vadnica*. Založba Obzorja Maribor, šesta dopolnjena in razširjena izdaja.