“Poišči in zamenjaj” v pregibnih jezikih

Primož Peterlin*, Agnes Pisanski†, Tomaž Erjavec‡

*Inštitut za biofiziko, Medicinska fakulteta, Univerza v Ljubljani
Lipičeva 2, SI-1000 Ljubljana, Slovenija
primoz.peterlin@biofiz.mf.uni-lj.si

†Oddelek za prevajanje in tolmačenje, Filozofska fakulteta, Univerza v Ljubljani
Aškerčeva 2, SI-1000 Ljubljana
agnes.pisanski@guest.arnes.si

‡Odsek ze inteligentne sisteme E8, Institut Jožef Stefan
Jamova 39, SI-1000 Ljubljana, Slovenija
tomaz.erjavec@ijs.si

Povzetek

Funkcija “Poišči in zamenjaj” (“Search and replace”), vdelana v večino urejevalnikov besedil, je zares uporabna samo v nepregibnih jezikih, kot je angleščina. Opisana je izvedba te funkcije za pregibne jezike, ki uporablja sezname oblikoslovno označenih besednih oblik. Uporabljeni sistem oblikoslovnega označevanja izvira iz projekta MULTEXT-East. Analiziramo primere, v katerih lahko program sam opravi zamenjavo, in navedemo tiste, v katerih je potrebna pomoč pisca. Kot posebni zgled pregibnega jezika je v prispevku uporabljena slovenščina.

1. Uvod

Funkcija “Poišči in zamenjaj” (angl. Search and replace), ki je del vseh sodobnih urejevalnikov besedil, je ena od velikih prednosti urejevalnikov besedil pred pisalnimi stroji. Žal pa ta velika prednost pride zares do izraza samo pri nepregibnih jezikih, kot je angleščina.

V čem je težava? V pregibnih jezikih lahko posameznemu geslu ustreza več deset besednih oblik. Denimo, da želimo povsod v besedilu, kjer je govora o ljubkem kužku, nadomestiti “ljubek kužek” s “prisrčen kužek”. Enostavno: vse nize “ljubek” zamenjamo s “prisrčen”, pa je! Ne, to je šele začetek. Prečesati moramo vse sklone (npr. “ljubkega” → “prisrčnega”) v vseh številih (npr. “ljubkih” → “prisrčnih”), in ne nazadnje, morda smo kje uporabili tudi določno obliko (“ljubki” → “prisrčni”). Ob tolikem delu za zamenjavo ene same besede se človek že vpraša, ali imajo urejevalniki besedil res *toliko* prednosti pred pisalnimi stroji.

Zakaj torej funkcija “Poišči in zamenjaj” v urejevalnikih besedil ne deluje tako, kot bi si želeli? Razlogov je več. Gotovo najpomembejši med njimi je ta, da se je računalništvo razvilo v angleškem govornem področju, kjer tega problema preprosto ni. Če odštejemo nepravilne glagole in izjeme pri tvorjenju množinske oblike, lahko vse besedne oblike v angleščini tvorimo iz osnove – ki je v angleščini praviloma kar enaka geslu – s končnicama -s in -ed. Z nekaj sreče lahko, kadar nobena od besed ni na seznamu izjem, tako enostavno zamenjamo osnovo (npr. *erase* → *delete*), pa bodo tvorjene oblike ostale pravilne (*erase*s → *delete*s, *erase*d → *delete*d).

Razmeroma nedavno, v večji meri šele zadnjih deset let, so programske hiše začele pripravljati tudi krajevno prilagojene (lokalizirane) različice programov. S tem pa izdelovalcem programske opreme počasi prihaja v zavest, da se struktura drugih jezikov lahko precej razlikuje od angleške.

S tem smo prišli še do drugega razloga. Za pregibno iskanje mora imeti program vgrajenega dovolj jezikovnega znanja za izbrani jezik – bodisi kot tvorbena pravila za pregibanje bodisi kot seznam besednih oblik z gesli (angl. lemma) in oblikoslovnimi značilkami[1] (angl. morphosyntactic description). Razvoj korpusnega jezikoslovja je ob hkratnem spuščanju cen in večanju zmogljivosti pomnilniških sredstev pripomogel, da smo danes prišli na prag možne rešitve po drugi od omenjenih poti. Pri tem nam je v veliko pomoč prvi javno dostopni seznam besednih oblik za slovenščino, ki je bil pripravljen v okviru projekta MULTEXT-East (Erjavec and Ide, 1998).

2. Klasifikacija MULTEXT-East

V klasifikaciji MULTEXT-East (Tufis et al., 1998) so besedne oblike taksonomsko razvrščene. Na najvišji ravni so razdeljene v 14 kategorij, ki ustrezajo posameznim besednim vrstam: samostalnikom (angl. nouns), glagolom (angl. verbs), pridevnikom (angl. adjectives), prislovom (angl. adverbs), zaimkom (angl. pronouns), predlogom (angl. adpositions), veznikom (angl. conjunctions), členkom (angl. particles), medmetom (angl. interjections), številnikom (angl. numerals), okrajšavam (angl. abbreviations), členom (angl. articles) ali določilnikom (angl. determiners), kot zadnjo kategorijo pa pušča neuvrščeno (angl. residual). V nadaljnji obravnavi se bomo omejili na besedne vrste, ki nastopajo v slovenščini.

Seznam besednih oblik MULTEXT-East sestavljajo trojke besedne oblike, gesla in oblikoslovne značilke:

```
stola   stol    Ncmsg
```

Za opis oblikoslovnih značilk bomo uporabili formalizem *regularnih izrazov* (Friedl and Oram, 1997), s katerimi lahko učinkovito opišemo vzorce znakov.

[1] Ponekod tudi *oblikoskladenjske* oznake.

2.1. Samostalniki

Samostalnike v klasifikaciji MULTEXT-East opišemo z naslednjim regularnim izrazom:

```
N[cp][mfn][sdp][ngdali](--[ny])?
```

Pri tem `N` pomeni samostalnik (angl. noun), ki mu sledi določilo vrste. To določa znak `c` (angl. common – občno ime) ali `p` (angl. proper – lastno ime). Temu sledi oznaka spola – eden od znakov iz množice `m` (angl. masculine – moški spol), `f` (angl. feminine – ženski spol) ali `n` (angl. neuter – srednji spol). Oznaki spola sledi oznaka števila: eden od znakov iz množice `s` (angl. singular – ednina), `d` (angl. dual – dvojina) ali `p` (angl. plural – množina). Oznaki spola sledi oznaka sklona: `n` (angl. nominative – imenovalnik), `g` (angl. genitive – rodilnik), `d` (angl. dative – dajalnik), `a` (angl. accusative – tožilnik), `l` (angl. locative – mestnik) ali `i` (angl. instrumental – orodnik). Neobvezna končna oznaka, ki sledima dvema določiloma, ki v slovenščini ne nastopata in sta zato označena s črtico `-`, določa živost samostalnika. V poštev pride samo pri tožilniku ednine samostalnikov moškega spola (prim. stol/gospoda). Možni izbiri sta `y` (yes, da) ali `n` (no, ne).

2.2. Glagoli

Glagole v klasifikaciji MULTEXT-East opišemo z regularnim izrazom:

```
V[moc][imcnpu][-pfs][-123][-spd]\
    [-mfn][-ap][ny]?(-----[pe])?
```

Oznaka `V` označuje glagole (angl. verb). Sledi določilo vrste, ki je lahko `m` (angl. main) za polnopomenski glagol, `o` (angl. modal) za naklonski (npr. *zmorem*) ali `c` (angl. copula) za vezni glagol (npr. pozno *je*). Sledi oznaka glagolske oblike: `i` (angl. indicative) za povednik, `m` (angl. imperative) za velelnik, `c` (angl. conditional) za pogojnik, `n` (angl. infinitive) za nedoločnik, `p` (angl. participle) za deležnik[2] ali `u` (angl. supine) za namenilnik.

Oznaki glagolske oblike sledi oznaka časa: `p` (angl. present) za sedanjik, `f` (angl. future) za prihodnjik (samo oblike glagola biti: bom, boš, bo, bova, bosta, bomo, boste, bodo in bojo), `s` (angl. past) za nesedanjik ali `-`, kadar oznaka časa nima pomena (pri nedoločni ali namenilni glagolski obliki). Tej sledi oznaka osebe: `1` za prvo, `2` za drugo in `3` za tretjo osebo, oziroma `-`, kadar oznaka osebe nima pomena (pri nesedanjiku).

Oznaki osebe sledi oznaka števila: `s` (angl. singular) za ednino, `d` (angl. dual) za dvojino, `p` (angl. plural) za množino ali `-`, kadar oznaka števila nima pomena (pri nedoločni ali namenilni glagolski obliki). Tej sledi oznaka spola: `m` (angl. masculine) za moški spol, `f` (angl. feminine) za ženski, `n` (angl. neuter) za srednji ali `-`, kadar oznaka spola nima pomena (pri nedoločni, velelni in namenilni glagolski obliki polnopomenskih glagolov).

Oznaki spola sledi oznaka načina: `a` (angl. active) za tvorni, `p` (angl. passive) za trpni način ali `-`, kadar oznaka nima pomena (pomen ima samo pri velelni glagolski obliki polnopomenskih glagolov in deležnike). Oznaki načina sledi neobvezna oznaka nikalnosti: `y` (angl. yes) za zanikane ali `n` (angl. no) za nezanikane glagole (pomen ima samo pri povedni glagolski obliki), tej pa lahko po petih oznakah, ki v slovenščini ne pridejo v poštev in so zato označene s črticami `-`, sledi neobvezna oznaka glagolskega vida: `p` (angl. progressive) za nedovršni ali `e` (angl. perfective) za dovršni vid.

[2] Deležniki na -l oziroma na -n/-t.

2.3. Pridevniki

Pridevniki so v klasifikaciji MULTEXT-East opisani z regularnim izrazom:

```
A[fso][pcse][mfn][spd]\
    [ngdali][-ny]?(-[ny])?
```

Pri tem pomeni `A` (angl. adjective) pridevnik, ki mu sledi oznaka vrste: `f` (angl. qualificative) za kakovostni,[3] `s` (angl. possessive) za svojilni ali `o` (angl. ordinal) za vrstni pridevnik. Oznaki vrste sledi oznaka stopnje: `p` (angl. positive) za osnovnik, `c` (angl. comparative) za primernik, `s` (angl. superlative) za presežnik ali `e` (angl. elative) za elativ.

Temu sledi oznaka spola – eden od znakov iz množice `m` (angl. masculine – moški spol), `f` (angl. feminine – ženski spol) ali `n` (angl. neuter – srednji spol). Oznaki spola sledi oznaka števila: eden od znakov iz množice `s` (angl. singular – ednina), `d` (angl. dual – dvojina) ali `p` (angl. plural – množina). Oznaki spola sledi oznaka sklona: `n` (angl. nominative – imenovalnik), `g` (angl. genitive – rodilnik), `d` (angl. dative – dajalnik), `a` (angl. accusative – tožilnik), `l` (angl. locative – mestnik) ali `i` (angl. instrumental – orodnik).

Oznaki sklona sledi neobvezna oznaka določnosti: `y` za določno, `n` za nedoločno obliko ali `-`, če oznaka nima pomena. Neobvezna končna oznaka, ki sledi oznaki, irelevantni za slovenščino in zato označeni s črtico `-`, določa živost pridevnika: `y` za živo ali `n` za neživo obliko.

2.4. Prislovi

Prislovi so v klasifikaciji MULTEXT-East opisani z regularnim izrazom:

```
Rg[pcse]
```

Pri tem oznaka `Rg` pomeni splošne prislove. V slovenščini drugih ni, zato je vrsta prislova vedno označena z `g` (angl. general). Oznaki vrste sledi oznaka stopnje: `p` (angl. positive) za osnovnik, `c` (angl. comparative) za primernik, `s` (angl. superlative) za presežnik ali `e` (angl. elative) za elativ.

2.5. Zaimki

Zaimke v klasifikaciji MULTEXT-East opišemo z regularnim izrazom:

```
P[pdisqrxzg][-123][-mfn][-spd][-ngdali]\
    [-spd][-mfn][-nyb][-ps][nar](-[ny])?
```

[3] Klasifikacija MULTEXT-East imenuje *kakovostne* pridevnike vse lastnostne pridevnike, torej kakovostne in merne.

Pri tem vodilna oznaka P (angl. pronoun) pomeni zaimek, ki mu sledi oznaka vrste:[4] p (angl. personal) za osebni, d (angl. demonstrative) za kazalni, i (angl. indefinite) za nedoločnostni, s (angl. possessive) za svojilni, q (angl. interrogative) za vprašalni, r (angl. relative) za oziralnostni, x (angl. reflexive) za povratni, z (angl. negative) za nikalni ali g (angl. general) za celostni.

Oznaki vrste pri osebnih in svojilnih zaimkih sledi oznaka osebe: 1 za prvo, 2 za drugo in 3 za tretjo osebo, ali - pri ostalih vrstah, kjer oznaka osebe nima pomena. Tej sledi oznaka spola: m (angl. masculine) za moški spol, f (angl. feminine) za ženski, n (angl. neuter) za srednji ali -, kadar oznaka spola nima pomena (povratni zaimki *se*, *si* ipd.; kazalni zaimki *le-tega*, *onega* ipd.; nedoločni zaimki *koliko*, *čigar* ipd.; vprašalni zaimki *koliko*, *kaj* ipd.; oziralnostni zaimki *kolikor*, *čigar* ipd.; nikalni zaimki *nič*, *noben* ipd. ter celostni zaimki *vsakdo*, *vsakogar* ipd.).

Oznaki spola sledi oznaka števila: s (angl. singular) za ednino, d (angl. dual) za dvojino, p (angl. plural) za množino ali -, kadar oznaka števila nima pomena (isti zgledi kot pri oznaki spola). Tej sledi oznaka sklona: n (angl. nominative – imenovalnik), g (angl. genitive – rodilnik), d (angl. dative – dajalnik), a (angl. accusative – tožilnik), l (angl. locative – mestnik), i (angl. instrumental – orodnik) ali – pri vseh vrstah zaimkov razen osebnih in svojilnih – tudi -, kadar oznaka sklona nima pomena.

Sledita oznaki števila in spola svojine. Prva je relevantna samo za svojilne zaimke, druga le za tretji sklon ednine svojilnih zaimkov (*njegov*, *njen*). Oznake so enake kot pri oznaki števila in spola. Sledi oznaka naslonke, ki je pomembna samo pri povratnih (*se*, *sebe*) in osebnih zaimkih (*njega*, *ga*). Oznaki naslonke sledi oznaka nanašanja, ki ima pomen edino pri povratnih zaimkih. Tam lahko zavzame vrednosti p (angl. personal) za osebno nanašanje (*se*) ali s (angl. possesive) za svojilno (*svoj*). Naslednja oznaka je oznaka skladenjske funkcije:[5] n (angl. nominal) za samostalniško, a (angl. adjectival) za pridevniško ali r (angl. adverbial) za prislovno funkcijo. Za oznako, ki v slovenščini nima pomena in je zato označena s črtico -, sledi zadnja oznaka, ki določa živost zaimka: y za živo ali n za neživo obliko.

2.6. Predlogi

Predloge v klasifikaciji MULTEXT-East opišemo z regularnim izrazom:

```
Sp[sc][gdali]?
```

Vsi predlogi v slovenščini se začenjajo z oznako Sp. Oznaka S označuje splošnejšo kategorijo "adposition", ki vključuje tako predloge (angl. prepositions, oznaka vrste p) kot tudi postpozicija (angl. postpositions). Zgleda za slednjo sta denimo "ago" in "away" v angleščini.

Oznaki vrste sledi oznaka sestavljenosti: s (angl. simple) za enostavne predloge (*z*, *pod*, *onkraj* ipd.) ali c (angl. compound) za priponske (*predenj*, *zase* ipd.). Oznaki sestavljenosti pri enostavnih predlogih sledi še oznaka sklona: g (genitive – rodilnik), d (dative – dajalnik), a (accusative – tožilnik), l (locative – mestnik) ali i (instrumental – orodnik).

2.7. Vezniki

Veznike v klasifikaciji MULTEXT-East opišemo z regularnim izrazom:

```
C[cs][sc]
```

Oznaka C (angl. conjunction) označuje kategorijo veznikov. Sledi ji oznaka vrste: c (angl. coordinating) za priredne veznike (*in*, *saj*, *vendar* ipd.) ali s (angl. subordinating) za podredne (*ker*, *če*, *najsi* ipd.). Zadnja oznaka označuje obliko: s (angl. simple) za enobesedne ali c (angl. compound) za večbesedne (*kljub temu da*, *medtem ko*).

2.8. Členki

Členke v klasifikaciji MULTEXT-East opišemo z regularnim izrazom:

```
Q
```

Oznaka Q označuje kategorijo členkov (angl. particle).

2.9. Medmeti

Medmete v klasifikaciji MULTEXT-East opišemo z regularnim izrazom:

```
I
```

Oznaka I (angl. interjection) označuje kategorijo medmetov.

2.10. Števniki

Števnike v klasifikaciji MULTEXT-East opišemo z regularnim izrazom:

```
M[coms][mfn][spd][ngdali][drl]\
    [-yn]?(--[ny])?
```

Oznaka M (angl. numerals) označuje kategorijo števnikov. Njej sledi oznaka vrste: c (angl. cardinal) za glavne (*en*, *dva*), o (angl. ordinal) za vrstilne (*tretji*, *četrti*), m (angl. multiple) za množilne (*peteren*, *šestkraten*) ali s (angl. special) za druge[6] številke (*sedmer*, *osmero*).

Temu sledi oznaka spola – m (angl. masculine) za moški spol, f (angl. feminine) za ženski ali n (angl. neuter) za srednji spol. Oznaki spola sledi oznaka števila: s (angl. singular) za ednino, d (angl. dual) za dvojino ali p (angl. plural za množino. Oznaki spola sledi oznaka sklona: n (nominative – imenovalnik), g (genitive – rodilnik), d (dative – dajalnik), a (accusative – tožilnik), l (locative – mestnik) ali i (instrumental – orodnik).

Naslednja oznaka označuje zapis: d (angl. digit) za številčni zapis (*1950*), r (angl. roman) za zapis z rimskimi

[4] Klasifikacija se nekoliko razlikuje od standardne (Toporišič, 1992): nedoločni, poljubnostni, mnogostni, istostni in drugostni zaimki so zbrani v kategorijo *nedoločnostni*; oziralni in oziralno poljubnostni pa skupaj v kategorijo *oziralnostni*.

[5] Za razliko od klasifikacije MULTEXT-East v ustaljeni (Toporišič, 1992) slovenski slovnici zaimki ne nastopajo kot vrhnja kategorija, pač pa so bodisi del širše skupine *samostalniških besed* (kamor spadajo samostalniki in samostalniški zaimki) bodisi *pridevniških besed* (te skupaj s pridevniki tvorijo še pridevniški zaimki, deležniki in števniki).

[6] Tu so mišljeni *ločilni* števniki.

številkami (*MCML*) ali `l` (angl. letter) za besedni zapis (*tisoč devetstopetdeset*). Oznaki zapisa sledi neobvezna oznaka določnosti: `y` za določno ali `n` za nedoločno obliko. Neobvezna končna oznaka, ki sledi oznakama, irelevantni za slovenščino in zato označenima s črtico `-`, določa živost števnika: `y` za živo ali `n` za neživo obliko.

2.11. Okrajšave

Okrajšave v klasifikaciji MULTEXT-East opišemo z regularnim izrazom:

```
Y
```

Oznaka `Y` označuje kategorijo okrajšav (angl. abbreviation).

3. Opis algoritma

Algoritem za pregibni "poišči in zamenjaj" lahko preprosto opišemo takole: preberemo izvirno in ciljno besedno in zanju ugotovimo geslo in oblikoslovno značilko. Če je iz oblikoslovnih značilk razvidno, da ni mogoča pregibna zamenjava, nadaljujemo z navadno zamenjavo nizov. Na navadno zamenjavo nizov preidemo tudi, kadar bodisi izvirne bodisi ciljne besede ni na seznamu in torej oblikoslovnih značilk ne moremo primerjati.

Če pa sta obe oblikoslovni značilki dostopni in dovoljujeta pregibno zamenjavo, pa iz izvirnega gesla sestavimo iskalni niz, ki pokrije vse besedne oblike izvirne besede. Ko najdemo katero od besednih oblik, jo nadomestimo z besedo, generirano iz *ciljnega gesla* in oblikoslovne značilke pravkar najdene *izvirne besedne oblike*.

Delovanje algoritma je ilustrirano s primerom na sliki 1. Sprehodimo se od koraka do koraka in si oglejmo, na kakšne težave utegnemo pri njem naleteti.

3.1. Branje izvirne in ciljne besede

Že pri branju izvirne in ciljne besede se lahko odločimo, ali nadaljujemo s pregibno funkcijo "poišči in zamenjaj" ali pa preidemo na navadno. Če se želimo izogniti besedni analizi izvirnega in ciljnega niza, se v tem koraku omejimo na *besede* in v primeru, ko podani izvir ali cilj sestavlja niz več besed, avtomatično preklopimo na navadni "poišči in zamenjaj".

3.2. Analiza izvirne in ciljne besede

Za izvirno in ciljno besedo moramo poiskati vnos(e) v seznamu, ki besednim oblikam prirejajo pare gesla in oblikoslovne značilke. Če za eno ali za drugo ne najdemo vnosa v seznamu besednih oblik, opustimo pregibni "poišči in zamenjaj" in preidemo na navadnega – zamenjujemo samo izvirni niz s ciljnim.

Prav tako na tej točki opustimo pregibni "poišči in zamenjaj", če izvirna in ciljna beseda slovnično nista zamenljivi. Zgled je zamenjava besede `sosed` z besedo `delal`. Tu nam slovnična pravila ne dajejo nobene možnosti, da bi uganili, kaj je uporabnik imel v mislih, zato ravnamo, kot da gre za preprosto zamenjavo nizov. O slovnični zamenljivosti bomo povedali več v razdelku "Ekvivalenčni razredi".

Druga težava je lahko, da tako pri izvirni kot pri ciljni besedi ena besedna oblika ustreza več geslom. Zgled: niz "soseda" lahko pomeni prvi sklon ednine samostalnika "soseda", prvi ali četrti sklon dvojine samostalnika "sosed" ali pa drugi in četrti sklon ednine tega samostalnika.

```
soseda =     Ncfsn
soseda sosed Ncmda
soseda sosed Ncmdn
soseda sosed Ncmsa
soseda sosed Ncmsg
```

V tem primeru primerjamo analizi izvirne in ciljne besede. Denimo, da obravnavamo zamenjavo niza "soseda" z nizom "Janeza". Ker je v primeru gesla "sosed" ujemanje s ciljno besedo dosti boljše, lahko domnevamo, da je uporabnik mislil to geslo, in ne zahtevamo potrditve.

```
soseda =     Ncfsn
soseda sosed Ncmda  Janeza Janez Npmda
soseda sosed Ncmdn  Janeza Janez Npmdn
soseda sosed Ncmsa  Janeza Janez Npmsa
soseda sosed Ncmsg  Janeza Janez Npmsg
```

Ni pa nujno, da bomo vedno imeli takšno srečo. V spodnjem zgledu, denimo, je že nimamo.

```
soseda =     Ncfsn  Petra =     Npfsn
soseda sosed Ncmda  Petra Peter Npmda
soseda sosed Ncmdn  Petra Peter Npmdn
soseda sosed Ncmsa  Petra Peter Npmsa
soseda sosed Ncmsg  Petra Peter Npmsg
```

Tu nimamo druge izbire, kot da od uporabnika zahtevamo, da se sam odloči, katero od možnosti je imel v mislih.

Enaka težava kot tu se ponovi še pozneje v osmem koraku, ko za najdeni niz preverjamo, ali ustreza samo iskanemu geslu.

3.3. Sestavljanje seznama izvirnih besednih oblik

Pri sestavljanju seznama besednih oblik je pripravna podatkovna struktura *asociativne tabele* – za njo se pravzaprav skriva razpršena tabela (angl. hash) – v programskem jeziku perl. Skladnja perla omogoča, da elemente prireditvene tabele `%mfd` naslavljamo tako, da kot ključ podamo kar besedno obliko:

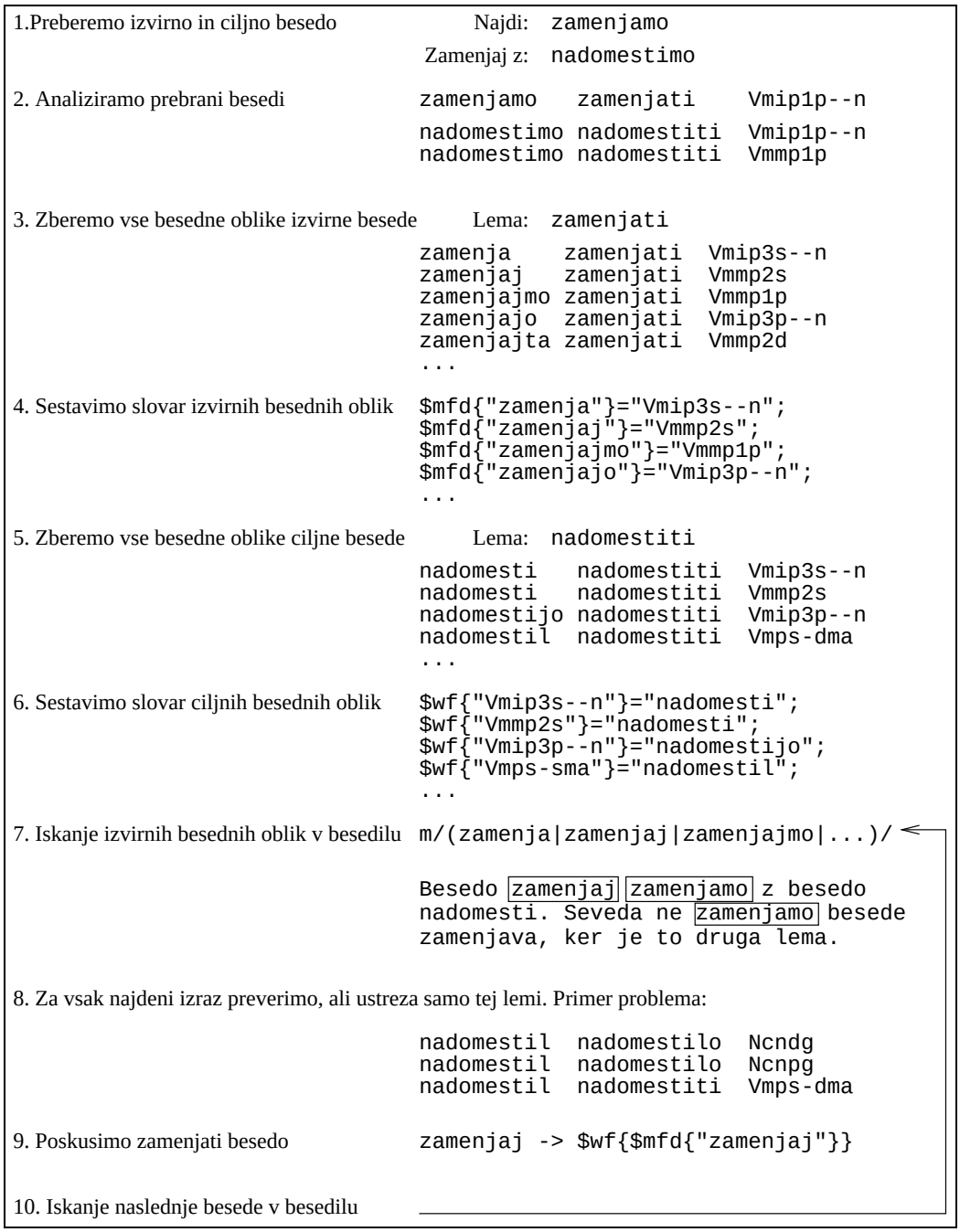
```
$mfd{"zamenja"}="Vmip3s--n";
```

Iz zbranega seznama besednih oblik izvirnega gesla zgradimo iskalni niz, ki ga uporabimo pri iskalnem operatorju `m//`.

3.4. Reakcija ob najdeni besedni obliki

Ko najdemo katero od besednih oblik izvirnega gesla, moramo najprej preveriti, ali besedna oblika ustreza samo iskanemu geslu. To ni vedno res: kot zgled bi posebej izpostavili besedne oblike "dela", "deli", "moči" in "vede"; vsako od njih lahko izpeljemo iz kar štirih različnih gesel.

Četudi je geslo nedvoumno, pa lahko še vedno ena sama besedna oblika ustreza množici različnih oblikoslovnih značilk. Ta korak zahteva ročen poseg pisca, kadar so besedne oblike na ciljni strani pri enakih oblikoslovnih značilkah pestrejša množica od izvirne, sicer pa lahko teče avtomatično. Pri tem koraku velja že povedano ob analizi izvirne in ciljne besede.



Slika 1: Ponazoritev algoritma za pregibni “poišči in zamenjaj” na zgledu zamenjave besede “zamenjati” z “nadomestiti”.

4. Ekvivalenčni razredi

V prejšnjem razdelku smo predvsem naštevali, pri katerih pogojih neujemanja ni mogoče pregibno iskanje in zamenjava. Vseeno pa bi bili prestrogi, če bi pri zamenjavi zahtevali, da zamenjamo samo besedi s popolnoma enak značilko, npr. `Ncmsa` (samostalnik, občno ime, moški spol, ednina, tožilnik) z drugo `Ncmsa`.

4.1. Samostalniki

Pri zamenjavi samostalnika s samostalnikom se se morata samostalnika ujemati v sklonu in številu. Zamenljivi parametri so vrsta: “Srečal sem soseda/Srečal sem Franceta” (soseda → Franceta, `Ncmsa` → `Npmsa`) in spol: “Srečal sem soseda/Srečal sem sosedo” (`Ncmsa` → `Ncfsa`) ter seveda kombinacija obeh: “Srečal sem Franceta/Srečal sem sosedo” (`Npmsa` → `Ncfsa`).

Zamenjava samostalnika s samostalnikom drugega spola dela težave, če je ob samostalniku še kaka druga pregibna besedna vrsta, denimo pridevnik ali zaimek. Več o tem spregovorimo še v razdelku “Težave z neujemanjem”.

Če dopustimo tudi spremembo števila, pa imamo povsod, kjer samostalnik v stavku nastopa kot osebek, dodatne težave še z neujemanjem glagola.

Sprememba sklona pri zamenjavi ni smiselna. Resda se pojavi pri nadomeščanju predložne zveze (npr. “oznaka za državo”) z rodilnikom (“oznaka države”), vendar pa gre tu pravzaprav za nepregibni “poišči in zamenjaj”.

4.2. Glagoli

Pri glagolih vrsta glagola ni zamenljiva (spreminjanje polnopomenskega glagola v naklonskega ipd.), lahko pa je smiselna sprememba glagolske oblike, npr. iz povedne v velelno ("primešate pecilni prašek" → "dodajte pecilni prašek", `Vmip2p--n` → `Vmmp2p`), ali pa iz povedne v pogojno ("bom" → "bi", `Vcif1s` → `Vcc`), zato to možnost dopuščamo.

Smiselna je tudi sprememba osebe: "Skozi okno vidimo grad" → "Skozi okno vidite grad", `Vmip1p--n` → `Vmip2p--n`, ali števila (tikanje/vikanje): "Pazi na čistočo" → "Pazite na čistočo", `Vmmp2s` → `Vmmp2p`.

Smiselna je tudi zamenjava spola v deležniku ("spremenjen" → "spremenjena", `Vmp--smp` → `Vmp--sfp`), kadar jo spremlja ustrezna sprememba ostalih pregibnih besednih vrst. Smiselna je tudi sprememba nikalnosti: "Danes je deževalo" → "Danes ni deževalo", `Vcip3s-an` → `Vcip3s-ay`. In končno, smiselna je tudi sprememba glagolskega vida: "Odpiram vrata" → "Odprem vrata".

4.3. Pridevniki

Za pridevnike velja vse povedano kot pri samostalnikih glede spola, sklona in števila. Dodatno lahko spreminjamo vrsto pridevnika: iz kakovostnega ("dober sir", `Afpmsn`) v vrstnega ("bohinjski sir", `Aopmsn`) ali svojilnega ("Mihov sir", `Aspmsn`). Smiselna je tudi sprememba stopnje: "dober" (`Afpmsn`), "boljši" (`Afcmsn`), "najboljši" (`Afsmsn`) in "predober" (`Afemsn`).

Sprememba določnosti slovnično sicer ni smiselna, v praksi lahko nanjo naletimo, še posebej znotraj istega pridevnika, pri lektorskih popravkih. Sprememba živosti se mora skladati s spremembo živosti samostalnika, na katerega se nanaša.

4.4. Zaimki

Pri zaimkih je smiselna zamenjava vrste zaimka ob nespremenjeni skadenjski funkciji. Zgled s samostalniškimi zaimki: iz osebnega zaimka ("On je prišel", `Pp3msn--n-n`) v nedoločnostnega ("Nekdo je prišel", `Pi---n----n`), vprašalnega ("Kdo je prišel", `Pq---n----n`) ali celostnega ("Vsakdo je prišel", `Pg---n----n`). Pri nikalnem zaimku ("Nihče", `Pz---n----n`) moramo obenem v zanikano obliko zamenjati tudi glagol. Zgled s pridevniškimi zaimki: kazalni ("Ta miza", `Pd-fsn----a`), nedoločnostni ("Neka miza", `Pi-fsn----a`), svojilni ("Moja miza", `Ps1fsns---a`), vprašalni ("Katera miza", `Pq-fsn----a`), oziralni ("Kakršna miza", `Pr-fsn----a`), povratni ("Svoja miza", `Px-fsn---sa`) in celostni ("Vsaka miza", `Pg-fsn----a`). Spet je enaka težava kot prej z nikalnim zaimkom: spremembo v ali iz nikalnega zaimka mora spremljati hkratna sprememba nikalnosti glagola. Zamenjava skladenjske funkcije ni smiselna.

Glede sklona velja enako, kot je povedano že pri samostalnikih: zamenjava sklona navadno ni slovnično smiselna. Pri osebnih in svojilnih zaimkih ima pomen tudi oznaka osebe. Pri svojilnih je oseba enostavno zamenljiva kategorija: "moj svinčnik" (`Ps1msns---a`) ali "tvoj svinčnik" (`Ps2msns---a`). Pri osebnih zaimkih, ki nastopajo v samostalniški skladenjski funkciji, pa moramo skupaj s spremembo osebe spremeniti tudi osebo glagola: "jaz pišem/ti pišeš".

Pri vseh zaimkih v pridevniški skladenjski funkciji je slovnično smiselna tudi sprememba spola, če jo obenem spremlja tudi sprememba spola drugih pregibnih besednih vrst. Prav tako velja tudi za spremembo spola samostalniških osebnih zaimkov, če jo obenem spremlja tudi sprememba spola pri glagolu. Enako kot za spremembo spola velja za spremembo števila zaimkov. O tem spregovorimo več v razdelku "Težave z neujemanjem".

Pri svojilnih zaimkih sta relevantni kategoriji še število in spol svojine. Za razliko od spremembe spola ("njegov/njegova", `Ps3msnsm--a`/`Ps3fsnsm--a`) pa je sprememba spola svojine ("njegov/njen", `Ps3msnsm--a`/`Ps3msnsf--a`) enostavno zamenljiva, in enako tudi število svojine ("njegov/njihov", `Ps3msnsm--a`/`Ps3msnp---a`).

4.5. Števniki

Zamenjava vrste številnikov – iz glavnega v vrstilnega, množilnega ali ločilnega – slovnično ni smiselna. Glede zamenjave spola ali števila velja enako, kot smo povedali za pridevnike. Zamenjava sklona tudi tu ni slovnično smiselna. Pač pa je načeloma zamenljiv zapis glavnih ("sto/100/C", `Mcmpnl`/`Mcmpna`/`Mcmpnr`) in vrstilnih ("stoti/100.", `Mompnl`/`Mompna`) številnikov.

4.6. Nepregibne oblike

Prislovi, vezniki, členki, medmeti in okrajšave so nepregibne oblike, katere povsem zadovoljivo obdela že navadna funkcija "poišči in zamenjaj".

Predlogi se vežejo z določenim sklonom. Ker zamenjava sklona samostalnika navadno ni slovnično smiselna, tudi prilagoditev predlogov spremenjenemu sklonu samostalnika ne bomo obravnavali.

4.7. Medkategorijske zamenjave

Do zdaj smo obravnavali zamenjave znotraj posamezne besedne vrste. S tem seveda nismo izčrpali vseh možnosti. Tako pomensko kot slovnično smiselne so zamenjave samostalnikov z osebnimi zaimki v tretji osebi: "sosed" → "on", `Ncmsn` → `Pp3msn--n-n`.

Pomensko in slovnično smiselne so tudi zamenjave pridevnikov s pridevniškimi zaimki ali številniki: "slovenski tekmovalec" (`Aopmsn`), "ta/naš/katerikoli/vsak tekmovalec" (`Pd-msn----a`/ `Ps1msnp---a`/ `Pi-msn----a`/ `Pg-msn----a`) in "prvi tekmovalec" (`Momsnl`).

Pri navedenih zgledih smo poleg ohranjanja sklona ohranjali tudi spol in število. Glede spremembe le-teh velja, kar smo povedali že prej pri samostalniku in pridevniku.

5. Težave z neujemanjem

Zavedati se je treba, da pregibna različica "poišči in zamenjaj" za slovenščino kljub temu, da omogoča menjavo različnih oblik iste besede, ne rešuje vseh težav, ki jih taka zamenjava povzroči. Eden od najbolj očitnih problemov, ki nastanejo ob zamenjavi, je neujemanje znotraj stavka ali

pogosto tudi znotraj besedila. Navedeni primeri so iz romana "1984" G. Orwella, ki je bil oblikoslovno opremljen v projektu MULTEXT-East (Dimitrova et al., 1998; Erjavec et al., 1998), ter iz besedil konference "Gore", prav tako dostopne prek konkordančnika laboratorija za naravne jezike na odseku za inteligentne sisteme Instituta Jožef Stefan (Erjavec, 1998).

Če izvirni samostalnik ni istega spola kot ciljni, pride do neujemanja s pridevnikom, ki se na ta samostalnik nanaša. Enako je tudi z deležnikom, z zaimkom in tako dalje. V prvem zgledu želimo zamenjati *sosed* s *soseda*: najbližji sosed → *najbližji soseda; Karnijski sosed → *Karnijski soseda; da ji je to sosed povedal → *da ji je to soseda povedal.

V naslednjem zgledu zamenjamo *beseda* z *izraz*: to je spet beseda, ki je že dolgo nisem slišal → *to je spet izraz, ki je že dolgo nisem slišal; Neka beseda je v Novoreku, je rekel Syme. Ne vem, ali jo poznaš → *Neka izraz je v Novoreku, je rekel Syme. Ne vem, ali jo poznaš.

V veliki večini primerov nastane neujemanje, ki ga je treba popraviti ročno.

Tovrstne težave niso omejene na samostalnike. Če se izraza, ki ju želimo zamenjati vežeta različno (z različnimi predlogi ali skloni), prav tako nastane neujemanje.

Če na primer zamenjamo glagola, ki se vežeta z različnimi predlogi, je zelo verjetno, da dobljeni stavki ne bodo pravilni. Za zgled zamenjajmo glagol *razmišljati* (ki se veže s predlogom "o") z glagolom *misliti* (ki se veže s predlogom "na"): Nehal je razmišljati o vojni → ?Nehal je misliti o vojni; Zdelo se je, da starec globoko razmišlja → ?Zdelo se je, da starec globoko misli.

Podobna neujemanja nastanejo, če zamenjamo dva glagola, od katerih je eden prehoden, drugi pa ne. Za zgled zamenjajmo glagol *plavati* z glagolom *preplavati*: tolmuni, v katerih plavajo race → *tolmuni, v katerih preplavajo race; Bilo je, kot bi plaval proti toku → *Bilo je, kot bi preplaval proti toku.

Podobna neujemanja nastajajo tudi pri pridevnikih ali številkih, ki se vežejo z različnimi skloni in predlogi.

6. Tehnična izvedba

Iskanje izvirne in ciljne besede v slovarju je časovno kritična točka. Trenutno je za shranjevanje uporabljen podatkovni sistem Berkeley DB (Olson et al., 1999; Seltzer and Yigit, 1991), ki uporablja linearno razpršeno tabelo (Litwin, 1980). Pri tem prva razpršena tabela povezuje besedne oblike z gesli in oblikoslovnimi značilkami, druga pa obratno. Zgled zapisov v prvi bazi:

```
nečim   1:1:nekaj:Pi---i----n
eden    2:1:eden:Mcmsnl:eden:Ncmsn
```

Celo v takšni izvedbi je iskanje niza v datoteki na disku zadovoljivo hitro. Pri tej velikosti seznama oblikoslovno označenih besednih oblik – približno 15 MB – bi v časovno kritičnih primerih prišlo v poštev celo hranjenje razpršenih tabel v pomnilniku.

V prototipni fazi je funkcija izvedena kot samostojni skript v perlu. Zares ima smisel šele ob vstavitvi v urejevalnik besedil, na primer Emacs (Stallman, 1997).

Običajno funkcija "poišči in zamenjaj" omogoča uporabniku, da se odloči bodisi za individualno potrjevanje vsakega posamičnega primera bodisi kot paketno zamenjavo vseh primerov. Vredno je opozoriti, da pri njeni pregibni izvedbi kljub morebitni izbiri paketne zamenjave vseeno zahtevamo individualno potrditev v dvoumnih situacijah, ki jih algoritem ne zmore razrešiti sam.

7. Načrti za nadaljnje delo

V opisani izvedbi algoritem v vseh dvomljivih situacijah zahteva odločitev uporabnika. Nekaterim takim primerom bi se lahko izognili z enostavno stavčno analizo.

Dvoumnim situacijam, ko ena besedna oblika lahko ustreza več različnim oblikoslovnim oznakam, bi se izognili s predhodnim oblikoslovnim označevanjem. Nekaj korakov pri slovenščini je v tej smeri že bilo storjenih (Džeroski et al., 1999).

Kot smo že omenili, je ena od prvenstvenih nalog tudi integracija programa v urejevalnik besedil. Emacs je prvi cilj zaradi svoje odprtosti in razširljivosti. Gotovo pa bi bila za širši krog uporabnikov zanimivejša integracija v katerega od pisarniških paketov. V tej zvezi je še posebej zanimiva najava podjetja Sun Microsystems, da 13. oktobra 2000 objavi izvorno kodo pisarniškega paketa StarOffice. Preliminarni pregled ustroja paketa (Sun, 2000) kažejo, da je modul za iskanje povsem ločen od krajevno prilagojenih delov. Integracija zato verjetno ne bo lahka, tako da nas ne sme presenetiti, če bo namesto tega pregibno iskanje prej zaživelo v kakem manjšem urejevalniku.

8. Zaključek

Urejanje besedil je dandanes tako pri nas kot v svetu zdaleč prevladujoča uporaba računalnika, in obratno: vsaj v razvitem svetu so praktično vsi spisi pripravljeni z urejevalniki besedil. Glede na to, za kako pomembno področje gre, moramo ugotoviti, da programi za pisanje besedil vsebujejo pretresljivo malo jezikovnega znanja, ki bi bilo lahko v pomoč pišočemu. Bolj ali manj vse jezikovno znanje navadno pomenita črkovalnik in pravila za razzlogovanje besed.

Pregibna funkcija "poišči in zamenjaj", opisana v tem prispevku, pomenita skromen korak naprej v smeri urejevalnikov besedil, ki bodo z dodatnim znanjem o jeziku olajšali pisanje. Pisca seveda ne odreši vnovičnega branja besedila po tem, ko je v njem zamenjal kakšno besedo. Za kaj takega ne bi bilo dovolj niti, če bi opravili stavčno analizo in znotraj stavka uskladili ujemanje vseh pregibnih besednih vrst, saj se lahko tudi drugi stavki pomensko navezujejo nanj. Zgled: če želimo v besedilu "Sredi sobe je stal vegast stol. Bil je star in umazan." nadomestiti "stol" z mizo, ni dovolj, da znotraj istega stavka popravimo še glagol in pridevnik, saj se nanj navezuje tudi naslednji stavek.

9. Literatura

Ludmila Dimitrova, Tomaž Erjavec, Nancy Ide, Heiki-Jan Kaalep, Vladimír Petkevič, in Dan Tufiş. 1998. Multext-East: Parallel and Comparable Corpora and Lexicons for Six Central and Eastern European Languages. V: *COLING-ACL '98*, str. 315–319, Montréal, Québec, Canada.

Sašo Džeroski, Tomaž Erjavec, in Jakub Zavrel. 1999. Morphosyntactic tagging of slovene : evaluating pos taggers and tagsets. Tehnično poročilo 8018, Inštitut Jožef Stefan, Ljubljana. http://nl.ijs.si/lll/bib/dzerza-report/.

Tomaž Erjavec in Nancy Ide. 1998. The MULTEXT-East corpus. V: Antonio Rubio, Natividad Gallardo, Rosa Castro, in Antonio Tejada, ur., *Proceedings / First International Conference on Language Resources and Evaluation, Granada, Spain, 28–30 May 1998*, str. 971–974, Paris. European Language Resources Association.

Tomaž Erjavec, Ann Lawson, in Laurent Romary. 1998. East meets West: A Compendium of Multilingual Resources. CD-ROM. ISBN: 3-922641-46-6.

Tomaž Erjavec. 1998. nl.ijs.si concordance service. Spletna stran. http://nl2.ijs.si/corpus/.

Jeffrey E. Friedl in Andy Oram, ur. 1997. *Mastering Regular Expressions: Powerful Techniques for Perl and Other Tools*. O'Reilly, Sebastopol, CA.

Witold A. Litwin. 1980. Linear hashing: A new tool for file and table addressing. V: *Sixth International Conference on Very Large Data Bases, October 1-3, 1980, Montreal, Quebec, Canada, Proceedings*, str. 212–223. IEEE Computer Society Press.

Michael Olson, Keith Bostic, in Margo Seltzer. 1999. Berkeley DB. V: *Proceedings of the Freenix Track of the 1999 Summer USENIX Technical Conference, June 6–11, 1999, Monterey, California*, str. 183–191, San Francisco, CA, USA. USENIX. http://www.sleepycat.com/docs/ref/refs/bdb_usenix.html.

Margo Seltzer in Ozan Yigit. 1991. A new hashing package for UNIX. V: *Proceedings of the Winter 1991 USENIX Conference: January 21-January 25, 1991, Dallas, TX, USA*, str. 173–184, San Francisco, CA, USA. USENIX. http://www.sleepycat.com/docs/ref/refs/hash_usenix.ps.

Richard Matthew Stallman, 1997. *GNU Emacs Manual*. Free Software Foundation, Boston, MA, 13. izd.

Sun Microsystems, Inc., Palo Alto, CA, 2000. *StarOffice API Reference Manual 1.0*. http://soldc.sun.com/staroffice/reference/.

Jože Toporišič. 1992. *Enciklopedija slovenskega jezika*. Cankarjeva založba, Ljubljana.

Dan Tufis, Nancy Ide, in Tomaž Erjavec. 1998. Standardized specifications, development and assessment of large morpho-syntactic resources for six central and eastern european languages. V: *Proceedings of the First International Language Resources and Evaluation Conference*, str. 233–240, Granada, Spain.