Sodobni prosti črkovalniki in baze pravilnih besednih oblik

Aleš Košir

ZASLON d.o.o, skupina HERMES SoftLab
Litijska 51, Ljubljana, Slovenija
ales.kosir@hermes.si

Računalniki so nesporno zelo uporabni, da nam olajšajo nekatera mukotrpna, a preprosta opravila. Med takšne naloge štejemo pregledovanje pravilnosti črkovanja. Postopek je v osnovi načeloma zelo preprost: zapisano besedilo mora program za črkovanje razdeliti na besede in za vsako besedo ugotoviti, ali je njen zapis pravilen. Pravilnost zapisa slovenske besede se ugotavlja kot skladnost s katerim od uveljavljenih naborov pravilnih besednih oblik, kakršen je aktualni Slovar slovenskega knjižnega jezika ali Slovenski pravopis. Z naprednejšimi tehnikami pregledovanja, ki vključujejo skladenjska pravila, se zaenkrat še ne ukvarjamo.

V zadnjem času je v uporabi vse več prostih črkovalnikov, od katerih jih je večina primerna tudi za rabo v slovenskem jeziku. Posebnost slovenskega jezika v primerjavi z angleščino, za katero je na voljo večina črkovalnikov, je bogata pregibnost besed, ki zahteva, da črkovalnik obvladuje veliko število besednih oblik. Če za osnovno rabo zadošča, da angleški črkovalniki hranijo približno 30 tisoč besed, je za slovenski jezik takšna številka nezadostna, potrebujemo vsaj desetkrat toliko besednih oblik. Do preloma tisočletja, ko so bili pomnilniki običajnih osebnih računalnikov preskromni, da bi v njih lahko obenem hranili celotno bazo slovenskih besednih oblik, je bilo treba uporabljati metode za krčenje slovarja:

- besede so bile shranjene v pomnilniku tako, da so bile leksikografsko urejene, a zapisane le s tistimi znaki, ki so se razlikovali od predhodne besede; predstavnik takih črkovalnikov je mspell,
- besede so bile okrajšane s ponskimi pravili, tako da je bilo treba ob besedi hraniti le njeno okrnjeno osnovo in ponsko pravilo; predstavnik takih črkovalnikov je ispell.

Z razširjenostjo dovolj velikih pomnilnikov so takšne težave vse manj izrazite in si lahko privoščimo shranjevanje celotnih razpršenih tabel obsežnih seznamov besednih oblik v pomnilnik. Če je do nedavna veljalo, da smo morali biti pri izboru črkovalnika zaradi opisanih težav s pomnilnikom previdni, se je izbira zdaj razširila. Primerni prosti programi za črkovanje, ki jih uporabljamo v zadnjem času in so neodvisni samostojni namenski programi ali moduli večjih sistemov, so:

- ispell kot eden prvih razširjenih prostih črkovalnikov, prilagojenih tudi za slovenščino,
- aspell kot sodobni prosti črkovalnik, ki je začenjal nadomeščati ispell,
- novi aspell, ki je od aprila 2002 priporočeni črkovalnik prostih programov,
- myspell kot del pisarniškega okolja OpenOffice.

Ker pa črkovalnik navadno uporabljamo kot del storitev katerega od urejevalnikov besedil, na spodnji tabeli prikazujemo, kako je urejevalnik mogoče povezati s črkovalnikom.

Urejevalniki	**Črkovalniki**			
	Ispell	**Aspell + pspell**	**New aspell**	**Myspell**
OpenOffice				√
KOffice			√	
abiword	√			
emacs	√	√	√	
vi	√	√	√	
pine	√	√	√	
sylpheed	√	√	√	

Preglednica prikazuje, kateri črkovalniki so vgrajeni v katerega od prostih urejevalnikov

Poglavitna ovira pri uporabi prostih črkovalnikov pri slovenščini je bilo pomanjkanje dobrih seznamov pravilnih besednih oblik. Dober seznam ima naslednje lastnosti:

- vsebovati mora vse besede iz najpogostejše dnevne rabe (časopisi, osebno dopisovanje, uradni dopisi),
- seznam mora vključevati sodobne pravilne besede (na primer *taliban*) in odsvetovati neustrezne besede (denimo *zgoščenka*),
- seznam mora vsebovati pogostejša lastna imena,
- v seznamu ne sme biti zavajajočih redkih besed, ki so morda pravilne, a podobne napačno pisanim pravilnim besedam (redka neobčevalna beseda *brezsen*).

Zaželeno je, da nam črkovalnik omogoča naše pogosto rabljene besede dodati v seznam pravilnih besednih oblik.

Od leta 1995 poteka v okviru ljubiteljskega projekta GNUsl zbiranje besednih oblik. Zbiranje je vključevalo pregledovanje dostopnih preverjenih elektronskih virov (internetne izdaje časopisov, elektronske knjige, lektorirana besedila…). Pokazalo se je, da na ta način sicer lahko zberemo večje število besednih oblik, a da ostajajo v seznamih sistematično luknje, zaradi katerih pri preverjanju črkovanja običajnega besedila naletimo na znaten delež (2-10 %) neprepoznanih besed. Te seznam je na voljo na strani http://nl.ijs.si/GNUsl.

V sodelovanju s podjetjem Amebis in s podporo Ministrstva za informacijsko družbo je skupina za slovenjenje pri društvu Lugos pripravila obsežen pregledani seznam besednih oblik, ki ga dopolnjuje seznam znanih napačnih besed (na primer *življenje*). V nasprotju s prejšnjim seznamom so v tem zajete vse besedne oblike pregibnih besed. Slovar je vgrajen v črkovalnike aspell, ispell in myspell ter preskušen z urejevalniki besedil, ki so našteti v zgornji tabeli. Pri ocenjevanju kakovosti seznama se pokaže, da je besedišče, ki obsega 1.163.826 besednih oblik (14.638.738 znakov), povsem primerno za vsakdanjo rabo in tudi za pregledovanje strokovnega jezika. V njegovo besedišče so vključeni popravki iz zadnje izdaje Slovenskega pravopisa, s čimer je ta seznam postal trenutno najsodobnejši. Seznam napačnih besed pa zajema 10.478 napačnih besed, zapisanih s skupaj 128.323 znaki.

Za naštete črkovalnike so pripravljeni distribucijski paketi za namestitev v operacijski sistem Linux, dostopni pa so na strani http://nl.ijs.si/GNUsl. Črkovalnike in slovensko bazo pravilnih besednih oblik je mogoče z nekaj osnovne spretnosti uporabiti tudi v drugih operacijskih sistemih, če so črkovalniki preneseni v njih za kak drug jezik. Vsi opisani izdelki so dostopni pod licenco za prosto programsko opremo.

Literatura

Košir A., (2002) "Slovenščina in računalniki" http://nl.ijs.si/GNUsl, Slovenija

Košir A., Peterlin P., Erjavec T., (1998) “GNUsl: prosto programje in slovenščina” Informacijska družba 1998, Jezikovne tehnologije, Ljubljana, Slovenija

Košir A., Peterlin P., (1997) “Slovenski črkovalniki in international Ispell”, 828. Seminar za numerično in racunalniško matematiko, IMFM, Slovenija

Peterlin P., (1996) “Zgodovina črkovalnikov za slovenščino” http://nl.ijs.si/GNUsl, Slovenija