

»SLONČEK«, PROGRAM ZA GESLENJE SLOVENSKEGA TEKSTA

INFORMATICA 3/91

Keywords: Spelling checker, Slovene language,
Word analysis

Jure Zupan
samostojni znanstvenik
Dragomer, 61351 Brezovica

ABSTRACT The principles of the spelling checker of the slovene language implemented in the program package 'SLONČEK' are described and discussed. Only the option called 'Analysis of words' in the plain text is described in detail. The analysis of words is based on the search for all combinations of 'roots' and 'sufices' that could give the valid words recognized in a dictionary. Besides the analysis of each word the program identifies 10 punctuation marks, hence giving a complete information for further syntax and semantic check of new words in the plain text. The dictionary of the program contains about 250 sets of sufices (approx. 20 sufices/set) and more than 32,000 'roots' of slovene words.

POVZETEK Opisan je koncept in izvedba programskega paketa SLONČEK za pregledovanje slovenskih tekstov. V prispevku je podrobneje obdelana le opcija za geslenje 'Analiza'. Geslenje je osnovano na iskanju in izpisu vseh možnih 'korenov' in pripadajočih 'pripon' za vsako besedo v tekstu. Program identificira še 10 ločil, tako da rezultirajoče geslenje daje polno izhodišče za nadaljnje sintaktično in semantično obdelavo besedil. Program vsebuje več kot 32.000 'korenov' slovenskih besed in približno 250 'nizov pripon' (v povprečju 20 pripon/niz) različnih sklanjatev in spregatev.

1 Uvod

Pregledovalnike slovenskih besedil bi glede na namen delovanja lahko razdelili v dve skupini. V prvi so pregledovalniki, ki preverjajo pravilnost zapísane besede in do neke mere kontrolirajo pravilnost sintakse (raba velike začetnice, raba s, z, h in k itd.). V drugi vrsti so pregledovalniki, ki poleg kontrole pravilnosti zapísane besede določijo pregledovani besedi vse možne **osnovne** oblike (nedoločnik, prvi sklon moške oblike itd.) in vse možne **ustrezne** oblike (sklon, oseba, čas itd.) v kateri najdena osnovna oblika lahko nastopa. Dodatno pregledujejo še vrsto in položaj ločil, ki so bistvena za razumevanje besedila ali za določanje sintaktičnih povezav.

Da bo povsem jasno, kaj je mišljeno z omenjenimi dejavnostmi, navajam v pojasnilo nekaj primerov. Oglejmo si naslednje stavke:

Pridi h meni.

Hvala, lep dete!

Župan je rekel: "To je dobro za vas".

V prvem stavku sta dve napaki, ki ju lahko odkrije že program, ki ni namenjen geslenju. Beseda 'priddi' je napačno napisana in program je ne sme imeti v slovarju. Poleg tega je namesto predloga 'k' uporabljen predlog 'h'. Tovrstno napako lahko vedno odkrijemo s preverjanjem prve črke sledeče besede.

Drugi stavek nima z gledišča pravilnosti pisanja besed nobene napake, pač pa ima očitno napačno sintakso med pridevnikom 'lep' in samostalnikom 'dete'. Če naj bi program opozoril na tovrstne napake, mora ugotoviti, da je beseda 'lep' pridevnik **moškega** spola v prvem sklonu ednine, naslednja beseda pa samostalnik **srednjega** spola v

prvem sklonu ednine, kar ni v skladu s pravili sintakse. Pri tem stavku je treba opozoriti še na dejstvo, da mora program preveriti sintakso v skladu z ločili. Če bi, na primer, namesto za besedo hvala stala veja za besedo lep, bi moral program pridevnik 'lep' uskladiti s samostalnikom 'hvala' in ne s samostalnikom 'deja'. Za pravilno izvajanje sintakse mora torej program poleg gesel in njihovih možnih oblik podati tudi vse potrebne podatke o vrsti in položaju ločil v stavku.

Tretji stavek je sicer pravilen tako v pravopisnem kot tudi v sintaktičnem smislu, je pa seveda popolnoma nedoločljiv s stališča semantike. Za razumevanje pomena tega stavka (n.pr. za prevod v tuj jezik) je nujno potrebna obdelava **vseh** gesel za **vse** besede v kontekstu sledečih oziroma predhodnih stavkov. Pri takih primerih mora orodje, ki naj omogoči pravilno razumevanje stavka (za prevod) zajeti **vse možne pomena (gesla) vseh besed** v posameznih stavkih. Zato mora najprej ločiti stavke med seboj (obdelava ločil) in nato nuditi vse možne variante osnovnih gesel.

Šele ko imamo na voljo orodje, ki v preiskovanem tekstu poda zgoraj navedene podatke (vsa gesla in z njimi vse možne oblike spregatev, spolov, časov in sklonov ter vrsto in položaje ločil), lahko nadaljujemo s sintaktično in kasneje s semantično obdelavo besedila.

V nadaljevanju sestavka je opisana osnovna ideja in algoritem programa SLONČEK, ki vrši vse zgoraj omenjene osnovne funkcije geslenja. Program SLONČEK je (med drugim) orodje za geslenje, podajanje slovničnih oblik in identifikacijo ločil. Rezultirajoča datoteka zaporedja gesel, slovničnih oblik in identifikacij ločil je osnova za nadaljnjo sintaktično in semantično obdelavo predloženega besedila. Podanih je tudi nekaj primerov (glej poglavje Primeri).

2 Koncept geslenja s programom SLONČEK

Osnovna zamisel geslenja pri programu SLONČEK je deljenje vsake besede na vrsto 'korenov' in 'pripon' in pregledovanje slovarja, če se posamezna delitev na 'koren' in 'pripono' ujema z ustreznimi podatki v bazi 'korenov' in 'standardnih pripon'. Deljenje preiskovane besede na 'koren' in 'pripono' poteka postopno - po posameznih

črkah. Vsota črk 'korena' in 'pripone' je ves čas stalna: kolikor krajši je 'koren', toliko daljša je 'pripona'. Najdaljša 'pripona' je lahko dolga šest, najkrajša pa nič črk. Program vedno prične preiskovati slovar za dano besedo s korenem, ki je enak celotni besedi (pripona dolga nič črk). Če program korena v slovarju ne najde, odreže 'korenu' zadnjo črko, jo doda na začetek stare 'pripone' in išče ponovno. Rezanje zadnje črke 'korena' in iskanje v slovarju, ponavlja program tako dolgo, da 'korena' ni več ali pa je bilo odrezanih že šest črk. To pomeni, da je najdaljša 'pripona' lahko dolga le šest črk. Glej Tabelo 1.

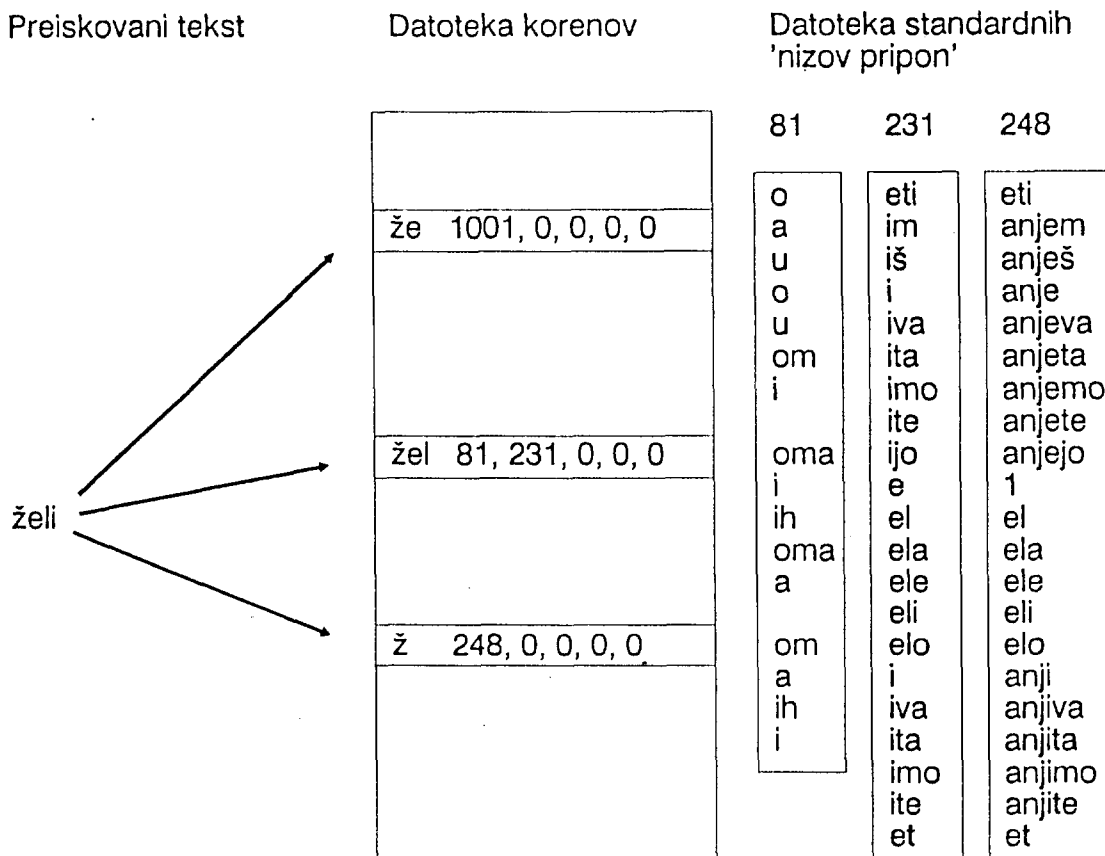
Tabela 1. Rezanje preiskovane besede na 'korene' in 'pripone'. Če je beseda krajša kot šest črk (primer: mati), se algoritem razbijanja zaključi že prej.

Korak	'Koren'	'Pripona'
1	pregledovati	-
2	pregledovat	(i)
3	pregledova	(ti)
4	pregledov	(ati)
5	pregledo	(vati)
6	pregled	(ovati)
7	pregle	(dovati)
1	mati	-
2	mat	(i)
3	ma	(ti)
4	m	(ati)

Slovar sestavljajo 'koreni' ob katerih so vpisani vsi možni nizi 'pripon'. Nizi 'pripon' so identificirani z zaporednimi številkami od 1-280. Ko algoritem odkrije navzočnost 'korena' v slovarju, prične preverjati ali je 'pripona', odrezana od preiskovane besede, vsebovana v 'nizu pripon' identificiranih ob korenu v slovarju. Če 'pripone' v standardnem nizu ni, ima preiskovana beseda status 'neznana'. V primeru, ko odrezano 'pripono' v nizu najdemo, postane preiskovana beseda 'znana'. Algoritem preveri še kolikokrat in na katerih mestih se v identificiranem 'nizu standardnih pripon' odrezana 'pripona' ponovi. Mesto, kjer se odrezana

'pripona' v nizu nahaja, namreč pove sklon, spol, število in/ali čas v katerem se preiskovana beseda nahaja. Slika 1 kaže kako algoritem pregleda besedico 'želi'.

Vse dobljene podatke o vsaki besedi (vsa gesla in mesta v nizu kjer se 'pripona' ujema s standardno pripono) algoritem izpiše na izhodno datoteko (glej poglavje Primeri).



Slika 1. Pregledovanje datoteke korenov in pripon za besedico 'želi'. Celotne besede 'želi' ni v datoteki. Koren 'žel' nastopa dvakrat, enkrat z nizom pripon št. 81 (samostalniki srednjega spola na 'o') in enkrat z nizom št. 231 (glagoli na 'eti'). Koren 'že' nastopa samo enkrat kot samostojna beseda (jedro) brez niza 'pripon' (oznaka 1001). Ker je koren 'že' jedro, torej po definiciji brez 'pripon', besedica 'že' v danem primeru ni potrjena kot možna. Koren 'ž' nastopa v datoteki z nizom pripon št. 248 (glagoli na 'eti').

Na sliki 1 se razločno vidi, da SLONČKOV algoritem pregleda za besedo 'želi' štiri možne 'korene': 'želi', 'žel', 'že' in 'ž'. Pri tem najde tri korene ('žel', 'že' in 'ž') v slovarju. Kljub temu, da imajo najdeni trije koreni skupno štiri 'nize pripon' (81, 231, 248 in 1001) so možna le tri gesla: želo, žeti in žeteli - koren 'že' z oznako 1001 (jedro) v danem primeru ne pride v poštev. Za ostale korene pa vedno najde ustrezno 'pripono' v 'nizu'. Pri geslu želo, najde tri možne oblike (prvi in četrti sklon dvojine ter šestli sklon množine), pri glagolu 'žeteli' tretjo osebo sedanjika in velelnik in končno pri glagolu 'žeti' najde 'pripono' 'eli' (koren je 'ž') v skupini kjer so množinska obrazila za prihodnji in pretekli čas. Osnovno obliko gesla algoritem dobi tako, da 'korenu' doda 'pripono', ki je v pridruženem 'nizu pripon' na prvem mestu.

3 Organizacija slovarja korenov in pripon

Slovar korenov. Kot je bilo že omenjeno, so gesla, ki jih podaja program SLONČEK, dobljena s pomočjo slovarja 'korenov' in ustreznih nizov 'pripon'. Slika 1 kaže datoteko z neposrednim dostopom v kateri so koreni dosegljivi preko 'hash' (razpršilnega) algoritma (1) s pomočjo praštevilskih dvojčkov (2). SLONČEK ima korene razporejene v treh datotekah. Vse tri datoteke so enako dolge (vsaka ima 20023 zapisov), ločijo se le po dolžini zapisa. V prvi datoteki so shranjeni koreni dolgi do vključno šest znakov, v drugi so koreni dolgi 7, 8 in 9 znakov, v tretji pa so shranjeni vsi daljši koreni. Največja dovoljena dolžina korena je 18

znakov. Vsak polni zapis vsebuje poleg korena tudi oznako (ali več oznak) niza(ov) pripon, ki z danim korenem tvorijo smiselna gesla. Kot se vidi s slike 1, ima datoteka z najkrajšimi koreni prostor za identifikacijo petih nizov, srednja datoteka ima lahko tri, ob najdaljših korenih pa je prostora za identifikacijo le dveh nizov.

Dolžina zaplsov je bila izbrana izkustveno, na podlagi statistike približno 15.000 besed. Med delom se je pokazalo, da bo v prihodnosti potrebno podaljšati zapise, da bi dobili možnost hranjenja večjega števila različnih nizov. Če sodi beseda v tako imenovano 'jedro' jezika (3), potem ima namesto identifikacije niza posebno oznako (1001).

Niz pripon. Z izrazom 'pripona' označujem tisti del besede, ki skupaj s 'korenem', ki je shranjen v eni od omenjenih treh datotek, tvori pravilno slovensko besedo v poljubni obliki, sklonu, spolu, času, osebi itd. 'Niz pripon' je torej skupina 'pripon', ki skupaj s 'korenem' tvorijo vse oblike sklanjatve ali spregatve. Najbolj značilno za 'niz pripon', ki ga uvaža programa SLONČEK je, da prva črka vseh 'pripon' v nizu ni nikoli enaka pri vseh priponah. Če bi namreč bila, bi že sodila h 'korenu'.

'Pripona' v konceptu programa SLONČEK torej nima slovnicega pomena, pač pa čisto računalniškega: to je tisti del besede (samostalnika, pridevnika, glagola itd.), ki pri sklanjatvi ali spregatvi vsaj enkrat zamenja začetno črko. Z drugimi besedami povedano, 'koren' je tisti del besede, ki pri celotni sklanjatvi ali spregatvi ostane neizpremenjen. Pravilo o neizpremenjanju korena je osnovna stalnica programa SLONČEK. Velja tako pri polglasniških 'e-jih', kot pri raznih mehčanjih in podobnih posebnostih slovenščine.

Program SLONČEK trenutno razlikuje 6 vrst pripon: štiri samostalniške (moške, ženske, srednje in množinske) in po eno vrsto pridevniških (sem sodijo tudi deležniške in druge podobne) in glagolskih. Zaradi omenjene togosti 'korena', pozna SLONČEK skoro 120 različnih samostalniških, 14 pridevniških in preko 70 glagolskih 'nizov pripon'.

Vsi samostalniški nizi pripon so sestavljeni iz 18 pripon, po 6 za vsak sklon ednine, dvojine in množine. Množinski samostalniki imajo, kar je razumljivo, le 6 pripon. Pridevniški

nizi imajo 13 pripon, glagolski pa 20. Tabela II kaže nekaj primerov 'nizov pripon'. Pomembno je opozoriti, da je istovrstna pripona, torej pripona istega sklona, časa, spola, ipd., v nizu vedno na istem mestu. S tem lahko program, ko pripono spozna, določi sklon, spol, čas itd., ki ga beseda v tekstu ima. V primeru, da je v nizu več enakih 'pripon' (to je možno samo pri samostalniki) poda računalnik vsa mesta (sklone, čase) v nizu kjer je prišlo do ujemanja.

Tabela II. Šest primerov različnih vrst nizov pripon. Pod vsakim nizom je podan tudi primer. Enica v nizu pomeni, da pri danem nizu samostojni koren ni možen.

1	41	85	129	167	207
-	a	e	je	ek	ovati
a	e	a	ij	ka	ujem
u	i	u	jam	ke	uješ
a	o	e	je	ki	uje
u	i	u	jah	ko	ujeva
om	o	em	jami	kega	ujeta
a	i	i	-----	kemu	ujemo
ov				kem	ujete
oma	ama	ema		kim	ujejo
a	i	i		kih	1
ih	ah	ih		kima	oval
oma	ama	ema		kimi	ovala
i	e	a		ak	ovale
ov				-----	ovali
om	am	em			ovalo
e	e	a			uj
ih	ah	ih			ujva
i	ami	i			ujmo
-----					ujta
					ujte
					ovat
vol	žen(a)	polj(e)	škar(je)	meh(ek)	k(ovati)

Kot primer lahko služi beseda 'žerjava', ki ga lahko sklanjamo (ob 'korenu' 'žerjav') s priponami iz niza št. '1' in niza št. '5'. Pripona 'a' je v nizu št. '1' (osebe in živali moškega spola) na mestih 2, 4, 7 in 10 (2. in 4. sklon ednine ter 1 in 4 sklon dvojine), v nizu s št. '5' pa le na 2, 7 in 10 mestu (žerjav

kot predmet v četrtem sklonu ednine nima obrazila 'a'). Ker sta ob korenu 'žerjav' v slovarju navedena **oba** niza (torej niz št. 1 in št. 5), bomo pri geslenju za kakršnokoli obliko besede 'žerjav' dobili podatke iz obeh nizov!

Posebej je treba poudariti, da je na **prvem mestu** v nizu vedno tista pripona s katero tvori 'koren' slovarsko **geslo**. Ker število nizov pripon še ni dokončno določeno, lahko uporabnik, v primeru da najde tako besedo, ki se ne pregiba po nobenem od obstoječih nizov, doda manjkajoči niz v

opciji 'Dodajanje posameznih besed' ali pri 'Interaktivnem popravljanju besedila'.

4 Primeri

Kot primer geslenja, je na sliki 2 navedena obdelava treh stavkov s programom SLONČEK. Analizirani stavki so

Pridi gori	priti (228) 16 gori (1001) 0 gora (41) 3 5 7 10 goreti (213) 4 16 k (1001) 0 meni (1001) 0 meniti (181) 4 10 16 mena (41) 3 5 7 10 ločilo št.: 1
Hvala . lep dete .	hvala (41) 1 ločilo št.: 4 lep (161) 1 dete (90) 1 4 ločilo št.: 1
župan je rekel : " Skrbi me za vas " .	župan (1) 1 je (1001) 0 jesti (229) 4 reči (234) 11 ločilo št.: 6 ločilo št.: 10 skrb (47) 2 3 5 7 8 10 13 14 16 skrbeti (213) 4 16 me (1001) 0 za (1001) 0 vas (47) 1 4 vas (145) 1 ločilo št.: 10 ločilo št.: 1

Slika 2. Trije stavki prostega teksta (na levi), obdelani z opcijo 'Analiza' programa SLONČEK. Gesla z ustreznimi podatki o možnih oblikah so v izhodni datoteki zapisana na desni strani dvojne črte. Glavna tri ločila (pika, klicaj in vprašaj) povzroče v izhodni datoteki poleg opisa še horizontalno delitev.

popravljeni stavki, ki sem jih navedel na začetku. Popravljene so samo tiskarske napake. Rezultat je dobljen z uporabo opcije 'ANALIZA', ki vrši geslenje, tako kot je opisano v zgornjih odstavkih. Program SLONČEK med geslenjem (sicer pa ne) razpozna 10 ločil in jih v izhodni datoteki opremi z ustreznimi zaporednimi številkami (pika = 1, klica = 2, vprašaj = 3, vejica = 4, podpičje = 5, dvopičje = 6, pomišljaj = 7, predklepaj = 8, zaklepaj = 9 in narekovaj = 10). Izhodna datoteka z originalnim tekstom na levi in z gesli na desni je povsem navaden alfanumeričen zapis in ga lahko kasneje obdelujemo s poljubnim programom.

5 Zaključek

Če kompleksno obdelavo teksta razdelimo na štiri oziroma pet faz relativno lahko ločljivih procedur (črkovanje, geslenje, sintaksa, semantika, prevod v tuj jezik), potem lahko rečemo, da opravlja program SLONČEK prvi dve dejavnosti. Mirno lahko trdimo, da je geslo (geslenje besed) bistveni podatek, ki je potreben za vse nadaljnje obdelave teksta v smeri njegovega razumevanja. Z opisanim programom SLONČEK (bolj natančno z njegovo opcijo 'Geslenje') je dano osnovno orodje, ki imogoča omenjena bolj zapletene dejavnosti (sintaksa, stavčna analiza, razumevanje konteksta in prevajanje). V opisani opciji je implementiran koncept, da mora biti programski del, ki opravlja geslenje teksta, ločen od delov programa, ki opravljajo zahtevnejše operacije. Bistveni razlog je v tem, da dopušča slovenščina veliko preveč dvomnosti in različnih interpretacij, da bi bilo tako zaltevno delo opravljeno sproti.

Poudariti je treba, da kljub temu, da število nizov pripon že presega število 250, iskanje novih pregibanj in dopolnjevanje pripon še ni končano. Posebej problematični so še številniki, vseh vrst zaimkov in oblike pomožnih glagolov, ki so trenutno dodani v slovar korenov, kot 'jedro' (oznake 1001, 1002 in 1003), oblikovno pa sploh še niso razdelani v nize pripon, razen nekaj izjem (glej sliko 2, primer besedice 'vas'). V opciji 'Dodajanje posameznih besed v slovar', si lahko uporabnik sam zgradi in doda poljubne nize pripon. Prav tako enostavno je dodajanje novih besed, saj program SLONČEK vodi uporabnika k odločitvi s prikazovanjem vseh možnih 'nizov pripon' za določeno vrsto besede.

Opisani program SLONČEK je operativen na osebnih računalnikih vrste IBM PC/ XT/AT/PS, ki imajo na voljo trdi disk. Delovanje iz gibkih diskov je zaradi velikega števila posegov na disk pri geslenju (6-12 posegov/besedo) zelo počasno. Popravljanje tekstov (zamenjevanje napačno napisanih besed) lahko opravimo interaktivno (sproti) ali pa s t.i. označevanjem besed, v 'batch modu'. SLONČEK vsebuje tudi hitro in enostavno pregledovanje slovarja in besed ter pripon, ki so v njem shranjene.

Slovar programskega paketa SLONČEK vsebuje približno 32.000 'korenov' zbranih iz različnih virov (4-7) in nekaj več kot 250 nizov pripon, kar ob povprečni dolžini 20 pripon na niz, da okrog 5000 pripon. Ker ostale opcije, ki jih program SLONČEK nudi, niso bile opisane natančneje, jih na tem mestu samo naštejemo:

- interaktivno popravljanje ASCII besedil,
- markiranje napak v ASCII besedilu ('batch-mode'),
- geslenje teksta in identifikacija ločil,
- dodajanje novih besed v slovar in
- pregledovanje in popravljanje slovarja

Pri testiranju dveh besedil dolgih vsak po okrog 20.000 besed (izpod peres dveh različnih avtorjev) se je pokazalo, da SLONČEK prepozna cca 97-98 % besed. Testiranje je bilo narejeno z opcijo 'Markiranje' neznanih besed.

Delovanje programa si vsak zainteresirani lahko ogleda pri avtorju na Kemijemskem Inštitutu "Boris Kidrič". Uporaba programa za raziskovalne namene, je možna po dogovoru z avtorjem.

ZAHVALA

Zahvaljujem se Kemijemskemu Inštitutu "Boris Kidrič" za finančno in organizacijsko pomoč pri izvedbi okrogle mize "Besedišče slovenskega jezika", dne 16. junija 1991. Glede na skrb, ki jo slovenski novinarji posvečajo slovenščini, je

seveda popolnoma jasno zakaj se, kljub obvestilom, ki so bila poslana vsem slovenskim časopisnim hišam, okroglo mize ni udeležil noben njihov sodelavec.

Pričujoča raziskava ni bila financirana niti s strani bivše RSS niti s strani RS ZRTD.

REFERENCE

1. D.E. Knuth, The Art of Computer Programming, Vol. 2., Seminumerical Algorithms, Addison-Wesley, Reading, Mass. 1975, Chapter 3.4
2. J. Zupan, Algorithms for Chemists, J. Wiley, Chichester, 1989, str. 34-42,
3. Okrogla miza 'Računalniška besedišča' na SAZU, 14. junij 1991, Ljubljana,
4. D. Debenjak, Slovensko-nemški slovar, CZ, Ljubljana, 1974 (cca. 15.000 besed)
5. F. Prešeren, Poezije, Ljubljana, 1848, (cca.25.000 besed)
6. S. Gregorčič, Poezije, DSM, Celovec, 1908 In Naša beseda, MK, Ljubljana, 1960, (cca. 26.000 besed),
7. J. Menart, Parizina, DZS, Ljubljana, 1960; Semafor mladosti, DZS, Ljubljana, 1963; Srednjeveške pridige in balade, CZ, 1990 (cca 26.500 besed)