mr. Mladen Sokele

Hrvatski spelling-checkeri

Dočekali smo! Imamo već šest javno dostupnih strojnih sricatelja za hrvatski jezik. Što nam nude, a što ne, kako rade i koje su im tajne - istražili smo za Vas ...

Otkad su krajem osamdesetih na hrvatskom tržištu postali dostupni programi za obradu teksta za PC računala (pod DOS-om), koji su u sebi imali spelling-checkere za engleski jezik, sa čežnjom smo priželjkivali dan kad će se pojaviti spelling-checkeri i za hrvatski jezik.

Kao što je poznato, u Hrvatskoj ne postoji standardna jezična literatura, što je između ostalog nužno i za određivanje nacionalnog identiteta. Dodatno, svjesni smo da se, kao posljedica općeg društvenog razvitka, riječi stvaraju tolikom brzinom (kako u običnom govoru tako i u znanosti) da je upitno dokle bi je tiskani jezični priručnici (i da postoje) mogli slijediti.

S druge strane, elektronička tehnologija je dokazala mogućnost arhiviranja, brzog pretraživanja i ažuriranja obimnih leksičkih baza najpoznatijih svjetskih jezika. Stoga je pred hrvatske jezikoslovce i informatičare postavljen zadatak da sustavno prionu na posao kako bi izgradili elektronički rječnik hrvatskog jezika. Uz pomoć elektroničkog rječnika, hrvatsko leksičko bogatstvo može biti dostupno korisnicima osobnih računala ili zahvaljujući općeprisutnim suvremenim telekomunikacijskim tehnologijama, ljudima širom svijeta, Hrvatima iz dijaspore ili znanstvenicima koji se bave kroatistikom.

Leksička baza u elektroničkom obliku osnova je za cijeli niz strojnih jezičnih pomagala. U prvom redu to su programi za korekciju teksta, sricatelji: spelling-checkeri. Zatim, nadgradnja na spelling-checkere jesu: rječnici sinonima (tezaurusi), stručni rječnici, rječnici žargona, etimologijski rječnici, enciklopedijski rječnici, programi za strojno prevođenje i razna druga pomagala za gramatičku i stilsku provjeru tekstova.

Tribina - "Hrvatski spelling-checker"

Dana 24.2.1997. s početkom u 17 sati, održana je tribina u Europskom domu u Zagrebu s temom "HRVATSKI SPELLING-CHECKER" na kojem su govorili naši priznati stručnjaci: dr. Šandor DEMBITZ, prof. dr. Josip SILIĆ, prof. Branko RANILOVIĆ i dipl. ing. Slaven BATNOŽIĆ.

Uvodnu riječ dao je dr. Dembitz: - "Spelling-checker je strojni alat koji pomaže da popravimo kvalitetu našeg teksta. Prvo su nastali na sjeverno-američkom prostoru sredinom sedamdesetih, a već su sredinom osamdesetih, gotovo skalpelski oštrim rezom, dijelili Europu na istočnu i zapadnu. Tada su već postojali spelling-checkeri npr. za francuski i finski jezik, dok za njima srodne jezike s druge strane "*željezne zavjese*" - rumunjski i mađarski, nisu".

Nacionalno značenje postojanja spelling-checkera

"Postojanje odnosno nepostojanje spelling-checkera za neki jezik u ovo informatičko doba postalo je kriterij za ocjenu kulturološkog stanja nacije" - istaknuo je prof. Ranilović.

Zašto su strojna jezična pomagala predmet od nacionalnog značaja? Evo primjera - u sljedećih 5 godina američka tvrtka ***SYSTRAN Software inc.*** (http://systranmt.com/naic.htm) izradit će programsku podršku za strojno prevođenje istočno-europskih jezika, među kojima i za "srpsko-hrvatski" *(First-ever Serbo-Croatian-into-English machine translation software)*. Investitor projekta je američki National Air Intelligence Center *(prethodna riječ nije slučajno podvučena)* sa sjedištem u američkoj zračnoj bazi u Daytonu, Ohio. Ugovor je "težak" 10 milijuna dolara!

Postoji opravdana bojazan da tuđi alati za prevođenje našeg jezika neće dosljedno slijediti osobitosti hrvatskog jezika. Ako sami ne sudjelujemo u realizaciji ove zadaće, moramo biti svjesni rizika da ćemo se pridružiti neslavnoj skupini malih naroda koji ne drže do sebe i koji su osuđeni na nestanak svojih nacionalnih obilježja.

S druge strane, uspješnim rješavanjem ove zadaće omogućit ćemo da hrvatska kultura svijetu postane lakše dostupna.

Nove tehnološke mogućnosti i stalna potreba da se ide naprijed omogućili su da ovaj iznimno opsežan posao u zadnjih par godina privede kraju i nekolicina domaćih entuzijasta (stručnjaka).



Tribina u Europskom domu (s lijeva na desno: dr. Dembitz, g. Batnožić, prof. dr. Silić i prof. Ranilović)

Prvo je u proljeće 1994. korisnicima *CARNet-a* postala javno dostupna usluga strojne provjere teksta - ***HAŠEK*** dr. Dembitza s Fakulteta elektrotehnike i računarstva iz Zagreba, a do kraja 1996. se na našem tržištu pojavilo još pet spelling-checkera za PC računala: ***HRVATSKA RIJEČ*** zadarskog poduzeća TTE d.o.o, ***HRVATSKI RAČUNALNI PRAVOPIS*** zagrebačkih autora prof. Ranilovića, prof. dr. Silića i g. Batnožića, ***HSC*** zagrebačkog autora g. Radovanovića, integrirani spelling-checker i tekst procesor ***LAPIS*** splitskog poduzeća Polar, te ***SITO*** zagrebačkog poduzeća FILOS d.o.o.

Rješenjem ovog, zapravo moralnog zadatka (naime zbog premalog tržišta ovo je komercijalno neprofitabilan posao u Hrvatskoj), stručnjaci su pokazali da "*ne moramo stalno gledati civiliziranom svijetu u leđa*". Oni su to vlastitim trudom i dokazali - ovaj složeni zadatak riješili su interdisciplinarnim radom informatičara i jezikoslovaca, ali nažalost bez pomoći odgovarajućih državnih institucija kako u organizacijskom tako i u financijskom smislu.

Problemi pri izradi hrvatskog spelling-checkera

U nastavku tribine bilo je govora o problemima s kojima su se stručnjaci susretali prilikom izrade hrvatskog spelling-checkera. U prvom redu to je nepostojanje odgovarajuće hrvatske jezične literature (niti u elektroničkom niti u tiskanom obliku). Tako da su autori, prema riječima prof. Silića, morali prvo sačiniti ***popis*** riječi služeći se raznovrsnim izvorima. Zatim su popisu dodavali ***opis*** riječi i na kraju su na osnovi popisa i opisa sačinili ***propis***.

Dosadašnje gramatike hrvatskog jezika su fenomenološke, pa nisu mogle pomoći autorima pri definiranju morfologije, što je osnova za računalno generiranje svih oblika riječi i time formiranje leksičke baze hrvatskog jezika. Informatički rječnik mora sadržavati ne samo natuknice, već i sve legalne gramatičke oblike kojih u hrvatskom ima izuzetno puno: 14 oblika imenica, 42 oblika pridjeva + 42 oblika komparativa + 42 oblika superlativa, 31 glagolski oblik itd.

Naime, najvažnije je prilikom izrade leksičke baze tako visokoflektivnog jezika kao što je hrvatski, gdje se riječ javlja u više oblika prema rodu, broju, padežu, licu, vremenu, načinu; bio rad na definiranju morfoloških modela *(za ovu primjenu se svi oblici moraju tretirati kao zasebne riječi, nasuprot engleskog, koji je izraziti primjer niskoflektivnog jezika, gdje riječ ima vrlo malo oblika)*. Nakon što su dobiveni precizni morfološki modeli koji su pridijeljeni kao dodatni atributi svakoj riječi u početnoj leksičkoj bazi, računalom su generirane riječi-izvedenice slijedeći zakonitosti pojedinih morfoloških modela.

Tako na primjer ***Hrvatski računalni pravopis*** ima 84.000 osnovnih riječi koje su, slijedeći zakonitosti oko 700 različitih morfoloških modela, dale ukupno oko 3 milijuna pojavnica. Informatički gledano, hrvatski jezik je oko 30 puta složeniji od engleskog, te je zadatak kompresije leksičke baze *(koja bi inače u nekomprimiranom obliku zauzimala oko 50 MB memorijskog prostora)* uspješno obavio g. Batnožić, dipl.ing., dobivši upotrebljivu indeksirano-komprimiranu leksičku bazu veličine oko 690 KB.

Pitanje novogovora

Kao korisnika strah me je novogovora jer nije nemoguće sljedeće: standardiziranje spelling-checkera na državnoj razini *(gdje bi se npr. odgovarajućom preporukom odredila uporaba nekog spelling-checkera u svim javnim djelatnostima)*. Ako bi taj "etalonski" spelling-checker bio novogovornog karaktera, novogovor bi mogao postati standardni jezik!

S druge strane, moramo priznati da je rad Bogoslava ŠULEKA *(Slovak rodom, srcem i dušom Hrvat - kako stoji na spomen-ploči)* na hrvatskom jezičnom fondu u prošlom stoljeću, djelomično danas prihvaćen kao standardni jezik, gdje je on doslovno izmišljao kovanice tražeći zamjene za razne tuđice. Stoga će vjerojatno sljedećim generacijama nešto ostati i od ovog današnjeg novogovora.

Prof. Ranilović: - "S metodološkog stanovišta u izradi spelling-checkera ne postoji problem novogovora. Postoji problem života jezika, odnosno, života nacije. Jezik kao iskaz jednog stanja društvene svijesti nacije jednako je toliko promjenjiv koliko su promjenjivi i uvjeti koji se u životu te nacije dešavaju. Jezik se mijenjao, mijenja se i mijenjat će se. S tog stanovišta niti jedan spelling-checker, kao niti jedan rječnik, nikad nije gotov. Što su društvene promjene veće, brže i dinamičnije, to će se i jezik toga naroda brže mijenjati. Nesumnjivo je da će se naš rječnik morati nadopunjavati jer taj posao nismo radili u jedno društveno stabilno vrijeme. Trudili smo se u njega staviti ono što se naziva standardni govor, dakle nije niti tehnološki usmjeren, niti znanstveno orijentiran nego opći govorni fond jezika u koji smo stavili ono što smatramo da je od novogovora već postalo dio tog standarda. Tu postoji doza subjektivne ocjene, ali mi priznajemo činjenicu da ćemo morati naš rječnik nadopunjavati, koliko se budu zapravo mijenjali društveni uvjeti."

Dr. Dembitz - "Neke novogovorne riječi, izrazito morfološke tvorbe, Hašek je svrstavao *(u nastavku pročitajte opis rada Hašeka koji može provjeriti i njemu nepoznate riječi - op.a.)* u najviše klase sumnje - jer nisu inherentne bazi iz koje Hašek crpi svoje jezično znanje o hrvatskom jeziku. Jednostavno - one nisu hrvatski jezik! S druge strane je primjer od prije dvije godine kad se pojavila *missica hrvatske neoleksike* - **suosnik** *(koaksijalni kabel - prev.a.)*. Tu riječ Hašek smatra mogućom hrvatskom riječi. Suosnik nije dio Hašekove leksičke baze, kao što nije ni zrakomlat jer naprosto nema tekstova koji to potvrđuju. Oni se nalaze samo u tekstovima koje treba uzeti s rezervom. Tako na primjer: **strojevina** za hardver, iako tvorbeno zadovoljava, nema adekvatnu opozicijsku kovanicu u hrvatskom jeziku jer semantički gledano, hardver funkcionira u engleskom u opoziciji sa softverom. Strojevina izolirana bez opozicijske kovanice u hrvatskom koja bi trebala biti jednako tako samorazumljiva kao odnos hardver/softver u engleskom, nema pravo egzistencije".

Prof. Silić: - "Kad je riječ o ovoj temi, treba reći da hrvatskim jezikom govorimo stoljećima, ne govorimo isključivo od jučer.

Hrvatski jezik treba shvatiti kao povijest i kao stanje, i povijest u stanju, odnosno, hrvatski jezik treba interpretirati u ime generacija.

Svaka prethodna ima sljedeću generaciju i ona prethodna mora ostaviti u nasljeđe sljedećoj generaciji. Kako se razvija društvo, tako se razvija i jezik, ali kako se u stvari sporo razvija društvo, tako se sporo razvija i jezik. Nema naglih promjena ni u društvu pa prema tome ni u jeziku, jer je jezik ono što prati društvo, i to u ovom trenutku treba imati u vidu. Prema tome sve nagle promjene koje se u govoru događaju nisu uvijek dobrodošle".

Spelling-checker ispravlja pravopisne i tipografske pogreške

Spelling-checkeri namijenjeni su prvenstveno za otkrivanje i ispravljanje ***pravopisnih i tipografskih pogrešaka*** u tekstovima. Za prve se može reći da su sustavne, a za druge da su slučajne pogreške. Pravopisne pogreške su najbliže riječima prirodnog jezika, dok su tipografske nešto dalje.

Koje su najčešće tipografske pogreške koje uspješno ispravljaju spelling-checkeri:

- *zamjena susjednih slova (npr. slovo - solvo)*
- *ispuštanje slova ili razmaka*
- *umetanje suvišnog slova ili razmaka*
- *dvije ili više istih riječi uzastopno*
- *pogrešno miješanje velikih i malih slova (npr. IVo)*
- *početak rečenice malim slovom.*

Nailaskom na nepoznatu riječ, spelling-checkeri, pored upozorenja, nude korisniku izbor najsličnijih riječi (onoj nepoznatoj) s ciljem njezine zamjene. Popis ponuđenih riječi najčešće se programski formira u tom trenutku i to po kriteriju za ispravak jednostruke pogreške. Spelling-checkeri ignoriraju nizove karaktera u tekstu koji nisu sastavljeni od slova (A-Ž) i apostrofa ('). Kombinacije slovnih nizova s brojkama ili simbolima prijavljuju kao pogrešku.

Na tribini su iznijeti podaci o vjerojatnostima da jedna tipografska pogreška pretvori početnu ispravnu riječ u drugu ispravnu riječ: u engleskom jeziku je oko 8%, a u hrvatskom jeziku, zbog svoje visoke flektivnosti, oko 12%.

Poseban problem su pogreške u leksičkoj bazi. One postoje i moraju se imati u vidu - spelling-checkeri nisu bezgrešni. Smatra se da leksička baza s oko 1 promil pogrešnih različnica, zadovoljava.

Po riječima prof. Silića, izrada rječnika sinonima (koja je u tijeku) omogućila je detekciju većeg broja pogrešaka u leksičkoj bazi ***Hrvatskog računalnog pravopisa***.



Odnos riječi i neriječi

Budući da ne postoji jednostavan sustav za provjeru leksičkih baza, svi autori se slažu da je izuzetno važan odaziv korisnika spelling-checkera i obavijesti o eventualnim pogreškama u leksičkoj bazi. Zauzvrat korisnicima obećavaju besplatnu distribuciju korigiranih leksičkih baza.

Kome je namijenjen spelling-checker

Spelling-checker nezamjenjivo je pomagalo za provjeru napisanih tekstova kako u svakodnevnom uredskom poslovanju tako i specijaliziranim primjenama, na primjer u DTP-u. Posebnu efikasnost pokazuju pri provjeri tekstova dobivenih OCR postupkom *(Optical Character Recognition - optičko prepoznavanje znakova na osnovi skeniranog tiskanog predloška)* na ispravljanju pogrešaka koje prave mnogi OCR programi zbog naših karakterističnih znakova.

Spomenimo još i rječnike sinonima (tezauruse) koji pomažu svima, a posebno prevoditeljima, u pronalaženju najpogodnije hrvatske riječi.

Hrvatski Akademski Spelling-Checker - HAŠEK

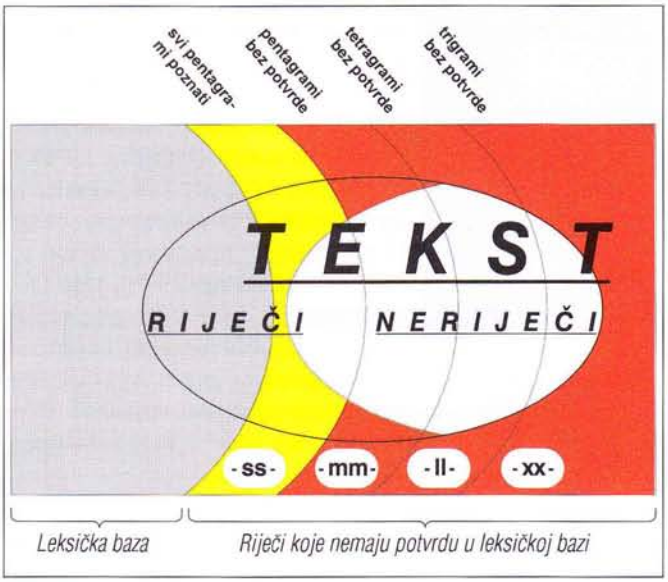
HACHECK (pohrvaćeno Hašek) je proradio u proljeće 1993., kada je strojno provjerena i prva knjiga pisana hrvatskim jezikom *(V. Bego, J. Butorac, ed.: "Josip Lončar - život i djelo", ETF-HAZU, 1993.)*. Njegov tvorac je dr. Šandor DEMBITZ, djelatnik Fakulteta elektrotehnike i računarstva iz Zagreba. Početna leksička baza derivirana je iz desne strane "Englesko-hrvatskog leksikografskog korpusa", kompjutorizirane verzije "Englesko-hrvatskog rječnika" R. Filipovića i suradnika te je imala oko 70.000 različnica.

Leksička baza dopunjava se različnicama iz tekstova koji stižu na obradu. Pored temeljne leksičke baze hrvatskih različnica, konstruirana je posebna leksička baza engleskih riječi (broji 70.000 različnica) te baza imena, kratica i sl. (broji 60.000 različnica), jer je praksa pokazala da su ove vrste riječi visoko zastupljene u hrvatskim tekstovima, naročito onima nastalim u akademskom (znanstvenom) miljeu.

Posebne leksičke baze, za razliku od temeljne, općejezično hrvatske, ne služe programu za deriviranje jezičnog znanja. Aktualna temeljna leksička baza broji 240.000 različnica. Sve one imaju potvrdu u tekstovima pisanim hrvatskim jezikom. Naučene su iz 1.400 tekstova ukupnog opsega od 8.000.000 pojavnica.

Hašek je namijenjen svima, javno je dostupan putem Interneta od proljeća 1994, a koristi se kao besplatna usluga raspoloživa na E-mail adresi: ***hacheck@fer.hr*** , ***Subject: provjera***, a u tijelo E-maila se upiše ili importira tekst. Ako je tekst izvorno nastao u nekom tekst-procesoru, treba ga eksportirati kao "glatki" txt-file, vodeći računa o eventualnoj konverziji znakova u kodnu stranicu CP437 (CROSCII). Hašek garantira diskreciju: tekstovi primljeni na provjeru brišu se po obavljenoj obradi. Engleske dijelove teksta (sažetak, literatura, citati izvornog nazivlja i sl.) ne treba brisati u pripremi za slanje jer je Hašek dobar i u engleskom.

Kao program se ne distribuira, niti postoji namjera da se tako nešto napravi. Hašek se svakodnevno mijenja, a stalno je na raspolaganju svojim korisnicima, odnosno, onima koji mu korisnicima žele postati.



Princip rada Hašeka

Njegovi mu korisnici daruju svoje znanje hrvatskog jezika, što im on vraća svojim strukturiranim upozorenjima na moguće pogreške u njihovom tekstu.

Hašekova temeljna leksička baza periodički se provjerava sa zadaćom da se njezina čistoća održi u granicama iznad 99,9%.

Hašek prepoznaje sve vrste tipografskih i pravopisnih pogrešaka. Tipografske pogreške u pravilu svrstava u klase visoke sumnje, dok neke pravopisne (ije/je), razvrstava po razdiobi sličnoj razdiobi nepoznatih riječi, što je deklarirana slabost programa. Takve je pogreške korisniku teže pronaći, makar su i one na check-listi.

Hašek provjerava tekst i formira izvještaj (check-listu) sa sumnjivim nizovima iz obrađenog teksta i putem E-maila automatski upućuje korisniku. Svakom slovnom nizu iz obrađivanog teksta, nizu koji nema izravnu potvrdu u leksičkoj bazi programa, Hašek asocijativnim postupkom pridružuje stupanj pripadnosti hrvatskom jeziku na osnovi analize trigrama, tetragrama i pentagrama.

Nepoznati se slovni nizovi razvrstavaju u 4 osnovne klase:

- `-xx-` ekstremno sumnjivi slovni nizovi koji sadrže trigrame bez potvrde;
- `-ll-` veoma sumnjivi slovni nizovi koji sadrže tetragrame bez potvrde;
- `-mm-` umjereno sumnjivi slovni nizovi koji sadrže pentagrame bez potvrde;
- `-ss-` gotovo nesumnjivi slovni nizovi koji sadrže pentagrame s potvrdom, ali sama riječ nije pronađena u leksičkoj bazi.

Pored navedenih osnovnih klasa, na check-listi pojavljuju se još 3 oznake klasa:

- `-kk-` odnosi se na kombinaciju slova i znamenaka u istom nizu (npr. A4);
- `-gg-` slovni nizovi s neuobičajenim rasporedom velikih i malih slova (npr. MicroSoft);
- `-cc-` kratki slovni nizovi (do 3 slova), koje Hašek svojim asocijativnim postupkom ne može razvrstati u osnovne klase (npr. HPT).

Činjenica je da program nepoznate riječi hrvatskoga jezika "gura" u klase niske sumnje, a pogreške u klase visoke sumnje, što olakšava pronalaženje pogrešaka. Vjerojatno je upravo zbog toga ova usluga, sa svim svojim manama, stekla popularnost potvrđenu stalnim porastom broja korisnika i opsega tekstova što se Hašeku šalju na provjeru. Spomenimo da je prva sustavna provjera leksikografske građe u slavenskim jezicima upravo izvršena pomoću Hašeka *(radi se o Hrvatskom leksikonu, glavnog urednika Dr. Vujića.)*

Hašek lako i brzo uči. Korisnik mu može poslati popis riječi za koje smatra da bi trebale biti u njegovoj leksičkoj bazi. Detaljnije upute možete dobiti upućivanjem E-maila na istu adresu: ***hacheck@fer.hr , Subject: uputa***.

U svojoj leksičkoj bazi ima novogovornih pojavnica samo u onoj mjeri u kojoj ga opravdavaju tekstovi pristigli na obradu.

Hašek je prilično brz - npr: 1 MB teksta obradi u 7 minuta i po brzini se može mjeriti s najbržim engleskim spelling-checkerima (program je instaliran na Sun SPARC radnoj stanici).

Planovi razvoja odnose se u prvom redu na proširenje leksičke baze koja će rasti kako je i do sada rasla, kroz obrade. Kada temeljna leksička baza prijeđe 300.000 različnica, treba očekivati neke promjene u postupku klasificiranja nepoznatog na sumnjivo (pogreške), odnosno, nesumnjivo (riječi).

Nadalje, uskoro će se pojaviti *suggest mod rada* za pogreške tipa -ije/je-, upravo radi svojstava već spomenute razdiobe ovog tipa pogrešaka na check-listi. Napravljen je brzi i djelotvorni algoritam (90-postotne točnosti), koji je u završnim fazama testiranja.

Sve specijalističke riječi (medicina, strojarstvo itd.), riječi su hrvatskog jezika i za njih ne trebaju specijalni rječnici. Loša je getoizacija leksika prema području uporabe. Ako pak neko ima potrebu da napravi spelling checker za svoje područje, neka se posluži općim alatom kakav je Hašek pri formiranju vlastite leksičke baze, a nju neka ugradi u ***ispell*** (freeware), kao što su već neki napravili za vlastite potrebe.

Za sada Hašeku nije planirano dodavanje rječnika sinonima i auto-korekcijski algoritmi opće namjene jer oni traže opsežnu temeljnu leksičku bazu, a i tada je pitanje jesu li na razini spelling checkera za hrvatski upotrebljivi *(Pogledajte raspored slova U, I i O na tipkovnici, te promislite koliko pridjeva tim slovima može završiti. Većina tipografskih pogrešaka koja pogađa završetak pridjeva na razini spelling checkinga, ne da se teorijski ni otkriti, a kamoli ispraviti).*

Strojno prevođenje bilo je Hašekovu autoru san otkada je na Hašeku počeo raditi (druga polovica 70-ih). I Hašek je nastao kao mali korak u ostvarivanju toga sna, da se pokaže da je ono za što se općenito misli da je nemoguće realizirati, ipak ostvarivo. Nitko nije vjerovao da se leksičkom bazom sumjerljivom bazama engleskih spelling checkera, mogu u razlučivanju riječi od neriječi (riječi pogođenih pogreškama) hrvatskog jezika, postići slični, praksom potvrđeni rezultati.

Asocijativne tehnike, na kojima se Hašek temelji, prodiru polako i u strojno prevođenje, polučuju zanimljive rezultate i te napore treba pratiti, skupa s tehnologijom koja stvara pretpostavke da te tehnike proizvode programe s primjerenim vremenskim odzivom. Bez toga, čeka nas ponor gubljenja identiteta jer već danas postoji bojazan da će hrvatski jezik, ako ne postane u sagledivo vrijeme strojno prevodiv, po "prirodnim" silama u sljedećem stoljeću nestati.

Nedostaci Hašeka: nije *user-friendly*, ne pruža komfor koji korisnik ima kada na osobnom računalu, u tekst-procesoru na koji je navikao, može pozvati spelling checker. Zahtijeva eksportiranje i slanje teksta te inverzne radnje po primitku check-liste, za što su mnogi korisnici nespremni. *(Taj posao se može automatizirati, npr. autor članka već pune dvije godine uspješno koristi Hašek tako da odgovarajućim macro-programom u Wordu, različitim bojama označi riječi u izvornom tekstu na koje je Hašek upozorio).*

Marketinških planova nema. Hašek je eksperiment s ciljem da se dobije uvid u to kako Hrvati pišu, bez ambicija da bude bilo kakav standard.

O nerazumijevanju hrvatske službene politike, i to kroz desetljeća, o strategijskoj važnosti strojne obrade hrvatskog jezika za Hrvate i Hrvatsku, Hašekov autor jednoga dana možda napiše knjigu. No, u njemu je još dovoljno istraživačkog naboja da bi svoje vrijeme tako trošio. Radije istražuje što se u svijetu na planu strojne obrade prirodnih jezika događa, provjerava koliko su tuđe ideje primjerene konkretnim hrvatskim prilikama, trudeći se da bude spreman za izazove ako se oni na ovim prostorima počnu artikulirati. Iskustva malih naroda govore da su proboji na planu strojne obrade nacionalnog jezika mogući samo ako je država za to ozbiljno zainteresirana. Pojava velikog broja hrvatskih spelling checkera u prošloj godini kazuje samo da bi se zainteresirana država ovdje imala kome obratiti, a da njezini zahtjevi ne ostanu tek puka želja.

Kontakt adresa:
Dr. Šandor DEMBITZ, Fakultet elektrotehnike i računarstva, Unska 3, 10000 Zagreb
E-mail: sandor.dembitz@fer.hr

Hrvatska riječ



g.Krešimir PINJATELA***HRVATSKA RIJEČ*** zadarske tvrtke TTE Inženjering d.o.o. prvi je hrvatski spelling-checker s tezaurusom sinonima, i razlikovnim rječnikom s elementima gramatičke provjere teksta. Djelo je g. Krešimira PINJATELE (apsolventa Fakulteta strojarstva i brodogradnje) i g. Branimira PINJATELE (diplomiranog inženjera brodogradnje). *Hrvatska riječ* program je koji se može koristiti s Microsoft Word 6.0, Windows Write, Lapis 2.0 i nekim drugim programima u Windows okruženju *(Notepad, CorelDraw itd.)*.

Nakon pokretanja pojavljuje se *Hrvatska riječ* u izborniku programa za pisanje kao dodatna stavka, tako



da se *Hrvatska riječ* koristi izravno iz njih, bez obzira koji i kakav spelling-checker već sadrže. Podržava hrvatska slova prema standardima CP1250, CP852, ISO8859-2 i CP437 (CROASCII). Spell-checker je integriran i u Newsroom Unisys-ov program za novinsko nakladništvo.

Hrvatska riječ namijenjena je korisnicima PC računala, radi pod Windows 3.1x i Windows 95 *(za sada samo sa 16-bitnim programima, npr. Word 6.0)*, a postoji i verzija za DOS za gotovo sve najpoznatije svjetske tekst procesore. Instaliran zauzima na disku oko 2 MB.

Priručnik uz Hrvatsku riječU paketu je solidno koncipiran priručnik (36 stranica) i dvije instalacijske diskete koje su britanskim programom *CopyControl* potpuno zaštićene od neovlaštenog kopiranja *(vjerujte: niti VgaCopy ni TeleDisk ovdje ne pomažu!)*. S druge strane, ovlašteni korisnici putem telefonske podrške mogu povećati broj licenci programa ili skinuti ograničenja ugrađena u demo verziju progama (upisivanjem dobivenih kodova poduzeća TTE d.o.o.), ako su za to stekli uvjete.

Na disketama se nalazi i test uzorak teksta koji na ilustrativan način demonstrira korisnicima brojne mogućnosti *Hrvatske riječi*.

Hrvatska riječ u potpunosti je hrvatski proizvod. Program, računalnu obradu hrvatske gramatike, te sažimanje rječničke baze riješio je g. Krešimir PINJATELA. Programerski dio posla trajao je, uz puno radno vrijeme i trajan produžni rad programera, tri godine. Usporedno su stvarane i obrađivane rječničke baze i hrvatski tezaurus.

Idejni projekt "Hrvatska riječ" započet je još prije uspostave demokracije i slobodne Hrvatske, a realizacija je počela prije četiri godine. Zamišljen je kao protuteža "hrvatskosrpskim" rječnicima i pravopisima (Pravopis hrvatskosrpskog jezika, Zagreb-Novi Sad 1960., Pravopisni priručnik hrvatskoga ili srpskoga jezika, Zagreb 1986.). Oslonac je bilo teško naći. Počelo se od Hrvatskoga pravopisa tzv. "Londonca".: Razlike između hrvatskoga i srpskoga književnog jezika, (Zagreb 1940.) Guberine i Krstića, te Brodnjakovog Razlikovnog rječnika srpskog i hrvatskog jezika (Zagreb 1991.).

Budući da je podloga za tu vrstu programa rječnik, postavljen je cilj da rječnik mora obuhvatiti što veći broj hrvatskih riječi. Ali, taj rječnik ne smije biti nekritičan prema riječima koje su iz političkih razloga i pod utjecajem hrvatskih jezikoslovaca "vukovaca" bile nametnute u skoro sve hrvatske i hrvatskosrpske rječnike tiskane u ovom stoljeću. Korisniku je trebalo pružiti na lagan način - automatski, koristeći prednosti računalne tehnologije, informaciju o mogućim zamjenama za neke riječi.

Rječnik je dopunjen osobnim imenima, toponimima, osnovnim stručnim riječima iz raznih disciplina: likovne umjetnosti, glazbe, matematike, fizike, kemije, strojarstva, brodogradnje, elektrotehnike, informatike, medicine, teologije...

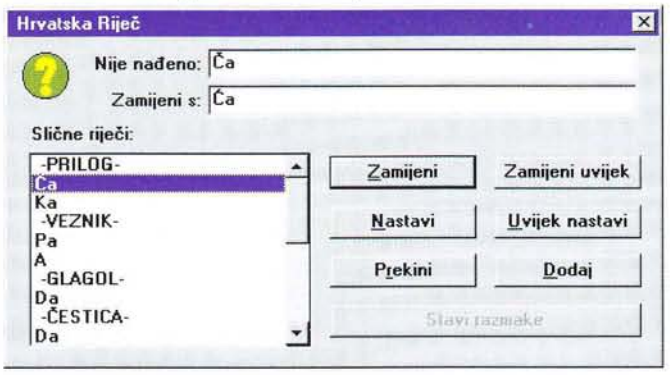
Za izradu tezaurusa (rječnika hrvatskih sinonima) autori nisu imali uzor, jer za hrvatski jezik, koliko je znano, nema tiskanoga rječnika sinonima. Oblici svih promjenjivih riječi stvaraju se prilikom pretraživanja rječnika. Za izradu rječnika sa svim oblicima poslužio je posebni pomoćni program GRAMATIKA. Takav pristup omogućio je sažimanje cijeloga rječnika u svega 324 KB.

Spell checker je provjeravan u radu na uređivanju i pripremi za tisak jednoga mjesečnika, uključujući i ispravljanje tekstova nakon obrade OCR. Testiranje programa je poslužilo i dodavanju nepoznatih riječi radi njihove obrade i uključivanja u glavni rječnik.

Programski paket *Hrvatska riječ* sastoji se od sljedećih programskih modula:

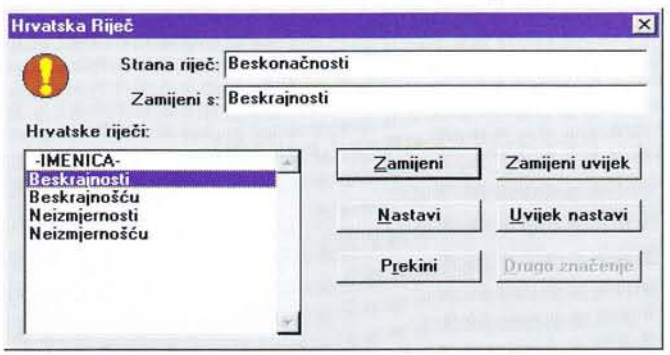
1. Hrvatski rječnik

Rječnik sadrži oko 55.000 najčešće korištenih riječi i sve oblike svih promjenjivih riječi. Ukupni broj riječi i njihovih oblika je oko 600.000. Dodatno, korisnik može sam dodavati riječi u vlastiti rječnik.



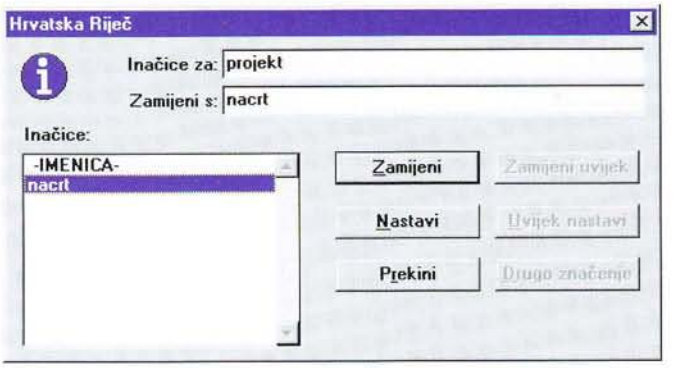
2. Razlikovni rječnik

Razlikovni rječnik je povezan s pretraživanjem osnovnoga rječnika, pa za stranu riječ (srbizme i neke druge tuđice, te riječi koje su zamijenile potisnute hrvatske riječi) automatski nudi hrvatske riječi. Ima 8.500 natuknica što čini oko 93.500 različitih oblika stranih riječi.



3. *Hrvatski tezaurus*

Rječnik hrvatskih inačica s oko 20.000 natuknica (što čini oko 220.000 različitih oblika riječi) vrlo je bogat istoznačnicama i sličnoznačnicama za svaku riječ. Treba naglasiti da se tu radi o prvom rječniku hrvatskih istoznačnica uopće. Odgovara onome što je poznato pod nazivom ***thesaurus*** kao dio mnogih svjetski poznatih tekst procesora.



4. Neke funkcije gramatike



Hrvatska riječ provjerava pisanje velikih i malih slova u riječi i rečenici, pogrešnu uporaba prijedloga *s/sa*, *k/ka*, veznika *da* u upitnim česticama (*da li ste*, umjesto *jeste li*), i *da* s prezentom.

Prepoznaje najčešće tipografske pogreške kao npr. pomiješana velika i mala slova, kratice bez točke, zamjena susjednih slova, izostanak razmaka iza znaka interpunkcije,



ispuštanje slova, ponavljanje iste riječi, ije/je. Po završetku provjere teksta, Hrvatska riječ daje statistički prikaz o "obavljenom poslu".

Ukratko, prednosti *Hrvatske riječi* su sljedeće:

- prepoznaje izuzetno velik broj riječi
- automatski razlikovni rječnik, kao hrvatski jezični savjetnik
- hrvatski tezaurus
- neke funkcije gramatike

Razvojni planovi za blisku budućnost:

- proširenje rječnika i tezaurusa,
- praćenje promjena hrvatskoga pravopisa i stvaranja hrvatskoga jezičnog standarda,
- razvoj strukovnih rječnika,
- otvaranje WEB stranice s koje će se moći "skinuti" nove inačice i demo.
- prilagoditi ga Wordu 7.0 i drugim 32-bitnim programima, s tim u vezi spomenimo da je prilikom pisanja ovog članka, isprobana *Hrvatsku riječ* i na 32-bitnim aplikacijama u Windows 95 okruženju i tvrdimo da su potrebne preinake toliko minorne, da bi se mogle nazvati kozmetičkim. Drugim riječima, Hrvatska riječ je za MS Office 95 *gotovo potpuno funkcionalna!*

Po riječima g. Pinjatele potrebna je koordinacija mjerodavnih državnih institucija za stvaranje hrvatskoga jezičnog standarda, kao i sufinanciranje razvojnoga projekta.

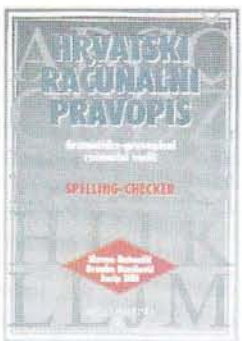
Projekt "Hrvatska riječ" rađen je bez ičije financijske potpore, samo iz osobnog entuzijazma i interesa za promicanje hrvatske samosvijesti preko djelatnoga doprinosa standardizaciji jezika.

Cijena programskog paketa *Hrvatska riječ* je 560 kn za jednog korisnika (pravne osobe) uz popust za više korisnika, a 299 kn za pojedince.

Kontakt adresa:

TTE Inženjering, d.o.o., Put Murata 22c, 23000 ZADAR
tel.: 023/435-553

Hrvatski računalni pravopis



*Priručnik Hrvatski računalni pravopis***Hrvatski računalni pravopis** djelo je zagrebačkih autora prof. Branka RANILOVIĆA *(diplomiranog psihologa, direktora tvrtke Infocentar)*, prof. dr. Josipa SILIĆA *(redovitog profesora na Katedri za hrvatski standardni jezik Odsjeka za kroatistiku Filozofskoga fakulteta Sveučilišta u Zagrebu)* i g. Slavena BATNOŽIĆA *(diplomiranog inženjera elektrotehnike i suvlasnika informatičke tvrtke MacroSoft)*

Prema riječima autora, proteklo je više od šest godina od početnih zamisli i računalnih gramatičkih modela do izrade *Hrvatskog računalnog pravopisa* u sadašnjem obliku. Leksička baza na razini natuknica dobivena je iz Anićevog *Rječnika hrvatskog jezika* iz 1991. U prvoj je fazi Anićev rječnik radikalno čišćen, jer je taj rječnik u smislu "hrvatske leksičke discipline" dosta slobodan, a u drugoj je fazi radikalno nadopunjavan raznim pomoćnim programima za tvorbu riječi.

Baza se dopunjavala i iz drugih rječnika, kao i testiranjem brojnih tekstova. Dragocjeno pomagalo za izradu morfoloških modela bio je rječnik profesora Josipa Matešića, takozvani *odostražnji rječnik*, gdje su riječi poredane abecednim redom s desna na lijevo *(tu su npr. svi infinitivi na jednom mjestu)*.

Sadašnji obujam leksičke baze je oko 3.000.000 riječi koje su generirane računalom na osnovi 84.000 natuknica slijedeći zakonitosti od oko 700 morfoloških modela.

Kako se u zadnje vrijeme dodavanjem novih natuknica nije povećao broj morfoloških modela, autori smatraju da je hrvatski jezik u tvorbenom pogledu do kraja opisan s oko 700 morfoloških modela *(autori očekuju da ga ni budući radovi na specijaliziranim rječnicima neće značajno mijenjati)*. Međutim, za potpuno prekrivanje hrvatskog jezika, autori smatraju da je potrebno oko 250.000 natuknica, odnosno, leksička baza od gotovo 10 milijuna riječi!

Za sad postoje dvije verzije *Hrvatskog računalnog pravopisa* za MS Word 6 i 7, a u tijeku dovršetka je verzija i za MS Word 8 (Word za MS Office 97) za koji je MicroSoft otkupio prava za hrvatsko tržište. Instaliran zauzima na disku oko 2,2 MB.

> Posebno je ugodno raditi s *Hrvatskim računalnim pravopisom* u Word-u kada je aktivna opcija **Automatic Spell Checking** *(Tools...Options...Spelling...Automatic_Spell_Checking)* jer se tada korisnik već tijekom psanja upozorava da je napravio griješku ili napisao dvije iste riječi riječi za redom.

Korisnici *Hrvatskog računalnog pravopisa* dobivaju u paketu: Gramatički tezaurus *(nezavisni programski alat koji za svaku unesenu riječ generira sve njezine gramatički ispravne oblike i knjižicu uputa: Hrvatski računalni pravopis)*, Speller i Hyphenator za Word 6.x i Word 7.x, priručnik s pravopisom, primjerima morfološke tvorbe i pravilima poslovnog dopisivanja (oko 150 str.)

Razvijen je i alat za pretvorbu bilo koje leksičke baze (npr. stručni medicinski rječnik) u *.lex oblik koji se brže pretražuje od korisničkog rječnika *(custom.dic)*, a sadrži sve morfološke oblike riječi.

Sljedeći planirani koraci su: izrada rječnika sinonima *(tezaurusa)* s oko 150.000 riječi i Grammar Checker *(u leksičku bazu su već ugrađeni atributi za gramatičku provjeru)*. Također je u izradi MS-Word 6.0. verzija za računala Apple MacIntosh.

Spomenimo da uskoro izlazi morfološki rječnik hrvatskog jezika u izdanju Matice hrvatske pod nazivom *Gramatika hrvatskog rječnika* (obujma oko 600 stranica), što je bila metodološka osnova za potpuno objašnjenje tvorbe 3.000.000 riječi.

Autori su tražili pomoć u koordinaciji istraživanja na ovom polju od raznih državnih institucija, ali na žalost bezuspješno. Rad su prijavili i Ministarstvu znanosti, od kojeg nisu dobili niti odgovor. Žalosno je da hrvatske znanstvene i državne institucije ne vide interes u takvim projektima koji se u svijetu smatraju strateškim istraživanjima na nacionalnoj razini.

Nabavna cijena Hrvatskog računalnog pravopisa je: 470 kn (132 DEM ili 86 USD)

O ostalim pojedinostima o Hrvatskom računalnom pravopisu pročitajte u posebnom članku.

Kontakt adresa:

josip.silic@ffzg.hr, tomislav.silic@public.srce.hr
URL: http://public.srce.hr/~ansilic/

HSC (Hrvatski Spelling-Checker)



HSC (Hrvatski Spelling-Checker) g. Branka RADOVANOVIĆA iz Zagreba, studenta četvrte godine FER-a, nastao je u lipnju 1995., i to prilagodbom jednog autorovog programa koji je čitao tekst i prepoznavao riječi koje nisu pronađene u datoteci s listom riječi. Leksička baza stvarana je doslovno od nule, dopunjavana je provjerom svih hrvatskih tekstova koji su autoru bili pri ruci (od analize vlastitih tekstova, tekstova s Interneta preko ručnog upisa riječi iz pravopisnih priručnika, enciklopedija, knjiga itd.).Na taj način leksička baza je za mjesec dana imala više od 40 tisuća riječi, a za šest mjeseci već oko 215 tisuća riječi.

Prva javno dostupna verzija HSC-a (HSC 1.00) pojavila se na Internetu u lipnju 1996. i raspolagala je s oko 325 tisuća riječi.

Leksikografska, pravopisna i gramatička verifikacija HSC-a oslanjala se na "Hrvatski pravopis" *(Babić - Finka - Moguš)*, "Pravopisni priručnik hrvatskoga ili srpskoga jezika" *(Anić - Silić)*, te tzv. Veliku gramatiku *("Povijesni pregled, glasovi i oblici hrvatskoga književnog jezika", Babić et al)*. Verifikacija stranih, osobnih i zemljopisnih imena izvršena je spelling-checkerima za engleski jezik.

Leksička baza aktualne verzije (HSC 1.60) ima oko 455 tisuća riječi i sadrži sve gramatičke oblike riječi, kratice, domaća i strana osobna imena i prezimena, te izvedene posvojne pridjeve. Zastupljena su uobičajena imena i prezimena, uz ograničenja koja nameće raspoloživi skup znakova, domaći i strani toponimi, razne druge vlastite imenice, uključujući imena institucija, zaštićena imena i neke strane riječi koje se u hrvatskom pišu ili se mogu pisati u izvornom obliku (npr. jazz). Instaliran zauzima na disku oko 700 KB.

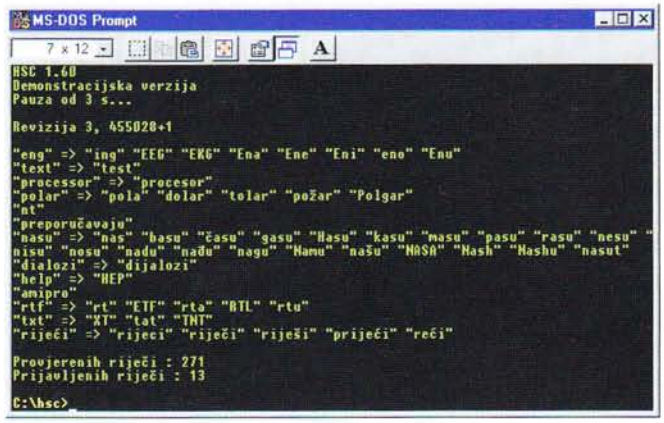
HSC radi pod DOS-om, čita isključivo obične ASCII tekstove u kodnim rasporedima CP 852, CP 1250, CP437 (CROSCII) ili ISO 8859-2, koje može automatski prepoznati. Osim riječi koje ne prepoznaje, HSC prijavljuje i pogrešno pisanje velikih slova, i to u četiri stupnja strogosti, prema izboru korisnika. Osim toga, HSC po želji korisnika može pronaći i sugerirati ispravne riječi slične pogrešnoj, i to tako da uzima u obzir i česte pogreške pri tipkanju (izostavljanje/ubacivanje slova, zamjena dvaju slova, i sl.). Vjerujte, ta opcija je izuzetno zanimljiva odraslima i djeci *(mala Dunja, kći autora ovog članka, stara 5 godina uživa izmišljati pogrešne riječi i gledati odziv HSC-a)*.

Odnos prema "novogovoru", odnosno "novopravopisu" je relativno konzervativan. Intenzitet i karakter novijih previranja u jeziku ne samo što često ne ulijeva povjerenje u njihovu održivost, već i praktički onemogućuje njihovo praćenje, tako da se rječnik u tom smislu uglavnom oslanja na "čvrste točke" jezika u leksikografskom smislu i pravopisa.

Za provjeru jednog tipičnog teksta od 5000 riječi *(petnaestak stranica s dvostrukim proredom)* HSC-u je, na 486DX2-66 računalu potrebno svega oko 10 sekundi.



HSC-ove opcije (help-screen)



HSC u DOS prozoru u suggest modu rada

HSC uspješno radi i u najskromnijim memorijskim uvjetima, već i na PC računalima s 512 kb RAM-a, i to praktički bez utjecaja na brzinu rada.

HSC je zamišljen univerzalno: nije vezan ni za koji tekst-procesor, kao što nije vezan ni za koji kodni raspored. Njegova jedinstvena mogućnost je provjera tekstova koji ne sadrže naše dijakritičke znakove, što bi uobičajenim načinom provjere bilo vrlo nepraktično. Mogućnost provjere HTML dokumenata, uz ISO 8859-2 podršku, čine ga vrlo pogodnim za pravopisnu provjeru hrvatskih tekstova na World Wide Webu.

Kod provjere "reprezentativnih" tekstova na hrvatskom jeziku (npr. novinskih članaka), efikasnost leksičke baze iznosi tipično od 98 do 99.5%. To znači da broj *ispravnih* riječi koje *nisu* prepoznate, iznosi od 0.5 do 2% u odnosu na broj ispravnih riječi u promatranom tekstu.

HSC omogućuje da korisnik nadopunjuje glavni rječnik. Dodane se riječi smještaju u zasebnu tekstualnu datoteku (čitljivu u bilo kojem tekst editoru): *PERS.DCT*. Njihov broj je ograničen samo raspoloživom memorijom; na PC računalima s 600 KB slobodne konvencionalne memorije broj riječi u pomoćnom rječniku može premašivati 10000.

HSC ima dva glavna nedostatka: nema "pravo" interaktivno korisničko sučelje i zasad još ne može čitati MS Word format, koji je u današnje vrijeme postao *de facto* standard na PC računalima.

Planovi razvoja HSC-a su ponajprije vezani uz otklanjanje spomenutih nedostataka, naravno uz stalan rad na proširenju leksičke baze. Budući da je rječnik zamišljen i razvijan obuhvatno, tj. tako da u načelu sadrži i stručne riječi iz svih grana znanosti i tehnologije *(u onoj mjeri u kojoj ih je bilo moguće prikupiti)*, razvoj zasebnih, specijalističkih rječnika nije u planu. HSC podržava dopunu rječnika, tako da je omogućeno prilagođivanje leksičke baze vlastitim specifičnim potrebama.

U planu je sljedeće: automatsko ubacivanje naših znakova u tekstove koji ih nemaju, provjera interpunkcije, verzije za druge operativne sustave (Linux i MS Windows).

Pred realizacijom su one mogućnosti koje drugi hrvatski spelling-checkeri ne nude, a to su enigmatske primjene *(pomoć pri rješavanju križaljki na osnovi pretraživanja leksičke baze regularnim izrazima, anagrami i sl.)*, generiranje rima za stihoklepce, edukativne primjene itd.

HSC je javno dostupan u, zasada, ograničenoj demo verziji, koja je *freeware*, a u planu je i distribucija neograničene komercijalne verzije. Može se preuzeti na Internetu, na lokaciji

ftp://gnjilux.cc.fer.hr/pub/msdos/ util/hscXXX.zip

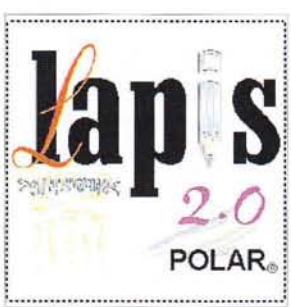
gdje je XXX oznaka aktualne verzije (u ovom trenutku 1.60).

Gospodin Radovanović je, osim svega, autor izvrsnog programa ***WCNT105*** *(freeware)* koji služi za filtriranje tekstova i njegovu statističku analizu (statistička analiza pojavnica/različnica, ispis histograma s frekvencijama učestalosti i drugo), a može se naći na Internetu u Simtelovoj bazi programa za MS-DOS. Program WCNT znatno je olakšao rad autoru ovog članka prilikom analize testova kojima su testirani ovdje opisani spelling-checkeri.

Gospodin Radovanović ističe dva ključna problema u odnosu javnosti prema problematici razvoja hrvatskih spelling checkera. Prvi je neprepoznavanje činjenice da program za strojnu provjeru tekstova na hrvatskom jeziku nije samo komercijalno i/ili tehnološko, "stručnjačko" pitanje, već prvenstveno *pitanje jezične i komunikacijske kulture jednog suvremenog društva*. Drugi leži u općenito nedovoljnom razumijevanju i poštovanju intelektualnog rada i intelektualne svojine. Na sreću, razvoj softvera te namjene je u Hrvatskoj već dovoljno poodmakao da se može reći da će temeljni problem provjere hrvatskih tekstova u skoroj budućnosti biti riješen, tako da preostaje uglavnom rad na nadgradnji, kao što je provjera gramatike/sintakse/stila, rječnik sinonima i sl.

Kontakt adresa:
Branko RADOVANOVIĆ
br31187@pinus.cc.fer.hr ili brad@fly.cc.fer.hr.

Lapis



LAPIS je program za obradu teksta u cijelosti na hrvatskom jeziku i jedini program za obradu teksta koji u sebi ima ugrađenu provjeru pravopisa za hrvatski jezik. Lapis je ponosno djelo firme Polar d.o.o. iz Splita. Trenutačno je dostupna verzija 2.0 za Win 3.x / Win 95, a u fazi beta testiranja je 32-bitna verzija - Lapis 3.0

Izrada samog spell-checkera intenzivno je počela krajem 1995., s početnom leksičkom bazom od oko 40.000 riječi, na osnovi koje su se po morfološkim pravilima računalom generirali ostali oblici riječi. Sadašnji opseg Lapisove leksičke baze je oko 368.000 riječi. Lapis 3.0 imat će proširenu leksičku bazu na oko 500.000 riječi.

Svi dijelovi Lapisa su na hrvatskom: naredbe, dijalozi, pomoć (help), knjiga, instalacijska procedura i ostalo. Do najsitnijih detalja vodilo se računa o specifičnostima našeg jezika; kodne stranice, automatska korekcija teksta, konvertiranje iz svih u sve kodne stranice.

Kupovinom programa Lapisa 2.0 dobiva se knjiga-priručnik (81 str.), 4 instalacijske diskete i trajna telefonska



Suggest mod rada prilikom pravopisne provjera teksta tekst procesorom Lapis

podrška za korisnike. Zanimljivo je da se po instalaciji programa mora telefonski kontaktirati firma Polar u Splitu, da bi se dobila odgovarajuća šifra s kojom se prvi put pokreće Lapis. Radi se o sofisticiranoj zaštiti programa od neovlaštenog kopiranja na principu vrlo sličnom PGP-u (javni ključ je serijski broj poznat odmah po instalaciji, a tajni ključ se dobiva na osnovi javnog - od firme Polar).

Minimalni konfiguracija računala na kojem se može pokrenuti LAPIS 2.0 je: računalo s procesorom 386 ili jačim, 4 MB RAM memorije i oko 7 MB prostora na tvrdom disku. Sve iznad toga omogućuje ugodniji rad s LAPIS-om. LAPIS radi na operativnom sistemu Windows 3.1; Windows 3.11; Windows 95; Windows NT. Preporučuju se Windowsi koji su prilagođeni za našu zemlju *(Windows CE, Pan Europe)*. LAPIS koristi sve fontove koji su ispravni i propisno instalirani na Windowsima. Čita MS Word format datoteka (2.0, 6.0), WordPerfect, AmiPro, WordStar i RTF format. Zapisivanje se obavlja u RTF, MS Word 2.0 i TXT format.

Posebno je zanimljiva *automatska korekcija teksta* koja je ugrađena u LAPIS. Njome se detektiraju pogreške koje drugi spelling-checkeri ne detektiraju (npr. *nemogu / ne mogu - ne* i *mogu* su, odvojeno napisane, gramatički ispravne riječi) ili naprosto ubrzavaju pravopisnu provjeru (šema - shema, indetificirati - identificirati, najednostavniji - najjednostavniji, i slično). Takvih parova najčešćih mogućih pogrešaka Lapis ima dvadesetak, a korisnici ih mogu jednostavno dopisivati. Ugrađena je dodatna funkcija automatskog prepravljanja prvog slova rečenice u veliko slovo.

Leksička baza je provjeravana ručno i/ili posebnim programima za provjeru riječi na principima heurističkih algoritama.

Neprepoznate riječi korisnici mogu jednostavno dodavati u Lapisov *custom.dic*, koji je jednostavna tekstualna datoteka, stoga se može editirati i *off-line*.

Izrađen je specijalistički rječnik za liječnike, a u planu je izrada dodatnih rječnika za engleski i njemački jezik. Za sada se ne planira izrada modula za sintaksnu i gramatičku provjeru integralnog teksta.

Još neke zanimljive karakteristike tekst procesora Lapis su:

- Sortiranje teksta uvažava redoslijed naših slova u potpunosti, čak i slova LJ i NJ.
- Konverzija teksta po svim kod nas upotrebljavanim kodnim stranicama.

- Sve funkcije formatiranja teksta *(mijenjanje oblika pisma, veličine stila i boje; centriranje, poravnavanje i uvlačenje odlomaka; rubovi oko teksta; sjenčanje odlomaka teksta; zaglavlja i podnožja; unos specijalnih simbola; unos slika raznih formata).*
- Poništavanje posljednje operacije nad tekstom u neograničenom broju koraka.
- Lapis podržava sofisticirane standarde za razmjenu podataka.
- Traženje datoteka na diskovima.
- Jednostavno kreiranje tablica.
- Ugrađeni program za crtanje grafikona (PolarGraf).
- Brojanje riječi i statistika teksta.
- Automatsko spremanje dokumenata i njihovo obnavljanje u slučaju rušenja operativnog sustava

Cijena programskog paketa LAPIS 2.0 za pravne osobe (poduzeća) je 1125 kn, a za fizičke osobe (građane) 1125 kn + porez 26,5% = 1423 kn. Kupovina programa može se izvršiti kod poduzeća POLAR ili njihovog ovlaštenog distributera.

I na kraju, Lapisov tim poručuje da je vrlo otvoren za suradnju, bilo ona stručne ili financijske prirode.

Kontakt adresa:

Polar d.o.o, Iločka 21, 21000 Split
tel./fax.:021-520599, 098-213-398
E-mail: polar@st.tel.hr
URL: http://www.dalmatianet.com/polar
http://www.polar.hr (uskoro)

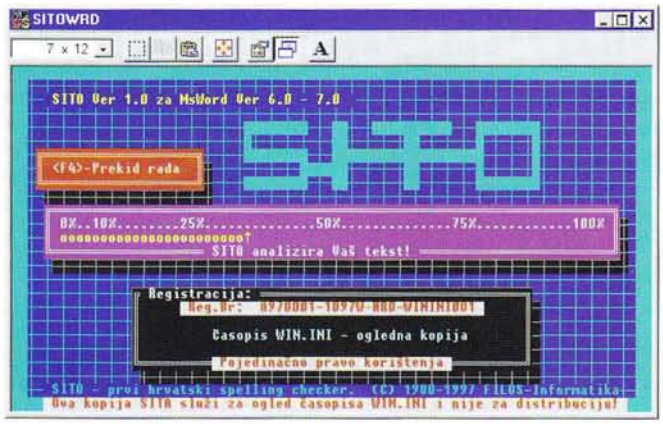
SITO



Autor *SITA* je prof. dr. Miroslav KRŽAK (profesor Filozofskog fakulteta Sveučilišta u Zagrebu). Njegov rad na izradi spelling-checkera može se smatrati pionirskim (naime teza doktorske disertacije obranjene na ETF-u 1983. bila je *Mogućnosti ispravljanja i uređivanja teksta na hrvatskom književnom jeziku.*Autor se počeo baviti izradom spelling-checkera za osobna računala još davne 1986., a javnosti je predstavljena radna verzija SITA za provjeru *MS Word for Windows* dokumenata početkom 1997. Autor je osnivač firme Filos d.o.o. čija jedna od grana djelatnosti je jezična: stvaranje morfološke baze hrvatskog književnog jezika pomoću programa MORF, te izrada programa vezanih za pretraživanje i korekturu tekstova - SITO.

Nakon pokretanja instalacijskog programa koji se nalazi na disketi (DOS program), na računalo se instaliraju sve potrebne datoteke (oko 520 KB). Zatim treba otvoriti dokument *sitoinst.doc* koji instalira dva WinWord macro-programa u normal.dot (SitoExp i SitoForWord, duljine oko 8 KB). Od tada je SITO uporabno i može se pozvati putem odgovarajućeg izbornika u Word-u.

Iniciranjem naredbe *Prosij!* započinje provjera *(cijelog)* teksta. SITO pokreće DOS program na čije izvršenje moramo čekati.



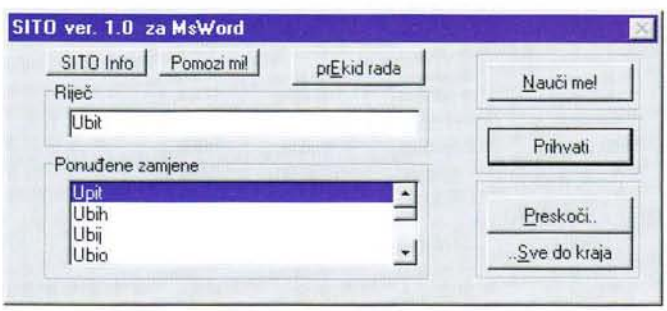
Prva faza SITO spelling-checkera izvršava se u DOS prozoru

Nakon toga slijedi pokretanje macro-programa koji prolazeći tekstom obavještava korisnika o rezultatima provjere - nudeći standardne spelling-checking opcije: *prihvati riječ, preskoči riječ, nauči riječ, te predlaže zamjene riječi.*

Program SITO ima vrlo solidnu leksičku bazu, ali neodgovarajuće programsko okruženje. Naime, jezgra SITA je DOS program koji obavlja sav posao oko pretraživanja leksičke baze, dok rezultate spelling-checkinga korisniku daje odgovarajući WinWord-macro program. Negativne strane su sljedeće: tekst se može provjeriti samo kao cjelina *(nije moguće analizirati samo dio teksta)* i dodatno, analiza teksta se ne može nastaviti od mjesta gdje je zaustavljena *(mora se ponovno pokrenuti analiza cijelog teksta).*

Dodatno, postoji ograničenje u dodavanju novih riječi (koje bi upisivali korisnici) Sito može naučiti samo oko 1000 novih različnica. Autori su doskočili navedenom ograničenju na taj način da nude korisnicima integriranje novih različnica zajedno s njihovim izvedenicama u postojeću leksičku bazu (o tome vodi brigu dodatni macro-program koji se pokreće unutar izbornika Sita, a on zapisuje nove riječi na disketu koja se mora poslati u Filos). Nakon učitavanja novog rječnika, SITO može učiti sljedećih 1000 različnica.

Leksička je baza SITA formirana iz različitih izvora, a najopsežniji izvor bio je hrvatski dio Drvodelićeva Hrvatsko-engleskog rječnika koji je (rječnik) dopunio prof. Željko Bujas. Također su ugrađene i riječi iz korpusa od oko 700.000 pojavnica koji je izradio jedan danski stručnjak i stavio ga na Internet *(ftp://aau.dk/pub/slav/ - op.a.).*



Druga faza SITO spelling-checkera izvršava se u Word-u

Konzultirani su i pravopisni rječnici (Babić, Moguš, Finka), te je rječnička baza dopunjena iz Rječnika hrvatskoga jezika prof. Vladimira Anića. Konačno, korišteni su hrvatski tekstovi HINE, Narodnih novina te drugih izvora koji se mogu naći u CARNetu.

Za generiranje svih morfoloških oblika neke riječi napravljen je poseban program MORF *(ne distribuira se korisnicima Sita)*, čime je formirana leksička baza od oko 1,080.000 riječi na osnovi oko 90.000 natuknica.

Nije pravljena nikakva razlika u riječima prema njihovu porijeklu, već su upisivane sve riječi koje realno čine hrvatski jezik kao što se to čini i u drugim kulturnim jezicima npr. engleskome. "Strana" riječ koja se rijetko koristi pisana je etimološki, a često korištena fonetički (hardver, softver). Toponimi, osobna imena, česte kratice i akronimi također su upisani u rječnik SITA.

SITO je namijenjeno za MS Word 6.0 i 7.0 i Code Page 1250 koji je podržan u Windows 95 Pan Eureopean verziji, te Windows 3.x Eastern European (EE). U završnom ispitivanju je verzija SITA 2.0W, koja je namijenjena 32-bitnim operativnim sustavima (Windows95 i NT). Planira se verzija za MacIntosh koja bi izašla paralelno sa SITO 2.0W inačicom. Također je moguće naručiti verziju za UNIX i druge operativne sustave. Postoji i stara verzija za DOS.

SITO prepoznaje sve tipkarske pogreške osim onih koje su se pretvorile u nove ispravne riječi izvan konteksta. SITO radi isključivo na principu prosijavanja riječi i nema ugrađenih algoritama za samoučenje. Novogovor ulazi u rječničku bazu poslije relativno kratkog 'praćenja' trajnosti novih riječi.

Filos nudi tehničku i lingvističku pomoć za registrirane korisnike Sita putem telefona ili Interneta (E-mail: filos@zg.tel.hr, te uskoro: www.filos.com). SITO se isporučuje na jednoj disketi s kratkim pisanim uputama za početak instaliranja.

Cijena jedne licence je 1.200,00 kn, a za kupovinu u paketu od više komada odobrava se popust. Verzija za mrežu isporučuje se prema vrsti licence i dogovoru.

Kontakt adresa:

FILOS d.o.o, Vince 2, 10000 Zagreb
Tel/fax: 01/43 84 90
E-mail: filos@zg.tel.hr

Usporedba hrvatskih spelling-checkera

Test tekstovi

Podaci o tekstovima kojima su izvršena testiranja hrvatskih spelling-checkera:

1. Uredski tekst - pravnički ugovor obujma 13 stranica A4, ima 2094 pojavnice od kojih su 830 različnice. Tekst ima točno 7 pogrešaka. Najčešće riječi su: *u, i, radova, izvoditelj, je, se, članak* itd.
2. Diplomski rad iz strojarstva - sadrži dosta stručnih i engleskih termina, te literaturnih citata, obujam 80 stranica A4, ima 11024 pojavnice od kojih su 2688 različnice. Tekst ima točno 19 pogrešaka. Najčešće riječi su: *i, za, je, se, u, na, pri* itd.
3. Dio Marinkovićeve zbirke novela ***Ruke*** - točnije prvih 165 stranica od ukupno 203, uspoređujući elektronički zapis s izdanjem Matice hrvatske iz 1981. *(što je 81,3%)*. Datoteka je dobivena putem Interneta *(ftp://aau.dk/pub/slav/)* i ima 49308 pojavnica od kojih su 13108 različnice. Tekst ima neodređeni broj pogrešaka, iz tiskanog predloška je dobiven OCR postupkom. Najčešće riječi su: *i, je, se, u, da, na, kao* itd.
4. Stari zavjet - Knjiga postanka dobivena putem Interneta *(http://jeronim.hbk.hr/biblija/sz)* koja u tiskanom obliku ima 43 stranice *(Biblija, Kršćanska sadašnjost, 1993)*, sadrži dosta arhaičnih riječi i osobnih imena, a ima 29200 pojavnice od kojih su 5741 različnice. Tekst ima neodređeni broj pogrešaka. Najčešće riječi su: *i, je, u, sa, da, a, na* itd.

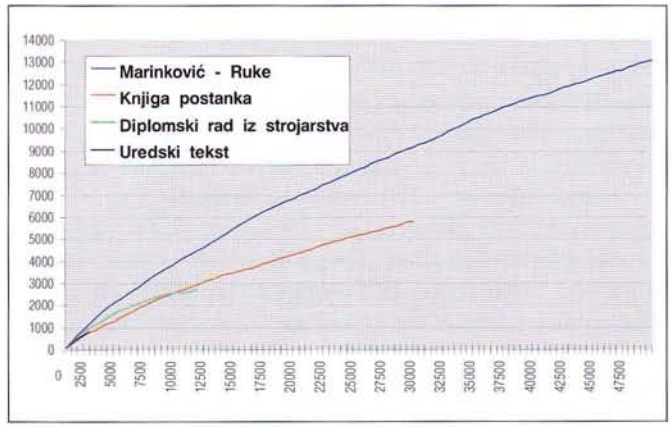
Različnice protiv pojavnica

Kako čitamo neki stvarni tekst hrvatskog jezika, vjerojatnost da će se ponovno pojaviti već pročitana riječ, stalno raste. Riječi se ponavljaju u tekstu, neke češće *(vidjeti dio: Test tekstovi)*, a neke rjeđe. Kao što je vidljivo na grafičkom prikazu, od 49308 riječi *(pojavnica)* u Marinkovićevom tekstu samo je 13108 različitih *(različnice)*. To najbolje znaju autori spelling-checkera koji su pokušali formirati svoje leksičke baze samo i jedino analizom različitih tekstova. Naime, ekstrapolacijom ovisnosti pojavnice/različnice na primjeru Marinkovićevog teksta *(koji inače ima izuzetno veliki broj različnica)*, može se izračunati da je potrebno oko 10^8 pojavnica *(što odgovara biblioteci od preko 1600 knjiga po 200 stranica)* da bi se formirala leksička baza veličine Silićeve leksičke baze *(3.000.000 riječi)*.

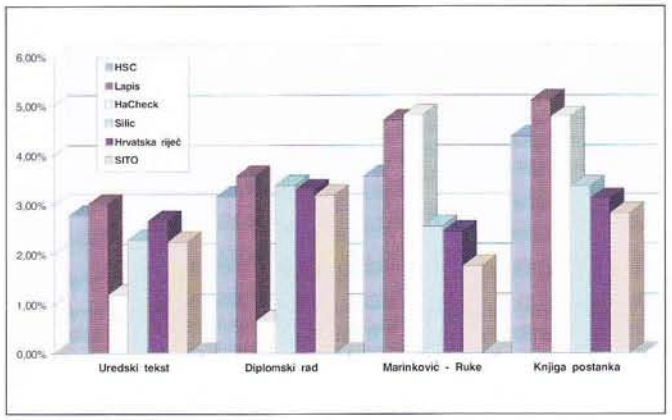
Ovisnost između broja pojavnica i različnica visokim indeksom regresije opisuje Zipf-Mandelbrotov zakon u integralnoj (kumulativnoj) formi:

$$N_{\text{različnica}} = a\ N^{\,b}_{\text{pojavnica}}$$

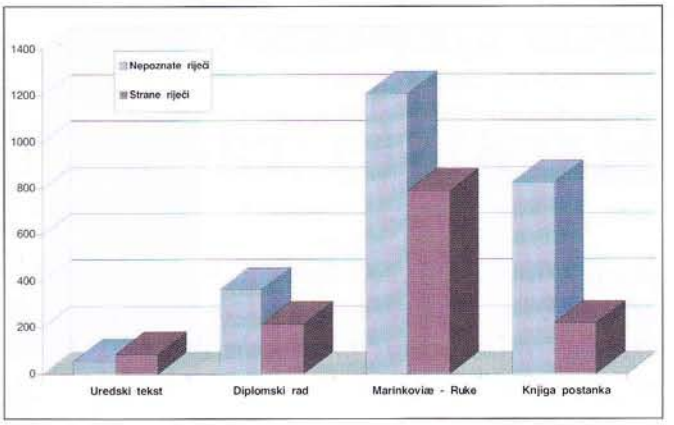
gdje je najbitniji faktor ***b*** koji se za hrvatske tekstove najčešće kreće u rasponu od 0,6 do 0,85 *(za engleske tekstove je nešto niži)*.



Grafički prikaz ovisnosti broja različnica (y) o broju pojavnica (x) na primjeru razmatranih tekstova



Grafički prikaz relativne ovisnosti detekiranih pogrešaka u odnosu na broj pojavnica



Grafički prikaz broja nepoznatih riječi (rezultat spelling-checkera) i broja stranih riječi (rezultat razlikovnog rječnika) programskog paketa Hrvatsko slovo

Zanimljiv je korolar ovog zakona: *brže će se popuniti leksička baza analiziranjem većeg broja kratkih tekstova nego manjim brojem velikih tekstova.* Ovo saznanje bi moglo dovesti do drastičnih promjena u praksi čitanja lektire (možda je efikasnije čitati poeziju, koja je, iako opsegom kraća, jezično bogatija).

Spelling-checkeri na djelu

U donjoj tablici prikazane su pogreške koje su otkrili spelling-checkeri.

Iz grafičkog prikaza ovisnosti broja detektiranih pogrešaka prema broju pojavnica, vidljivo je da su granice prekrivanja teksta najčešće u intervalu od 3% do 5%, što potvrđuje da su svi razmatrani spelling-checkeri realno upotrebljivi. Za prvo mjesto natječu se ***Hrvatska riječ*** i ***SITO*** *(ova konstatacija je utoliko netočna što je poznato da je SITO naučeno Marinkovićevim tekstom).* Ističe se dobra prilagođenost ***Hašeka*** pri provjeri stručnih tekstova *(Hašek je najčešće korišten u akademskim krugovima),* a u svojoj bazi ima i oko 70.000 engleskih riječi koje su najčešći pratilac stručnih tekstova.

Zanimljivo je pogledati učinak programskog paketa *Hrvatska riječ* s aspekta funkcije razlikovnog rječnika u analizi hrvatskih tekstova:

Brzina rada svih opisanih spelling-checkera zadovoljavajuća je za rad na suvremenim osobnim računalima. Većina ih je brza koliko i spelling-checker za MS Word, a

Tablica s brojem detektiranih pogrešaka

Broj detektiranih pogrešaka	Uredski tekst	Diplomski rad iz strojarstva	Marinković, Ruke	Knjiga postanka
HSC	58	348	1764	1280
Lapis	63	395	2324	1493
Hašek	25	73	2373	1393
Hr. rač. pravopis	48	373	1261	984
Hrvatska riječ	56	364	1210	909
SITO	47	352	872*	826

* *Napomena: prema riječima autora, SITO je već naučeno leksikom razmatranog teksta*

radi usporedbe, spomenimo da engleski spelling-checker koji je ugrađen u MS Word 7, ima leksičku bazu od samo 106.000 riječi, a smatra se da ima vrlo dobru prekrivenost engleskog jezika. Odgovarajući rječnik sinonima *(tezaurus)* ima 440.000 ključnih riječi *(podaci dobiveni od MicroSoft-Hotlinea za Hrvatsku).*

Umjesto zaključka, ponovimo: ***na hrvatskom tržištu imamo već šest zadovoljavajuće kvalitetnih spelling-checkera!*** Autorima želimo što veći komercijalni uspjeh, divimo se njihovom entuzijazmu i zahvaljujemo im na uloženom trudu. Budući korisnici - pokažite svijest o negativnostima koje uzrokuje softversko piratstvo - legalno nabavite hrvatske programe za strojnu provjeru teksta.

Literatura:

Dembitz, Š.: *Automatizacija postupka otkrivanja grešaka u tekstu u novim telekomunikacijskim službama,* doktorska disertacija, FER Zagreb, 1993.

Napomena čitateljima:

Na CD-ROM-u u ovom izdanju WIN.INI-ja potražite programe:

- Hrvatski spelling-checker Ver1.60 u arhivi HSC160.ZIP,
- Word Count Ver1.05 u arhivi WCNT105.ZIP i
- demo verziju programa SITO.

msokele@public.srce.hr
URL:http://public.srce.hr/~msokele

Zahvaljujemo poduzećima: Polar iz Splita, TTE Inženjering d.o.o. iz Zadra i FILOS d.o.o. iz Zagreba za ustupljene programe, te Katedri za preradu polimera Fakulteta strojarstva i brodogradnje Sveučilišta u Zagrebu za korištenje programa Hrvatski računalni pravopis. Posebne pohvale autorima Hašeka i HSC-a koji su svima omogućili besplatno korištenje svojih programa.

❑