Damir Boras

RJEČNIČKA BAZA KAO OSNOVA ZA IZRADU AUTOMATSKOG DETEKTORA POGREŠAKA TEKSTA NA HRVATSKOM JEZIKU PISANOG POMOĆU KOMPJUTORA

1. Općenito o detektorima pogrešaka

Detektori, tj. korektori pogrešaka u pisanom tekstu (engleski: spelling checker) kompjutorski su programi koji omogućuju provjeru pravopisne ispravnosti teksta pisanog (priređenog) na kompjutoru. To znači da ne provjeravaju njegovu smislenost niti ispravnost pisanja interpunkcija, već samo ispravnost pisanja pojedinih riječi prema pravopisnim pravilima jezika teksta.

Takvi programi pojavili su se najranije na engleskom govornom području (SAD) gdje je prvo i došlo do masovnije upotrebe kompjutora za upis i uređivanje teksta.

Danas već postoji na tržištu niz programa za detektiranje ispravnosti pisanja riječi za razne jezike (engleski, francuski, njemački), pa čak i različite verzije programa za britanski i američki engleski jezik. Ti programi omogućuju da se pri pisanju/prepisivanju teksta jednostavno i brzo pronađu moguće pogreške u pisanju, bilo da se piše na vlastitom ili stranom jeziku. Većina tih programa radi tako da uspoređuje cijele riječi s riječima u svom ugrađenom rječniku koji sadrži dovoljan broj riječi jezika teksta. Uz rječnik ti programi imaju obično ugrađen i tzv. tezaurus, odnosno strukturirani popis riječi, koji sadrži unaprijed priređen popis riječi sličnih po značenju ili po načinu pisanja. Takav tezaurus omogućuje da se za zadanu riječ pronađu sve riječi iste ili slične po značenju ili slične po načinu pisanja. To omogućuje korisniku koji piše na stranom jeziku da brzo dobije popis sinonima za neku određenu riječ, pa je ne mora provjeravati u klasičnom rječniku. Za jezike koji nemaju fonetski pravopis (engleski i francuski npr.) često postoji i mogućnost pronalaženja riječi po zvučnoj sličnosti.

2. Svrha istraživanja

Iako postoji potreba za detektorima pogrešaka u tekstu pisanom na kompjutoru, za hrvatski standardni jezik takvih programa još uvijek nema na tržištu. Tim se problemom u nas bavi više istraživača: Kržak (1983), Bohaček i Dembitz (1978), Urošević. Svrha je ovog rada da utvrdi koje bi općenite uvjete ti programi trebali zadovoljavati da bi se mogli komercijalno upotrebljavati, te da provjeri da li rječnički pristup, odnosno upotreba postojeće Rječničke baze hrvatskog književnog jezika, koju su ranije opisali i izradili Kržak i Boras (1985), može zadovoljiti te uvjete.

3. Pristupi izvedbama detektora pogrešaka

Pristupi izvedbama detektora pogrešaka pojedinih riječi pisanog teksta brojni su. Pregled tih pristupa dao je Kržak u svojoj disertaciji (Kržak, 1983). Naprimjer: Sitar i Zamora koriste tablice digrama i trigrama, Cornew referentni rječnik s najčešćim rječima jezika teksta, a Vossler i Branston kombinaciju tih dviju metoda. Kržak je provjeravao upotrebu tablica slogova ili "pseudoslogova", dok su Bohaček i Dembitz (1978) razvili specijalnu metodu m-gramskih težina riječi.

Svi ovi pristupi svode se na dva osnovna, koje uvjetno možemo nazvati: 1) "statistički" i 2) "rječnički" pristup.

3.1 Statistički pristup

Ovaj se pristup zasniva na analizi teksta na razini čestote pojavljivanja pojedinih kombinacija znakova (digrama, trigrama, n-grama, odnosno slogova ili pseudoslogova), koja se za pojedini jezik može lako odrediti iz dovoljno velikog korpusa. Prilikom pisanja teksta program ukazuje na mjesta gdje se pojavljuju kombinacije znakova ili slogovi koji se ne nalaze u ugrađenoj tablici ili im je vjerojatnost pojavljivanja manja od neke unaprijed zadane vrijednosti. Naprimjer, prema tablici raspodjele digrama (Boras i Kržak, 1984) digram "DJ" pojavljuje se u 1.004.156 digrama 207 puta, a digram "JD" samo 10 puta. Program će ukazivati na sva mjesta gdje se pojavljuje digram "JD" jer mu je vjerojatnost pojavljivanja vrlo mala. Koja je granica vjerojatnosti na kojoj se neki digram prihvaća ili odbacuje, određuje se prema željenoj točnosti detektora.

Kada naiđe na takovo "sumnjivo" mjesto, program traži da prihvatimo ili promijenimo riječ odnosno tu kombinaciju znakova. Kod ovog, "statističkog" pristupa jedan od problema u izvedbi detektora jest postotak nedetektiranih pogrešaka, budući da su mnoge kombinacije dvoslova i troslova dosta česte, ali u nekim riječima jednostavno ne dolaze, a lako se mogu pojaviti kao rezultat statistički najčešćih pogrešaka pri strojnom pisanju (Kržak, 1983).

Drugi problem koji se ovdje pojavljuje jest broj tzv. "lažnih pogrešaka", tj. ukazivanje na pogrešku koja to nije. Taj problem na statističkoj razini nije lako riješiti jer se ne može unaprijed reći koje su riječi ispravne iako sadrže "zabranjene" kombinacije znakova. U prethodnom primjeru takva bi se pogreška javila u svim riječima koje sadrže digram "JD", što znači i u svim superlativima pridjeva koji počinju sa slovom "D" (osim, naravno, pridjeva "dobar"). Rješenje za to jest ugrađivanje popisa "iznimaka" u odnosu na statistička svojstva pojavljivanja kombinacija znakova. To dovodi do "rječničkog" pristupa u izvedbi korektora pogrešaka, tj. do upotrebe rječnika, odnosno rječničke baze. U statističkom pristupu popis iznimaka predstavlja prijelaz prema rječniku. Statistički se pristup i upotrebljavao iz "straha" da bi rječnik bio prevelik. Međutim, pokazuje se (prema Kržak, 1983) da veličina dodatnih podataka u statističkom pristupu ubrzo dostiže red veličine rječnika.

3.2 Rječnički pristup

Pretpostavka "rječničkog" detektora jest da ima ugrađen dovoljno velik rječnik koji sadrži većinu riječi koje se pojavljuju u tekstu te da ispravnima smatra one riječi koje se nalaze u rječniku, a uvjetno "neispravnima" one kojih u tom rječniku nema. U takvom detektoru nedektiranih pogrešaka na pravopisnoj razini nema, a broj "lažnih pogrešaka" ovisi o veličini rječnika kao i sposobnosti programa da gradi dodatne "korisničke" rječnike najčešćih riječi koje pojedini korisnik upotrebljava i koje se ne nalaze u glavnom rječniku.

4. Poteškoće pri izradi "rječničkog" detektora pogrešaka za hrvatski standardni jezik

Izvedba rječničkih detektora pogrešaka prilično je jednostavna za pretežno analitičke (neflektivne) jezike, kao što su engleski i francuski, jer broj riječi takvog rječnika uglavnom odgovara i broju rječničkih natuknica uz mali broj iznimaka. Naprimjer, u engleskom jeziku većina riječi tvori plural tako da im se doda "s" na kraju riječi uz manji broj pravila koja se dadu ugraditi u program, dok je broj nepravilnih plurala zanemariv (npr. mouse/mice) - tako da se takve riječi jednostavno ugrađuju u rječnik kao nove, zasebne riječi. Takovi rječnici, odnosno tezaurusi, sadrže i preko 100 tisuća riječi, pri čemu po obimu sadrže 250 do 300 tisuća znakova. Očito je da jednostavna rječnička struktura omogućuje i sažimanje rječnika.

Slijedi pregled podataka o veličini rječnika i broju riječi za neke od najčešće korištenih programa za upis i uređivanje teksta, koji sadrže i detektore pogrešaka za engleski, francuski i njemački jezik.

Tablica 1.

Program za upis i uređenje teksta	**Jezik**	**Veličina rječnika**	
		br. riječi	**br. znakova**
WordPerfect 4.1	Am. engl.	85 000	216 663
WordPerfect 4.2	Am. engl.	100 000	290 304
	Njemački	100 000	284 564

WordPerfect 5.0	Engleski	115 000	292 095
	Francuski	95 000	271 933
Word 4.0	Engleski	120 000	175 000
	Njemački	115 000	360 000
	Francuski	100 000	240 830
Word 5.0	Engleski	120 000	167 869

Kod sintetičnih (flektivnih) jezika, kao što je hrvatski, čisto rječnički pristup, gdje se bilježi svaka riječ onako kako se pojavljuje u tekstu, nije ekonomičan jer već i manji broj riječi daje vrlo velik broj oblika pa bi takav rječnik zauzimao previše mjesta. Prema Kržaku (1983), odnos između broja riječi i broja oblika za standardne hrvatske rječnike iznosi 1 : 10. Kržak i Boras (1985) na primjeru rječničke baze hrvatskog književnog jezika pokazali su da je samo 4009 riječi dalo 36965 oblika. To se odnosi na rječničku bazu u kojoj je definirano 28 vrsta riječi. Ako bi se taj omjer računao prema tradicionalnoj gramatičkoj podjeli na 9 vrsta riječi, iznosio bi tada 1 : 16.

Drugi pristup izradi rječničkog detektora za sintetične jezike jest upotreba rječničke baze podataka koja bi sadržavala posebno početke riječi i posebno nastavke, pri čemu bi uz svaki početak bilo označeno koji nastavci se uz taj početak mogu upotrijebiti. Na taj način uštedilo bi se na prostoru, a brzina pronalaženja riječi u bazi ovisila bi o organizaciji podataka u bazi i načinu pretraživanja.

U ovom radu ispitat će se da li se takva rječnička baza, kakvu su priredili Kržak i Boras (1985), može efikasno upotrijebiti kao detektor pogrešaka pisanja pojedinih riječi u tekstu.

Stoga je prethodno potrebno analiziranjem postojećih najčešće korištenih programa utvrditi koje uvjete mora općenito zadovoljavati program za pronalaženje pogrešaka u tekstu pisanom na kompjutoru.

5. Uvjeti koje općenito mora zadovoljavati program za pronalaženje pogrešaka u tekstu pisanom na kompjutoru

Analizirano je i ispitano nekoliko ranije navedenih najčešće korištenih programa za upis i uređivanje teksta koji uključuju i programe za provjeru ispravnosti pisanja pojedini riječi (WordPerfect 4.1, 4.2, 5.0; Word 4.0, 5.0; WordStar 4.0), te zaseban program Turbo Lightning Spelling Checker koji se može upotrebljavati i uz svaki od prethodno navedenih programa. Svi se oni koriste na PC kompatibilnim računalima zajedno s pripadajućim rječnicima.

Pomoću navedenih programa provjeravani su tekstovi na engleskom, francuskom i njemačkom jeziku, pri čemu je ustanovljeno da su svi prekrivali oko 99,5% riječi u provjeravanim tekstovima. Brzina analiziranih programa iznosila je za te tekstove oko 20 do 35 riječi/sek. na PC XT/4MHz, oko 30 - 50 riječi/sek. na Turbo XT/8Mhz te oko 200 riječi/sek na PC AT386/16MHz kompatibilnom računalu. Brojčani rezultati te analize dani su u tablici 2.

Prema tome, nužno je da rječnička baza zadovoljava slijedeće uvjete da bi imala karakteristike slične onima u analiziranim programima:

1. Rječnička baza mora sadržavati najmanje 99% riječi koje se pojavljuju u prosječnom tekstu (to se odnosi na riječi u svim njihovim oblicima).

2. Program mora omogućavati definiranje vlastitog, pomoćnog ("korisničkog") rječnika koji će uz osnovni, glavni rječnik, za svakog korisnika pokrivati preostale riječi koje on najčešće upotrebljava.

I ovdje su analitički jezici u prednosti pred sintetičnim. Kod sintetičnih jezika tu se opet javlja problem oblika. Pri definiranju pomoćnog rječnika moguća su kod nas dva pristupa: (1) unosi se samo oblik koji se pojavio u tekstu i (2) unose se svi oblici za pojavljenu riječ, pri čemu se opet javlja problem verifikacije unesenih oblika koji dovodi do mogućnosti upotrebe generatora morfoloških oblika (Kržak, 1988). Koji je ovdje najprikladniji pristup, trebalo bi posebno istražiti.

3. Program mora omogućavati da se pojedine riječi koje se u nekom tekstu vrlo često javljaju tokom samog provjeravanja privremeno označe kao ispravne iako se ne nalaze u glavnom rječniku, a korisnik ih ne želi unositi u svoj pomoćni (tzv. korisnički) rječnik. Npr. vlastito ime koje se pojavljuje samo u nekom određenom tekstu, koji se upravo provjerava, i koje se ne želi unositi u pomoćni rječnik, korisnik može označiti kao ispravnu

riječ, tako da ne mora svaki puta kada program naiđe na takovu riječ označavati da je to ispravna riječ.

4. Mora prepoznavati način pisanja riječi (velikim ili malim slovima, isključivo velikim slovima, ili početnim velikim slovom kada se odgovarajuća riječ piše malim slovima).

5. Mora se izravno upotrebljavati iz nekog od postojećih programa za obradu teksta, samo na pritisak tipke.

6. Moraju biti dovoljno brzi, tj. morali bi na najsporijem računalu provjeravati oko 20 riječi/sek.

7. Moraju omogućavati provjeru pojedine riječi, pojedine stranice i cijelog teksta.

8. Poželjno je da baza sadrži tezaurus koji povezuje riječi ili skupine riječi u značenjske ili zvukovne skupove.

9. Rječnik mora biti u dovoljno kompaktnom obliku kako ne bi zauzimao mnogo prostora, budući da su takvi programi uglavnom namijenjeni korisnicima koji rade na osobnim računalima. Iz tablice 1. vidi se da neki rječnici sadrže čak 120 tisuća riječi, sažetih u samo 170 tisuća znakova.

6. Rječnička baza hrvatskog standardnog jezika

Za analizu je uzeta postojeća Rječnička baza hrvatskog književnog jezika (Kržak i Boras, 1985). Ta baza provobitno je izgrađena kao model na kojem je pokazano da je moguće izraditi rječničku bazu podataka koja će oblike riječi izvoditi iz nekih osnovnih oblika. Ti su oblici početak riječi i paradigma završetaka. Posebno izgrađenim sustavom veza u bazi svaka se riječ može ne samo jednoznačno upisati u bazu već i sigurno pronaći. U bazu su upisane najčešće riječi uzete iz frekvencijskog rječnika hrvatskog jezika iz doktorske disertacije I. Furlana (1961).

Podaci iz baze koja je prvobitno napravljena na kompjutorskom sistemu Hewlett-Packard 2000 i na programskom jeziku HP BASIC 2000, preneseni su na PC XT kompatibilno računalo.

Prva verzija baze (na HP 2000 sistemu) bila je organizirana tako da je svaka natuknica imala do pet različitih početaka i jednu skupinu nastavaka. Skupine nastavaka uglavnom su odgovorale gramatičkim nastavcima. Tako je naprimjer, riječ tetak u bazi imala za sve svoje oblike četiri različita početka, i to: tetak-, tetk-, tec- i teč- i na njih vezanu skupinu nastavaka (paradigmu): -0, -a, -u, -a, -e, -om, -i, -a, -ima, -e. U posebnoj datoteci zapisivane su veze između pojedinog početka i odgovarajućeg nastavka (završetka). To znači, ako bismo nastavke označili redom od 1 do 10, tada su uz početke označivani i odgovarajući nastavci: tetak- 1, 8; tetk- 2, 3, 4, 6, 10; tec- 7, 9; teč- 5. U toj posebnoj datoteci bila je zapisana i vrsta riječi.

U tu bazu unesene su riječi za prvih 2617 natuknica iz spomenutog frekvencijskog rječinka I. Furlana. Budući da je u Rječničkoj bazi, različito od klasične gramatičke podjele, definirano 28 vrsta riječi, pojedina natuknica iz frekvencijskog rječnika dala je više riječi. Naprimjer, za glagol "napraviti" upisane su zapravo četiri riječi: glagol "napraviti", glagolski pridjev prošli "napravio,-la,-lo", glagolski pridjev trpni "napravljen" i glagolski prilog "napravivši". Tako je u bazu upisano oko 4800 riječi. Još oko 2000 riječi dodano je tokom ispitivanja baze, tako da je ona ukupno sadržavala 6328 riječi i 74 paradigme nastavaka.

Sve paradigme nalaze se u tzv. reduciranom obliku budući da se u promjeni riječi u pisanom obliku neki oblici za različite padeže nikad ne razlikuju. Tako se, naprimjer, za imenice u pisanom obliku nikad ne razlikuju dativ i lokativ jednine, zatim nominativ i vokativ množine te dativ, lokativ i instrumental množine. Zbog toga se paradigma imenskih riječi (7 jednine i 7 množine) mogla reducirati na ukupno 10 različitih oblika. Na sličan su se način paradigme pridjevskih riječi, koje obuhvaćaju ukupno osamdesetak oblika, reducirale na dvije: nenepčanu s 15 oblika i nepčanu s 14 oblika. Glagolska paradigma reducirana je na 20 oblika prostih vremena.

Ova verzija baze nije bila laka za održavanje jer se nalazila u 3 fizičke datoteke: datoteka nastavaka, datoteka početaka i datoteka veza, pri čemu su kao veze služili redni brojevi početaka u datoteci početaka, redni brojevi paradigmi u datoteci nastavaka i redni broj sloga u datoteci veza, koji su se prilikom dodavanja novih riječi i sortiranja baze morali svi mijenjati.

6. 1 Nova struktura rječničke baze

Zbog toga je na PC XT računalu odmah promijenjena struktura podataka u bazi na principima koje je opisao Kržak (1988), i to tako da svaka riječ može imati samo jedan početak na koji se nadovezuju nastavci iz odgovarajuće paradigme. To znači da početak može biti i prazan (npr. riječi čovjek/ljudi, slati/šaljem, biti/jesam/sam), pa se u tom slučaju svi oblici nalaze zapravo u paradigmi nastavaka. Kržak je takvu organizaciju zamislio

intuitivno očekujući da će pretraživanje baze biti brže, ali nije proveo ispitivanje. Uz takvu promjenu strukture rječničke baze, koju smo izvršili, od ukupno 6328 riječi unesenih u bazu nastalo je 386 različitih paradigmi. Pritome je struktura paradigmi ostala ista kao u prvoj verziji baze koja je opisana u Kržak i Boras (1985) te svaka od tih paradigmi sadrži određeni broj nastavaka, i to 10 za imenske riječi, 14 ili 15 za pridjevske riječi te 20 za glagole. Pojedini nastavci mogu biti i prazni (u bazi označeno nastavkom "0"), što znači da se za taj oblik uzima samo početak, ili na odgovarajućem mjestu paradigme nema ničega, što znači da ta riječ nema odgovarajućeg oblika.

Tako preoblikovana rječnička baza sadrži tri glavne datoteke, i to datoteku početaka s nazivom "POČECI", koja sadrži početke riječi i datoteku paradigmi s nazivom "PARADIGME", koja sadrži sve nastavačke paradigme te datoteku "NASTOBR" koja sadrži samo različite nastavke koji se pojavljuju u paradigmama, i to s preokrenutim redoslijedom slova. Ta je datoteka početna datoteka od koje se polazi pri pretraživanju baze. U datoteci početaka riječi nalaze se podaci o vrsti riječi, broj nastavačke paradigme, te podatak o rodu za imenice, odnosno o vidu za glagole.

Osim tih dviju glavnih datoteka postoje i neke pomoćne datoteke. To su datoteka naziva vrsta riječi "VRIJECI", datoteka "NASTAVCI PRAZNIH" koja sadrži sve nastavke onih paradigmi na koje su vezane riječi s praznim počecima. Sve te datoteke u konkretnoj izvedbi imaju nešto drugačija imena zbog ograničenja operativnog sistema te bi se stoga i teže pamtila.

6.2 Pretraživanje rječničke baze

Pri pretraživanju rječničke baze najveći je problem odrediti što u nekom obliku koji se želi pronaći u bazi predstavlja početak, a što završetak. Početaka ima u postojećoj bazi 6328, pri čemu bi za pojedinu riječ koja se traži trebalo ispitati u bazi sve početke riječi koji počinju istim slovom kao tražena riječ i koji bi po abecednom redoslijedu bili jednaki toj riječi ili bi došli prije te riječi. U slučaju kada se početak iz baze poklapa s početkom tražene riječi potrebno je još provjeriti da li se ostatak tražene riječi nalazi u paradigmi na koju ukazuje početak pročitan u bazi. Pri takvom načinu pretraživanja nepotrebno se ispituje jako mnogo podataka jer se naprimjer ispituju paradigme i svih jednoslovnih početaka koji se nalaze u rječničkoj bazi, a jednaki su početnom slovu tražene riječi. U konkretnoj se izvedbi radi ubrzanja pretraživanja nakon provjeravanja svih jednoslovnih početaka s tim slovom provjeravaju još samo počeci koji se s traženom riječi slažu i u drugom slovu. Pristup do tih početaka osiguran je posebnom indeksnom datotekom.

S porastom baze tako je i linearno padala i brzina pretraživanja baze, te je brzina i već kod veličine baze od 6000 početaka pala čak ispod 1 riječi u sekundi za neke riječi koje su počinjale slovom "P" na računalu PC-XT/8Mhz. Pritom je program za pretraživanje baze pisan u programskom jeziku TurboBasic.

Stoga je bilo potrebno pronaći algoritam pretraživanja koji bi smanjio broj nepotrebnih provjeravanja podataka u bazi. Drugim riječima trebalo je bolje u zadanoj riječi određivati što može predstavljati početak, a što završetak riječi. S druge strane, u 386 nastavačkih paradigmi nalazi se ukupno 4615 nastavaka, od kojih su samo 1033 različita. Kako se za nove riječi koje se dodaju u bazu broj novih nastavačkih paradigmi i broj novih nastavaka vrlo sporo povećava, algoritam pretraživanja u bazi je preokrenut: prvo se odredi koji nastavak može odgovarati završetku tražene riječi, pa se tada u datoteci početaka binarnim pretraživanjem provjerava da li se upravo takav početak u njoj nalazi. Ako se nalazi, potrebno je još samo provjeriti da li se i pretpostavljeni završetak riječi nalazi u paradigmi na koju je vezan pronađeni početak. Na taj način bitno je povećana brzina pretraživanja u bazi tako da sada iznosi na PC-XT računalu oko 6,7 riječi u sekundi, a na PC-AT računalu oko 36 riječi u sekundi.

Program za pretraživanje rječničke baze pronalazi u bazi sve riječi koje odgovaraju traženoj, te daje i njen gramatički oblik. Naprimjer za riječ "SAM" dobiva se kao rezultat glagol "BITI" i pridjev "SAM, SAMA, SAMO". Pri upotrebi baze kao detektora pogrešaka pisanja pojedinih riječi u tekstu pretraživanje se može prekinuti čim se pronađe prva riječ koja odgovara traženoj, jer se u ovoj primjeni ne ispituju veze među riječima i smislenost teksta, nego samo ispravnost pisanja pojedine riječi.

7. Rječnička baza kao program za provjeru ispravnosti pisanja riječi

Očito je da se rječnička baza može upotrijebiti kao program za pronalaženje pogrešaka, pri čemu program za pretraživanje baze treba tako preurediti da zadovolji sve ranije navedene uvjete koji se postavljaju na takve programe.

Sve uvjete, osim 1, 5. i 9. lako je zadovoljiti jer zahtijevaju samo tehničku izvedbu u programiranju,

dok 1, 5. i 9. uvjet treba provjeriti na stvarnom uzorku.

Ovo je istraživanje zbog toga posebno usmjereno na provjeru kako rječnička baza zadovoljava 1. uvjet - uvjet potpunosti baze, 5. uvjet - uvjet dovoljne brzine pretraživanja i 9. uvjet - uvjet dovoljne sažetosti baze.

S postojećim podacima rječnička baza prepoznaje oko 70% riječi koje se pojavljuju u različitim tekstovima uz brzinu od oko od 33 - 36 riječi u sekundi na računalu PC AT-386, za program pisan u programskom jeziku TurboBasic.

Da bi se moglo procijeniti kako će se brzina pretraživanja mijenjati s povećanjem udjela poznatih riječi u provjeravanim tekstovima, potrebno je izmjeriti brzinu pretraživanja na tekstovima koji sadrže točno određeni postotak prepoznatih riječi, i to za različite postotke prepoznatih riječi u tekstu te tada izračunati ovisnost brzine pretraživanja o broju poznatih riječi.

Tako izračunatu brzinu pretraživanja treba usporediti s brzinom već postojećih programa za provjeru ispravnosti pisanja za druge jezike.

8. Tekstovi u uzorku za mjerenje karakteristika programa

S jedne je strane trebalo izmjeriti karakteristike postojećih programa koji su provjeravali ispravnost pisanja za američki odnosno britanski engleski, te francuski i njemački jezik, te s druge strane karakteristike Rječničke baze za hrvatski jezik, ali na tekstovima koji sadrže različite udjele poznatih riječi.

Kao uzorak stranih tekstova uzeti su tekstovi udžbenika engleskog, francuskog i njemačkog jezika Škole stranih jezika u Zagrebu. Pretpostavili smo da takvi tekstovi sadrže uobičajen broj riječi koje se upotrebljavaju u svakom od tih jezika, te da za potrebe ovog istraživanja mogu predstavljati standardne tekstove. Tekstovi udžbenika uzeti su u potpunosti, jedino su izbačeni znakovi koji se u programima za upis i uređenje teksta upotrebljavaju za grafičko oblikovanje teksta, tako da je zadržan samo čisti tekst svakog udžbenika. Njihova veličina, broj poznatih i nepoznatih riječi i rezultati mjerenja navedeni su u tablici 2.

Kao osnova za uzorak hrvatskih tekstova, u nedostatku pravog, po znanstvenim kriterijima izabranog korpusa, uzeti su novinski tekstovi iz pet različitih brojeva časopisa "Start", koje smo već upotrebljavali u prethodnim istraživanjima (Kržak, Boras, 1984, 1985). Takav uzorak, s jedne strane, ograničava općenitost mjerenja, ali se, s druge strane, može uzeti da se jezikom kojim se piše u časopisu "Start" koristi najveći broj pretpostavljenih korisnika programa za pronalaženje pogrešaka. Osim toga, ozbiljnost i širina tema u časopisu "Start" mnogo je veća nego u običnim novinama, tako da se i stoga ti tekstovi mogu uzeti kao prilično dobra aproksimacija standardnog hrvatskog jezika.

Pregled podataka o osnovnom uzorku hrvatskih tekstova nalazi se u tablici 3.

Svi ovi uzorci višestruko su lektorirani tako da je u njima ostalo vrlo malo pogrešaka, te se može smatrati da preostale greške neće utjecati na točnost mjerenja.

Kao što se vidi iz tablice 3, Rječnička baza u osnovnom uzorku prepoznaje u prosjeku 71,94% riječi. Iz tog osnovnog teksta priređeno je 19 različitih uzoraka teksta različitih veličina i različitog udjela prepoznatih riječi. Pritome je upotrijebljena metoda slučajnog izbora pojedinih riječi iz stvarnog teksta polazeći od činjenice da se detektorom pogrešaka ne ispituje smislenost ispitivanog teksta, nego samo ispravnost pisanja riječi te da se stoga tekstovi mogu sastoji od bilo kakvog niza riječi uzetog iz stvarnog teksta, a da će se pritom zadržati i osnovne statističke karakteristike stvarnog teksta.

Uzorci su izrađeni tako da su u osnovnom uzorku prethodno označene sve nepoznate riječi, a tada se posebnim programom s ugrađenim generatorom slučajnih brojeva generirao uzorak s točno određenim karakteristikama.

Za još 5 uzoraka uzeti su stvarni tekstovi, i to 3 uzorka iz časopisa "Start" (uzorci T14, T15 i T16 u tablici 4) te 2 uzorka iz Vodiča kroz studij za šk. godinu 1989/90. Filozofskog Fakulteta Sveučilišta u Zagrebu (uzorci T17 i T18). Karakteristike pojedinih uzoraka i rezultati mjerenja prikazani su u tablici 4. Uzorci T19 - T24 dodani su posljednji da bi se vidjelo slaganje rezultata mjerenja i s manjim uzorcima tekstova.

9. Rezultati mjerenja

Tablica 2.

Karakteristike uzoraka stranih tekstova i brzine različitih detektora pogrešaka na različitim kompjutorima

Tekst	Računalo*	Tekst proc.	Velič. u znak.	Broj riječi	Broj nepoznatih riječi	Vrijeme (sek.)	Brzina zn/sek	rij/sek
Engl.1	AT386	WP 4.2	272155	45681	246	225	1209,6	203
	XT/8					891	305,5	51,3
	XT/8	WP 5.0	275676	47404	253	1304	211,4	36,4

* Računala imaju slijedeće karakteristike:

AT386, PC AT-386, 16 Mhz, 15,5 (jačina u odnosu na PC-XT)
XT/8, PC XT-Turbo, 8 Mhz, 1,7

** Oznake tekstprocesora:

WP 4.2 - WordPerfect 4.2
WP 5.0 WordPerfect 5.0

Iz gornje tablice vidljivo je da je pokrivanje oko 99,5 %

Tablica 3.

Karakteristike osnovnog uzorka hrvatskog teksta u časopisu "Start"

Tekst	Velič. (znak.)	Broj riječi riječi	Broj pozn. riječi	Broj nepoz. riječi	% pozn. riječi	% nepozn.	
Start 1	146756	21148	15337	5811	72,52	27,48	
Start 2	392262	57476	41548	15928	72,29	27,71	
Start 3	554298	81038	58245	22793	71,07	28,13	
Start 4	446640	65471	47632	17839	72,75	27,25	
Start 5	462304	67293	47619	19674	70,76	29,24	
UKUPNO		2012260	292426	210381	82045	71,94	28,06

Iz ove tablice lako se dade izračunati koliko znakova otpada u prosjeku u ovom uzorku na jednu riječ:

Broj znakova po riječi = 2012260/292426 = 6,881 znakova

U taj broj ugrađene su i sve interpunkcije i razmaci među riječima. To je zapravo broj znakova koji u uzorku otpada na jednu riječ. Stvarna prosječna duljina riječi u ukupnom uzorku izračunata je posebnim programom i ona iznosi:

Prosječna duljina riječi u uzorku = 5,281

Izračunata je i prosječna duljina riječi koje sa sada ugrađenim rječnikom prepoznaje rječnička baza te prosječna duljina neprepoznatih riječi:

Prosječna duljina poznatih riječi = 4,280

Prosječna duljina nepoznatih riječi = 7,849

Ove razlike proistječu iz karaktera podataka u bazi, u kojoj se sada nalaze riječi prvih 2600 natuknica iz frekvencijskog rječnika I. Furlana (1961). Među najčešće riječi dolazi najveći broj kratkih pomoćnih riječi koje utječu na njihovu prosječnu duljinu.

Tablica 4.

Karakteristike generiranih uzoraka tekstova i brzina pretraživanja pomoću rječničke baze

Tekst (znak.)	Velič. riječi	Broj pozn. rij.	Broj nepoz. rij.	Broj pozn. rij.	% nepozn. riječi	%	Vri-jeme (sek.)	Brzina znak/ sek.	riječi/ sek.
T01	62930	10013	10013	0	100,00	0,00	260,06	241,98	38,50
T02	64387	9913	9415	498	94,98	5,02	259,49	248,13	38,20
T03	66906	9914	8904	1010	89,81	10,19	263,89	253,54	37,57
T04	69351	9971	8467	1504	84,92	15,08	269,62	257,21	36,98
T05	72465	10004	7951	2053	79,48	20,52	275,02	263,49	36,38
T06	74936	9814	6798	3016	69,27	30,73	273,14	274,35	35,93
T07	81337	10060	6038	4022	60,02	39,98	288,29	282,14	34,90
T08	85966	10077	5026	5051	49,88	50,12	296,17	290,26	34,02
T09	90490	9982	4023	5959	40,3	59,7	301,88	299,76	33,07
T10	94235	9974	3012	6962	30,2	69,8	306,35	307,60	32,56
T11	99113	10032	2005	8027	19,99	80,01	313,25	316,40	32,03
T12	104445	10039	1029	9010	10,25	89,75	323,28	323,08	31,05
T13	107929	9919	0	9919	0,00	100,00	327,35	329,70	30,30
T14	146756	21148	15337	5811	72,52	27,48	596,92	245,86	35,43
T15	13732	1831	1278	553	69,8	30,2	50,53	251,97	36,24
T16	12307	1831	1278	553	69,8	30,2	50,78	242,36	36,06
T17	13007	1698	1187	511	69,91	30,09	48,48	268,3	35,02
T18	53014	6585	4019	2566	61,03	38,97	185,22	286,22	35,55
T19	792	102	66	36	64,71	35,29	2,94	269,39	34,74
T20	1443	204	166	38	81,37	18,63	5,46	264,29	37,37
T21	2542	307	167	140	54,40	45,60	8,93	284,66	34,38
T22	2920	414	368	46	88,89	11,11	11,44	255,24	36,18
T23	3241	502	478	24	95,22	4,78	12,98	249,69	38,67
T24	4022	638	638	0	100,00	0,00	16,66	241,42	38,30

Podaci iz ove tablice dobiveni su mjerenjem na PC-AT 386/16Mhz kompatibilnom računalu. Snaga tog računala je u odnosu na PC-XT/4MHz računalo, prema programu NORTON UTILITIES slijedeća:

Indeks brzine računanja	17,6
Indeks disk jedinice	11,3
Ukupni indeks snage	15,5

Ti su podaci analizirani pomoću programa MULREG koji metodom najmanjih kvadrata izračunava višestruku linearnu regresiju, tj. za zavisnosti općeg oblika:

$$Y = A0 + A1 * X1 + A2 * X2 + ... + An * Xn$$

Kao nezavisne varijable uzeti su broj poznatih odnosno nepoznatih riječi te veličina datoteke. Dobiveni su slijedeći rezultati, pri čemu oznake u formulama znače slijedeće:

T - vrijeme u sekundama
L - duljina datoteke u znakovima (byte)
R - broj riječi
Rzna - broj poznatih riječi
Rnezna - broj nepoznatih riječi
Vr - brzina u riječima/sek
Vz - brzina u znakovima/sek
Pzna - Postotak poznatih riječi

$$T = -0,357559 - 0,000606\ L + 0,029746\ Rzna + 0,039465\ Rnezna \quad (1)$$

uz indeks determinacije 0,99993

odnosno

$$T = -1{,}027695 + 0{,}026150\ Rzna + 0{,}032966\ Rnezna \qquad (2)$$

uz indeks determinacije 0,99986

odnosno

$$T = 2{,}893915 + 0{,}003522\ L \qquad (3)$$

uz indeks determinacije 0,96781

odnosno

$$Vr = 30{,}098885 + 0{,}081617\ Pzna \qquad (4)$$

uz indeks determinacije 0,96090

odnosno

$$Vz = 331{,}94 - 0{,}911935\ Pzna \qquad (5)$$

uz indeks determinacije 0,89787

10. Analiza izmjerenih podataka

Iz navedenog se vidi da vrijeme prolaženja kroz datoteku ovisi gotovo isključivo o broju poznatih i nepoznatih riječi (1). Međutim, s obzirom na to da i duljina datoteke ovisi o broju riječi, a s obzirom na to da na jednu riječ otpada u uzorku 6,881 znakova, moguće je u formulu (2) uvrstiti tu vrijednost. Uz pretpostavku da se u bazu ugrade sve nepoznate riječi, formula (2) se transformira u:

$$T = -1{,}027695 + 0{,}026150\ Rzna \qquad (6)$$

Prosječna brzina prolaženja kroz bazu bila bi tada:

$$V = L/T = L/(-1{,}027695+0{,}026150\ Rzna), \qquad (7)$$

odnosno ako se u to uvrsti da je

$$L = 6{,}881 * Rzna$$

$$V = \frac{6{,}881 \times Rzna}{-1{,}027695 + 0{,}026150\ Rzna} \qquad (8)$$

Za broj riječi veći od 2000 konstanta u nazivniku doprinosit će manje od 2% vrijednosti nazivnika, pa se može zanemariti, te formula (8) prelazi u:

$$V = \frac{6{,}881 \times Rzna}{0{,}026150\ Rzna} = \frac{6{,}881}{0{,}02615} = 263{,}136\ zn/sek \qquad (9)$$

Vrijednost iz (9) je ekstrapolirana veličina brzine provjeravanja baze za slučaj kada bi se u bazu ugradile sve nepoznate riječi.

To je više od vrijednosti koja se dobiva uvrštavanjem u formulu (5) vrijednosti 100 za postotak poznatih riječi:

$$Vz = 331{,}94 - 0{,}911935 * 100 = 331\text{-}91 = 240$$

Razlika nastaje zbog toga što duljini prolaza sada doprinose sve riječi, dakle s prosječnom duljinom većom nego što je prosječna duljina riječi u uzorku.

Ovaj izvod ujedno i pokazuje da je bilo opravdano generirati uzorke slučajnim izborom riječi, jer se vidi da sama duljina datoteke ne utječe na vrijeme provjeravanja, nego da ona ovisi praktički samo o broju riječi u uzorku. Duljina datoteke, i sama, naravno ovisi o broju riječi.

Dobivena ekstrapolirana vrijednost od 263 zn/sek za brzinu provjeravanja baze, za slučaj kada bi baza sadržavala veći broj riječi govori da je brzina rječničke baze za odgovarajuće računalo oko 4,5 puta manja od brzine sličnih programa za druge jezike. Uzrok je tome činjenica da je mjerenje izvršeno na modelu Rječničke baze napisanom u programskom jeziku TurboBasic, dok su ostali ispitivani programi napisani u asembleru ili

programskom jeziku C.

Brzina izvođenja programa pisanih u jeziku TurboBasic, za sve vrste aplikacija, manja je za red veličine, odnosno 10-15 puta od istih takvih programa pisanih u asembleru ili C-u. Stoga se može s punim pravom pretpostaviti da će brzina Rječničke baze, kada se, što se i planira, izvede u programskom jeziku C, čak i uz faktor sigurnosti predviđanja od 2 - 2,5 biti usporediva s brzinom postojećih programa za provjeru ispravnosti riječi za druge jezike.

11. Procjena potrebnog broja riječi i veličine rječničke baze

Da bi baza pokrivala više od 99% riječi u prosječnim tekstovima potrebno ju je dopuniti. U postojećem uzorku Rječnička baza pokriva 71.94% riječi, odnosno prepoznaje 210381 od ukupno 292426 riječi koje su se pojavile u uzorku. U 210381 prepoznatoj riječi ima ih samo 13399 različitih, a rječnička baza ih je prepoznala izvođenjem oblika od 3865 početaka u bazi. Taj podatak dobiven je automatskom lematizacijom poznatih riječi u uzorku pomoću posebnog programa. Pritom se nije ulazilo u kontekstualnu provjeru pojedinih pojavljenih riječi koje bi po obliku mogle pripadati više od jednoj osnovi (npr. "da" može biti veznik, čestica i treće lice prezenta od "dati"), nego je takva riječ pridodana samo jednoj osnovi (početku) u rječničkoj bazi, i to onoj koja se u bazi prva pojavljuje (u primjeru je to veznik "da"). Budući da se ovdje radi o upotrebi rječničke baze kao detektora pogrešaka, nije važno o kojoj se riječi radi, nego da li ta riječ postoji u rječničkoj bazi ili ne. Na taj način dobiven je popis frekvencija pojavljivanja riječi koje pripadaju određenoj osnovi (početku) u rječničkoj bazi. Dio tog popisa prikazan je u tablici 5. Samo za ilustraciju prikazano je i o kojim se riječima radi te kako je baza prepoznala vrstu riječi.

Tablica 5.

Frekvencijski popis osnova (početaka) Rječničke baze dobiven automatskom lematizacijom riječi u ispitivanom uzorku

R.br.	Riječ	Vrsta riječi	Frekv.	Kumul. frekv.	Kum. % prekrivanja
1.	biti	spona	15744	15744	5,38
2.	i	veznik	9828	25572	8,74
3.	u	veznik	9129	34701	11,87
4.	da	veznik	5467	40168	13,74
5.	se,si	imen. zamj.	5412	45580	15,59
6.	taj,ta,to	pridj.zamj.	4378	49958	17,08
7.	na	prijedlog	4274	54232	18,55
8.	koji,-a,-e	pridj.zamj.	3664	57896	19,8
9.	uobičajio,-la,-lo	glag.pridj.	2810	60706	20,76
10.	bio,-la,-lo glag.	pridj.	2090	62796	21,47
.....					
27.	onaj,-a,-o	pridj.zamj. ili			
	on,ona,ono	imen. zamj.	1028	88661	30,32
.....					
70.	put	im. m.roda	346	117022	40,02
.....					
212.	rat	im. m.roda	135	146289	50,03
.....					
561.	zlatan,-na,-no	pridjev	53	175471	60,01
.....					
1966.	složiti	glagol	8	204703	70,00
.....					
3865.	čuvao,-la,-lo	glag.pridj.	1	210381	71,94

Prema ovim podacima u tablici procijenjeno je s koliko riječi treba dopuniti rječničku bazu. Pomoću programa CURFIT izvršeno je prilagođavanje krivulje metodom najmanjih kvadrata te je za redni broj riječi (R) i kumulativni postotak prekrivanja (PP) dobivena slijedeća zavisnost, uz indeks determinacije 0.99582:

$$PP = A + B * \ln(R)$$

gdje je: A - konstanta, i iznosi 2,31102

B - konstanta, i iznosi 8.73528

ln(R) - prirodni logaritam broja riječi

Iz ove zavisnosti dobiva se da je za postotak prekrivanja od 99% potrebno 64138 riječi, za 99.5% prekrivanja 67916 riječi, te za 100% prekrivanje 71917 riječi. Budući da je ovisnost dobivena na konačnim uzorcima, a stvaran broj riječi u nekom jeziku je teoretski neograničen, može se reći da su ovi iznosi **najmanji** potreban broj riječi da bi se postiglo željeno prekrivanje. Ti brojevi se po redu veličine dobro slažu i s brojem riječi u programima za druge jezike.

Brzina pretraživanja neće se u tom slučaju previše smanjiti jer se radi o logaritamskoj ovisnosti. Naime, stvarna brzina pretraživanja za pojedinu riječ ovisi i o broju podataka koji se pretražuju. Ta brzina može se grubo prikazati formulom:

$$T = K * \log(N)$$

gdje je T vrijeme pretraživanja, K je konstanta, a N je broj podataka. Prema tome, za neki drugi broj podataka odnos T1/T2 iznosit će:

$$T1/T2 = \log(N1)/\log(N2)$$

pa se za vrijednosti 6328 koliko ima riječi u bazi koja se ispituju i pretpostavljene veličine od približno 65000 riječi dobiva:

$$T1/T2 = \log(6328)/\log(65000) = 8.75/11.08 = 0.79$$

što znači da će se brzina pretraživanja u tom slučaju smanjiti oko 21%, što još uvijek ulazi u zadane okvire.

Što se tiče veličine baze, ona u sadašnjem obliku iznosi 85305 znakova za 6328 početaka, što iznosi u prosjeku 13.481 znak po pojedinom početku. Za 65000 početaka veličina baze iznosila bi 876265 znakova, što znači da se u ovom obliku baza ne bi mogla smjestiti u radnu memoriju računala. Rješenje za to jest da se u radnu memoriju smjesti dio baze koji bi pokrivao recimo 90% riječi u tekstu, za što bi prema gornjoj formuli trebalo oko 23000 riječi koje bi zauzele oko 310000 znakova, što bi lako stalo u radnu memoriju, dok bi se za ostalih 10% riječi trebala pretraživati baza na disku. Sve ove veličine odnose se na bazu u kojoj uopće nije izvršeno dodatno sažimanje podataka. Tzv. "pakiranjem" podataka mogla bi se potrebna veličina smanjiti i do 50%.

12. Zaključak

Prema svemu navedenom, Rječnička baza predstavlja pogodnu osnovu za izradu programa za traženje pogrešaka u tekstu, ali uz slijedeće uvjete:

- potrebno ju je napisati u bržem programskom jeziku (jeziku C ili asembleru).
- potrebno ju je dopuniti novim riječima. Približno to iznosi ukupno oko 65000 riječi. Da bi se napravio izbor tih riječi, potrebno je dodatno istraživanje.
- potrebno je istražiti na koji način sažeti podatke u bazi, da bi za pretpostavljeni broj riječi i veličinom odgovorala PC kompatibilnom računalu, koje za običnu radnu memoriju ima ograničenje od 640 KB.
- potrebno je istražiti najpogodniji način za unošenje dodatnog korisničkog rječnika i "privremenih" riječi.

Literatura:

1. Kržak, M. Mogućnosti ispravljanja i uređivanja teksta na hrvatskom književnom jeziku uz pomoć elektroničkog računala. Doktorska disertacija. - Zagreb: Elektrotehnički fakultet, 1983.
2. Bohaček, Z. i Dembitz, Š. Prepoznavanje riječi prirodnog jezika. - Informatica 78, 1114, Bled

1978.

3. Urošević Z. Statistička metoda otrkivanja i korekcije slovnih grešaka supstitucionog tipa u tekstu na srpsko-hrvatskom jeziku. - Beograd: BIGZ,
4. Kržak, M. i Boras, D. Rječnička baza hrvatskog književnog jezika. - Informatologia Yugoslavica, 17, 1985, 3-4, str. 223-242.
5. Boras, D. i Kržak, M. Distribution of Digrams, Trigrams and Word Forms in Croatian Literary Langugage. - IRCIHE Bulletin, 10, 1984, 2-3, str. 46-61.
6. Kržak, M. Serbo-Croatian morpho-spelling. - Proceedings of the IV-th Conference *Computer Processing of Language Data*. - Portorož, 1988.
7. Furlan, I. Raznolikost rječnika. Struktura govora. Doktorska disertacija. - Zagreb 1961.