UNIVERZITET U BEOGRADU
PRIRODNO-MATEMATIČKI FAKULTETI
MATEMATIČKI FAKULTET
BEOGRAD

Virtual Library of Faculty of Mathematics - University of Belgrade
elibrary.matf.bg.ac.rs

CVETANA J. KRSTEV

PROGRAMSKI SISTEMI ZA UREĐIVANJE TEKSTA

magistarska teza

Beograd, 1989.

Virtual Library of Faculty of Mathematics - University of Belgrade
elibrary.matf.bg.ac.rs

Mentor: dr Nedeljko Parezanović, red. prof.
Prirodno-matematički fakulteti, Beograd

Članovi Komisije: dr Zoran Žiletić, red. prof.
Filološki fakultet, Beograd

dr Žarko Mijajlović, van. prof.
Prirodno-matematički fakulteti, Beograd

Branjeno 31.05.1989

Beograd, maj 1989.

Virtual Library of Faculty of Mathematics - University of Belgrade
elibrary.matf.bg.ac.rs

Z A H V A L N O S T

Istraživanje opisana u ovoj tezi su ostvarena kroz rad **Grupe za računarsku obradu prirodnih jezika** Računarske laboratorije Prirodno-matematičkih fakulteta Univerziteta u Beogradu.

U izradi teze, prateći neposredno razvoj moga rada, neprocenjivu pomoć, podršku i savete mi je pružio mentor profesor dr **Nedeljko Parezanović**, na čemu mu duboko zahvaljujem.

Molim, takođe, članove Komisije, profesora dr **Zorana Žiletića** i profesora dr **Žarka Mijajlovića**, koji su svojim primedbama doprineli jasnosti konačnog teksta teze, da uvaže izraze moje osobite zahvalnosti.

Tokom rada na tezi, u razjašnjavanju mehanizma obrazaca za programski sistem T_EX, što je značajno uslovilo izgled izložene konstrukcije obrazaca za srpskohrvatski jezik, pomoć i informacije su mi pružili i profesor dr **Janusz S. Bien** sa Instituta za informatiku Varšavskog univerziteta i profesor **Jacques Desarmenien** sa Instituta za informatiku Univerziteta u Strazburu, te ih molim da i na ovom mestu nađu izraz moje zahvalnosti.

Za deo realizovanih programskih eksperimenata nad korpusom srpskohrvatskog jezika, zahvalu dugujem **Računskom centru Republičkog zavoda za statistiku SR Srbije**, bez čije računarske podrške ovi eksperimenti ne bi bili mogući.

SADRŽAJ

Virtual Library of Faculty of Mathematics - University of Belgrade
elibrary.matf.bg.ac.rs

Virtual Library of Faculty of Mathematics - University of Belgrade
elibrary.matf.bg.ac.rs

Uvodna reč

Istraživanja koja su predmet ovog rada pripadaju, u svom najopštijem vidu, računarskoj metodologiji **automatska obrada teksta**. Tekst, kao predmet računarske obrade, premda može biti formalno definisan, je, u svojoj biti, supstancija organizovana prirodnim jezikom. Šta više, tekst koji se javlja kao predmet konkretne računarske obrade je uvek organizovan konkretnim prirodnim jezikom. Iz ove činjenice sledi da se na nivou konstrukcije programskih alata za obradu konkretnog prirodnog jezika mora prepoznati njegova prirodno-jezička strukturiranost. Osnovni cilj teze je, stoga, da opiše i detaljno istraži neke probleme koji se javljaju u automatskoj obradi teksta a indukovani su njegovom organizovanošću srpskohrvatskim jezikom. U tom svetlu su razmotreni različiti problemi u obradi teksta koji se javljaju na nivou II artikulacije (u Martineovom smislu) prirodnog jezika. Posebna pažnja je posvećena problemu rastavljanja reči na kraju retka što predstavlja centralni problem u projektovanju kvalitetnih algoritama za oblikovanje pasusa. Rezultati ovih istraživanja su utemeljeni na računarskoj analizi korpusa savremenog srpskohrvatskog jezika pa se stoga njihova preciznost duguje eksperimentalnoj sredini u kojoj su nastali (za razliku, na primer, od postojećih Pravopisnih rešenja). Na osnovu ovih znanja, realizovane su različite strategije rastavljanja reči na kraju retka i ugrađene u eksperimentalnu programsku okolinu u kojoj su iscrpno testirane.

Napomena: Rezultati koji su izloženi u ovoj tezi su ostvareni kroz istraživanja vezana za područje automatske obrade teksta na srpskohrvatskom jeziku. U tom svetlu, u odnosu na taksonomiju koju je usvojila *Association for Computing Machinery* [Kategorije i predmetni opis *ACM Computing Reviews*] ova istraživanja pripadaju oblasti **I Računarske metodologije**: I.7 Obrada teksta (H.4); I.7.1 Uređivanje teksta i I.7.2 Oblikovanje dokumenta. Dodatne ključne reči i fraze su: **rastavljanje reči, srpskohrvatski jezik, korekcija teksta**.

PRILOZI

Virtual Library of Faculty of Mathematics - University of Belgrade
elibrary.matf.bg.ac.rs

1 UVOD

U vreme nastanka i razvoja računara i njihovog postepenog uključivanja u mnoge oblasti ljudske delatnosti, računari su se prvenstveno koristili za obradu numeričkih podataka. Grubo govoreći, korisnici su računaru saopštavali numeričke podatke kodirane skupom dekadnih cifara a računar je, pošto obavi predviđena izračunavanja, čoveku saopštavao rezultate u obliku brojeva, jednostavnih tabela uz, eventualno, neka kraća tekstualna objašnjenja. Danas nam izgleda sasvim prirodno da se ovakvo stanje nije moglo dugo održati. S jedne strane, pisani tekst, uz sliku i govor, predstavlja osnovno komunikacijsko sredstvo među ljudima koje ima dugu istoriju te se nije moglo očekivati da će ga se ljudi u novoj informatičkoj eri odreći. S druge strane, ubrzo se pokazalo da mogućnosti računara daleko prevazilaze sposobnost računanja i memorisanja numeričkih podataka.

Iako tekst predstavlja osnovno komunikacijsko sredstvo među ljudima, definisanje pojma *tekst* nije nimalo lako. Tekst je, u najopštijem smislu, informacija kodirana karakterima ili sekvencijama karaktera, pri čemu se pod karakterima podrazumevaju slova, cifre, interpunkcijski i specijalni znaci. Međutim, da bi ova definicija bila preciznija mora se dodati da ova sekvencija karaktera mora da zadovoljava izvesne (gramatičke, ortografske, grafičke, idr.) zahteve određenog jezika.

Bliže određenje pojma tekst dobijamo ako ga razložimo na njegove tri osnovne karakteristike [Boitet85]:

- *Forma* (ili *izgled*) teksta predstavlja sve što karakteriše spoljašnju pojavu teksta na papiru, od izgleda i veličine slova, širine redaka, izgleda margina, razmaka između redaka, pa do načina isticanja pojedinih delova teksta.

- *Struktura* teksta predstavlja logičku podelu teksta u hijerarhijski povezane delove, kakvi su: tomovi, delovi, glave, odeljci, tačke, pasusi, liste, tabele, slike, dijagrami, itd. Struktura teksta zavisi od tipa teksta: za tekst drame karakteristična je hijerarhijska podela na činove, scene i lica dok je za tekst pesme karakteristična podela u strofe i retke.

- *Sadržaj* teksta predstavlja sekvenciju reči, ili rečenica, određenog jezika koja nosi neku informaciju.

Premda je u ovakvom određenju pojma tekst, izdvojen sadržaj teksta kao onaj njegov deo, ili bolje rečno vid, koji prenosi informaciju, značajno je istaći da su za prenošenje poruke čitaocu izuzetno značajni i forma i sadržaj tekst. Zamislimo da iz nekog teksta uklonimo sve promene pismovne vrste. Takav tekst čitaocu više ne nosi istu poruku kao originalni tekst, jer iz njega on ne može razabrati kojim delovima teksta je autor želeo da da posebno značenje. Takođe, istom sadržaju se može dati različta struktura a ista struktura može imati različitu spoljašnju reprezentaciju. Sve ove promene sadržaja i strukture, međutim, menjaju poruku autora.

Od kada je računarima počeo da se obrađuje tekst, razvile su se mnoge, relativno udaljene oblasti računarstva kojima je zajedničko svojstvo da su im osnovni podaci koje obrađuju zapisani u nekom prirodnom jeziku. Tako su se razvile zasebne oblasti, kakva je automatsko prevođenje s jednog na drugi prirodni jezik ili pronalaženje

Virtual Library of Faculty of Mathematics - University of Belgrade
elibrary.matf.bg.ac.rs

dokumenata prema sadržaju kao i različite primene računara u onim oblastim ljudskog delovanja u kojima se tekst javlja kao najpodesniji nosilac informacija (bibliotekarstvo, leksikografija, i sl.). Jedna od podoblasti računarstva u kojoj je tekst osnovni oblik podataka je *automatska obrada teksta*, koja se bavi problemima vezanim za beleženje teksta na računarskom medijumu u obliku koji obezbeđuje održavanje sva tri njegova vida i omogućava njegovu reprodukciju u papirnom obliku.

1.1 Programska podrška automatskoj obradi teksta

Tipovi programske opreme namenjene automatskoj obradi teksta, njihove osnovne funkcije i mogućnosti najlakše se mogu predstaviti kroz kraći istorijski pregled nastanka i razvoja ove programske opreme.

Uređivač (engl. *editor)* je interaktivan računarski program pomoću koga korisnik može da kreira, menja i memoriše željeni dokument. Dokument može da sadrži tekst na prirodnom jeziku, ali i računarski program, matematičke formule, tabele, linijsku grafiku, fotografiju - dakle, sve ono što se može naći zabeleženo na papiru [Meyrow82]. Proces uređivanja se sastoji od interaktivnog dijaloga korisnika sa računarom u kome se, pre svega, izdvaja deo dokumenta koji treba da se nalazi u vidnom polju korisnika i koji treba na neki način izmeniti. Zatim se određuje kako će se ovaj izdvojeni deo dokumenta trenutno oblikovati i prikazati na izlaznom medijumu. U sledećem koraku se navode i izvršavaju operacije koje menjaju dokument a na kraju se na odgovarajući način menja izgled izmenjenog dela dokumenta.

Ovako definisan proces uređivanja uključuje, pre svega, operaciju kretanja kroz dokument da bi se u njemu pronašao onaj deo koji je u tom trenutku od interesa za korisnika. Iz ovog dela dokumenta izdvaja se njegov relevantni podskup, koji se zatim oblikuje i prikazuje na izlaznom medijumu. Uređivačke operacije kreiraju i menjaju dokument. Elementi nad kojim operišu uređivačke operacije zavise od namene uređivača. Uređivači namenjeni pripremi rukopisa, na primer, operišu nad karekterima, rečima, pasusima, itd. dokuređivači namenjeni pripremi programa operišu nad ključnim rečima i konstrukcijama programskog jezika.

Korisnik komunicira sa uređivačem preko ulaznih i izlaznih uređaja koristeći interaktivni jezik sistema. *Ulazni uređaji* uređivača zadovoljavaju sledeće potrebe korsinika: unos elemenata dokumenta, unos komandi i obeležavanje elemenata nad kojem te komande deluju. Prema njima se i ulazni uređaji mogu podeliti u tri kategorije:

- *tekstualni uređaji*, a to su najčešće alfanumerički tasteri na tastaturi terminala,
- *izborni uređaji*, a to su najčešće funkcionalni tasteri ili, alternativno, odgovarajuće kontrolne sekvencije, i
- *pozicioni uređaji*, a najpoznatiji među njima su miš, komandna palica i sl. ili, alternativno, tasteri za pomeranje kursora, navijanje ekrana i sl.

Na *izlaznom uređaju* uređivača korisnik može da vidi elemente dokumenta koji kreira ili menja, kao i rezultat uređivačkih operacija. Danas su najčešći izlazni uređaji uređivača ekrani terminala ili radnih stanica.

Virtual Library of Faculty of Mathematics - University of Belgrade
elibrary.matf.bg.ac.rs

Interaktivni jezik uređivača se, kao i svaki drugi jezik sastoji od tri komponente: semantičke, sintaksičke i leksičke. Njegova semantička komponenta definiše značenje pojednih operacija uređivača, nad kojim elementima se pojedine operacije mogu izvršiti i kakav je njihov rezultat. Dok su semantičke komponente jezika većine uređivača slične i obuhvataju operacije kao što su *umetni*, *izbaci*, *kopiraj*, *premesti*, *pronadi*, *zameni*, itd., njihove sintaksičke i leksičke komponente se mogu veoma razlikovati. Tako se, na primer, lekseme jezika (odnosno, komande uređivača) mogu zadavati kao karakterske niske, kontrolne sekvencije ili pomoću funkcionalnih tastera. Druga mogućnost je da se komanda izabere iz ponuđenog skupa komandi, takozvanog menija.

Važno je uočiti da su prvi uređivači ušli u upotrebu pre nego što se računarom počeo uređivati tekst na prirodnom jeziku. Ti uređivači su, prvenstveno, bili namenjeni kreiraju i menjaju programa i programskih podataka. Ovakva prvobitna uloga uređivača odrazila se i na sve kasnije uređivače: Oni su namenjeni beleženju, prvenstveno, *sadržaja* teksta. Pomoću uređivača mogu se, takođe, kreirati i menjati *obležja* njegove strukture i forme ali njihova namena nije da iz ovih obeležja reprodukuju željenu formu teksta na izlaznom medijumu.

Ovako definisani uređivači, ili kako se ponekad nazivaju "čisti" uređivači, su i danas u širokoj upotrebi i prisutni su gotovo na svim računarima. **EDLIN** i **Professional Editor** su najpoznatiji takvi uređivači za personalne računare. Danas se oni, međutim, koriste samo za pripremanje programa ili za pripremanje teksta dokumenta za neinteraktivne formatere.

Formateri (engl. *formatter*) su programski paketi koji na papiru ili zaslonu generišu tvrdu, odnosno meku, kopiju dokumenta na osnovu obeležja kojima se specifikuje njegov izgled [Furuta82]. Dokument se, kao subjekat formatiranja, može opisati kao objekat koga čini hijerarhijska struktura primitivnih objekata. Svaki objekat je, pri tome, primer iz klase objekata. Tipične klase dokumenata su poslovno pismo, članak za određeni časopis, teza, i sl. a tipične klase nižeg nivoa su odeljak, pasus, fusnota, jednačina, pismovna vrsta i sl. Objekti se, dalje, klasifikuju kao apstraktni ili konkretni objekti. Jedan *apstraktni objekat* se označava imenom i klasom kojoj pripada. Na primer, ime "tekst" iz klase *reč* označava apstraktni objekat tekst. *Konkretni objekti* se definišu na prostoru jedne ili više stranica i predstavljaju jednu moguću oblikovanu sliku apstraktnog objekta. Na primer, određeni pasus dokumenta, koji je apstraktni objekat iz klase pasus, može se konkretno predstaviti na mnogo različitih načina u zavisnost od izabrane pismovne vrste, širine retka, vrste margina itd.

Osim toga, objekti mogu biti uređeni i neuređeni. Objekti, kao što su reč i pasus, su *uređeni* što znači da možemo govoriti o prvom, poslednjem, prethodnom, sledećem, i sl. Proces formatiranja se nad uređenim objektima odvija sekvencijalno, dok se *neuređeni* objekti, kao što su slike, tabele, formule, itd. mogu nasumično birati.

Operacije formatiranja se nad ovako postavljenim modelom dokumenta mogu definisati kao preslikavanje apstraktnih u konkretne

Virtual Library of Faculty of Mathematics - University of Belgrade
elibrary.matf.bg.ac.rs

objekte. Primeri su transformacija apstraktnog karaktera u njegov opis u konkretnoj pismovnoj vrsti, preslikavanje pasusa u sekvenciju redaka, podela apstraktnog dokumenta na stranice, izrada dvodimenzionalnog matematičkog objekta iz jednodimenzionalnog zapisa, itd. Ovako definisane operacije formatiranje mogu se razvrstati u pet osnovnih grupa:

(1) Izbor konkretnog primitivnog objekta, kao što je pronalaženje određenog karaktera u konkretnoj pismovnoj vrsti;

(2) Horizontalno i vertikalno postavljanje objekata. Primer ovih operacija su uvlake, tabulacija, centriranje, prored, postavljanje supskripta i superskripta i sl;

(3) Horizontalno i vertikalno podešavanje objekata, pod čime se podrazumeva relativno postavljanje objekata u odnosu na druge objekte. Primer je podešavanje znaka jednakosti u matematičkoj formuli, centriranje elementa u tabeli, itd.;

(4) Podela apstraktnih objekata u konkretne objekte. Ove operacije obuhvataju podelu objekata u retke i stranice, baratanje zaglavljima stranica i fusnotama, itd.;

(5) Određivanje razmere objekata. Ove operaciji smanjuju ili povećavaju objekte da bi se smestili u za njih alociran prostor i da bi odgovarali veličini ostalih objekata dokumenta.

Iako su svi formateri snabdeveni ovim operacijama, realizacije konkretnih formatera se veoma razlikuju i možemo ih razvrstati prema sledeće tri osnovne karakteristike:

- *paketni* prema *interaktivnim* formaterima. Paketni formateri, koji se često nazivaju i "čisti" formateri ili kompilatorski formateri, prihvataju opis dokumenta koji je prethodno pripremljen pomoću nekog uređivača. Ove formatere karakteriše da se posle svake izmene dokument mora ponovo formatirati da bi korisnik mogao da vidi rezultat te izmene. Interaktivni formater, koji se još naziva i integrisani uređivač/formater ili interpretativni formater, korisniku prikazuje vidljivu sliku konkretnog dokumenta u toku kreiranja i menjanja dokumenta. Drugim rečima, korisnik može odmah da vidi efekat svake izmene dokumenta na njegov izgled. Primer paketnog formatera je **TeX** (tm. *AMS*) a interaktivnog **Ventura Publisher** (tm. Xerox).

- korisnik komunicira sa formaterom pomoću *proceduralnog* ili *deklarativnog* jezika. U formaterima koji koriste proceduralni jezik, korisnik upotrebljava komande pomoću kojih sam izabira konkretne objekate i raspoređuje ih na prostoru stranice. Formateri koji koriste deklarativni jezik omogućavaju odvajanje sadržaja dokumenta od formaterskih akcija. Na taj način se, dajući drugo značenje deklarativnim komandama, iz istog dokumenta mogu proizvesti potpuno različiti konkretni dokumenti. Primer primene proceduralnog jezika je **TeX** a deklarativnog **GML** (General **M**ark-up **L**anguage, tm. IBM).

- formateri koji obrađuju *samo tekst* prema formaterima koji obrađuju *tekst i sliku*. Formateri koji obrađuju i tekst i sliku su nastali sa ulaskom u široku upotrebu izlaznih uređaja kod kojih je svaka tačka (piksel) adresibilna, kakvi su ekrani visoke rezolucije, laserski štampači, uređaji za fotoslog. Primer formatera koji obrađuje samo tekst je **TeX** a formatera koji obrađuje i tekst i sliku je **Ventura Publisher** (tm. Xerox).

Virtual Library of Faculty of Mathematics - University of Belgrade
elibrary.matf.bg.ac.rs

Razvoj formatera je bio uslovljen razvojem izlaznih uređaja koji su mogli da ponude kvalitet uporedljiv bar sa kvalitetom pisaće mašine (mala i velika slova, mogućnost isticanja delova teksta promenom pismovne vrste, itd.). Dalji ubrzan razvoj izlaznih uređaja uslovio je razvoj formatera u dva osnovna pravca. S jedne strane, ulaskom u široku upotrebu personalnih računara nastali su procesori reči (engl. *word processor*) koji pripadaju klasi integrisanih uređivača i formatera. Pretpostavka, i osnovna karakteristika, prvih procesora reči, od kojih je najpoznatiji **WordStar** (tm. MicroPro), je rad sa neproporcionalnim pismovnim vrstama. Alat za kreiranje novih pismovnih vrsta korisniku nije bio na raspolaganju. Oni su se stoga uglavnom koristili samo kao zamena za pisaću mašinu. Dalji razvoj izlaznih uređaja (ekrani visoke rezolucije, igličasti štampači sa kvalitetnim otiskom) doneo je celu novu generaciju procesora reči koji se popularno nazivaju WYSIWYG sistemi (od engleskog *What You See Is What You Get*). Njihova osnovna karakteristika je da oni korisniku u svakom trenutku prikazuju kako će izgledati dokument odštampan na papiru. Pomenimo samo neke od poznatijih WYSIWYG sistema: **Word** (tm. Microsoft), **ChiWriter** (tm.) i **T**^{3} (tm. TCI Software Research).

Drugi pravac razvoja formatera kretao se ka primeni u grafičkoj industriji. Revolucija u grafičkoj industriji je nastala prelaskom sa "vrućeg" (olovnog) sloga na "hladan" slog, kod koga slika slova nastaje fotografskim putem. Prvi uređaji za fotoslog koristili su rotirajuće diskove na kojima su bile uskladištene slike slova. Ploča potrebna za štampu dobijena je osvetljavanjem filma preko negativa odgovarajućeg slova. Premda je rad ovih uređaja bio podržan računarom, čija je jedina funkcija bila upravljanje fotoslogom, programska podrška slaganju teksta razvijala se nezavisno od programa za uređivanje i formatiranje teksta. Sledeća generacija uređaja za fotoslog zamenila je "analogni" proces osvetljavanja slova "digitalnim". Takav postupak omogućio je, između ostalog, samostalan razvoj pismovnih vrsta, nezavisno od grafičke industrije. Počeli su se razvijati slagači teksta (engl. *typesetter*), koji su kao jedini, ili jedan od mogućih izlaza, nudili fotoslog. Rezolucija uređaja za fotoslog od, na primer, 2500 adresibilnih tačaka po inču otvorila je nove mogućnosti ali je pred formatere postavila nove zahteve: dokument složen ovakvim formaterom ne bi smeo po kvalitetu da zaostaje za "ručno" složenim tekstom. Jedan od najpoznatijih slagača koji zadovoljava ovaj zahtev je **TeX** (tm: AMS)

Poslednjih godina svedoci smo nagle demokratizacije procesa izdavanja. Nju je donelo malo izdavaštvo (engl. *desk-top publishing*) u kome se, na izvestan način, objedinjuju dva do tada razdvojena pravca razvoja formatera: procesori reči i slagači teksta. Prekretnicu su doneli laserski štampač **Apple Laser Writer** i programski paket **PageMaker**. Ulaskom laserskih štampača u široku upotrebu, razvijeni su mnogi drugi paketi koji podržavaju malo izdavaštvo. Da bi se moglo reći da neki programski paket za formatiranje teksta podržava malo izdavaštvo, ono mora da zadovoljava sledeće kriterijume [Seybold87]:

(1) Paket mora da obezbedi takvo slaganje teksta koje po kvalitetu bitno ne odstupa od slagača teksta;

(2) Paket mora da omogući korisniku obavljanje slagačih funkcija na znatno jednostavniji način od onog koji nude sistemi za slaganje teksta u grafičkoj industriji;

Virtual Library of Faculty of Mathematics - University of Belgrade
elibrary.matf.bg.ac.rs

(3) Paket mora da omogući uključivanje u dokument crteža i slika koji su nastali ili skaniranjem postojećeg crteža ili su generisani korišćenjem nekog grafičkog programa;

(4) Paket mora da poseduje dobre uređivačke mogućnosti, koje se mogu meriti sa mogućnostima najboljih procesora reči a primenjuju se na tekst u interaktivnom procesu slaganja.

(5) Paket mora da obezbeđuje izlaz na uređaju čija je rezolucija najmanje 200 tačaka po inču.

Među najpoznatije pakete za malo izdavaštvo, osim pomenutog **PageMaker**-a, spadaju **Ventura Publisher** (tm. Xerox) i **PC TeX** (tm. Personal TeX).

1.2 Programski paketi za obradu teksta kao alat za obradu teksta na prirodnom jeziku

Većina programskih paketa za obradu teksta namenjena je engleskom govornom području tako da čak i oni paketi koji pretenduju na opštu primenljivost, podržavaju prvenstveno obradu tekstova na engleskom jeziku. Ovo je posledica činjenice da je svaki tekst zapisan u određenom prirodnom jeziku koji se, kroz radne pretpostavke u izgradnji paketa, specificira kao engleski jezik. Ovakvoj pretpostavci doprinosi i jednostavnost engleskog alfabeta kao i jednostavnost morfologije. Njihova primena u neengleskoj sredini dovodi, stoga, do različitih neočekivanih problema.

Teškoće, pre svega, stvaraju alfabeti evropskih jezika koji se svi, u manjoj ili većoj meri, razlikuju od engleskog alfabeta. Prve teškoće nastaju pri izboru adekvatne tastature za unos teksta u računar. Premda je korišćenje tastatura inspirisanih rasporedom na QWERTY-tastaturi široko rasprostranjeno, u zavisnosti od jezika kome su namenjene, među njima postoje određene razlike. Za potrebe unosa teksta je preporučljivo korišćenje tastature koja poseduje iste tastere, i u istom redosledu, koje poseduje i tastatura nacionalne pisaće mašine. Sadržaj i izgled nacionalne tastature je najčešće propisan odgovarajućim nacionalnim standardom. Uprkos postojanju odgovarajućih standarda u Jugoslaviji [JUS.K1.002, JUS.K1.003], tastature sa predviđenim rasporedom nisu uvek dostupne. Čak i kada na tastaturi postoje svi potrebni tasteri, dešava se da oni nisu u standardom predviđenom redosledu. Neretko se, stoga, za potrebe unosa srpskohrvatskog teksta koriste američka, nemačka ili engleska tastatura.

Problem tastature je tesno povezan sa izborom skupa karaktera i njihovih binarnih kodova. Većina programskih paketa za obradu teksta zasniva se na predstavljanju teksta u ASCII-kodu kao varijanti ISO 7-bitnog koda ISO 646 [ISO 646]. Ovaj kod sadrži 128 karaktera, od kojih su prva 33 i 127. kontrolni karakteri dok su preostalih 94, grafički karakteri. U ovoj kodnoj tablici slova su samo mala i velika slova engleskog alfabeta A-Z. Pozicijama 64, 91, 92, 93, 94, 96, 123, 124, 125 i 126 nije pridružen nikakav određeni grafički karakter. Ovaj standard propisuje da se ovih deset pozicija može koristiti za nacionalne karaktere i posebne aplikacije. Ukoliko za time ne postoji potreba, preporučuje se da na ovim pozicijama budu karakteri Međunarodne referentne verzije ISI 646 standarda (skraćeno IRV). U sledećoj tabeli je prikazano kako se mala i velika slova *Č*, *Ć*, *Đ*, *Š* i *Ž*

Virtual Library of Faculty of Mathematics - University of Belgrade
elibrary.matf.bg.ac.rs

srpskohrvatskog alfabeta smeštaju u ovih 10 pozicija 7-bitnog koda [JUS.B1.002].

Međutim, kodove koji prema ovom standardu odgovaraju malim i velikim slovima *Č*, *Ć*, *Đ*, *Š* i *Ž* na anglosaksonskim tastaturama generišu tasteri čiji raspored nije pogodan za unos teksta na srpskohrvatskom jeziku. Tako se, na primer, malo i veliko slovo *Ž* i *Č* ne dobijaju pritiskom na isti taster dok se malo slovo *š* dobija u gornjem a veliko slovo *Š* u donjem redu tastature.

Osim toga, proizvođači programskih paketa, usled skučenih mogućnosti koje nudi 7-bitni kod, često ne vode računa o tome da je gornjih 10 karaktera preporučeno za kodiranje karaktera nacionalnih alfabeta, pa im dodeljuju različite kontrolne funkcije.

pozicija	IRV grafički karakter		Srpskohrvatska latinica grafički karakter	
64	@	komercijalno 'at'	Ž	veliko slovo *Ž*
91	[	uglasta zagrada, otvorena	Š	veliko slovo *Š*
92	\	obratna kosa crta	Đ	veliko slovo *Đ*
93	]	uglasta zagrada, zatvorena	Ć	veliko slovo *Ć*
94	^	kapica	Č	veliko slovo *Č*
96	`	kratki akut iznad	ž	malo slovo *Ž*
123	{	vitičasta zagrada, otvorena	š	malo slovo *Š*
124	|	vertikalna spojnica	đ	malo slovo *Đ*
125	}	vitičasta zagrada, zatvorena	ć	malo slovo *Ć*
126	~	titla (ili nadvlaka)	č	malo slovo *Č*

Paketi za obradu teksta nude ponekad korisniku i dodatne mogućnosti, kao što su automatsko formiranje indeksa, sortiranje, itd. Raspored nacionalnih karaktera u kodnoj tablici onemogućava korišćenje ovih dodatnih mogućnosti korisnicima izvan engleskog govornog područja. Kao što je poznato, za ASCII kod, kao i za druge kodove (npr. EBCDIC kod), važi sledeće:

- Unutar kolacione sekvencije velika slova *A* - *Z* raspoređena su u alfabetskom redosledu;

- Unutar kolacione sekvencije mala slova *a* - *z* raspoređena su u alfabetskom redosledu;

- Cifre 0 - 9, kao ni specijalni karakteri, nisu između kodova velikih ili malih slova u kolacionoj sekvenciji.

Kako se prilikom sortiranja ovi programi najčešće oslanjaju na kolacionu sekvenciju, sortirajući zapravo vrednosti celobrojnih ekvivalenata svakog karaktera, tj. kodove. Dok je rezultat sortiranja za engleski jezik ispravan, dotle je za većinu drugih evropskih jezika on neprihvatljiv: tako je, na primer, u verziji ISO 646 koda za srpskohrvatsko latinično pismo slovo *Ž* uvek ispred slova *A* dok su slova *Š*, *Đ*, *Č* i *Ć* iza slova *Z*. Iz istog razloga ne može se dobiti ni ćirilični poredak ključeva koji se sortiraju.

Za unos srpskohrvatskog teksta u računar češće se koristi latinična od ćirilične tastature i odgovarajući 7-bitni kod za

Virtual Library of Faculty of Mathematics - University of Belgrade
elibrary.matf.bg.ac.rs

srpskohrvatsko latinično pismo. Digrafi srpskohrvatskog latiničnog pisma *Nj*, *Lj*, *Dž* i *Dj* se, stoga, unose kao što je to uobičajeno na pisaćoj mašini: na primer, slovo *Nj* će u tekstu biti zapisano sa dva koda, kodom slova *N* i kodom slova *j*. Premda za prikazivanje teksta na ekranu i njegovo štampanje na papiru ovakav unos obično ne stvara veće teškoće, korišćenje dodatnih mogućnosti paketa može biti onemogućeno. Imajući u vidu da digrafi *Nj*, *Lj*, *Dž* i *Dj* zauzimaju zasebno mesto u srpskohrvatskom alfabetu (npr. slovo *Nj* nalazi se u alfabetu između slova *N* i *O*), korišćenje dodatnih mogućnosti paketa koje podrazumevaju sortiranje zasnovano na kolacionoj sekvenciji ne obezbeđuje željeni redosled: na primer, slovo *Nj* se uvek nalazi u okviru slova *N* između reči na *Ni* i *Nk*. Osim toga, često se previđa činjenica da se prilikom unosa teksta velika slova digrafa mogu uneti sa dve različite komibanacije kodova (npr. *Nj* i *NJ*) što kod sortiranja može da stvori dodatne teškoće. Ovakav zapis digrafa stvara teškoće i rutini za automatsko rastavljanje reči na kraju retka koja za svaki takav zapis mora da odluči da li je u pitanju konsonantska grupa ili digraf. I na kraju, ovakav unos digrafa otežava štampanje istog teksta korišćenjem jednom latiničnog a drugi put ćiriličnog pisma. Problem se može rešiti posredstvom postprocesora koji neće svako *Nj* prevesti u Њ.

Dok se za kodove koje koriste paketi za obradu teksta, kao i za pomenute dodatne mogućnosti paketa, može govoriti o implicitnoj pretpostavci da je tekst koji se obrađuje zapisan na engleskom jeziku, u veliki broj ovih paketa ugrađene su komponente koje su eksplicitno namenjene tekstu na engleskom jeziku. Pomenimo samo neke.

Komponenta za automatsko rastavljanje reči na kraju retka, koja je ugrađena u većinu paketa za obradu teksta, obezbeđuje oblikovanje teksta u pasuse koji imaju ujednačenu dužinu retka i optimalan i ravnomeran razmak između reči u retku. Svaka ovakva komponenta može, u principu, na zadovoljavajući način da rastavi reči samo jednog prirodnog jezika. To znači da se za svaki prirodan jezik mora razviti posebna komponenta za rastavljanje reči koja se zasniva na znanju o tom jeziku. Veći deo ovog rada posvećen je mogućim načinima za realizaciju ove komponente za srpskohrvatski jezik.

U neke od programskih paketa za obradu teksta ugrađena je komponenta za otkrivanje i korekciju grešaka u tekstu. Pod greškom u tekstu se ovde podrazumeva odstupanje reči iz teksta od znanja ugrađenog u ovu komponentu. Dakle, greške u tekstu se otkrivaju i koriguju na nivou reči, ne ulazeći u sintaksu i semantiku teksta. Dva su najčešća izvora grešaka u tekstu: tipografske greške i neznanje autora. Grubo govoreći, moguće strategije za realizaciju ove komponente su [Peterson80b]:

- Frekvencijski rečnik svih reči iz teksta koji se koriguje. Ovakav rečnik predstavlja samo pomoć autoru ili korektoru jer se tipografske greške nalaze, verovatno, među rečima sa najmanjom frekvencijom.

- Komponenta se zasniva na kvantitativnim obeležjima određenog prirodnog jezika. Ovakva komponenta se najčešće realizuje bilo preko rečnika digrama, trigrama, konsonantskih grupa, slogova, itd. formiranih nad korpusom raznorodnih tekstova bilo preko sličnih rečnika formiranih nad tekstom koji se koriguje.

Virtual Library of Faculty of Mathematics - University of Belgrade
elibrary.matf.bg.ac.rs

- Komponenta koristi rečnik reči određenog prirodnog jezika. Svaka reč izdvojena iz teksta koji se koriguje traži se u rečniku. Reč koja je u rečniku proglašava se ispravnom a svaka reč koja nije u rečniku je potencijalno neispravna.

Jasno je da se sve ove strategije, osim prve koja je dosta primitivna, zasnivaju na znanju o određenom prirodnom jeziku. Međutim, ako je osnovni kriterijum kvaliteta ovih komponenta da one treba što manje neispravnih reči da proglase ispravnim i što manje ispravnih reči da označe za potencijalno neispravne onda strategija koja se zasniva na ugrađenom rečniku pokazuje najbolje rezultate. No, ova strategija, koja se za engleski jezik može bez većih teškoća primeniti postavlja pred jezike sa bogatom morfologijom, kakav je srpskohrvatski, niz složenih problema, kakvi su: izgradnja rečnika imajući u vidu razne oblike reči, dovođenje dimenzija rečnika u granice koje dopuštaju efikasno skladištenje i pretraživanje rečnika, itd.

U neke pakete za obradu teksta ugrađene su komponente za slanje pošte. U ove komponente nije ugrađeno nikakvo znanje o prirodi ponašanja imena u jeziku, što ne remeti njihovo funkcionisanje za engleski jezik. Za srpskohrvatski jezik su ovako jednostavno zamišljene komponente u određenim slučajevima neprimenljive a, pre svega, u slučajevima kada je potrebno deo adrese (npr. ime) uključiti u vezani tekst.

Ovim se lista komponenta paketa za obradu teksta u koje je, implicitno ili eksplicitno, ugrađeno znanje o nekom prirodnom jeziku ne iscrpljuje. Postoje, tako, još i komponente koje pružaju autoru pomoć pri pisanju: komponente za stilsku analizu, rečnici sinonima, terminološki rečnici, tezaurusi i sl. [Vitas85b].

Virtual Library of Faculty of Mathematics - University of Belgrade
elibrary.matf.bg.ac.rs

2 UREĐIVANJE PASUSA U RETKE I PROBLEM RASTAVLJANJA REČI NA KRAJU RETKA

2.1 PODELA PASUSA U RETKE

U tradiciji pisane reči je predstavljanje teksta u dvodimenzionalnom obliku: Tekst je podeljen na stranice a na svakoj stranici tekst je podeljen na retke. Srednjevekovni prepisivači koji su, prepisujući tekstove rukom, stvarali rukopise izvanredne lepote vodili su računa o tome da tekst podele u retke tako da desna ivica teksta bude što je moguće više poravnata. I u to vreme se smatralo da podela teksta u retke približno iste širine čitaocu olakšava čitanje.

Sa pojavom štamparija, prvi štampari su se trudili da održe ovu tradiciju. U [Knuth81] je dat istorijski pregled koji pokazuje na koje se sve načine razrešavanjo problem podele pasusa u retke i poravnavanja ivica teksta. Zanimljiv je primer prve višejezične Biblije, nastale u Španiji u vremenu od 1514. do 1517. godine, u kojoj je na jednoj stranici u više uskih kolona slagan tekst Biblije na raznim jezicima. U latinskoj i hebrejskoj koloni je,. na primer, primenjen neobičan pristup: redak je s desne, odnosno, leve strane popunjavan nekim specijalnim simbolom do pune širine retka. U drugim kolonama iste Biblije primenjivan je drugi pristup: razmak između reči je podešavan da bi se postigla puna širina retka. Reči koje ne bi mogle da stanu u redak, deljene su crticom na kraju retka, ali, kada crtica nije mogla da stane u redak, ona je jednostavno izostavljana.

Višejezična verzija Psalma koja je nastala u približno isto vreme je slagana u osam kolona: u sedam kolona je bio tekst Psalma na raznim jezicima i on je, jer se radi o poeziji, složen sa jednom iskrzanom ivicom dok je osma kolona, u kojoj su date napomene, složena sa poravnatim ivicama. Ravnomerna širina retka je, uglavnom, postignuta podešavanjem razmaka između reči, ali štampar je koristio i neke druge "trikove". Tako je, na primer,u to vreme bilo uobičajeno da se u tekstu na latinskom jeziku *m* i *n* zamenjuju titlom iznad prethodnog vokala: na primer, *premium* je zamenjivano sa *premiu͂*.

U većini ovih starih knjiga se uočava da, iako su se štampari trudili da poravnavaju ivice teksta podešavanjem razmaka između reči, oni se uglavnom se vraćali unazad da ponovo slože početni deo pasusa, iako bi to u nekim slučajevima poboljšalo podelu pasusa u retke u celini. Tako se često sreću "gusto" složeni redak iza koga sledi "labavo" složeni redak ili veliki broj redaka koji se završavaju sa podeljenom reči. Na primer, u višejezičnoj Bibliji nastaloj 1572. godine u Antverpenu, 40% svih redaka se završava podeljenom reči.

Sa napretkom štamparske tehnike nastajale su i utvrđivane norme dobre tipografije. Ove norme obično nisu bile sistematizovane već su najčešće bile iskustvene i prenosile su se, u skladu sa tradicijama zanatstva, s kolena na koleno. Tako su nastale i razne tipografske škole, pa se i danas razlikuju anglosaksonske od francuskih tipografskih normi. Ljudima izvan struke ove norme nisu bile poznate, o njima nisu mogli da se informišu iz knjiga a najčešće nisu imali ni potrebe da o njima nešto znaju. Danas, kada preovlađuje uverenje da autor, uz pomoć svog personalnog računara, može i sam da bude štampar (ili bar slagač) svog teksta, izražena je neophodnost da se ove

Virtual Library of Faculty of Mathematics - University of Belgrade
elibrary.matf.bg.ac.rs

iskustvene tipografske norme sistematizuju i formulišu tako da ih i laici mogu razumeti i primeniti.

Osnove tipografskih normi koje se odnose na podelu pasusa u retke, mogle bi se formulisati na sledeći način [Mesaroš83]:

1. Preporučuje se da se pasus deli u retke približno iste širine. Ova preporuka se odnosi na pasuse koji se slažu sa poravnatim ivicama, ali i na one koji se slažu sa jednom iskrzanom ivicom (osim, naravno, poezije i drugih tekstova kod kojih je podela pasusa u retke unapred definisana). Iz ovoga sledi da su problem podele pasusa u retke i problem poravnavanja ivica u tipografiji dva odvojena, iako međusobno povezana, problema.

2. Preporučuje se da se kod poravnavanja ivica teksta, puna širina teksta postigne podešavanjem širine razmaka između *reči.* Alternativa ovom rešenju je podešavanje razmaka između *slova*, međutim to se retko smatra poželjnim.

3. Preporučuje se da se svi reci pasusa slože sa razmakom između reči čija je širina što približnija *optimalnoj.* U svakom slučaju, širina razmaka između reči ne bi trebala da bude manja od *minimalne* niti veća od *maksimalne.* Veličina optimalne, minimalne i maksimalne širine između reči obično zavisi od upotrebljene pismovne veličine, ali i od toga kakav kvalitet se želi postići, željene širine retka, i sl. Neke norme, na primer, preporučuju da širina optimalnog razmaka između reči bude jednaka trećini *ema*, gde je *em* veličina upotrebljenog pisma i odgovara širini velikog slova M, da širina minimalnog razmaka bude jednaka četvrtini *ema* a širina maksimalnog razmaka tri četvrtine *ema*.

4. Klasični uređaji za slaganje teksta, a i većina programskih paketa za slaganje teksta, primenjuju postupak za podelu pasusa u retke koji se može neformalno opisati na sledeći način:

- Redak se popunjava rečima sa optimalnim razmakom između reči i posle svake dodate reči proverava se da li je širina retka dostigla ili premašila željenu širinu retka;
- Ako jeste, širina razmaka između reči se smanjuje da bi se postigla željena širina retka a zatim se proverava da li je manja od minimalne;
- Ako jeste, poslednja reč iz retka se prebacuje u sledeći redak a širina razmaka između reči se povećava. Zatim se proverava da li je širina razmaka veća od maksimalne;
- Ako jeste, poslednja reč se crticom deli na kraju retka;
- Ako se reč ne može podeliti, redak se slaže sa razmakom čija je širina veća od maksimalne.

Treba napomenuti da se primenom ovog postupka reci slažu jedan po jedan, i jednom složeni redak (*isključeni redak*, u tipografskoj terminologiji) se više nikada ponovo ne slaže da bi se ostvarilo povoljnije rešenje. Može se dokazati da se primenom ovog postupka pasus slaže u *minimalan* broj redaka [Achugbue81] ali ne i na

Virtual Library of Faculty of Mathematics - University of Belgrade
elibrary.matf.bg.ac.rs

optimalan način (pri čemu sam štampar, odnosno, slagač treba da definiše šta je *optimalno* složeni pasus [Knuth81]).

5. Osim ovih opštih normi vezanih za podelu pasusa u retke, postoje i mnogi drugi zahtevi koji predstavljaju deo tipografskih normi. Pomenimo samo neke:

- nije poželjno da u složenom pasusu "labavo" složeni redak (kod koga je širina razmaka između reči značajno veća od optimalne) i "gusto" složeni redak (kod koga je širina razmaka između reči značajno manja od optimalne) budu jedan do drugog;
- nije poželjno da se dva uzastopna retka završavaju podeljenom reči, jer se smatra da svaki redak završen podeljenom reči bar na trenutak zaustavlja proces čitanja;
- nije poželjno da se pretposlednji redak pasusa završava podeljenom reči.

U odnosu na staru štamparsku tehniku kaja je koristila olovni (ili "vrući") slog, za programske sisteme za slaganje teksta podešavanje ivica teksta ne predstavlja veći problem. Problem se u osnovi svodi na jednostavno izračunavanje širine razmaka između reči, koja se izražava u nekoj veličini nezavisnoj od štampača, i njeno pretvaranje, u trenutku štampanja, u veličinu koja zavisi od rastera štampača.

Međutim, ostaje da se reši problem podele pasusa u retke. Ručni slagači su najčešće koristili gore opisanu proceduru za podelu pasusa u retke a i mnogi programski sistemi za slaganje teksta su takođe usvojili istu proceduru. Mogućnosti programskih sistema su, međutim, mnogo veće. Imajući u vidu sposobnost računara da memoriše veliki broj podataka, da veoma brzo obavi veliki broj računskih operacija kao i mogućnost programskih paketa da odvoji procese unosa teksta, slaganja i štampanja mogu se konstruisati algoritmi koji *optimizuju* proces podele pasusa u retke [Achugbue81, Knuth81, Pringle81, Samet82]. Proces optimizacije ovde treba samo uslovno shvatiti jer se većina kriterijuma optimizacije zasniva na estetskim kategorijama. Jedan posebno sofisticiran algoritam za optimizaciju procesa podele pasusa u retke je primenjen u programskom sistemu za slaganje teksta TeX [Knuth84].

Kod ručnog slaganja teksta, o rastavljanju reči na kraju retka su vodili računa sami slagači, oslanjajući se na sopstveno znanje i često intuitivan osećaj o tome kako reči treba rastavljati. Namena programskih sistema za slaganje teksta je da preuzme sve, ili bar većinu, funkcija ručnog slagača. Tako je nastala potreba da se u ove sisteme ugradi potrebno znanje o rastavljanju reči na kraju retka koje bi omogućilo potpunu automatizaciju procesa slaganja teksta. Potrebno je, takođe, da automatski rastavljač postiže zadovoljavajući kvalitet koji ne bi kompromitovao vrednost programskog paketa za slaganje teksta.

2.2 RASTAVLJANJE REČI NA KRAJU RETKA

Čovek može jednostavno da rastavi reč ako je polako izgovori i u sebi označi kao moguće tačke na kojima se reč može rastaviti svaki

Virtual Library of Faculty of Mathematics - University of Belgrade
elibrary.matf.bg.ac.rs

prekid u izgovoru reči. Stoga se može reći da je problem rastavljanja reči na kraju retka usko povezan sa problemom rastavljanja reči na slogove. Kako se slog, obično, shvata kao najmanja artikulaciona jedinica govora, rastavljanje reči na granici dva sloga omogućava lak izgovor prenesenog dela reči i ne remeti ritam čitanja.

Međutim, s jedne strane, pogramski sistem za slaganje teksta mora iz pisane, a ne iz izgovorene reči da dokuči moguće tačke na kojima se reč može rastaviti. S druge strane, fonematska struktura sloga je određena skupom pravila koja su specifična za svaki prirodan jezik. Tako se, na primer, reči koje se u dva prirodna jezika pišu na istovetan način i imaju isto značenje mogu različito rastavljati na slogove. Primer je reč *magnificence* u engleskom (*mag-nif-i-cence*) i francuskom jeziku (*ma-gni-fi-cence*).

Pravila za rastavljanje reči na kraju retka mogu se, grubo govoreći, podeliti u tri grupe:

- ortografska pravila, koja se mogu definisati na osnovu zapisa reči;
- fonetska pravila, koja se definišu na osnovu izgovora reči;
- semantička pravila, koja zavise od etimologije i značenja reči.

U većini prirodnih jezika među pravilima za rastavljanje reči na kraju retka preovlađuju pravila iz jedne od ovih grupa. Pravila za rastavljanje reči na kraju retka i preovlađujući uticaj pojedinih grupa pravila ćemo, u svetlu realizacije automatskog rastavljača reči na kraju retka, analizirati na primeru engleskog i srpskohrvatskog jezika.

2.2.1 Rastavljanje reči engleskog jezika na kraju retka

U engleskom jeziku prisutna su pravila iz svake od navedenih grupa, koja se ovde analiziraju na osnovu definicije date u uvodu Vebsterovog rečnika [Webster61].

(a). Jedino čisto ortografsko pravilo u engleskom jeziku je:

Diftonzi (*oi*, *ay*, itd.), digrafi (*th*, *sh*, itd.) i trigrafi (*eau*, itd.) se ne rastavljaju. Udvojeni konsonanti se obično rastavljaju.

(b). Primer čistog fonetskog pravila je:

Dva susedna vokala koji se odvojeno izgovaraju se rastavljaju.

Postoje, takođe, i pravila koja su kombinacija ortografskog i fonetskog pravila:

Konsonant (ili digraf ili trigraf) između dva vokala obično se pridružuje drugom vokalu (*ortografsko pravilo*). Međutim, konsonant ostaje sa prethodnim vokalom ako je:

- prethodni vokal kratak i naglašen;

Virtual Library of Faculty of Mathematics - University of Belgrade
elibrary.matf.bg.ac.rs

- prethodni vokal nije naglašen a izgovara se kao *u* u *sun* ili ... (*fonetsko pravilo*).

(c). Primer čistog semantičkog pravila je:

Delovi složene reči se rastavljaju.

Postoje, takođe, i pravila koja su kombinacija ortografskih i semantičkih pravila. Primer je:

Uobičajeni prefiksi i sufiksi odvajaju se od osnove ako su na nju direktno dodati ili ako sufiks zamenjuje muklo *e* (*ortografsko pravilo*) osim ako je dodavanjem afiksa promenjen zapis reči ili ako reč dobija specijalno značenje (*semantičko pravilo*).

Većina pravila za rastavljanje reči engleskog jezika je iz grupe fonetskih pravila, što potiče od nefonetičnosti engleskog pisma. Tako se, na primer, slovo *a* izgovara različito u rečima *hat*, *pass*, *came*, *water*, *dare* i *ago*. Može se reči da se iz zapisa reči ne može zaključiti kako se reč izgovara a rastavljanje reči zavisi, u mnogim slučajevima, od izgovora.

Naglašenost slogova u nekim slučajevima takođe ima uticaja na rastavljanje. Na primer, reč *metal* ima naglasak na prvom slogu, dok je u pridevu *metallic* naglasak na drugom slogu. Rezultat toga je da se imenica rastavlja *met-al* dok se pridev rastavlja *me-tal-lic*. Još karakterističniji je primer imenice *dem-on-stra-tion* prema pridevu *de-mon-stra-tive* gde je promena samo dva slova dovela do promene naglaska slogova i drukčije podele reči na poziciji koja je udaljena čak za devet mesta.

Gotovo sva pravila za rastavljanje reči engleskog jezika imaju izuzetke. Tako je, na primer, *th* gotovo uvek digraf, i kao takav se ne rastavlja, međutim postoje reči u kojima je *th* konsonantska grupa: primer je reč *pothole*. Takođe, sufiks *-ing* je veoma frekventan u engleskom jeziku i, prema pravilu koje važi za uobičajene afikse, odvaja se od osnove. Međutim, postoje reči, kao na primer, *ring*, u kojima *-ing* nije sufiks i za njih ovo pravilo ne važi.

Važan je, takođe, redosled primene pravila. Udvojeni konsonanti se rastavljaju, kao na primer u reči *let-ter*, ali postoje situacije u kojima se ovo pravilo poništava primenom nekog pravila većeg prioriteta: primer su reči *pass-ing* i *pass-port*. Ovi primeri govore da paravilo o odvajanju afiksa od osnove i pravilo o rastavljanju složene reči na njene delove imaju veći prioritet od pravila o rastavljanju udvojenih konsonanata.

2.2.2 Rastavljanje reči srpskohrvatskog jezika na kraju retka

Kao što je već rečeno, u srpskohrvatskom jeziku, kao i u većini drugih jezika, reči se na kraju retka, najčešće, rastavljaju na granici sloga. Stoga su pravila za rastavljanje reči na kraju retka tesno povezana sa pravilima za određivanje granice sloga.

Slog se sastoji iz centralne foneme i marginalnih fonema koje se grupišu oko centralne foneme. Centralne foneme su najčešće vokali,

Virtual Library of Faculty of Mathematics - University of Belgrade
elibrary.matf.bg.ac.rs

ali u srpskohrvatskom jeziku to mogu biti i foneme *r* i *l*. Marginalne foneme su konsonanti. Razliku se dve vrste slogova: *otvoreni slog* (sa strukturom **-V-** ili **-CV-**) i *zatvoreni slog* (sve ostale strukture). Postojanje zatvorenih slogova u srpskohrvatskom jeziku povlači postojanje konsonantskih grupa. Problem određivanje granice sloga se, stoga, u osnovi svodi na identifikovanje zatvorenih slogova, odnosno, na utvrđivanje pravila za rastavljanje konsonantskih grupa.

Za srpskohrvatski jezik su, u više navrata, predloženi skupovi pravila za određivanje granice sloga, odnosno rastavljanje konsonantskih grupa, [Belić34], [Stevanović64]. Međutim, nijedan od ovih skupova pravila nije od lingvista dobio punu podršku. Sam Pravopis srpskohrvatskog književnog jezika @Pravopis60# daje određene preporuke za rastavljanje reči srpskohrvatskog književnog jezika na kraju retka, u kojima se, na izvestan način, zaobilazi precizan odgovor na problem identifikovanja zatvorenih slogova. Te preporuke su sledeće:

a) Dva vokala koja se u reči nalaze jedan do drugoga rastavljaju se, ali nije pogrešno ni ako se ne rastave. Na primer, *ža-oka* ili *žao-ka*.

b) Dva ili više konsonanata sa kraja reči nije dopušteno bez vokala preneti u novi redak. Na primer, ne *mlado-st* već *mla-dost*.

c) Ako se između dva vokala nalazi jedan konsonant, on pripada drugom vokalu i sa njim se prenosi u drugi redak. Na primer, *bo-rac*.

d) Ako se između dva vokala nalaze dva ili više konsonanata, u novi redak se prenose oni konsonanti koji se sa vokalom iz sebe mogu lako izgovoriti (na primer, *ze-mlja* ili *zem-lja*) dok se ne preporučuje da se konsonantska grupa teška za izgovor prenosi u novi redak. Na primer, ne *bu-mbar* već *bum-bar*.

e) Ako se u složenoj reči razaznaju delovi od kojih je sastavljena, ona se rastavlja na te delove, a svaki od njih se rastavlja kao reč za sebe (na primer, *ob-istiniti*). Ako se, pak, ovi delovi ne razaznaju, reč se rastavlja kao da nije složena. Na primer, ne *raz-um* već *ra-zum*.

Preporuke a) i c) odnose se na identifikovanje otvorenih slogova i njihova primena, uglavnom, ne stvara nedoumice. Preporuka b) se takođe odnosi na identifikovanje otvorenog sloga, jer dva ili više konsonanata na kraju reči, kao marginalne foneme, ne mogu obrazovati slog. Međutim, ovako formulisana preporuka, ima izuzetaka: Na primer, *r* se na kraju reči iza konsonanta pojavljuje u ulozi vokala i može biti centralna fonema (npr. *masakr*) a ulogu centralne foneme mogu imati i glasovi *l*, *m* i *n* na kraju reči iza konsonanta (npr. *bicikl*).

Preporuka d) odnosi se na identifikovanje zatvorenih slogova, odnosno, na rastavljanje konsonantskih grupa. Pravopis, međutim, samo uvodi intuitivne pojmove "lako" i "teško" izgovorljivih grupa i daje nekoliko primera a ne daje nikakve dalje sugestije za razlikovanje jedne "vrste" konsonantskih grupa od drugih.

Preporuka e) odnosi se na semantička pravila za rastavljanje reči na kraju retka. Kao što se videlo i na primeru engleskog jezika, kod automatske primene ovih pravila nailazi se na dosta teškoća. Tako

Virtual Library of Faculty of Mathematics - University of Belgrade
elibrary.matf.bg.ac.rs

je, na primer, *ob-* prefiks u reči *obisitniti* ali ne i u reči *običaj*. Poseban problem predstavlja prosuđivanje da li se u nekoj složenoj reči razaznaju delovi od kojih je nastala. Na primer, kako se može prosuditi da li se u reči *izdajnik* razaznaju delovi od kojih je ona nastala. Iz ove preporuke bi se moglo zaključiti da se apsolutna pravila za rastavljanje reči na kraju retka ne mogu propisati već da ona, u izvesnoj meri, zavise i od osećaja svakog pojedinca.

Dodatni problem kod primene ove preporuke stvaraju i homografi. Tako je, na primer, u reči *podići* (*podiđem*) prefiks *pod-* dok je u reči *podići* (*podignem*) prefiks *po-*. Još jedan primer je reč *podaviti* (*podavijem*) u kojoj je prefiks *poda-* dok je u reči *podaviti* (*podavim*) prefiks *po-*.

Prilikom korišćenja latiničnog pisma, poseban problem kod primene svih ovih preporuka čini razlikovanje digrafa *nj*, *lj*, *dž* i *dj* od odgovarajućih konsonantskih grupa. Tako je, na primer, *nj* u reči *konjina* digraf dok je u reči *konjunkcija* konsonantska grupa. Isto tako je *dž* digraf u rečima *podžaveljati se* i *nadžakbaba* a konsonantska grupa u rečima *podžupan* i *nadživeti*.

2.3 STRATEGIJE ZA REALIZACIJU AUTOMATSKOG RASTAVLJANJA REČI

Nezavisno od prirodnog jezika kome su namenjene, moguće su sledeće strategije za realizaciju procedura za automatsko rastavljanje reči na kraju retka:

2.3.1 Korišćenje rečnika reči

Procedure koje se zasnivaju na ovoj strategiji koriste rečnik reči nekog prirodnog jezika, pri čemu je svaka reč iz rečnika obeležena na mestima na kojima se može rastaviti. Svaka reč, za koju se u procesu oblikovanja pasusa ustanovi da se mora rastaviti, traži se u rečniku. Ako se reč nalazi u rečniku, bira se ona tačka podele reči koja je najpovoljnija sa stanovišta podele pasusa u retke.

Ova strategija je za engleski jezik primenjena već 1962. godine u okviru prvog komercijalnog programa za oblikovanje pasusa koji je primenjivan za pripremu za štampu časopisa *Palm Beach Post Times*. Program je realizovan na računaru RCA 301 i koristio je rečnik od 30.000 reči koji se nalazio na magnetnoj traci.

Da bi se primenila ova strategija moraju biti obezbeđeni određeni uslovi. Pre svega se postavlja pitanje da li za prirodan jezik za koji treba da konstruisati proceduru za rastavljanje reči traženi rečnik postoji. Za engleski jezik, na primer, postoji više opštih rečnika u kojima su odrednice obeležene na mestima na kojima se mogu rastaviti [Hornby86], [Longman]. Međutim, za većinu drugih jezika, među koje spada i srpskohrvatski jezik, ovakvi rečnici ne postoje.

Sledeći problem koji treba razrešiti je da li postoji rečnik reči, obeleženih za rastavljanje ili ne, u mašinski čitljivom obliku. Za mnoge jezike danas postoje rečnici u mašinski čitljivom obliku koji su nastali bilo iščitavanjem postojećeg rečnika pomoću optičkog čitača karaktera, bilo korišćenjem računara u samom procesu izrade rečnika.

Virtual Library of Faculty of Mathematics - University of Belgrade
elibrary.matf.bg.ac.rs

Ukoliko ovakvi rečnici postoje, oni se naknadnom obradom, reč po reč, mogu snabedeti informacijama potrebnim za rastavljanje na kraju retka.

Uobičajeno je da se u rečnicima, kao odrednica, navodi samo osnovni oblik reči ili, najviše, deo njene paradigme. To, pak, znači da se u mašinski čitljivom rečniku ne nalaze svi oblici reči, pa se sve reči koje se mogu pojaviti u tekstu neće moći rastaviti korišćenjem ovakvog rečnika. Moguće rešenje je da se od rečnika odrednica, ručno ili automatski, generiše rečnik koji će sadržati sve oblike reči.

Čak i kada se za neki prirodni jezik mogu zadovoljiti svi navedeni uslovi, izbor ove strategije za rastavljanje reči na kraju retka ima određene nedostatke. Ukoliko se želi postići potpuna tačnost procedure za automatsko rastavljanje reči, rečnik mora biti dovoljno velik (100.000, i više, osnovnih oblika reči). U svetlu primene u proceduri za rastavljanje reči, koja se u procesu automatskog slaganja teksta poziva relativno često, postavlja se pitanje efikasnog pretraživanja ovako velikog rečnika. Osim toga, u jednom velikom opštem rečniku naći će se mnoge arhaične reči čija je verovatnoća pojavljivanja u tekstu veoma mala, dok mnogih ličnih imena i termina specifični za neku usku naučnu oblast neće biti. Svaki prirodan jezik se svakodnevno obogaćuje novim rečima, pa se postavlja i pitanje redovnog ažuriranja rečnika novim rečima.

2.3.2 Korišćenje pravila

Alternativna strategija se zasniva na pretpostavci da je za zadati prirodan jezik moguće utvrditi koje se niske slova, a u najjednostavnijem slučaju, to su kombinacije dva slova, obavezno rastavljaju, odnosno koje se nikad ne rastavljaju. Za razliku od prethodne strategije, kod koje je svo znanje o prirodnom jeziku za koji se konstruiše rastavljač sadržano u rečniku, u ovoj strategiji je svo znanje o prirodnom jeziku sadržano u pravilima, tj. u samoj proceduri.

Jedan ovakav pristup je za engleski jezik opisan u [Rich65]. U opisanu proceduru je ugrađen sledeći jednostavan skup pravila za koji autori smatraju da daju zadovoljavajuće rezultate:

- Bar jedan vokal, ne računajući *E* iz niski *-E*, *-ES* i *-ED* na kraju reči, mora se preneti u naredni redak;

- Bar jedan vokal mora ostati u tekućem retku;

- Novi redak ne može početi vokalom ili udvojenim konsonantom;

- Sledeći parovi slova se ne rastavljaju: *SH*, *GH*, *PH*, *CH*, *TH*, *WH*, *GR*, *PR*, *CR*, *TR*, *WR*, *BR*, *FR*, *DR*, vokal + *R*, vokal + *N*, *OM*.

Za srpskohrvatski jezik je jedna procedura zasnovana na ovoj strategiji opisana u [Vitas81]. U proceduru su ugrađena sledeća pravila za rastavljanje reči:

- Ako reč počinje konsonantskom grupom, onda ta konsonantska grupa pripada slogu oko prve centralne foneme u reči. Pri tome su centralne foneme u srpskohrvatskom jeziku, kao što je rečeno u tački 2.2.2 svi vokali i glas r u interkonsonantskom položaju, na početku

Virtual Library of Faculty of Mathematics - University of Belgrade
elibrary.matf.bg.ac.rs

reči ispred konsonanta, na kraju reči iza konsonanta kao i ispred *o* postalog od *l*.

- Ako sa **V** označimo centralne foneme a sa **C** marginalne foneme, u segmentima oblika **VV** i **VCV** granica sloga je iza prve centralne foneme: **V-V**, odnosno **V-CV**.

- Da bi se rastavila konsonantska grupa, tj. segemenati oblika VC^nV (n>1), ispituju se samo prva dva konsonanta grupe, C_1 i C_2. Odluku o rastavljanju ova dva konsonanta procedura zasniva, opciono, na jednom od dva skupa pravila koja potiču od preporuka Belića [Belić34], odnosno Stevanovića [Stevanović64]. Prema prvom skupu pravila konsonanti C_1 i C_2 se ne rastavljaju ako:

C_1 je strujni frikativ (*s*, *š*, *z*, *ž*)
ili
C_1 nije pravi sonant (*v*, *j*, *l*, *lj*, *r*) a C_2 je pravi sonant

Prema drugom skupu pravila konsonanti se rastavljaju ako:

C_1 je eksplozivni konsonant (*b*, *d*, *g*, *p*, *t*, *k*) a C_2 nije pravi sonant
ili
C_1 je sonant (*m*, *n*, *nj*, *v*, *j*, *l*, *lj*, *r*)

I jedna i druga opisana primena ove strategije zasniva se na nepromenljivim karakteristikama određenih slova (izuzetak je, donekle, samo slovo *r* u srpskohrvatskom jeziku). Za parove, ili veće grupe slova, koja imaju određene karakteristike, pravila ugrađena u ove procedure propisuju da li, i na kom mestu, se uvek rastavljaju ili se nikad ne rastavljaju. Ova odluka ne zavisi od slova koja okružuju grupu, od njenog izgovora niti od značenja reči.

Alternativni pristup se zasniva na statističkim podacima iz kojih se određuje verovatnoća da je rastavljanje reči u određenoj tački ispravno. Jedan ovakav pristup je za engleski jezik primenjen na računaru IBM 1620 i korišćen je u *Oklahoma Publishing Company* [System360]. Da bi se prikupili potrebni statistički podaci analizirano je 40.000 različitih reči i tačke u kojima se one mogu rastaviti. Zbrajana su pojavljivanja svih parova slova koji sadrže tačku rastavljanja ili su joj susedni. Korišćenjem odgovarajućih težinskih faktora za frekvenciju pojavljivanja svake reči, za svaki par slova iz alfabeta su izračunate verovatnoće da se tačka rastavljanja nalazi levo od njega, između slova ili desno od njega, i one su zapamćene u odgovarajućoj matrici.

Ovi statistički podaci su, zatim, korišćeni da bi se za svaku poziciju u reči procenila verovatnoća da je to tačka u kojoj se reč može rastaviti. Za parove slova koji prethode, okružuju i slede određenu poziciju u reči iz matrice verovatnoća uzimane su verovatnoće da je tačka rastavljanja desno, između, odnosno levo od njih. Ove verovatnoće su, zatim, pomnožene i tako se dobijala aproksimacija ukupne verovatnoće da se na toj poziciji u reči reč može rastaviti. Da bi se reč rastavila, ovaj proces se ponavljao za svaku poziciju, a kao tačka rastavljanja birala se pozicija čija je ukupna verovatnoća najveća.

Virtual Library of Faculty of Mathematics - University of Belgrade
elibrary.matf.bg.ac.rs

2.3.3 Korišćenje pravila i rečnika izuzetaka

Procedure koje se zasnivaju na strategiji pravila obično ne rastavljaju sve reči ispravno, već je procenat ispravno rastavljenih reči obično između 85 i 95%. Testiranje ovih procedura pokazuje da se bolji rezultat, korišćenjem isključivo strategije pravila, ne može postići. S jedne strane, u svakom prirodnom jeziku postoje reči koje odstupaju od uobičajenih obrazaca za rastavljanje (reči stranog porekla, i sl.). S druge strane, semantička pravila za rastavljanje reči se ne mogu izraziti preko ovih "mehaničkih" pravila koja se zasnivaju isključivo na karakteristikama pojedinih slova, vodeći pri tom računa, eventualno, još samo o njihovoj poziciji u reči.

Da bi se realizovao automatski rastavljač koji ispravno rastavlja sve reči nekog jezika, često se koristi strategija koja predstavlja mešavinu strategije rečnika i strategije pravila. U rečnik se smeštaju samo reči koje se korišćenjem pravila ne rastavljaju ispravno, sa naznakom kako ih treba rastaviti. Kako u rečniku sada ne moraju da se nalaze sve reči nekog jezika, dimenzija rečnika se značajno smanjuje. Takav rečnik se, obično, naziva *rečnik izuzetaka*.

Osim rečnika izuzetaka, koji sadrži reči, ili osnove reči iz kojih se može generisati njihova paradigma, često se koriste i *rečnici niski-izuzetaka*. Ovi rečnici obično sadrže one prefiksne i sufiksne niske[1] čije prepoznavanje na početku, odnosno, kraju reči ima posebno značenje. Često se iza, odnosno, ispred prepoznate afiksne niske nalazi tačka u kojoj se reč može rastaviti a prepoznavanje ovih niski može i da ukaže na neophodnost primene neke specifične grupe pravila. Važno je napomenuti da, iako ovi rečnici obično sadrže prave prefikse i sufikse, u njima se ponekad nalaze i niske koje nisu afiksi u lingvističkom smislu ali su zajedničke nekoj većoj grupi reči koje se sve rastavljaju po sličnom obrascu.

Za engleski jezik, jedna jednostavna procedura koja koristi ovu strategiju je opisana u [Gimpel76]. Procedura je napisana na programskom jeziku SNOBOL4 i primenjena je u jednostavnom formateru koji je, takođe, napisan u SNOBOL-u da bi se na njemu u knjizi ilustrovali određeni algoritmi. Sama knjiga je pripremljena za štampu korišćenjem istog formatera.

U proceduru su ugrađena jednostavna pravila. Definisani su svi parovi slova (digrami) koji se smeju rastaviti. Osim ovog pravila koristi se samo još pravilo da bar jedan vokal mora ostati u tekućem retku i da se bar jedan vokal mora preneti u sledeći redak. Osim pravila, procedura koristi i rečnik sufiksa. Ovi sufiksi su u proceduru ugrađeni, u skladu sa programskim jezikom SNOBOL4, u obliku obrasca. Svi ugrađeni sufiksi su grupisani u tri kategorije. Prvu kategoriju čine *sufiksi koji se mogu rastaviti.* Ispred ovih sufiksa reč se može rastaviti. Primer su sufiksi *-ness* i *-less*. Drugu kategoriju čine *sufiksi koji se ne smeju rastaviti.* Kada se prepozna sufiks iz ove kategorije, sufiks se odbacuje i prelazi se na ispitivanje digrama. Oni apsolutno zabranjuju rastavljanje ispred sufiksa, čak iako tabela digrama ukazuje da se digram sme rastaviti. Primer je sufiks *-ing*. Poslednju kategoriju čine *neutralni sufiksi.* Ovi sufiksi se niti

[1] Termini prefiksna i sufiksna niska označavaju nisku sa kojom počinje, odnosno, završava reč. Prefiksi i sufiksi su lingvističke kategorije.

Virtual Library of Faculty of Mathematics - University of Belgrade
elibrary.matf.bg.ac.rs

rastavljaju niti lansiraju testiranje digrama. Oni ukazuju da je moguća superpozicija sufiksa, pa se prepoznati sufiks odbacije i ispituje se dalje prisustvo sufiksa.

Dalje poboljšanje ovog algoritma je opisano u [Moitra79]. Osim tabele digrama i rečnika sufiksa, grupisanih u opisane tri kategorije, ova procedura koristi i rečnik prefiksa i rečnik izuzetaka. Redosled pretraživanja je sledeći:

rečnik izuzetaka;
rečnik sufiksa;
rečnik prefiksa;
tabela digrama.

Po samoj prirodi rečnika izuzetaka, on se prvi pretražuje. Rečnik sufiksa se pretražuje pre rečnika sufiksa jer je postojanje sufiksa u engleskom jeziku češće od postojanja prefiksa.

Međutim, primena ovog algoritma pokazuje da se jedan broj reči i dalje rastavlja neispravno. Jedan broj problema nastaje kada se prepozna prefiksna, odnosno sufiksna, niska u rečima u kojima ona nije prefiks, odnosno, sufiks. Primer je prefiks *pro-* koji je veoma čest u engleskom jeziku ali nije prefiks u reči *problem* koja se rastavlja *prob-lem*. Slično je i sa prefiksom *co-* u rečima *court* i *count*.

Rečnici prefiksa i sufiksa se pretražuju tako da se uvek pronalazi najduža niska. Tako su, na primer, sufiksi i *-ate* i *-rate*. Kada procedura sravni sufiks *-ate* ona se ne zaustavlja i pokušava da sravni sufiks *-rate*. Međutim, postoje reči koje se završavaju niskom *-rate* ali imaju sufiks *-ate* ispred koga se reč rastavlja. Na primer, u reči *incorporate* je sufiks - *rate* dok je u reči *generate* sufiks *-ate*.

Svi ovi, i niz drugih problema, mogu se rešiti proširivanjem rečnika izuzetaka. Tako se, na primer, u slučaju sufiksa *-ate* i *-rate* u rečnik sufiksa može staviti samo jedan od njih dok se sve reči u kojima je sufiks onaj drugi smeštaju u rečnik izuzetaka. Realna je opasnost, međutim, da rečnik izuzetaka postane prevelik i neefikasan za pretraživanje.

U [Ocker75] je opisana još jedna procedura koja se zasniva na rečnicima i tabeli digrama. Tabela digrama je ovde znatno složenija. Svakom paru slova je, umesto binarne vrednost koja govori da li se digram može, ili ne može, rastaviti, pridružen težinski faktor koji određuje njegovu izgovorljivost. Reč se rastavlja na poziciji koja ima najveći težinski faktor. Tako je, na primer, digrafu *ou* koji se ne rastavlja pridružen težinski faktor 0 dok je konsonatskoj grupi *tb* koja je teška za izgovor pridružen najveći težinski faktor 128. Ovi težinski faktori regulišu i prioritete primene pravila. Na primer, udvojeni konsonant ima težinski faktor 33 dok konsonantska grupa *sb* teška za izgovor ima težinski faktor 128. Stoga će u reči *passbook*, pravilo da se konsonantska grupa teška za izgovor, koja je često, kao i u ovom slučaju, posledica složene reči, poništiti pravilo da se udvojeni konsonanti rastavljaju.

Rečnici prefiksa i sufiksa takođe sadrže težinske faktore koji, kada se u reči prepozna prefiksna, odnosno sufiksna niska, mogu da promene težinske faktore zasnovane na tabeli digrafa. Svakom afiksu

Virtual Library of Faculty of Mathematics - University of Belgrade
elibrary.matf.bg.ac.rs

iz rečnika pridružena je i lista izuzetaka tog afiksa.

U engleskom jeziku je i naglašenost slogova u nekim slučajevima od značaja za rastavljanje (videti tačku **2.2.1**). U ovoj proceduri je i naglašenost sloga postala element za odlučivanje. U proceduru je ugrađen *rečnik akcenatskih obrazaca*, u kome su dati uobičajeni obrasci naglašavanja engleskih reči u zavisnosti od broja slogova. Na osnovu broja slogova u reči i poznate naglašenosti nekih slogova - na primer, uz svaki afiks je u rečniku pridružena informacija o tome da li je naglašen - iz rečnika akcenatskih obrazaca se određuje najverovatnija naglašenost ostalih slogova.

Za srpskohrvatski jezik je predložena procedura koja se zasniva na pravilima i rečnicima izuzetaka, prefiksa i sufiksa [Krstev85]. O pravilima za rastavljanje reči srpskohrvatskog jezika i rečnicima izuzetaka biće govora u glavama **3** i **4** ovog rada.

2.3.4 Korišćenje rečnika obrazaca

U programski paket TeX je ugrađena procedura za rastavljanje reči na kraju retka koja se zasniva na drukčijoj strategiji, koju je u periodu od 1980.-82 godine razvio F.M.Liang [Liang83]. Ova strategija se zasniva na *obrascima*, koji predstavljaju delove reči nekog jezika i snabdeveni su informacijama na kojim mestima je dozvoljeno a na kojim zabranjeno reč, čija podniska je obrazac, rastaviti. Funkcija ovih obrazaca i funkcionisanje odgovarajuće procedure se najlakše mogu objasniti na primeru. Uzmimo reč iz engleskog jezika *hyphenation*. Ovu reč prvo treba proširiti sa obe strane markerima koji označavaju kraj reči:

+hyphenation+

Dužina ove reči je, ako računamo i njene krajnje tačke, 15. Izvadimo sve podniske ove reči koje su dužine od 1 do 15:

+ h y p h e n a t i o n +

su sve podniske dužine 1,

+h hy yp ph he en na at ti io on n+

su sve podniske dužine 2,

+hy hyp yph phe hen ena nat ati tio ion on+

su sve podniske dužine 3, itd. U jednoj podniski dužine k postoji $k+1$ pozicija između slova na koju se postaviti koeficijent od 0 do 9 koji govori da li je poželjno reč rastaviti na tom mestu ili ne. Jedna takva sekvencija slova i cifara je *obrazac*. Tako su obrasci, na primer, *o1h0e2n0* i *1n0a0* odnosno *he2n* i *1na*, jer se podrazumeva da je koeficijent, ako ne postoji, jednak 0. U stvari od značaja su samo oni obrasci u kojima je bar jedan koeficijent različit od 0 i samo takvi obrasci se smeštaju u rečnik.

Procedura za rastavljanje reči ugrađena u TeX traži u *rečniku obarazaca* sve podniske reči koju treba rastaviti. Tako se za reč *hyphenation* u rečniku obrazaca engleskog jezika pronalaze sledeći obrasci:

Virtual Library of Faculty of Mathematics - University of Belgrade
elibrary.matf.bg.ac.rs

hy3ph	*he2n*	*hena4*	*hen5at*
1*na*	*n2at*	1*tio*	2*io*

Koeficijenti koji se nalaze u ovim obrascima se smeštaju na odgovarajuće pozicije između slova polazne reči. Kako više koeficijenata, koji dolaze iz različitih obrazaca, treba smestiti na istu poziciju u reči, bira se najveći od njih. U našem slučaju, između slova *e* i *n*, prema obrascu 1*na* treba smestiti koeficijent 1 a prema obrascu *he2n* koeficijent 2. Bira se veći koeficijent, i to je 2. Rezultat ovakvog procesa je:

+hy3phe2n5a4t2ion+

Neparni brojevi označavaju pozicije na kojima je rastavljanje dozvoljeno, dok parni brojevi, uključujući i 0 koja je izostavljena, označavaju pozicije na kojima je rastavljanje reči zabranjeno. Prema tome, u reči *hyphenation* postoje dve pozicije na kojima se reč može rastaviti i to su: *hy-phen-ation*.

Glavna ideja ove procedure je sledeća: Obrasci koji sadrže koeficijent 1 označavaju pozicije na kojima je, u opštem slučaju, dozvoljeno rastaviti reč. U onim slučajevima kada na toj poziciji rastavljanje ipak nije dozvoljeno, drugi obrasci koji ga delimično prekrivaju poništavaju koeficijent 1 pomoću bar jednog parnog koeficijenta. U našem slučaju, obrazac 1*na* govori da se reč često može rastaviti ispred *n* iza koga sledi *a* dok obrazac *he2n* govori da nikada nije dozvoljeno rastaviti reč između *he* i *n*.

Rečnik obrazaca koji je ugrađen u TeX je konstruisao Liang pomoću svog programa PATGEN, koji je automatski, iz postojećeg rečnika (*Webster's Pocket Dictionary*) u kome su bile označene pozicije na kojima se reč može rastaviti, generisao rečnik obrazaca za engleski jezik. Ovaj rečnik sadrži 4447 obrazaca i obezbeđuje da se pomoću njega u engleskom tekstu pronađe približno 90% dozvoljenih pozicija a da se ne načini ni jedna greška.

Kao što je rečeno u tački **2.3.1**, za većinu jezika ne postoje rečnici u kojima su označene pozicije na kojima se reč može rastaviti, što, pak, znači da se ne može primeniti program PATGEN za generisanje obrazaca. Za mnoge jezike, ni postojanjem ovakvog rečnika problem se ne bi mogao rešiti, jer se u rečniku moraju naći svi oblici reči. Da bi generisao obrasce za engleski jezik, Liang je, kada je dodao sve oblike reči, proširio postojeći rečnik od 31.036 reči na 49.858 reči. Za flektivne jezike, ovo proširenje bi bilo mnogo veće, pa bi rečnik mogao da postane tri, i više, puta veći.

Za francuski jezik, takođe, ne postoje rečnici u kojima su označene pozicije na kojima se reč može rastaviti pa je, stoga, odabran drugačiji pristup generisanju obrazaca: obrasci su generisani direktno iz pravila za rastavljanje reči francuskog jezika [Desarm84e]. Na taj način je formiran rečnik od 775 obrazaca. Ovakav način, ne samo da je za jezike bez odgovarajućeg rečnika efikasniji već, takođe, čini jasnom funkciju svakog obrasca u rečniku i olakšava promenu rečnika obrazaca u slučaju izmene ili dodavanja nekog pravila. O formiranju rečnika obrazaca za srpskohrvatski jezik biće reči u glavi **5** ovog rada.

Virtual Library of Faculty of Mathematics - University of Belgrade
elibrary.matf.bg.ac.rs

Virtual Library of Faculty of Mathematics - University of Belgrade
elibrary.matf.bg.ac.rs

3 PRAVILA ZA RASTAVLJANJE REČI SRPSKOHRVATSKOG JEZIKA

Kao što je u glavi 2 već rečeno, za srpskohrvatski jezik je formulisano više skupova pravila za rastavljanje reči na kraju retka, ali ni jedan od ovih skupova pravila nije dobio punu legitimnost. Preporuke koje su date u Pravopisu srpskohrvatskog književnog jezika su autorima jasne i pružaju im dosta slobode, ali su sa stanovišta konstruisanja procedure za automatsko rastavljanje reči na kraju retka nedovoljno precizne. Problemi koji se moraju rešiti pre nego što se pristupi izradi ovakve procedure su sledeći:

1 *Problem transliteracije.* Ovaj problem se, u osnovi, svodi na precizno razgraničavanje pojavljivanja niski *lj*, *nj*, *dj* i *dž* u ulozi digrafa, odnosno, konsonantskih grupa. Ovaj problem se postavlja samo onda kada se koristi latinično srpskohrvatsko pismo i kada se za unošenje i zapis teksta koriste takve tastature, odnosno, kodne tablice u kojima za srpskohrvatske digrafe ne postoje odvojeni tasteri, odnosno, kodovi [Krstev88]. Kako je ovo dovoljno čest slučaj, rešenje ovog problema zahteva pažljivu brigu.

2 *Klasifikacija konsonantskih grupa.* Iz diskusije izložene u glavi 2 (tačke 2.2.2 i 2.3.2) proizlazi da su glavne razlike između predloženih skupova pravila za rastavljanje reči srpskohrvatskog jezika u pristupu rastavljanju konsonantskih grupa. Tako se u preporukama Pravopisa srpskohrvatskog književnog jezika uvode intuitivni pojmovi "lako" i "teško" izgovorljivih konsonantskih grupa. Da bi se ovi intuitivni pojmovi jasnije precizirali, potrebno je, pre svega, formirati rečnik konsonantskih grupa koje se realizuju u srpskohrvatskom jeziku a zatim izvesti precizna pravila za njihovo rastavljanje.

3 *Definisanje semantičkih pravila.* Da bi se precizirala semantička pravila za rastavljanje reči na kraju retka, u smislu odgovarajućih preporuka iz Pravopisa, potrebno je, pre svega, ustanoviti koje su to reči, odnosno morfemi, koje učestvuju u stvaranju složenih reči. Prvi korak je utvrđivanje liste prefiksa i sufiksa koji se u srpskohrvatskom jeziku koriste za gradnju novih reči i analiza njihovog korišćenja. U drugom koraku, treba izvršiti analizu složenica u srpskohrvatskom jeziku koja bi omogućila izradu rečnika reči koje učestvuju u gradnji složenica.

Da bi se rešili prvi i drugi navedeni problem, nad korpusom je izrađen rečnik konsonantskih grupa, dok je za delimično rešenje trećeg probleme izvršena analiza pojavljivanja prefiksa.

3.1 Rečnik konsonantskih grupa srpskohrvatskog jezika

Konsonantske grupe u srpskohrvatskom jeziku su u više navrata bile predmet teorijskih istraživanja, najčešće kao deo komparativnih proučavanja konsonantskih grupa u slovenskim jezicima. Na osnovu više rečnika i Pravopisa srpskohrvatskog jezika, sačinjen je pregled inicijalnih i finalnih sistema konsonantskih grupa u srpskohrvatskom jeziku, proučena je poveznost ovih sistema kao i fonološke karakteristike konsonantskih grupa [Tolstoja68a, Tolstoja68b,

Tolstoja72]. S obzirom da za ova istraživanja nije korišćen računar, izostao je pregled medijalnog sistema konsonantskih grupa. Kako ova istraživanja nisu vršena nad vezanim tekstom, date su samo pretpostavke o upotrebi pojedinih konsonantskih grupa u pisanom tekstu i govoru.

Da bi se sačinili rečnici konsonantskih grupa koje se pojavljuju u inicijalnom, medijalnom i finalnom položaju u reči i utvrdila frekvencija njihovog pojavljivanja u vezanom tekstu korišćen je programski sistem **AURORA** za automatsko generisanje konkordanci i ostalih vrsta indeksa [Vitas79, Vitas82]. Za izradu rečnika je korišćen deo postojećeg korpusa veličine oko 100.000 reči, koji se sastoji od tekstova navedenih u Prilogu **A.1**. Tekstovi ovog korpusa su kodirani tako da, osim alfabetskih karaktera, sadrže interpunkcijske i specijalne karaktere kao i neka obeležja strukture teksta. Posebnu pogodnost, sa stanovišta ovog istraživanja, predstavljala je činjenica da su u tekstovima ovog korpusa digrafi *lj*, *nj*, *dj* i *dž* kodirani posebnim dvoslovnim oznakama *LX*, *NX*, *DX*, odnosno *DY* čime je izbegnuta nedoumica oko razlikovanja digrafa od odogovarajuće konsonantske grupe.

Za potrebe izrade rečnika konsonantskih grupa, izmenjen je samo onaj deo programskog sistema **AURORA** kojim se definiše *reč*, odnosno, koji iz teksta izdvaja *reč* koja se smešta u rečnik. Za potrebe ovog istraživanja *reč* je definisana kao svako pojavljivanje u tekstu niske oblika C^n gde je C konsonant i $n>=2$ koja je podniska niske oblika $V_1C^nV_2$, gde su V_1 i V_2 iz skupa vokala, blanko karakter ili neki drugi karakter kojim se reč (u uobičajenom značenju) omeđava. Pri tome su kao različite reči tretirane iste niske C^n u zavisnosti od okruženja V_1 i V_2. Na primer, u korpusu su identifikovana četiri razna pojavljivanja niske *RT*:

- u reči *RT*ANj, V_1 je blanko karakter a V_2 je vokal;
- u reči FLE*RT*, V_1 je vokal a V_2 je blanko karakter;
- u reči HA*RT*IJA, i V_1 i V_2 su vokali;
- u reči *RT*, i V_1 i V_2 su blanko karakteri.

Na taj način su identifikovana i razdvojena pojavljivanja konsonantskih grupa u inicijalnom, finalnom i medijalnom položaju što je omogućilo automatsku izradu odgovarajućih rečnika.

Kao što se vidi iz definicije reči koja je za ovu priliku ugrađena u sistem **AURORA**, nije identifikovan glas *r* u ulozi vokala iako se i to moglo automatski učiniti. Razlozi za to su sledeći: Na ovaj način je bilo moguće izvršiti i dodatnu analizu pojavljivanja glasa *r* u ulozi vokala, kao i analizu fonoloških karakteristika konsonanata koji ga okružuju.

Sistemom **AURORA** su, zatim, nad izabranim korpusom i sa ovako definisanom reči izrađeni sledeći rečnici:

- Konkordance u ćiriličnom poretku;
- Rečnik konsonantskih grupa prema njihovoj poziciji u reči;
- Frekvencijski rečnik konsonantskih grupa;
- Rečnik konsonantskih grupa prema njihovoj dužini.

U konkordancama je za svako pojavljivanje konsonatske grupe u tekstu izlistan kontekst u kome se ona pojavila od tri reči ulevo i udesno od reči koja sadrži konsonantsku grupu. Razlozi sa izradu konkordanci konsonantskih grupa bili su sledeći:

Virtual Library of Faculty of Mathematics - University of Belgrade
elibrary.matf.bg.ac.rs

- analiza konteksta u kome se pojavila neka konsonantska grupa omogućila je otklanjanje grešaka u samom korpusu, koje su nastale kao posledica grešaka u kucanju;

- kontekst je, osim toga, omogućio da se u kasnijim istraživanjima izvrši analiza strukture konsonantskih grupa (konsonantske grupe kao posledica spoja prefiksa, ili sufiksa, sa osnovom, i sl.).

Posle ispravke svih grešaka identifikovanih analizom konkordanci i ručnog ažuriranja svih frekvencija sačinjene su dve grupe frekvencijskih rečnika: u prvoj grupi nije vršena identifikacija glasa *r* u ulozi vokala (videti [Krstev82]), dok je u drugoj grupi, koja je nastala ručnom doradom prve grupe rečnika identifikovan glas *r* u ulozi vokala i izvršena su odgovarajuća ažuriranja frekvencija (videti Prilog **B.1** ovog rada).

Analiza dobijenih rezultata pokazala je sledeće: Identifikovano je ukupno 381 različitih konsonantskih grupa koje ne sadrše *r* u ulozi vokala i one su se pojavile u korpusu 49.495 puta. Različitih konsonantskih grupa koje sadrže *r* u ulozi vokala je identifikovano ukupno 519 i one su se pojavile 51.178 puta.

Konstatovano je da se u tekstu od oko 100.000 reči pojavio relativno mali broj različitih konsonantskih grupa. Analiza većeg korpusa verovatno ne bi značajnije uticala na ovaj zaključak jer broj identifikovanih konsonantskih grupa ne raste linearno sa dužinom teksta. Tako je već posle polovine analiziranog korpusa identifkovano 441 različitih konsonantskih grupa, tj. 85% od ukupnog broja.

Konstatovano je, takođe, da se mnoge od konsonantskih grupa koje su navedene u teoretskim radovima ne pojavljuju u vezanom tekstu. S druge strane, identifikovano je samo nekoliko konsonantskih grupa koje se u teoretskim radovima ne pominju i one su se, najčešće, pojavile u rečima stranog porekla. Na primer, finalna grupa *-rks* (u reči *Marks*) i inicijalne grupe *dm* (u imenu *Dmitar*), *dj* (u reči *djed*) i *dnj* (u reči *Dnjepar*).

Grupe "lake" za izgovor		Grupe "teške" za izgovor	
Primer	frekvencija	Primer	frekvencija
sedlo	134	zamka	7
vježba	53	borba	174
mišji	0	bumbar	24
sestra	232	sunce	321
zemlja	975	bratstvo	72
slavlje	368	školski	20
živjeti	0	sredstvo	102
		sudski	93

Tabela 3.1 Prikaz primera iz Pravopisa "lako" i "teško" izgovorljivih konsonantskih grupa sa frekvencijom njihovog pojavljivanja u korpusu pisanih tekstova

Virtual Library of Faculty of Mathematics - University of Belgrade
elibrary.matf.bg.ac.rs

Posebno značajne rezultate dala je analiza pojavljivanja konsonantskih grupa *lj*, *nj*, *dj* i *dž*. U korpusu je identifikovana samo konsonantska grupa *dj* i to u inicijalnoj poziciji jedanput u reči *djed* a u medijalnoj poziciji ukupno 17 puta u rečima *odjek*, *odjednom*, *odjedanput*, *podjednako* i *ovdje*.

U Pravopisu je navedeno samo nekoliko primera "lako" i "teško" izgovorljivih konsonantskih grupa. Analiza korpusa je pokazala da se ne može uspostaviti veza između "težine" grupe i frekvencije njenog pojavljivanja (videti Tabelu **3.1**). Iz toga sledi da se problem identifikacije "teško" izgovorljive konsonantske grupe ne može rešiti bez dodatnog lingvističkog istraživanja koje bi se zasnivalo na ovoj analizi korpusa (videti glavu **4** ovog rada).

				Sonanti		
	Eksploz.	Frikativi	Afrikate	Pravi	Nazalni	Ukupno
B P G K D T	*20* 974	*27* 1115	*8* 102	*26* 12422	*21* 3770	*102* 18383
Ž Š Z S	*33* 8958	*7* 100	*7* 289	*16* 5149	*9* 3012	*72* 17508
F H	*2* 52	*1* 103	*0*	*6* 422	*4* 28	*13* 605
Č Ć C D Dz	*4* 949	*2* 13	*1* 2	*7* 249	*8* 463	*22* 1667
R L V J Lj	*37* 1575	*41* 1250	*12* 412	*13* 1902	*11* 2149	*112* 7288
Nj M N	*22* 913	*16* 1044	*8* 435	*7* 1208	*5* 444	*58* 4044
UK UP NO	*118* 13421	*94* 3625	*36* 1240	*75* 21343	*58* 9866	*381* 49495

Legenda: Linija ——— i Belić i Stevanović rastavljaju
Linija ····· samo Belić rastavlja
Linija ——— samo Stevanović rastavlja

Tabela 3.2 Pregled broja i frekvencije pojavljivanja prva dva konsonanta u konsonantskim grupama na korpusu pisanih tekstova prema njihovim fonetskim osobinama

Kako se sistemi pravila za rastavljanje konsonantskih grupa koji potiču od Belića i Stevanovića zasnivaju na fonetskim osobinama prva dva konsonanta u grupi, sačinjena je tabela koja treba da ukaže na

Virtual Library of Faculty of Mathematics - University of Belgrade
elibrary.matf.bg.ac.rs

stvarne razlike između ova dva sistema, imajući u vidu pojavljivanje konsonantskih grupa u vezanom tekstu. U Tabeli 3.2 je dat pregled broja različitih konsonantskih grupa, kao i frekvencija pojavljivanja tih grupa, prema fonetskim osobinama prva dva konsonanta grupe. Konsonanti su prema fonetskim osobinama svrstani u sledeće kategorije:

Esplozivni (ili praskavi):	*B, P, G, K, D, T*
Frikativi (ili strujni):	*Ž, Š, Z, S, F, H*
Afrikate (ili sliveni):	*C, Ć, Dž, Č, Đ*
Pravi sonanti (ili glasnici):	*R, L, Lj, V, J*
Nosni sonanti (ili glasnici):	*M, N, Nj*

Iz Tabele 3.2 se vidi da konsonantske grupe koje se ni po Belićevom ni po Stevanovićevom sistemu pravila ne rastavljaju, te bi se mogle svrstati u grupe "lake" za izgovor, čine 60% od ukupnog broja pojavljivanja konsonantskih grupa, odnosno 22% od ukupnog broja različitih grupa. S druge strane, konsonantske grupe koje se rastavljaju i po jednom i po drugom sistemu pravila, te bi se mogle svrstati u grupe "teške" za izgovor, čine 32% od ukupnog broja pojavljivanja konsonantskih grupa, odnosno 63% od ukupnog broja različitih grupa. Zaključak bi mogao biti da preostaje samo da se u dodatnoj lingvističkoj analizi 17% identifikovanih konsonantskih grupa razvrsta u "lake", odnosno "teške", za izgovor. Međutim, osim identifikacije prefiksa, sufiksa i složenica, preostaje, takođe, da se preispita adekvatnost strategije rastavljanja konsonantskih grupa uvek iza prvog konsonanta.

U cilju provere rezultata dobijenih analizom korpusa pisanih tekstova, cela analiza je ponovljena na korpusu govora učenika IV razreda osnovne škole [Cvijov84]. Ovaj korpus od približno 100.000 reči formiran je sa ciljem da se opišu leksičke, sintaksičke i semantičke karakteristike dečjeg govora. Analiza pojavljivanja konsonantskih grupa nad ovim korpusom je bila zasnovana na istim principima i izvršena je korišćenjem iste programske podrške kao i opisana analiza korpusa pisanih tekstova.

Analiza dobijenih rezultata pokazala je sledeće: Identifikovano je ukupno 321 različitih konsonantskih grupa koje ne sadrže *r* u ulozi vokala i one su se pojavile u korpusu 34515 puta. Različitih konsonantskih grupa koje sadrže *r* u ulozi vokala je identifikovano ukupno 420 i one su se pojavile 35715 puta. U Prilogu B.2 ovog rada dati su frekvencijski rečnici konsonantskih grupa koje ne sadrže *r* u ulozi vokala prema poziciji konsonantske grupe u reči izrađeni na korpusu dečjeg govora.

Analiza korpusa dečjeg govora potvrdila je rezultate analize korpusa pisanih tekstova u odnosu na pojavljivanje konsonantskih grupa *lj, nj, dj* i *dž*. U korpusu su identifikovane konsonantske grupa *dj* i *nj* i to grupa *dj* samo u medijalnoj poziciji ukupno 21 put u rečima *odjednom* i *odjedanput* a grupa *nj*, takođe samo u medijalnoj poziciji, ukupno 12 u reči *injekcija*. Potvrđen je rezultat analize korpusa pisanih tekstova da se ove konsonantske grupe u ekavskoj varijanti srpskohrvatskog jezika uvek realizuju na spoju prefikasa sa osnovom.

Virtual Library of Faculty of Mathematics - University of Belgrade
elibrary.matf.bg.ac.rs

Pozicija	Pisani tekstovi		Dečji govor	
	Ukupno	Eks.	Ukupno	Eks.
Medijalna	350	90 (25%)	304	44 (14%)
Inicijalna	95	20 (21%)	87	12 (14%)
Finalna	24	12 (50%)	21	9 (43%)

Tabela 3.3 Pregled različitih konsonantskih grupa prema poziciji u reči u korpusu pisanih tekstova i korpusu dečjeg govora

Upoređivanjem rezultata analize korpusa pisanih tekstova i dečjeg govora, konstatovane su razlike u pojavljivanju konsonantskih grupa, koje su predstavljene u koloni **da/ne** Tabele **3.3**, u kojoj je dat broj konsonantskih grupa koje se pojavljuju samo u jednom od korpusa.

				Sonanti		
	Eksploz.	Frikativi	Afrikate	Pravi	Nazalni	Ukupno
B P G K D T	*19* 900	*23* 572	*6* 39	*27* 9409	*15* 1967	*90* 12887
Ž Š Z S	*29* 6892	*4* 11	*5* 88	*19* 4447	*12* 3341	*69* 14779
F H	*2* 133	*1* 1	*0*	*5* 226	*1* 11	*9* 371
Č Ć C D Dz	*5* 313	*0*	*0*	*5* 63	*7* 202	*17* 578
R L V J Lj	*30* 1220	*21* 219	*10* 369	*12* 865	*13* 844	*86* 3517
Nj M N	*21* 1319	*13* 143	*7* 264	*8* 338	*1* 319	*50* 2383
UK UP NO	*106* 10777	*62* 946	*28* 760	*76* 15348	*49* 6684	*321* 34515

Legenda: Linija ——— i Belić i Stevanović rastavljaju
Linija ..... samo Belić rastavlja
Linija ——— samo Stevanović rastavlja

Tabela 3.4 Pregled broja i frekvencije pojavljivanja prva dva konsonanta u konsonantskim grupama na korpusu dečjeg govora prema njihovim fonetskim osobinama

Virtual Library of Faculty of Mathematics - University of Belgrade
elibrary.matf.bg.ac.rs

Tabela **3.4** pokazuje izvesno odstupanje od rezultata dobijenih analizom korpusa pisanih tekstova. Konsonantske grupe koje se ne rastavljaju ni po Belićevom ni po Stevanovićevom sistemu pravila čine 71% od ukupnog broja pojavljivanja konsonantskih grupa, odnosno 33% od ukupnog broja različitih grupa, dok konsonantske grupe koje se rastavljaju i po jednom i po drugom sistemu pravila čine 26% od ukupnog broja pojavljivanja konsonantskih grupa, odnosno 59,5% od ukupnog broja različitih grupa. U korpusu dečjeg govora je, dakle, od ukupnog broja različitih konsonantskih grupa samo 7,5% onih koje se prema ova dva sistema pravila različito rastavljaju.

	B	*P*	*G*	*K*	*D*	*T*	*Ž*	*Š*	*Z*	*S*	*F*	*H*
B					xo				x			
P				xo		xo		xo		xo		x
G					xo				x			
K	x					xo		xo		xo		x
D	xo		xo	xo				o	xo	xo		
T		xo		xo	o				x	xo		xo
Ž	xo				xo							
Š		xo		xo		xo						
Z	xo		xo		xo	o						
S		xo		xo		xo					xo	xo
F						o				xo		
H						xo						
C				xo								
Č				xo		o						
Dž	xo											
Ć				xo						x		
D										x		
R	xo	xo	xo	xo	xo	xo	x	xo	xo	xo	x	xo
L	xo	x	o	xo	xo	xo		x	xo	xo	xo	
Lj				xo	xo			x		xo		
V	x		xo	xo	xo	xo		x		xo		
J	xo	xo	xo	xo	xo	xo		xo	xo	xo	x	o
M	xo	xo		xo	o	xo		o	xo	xo	x	
N	xo	xo	xo	xo	xo	xo	x	xo	xo	xo	o	o
Nj										xo		

Legenda: x - korpus pisanih tekstova
o - korpus dečjeg govora

Tabela 3.5 Pregled pojavljivanja parova konsonanata na početku konsonantske grupe u korpusu pisanih tekstova i korpusu dečjeg govora

Virtual Library of Faculty of Mathematics - University of Belgrade
elibrary.matf.bg.ac.rs

Iz zahteva, koji je sadržan i u Belićevom i u Stevanovićevom sistemu pravila, da se konsonantske grupe koje se rastavljaju, rastavljaju uvek iza prvog konsonanta, kao i činjenice da konsonantske grupe od dva konsonanta preovlađuju, proističe potreba za analizom početaka konsonantskih grupa u vezanom tekstu. U Tabeli 3.5 dat je pregled pojavljivanja parova konsonanata na početku konsonantske grupe, nezavisno od njenog položaja u reči, u korpusu pisanih tekstova i dečjeg govora. Od 625 mogućih parova konsonanata, u korpusu pisanih tekstova se na početku grupe pojavilo 258 a u korpusu dečjeg govora 237 parova konsonanata. Dakle, u bar jednom korpusu se pojavljuje 288 parova konsonanata, dok se 337 parova ne pojavljuje na početku ni jedne konsonantske grupe.

	C	Č	Dž	Ć	Đ	R	L	Lj	V	J	M	N	Nj
B						xo	xo	xo		xo	x	xo	x
P	xo	xo		xo		xo	xo	xo		x		xo	xo
G						xo	xo	xo	xo		xo	xo	xo
K	xo	xo		xo		xo	xo	xo	xo	o	xo	xo	xo
D						xo	xo	xo	xo	xo	xo	xo	xo
T	x	x				xo	xo	xo	xo	xo	xo	xo	xo
Ž					xo	x	x	xo	o	xo	o	xo	xo
Š	xo	xo		xo		o	xo	xo	x		xo	xo	xo
Z						xo	xo	xo	xo	xo	xo	xo	
S	xo					xo	xo		xo	xo	xo	xo	
F						xo	xo			x		x	
H						xo	o		xo		x	xo	
C						o			xo		xo	xo	o
Ć							xo	x	xo	xo	xo	xo	x
Dž													
Đ	x											xo	xo
R	xo	xo	xo		xo		xo	xo	xo	xo	xo	xo	xo
L	xo	o	x						x		xo	xo	o
Lj	x											xo	
V	xo	xo		x	x	xo	xo	xo		xo		xo	x
J	xo	xo				xo	xo		xo	xo	xo	xo	
M	xo	xo		x		xo	xo	xo	xo	o		xo	x
N	xo	xo	xo	x	xo		x	xo	xo	o			
Nj	o												

Tabela 3.5 - Nastavak

3.2 Analiza pojavljivanja prefiksa u srpskohrvatskom jeziku

Kao prvi korak u definisanju semantičkih pravila za rastavljanje reči na kraju retka nameće se analiza korišćenja prefiksa u srpskohrvatskom jeziku u vezanom tekstu.

Virtual Library of Faculty of Mathematics - University of Belgrade
elibrary.matf.bg.ac.rs

3.2.1 Prefiksi u srpskohrvatskom jeziku i ciljevi analize njihovog pojavljivanja u vezanom tekstu

Prefiksi spadaju u grupu *afiksalnih morfema* (ili *afiksa*) i kao takvi se ne pojavljuju u svakoj reči. Oni predstavljaju, za razliku od *korenskih* (ili *leksičkih*) *morfema* koje sadrže osnovno (leksičko) značenje reči, značajna tvorbena sredstva pomoću kojih se izražava tvorbeno značenje reči. Stoga se oni nazivaju i *funkcionalnim morfemima*. Tvorbeno značenje reči predstavlja modifikovano leksičko značenje osnovne reči, pri čemu se prefiks pojavljuje kao modifikator. Karakteristika prefiksa je da oni ne menjaju *gramatička svojstva* osnovne reči već samo deluju na leksičko značenje reči: na primer, *brzo - prebrzo*, *čitati - pročitati*, *brat - nebrat*, i sl.

Prefiksi učestvuju u tvorbi različitih vrsta reči: imenica, glagova, prideva i priloga. Neki od prefiksa učestvuju u tvorbi više vrsta reči. Na primer, prefiks *pre-* koristi se u tvorbi prideva, priloga i glagola sa istim, augmentativnim, tvorbenim značenjem:

Pridevi:	*krasan*	*prekrasan*
	širok	*prešírok*
Prilozi:	*daleko*	*predaleko*
	obimno	*preobimno*
Glagoli:	*soliti*	*presoliti*
	platiti	*preplatiti*

Svi prefiksi, međutim, nisu podjednako plodni. Tako se, na primer, prefiks *pa-* koristi samo u tvorbi imenica, i to samo u jednom značenju, za tvorbu imenica sa sporednim značenjem: *jeka - pajeka*, *perje - paperje*.

S druge strane, prefiks *po-*, na primer, se koristi u tvorbi glagola u četiri različita značenja:

Distributivno:	*razbijati*	*porazbijati*
	dizati	*podizati*
Deminutivno:	*sedeti*	*posedeti*
	pričati	*popričati*
Finitivno:	*piti*	*popiti*
	tonuti	*potonuti*
Transformativnost:	*čupati*	*počupati*
	pljuvati	*popljuvati*

Kako se prefiksalna tvorba reči sastoji od dopisivanja prefiksa ispred osnove ili samostalne leksičke jedinice (leksema), ponekad na spoju prefiksa i osnove ili leksema dolazi do fonoloških promena. Tako se, osim osnovnih oblika prefiksa, sreću i izmenjeni oblici prefiksa, koji se od osnovnih oblika razlikuju samo u izrazu i rezultat su alternacije fonema na tvorbenom spoju. Tako nastaju sledeće alternante (videti Tabelu 1 Priloga C):

a) **ozvučeni prefiksi**

Prema pravilu o raspodeli konsonanata prema zvučnosti po kome se ispred *zvučnog* konsonanta realizuje samo *zvučni* konsonant, bezvučni

Virtual Library of Faculty of Mathematics - University of Belgrade
elibrary.matf.bg.ac.rs

konsonant se ispred zvučnog konsonanta zamenjuje svojim zvučnim parom. Tako je *z-* ozvučena varijanta prefiksa *s-* koja se realizuje ispred osnova ili leksema koje počinju sa *B, D, G, Dž* i *Đ*. Na primer,

spasti		*zbiti*
i	ali	i
slepiti		*zbaciti*

b) **obezvučeni prefiksi**

Prema pravilu o raspodeli konsonanata prema zvučnosti, po kome se ispred *bezvučnog* konsonanta realizuje samo *bezvučni* konsonant, zvučni konsonant se ispred bezvučnog konsonanta zamenjuje svojim bezvučnim parom. Tako, na primer,

is- nastaje od *iz-* ispred *P, T, K, F, C, H*
izleteti ali *iskočiti*

op- nastaje od *ob-* ispred *T, K, Č, Ć, Š, S, F, C, H*
obložiti ali *opkoliti*

ot- nastaje od *op-* ispred *P, K, Č, Ć, F, C, H*
odmaknuti ali *otploviti*

c) **izmenjeni prefiksi**

Prema pravilu o raspodeli konsonanata po mestu nastanka *nenepčani* konsonanti *s* i *z* ispred *nepčanih* konsonanata zamenjuju se nepčanim konsonantima *ś* i *ź* a nenepčani konsonant *h* nepčanim *ś*. Tako, na primer,

ś- nastaje od *s-* ispred *Č, Ć*
sleteti ali *ščešljavati*

iź- nastaje od *iz-* ispred *Đ, Dž*
izleteti ali *iždavoliti se*

Događa se, takođe, da na spoju prefiksa i osnove ili leksema deluju dva pravila: pravilo o raspodeli konsonanata po zvučnosti i pravilo o raspodeli konsonanata po mestu nastanka. Tako, na primer,

iš- nastaje od *iz-* ispred *Č, Ć* umesto *is-*
išćupati a ne *izćupati* ni *isćupati*

d) **okrnjeni prefiksi**

Prema pravilu o raspodeli konsonanata s obzirom na istovetnost ili sličnost konsonanata, poslednji konsonant prefiksa ispred istovetnog ili sličnog konsonanta osnove ne menja se već se jednostavno gubi. Tako, na primer,

ra- nastaje od *raz-* ispred *Z, S, Ž, Š*
razdeliti ali *raširiti*

o- nastaje od *od-* ispred *Đ*
odlutati ali *odeliti* a ne *oddeliti*

e) **prošireni prefiks**

Prema pravilu o raspodeli konsonanata po kome se između prefiksalnog *s* i konsonanata *s, z, š, ž* realizuje vokal *a*, prefiks *s-* ima alternantu *sa-* koja se realizuje kada je prvi konsonant osnove *s, z, š, ž*. Na primer,

Virtual Library of Faculty of Mathematics - University of Belgrade
elibrary.matf.bg.ac.rs

skiseliti se		*sašiti*		*sšiti*
i	ali	i	a ne	i
stresti		*saznati*		*sznati*

Mnogi prošireni prefiksi su, međutim, uslovljeni leksički i nisu nastali usled fonoloških promena. Tako, na primer, za mnoge reči koje su nastale korišćenjem proširene alternante postoje odgovarajuće reči istog značenja koje su nastale od osnovnog oblika prefiksa i iste osnove. Na primer,

uza- nastaje od *uz-* : *uzavreti* ali i *uzvreti*

oba- nastaje od *ob-* : *obaviti* ali i *oviti* i *obviti*

Da bi se formulisala semantička pravila za rastavljanje reči na kraju retka nije, međutim, dovoljna samo lista osnovnih prefiksa srpskohrvatskog jezika i njihovih alternanti. Neophodno je, takođe, analizirati njihovo pojavljivanje na reprezentativnom korpusu pisanih tekstova. Ova analiza se sprovodi sa sledećim ciljevima:

- Utvrđivanje frekvencije pojavljivanja svakog pojedinačnog prefiksa u pisanom tekstu;

- Identifikacija prefiksa za koje postoje reči koje počinju prefiksnom niskom koja u toj reči, međutim, nije prefiks (na primer, prefiksna niska *ob-* je prefiks u reči *obuhvatiti* ali ne i u reči *obućari)*. Treba, takođe, utvrditi frekvenciju pojavljivanja ovakvih prefiksnih niski kao prefiksa u odnosu na ukupno pojavljivanje prefiksa.

- Identifikacija prefiksa za koje postoje reči koje počinju prefiksnom niskom koja u toj reči nije prefiks, već je prefiks kraća niska. (na primer, niska *naj-* je prefiks u reči *najelegantniji* dok je u reči *najednom* prefiks *na-*). Treba takođe utvrditi frekvenciju pojavljivanja dužih prefiksnih niski kao prefiksa u odnosu na ukupno pojavljivanje prefiksne niske.

- Identifikacija prefiksa koji se kombinuju sa drugim prefiksim i utvrđivanje liste prefiksa sa kojima se kombinuju. Na primer, u reči *do-pri-nositi* prefiks *do-* se kombinuje sa prefiksom *pri-* a u reči *is-po-raz-boljevati se* prefiks *is-* se kombinuje sa prefiksom *po-* a ovaj sa prefiksom *raz-*.

3.2.2 Formiranje frekvencijskog rečnika prefiksa

Da bi se izvršila analiza pojavljivanja prefiksa u korpusu srpskohrvatskih tekstova, formirana je, pre svega, lista osnovnih prefiksa i njihovih alternanata. Ova lista se sastoji od svih prefiksa datih u [Gramatika] čija je dužina najmanje dva slova (videti Tabelu 1 Priloga C.1 ovog rada). Ovo ograničenje je uvedeno iz dva razloga:

- jednoslovni prefiksi su manje zanimljivi sa stanovišta rastavljanja reči na kraju retka (izuzev u slučaju kombinavanja prefiksa) jer se reč na kraju retka nikada ne rastavlja tako da u retku ostane samo jedno slovo;

- jednoslovni prefiksi u srpskohrvatskom su *a-*, *i-*, *o-*, *u-* i *s-* (i njegove alternante *š-* i *z-*). Kako prema [Tomić78], reči koje

Virtual Library of Faculty of Mathematics - University of Belgrade
elibrary.matf.bg.ac.rs

počinju sa jednim od ovih ovih 7 slova čine 43% celog uzorka, identifikacija ovih prefiksa među svim rečima koje počinju nekim od ovih 7 slova bila bi na većem uzorku veoma dugotrajan posao podložan greškama. S druge strane, konsultovanje rečnika [Pravopis] pokazuje da je relativna frekvencija ovih prefiksa u odnosu na sve reči koje počinju odgovarajućim slovom mala.

Da bi se ostvarili ciljevi analize naznačeni u tački **3.2.1** i ovog puta je korišćen programski sistem **AURORA** za automatsko generisanje konkordanci i ostalih vrsta indeksa [Vitas79, Vitas82]. Za izradu konkordanci korišćen je deo postojećeg korpusa veličine 44.311 reči koji se sastoji od tekstova navedenih u Prilogu **A.2.**[1] Tekstovi ovog korpusa su kodirani na način opisan u tački **1.** ove glave.

Za potrebe izrade ovih konkordanci, *reč* je definisana na uobičajen način, kao niska slova između graničnika. U programskom sistemu **AURORA** izmenjen je samo onaj deo u kome se donosi odluka koje reči iz teksta se isključuju (ignorišu) a za koje od preostalih reči se izrađuju konordance.

Za potrebe analize pojavljivanja prefiksa u tekstovima na srpskohrvatskom jeziku ignorisane su sve funkcionalne (neznačeće) reči. Lista ovih reči se sastojala od 77 oblika pomoćnih glagola *biti* i *hteti*, 186 oblika zamenica, 93 predloga i 140 veznika, priloga, brojeva, rečci i uzvika. Na taj način je izbegnuta izrada konkordanci za visokofrekventne funkcionalne reči koje sadrže prefiksne niske (na primer, *nisam*, *neću*, *neko*, *niko* i sl.). Kako su rečnici funkcionalnih reči organizovani u binarna drveta koja se efikasno pretražuju [Sedgewick83][2], isključivanje ovih reči ne opterećuje prvi deo obrade (formiranje rečnika) dok znatno ubrzava drugi deo obrade (izdavanje konkordanci). Međutim, iako su ove reči isključene iz izrade konkordanci, sabirana je frekvencija njihovog pojavljivanja pa je kao dodatni rezultat ove analize dobijen i frekvencijski rečnik funkcionalnih reči koji je dat u Prilogu **D.**

Programski sistem **AURORA** je za potrebe ove analize uključivao u rečnik samo one značeće reči koje počinju jednom od 79 prefiksnih niski (videti Prilog **C.1**). Program je, pri tome, uveo dodatna ograničenja:

- za alternante prefiksa vodilo se računa o tome u kojim se slučajevima one mogu realizovati. Na primer, od reči koje počinju prefiksnom niskom *raš-* u rečnik su ulazile samo one reči u kojima iza prefiksne niske sledi slovo *Č* ili *Ć*. Tako je u rečnik uključena reč *rašćišćavanje* dok, na primer, reči *Raška* i *Rašica* ne bi bile uključene u rečnik.

---

[1] Veličina korpusa je bila diktirana računarom na kome je obrada izvršena - Računar IBM360/44 Računarske laboratorije PMF-a je imao unutrašnju memoriju od 128Kb.

[2] O organizaciji rečnika za efikasno pretraživanje biće više reči u Glavi 4 ovog rada.

Virtual Library of Faculty of Mathematics - University of Belgrade
elibrary.matf.bg.ac.rs

- za prefiksne niske koje se završavaju vokalom, u rečnik su ulazile samo one reči u kojima iza prefiksne niske sledi konsonantska grupa. Osnovni razlog za ovo ograničenje je činjenica da mnogi od ovih prefiksa predstavljaju vrlo frekventne slogove u srpskohrvatskom jeziku (npr, *pa-*, *po-*, *pri-* itd.) pa bi uključivanje svih ovih reči u rečnik znatno opteretilo analizu. S druge strane, slučajevi u kojima iza ovi prefiksnih niski sledi **V** ili **CV** ionako nisu sporni u svetlu postavljenih ciljeva.

Na ovaj način je među 44311 analiziranih reči identifikovano 4653 reči koje, pod gornjim ograničenjima, počinju prefiksnom niskom, od kojih su 1875 različite reči.

Sistemom **AURORA** su, zatim, za sve reči koje su uključene u rečnik izrađene konkordance u ćiriličnom poretku. U konkordancama je za svako pojavljivanje reči iz rečnika u tekstu izlistan kontekst u kome se ona pojavila od pet reči ulevo i udesno. Konkordance su i ovog puta omogućile da se identifikuju i otklone eventualne greške u korpusu koje su nastale kao posledica grešaka u kucanju.

Izrađene konkordance su osim toga omogućile ostvarivanje ciljeva analize koji su postavljeni u tački **3.2.1** i kategorizaciju prefiksa u cilju definisanja semantičkih pravila. Rezultati analize su sumirani u Prilogu **C.2**. U ovoj tabeli su za svaki prefiks dati sledeći podaci:

1. Dužina prefiksa;

2. Tip prefiksa (kategarizacija prefiksa biće objašnjena kasnije);

3. Da li je prefiks višesložni;

4. Za alternante prefiksa, vrsta fonološke promene koja je dovela do alternacije (prema podeli iz tačke **3.2.1**);

5. Pojavljivanje prefiksa u uzorku:
- (1) Prefiksna niska se ne pojavljuje u uzorku;
- (2) Prefiksna niska se pojavljuje u uzorku, ali nikada kao prefiks;
- (3) Prefiksna niska se pojavljuje u uzorku, nekada, ali ne uvek, kao prefiks;
- (4) Prefiksna niska se pojavljuje u uzorku uvek kao prefiks;

6. Za slučajeve **5**(3), relativna frekvencija pojavljivanja prefiksne niske kao prefiksa u odnosu na ukupno pojavljivanje prefiksne niske;

7. Da li prefiks sadrži kraći prefiks (redni broj kraćeg prefiksa, ili sam prefiks ako je jednoslovni);

8. Za slučajeve **5**(2) i **5**(3), ako prefiksna niska sadrži kraći prefiks, relativna frekvencija pojavljivanja kraćeg prefiksa u odnosu na ukupno pojvaljivanje prefiksne niske.

Virtual Library of Faculty of Mathematics - University of Belgrade
elibrary.matf.bg.ac.rs

3.2.3 Kategorizacija prefiksa i semantička pravila

Analiza pojavljivanja prefiksa u vezanom tekstu za potrebe definisanja semantičkih pravila za rastavljanje reči na kraju retka treba da pokaže koji se prefiksi slabo frekventni ili se uopšte ne pojavljuju u korpusu pa se mogu ignoristi, koje prefiksne niske su u korpusu uvek prefiksi, a za koje prefikse niske su nam neophodne dodatne informacije da bismo potvrdili, ili odbacili, pretpostavku da je ona prefiks.

U cilju definisanja semantičkih pravila izvršena je, na osnovu ove analize, kategorizacija prefiksa. Prilikom kategorizacije prefiksa vodilo se računa o tome da identifikacija prefiksa u reči, u cilju rastavljanja reči na kraju retka, treba da zadovolji sledeće uslove:

- Prvenstveni zadatak je da se reč ne rastavi na pogrešnom mestu. Prema tome, reč se rastavlja iza prefiksa samo ako ga sa sigurnošću možemo identifikovati;

- Proces identifikacije prefiksa u reči treba da bude nezavisan od pravila za rastavljanje reči, jer, kao što smo videli, ova pravila nisu čvrsto postavljena i mogu se menjati;

- Preporučljivo je reč rastaviti iza prefiksa, ali samo ako smo sigurni da smo identifikovali prefiks, a, naročito, ako je prefiks višesložni.

Kategorizacija prefiksa je učinjena s obzirom na sledeće osobine prefiksa:

- Višesložni prefiksi. Reč je poželjno rastaviti iz prefiksa a ne na nekom drugom mestu. Na primer, bolje je *posle-ratni* nego *po-sleratni*.

- Alternante prefiksa. Ove prefiksne niske mogu biti prefiksi u reči samo ako iza njih sledi neki iz unapred definisanog skupa slova. Osim toga, ove prefiksne niske nikad ne mogu biti prefiksi ako iza njih sledi vokal.

- Prefiks u sebi sadrži, kao prefiksnu nisku, drugi prefiks. U ovom slučaju potrebne su dodatne informacije da bismo kao prefiks identifikovali duži ili kraći prefiks (ili odbacili oba). Na primer, *uz-buditi*, *u-zidati*, *uzorak*.

- Prefiks se završava vokalom. Za prefikse koji se završavaju vokalom mogu se razlikovati sledeće mogućnosti:

  - iza prefiksa je **V**. Reč se može rastaviti iza prefiksne niske u svakom slučaju (**V-V**). Npr., *preko-okeanski*;
  - iza prefiksa je **CV**. Reč se može rastaviti iza prefiksne niske u svakom slučaju ((**C**)***V-CV**). Npr., *preko-potrebno*;
  - iza prefiksa je $\mathbf{C^nV}$, n>1. Reč se može rastaviti iza prefiksne niske ako iza prefiksa sledi neka od inicijalnih konsonantskih grupa. Npr., *ekstra-profit*, ali *eks-trakti* a ne *ekstra-kti*.

Virtual Library of Faculty of Mathematics - University of Belgrade
elibrary.matf.bg.ac.rs

Dakle, da bismo identifikovali prefiks potrebne su nam dodatne informacije samo ako iza prefiksne niske sledi C^nV, $n>1$ i C^n nije inicijalna konsonantska grupa.

- Prefiks se završava konsonantom. Za prefikse koji se završavaju konsonantom mogu se razlikovati dve mogućnosti:

  - iza prefiksa je vokal. Reč se rastavlja iza prefiksne niske samo kada je ona prepoznata kao prefiks. Npr., *ob-uhvatiti*, ali *o-bućari* a ne *ob-ućari*.
  - iza prefiksa je konsonant. Reč se može rastaviti iza prefiksne niske ako ona ne sadrži kraći prefiks. Npr., *pod-voditi*, ali *po-dvojiti* a ne *pod-vojiti*.

Dakle, da bismo identifikovali prefiks potrebne su nam dodatne informacije ako iza prefiksne niske sledi vokal ili ako prefiksne niske sadrži kraći prefiks.

- Pojavljivanje prefiksa u korpusu. Za definisanje semantičkih pravila su od značaja četiri mogućnosti:

  - Prefiksna niska se ne pojavljuje u korpusu. Ove prefikse niske možemo ignorisati jer su ti prefiksi slabo frekventni. Npr., prefiks *nadri-*;
  - Prefiksna niska se pojavljuje u korpusu ali nikad kao prefiks. Možemo smatrati da ove prefiksne niske u reči nikada nisu prefiks. Npr., prefiks *pa-*;
  - Prefiksna niska se pojavljuje u korpusu, nekada, ali ne uvek, kao prefiks. Za ove prefiksne niske su nam potrebne dodatne informacije da bismo identifikovali prefiks. Npr., *oba-sjati*, ali i *obala* i *o-baviti*;
  - Prefiksna niska se pojavljuje u korpusu i uvek je prefiks. Možemo smatrati da su ove prefiksne niske uvek prefiksi. Npr., prefiks *među-*.

Na osnovu ovih osobina prefiksa i analize njihovog uticaja na rastavljanje reči na kraju retka izvršena je sledeća kategorizacija prefiksa:

- Prefiksi tipa 1. Reč se uvek rastavlja iza prefiksnih niski ovog tipa jer važi jedno od sledećih tvrđenja:

  - Prefiksna niska je u korpusu uvek prefiks:

anti-	*beš-*	*iš-*	*među-*	*nat-*
ot-	*pot-*	*pret-*	*protiv-*	*raza-*
raš-	*trans-*	*ultra-*	*vele-*	

  - Prefiksna niska se u korpusu ne pojavljuje, pa smatramo da nećemo pogrešiti ako iza nje reč rastavimo:

iz-	*kontra-*	*kvazi-*	*mimo-*	*nadri-*
nazovi-	*nuz-*	*nus-*	*pseudo-*	*raž*
izvan-				

  - Prefiksna niska je alternanta prefiksa koja se završava konsonantom (dakle, iza nje obavezno sledi konsonant) a kraća alternanta istog prefiksa se u korpusu ne pojavljuje:

op-	*uš-*

Virtual Library of Faculty of Mathematics - University of Belgrade
elibrary.matf.bg.ac.rs

- Prefiksi tipa 2. Reč se iza prefiksne niske ovog tipa uvek rastavlja, osim ako iza nje sledi konsonantska grupa koja je "teška" za izgovor, odnosno nije u rečniku inicijalnih grupa. (To, pak, znači da se reč rastavlja na isti način kada je u njoj identifikovan prefiks kao i kada se identifikacija prefiksa ne vrši).

- Prefiksna niska je jednosložna i završava se vokalom:

de-	*do-*	*na-*	*ne-*	*ni-*
pa-	*po-*	*pra-*	*pre-*	*pri-*
pro-	*sa-*	*su-*	*za-*	*be-*
ra-				

- Prefiksna niska je višesložna i predstavlja prošireni prefiks osnovnog prefiksa. Ali, taj osnovni prefiks iza koga bi sledio poslednji karakter prefiksne niske se ne pojavljuje u korpusu. Okrenjeni prefiks, iza koga slede preostali karakteri prefiksne niske se pojavljuje u korpusu, ali okrnjeni prefiks je jednoslovni i iza njega se reč ne rastavlja:

oba-	*oda-*	*uza-*

- Prefiksi tipa 3. Prefiksi ovog tipa se završavaju konsonantom. Ako iza prefiksne niske nije vokal, reč se iza nje rastavlja jer rastavljanje konsonantske grupe ne može biti pogrešno pošto, iz raznih razloga, prefiksna niska bez poslednjeg konsonanta ne može biti prefiks:

eks-	*ek-* nije prefiks, ali se *eks-* pojavljuje u *ekser*;
dis-	*di-* nije prefiks, ali se *dis-* pojavljuje u *disanje*;
naj-	*na-* jeste prefiks, ali sa *J-* ne počinje ni jedna inicijalna KG, ali se *naj-* pojavljuje u *na-javiti*;
niz-	*ni-* jeste prefiks, ali se, osim u zamenicama, retko javlja. Ali zato se *niz-* pojavljuje u *nizanje*;
nis-	važe isti razlozi kao i za prefiks *niz-*;
pred-	*pre-* jeste prefiks, ali iza njega se u korpusu nikad ne pojavljuje DC^n, n>0. Ali zato se *pred-* pojavljuje u *pre-davati*;
raz-	*ra-* jeste prefiks, ali iza njega se u korpusu nikad ne pojavljuje ZC^n, n>0. Ali zato se *raz-* pojavljuje u *raz-obličiti* i *ra-zoren*.
uz-	*u-* jeste prefiks, ali iza njega se u korpusu nikad ne pojavljuje ZC^n, n>0. Ali zato se *uz-* pojavljuje u *uz-iskati* i *u-zidati*;
iz-	*i-* jeste prefiks, ali iza njega se u korpusu nikad ne pojavljuje ZC^n, n>0. Ali zato se *iz-* pojavljuje u *iz-okola* i *i-zaći*.

- Prefiksi tipa 4. Prefiksne niske su višesložne i u korpusu nisu uvek, ili nisu uopšte, prefiks. Potrebne su nam dodatne informacije da bismo identifikovali prefiks ukoliko želimo da rastavljanju iza prefiksa damo veći prioritet. Ako prefiksna niska ovog tipa nije prefiks u reči, propuštamo jednu moguću, ili čak preporučljivu, tačku u kojoj se reč rastavlja.

arhi- u reči *arhi-dakon*	ali i u reči *arhivirati*
beza- u reči *beza-zoran*	*bez-* u reči *bez-alkoholni*

Virtual Library of Faculty of Mathematics - University of Belgrade
elibrary.matf.bg.ac.rs

iza-	u reči *iza-brati*	*iz-*	u reči *iz-analizirati*
nada-	u reči *nada-sve*	*na-*	u reči *na-daleko*
novo-	u reči *novo-nastalo*	ali i u reči *novosti*	
poda-	u reči *poda-viti*	*po-*	u reči *po-daviti* [3]
polu-	u reči *polu-pismen*	*po-*	u reči *po-lupati*
posle-	u reči *posle-ratni*	*po-*	u reči *po-slednji*
protu-	u reči *protu-požarni*	*pro-*	u reči *pro-turiti*
samo-	u reči *samo-hodni*	ali i u reči *samoća*	
preko-	u reči *preko-potreban*	*pre-*	u reči *pre-koračiti*
ekstra-	u reči *ekstra-dohodak*	ali i u reči *ekstrahovati*	
super-	u reči *super-provodnik*	ali i u reči *superioran*	
inter-	u reči *interkontinentalni*	ali i u reči *interesantno*	

- Prefiksi tipa 5. Prefiksi ovog tipa završavaju se konsonantom, a iza njih se može pojaviti i vokal i konsonant. Prefiksna niska bez poslednjeg konsonanta je takođe prefiks koji se pojavljuje u korpusu, pa možemo identifikovati duži prefiks, ili jednostavno rastaviti reč iza njega, samo ako poslednji konsonant prefiksne niske i konsonanti koji slede formiraju konsonantsku grupu "tešku" za izgovor, odnosno konsonantsku grupu koja nije iz rečnika inicijalnih grupa. U svim ostalim slučajevim su nam za odlučivanje potrebne dodatne informacije.

bez-	*be-* je prefiks i pojavljuje se u rečima *be-zuban* i *be-zakonje*;
nad-	*na-* je prefiks i pojavljuje se u rečima *na-dvoje*, *na-dole* i *na-daleko*;
ob-	*o-* je prefiks i pojavljuje se u rečima *o-bližnji* i *o-beležiti*;
od-	*o-* je prefiks i pojavljuje se u rečima *o-dvajati* i *o-dignuti*;
pod-	*po-* je prefiks i pojavljuje se u rečima *po-dvojiti* i *po-dići*;
prek-	*pre-* je prefiks i pojavljuje se u rečima *pre-kaliti* i *pre-kvalifikovati*.

- Prefiksi tipa 6. Prefiksi ovog tipa završavaju se konsonantom, a iza njih se može pojaviti samo konsonant. Prefiksna niska bez poslednjeg konsonanta je takođe prefiks koji se pojavljuje u korpusu, pa možemo identifikovati duži prefiks, ili jednostavno rastaviti reč iza njega, samo ako poslednji konsonant prefiksne niske i konsonanti koji slede formiraju konsonantsku grupu "tešku" za izgovor, odnosno konsonantsku grupu koja nije iz rečnika inicijalnih grupa. U svim ostalim slučajevim su nam za odlučivanje potrebne dodatne informacije.

bes-	*be-* je prefiks i pojavljuje se u reči *be-skrupulozan*;
is-	*i-* je prefiks i pojavljuje se u reči *i-skočiti*;

---

[3] Ovo je primer homografije koja se na nivou reči, bez odgovarajuće sintaksne analize, ne može razrešiti.

Virtual Library of Faculty of Mathematics - University of Belgrade
elibrary.matf.bg.ac.rs

ras- *ra-* je prefiks i pojavljuje se u rečima *ra-stezati* i *ra-staviti*;

us- *u-* je prefiks i pojavljuje se u rečima *u-spavati se* i *u-stati*;

Da bi se ostvario i poslednji cilj analize pojavljivanja prefiksa u korpusu, izvršena je analiza pojavljivanja kombinacija prefiksa u korpusu koja je data u Tabeli 3.6. U tabeli su date sve moguće kombinacije dva prefiksa[4], pri čemu su podvučene one koje su se u korpusu ostvarile. Jedan od razloga za relativno malu učestanost pojavljivanja kombinacija prefiksa u korpusu leži, svakako, i u načinu na koji je korpus analiziran:

- Jednoslovni prefiksi nisu u listi prefiksa, pa se nije moglo analizirati kombinovanje ovih prefiksa sa drugima prefiksima;

- Za prefikse koji se završavaju vokalom, u rečnik su ulazile samo one reči kod kojih iza prefiksa sledi konsonantska grupa. Stoga se nije moglo analizirati kombinovanje ovih prefiksa sa prefiksima koji počinju vokalom ili segmentom oblika CV.

Izvršena analiza pojavljivanja kombinacija prefiksa u korpusu može se, stoga, smatrati relevantnom samo za potrebe definisanja semantičkih pravila za potrebe rastavljanja reči na kraju retka.

Kombinacije od tri prefiksa su u srpskohrvatskom retke (na primer, *is-po-razboljevati se* i *po-iz-o-stavljati*). U korpusu nije identifikovano ni jedno pojavljivanje kombinacije od tri prefiksa. Stoga smatramo da se, za potrebe rastavljanja reči na kraju retka, prilikom identifikovanja prefiksa možemo zaustaviti na drugom prefiksu.

Uzimajući u obzir izrađenu kategorizaciju prefiksa i pod pretpostavkom da se prvi prefiks sa sigurnošću može identifikovati, može se zaključiti da su sa stanovišta rastavljanja reči na kraju retka sporne samo one kombinacije kod kojih je zadovoljen jedan od sledeća dva uslova:

- Prvi prefiks je jednoslovni pa kao takav i nije u listi prefiksa koje identifikujemo jer posle njega se reč neće rastavljati. S druge strane, identifikovanje ovih prefiksa može biti od značaja za rastavljanje reči na nekom drugom mestu, ukoliko se on kombinuje sa nekim prefiksom (npr. *obespravljen*);

- Drugi prefiks je jednoslovan ili pripada kategoriji onih prefiksa koji zahtevaju neke dodatne informacije da bismo ih mogli identifikovati (prefiksi tipa 3-6).

Kombinacije prefiksa na koje je prilikom rastavljanja reči na kraju retka potrebno obratiti pažnju su u Tabeli 3.6 obeležene zvezdicom.

---

[4] Moguce kombinacije prefiksa je za glagole i prideve definisao prof. dr Ljubomir Popović. Prefiks NE- nije uziman u obzir.

Virtual Library of Faculty of Mathematics - University of Belgrade
elibrary.matf.bg.ac.rs

Prvi prefiks (tip)	Drugi prefiks (tip)				
DO (2)	PRI (2)				
IZ (3)	DO (2)	NA (2)	O*	UZ* (3)	
	U*				
IS (6)	PO (2)	PRE (2)	PRI (2)	PRO (2)	
NA (2)	DO (2)	PO (2)	PRE (2)	UZ* (3)	ZA (2)
OB (5)	U*	UZ* (3)			
OD (5)	U*	UZ* (3)			
OT (1)	PO (2)				
O	BEZ* (5)	BES* (6)	BEŠ* (6)		
PO (2)	IZ* (3)	NA (2)	NAJ* (3)	OD* (5)	PRI (2)
	PRO (2)	S*	UZ* (3)	U*	Z*
	ZA (3)				
POD (5)	NA (2)	PO (2)	U*		
PRE (2)	DO (2)	NA (2)	OB* (5)	O*	PO (2)
	PRO (2)	S*	UZ* (3)	U*	
PRI (2)	DO (2)	NA (2)	O*	PO (2)	U*
RAZ (3)	U*				
RAS (6)	PO* (2)	PRO* (2)			
S	PO* (2)	POD* (2)	UZ* (3)		
US (6)	PO* (2)				
U	PO* (2)				
ZA (2)	DO (2)	O*	PO (2)	U*	UZ* (3)

Tabela 3.6 Moguće kombinacije dva prefiksa u srpskohrvatskom jeziku

Virtual Library of Faculty of Mathematics - University of Belgrade
elibrary.matf.bg.ac.rs

4 KONSTRUISANJE RUTINE ZA RASTAVLJANJE REČI SRPSKOHRVATSKOG JEZIKA NA KRAJU RETKA

Cilj analize pojavljivanja konsonantskih grupa i prefiksa u srpskohrvatskom jeziku, koja je opisana u glavi 3 ovog rada, je konstruisanje rutine za rastavljanje reči srpskohrvatskog jezika koja se, u skladu sa diskusijom iz glave 2, zasniva na pretraživanju rečnika izuzetaka i primeni pravila. Tako konstruisana rutina bi se mogla ugraditi u neki program za uređivanje i oblikovanje teksta. S druge strane, cilj analize je izrada rečnika obrazaca za rastavljanje reči koji bi se ugradio u programski sistem za slaganje teksta TEX [Knut82], danas nesumnjivo jedan od vodećih produkata u domenu slaganja matematičkog teksta. Izrada rečnika obrazaca za srpskohrvatski jezik za TEX biće opisana u glavi 5 ovog rada.

U ovoj glavi biće opisana rutina za rastavljanje reči srpskohrvatskog jezika koja se zasniva na primeni sledećih koraka, u datom redosledu:

- *pretraživanje rečnika izuzetaka*, čime se razrešavaju problemi transliteracije;

- *pretraživanje rečnika prefiksa*, koje obezbeđuje potrebne informacije za odlučivanje da li je prefiksna niska u reči prefiks ili ne;

- *primena pravila* po kojima se osnova reči, dobijena odvajanjem prefiksa, dalje rastavlja.

Da bi se, dakle, realizovala željena procedura bilo je potrebno, pre svega, formirati rečnike izuzetaka i prefiksa a zatim i definisati pravila za rastavljanje reči.

4.1 Rečnik izuzetaka

Analiza pojavljivanja konsonantskih grupa u korpusu srpskohrvatskog jezika je pokazala da je frekvencija projavljivanja konsonantskih grupa *dj*, *nj*, *lj* i *dž* veoma mala. Stoga je odlučeno da se problem transliteracije razreši pomoću rečnika izuzetaka koji bi sadržao sve poznate reči u kojima se javljaju ove konsonantske grupe.

S druge strane, analiza je takođe pokazala da se ove konsonantske grupe u ekavskoj varijanti srpskohrvatskog jezika javljaju samo na spoju prefiksa i osnove. Tako se, na primer, konsonantska grupa *dj*, koja ima najveću frekvenciju od ove četiri konsonantske grupe, u korpusu pojavila u rečima *ovdje* (u pesmi A. Šantića "Ostajte ovdje") i *djed* koje obe pripadaju ijekavskoj varijanti srpskohrvatskog jezika. Da bi se izradio što potpuniji rečnik izuzetaka, analiziran je rečnik srpskohrvatskog jezika iz [Pravopis] i u njemu su pregledane sve reči koje počinju prefiksnim niskama koje se završavaju slovom *d* (*nad-*, *od-*, *pod-* i *pred-*) i *n* (*izvan-*, *in-* i *kon-*). Slovom *l* se ne završava ni jedan prefiks. Prefiksi *in-* i *kon-* se ne nalaze u našoj listi prefiksa (videti Prilog C), ali kako je analiza pokazala da se konsonantska grupa *nj-* pojavljuje upravo na spoju ovih prefiksa sa osnovom, oni su za potrebe transliteracije pridodati listi prefiksa.

Virtual Library of Faculty of Mathematics - University of Belgrade
elibrary.matf.bg.ac.rs

Iz reči identifikovanih u korpusu i pronađenih u rečniku nastao je sledeći rečnik izuzetaka:

Konsonantska grupa	Prefiks	Reči
DJ	*nad-*	*nadjačati nadjunačiti*
	od-	*odjahati odjava odjaviti*
		odjedanput odjednom odjedared
		odjedriti odjek odjeknuti
		odjekivati odjuriti odjužiti
	pod-	*podjarmiti podjarmljivati*
		podjednak podjesti podjedati
		podjezični
	pred-	*predjelo*
NJ	*in-*	*injekcija injunktiv*
	kon-	*konjugacija konjugirati*
		konjunkcija konjunktiv
		konjunktivitis konjunktura
Dž	*nad-*	*nadžeti nadžnjeti nadživeti*
		nadživljavati nadžupan
	pod-	*podžeći podžizati podžupan*

Kako je u konkretnoj programskoj realizaciji rečnik izuzetaka deo rečnika prefiksa, o samoj strukturi rečnika biće govora u tački 4.2.1 ove glave.

4.2 Rečnik prefiksa

Kao što je rečeno u tački 3.2.3, postoje prefiksne niske koje se ne mogu sa sigurnošću identifikovati kao prefiks u reči samo na osnovu pojavljivanja prefiksne niske na početku reči i fonoloških osobina glasova koji slede. Da bi se one sa sigurnošću identifikovale potrebne su dodatne informacije. Takvi su prefiksi tipa od 3 do 6.

Jedan način da se ovaj problem razreši je izgradnja *rečnika prefiksa*, koji će sadržati "sve" reči koje počinju prefiksnim niskama tipa od 3 do 6 a uz svaku od ovih reči i informaciju o tome da li je prefiksna niska u toj reči prefiks ili ne. Ovakva koncepcija sadržaja rečnika prefiksa motivisana je dvostrukom namenom takvog rečnika:

- rečnik je deo samostalne rutine za rastavljanje reči na kraju retka;

- rečnik prefiksa treba da pruži potrebne informacije za formiranje semantičkih obrazaca koji se uključuju u rečnik obrazaca programskog paketa T_EX.

Ovako zamišljen rečnik prefiksa se može izgraditi na dva načina:

- automatskim izdvajanjem svih reči koje počinju prefiksnim niskama tipa od 3 do 6 iz nekog postojećeg opšteg rečnika

Virtual Library of Faculty of Mathematics - University of Belgrade
elibrary.matf.bg.ac.rs

srpskohrvatskog jezika koji je u mašinski čitljivom obliku. Sve izdvojene reči čovek zatim može da snabde potrebnim informacijama;

- izdvajanjem svih reči koje počinju prefiksnim niskama tipa od 3 do 6 iz korpusa pisanih tekstova na srpskohrvatskom jeziku. U interaktivnom dijaloga sa čovekom se svaka izdvojena reč snabdeva potrebnim informacijama i smešta u rečnik.

Kao što je već rečeno u Glavi 2 ovog rada, ne postoji opšti rečnik srpskohrvatskog jezika u mašinski čitljivom obliku, pa se, kao jedino realno moguć, nametnuo drugi način za izgradnju ovako koncipiranog rečnika izuzetaka.

4.2.1 Struktura rečnika

Pri koncipiranju strukture rečnika prefiksa, kao rečnika svih reči koje počinju prefiksnom niskom, prvenstveno treba razrešiti problem efikasnog zapisa rečnika, u smislu korišćenja memorijskog prostora, a zatim i efikasnog pretraživanja rečnika. Pri tome, u srpskohrvatskom jeziku osnovni problem predstavljaju oblici reči koji se svi, a ne samo osnovni oblik kakav se sreće u klasičnim rečnicima, moraju naći u rečniku prefiksa. Punjenje rečnika svim oblicima reči znatno utiče na veličinu rečnika a, takođe, i na efiksanost pretraživanja. Tako je, na primer, engleski rečnik koji je Liang koristio za formiranje obrazaca naraštao sa 31.836 na 49.858 reči kada su rečniku dodati svi oblici reči. Desarmenien [Desarm87] ističe da je za francuski jezik rečnik koji sadrži sve oblike reči pet puta veći od rečnika koji sadrži samo osnovne oblike reči. Za srpskohrvatski jezik bi ovaj odnos, verovatno, bio i veći.

Stoga je za potrebe konstruisanja rečnika prefiksa primenjen postupak segmentacije reči srpskohrvatskog jezika koji je opisan u [Vitas85, AOT85]. Ovde ćemo ga ukratko opisati. Reč srpskohrvatkog jezika može se predstaviti u obliku:

p/r/s

gde je *p* prefikna niska, *s* je oblični nastavak a *r* je oblična (morfološka) osnova reči. Osnova upućuje na leksičko značenje reči, dok nastavci služe da se oblik reči uskladi sa drugim rečima u rečenici.

Predložen je sledeći postupak segmentacije:

- Proverava se da li reč počinje nekom prefiksnom niskom. Pri tome se u reči uvek pronalazi *najduži* prefiks. Ako ne počinje, dodeljuje joj se nulti prefiks (tj., prefiks te reči je prazna niska);

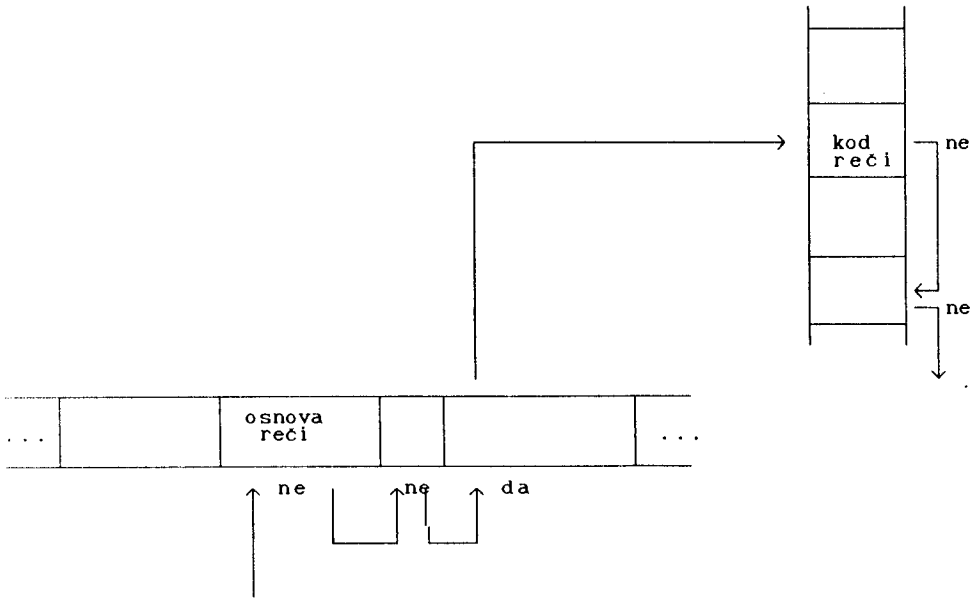
- Proverava se da li se reč završava obličnim nastavkom. (Lista obličnih nastavaka data je u Prilogu **E**). Pri tome se u reči uvek pronalazi *najduži* nastavak. Ako se ne završava, dodeljuje joj se nulti nastavak;

- Podniska *r* se određuje tako što se iz reči izdvajaju prepoznati prefiks i nastavak. Podniska *r* ne sme biti prazna.

Ovakvim postupkom će, na primer, reč *prava* biti segmentirana na *pra/v/a* a reč *prastar* na *pra/star/ø*. Reč *prede* se ne može segmentirati kao *pred/ø/e*, jer je osnova prazna niska, već kao *pre/d/e*.

Virtual Library of Faculty of Mathematics - University of Belgrade
elibrary.matf.bg.ac.rs

U reči *prekriti* se kao prefiks prepoznaje duža prefiksna niska *prek-*, a ne *pre-*, pa se ova reč segmentira kao *prek/ri/ti*. Kao što se već i iz ovih primera vidi, ovaj postupak segmentacije je mehanički: *pra-* je prefiks u reči *prastar* ali ne i u reči *prava*, dok u reči *prede* prefiks nije ni *pred-* ni *pre-*. Takođe, izdvojena osnova ne odgovara, osim izuzetno, korenskoj morfemi date reči. Bez obzira na ove nedostatke, ovakav postupak segmentacije se pokazuje korisnim pri konstruisanje rečnika prefiksa.



Slika 4.1 Pretraživanje rečnika. Osnova reči izvađene iz korpusa traži se u rečniku osnova koji je uređen po prvom slovu osnove. Za pronađenu osnovu traži se *kod reči* koji odgovara prefiksu reči iz korpusa. U listi kodova su olančani kodovi reči koji odgovaraju istoj osnovi.

Struktura samog rečnika je sledeća. Reč koju treba smestiti u rečnik se segmentira, a u rečnik se smešta samo podniska *r*, ako u rečniku već ne postoji. Svakoj ovakvoj reči iz rečnika pridružuje se lista *kodova reči*. Kod reči sastoji se iz:

- koda prefiksne niske *p* i
- informacije da li je ta prefiksna niska u reči prefiks.

Dakle, iz reči u rečniku i svih kodova koji su joj pridruženi mogu se reprodukovati sve polazne reči, do obličnog nastavka *s*. Na ovaj način se veličina rečnika redukuje a pretpostavka je da potpuno isključivanje obličnog nastavka u procesu pretraživanja neće bitnije uticati na tačnost rezultata. Ovako koncipirana struktura rečnika je, kao što ćemo kasnije videti, od velike pomoći prilikom formiranja semantičkih obrazaca.

Informacija u kodu reči koja govori da li je prefiksna niska *p* u reči prefiks može imati tri vrednosti:

- Niska *p* je u reči *p/r/x* prefiks;
- Niska *p* u reči *p/r/x* nije prefiks;
- Niska *p* u reči *p/r/x* može biti prefiks, ali ne mora.

Postupkom mehaničke segmentacije reči koji se primenjuje za konstrukciju i pretraživanje rečnika prefiksa ne mogu se rešiti problemi homografije, pa je stoga otvorena mogućnost da za neku reč na

Virtual Library of Faculty of Mathematics - University of Belgrade
elibrary.matf.bg.ac.rs

ovom nivou analize ne možemo sa sigurnošću identifikovati prefiks.

Na slici **4.1** je shematski prikazana struktura rečnika i postupak njegovog pretraživanja.

4.2.2 Program za konstruisanje rečnika prefiksa

Da bi se nad korpusom pisanih tekstova formirao rečnik prefiksa, čija je struktura opisana u prethodnoj tački, razvijen je odgovarajući program, koji mora da zadovolji sledeće zahteve:

a) *Program radi nad proizvoljnim tekstom.* Program može da prihvati proizvoljan tekst na srpskohrvatskom jeziku koji je u mašinski čitljivom obliku. Dakle, program je nezavistan od ulazne azbuke kojom je tekst kodiran;

b) *Definicija reči i liste prefiksa i nastavaka nisu deo programa.* Definicija reči i liste prefiksa i nastavaka mogu se menjati, u zavisnosti od teksta koji se obrađuje i u skladu sa novim potrebama i zahtevima;

c) *Program je interaktivan.* Kad god se u tekstu naiđe na reč koja je kandidat za rečnik, čovek odlučuje o tome da li reč treba uneti u rečnik i koji je njen kod;

d) *Program se obučava u toku rada.* Rečnik se popunjava u toku rada programa, i čovek ne mora ponovo da odlučuje o rečima koje su već u rečniku.

Struktura izrađenog programa koji zadovoljava gornje uslove je sledeća:

```
program {program za formiranje rečnika prefiksa}
učitaj_liste_prefiksa,_nastavaka,_inicijalnih_konsonantskih
     grupa_i_kodova_ulaznog_teksta;
if dopuna_postojećeg_rečnika
   then učitaj_parametre_postojećeg_rečnika;
repeat
   izdvoji_reč_i_prekodiraj_je;
   if kandidat_za_transliteraciju then
      if reč_pogrešno_kodirana then prekodiraj_je;
   if postoji_nastavak then odbaci_nastavak;
   if postoji_prefiks then
       begin odbaci_prefiks;
             if kod_prefiksa_od_3_do_7 and
                reč_je_kandidat_za_rečnik then
                begin if reč_je_u_rečniku and novi_prefiks
                      then unesi_novi_kod_reči;
                      if reč_nije_u_rečniku then
                      unesi_reč_i_kod_reči_u_rečnik
                end
       end;
until nema_više_reči or želim_kraj;
zapamti_parametre_rečnika;
izdaj_statistiku_rada_programa.
```

Virtual Library of Faculty of Mathematics - University of Belgrade
elibrary.matf.bg.ac.rs

Opišimo sada detaljnije kako su realizovani pojedinačni zahtevi koji su pred ovaj program postavljeni. Da bi se zadovoljio prvi zahtev uvedena je *interna azbuka programa*. Program izdvaja reči iz ulaznog teksta, koji može biti zapisan proizvoljnom azbukom, i svaku reč ponovo kodira internom azbukom programa. Nadalje se u programu koristi samo reč kodirana internom azbukom kojom su takođe zapisane i sve liste koje program koristi. Interna azbuka programa ne razlikuje mala i velika slova. Osnove reči u rečniku su takođe zapisane korišćenjem interne azbuke, pa je tako sam rečnik nezavisan od izvora iz koga je nastao. Na taj način su postignuta dva cilja:

- Rečnik može da se izgrađuje nad proizvoljnim tekstom;
- Zapis reči korišćenjem interne azbuke je nedvosmislen.

Definicija reči je izdvojena iz programa. Rutina programa koja izdvaja reč iz teksta (potprogram **REC**), pretpostavlja da je klasa izdvojenog simbola, koji se može sastojati iz jednog ili više karaktera, jedna od sledećih predviđenih klasa:

0. Nepoznati simbol (npr. simbol **X** u srpskohrvatskoj azbuci);

1. Simbol koji nikada nije deo reči (npr. interpunkcijski simboli, specijalni simboli, itd.);

2. Simbol koji je uvek deo reči (npr. slova);

3. Simbol koji se ignoriše (npr. karakter "+" se u azbuci teksta, koja nema mala i velika slova, može koristiti za označavanje velikih slova);

4. Simbol koji je deo reči samo ako se nalazi na kraju reči (npr. apostrof);

5. Simbol koji označava početak dela teksta koji se ignoriše, tj. iz koga se reči ne izdvajaju, koji se završava ponovnom pojavom istog simbola (npr. deo teksta na drugom jeziku, matematička formula, itd.);

6. Simbol kojim je reč rastavljena na kraju reda. Ovaj simbol se preskače a reč rekonstruiše (npr. u nekim azbukama teksta ovaj simbol mogu činiti tri karaktera: "-", CR i LF);

7. Simbol koji je deo reči samo ako se ne nalazi na početku reči (npr. cifre);

8. Simbol koji označava kraj reda i koji uvek završava reč;

9. Simbol koji je deo reči samo ako je unutar reči (npr. crtica);

Svakom simbolu azbuke ulaznog teksta se, u opisu azbuke, pridružuju sledeće informacije:

- klasa simbola (u smislu gornje definicije);
- brojčani kod simbola kojim se reguliše redosled simbola u azbuci;
- zapis simbola u internoj azbuci;
- dužina (u karakterima) zapisa simbola u internoj azbuci;

Na ovaj način se jednostavnom promenom klase određnih simbola u azbuci ulaznog teksta može promeniti definicija reči. Tako, na primer, cifre mogu biti u klasi 7 i na taj način deo reči, ako nisu na

Virtual Library of Faculty of Mathematics - University of Belgrade
elibrary.matf.bg.ac.rs

njenom početku, ili se, pak, mogu svrstati u klasu 1 što znači da nikada nisu deo reči. Jednostavnom promenom brojčanog koda simbola iz klase 2 se, takođe, može dobiti ćirilični, odnosno, latinični ili neki drugi redosled odrednica u rečniku. Struktura rutine za izdvajanje reči je:

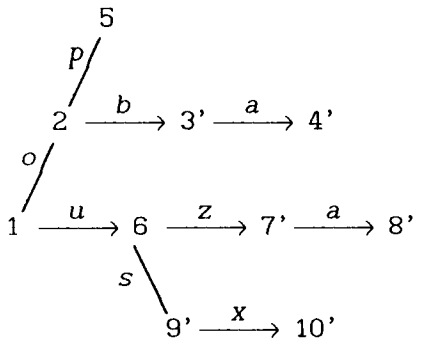
```
{Rutina za izdvajanje reči}
{i je pozicija simbola u ulaznom tekstu}
smer_skaniranja:=s_leva_udesno;
repeat {Traži se početak reči}
   kod:=kod_simbola(i); {Određuje se kod simbola na tekućoj poziciji}
   kraj:=dužina_simbola(i)*smer_skaniranja + i;
   case klasa(kod) of
      0: writeln('Nepoznati simbol');
      1: {Nije deo reči}
      2: begin ipoc:=i; pocetak_reci:=true end;
      3: {Ignoriše se}
      4: {Nije početak reči}
      5: while kod <> kod_simbola(tekst(i)) do
            i:=i+dužina_simbola(i);
      6: writeln('Crtica za rastavljanje nije u reci');
      7: {Nije početak reči}
      8: begin ucitaj_u_ulazni_bafer_sledeci_red_teksta;
            kraj:=0 end;
      9: {Nije početak reči}
   i:=i + smer_skaniranja {Pokazivač pokazuje sledeći simbol}
   if i > veličina_ulaznog_bafera then
      učitaj_u_ulazni_bafer_sledeći_red_teksta;
until not pocetak_reci;
repeat {Traži se kraj reči}
   kod:=kod_simbola(i); {Određuje se kod simbola na tekućoj poziciji}
   kraj:=i + dužina_simbola(i)*smer_skaniranja;
   case klasa(kod) of
      0: begin writeln('Nepoznati simbol'); kraj_reči:=true end;
      1: begin kraj_reči:=true; ikraj:=kraj end;
      2: begin prekodiraj_simbol_teksta_u_internu_azbuku;
            i:=kraj+smer_skaniranja end;
      3: i:=kraj+smer_skaniranja;
      4: begin ikraj:=i-1;
         kraj_reči:=true;
         if sledeći_simbol_nije_deo_reči then begin
            prekodiraj_simbol_teksta_u_internu_azbuku;
            ikraj:=kraj end;
      5: begin kraj_reči:=true; ikraj:=i-1 end;
      6: i:=kraj+smer_skaniranja;
      7: begin prekodiraj_simbol_teksta_u_internu_azbuku;
            i:=kraj+smer_skaniranja end;
      8: begin kraj_reči:=true; ikraj:=i-1 end;
      9: if sledeći_simbol_je_deo_reči then
            prekodiraj_simbol_teksta_u_internu_azbuku;
         else begin kraj_reči:=true; ikraj:=i-1 end;
   if (i > veličina_ulaznog_bafera) and not kraj_reči then
      učitaj_u_ulazni_bafer_sledeći_red_teksta
until not kraj_reči;
```

Virtual Library of Faculty of Mathematics - University of Belgrade
elibrary.matf.bg.ac.rs

Za izdvajanje simbola iz ulaznog teksta, koristi se opšta rutina za pretraživanje drveta (potprogram **DRVO**) [Peterson80]. Osim azbuke ulaznog teksta, strukturom drveta su predstavljene i ostale liste koje program koristi, a to su:

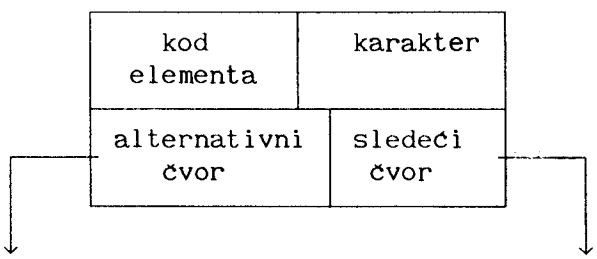
- lista prefiksa;
- lista nastavaka;
- lista inicijalnih konsonantskih grupa;
- lista početaka reči koje su, potencijalno, pogrešno kodirane u internoj azbuci.

Funkcija ovih drveta je efikasno sravnjivanje teksta sa elementima iz liste. Iz korena drveta polazi grana za svaki karakter kojim počinje neki element liste. One se dalje granaju na podrveta, pri čemu se sledeći karakter elementa liste koristi za dalje grananje, itd. Na primer, *ob-*, *op-*, *oba-*, *uz-*, *us-*, *uza-* i *usx-* su u listi prefiksa i deo drveta za pretraživanje ovih prefiksa izgleda:



Da bi prepoznali prefiks *usx-* polazi se iz čvora 1, a zatim se granom obeleženom sa *u* stiže u čvor 6. Zatim se granom *s* stiže u čvor 9 a odatle granom *x* u čvor 10. Čvor 10 je list drveta pa je prefiks *usx-* prepoznat. Za nisku *uzx-* se iz čvora 1 preko čvora 6 stiže u čvor 7 a iz tog čvora ne polazi grana obeležena sa *x* pa *uzx-* nije prefiks. Drvo na gornjoj slici, koje odgovara delu liste prefiksa, je binarno ali u opštem slučaju broj grana koje izlaze iz jednog čvora odgovara broju karaktera koji mogu da slede do tada prepoznatu nisku. Listovi drveta odgovaraju prepoznatom elementu liste, ali kako neki prefiks elementa liste može takođe da bude element liste u drvetu su obeleženi oni čvorovi koji predstavljaju kraj nekog elementa liste.

Sama drveta su predstavljena pomoću povezane liste. Svaki čvor i list drveta u povezanoj listi kojom je drvo predstavljeno ima oblik:



Kod elementa liste predstavlja identifikaciju elementa liste i ima funkciju uređivanja elemenata liste a, osim toga, predstavlja obeležje kraja elementa liste: nenulta vrednost koda elementa liste označava da je prepoznat element liste. Međutim, ako *sledeći čvor*

Virtual Library of Faculty of Mathematics - University of Belgrade
elibrary.matf.bg.ac.rs

pokazuje neki čvor drveta, pretraživanje drveta se nastavlja jer uvek pokušavamo da pronađemo najduži element liste.

Drveta su organizovana prema smeru skaniranja teksta. Drveta za sravnjivanje prefiksa, inicijalnih konsonantskih grupa i početaka reči koje su, potencijalno, pogrešno kodirane su organizovana za skaniranje teksta s desna ulevo, drvo za sravnjivanje nastavaka je organizovano za skaniranje teksta s leva u desno dok za azbuku ulaznog teksta postoje dva drveta: jedno za skaniranje teksta s desna ulevo a drugo za skaniranje teksta s leva u desno.

Rutina za pretraživanje drveta koje je predstavljeno ovakvom povezanom listom ima sledeću strukturu:

```
procedure drvo(tekst,ipoc,izav,smer,ikrj,sravnjen_element);
if karakterom_tekst(ipoc)_počinje_neki_element_u_listi then
   begin
      i:=ipoc; ikrj:=i; kod_elementa_liste:=0;
      p:=čvor(kod_prvog_karaktera);
      while postoji_čvor and not kraj_tekst do
         begin
            if kod_elementa(p) <> 0 then
               begin kod_elementa_liste:=kod_elementa(p);
                     ikrj:=i+smer_skaniranja end;
            i:=i+smer_skaniranja;
            p:=sledeći_čvor(p);
            if ima_još_čvorova and not kraj_teksta do
               then while postoji_čvor and
                          (karakter(p) <> tekst(i)) do
                          p:=alternativni_čvor(p);
         end;
      sravnjen_element_liste:= kod_elementa_liste <> 0;
   end;
```

Početak i kraj dela teksta koji se sravnjuje sa konkretnim drvetom imaju različito značenje koje zavisi od toga koja lista je drvetom predstavljena:

- *lista prefiksa*:	početak teksta je početak reči kraj teksta je kraj reči;
- *lista nastavaka*:	početak teksta je kraj reči kraj teksta je početak reči;
- *lista inicijalnih konsonantskih grupa*:	početak teksta je prvi konsonant kraj teksta je poslednji konsonant;
- *lista početaka potencijalno pogrešno kodiranih reči*:	početak teksta je početak reči kraj teksta je kraj reči;
- *azbuka ulaznog teksta*:	početak teksta je tekući karakter kraj teksta je kraj ulaznog bafera.

Za svaku reč koja se izdvoji iz teksta i kodira internom azbukom programa se proverava da li je dobro kodirana. Većina tekstova na srpskohrvatskom jeziku zabeleženih na magnetnom medijumu je uneta korišćenjem tastature koja nema posebne tastere za digrafe *dj*, *nj*, *lj* i *dž* [Krstev88]. Ovi digrafi, koji u srpskohrvatskom jeziku predstavljaju

Virtual Library of Faculty of Mathematics - University of Belgrade
elibrary.matf.bg.ac.rs

zasebna slova u alfabetu, su stoga u zapisu teksta najčešće zabeleženi pomoću dva koda: digraf *dj* je, na primer, zabeležen kodom slova *d* i kodom slova *j*. Kako je analiza pojavljivanja konsonantskih grupa (tačka 3.1) pokazala da je frekvencija pojavljivanja konsonantskih grupa *dj*, *nj*, *lj* i *dž* mala, odgovarajući parovi kodova u tekstu se u internoj azbuci uvek kodiraju kao jedno slovo, na sledeći način:

dj	i	*Dj*	i	*DJ*	⇒	*DX*
nj	i	*Nj*	i	*NJ*	⇒	*NX*
lj	i	*Lj*	i	*LJ*	⇒	*LX*
dž	i	*Dž*	i	*DŽ*	⇒	*DY*

Prema onome što je rečeno o problemu transliteracije (tačka **4.1**), pogrešno mogu biti kodirane samo one reči koje počinju niskama *NADX-*, *ODX-*, *PODX-*, *PREDX-*, *INX-*, *KONX-*, *NADY-* i *PODY-*. U procesu konstruisanja rečnika čovek treba da odluči Za sve reči koje počinju ovim niskama da li su dobro kodirane. Ako nisu, rutina za proveru ispravnosti kodiranja reči ih ponovo kodira i ove reči se zatim smeštaju u rečnik izuzetaka, ako se u njemu već ne nalaze. Kao što je rečeno u tački **4.1** rečnik izuzetaka je, u stvari, sastavni deo rečnika prefiksa. To znači da se i ove reči segmentiraju na opisani način, pošto svaka od njih sadrži prefiks iz liste prefiksa.

Svakom čvoru drveta za sravnjivanje gornjih niski koji je završni, su uz *kod elementa liste* pridružene još dve informacije:

- *klasa elementa liste*, koja definiše koja konsonantska grupa je pogrešno kodirana;
- *kod sadržanog prefiksa*;

Struktura rutine koja proverava kodiranje reči (potprogram TRLIT) je sledeća:

```
smer:=s_desna_u_levo;
drvo(tekst,ipoc,izav,smer,ikrj,sravnjen_element_liste);
if sravnjen_element_liste then
   begin
      writeln('Da li je reč ispravno kodirana?');
      for i:=ipoc to izav do write(tekst(i));
      readln(odgovor);
      if odgovor='ne' then
         case klasa_elementa_liste of
            1: zameni_kod_slova_DX_kodovima_slova_D_i_J;
            2: zameni_kod_slova_NX_kodovima_slova_N_i_J;
            3: zameni_kod_slova_LX_kodovima_slova_L_i_J;
            4: zameni_kod_slova_DY_kodovima_slova_D_i_Ž;
         end;
      prefiks_u_reči:=kod_sadržanog_prefiksa;
      tip_prefiksa:= 7; {Prefiksi tipa 7 obavezno ulaze
                              u rečnik}
   end;
```

Deo programa koji utvrđuje da li je reč koja sadrži prefiks čiji je tip broj od 3 do 7 kandidat za rečnik ne traži posebno objašnjenje. Ovaj deo programa se zaniva na klasifikaciji prefiksa koja je data u tački **3.2.3**. Struktura ovog dela programa je:

Virtual Library of Faculty of Mathematics - University of Belgrade
elibrary.matf.bg.ac.rs

```
case tip_prefiksa of
     1: nije_kandidat_za_rečnik;
     2: nije_kandidat_za_rečnik;
     3: if iza_prefiksa_je_vokal then
           jeste_kandidat_za_rečnik;
     4: jeste_kandidat_za_rečnik;
     5: if iza_prefiksa_je_vokal or
           poslednji_konsonant_prefiksa_i_konsonanti_koji_slede
           čine_inicijalnu_konsonantsku_grupu then
           jeste_kandidat_za_rečnik;
     6: if poslednji_konsonant_prefiksa_i_konsonanti_koji_slede
           čine_inicijalnu_konsonantsku_grupu then
           jeste_kandidat_za_rečnik;
     7: jeste_kandidat_za_rečnik;
end;
```

Za svaku reč koja je kandidat za rečnik a koja se još ne nalazi u rečniku, čovek koji popunjava rečnik treba, pre svega, da odluči da li je reč ispravna. Stoga se na zaslonu računara osim same reči prikazuje i cela rečenica u kojoj se ta reč pojavljuje. Na taj način se izbegava opasnost da u rečnik uđu neispravne reči koje se u tekstu pojavljuju, najverovatnije, kao posledica grešaka u kucanju. Ako je reč ispravna, čovek zatim treba da odluči da li je sravnjena prefiksna niska, a to je uvek najduža moguća prefiksna niska, prefiks u reči ili ne. Osnova reči i njen kod, koji čine kod prefiksa i informacija da li je prefiksna niska u reči prefiks, se zatim smeštaju u rečnik.

U svakom trenutku u kome čovek komunicira sa programom, on može da zahteva prekid rada programa. U svakom slučaju, bilo da je program završio sa radom jer je analiziran ceo tekst ili zato što je čovek to zahtevao, pamte se svi parametri rečnika koji omogućavaju njegovo pretraživanje. Taj rečnik se može, u nekoj sledećoj sesiji rada programa, dopunjavati rečima iz nekog drugog teksta a može se i koristiti u okviru rutine za rastavljanje reči na kraju retka (tačka 3 ove glave).

Posle svake sesije rada programa izdaje se osnovna statistika rada programa koja sadrži sledeće podatke:

- broj analiziranih redova teksta;
- broj analiziranih reči teksta;
- broj reči teksta koje su kandidat za rečnik;
- broj osnova reči pronađenih u rečniku;
- broj osnova koje su sa istim prefiksom pronađene u rečniku;
- broj osnova u rečniku pre početka sesije;
- broj osnova dodatih rečniku u toku sesije;

Osim toga, za svaki prefiks iz liste prefiksa se izdaju sledeći podaci:

- koliko puta je prefiksna niska sravnjena na početku reči;
- koliko puta je reč koja počinje tom niskom bila kandidat za rečnik;
- koliko puta je sravnjena prefiksna niska bila prefiks u reči koja nije pronađena u rečniku;
- koliko puta sravnjena prefiksna niska nije bila prefiks u reči koja nije pronađena u rečniku.

Virtual Library of Faculty of Mathematics - University of Belgrade
elibrary.matf.bg.ac.rs

Program za formiranje rečnika prefiksa je napisan na programskom jeziku FORTRAN77 a realizovan je i primenjen na IBM-PC kompatibilnom personalnom računaru. Dva razloga su uticala na izbor programskog jezika. U vreme početka rada na ovom problemu, autoru je bio na raspolaganju samo programski jezik FORTRAN IV [Parezanović72], pa je prva verzija ovog programa napisana na FORTRAN-u IV i implementirana na računaru IBM 360/44. Osim toga, najveći deo postojećeg programskog ambijenta u kome su vršene analize i provere dobijenih rezultata takođe je realizovan na programskom jeziku FORTRAN. Iako programski jezik FORTRAN, kao i većina drugih opštenamenskih jezika, nije posebno pogodan za obradu tekstualnih informacija [Day84], postojanje odgovarajućeg programskog ambijenta i kao razvoj mnogih opštih funkcija namenjenih obradi tekstualnih informacija koje u FORTRAN nisu ugrađene omogućilo je uspešno prevazilaženje svih teškoća.

4.2.3 Analiza rečnika prefiksa konstruisanog nad korpusom

Korišćenjem opisanog programa konstruisan je rečnik prefiksa nad korpusom pisanih tekstova (Prilog **A.3**). Iz korpusa je izdvojeno 212.816 reči, od kojih su 5.763 reči bile kandidat za rečnik. Konstruisani rečnik sadrži 882 različite osnove, dok osnova sa različitim prefiksom ima 978. Frekvencija reči u rečniku je 5.657. Preostali kandidati za rečnik, njih 106, su odbačeni iz jednog od dva razloga:

- reč je pogrešno zapisana, kao posledica greške u kucanju;
- reč predstavlja vlastito ime. Rečnik prefiksa je zamišljen kao opšti rečnik, pa stoga vlastita imena u njega nisu uvrštena.

Tabela iz Priloga **F.1** takođe pokazuje da ni za jednu prefiksnu nisku, pa čak ni za prefiksne niske tipa 4, nisu sve identifikovane reči iz korpusa predstavljale kandidat za rečnik. Osim ova dva razloga, na smanjivanje broja kandidata za rečnik utiče i postupak segmentacije: Ako je osnova reči koja se dobija odbacivanjem prefiksa i nastavka prazna, ona ne ulazi u rečnik.

Od 882 osnove u rečniku, 55 su osnove koje pripadaju rečniku izuzetaka kojim se razrešavaju problemi transliteracije. Ovaj deo rečnika nije popunjen nad korpusom, već je on popunjen rečima za koje je bilo unapred poznato da zahtevaju razrešavanje problema transliteracije (vidi tačku **4.1**).

U rečniku prefiksa na svaku osnovu prosečno dolazi 1,109 prefiksa. Iz ovog podatka se može zaključiti da se, za potrebe rečnika prefiksa, primenjenim postupkom segmentacije ne postiže veliki učinak u prostornom sažimanju rečnika. S druge strane se, pak, u velikoj meri smanjuje "eksplozija" oblika reči u rečniku. Uzmimo za primer glagole *odvesti* i *odvoditi*, u kojima je *od-* prefiks. U rečnik je uvršteno 9 osnova koje odgovaraju paradigmi ovih glagola. Ovih 9 osnova, pak, pokriva 52 oblika glagola *odvoditi*. U rečniku nedostaje samo osnova *voda* (za 1. i 2. lice množine imperfekta *od/voda/smo* i *od/voda/ste*) da bi bila pokrivena celokupna paradigma ovih glagola.

Virtual Library of Faculty of Mathematics - University of Belgrade
elibrary.matf.bg.ac.rs

osnova	oblici		
1.*ve*	*od/ve/o*		
2.*veden*	*od/veden/ø* *od/veden/i* *od/veden/e* *od/veden/u*	*od/veden/og* *od/veden/im* *od/veden/a* *od/veden/og*	*od/veden/om* *od/veden/ih* *od/veden/oj*
3.*vel*	*od/vel/i* *od/vel/o*	*od/vel/e*	*od/vel/a*
4.*ves*	*od/ves/ti*		
5.*vod*	*od/vod/i* *od/vod/imo* *od/vod/ah* *od/vod/eći*	*od/vod/e* *od/vod/ite* *od/vod/aše*	*od/vod/im* *od/vod/ih* *od/vod/ahu*
6.*vodi*	*od/vodi/o* *od/vodi/še*	*od/vodi/smo* *od/vodi/vši*	*od/vodi/ste*
7.*vodil*	*od/vodil/a* *od/vodil/e*	*od/vodil/o*	*od/vodil/i*
8.*vođen*	*od/vođen/ø* *od/vođen/i* *od/vođen/e* *od/vođen/u*	*od/vođen/og* *od/vođen/im* *od/vođen/a* *od/vođen/og*	*od/vođen/om* *od/vođen/ih* *od/vođen/oj*
9.*vođenj*	*od/vođenj/e* *od/vođenj/em*	*od/vođenj/a* *od/vođenj/ima*	*od/vođenj/u*

Od 882 različite osnove u rečniku, uz 78 osnova su pridružene dve ili više prefiksnih niski. U slučajevima kada sve prefiksne niske pridružene osnovi predstavljaju prefiks u reči, osnove iz rečnika u većini, ali ne u svim, slučajevima predstavljaju korenske morfeme. Na primer,

ob/lete/ti
od/lete/ti

bes/pravn/o
is/pravn/o
us/pravn/o

ali

preko/morsk/a
nad/morsk/a

gde *morsk* ne predstavlja korensku morfemu jer reči *prekomorska* i *nadmoraska* sadrže tvorbeni sufiks *-ski*. U slučajevima kada, pak, sve, ili neke, prefiksne niske ne predstavljaju prefiks u reči, osnova iz rečnika ne predstavlja korensku morfemu. Na primer,

posle/ratn/im
ob/ratn/o

uz/el/i
od/el/a

pod/el/i
pred/el/u

Virtual Library of Faculty of Mathematics - University of Belgrade
elibrary.matf.bg.ac.rs

Za samo dve reči iz korpusa se samo na osnovu oblika reči ne može odlučiti da li je prefiksna niska u reči prefiks ili ne. To su infinitivi glagola *podići* (*pod-iđem*) i *podići* (*po-dignem*) i dativ imenice *uzor* (*u-zoru*), odnosno, treće lice množine prezenta glagola *uzorati* (*uz-oru*).

Tabela iz Priloga **F.1** pokazuje da se od 81 prefiksnih niski u korpusu nije pojavilo 17 niski, i to 11 niski tipa 1, 2 niske tipa 2 i 4 niske tipa 4. Značajno je uočiti da su samo 5 prefiksnih niski u rečima - kandidatima za rečnik bile uvek prefiksi, odnosno, nikad nisu bile prefiksi. Tako su *bes-* i *novo-* u rečima - kandidatima za rečnik uvek prefiksi a *niz-*, *nada-* i *poda-* nikad nisu prefiksi. Ovi rezultati potvrđuju izvršenu tipizaciju prefiksa.

Većinu od 19 različitih tekstova koji su sačinjavali korpus nad kojim je popunjavan rečnik prefiksa, njih 13, čine udžbenici za osnovnu školu. Od svih udžbenika za osnovnu školu, 8 predstavljaju udžbenike iste vrste. To su udžbenici za predmet "Poznavanje prirode i društva" za treći razred osnovne pod zajedničkim nazivom "Naš kraj". Tabela iz Priloga **F.2**, u kojoj su tekstovi navedeni u redosledu u kome su analizirani, pokazuje za ove tekstove tendenciju smanjivanja broja novih osnova u rečniku, kako u odnosu na ukupnu dužinu teksta tako i prema broju kandidata za rečnik. Kod udžbenika druge vrste (na primer, tekstovi 12. i 13.) uočava se značajnije povećanje broja novih osnova u rečniku. Iako su tekstovi koji ne pripadaju udžbeničkoj literaturi u korpusu bili slabije zastupljeni i njihova ukupna dužina je značajno manja, može se ipak zaključiti (na primer, tekst 14.) da je bogatstvo jezika u ovim tekstovima takvo da utiče na značajno povećanje broja novih osnova u rečniku. Za zadovoljavajuće popunjavanje rečnika prefiksa značajnu ulogu stoga ima proučavanje podjezika pojedinih struka te nadgradnja opšteg rečnika prefiksa nad tekstovima koji pripadaju određenim specifičnim oblastima.

4.3 Rutina za rastavljanje reči srpskohrvatskog jezika

Da bi se konstruisala rutina za rastavljanje reči srpskohrvatskog jezika koja se zasniva na kombinaciji strategije rečnika i strategije pravila potrebno je još definisati pravila po kojima će se rastavljati reči koje nisu sadržane u rečniku izuzetaka. Ova pravila moraju da zadovoljavaju sledeće uslove:

- Pravila se mogu automatski primeniti;
- Pravila su u skladu sa preporukama koje daje Pravopis;
- Pravilima se definiše rastavljanje svih reči koje su se pojavile u analiziranim korpusima.

4.3.1 Definisanje pravila

Kao što je već rečeno u tački **2.2.2**, najveći problem za rastavljanje reči srpskohrvatskog jezika predstavljaju konsonantske grupe. Stoga su, pre svega, utvrđeni principi za rastavljanje

Virtual Library of Faculty of Mathematics - University of Belgrade
elibrary.matf.bg.ac.rs

konsonantskih grupa u srpskohrvatskom jeziku.[1] Ako sa **V** označimo slovo koje pripada skupu vokala, sa **C** slovo koje pripada skupu konsonanata a sa **X** proizvoljno slovo, onda ove principe možemo formulisati na sledeći način:

1. Na početku sloga su prihvatljive konsonantske grupe koje su u srpskohrvatskom jeziku uobičajene na početku i kraju reči;

2. a) Idealna struktura sloga u srpskohrvatskom jeziku je **CV**, odnosno, najpoželjnije je reč rastaviti **X-CV-X**.

b) Reč se ne rastavlja $(C)^n$**-V**, tj. reč se ne rastavlja tako da se deo reči koji ostaje u retku završava konsonantima a deo reči koji se prenosi u sledeći redak počinje vokalom;

3. Početak dela reči koji je prenet u sledeći redak ima oblik $-(C)^nV$, pri čemu je $n \leq 3$, tj. ispred vokala se nalazi najviše tri konsonanta;

4. Završetak dela reči koji ostaje u retku ima oblik $-V(C)^n$, pri čemu je n jednako 0, 1 ili 2 a samo izuzetno 3;

5. Reč se ne rastavlja na $-C_1C_2V$, ako je:
$C_1 \in \{R, L, Lj, V, J, M, N, Nj\}$ (C_1 je sonant) i
$C_2 \notin \{R, L, Lj, V, J, M, N, Nj\}$ (C_2 nije sonant),
tj. sonante ne rastavljaju od vokala nesonanti;

6. Reč se može rastaviti na $-C_1C_2C_3V$ samo ako je konsonantska grupa $C_1C_2C_3$ uobičajena u srpskohrvatskom jeziku, tj. ako je:
$C_1 \in \{S, Z, Š, Ž\}$ i
C_2 je proizvoljan konsonant i
$C_3 \in \{R, L, Lj, V, J\}$ (C_3 je sonant);

7. Prihvatljivo je rastaviti reč u obliku C_1C_2- ako je:
1. $C_1 \in \{R, L, Lj, V, J, M, N, Nj\}$ (C_1 je sonant) i
C_2 je proizvoljni konsonant, ili
2. $C_1 \in \{S, Z, Š, Ž, F, H\}$ (C_1 je frikativ) i
$C_2 \notin \{R, L, Lj, V, J, M, N, Nj\}$ (C_2 nije sonant), ili
3. $C_1C_2 \in \{PS, BZ, KS, GZ, PT, BD, KT, GD\}$;

8. Sufiksi $-SK(I)$ i $-STV(O)$ se ne rastavljaju, to jest, reči koje sadrže ove sufikse se uvek rastavljaju kao $-SK$, odnosno, $-STV$.

9. Poželjno je reč rastaviti $-C_1C_2$ ako je:
1. $C_1 \in \{S, Z, Š, Ž\}$ (C_1 je frikativ) i
C_2 je proizvoljan konsonant, ili
2. $C_1 \notin \{R, L, Lj, J, N, Nj\}$ i
$C_2 \in \{R, L, Lj, V, J, M, N, Nj\}$ (C_2 je sonant).

U odnosu na preporuke koji su postavili Belić i Stefanović (vidi **2.3.2**), ovi principi se mogu, pre svega, grupisati u pozitivne principe koji preporučuju određeno ponašanje (1, 2a, 3, 4, 5, 6, 8, 9) i negativne principe koji zabranjuju određeno ponašanje (2b, 5, 8). Osim toga, za razliku od preporuka Belića i Stevanovića ovi principi

---

[1] Principe za rastavljanje konsonantskih grupa u srpskohrvatskom jeziku je, koristeći rezultate analize pojavljivanja konsonantskih grupa u korpusu pisanih tekstova, utvrdio dr Ljubomir Popović, profesor Filološkog fakulteta u Beogradu.

Virtual Library of Faculty of Mathematics - University of Belgrade
elibrary.matf.bg.ac.rs

posmatraju konsonantsku grupu kao celinu pa se odluka ne donosi samo na osnovu osobina prva dva konsonanta u grupi.

Svi ovi principi, osim principa 8, se prema diskusiji iz 2.2, mogu svrstati u ortografska pravila. Međutim, princip 2a, na primer, shvaćen kao ortografsko pravilo ima izuzetke (npr. *raz-uveriti*). To dalje znači da se pri izradnji rutine za automatsko rastavljanje reči na kraju retka pravila zasnovana na ovim principima moraju kombinovati sa rečnicima.

Principi 5 i 9 koji definišu koji se konsonanti, i u kojim situacijama, ne rastavljaju predstavljaju modifikaciju Belićevog skupa pravila za rastavljanje konsonanata. Podsetimo se da Belićeva pravila, kao i ovde postavljeni principi, definišu kada se konsonanti ne rastavljaju. Važno je uočiti da od inicijalnih konsonantskih grupa koje su se pojavile u korpusima, oblik koji navodi princip 9 nemaju samo grupe *GD*, *PS*, *HT*, *PT*, *TK*, *PŠ*, *PČ*, *KĆ*.

Princip 6 je proizašao iz analize pojavljivanja konsonantskih grupa u korpusima koja je pokazala sledeće:

- tročlane inicijalne konsonantske grupe imaju u najvećem broju oblik koji navodi princip 6 (od 20 različitih grupa 19 ima taj oblik). Grupe koje nisu tog oblika se pojavljuju u rečima stranog porekla (npr., *BLŠ* u poljskom imenu *Blšeslav*).
- Od 110 medijalnih tročlanih konsonantskih grupa koje se pojavljuju u korpusima, 23 ima oblik koji navodi princip 6. Grupe koje nisu tog oblika pojavljuju se, najčešće, u sledećim situacijama koje zahtevaju rastavljanje konsonantske grupe:

  - 12 grupa u rečima koje su preuzete iz drugih jezika (npr., *BSB* u nazivu dinsatije *Habsburg* ili *JBN* u reči *šoferšajbna*);
  - 32 grupe na spoju prefiksa i osnove (npr., većina tročlanih konsonantskih grupa, njih 17, koje počinju sa *J* su nastale su na spoju prefiksa *naj-* sa osnovom);
  - 12 grupa na spoju sufiksa *-ski* i osnove (npr., *DSK* u pridevu *beogradski*);
  - u složenicama 4 grupe (npr., *NGR* u imenu *Ivangrad*).

Od preostalih 25 grupa, najveći broj, njih 17, se javlja u rečima koje su odomaćene u srpskohrvatskom jeziku ali su, u osnovi, stranog porekla (npr. *NKC* u reči *funkcija* ili *KTR* u reči *elektrika*). Dve grupe se javljaju na spoju modifikovane osnove i sufiksa *-ki* (*NŠK* u *Karolinški* i *RŠK* u *Habsburški*) a dve grupe su posledica pojavljivanja glasa *a* u genitivu množine reči koje sadrže sufiks *-stvo* (*NST* u *domaćinstava* i *TST* u *bratstava*). Preostale 4 grupe pojavljuju se u domaćim rečima (*JSC* u *vojsci*, *JMLj* u *pozajmljivati*, *PŠT* u *opšte* i *opština* i *TKV* u *rotkva*).

Princip 7 je takođe proizašao iz analize pojavljivanja konsonantskih grupa u korpusima koja je pokazala sledeće:

- većina finalnih konsonantskih grupa ima oblik koji navodi princip 7 (npr., *LM* u *film*, *KS* u *keks*). S druge strane, kada se *ST*, kao najfrekventnija finalna grupa, nalazi na početku tročlane medijalne grupe, ta grupa uvek ima oblik koji definiše princip 6;

- pojedine tročlane konsonantske grupe koje nemaju oblik koji navodi princip 6 i ne rastavlju se obavezno zbog odbacivanja prefiksa ili sufiksa, nastaju kao posledica tvorbe prideva (npr., *KTN* u

Virtual Library of Faculty of Mathematics - University of Belgrade
elibrary.matf.bg.ac.rs

direktno, *LTN* u *asfaltni*, *NSN* u *renesansni*, *NTN* u *sedimentni*, *RSN* u *konkursni* i *RFN* u *metamorfni*).

Analiza pojavljivanja konsonantskih grupa u korpusima je pokazala da se sve tročlane grupe oblika *CSK*, kao i sve četvoročlane grupe oblika *CCSK*, javljaju na spoju osnove i sufiksa *-ski*. Izuzetak su samo grupe *JSK* i *KSK* koje se javljaju i na spoju prefiksa i osnove (*najskuplje* i *majski*, odnosno *ekskomunicirati*). Takođe, najveći broj četvoročlanih grupa oblika *CSTV* se javlja na spoju osnove i sufiksa *-stv*. Izuzetak je samo grupa *PSTV* u *sopstvenik*. Treba takođe napomenuti da od 22 različite četvoročlane grupe samo 3 grupe nemaju oblik *CCSK*, odnosno *CSTV*. To su grupe *KSPL* (*eksplozija*), *NSTR* (*demonstracija*) i *JSTR* (*najstrašniji*).

Iz ovako postavljenih principa za rastavljanje konsonantskih grupa definisana su pravila za rastavljanje konsonantskih grupa koja, između ostalog, definišu i prioritete primene pojedinih pravila.

Dvočlane konsonantske grupe. 1. $-C_1C_2V$, ako konsonantska grupa C_1C_2 zadovoljava uslov 9 ili, da bi bio zadovoljen princip 1, $C_1C_2 \in \{GD, PS, HT, PT, TK, PŠ, PČ, KĆ\}$.

2. C_1-C_2V, u svim ostalim slučajevima.

Tročlane konsonantske grupe. 1. $C_1-C_2C_3V$, ako je $C_2C_3=SK$ (princip 8).

2. $-C_1C_2C_3V$, ako konsonantska grupa $C_1C_2C_3$ zadovoljava princip 6.

3. $C_1-C_2C_3V$, ako konsonantska grupa C_2C_3 zadovoljava princip 9 ili, da bi bio zadovoljen princip 1, $C_2C_3 \in \{GD, PS, HT, PT, TK, PŠ, PČ, KĆ\}$.

4. $C_1C_2-C_3V$, ako konsonantska grupa C_1C_2 zadovoljava princip 7.

Napomena: Viši prioritet primene principa 9 i 1 nad principom 7 će na zadovoljavajući način rešiti većinu, ali ne sve, slučajeve kada tročlana grupa ne zadovoljava princip 6 i ne rastavlja se odbacivanjem prefiksa, odnosno sufiksa: npr. bolje je *konkurs-ni* nego *konkur-sni*. Na zadovoljavajući način biće rešen i veći broj složenica jer se konsonantske grupe češće javljaju u inicijalnoj nego finalnoj poziciji (*NGR* u *Ivan-grad*, *VGR* u *Danilov-grad*, *RPL* u *Šar-planinac* i *NBR* u *mašin-bravar*).

Četvoročlane konsonantske grupe. 1. $C_1C_2-C_3C_4V$, ako je $C_3C_4=SK$, odnosno $C_1-C_2C_3C_4V$ ako je $C_2C_3C_4=STV$ (princip 8).

2. $C_1-C_2C_3C_4V$, ako $C_2C_3C_4$ zadovoljava uslov 6.

3. $C_1C_2-C_3C_4V$, ako C_1C_2 zadovoljava uslov 7 a C_3C_4 uslov 9.

Napomena: Kao što je rečeno, samo 3 četvoročlane grupe pronađene u korpusu se ne rastavljaju po pravilu 1 za čevoročlane grupe. Od njih će dve grupe biti rastavljanje odbacivanjem prefiksa (*eks-plozija* i *naj-strašniji*), dok će treća biti rastavljena[2] na zadovoljavajući način (*demon-stracija*, *in-strukcija*, *kon-strukcija*).

---

[2] Prefiksi in- i kon- se analiziraju samo ako iza njih sledi slovo J, radi razrešavanja problema vezanih uz transliteraciju.

Virtual Library of Faculty of Mathematics - University of Belgrade
elibrary.matf.bg.ac.rs

4.3.2 Struktura rutine

Prilikom konstruisanja rutine za rastavljanje reči na kraju retka koja koristi rečnike i pravila vodilo se računa o sledećim zahtevima koje ona mora da zadovolji:

- *Rutina se može koristiti za proizvoljan tekst na srpskohrvatskom jeziku*, što znači da rutina mora biti nezavisna je od ulazne azbuke kojom je tekst kodiran;

- *Rutina neće pogrešno rastaviti reč.* Na pozicijama na kojima, na osnovu u rutinu ugrađenog znanja, nije jasno kako treba rastaviti reč, reč neće biti rastavljena. Rastavljanje reči se nastavlja od pozicije sledećeg vokala. U tome je smisao informacije koja je sadržana u rečniku prefiksa. Na primer, za reč *obuhvatiti*, ako je reč u rečniku sa informacijom da je *ob-* prefiks u reči, tačke u kojima se ona može rastaviti su *ob-u-hva-ti-ti*. Ako je ova reč u rečniku sa informacijom da *ob-* nije prefiks, tačke u kojima se ona može može rastaviti su *o-bu-hva-ti-ti*. Ako, pak, ova reč nije u rečniku ili je u rečniku sa informacijom da ne znamo da li je *ob-* prefiks, tačke u kojima se ona može rastaviti su *obu-hva-ti-ti*.

- *Rutina se može prilagoditi specifičnim potrebama.* Rutina neće rešavati probleme pogrešne transliteracije ako je, na primer, ulazna azbuka takva da su digrafi *dj*, *nj*, *lj* i *dž* u njoj nedvosmisleno kodirani. Rutina neće pretraživati rečnik prefiksa ako želimo da povećamo efikasnost rutine na račun kvaliteta.

Struktura rutine koja zadovoljava gornje uslove je sledeća:

```
{Rutina za rastavljanje reči na kraju retka}
kodiraj_reč_internom_azbukom;
nastavak := pocetak_reči;
vokal1 := prvi_vokal_posle_nastavak;
if transliteracija_je_problem and
   reč_je_kandidat_za_transliteraciju then
   reši_problem_transliteracije;
if rečnik_prefiksa_se_pretražuje and
   reč_počinje_prefiksnom_niskom then
   utvrdi_da_li_je_prefiksna_niska_prefiks_u_reči;
case kod_prefiksne_niske of
   prefiksna niska je prefiks: begin tačka_podele_iza_prefiksa;
                                     nastavak := početak_osnove;
                              end
   prefiksna niska nije prefiks: nastavak := početak_reči;
   ne znam da li je prefiksna niska prefiks:
      nastavak := prvi_vokal_iza_prefiksne_niske;
end;
vokal2 := prvi_vokal_iza_vokal1;
while postoji_vokal2 do
   begin if između_vokal1_i_vokal2_konsonantska_grupa
         then begin
                 rastavi_konsonantsku_grupu;
                 tačka_podele_iza_vokal1 + broj_konsonanata;
              end
         else tačka_podele_iza_vokal1;
         vokal1 := vokal2;
         vokal2 := prvi_vokal_iza_vokal1
   end;
```

Virtual Library of Faculty of Mathematics - University of Belgrade
elibrary.matf.bg.ac.rs

Opišimo sada detaljnije pojedine funkcije ove rutine. Kodiranje reči internom azbukom i ulogu interne azbuke smo opisali u tački **4.2.2**. Kako rutina za rastavljanje reči koristi samo reč kodiranu internom zabukom, neophodno je da se prilikom kodiranja reči internom azbukom uspostavi takva veza između reči kodirane internom azbukom i originalnog zapisa reči koja omogućava da se reč, zapisana ulaznom azbukom, rastavi u tački koju je rutina odredila za reč zapisanu internom azbukom.

Pretraživanjem odgovarajućeg drveta, kako je opisano u tački **4.2.2**, proverava se da li je reč kandidat za transliteraciju. Detaljnija struktura dela rutine koja rešava problem transliteracije je:

```
{Procedura koja rešava problem transliteracije}
begin
Privremeno_kodiraj_reč;
case klasa_transliteracije of
   1: privremeno_zameni_kod_slova_DX_kodovima_slova_D_i_J;
   2: privremeno_zameni_kod_slova_NX_kodovima_slova_N_i_J;
   3: privremeno_zameni_kod_slova_LX_kodovima_slova_L_i_J;
   4: privremeno_zameni_kod_slova_DY_kodovima_slova_D_i_Ž;
end;
proveri_da_li_je_reč_u_rečniku_izuzetaka;
if reč_je_u_rečniku_izuzetaka then
   begin
      zameni_kod_reči_privremenim_kodom_reči;
      {sve reči koje su pogrešno kodirane sadrže prefiks}
      reč_počinje_prefiksnom_niskom := true;
      prefiks_u_reči := kod_sadržanog_prefiksa;
      tip_prefiksa := 7
   end
end;
```

Provera da li reč počinje prefiksnom niskom se vrši pretraživanjem liste prefiksa koja je predstavljena strukturom drveta, kako je to opisano u tački **4.2.2**. Za reči koje su rutinom za rešavanje problema transliteracije ponovo kodirane se ne proverava da li počinju prefiksnom niskom jer je za njih taj podatak već utvrđen. Deo rutine za rastavljanje reči na kraju retka koji utvrđuje da li je prefiksna niska u reči prefiks ima sledeću strukturu:

```
{Procedura koja proverava da li je prefiksna niska prefiks}
begin
   if deo_reči_iza_prefiksa_duži_od_dva_karaktera then
      case tip_prefiksa of
         1: begin proveri_u_rečniku := false;
               kod_prefiksne_niske :=prefiksna_niska_je_prefiks
            end;
         2: begin proveri_u_rečniku := false;
               kod_prefiksne_niske :=prefiksna_niska_nije_prefiks
            end;
         3: if iza_prefiksne_niske_je_vokal then
               proveri_u_rečniku := true;
         4: proveri_u_rečniku := true;
```

Virtual Library of Faculty of Mathematics - University of Belgrade
elibrary.matf.bg.ac.rs

```
            5: if iza_prefiksne_niske_je_vokal or
                  poslednji_konsonant_prefiksne_niske_i_konsonanti
                  koji_slede_čine_inicijalnu_konsonantsku_grupu
                  then proveri_u_rečniku := true;
            6: if poslednji_konsonant_prefiksne_niske_i_konsonanti
                  koji_slede_čine_inicijalnu_konsonantsku_grupu
                  then proveri_u_rečniku := true;
            7: proveri_u_rečniku := true
         end;
      if proveri_u_rečniku then
         begin odbaci_nastavak;
               proveri_da_li_je_reč_u_rečniku_i_utvrdi_kod
               prefiksne_niske;
               if reč_nije_u_rečniku then
                  kod_prefiksne_niske :=
                        ne znam da li je prefiksna niska prefiks
         end
   end;
```

Opišimo detaljnije još samo proceduru za rastavljanje konsonantske grupe. Ova procedura se u potpunosti zasniva na pravilima postavljenim u tački **4.3.1**.

```
   {Procedura za rastavljanje konsonantske grupe}
   begin
      case dužina_konsonantske_grupe of
         2: if zadovoljen_princip_9 then
               broj_konsonanata := 0;
         3: if zadovoljen_princip_6 then
               broj_konsonanata := 0
            else if zadovoljen_princip_9 then
               broj_konsonanata := 1
                 else if zadovoljen_princip_7 then
                    broj_konsonanata := 2;
         4: if zadovoljen_princip_6 then
               broj_konsonanata := 1
            else zadovoljen_princip_7 and zadovoljen_princip_9 then
               broj_konsonanata := 2
      end;
```

4.3.3 Analiza rada rutine za rastavljanje reči

Rad opisane rutine za rastavljanje reči na kraju retka je testiran nad tekstom (Prilog **A.4**) koji je zadovoljavao sledeće uslove:

- ovaj tekst nije deo korpusa nad kojim je analizirano pojavljivanje konsonantskih grupa u pisanim tekstovima;

- ovaj tekst nije deo korpusa nad kojim je analizirano pojavljivanje prefiksa u pisanim tekstovima;

- ovaj tekst nije deo korpusa nad kojim je konstruisan rečnik prefiksa.

Za testiranje rutine za rastavljanje reči na kraju retka korišćen je programski sistem **AURORA** za automatsko generisanje konkordanci i ostalih vrsta indeksa [Vitas79, Vitas82]. U prvoj fazi je nad izabranim tekstom konstruisan rečnik svih različitih oblika reči koje su se u tekstu pojavile i koje imaju dužinu veću od 4 karaktera.

Virtual Library of Faculty of Mathematics - University of Belgrade
elibrary.matf.bg.ac.rs

Dužina izabranog teksta je 8.952 pojavljivanja reči, dok je konstruisani rečnik imao 3.075 različitih oblika reči. U drugoj fazi su za ovako formiran rečnik izrađene konkordance sa najviše tri konteksta za svaki oblik reči a za svaku reč su, osim toga, navedene tačke u kojima se reč može rastaviti. Kontekst u kome se reč pojavila je omogućio da se identifikuju greške u kucanju teksta, kao i oni homografi koji utiču na rastavljanje reči. Tačke u kojima se reč može rastaviti je odredila rutina za rastavljanje reči na kraju retka koja je koristila rečnike izuzetaka i prefiksa opisane u tački **4.2.3.**

Od 3.075 različitih oblika reči koje je rutina rastavljala, pogrešno je rastavljena 21 reč. Sve pogrešno rastavljene reči potiču od kombinovanja prefiksa o čemu rutina za rastavljanje reči nije vodila računa. To su sledeće reči:

ne-i-zvr-še-nje	umesto	*ne-iz-vr-še-nje*	
ne-o-bra-di-va	umesto	*ne-ob-ra-di-va*	
pro-i-zvo-di-ti	umesto	*pro-iz-vo-di-ti* (19 oblika)	

123 različitih reči je rastavljeno konsultovanjem rečnika prefiksa, dok su u 24 reči propuštene jedna, dve ili tri tačke u kojima se reč može rastaviti jer reči nisu bile u rečniku prefiksa. Na primer,

ispi-si-va-ti	umesto	*is-pi-si-vati*	
bezalko-hol-na	umesto	*bez-al-ko-hol-na*	
supermar-ket	umesto	*super-mar-ket*	ili
		su-per-mar-ket	

Strategiju po kojoj se, ako ne možemo da utvrdimo da li je prefiksna niska u reči prefiks, rastavljanje reči nastavlja posle prvog vokala iza prefiksne niske su nametnuli prefiksi koji imaju i proširenu i okrnjenu alternantu (npr., *beza-*, *bez-* i *be-*). Za ostale prefikse bi strategija mogla biti da se rastavljanje reči nastavlja od prvog vokala iza prefiksne niske. Na primer, ako reč *nadirati* nije u rečniku prefiksa, rutina bi mogla reč da rastavi kao *nadi-ra-ti* umesto sadašnjeg *nadira-ti*.

Analiza rada rutine za rastavljanje reči je, takođe, pokazala da bi se daljim profinjavanjem pravila za rastavljanje reči datih u tački **4.3.1** mogli postići bolji rezultati. Modifikacijom sledećih pravila mogao bi se poboljšati rad rutine:

- Ako je konsonant $C_1 = M$ i $C_2 \in \{N, L, R, Lj\}$, ili $C_1 = V$ i $C_2 \in \{L, R\}$ (što znači da C_1C_2 čine inicijalnu konsonantsku grupu) ili $C_1 \notin \{R, L, Lj, V, J, M, N, Nj\}$ i $C_2 \in \{R, L, Lj, V, J, M, N, Nj\}$, poželjno je reč rastaviti -C_1C_2. Ako se pravilo 9.2 modifikuje na ovaj način biće:

tram-vaj	umesto	*tra-mvaj*
za-bav-nik	umesto	*za-ba-vnik*

- Ako za tročlanu konsonantsku grupu ne važi pravilo 6, i ako je $C_1C_2 = KS$, onda se reč rastavlja C_1C_2-C_3. U protivnom, ako C_2C_3 zadovoljava pravilo 9 reč se rastavlja C_1-C_2C_3. U protivnom, ako C_1C_2 zadovoljava pravilo 7 reč se rastavlja C_1C_2-C_3. Ako se redosled primene pravila za tročlane konsonantske grupe promeni na ovaj način dobićemo:

teks-til	umesto	*tek-stil*

Virtual Library of Faculty of Mathematics - University of Belgrade
elibrary.matf.bg.ac.rs

5 RUTINA ZA RASTAVLJANJE REČI SRPSKOHRVATSKOG JEZIKA ZASNOVANA NA REČNIKU OBRAZACA

Kao što je već rečeno u tački **2.3.4**, za srpskohrvatski jezik ne postoje rečnici u kome su reči označene na mestima na kojima se mogu rastaviti. Stoga se za srpskohrvatski jezik ne može primeniti program PATGEN [Liang83], ili neki sličan program, koji bi generisao potrebne obrasce direktno iz informacija u rečniku. Stoga su, na primer, za francuski jezik obrasci generisani ručno, iz samih pravila [Desarm84f], [Desarm87]. Za poljski jezik, za koji takođe ne postoje rečnici sa potrebnim informacijama, rečnici obrazaca su generisani koristeći isti pristup [Kolodz87], [Kolodz88].

Znanja stečena analizom korpusa kao i na tom osnovu formulisana pravila i konstruisani rečnici izuzetaka predstavljaju potrebnu bazu za generisanje rečnika obrazaca za srpskohrvatski jezik. Ovi obrasci su, kao i za francuski i poljski jezik, generisani direktno iz samih pravila. Formiranje rečnika obrazaca je izvršeno u dve faze. U prvoj fazi su generisana pravila koja su zatim u drugoj fazi testirana nad korpusom. Analiza rezultata dobijenih u drugoj fazi je uticala na izmenu i dopunu rečnika obrazaca.

5.1 Generisanje obrazaca za srpskohrvatski jezik

Rečnik obrazaca za srpskohrvatski jezik je generisan postupno, pri čemu su u svakom koraku rečniku dodavani obrasci koji odgovaraju jednoj grupi pravila za rastavljanje reči srpskohrvatskog jezika. Prvo su generisani obrasci koji odgovaraju pravilima za rastavljanje reči srpskohrvatskog jezika (tačka **4.3.1**) a zatim obrasci koji odgovaraju semantičkim pravilima (tačka **3.2.3** i **4.2.3**).

(a) Dva vokala koja se u reči nalaze jedan do drugoga se rastavljaju. Iz ovog pravila je generisano 25 obrazaca oblika

V1V

gde **V** označava slovo iz skupa vokala {*a*, *e*, *i*, *o*, *u*}.

(b) Ako se između dva vokala nalazi jedan konsonanat on pripada drugom vokalu. Iz ovog pravila je generisano 105 obrazaca oblika

1CV

gde **C** označava slovo koje ne pripada skupu {*a*, *e*, *i*, *o*, *u*, *lj*, *nj*, *dj*, *dž*} a **V** slovo iz skupa vokala.

Alternativni zapis ovog pravila su obrasci oblika **CV1**. Pogledajmo koje su prednosti izabranog oblika obrazaca. Obrasci ovog oblika rastavljaju konsonantsku grupu, osim ako ne postoji neki drugi obrazac koji to zabranjuje, a to je pristup koji preporučuje i Pravopis (konsonantska grupa "laka" za izgovor se može, ali ne, mora rastavljati). Osim toga, ovi obrasci sprečavaju da se u sledeći redak prenesu samo konsonant, ili konsonanti, bez vokala.

(c) Niske *dj*, *nj*, *lj* i *dž* u najvećem broju slučajeva predstavljaju digrafe. Kako su ove niske u najvećem broju slučajeva

Virtual Library of Faculty of Mathematics - University of Belgrade
elibrary.matf.bg.ac.rs

digrafi uvodimo sledeća 4 obrasca:

1d2j 1n2j 1l2j 1d2ž

Kao što je rečeno u tački **4.1** ove niske se pojavljuju kao konsonantske grupe samo na spoju prefiksa sa osnovom, pa će rastavljanje ovih konsonantskih grupa obezbediti obrasci koji odražavaju odgovarajuća semantička pravila (vidi tačke (k) - (r) ovog poglavlja). Na ovom mestu ćemo uključiti samo 4 obrasca za identifikovanje prefiksa *in-* i *kon-*.

.in3jekc .in3junkt

.ko2n3jug .ko2n3junk

(d) Dvočlane konsonantske grupe se ne rastavljaju ako zadovoljavaju uslov 9. Kao što je rečeno, obrasci iz (b) rastavljaju sve konsonantske grupe ispred poslednjeg konsonanta, pa je potrebno uvesti nove obrasce koji obezbeđuju da se konsonantske grupe koje zadovoljavaju uslov 9 ne rastave. Tako dobijamo 96 novih obrazaca oblika:

$1C_1 2C_2$

gde je $C_1 \in \{S, Z, Š, Ž\}$ a C_2 je proizvoljan konsonant i 109 obrazca oblika:

$1C_3 2C_4$

gde je $C_3 = M$ i $C_4 \in \{N, L, R\}$ ili $C_3 = V$ i $C_4 \in \{L, R\}$ ili $C_3 \notin \{R, L, Lj, V, J, M, N, Nj\}$ a $C_4 \in \{R, L, Lj, V, J, M, N, Nj\}$. Ovim obrascima se dodaje još 8 obrazaca:

1g2d 1p2s 1h2t 1p2t
1t2k 1p2š 1p2č 1k2č

Ovim obrascima se dodaju još dva obrasca koja obezbeđuju rastavljanje konsonantskih grupa *mnj* i *vlj*:

2m3nj 2v3lj

Pogledajmo kako bi se na osnovu do sada formiranog rečnika obrazaca rastavila reč *moljci*.

. *m o l j c i* .
1m o
1l2j
1c i

Tako dobijamo mo1l2j1ci a to znači da se reč rastavlja *mo-lj-ci*. Dok obrasci oblika 1l2j dobro funkcionišu u slučaju pojavljivanja digrafa ispred vokala, da bi se onemogućilo rastavljanje reči ispred digrafa u slučajevima kada iza digrafa sledi konsonant, uvodi se još 11 obrazaca:

2ljn 2ljk 2ljs 2ljš 2ljc 2ljd
2ljb 2njs 2njc 2djs 2dž3b

Ovi obrasci su odabrani na osnovu analize pojavljivanja konsonantskih grupa (videti Tabelu **3.5**).

(e) Glas r između dva konsonanta preuzima ulogu vokala. To znači da se reč može rastaviti iza *r* ako se *r* nalazi između dva konsonanta. Iz ovog pravila se može generisati 576 obrazaca oblika:

Virtual Library of Faculty of Mathematics - University of Belgrade
elibrary.matf.bg.ac.rs

1C2rC

gde C označava slovo iz skupa konsonanata, bez slova *r* (C∉ {*a*, *e*, *i*, *o*, *u*, *r*}). Analiza pojavljivanja glasa r u interkonsonantskom položaju, kao deo analize pojavljivanja konsonantskih grupa, nam, međutim, pokazuje da su se u korpusu pisanih tekstova realizovalo samo 106 niski oblika C*R*C, i to:

brv	brg	brd	brž	brz	brk	brlj
	brn	brs	brt	brć	brč	
vrb	vrv	vrg	vrd	vrđ	vrk	vrl
	vrn	vrnj	vrs	vrt	vrć	vrh
	vrš					
grb	grd	grk	grl	grm	grn	grt
	grć	grč	grš			
drv	drz	drž	drl	drlj	drh	drš
zrn						
žrv	žrt					
krb	krv	krg	krž	krk	krm	krn
	krnj	krp	krs	krt	krc	krč
	krš					
mrv	mrd	mrž	mrz	mrk	mrl	mrlj
	mrnj	mrt	mrć	mrš		
prv	prž	prk	prl	prlj	prs	prt
	prš					
srb	srp	srć	src	srč		
trv	trg	trd	trž	trz	trk	trl
	trlj	trm	trn	trp	trs	trt
	trč	trš				
hrv						
crv	crk	crn	crp	crt		
črn						

Iz ovog pregleda se vidi da su sve niske oblika C_1RC_2 koje se realizuju u srpskohrvatskom jeziku takvog oblika da je obrazac $1C_{12}R$ generisan u tački (d) a da obrazac $1R2C_2$ nije generisan. Dakle, kada se obrazac $1C_{12}R$ primeni na nisku C_1RC_2 dobijamo

$1C_{12}RC_2$

što je i potrebno. Pogledajmo sada kako se na osnovu do sada formiranog rečnika rastavlja reč *zabrljati*.

```
. z a b r l j a t i .
 1z a
    1b2r
        112j
          1j a
              1t i
```

Virtual Library of Faculty of Mathematics - University of Belgrade
elibrary.matf.bg.ac.rs

Na ovaj način dobijamo 1za1br1l2ja1ti, a to znači da se reč rastavlja *za-br-lja-ti*.

(f) Tročlana konsonantska grupa se ne rastavlja ako je uobičajena u srpskohrvatskom jeziku. Kao što smo videli tročlane konsonantske grupe koje su uobičajene u srpskohrvatskom jeziku su oblika:

$$C_1C_2C_3$$

gde je $C_1 \in \{S, Z, Š, Ž\}$, C_2 je proizvoljan konsonant a $C_3 \in \{R, L, Lj, V, J\}$. Na prvi pogled izgleda da bi za primenu ovog pravila bilo potrebno uvesti novih 460 obrazaca. Međutim, analiza pojavljivanja konsonantskih grupa pokazuje da se konsonant C_2 realizuje samo kao eksplozivni konsonant, frikativ *H*, afrikat *C* ili *Č* i sonant *M* ili V. To, pak, znači da se C_1C_2 sravnjuje sa obrascem oblika (d) jer je $C_1 \in \{S, Z, Ž, Š\}$. Međutim, i C_2C_3 se takođe sravnjuje sa obrascem oblika (d) jer $C_2 \notin \{R, L, Lj, J, N, Nj\}$ a C_3 je sonant. Kada se primene oba ova obrasca dobijamo:

$$1C_12C_22C_3$$

što je i potrebno. Pogledajmo kako bi se na osnovu do sada formiranog rečnika obrazaca rastavila reč *presvlačiti*.

```
. p r e s v l a č i t i .
 1p2r
   1r e
       1s2v
         1v2l
           1l a
               1č i
                   1t i
```

Tako se dobija 1p2re1s2v2la1či1ti a to znači da se reč rastavlja *pre-svla-či-ti*.

(g) Tročlana konsonantska grupa koja se ne može rastavit C-CC rastavlja se kao CC-C ako je zadovoljen princip 7. Tročlana konsonantska grupa $C_1C_2C_3$ koja se ne sravnjuje sa obrascem oblika (f), niti se niska C_2C_3 sravnjuje sa obrascem oblika (d) biće rastavljena kao C_1C_2-C_3 jer se C_3V sravnjuje sa obrascem oblika (b). Međutim, kao što smo napomenuli u tački **4.3.1**, ponekad je bolje, čak i kada se C_2C_3 sravnjuje sa obrascem oblika $1C_22C_3$, tročlanu grupu rastaviti kao C_1C_2-C_3, nego kao C_1-C_2C_3, pa dodajemo sledećih 6 obrazaca:

ur2s3n k2s3t k2t3n l2t3n n2t3n or2f3n

Pogledajmo kako se na osnovu do sada formiranog rečnika rastavlja reč *kontaktni*.

```
. k o n t a k t n i .
 1k o
      1t a
          k2t3n
           1t2n
             1n i
```

Tako dobijamo 1kon1tak2t3ni što znači da se reč rastavlja *kon-takt-ni*.

(h) Sufiksi *SK* i *STV* se ne rastavljaju. Niska *sk* se sravnjuje sa obrascem oblika (d) dok se niska *stv* sravnjuje sa niskom oblika (f)

Virtual Library of Faculty of Mathematics - University of Belgrade
elibrary.matf.bg.ac.rs

što znači da se ove konsonantske grupe ne rastavljaju. S druge strane, tročlana ili četvoročlana konsonantska grupa koja sadrži sufiks *sk*, i dakle sravnjuje se sa obrascem 1s2k, nikad se neće moći sravniti sa obrascem oblika (f) koji bi poništio dejstvo ovog obrasca.

(i) Najveći broj četvoročlanih konsonantskih grupa nastaje na spoju sufiksa i osnove. Analiza pojavljivanja konsonantskih grupa na korpusu je pokazala da najveći broj četvoročlanih konsonantskih grupa nastaje na spoju sufiksa *-sk(i)* ili *-stv(o)* sa osnovom. Rastavljanje reč na spoju sufiksa sa osnovom obezbeđuju, dakle, obrasci oblika (d) i (f). Preostale identifikovane konsonantske grupe će biti ispravno rastavljene sravnjivanjem sa obrascem oblika (f), vodeći računa da kod rastavljanja četvoročlanih gupa princip 6 ima prioritet nad principima 7 i 9 (videti **4.3.1**).

(j) Dva ili više konsonanata sa kraja reči nije dopušteno bez vokala preneti u novi redak. Obrasci oblika (b) obezbeđuju da se prilikom rastavljanja reči u novi redak mora preneti bar jedan vokal. Međutim, obrasci oblika (d) i (f), koji obezbeđuju da se konsonantske grupe koje zadovoljavaju princip 9, odnosno 6, ne rastavljaju, omogućili bi prenošenje ovih finalnih konsonantskih grupa bez vokala u novi redak. Analiza pojavljivanja finalnih konsonantskih grupa u korpusu pokazuje da se od svih identifikovanih finalnih grupa samo njih četiri se sravnjuje sa obrascima oblika (e). Stoga je potrebno uvesti još 4 obrasca:

2st. 2sl. 2št. 2kl.

Pomenimo još glas r koji na kraju reči iza konsonanta ima ulogu vokala. U tom slučaju se konsonanti sa kraja reči mogu preneti u novi redak. Takva je, na primer, finalna grupa *kr* reči *masakr*. Ali ova grupa se sravnjuje sa obrascem oblika (e), pa će pomoću do sada formiranog rečnika obrazaca reč biti ispravno rastavljena: *ma-sa-kr*.

Rutina za rastavljanje reči programskog sistema *TeX* reči rastavlja tako da u redu ostaju najmanje dva slova reči a da se u novi redak prenose najmanje tri slova. Stoga u rečnik obrazaca za rastavljanje reči srpskohrvatskog jezika ovi obrasci nisu uvršteni (videti Prilog G).

Pre nego što opišemo obrasce koji definišu rastavljanje reči na spoju prefiksa i osnove, primetimo da je do sada formiran rečnik od 370 obrazaca. Pogledajmo kako se na osnovu ovog rečnika rastavlja reč *zakržljao*.

```
. z a k r ž l j a o .
 1z a
     1k2r
         1ž21 j
           112j
             1j a
                a1o
```

Tako dobijamo 1za1k2r1ž212ja1o, što znači da se reč rastavlja *za-kr-žljao*. (Samo jedno slovo, ovde slovo *o*, se ne prebacuje u sledeći redak.)

Virtual Library of Faculty of Mathematics - University of Belgrade
elibrary.matf.bg.ac.rs

(k) Prefiksi tipa 1. Prefiksne niske tipa 1 u rečima srpskohrvatskog jezika su uvek prefiksi iza kojih se reči rastavljaju. Iz ovog pravila se generiše 36 obrazaca:

.an2ti
.be2š3č .be2š3ć
.iž3dj .iž3dž .iš3č .iš3ć
.iz4van .kon2tra .kva2zi .me2dju .mi2mo .na2dri
.na2zo2vi
.nu2z3 .nu2s3p .nu2s3t .nu2s3k .nu2s3č .nu2s3ć
.nu2s3š .nu2s3f .nu2s3h
.pro2ti2v3l .pro2ti2v3r .pse2u2do
.ra2za .ra2ž3dj .ra2ž3dž .ra2š3č .ra2š3ć
.uš3č .uš3ć
.tran2s3 .ul2tra .ve2le

Za prefikse tipa 1 *nat-*, *op-*, *ot-*, *pot-* i *pret-* koji su alternante odgovarajućih osnovnih prefiksa nije generisan ni jedan obrazac jer rastavljanje reči iza ovih prefiksa regulišu obrasci za rastavljanje konsonantskih grupa.

(l) Prefiksi tipa 2. Reč se iza prefiksne niske tipa 2 rastavlja na isti način kada prefiksna niska jeste, odnosno nije, identifikovana kao prefiks. Prefiksne niske ovog tipa se, dakle, ignorišu pa se za ove prefikse ne generiše ni jedan obrazac.

(m) Prefiksi tipa 3. Reč se iza prefiksne niske tipa 3 obavezno rastavlja ako iza prefiksne niske ne sledi vokal. U protivnom su za identifikaciju prefiksa potrebne dodatne informacije. Za generisanje obrazaca koji sadrže ove dodatne informacije konsultovan je rečnik prefiksa koji je konstruisan nad korpusom (vidi **4.2.3**) i opšti rečnik [Rečnik]. Generisanje obrazaca za prefikse tipa 3 ilustrovaćemo na primeru prefiksa *naj-*. Kao što se vidi u Prilogu **F.1**, u rečnik prefiksa je uvršteno 27 reči sa prefiksnom niskom *naj-*. Od ovih reči, 21 reč sadrži prefiks *naj-*, i to:

najefikasnije	*najizrazitiji*	*najiskorišćenija*
najistaknutiji	*najogorčenije*	*najodgovornije*
najozbiljnije	*najomiljenije*	*najopasniji*
najopremljeniji	*najosnovniji*	*najočuvanija*
najudaljenija	*najuži*	*najužasnijim*
najuobičajenije	*najuočljivije*	*najuređeniji*
najuspešnije	*najuticajnijih*	*najučenijeg*

Šest reči iz rečnika prefiksa koje počinju prefiksnom niskom *naj-* ne sadrži prefiks *naj-*. To su reči:

najavljuju	*najamnika*	*najamnička*
najezda	*naježeni*	*najurio*

U rečniku [Rečnik] se, osim ovih reči, pojavljuje još 13 odrednica koje sadrže prefiksnu nisku *naj-* koja u njima nije prefiks. To su sledeće reči:

najahati	*najam*	*najamnina*
najava	*najaviti*	*najedanput*
najedati	*najediti*	*najednako*
najednom	*najedriti*	*najesti*
najezda		

Virtual Library of Faculty of Mathematics - University of Belgrade
elibrary.matf.bg.ac.rs

Na osnovu ovih informacija, a posebno vodeći računa da većina opisnih (kvalitativnih) prideva ima superlativ, te da se superlativi prideva ne pojavljuju kao odrednice u rečniku, generisani su sledeći obrasci:

```
.na2j3          .na3j4av    .na4j5avet .na3j4amn    .na3j4ah
                .na3j4ez    .na3j4ež   .na3j4ed     .na3j4eo
                .na3j4eš    .na3j4el   .na4j5elem   .na3j4esti
                .na3j4uri   .na3j4amn  .na3j4ama    .na3j4ame
                .na3j4amo   .na3j4amu
```

Pogledajmo kako se na osnovu do sada formiranog rečnika obrazaca rastavlja reč *najurgentnije*:

```
. n a j u r g e n t n i j e .
 1n a
. n a2j3
    1j u
        1g e
            n2t3n
             1t2n
               1n i
                  1j e
```

Na ovaj način dobijamo 1na2j3ur1gen2t3ni1je, što znači da se reč rastavlja *naj-ur-gent-ni-je.* Reč *najavljivačica* se sravnjuje sa sledećim obrascima:

```
. n a j a v l j i v a č i c a .
 1n a
. n a2j3
. n a3j4a v
    1j a
        1v2l
        2v3l j
          1l2j
            1j i
                1v a
                    1č i
                        1c a
```

Na ovaj način dobijamo 1na3j4av3l2ji1va1či1ca, što znači da se reč rastavlja *na-jav-lji-va-či-ca.*

Na sličan način su generisani potrebni obrasci i za ostale prefikse tipa 3.

```
.di2s3          .di3s4a   .di3s4e   .di3s4i   .di3s4o   .di3s4u
.ek2s3          .ek3s4a   .ek3s4e   .ek3s4i   .ek3s4o   .ek3s4u
.iz3            .iz4iš    .iz4uzeta .iz4uzetn .iz4uzetk .iz4e
                .iz4id    .iz4oba   .iz4of    .iz4og    .iz4ol
                .iz4om    .iz4ote   .iz4oto   .iz4uva   .iz4uma
                .iz4ume   .iz4umi   .iz4umo   .iz5umir  .iz4ut
                .iz5umel  .iz5umet  .iz5umeć  .iz5umeh  .iz4umeš
.uz3            .uz4e       .uz4im       .uz4id       .uz4ic
                .uz4orak    .uz4orc      .uz4ork      .uz4orn
                .uz4orit
```

Virtual Library of Faculty of Mathematics - University of Belgrade
elibrary.matf.bg.ac.rs

	.ni2z3br	.ni2z3v		
.pre2d3	.pre3d4a	.pre3d4e	.pre3d4i	.pre3d4o
	.pre4d5ig	.pre4d5isp	.pre4d5ist	.pre4d5izb
	.pre4d5odr	.pre4d5ose		
	.pre3d4ubo	.pre3d4ug	.pre3d4uh	.pre3d4uj
	.pre3d4um	.pre3d4up	.pre3d4usr	.pred4uzeć
	.pre3d4uzimlj		.pre3d4uzetn	
	.pre3d4ja	.pre3d4je	.pre4d5jel	.pre3d4ji
	.pre3d4jo	.pre3d4ju		
.ra2z3	.ra3z4in	.ra3z4ilaza	.ra3z4ilas	.ra3z4on
	.ra3z4oren	.ra3z4ori	.ra3z4udj	.ra3z4um
	.ra3z4uzdan	.raz4oran	.ra3z4orn	

Za identifikovanje prefiksa tipa 3 generisano je 108 obrazaca.

(n) Prefiksi tipa 4. Za prefikse ovog tipa su uvek potrebne dodatne informacije da bi se prefiks identifikovao. Za generisanje obrazaca koji sadrže ove dodatne informacije konsultovan je rečnik prefiksa koji je konstruisan nad korpusom (vidi **4.2.3**) i opšti rečnik [Rečnik]. Generisanje obrazaca za prefikse tipa 4 ilustrovaćemo na primeru prefiksa *preko-*. Kao što se vidi u Prilogu **F.1**, u rečnik prefiksa je uvršteno 6 reči sa prefiksnom niskom *preko-*. Od ovih reči, 3 reči sadrži prefiks *preko-*, i to:

prekomerno *prekomorska* *prekookeanski*

U rečniku [Rečnik] se, osim ovih reči, pojavljuje još 7 odrednica koje sadrže prefiks *preko-*. To su sledeće reči:

prekoatlantski *prekobrojan* *prekonoć* *prekoputa*
prekosutra *prekosutrašnji* *prekovremeni*

Tri reči iz rečnika prefiksa koje počinju prefiksnom niskom *preko-* ne sadrži prefiks *preko-*. To su reči:

prekor *prekoračeni* *prekoračena*

U rečniku [Rečnik] se, osim ovih reči, pojavljuje još 13 odrednica koje sadrže prefiksnu nisku *preko-* koja u njima nije prefiks. To su sledeće reči:

prekomandovati *prekopati* *prekopirati* *prekoran*
prekoriti *prekositi* *prekovati*

Na osnovu ovih informacija generisani su sledeći obrasci:

.pre2kome	.pre2komo	.pre2koo	.pre2koat	
.pre2kobr	.pre2kono	.pre2kopu	.pre2kosu	.pre2kovr

Pogledajmo kako se na osnovu do sada formiranog rečnika obrazaca rastavlja reč *prekomeran*.

```
. p r e k o m e r a n .
 1p2r
. p r e2k o m e
   1r e
      1k o
         1m e
            1r a
```

Virtual Library of Faculty of Mathematics - University of Belgrade
elibrary.matf.bg.ac.rs

Na ovaj način dobijamo 1p2re2ko1me1ran, što znači da se reč rastavlja *preko-me-ran*. Reč *prekomanda* se sravnjuje sa sledećim obrascima:

```
. p r e k o m a n d a .
 1p2r
   1r e
       1k o
           1m a
                1d a
```

Na ovaj način dobijamo 1p2re1ko1man1da, što znači da se reč rastavlja *pre-ko-man-da*.

Na sličan način su generisani potrebni obrasci i za ostale prefikse tipa 4.

.bez4a	.bez5al	.be3zak		
.in2ter	.in3tere	.in3teri		
.iz4a	.iz5an			
.no2vo	.no3vosti	.no3vošć		
.na4dasv				
.po2d4av	.po3davim			
.po2lu	.po3ludeo	.po3ludel	.po3ludeć	.po3ludeh
	.po3ludeš	.po3ludi	.po3ludet	.po3lupa
	.po3lutan	.po3lutar	.po3lutk	.po3luga
	.po3luge	.po3luzi	.po3lugu	.po3lugom
.po2sle	.po3sledic	.po3slednj	.po3sledič	.po3sleni
.su2pers				
.sa2mo	.sa3moć	.sa3mostan	.sa3mota	.sa3motn
	.sa3moti	.sa3movat	.sa3movao	.sa3moval
	.sa3movać	.sa3movah	.sa3movas	.sa3movaš

Za identifikovanje obrazaca tipa 4 generisano je 59 obrazaca.

(o) Prefiksi tipa 5. Prefikse ovog tipa možemo identifikovati ako poslednji konsonant prefiksne niske i konsonanti koji slede formiraju konsonantsku grupu koja je "teška" za izgovor, tj. konsonantsku grupu koja se uvek rastavlja. U svim ostalim slučajevima su za identifikovanje prefiksa potrebne dodatne informacije. Za generisanje obrazaca koji sadrže ove dodatne informacije konsultovan je rečnik prefiksa koji je konstruisan nad korpusom i opšti rečnik [Rečnik]. Generisanje obrazaca za prefikse tipa 5 ilustrovaćemo na primeru prefiksa *bez-*. Kao što se vidi u Prilogu **F.1**, u rečnik prefiksa je uvršteno 14 reči sa prefiksnom niskom *bez-*. Od ovih reči, 12 reči sadrži prefiks *bez-*, i to:

bezbedno	*bezbednije*	*bezbedniji*
bezbednosti	*bezbojna*	*bezbroj*
bezbrižna	*bezbrojna*	*bezvazdušan*
bezvodni	*bezizlaznoj*	*bezobličan*

U rečniku [Rečnik] se, osim ovih reči, pojavljuju još 62 odrednice koje sadrže prefiks *bez-* i koje ovde nećemo navoditi. Dve reči iz rečnika

Virtual Library of Faculty of Mathematics - University of Belgrade
elibrary.matf.bg.ac.rs

prefiksa koje počinju prefiksnom niskom *bez-* ne sadrži prefiks *bez-*. To su reči:

bezemljaši *beznačajna*

U rečniku [Rečnik] se, osim ovih reči, pojavljuje još 5 odrednica koje sadrže prefiksnu nisku *bez-* koja u njima nije prefiks. To su sledeće reči:

bezistan *bezloban* *bezračan*
bezub *bezvučan*

Na osnovu ovih informacija generisani su sledeći obrasci:

.be2z3 .be3z4e .be3z4is .be3z4lo
.be3z4nača .be3z4rač .be3z4ub
.be3z4vu

Pogledajmo kako se na osnovu do sada formiranog rečnika obrazaca rastavlja reč *beznačelan.*

```
. b e z n a č e l a n .
 1b e
. b e2z3
    1z2n
      1n a
          1č e
              1l a
```

Na ovaj način dobijamo 1be2z3na1če1lan, što znači da se reč rastavlja *bez-na-če-lan.* Reč *beznačajnost* se sravnjuje sa sledećim obrascima:

```
. b e z n a č a j n o s t .
 1b e
. b e2z3
. b e3z4n a č a
    1z2n
      1n a
          1č a
               1n o
                   1s2t
```

Na ovaj način dobijamo 1be3z4na1čaj1no1s2t, što znači da se reč rastavlja *be-zna-čaj-nost* (imajući u vidu da se u sledeći red prenose najmanje dva slova).

Na sličan način su generisani potrebni obrasci i za ostale prefikse tipa 5.

.na2d3 .na3d4a .na3d4e .na3d4i .na4d5igr
.na3d4o .na3d4u .na3d4r .na4d5rea
.na4d5red [.na4d5redj] .na3d4voj .na3d4vor
.na3d4ji .na3d4je .na3d4ju .na4d5jun
.na3d4jo .na3d4ža .na3d4židž

.ob3 .ob4e .ob4i .ob5igr .ob5ist
.ob4jek .ob4l .ob5lag .ob5laž
.ob5let .ob5leć .ob5lik .ob5lic
.ob5liz .ob5lep .ob5lit .ob5lij
.ob5lio .ob5lil .ob5liš .ob5lić
.ob5lič .ob5liv .ob4o .ob4r
.ob5rad [.ob5radj] .ob5rasl .ob5rast

Virtual Library of Faculty of Mathematics - University of Belgrade
elibrary.matf.bg.ac.rs

	.ob5rač	.ob5ruč	.ob4uć	.ob4uk
	.ob4uv	.ob4uo	.ob4ul	.ob4uš
	.ob4učen	.ob4us	.ob4uc	.ob4učić
.od3	.od4vaj	.od4voj	.od4e	.od4i
	.od5igr	.od5is	.od4o	.od5oka
	.od5onud	.od5ovud	.od4rž	.od4rv
	.od4rp	.od4rt	.od4reš	.od4rat
	.od4rać	.od4rah	.od4raš	.od4rao
	.od4ral	.od4rasm	.od4rap	.od4uš
	.od4ugo	.od4už	.od4uh	.od4ulj
	.od4ur	.od4uk	.od4ust	.od4uva
	.od4un	.od4ža		
.po2d3	.po3d4a	.po4d5adm	.po3d4vig	.po3d4viz
	.po3d4vor	.po3d4vost	.po3d4e	.po3d4i
	.po4dići	.po4d5il	.po4d5idj	.po4d5iš
	.po3d4o	.po4d5of	.po3d4nev	.po4d5oč
	.po4d5odb	.po3d4rinj	.po3d4rum	.po3d4raž
	.po3d4rem	.po3d4rht	.po3d4rob	.po3d4rp
	.po3d4rš	.po3d4rt	.po3d4ruš	.po3d4ruž
	.po3d4rž	.po3d4ud	.po3d4un	.po3d4ug
	.po3d4ulj	.po3d4už	.po3d4uk	.po3d4uc
	.pod4uzeć	.po3d4uzetn	.pod4uh	.po3d4jo
	.po3d4jo	.po3d4ju	.po4d5jed	.po4d5jes
	.po4d5jez	.po3d4ji		
.pre2k3lan	[.pre2k3lanj]	.pre2k3j		

Za identifikovanje prefiksa tipa 5 generisano je 151 obrazaca.

(p) Prefiksi tipa 6. Prefikse ovog tipa, kao i prefikse tipa 5, možemo identifikovati ako poslednji konsonant prefiksne niske i konsonanti koji slede formiraju konsonantsku grupu koja je "teška" za izgovor, tj. konsonantsku grupu koja se uvek rastavlja. U svim ostalim slučajevima su za identifikovanje prefiksa potrebne dodatne informacije. Za generisanje obrazaca koji sadrže ove dodatne informacije konsultovan je rečnik prefiksa koji je konstruisan nad korpusom i opšti rečnik. Generisanje obrazaca za prefikse ovog tipa ilustrovaćemo na primeru prefiksa *bes-*. Kao što se vidi u Prilogu **F.1**, u rečnik prefiksa je uvršteno 12 reči sa prefiksnom niskom *bes-*. U svim ovim rečima prefiksna niska *bes-* je prefiks:

beskonačnom	*beskraj*	*beskraju*
beskraja	*beskrajan*	*beskrajni*
besplatno	*bespomoćno*	*besposličili*
bespravan	*bespravni*	*besprimerno*

U rečniku se, osim ovih reči, pojavljuje još 36 odrednica koje sadrže prefiks *bes-* i koje ovde nećemo navoditi. Osim toga, u rečniku se pojavljuje i 9 odrednica koje sadrže prefiksnu nisku *bes-* koja u njima nije prefiks. To su sledeće reči:

beskrupulozan	*bespokojan*	*besporan*
bestidan	*bestidnik*	*bestija*
bestijalan	*bestrasan*	*bestseler*

Na osnovu ovih informacija generisani su sledeći obrasci:

Virtual Library of Faculty of Mathematics - University of Belgrade
elibrary.matf.bg.ac.rs

.be2s3p .be3s4pok .be3s4por
.be2s3k .be3s4kru
.be2s3t .be3s4ti .be3s4tras .bes4ts
.be2s3h
.be2s3c
.be2s3f

Pogledajmo kako se na osnovu do sada formiranog rečnika obrazaca rastavlja reč *bestragija.*

. b e s t r a g i j a .
1b e
. b e2s3t
1s2t
1t2r
1r a
1g i
1j a

Na ovaj način dobijamo 1be2s3t2ra1gi1ja, što znači da se reč rastavlja *bes-tra-gi-ja.* Reč *bestrasna* se sravnjuje sa sledećim obrascima:

. b e s t r a s n a .
1b e
. b e2s3t
. b e3s4t r a s
1s2t
1t2r
1r a
1s2n
1n a

Na ovaj način dobijamo 1be3s4t2ra1s2na, što znači da se reč rastavlja *be-stra-sna.*

Na sličan način su generisani potrebni obrasci i za ostale prefikse tipa 6.

.is3p .is3k .is3t .is3f .is3c .is3h
.is4kak .is4koč .is4kač .is4tup
.is4toro .is4tob .is4tod .is4toi
.is4tom .is4top .is4tos .is4tove
.is4tovr .is4krenu

.ra2s3p .ra2s3k .ra2s3t .ra2s3f .ra2s3c .ra2s3h
.ra3s4koš .ra3s4ta .ra2s4tinj .ra2s3tuć
.ra3s4toj .ra3s4te
.ra4s5tanj .ra4s5tap .ra4s5tak .ra4s5teg
.ra4s5teć .ra4s5ter .ra4s5tez .ra4s5tež

.us3p .us3k .us3f .us3c .us3h
.us3talas .us3traj .us3traž .us3trč .us3treb
.us3trep .us3trg .us3trp .us3tuk .us3tum
.us3tv
.us4kla .us4koč .us4kak .us4kog
.us4kot .us4kol .us4koro .us4porav
.us4koć .us4kos .us4pav .us4poren
.us4pori .us4pok

Za identifikovanje prefiksa tipa 6 generisano je 82 obrazaca.

Virtual Library of Faculty of Mathematics - University of Belgrade
elibrary.matf.bg.ac.rs

(r) Kombinacije prefiksa. Kao što je rečeno u tački 3.2.3, za rastavljanje na kraju retka su od značaja one kombinacije dva prefiksa, u kojima su za identikovanje drugog prefiksa potrebne dodatne informacije (tj., drugi prefiks je tipa 3-6). Kako je prilikom konstruisanja rečnika prefiksa vršena identifikacija samo prvog prefiksa, da bi se generisali potrebni obrasci konsultovan je opšti rečnik [Rečnik]. Generisanje obrazaca koji su potrebni za identifikovanje kombinacije dva prefiksa ilustrovaćemo na primeru kombinacije *o-bez-*. U rečniku se nalazi 9 odrednica koje počinju niskom *obez*. Od toga 8 reči sadrži kombinaciju prefiksa *o-bez-*:

obezbediti	*obezglaviti*	*obeznaditi*
obeznaniti	*obezoružati*	*obezumiti*
obezvrediti	*obezobraziti*	

a jedna reč ne sadrži kombinaciju prefiksa:

obezubiti

Na osnovu ovih informacija, i na osnovu obrazaca generisanih u tački (o) (.ob3 i .ob4e), generisani su sledeći obrasci:

.obe2z3b	.obe2z3g	.obe2z3n	.obe2z3o
.obe2z3um	.obe2z3v		

U rečniku su identifikovane i kombinacije *o-bes*, *o-beš*, *ob-uz*, *po-naj*, *po-od*, *po-uz*, *pro-iz*, *za-uz* i *novo-iz* na osnovu kojih su generisani sledeći obrasci:

.obe2s3h	.obe2s3c	.obe2s3k	.obe2s3pr
.obe2š3č			
.obu2z3d			
.poo2d3m	.poo2d3r		
.pou2z3d			
.proi2z3v	.proi2z3n	.proi2s3	
.zau2z3d			
.novoi2z3g			

Poseban problem predstavlja kombinovanje izuzetno plodnih prefiksa *naj-* i *ne-* sa prefiksima tipa 3 - 6. Kako se u opštem rečniku nalazi samo manji broj reči koje sadrže ove prefikse, rečniku obrazaca su prikljućeni samo obrazci koji zadovoljavaju sledeće uslove:

- za prefiks *naj-* su generisani obrasci koji odgovaraju kombinacijama prefiksa *naj-* sa prefiksima koji učestvuju u pridevskoj tvorbi. Pri tom su generisani samo obrasci koji odražavaju pravilo, a ne i izuzetke od pravila. Na primer, za kombinaciju prefiksa *naj-bez-* je generisan obrazac

.najbe2z3

dok obrasci

.najbe3z4e	.najbe3z4is	.najbe3z4lo	.najbe3z4nača
.najbe3z4rač	.najbe3z4ub	.najbe3z4vuč	

nisu uključeni u rečnik obrazaca.

- za prefiks *ne-* su generisani obrasci koji odgovaraju kombinacijama prefiksa *ne-* sa prefiksima koji učestvuju u tvorbi

Virtual Library of Faculty of Mathematics - University of Belgrade
elibrary.matf.bg.ac.rs

prideva, imenica i priloga. I za ovaj prefiks su generisani samo obrasci koji odražavaju pravilo a ne i izuzetke od pravila.

Rečniku obrazaca su dodati sledeći obrasci:

.najbe2z3	.najbe3z4a		
.najbe2s3p	.najbe2s3k	.najbe2s3t	
.najbe2s3č			
.najek2s3t			
.naji2z3	.naji3z4a		
.naji2s3p	.naji2s3k	.naji2s3t	
.najo2b3	.najo3b4a		
.najo2d3	.najo3d4a		
.najpo2d3	.najpo3d4a		
.najpre2d3			
.najra2z3	.najra3z4a		
.najra2s3p	.najra2s3k	.najra2s3t	
.naju2z3	.naju3z4a		
.naju2s3p	.naju2s3k	.naju2s3t	.naju2s3h
.nei2z3	.nei3z4a	.nei3z4e	.nei3z4i
	.nei3z4o	.nei3z4u	
.nei2s3p	.nei2s3k	.nei2s3t	
.neo2b3	.neo3b4a	.neo3b4e	.neo3b4i
	.neo3b4o	.neo3b4u	
.neo2d3	.neo3d4a	.neo3d4e	.neo3d4i
	.neo3d4o	.neo3d4u	
.nepo2d3	.nepo3d4a	.nepo3d4e	.nepo3d4i
	.nepo3d4o	.nepo3d4u	
.nepre2d3			
.nera2z3	.nera3z4u	.nera3z4o	
.nera2s3p	.nera2s3k	.nera2s3t	
.neu2z3	.neu3z4a		
.neu2s3p	.neu2s3k	.neu2s3t	.neu2s3h

Za identifikovanje kombinacija prefiksa generisano je 90 obrazaca. Primena ovako formiranog rečnika obrazaca treba da pokaže da li je potrebno dopuniti rečnik obrascima koji za kombinacije sa prefiksima *naj-* i *ne-* odražavaju odstupanje od uobičajenog.

Rečnik obrazaca za srpskohrvatski jezik, čije generisanje je opisano u ovom odeljku, sadrži 896 obrazaca. Uporedimo veličnu dobijenog rečnika sa sličnim rečnicima generisanim za druge evropske jezike. Tako rečnik obrazaca za engleski jezik koji je generisao Liang ima 4447 obrazaca [Liang83]. Rečnik obrazaca za francuski jezik koji je generisao J. Desarmenien ima 804 obrazaca [Desarm87]. Rečnik obrazaca za poljski jezik je generisala H. Kolodziejska i on ima 2168 obrazaca. Problem je rešavan i za nemački [Appelt85] i švedski jezik [Romberger85]. Primetimo još da za italijanski jezik (obrasce je generisao J. Desarmenien) rečnik ima samo 88 obrazaca dok za nemački jezik, koji je još složeniji od engleskog, rečnik ima oko 7000 obrazaca.

Virtual Library of Faculty of Mathematics - University of Belgrade
elibrary.matf.bg.ac.rs

5.2 Analiza rada rutine za rastavljanje reči

U programski sistem *TeX* ugrađena je rutina za rastavljanje reči koja se zasniva na rečniku obrazaca. Rečnik obrazaca je izdvojen iz programskog sistema i smešten u tekstualnu datoteku HYPHEN.TEX. Sami obrasci su argumenti komande *TeX*-a **\patterns**. U vreme oblikovanja pasusa u retke *TeX*, međutim, ne pretražuje obrasce iz tekstualne datoteke HYPHEN.TEX. Da bi se pretraživanje rečnika obrazaca učinilo dovoljno efikasnim, program INITEX koji se koristi za instaliranje programskog sistema *TeX* učitava rečnik obrazaca i iz njega formira strukturu koja je efikasna za pretraživanje. U procesu inicijalizacije INITEX prvo konstruiše trai (engl. *trie*) [Knuth73] koji nije spakovan u sekvencijski memorijski prostor već je povezan na način koji omogućava lako umetanje novih elemenata. U sledećem koraku se vrši kopresija traia identifikovanjem zajedničkih podtraia. Na kraju se trai na efikasan način pakuje u sekvencijalni oblik koji rutina za rastavljanje reči sistema *TeX* i koristi.

Da bi se proverila valjanost izrađenog rečnika obrazaca za rastavljanje srpskohrvatskog jezika, formirana je nova tekstulana datoteka HYPHEN.TEX (videti Prilog G). U ovu datoteku je uz rečnik obrazaca uvršten i rečnik izuzetaka koji se sastoji od reči koje rutina za rastavljanje reči ne bi ispravno rastavila. Ovaj rečnik se sastoji od 7 reči koje su argumenti komande *TeX*-a **\hyphenation**:

na-jam	no-vost	odra-ste	po-dne
po-dno	uzore	uzori	

U ovom rečniku se nalaze i homografi *odraste*, *uzore* i *uzori* za koje je, kao i u drugim sličnim slučajevima, primenjen sledeći princip: da reč ne bi bila pogrešno rastavljena, na kritičnom mestu reč se ne rastavlja.

Korišćenjem programa INITEX i ovako formirane datoteke HYPHEN.TEX instaliran je programski sistem *TeX*. Komanda *TeX*-a **\showhyphens** je zatim korišćena za proveru valjanost izrađenog rečnika. Kao argumenti komande **\showhyphens** korišćene su sledeće reči:

- sve reči iz rečnika izuzetaka (videti tačku **4.1**);
- sve reči iz rečnika prefiksa (videti tačku **4.2.3**);
- reči koje ilustruju pojavljivanje konsonantskih grupa;
- izbor reči iz [Rečnik] koje ilustruju pojavljivanje određenih prefiksnih niski, prefiksa i kombinacija prefiksa.

U slučajevima koji su zahtevali posebnu pažnju (npr. glagoli *poći* i *naći*) analizirano je rastavljanje svih oblika reči. Ova analiza je ukazala na izvesne nedostatke prvobitno izrađenog rečnika koje su zatim otklonjene a ceo postupak instaliranja *TeX*-a i analiziranja dobijenih rezultata je ponovljen više puta dok nije dobijen rečnik obrazaca i izuzetaka predstavljen u Prilogu G. Tako je, na primer, u prvobitnom rečniku postojao obrazac r2s3n koji je obezbeđivao ispravno rastavljanje reči *kon-kurs-ni*. Kako se, međutim, kombinacija *rsn* pojavljuje i u drugačijem kontekstu, reč *raskrsnice* je bila rastavljena *ras-krs-nice*, što možda nije neispravno ali odstupa od uobičajenog.

Virtual Library of Faculty of Mathematics - University of Belgrade
elibrary.matf.bg.ac.rs

Stoga je obrazac r2s3n zamenjen obrascem ur2s3n a, iz istog razloga, obrazac r2f3n obrascem or3f3n.

Da bi formirani rečnik obrazaca dobio punu potvrdu, svakako je potrebna njegova provera na vezanom tekstu, i to na tekstovima iz raznovrsnih oblasti. Osim toga, postavljeni principi za rastavljanje reči se, takođe, moraju analizirati na širem uzorku a, posebno, "semantički" principi čije definisanje u značajnoj meri zavisi od jezičkog osećaja autora.

5.3 Odnos rutine zasnovane na pravilima i rečnicima i rutine zasnovane na rečniku obrazaca

Rečnik obrazaca za rastavljanje reči srpskohrvatskog jezika opisan u tački **5.1** formiran je na takav način da obezbedi rastavljanje reči u skladu sa pravilima definisanim u tački **4.3.1**. U rutinu za rastavljanje reči ugrađena su ista pravila. Može se, prema tome, očekivati da će se primenom rutine zasnovane na rečniku obrazaca i rutine zasnovane na pravilima i rečnicima postići približno isti rezultati u rastavljanju reči srpskohrvatskog jezika. To potvrđuju i analize rada ovih rutina date u tačkama **4.3.3** i **5.2**.

Međutim, osim po rezultatima rada koji su postignuti za konkretna pravila i konkretan jezik, strategije ugrađene u ove dve rutine mogu se porediti i po drugim kriterijumima. Pomenimo ovde neka od njih.

Opštost. U rutini za rastavljanje reči koja se zasniva na pravilima i rečnicima, svi izuzeci od pravila (ili, bolje rečeno, odstupanja od uobičajenog) moraju se posebno tretirati, tj. zahtevaju dodatno programiranje. Tako u rutini za rastavljanje reči srpskohravtskog jezika probleme transliteracije rešava rutina TL, a identifikaciju prefiksa vrše rutine DALIUR i IMAPF. Kao što je rečeno, ova rutina ne prepoznaje kombinacije prefiksa. Da bi se prepoznavanje kominacija prefiksa uključilo u rad rutine treba intervenisati u programu. S druge strane, u rutini za rastavljanje reči koja se zasniva na rečniku obrazaca svo znanje o jeziku koje je potrebno za rastavljanje reči tog jezika sadržano je u rečniku dok rutina sadrži samo "znanje" o pretraživanju rečnika koji je pretstavljen jednom opštom strukurom podataka (trai). U rečniku obrazaca, na taj način, ne postoji nikakva razlika između obrazaca koji definišu pravila i obrazaca koji odražavaju odstupanja od pravila i rutina za rastavljanje reči ih primenjuje na isti način. Za identifikovanje kombinacija prefiksa ovom rutinom bilo je potrebno samo generisati odgovarajuće obrasce (tačka (r) iz odeljka **5.1**) iz znanja o jeziku.

Prilagodljivost. U vezi sa raspodelom znanja o jeziku između rečnika (odnosno, opšte govoreći, podataka) i programa je i fleksibilnost ovih rutina. Svaka promena pravila za rastavljanje reči se u rutini koja se zasniva na rečniku obrazaca odražava samo u izmeni rečnika obrazaca i nikakva promena pravila ne zahteva izmenu programa. Neke promene pravila bi, međutim, zahtevale intervenciju u rutini zasnovanoj na pravilima i rečnicima. Pogledajmo dva primera. Rečnici obrazaca su generisani pod pretpostavkom da se u ulaznom tekstu slovo *D* srpskohrvatskog alfabeta kodira sa dva slova *DJ*. Međutim, kako to nije uvek slučaj, tj. ovo slovo se u ulaznom tekstu često kodira i zasebnim

Virtual Library of Faculty of Mathematics - University of Belgrade
elibrary.matf.bg.ac.rs

kodom, možemo formirati opštiji rečnik obrazaca koji dopušta da se slovo *D* kodira bilo zasebnim kodom bilo kodom *DJ*. Slična intervencija može se učiniti i u rutini zasnovanoj na pravilima i rečnicima ne menjajući ništa u samom programu. Kao što je rečeno u odeljku **4.2**, treba samo izmeniti ulaznu azbuku, koja se definiše izvan programa, na takav način da se i jedninstven kod *D* i kod *DJ* prevode u kod *DX* interne azbuke programa.

Pravila za rastavljanje reči srpskohrvatskog jezika definišu da se reč rastavlja između dva susedna vokala, ali se reč može rastaviti i posle drugog vokala (vidi odeljak 2.2.2). Na primer, dozvoljeno je i *po-uka* i *pou-ka*. Mogli bismo ovo pravilo postaviti strožije i dozvoliti rastavljanje reči samo između dva susedna vokala. U rutini zasnovanoj na rečniku obrazaca treba samo zameniti obrasce oblika **V1V** obrascima oblika **V1V2**, gde je **V**$\in\{a, e, i, o, u\}$. S druge strane, ovakva promena pravila se u rutinu zasnovanu na pravilima i rečniku ne bi mogla ugraditi bez intervenicije u programu.

Višejezičnost. Pretpostavka ugrađena i u jednu i u drugu rutinu je da je ulazni tekst zapisan na jednom prirodnom jeziku. Ukoliko je ulazni tekst zapisan na dva ili više prirodnih jezika u formateru nam je potrebna rutina za rastavljanje reči koja može da rastavi reči svih jezika koji se u tekstu koriste. Ni jedna ni druga opisana rutina ne nude ovu mogućnost. Međutim, ukoliko formater koristi rutinu koja se zasniva na rečniku obrazaca potrebno je samo ugraditi mogućnost da rutina koristi, prema potrebi, različite rečnike obrazaca. Ukoliko formater koristi rutinu zasnovanu na pravilima i rečnicima u njega je potrebno ugraditi, pored rečnika, i rutine za sve prirodne jezike koji se u ulaznom tekstu mogu koristiti.

Poređenje strategija urgađenih u rutinu zasnovanu na rečniku obrazaca, odnosno, rutinu zasnovanu na pravilima i rečnicima prema kriterijumima opštosti, prilagodljivosti i mogućnosti primene na višejezični tekst ističe prednost rutine zasnovane na rečniku obrazaca. Poređenja po nekim kriterijumima, ipak, pokazuju prednosti strategije ugrađene u rutinu zasnovane na pravilima i rečnicima. Rutina zasnovana na rečniku obrazaca može da reši samo zadatak rastavljanja reči postavljen na način opisan u ovom radu. I najmanja promena u samoj postavci zadatka mogla bi da zahteva suštinsku promenu u koncepciji rutine. Na primer, promena koja bi podeli reči u nekim tačkama dala veći prioritet (npr. iza prefiksa) ne može se ugraditi u rutinu zasnovanu na rečniku obrazaca intervencijom u samom rečniku, već bi takva promena zahtevala suštinsko redefinisanje strategije ugrađene u ovu rutinu. Takva promena se, pak, može ugraditi u rutinu zasnovanu na pravilima i rečniku manjom intervencijom u programu.

Osim toga, rečnik obrazaca formiran za potrebe rastavljanja reči može se koristiti samo za te potrebe. S druge strane, rečnik izuzetaka i rečnik prefiksa koje koristi rutina zasnovana na pravilima i rečnicima mogu, pak, biti deo, ili izvod, jednog opštijeg rečnika koji bi se mogao koristiti i za razne druge svrhe (npr. u programima za otkrivanje i korigovanje grešaka u tekstu). Potrebno je samo da odrednice ovog opšteg rečnika budu snabdevene informacijama koje su potrebne za rastavljanje reči na kraju retka.

Virtual Library of Faculty of Mathematics - University of Belgrade
elibrary.matf.bg.ac.rs

6 PRAVCI DALJEG RADA

Problem rastavljanja reči na kraju retka je jedan od one vrste problema za koje se približno, ili naivno, rešenje može jednostavno formulisati. Međutim rešenje koje bi bilo dovoljno dobro da bi se moglo primeniti u svim domenima uređivanja teksta uz pomoć računara, od poslovne prepiske do slaganja knjiga, zahteva da se problem detaljno prouči, kako sa računarske, tako i sa lingvističke strane. To potvrđuje i raznovrsnost ponuđenih rešenja ovog problema za mnoge evropske jezike.

Ovaj rad pokušava da ponudi jedno takvo rešenje za srpskohrvatski jezik. Kao i za mnoge druge jezike, rastavljanje reči srpskohrvatskog jezika na kraju retka je jednostavan zadatak za čoveka ali pravila najčešće nisu dovoljno precizno definisana da bi se mogla automatski primeniti. Prvenstveni cilj ovog rada je da prouči one kvantitativne osobenosti srpskohrvatskog jezika koje predstavljaju osnovu za definisanje pravila za rastavljanje reči na kraju retka. Kao rezultat nastojanja u postizanju ovog cilja dobijani su rezultati i izgrađen je programski alat koji se dalje može koristiti u rešavanju problema iz oblasti obrade teksta, računarske lingvistike, itd. Pomenimo samo neke od njih.

Za tekstove napisane na srpskohrvatskom jeziku često se javlja potreba da budu štampani korišćenjem i latiničnog i ćiriličnog pisma. Dok štampanje teksta koji je u računar unet korišćenjem ćirilične tastature u latiničnom pismu obično ne predstavlja veći problem, obrnuta situacija je obavezno nalagala naknadno uređivanje teksta koje je, po pravilu, bilo podložno greškama. Izrađeni rečnik izuzetaka, nastao kao rezultat analize pojavljivanja digrafa i odgovarajućih konsonantskih grupa u tekstovima na srpskohrvatskom jeziku omogućava da se ovaj problem na automatski način sa potpunom tačnošću reši.

Kao što je pomenuto u uvodu ovog rada, rastavljanje reči na kraju retka predstavlja samo jednu od mnogih prirodno-jezičkih komponenti koje su ugrađene u većinu uređivača i formatera teksta. Rezultati ovog rada, kao što su frekvencijski rečnici konsonantskih grupa, frekvencijski rečnik funkcionalnih reči, tipologija prefiksa, rečnici izuzetaka i prefiksa predstavljaju osnovu za izradu programa za otkrivanje i korekciju grešaka u tekstu na srpskohrvatskom jeziku, koji bi se zasnivao bilo na ugrađenom rečniku bilo na utvrđenim kvantitativnim zakonitostima jezika [Paterson80].

Rezultati primene segmentacije reči srpskohrvatskog jezika u izradi rečnika izuzetaka i prefiksa koja je predložena u [Vitas85] predstavljaju dobru osnovu za dalje profinjavanje predloženog postupka u cilju daljeg sažimanja rečnika i izbegavanja grešaka.

Izrađeni frekvencijski rečnik funkcionalnih reči predstavlja prvi korak u izradi rečnika najfrekventnijih reči srpskohrvatskog jezika. Takav rečnik predstavlja, s jedne strane, polaznu tačku u izradi (hijerarhijskog) rečnika srpskohrvatskog jezika a, s druge strane, predstavlja nezaobilazno sredstvo u raznim domenima obrade teksta i računarske lingvistike, kao što su program za otkrivanje grešaka, izrada konkordanci, stilska i leksička analiza teksta, dokazivanje autorstva, kritička analiza teksta pa sve do automatskog

Virtual Library of Faculty of Mathematics - University of Belgrade
elibrary.matf.bg.ac.rs

PRILOG -A-

1. Korpus pisanih tekstova nad kojim su izrađeni frekvencijski rečnici konsonantskih grupa

1. Poznavanje društva za IV razred osnovne škole, radna verzija, Zavod za udžbenike i nastavna sredstva, Beograd, 1982, (oko 90 pp)

2. dr Ivan Božić: Istorija za V razred osnoven škole, Zavod za udžbenike i nastavna sredstva, Beograd, 1978, (oko 110 pp.)

3. B. Nikodijević, dr M. Marjanović, M. Latković: Matematika za V razred osnovne škole, 10% uzorak, Zavod za udžbenike i nastavna sredstva, Beograd, 1979, (oko 11 pp.)

4. M. Potkonjak, M. Milošević, T. Rakićević: Geografija sa geografskom čitankom za V razred osnovne škole, Zavod za udžbenike i nastavna sredstva, Beograd, 1978, (oko 150 pp.)

5. dr Ivan Božić: Istorija za IV razred osnovne škole, Zavod za udžbenike i nastavna sredstva, Beograd, 1979, (oko 200 pp.)

6. Momčilo Nastasijević: Sedam lirskih krugova, ciklus Bdenja, Prosveta, 1962, (pesme, oko 20 pp.)

7. Momčilo Nastasijević: Beleška za stvarnu reč, Ibid, (pp. 159-162)

8. Vasko Popa: Sporedno nebo

9. Zakon o udruženom radu, Savremena administracija, 10% uzorak, (oko 80 članova)

10. Zakon o zapošljavanju, Službeni glasnik SRS, 31/77, (prva 24 člana)

2. Korpus pisanih tekstova nad kojim je analizirano korišćenje prefiksa

1. Zakon o penzijskom i invalidskom osiguranju radnih ljudi koji samostalno obavljaju delatnost..., Savremena adminstracija, (156. članova)

2. Poznavanje prirode i društva za III razred - Beograd, Zavod za udžbenike i nastavna sredstva

3. Poznavanje prirode i društva za III razred - Južnomoravski region, Zavod za udžbenike i nastavna sredstva

Virtual Library of Faculty of Mathematics - University of Belgrade
elibrary.matf.bg.ac.rs

3. Korpus pisanih tekstova nad kojim je konstruisan rečnik prefiksa

1. Poznavanje prirode i društva za III razred - Kraljevo, Zavod za udbenike i nastavna sredstva, (10.874 reči)
2. Poznavanje prirode i društva za III razred, Zavod za udbenike i nastavna sredstva, (24.981 reči)
3. Vasko Popa: Sporedno nebo (2.120 reči)
4. Poznavanje prirode i društva za III razred - Beograd, Zavod za udbenike i nastavna sredstva, (12.349 reči)
5. Poznavanje prirode i društva za III razred - Titovo Uzice, Zavod za udbenike i nastavna sredstva, (15.744 reči)
6. Poznavanje prirode i društva za III razred - Šumadija, Zavod za udbenike i nastavna sredstva, (8.752 reči)
7. Poznavanje prirode i društva za III razred - Podrinje, Zavod za udbenike i nastavna sredstva, (12.512 reči)
8. Poznavanje društva za IV razred, Zavod za udbenike i nastavna sredstva, (16.888 reči)
9. Laza Kostić: Santa Maria della Salute, (652 reči)
10. Poznavanje prirode i društva za III razred - Leskovac, Zavod za udbenike i nastavna sredstva, (11.688 reči)
11. Poznavanje prirode i društva za III razred - Niš, Zavod za udbenike i nastavna sredstva, (14.377 reči)
12. Biologija za V razred, Zavod za udžbenike i nastavna sredstva (2.232 reči)
13. dr Ivan Božić: Istorija za V razred osnovne škole, Zavod za udžbenike i nastavna sredstva, Beograd, 1978, (16.812 reči)
14. Rastko Petrović: Otkrovenja, (4.809 reči)
15. Ivo Andrić: Prozori, (1.705 reči)
16. Poznavanje prirode i društva za III razred - Zaječar, Zavod za udbenike i nastavna sredstva, (11.010 reči)
17. M. Potkonjak, M. Milošević, T. Rakićević: Geografija sa geografskom čitankom za V razred osnovne škole, Zavod za udžbenike i nastavna sredstva, Beograd, 1978, (17.434 reči)
18. Žan Pol Sartr: Đavo i gospod bog, (897 reči)
19. Kolin Dej: Obrada tekst, Nolit, 1989, (26.958 reči)

4. Korpus pisanih tekstova nad kojima je analiziran rad rutina za rastavljanje reči srpskohrvatskog jezika na kraju retka

1. Poznavanje prirode i društva za III razred - Podunavlje, Zavod za udbenike i nastavna sredstva, (? reči)

Virtual Library of Faculty of Mathematics - University of Belgrade
elibrary.matf.bg.ac.rs

PRILOG -B-

1 Frekvencijski rečnici konsonantskih grupa - korpus pisanih tekstova

INICIJALNE KONSONANTSKE GRUPE					
Kons. gr	frekven.	u B.2 (da/ne)	Kons. gr	frekven.	u B.2 (da/ne)
1. PR	3389		41. ZL	36	
2. SV	1609		42. ZR	35	
3. ST	951		43. MR	33	
4. DR	759		44. GV	27	
5. SL	738		45. DN	26	
6. KR	703		46. ZGR	25	
7. GR	629		47. PLj	21	
8. PL	488		48. SKR	21	
9. VL	429		49. KV	18	
10. BR	400		50. SHV	18	
11. TR	388		51. DL	17	
12. ŠT	380		52. TL	17	
13. SR	360		53. SPL	16	
14. MN	297		54. KM	15	ne
15. VR	286		55. PS	15	
16. SP	253		56. SJ	15	
17. SM	243		57. HT	15	
18. STR	232		58. PT	14	
19. ZN	227		59. ŠP	14	
20. GL	177		60. TK	13	
21. DV	172		61. ZDR	11	
22. STV	148		62. ČV	10	
23. SN	142		63. HV	9	
24. SK	141		64. ZBL	8	ne
25. ČL	140		65. ZM	8	
26. ZB	134		66. PŠ	8	
27. HR	126		67. CV	8	
28. KL	124		68. ZBR	7	ne
29. ML	95		69. FL	7	
30. GD	88		70. ŠLj	6	
31. SKL	84		71. SF	5	ne
32. FR	81		72. SC	5	ne
33. ZV	78		73. ZG	4	
34. KN	77		74. SH	4	ne
35. BL	70		75. ŠTR	4	ne
36. ŠK	60		76. PJ	3	ne
37. KNj	57		77. PČ	3	
38. TV	56		78. SPLj	3	ne
39. HL	50		79. ŠV	3	
40. SPR	44		80. GM	2	ne

Virtual Library of Faculty of Mathematics - University of Belgrade
elibrary.matf.bg.ac.rs

INICIJALNE KONSONANTSKE GRUPE					
Kons. gr	frekven.	u B.2 (da/ne)	Kons. gr	frekven.	u B.2 (da/ne)
81. DM	2	ne	91. ŽL	1	ne
82. ŽB	2		92. ŽR	1	ne
83. ŽNj	2	ne	93. ZGL	1	ne
84. KLj	2		94. MLj	1	ne
85. FJ	2	ne	95. TM	1	
86. ŠKR	2				
87. BLŠ	1	ne			
88. GN	1				
89. DJ	1	ne			
90. DNj	1	ne			

Dužina	Broj	frekvencija
2	79	14840
3	16	625

FINALNE KONSONANTSKE GRUPE					
Kons. gr	frekven.	u B.2 (da/ne)	Kons. gr	frekven.	u B.2 (da/ne)
1. ST	367		13. KST	1	
2. NT	20		14. LT	1	
3. ND	7		15. MF	1	ne
4. LS	5		16. NK	1	ne
5. KT	5		17. NC	1	ne
6. RS	4	ne	18. PS	1	
7. LF	3		19. RD	1	
8. JŠ	2	ne	20. RKS	1	ne
9. RG	2		21. RH	1	ne
10. RT	2		22. SL	1	ne
11. VS	1	ne	23. ŠT	1	ne
12. JN	1	ne	24. RN	1	ne

Dužina	Broj	frekvencija
2	22	429
3	2	2

Virtual Library of Faculty of Mathematics - University of Belgrade
elibrary.matf.bg.ac.rs

MEDIJALNE KONSONANTSKE GRUPE					
Kons. gr	frekven.	u B.2 (da/ne)	Kons. gr	frekven.	u B.2 (da/ne)
1. ST	3199		43. ŠĆ	207	
2. DN	2322		44. TSK	194	
3. VN	1050		45. DNj	185	
4. MLj	986		46. ŠNj	178	
5. SN	941		47. RB	174	
6. SL	905		48. RT	171	
7. ČK	835		49. STR	169	
8. PR	628		50. STV	166	
9. NSK	591		51. ZL	162	
10. ZV	540		52. BLj	158	
11. ŠT	528		53. SV	155	
12. DR	520		54. ŠK	154	
13. TR	509		55. ND	148	
14. ZN	476		56. SM	148	
15. RG	459		57. PŠT	143	
16. RSK	457		58. ZR	140	
17. SP	448		59. RC	140	
18. BR	408		60. LT	139	
19. GR	388		61. MN	139	
20. TN	371		62. NK	139	
21. VLj	370		63. PL	139	
22. ČN	364		64. KL	138	
23. BL	350		65. MSK	138	
24. SK	330		66. DL	134	
25. NT	328		67. PN	134	
26. NC	322		68. ŠN	134	
27. VR	314		69. DST	131	
28. RN	310		70. VL	126	
29. GL	303		71. NSTV	124	
30. ŠTV	300		72. ZB	123	
31. KV	299		73. VC	119	
32. JSK	279		74. ŠL	118	
33. ŽN	264		75. RSTV	110	
34. ZD	264		76. LjN	109	
35. JV	248		77. KT	103	
36. TV	248		78. FSK	103	
37. LN	247		79. DSTV	102	
38. KR	242		80. DM	100	
39. TK	234		81. TL	100	
40. BN	226		82. PT	97	
41. JN	219		83. DSK	96	
42. ZM	210		84. RM	95	

Virtual Library of Faculty of Mathematics - University of Belgrade
elibrary.matf.bg.ac.rs

MEDIJALNE KONSONANTSKE GRUPE					
Kons. gr	frekven.	u B.2 (da/ne)	Kons. gr	frekven.	u B.2 (da/ne)
85. HV	95		127. ZDV	36	
86. LK	94		128. FR	36	
87. DV	93		129. GD	35	
88. ŠLj	93		130. KŠ	35	
89. ZGR	88		131. LjK	35	
90. PLj	87		132. ML	35	
91. DG	85		133. NS	35	ne
92. PSK	85		134. TLj	35	
93. MP	81		135. VČ	34	
94. JT	78		136. RV	34	
95. TSTV	71		137. MC	33	
96. DB	69		138. PSTV	32	
97. NČ	64		139. JM	31	
98. VSK	63		140. KS	31	
99. KN	62		141. PS	31	
100. BJ	60		142. RKS	31	ne
101. LM	59		143. HT	31	
102. TP	59		144. BSB	30	ne
103. ĆN	59		145. DS	30	
104. RK	56		146. LB	30	
105. GN	55		147. NZ	30	
106. SR	54		148. NTR	30	
107. ŽB	53		149. VD	28	
108. KLj	51		150. ČV	28	
109. MSTV	50	ne	151. ZJ	27	
110. SF	50		152. TM	27	
111. ČJ	49		153. ŽNj	26	
112. DBR	48		154. ZGL	26	
113. LC	48		155. JD	26	
114. RH	48		156. JP	26	
115. NjSK	44		157. RS	26	
116. RD	44		158. MR	25	
117. RJ	42		159. ŽD	24	
118. JPR	41		160. KC	24	
119. JST	41		161. MB	24	
120. SPR	41		162. NGL	23	
121. LjSK	39		163. PK	22	
122. SC	39		164. VK	21	
123. TČ	39	ne	165. DZ	21	
124. BD	38		166. JČ	21	
125. ŽJ	38		167. NG	21	
126. ZG	36				

Virtual Library of Faculty of Mathematics - University of Belgrade
elibrary.matf.bg.ac.rs

MEDIJALNE KONSONANTSKE GRUPE					
Kons. gr	frekven.	u B.2 (da/ne)	Kons. gr	frekven.	u B.2 (da/ne)
168. TNj	21		210. KSPL	10	
169. TST	21		211. SJ	10	
170. LSK	20		212. SKL	10	ne
171. TKR	20		213. SKR	10	
172. RNj	19		214. TJ	10	
173. TH	19		215. VG	9	ne
174. ČNj	19	ne	216. KM	9	
175. GNj	18		217. LD	9	
176. JK	18		218. NLj	9	
177. DJ	17		219. VSTV	8	ne
178. JB	17		220. ŽLj	8	
179. LV	17	ne	221. SPL	8	
180. NL	17	ne	222. DžB	8	
181. PC	17		223. ŠČ	8	
182. HN	17		224. VT	7	
183. JZN	16	ne	225. ŽD	7	
184. NDR	16		226. JSC	7	
185. RP	16		227. LF	7	
186. GM	15		228. MK	7	
187. NTSK	15		229. PĆ	7	
188. RL	15		230. RŽ	7	ne
189. RČ	15		231. DLj	6	
190. JC	14		232. KČ	6	
191. HR	14		233. LjŠ	6	ne
192. ŠTR	14		234. NĐ	6	
193. NKC	13	ne	235. NTN	6	
194. VŠ	12	ne	236. FT	6	
195. GLj	12		237. BZ	5	ne
196. JL	12		238. JML	5	
197. KTR	12		239. KST	5	
198. LP	12	ne	240. LS	5	
199. LjSTV	12		241. PČ	5	
200. MBR	12		242. RSN	5	ne
201. NST	12		243. RŠK	5	ne
202. PH	12	ne	244. SHR	5	
203. PŠ	12		245. TPR	5	
204. RĐ	12		246. HM	5	ne
205. RŠ	12		247. VNj	4	ne
206. ĐSTV	11	ne	248. ZDR	4	
207. ZBR	11		249. JG	4	
208. ĆK	11		250. JZ	4	
209. JSTV	10	ne	251. JKR	4	ne

Virtual Library of Faculty of Mathematics - University of Belgrade
elibrary.matf.bg.ac.rs

MEDIJALNE KONSONANTSKE GRUPE					
Kons. gr	frekven.	u B.2 (da/ne)	Kons. gr	frekven.	u B.2 (da/ne)
252. JS	4		293. NB	2	ne
253. JSL	4	ne	294. NSN	2	ne
254. LG	4		295. PNj	2	
255. MZ	4		296. RKT	2	ne
256. MĆ	4	ne	297. RLj	2	
257. NV	4		298. SHL	2	ne
258. NŽ	4	ne	299. SCV	2	ne
259. RZ	4		300. TKLj	2	
260. RFN	4	ne	301. ĆNj	2	
261. SKLj	4	ne	302. ĆSTV	2	ne
262. TKL	4	ne	303. ĆC	2	ne
263. CK	4		304. FN	2	ne
264. CN	4		305. HL	2	
265. DVL	3		306. ŠC	2	
266. DGR	3	ne	307. VB	1	ne
267. JBL	3	ne	308. VDJ	1	ne
268. JJ	3		309. VJ	1	
269. JR	3		310. GZ	1	ne
270. LjC	3	ne	311. DKR	1	ne
271. MV	3		312. DML	1	ne
272. MNj	3	ne	313. JSN	1	ne
273. MF	3	ne	314. JSP	1	ne
274. NSTR	3		315. JSTR	1	ne
275. NDž	3		316. JTV	1	ne
276. RTSK	3		317. JF	1	ne
277. SH	3		318. JŠ	1	
278. ŠM	3		319. KB	1	ne
279. BM	2	ne	320. KJ	1	
280. BNj	2	ne	321. KĆ	1	ne
281. VĆ	2	ne	322. KH	1	ne
282. DSL	2	ne	323. LZ	1	
283. ZVL	2		324. LPSK	1	ne
284. JPL	2	ne	325. MPL	1	ne
285. JSV	2	ne	326. MS	1	ne
286. KSTV	2	ne	327. MĆ	1	
287. LSTV	2	ne	328. NGR	1	
288. LTSK	2	ne	329. NGC	1	ne
289. LŠ	2	ne	330. NKV	1	ne
290. LDž	2	ne	331. NP	1	
291. LjD	2		332. NSP	1	ne
292. MT	2		333. NĆ	1	ne
			334. NŠK	1	ne

Virtual Library of Faculty of Mathematics - University of Belgrade
elibrary.matf.bg.ac.rs

MEDIJALNE KONSONANTSKE GRUPE					
Kons. gr	frekven.	u B.2 (da/ne)	Kons. gr	frekven.	u B.2 (da/ne)
335. PST	1		344. STJ	1	ne
336. RGL	1	ne	345. SHV	1	ne
337. RTR	1	ne	346. TC	1	ne
338. RF	1	ne	347. CV	1	ne
339. RDŽ	1		348. ČLj	1	ne
340. RŠT	1	ne	349. ČM	1	
341. SD	1	ne	350. ŠČL	1	ne
342. SKV	1				
343. STLj	1	ne			

Dužina	Broj	frekvencija
2	242	29076
3	88	3742
4	20	571

Virtual Library of Faculty of Mathematics - University of Belgrade
elibrary.matf.bg.ac.rs

2 Frekvencijski rečnici konsonantskih grupa - korpus dečjeg govora

INICIJALNE KONSONANTSKE GRUPE					
Kons. gr	frekven.	u B.1 (da/ne)	Kons. gr	frekven.	u B.1 (da/ne)
1. PR	1867		44. DN	16	
2. SM	1776		45. DL	15	
3. SV	1204		46. ŠP	15	
4. DR	1095		47. MR	14	
5. ŠT	787		48. SKL	13	
6. TR	667		49. KV	9	
7. ST	655		50. SVR	9	ne
8. ŠK	651		51. SKR	9	
9. DV	445		52. TK	9	
10. KR	441		53. FR	9	
11. SL	419		54. ZM	8	
12. BR	404		55. PLj	8	
13. ZN	381		56. ČL	8	
14. VR	320		57. FL	7	
15. MN	279		58. ŽV	6	ne
16. GL	274		59. ZG	6	
17. PL	213		60. PŠ	6	
18. GR	197		61. TV	6	
19. SK	174		62. ZDR	5	
20. SPR	165		63. KLj	5	
21. SP	150		64. ŠLj	5	
22. ZV	138		65. KN	4	
23. STR	134		66. SHV	4	
24. HT	128		67. GNj	3	ne
25. GD	124		68. ZR	3	
26. PS	119		69. SPL	3	
27. SR	111		70. CR	3	ne
28. KNj	109		71. ŠĆ	3	ne
29. HR	85		72. BJ	2	ne
30. SN	80		73. ŽB	2	
31. STV	77		74. KĆ	2	ne
32. ML	62		75. SMR	2	ne
33. BL	55		76. ČV	2	
34. KL	53		77. ŠV	2	
35. ŽM	47	ne	78. ŠKR	2	
36. CV	44		79. GV	1	
37. ZB	38		80. GN	1	
38. HL	29		81. PČ	1	
39. ZL	26		82. SJ	1	
40. VL	25		83. SML	1	ne
41. ZGR	23		84. STJ	1	ne
42. HV	20		85. TL	1	
43. PT	17		86. TM	1	
			87. ŠR	1	ne

Dužina	Broj	frekvencija
2	73	13890
3	14	448

Virtual Library of Faculty of Mathematics - University of Belgrade
elibrary.matf.bg.ac.rs

FINALNE KONSONANTSKE GRUPE					
Kons. gr	frekven.	u B.1 (da/ne)	Kons. gr	frekven.	u B.1 (da/ne)
1. ST	200		12. RD	3	
2. LM	44	ne	13. LT	2	
3. NT	24		14. LF	2	
4. RT	10		15. RG	2	
5. RK	9	ne	16. TD	2	ne
6. KST	8		17. LS	1	
7. KS	7	ne	18. NG	1	ne
8. KL	4	ne	19. ND	1	
9. JZ	3	ne	20. RŠ	1	ne
10. KT	3		21. FT	1	ne
11. PS	3				

Dužina	Broj	frekvencija
2	20	323
3	1	8

Virtual Library of Faculty of Mathematics - University of Belgrade
elibrary.matf.bg.ac.rs

MEDIJALNE KONSONANTSKE GRUPE					
Kons. gr	frekven.	u **B.1** (da/ne)	Kons. gr	frekven.	u **B.1** (da/ne)
1. ST	1574		43. RK	113	
2. GR	1206		44. VC	110	
3. DN	1169		45. JK	108	
4. SL	1125		46. NK	101	
5. ND	873		47. RN	98	
6. ŠL	841		48. LN	96	
7. ŠT	734		49. KR	94	
8. BR	423		50. GD	92	
9. RT	417		51. SKR	89	
10. VN	377		52. BL	86	
11. JV	358		53. VD	84	
12. TR	341		54. DR	83	
13. PR	322		55. RSK	79	
14. ČK	295		56. SV	71	
15. ZN	283		57. NC	69	
16. KL	263		58. DL	69	
17. SP	261		59. SPR	68	
18. SM	236		60. ZGL	67	
19. GL	232		61. MB	67	
20. STR	214		62. ZG	66	
21. TK	208		63. KT	63	
22. DM	204		64. TSK	63	
23. SN	203		65. NT	61	
24. ZR	195		66. ZL	59	
25. MLj	192		67. ZB	58	
26. ZM	185		68. MP	58	
27. ŠK	176		69. RC	58	
28. JB	176		70. HV	58	
29. DB	166		71. RB	57	
30. MC	152		72. ŽB	56	
31. TN	149		73. PŠT	55	
32. ČN	148		74. ŠĆ	55	
33. TV	147		75. ŽN	51	
34. GN	134		76. ŠN	51	
35. JČ	132		77. ŠTV	50	
36. JN	129		78. PL	47	
37. PSK	128		79. ŽD	46	
38. PT	125		80. VR	45	
39. DV	124		81. NSK	45	
40. SK	123		82. KV	44	
41. ZD	122		83. PŠ	42	
42. ZV	117		84. MN	40	

Virtual Library of Faculty of Mathematics - University of Belgrade
elibrary.matf.bg.ac.rs

MEDIJALNE KONSONANTSKE GRUPE					
Kons. gr	frekven.	u B.1 (da/ne)	Kons. gr	frekven.	u B.1 (da/ne)
85. ĆN	39		127. LjK	18	
86. RG	38		128. MR	18	
87. RD	38		129. GNj	17	
88. ŠLj	37		130. DG	17	
89. JL	33		131. JC	17	
90. ML	31		132. KSK	17	ne
91. NČ	31		133. LjN	17	
92. TP	31		134. NSTV	17	
93. KM	30		135. RLj	17	
94. TNj	30		136. NG	16	
95. LSK	29		137. NZ	16	
96. MSK	29		138. NTR	16	
97. MBR	28		139. RZ	16	
98. TJ	28		140. VČ	15	
99. BJ	27		141. KS	15	
100. BN	26		142. KC	15	
101. DNj	26		143. NTN	15	
102. TL	26		144. PN	15	
103. JM	25		145. RŠ	15	
104. ŠNj	25		146. ŽLj	14	
105. VL	24		147. LK	14	
106. VLj	24		148. MŠ	14	ne
107. RM	24		149. DBR	13	
108. SC	24		150. LjD	13	
109. VK	23		151. NJ	12	ne
110. ZJ	22		152. PC	11	
111. LM	22		153. RS	11	
112. RD	22		154. ĆK	11	
113. TST	22		155. HN	11	
114. DJ	21		156. VSK	10	
115. DS	21		157. ŽNj	10	
116. ZDR	21		158. ZDV	10	
117. JD	21		159. KLj	10	
118. KŠ	21		160. PK	10	
119. BLj	20		161. GM	9	
120. ŽJ	19		162. ZGR	9	
121. JT	19		163. LS	9	
122. MD	19	ne	164. DST	8	
123. PLj	19		165. JSK	8	
124. STV	19		166. KST	8	
125. KN	18		167. NP	8	
126. LT	18		168. RČ	8	

Virtual Library of Faculty of Mathematics - University of Belgrade
elibrary.matf.bg.ac.rs

MEDIJALNE KONSONANTSKE GRUPE					
Kons. gr	frekven.	u B.1 (da/ne)	Kons. gr	frekven.	u B.1 (da/ne)
169. FR	8		211. JR	3	
170. ŽV	7	ne	212. LjSK	3	
171. JPR	7		213. MT	3	
172. MK	7		214. MČ	3	
173. RV	7		215. NDž	3	
174. RSTV	7		216. PKL	3	ne
175. ĆNj	7		217. PNj	3	
176. KČ	6		218. PČ	3	
177. NSTR	6		219. RJ	3	
178. PS	6		220. RPL	3	ne
179. RNj	6		221. RDž	3	
180. SPL	6		222. SVL	3	ne
181. SR	6		223. SJ	3	
182. TLj	6		224. ŠTR	3	
183. TPR	6		225. ŠČ	3	
184. HR	6		226. DVL	2	
185. CK	6		227. DSTV	2	
186. ŠM	6		228. ZLj	2	ne
187. JZ	5		229. JDR	2	ne
188. JST	5		230. JP	2	
189. NV	5		231. JH	2	ne
190. NGL	5		232. KNj	2	ne
191. NGT	5	ne	233. KTR	2	
192. NĐ	5		234. LB	2	
193. NjSK	5		235. LD	2	
194. SF	5		236. LTN	2	ne
195. ČJ	5		237. LC	2	
196. BD	4		238. LČ	2	ne
197. VT	4		239. LjSTV	2	
198. DSK	4		240. MV	2	
199. JG	4		241. MZ	2	
200. LZ	4		242. NBR	2	ne
201. NDR	4		243. NGR	2	
202. TSTV	4		244. NST	2	
203. TH	4		245. NF	2	ne
204. HL	4		246. NH	2	ne
205. CN	4		247. NjSTV	2	ne
206. ŠP	4	ne	248. PSTV	2	
207. VJ	3		249. RL	2	
208. DLj	3		250. RP	2	
209. ZBR	3		251. RTSK	2	
210. JMLj	3	ne	252. TKR	2	

Virtual Library of Faculty of Mathematics - University of Belgrade
elibrary.matf.bg.ac.rs

MEDIJALNE KONSONANTSKE GRUPE

Kons. gr	frekven.	u B.1 (da/ne)	Kons. gr	frekven.	u B.1 (da/ne)
253. HT	2		280. LKL	1	ne
254. ČM	2		281. LMSK	1	ne
255. DžB	2		282. LNj	1	ne
256. ŠC	2		283. LF	1	
257. VGR	1	ne	284. MJ	1	ne
258. GLj	1		285. NDV	1	ne
259. DZ	1		286. NDJ	1	ne
260. DK	1	ne	287. NLj	1	
261. DSP	1	ne	288. NTSK	1	
262. DŠK	1	ne	289. NŠP	1	ne
263. DŠT	1	ne	290. NjC	1	ne
264. ŽĐ	1		291. PĆ	1	
265. ZVL	1		292. RH	1	
266. ZGN	1	ne	293. SKV	1	
267. ZT	1	ne	294. SH	1	
268. JBN	1	ne	295. SHR	1	
269. JJ	1		296. TKV	1	ne
270. JML	1		297. TKLj	1	
271. JS	1		298. TM	1	
272. JSC	1		299. FSK	1	
273. JHR	1	ne	300. FT	1	
274. JŠ	1		301. CM	1	ne
275. KJ	1		302. CNj	1	ne
276. KSP	1	ne	303. ČV	1	
277. KSPL	1		304. ČT	1	ne
278. KTN	1	ne			
279. LG	1				

Dužina	Broj	frekvencija
2	221	18599
3	71	1197
4	12	47

Virtual Library of Faculty of Mathematics - University of Belgrade
elibrary.matf.bg.ac.rs

PRILOG -C-

1 Lista prefiksa i njihovih alternacija

Prefiks	Uslov za alternaciju	Prefiks	Uslov za alternaciju
1. ANTI		41. OD	
2. ARHI		42. OT	*P K Č Ć F C H*
3. BEZ		43. ODA	
4. BES	*P K T F C H*	44. PA	
5. BE	Z S Ž Š	45. PO	
6. BEŠ	*Č Ć*	46. POD	
7. BEZA		47. POT	*P K Č Ć F C H*
8. DE		48. PODA	
9. DIS		49. POLU	
10. DO		50. POSLE	
11. EKS		51. PRA	
12. EKSTRA		52. PRE	
13. IZ		53. PRED	
14. IS	*P T K F C H*	54. PRET	*P K Č Ć F C H*
15. IŽ	*D Dž*	55. PREKO	
16. IŠ	*Č Ć*	56. PREK	
17. IZA		57. PRI	
18. IZVAN		58. PRO	
19. INTER		59. PROTIV	
20. KONTRA		60. PROTU	
21. KVAZI		61. PSEUDO	
22. MEĐU		62. RAZ	
23. MIMO		63. RAS	*P T K F C H*
24. NA		64. RAZA	
25. NAD		65. RA	
26. NAT	*P K Č Ć F C H*	66. RAŽ	*D Dž*
27. NADRI		67. RAŠ	*Č Ć*
28. NAJ		68. SA	
29. NAZOVI		69. SU	
30. NE		70. SUPER	
31. NI		71. SAMO	
32. NIZ		72. TRANS	
33. NIS	*P T K Č Ć Š S F C H*	73. UZ	
		74. US	*P T K F C H*
34. NOVO		75. UZA	
35. NADA		76. UŠ	*Č Ć*
36. NUZ		77. ULTRA	
37. NUS	*P T K Č Ć Š S F C H*	78. VELE	
		79. ZA	
38. OB			
39. OP	*T K Č Ć Š S F C H*		
40. OBA			

Virtual Library of Faculty of Mathematics - University of Belgrade
elibrary.matf.bg.ac.rs

2 Rezultati analize pojavljivanja prefiksa u korpusu (Prilog A.2)

Prefiks	1	2	3	4	5	6	7	8
1. ANTI	4	1	da		4		*a-*	
2. ARHI	4	4	da		2		*a-*	0
3. BEZ	3	5					5	
4. BES	3	6		b			5	
5. BE	2	2		d				
6. BEŠ	3	1		c			5	
7. BEZA	4	4	da	e			5	
8. DE	2	2			2			
9. DIS	3	3			3	2/10		
10. DO	2	2			3	102/136		
11. EKS	3	3			3	1/2		
12. EKSTRA	6	4	da		1		11	
13. IZ	2	3			3	320/328	*i-*	0
14. IS	2	6		b	3	133/206	*i-*	1/206
15. IŽ	2	1		c	1		*i-*	
16. IŠ	2	1		c	4		*i-*	
17. IZA	3	4	da	e	3	8/13	*i-*	0
18. IZVAN	5	1	da		1		13	
							i-	
19. INTER	5	4	da		2		*i-*	0
20. KONTRA	6	1	da		1			
21. KVAZI	5	1	da		1			
22. MEĐU	4	1	da		4			
23. MIMO	4	1	da		1			
24. NA	2	2			3	94/185		
25. NAD	3	5			3	39/44	24	3/44
26. NAT	3	1		b	4		24	
27. NADRI	5	1	da		1		24	
28. NAJ	3	3			4		24	
29. NAZOVI	6	1	da		1		24	
30. NE	2	2			3	93/113		
31. NI	2	2			2			
32. NIZ	3	3			2		31	0
33. NIS	3	3		b	2		31	0
34. NOVO	4	4	da		2			
35. NADA	4	4	da		1			
36. NUZ	3	1			1			
37. NUS	3	1		b	1			
38. OB	2	5			3	98/266	*o-*	79/266
39. OP	2	1		a	3	8/83	*o-*	0
40. OBA	3	2	da	e	3	31/185	38	0
							o-	136/185

Virtual Library of Faculty of Mathematics - University of Belgrade
elibrary.matf.bg.ac.rs

Prefiks	1	2	3	4	5	6	7	8
41. OD	2	5			3	329/541	*o-*	22/541
42. OT	2	1		b	4		*o-*	
43. ODA	3	2	da	e	3	8/9	41	0
							o-	0
44. PA	2	2			2			
45. PO	2	2			3	327/478		
46. POD	3	5			3	50/126	45	71/126
47. POT	3	1		b	4		45	
48. PODA	4	4	da	e	2		45	0
							46	0
49. POLU	4	4	da		4		46	
50. POSLE	5	4	da		3	2/16	45	0
51. PRA	3	2			2			
52. PRE	2	2		d*	4			
53. PRED	4	3			3	41/122	52	11/122
54. PRET	4	1		b	4		52	
55. PREKO	5	4	da		4		52	
							56	
56. PREK	4	5			2		52	0
57. PRI	3	2			3	69/99		
58. PRO	3	2			3	25/100		
59. PROTIV	6	1	da		4		58	
60. PROTU	5	4	da		1		58	
61. PSEUDO	6	1	da		1			
62. RAZ	3	3			3	169/215	65	0
63. RAS	3	6		b	3	21/68	65	20/68
64. RAZA	4	1	da	e	4		62	
							65	
65. RA	2	2		d	3			
66. RAŽ	3	1		c	1		65	
67. RAŠ	3	1		c	4		65	
68. SA	2	2		e	3	26/53	*s-*	0
69. SU	2	2			3			
70. SUPER	5	4	da		1		69	
71. SAMO	4	4	da		3	106/109	68	0
72. TRANS	5	1			4			
73. UZ	2	3			3	17/34	*u-*	1/34
74. US	2	6		b	3	5/118	*u-*	14/118
75. UZA	3	2	da	e	3	2/7	73	0
							u-	0
76. UŠ	3	1		c	2		*u-*	0
77. ULTRA	5	1	da		4		*u-*	
78. VELE	4	1	da		4			
79. ZA	2	2			3	149/158		

Virtual Library of Faculty of Mathematics - University of Belgrade
elibrary.matf.bg.ac.rs

PRILOG -D-

FREKVENCIJSKI REČNIK FUNKCIONALNIH REČI

Pomoćni glagoli

uzorak = 44311 ukupno = 2588

Reč	frekv.	Reč	frekv.	Reč	frekv.
1 JE	1121	27 JESTE	5	53 BEŠE	0
2 SU	587	28 TREBALO	5	54 BESMO	0
3 IMA	106	29 ĆEŠ	4	55 BESTE	0
4 SI	97	30 BUDI	3	56 BEHU	0
5 BI	79	31 JESI	3	57 BUDEM	0
6 BIO	79	32 NEĆEMO	3	58 BUDEMO	0
7 NIJE	58	33 ĆETE	3	59 BUDETE	0
8 TREBA	52	34 BIĆE	2	60 BUDIMO	0
9 ĆE	47	35 BUDITE	2	61 IMATE	0
10 BILA	41	36 JESU	2	62 IMAM	0
11 IMAJU	39	37 NISI	2	63 IMAŠ	0
12 SMO	36	38 NISMO	2	64 JESAM	0
13 BITI	30	39 BISTE	1	65 JESMO	0
14 NISU	29	40 BUDEŠ	1	66 NEĆU	0
15 NEMA	28	41 NISAM	1	67 NEĆEŠ	0
16 BILI	25	42 NISTE	1	68 NEĆETE	0
17 BUDE	17	43 ĆU	1	69 NEMAM	0
18 STE	17	44 HOĆE	1	70 NEMAMO	0
19 NEĆE	9	45 BIH	0	71 NEMAŠ	0
20 SAM	9	46 BIĆEŠ	0	72 NEMATE	0
21 BISMO	8	47 BIĆEMO	0	73 NEMAJU	0
22 BILE	8	48 BIĆETE	0	74 HOĆU	0
23 ĆEMO	7	49 BIĆU	0	75 HOĆEŠ	0
24 BUDU	6	50 BEJAH	0	76 HOĆEMO	0
25 IMATI	6	51 BEJAHU	0	77 HOĆETE	0
26 IMAMO	5	52 BEJASMO	0		

Virtual Library of Faculty of Mathematics - University of Belgrade
elibrary.matf.bg.ac.rs

veznici, prilozi, brojevi, rečce, uzvici, upitne reči

uzorak = 44311 ukupno = 5139

Reč	frekv.	Reč	frekv.	Reč	frekv.
1 I	2182	48 KO	11	95 VALjDA	0
2 DA	611	49 VRLO	10	96 DVOJE	0
3 ILI	269	50 KOLIKO	9	97 DVOJICA	0
4 A	210	51 PET	9	98 DVOJI	0
5 NE	178	52 HILjADA	9	99 DVOJA	0
6 KAO	145	53 TAMO	8	100 DEVET	0
7 KADA	78	54 MOŽDA	7	101 DESETI	0
8 SAMO	67	55 OBA	7	102 DRUGDE	0
9 ŠTO	67	56 ONDA	7	103 DOVDE	0
10 POSLE	65	57 STO	7	104 DAKAKO	0
11 LI	64	58 ČETIRI	7	105 DAŠTA	0
12 ŠTA	61	59 ŠEST	7	106 ZAISTA	0
13 ALI	57	60 LEPO	6	107 JE	0
14 ZBOG	56	61 SVUDA	6	108 JAKO	0
15 TOGA	54	62 EVO	5	109 KAMO	0
16 GDE	53	63 JEDNI	5	110 KOJEKUDA	0
17 KAKO	47	64 KUDA	5	111 KUKU	0
18 DRUGE	44	65 GORE	4	112 LELE	0
19 DRUGA	42	66 ETO	4	113 MILIJARDA	0
20 JOŠ	38	67 MALO	3	114 MAKAR	0
21 MNOGO	37	68 NITI	3	115 NIPOŠTO	0
22 JEDAN	36	69 OSAM	3	116 ONOMAD	0
23 PRE	36	70 OBE	3	117 ONAMO	0
24 TU	35	71 BAŠ	2	118 OVAMO	0
25 ZATO	32	72 DEVETI	2	119 ODANDE	0
26 NEKOLIKO	29	73 DESET	2	120 OJ	0
27 JER	27	74 NEKAD	2	121 OHO	0
28 SVAKAKO	25	75 OSMI	2	122 PETORO	0
29 ZAŠTO	23	76 SEDAM	2	123 PETORICA	0
30 TRI	23	77 STOTINA	2	124 POZADI	0
31 DVA	21	78 STOGA	2	126 SEDMI	0
32 MEDUTIM	21	79 SASVIM	2	127 SVAKAD	0
33 DRUGI	20	80 TREĆI	2	128 SILNO	0
34 TADA	18	81 DOLE	1	129 TREĆE	0
35 JEDNA	17	82 ENO	1	130 TROJICA	0
36 NEŠTO	17	83 ZAR	1	131 TROJI	0
37 DRUGO	16	84 MILION	1	132 TROJA	0
38 DOBRO	16	85 NAPRED	1	133 TISUĆA	0
39 DVE	15	86 NAZAD	1	134 TUDA	0
40 NI	15	87 NO	1	135 UH	0
41 POTPUNO	15	88 ODAVDE	1	136 ČETVRTI	0
42 TE	15	89 PETI	1	137 ČETVORICA	0
43 PRVI	14	90 TREĆA	1	138 ČETVORI	0
44 DOSTA	13	91 TROJE	1	139 ČETVORA	0
45 SADA	13	92 ČETVORO	1	140 ŠESTI	0
46 JEDNE	12	93 AH	0		
47 JEDNO	11	94 BUDUĆI	0		

Virtual Library of Faculty of Mathematics - University of Belgrade
elibrary.matf.bg.ac.rs

zamenice

uzorak = 44311 ukupno = 4040

Reč	frekv.	Reč	frekv.	Reč	frekv.
1 SE *	1168	48 MI *	17	95 KOLIK *	5
2 KOJ *	656	49 KAKV *	17	96 JA	4
3 TO	188	50 ČIJ	16	97 JOJ	4
4 NAŠE	105	51 NjOJ	15	98 OVU	4
5 OVOG	99	52 OVIH *	15	99 SVAKA	4
6 SVE	95	53 TVOJ	15	100 SVOJOJ	4
7 SVOJE *	52	54 NjEMU	14	101 TVOGA *	4
8 NjEG	50	55 SVAKI *	14	102 NEKOJ	4
9 SVI	44	56 MOJ	14	103 ME	3
10 IH	43	57 NAŠOJ	13	104 NjOM	3
11 ONA	43	58 NEKA	13	105 NjU	3
12 ONI *	41	59 OVO	13	106 ONIH	3
13 NjIHOV *	37	60 OVOM *	13	107 SVAKIM	3
14 NjIH	36	61 OVAK *	13	108 SVAKU	3
15 ON	36	62 TOG	13	109 ČIM	3
16 SVOJIM	33	63 NAŠU	12	110 VAŠA	2
17 MU	31	64 NEKI	12	111 SVAKIH	2
18 NAŠEM	30	65 SVAKOG *	12	112 SEBI	2
19 OVE	30	66 TVOM	12	113 TAKAV *	2
20 KOME	29	67 IM	11	114 VAŠI *	2
21 GA	28	68 NEKIM	11	115 NIKAK *	2
22 ONE	28	69 SVOGA	11	116 TEB	2
23 ONO	28	70 SVOJIH	11	117 VAMA	1
24 SVOJ *	28	71 NAM	10	118 VAŠ	1
25 NjIM *	28	72 NEKO	10	119 GOD	1
26 TAKV	28	73 NEKU	10	120 MOGA	1
27 KOGA	25	74 SVAKE	10	121 NEKOGA	1
28 TOME	25	75 SVAKO	10	122 NIKO	1
29 NAŠA	24	76 SEBE	10	123 ONAJ	1
30 NAŠIH	24	77 OVOJ	9	124 ONOJ	1
31 OVAJ	24	78 SVA	9	125 ONU	1
32 NAŠ	23	79 ČEMU	8	126 SVOME	1
33 OVIM	23	80 OVI	7	127 SOBOM	1
34 TAJ	23	81 SVAKOM	7	128 TVOG *	1
35 NAS	22	82 SVOJOM	7	129 ONOLIK *	1
36 NAŠEG	22	83 TOJ *	7	130 SVAČ	1
37 TIH *	22	84 PONEK *	7	131 VASKOLIK	0
38 NjEN *	22	85 VAS	6	132 VAŠU	0
39 NAŠIM	21	86 VI	6	133 GDEKO	0
40 NEKE	21	87 NjE *	6	134 GDEŠTA	0
41 OVA	21	88 VAŠE *	6	135 IKO	0
42 SVOJU	21	89 VAM	5	136 IKOGA	0
43 TI	21	90 NAMA	5	137 IKOME	0
44 TIM	21	91 NEKIH	5	138 IČEGA	0
45 TOM	20	92 NEKOG	5	139 IČEMU	0
46 NAŠI	19	93 NEKOM	5	140 IŠTA	0
47 TA	18	94 SVOJA	5	141 JE	0

Virtual Library of Faculty of Mathematics - University of Belgrade
elibrary.matf.bg.ac.rs

zamenice					
Reč	frekv.	Reč	frekv.	Reč	frekv.
142 KAKAV	0	157 Nj	0	172 TE	0
143 KIM	0	158 NjEZIN	0	173 TOBOM	0
144 KO	0	159 ONIM	0	174 TU	0
145 KOGOD	0	160 ONOG	0	175 ŠTA *	0
146 MA	0	161 ONOM	0	176 VAŠO*	0
147 MENE	0	162 SAV	0	177 IKAK*	0
148 MENI	0	163 SVAKOGA	0	178 IKOJ *	0
149 MNOM	0	164 SVAKOJ	0	179 IKOLIK *	0
150 MOG	0	165 SVAKOME	0	180 NEKOLIK*	0
151 MOM	0	166 SVAŠTA	0	181 NEČIJ*	0
152 MOME	0	167 SVO	0	182 NIKOJ*	0
153 NAŠOM	0	168 SVOJEGA	0	183 NIČIJ *	0
154 NEKOME	0	169 SVOJEMU	0	184 OVOLIK*	0
155 NI	0	170 SVOJI	0	185 SVAKAK *	0
156 NIKOGA	0	171 SI	0	186 SVAKOLIK	0

* Pod ovom niskom su grupisane sve zamenice koje počinju tom niskom. Na primer, pod KOJ su grupisane zamenice:
KOJI, KOJEM, KOJEG, KOJIM, KOJIH, KOJE, KOJA, KOJOJ, KOJU, KOJOM i KOJOJ

Virtual Library of Faculty of Mathematics - University of Belgrade
elibrary.matf.bg.ac.rs

predlozi

uzorak = 44311 ukupno = 4859

Reč	frekv.	Reč	frekv.	Reč	frekv.
1 U	1364	32 PRI	12	63 NASRED	0
2 NA	777	33 NAKON	10	64 NASPRAM	0
3 ZA	530	34 BLIZU	9	65 NASPRAMA	0
4 OD	407	35 IME	9	66 NAMESTO	0
5 SA	239	36 MIMO	7	67 NIZA	0
6 IZ	209	37 PRED	7	68 NIŽE	0
7 DO	159	38 UMESTO	6	69 OKROM	0
8 PO	146	39 ISPRED	5	70 PORADI	0
9 O	143	40 IZA	3	71 POKRAJ	0
10 PREMA	123	41 NIZ	3	72 POVRH	0
11 KOD	62	42 VRH	2	73 PODNO	0
12 VIŠE	61	43 NAD	2	74 POSLE	0
13 BEZ	56	44 USLED	2	75 POSRED	0
14 POD	44	45 ČELO	2	76 POPUT	0
15 S	44	46 DUŽ	1	77 PRE	0
16 KROZ	43	47 IZVAN	1	78 RAZMA	0
17 IZMEĐU	37	48 KA	1	79 SAVRH	0
18 PREKO	35	49 NASUPROT	1	80 SADNO	0
19 PROTIV	33	50 OKOLO	1	81 SVRH	0
20 OKO	32	51 SPREMA	1	82 SRED	0
21 MESTO	31	52 UPRKOS	1	83 SPRAM	0
22 PUT	31	53 DNO	0	84 SPREM	0
23 KRAJ	26	54 ZBOG	0	85 SUPROT	0
24 RADI	26	55 ZARADI	0	86 SUSRET	0
25 OSIM	23	56 ISKRAJ	0	87 UPRKOS	0
26 PORED	23	57 KROM	0	88 USRED	0
27 UZ	18	58 K	0	89 UKRAJ	0
28 IZNAD	14	59 NAKRAJ	0	90 UZDUŽ	0
29 ISPOD	13	60 NAVRH	0	91 UZA	0
30 VAN	12	61 NADNO	0	92 UVRH	0
31 MEĐU	12	62 NADOMAK	0	93 UDNO	0

Virtual Library of Faculty of Mathematics - University of Belgrade
elibrary.matf.bg.ac.rs

PRILOG -E-

LISTA OBLIČNIH NASTAVAKA

1. A
2. IMA
3. AMA
4. OGA
5. EGA
6. U
7. AMU
8. EMU
9. JU
10. IJU
11. AJU
12. AHU
13. I
14. TI
15. ĆI
16. UĆI
17. EĆI
18. VŠI
19. O
20. EMO
21. AMO
22. IMO
23. SMO
24. MO
25. JMO
26. E
27. ONE
28. ETE
29. ATE
30. ITE
31. STE
32. TE
33. JTE
34. JE
35. ŠE
36. OŠE
37. AŠE
38. IM
39. OM
40. AM
41. EM
42. JEM
43. H
44. IH
45. OH
46. AH
47. JAH
48. J
49. OJ
50. OG
51. EG
52. V
53. AV
54.
55. *prazna niska*

Virtual Library of Faculty of Mathematics - University of Belgrade
elibrary.matf.bg.ac.rs

PRILOG -F-

1. Analiza rada programa za konstruisanje rečnika prefiksa nad korpusom (Prilog A.3) - pojavljivanje prefiksa

Ukupno redova teksta:	28.265	Ukupno osnova u rečniku:	882
Ukupno reči teksta:	212.816	Ukupno osnova sa različitim prefiksom:	978
Traženo reči u rečniku:	5.763		
Pronađeno reči u rečniku:	4.778	Prosečno prefiksa po osnovi:	1,109
Od toga sa istim prefiksom:	4.672		

Prefiksna niska	Pronađeno u korpusu	Kandidat za rečnik	Jeste prefiks	Nije prefiks
1. ANTI	30	0	0	0
2. ARHI	0	0	0	0
3. BEZ	30	26	11	3
4. BES	32	29	12	0
5. BE	0	0	0	0
6. BEŠ	0	0	0	0
7. BEZA	1	0	0	0
8. DE	1287	0	0	0
9. DIS	26	12	0	2
10. DO	2739	0	0	0
11. EKS	40	1	0	1
12. EKSTRA	0	0	0	0
13. IZ	2680	69	17	4
14. IS	1150	1013	134	19
15. IŽ	0	0	0	0
16. IŠ	5	0	0	0
17. IZA	116	75	15	3
18. IZVAN	21	0	0	0
19. INTER	82	81	7	6
20. KONTRA	0	0	0	0
21. KVAZI	0	0	0	0
22. MEĐU	294	0	0	0
23. MIMO	1	0	0	0
24. NA	8784	0	0	0
25. NAD	166	93	12	14
26. NAT	8	0	0	0
27. NADRI	0	0	0	0
28. NAJ	1055	41	21	6
29. NAZOVI	0	0	0	0
30. NE	3119	0	0	0
31. NI	993	0	0	0
32. NIZ	163	53	0	5
33. NIS	378	0	0	0
34. NOVO	81	11	7	0
35. NADA	13	3	0	2
36. NUZ	0	0	0	0
37. NUS	0	0	0	0
38. OB	1287	1010	42	90
39. OP	261	0	0	0
40. OBA	326	0	0	0

Virtual Library of Faculty of Mathematics - University of Belgrade
elibrary.matf.bg.ac.rs

Prefiksna niska	Pronađeno u korpusu	Kandidat za rečnik	Jeste prefiks	Nije prefiks
41. OD	3277	1125	112	32
42. OT	206	0	0	0
43. ODA	53	0	0	0
44. PA	1858	0	0	0
45. PO	6362	0	0	0
46. POD	825	515	32	42
47. POT	85	0	0	0
48. PODA	74	71	0	7
49. POLU	78	72	7	4
50. POSLE	344	127	2	2
51. PRA	493	0	0	0
52. PRE	1140	0	0	0
53. PRED	709	280	12	21
54. PRET	85	0	0	0
55. PREKO	204	9	3	3
56. PREK	29	24	1	12
57. PRI	1712	0	0	0
58. PRO	2426	0	0	0
59. PROTIV	203	0	0	0
60. PROTU	0	0	0	0
61. PSEUDO	0	0	0	0
62. RAZ	1128	64	12	15
63. RAS	286	234	57	13
64. RAZA	9	0	0	0
65. RA	0	0	0	0
66. RAŽ	0	0	0	0
67. RAŠ	2	0	0	0
68. SA	2815	0	0	0
69. SU	4680	0	0	0
70. SUPER	0	0	0	0
71. SAMO	428	51	17	1
72. TRANS	16	0	0	0
73. UZ	288	54	1	12
74. US	636	620	30	35
75. UZA	25	0	0	0
76. UŠ	44	0	0	0
77. ULTRA	0	0	0	0
78. VELE	13	0	0	0
79. ZA	4764	0	0	0
80. IN	612	0	0	0
81. KON	319	0	0	0

Virtual Library of Faculty of Mathematics - University of Belgrade
elibrary.matf.bg.ac.rs

2. Analiza rada programa za konstruisanje rečnika prefiksa nad korpusom (Prilog A.3) - popunjavanje rečnika

Tekst	Ukupno reči	Ukupno kandidata	Dodato osnova u rečnik		
			Ukupno	% od reči	% od kand.
1.	10.874	265	109	1,0024	41,1
2.	24.981	582	157	0.6285	27,0
3.	2.120	52	27	1,1792	48,1
4.	12.349	262	60	0,4859	22,9
5.	15.744	353	50	0,3176	14,2
6.	8.752	213	21	0,2399	9,9
7.	12.512	398	22	0,1758	5,5
8.	16.888	378	60	0,3553	15,9
9.	652	5	1	0,6135	80,0
10.	11.688	301	28	0,2396	9,3
11.	14.377	322	27	0,1878	8,4
12.	2.232	72	12	0,5376	16,7
13.	16.812	421	62	0,3688	14,7
14.	4.809	99	34	0,7070	34,3
15.	1.705	25	8	0,4692	32,0
16.	11.010	304	21	0,1907	6,9
17.	17.434	562	35	0,2008	6,2
18.	897	10	3	0,3344	30,0
19.	26.958	1026	89	0,3301	8,7

Virtual Library of Faculty of Mathematics - University of Belgrade
elibrary.matf.bg.ac.rs

PRILOG -G-

Rečnika obrazaca za rastavljanje reči srpskohrvatskog jezika

.an2ti	.be2s3c	.be2s3f	.be2s3h	.be2s3k
.be3s4kru	.be2s3p	.be3s4pok	.be3s4por	.be2s3t
.be3s4ti	.be3s4tras	.bes4ts	.be2š3č	.be2š3ć
.be2z3	.bez4a	.be3zak	.bez5al	.be3z4e
.be3z4is	.be3z4lo	.be3z4nača	.be3z4rač	.be3z4ub
.be3z4vu	.di2s3	.di3s4a	.di3s4e	.di3s4i
.di3s4o	.di3s4u	.ek2s3	.ek3s4a	.ek3s4e
.ek3s4i	.ek3s4o	.ek3s4u	.in3jekc	.in3junkt
.in2ter	.in3tere	.in3teri	.is3c	.is3f
.is3h	.is3k	.is4kač	.is4kak	.is4koč
.is4krenu	.is3p	.is3t	.is4tob	.is4tod
.is4toi	.is4tom	.is4top	.is4toro	.is4tos
.is4tove	.is4tovr	.is4tup	.iš3č	.iš3ć
.iz3	.iz4a	.iz5an	.iz4e	.iz4id
.iz4iš	.iz4oba	.iz4of	.iz4og	.iz4ol
.iz4om	.iz4ote	.iz4oto	.iz4van	.iz4uma
.iz4ume	.iz4umeh	.iz4umel	.iz4umet	.iz4umeć
.iz4umeš	.iz4umi	.iz5umir	.iz4umo	.iz4ut
.iz4uva	.iz4uzeta	.iz4uzetk	.iz4uzetn	.iž3dj
.iž3dž	.ko2n3jug	.ko2n3junk	.kon2tra	.kva2zi
.me2dju	.mi2mo	.na2d3	.na3d4a	.na4dasv
.na3d4e	.na3d4i	.na4d5igr	.na3d4ji	.na3d4je
.na3d4jo	.na3d4ju	.na4d5jun	.na3d4o	.na3d4r
.na4d5rea	.na4d5red	.na4d5redj	.na4dri	.na3d4u
.na3d4voj	.na3d4vor	.na3d4ža	.na3d4židž	.na2j3
.na3j4ah	.na3j4ama	.na3j4ame	.na3j4ami	.na3j4amn
.na3j4amo	.na3j4amu	.na3j4av	.na4j5avet	.najbe2s3k
.najbe2s3p	.najbe2s3t	.najbe2š3č	.najbe2z3	.najbe3z4a
.na3j4ed	.na3j4el	.na4j5elem	.najek2s3t	.na3j4eo
.na3j4esti	.na3j4eš	.na3j4ez	.na3j4ež	.naji2s3k
.naji2s3p	.naji2s3t	.naji2z3	.naji3z4a	.najo2b3
.najo3b4a	.najo2d3	.najo3d4a	.najpo2d3	.najpo3d4a
.najpre2d3	.najra2s3k	.najra2s3p	.najra2s3t	.najra2z3
.najra3z4a	.na3j4uri	.naju2s3h	.naju2s3k	.naju2s3p
.naju2s3t	.naju2z3	.naju3z4a	.na2zo2vi	.nei2s3k
.nei2s3p	.nei2s3t	.nei2z3	.nei3z4a	.nei3z4e
.nei3z4i	.nei3z4o	.nei3z4u	.neo2b3	.neo3b4a
.neo3b4e	.neo3b4i	.neo3b4o	.neo3b4u	.neo2d3
.neo3d4a	.neo3d4e	.neo3d4i	.neo3d4o	.neo3d4u
.nepo2d3	.nepo3d4a	.nepo3d4e	.nepo3d4i	.nepo3d4o
.nepo3d4u	.nepre2d3	.nera2s3k	.nera2s3p	.nera2s3t
.nera2z3	.nera3z4o	.nera3z4u	.neu2s3h	.neu2s3k
.neu2s3p	.neu2s3t	.neu2z3	.neu3z4a	.ni2z3br
.ni2z3v	.no2vo	.no2voi2z3g	.no3vosti	.no3vošć
.nu2s3f	.nu2s3h	.nu2s3k	.nu2s3č	.nu2s3ć
.nu2s3p	.nu2s3t	.nu2s3š	.nu2z3	.ob3
.ob4e	.obe2s3c	.obe2s3h	.obe2s3k	.obe2s3pr
.obe2š3č	.obe2z3b	.obe2z3g	.obe2z3n	.obe2z3o
.obe2z3um	.obe2z3v	.ob4i	.ob5igr	.ob5ist
.ob4jek	.ob4l	.ob5lag	.ob5laž	.ob5lep
.ob5let	.ob5leć	.ob5lic	.ob5lič	.ob5lić
.ob5lij	.ob5lik	.ob5lil	.ob5lio	.ob5liš

Virtual Library of Faculty of Mathematics - University of Belgrade
elibrary.matf.bg.ac.rs

.ob5lit	.ob5liv	.ob5liz	.ob4o	.ob4r
.ob5rač	.ob5rad	.ob5radj	.ob5rasl	.ob5rast
.ob5ruč	.ob4uc	.ob4učen	.ob4učić	.ob4uć
.ob4uk	.ob4ul	.ob4uo	.ob4us	.ob4uš
.ob4uv	.obu2z3d	.od3	.od4e	.od4i
.od5igr	.od5is	.od4o	.od5oka	.od5onud
.od5ovud	.od4rać	.od4rah	.od4ral	.od4rao
.od4rap	.od4rasm	.od4raš	.od4rat	.od4reš
.od4rp	.od4rt	.od4rv	.od4rž	.od4ugo
.od4uh	.od4uk	.od4ulj	.od4un	.od4ur
.od4ust	.od4uš	.od4uva	.od4už	.od4vaj
.od4voj	.od4za	.po2d3	.po3d4a	.po4d5adm
.pod4av	.po3davim	.po3d4e	.po3d4i	.po4dići
.po4d5iš	.po4d5idj	.po4d5il	.po3d4je	.po4d5jed
.po4d5jes	.po4d5jez	.po3d4ji	.po3d4jo	.po3d4ju
.po3d4nev	.po3d4o	.po4d5odb	.po4d5of	.po4d5oč
.po3d4raž	.po3d4rem	.po3d4rht	.po3d4rinj	.po3d4rob
.po3d4rp	.po3d4rš	.po3d4rt	.po3d4rum	.po3d4ruš
.po3d4ruž	.po3d4rž	.po3d4uc	.po3d4ud	.po3d4ug
.pod4uh	.po3d4uk	.po3d4ulj	.po3d4un	.po3d4uzetn
.pod4uzeć	.po3d4už	.po3d4vig	.po3d4viz	.po3d4vor
.po3d4vost	.po2lu	.po3ludeć	.po3ludeh	.po3ludel
.po3ludeo	.po3ludeš	.po3ludet	.po3ludi	.po3luga
.po3luge	.po3lugom	.po3lugu	.po3lupa	.po3lutan
.po3lutar	.po3lutk	.po3luzi	.poo2d3m	.poo2d3r
.po2sle	.po3sledic	.po3sledič	.po3slednj	.po3sleni
.pou2z3d	.pre2d3	.pre3d4a	.pre3d4e	.pre3d4i
.pre4d5ig	.pre4d5isp	.pre4d5ist	.pre4d5izb	.pre3d4ja
.pre3d4je	.pre4d5jel	.pre3d4ji	.pre3d4jo	.pre3d4ju
.pre3d4o	.pre4d5odr	.pre4d5ose	.pre3d4ubo	.pre3d4ug
.pre3d4uh	.pre3d4uj	.pre3d4um	.pre3d4up	.pre3d4usr
.pre3d4uzeć	.pre3d4uzetn	.pre3d4uzimlj	.pre2k3j	.pre2k3lan
.pre2koa	.pre2kobr	.pre2kome	.pre2komo	.pre2kono
.pre2koo	.pre2kopu	.pre2kosu	.pre2kovr	.proi2s3
.proi2z3n	.proi2z3v	.pro2ti2v3l	.pro2tiv3r	.pse2u2do
.ra2s3c	.ra2s3f	.ra2s3h	.ra2s3k	.ra3s4koš
.ra2s3p	.ra2s3t	.ra3s4ta	.ra4sptak	.ra4s5tanj
.ra4s5tap	.ra3s4te	.ra4s5teć	.ra4s5teg	.ra4s5ter
.ra4s5tez	.ra4s5tež	.ra3s4tinj	.ra3s4toj	.ra3s4tuć
.ra2s3č	.ra2š3ć	.ra2z3	.ra2za	.ra3z4ilas
.ra3z4ilaza	.ra3z4in	.ra3z4on	.raz4oran	.ra3z4oren
.ra3z4ori	.ra3z4orn	.ra3z4udj	.ra3z4um	.ra3z4uzdan
.ra2ž3dj	.ra2ž3dž	.sa2mo	.sa3moć	.sa3mostan
.sa3mota	.sa3moti	.sa3motn	.sa3movać	.sa3movah
.sa3moval	.sa3movao	.sa3movas	.sa3movaš	.sa3movat
.su2pers	.tran2s3	.ve2le	.ul2tra	.us3c
.us3f	.us3h	.us3k	.us4kak	.us4kla
.us4koč	.us4koć	.us4kog	.us4kol	.us4koro
.us4kos	.us4kot	.us3p	.us4pav	.us4pok
.us4porav	.us4poren	.us4pori	.us3talas	.us3traj
.us3traž	.us3trč	.us3treb	.us3trep	.us3trg
.us3trp	.us3tuk	.us3tum	.us3tv	.uš3č
.uš3ć	.uz3	.uz4e	.uz4ic	.uz4id
.uz4im	.uz4orak	.uz4orc	.uz4orit	.uz4ork
.uz4orn	.zau2z3d	a1a	a1e	a1i
a1o	a1u	1ba	1be	1bi

Virtual Library of Faculty of Mathematics - University of Belgrade
elibrary.matf.bg.ac.rs

1b2j	1b2l	1b2lj	1b2m	1b2n
1b2nj	1bo	1b2r	1bu	1b2v
1ca	1ce	1ci	1c2j	1c2l
1c2lj	1c2m	1c2n	1c2nj	1co
1c2r	1cu	1c2v	1ča	1če
1či	1č2j	1č2l	1č2lj	1č2m
1č2n	1č2nj	1čo	1č2r	1ču
1č2v	1ća	1će	1ći	1ć2j
1ć2l	1ć2lj	1ć2m	1ć2n	1ć2nj
1ćo	1ć2r	1ću	1ć2v	1da
1de	1di	1d2j	1dj2j	1dj2l
1dj2lj	1dj2m	1dj2n	1dj2nj	1dj2r
2djs	1dj2v	1d2l	1d2lj	1d2m
1d2n	1d2nj	1do	1d2r	1du
1d2v	1d2ž	2dž3b	1dž2j	1dž2l
1dž2lj	1dž2m	1dž2n	1dž2nj	1dž2r
1dž2v	e1a	e1e	e1i	e1o
e1u	1fa	1fe	1fi	1f2j
1f2l	1f2lj	1f2m	1f2n	1f2nj
1fo	1f2r	1fu	1f2v	1ga
1g2d	1ge	1gi	1g2j	1g2l
1g2lj	1g2m	1g2n	1g2nj	1go
1g2r	1gu	1g2v	1ha	1he
1hi	1h2j	1h2l	1h2lj	1h2m
1h2n	1h2nj	1ho	1h2r	1h2t
1hu	1h2v	i1a	i1e	i1i
i1o	i1u	1ja	1je	1ji
1jo	1ju	1ka	1k2ć	1ke
1ki	1k2j	1k2l	1k2lj	1k2m
1k2n	1k2nj	1ko	1k2r	k2s3t
k2t3n	1ku	1k2v	1la	1le
1li	1l2j	2ljb	2ljc	2ljd
2ljk	2ljn	2ljs	1lo	l2t3n
1lu	1ma	1me	1mi	1m2l
1m2n	2m3nj	1mo	1m2r	1mu
1na	1ne	1ni	1n2j	2njc
2njs	1no	n2t3n	1nu	o1a
o1e	o1i	o1o	or2f3n	o1u
1pa	1p2č	1pe	1pi	1p2j
1p2l	1p2lj	1p2m	1p2n	1p2nj
1po	1p2r	1p2s	1p2š	1p2t
1pu	1p2v	1ra	1re	1ri
1ro	1ru	1sa	1s2b	1s2c
1s2č	1s2d	1s2dj	1s2dž	1se
1s2f	1s2g	1s2h	1si	1s2j
1s2k	1s2l	1s2lj	1s2m	1s2n
1s2nj	1so	1s2p	1s2r	1s2š
1s2t	1su	1s2v	1s2ć	1s2z
1s2ž	1ša	1š2b	1š2c	1š2č
1š2ć	1š2d	1š2dj	1š2dž	1še
1š2f	1š2g	1š2h	1ši	1š2j
1š2k	1š2l	1š2lj	1š2m	1š2n
1š2nj	1šo	1š2p	1š2r	1š2s
1š2t	1šu	1š2v	1š2z	1š2ž
1ta	1te	1ti	1t2j	1t2k
1t2l	1t2lj	1t2m	1t2n	1t2nj

Virtual Library of Faculty of Mathematics - University of Belgrade
elibrary.matf.bg.ac.rs

1to	1t2r	1tu	1t2v	u1a
u1e	u1i	u1o	ur2s3n	u1u
1va	1ve	1vi	1v2l	2v3lj
1vo	1v2r	1vu	1za	1z2b
1z2c	1z2č	1z2ć	1z2d	1z2dj
1z2dž	1ze	1z2f	1z2g	1z2h
1zi	1z2j	1z2k	1z2l	1z2lj
1z2m	1z2n	1z2nj	1zo	1z2p
1z2r	1z2s	1z2š	1z2t	1zu
1z2v	1z2ž	1ža	1ž2b	1ž2c
1ž2č	1ž2ć	1ž2d	1ž2dj	1ž2dž
1že	1ž2f	1ž2g	1ž2h	1ži
1ž2j	1ž2k	1ž2l	1ž2lj	1ž2m
1ž2n	1ž2nj	1žo	1ž2p	1ž2r
1ž2s	1ž2š	1ž2t	1žu	1ž2v
1ž2z				

Rečnik izuzetaka

na-jam
no-vost
odra-ste
po-dne
po-dno
ra-sti
ra-stu
uzore
uzori

Virtual Library of Faculty of Mathematics - University of Belgrade
elibrary.matf.bg.ac.rs

LITERATURA

[Achugbue81] Achugbue, J. O.: *On the Line Breking Problem in Text Formatting*, Proceedings of the ACM SIGPLAN SIGOA Symposium on Text Manipulation, Oregon 1981, ACM SIGPLAN Notices, Vol. 16, No. 6, pp. 117-121, (1981)

[AOT1985] ------- : *Automatska obrada teksta*, Projekat OZN Beograd, Koordinator Matematički Institut, Beograd, (1985)

[Appelt85] Appelt, W: *The hyphenation of non-English words xith TeX*, TeX for Scientific documentation, Proceedings of the Como Conference, May 1985, Addison-Wesley, Reading, Mass., USA, (1985)

[Belić34] Belić, A.: *Pravopis srpskohrvatskog književnog jezika*, Beograd, pp. 15-18, (1934)

[Boitet85] Boitet, C., Verastegui, N., Bachut, D.: *Various Representations of Text for EUROTRA*, Second Conference of the European Chapter of the ACL, Proceedings of the Conference, Geneva, March 1985, University of Geneva, (1985)

[Cvijov84] Cvijović, M.: *Dečji govor, rečnik i rečenica - četvrti razred osnovne škole*, Zavod za udžbenike i nastavna sredstva, Beograd, (1984)

[Day84] Day, A. C.: *Text Processing*, Cambridge University Press, Cambridge, (1984)

[Desarm84e] Desarmenien, J.: *How to Run TeX in French*, Department of Computer Science, Stanford University, Report No. STAN-CS-1013, (1984)

[Desarm84f] Desarmenien, J.: *La division par ordinateur des mots francais avec le logiciel TeX*, (deo rada [Desarm84e]), (1984)

[Desarm87] Desarmenien, J.: *French hyphenation by computer: application to TeX*, Technology and Science of Informatics, Gauthier-Villars & John Wiley & Sons, Vol. 6, No. 1. (1987)

[Furuta82] Furuta, R., Scofield, J., Shaw, A.: *Document Formatting Systems: Survey, Concepts, and Issues*, ACM Computing Sutveys, Vol. 14, No. 3, pp. 417-472, (1982)

[Gimpel76] Gimpel, J. F.: *Algorithms in SNOBOL4*, John Wiley & Sons, pp. 211-215, (1976)

[Gramatika] ------- : *Priručna gramatika hrvatskoga književnog jezika*, (grupa autora), Školska knjiga, Zagreb, (1979)

[Hornbey86] ------- : *Oxford Advanced Learner's Dictionary of Current English*, Oxford University Press, 23rd Impression, (1986)

[ISO646] ------- : *Information processing - ISO 7-bit coded character set for information interchange*, ISO 646, International Organization for Standardisation, (1983)

Virtual Library of Faculty of Mathematics - University of Belgrade
elibrary.matf.bg.ac.rs

[JUS.B1.002] ----- : *Skup znakova za razmenu podataka kodiranih sa 7 bitova za srpskohrvatsko latinično pismo*, JUS I.B1.002, Savezni zavod za standardizaciju, (1986)

[JUS.K1.002] ------- : *Jedinice za unos podataka - Tastatura sa 47 tipki za slovenačko i hrvatsko latinično pismo*, JUS I.K1.002, Savezni zavod za standardizaciju, (1986)

[JUS.K1.003] ------- : *Jedinice za unos podataka - Tastatura sa 47 tipki za srpskohrvatsko ćirilično pismo*, JUS I.K1.003, Savezni zavod za standardizaciju, (1986)

[Knuth73] Knuth, D. E.: *The Art of Computer Programing*, Vol. 3, Addison-Wasley, Reading, Mass., USA, (1973)

[Knuth81] Knuth, D. E., Plass, M. F.: *Breaking Paragraphs into Lines*, Software - Practice and Experience, Vol. 11, pp. 1119-1184, (1981)

[Knuth84] Knuth, D. E.: *TeXbook*, Addison-Wesley, Reading, Mass., (1984)

[Knuth86] Knuth, D. E.: *TeX: The Program*, Addison-Wasley, Reading, Mass., (1986)

[Kolodz87] Kolodziejska, Hanna: *Dzielenie wyrazow polskich w systemie TeX*, Report No. 165, Institut of Informatics, Warsaw University, (1987)

[Kolodz88] Kolodziejska, Hanna: *Le traitement des textes polonais avec le logiciel TeX*, Les Cahiers de GUTenberg, No. 0, IRISA, Rennes, (1988)

[Krstev82] Krstev, C.: *Frekvencijski rečnik konsonantskih grupa u srpskohrvatskom jeziku i problem rastavljanja na slogove*, Zbornik II naučnog skupa "Računarska obrada lingvističkih podataka", Institut "Jožef Stefan", Bled, pp. 389-404, (1982)

[Krstev85] Krstev, C.: *Rastavljanje reči srpskohrvatskog jezika na kraju retka*, Zbornik III naučnog skupa "Računarska obrada jezičkih podataka", Institut "Jožef Stefan", Bled, pp. 289-301, (1985)

[Krstev88] Krstev, C.: *O problemima primene i adaptiranja programske opreme za obradu teksta u neengleskim govornim sredinama*, Zbornik IV naučnog skupa "Računarska obrada jezičkih podataka", Institut "Jožef Stefan", Portorož, pp. 235-244, (1988)

[Liang83] Liang, F. M.: *Word Hy-phen-a-tion by Com-put-er*, Ph. D. Thesis, Departement of Computer Science, Stanford University, Report No. STAN-CS-83-977, (1983)

[Longman] ------- : *Longman Dictionary of Contemporary English*,

[Mesaroš83] Mesaroš, F.: *Fotoslog - automatska obrada teksta*, Viša grafička škola, Zagreb, (1983)

[Meyrow82] Meyrowitz, N., Van Dam, A.: *Interactive editing systems*, ACM Computing Surveys, Vol. 14, No. 3, pp. 321-415, (1982)

Virtual Library of Faculty of Mathematics - University of Belgrade
elibrary.matf.bg.ac.rs

[Moitra79] Moitra, A., Mudur, S.P., Narwekar, A. W: *Design and Analysis of a Hyphenation Procedure*, Software - Practice and Experience, Vol. 9, pp. 325-337, (1979)

[Ocker75] Ocker, W. A.: *A program to Hyphenate English Words*, IEEE Transactions on Professional Communication, Vol. PC-18, No. 2, (1975)

[Parezanović72] Parezanović, N.: *Algoritmi i programski jezik FORTRAN IV*, Matematički institut, Beograd, (1972)

[Peterson80] Peterson, J. L.: *Computer Programs for Spelling Correction: An Experiment in Program Design*, Springer-Verlag, Berlin Heidelberg New York, (1980)

[Pravopis60] ------- : *Pravopis srpskohrvatskog književnog jezika*, Novi Sad - Zagreb, 130-131, (1960)

[Pringle81] Pringle, A. M.: *Justification with Fewer Hyphens*, The Computer Journal, Vol. 24, No. 4, pp. 320-323, (1981)

[Rečnik] Benton, M.: *Srpskohrvatsko-engleski rečnik*, Prosveta, Beograd, (1978)

[Rich65] Rich, R. P., Stone, A. G.: *Method for Hyphenating at the End of a Printed Line*, Communications of the ACM, Vol. 8, No. 7, pp. 444-445, (1965)

[Romberger85] Romberger, S., Sundblad, Y.: *Adapting TeX to Languages that Use Latin Alphabetic Characters*, TeX for Scientific Documentation, Proceedings of the Como Coference, May 1985, Addison-Wasley, Reading, Mass., USA, (1985)

[Samet82] Samet, H.: *Heuristics for the Line Division Problem in Computer Justified Text*, Communications of th ACM, Vol. 25, No. 8, pp. 564-571, (1982)

[Sedgewick83] Sedgewick, R.: *Algorithms*, Addison-Wesley, Reading, Mass., (1983)

[Seeber72] Seeber, E. D.: *A Style Manual for Students*, Bloomington & London, Indiana University Press, pp. 28-31, (1972)

[Seybold87] Seybold, J. W.: *The Desktop-Publishing Phenomenon*, BYTE, May 1987

[System360] -------, *System/360 Text Processor - Hyphenation/360, Application Description Manual*, 2nd ed., IBM Corp.

[Stevanović64] Stevanović, M.: *Savremeni srpskohravtski jezik*, Beograd, pp. 152-156, (1964)

[Tolstoja68a] Tolstoja, S. M.: *Sočetaemost' soglasnyh v svjazi s fonologičeskoj strukturi slova v slavjanskih jazykah*, Sovetskoe slavjanovedenie 1, pp. 41-55, (1968)

Virtual Library of Faculty of Mathematics - University of Belgrade
elibrary.matf.bg.ac.rs

[Tolstoja68b] Tolstoja, S. M.: *Fonologičeskoe rasstojanie i soćetaemost' soglasnyh v slavjanskih jazykah* , Voprosy jazikoznanija, vol. 3, pp. 66-82, (1968)

[Tolstoja72] Tolstoja, S. M.: *Načan'nye i konačnye soćetanija soglasnyh v serbsko-horvatskom jazyke*, Isledovanija po serbsko-horvatskom jazyku, Moskva, pp. 3-38, (1972)

[Tomić78] Tomić, T.: *Statistička analiza srpskohrvatskog teksta pomoću računara*, Zbornik "Kompjuterska obrada lingvističkih podataka", ed. M. Šipka, Institut za jezik i književnost u Sarajevu, Sarajevo, (1978)

[Vitas79] Vitas, D.: *Prikaz jednog sistema za automatsku analizu teksta*, INFORMATICA '79, Bled, 7101, (1979)

[Vitas81] Vitas, D.: *Podela na slogove srpskohrvatskih reči*, Informatika, Ljubljana, 3/1981

[Vitas82] Vitas, D.: *Prikaz jednog programskog sistema za automatsku obradu teksta*, Zbornik II naučnog skupa "Računarska obrada lingvističkih podataka", Institut "Jožef Stefan", Bled, pp. 457-465, (1982)

[Vitas85] Vitas, D.: *Jedan postpuak automatske segmentacije srpskohrvatskih reči i njegove primene*, Zbornik III naučnog skupa "Računarska obrada jezičkih podataka", Bled, oktobar 1985, Institut "Jožef Stefan", pp. 303-313, (1985)

[Vitas85b] Vitas, D., Tancig, P.: *Prirodno-jezičke komponente u automatskoj obradi teksta*, Zbornik III naučnog skupa "Računarska obrada jezičkih podataka", Bled, oktobar 1985, Institut "Jožef Stefan", pp. 15-20, (1985)

[Webster61] -------, *Rules for syllabic division of words in writing or printing*, Webster's New International Dictionary, 2nd ed. Springfield, Mass: Merriam, pp. LVIII-LIX, (1961)

Virtual Library of Faculty of Mathematics - University of Belgrade
elibrary.matf.bg.ac.rs