PRIKAZ JEDNOG SISTEMA ZA AUTOMATSKU ANALIZU TEKSTA

D. Vitas
Matematički institut, Beograd, Knez Mihailova 35, Jugoslavija

Svrha rada je da se opiše struktura jednog sistema za generisanje konkordanci i drugih oblika rečnika za dati tekst, napisan na prirodnom jeziku, kao i za odredjivanje kvantitativnih parametara teksta. Opisani su elementi modela teksta i osnovne karakteristike realizovanog sistema. Prikazana su prva iskustva u primeni sistema na tekstove različite prirode kao i mogućnosti daljih istraživanja.

A REVIEW OF A SYSTEM FOR AN AUTOMATIC TEXT ANALYSIS. The present paper is intented to describe the structure of a system for generating concordances and other forms of dictioneries for a given text, written in a natural language, and for determinating of quantitative parameters of the text. The elements of the text model and the characteristics of the realized system are described. First experiences in the application of system to texts of different kinds are displayed, and the further investigations are outlined.

U V O D

Problem automatske izrade rečnika nad datim tekstom nije nov. Iscrpan pregled ostvarenih rezultata u ovoj oblasti nalazi se u (Tancig/77/, Tancig/78/).

Prilikom konstruisanja sistema koji je opisan u ovom radu, namera je bila da se omogući objedinjavanje nekih do sada ostvarenih rezultata i tekućih istraživanja u jedan složeni model srpskohrvatskog (kraće,SH-)jezika. Naime, tokom poslednje decenije, uz pomoć računara, u beogradskoj sredini su vršena obimna ispitivanja nad različitim leksičkim materijalom(npr.Lukić /70/,Vasić/73/,Ničković,Cvijović/77/,Plavšić /78/, itd). Ovi radovi nisu bili praćeni odgovarajućim razvojem metodologije prikupljanja i obrade lingvističkih podataka, a shodno tome, ni razvojem prilagodjene programske opreme, što je od uticaja kako na pouzdanost tako i na cenu izvršenih istraživanja. Svrha sistema, koji se opisuje u ovom radu, je, da osim uključivanja dobijenih rezultata u model SH-jezika, obezbedi standardnu programsku podršku budućih istraživanja, utoliko pre što nam predstoje neki obimni i značajni lingvistički (npr.obrtanje SH-rečnika, izrada minimalnih frekvencijskih rečnika SH-jezika, itd.) i računarski projekti(Algoritmi morfo-sintaksičke analize SH-jezika).

MODEL TEKSTA

Tekst se, u lingvističkoj praksi, obično definiše kao skup rečenica nekog prirodnog jezika, koje mogu biti podvrgnute lingvističkoj analizi. Za potrebe automatske analize teksta, ovako uveden pojam teksta treba ograničiti i, takodje, bliže odrediti njegovu formalnu strukturu. U daljem izlaganju ćemo pod pojmom tekst podrazumevati pisani tekst, kojem je, osim lingvističke interpretacije, dodeljena i izvesna grafička interpretacija. Grafička interpretacija teksta nema veći lingvistički značaj, ali nam dopušta da tekstu pridružimo odredjenu grafičku formu, što je u nekim slučajevima(npr.poetski tekstovi) od posebne važnosti.Ovako shvaćen pojam teksta je moguće formalno opisati na sledeći način.

Neka je A azbuka prirodnog jezika L. Elemente azbuke zovemo simboli azbuke. Neka je na azbuci A definisana relacija poretka R medju simbolima azbuke. Na uobičajeni način se uvode pojmovi lanca, podlanca, prefiksa lanca i sufiksa lanca nad azbukom A. Skup svih lanaca nad azbukom A

obeležimo sa A^*. Neka je $A^+ = AA^*$. Neka je V skup svih oblika reči jezika L. Tada je $V \subseteq A^*$. Skup V zovemo rečnik oblika jezika L. Neka je I neprazan skup i $A \cap I = \phi$. Elemente skupa I ćemo zvati interpunkcijski simboli. Lanac interpunkcijskih simbola ćemo zvati interpunkcijski lanac. Na osnovu uobičajenih pravila ortografije, interpunkcijski lanac je prefiks ili sufiks lanca nad azbukom A. Elemente skupa

$$I^*A^+I^* \ldots . (1)$$

zvaćemo reč teksta. Označimo sa K´neprazan skup čije ćemo elemente zvati simboli grafičke reprezentacije teksta, ili kraće, grafički simboli. Skupovi A,I,K´ su dva po dva disjunktni.

Jedan istaknuti element skupa K´ zvaćemo razmak (ili blanko-simbol), u oznaci $\smile$. Neka je $K=K' - \{\smile\}$. Neka je razmak prefiks svake reči teksta (1). Skup tekstova T nad azbukom A bi, tada, mogao da se definiše kao podskup sledećeg skupa:

$$(K^*(\{\smile\}I^* A^+ I^*)^+)^+ \ldots \quad (2)$$

Jedan način definisanje skupa tekstova je konstruisanje gramatike G takve da je L(G)=T, gde je L(G) jezik generisan gramatikom G. Jedna takva gramatika je implicitno definieana programskim sistemom opisanim u daljem izlaganju. Eksplicitno navodjenje ove gramatike prelazi obim ovog rada.

Za dati tekst nad azbukom A, koji je element skupa (2), zadatak automatskog generisanja rečnika se svodi na izdvajanje iz skupa (2) elemenata oblika A^*. Poredak reči u rečniku je obično indukovan poretkom R medju simbolima azbuke A.

U izrazu (1) koji definiše reč teksta, skup A^+ se može zameniti rečnikom oblika V, no na taj način tekst bi mogao biti sastavljen isključivo od "gramatičnih" reči, čime bi se zadatak mehaničkog izdvajanja reči nad azbukom A iz skupa (2) nepotrebno komplikovao. Pitanje sintaksičkih i semantičkih relacija u tekstu je ostavljeno po strani iako se neki jezički fenomeni mogu ispitivati i u okviru ovako zamišljenog modela teksta. Jedan pokušaj formalnog opisivanja sintaksičko-semantičkih relacija u tekstu se nalazi npr. u (Petöfi/77/).

Grafički simboli u modelu teksta omogućavaju očuvanje originalne grafičke reprezentacije teksta, čime se pojednostavljuje korigovanje i prikazivanje tekstova. S druge strane, u kritičkoj analizi tekstova (Froger/68/), grafička reprezentacija je od posebnog značaja prilikom uredjivanja kritičkih izdanja. Na osnovu pojedinih grafičkih simbola moguće je definisati poredak medju delovima teksta na način koji ćemo kasnije izložiti.

OSOBINE REALIZOVANOG SISTEMA

U daljem izlaganju ćemo opisati neke karakteristike realizovanog programskog sistema za automatsko generisanje konkordanci i drugih rečnika nad datim tekstom. Sistem je konstruisan u skladu sa opisanim modelom teksta. Programi sistema su napisani na programskom jeziku FORTRAN IV (IBM proširena verzija) i testirani u RC Matematičkog instituta. Sistem je koncipiran tako d se lako može prilagoditi potrebama korisnika.

Kako se sistem primenjuje na proizvoljan tekst napisan na prirodnom jeziku, tok obrade teksta ne zavisi od izbora azbuka A,I i K, pa ih korisnik može sâm odrediti prema prirodi teksta. Kak se za opis teksta obično koristi više od 49 dopuštenih FORTRAN-skih karaktera, neki simboli azbuka A,I i K se kodiraju grupom FORTRAN-skih karaktera. Kodovi simbola azbuka zadaju se u potprogramu KODER, u redosledu koji je odredjen frekvencijom njihovog pojavljivanja u jeziku. Tako,napr, slovo Ž se može kodirati sa ZX ili ZZ a interpunkcijski znak ... je prirodno predstaviti sa ... Prilikom definisanja azbuka A,I i K potrebno je voditi računa o uslovu disjunktnosti azbuka i odstraniti eventualno ambiguitetne simbole. Potprogramom KODER se definiše i poredak medju simbolima azbuke A (ćirilični i latinični poredak su dva moguća poretka za azbuku SH-jezika). Kako se u toku obrade teksta prvo razvrstavanje reči vrši po prvom ili poslednjem slovu reči, moguće je, radi efikasnije obrade obimnih tekstova, pojedina frekventna slova rastaviti na manje frekventne grupe slova Tako, umesto po inicijalnom p-, jedno razvrstavanje reči se može izvršiti i po sledećim inicijalnim grupama: pa-, po-, pre-, pri-, pro-, pu-, p-. Dekodiranje karaktera u tekstu rutinom KODER se može vršiti sleva na desno ili sdesna na levo. Ako se u tekstu upotrebi karakter koji ne pripada definisanim azbukama, štampa se poruka o grešci. U primerima na kraju rada, poredak reči je ćirilični s tim što je azbuci od 30 ćiriličnih slova, dodat i znak elizije (´).

Skup dopuštenih interpunkcijskih simbola je definisan Pravopisom a bez teškoća se može proširiti tako da obuhvati i pravopisne simbole. Skup grafičkih simbola je odredjen potrebom reprodukovanja izvornog grafičkog oblika teksta. Za opisivanje grafičke strukture literarnog

teksta, sistem raspolaže sledećim simbolima:
-razmak,
-simbol kraja reda u izvornom tekstu (upotrebljeni karakter je /).Dozvoljeni oblici upotrebe ovog simbola su ili /␣ ili /n␣ (gde je n ⩾2). Drugi dozvoljeni oblik označava da u izvornom tekstu sledi n-1 prazan red.
-simbol paginacije (oznaka &Pn, gde je n broj strane u izvornom tekstu).
-simbol komentara(&), koji dopušta da se delovi teksta navedeni izmedju dva simbola komentara, isključe iz obrade teksta.
-simbolom naslova ($), osim obeležavanja naslova u tekstu, se definiše medjusobni odnos delova jednog ili više tekstova. Simbol naslova se upotrebljava u dva oblika:

$␣ *naslov* ␣ $ ... (a)

$n␣ *naslov* ␣ $ ... (b)

gde *naslov* označava bilo kakav tekst koji ne sadrži znak $ a n prirodan broj, koji ćemo nazvati brojač dubine. Kod književnog teksta, brojač dubine je odredjen kompozicionom strukturom dela (v.npr. Kajzer/73/), tako što se svakoj kompozicionoj jedinici dodeli broj kojim je odredjen položaj kompozicione jedinice prema drugim kompozicionim jedinicama teksta. Tako u kompozicionoj strukturi drame, jedan niz kompozicionih jedinica je drama-čin-scena, kojima se redom dodeljuju vrednosti brojača dubine 1,2,3. Vrednost brojača dubine se odredjuje prilikom pripreme teksta za automatsku obradu na osnovu naslova u tekstu. Ukoliko ih nema poželjno je simulirati ih koristeći paginaciju ili neko drugo obeležje. Ako se pri navodjenju naslova izostavi brojač dubine(oblik (a)), onda se podrazumeva da on ima onu vrednost koju je dobio poslednjim navodjenjem naslova u obliku (b). Inicijalno, brojač dubine je 1.

Na osnovu vrednosti brojača dubine, pot'program NASLOV odredjuje niz brojeva kojim je opisana kompoziciona struktura teksta. S druge strane, na osnovu brojača dubine, brojača redova i brojača reči u redu se formira jedinstvena šifra konteksta pojavljivanja reči u tekstu. Ako je svakoj analiziranoj reči teksta dodeljena šifra konteksta, moguće je jednoznačno rekonstruisati tekst. Broj šifara konteksta koje će se upamtiti uz reč moguće je ograničiti.

Ažuriranje datoteke izvornih tekstova i editovanje tekstova vrši rutina EDIT. Listing ulaznog teksta je prikazan na sl.1 a na sl.2 je editovan deo teksta sa sl.1(a). Korektura tekstova se vrši na osnovu podataka, koje uz editovani tekst, obezbedjuje rutina EDIT.

Kada se u toku analize teksta, izoluje pojavljivanje neke reči, odredjuju se parametri pojavljivanja (dužina, kodovi početnog i završnog slova, šifra konteksta,itd.) da bi se, zatim, u matrici reči, to pojavljivanje ažuriralo(frekvencija, karakteristike konkretnog pojavljivanja kao veliko slovo, rima, itd.). Matrica reči se olančava po parametrima koji su od značaja za dobijanje odredjenih rečnika (osim olančavanja matrice reči po početnim simbolima reči, lančanje se može vršiti i po završnim simbolima reči, dužini, frekvenciji ili po pojavljivanju odredjenih morfema u reči, itd.).

U toku formiranja rečnika, moguće je izvesne unapred zadane reči isključiti iz obrade (npr. vrlo frekventne reči ili funkcionalne reči) ili formirati podrečnike sa odredjenim svojstvom (npr.. rečnik slogova za dati tekst), što se postiže skupom rutina FIXR.

U sistem su inkorporirane i rutina SLOG, za odredjivanje broja slogova u reči i dužine reči u glasovima, i rutina SLOG1, koja reč rastavlja na slogove opciono prema pravilima predloženim u Pravopisu A.Belića, odn. Gramatici srpskohrvatskog jezika M.Stevanovića. Podela reči na slogove rutinom SLOG1 obuhvata prepoznavanje prefiksa ali, naravno, ne i semantičku podelu.

U toku obrade teksta, izračunavaju se oni kvantitativni parametri koji su globalne prirode (dužina rečenice u rečima, dužina stiha, itd.)

Kada je analizirani tekst obradjen i formirana matrica reči konstruišu se pojedini rečnici tako što se lančanjem povezane reči sortiraju. Poredak medju rečima se utvrdjuje logičkom funkcijom RED. Ako su x i y dve reči analiziranog teksta a R poredak u rečniku onda je

$$RED(x,y) = \begin{cases} T, & \text{ako } x \; R \; y \\ F, & \text{inače} \end{cases}$$

Sa ovako ugradjenom funkcijom RED postupak sortiranja reči je nezavisan od poretka koji se zahteva u rečniku. Primer jednog dela direktno i obrnuto uazbučenih konkordanci je prikazan na sl. 3.

Štampanje izlaznih rezultata se može urediti na način koji je korisniku najpogodniji. Pitanje opcija u štampanju kao i lokalnih kvantitativnih analiza (dužina reči, broj slogova, itd.) se rešava u okviru rutine za štampanje.

PRVE PRIMENE I DALJA ISTRAŽIVANJA

Sistem je do sada testiran na više različitih tekstova:

- ciklus pesama "Bdenja" iz zbirke "Sedam lirskih krugova " Momčila Nastasijevića,
- rečnici dečijih definicija najfrekventnijih imenica dečijeg govora (nastavak istraživanja iz (Vasić/73/)). U okviru ovog istraživanja je ispitana primena sistema na ankete sa odgovorima na prirodnom jeziku.
- konkordance i frekvencijski rečnik nad dečijim tekstovima (u toku izrade; oko 150.000 reči)

Za tekst od oko 1000 reči potrebno je oko 20 sekundi procesorskog vremena za izdvajanja reči iz teksta i njihovo uredjivanje u leksikografskom i antileksikografskom poretku. Sistem je stavljen na raspolaganje članovima Seminara za matematičku lingvistiku Matematičkog instituta u cilju objedinjavanja dosadašnjih istraživanja na polju leksikoloških i statističkih istraživanja SH-jezika. Sledeća faza u dogradnji sistema bi trebalo da obuhvati problem modeliranja fonoloških i morfoloških zakonitosti SH-jezika.

BIBLIOGRAFIJA

1 Dom J.Froger /68/: La critique des textes et son automatisation, Dunod, Paris, 1968.

2 Kajzer V./73/: Jezičko umetničko delo,(prevod), SKZ, Beograd, 1973.

3 Lukić V./70/: Aktivni pisani rečnik učenika na osnovnoškolskom uzrastu,Inst.za pedagoška istraživanja, Beograd, 1970.

4 Ničković R, Cvijović M./77/:Dečiji govor,Zavod za udžbenike i nast.sredstva,Beograd, 1977.

5 Plavšić P./78/: Leksičke i semantičke analize frekvencijskih rečnika TV dnevnika i TV drame, u Šipka M./78/

6 Petőfi J./77/: Semantics-pragmatics-text theory, Journal for Descriptive Poetics and Theory of Literature, 2(1977), pp.119-149.

7 Tancig P./77/: Pristup k računalniškemu obravnavanju slovensega jezika,IJS Delovno poročilo,DP-1159, Ljubljana, 1977.

8 Tancig P./78/: Računalniška lingvistika v Sloveniji, u Šipka M./78/

9 Šipka M./78/:(ed.):Kompjuterska obrada lingvističkih podataka,Inst.za jezik i književnost, Sarajevo,1978.

10 Vasić S./73/: Razvojne govorne norme u naše dece,Interno izdanje,Inst.za eksp.fonetiku i patologiju govora,Beograd, 1973.


```
$2 SEDAM LIRSKIH KRUGOVA $ W
$3 BDENJA $ W
$4 MOLITVA $ W
& CVC SU BILA                    TESTIRANJA & W
+SMAGNEM LI CVC DUBINOM U VECERNJU,/ ILI JE TIHI POJ,/ W
IL' DUBINA SE CTVCRI GDE BOLELO=/ +TIHO PO MLCI BRODIM SMERNI RAB.//2 W
+PAKAC MENI CCXE,/ BCLXEZAN, CRACXU NA PLT./ +RASTCCXI C RASTCCXI RABA./ W
+DUBLXE DNC CUSXI NC STRADANXU,/ BEZ DNA RECX GVA SMERNA U VECERNJU./2 W
```

Sl.1(a).Listing ulaznog materijala-literarni tekst

```
$ MAJKA 3 $                                                    $
JE 139 MAJKA 125 NAS 33 OCMACJICA 24 KOJA 24 I 22 BICJE 21 ZSENA 21 RODILA 19
ZSIVC 17 RODITELJ 15 CSUVA 13 ONA 12 ZA 10 DECL 9 NASS 9 SSTC 8 DETE 7 NE 5
```

Sl.1(b).Listing ulaznog materijala-frekvencijski rečnik

```
1. 1. 1. 1.                    MOLITVA


      +SMAGNEM'LI CVC DUBINOM U VECERNJU,
      ILI JE TIHI POJ,
      IL' DUBINA SE CTVCRI GDE BOLELO=
      +TIHO PO MLCI BRODIM SMERNI RAB.

      +PAKAC MENI CCXE,
      BCLXEZAN, CRACXU NA PLT.
      +RASTCCXI C RASTCCXI RABA.
      +DUBLXE DNC CUSXI NC STRADANXU,
      BEZ DNA RECX GVA SMERNA U VECERNJU.
```

Sl.2. Tekst sa sl.1(a) reprodukovan rutinom EDIT

```
1. 1. 1. 1. 1. 1.      $2 SEDAM LIRSKIH KRUGOVA $
1. 1. 1. 1. 1. 1.      $3 BDENJA $
1. 1. 1. 1. 1. 1.      $4 MOLITVA $
1. 1. 1. 2. 1. 1.      $4 GOSPI $
1. 1. 1. 3. 1. 1.      $4 BUZKJAK $
1. 1. 1. 4. 1. 1.      $4 DVE RANE $
1. 1. 1. 4. 1. 1.      $5 - $
1. 1. 1. 4. 2. 1.      $5 - $
1. 1. 1. 5. 1. 1.      $4 OSAMA NA TRGU $
```

Sl.4. Naslovi sa brojačima dubina i odgovarajućim šiframa konteksta

```
 134.  91.IZVIJE                F=    1
...   1.          1 1 2.....1.       +ZAVAPIM,/ AL' IZVIJE SE GLAS.. W
 135.  313.IZGUBE               F=    1
...   1.          101 7....14        IZGUBE STVORENXA,/2 W
 136.    7.ILI                  F=    1
...   1.          1 1 1.....2        +SMAGNEM LI CVO DUBINOM L VICENXU,/ ILI JE TIHI POJ,/ W
 137.  11.IL'                   F=    3
...   1.          1 1 7.2.... 1      +CXEKNI TU PRESANNUTI L PLODU,/ IL' IZA PLAMENA CAR/ W
...   2.          101.6.....3        I OCXI.. BLUDE LI, PAGUBI NA RUB/ IL' SKRUSXENXU./2 W
...   3.          1 1.1.....3        IL' DUBINA SE OTVORI GDE BOLELO=/ +TIHO PO MUCI BRODIM SMERNI RAB./2
 138.  209.IM                   F=    1
...   1.          1 1.59.....1       +TO KUDA NEPROHOD IM,/ CXUDNO MI SE OTVORI PUT./2  W

1004.  401.GRESX               F=    1
...   1.          1010900.....09     +I KOJI ZA MNOM,/ I U NEZNANXU,/ GRESX CVAJ CYDNI PUT../2 W
1005.  462.POHODISX            F=    1
...   1.          1010220200.002     +SVE SAMLXI./ SNOM POHODISX ME TUCXA./ GRESXNIJI KAD SAMOTAN W
1006.  462.PREGORISX           F=    1
...   1.          1011002230.005     +ZXIV PREMREM TOBOM,/ PETVA PREGORISX MNOME/ DO SPASENXA./2 W
1007.  336.NAVESTISX           F=    1
...   1.          1010702000.007     PAKLENI PLAMEN NAVESTISX,/ TI, O JEDINA TI./2 W
```

Sl.3. Konkordance u direktnom i obrnutom ćiriličnom poretku literarnog teksta

```
237.MAJAKA                      3
          1.                  814
          2.                  421
          3.                   83
236.MAJK                        1
          1.                   84
239.MAJKA                     717
          1.                  913
          2.                  726
          3.                  521
          4.                  359
          5.                  137
          6.                    2
240.MAJKAJE                     4
          1.                  369
          2.                   45
241.MAJKE                       2
```

Sl.5. Deo rečnika dobijenog objedinjavanjem 6 frekvencijskih rečnika(sl.1(b)).Prikazane greške su iz izvornog materijala