Željko Bujas
Zagreb

ZAŠTO I KAKO JE NASTALA KOMPJUTERSKA KONKORDANCA MARULIĆEVIH HRVATSKIH DJELA

(Referat održan na Znanstvenom razgovoru o Marku Maruliću u Hrvatskom filološkom društvu u Zagrebu, 17. 12. 1974)

1

Kao što je poznato ovom skupu, vrijeme kad se kompjuter još senzacionalistički nazivao »elektronskim mozgom« već pripada prošlosti. Danas, doznačujući račune za struju, obračunavajući porez na osobne prihode i upućujući opomene zbog zaostalih rata otplate, on se spustio na prozaičnu razinu svakodnevice, postavši u svijesti običnog građanina nekom vrsti elektronskog pandura.

Ponešto od svoje nekadašnje tajanstvenosti kompjuter je sačuvao na jednom području koje, premda znanstveno, ostaje i dalje unutar kruga zanimanja šire javnosti, ostaje dijelom »opće kulture« — na području jezika. Ali dok činjenicu da kompjuter može upravljati tako složenim tehničkim sustavom kakav je svemirsko letalo primamo bez naročitog uzbuđenja, vijesti o pokušajima i relativnom uspjehu kompjutera u prevođenju s jednog jezika na drugi izazivlju u pravilu vrlo određenu reakciju — obično skepsu, a kod »upućenijih«, »senzibilnijih« osoba i porugu. Pogotovu, bojim se, a priori je odbojna obavijest da je kompjuter upotrijebljen za analize književnog teksta, da je taj bešćutni elektronički robot zagazio u takav komorni prostor, kidajući gustu krhku pređu stvaralačkog tkiva. Pa jezik i književni tekst — rezonira se često — nisu matematika i tehnika: njihova je struktura nepravilna i nepredskaziva. Što tu kompjuter može uopće učiniti?

Zaboravlja se pritom da u jezičnoj strukturi ima neslućeno mnogo skrivenih pravilnosti i zakonitosti, bez kojih nije ni književni tekst (sjetimo se samo na formalnu stilističku analizu). A ako je kompjuter pri-

znato djelotvorno pomagalo u egzaktnim analizama s njihovim zakonitim ustrojstvom, nema razloga zašto on ne bi bio jednako djelotvoran u analizi zakonitosti u jezičnoj i književnoj građi.

Ne zaboravimo dalje — a to se obično ne zna — da su jezična, a ponekad i književna istraživanja najvećim dijelom dugotrajno i zamorno, repetitivno, polumehaničko prikupljanje, sortiranje i kategoriziranje građe. Doduše to su sve neophodni poslovi i faze obrade, ali u biti tek pripremni radovi, na koje se nažalost često utroši i 80—90 posto vremena. A za prave kreativne analize s mogućim konačnim sintezama preostane onda najmanje vremena, pa i osjetno smanjena energija. Međutim upravo kompjuter, taj elektronički sluga suvremene civilizacije, sa svojom gotovo neshvatljivom sposobnošću da obavi astronomski broj repetitivnih logičkih operacija prepoznavanja, usporedbi i kategorizacija u jednako nevjerojatno kratkom vremenu, jedinstveno je iskoristiv za spomenute neizbježne predradnje.

Oslobađajući nas dakle najvećeg dijela napora oko pripremnih poslova i početnog mehaničkog razvrstavanja građe, kompjuter nam omogućuje da se ili mnogo ranije i bezbolnije pozabavimo pravim kreativnim analizama, ili da se upustimo u sve one opsežne istraživačke pothvate koje prije nismo ni počinjali, znajući da ih sami, ručnom obradom, ne bismo nikad dovršili. Dakle upravo taj mehanički robot demehanizira istraživača, omogućujući mu da bude istinski kreativnim, humanim.

2

Ja nisam kroatist, i moja prva iskustva s kompjuterom kao pomagalom za istraživanje teksta bila su još 1963, pri morfosintaktičkoj analizi dvojezičnog stručnog teksta (engleskog izvornika i hrvatskog prijevoda). Radeći na doktorskoj tezi, šifrirao sam tada uzorak od 32.432 riječi teksta pomoću 169 slovčano-brojčanih kodova. A budući da je posebni naglasak analize bio na redu riječi — kao problemu hipotetskog kompjuterskog prevođenja stručnog teksta s engleskog na hrvatski — trebalo je ispitivati strukture, koje su ti kodovi predstavljali, na svakom od mogućih rednih mjesta u 1302 rečenice uzorka. Prosječna dužina rečenice bila je doduše 20 riječi, ali najduža od njih imala je čak 98 mjesta. Ručno tabeliranje takve građe potrajalo bi vjerojatno jednu godinu. Tabulator mehanografskog centra u Radi Končaru — tada naime u Zagrebu nije još postojao niti jedan pravi, potpuni, kompjuterski sistem — izradio je međutim za samo pet dana 4.308 tabelarnih pregleda građe. Ove tabelarne liste su dokraja iscrpno i precizno — abecednim i brojčanim redom — inventirale sve strukture teksta prema osam mogućih kategorija promjena u redu riječi nastalih prevođenjem, na svakom mogućem rednom mjestu u rečenici i prema strukturi bližeg konteksta.

Sve to nepokolebivo me je uvjerilo u velike prednosti sličnog inventiranja svakog opsežnog teksta koji želimo podvrgnuti temeljitoj analizi. Već tada učvrstila se u meni zamisao o pokušaju da se tako obrade naj-

važniji tekstovi hrvatske književnosti i otvore intenzivnim jezičnim i književnim istraživanjima. I ranije bili su mi poznati slični, ručno izrađeni, inventari velikana engleske književnosti, Šekspira i Chaucera. Radilo se dakako o *konkordanc(ij)ama*, dakle o tako inventiranim riječima nekog teksta da je svaka riječ (bez obzira na to koliko se puta ponavljala, gdje se javljala i u kojom obliku dolazila) navedena abecednim redom i uz ograničeni kontekst te naznaku lokacije.

Moguće primjene takvog pregleda postaju jasne svakom tko ga uzme u ruku. Svaka se riječ konkordiranog teksta može vrlo brzo i — bez opasnosti ikakvog previda — pronaći sa svim svojim ponavljanjima u cijelom, pa i najopsežnijem, tekstu. Ona se zatim, jednako brzo i jednostavno (prelijećući abecedirane liste očima), može uspoređivati s drugim riječima i kategorizirati na najrazličitije načine prema njezinim kolokacijama, to jest prema neposrednom, najužem kontaktu. Isto tako mogu se, uz minimalni napor i utrošak vremena, mogu se nepogrešivo pronaći sve reference u konkordiranom tekstu na osobe, mjesta, tematske riječi. Cijeli niz sličnih analitičkih postupaka upravo se nameće korisniku konkordance.

Međutim, konkordance Šekspira i Chaucera, izrađivane ručno u proteklom stoljeću, bile su nevjerojatno dugi i mukotrpni pothvati. Prva je nastajala 16 a druga čak 52 godine (1872—1924). Danas, zahvaljujući kompjuteru, takvi pothvati u pravilu ne traju više od nekoliko mjeseci (kod tekstova prosječne dužine). Izrada kompjuterskih konkordanca je danas jedan od najčešćih oblika kompjuterske obrade teksta, i u proteklih deset godina izrađeno je bar 150 opsežnih kompjutrskih konkordanca.

3

Prilika da izradim prve kompjuterske konkordance hrvatskih tekstova pružila mi se u toku školske godine 1967/68, kad sam se, kao stipendist zaklade Ford Foundation našao na University of Texas (u gradu Austinu), u Institutu Linguistics Research Center (u daljem tekstu: LRC). Svrha mog sedmomjesečnog boravka u LRC — čije se osoblje pretežno bavilo problemima kompjuterskog prevođenja — bila je doduše razrađivanje metodologije kompjuterskog konkordiranja za potrebe zagrebačkog projekta kontrastivne analize. Ali iznimno povoljne prilike za rad u tom institutu, te susretljivost direktora LRC-a, prof. W. P. Lehmana i ostalog osoblja, nametale su mi i moralnu obavezu da predviđeni boravak na University of Texas što šire iskoristim, posebno za hrvatska jezična i srodna istraživanja. Najkorisnije što sam u danim okolnostima mogao učiniti, zaključio sam, bilo je da pokušam izraditi prve kompjuterske konkordance što značajnijih hrvatskih tekstova. Time bi se ujedno pridonijelo stvaranju povoljnije klime za brže i šire prihvaćanje kompjutera i na području humanističkih znanosti u nas.

Od fakultetskog kolege prof. Milana Moguša ubrzo sam, prema ranijem dogovoru, dobio tekst Marulićeve *Susane* koga nije bilo u sveučilišnoj knjižnici University of Texas. Iz skučenog izbora hrvatskih djela u toj knjižnici lako sam odredio još dva teksta za kompjutersko konkordiranje: Gundulićevog *Osmana* i, kao prozni suvremeni tekst, Krležin roman *Povratak Filipa Latinovicza* (uz njegov prijevod na engleski radi kasnijih kontrastivnih analiza).

Gledano unatrag, radilo se zaista o velikom zalogaju, jer se ukupno konkordiralo 202.583 riječi teksta, od čega 126.626 (ili 62,5%) na hrvatskom — što je za tamošnje osoblje predstavljalo posebne teškoće. Kao što je ovom skupu možda poznato, svaki tekst koji se želi kompjuterski obraditi treba najprije preobličiti tako da ga kompjuter može uopće primiti. To se najčešće radi tako da se tekst prepisuje na bušene kartice. Svaka bušena kartica, na koju stane desetak riječi teksta i oznaka mjesta u tekstu, nekoliko se puta provjerava dok se ne uklone sve greške nastale pri samom ubušavanju. Zanimljivo je da tih grešaka nije bilo više nego obično, premda je tamošnjim bušačima materinji jezik bio engleski odnosno španjolski (dvije teksaške Meksikanke). Poseban problem i teškoća bili su u tome što su te djevojke u toku ubušavanja morale same transliterirati hrvatski tekst. Američki kompjuteri, prilagođene engleskom jeziku, nisu imali posebnih znakova za hrvatske dijakritike, pa su se oni zamjenjivali dvoslovima prema formuli: č = cy, ć = cz, đ = dz, š = sz i ž = zz.

Susana, sa svojih 5.333 riječi teksta — dakle ni 3% ukupne tekstualne mase koja se ubušavala u LRC — bila je najranije pripremljena za konkordiranje (u ožujku 1968), ali je pretvorena u kompjutersku konkordancu tek potkraj svibnja iste godine. Tada su u jednoj noćnoj smjeni odjednom izrađene konkordance ovih tekstova:

1. *Susana,* 5.333 riječi (89 str. konkordance)
2. *Susana,* odostražno
3. *Osman,* 56.134 riječi (936 str.)
4. *Povratak Filipa Latinovicza,* 65.159 riječi (1.086 str.)
5. *The Return of Philip Latinovicz,* 75.957 riječi (1.266 str.)

Ukupno je dakle konkordirano 207.916 riječi teksta i dobiveno 3.466 stranica kompjuterskih konkordanca.

4

Sva ta gomila papira stigla je u Zagreb u studenom 1968, izazvala čuđenje carinika i bila pokazana većem broju kolega na Filozofskom fakultetu. Uz razumljivo zanimanje koje je taj materijal izazvao, pale su i prve napomene da te konkordance neće odigrati punu ulogu sve dok ne postanu pravom svojinom hrvatske, i šire, znanstvene i kulturne javnosti — jednom rječju dok se ne objave.

Tri slijedeće godine, 1969—1971, bio sam međutim intenzivno zauzet vođenjem kompjuterske obrade (s konkordiranjem kao konačnom fazom) u već spomenutom zagrebačkom projektu kontrastivne analize. Za potrebe tog projekta izrađena je naime konkordanca gramatički šifriranog engleskog korpusa od 505.823 riječi. (U toku 1970. ponudio sam doduše kompjutersku konkordancu *Osmana* Jugoslavenskoj akademiji da je izda, ali je konačni odgovor bio negativan.)

U toku 1972. koncipirana je, s kolegom Mogušem, kao znanstveni projekt zajednička zamisao da se glavni tekstovi hrvatske pismenosti i književnosti — od Baščanske ploče do Hrvatskog preporoda — prirede za kompjutersku obradu i od njih izrade kompjuterske konkordance. Konačan cilj takvog projekta bio bi niz izrađenih i objavljenih konkordanca, čije bi samo postojanje umnogostručilo realni potencijal znanstvenog istraživanja te građe, nametnuvši ujedno strogu egzaktnost pristupa i obavezu znatno više akribijske razine analiza. Zajednički elaborat podnijet je Republičkom savjetu za naučni rad SRH u okviru petogodišnjeg plana (1971—1975) Instituta za lingvistiku Filozofskog fakulteta u Zagrebu (u daljem tekstu: ILFFZ). Elaborat je prihvaćen i u rujnu 1972. pokrenut je sredstvima iz Fonda za naučni rad SRH znanstveni projekt *Kompjuterska analiza tekstova stare hrvatske književnosti.* Osim povoljnoj klimi za kompjuterska istraživanja, projekt — što treba posebno istaknuti — duguje mnogo i neumornom zagovaranju i širini pogleda direktora ILFFZ, prof. dr Rudolfa Filipovića.

Sunosioci projekta su Ž. Bujas (koordiniranje kompjuterske obrade) i M. Moguš (kroatistička strana posla), a od početka rada postoji i stalni asistent projekta. Prvi asistent — od rujna 1972 do rujna 1973 — bila je dr Dunja Jutronić, sada docent na Katedri za engleski jezik i književnost na Filozofskom fakultetu u Zadru. Drugi asistent, Maja Bratanić-Čimbur, diplomirani anglist i hispanist, a sada radi na projektu.

5

Kako je dio Marulićevog opusa već bio zahvaćen izradom ranije spomenutih konkordanca *Susana* (normalne i odostražne), rad u novom projektu počeo je priređivanjem za kompjutersku obradu ostalih hrvatskih djela »oca hrvatske književnosti«. Opis tog postupka, kao i same kompjuterske obrade zavređuju poseban osvrt. Takav se osvrt upravo priprema i uskoro će se pojaviti u časopisu »Suvremena lingvistika«. Ovom prilikom prikazat ću te poslove tek u glavnim crtama.

Pošli smo od teksta Marulićevih hrvatskih djela iz edicije »Pet stoljeća hrvatske književnosti« (priredio Ivan Slamnig). Tom korpusu je kolega Moguš dodao Marulićeve dvije prozne poslanice te njegove bilješke i kratak prozni uvod u *Susanu.*

U pripremnoj fazi, prije početka samog ubušavanja moralo se riješiti mnoštvo tehnički detalja. Među njima: da li numerirati stihove u konti-

nuitetu cijelog sveska ili unutar pojedinih, pa i kratkih pjesama? Kako osigurati da dugi opisni naslovi i podnaslovi budu u konačnom konkordiranom tekstu uočljivo različiti od numeriranih stihova? Kako uklopiti u ostali kontekst lica povremenog dijaloga? Kako odrediti i numerirati retke proznih dijelova građe (uvoda u *Juditu,* poslanica, autorovih bilježaka)?

Sva ta pitanja i mnoštvo drugih tehničkih detalja rješavani su u stalnom dogovoru s dipl. ing. Milutinom Cihlarom, rukovodiocem sektora za programiranje u Elektroničkom računskom centru Centra za ekonomski razvoj grada Zagreba (u daljem tekstu: ERC CER), gdje se vršila kompjuterska obrada ubušenog teksta i konačno konkordiranje. Moja ranija trogodišnja (1969—1971) suradnja s njim na izradi spomenute mamut-konkordance (10 svezaka po 1.000 str.) za zagrebački kontrastivni projekt, nesumnjivo je znatno olakšala suradnju na ovom sličnom, premda u mnogočem specifičnom, pothvatu. Ipak ne može se prenaglasiti spremnost na suradnju, bezrezervno uživljavanje u problem i širina gledanja ing. Cihlara kome *Kompjuterska konkordanca hrvatskih djela Marka Marulića* mnogo duguje.

Samo pripremanje teksta za kompjutersku obradu teklo je ovim redom:

1) Asistent projekta priredio bi tekst za ubušavanje, unoseći rukom oznake početka i završetka grafičkog bloka teksta i redni broj stiha. (Prozne dijelove teksta sam kopjuter razlaže u retke ograničene dužine i mehanički ih numerira.) Istovremeno se jednoobrazno označavaju i podnaslovi, tekst bilježaka, lica u dijalozima i slični marginalni dijelovi teksta.

2) Tako priređen tekst prepisivao se zatim na fleksorajteru (Flexowriter), što je poseban električni pisaći stroj, koji automatskih buši papirnatu traku sličnu teleprinterskoj. Tako se identičan prepisani tekst nalazi na papiru — kao i kod svakog drugog pisaćeg stroja — i na bušenoj papirnoj traci. Svrha dobivanja bušene trake je u tome što tako prepisan tekst kompjuter može čitati. Tekst se može i drukčije prirediti za kompjutersku obradu, na primjer (a to je i najčešći slučaj) prepisivanjem na bušene kartice. To međutim zahtijeva obučeno osoblje i posebne ormare za pohranu kartica, dok je prednost fleksorajtera za mali institut kakav je ILFFZ upravo u njegovoj kompaktnosti i lakoći s kojom se pohranjuju koluti bušene trake. Svi prepisivači Marulićevog teksta na institutskom fleksorajteru bili su studenti Filozofskog fakulteta u Zagrebu: Iskra Devčić, Ivo Lozica, Ljiljana Ostojić i Zlatica Skender.

3) Izbušene trake nosile su se zatim u ERC CER. Tu su se one najprije presnimile na magnetsku traku radi znatno brže i fleksibilnije kasnije obrade, te neusporedivo veće kompaktnosti podataka na tom mediju. S dobivene magnetske trake tekst se tada otiskivao kao kompjuterski ispis (široke liste papira) i vraćao u ILFFZ.

4) Svrha toga bila je da asistent projekta sada kolacionira kompjuterski ispis iz ERC CER — dakle ubušeni tekst — s izvornim tiskanim tekstom iz edicije »Pet stoljeća«. Sve uočene greške označavane su tada na kompjuterskom ispisu (slično tiskarskim korekturama). Korigirani ispis vraćao se zatim u ERC CER.

5) Sustavne greške — na primjer zaboravljeni razmak nakon znakova interpunkcije — ispravljale su se sada u ERC CER na samoj magnetskoj traci pomoću posebnih kompjuterskih programa.

6) Sadržaj tako ispravljene trake prenosio se tada na bušene kartice. Tako su se doduše dobili identični tekstovi na dva razna medija, ali to ima svoje puno praktično opravdanje. Radi se o tom da se uočene *nesustavne* greške najprikladnije uklanjaju tako da se bušena kartica s odsječkom teksta koji sadrži grešku izvadi fizički iz sekvence ostalih kartica, umjesto nje izbuši nova i uloži na njezino mjesto.

7) Sadržaj svih bušenih kartica ponovo se zatim otiskivao kao kompjuterski ispis. Ovaj se opet slao u ILFFZ gdje se kolacionirao s ranijim ispisom (prvom korekturom), da se vidi nije li pri bušenju novih, popravljanih, kartica došlo ponovo do greške u samom bušenju. Na tom drugom ispisu označavale su se sada te nove greške, još jednom su u ERC CER zamjenjivane kartice — i čitav se postupak ponavljao sve dok asistent projekta i prof. Moguš nisu dali konačni »imprimatur«.

8) Definitivno ispravljen tekst vratio se sada u ERC CER na magnetsku traku; uključen je program konkordiranja i, za nekoliko sati, tekst hrvatskih djela Marka Marulića razložen je i zatim preustrojen u svoj konačni oblik kompjuterske konkordance. Nakon toga uključen je program tiskanja i brzotiskač kompjuterskog sistema IBM 360/170 — da konačno spomenemo ime i tog skromnog suautora — otisnuo je za svega 83 minute *kompjutersku konkordancu hrvatskih djela Marka Marulića.*

6

U tom svom konačnom obliku, koji možete vidjeti na ovoj maloj izložbi, ta konkordanca obuhvaća 598 stranica velikog kompjuterskog formata. Središnjim stupcem svake stranice teku abecednim redom konkordirane riječi (»stožernice«), praćene lijevim i desnim neposrednim kontekstom ograničenim maksimalnom širinom kompjuterskog ispisa (132 znaka), dok posebni stupac na lijevom rubu stranice pokazuje točno mjesto u izvornom tekstu gdje se stožernica (ključna riječ retka) javlja. Kako na svakoj takvoj stranici imamo 60 redaka, osim zadnje, nepotpune, konačni točan broj stožernica — dakle ukupni broj riječi u potpunom tekstu Marulićevih hrvatskih djela iznosi 35.914.

Ovaj svezak ima zapravo 649 str. jer uključuje 51 str. uvoda koji sadrži puni tekst što se konkordira, popraćen točnom numeracijom svakog stiha i proznog retka. To je učinjeno zato da se olakša konzultiranje šireg

konteksta od onog što ga pruža prostorno ograničen redak konkordance. Ovo je bilo potrebno u raznorodnom tekstu kakav je Marulićev, dok se to moglo zanemariti u jedinstvenom tekstu Gundulićevog *Osmana* s njegovim standardno numeriranim recima. Ta se kompjuterska konkordanca i objavljuje zato bez takvog uvodnog dijela.

Tok svih upravo opisanih postupaka i poslova prikazan je ovdje, nadam se na shvatljiv način, s nekoliko eksponata svaki od kojih ilustrira po jednu fazu obrade. Premda je čitav proces nužno simplificiran, vjerujem da ćete ipak steći stanovitu predodžbu o opsežnosti poslova. Posebno su pojednostavljene faze obrade u ERC CER, dakle na samom kompjuterskom sistemu. Svakoj od njih je, ne zaboravimo, predstojalo zamorno pisanje programa i njihovo često dugotrajno testiranje. Ovaj debeli svezak kompjuterskih ispisa — dvaput deblji od same konkordance uz koju stoji — sadrži isključivo glavne i pomoćne programe, ispise pokusnih obrada, dorade i dopune programa i slično. On sam najrječitije govori o svoj toj skrivenoj, malo poznatoj, a nezaobilaznoj strani svakog kompjuterskog pothvata.

7

Pred nama je dakle, i fizički prisutan, konačni proizvod svih tih ne tako kratkih i ne tako jednostavnih poslova o kojima sam govorio — *Kompjuterska konkordanca hrvatskih djela Marka Marulića.* Nastajala je doduše preko godinu dana, trpeći od većine dječjih bolesti svakog sličnog inovativnog pothvata. Ali tu je. Koje su njezine svestrane mogućnosti, kakve kvalitetno nove analize ona omogućuje, najbolje će pokazati kroatisti i drugi stručnjaci kojima je namijenjena kao djelotvorno pomagalo. Izlaganja kolege Moguša »Je li Marko Marulić autor Firentinskog zbornika?« — koje kronološkom i metodološkom logikom slijedi odmah za mojim — prikazat će to na najbolji mogući način. Ono će, uvjeren sam, i ovom znanstvenom skupu jasno predočiti neizbježnost naredne faze našeg započetog posla — objavljivanja izrađenih kompjuterskih konkordanca. Ranije dovršena, i u toku 1974. u ERC CER dorađena (vraćanjem dijakritika) kompjuterska konkordanca Gundulićevog *Osmana* nalazi se pred objavljivanjem — tehnikom pretiska — u zagrebačkom Liberu, zahvaljujući značajnoj novčanoj pomoći Republičkog savjeta za naučni rad SRH. Stojimo pred potpisivanjem ugovora za slično izdavanje Marulićeve hrvatske konkordance.

U ovom času u ILFFZ pripremljena su za kompjutersku obradu, i praktički se nalaze pred konkordiranjem, sva djela starih hrvatskih pisaca hvarskog i zadarskog kruga. Završeno je ubušavanje ostalih Gundulićevih djela. čitavog Bunićevog opusa, svih djela Marina Držića. Počinje ubušavanje Ranjininog zbornika. Projekt *Kompjuterske analize tekstova stare hrvatske književnosti* ulazi, kako vidite, u puni zamah. Kolega Moguš i ja ne ćemo posustati. Svjesni smo ipak da od toga kako će naša

znanstvena, kultura i stručna javnost prihvatiti prve objavljene kompjuterske konkordance *Osmana* i, neposredno za njom, *Hrvatskih djela Marka Marulića* u mnogome zavisi brzina izlaženja drugih svezaka. A bez niza *objavljenih* kompjuterskih konkordanca teško će doći do onog kvalitetnog, trajnog učinka u jezičnim i književnim istraživanjima središnjim za hrvatsku nacionalnu kulturu, što ga taj novi priručnički rod omogućuje.