UDK 808.62-323.1
Izvorni znanstveni članak
Primljen 29. XI. 1997.
Prihvaćen za tisak 2. III. 1998.

Milan Moguš
Hrvatska akademija znanosti i umjetnosti
Zrinski trg 11, HR-10000 Zagreb

O HRVATSKOM ČESTOTNOM RJEČNIKU

Autor iznosi osnovne probleme koji su se javljali pri izradi *Hrvatskoga čestotnoga rječnika* što je nedavno izrađen u Zavodu za lingvistiku Filozofskoga fakulteta u Zagrebu.

Čestotni rječnici već imaju svoju stogodišnju povijest: prvi je takvo djelo izradio F. W. Kaeding pod naslovom *Häufigkeitswörterbuch der deutschen Sprache* i objavio ga u Steglitzu 1897. godine. U Hrvatskoj je također izrađen jedan čestotni rječnik. Riječ je o djelu Ivana Furlana koje je bilo sastavnim dijelom njegove veće radnje *Raznolikost rječnika i struktura govora*.[1] Furlanov čestotnik nije velik: ima deset uzoraka po deset tisuća riječi, tj. korpus mu je obuhvaćao sto tisuća riječi. »Težište je Furlanove radnje istraživanje raznolikosti rječnika i strukture govora, zapravo pismenog izražavanja, modelom frekvencije kod djece školskoga uzrasta, u odnosu na model rječnika«.[2] Nakon Furlanova čestotnoga rječnika pojavio se još jedan na ograničenu korpusu (*Vjesnikov Čestotni rječnik*, 1983), a nakon njega nisu se u Hrvatskoj izrađivali takvi rječnici iako je potreba za njima bila velika.[3]

U odnosu na tadašnja istraživanja Furlanov je frekvencijar imao jednu prednost — izrađen je na osnovi stvarnih, potvrđenih (pisanih) tekstova. Bila je to u neku ruku novost jer su se prije tridesetak godina gramatike (pogotovu školske) izrađivale "napametno", tj. bez stvarnih, potvrđenih primjera, a ni s rječnicima nije bilo mnogo bolje. Zato mi se činilo sasvim prirodnim da kao jedan od plodova projekta *Korpus suvremenoga hrvatskoga književnog jezika*, na kojem sam, zajedno s glavnim suradnicima dr. Majom Bratanić i dr. Markom Tadićem, radio sedamdesetih godina u Zavodu za lingvistiku Filo-

[1] Ivan Furlan, *Raznolikost rječnika i struktura govora*, Zagreb 1961.

[2] Mijo Lončarić, O čestotnim rječnicima i čestotniku hrvatskog književnog jezika, *Suvremena lingvistika* 15—16, Zagreb 1977, str. 43.

[3] Bilo je potrebe i za izradom drugih rječnika suvremenoga hrvatskoga jezika, jednosveščanoga ponajprije, ali političke okolnosti nisu tomu pogodovale.

zofskoga fakulteta u Zagrebu, bude izrađen čestotni rječnik hrvatskoga književnog jezika na temelju korpusa.[4] U ovom ću se referatu usredotočiti na metodologiju rada i na neke spoznaje do kojih smo u tijeku rada došli.

Trebalo je, ponajprije, odrediti neke parametre našega budućega čestotnika. To su: 1. ukupna veličina korpusa, 2. vrsta tekstova koji ulaze u korpus, 3. vremenski raspon tekstova, 4. problem potkorpusa. Nijedan od spomenutih parametara nije se razmatrao zasebno: jedna je odluka gotovo uvijek povlačila za sobom drugu.

U rješavanju naznačenih problema činilo nam se, s obzirom na iskustva drugih sredina tada, najprihvatljivijim da naš korpus obuhvati 1 milijun riječi-pojavnica, tj. 1 milijun riječi koje se u tekstu pojavljuju slijedeći jedna drugu. To je u to doba bio dovoljno reprezentativan korpus, a, usput govoreći, bio je 10 puta veći od spomenutoga Furlanova čestotnoga rječnika hrvatskih riječi. Takva je odluka o veličini korpusa bila povezana s odlukom o izboru potkorpusnih uzoraka. Naime, budući da smo tada, sedamdesetih godina našega stoljeća, stvarali u Zavodu za lingvistiku korpus suvremenoga hrvatskoga književnoga jezika koji bi mogao biti temeljem svekolikomu našemu budućem proučavanju, smatrali smo najprikladnijim da i naš čestotni rječnik bude odraz toga korpusa. A to je pak značilo, u konkretnomu slučaju, da smo krenuli od Krležine književne pojave tridesetih godina dvadesetoga stoljeća, od njegova *Povratka Filipa Latinovicza*, do godine pravoga početka našega rada, tj. do 1975. Omeđili smo dakle korpus na tekstove u rasponu od 45 godina veoma plodnoga rada hrvatskih književnika. Nije bilo sumnje da to jest hrvatski književni jezik. Time je ujedno bila riješena dilema koja se tada mogla postaviti: naime, treba li naš rječnik obuhvatiti građu cjelokupnoga »hrvatskoga ili srpskoga jezika« ili samo građu tzv. »hrvatske varijante«. Čak i oni koji su smatrali, doduše najčešće teoretski, da bi bilo »idealno kad bi se paralelno mogli raditi čestotnici svih standardnih idioma kojima je osnova štokavsko narječje« jer bi to »pružalo mogućnosti zanimljivim usporedbama i saznanjima«,[5] sugerirali su ipak da je »u našim prilikama« bolje »izraditi čestotnik za jedan od tih kodova koji nama i pripada, tj. čestotnik hrvatskoga književnoga jezika«.[6] Za nas ta dilema uopće nije postojala i odluka je bila jasna. Nakon te odluke slijedile su druge.

Smatrali smo da korpus od 1 milijun riječi-pojavnica može biti pogodan za raznovrsna lingvistička istraživanja ako bude stilski što raznolikiji. Jer, i drugi su pripominjali da je pri utvrđivanju i proučavanju frekvencije riječi u ko-

[4] Reprezentativni korpus hrvatskoga književnoga jezika omogućio bi egzaktnija jezična istraživanja na raznovrsnim jezičnim razinama, pa tako i za leksikografske studije.

[5] M. Lončarić, n.dj., str. 44.

[6] Isto.

jemu jeziku važno uzimati građu iz raznovrsnih stilskih područja. Zato smo svoj milijunski korpus (nazvan »Mogušev korpus«) podijelili na 5 jednakih potkorpusa, tako da svaki potkorpus obuhvaća 200.000 riječi-pojavnica. To su: proza, poezija, drama, udžbenički tekstovi i novine.[7] Svaki je potkorpus dobio svoju oznaku, i to: *D* (drama), *N* (novine), *P* (proza), *S* (stihovi), *U* (udžbenici).

Javio se još jedan problem: kako ispuniti količinu od 200.000 riječi-pojavnica za svaki potkorpus. Moralo se ići u prethodna istraživanja. Kao pokusni tekst odabrali smo dio romana *Divota prašine* Vjekoslava Kaleba. Upisan je u računalo, konkordiran i frekvencijski obrađen najprije uzorak od 5.000 kontinuiranih riječi-pojavnica, zatim isto tako uzorak od 10.000 riječi-pojavnica i onda uzorak od 20.000 riječi-pojavnica. Prvi se uzorak pokazao sasvim nedostatan za iole ozbiljnije zaključivanje o frekvenciji riječi nekoga autora. Drugi, od 10.000 riječi-pojavnica, davao je daleko vjerodostojniju sliku uporabe pojedinih riječi. Međutim, povećanje je uzorka na 20.000 riječi, tj. za dvostruko više od 10.000, pokazalo da je, unatoč stopostotnomu povećanju riječi-pojavnica, broj novih natuknica povećan tek za nešto manje od deset posto. Taj je zaključak bio važan jer je pružao mogućnost da se u potkorpusima proze, poezije i drame poveća broj autora na 20, svaki sa po 10 tisuća riječi kontinuiranoga teksta. Time je ujedno bila osigurana u većoj mjeri različitost tekstova s obzirom na još jednu odluku: da se svaki autor, u pravilu, pojavi u jednomu potkorpusu svojim tekstom samo jedanput. Na taj su se način, iznimno, neki autori mogli naći i u dva, stilski različita, potkorpusa.

Naš se izbor od 20 autora kretao u prozi od Miroslava Krleže do Krste Špoljara i Čede Price, u poeziji od Dragutina Tadijanovića do Josipa Pupačića, a u drami od Milana Begovića do Ivana Raosa. U računalo je upisano 10 tisuća riječi kontinuiranoga teksta svakoga od ovih pisaca.[8] Navodimo njihov poredak onako kako su uzorci poredani u potkorpuse.

Prozni tekstovi:

1. Cesarec, August: Novele
2. Desnica, Vladan: Proljeća Ivana Galeba
3. Dončević, Ivan: Mirotvorci
4. Jeličić, Živko: Ljetnih večeri
5. Jelić, Vojin: Anđeli lijepo pjevaju
6. Kaleb, Vjekoslav: Na kamenju
7. Kolar, Slavko: Mi smo za pravicu
8. Kozarčanin, Ivo: Sam čovjek

[7] Za takvu su se ili sličnu raspodjelu na potkorpuse zalagali i drugi autori čestotnih rječnika.

[8] Ovim se popisom nipošto ne izriče vrijednosni sud o upisanim tekstovima.

9. Krleža, Miroslav: Povratak Filipa Latinovicza
10. Majer, Vjekoslav: Život puža
11. Marinković, Ranko: Ruke
12. Simić, Novak: Braća i kumiri
13. Slamnig, Ivan: Neprijatelj
14. Novak, Slobodan: Mirisi, zlato i tamjan
15. Šoljan, Antun: Izdajice
16. Šegedin, Petar: Djeca božja
17. Špoljar, Krsto: Vrijeme i paučina
18. Prica, Čedo: Nekoga moraš voljeti
19. Božić, Mirko: Kurlani
20. Franičević-Pločar, Jure: Raspukline

Stihovi:

1. Tadijanović, Dragutin: Sabrane pjesme
2. Cesarić, Dobriša: Sabrane pjesme
3. Stamać, Ante: Dešifriranje vage
4. Ujević, Tin: Žedan kamen na studencu
5. Šop, Nikola: Dok svemiri venu
6. Parun, Vesna: Zore i vihori
7. Krklec, Gustav: Pjesme, epigrami i basne
8. Milićević, Nikola: Ruke pune mošta
9. Vučetić, Šime: Pjesništvo
10. Kovačić, Ivan Goran: Jama
11. Vitez, Grigor: Kad bi drveće hodalo
12. Popović, Vladimir: Oči
13. Franičević, Marin: Nastanjene uvale i dr.
14. Kaštelan, Jure: Crveni konj i druge zbirke
15. Mađer, Miroslav Slavko: Lelije, zelene lelije
16. Mihalić, Slavko: Izabrane pjesme
17. Petrak, Nikica: Suho slovo
18. Ivanišević, Drago: Ljubav
19. Slaviček, Milivoj: Predak
20. Pupačić, Josip: Sabrane pjesme

Dramski tekstovi

1. Begović, Milan: Bez trećega
2. Mesarić, Kalman: Gospodsko dijete
3. Senečić, Geno: Slučaj s ulice
4. Muradbegović, Ahmed: Bijesno pseto
5. Feldman, Miroslav: U pozadini
6. Budak, Pero: Mećava
7. Gervais, Drago: Karolina riječka
8. Raos, Ivan: Dvije kristalne čaše
9. Marinković, Ranko: Glorija
10. Kolar, Slavko: Svoga tela gospodar

11. Matković, Marijan: Heraklo
12. Krleža, Miroslav: Aretej
13. Božić, Mirko: Pravednik
14. Ivanac, Ivica: Odmor za umorne jahače
15. Šoljan, Antun: Brdo
16. Fabrio, Nedjeljko: Reformatori
17. Hadžić, Fadil: Političko vjenčanje
18. Supek, Ivan: Heretik
19. Brešan, Ivo: Predstava Hamleta u selu Mrduša Donja
20. Kušan, Ivan: Vaudeville

Potkorpus stručnoga (udžbeničkoga) teksta obuhvatio je kontinuirane tekstovne izvatke od po 3.390 riječi-pojavnica iz 58 udžbenika koji su se tada upotrebljavali u završnim (maturalnim) razredima srednjih škola.

U potkorpus novina ušli su tekstovi od po 25 tisuća riječi tiskani istoga dana (21. u mjesecu) i iz iste godine, tj. 1975., i to iz zagrebačkoga *Vjesnika* (travanj, lipanj, rujan, prosinac), iz splitske *Slobodne Dalmacije* (ožujak), iz riječkoga *Novoga lista* (ožujak), iz osječkoga *Glasa Slavonije* (ožujak), iz zagrebačkoga izdanja *Borbe* (ožujak). Budući da je samo zagrebački *Vjesnik* ispunjavao spomenutu kvotu (ostale novine nisu toga dana imale 25.000 riječi), potkorpus novina dopunjen je razlikom do 200.000 iz *Vjesnika* (ožujak 1977.) i *Večernjega lista* (kolovoz 1977. godine).

Usput treba spomenuti da je u potkorpusu stihova, s obzirom na utvrđenu količinu teksta, bilo iznenađenja. Naime, potpuni opus pojedinih pjesnika, npr. Dobriše Cesarića, bio je manji od 10.000 riječi-pojavnica. Budući da se nije radilo o velikim "manjkovima", nadoknađivali smo ih razmjernim povećanjem broja riječi kod drugih pjesnika koji su ušli u popis. A dobili smo i jedan dodatni proizvod – kompletne konkordancije za one pjesnike čiji opus nije prelazio 10.000 riječi. Tako je, između ostalih, izrađena konkordancija potpunih Cesarićevih i tadašnjih Tadijanovićevih pjesničkih djela.

Pošto je sva spomenuta građa od nešto više od 1 milijun riječi (točno 1,001.748) upisana u računalo, nastupio je dugotrajan i mukotrpan posao lematiziranja. Razumije se da je najprije trebalo smisliti programe za takve računalne obrade,[9] a potom tu obradu i provesti.

Radeći na lematizaciji, susretali smo se s nizom problema. Ne rangirajući ih po važnosti, iznosimo ih više-manje onim redom kojim su navirali.

Smatrali smo da u rječnik hrvatskoga jezika, pa bio on i čestotni, ne treba unositi strane riječi, izričaje i fraze, npr. u Krležinim njemačkim ili latinskim rečenicama. Međutim, u korpus su ušle fonološki i morfološki adaptirane posuđenice (npr. *jahta, pamflet*). Izostavili smo također sve domaće i strane antroponime i toponime, odnosno one leksičke jedinice koje se mogu smatrati

[9] Programe za računalne obrade izradio je kolega Marko Tadić.

vlastitim imenima, a prepoznaju se velikim početnim slovom. Nisu lematizirani ni pridjevi izvedeni od osobnih imena i prezimena. Kad se i to uklonilo, korpus je još uvijek bio gotovo milijunski, točnije: u obradu je nakon spomenutoga kriterija ušlo 952.349 pojavnica. Međutim, trebalo je riješiti neke dileme. Navodimo, ilustracije radi, samo neke.

1. Problem istopisnica. — Iako različita značenja pojedinih riječi (npr. *akademija*) nismo, u načelu, posebno razlučivali, ipak se problem istopisnica u mnogo slučajeva nametao. To se osobito vidjelo kod pravih istopisnice, tj. onih koje se podudaraju ne samo grafemima nego i naglaskom (npr. *atlas, biti, list* i sl.). Tu smo, veće razumljivosti radi, dodavali natuknici značenje ili odrednicu, kao npr. *atlas* 'tkanina' i *atlas* 'zem(ljopis)', *biti* 'jesam' i *biti* 'tući', *listati* 'zelenjeti' i *listati* 'prelistavati', *lučiti* 'dijeliti' i *lučiti* 'izlučivati', *glasina* 'aug. od *glas*' i *glasina* 'nepouzdana vijest', *glasnik* 'glasonoša' i *glasnik* 'glasilo', *list* 'dio biljke' i *list* 'pismo', *gotovo* 'gotovina' i *gotovo* 'dovršeno' itd. Kod nepravih istopisnica označen je naglasak, npr. *grȁd* i *grâd*, *lȕk* i *lûk*, *mȁna* i *mána*, *trȅšnja* i *tréšnja*, *pȁra* 'plinovito stanje vode' i *pàra* 'novac'.

2. Problem dijalektalizama i uopće nestandardnih oblika. — Dramski su tekstovi, odnosno oni prilagođeni za dramsku izvedbu, od Slavka Kolara (*Svoga tela gospodar*) preko Pere Budaka (*Mećava*) do Ive Brešana (*Predstava Hamleta u selu Mrduša Donja*) puni dijalektalizama. Dijalektalizme koji se od standardnoga oblika razlikuju samo fonološki, svrstali smo, ne mijenjajući im dijalekatni oblik, pod standardnu lemu uz zvjezdicu kao oznaku nestandardnosti (*), npr. oblici *snig** i *sneg** navedeni su upravo tako pod natuknicom *snijeg*, imenički oblik *telo** i *tilo** nalaze se pod natuknicom *tijelo*, pridjev *gladen** zabilježen je kao nestandardni i pribrojen natuknici *gladan*, *lepi** i *lip** nalaze se kod natuknice *lijep*, *denes** se nalazi *sub voce danas*, *gri** je priključen *grijehu*, *laditi** je pod lemom *hladiti*, oblik *goricah** priključen je natuknici *gorica*, frekvencija oblika *govoril** i *govorija** nalazi se kod natuknice *govoriti* itd. Leksički dijalektalizmi (npr. *kaj**, *gda**, *bogek**, *cucek**, *žnora**, *dišpet**, *brontulon**, *kantun**, *žmul**, *letrika** i dr.) navedeni su kao posebne natuknice s oznakom nestandardnosti. Tako je *pes** došao pod lemu *pas*, a za *cucek** je uspostavljena nova lema. Neki standardni oblici, poput *dobi* (aorist od *dobiti*), mogu biti označeni i kao nestandardni kad u tekstu to zaista jesu, npr. *dobi** kad stoji umjesto *dobio* (u tekstu: *ja ne dobih a ti dobi*, odnosno *ja san dobi*). Isto tako, infinitiv bez *-i* zabilježen je kao standardan kad je u tekstu napisan kao dio futura (npr. *analizirat će* ili *bacat ću*), ali je nestandardan u primjerima poput *ja ću bacit, ti se moraš prignut, on će leć* i sl. Kao nestandardni oblici označene su i grafijske varijante (npr. *babske** mj. *bapske*, *avijon** *mj. avion*).

3. Problem pridjeva i glagolskoga pridjeva. — U primjerima tipa *cijeli mu je svijet bio uzdrman* oblik *uzdrman* lematiziran je kao oblik glagola *uzdrmati* jer je dio predikata, a u primjeru *gledao je njegovu uzdignutu ruku* oblik *uzdignut* lematiziran je kao pridjev jer je atribut. Što se pak tiče poime-

ničenih pridjeva (npr. *debeli ustane* ili *debeloga još nema)*, lematizirani su kao imenice, tj. stavljena im je gramatička oznaka *m, f* ili *n*. Takve su imenice dobile uvijek natuknički oblik određenoga oblika pridjeva (npr. *debeli*, m) – za razliku od pridjevske natuknice (*debeo*, adj). Zbog svega je spomenutoga trebalo katkada i po nekoliko puta zagledati u veće cjeline teksta. Jer, polazili smo od toga da je korpus naše jedino čvrsto uporište i da ga treba poštivati. Upravo zbog toga načela neki su oblici prezentirani onako kako su zapisani u tekstu, npr. *a-lu-zi-ja-ma* (pored, dakako, potvrđenoga oblika *aluzijama*), *mraaak* i dr.

Kad su, nakon svih provjera, i ti problemi bili riješeni, izrađena su tri dijela *Hrvatskoga čestotnoga rječnika*: 1. Čestotni rječnik (čestotni popis lema), 2. Abecedni rječnik (abecedni popis lema uz oznaku čestote) i 3. Abecedni rječnik s pojavnicama (abecedni popis lema uz pripadajuće pojavnice s njihovim čestotama).

Razmatrajući podatke što se nalaze u prvom dijelu, po čestoti složenom korpusu, može se dodati i poseban komentar. Zacijelo nećemo nikoga iznenaditi ako se kaže da je i u ovom milijunskomu korpusu najfrekventnija natuknica glagol *biti* 'jesam' (što znači svi oblici toga glagola od *bi* do *su*) sa 56.194 potvrde, a to predstavlja relativnu čestotu od 5.6112 ili 5,6112% cijeloga korpusa, s rangom broj 1 i s oznakom zastupljenosti u svim potkorpusima. Budući da se radi o čak 85 različitih oblika glagola *biti*, tako visoka čestota ne iznenađuje. Zanimljiviji je stoga veznik *i*, koji je upotrijebljen čak 41.865 puta, što predstavlja relativnu čestotu od 4.1802 ili 4,1805% cijeloga korpusa, s rangom broj 2 i s oznakom zastupljenosti u svim potkorpusima. To ujedno znači da bi se moglo uzeti kao pravilo da veznik *i* zaprema 4,1% svakoga hrvatskoga teksta. Na trećemu je mjestu prijedlog *u* (27.087 potvrda ili 2,7047% teksta u svim potkorpusima), na četvrtome je veznik *da*, na petome zamjenica *on*, na šestome prijedlog *na*, na sedmome zamjenica *koji*, na osmome zamjenica *ja*, na devetomu prijedlog *za*, na desetomu partikula *ne* itd. Zanimljivo je također pripomenuti da je najfrekventnija imenica *godina* sa 1.750 potvrda, zatim *ruka* sa 1.599 potvrda, *zemlja* sa 1.506 potvrda, *čovjek* sa 1.341 potvrdom, pa *rad, vrijeme, život* itd.

Najfrekventnije riječi našega korpusa od glagola *biti* 'jesam' do pridjeva *crn* (sa 610 potvrda) pojavljuju se u svim potkorpusima. Tada nastaje prekid jer se uzvik *o* (sa 608 potvrda) nalazi u drami, prozi i stihovima (nije zabilježen u novinama i udžbenicima). Razumije se da svaki potkorpus ima, poglavito kod glagola i imenica, drugačiji čestotni red. To se osobito dobro vidi u čestotnicima pojedinih potkorpusa.

U druge dvije cjeline nalaze se natuknice poredane abecednim redom. Iako je i u njima frekvencija u prvome planu, jednako je tako zanimljiv podatak u kojemu se potkorpusu riječ javlja. Tako npr. imenica *galerija* ima rang čestote 79 sa 19 potvrda (i to u oblicima *galerija* 8, *galerijama* 3, *galerije* 4,

galeriji 3, *galeriju* 1) javlja se u svim potkorpusima, za razliku od imenice *galeb* (s oblicima *galeb* 9, *galebe* 4, *galebi* 1, *galebom* 1, *galebova* 3, *galebove* 1, *galebovi* 6) koja ima 25 potvrda, s rangom čestote 73, ali samo u prozi i stihovima, *gazde* nema u novinama, *gdjekada* je samo u drami, *grcaj* samo u stihovima, *gregorijanski* samo u novinama, pridjev *grijaći* samo u udžbeniku, *grk* 'vino' samo u prozi itd. Ostaloj vašoj znatiželji možete udovoljiti pogledate li u *Hrvatski čestotni rječnik*, koji će uskoro izaći iz tiska.

On Croatian frequency dictionary

Summary

The author discusses the main problems involved in the compilation of the *Croatian Frequency Dictionary* which has recently been made at the Linguistic Institute at the Zagreb Faculty of Philosophy.

Ključne riječi: rječnik, riječ, čestota, čestotni rječnik, hrvatski jezik
Key words: dictionary, word, frequency, frequency dictionary, Croatian