Željko Bujas i Milan Moguš — *Zagreb*

O MOGUĆNOSTIMA KOMPJUTORSKE OBRADE DIJALEKATSKIH PODATAKA

Ovo nije prvi put da se govori o kompjutoru kao pomagalu pri obradi dijalekatskih podataka.[1] To je i razumljivo jer — gdje je vrlo mnogo podataka, odmah se javlja misao o njihovu kompjutorskom sređivanju. Na temelju svoga dugogodišnjeg iskustva nastojat ćemo pokazati da se kompjutor može veoma korisno upotrijebiti i u dijalektologiji. Ograničit ćemo se zasada na tri vrste pomoći.

I. — Jedan od najvažnijih zadataka naše dijalektologije je rad na dva golema projekta. Prvo je izrada Općeslavenskog dijalektološkog atlasa (OLA), a drugo Jugoslavenskog dijalektološkog atlasa (JDA). Budući da je, s jedne strane, za prvi projekt izrađena znanstvena metodologija koja se već nekoliko godina primjenjuje u svim slavenskim zemljama, a, s druge, da je rad toliko uznapredovao da se uskoro mogu očekivati prvi rezultati, nećemo govoriti o upotrebi kompjutora u vezi s tim pothvatom. Drugi je projekt još uvijek u pripremnoj fazi, još se uvijek, naime, prikuplja građa s terena (čakavskog, kajkavskog i štokavskog),[2] pa je i eventualna mogućnost kompjutorske obrade dobivenih podataka tek pred nama. Zato ćemo radu u vezi s upravo tim projektom posvetiti veću pažnju.

Što preciznije određivanje fizičkog opsega podataka, s posebnim obzirom na njihovu strukturiranost, neophodna je predradnja za svako planiranje kompjutorske obrade. Posebno je tako kad se podvrgavamo koordinatama tako specifične i opsežne građe kakva je dijalektološka. Pokušat ćemo, dakle, biti onako precizni koliko nam dopuštaju postojeće stanje i podaci u hrvatskoj dijalektologiji.

[1] Uspor. M. Moguš, *Metode suvremene lingvistike u prikupljanju i obradi dijalektološkog materijala*, Institut za jezik i književnost u Sarajevu, Posebna izdanja, sv. 2, Sarajevo 1974, str. 101—103.

[2] Prikupljanje građe s makedonskog i slovenskog jezičnog područja uglavnom je (ili potpuno) završeno.

Predviđenih 500 lokacija (punktova) u SRH treba da budu pokriveni s istim brojem upitnika. Kako svaki upitnik sadrži 1073 pitanja (uzimajući u obzir da su neka od njih višestruka), treba računati s ukupno 536.500 pitanja. Budući da svako pitanje uključuje prosječno dva morfološka potpitanja — na desnoj strani upitnika — moramo realno računati s otprilike 1,000.000 pitanja/potpitanja, i bar isto toliko odgovora. Naglasili smo »bar« misleći na slučajeve gdje će se odgovori informanata razlikovati.

Brojka od 1,000.000 odgovora ne pruža ni u kojem slučaju dovoljno iscrpnu predodžbu o *potencijalu* lingvističkog sadržaja takve građe. Za totalni analitički pristup toj građi — bez kojeg je nemoguća istinska sinteza — važna je naime spoznaja o svim mogućim »pretincima« u koje mogu pasti odgovori iz upitnika. Bez sustavnog predviđanja i kategoriziranja takvih pretinaca nemoguće je izgraditi suvislu metodologiju analize tako opsežne i tako strukturirane građe kao što je ova o kojoj je riječ.

Fonetsko-fonološki odgovori uključuju tako 20 predviđenih vokalnih znakova plus 9 općevokalnih, a konsonantski preko 60 znakova plus 6 općih, ili ukupno stotinjak znakova. *Prozodijski* podaci usložnjavaju to time što za svaki vokal predviđaju 10 suprasegmentala, odnosno oznake za dužinu ispred i iza naglaska.

Morfološki odgovori zahtijevaju 17 pretinaca/kategorija za imenske oblike. Glagolski oblici traže bar 100 takvih pretinaca. Oba se broja, naravno, množe s tri ako želimo posebne pretince za pripadnost rodu. Kategorije u *tvorbi riječi* lako je zamisliti: sama osnova, afiksalna tvorba (prefiks, sufiks, prefiks i sufiks), složenice itd. Ukoliko ih leksički razložimo (na pojedine sufikse, prefikse, osnove), i te će kategorije lako dostići i premašiti brojku 100.

Moguće je zamisliti i *leksičke (vokabularne)* kategorije, premda neuobičajene u većini dijalektoloških analiza: identičan s koine svog narječja; manje (tvorbeno) odstupa od koine; potpuno druga leksička osnova; i slično. Lako će se stvoriti i, jednako neuobičajene, *semantičke* i *stilističke* kategorije: puna ili djelomična odsutnost/prisutnost homonimije (sinonimije, antonimije, polisemije); neutralnost i afektivnost izraza (augmentativ, deminutiv, pejorativ, hipokoristik, ironija, humor i slično). Zanimljivo bi bilo proširiti potencijal stilističkih kategorija i upotrebnim registrima: samo djeca; samo među ženama; »da drugi ne razumiju«; žargon; i drugo.

Konačno, moguće je i poželjno uvesti i kategorije »kvalitete podatka/odgovora«, koje bi se pridavale svim odgovorima bez obzira na razinu i područje jezične analize. Na primjer: ne zna; koleba se; brka s drugim; neprecizna artikulacija; nesiguran ispitivač (da li je dobro čuo); nepostavljeno pitanje; uskraćen odgovor.

Naših 1,000.000 odgovora umnoženih za puni potencijal kategorija koji smo upravo prikazali pretvara se, sada je jasno, u *nekoliko stotina milijuna mogućih specifičnih odgovora.* Ako dakle želimo iscrpne analize takva dijalektološkog potecijala — kompjutor je jedini odgovor.

II. — Drugi izvor pri izradi karata za dijalektološki atlas predstavljaju objavljene monografije o pojedinim mjesnim govorima. Ako se i ovdje ograničimo samo na SRH, onda u fond podataka ulaze primjeri iz raspravâ o mnogim štokavskim govorima te svi podaci koji se odnose na čakavske i kajkavske govore.

Koliki je opseg te građe, teško je procijeniti. Ako je, možda, u fonološko-morfološkom bloku objavljenih rasprava manje podataka od upitničkih, više ih je zacijelo u tvorbenom, sintaktičkom, leksičkom i stilističkom dijelu jer su tim problemima posvećene, npr. na čakavskom terenu, čak čitave rasprave ili knjige.[3] Može se dakle računati s približno istim brojem podataka kao i kod upitnikâ, iako jedni i drugi podaci neće biti istovrsni. Ukoliko se i kod ove građe želimo obavijestiti koji se sve podaci odnose na koji punkt i u kojoj se raspravi to nalazi (a iz takvih će se totalnih pregleda vidjeti i tzv. prazna mjesta, tj. odsutnost podatka), onda i opet dolazimo do spoznaje da bi kompjutorsko sređivanje bilo najdjelotvornije.

Svojstva su današnjih kompjutora uglavnom poznata, pa ćemo istaknuti samo ona posebno relevantna za dijalektološki zadatak o kojem govori ovaj referat. To su (1) velika brzina i pouzdanost u sortiranju građe, (2) neiscrpne mogućnosti permutiranja građe, (3) najrazličitije oblikovanje građe i rezultata analiza (konkordancije, indeksiranje, selektivni ispisi) i (4) matriciranje podataka.

Kompjutor će tako, na primjer, s lakoćom izvući slijedeće podatke iz fonetsko-fonološko-prozodijskog sadržaja jednog milijuna odgovora naše dijalektološke građe:

1) opću distribuciju pojedinog glasa (ili akcenta);
2) kombinirane distribucije pojedinih glasova s akcentom i dužinom;
3) distribucije kombiniranih glasova (skupina suglasnika i sl.);
4) pozicijske varijante (glas na početku/sredini/kraju riječi);
5) akcenatske i druge varijante iste leksičke jedinice;
6) skupine glasova po artikulaciji;
7) artikulacijske osobine korelirane s prozodijskim elementom.

Izvukavši te podatke, kompjutor ih može prikazati u obliku slijedećih mogućih popisa: puni abecedni, selektivni abecedni, odostražni puni i selektivni (za pojedini sufiks ili glasovnu skupinu), po dužini riječi, po učestalosti javljanja, itd.

III. — Posebno je značajna, i za ekonomičnost dijalektoloških istraživanja važna, činjenica da kompjutor može prikazati dijalektološke podatke iz svoje memorije i u obliku terenske distribucije, dakle kao kartu.

[3] Uspor. M. Hraste, *Sufiksi za tvorbu deminutiva i augmentativa u čakavskim govorima srednje Dalmacije*, Zbornik radova Filozofskog fakulteta, sv. 2, Zagreb 1954, str. 57—66; B. Finka, *Čakavske stilističke studije*, »Suvremena lingvistika«, sv. 5—6, Zagreb 1972, str. 15—18; B. Jurišić, *Rječnik govora otoka Vrgade*, Biblioteka Hrvatskog dijalektološkog zbornika, Zagreb 1973.

Kako je to moguće i kako izgleda takva karta?

U svojoj memoriji kompjutor formira gustu koordinatnu mrežu. Na nju on zatim nanosi jednostavne ili združene, permutirane dijalektološke podatke prema lokaciji iz upitnika. Takav »raster« podataka sada printer kompjutorskog sistema otiskuje kao običan kompjutorski ispis na slijepe karte SRH. Gustoća takva ispisa je standardna: 60 redaka po 132 znaka, ili 7920 znakova, na veliki list kompjutorskog ispisa — odnosno 7920 rasterskih pozicija na četvorini od 28 × 38 cm. Na svaku od 500 upitničkih lokacija otpalo bi tako 16 pozicija, što bi se zbog nepravilnog oblika SRH vjerojatno svelo na 8—10 (druge bi padale izvan republičkih i državnih granica, na vodene površine i slično). Kapacitet takve karte može se lako umnogostručiti — a s njime i preciznost »rastriranja« podataka — tako da na svaki list kompjutorskog ispisa dođe samo jedna četvrtina ili šesnaestina SRH, što znači 70 odnosno 250 rasterskih pozicija na svaku od 500 upitničkih lokacija. Izbor znakova koji bi se utiskivali na relevantne lokacije dosta je bogat (cijela abeceda, interpunkcija i posebni znakovi, npr. +, *, %, &, $ itd.), pa bi to činilo takvu kompjutorsku dijalektološku kartu preglednom i fleksibilnom. Brzina tiskanja takvih karata je 18 na minutu, a prednost je postupka u tome što ih, kako smo već istakli, otiskuje običan printer, redovit dio svakog kompjutorskog sistema. Za prave, precizne dijalektološke karte može se upotrijebiti i poseban skupi stroj, tzv. ploter (nanositelj), koji je neposredno vezan na memoriju kompjutora i kartografski uobličuje podatke iz te memorije. Ovisno o preciznosti i gustoći tih podataka, takav ploter može ucrtavati i izoglose, odlučivati o dominantnim znakovima, mijenjati znakove pri preklapanju distribucije i slično.

Posebno je pitanje koliko bi ljudi i koliko sredstava zahtijevao znanstveni projekt koji bi kompjutorski obradio sve dijalektološke podatke iz SRH (dakle, uz pretpostavku da je pokriveno svih 500 lokacija u Republici). Ovdje moramo odmah naglasiti da bi se najveći dio posla utrošio na pripemnu fazu: prenošenje podataka na najprikladniji medij za ovakvu obradu — bušene kartice. Na osnovi našeg dosadašnjeg iskustva u kompjutorskoj obradi tekstova i jezičnih podataka, trebalo bi računati s ekipom od 2—3 rukovodeće osobe (vođe projekta, glavni koordinatori), 3—4 lingvista prireditača upitničke i druge građe za bušenje, 2—4 bušačice-verifikatorke (ovisno o ritmu poslova) i 1—2 programera. Bušačice i programeri mogli bi biti vanjski suradnici. Ostali troškovi, uz pretpostavku da je kompjutorsko vrijeme besplatno (radi se o znanstvenom projektu), bili bi znatno niži, a sastojali bi se pretežno od izdataka za same kartice i nešto uredske opreme. Računajući s pet godina trajanja projekta, vjerojatno ukupna potrebna sredstva ne bi premašila iznos od 1,500.000 din.