UDK: 004:038
Stručni članak
Primljeno: 12. 6. 2008.
Prihvaćeno: 26. 6. 2008.

INFORMATIČKA POTPORA OČUVANJU HRVATSKE JEZIČNE BAŠTINE NA PRIMJERU HRVATSKO-TALIJANSKO-LATINSKOG RJEČNIKA DRAGUTINA PARČIĆA

Mira Zokić

Sažetak: *Stari hrvatski rječnici vrijedan su dio naše jezične baštine. Analiza strukture takvog rječnika daje uvid u sliku kulturnih, socijalnih i ekonomskih prilika vremena u kojem je nastao i uspoređuje tadašnji hrvatski jezik sa suvremenim. No tekstovi pisani na papiru s vremenom blijede i propadaju. Digitalizacija se vrši upravo radi očuvanja rječnikâ, ali i analize građe, dostupnosti većem broju ljudi i lakše upotrebe.*
Digitalizacijom staroga Parčićeva rječnika sačuvano je vrijedno djelo hrvatske kulturne baštine. Osim što se digitalizacijom i obradbom građe ovoga rječnika pružaju mogućnosti olakšanja daljega procesa obradbe, istraživanja i samog očuvanja, omogućuje se također usporedba strukture jezika i života iz vremena nastanka rječnika sa suvremenim rječnicima i sadašnjicom.

Ključne riječi: *hrvatski rječnik, digitalizacija, jezična baza, jezična baština, suvremeni riječnik*

1. Uvod

Na prvi se pogled može učiniti da informacijska tehnologija i jezik nisu povezani pojmovi. Međutim, stvarnost je upravo suprotna. Informacijska tehnologija umnogome se uvukla u lingvistiku. Danas se u računalima prikupljaju i pohranjuju značajne količine jezične građe u obliku korpusa i rječnika.

Stari hrvatski rječnici vrijedan su dio naše jezične baštine. Analiza strukture takvog rječnika daje uvid u sliku kulturnih, socijalnih i ekonom-

skih prilika vremena u kojem je nastao i uspoređuje tadašnji hrvatski jezik sa suvremenim. No tekstovi pisani na papiru s vremenom blijede i propadaju. Upravo radi očuvanja rječnikâ, analize njihove građe te dostupnosti većem broju ljudi i lakše upotrebe, vrši se digitalizacija.

Tako se može digitalizirati i poprilično star rječnik iz 17. stoljeća i time ovjekovječiti njegova vrijednost. Na osnovi znanstveno utemeljenih načela računalne obradbe cjelokupnoga rječnika može se ući u strukturu tadašnjega hrvatskog jezika i pogledom iznutra usporediti stanje s onim u suvremenim rječnicima hrvatskoga jezika, posebice u čestotnom rječniku.

Ovaj rad temelji se na računalnoj obradbi prvoga sveska *Hrvatsko-talijansko-latinskog rječnika* iz 1901. godine. Autor je hrvatski leksikograf Dragutin Parčić. Proces digitalizacije sastojao se od nekoliko koraka. Prvi korak bilo je skeniranje kako bi se tekst prebacio u elektronički oblik. Zatim je slijedilo ručno ispravljanje pogrešno prepoznatih znakova i na kraju izradba baze podataka koja odgovara strukturi rječnika.

Moderna informacijska i komunikacijska tehnologija golemi je potencijal koji snažno utječe na promjenu načina života te na različite sfere društva. Mogućnost prevođenja dokumenata u elektronički oblik za jezik znači očuvanje vrijedne hrvatske jezične baštine za buduće naraštaje.

Danas je računalo nezaobilazno pomagalo u raznim područjima znanosti jer omogućuje lakšu i bržu obradbu podataka, mogućnost dugotrajnije i jednostavnije njihove pohrane, brži prijenos i primanje podataka te informacija. Na isti se način u lingvistici prikupljaju i pohranjuju velike količine jezične građe u obliku rječnika i korpusa. Današnja lingvistička istraživanja nezamisliva su bez računalne potpore.

2. Značenje očuvanja jezične baštine

Iz sažeto izložene povijesti hrvatskoga jezika i njezine periodizacije izdvaja se nekoliko problema. U prvome se redu uočava izvanredno oštar kontrast između prvoga i drugog razdoblja. Prvo je trajalo šest stotina godina (10.-15. st.), drugo samo jedno (16.) stoljeće. Prvo je izgradilo čakavsko-hrvatskocrkvenoslavenski amalgam, koji se već počeo otvarati prema kajkavštini i štokavštini te širiti svoje funkcionalne potencijale

pa je već na koncu 15. stoljeća, dakle na pragu novoga vijeka, mogao (mutatis mutandis) preuzeti ulogu kakvu je odigrao sličan srednjoruskо--ruskocrkvenoslavenski amalgam tek pri koncu 18. stoljeća – postati materijalnom osnovicom jezičnoga standarda. No, to se nije dogodilo – naprotiv, djelokrug opisanoga hrvatskog jezičnog amalgama drastično se teritorijalno sužava (otprilike na trokut zadarsko područje – šira kvarnerska area – Pokuplje), veze među hrvatskim krajevima slabe zbog rascjepkanosti na razne državne formacije (Hrvatsko-Ugarsko Kraljevstvo, Mletci, Austrija) i osobito zbog prijetnje otomanskih osvajanja, a na raznim se područjima počinje razvijati nova pismenost na nekoliko čakavskih, štokavskih i kajkavskih dijalekata.

Iako je leksikografija u Hrvata od kraja 16. stoljeća bila bogata rječnicima, to ipak nije bilo dovoljno za potrebe književnosti u 19. stoljeću, u doba sve razvijenijega kulturnog i znanstvenog života, a pogotovo za potrebe Dalmacije, u kojoj takvi rječnici nisu bili od veće koristi budući da su pretežno bili tumačeni njemačkim jezikom.

Preporoditeljski posao koji su započeli Gaj i ilirci u Zagrebu i sjevernoj Hrvatskoj pod utjecajem češkoga panslavizma i otpora germanizaciji u okrilju germanskog imperija i kulture ubrzo je naišao na odjek u Dalmaciji, u okrilju istoga imperija ali u ozračju italofone kulture. Dalmacija je u to doba administrativno i kulturno odijeljena od ostalih dijelova Hrvatske. Italofona kultura u Dalmaciji posljedica je četiri stoljeća mletačke i kratke ali učinkovite francuske uprave. Stoga je u Dalmaciji kao i u sjevernoj Hrvatskoj glavno sredstvo narodnoga buđenja bilo promicanje hrvatskog jezika u kulturi i javnom životu. Veliki dio preporoditeljskoga napora bio je usmjeren prema onim Hrvatima koji su se školovali na talijanskom jeziku, pa su i gramatike hrvatskoga jezika pisane na talijanskom, kao što su i rječnici bili talijansko-hrvatski odnosno hrvatsko-talijanski. Upravo je Parčić osjetio potrebu takve sredine, napisavši vrlo cijenjeni rječnik koji je u to doba slovio kao najopsežnije djelo takve vrste na ovim prostorima.

Parčićev život i rad vezani su uz vrlo snažne hrvatske kulturne i nacionalne pokrete u Dalmaciji, osobito uza zadarsku sredinu. Zadar je u Parčićevo doba bio vrlo značajno glagoljaško središte, kao i središte modernih hrvatskih kulturnih preporoditelja. Na političkom planu, tu je borba za ujedinjenje Dalmacije s Hrvatskom, dok su istovremeno izrazito važna bila nastojanja na polju hrvatskoga jezika i književnosti. To je

doba jezičnih rasprava oko hrvatskoga pravopisa – vodila se borba oko etimološkoga ili fonetičkog pravopisa – kao i borbe za ikavsku štokavštinu kao književni jezik.

Parčićeva znanstvena djelatnost u tom je vremenu bila dvostruka: s jedne je strane on strastveni glagoljaš i čuvar te osebujne hrvatske kulturne baštine, a s druge strane vrlo značajan proučavatelj suvremenoga hrvatskog jezika i veliki moderni leksikograf. Njegov leksikografski rad značajan je upravo zbog pisanja dvojezičnih, talijansko-hrvatskih i hrvatsko-talijanskih, rječnika, koji su dali neprocjenjiv prilog hrvatskoj rječničkoj zakladi. (Uz hrvatsko-talijanske rječnike, Parčić je priredio i tiskao i talijansko--hrvatski rječnik, godine 1868. u Zadru, koji je imao više izdanja kao i hrvatsko-talijanski.)

Parčić se rano počeo baviti leksikografskim radom. U izradbi svojih, oslanjao se na već postojeće rječnike Mikalje, Della Belle kao i na Stullijevo *Rječosložje*, obilato se služeći riječima koje je skupljao preko raznih suradnika, jezikoslovaca. Kao profesor zadarske realke, školske godine 1857./58. uredio je mali talijansko-hrvatski rječnik, *Rječnik ilirsko-talijanski – polag najnovijih izvorah*, koji je tiskan 1858. Taj rječnik on sam zove „mali rječnik", u skladu sa svojom željom da jednom napravi potpun i reprezentativan hrvatsko-talijanski rječnik. To se nastojanje i ostvarilo godine 1874. drugim izdanjem hrvatsko-talijanskoga rječnika s naslovom *Rječnik slovinsko-talijanski* (Vocabolario slavo-italiano), tiskanim u Zadru kod braće Battara. To drugo izdanje Parčićeva hrvatsko-talijanskoga rječnika znatno je prošireno i obogaćeno novim riječima i izrazima. Parčić će u svom trećem izdanju, kojim se bavi ovaj rad, posve izostaviti izraze *slovinski* ili *ilirski* pa to treće izdanje nosi naslov *Rječnik hrvatsko-talijanski*.

Tiskan je u Zadru u *Narodnom listu* 1901., godinu dana prije Parčićeve smrti, i vrhunac je njegova leksikografskog rada. Ovo izdanje bogatije je od drugoga za više od petnaest tisuća novih riječi, prikupljenih što iz pučkog govora, što iz školskih i poučnih knjiga, kaže sam autor u Predgovoru. Predgovor ovog izdanja bitan je i po tome što u njemu Parčić jasno iznosi svoju poziciju glede pravopisa. „Ipak mi je nješto reći glede pravopisa. Još nismo na čistu s otim blaženim pitanjem. Od nedavna se pojavila nova struja, i ta kao da sve to više preotimlje mah zbog toga, što je s višeg mjesta propisano, da se školske knjige imaju držati posve fonetičkog pravopisa, da se ima uvesti njekakvo nejasno i klimavo načelo, 'piši kako

govoriš'. Ja u rječničkom poredanju rieči niesam mogao prihvatiti toga načela, dapače sam osvjedočen, da je bolje ovo drugo: 'piši za oko, govori za uho'. U tom me bodre i rieči sv. Augustina: sermo debetur auribus, i primjer, kako se služe danas svi izobraženi narodi u pisanju svoga jezika. Ja se dakle držim ponajveć etimološkog pravopisa, udešena prema umjerenoj fonetici." Treće izdanje *Hrvatsko-talijanskoga rječnika* ima više od 1200 stranica s oko 90 tisuća riječi, a dodana su mu „Osobna imena muška i ženska" i „Zemljopisna imena". Unio je u rječnik i riječi „koje se raznih strukah javnog uređivanja tiču", ekscerpirao je pravni časopis *Pravdonošu i Pravoslovno-političku terminologiju* iz 1853. Strane je riječi označio zvjezdicom. Tako je posebno bilježio turske i neke riječi iz ruskoga, talijanskog i grčkog izvora, kako bi se lakše spoznali neki pučki izrazi.

3. Digitalizacija rječnika

Ovaj rad temelji se na računalnoj obradbi trećega izdanja *Hrvatsko-talijansko-latinskoga rječnika* iz 1901. godine. Autor je hrvatski kanonik Dragutin Antun Parčić.

Proces digitalizacije građe Parčićeva rječnika proveden je u četiri koraka:

1) skeniranje teksta
2) prepoznavanje odnosno čitanje teksta pomoću računalnog programa za optičko prepoznavanje znakova (OCR)[1]
3) ručno ispravljanje teksta u odgovarajućem programu za obradbu teksta
4) pohranjivanje digitalne građe

Zadnji je korak u obradbi pohranjivanje građe. Obrađeni se materijal pohranjuje na odgovarajući medij, kako bi se omogućila njegova kasnija upotreba i pretraživanje.

Kao što je poznato, OCR ima teškoća pri prepoznavanju starih tekstova zbog njihove istrošenosti i lošeg stanja. Tako je program određene znakove pogrešno prepoznao i netočno „prepisao".

Prilikom ručnog ispravljanja znakova koje program nije točno prepoznao uočene su sljedeće pogreške:

[1] Engl. Optical Character Recognition.

Slovo b često je prepoznato kao h, a B kao broj 3.

Slovo c često je prepoznato kao e, i obratno.

Slovo d često je prepoznato kao cl, c-, đ.

Slovo f često je prepoznato kao t.

Slovo h često je prepoznato kao b, k, li, a kurzivno slovo h (h) gotovo uvijek kao b.

Slovo i često je prepoznato kao ì, í, î, e'.

Slovo k često je prepoznato kao h, iz.

Slovo l često je prepoznato kao i, i', !, I ili kao broj 1.

Slovo M često je prepoznato kao 31, 3II.

Slovo m često je prepoznato kao rn, in, nn, rr, nì, nu, ni, ur.

Slovo n često je prepoznato kao ri, r, u.

Slovo o često je prepoznato kao c, e, O.

Slovo r često je prepoznato kao n, a slovo R kao broj 12.

Slovo s često je prepoznato kao r, 5.

Slovo z često je prepoznato kao broj 7.

Slova ā, ē, ī, ō, ū najčešće su prepoznata kao da su bez naglaska, dok su à, è, ì, ò, ù kao znakovi bez naglasaka prepoznati nešto rjeđe.

4. Rječnička baza

Digitalni bi se rječnici bez specijalnih alata za pisanje, pohranu, pretraživanje i distribuciju sveli na istu razinu uporabivosti na kojoj se nalaze klasični papirnati rječnici. Računalna obradba jezika područje je bliske suradnje jezikoslovaca i informatičara, u kojoj je informatičarima na prvom mjestu što učinkovitija i brža obradba podataka uz što manji utrošak računalnih resursa. Rezultati nerijetko nadilaze ponekad prestroge lingvističke okvire i čak daju uvide u neke nove jezične činjenice. Računalni modeli za analizu jezične strukture postaju tako izrazito učinkoviti i ponekad pojednostavnjuju iznimno složene lingvističke pristupe. Uključivanjem računala u suvremenu leksikografiju dosadašnje sastavljanje rječnika doživjelo je prekretnicu jer su iznimno velike količine jezične građe lako, brzo i jeftino dostupne.

Hrvatski jezik ima iznimno razrađenu morfologiju pa je morfološka obradba nezaobilazno polazište za obradbu na ostalim jezičnim razina-

ma. Važnu ulogu među alatima za obradbu jezika na razini riječi imaju alati specijalizirani za tu svrhu. Oni su potrebni za morfološki bogate jezike, točnije flektivno bogate jezike. Flektivni jezici, kao što je hrvatski, gramatičko značenje izražavaju dodavanjem različitih gramatičkih elemenata na osnovu.

Dosada je za hrvatski jezik u potpunosti ostvaren samo pristup koji se temelji na uporabi leksičke baze podataka pomoću koje se mogu, iz danih osnovnih oblika, izvoditi svi oblici riječi. U hrvatskom jeziku problem su višeznačnosti. Da bi se prema modelu za segmentaciju teksta one mogle razriješiti, nužne su rječničke baze koje mogu generirati sve postojeće riječi u svim oblicima. Potrebno je izgraditi leksikon koji iscrpno obuhvaća sve kombinacije morfema prema pravilima hrvatskoga jezika i pohraniti ga u bazu podataka. Cilj je automatske morfološke obradbe generiranje oblika riječi, i to prepoznavanjem određenih gramatičkih osobina te riječi.

Tablica 1: *pod nazivom Parcic_Corpus ima sljedeću strukturu:*

Ime polja	**Vrsta polja**	**Parcic_Corpus - *Opis***
ID	AutoNumber	Redni broj teksta
Tekst	Memo	Pojedini paragraf (odlomak) – natuknica u rječniku – cjelokupni tekst natuknice s HTML oznakama za masna i kurzivna slova
Trans_Tekst	Memo	Hrvatski tekst uz pomoć funkcije transliteriran na današnji hrvatski grafički sustav, talijanski i latinski ostaju nepromijenjeni

Na prethodnu je tablicu vezana tablica Parcic_Pojavnice, koja sadrži sve riječi što se pojavljuju u tekstu rječnika (pojavnice) zajedno s uputnicama na tablicu Parcic_Corpus:

Ime polja	Vrsta polja	Parcic_Pojavnice - *Opis*
Rbp	AutoNumber	Redni broj pojavnice
Rbt	Number	Broj odlomka – odgovara polju ID u tablici Parcic_Corpus
Lok	Number	Pojavnica u odlomku (paragrafu) počinje na mjestu Lok
Jez	Text	Jezik – nije se popunjavalo
Pojavnica	Text	Pojavnica kako se pojavljuje u tekstu
LemaP	Text	Lema pojavnice – osnovni oblik (nije se popunjavalo u ovoj fazi obradbe)
B	Yes/No	Je li masno
I	Yes/No	Je li koso (italic)
U	Yes/No	Je li podcrtano
V	Yes/No	Počinje li velikim slovom
K	Yes/No	Je li kratica (nalazi li se u popisu kratica)
S	Yes/No	Je li "zabranjena" ("stop") riječ (nalazi li se u popisu riječi koje se ne obrađuju dalje)
ImeT	Text	Ime teksta
LemaU	Text	Lema pojavnice, viša razina

U rječničku je bazu tekst, bez praznih redova, unesen u tablicu Parcic_Corpus, iz koje je posebnim programskim modulom, napisanim u Visual Basicu, oblikovana tablica Parcic_Pojavnice, u kojoj je svaka pojavnica označena rednim brojem. Numeriranje pojavnica služi kako bi se znalo iz kojega je dijela teksta svaka od njih uzeta; dakle, tako se utvrđuje odlomak i lokacija.

5. Zaključak

Digitalizacija Parčićeva rječnika nije ga samo sačuvala za buduće naraštaje nego je omogućila analizu strukture samoga rječnika i hrvatskog jezika iz vremena u kojemu je on nastao. Da bi se mogao pretraživati tekst rječnika, prebačen je u rječničku bazu koja je prethodno oblikovana pomoću programa MC Access 2000.

6. Literatura:

1. Marko Tadić, *Jezične tehnologije i hrvatski jezik*, Ex libris, Zagreb 2003.
2. Damir Boras, *Teorija i pravila segmentacije teksta na hrvatskom jeziku*, doktorska disertacija, Filozofski fakultet Sveučilišta u Zagrebu – Odsjek za informacijske znanosti, Zagreb 1998.
3. Blaženka Knežević, *Elektroničko uredsko poslovanje i odvijanje poslovnih procesa*, magistarski rad, Ekonomski fakultet Zagreb, 2002.
4. Zoran Stipanović, *Primjena OCR-a u vektorizaciji katastarskih planova*.
5. Miroslav Kiš, *Informatički rječnik*, Naklada Ljevak, Zagreb 2000.
6. Marko Samardžija – Ante Selak, *Leksikon hrvatskog jezika i književnosti*, Pergamena, Zagreb 2001.
7. Davor Virkes, *Multimedijske komunikacije namijenjene slijepim osobama*, magistarski rad, FER Zagreb, 2004.
8. Matica hrvatska, *Vienac* br. 171, Digitalizacija starih izdanja.
9. Informatički časopisi *Bug* (2005., 2006.) i *infoTrend* (2006.).

UDC: 004:038
Professional article
Accepted: 12. 6. 2008.
Confirmed: 26. 6. 2008.

INFORMATIC SUPPORT IN PRESERVATION OF CROATIAN LINGUISTIC HERITAGE ON EXAMPLE OF CROATIAN-ITALIAN-LATIN DICTIONARY BY DRAGUTIN PARČIĆ

M. ZOKIĆ, Split Šk. vjesn. 57 (2008.), 1-2

Summary: *Old Croatian languages make a valuable part of our language heritage. Analysis of the structure of such dictionary offers an insight into cultural, social and economic opportunities of that time, thus comparing the then time Croatian language with the modern one. Still, the texts written in paper have timely faded and got ruined. Digitalization is carried out just because of preservation of the dictionary, but also for structural analysis, availability to a larger number of people and easier application.*
By digitalization of old Parcic Dictionary, a valuable work of the Croatian cultural heritage has been preserved. Except that by digitalization and structural analysis of this dictionary, many possibilities have been offered for facilitation of subsequent analytical process, research and preservation itself, also comparison of the language structure and life, from the time of the occurrence of the dictionary, with modern dictionaries and present time, has been additionally provided.

Key words: *Croatian language, digitalization, language base language heritage, modern dictionary*