

Marko Tadić
Zavod za lingvistiku, Filozofski fakultet, Zagreb

Problemi računalne obrade imeničnih oblika u hrvatskome

Cilj je ovoga rada prikazati neke od problema koji su se uzimali u obzir i pokušali riješiti pri ostvarivanju sustava za generiranje imeničnih oblika GENIMEN u Tadić (1992).

Cilj je ovoga rada prikazati neke od problema koji su se uzimali u obzir i pokušali riješiti pri ostvarivanju sustava za generiranje imeničnih oblika GENIMEN u Tadić (1992).¹ Za te se probleme može pretpostaviti da su s jedne strane nezaobilazni u računalnoj obradi imenične morfologije hrvatskoga, a s druge se strane dio njih pojavljuje i u računalnoj obradi oblika ostalih vrsta promjenljivih riječi. Izvedba sustava GENIMEN odvijala se u dva koraka: a) stvaranje lingvističkoga modela (tj. opisa) deklinacije imenica temeljenog s jedne strane na tradicionalnoj gramatici, a s druge, na zasadama generativizma tj. generativne fonologije; b) provjera toga opisa u obliku računalnoga sustava kadrog proizvoditi sve potencijalne oblike imenica. Tako se koncipirana izvedba neminovno morala susresti s nizom ograničenja. O tim će ograničenjima ovdje također biti riječi.

1. Problem pristupa modeliranju sustava

Dosada su poznati razni formalizmi ili računalni sustavi² koji su se okušali u obradi morfologije pojedinoga prirodnog jezika. Dio njih je, s jedne strane, donekle prisiljen ograničenjima računalne prirode, nerijetko posezao za rješenjima koja nisu uvijek u potpunosti poštivala jezične zakonitosti, barem ne u

- 1 O pojedinostima same računalne izvedbe sustava GENIMEN potanko se raspravlja u Tadić (1992), stoga se time ovaj rad neće detaljnije baviti, već će pokušati upozoriti na probleme pretežito lingvističke provenijencije s kojima se taj pokušaj računalne obrade imeničnih oblika u hrvatskome susreo.
- 2 Kay (1977), Koskenniemi (1983), Winograd (1983), Jäppinen-Ylilammi (1986), Finkler-Neumann (1988), Kržak (1988), Kržak (1990) samo su neki od cijeloga niza sustava koji na neki način uključuju obradu morfologije.

onom obliku u kojem ih iskazuju gramatike. S druge se strane, dio njih, unatoč lingvistički relevantnim postavkama, služio metodama za koje se u lingvistici ne bi moglo naći opravdanje.³ Takvi su pristupi (najčešće zahvaljujući anglocentričnosti istraživača, ali i nezainteresiranosti za morfologiju – tà sintaksa je dugo bila u središtu zanimanja) obradu morfologije najčešće svodili na različite oblike popisâ. Dakle, propisa, kao ni obrade, nije ni bilo – reducirao se na popis. Takvi su popisi bili ili neskrraćeni (gdje se, najvećma za potrebe računalne obrade drugih jezičnih razina, morfološka razina rješavala jednostavnim popisom svih mogućih oblika riječi) ili pak skraćeni nekom od informatičkih metoda (npr. krnjenje (*truncating*) gdje se za početak riječi uzimlje dio nepromjenljiv u pismu, a za dočetke promjenljivi dijelovi: *vu-* + *-k/-ka/-ku/-če/-kom/-ci/-ovi...* itd.). Postupak je to za informatičara potpuno legitiman i njime se krati vrijeme obrade kao i kapacitet potrebne memorije, ali za lingvistiku je to metoda segmentacije koja ruši same njezine temelje.

Riječ u nekom obliku rezultat je ili polazište svakog sustava za obradu morfologije. Polazna je pretpostavka svih takvih sustava postojanje jedinica nižih od razine riječi koje se jedinice kombiniraju na neki način ne bi li oformile bazičnu strukturu riječi. Nazovimo tu razinu u najširem smislu *morfotaktičkom razinom*.⁴ Neki sustavi već u ishodištu jedinice s razine ispod razine riječi ne izjednačuju s jezičnim jedinicama. Takvi su pretežito sustavi koji kreću s informatičkim teorijskim postavkama, a one nerijetko rezultiraju više ili manje čistim modelom relacijske datobaze. Radi se o relacijskom modelu riječi (ili, točnije rečeno, relacijskome elementu u modelu riječi) kojim je moguće prikazati popis u skrćenom obliku – krnjenje je zapravo primjer takva pristupa.

Razina riječi na kojoj se ostvaruju tek kombinacije nižih jezičnih jedinica, razina je koja se može prikazati čistim relacijskim modelom, tj. relacijom među apstraktnim entitetima koji bi u lingvističkom modelu bili ekvivalenti nižih jezičnih jedinica (najčešće morfemâ ili njihovih sklopova). Međutim, da bi se dobila riječ u površinskom obliku, potrebno je povezati morfotaktički niz s površinskim, primjenom algoritama, tj. transformacija. Tako smo došli do algoritamskoga elementa u modelu riječi.

Na ovome se mjestu pokazuje potreba za lučenjem dviju razina obrade i dviju razina podataka:

1. *morfotaktička razina* na kojoj su jedinice ispod razine riječi tek povezane relacijom u morfotaktički niz. Ova razina služi kao ulazna za transformacije.

2. *preobličena razina* na kojoj su neke jedinice u nizu transformirane i koji niz predstavlja gotovu riječ. Riječ je rezultat djelovanja transformacija na morfotaktički niz.

3 Stohastički se pristupi rješavanju problema (v. u Kržak 1990) ovdje uopće i ne uzimlju u obzir jer bi lingvista prije svega trebale zanimati (algebarsko-)lingvističke (kojima algoritmi i relacije ne bi smjeli biti strani), a ne isključivo matematičke (statističke, aritmetičke) metode rješavanja problema.

4 U Babić (1991), str. 27, unutar teorijskoga okvira generativne fonologije predlažu se nazivi tih dviju razina: 1. apstraktna ili ishodišna, tj. fonološko-sustavna; 2. izvedena ili konkretna, tj. fonetsko-sustavna.

Od jedinica koje formiraju niz na prvoj razini postoji barem jedna koja je element zatvorenoga skupa (nastavci) i barem jedna koja je element otvorenoga skupa (leksički morfemi i(li) osnove koje su smještene u leksikon). Zbog te pripadnosti barem jedne jedinice otvorenome skupu ni za jedan se sustav ne može tvrditi da opisuje oblike svih riječi nekoga jezika već samo one oblike čije su sastavnice uvrštene u popis elemenata toga potencijalno otvorenoga skupa, a koji je popis *via facti* uvijek konačan. No zato se mogu pokušati opisati svi tipovi generiranja oblika riječi, a njih je u nekom jeziku zacijelo konačan broj.

Polazna je zamisao GENIMEN-a bila izvesti sustav koji bi u najvećoj mogućoj mjeri poštivao propis što ga na razini deklinacije imenica daje gramatika hrvatskoga jezika poštujući pri tom segmentaciju riječi na sastavnice na nižoj razini, tradicionalno nazvane osnovom i nastavcima, uz primjenu transformacija kojima se osnove preobličuju.

2. Problem ograničenja područja

Razgraničenje između derivacije i fleksije kod imenica oštro je i stoga nema razloga ubrajati ga u probleme kao što bi to zacijelo bilo potrebno da je umjesto imenične obrađivana glagolska promjena gdje zahvaljujući prije svega glagolskome aspektu to razgraničenje nije više toliko čvrsto. GENIMEN je ograničen na fleksiju, tj. deklinaciju imenica.

Temelj za lingvistički model izložen je u popisu deklinacijskih uzoraka u dijelu o morfologiji u Babić–Brozović–Moguš–Pavešić–Škarić–Težak: *Povijesni pregled, glasovi i oblici hrvatskoga književnog jezika*⁵ s time što je on doraden i proširen jer tako izloženi uzorci fleksije ne pokrivaju polje svih imeničkih deklinacijskih uzoraka. Predložak je utoliko manjkav jer nigdje eksplicitno ne izlaže uzorke deklinacije supstantiviziranih pridjeva kao što su *slatko, Hrvatska, Stari* itd.⁶ Takve su riječi nesumnjivo imenice što je lako provjerljivo na sintaktičkoj razini, a istodobno se uzorci njihove deklinacije ne navode u imeničkom nego u pridjevskom dijelu. Stoga GENIMEN uključuje pridjevsku deklinaciju u onoj mjeri u kojoj je potrebno da se opiše deklinacija supstantiviziranih pridjeva.

S obzirom na to da je jedno od ograničenja bilo upotreba standardne grafije, iz njega je proizašlo isključivanje suprasegmentalnih obilježja iz modela, koja bi zapravo obilježja mogla činiti paralelni sustav jer je na istom obliku riječi moguće ostvariti više njihovih kombinacija. Time je ujedno otpala mogućnost razlikovanja oblika maline od oblika jednine i(li) množine te tako taj dio paradigme nije uključen u model, tj. ne tretira se kao zasebna kategorija broja već su oblici »utopljeni« u oblike jednine i(li) množine. Model koji bi suprasegmentalna obilježja uzimao u obzir svakako bi morao voditi računa i o razlikovanju oblika maline.

5 Dio koji obrađuje morfologiju posebno se navodi kao Pavešić–Težak–Babić (1991). Deklinacijski uzorci uzeti u obzir ovdje obrađuju se na str. 489–612 i dijelom str. 618–628.

6 *ibid.*, str. 572, zadnja napomena u t. 278. samo spominje da se imenice nastale od posvojnih pridjeva dekliniraju kao pridjevi.

Problem za sebe predstavlja osobit oblik imenične fleksije, a to je fleksija kratica koja počesto sasvim izmiče pravilima. Ovdje se terminom *kratica etiketa*, tj. ono na što se u sklonidbi nastavak dodaje crticom. Naime, u tim je slučajevima rijetko kada potpuno jasno koja se od tih pokrata može, a koja ne može deklinirati. Ako se već može deklinirati, često je nejasan i rod koji koja od kratica može imati itd. To područje niti jedna gramatika hrvatskoga jezika ne dodiruje premda bi, bar kao napomenu, valjalo navesti da se, npr. podatak o rodu i(li) deklinacijskom uzorku, može pronaći u leksikonu.

3. Problem izvedbe sustava

Sâm se opis u onom svom dijelu u kojem nasljeđuje tradicionalnu gramatiku, sastoji od klasifikacije deklinacijskih uzoraka koja je toliko fina da hvata sve nijansne razlike prisutne među njima. Takva klasifikacija mora biti osjetljiva na razne morfološke obavijesti⁷ i one se tretiraju kao relevantni kriteriji klasifikacije uzoraka. Temeljni je kostur klasifikacije preuzet iz Pavešić–Težak–Babić (1991). No, osnovni je problem⁸ toga rada što je, čini se, neki uzorak uključen u klasifikaciju tek ako je za nj pronađena potvrda. Takvu je modelu imenične deklinacije ograničena proizvodnost jer, premda vjerojatno ne postoji potvrda za imenicu s nepostojanim *-a-* npr. *N pacac*, *G paca* (ili *pacca*) itd., nema razloga da se takav deklinacijski uzorak ne uvrsti u klasifikaciju ukoliko je ta kombinacija fonema u hrvatskome moguća. To što ona nije leksički iskorištena ne bi smjelo utjecati na opis imenične deklinacije jer ne znači da jednoga dana, možda i uskoro, ta kombinacija fonema neće postati leksički iskorištena. Stoga bi postojećim kriterijima klasifikacije deklinacijskih uzoraka valjalo dodati novi koji bi pokrивao (morfo)fonološku⁹ tipologizaciju svih potencijalnih, a ne samo potvrđenih završetaka osnovâ.

Sadašnja je verzija sustava GENIMEN utoliko manjkava jer se uglavnom držala klasifikacije navedene u Pavešić–Težak–Babić (1991). Razlika je jedino u primjeni istih onih kriterija klasifikacije koji su većim dijelom izloženi već u istom radu u odjeljku pred popisom uzoraka, a onda u samom popisu uzoraka počesto primjenjivani nedovoljno dosljedno. GENIMEN je težio dosljednosti u primjeni kriterija tako da je broj uzoraka narastao na 404.¹⁰ Sam je koncept uzorka ključan za GENIMEN jer su svi podaci o fleksiji imenica »zgzusnuti« u deklinacijski uzorak kojim se određuju tri komponente propisa nužne za generiranje nekog oblika neke imenice:

- 7 cf. Mihaljević (1991), str. 85, bilješka 122: »U morfološke obavijesti spadaju podaci kao što su: prisustvo ili odsustvo morfološke granice među sastavnicama, obilježja [strano], [domaće] /koja su svakako relevantna za GENIMEN/ i sl. i obilježja izuzetnosti.« (primjedba moja).
- 8 O mnogim tiskarskim, ali i strukturnim pogreškama u klasifikaciji – posebno u obročavanju granâ – valjalo bi napisati poseban članak kao uputu za prepravak pri izdavanju drugoga izdanja.
- 9 O razmeđima fonologije, morfonologije i morfologije v. raspravu J. Silića u istome broju.
- 10 v. dopunjenu klasifikaciju u Tadić (1992:51–65). Ta klasifikacija nikada nije ni zamišljena kao konačna te je već gore iznesena primjedba karakterizira kao podložnu promjenama.

1. odabir nastavaka
2. alternacijska pravila za preobliku osnove (tj. transformacije)
3. redosljed primjene tih pravila.

Nastavci i pravila u modelu čine propis, a imeničke osnove, smještene u leksikonu, popis.

Koji su kriteriji za klasifikaciju imeničnih deklinacijskih uzoraka? Svakako ih valja potražiti unutar pojava u imeničnoj deklinaciji jer njihova prisutnost ili odsutnost može služiti za razlikovanje deklinacijskih uzoraka. To su: višestruki oblici za isti padež, dva oblika množine (samo »kratka« ili obje množine¹¹), skraćena množina, nepostojano *-a-*, alomorfija – primjena transformacija i njihova neprimjena, kriterij roda, živo/neživo–deklinacija, domaće/strano–deklinacija¹² i slični, zapravo semantički, kriteriji kao što su nazivi mjernih jedinica ili imena naroda (koji imaju množinu bez umetka). Ovime popis kriterija zacijelo nije iscrpljen.

Daljnji je problem redosljed primjena kriterija u klasifikaciji. Kriteriji zabilježeni u Pavešić–Težak–Babić (1991) mogu se u grubo podijeliti u četiri grupe:

3.1. Morfotaktički kriteriji

Morfotaktički kriterij globalno definira kojim se nastavcima formiraju oblici paradigme, tj. koji nastavci sudjeluju u stvaranju morfotaktičkoga niza. Taj se kriterij pojavljuje na najvišoj razini – razdiobi na vrste. No postoji i stanovit broj uzoraka kojima se opisuju višestruki oblici istoga padeža i jedini kriterij kod takvih, najčešće iznimnih imenica, za razlikovanje na nižim razinama je ponovno morfotaktički. Takav je primjer imenica *sinak* s Vjd na *-o* za razliku od ostalih uzoraka iz istoga neposrednog čvora u klasifikaciji koje imaju Vjd na *-e*.

3.2. Morfonološki kriteriji

Morfonološki kriteriji čine najveći broj kriterija klasifikacije i njima se pojedine klase uzoraka razlikuju na više razina. Oni su dvovrsni, a određuje ih:

1. (mor)fonološki sastav osnove što zapravo uključuje tip završetka osnove i duljinu osnove u slogovima;
2. postojanje nekoga skupa transformacija.

Primjer za klasifikaciju prema završecima osnove: završetak u Njd a–vrste muškoga roda na prednepčani, nepčani ili mekonepčani suglasnik (itd.) gdje je kod mekonepčanih daljnja podjela na završetke na *-k*, *-g*, *-h*, odnosno završetak na *-k*, na *-Vk*, *-ak* (nepostojano a), *-Ck* itd. Kako su same transformacije vezane uz uzorak, sasvim je prirodno da razlika između primjene i neprimjene neke transformacije također posluži kao kriterij za klasifikaciju uzoraka.

11 Uopće su termini »kratka« i »duga« množina problematični. Precizniji su termini »množina s umetkom« i »množina bez umetka«.

12 cf. Tadić (1992:44–47).

3.3. Strukturno–gramatički kriterij

Strukturno–gramatički kriteriji primjenjuju se (gdje je to potrebno) na jednoj od najviših razina razdiobe – kriterij roda, ili na najnižim razinama – kriterij žive/nežive deklinacije¹³ i kriterij množina bez umetka/obje množine.

3.4. Kriterij domaće/strano

U nevelikom broju slučajeva primjenjuje se kriterij kojim se osobito obilježuju posuđenice. Najčešće je na daljnjim, nižim razinama riječ o zajedničkoj primjeni morfotaktičkih i morfonoloških kriterija jer se takvi uzorci od uzoraka za »domaće« imenice razlikuju odabirom nastavaka i(li) neprimjenom transformacije koja se u »domaćim« uzorcima primjenjuje. Gdje doista pre-staje »strano«, a počinje »domaće«?¹⁴

Ne čini se nemogućim pokušati provesti klasifikaciju deklinacijskih uzoraka primjenjujući kriterije drugim redoslijedom uz uključivanje novih kriterija. Zapravo pravi posao za buduće inačice sustava GENIMEN tek predstoji u pronalaženju što manjeg broja relevantnih kriterija kojima se što gospodarnije može izvesti klasifikacija imeničnih deklinacijskih uzoraka.

Literatura

- Babić, Stjepan (1986) *Tvorba riječi u hrvatskom književnom jeziku – Nacrt za gramatiku*, JAZU–Globus, Zagreb.
- Babić, Stjepan–Finka, Božidar–Moguš, Milan (1990) *Hrvatski pravopis*, Školska knjiga, Zagreb, (2. izdanje – pretisak).
- Babić, Stjepan–Brozović, Dalibor–Moguš, Milan–Pavešić, Slavko–Škarić, Ivo–Težak, Stjepko (1991) *Povijesni pregled, glasovi i oblici hrvatskog književnog jezika*, HAZU–Globus, Zagreb.
- Babić, Zrinka (1991) *Generativni opis konjugacijskih oblika*, Znanstvena biblioteka HFD–a, Zagreb.
- Di Sciullo, Anna–Maria–Williams, Edwin (1987) *On the Definition of the Word*, MIT Press, Cambridge MA.
- Erjavec, Tomaž–Tancig, Peter (1987) »Pregled nekaterih računalniških pristopov k morfološki analizi jezika« u *Vitas* (1990), str. 118–122.
- Erjavec, Tomaž–Tancig, Peter (1988) »Dvo-nivojski model kot teorija in program za morfološko analizo in sintezo« u *ROJP* 4, str. 199–206.
- Finkler, Wolfgang–Neumann, Günter (1988) »MORPHIX, Fast Realization of a Classification–Based Approach to Morphology«, interna publikacija projekta SFB 314 (XTRA), Bericht Nr. 40, Juni 1988.
- Jäppinen, Harri–Ylilampi, Matti (1986) »Associative Model of Morphological Analysis: An Empirical Inquiry« u *Computational Linguistics*, Vol. 12, No. 4, str. 257–272
- Kay, Martin (1977) »Morphological and Syntactic Analysis« u Zampolli (1977), str. 131–233.

13 Poseban problem u ovome slučaju čine imenice koje označuju nešto živo, ali služe kao ime skupa živih bića npr. »Hajduk«, »Željezničar« itd. te mogu Ajd imati jednak i Njd i Gjd.

14 Npr. Vjd imenice *nec* je *necu*, ali *perec*: *perecu/pereče*.

- Koskenniemi, Kimmo (1983) *Two-Level Morphology: A General Computational Model for Word-Form Recognition and Production*, Publications of the Dept. of General Linguistics, Univ. of Helsinki.
- Kržak, Miroslav (1988) »Serbo-Croatian Morpho-spelling« u *ROJP 4*, str. 207–214.
- Kržak, Miroslav (1990) »Opisna, stohastička i relacijska gramatika na primjeru morfologije hrvatskog književnog jezika« u Tkalac-Tuđman (1990), str. 39–55.
- Kržak, Miroslav–Boras, Damir (1985) »Rječnička baza hrvatskog književnog jezika« u *Informatologia Yugoslavica 17 (3–4)*, str. 223–242.
- Lopina, Vjera (1990) »Jezično znanje na primjeru tvorbe riječi« u Tkalac-Tuđman (1990), str. 33–38.
- Mihaljević, Milan (1991) *Generativna i leksička fonologija*, Školska knjiga, Zagreb.
- Pavešić, Slavko–Težak, Stjepko–Babić, Stjepan (1991) »Oblici hrvatskoga književnog jezika (morfologija)« u Babić–Brozović–Moguš–Pavešić–Škarić–Težak (1991), str. 489–612.
- ROJP 4*, Zbornik radova 4. konferencije Računalniška obdelava jezikovnih podataka, Portorož, 3. 10.–7. 10. 1988.
- Samardžija, Marko (1988) »Duga i kratka množina u hrvatskom književnom jeziku« u *Jezik*, god. 35, br. 5, str. 129–136.
- Srhoj-Čerina, Ljubica (1986) »Kolebanja u dugoj množini« u *Jezik*, god. 33, br. 5, str. 148–150.
- Tadić, Marko (1992) *Kompjutorska obrada morfologije hrvatskoga književnog jezika na imeničnom potkorpusu*, magistarski rad, Sveučilište u Zagrebu.
- Tkalac, Slavko–Tuđman, Miroslav (ur.) (1990) *Informacijske znanosti i znanje*, Zavod za informacijske studije, Zagreb.
- Vukušić, Stjepan (1991) »Skлонidba imenica vuk, vrag, rog, bog u kratkoj množini« u *Jezik*, god. 38, br. 3, str. 70–72.
- Winograd, Terry (1983) *Language as a Cognitive Process*. Vol 1: Syntax, Addison-Wesley, Reading-Menlo Park-London-Amsterdam-Don Mills-Sydney.
- Zampolli, Antonio (1977) *Linguistic Structures Processing*, North-Holland Publishing Co, Amsterdam-New York-Oxford.

Problems of Computational Generation of Noun Forms in Croatian

The aim of this contribution is to deal with some of the problems that appeared during the linguistic and computational development of a system for noun-forms generation in Croatian called GENIMEN and further elaborated in Tadić (1992).