Izvorni znanstveni rad
Rukopis primljen 19. 5. 2015.
Prihvaćen za tisak 23. 10. 2015.

Ana Meštrović, Sanda Martinčić-Ipšić
amestrovic@uniri.hr, smarti@uniri.hr
Odjel za informatiku Sveučilišta u Rijeci
Hrvatska

Mihaela Matešić
mmatesic@ffri.hr
Filozofski fakultet Sveučilišta u Rijeci
Hrvatska

Postupak automatskoga slogovanja temeljem načela najvećega pristupa i statistika slogova za hrvatski jezik

Sažetak

Analiza slogova kao osnovnih elemenata jezika važna je za različite postupke u domeni računalne analize prirodnog jezika i govornih tehnologija. Cilj je rada prikazati i evaluirati automatski postupak slogovanja (silabifikacije) za hrvatski jezik te prikazati statističke rezultate raspodjele slogova za hrvatski jezik. Statistička analiza slogova provedena je za dva različita korpusa: korpus RJEČNIK, koji obuhvaća popis hrvatskih leksema u tzv. kanonskom obliku, dobiven iz rječnika hrvatskoga jezika, te korpus SOBiR, koji sadrži popis svih oblika riječi hrvatskoga jezika. Statistički rezultati za opisani postupak uspoređeni su s rezultatima (distribucijama slogova) dobivenima za hrvatski jezik iz postojećih izvora. Provedena je usporedba automatskog postupka s ručnim postupkom i prikazani su rezultati, u okviru čega je određena (aproksimativna) pogreška automatiziranog postupka slogovanja.

Ključne riječi: slog, slogovanje, silabifikacija, distribucija slogova, najveći pristup

1. UVOD

Slog je najmanja jedinica koju je moguće izgovoriti /[1]/ te je stoga i temeljna govorna jedinica (Škarić, 1991: 82). S tog je aspekta analiza jezika na slogovnoj razini važna za različite postupke odnosno metode razvijene u području računalne analize prirodnoga jezika (NLP). Postupak automatskoga slogovanja /[2]/ i podaci o distribuciji slogova važni su zbog potencijalne primjene u postupcima automatskog raspoznavanja govora (Ganapathiraju i sur., 2001; Shafran i Ostendorf, 2003; De Jong i Wempe, 2009), sinteze govora (TTS) (Žganec Gros i sur., 2002, 2006; Sečujski, 2005; Marinčič i sur., 2009) i korpusne statistike.

Kako su govornici u stanju intuitivno segmentirati svoj govor na slogove /[3]/ – jedinice kojima je izraz uglavnom manji od riječi i širi od glasa – lingvistiku i posebice fonetiku vrlo je rano zaintrigiralo pitanje slogova. Lingvistika, još od razdoblja strukturalizma, posvećuje svoje zanimanje ponajprije proučavanju strukture i granica sloga, ali i njegove funkcije. Iako se u fonologiji pritom slog promatra najčešće kao struktura ključna za opis fonotaktičkih svojstava ovoga ili onog jezika, tj. distribucije fonema (usp. npr. Muljačić, 1972: 151), razvijene su i takve fonološke teorije prema kojima je slog, a ne fonem ili distinktivno obilježje, temeljna jedinica u jeziku (usp. npr. Trask, 2005: 323).

Konačno, za istraživanje slogova zanimanje pokazuje i teorija jezičnoga usvajanja. Poznato je naime da se usvajanje jezika kod djece u najranijoj dobi temelji na slogovima, iz slogova se postupno počinju tvoriti riječi, koje se zatim povezuju u mrežu mentalnog leksikona (Oudeyer, 2001).

Cilj je ovoga istraživanja definirati i implementirati postupak (fonološkoga) slogovanja za hrvatski jezik prema pravilima propisanima i opisanima u dostupnoj literaturi. Definirani postupak temelji se na postojećim pravilima slogovanja za hrvatski jezik temeljem načela najvećega pristupa. U radu su prikazani statistički rezultati analize raspodjele slogova i raspodjele različitih modela slogova te raspodjele broja slogova po riječima. Statistička analiza provedena je za dva različita korpusa: prvi korpus obuhvaća lekseme iz hrvatskog rječnika (RJEČNIK); drugi korpus obuhvaća sve oblike riječi (SOBiR). Nakon toga provedena je evaluacija postupka, odnosno procjena pogreške automatskog postupka slogovanja. Uspoređeni su rezultati automatskoga postupka s rezultatima ručnoga postupka slogovanja.

U drugome poglavlju prikazana je struktura sloga u hrvatskome jeziku. Tu su objašnjena pravila i ograničenja koja su uzeta u obzir prilikom izgradnje automatskoga

postupka. U trećemu poglavlju opisan je automatski postupak i dodatna pravila koja su uzeta u obzir kako bi se ispravile pogreške u automatiziranom postupku rastavljanja na slogove. U četvrtome poglavlju opisani su rezultati. U petome poglavlju opisana je evaluacija postupka. U zadnjem poglavlju iznesena su zaključna razmatranja i zamisli vezane za buduća istraživanja.

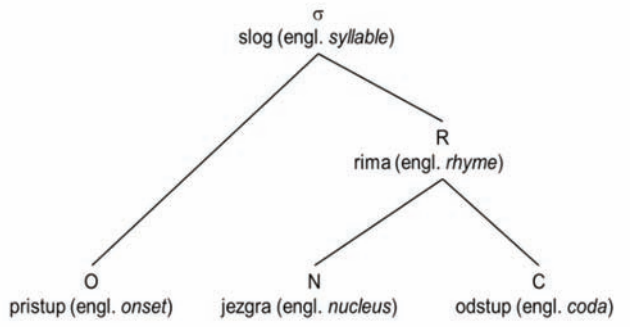
2. SLOGOVI U HRVATSKOME JEZIKU

Jezikoslovna literatura bilježi mnogobrojne pokušaje definiranja pojma sloga (pri čemu se nastoji opisati struktura sloga) i identificiranja sloga u glasovnom lancu (konkretno, kako raščlaniti riječi na slogove). Pitanje opisa strukture sloga i utvrđivanja slogovnih granica u riječima središnje je pitanje i za temu koja se istražuje u ovome radu.

2.1. Struktura sloga

Struktura sloga opisuje se u literaturi na različite načine. Uz fiziološki i psihički (psiholingvistički) pristup određenju sloga /[4]/ najčešće se još spominju fonetsko i fonološko poimanje, koji su u lingvističkim raspravama i najzastupljeniji pogledi na slog. Fonetski pristup slogu, u okviru vrlo poznate teorije prominentnosti ili čujnosti, temelji se na postavci da se glasovi međusobno razlikuju po stupnju čujnosti (koja odgovara pojmu zvonkosti ili sonornosti). Slog se dakle sastoji od dijelova koji su više ili manje zvonki, pri čemu je dio s najvišim stupnjem zvonkosti ujedno i središnji dio sloga, dok su rubni dijelovi sloga manje zvonki. U okviru te teorije poznata je Jespersenova ljestvica zvonkosti /[5]/, koja je u modificiranu obliku prihvaćena i u drugim teorijama. Fonološki pristup slogu susreće se, kao i fonetski, još u strukturalističkim raspravama, koje strukturu sloga opisuju kao kontrastiranje vokala (V) i konsonanata (C) u vremenskome slijedu, ili preciznije silabema odnosno samoglasnika (koji čini jezgru, obavezni dio sloga) i asilabema odnosno suglasnika (koji popunjavaju neobavezni, rubni dio sloga). I u strukturalističkim teorijama i u ranoj generativnoj teoriji fonološki prikazi zaustavljali su se na odsječcima (segmentima) kao konstitutivnim jedinicama kompleksnih jedinica, tj. prikazi su se sastojali od linearnih nizova neovisnih elemenata veličine odsječka. Razlikovna obilježja pak shvaćala su se, kao i u Jakobsonovoj teoriji, sastavnim dijelovima pojedinoga odsječka – oni su vezani za pojedini odsječak (zarobljeni u njemu), a unutar odsječka nisu formalno organizirani (poimani su, naime, kao svežnjevi).

Razvojem generativne teorije oblikuje se novo gledište: razlikovna obilježja ostvaruju se u domenama i manjim i većim od odsječka (tj. neka se obilježja mogu protezati na domene veće od jednog odsječka), a odsječci imaju unutrašnju formalnu strukturu. To se novo shvaćanje, oblikovano u autosegmentnoj (višeglasnoj) /[6]/ generativnoj fonološkoj teoriji (Goldsmith, 1976), počinje primjenjivati i na problematiku sloga (Kahn, 1976 [1980]; Halle i Vergnaud, 1980; Clements i Keyser, 1983) te je sredinom 1980-ih godina u generativnoj lingvistici prevladalo mišljenje da je struktura sloga hijerarhijski organizirana. Slog se dakle na početnoj razini sastoji od pristupa (engl. *onset*) i rime (engl. *rhyme*), a zatim se na nižoj razini od početne rima dijeli na jezgru, koja je obavezni element (engl. *nucleus* ili *peak*) i odstup (engl. *coda*), koji je fakultativan (Slika 1 /[7]/).



Slika 1. Univerzalna shema sloga u autosegmentnoj fonologiji
Figure 1. Universal syllable template in autosegmental phonology

U razmatranju sloga hrvatski lingvisti do ranih 1990-ih godina mahom daju prednost strukturalističkome fonološkom pristupu (Muljačić, 1972; Junković, 1973; Škarić, 1991; Turk, 1991), koji u unutrašnjem ustrojstvu sloga ne pronalazi hijerarhijsko ustrojstvo, nego ustrojstvo linearnoga slijeda. Od 1989. godine nadalje pojavljuju se i u Hrvatskoj autori koji predstavljaju generativni pristup fonologiji: Babić /[8]/ (1989), tj. Jelaska (1997) te Mihaljević (1991).

U ovome radu analizi strukture sloga pristupa se s fonološkoga aspekta, uzimajući u obzir zaključke različitih fonoloških teorija po načelu njihove primjenjivosti na razvoj postupka slogovanja. Za takvu primjenu kao najpraktičniji pogled na strukturu sloga odabran je onaj koji kao sastavne dijelove sloga promatra odsječke (foneme). U skladu s time kao nositelji sloga (jezgra sloga) identificiraju se oni fonemi koji se u

hrvatskoj terminološkoj tradiciji nazivaju samoglasnicima (tj. silabemima /[9]/). U hrvatskome jeziku status nositelja sloga imaju svi vokali i neki sonanti (među sonantima to su: /r̥/, /l/ i /n/: *vrh*, *bicikl*, *njutn*). Pritom su sonanti /l/ i /n/ nositelji sloga samo u poziciji tzv. bočne slogovnosti /[10]/ (u primjerima poput *bicikl*, *njutn*). Metodološka je dosljednost zahtijevala da se za potrebe ciljeva ovoga istraživanja napusti mogućnost tretiranja sonanata /l/ i /n/ u tim pozicijama kao jezgre sloga. Naime kao nositelji sloga identificiraju se samo glasovi sa statusom fonema, dok se pri bočnoj slogovnosti ostvaruju alofoni (bočna je slogovnost naime pojava vezana uz fonetski, a ne fonološki slog). Jezgru sloga u poziciji bočne slogovnosti čini i sonant /r/ (npr. *masakr*), a on će se, u skladu s primijenjenom metodologijom, ipak smatrati slogotvornim budući da je njegovu slogotvornu alofonsku realizaciju moguće pridružiti fonemu /r̥/.

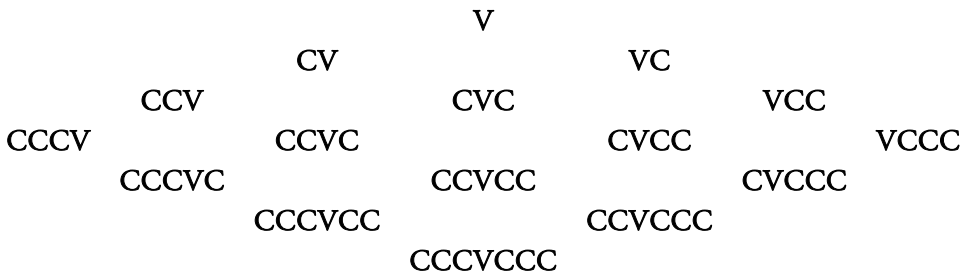
Kad je riječ o završetku sloga, razlikuju se: otvoreni slogovi (završavaju vokalom) i zatvoreni slogovi (završavaju konsonantom), o čemu će također više riječi biti u poglavlju o postupku rastavljanja na slogove (v. poglavlje 2.2.).

Pojam središnjega i rubnoga dijela sloga, kakav poznaju strukturalistički pristupi slogu, moguće je u osnovnoj zamisli pratiti i u autosegmentnoj fonologiji, samo što su u toj teoriji oni u drugačijem hijerarhijskom odnosu, o čemu je već bilo riječi. Središnji dio (pojas) sloga (nukleus ili jezgra) zvonak je i obavezan. Rubni dio (pojas) može biti na početku i na kraju sloga – neobavezan je i šuštav.

U literaturi se najčešće nailazi na stav da u hrvatskome slogovnom pristupu mogu biti najviše tri suglasnika (Junković, 1973; Škarić, 1991; Turk, 1992). Zrinka Jelaska još 1989. godine potpuno opravdano spominje i mogućnost četiriju članova u slogovnome pristupu. Taj se četvrti član ostvaruje u riječima s izgovornom vrijednosti [je] za dugu alternantu kontinuante nekadašnjega glasa jata, tj. za dvoglasnik /ie/ /[11]/: npr. [*sprjéčiti*] (Babić, 1989: 70). Na pravopisnome planu četvrti član u toj poziciji nije uočljiv zbog hrvatske pravopisne tradicije, prema kojoj se duga alternanta odnosno dvoglasnik /ie/ zapisuje s *ije*. Kako od devedesetih godina 20. stoljeća pravopisne knjige što dopuštaju što propisuju pisanje (i izgovor) riječi s alternantom /je/ u kratkim slogovima, npr. *strjelica*, *ždrjebad*, *sprječavati*, četvrti je član postao "vidljiv" i na pravopisnome planu, pa će za istraživanja na korpusima tekstova pisanih od 1990-ih godina naovamo biti potrebno predvidjeti i takve slogove. Oni se mogu opisati kao *SCrje* (pri čemu *S* predstavlja *s*, *š*, *z* ili *ž*, *C* bilo koji šumnik ili sonante *m*, *v*, treći je sonant *r*, a četvrti je *j* /[12]/ – pritom jezgreni vokal može biti samo *e*). U ovome

se istraživanju dakle *SCr* ispred "jatovskih" izgovornih [*jē*] i [*je*] prevodi u model CCCCV. Iz toga slijedi da se *r* u *SCrje* ne smatra slogotvornim.

Prema broju predjezgrenih i zajezgrenih fonema, u hrvatskome jeziku ostvaruje se prema Turk (1992: 35) 14 tipova slogova (11 se tipova ostvaruje u riječima hrvatskoga podrijetla i još dodatna tri tipa u riječima stranog podrijetla: VCC, CVCCC, CCVCCC), dok se tipovi VCCC i CCCVCCC ne ostvaruju (v. Sliku 2).



Slika 2. Mogući modeli slogova u hrvatskome jeziku prema broju predjezgrenih i zajezgrenih fonema (prikaz prema Turk, 1992: 35)

Figure 2. Possible syllable models in Standard Croatian according to the number of phonemes at syllable margins (modified scheme based on Turk, 1992: 35)

Prema našem istraživanju, koje se temelji na načelu najvećega pristupa slogu (v. poglavlje 2.2.2. u ovome radu), ostvaruje se međutim 16 tipova slogova, i to 12 u riječima hrvatskoga podrijetla (jer shemi dodajemo i slog CCCCV /[13]/ u primjerima poput *sprječavati* /[14]/ i sl.), a uz tipove VCC, CVCCC, CCVCCC u riječima stranog podrijetla uočavamo i tip VCCC (u primjeru *triangl*). Ni u našem se istraživanju ne ostvaruje tip sloga sa shemom CCCVCCC.

2.2. Rastavljanje na slogove

Poznavanje strukture sloga uvjet je za provedbu slogovanja (dijeljenja, tj. rastavljanja riječi na slogove). Kao što se razlikuju fonetski i fonološki pogledi na slog, moguće je govoriti i o fonetskome odnosno fonološkom slogovanju (usp. Turk, 1992: 34) /[15]/. Za potrebe cilja ovoga istraživanja primijenit će se fonološko slogovanje (tj. u primjerima u kojima se fonetsko i fonološko slogovanje ne podudaraju, odabrano je fonološko, npr. *bicikl* se rastavlja: *bi-cikl*).

2.2.1. Sljedovi i spojevi fonema

Određivanje slogovne granice u ovome istraživanju provodi se na morfološkoj riječi. Glasovna sekvencija koja odgovara jednoj (morfološkoj) riječi predstavlja slijed fonema. Slogovanje se odvija tako da se u tome slijedu asilabemi pridružuju silabemima, što je jednostavan postupak kad su silabemi i asilabemi raspoređeni u riječi uzastopno po jedan. Ako se u slijedu nađu dva ili više asilabema, potrebno je odlučiti hoće li se oni pridružiti različitim slogovima (pa će tada slogovna granica prolaziti između dvaju asilabema) ili će se pojaviti zajedno kao spoj u pristupu odnosno odstupu jednoga sloga. Drugim riječima, na razini (morfološke) riječi govorit će se o slijedu fonema, a na razini sloga o spoju asilabema odnosno suglasnika (sonanata i konsonanata). Pojam slijeda i spoja razlikujemo zato što takvim terminološkim određenjima dobivamo mogućnost razlikovanja onoga niza fonema koji je realiziran u (morfološkoj) riječi od onoga niza fonema koji je moguć pri slogovanju /[16]/. Drugim riječima, dok se npr. slijed *rgl* u *izverglati* zaista ostvaruje kao takav slijed, dotle se on na razini sloga može percipirati kao jednak cjelini slijeda u (morfološkoj) riječi (pri mogućem rastavljanju *iz-vergl-a-ti*) ili dijelovima te cjeline (ako se rastavi *iz-verg-la-ti*, percipira se spoj *rg*; a ako se rastavi *iz-ver-gla-ti* ili *i-zver-gla-ti*, percipira se spoj *gl*). Razlikovanje tih dvaju tipova nizova – sljedova i spojeva – u ovome je radu potrebno i zato što će se u statističkoj obradi posebno promatrati sljedovi, a posebno spojevi suglasnika.

2.2.2. Provedba slogovanja prema načelu najvećega pristupa

Postoje opća i posebna pravila za provedbu silabifikacije budući da ona ima "univerzalne težnje, ali je ovisna i o obilježjima pojedinoga jezika" (Jelaska, 2004: 172). Primjerice, među općim pravilima obično se navode ova: 1) jedan suglasnik pred samoglasnikom pripada uvijek prvom sljedećem slogu, 2) slijed kojim riječ može početi može stajati i na početku sloga, 3) slijed kojim riječ može završiti može stajati i na kraju sloga, 4) medijalni elementi kojima riječ ne može početi trebaju se podijeliti u dva sloga (Kuryłowicz, 1948, prema Muljačić, 1972: 162) /[17]/.

Autosegmentna fonologija poznaje u slogovanju i načelo najvećega pristupa sloga (Jelaska, 2004: 173–174). Prema tome se načelu pristupu dodjeljuju svi zatvornici u riječi za koje je to moguće (a moguće je ako se pritom ne krše pravila o broju pristupnih mjesta, sonornosti te ograničenju sljedova).

Tome načelu dat će se prednost i u ovome istraživanju. To znači da će se u primjerima poput *ispitati* primijeniti rastavljanje (1) *i-spi-ta-ti* umjesto jednako tako

mogućega rastavljanja (2) *is-pi-ta-ti*. U (1) primijenjeno je načelo najvećega pristupa, dok je u (2) pri izdvajanju prvoga sloga primijenjeno načelo najvećega odstupa. U lingvistici je utvrđeno da se jezici ne ravnaju po načelu najvećega odstupa, već po načelu najvećega pristupa (Jelaska, 2004: 174). Napomenimo i to da bi rastavljanjem kao u primjeru (2) bila uzeta u obzir morfemska granica (*iz+pitati*), što je povezano s potrebom za semantičkom prozirnosti, ali samo u određenim komunikacijskim situacijama, kao što to objašnjava npr. Škarić (1991: 328). Ugrađivanje takvih, semantičkih polazišta u automatski postupak slogovanja pretpostavljalo bi dodatne korake poput utvrđivanja korpusa morfema i sl.

2.2.3. Ograničenja u sljedovima suglasnika

Za razvoj automatskoga postupka slogovanja u ovome istraživanju odsudna su pravila za ograničavanje mogućih suglasničkih spojeva u jednom slogu jer o tim ograničenjima ovisi hoće li: a) slijed suglasnika u medijalnoj poziciji u riječi čitav pripasti pristupu sljedećega sloga ili će se b) neki suglasnici iz slijeda dodijeliti odstupu prethodnoga sloga, a neki pristupu sljedećega sloga. Pravilo da se medijalni elementi kojima riječ ne može početi dijele u dva sloga zahtijeva odgovor na pitanje: kojim to skupovima riječ može početi – samo onima koji su doista i ostvareni u leksiku ili i onima koji kad bi se u leksiku ostvarili, ne bi bili nepodesni?

Zaključci o mogućem početku riječi, pa i onom (još) neostvarenom, mogu se izvoditi na temelju postojećih ostvarenih početaka. Primjerice, budući da su mogući spojevi sonanata u kojima na prvome mjestu stoji usneni, a na drugome neusneni sonant (*mrav*, *mlad*, *mljeti, vruć*, *vlas*), iz toga se zaključuje da je moguć i spoj *vlj-* na početku sloga iako nije ostvaren – jedino uz tu pretpostavku moguće je naime rastaviti *u-mrt-vljen* (Junković, 1973: 42). Istoj kategoriji spojeva pripada i spoj *vn* /[18]/. Izvođenje zaključka moguće je i na temelju ovakve usporedbe: "Možda bi mogle biti prihvatljive riječi što počinju npr. skupinom *dg*, jer postoji *tk* u pristupu (...)" (Jelaska, 2004: 160). Iscrpan popis svih mogućih i nemogućih spojeva koji bi se mogao prenijeti u algoritam za automatsko slogovanje ne postoji jer su lingvistički radovi o slogu u hrvatskome jeziku pisani s drugim ciljem (a to je teorijsko utemeljenje opisa strukture sloga). Ipak, najobuhvatniji je popis predstavila Jelaska /[19]/ tako što je na temelju pravila o generativnim razlikovnim obilježjima (Chomsky i Halle, 1968) opisala ograničenja fonema u spojevima u hrvatskome slogu (2004: 165–171). Zahvaljujući toj analizi ograničenja u algoritam su uza sve ostvarene spojeve na početcima riječi uvršteni i neki dodatni mogući spojevi. Kao mogući spojevi u pristupu sloga uvršteni su: *vlj*, *vnj* i *vn* jer su riječi koje sadrže te sljedove frekventne, a među ograničenjima nema takvih koja bi sprečavala navedene spojeve. Uvršten je i spoj *žđ*

iako nije frekventan, ali je korespondentan spoju *šć*, koji se – premda ni on nije frekventan – ipak ostvaruje na početku riječi (npr. *šćućuriti*, *šćavet*, *šćakavizam*). Neki se od spojeva nazivaju u literaturi snošljivima (Turk, 1992) i dopuštenima (termin kojem Jelaska (2004) daje blagu prednost), a riječ je o onim spojevima koji su potvrđeni u malome broju primjera. Među takvim spojevima Jelaska (2004: 169) navodi i spojeve *ct* te *čt*. U naš algoritam uvršten je spoj *ct* jer je ostvaren u *octen*, *octa*, dok spoj *čt* nije uvršten jer se on ne pojavljuje u korpusu /[20]/.

Kao najčvršća potvrda da se ovaj ili onaj spoj ostvaruje uzeto je njegovo pojavljivanje na početku riječi. Takvo pojavljivanje svakako je u vezi i s jezičnim posuđivanjem: naprimjer integracijom posuđenice *pneumatski* u hrvatski jezik spoj *pn*, za koji nema ograničenja, prešao je iz mogućega spoja u ostvareni spoj te se za slogovno rastavljanje riječi *vapno* kao *va-pno* spoj *pn* u pristupu drugoga sloga može i potvrditi (kad takve leksičke potvrde ne bi bilo, rastavljanje *va-pno* ugradilo bi se u algoritam iz dvaju razloga: prvi je što za taj spoj nema ograničenja, a drugi je da je on zastupljen kao slijed u većem broju riječi i oblika).

U razmatranju koji su spojevi ostvareni uočava se važnost podrijetla riječi. U svih se hrvatskih autora naime pronalaze napomene o tome da su neki spojevi ostvareni samo u posuđenicama ili pak u dijalektima /[21]/. Na temelju takvih podataka može se zaključiti o tipičnoj strukturi sloga u riječima hrvatskoga podrijetla i o tome kakve je inovacije integracija posuđenica u hrvatski jezik donijela u sljedovima i spojevima.

2.2.4. Pravopisni plan i automatski postupak slogovanja

Za slogovanje poseban problem predstavlja odnos izgovorne stvarnosti i pravopisnoga plana, pri čemu ponajprije mislimo na provođenje onih pravopisnih pravila koja su utemeljena na morfonološkome i tradicijskome /[22]/ pravopisnom načelu. Primjerice riječ *azbestni* [àzbesnī] u govoru bi se rastavljala *a-zbe-sni* (ako se poštuje načelo najvećega pristupa). Međutim *t*, koje se u pismu zadržava, ne može se samo "ubaciti" u taj prikaz (*a-zbe-stni*) jer ne postoji slog koji bi u pristupu imao *stn*). U takvim će se primjerima odstupiti od dosljedne primjene načela najvećega pristupa i to će se načelo kombinirati s načelom najvećega odstupa: *a-zbest-ni*.

Kako novija pravopisna praksa dopušta (ili propisuje) pisanje imenica na *-tak*, *-dak*, *-tac* i *-dac* morfonološki, postavlja se pitanje hoće li se riječ *podatci* [podáci] u pismu predstavljati rastavljanjem s prikazom *t* ili bez prikaza *t*. Kao i u prethodnome primjeru, odluka da se prikaže s *t* podrazumijeva da je moguće samo *po-dat-ci* jer *po-da-tci* prikazuje u pristupu trećega sloga nemogući spoj suglasnika.

Naravno, slog koji u odstupu ima *t* ili *d* ispred sloga koji u pristupu ima *c* ulazi u statističku obradu uz poštovanje izgovorne stvarnosti, tj. kao otvoreni slog.

U riječima u kojima se pojavljuju sljedovi koji se prema hrvatskome pravopisu pišu *ds* i *dš*, slogovna granica prolazi između fonema u tim sljedovima jer kroz njih prolazi morfemska granica. Utvrđeno je naime da u riječima s takvom granicom *d* pripada odstupu sloga, a *s* i *š* pripadaju pristupu sljedećega sloga: npr. *grad-ski*. Iako se u govoru *ds* može ostvariti kao afrikatni suglasnik [c], a *dš* kao afrikatni suglasnik [č], u fonološkome smislu nije riječ o afrikatnim suglasnicima (Jelaska, 2004: 176–177). Slog koji u odstupu ima *d* ili *t* ispred sloga koji u pristupu ima *s* ili *š* ulazi stoga u statističku obradu kao zatvoreni slog.

Riječi koje sadrže slijed koji se bilježi kao *-naest-* (*dvanaest*, *dvanaestica* i sl.) rastavljaju se tako da spomenuti slijed čini jedan slog. Tako naime preteže u izgovoru (Jelaska, 1997), a i u novijoj gramatici izričito se propisuje takav izgovor: "Slog se *-aēst* izgovara [ai̯̯st]" (Silić i Pranjković, 2005: 141).

Nadalje, duga je alternacija kontinuante nekadašnjega glasa jata (koja se u suvremenim hrvatskim normativnim priručnicima naziva i dvoglasnikom *ie*) jednosložna, npr. *cvijet*, *snijeg*, *promijeniti*, no ona se u pismu prikazuje jednakim grafemskim slijedom kao i dvosložne sekvencije *ije* u npr. *pijem*, *pijemo*, *dvije*, *prije*. U algoritam su unesena pravila prema kojima se u riječima s jednosložnim izgovorom te sekvencije ona čitava nalazi u jednom slogu, a ako je izgovor dvosložan, sekvenciju razdvaja slogovna granica između *i* i *je*. Posuđenice poput *pacijent*, *koeficijent*, *dijeta* također se rastavljaju prema pravilu o dvosložnosti sekvencije *ije*, iako se u izgovoru hrvatskih govornika u tim riječima tolerira i jednosložan izgovor (o čemu će biti više riječi u jednome budućem radu).

Sonant /r/ smatrat će se slogotvornim u poziciji između dvaju suglasnika (osim kad je drugi suglasnik /j/, o čemu se govori također u potpoglavlju 2.1. Struktura sloga i poglavlju 3. Postupak automatskog rastavljanja na slogove), tj. postupit će se prema fonološkome kriteriju, a ne tvorbenome (smatrat će se da se u primjerima poput *umro*, *istro* dogodila fonološki uvjetovana alternacija: slogotvorni /r̥/ zamijenjen je suglasnikom /r/ u poziciji ispred vokala). Tako će se postupiti i zato što smo mišljenja da se u suvremenoj hrvatskoj ortoepiji navedeni primjeri u neutralnom izgovoru ostvaruju [ìstro] i [ùmro], a ne [ìstr̥o] i [ùmr̥o]. Slijed vokala i /r̥/ pojavljuje se u rijetkim primjerima na prefiksalno-korijenskoj granici (npr. *zarđati*, *porvati*, *zarzati*), a uz to svi se ti leksemi u suvremenome hrvatskom jeziku susreću i u inačici izvedenoj iz istih osnovnih riječi, ali s tzv. protetskim *h* (*zahrđati*, *pohrvati*, *zahrzati*) – ta je inačica u nekim izvorima ocijenjena i normativno preporučljivijom. U literaturi je komentiran i jedan (prema našim spoznajama, ujedno i jedini) primjer slijeda vokala

i /r̥/ na dvokorijenskoj granici: *zelenortski* (Brozović, 2001). U algoritmu su takvi primjeri zanemareni zbog njihove niske frekventnosti. Primjere s /r/ u tzv. bočnom slogu algoritam rastavlja tako da se /r/ tretira kao silabem: rastavlja se npr. *ma-sa-kr*, *ža-nr*.

3. POSTUPAK AUTOMATSKOG RASTAVLJANJA NA SLOGOVE

U ovome poglavlju opisan je postupak automatskog rastavljanja na slogove /[23]/. Postupak je definiran prema općim načelima rastavljanja na slogove te prema načelu najvećeg pristupa. Također je bilo potrebno uzeti u obzir specifičnosti i iznimke koje se pojavljuju u hrvatskome jeziku. Prema razmatranjima opisanima u prethodnim poglavljima, izdvojeno je 14 pravila na temelju kojih je definiran postupak automatskoga slogovanja:

P1. jedno mjesto u slogu obavezno zauzima jedan od samoglasnika (silabema);

P2. prije silabema moguć je slijed od najviše četiri suglasnička fonema;

P3. nakon silabema moguć je slijed od najviše tri suglasnička fonema;

P4. jedan suglasnik pred samoglasnikom pripada uvijek prvom slogu;

P5. suglasnički slijed kojim riječ može početi može stajati i na početku sloga;

P6. medijalni elementi kojima riječ ne može početi dijele se u dva sloga tako da se provjerava može li riječ započeti slijedom koji je početak sloga ili slijedom koji smo dodatno uvrstili kao dopušten slijed;

P7. načelo najvećeg pristupa: ako se primjenom pravila P6 dogodi da dolazi u obzir više mogućnosti, odabire se ono rastavljanje koje će rezultirati najvećim pristupom;

P8. ako se fonem /r/ nalazi između dvaju suglasnika, proglašava se slogotvornim /r̥/, osim ako iza njega slijedi /j/ (npr. u *vrtovi* je /r̥/, a u *vrjednovati* je /r/ – više o slijedu *SCrje* v. u poglavlju 2.1. Struktura sloga);

P9. ako se fonem /r/ nalazi na početku riječi, a nakon njega slijedi suglasnik, proglašava se slogotvornim /r̥/, osim ako iza njega slijedi /j/ (npr. u *rzati* je /r̥/, a u *rješenje* je /r/);

P10. ako se fonem /r/ nalazi na kraju riječi iza suglasnika i ako takav slijed odgovara slijedu potvrđenome na početku riječi, proglašava se slogotvornim /r̥/;

P11. ako je slijed *-ije-* od *jata* (osim u *dvije* i *prije*), čitav pripada istome slogu;

P12. slijed koji se bilježi kao *-naest-* (*dvanaest*, *dvanaestica* i sl.) ne rastavlja se na dva sloga;

P13. slijed *-nj-* u nekim se iznimkama tretira kao dvofonemski, a ne kao /ń/ /[24]/;

P14. slijed *-dž-* u nekim se slučajevima tretira kao dvofonemski, a ne kao /ʒ́/ /[25]/.

Pravila P1 te P4–P7 opća su pravila slogovanja koja vrijede i za druge jezike. Pravila P2, P3 te P8–P14 odnose se na riječi koje se prema hrvatskome pravopisu pišu prema morfonološkome načelu i tradicijskome načelu (za pojam tradicijskoga načela v. Matešić, 2014). Pravila P6 i P7 posebno se odnose na problem određivanja granice između dva sloga ako se u sredini riječi nakon nositelja sloga naiđe na slijed s više od jednog suglasnika (VCCV, VCCCV, VCCCCV, VCCCCCV). Svi modeli rastavljanja za sve moguće sljedove (uključivši i slijed VCV koji ima jednoznačno rastavljanje) prikazani su u Tablici 1.

Tablica 1. Mogući modeli granica između slogova
Table 1. Possible models of the syllable boundaries

Slijed u riječi	Mogući modeli rastavljanja	Ostvareni modeli rastavljanja	Broj ostvarenih od mogućih modela rastavljanja	Modeli definirani prema pravilu
VCV	V–CV VC–V	V–CV	1 od 2 moguća modela rastavljanja	Neostvareni model zbog P4
VCCV	V–CCV VC–CV VCC–V	V–CCV VC–CV	2 od 3 moguća modela rastavljanja	Neostvareni model zbog P7
VCCCV	V–CCCV VC–CCV VCC–CV VCCC–V	V–CCCV VC–CCV VCC–CV	3 od 4 moguća modela rastavljanja	Neostvareni model zbog P7
VCCCCV	V–CCCCV VC–CCCV VCC–CCV VCCC–CV VCCCC–V	V–CCCCV VC–CCCV VCC–CCV VCCC–CV	4 od 5 mogućih modela rastavljanja	Neostvareni model zbog P3
VCCCCCV	V–CCCCCV VC–CCCCV VCC–CCCV VCCC–CCV VCCCC–CV VCCCCC–V	VC–CCCCV VCC–CCCV VCCC–CCV	3 od 6 mogućih modela rastavljanja	Neostvareni modeli zbog P2 i P3

Uzevši u obzir pravila P2, P3 i P4, dobiva se 13 ostvarivih modela rastavljanja od ukupno 20 mogućih. Primjenom pravila P6, prema kojem su moguća ona rastavljanja kojima se dobiva takav slijed suglasnika koji može biti početak riječi, ne osigurava se jednoznačnost postupka. Naprimjer, u riječi *istražiti* u njezinu početnom slijedu VCCCV (*istra*) moguća su rastavljanja: *i-stra* (V–CCCV), *is-tra* (VC–CCV) te *ist-ra* (VCC–CV). Iz tog je razloga uvedeno pravilo P7, kojim se zadaje primjena rastavljanja po načelu najvećega pristupa i tako se osigurava jednoznačnost postupka.

Postupak za rastavljanje na slogove definiran je tako da se u znakovnom nizu automatski pronalazi nositelj sloga. Posebno se za svako pojavljivanje fonema /r/ provjerava radi li se o slogotvornome /r̥/.

U idućem koraku, ovisno o broju suglasnika koji slijede nakon nositelja sloga, ispituju se sve mogućnosti rastavljanja redoslijedom koji je prikazan u Tablici 1 i osigurava se najveći pristup. Za moguću se granicu provjerava postoji li pojavljivanje toga slijeda fonema na početku riječi. Ako postoji, definirana je slogovna granica i nastavlja se s preostalim nizom fonema. Ako promatrani niz fonema ne može biti početak riječi, provjerava se sljedeća moguća granica.

Kako se kontinuanta nekadašnjega dugoga jata na pravopisnome planu predstavlja slijedom *ije*, a pri slogovanju je treba smatrati jednosložnim slijedom, napravili smo bazu "dugojatovskih" primjera, prilagođenu potrebama našega postupka automatskog slogovanja. Baza sadrži sljedove s "dugim jatom", koji uglavnom odgovaraju korijenskim morfemima i alomorfima. Međutim u slučajevima kad je prijetila homografija prema slijedu *ije* koji nije dugi jat, to se razrješavalo drugačije, tj. navođenjem najmanjega razlikovnog slijeda odnosno najmanjih razlikovnih sljedova. Tako primjerice za *smiješiti se* nije bilo dovoljno navesti korijen /*smiješ*/ zbog homografije s 2. licem prezenta glagola *smjeti* (u kojem *ije* nije "dugojatovsko", tj. od jata potječe tek alternacija *i* ispred *je*), nego su u takvim slučajevima raspisani svi mogući sljedovi – za spomenuti primjer to su: *smiješi*, *smiješe*). Ta je baza javno dostupna na poveznici http://langnet.uniri.hr/resources.html.

Opisani postupak automatskog slogovanja implementiran je u programskom jeziku Python te je javno dostupan na poveznici http://langnet.uniri.hr/products.html.

4. REZULTATI

U ovome poglavlju prikazani su rezultati statističke analize koji se dobivaju primjenom implementiranoga postupka za rastavljanje na slogove na dvama odabranim

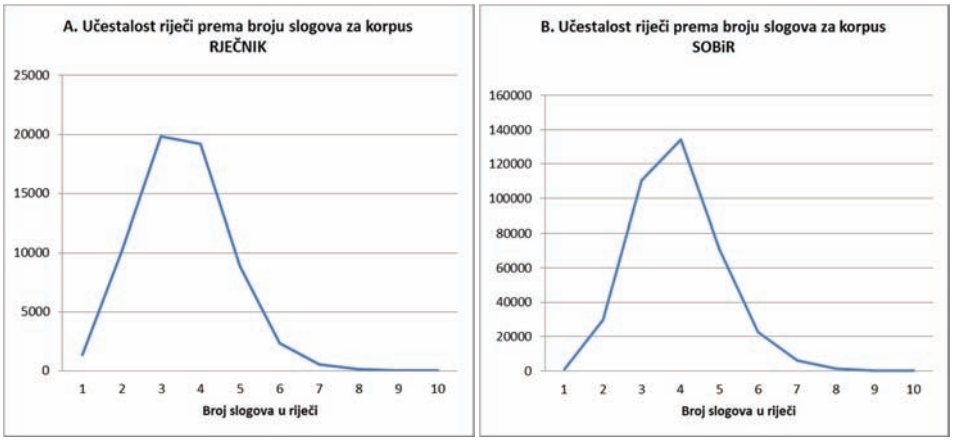
korpusima: RJEČNIK i SOBiR. Korpus RJEČNIK obuhvaća lekseme iz Anićeva rječnika (2003) i sadrži 62 387 leksema, što nije potpuni sadržaj Anićeva rječnika jer su izostavljeni toponimi /[26]/ i homografi. Korpus SOBiR obuhvaća sve paradigmatske oblike riječi i sadrži 377 143 jedinice (korpus SOBiR također ne sadrži homografe ni toponime) /[27]/. U obama korpusima nalaze se, naravno, i riječi stranoga podrijetla, uz uvjet da su pravopisno integrirane u hrvatski jezik (tako se u korpusima primjerice zadržavaju riječi poput: *bicikl*, *monokl*, *pneumatika* i sl., a uklonjene su riječi poput *pianissimo*, *pizza* ili *defaultni* i sl.).

Budući da je jedan od ciljeva u ovome istraživanju bio dobiti što je moguće iscrpniji popis slogova, odluka o odabiru resursa donesena je na sljedeći način: tekstni korpusi nisu došli u obzir kako rezultati ne bi ovisili o uzorku tekstova, već su odabrani opisani leksički resursi, koji su dali potpuniji inventar slogova nego što bi ga dao ikoji korpus tekstova.

Rezultati u obliku slogovanog popisa riječi iz korpusa RJEČNIK i SOBiR, pri čemu su riječi rastavljene na slogove temeljem načela najvećega pristupa, javno su dostupni na poveznici http://langnet.uniri.hr/products.html.

Analizirana je učestalost riječi prema duljini slogova, odnos otvorenih i zatvorenih slogova, zastupljenost svih modela slogova za oba korpusa. Nadalje, za oba korpusa analizirana je zastupljenost svakoga pojedinog slogovnog modela prema poziciji sloga u riječi, odnosno zastupljenost slogovnog modela u jednosložnim riječima, na početku riječi, u sredini riječi i na kraju riječi te složenost jezgre sloga. Ta je statistička analiza provedena prema parametrima relevantnima za lingvistički opis hrvatskoga sloga sa svrhom doprinosa statističkome opisu strukture hrvatskoga sloga.

Na Slici 3 prikazane su distribucije riječi prema duljini slogova. Prvi graf prikazuje distribuciju za korpus RJEČNIK, a drugi graf za korpus SOBiR. U oba slučaja najfrekventnije su riječi s duljinom između dva sloga i pet slogova, s time da su riječi iz korpusa SOBiR u prosjeku duže, što je i očekivano jer su obuhvaćeni svi paradigmatski oblici riječi. Prema Škariću (1991), prosječna govorna riječ (na korpusu govora javnih glasila) ima 3,12 slogova, a jezična 2,25, međutim zbog različitosti korpusa rezultate nije moguće uspoređivati. U korpusu RJEČNIK riječ ima prosječno 3,54 sloga, a u korpusu SOBiR riječ ima prosječno 3,92 sloga.



Slika 3. Učestalost pojavljivanja riječi prema broju slogova za korpus RJEČNIK (A) i korpus SOBiR (B)

Figure 3. Word frequencies according to the number of syllables for RJEČNIK corpora (A) and SOBiR corpora (B)

Nadalje, provedena je analiza odnosa između zastupljenosti otvorenih i zatvorenih slogova, što je prikazano u Tablici 2. Statistika slogova "s ponavljanjem" uzima u obzir ukupan broj pojavljivanja slogova, dok statistika "samo različiti" uzima u obzir jedinstveno pojavljivanje pojedinoga sloga. Naprimjer, slog *da* "s ponavljanjem" uzet je u obzir 18 528 puta, dok je postupkom "samo različiti" uzet u obzir samo jednom.

Ako se promatraju sva pojavljivanja slogova "s ponavljanjem" za korpus RJEČNIK, taj odnos je 81:19 u korist otvorenih slogova. Ako se pak promatraju sva pojavljivanja slogova za korpus SOBiR, odnos je otvorenih prema zatvorenim slogovima 84:16 (Tablica 2). Ti su odnosi vrlo blizu već objavljenih rezultata (Škarić, 1991) koji navode odnos 85:15 u korist otvorenih slogova. Iz same naravi uspoređenih korpusa sasvim je jasno da su rezultati za korpus SOBiR bliži Škarićevim rezultatima jer je on svoje istraživanje proveo na korpusima tekstova, a ne na rječniku.

Međutim, ako analiziramo "samo različite" slogove za oba korpusa, odnos otvorenih i zatvorenih slogova bitno je drugačiji. Očekivano, zbog manjeg broja različitih jedinica u skupu samoglasnika u hrvatskome jeziku u odnosu na veći broj različitih jedinica u skupu suglasnika rezultati potvrđuju veću raznolikost zatvorenih slogova, kao što je prikazano u Tablici 2 (84:16 u korist zatvorenih slogova za korpus RJEČNIK i 83:17 u korist zatvorenih slogova za korpus SOBiR).

Tablica 2. Zastupljenost otvorenih i zatvorenih slogova u korpusima
Table 2. Open and closed syllables contribution in the corpora

	Korpus RJEČNIK				Korpus SOBiR			
	S ponavljanjem		Samo različiti		S ponavljanjem		Samo različiti	
Vrsta sloga	Broj slogova	%	Broj slogova	%	Broj slogova	%	Broj slogova	%
Otvoren slog	177 941	81	841	16	1 240 097	84	905	17
Zatvoren slog	43 022	19	4 279	84	236 421	16	4 350	83
Ukupno slogova	220 963		5 120		1 476 518		5 255	

Zatim je napravljena i podrobna analiza zastupljenosti svakoga pojedinog modela sloga za oba korpusa. Numerički rezultati u Tablici 3 pokazuju da je najzastupljeniji model CV s 57% za korpus RJEČNIK i 60% za korpus SOBiR, a nakon toga slijede oblici CCV, CVC i V. Svi ostali oblici imaju vrlo malu zastupljenost (manje od 3%). Postoci zastupljenosti modela približno su isti za oba korpusa. Postoji neznatna razlika kod modela slogova koji imaju nisku zastupljenost: između 2% i 1%. Naprimjer, za korpus RJEČNIK na šestom mjestu pojavljuje se model CVCC s 2%, a u korpusu SOBiR takav se model pojavljuje s 0,3% i nalazi se na osmom mjestu u redoslijedu pojavljivanja.

Pojavljivanje modela VCCC, koje nije uobičajeno za hrvatski jezik (Turk, 1992), može se naći u vrlo malom postotku u oba korpusa budući da oba korpusa sadrže i riječi stranoga podrijetla.

Mali je postotak (0,02%) u automatskome postupku rastavljanja dobivenih oblika u korpusu SOBiR pogrešan, tj. ne odgovara dopuštenim slogovnim modelima. Takvi su naprimjer oblici C, CC, CCC i sl., koji potječu od veznika i prijedloga koji u izrazu imaju samo jedan konsonant (npr. *s*, *k*) te od onomatopejskih izraza i konvencionalnih uzvika (npr. *ššš*, *hm*, *pst*, *khm* i sl.). Još je manji postotak takvih slogova u korpusu RJEČNIK.

Tablica 3. Zastupljenost slogovnih modela za slogove iz korpusa "s ponavljanjem"
Table 3. Syllable models contribution for two corpora "with repetition"

Model sloga		Korpus RJEČNIK		Korpus SOBiR	
		Broj slogova	%	Broj slogova	%
1	CV	125 845	56,953	883 749	59,8536
2	CCV	37 746	17,082	250 064	16,9361
3	CVC	28 585	12,937	166 224	11,2578
4	V	11 404	5,161	89 427	6,0566
5	CCVC	6 358	2,877	45 152	3,0580
6	CVCC	4 093	1,852	4 489	0,3040
7	CCCV	2 946	1,333	16 857	1,1417
8	VC	2 435	1,102	16 793	1,1373
9	CCVCC	962	0,435	1 114	0,0754
10	CCCVC	478	0,216	2 430	0,1646
11	VCC	44	0,020	81	0,0055
12	CVCCC	39	0,018	70	0,0047
13	CCCVCC	18	0,008	51	0,0035
14	CCVCCC	8	0,004	13	0,0009
15	VCCC	2	0,001	4	0,0003
16	CCCVCCC	0	0,000	0	0,0000
	Drugi modeli	0	0,000	240	0,0163
	Ukupno	220 963		1 476 518	

Nakon toga ponovljena je analiza zastupljenosti slogovnih modela samo za različite slogove, dakle analizirani su slogovi bez ponavljanja. Rezultati prikazani u Tablici 4 pokazuju da je najfrekventniji oblik CVC s 34% u korpusu RJEČNIK i 32% u korpusu SOBiR, a potom slijede oblici CCVC, CCV te CVCC. Dakle, kao što je već prikazano u odnosima otvorenih i zatvorenih slogova, ako se analiziraju samo različiti slogovi, bitno je veća zastupljenost zatvorenih slogova.

Tablica 4. Zastupljenost slogovnih modela samo za "različite slogove" u korpusima
Table 4. Syllable models contribution for two corpora "unique syllables"

Model sloga		Korpus RJEČNIK		Korpus SOBiR	
		Broj slogova	%	Broj slogova	%
1	CVC	1 742	34,02	1 677	31,91
2	CCVC	1 598	31,21	1 869	35,57
3	CCV	557	10,88	606	11,53
4	CVCC	411	8,03	323	6,15
5	CCVCC	213	4,16	150	2,85
6	CCCVC	168	3,28	159	3,03
7	CV	155	3,03	163	3,10
8	CCCV	123	2,40	130	2,47
9	VC	82	1,60	84	1,60
10	CVCCC	26	0,51	13	0,25
11	VCC	21	0,41	34	0,65
12	CCCVCC	9	0,18	8	0,15
13	CCVCCC	7	0,14	4	0,08
14	V	6	0,12	6	0,11
15	VCCC	2	0,04	4	0,08
16	CCCVCCC	0	0,00	0	0,00
	Drugi modeli	0	0,00	25	0,48
	Ukupno	5 120		5 255	

Nadalje analizirana je zastupljenost određenih modela slogova prema poziciji u riječi. Postoje četiri mogućnosti: model se pojavljuje kao jedini član jednosložne riječi, model se pojavljuje na početku višesložne riječi, model se pojavljuje u sredini višesložne riječi i model se pojavljuje na kraju višesložne riječi. Rezultati su prikazani u Tablici 5.

Tablica 5. Zastupljenost različitih modela (%) prema poziciji sloga po jednosložnim i višesložnim riječima za korpuse RJEČNIK i SOBiR

Table 5. Different syllable models (%) according to the position in the monosyllabic and polysyllabic words for the corpora RJEČNIK and SOBiR

Model sloga	% modela u riječima – RJEČNIK				% modela u riječima – SOBiR			
	Jednosložne riječi	Višesložne riječi			Jednosložne riječi	Višesložne riječi		
	MONO	INIT	MED	FINAL	MONO	INIT	MED	FINAL
VCCC	0,07	0,00	0,00	0,00	0,17	0,00	0,00	0,00
VCC	0,36	0,01	0,00	0,05	1,34	0,01	0,00	0,00
VC	2,39	3,00	0,37	0,34	2,25	3,67	0,38	0,06
V	0,07	**13,93**	2,82	0,26	0,08	**17,12**	2,58	1,69
CVCCC	1,30	0,01	0,01	0,02	0,50	0,01	0,00	0,00
CVCC	**9,78**	0,13	0,26	**5,94**	**8,60**	0,11	0,28	0,52
CVC	**40,87**	**8,14**	**9,03**	**23,35**	**36,89**	**7,75**	**8,00**	**20,95**
CV	3,04	**51,92**	**65,89**	**48,91**	4,92	**49,03**	**66,21**	**58,58**
CCVCCC	0,22	0,00	0,00	0,00	0,08	0,00	0,00	0,00
CCVCC	**5,14**	0,02	0,05	1,35	4,09	0,03	0,06	0,14
CCVC	**30,36**	2,19	2,19	4,04	**29,97**	1,86	1,99	**6,21**
CCV	2,39	**19,11**	**17,79**	**14,25**	**6,18**	**19,04**	**18,77**	**11,33**
CCCVCC	0,22	0,01	0,00	0,01	0,33	0,00	0,00	0,00
CCCVC	3,70	0,19	0,18	0,22	3,17	0,15	0,19	0,12
CCCV	0,07	1,34	1,40	1,25	1,34	1,19	1,52	0,37
Drugi modeli	0,00	0,00	0,00	0,00	0,08	0,03	0,01	0,03

U korpusu RJEČNIK u jednosložnim se riječima (MONO) najučestalije pojavljuju slogovni modeli CVC, CCVC, CVCC i CCVCC (86%). Na početku višesložnih riječi (INIT) najučestaliji su CV, CCV, V i CVC (93%). U središnjem su dijelu višesložnih riječi (MED) najzastupljeniji: CV, CCV i CVC (93%). U finalnome dijelu višesložnih riječi (FINAL) najčešći su: CV, CVC, CCV i CVCC (92%). U korpusu SOBiR učestalost je simetrična onoj u RJEČNIKU u pozicijama INIT i

MED, dok je neznatno različita u poziciji MONO (prva tri modela su podudarna, no četvrti je CCV) i poziciji FINAL (ponovo su prva tri modela jednaka, a četvrti je različit, CCVC). Modeli u čijoj se statističkoj zastupljenosti razlikuju korpusi RJEČNIK i SOBiR u svakome su korpusu zastupljeni s 5–6%, a razlika ide u smjeru manje kompleksnosti odstupnoga dijela sloga u SOBiR-u, što se može protumačiti fleksijskim razlozima, a što potvrđuju ove pojave: pojednostavnjivanje odstupa u finalnome slogu (slogu na koji u hrvatskome jeziku utječe fleksija) te porast broja slogova s nepopunjenim odstupom u jednosložnih riječi (pojedini paradigmatski oblici jednosložne su riječi, npr. neki oblici prezenta pomoćnih glagola, *bi* kao dio kondicionala, nenaglašeni zamjenički oblici itd.).

Oprečna gledišta u suvremenome hrvatskom jezikoslovlju na status tzv. dugoga jata (sekvencija *ije* kad je njome u pismu predstavljena jednosložna alternanta kontinuante nekadašnjega dugoga jata u suvremenome hrvatskome standardnom jeziku), ponovimo, glase: 1) tzv. dugi jat ima vrijednost [j]/[i̯] + [e] te 2) tzv. dugi jat ima vrijednost dvoglasnika [ie̯]. Ta se razlika u gledištu može odraziti na analizu kompleksnosti jezgre sloga ako se na dvoglasničku jezgru gleda kao na dvočlanu. Dvočlanost pritom treba shvatiti kao dva člana na razini jednoga te istoga sloga, što je važno napomenuti zato što je u znanosti utvrđeno da je tzv. dugi jat u ortoepiji hrvatskoga standardnog jezika jednosložan, što je prihvaćeno i u obama navedenim gledištima. Dvočlanost jezgre koju čini tzv. dugi jat proizlazi iz činjenice da takav jat ima vrijednost dviju kratkih jedinica, što se u teoriji sloga često prikazuje grananjem jezgre. Naravno, i svaka druga duga vokalska jedinica sastoji se od dviju kratkih podjedinica, ali su one uvijek jednake (dugi *a* grana se u dva kratka *a*, dugi *e* grana se u dva kratka *e* itd.), za razliku od tzv. dugoga jata koji se grana u dvije različite podjedinice (*i* i *e*). U okviru teorije o (statističkoj) kompleksnosti jezgre to znači da se jezgra koju čini dugi *a* neće smatrati kompleksnom, dok će ona koju čini tzv. dugi jat biti kompleksna. Drugim riječima, prema gledištu 1) jezgra svih slogova u hrvatskome standardnom jeziku je nekompleksna, a prema gledištu 2) dio slogova ima kompleksnu jezgru. Stoga je statistika jezgre sloga u ovome radu dana za sve jedinice koje se mogu smatrati nositeljima sloga (V) bilo prema gledištu 1) bilo prema gledištu 2). U Tablici 6 prikazana je distribucija pojedinih vokala u slogovima koji se nalaze u jednosložnim riječima (MONO), na početku višesložne riječi (INIT), u sredini višesložne riječi (MED) i na kraju višesložne riječi (FINAL). Statistika koja se temelji na gledištu 1) pokazuje da su najčešće jezgre sloga u jednosložnih riječi *a* i *e* (28% i 18%), u inicijalnim slogovima najčešće su jezgre *a* i *o* (26,6% i 23%), u medijalnima

a i *i* (svaka po 30%), a u finalnima *i* i *a* (42% i 34%). Statistika koja se temelji na gledištu 2) pokazuje da jezgra *e* s drugoga mjesta po zastupljenosti u poziciji MONO pada na posljednje, sedmo mjesto, te je u toj poziciji druga najzastupljenija jezgra *o* (dvoglasnik je peti po zastupljenosti). I u ostalim pozicijama zastupljenost *e* pada s trećega (u INIT) i četvrtoga (u MED i FINAL) na posljednje mjesto. Ukupno gledano nema razlike u čestotnosti za prve tri najzastupljenije jezgre (usp. "ukupno" u Tablici 6), no od četvrtoga mjesta naniže razlike su velike, i to najveće za jezgru *e*. Njezina zastupljenost naime ukupno pada s četvrtoga na posljednje, sedmo mjesto.

Tablica 6. Statistika jezgre sloga
Table 6. Nuclei statistics

Silabem	Ukupno	%	MONO	%	INIT	%	MED	%	FINAL	%
a /[28]/	66 616	30,15	383	27,75	16 218	26,58	29 361	30,09	20 654	33,86
1) e+ije /[29]/	30 868	13,97	254	18,41	11 835	19,40	14 090	14,44	4 689	7,69
2) e /[30]/	986	0,45	56	4,06	322	0,53	544	0,56	64	0,10
2) ie̯	29 882	13,52	198	14,35	11 513	18,87	13 546	13,87	4 625	7,58
i	64 736	29,30	230	16,67	9 288	15,22	29 400	30,13	25 818	42,32
o	39 247	17,76	239	17,32	14 290	23,42	15 805	16,20	8 913	14,61
u	16 149	7,31	211	15,29	7 522	12,33	7 542	7,73	874	1,43
r̥	3 347	1,51	63	4,57	1 854	3,04	1 371	1,41	59	0,10
Zbroj	220 963		1 380		61 007		97 569		61 007	

5. EVALUACIJA POSTUPKA I PROCJENA POGREŠKE

Korpus RJEČNIK sadrži 38% riječi kod kojih je određivanje granice slogova jednoznačno jer se u njima ne ostvaruju uzastopni sljed ovi s više od jednog suglasnika. U preostalih 62% riječi potrebno je uzeti u obzir složenija pravila za određivanje granice (P2–P14).

Postupak slogovanja provjeren je na 500 riječi iz korpusa RJEČNIK. Skup za evaluaciju od 500 riječi dobiven je slučajnim odabirom riječi koje sadrže slijed dvaju i više suglasnika (CC, CCC...). Udio riječi u kojima je slijed konsonanata i vokala izmjeničan (CV, CVCV, CVCVCV… te CVC) u korpusu RJEČNIK iznosi 38% (a

slično je i u korpusu SOBiR budući da u hrvatskome jeziku promjena izraza u paradigmatskim oblicima uglavnom pridonosi izmjeničnosti slijeda konsonanata i vokala). Kako se riječi s takvim, izmjeničnim slijedom konsonanata i vokala sloguju sa stopostotnom točnosti, izostavili smo ih iz postupka provjere te smo se usmjerili k dijelu gdje je očekivana pogreška. Postupak provjere temeljili smo na usporedbi rezultata automatskog postupka slogovanja s ručno rastavljenim rezultatom (rastavljanje je proveo lingvist). Ustanovili smo pogrešku kod 11 riječi (naprimjer: prijedlog *k*, prijedlog *s*, u nekim je riječima ovim postupkom nemoguće detektirati *r* kao slogotvorno, npr. *zelenortski*, *zarzati*, *porvati* i sl.). Dakle, ukupna pogreška na 62% riječi iz korpusa RJEČNIK je oko 2%, čime je procjena ukupne pogreške na 100% riječi ispod 2% (pritom treba imati na umu kako su korpusi priređeni – v. početak poglavlja Rezultati). U skladu s time iz ovoga se rezultata može očekivati pogreška sličnoga ranga i na korpusu SOBiR. Ocjenu pogreške moguće je dodatno aproksimirati temeljem pronalaska 240 slogova koji nisu mogući prema hrvatskom modelu rastavljanja, a dobiveni su iz riječi stranoga podrijetla (zadnji redak u Tablici 3), što povećava pogrešku za najmanje 0,02%.

6. ZAKLJUČAK

Analiza slogova kao osnovnih elemenata jezika važna je za različite postupke u domeni računalne analize prirodnog jezika i govornih tehnologija. U ovome je radu predložen postupak automatskoga slogovanja hrvatskih riječi temeljen na načelu najvećega pristupa, koji osigurava jednoznačnost postupka rastavljanja. Prema tome se načelu pristupu dodjeljuju svi suglasnici u riječi za koje je to moguće, uz uvjet da se pritom ne krše pravila o broju pristupnih mjesta, sonornosti te ograničenju sljedova.

Na rezultatima provedenoga automatskog postupka slogovanja napravljena je statistička analiza kako bi se dobio statistički opis strukture hrvatskoga sloga, što je učinjeno prema parametrima relevantnima za lingvistički opis hrvatskoga sloga.

Statistička analiza distribucije slogova provedena je dakle za dva korpusa, različita po morfološkome opsegu jedinica: korpus RJEČNIK, koji obuhvaća popis hrvatskih leksema u tzv. kanonskom obliku, dobiven iz rječnika hrvatskoga jezika, te korpus SOBiR, koji sadrži popis svih oblika riječi hrvatskoga jezika. U njima su promatrani slogovni modeli – tj. strukture mogućih slogova – pri čemu je jezgra sloga predstavljena slogotvornim fonemom (V), a rubni dijelovi sloga suglasnikom ili suglasnicima (C).

Analizirani su: učestalost riječi prema duljini slogova, odnos otvorenih i zatvorenih slogova te zastupljenost svih modela slogova za oba korpusa. Nadalje, za oba je korpusa analizirana zastupljenost svakoga pojedinog slogovnog modela prema poziciji sloga u riječi, odnosno zastupljenost slogovnoga modela u jednosložnim riječima, na početku riječi, u sredini riječi i na kraju riječi te složenost jezgre sloga.

Usporedna analiza korpusa pokazala je:

- da su otvoreni slogovi učestaliji od zatvorenih slogova ako se promatraju ponavljanja;
- ako se promatraju slogovi bez ponavljanja, učestalost se u istom omjeru povećava u korist zatvorenih slogova;
- u korpusu koji sadrži sve oblike riječi udio otvorenih slogova povećava se za 3% zahvaljujući tome što gramatemi u velikom broju završavaju vokalom;
- najzastupljeniji model sloga s ponavljanjem je CV (57% za korpus RJEČNIK i 60% za korpus SOBiR), a nakon toga slijede oblici CCV (17% i 17%), CVC (13% i 11%) i V (5% i 6%);
- najzastupljeniji je model sloga za različite slogove CVC (34% za RJEČNIK i 32% za SOBiR), a potom slijede oblici CCVC (31% i 36%), CCV (11% i 12%) te CVCC (8% i 6%);
- u jednosložnim riječima u oba se korpusa najučestalije pojavljuju CVC, CCVC i CVCC (81% u RJEČNIKU, 76% u SOBiR-u); četvrti po učestalosti u RJEČNIKU je model CCVCC (5%), a u SOBiR-u CCV (6%);
- na početku višesložnih riječi u oba su korpusa najučestaliji: CV, CCV, V i CVC (93%);
- u središnjem dijelu višesložnih riječi u oba su korpusa najučestaliji: CV, CCV i CVC (93%);
- u finalnom dijelu višesložnih riječi u oba su korpusa najučestaliji modeli: CV, CVC i CCV (86% u RJEČNIKU, 91% u SOBiR-u); četvrti po učestalosti u RJEČNIKU je model CVCC (6%), a u SOBiR-u CCVC (6%).

Kompleksnost jezgre sloga prema nositelju sloga prikazana je na razini jednosložnih riječi te prema položaju na početku, sredini i završetku višesložnih riječi. Izračunata je posebno kompleksnost jezgre za dva slučaja koji odgovaraju sukobljenim gledištima na status te jedinice u suvremenome hrvatskom standardnom jeziku. Ako se tzv. dugi jat kao nositelj sloga smatra kompleksnom jedinicom, to se odražava na zastupljenost jezgre *e*, koja ukupno pada s četvrtoga na posljednje, sedmo mjesto.

Drugim riječima, i pojavnost i zastupljenost jezgre *e* znatno se povećava ako se slogovi koji sadrže "jat" smatraju slogovima s jezgrom *e*.

Analiza jezika na slogovnoj razini važna je za različite postupke u području računalne analize prirodnoga jezika. Predstavljeni postupak automatskoga slogovanja izravno je primjenjiv u postupcima automatskog raspoznavanja govora (Shafran i Ostendorf, 2003; De Jong i Wempe, 2009; Martinčić-Ipšić i sur., 2011) i sinteze govora (Žganec Gros i sur., 2002, 2006; Sečujski, 2005; Pobar i sur., 2012). Budući da slogovi mogu tvoriti akustičke jedinice (Ganapathiraju i sur., 2001), postupak automatskoga slogovanja može se primijeniti u automatskom određivanju prozodijskog modela (Načinović-Prskalo i Martinčić-Ipšić, 2013) i u automatskom određivanju vrste i mjesta naglaska (Žganec Gros i sur., 2006; Marinčič i sur., 2009) te također u analizi strukture slogovnog sustava pomoću kompleksnih mreža (Ban i sur., 2013).

Pored navedene tehnološke primjene ovo je istraživanje moguće nastaviti u nekoliko smjerova. U planu nam je razviti postupak automatskoga slogovanja temeljenoga na suprotnome načelu – načelu najvećega odstupa. Pored fonološkoga kanimo istražiti i postupke fonetskoga slogovanja u kojima bi se upotrijebili u okviru sustava za sintezu hrvatskoga jezika već razvijeni automatski postupci grafemsko-fonemske pretvorbe (Načinović i sur., 2009). Na temelju dobivenih rezultata za različite pristupe slogovanju namjeravamo razviti hibridni model koji bi udružio oba načela u jedinstven postupak i koji bi se približio modelu govornikove percepcije slogovanja, odnosno koji bi mogao oponašati prosječni kognitivni model hrvatskoga govornika pri slogovanju.

BILJEŠKE

[1] "Ako se hoće izgovoriti manje od sloga, npr. *s*, *š*, *ž* i sl., izgovaraju se puni slogovi [sə], [šə], [žə] ili se uopće ne izgovaraju, već se proizvode zvukovi *sss*, *ššš*, *žžž*" (Škarić, 1991: 82).

[2] Podjela/rastavljanje/raščlamba riječi na slogove naziva se još i *silabacija* (Muljačić, 1972: 149; Turk, 1992: 34) te *silabifikacija* (u nizu izvornih i prijevodnih djela iz lingvistike, fonetike, psihologije itd.). U ovome se radu rabi termin *slogovanje* koji je ponudila Jelaska (2004: 172).

[3] To je nezaobilazna tvrdnja u literaturi (npr. Malmberg, 1954/1995: 65; Simeon, 1969: 430; Trask, 2005: 323 itd.); Jelaska (2004: 100) također: "... raščlamba na slogove jedna [je] od osnovnih govornih sposobnosti, najvjerojatnije povezana s kognitivnim razvojem (...)".

[4] Fiziološki pristupi opisuju slog opažanjem načina na koji se zrak potiskuje iz pluća kroz govorne organe. Tako se za nastanak sloga odsudnim smatra naprimjer niz mišićnih stezanja i opuštanja pri procesu disanja (tzv. pulsna teorija R. H. Stetsona iz 1928. godine), sudjelovanje disanja i grkljana (Ladefoged, 1982, 1993), trzajni potisak zraka iz pluća pri čemu se potisku naizmjenično suprotstavljaju artikulatori i fonatori, ponekad i udisajni mišići (Škarić, 1991: 328). Psiholingvistički opisi ističu da je slog i apstraktna jedinica i najmanja opažajna jedinica govora (Jelaska, 2004: 102). To potvrđuju naprimjer pojave kao što su: različito intuiranje granice sloga ovisno o polaznom jeziku govornika, različito rastavljanje homofona na slogove kako bi se i izrazom ostvarila razlika koja postoji u sadržaju, a i pojava da se pri reverzijskim izgovornim pogreškama (tzv. spunerizmima, engl. *spoonerisms*, v. Horga, 1996: 41–42) ne radi o zamjeni pojedinačnih glasova nego o zamjeni dijelova slogova, npr. *plazi jezik* > *jazi plezik* (v. npr. Jelaska, 2004: 102).

[5] Prema toj ljestvici, najzvonkiji su samoglasnici, a najmanje su zvonki bezvučni suglasnici. Ostali glasovi grupiraju se i raspoređuju između tih dviju krajnosti. Raspoređivanje se odvija na temelju stavljanja glasova u međusobnu relaciju s obzirom na to od kojih su glasova više, a od kojih manje zvonki. Različiti autori identificiraju i različit broj stupnjeva na toj ljestvici – podrobnije o toj temi progovara Jelaska u ovim svojim radovima: Babić (1989: 67) te Jelaska (2004: 132–144).

[6] Hrvatski naziv preuzet iz članka Babić (1989).

[7] Na slici su prikazane međunarodne oznake za slogovne dijelove. Hrvatske oznake, prema prijedlogu Jelaske (2004), odgovaraju početnome slovu hrvatskoga naziva za pojedini slogovni dio. Uzgred, ista autorica u svojem ranijem radu (Babić, 1989) rabi za *rimu* naziv *srok* (i u skladu s time oznaku *S*).

[8] Uzgred, autorica je generativni pristup primijenila i pet godina prije u magistarskome radu *Generativni prikaz prozodema plodnih glagola hrvatskoga književnog jezika* (Babić, 1984).

[9] Strukturalističke fonološke teorije o njima govore kao o fonemima s razlikovnim obilježjem [+*silabičan*].

[10] Termin prema Turk (1992: 42–43).

[11] Za potrebe ovoga istraživanja neće se problematizirati nazivi "alternanta *ije*", "alternanta [*jē*]" te "dvoglasnik /ie/" koji se svi susreću u suvremenim hrvatskim normativnim priručnicima (više o tome v. npr. Matešić, 2008 te Martinović, 2009). Valja međutim jasno reći nešto drugo, a to je da će se u ovoj raspravi i istraživanju smatrati da je početak te jedinice bliži suglasničkome ostvaraju (jer smo mišljenja da se on upravo tako čuje u suvremenome neutralnom općehrvatskom govoru) te se stoga i prihvaća da u primjerima poput [*sprjéčiti*] prvi slog ima strukturu CCCCV – za razliku od strukture CCCV, koju sugerira diftonški izgovor "jatovskoga" [*jē*].

[12] Jelaska se bavi i pitanjem je li taj *j* zapravo *suglasnički i* te daje generativni pogled na taj problem (2004: 124).

[13] Na Slici 2 nije prikazan tip CCCCV.

[14] Također, ako se *sprječavati* rastavi kao *sprječ-a-va-ti*, moguće bi bilo uočiti i tip CCCCVC, no njega ovdje ne uočavamo zato što smo odabrali rastavljanje prema načelu najvećega pristupa, prema kojem je moguće samo *sprje-ča-va-ti* (više o načelu najvećega pristupa v. u poglavlju 2.2.2. ovoga rada).

[15] Jelaska (2004: 172) upozorava i na razliku između govornog i pismovnog dijeljenja na slogove budući da su oba prisutna u iskustvu hrvatskih govornika.

[16] U literaturi se pri spominjanju distribucije fonema nailazi na nazive *skup*, *skupina*, *slijed* i *spoj* suglasnika. Na temelju uporabe tih naziva primjećuje se da se slijedom uglavnom naziva niz bilo kojih fonema, a skupom, skupinom i spojem uglavnom slijed suglasnika koji mogu imati status morfonema. U radovima je ipak uočljiva i povremena sinonimska uporaba tih termina (usp. npr. kontekste u kojima se sekvencije istoga ranga nazivaju slijedom, skupom i skupinom u Turk, 1992: 37–39, odnosno spojem, skupinom i slijedom npr. u Jelaska, 2004: 160). Za potrebe istraživanja u ovome radu upotrebljavaju se pojmovi *slijed* i *spoj* s jasno razgraničenim terminološkim značenjem, a neće se upotrebljavati termin *skupina*, ponajviše zato što se u literaturi često tako nazivaju one suglasničke sekvencije koje imaju status morfonema (npr. *st* u *premostiti* je morfonem, dok je *st* u *istisnuti* suglasnički slijed, kroz koji prolazi tvorbena granica). Je li nešto od slijeda ili spoja ujedno i skupina u značenju dvaju suglasnika koji čine morfonem, to za istraživanje slogova u ovome radu neće biti odsudno.

[17] U pravilima 2) i 3) izvorno se govori o *skupinama* kojima riječ može početi odnosno završiti, no ovdje smo primijenili drukčije nazivlje, kao što je prethodno objašnjeno.

[18] Ni on se ne ostvaruje na početku nijednog apelativa u hrvatskome standardnome jeziku, pa se kao oprimjerenja spominju početno *vn-* u prezimenu *Vnuk* (Junković, 1973: 47) ili nestandardnojezični, čakavski apelativ *vnuk* (Turk, 1992: 39), no ne bez metodoloških poteškoća povezanih s pitanjem rubnosti i autonomije imena ili pitanjem opsega utjecaja sustava na standard.

[19] Popisi koje donosi Turk (1992) zapravo informiraju o fonotaktičkim pravilima u hrvatskome jeziku, tj. usredotočeni su na odgovor na pitanje koliko se i kojih suglasnika može naći u hrvatskome jeziku u ovoj ili onoj poziciji.

[20] Taj se spoj ne susreće u suvremenome hrvatskom standardnom jeziku, nego samo u arhaizmima *počten*, *počtovan* i njihovim izvedenicama.

[21] Zamijećeno je primjerice da bogatiji odstup sloga imaju posuđenice, dok se u hrvatskih riječi u odstupu nalazi mali broj mogućnosti (usp. Turk, 1992: 37; usp. Jelaska, 2004: 150).

[22] Tradicijskome načelu pripadalo bi pisanje *ije* za dugu alternaciju [je] odnosno dvoglasnik *ie* u suvremenome hrvatskom standardnom jeziku (o tome terminu u vezi s hrvatskim jezikom v. Matešić, 2014).

[23] Postupak automatskog rastavljanja na slogove implementiran je u programskome jeziku Python. Implementacijske pojedinosti ne navodimo u ovome radu jer je ovdje cilj bio jasno predstavljanje zamisli i rezultata, pa bi tehnički detalji samo opteretili tekst. Autori planiraju

spomenuti postupak podrobno opisati u zasebnome radu i staviti ga putem mreže na otvoreno raspolaganje akademskoj zajednici.

[24] To se pravilo odnosi na ove primjere: *konjunkcija, konjunktivitis, konjugacija, konjugacijski, konjugata, konjugiran, konjugirati, konjunkcija, konjunktiv, konjunktura, konjunkturan, konjunkturist, konjunkturistički, injekcija, injekcijski, injektor, injektirati, injicirati*. Dakako, za potrebe uključivanja u algoritam izdvojeni su njihovi korijenski morfemi kako bi se na željeni način rastavili i pojedini oblici tih riječi u korpusu SOBiR.

[25] To se pravilo odnosi na ove primjere: *odživjeti, nadživjeti, nadživljavati, podžanr, podžeći, podžizati, podžupan*. Da bi se na željeni način rastavili i pojedini paradigmatski oblici tih riječi u korpusu SOBiR, u algoritmu su za te riječi uzete u obzir sve veze prefiksā i korijenskih morfema (kao i alomorfa korijenskih morfema, naravno).

[26] Iz RJEČNIKA su izostavljeni i oni leksemi koji se prema hrvatskome pravopisu pišu načelom transliteracije. U cjelini standardnoga jezika to je dio toponima iz stranih jezika i to su beziznimno antroponimi iz stranih jezika koji se služe latinicom, a konkretno iz RJEČNIKA takav je bio dio toponima (npr. Calcutta, Shangai, München itd.). Da bi se primjeri zapisani transliteracijom rastavili ispravno na slogove, potrebno bi bilo poštovanje posebnih pravila. Primjerice antroponim *Shakespeare* ovim bi se postupkom rastavljao kao *Sha-ke-spe-a-re* umjesto ispravnoga *Shake-speare* (ili *Shakes-peare*, što je također ispravno rastavljeno, no uz primjenu načela najvećega odstupa). Takav se problem ne pojavljuje, naravno, u primjerima koji su pravopisno integrirani u hrvatski jezik, tj. koji se prema pravopisnim pravilima pišu prema načelu transkripcije – primjer *šekspirolog* našim se automatskim postupkom rastavlja *šek-spi-ro-log*, što je ispravno.

[27] SOBiR je nastao na temelju korpusa u slobodnome dostupu, kojem je autor Goran Igaly. Riječ je o izvoru koji autor kontinuirano dopunjuje: http://www.igaly.org/rjecnik-hrvatskih-jezika/pages/.php?lang=HR.

[28] Slogovima koji sadrže jezgru *a* smatrani su i oni koji sadrže sekvenciju *naest*.

[29] Slogovi koji sadrže jezgru *e* (s ubrajanjem i onih slogova koji sadrže sekvenciju *ije* kad je njome u pismu predstavljena jednosložna alternanta kontinuante nekadašnjega dugoga jata u suvremenome hrvatskome standardnom jeziku).

[30] Slogovi koji sadrže jezgru *e* (bez ubrajanja onih slogova koji sadrže sekvenciju *ije* kad je njome u pismu predstavljena jednosložna alternanta kontinuante nekadašnjega dugoga jata u suvremenome hrvatskome standardnom jeziku).

REFERENCIJE

Anić, V. (2003). *Veliki rječnik hrvatskoga jezika*, 4. izdanje. Zagreb: Novi Liber.

Babić, Z. (1984). *Generativni prikaz prozodema plodnih glagola hrvatskoga književnog jezika*. Magistarski rad, Zagreb.

Babić, Z. (1989). Slogovna struktura hrvatskoga književnog jezika. *Jezik* **36**, 65–71, 123–128, 133–146.

Ban, K., Ivakić, I., Meštrović, A. (2013). A preliminary study of Croatian language syllable networks. *Proceedings MIPRO Junior – Student Papers* (ur. P. Biljanović), 1697–1701.

Brozović, D. (2001). Pravopisi i slovopisi u hrvatskoj povijesti. *Vijenac* **184**, 22. ožujka 2001.

Chomsky, N., Halle, M. (1968). *The Sound Pattern of English*. New York: Harper & Row.

Clements, G. N., Keyser, S. J. (1983). *CV Phonology: A Generative Theory of the Syllable*. Cambridge, Mass.: MIT Press.

De Jong, N. H., Wempe, T. (2009). Praat script to detect syllable nuclei and measure speech rate automatically. *Behavior Research Methods* **41**, 2, 385–390.

Ganapathiraju, A., Hamaker, J., Picone, J., Ordowski, M., Doddington, G. R. (2001). Syllable-based large vocabulary continuous speech recognition. *IEEE Transactions on Speech and Audio Processing* **9**, 4, 358–366.

Goldsmith, J. (1976). An overview of autosegmental phonology. *Linguistic Analysis* **2**, 23–68.

Halle, M., Vergnaud, J.-R. (1980). Three dimensional phonology. *Journal of Linguistic Research* **1**, 4, 83–105.

Horga, D. (1996). *Obrada fonetskih obavijesti*. Zagreb: HFD.

Igaly, G. (2014). portal projekta *Jedinstveni otvoreni rječnik HJ*, http://www.igaly.org /rjecnik-hrvatskih-jezika/pages/.php?lang=HR (posjet 10/2013).

Jelaska, Z. (1997). *Poredbeni opis glasovne strukture hrvatskoga jezika*. Doktorski rad, Zagreb: Filozofski fakultet.

Jelaska, Z. (2004). *Fonološki opisi hrvatskoga jezika: glasovi, slogovi, naglasci*. Zagreb: Hrvatska sveučilišna naklada.

Junković, Z. (1973). Struktura sloga i fonološka vrijednost suglasnika *v* u književnom jeziku. *Jezik* **21**, 1–5 i 37–52.

Kahn, D. (1976 [1980]). *Syllable-Based Generalizations in English Phonology*. Doktorska disertacija, New York: Garland.

Ladefoged, P. (1982). *A Course in Phonetics*. New York: HBJ.

Ladefoged, P. (1993). *A Course in Phonetics*, 4. izdanje. New York: HBJ.

Malmberg, B. (1995 [1954]). *Fonetika*. Zagreb: Ivor.

Marinčič, D., Tušar, T., Gams, M., Šef, T. (2009). Analysis of automatic stress assignment in Slovene. *Informatica* **20**, 1, 35–55.

Martinčić-Ipšić, S., Pobar, M., Ipšić, I. (2011). Croatian large vocabulary automatic speech recognition. *Automatika* **52**, 2, 147–157.

Martinović, B. (2009). Izgovor i pisanje imeničnih jednosložnica s jatom. *Jezik* **56**, 133–144.

Matešić, M. (2008). Jat – prilog za leksikografsku natuknicu. *Riječki filološki dani – Zbornik radova 7*, 491–505.

Matešić, M. (2014). Pravopis i tradicija: teorijsko-metodološki pristup jednome normativnom načelu u suvremenoj hrvatskoj pravopisnoj normi i praksi. *Riječki filološki dani – Zbornik radova 9*, 551–562.

Mihaljević, M. (1991). *Generativna i leksička fonologija*. Zagreb: Školska knjiga.

Muljačić, Ž. (1972). *Opća fonologija i fonologija suvremenoga talijanskog jezika*. Zagreb: Školska knjiga.

Načinović, L., Pobar, M., Ipšić, I., Martinčić-Ipšić, S. (2009). Grapheme-to-phoneme conversion for Croatian speech synthesis. *32. međunarodni skup MIPRO, zbornik radova, sv. 3: Računala u tehničkim sustavima. Inteligentni sustavi*, 318–323.

Načinović Prskalo, L., Martinčić-Ipšić, S. (2013). An overview of prosodic modelling for Croatian speech synthesis. *5th International Conference on Information Technologies and Information Society – ITIS 2013* (ur. Z. Levnajić), 105–112.

Oudeyer, P. Y. (2001). The origins of syllable systems: An operational model. *Proceedings of the International Conference on Cognitive Science*, *COGSCI*, 27–31.

Pobar, M., Martinčić-Ipšić, S., Ipšić, I. (2012). Optimization of cost function weights for unit selection speech synthesis using speech recognition. *Neural Network World* **22**, 5, 429–441.

Sečujski, M. S. (2005). Obtaining prosodic information from text in Serbian language. *Computer as a Tool, EUROCON-2005*, vol. 2, 1654–1657.

Shafran, I., Ostendorf, M. (2003). Acoustic model clustering based on syllable structure. *Computer Speech & Language* **17**, 4, 311–328.

Silić, J., Pranjković, I. (2005). *Gramatika hrvatskoga jezika: za gimnazije i visoka učilišta*. Zagreb: Školska knjiga.

Simeon, R. (1969). *Enciklopedijski rječnik lingvističkih naziva*. Zagreb: Matica hrvatska.

Škarić, I. (1991). Fonetika hrvatskoga književnog jezika. U: Povijesni pregled, glasovi i oblici hrvatskoga književnog jezika (Nacrti za gramatiku), 61–378. Zagreb: Hrvatska akademija znanosti i umjetnosti i Nakladni zavod Globus.

Trask, R. L. (2005 [1999]). *Temeljni lingvistički pojmovi*. Zagreb: Školska knjiga.

Turk, M. (1992). *Fonologija hrvatskoga jezika (raspodjela fonema)*. Rijeka–Varaždin: ICR i Tiskara Varaždin.

Žganec Gros, J., Cvetko-Orešnik, V., Jakopin, P. (2006). SI-PRON pronunciation lexicon: A new language resource for Slovenian. *Informatica* **30**, 4, 447–452.

Žganec Gros, J., Žganec, M., Mihelič, A., Knez, M., Merčun, A., Marinčič, D. (2002). The phonetic family of voice-enabled products. *Jezikovne tehnologije / Language Technologies V* (ur. T. Erjavec i J. Žganec Gros), 127–131.

Internetski izvori

http://langnet.uniri.hr/products.html
http://langnet.uniri.hr/resources.html

Ana Meštrović, Sanda Martinčić-Ipšić
amestrovic@uniri.hr, smarti@uniri.hr
Department of Informatics, University of Rijeka
Croatia

Mihaela Matešić
mmatesic@ffri.hr
Faculty of Humanities and Social Sciences, University of Rijeka
Croatia

Syllabification based on maximal onset principle for Croatian

Summary

Syllables are important units in natural language processing and speech applications. Syllabification of Croatian proposed in this paper is based on the maximal onset, which incorporates maximal number of consonants (C) considering the number of possible access points, sonority and allowed sequences. The syllabic model is the model of nuclei (V) and consonants (C) combinations. All possible combinations are reduced to the set of model realizations and the uniqueness of syllabification is ensured by the rules (P1–P14).

The maximal onset algorithm is tested on two different corpora: RJEČNIK – comprises lexemes from Croatian dictionary in canonical form, and SOBiR – the list of all flective word forms.

Results of the maximal onset syllabification algorithm are presented in the form of different statistical distributions: the frequency of words by the number of syllables, the number of open vs. closed syllables, all possible (and observed) syllabic models, the frequencies of syllables according to the position in a word (monosyllabic, initial, medial and final) and nuclei complexity. Results thus achieved are consistent with previously reported findings. The relation between open and closed syllables in lemmatized word forms is inverted in all flective forms. Additionally, we analyzed the models according to the repetition: "with repetition" and "unique syllables". The CV model is the most frequent (57% per RJEČNIK and 60% per SOBiR) considering syllables "with repetition" and CVC model is the most frequent (34% per RJEČNIK and 32% per SOBiR) considering "unique syllables". In monosyllabic words the prevailing models are CVC, CCVC, CVCC and CCVCC (86%); in polysyllabic words

prevailing models in the initial part are CV, CCV, V i CVC (93%), in the medial part are CV, CCV and CVC (93%) and in the final part are CV, CVC, CCV and CVCC (92%). Finally, algorithm was manually tested, and evaluated (approximated error is below 2%).

Key words: syllable, syllabification, distribution of syllables, maximal onset principle