

Automatinis skiemenavimas: problemos ir jų sprendimas

Gailius Raškinis, Asta Kazlauskienė

Anotacija. Natūraliai pasakytame žodyje tarp skiemenų nėra ryškių fonetinių ribų. Kalbai svarbesnis skiemenų skaičius, o ne tikslios ribos tarp jų. To užtenka ir kuriant skiemenavimo įrankį kalbos kompiuterinio apdorojimo reikmėms. Tačiau fundamentiniams skiemens struktūros tyrimams reikia tikslų skiemens ribų. Tam ir buvo kuriama automatinio skiemenavimo programa. Skiemens ribos čia nustatomos remiantis funkcinę skiemens teorija: naujas skiemuo pradedamas maksimaliu priebalsių junginiu, kurio modelis galimas žodžio pradžioje. Tačiau ne visada šis modelis (pučiamasis, sprogstamasis arba afrikata, sklandusis priebalsis) gali būti be išlygų taikomas. Straipsnyje aprašomi probleminiai skiemenavimo atvejai ir jų sprendimas: dviejų balsių samplaikos interpretavimas (dvibalsis, hiatas ar papriešakėjęs balsis), priešdėlių nustatymo sunkumai, morfemų sandūra sudurtiniuose žodžiuose. Skiemenavimo programa klaidų nedaro. Jų gali atsirasti dėl labai retų homografų, vienodai rašomų, bet skirtingai tariamų ir skiemenuojamų žodžių. Tokius atvejus reikia atskirai rankiniu būdu peržiūrėti.

Aprašytas skiemenavimo algoritmas naudojamas VDU sukurtose automatinio kirčiavimo ir transkribavimo programose bei fonotaktikos ir ritmikos tyrimuose.

Reikšminiai žodžiai: skiemuo; algoritmas; priešdėlis; skiemens centras; priebalsis.

Įvadas

Skiemuo – kalbėjimo srauto atkarpa, kurią daugelis intuityviai jaučiame, tačiau sunkiai apibrėžiame. Kalboje izoliuotų garsų ir skiemenų nėra. Natūraliai pasakytame žodyje tarp skiemenų dažnai nėra ryškių ribų. Dėl to egzistuoja įvairių skiemens ribų nustatymo teorijų. Vienos jų remiasi artikuliacinėmis ypatybėmis (kalbos padargų raukšlės įtampa, iškvepiamo oro kiekiu); kitos – akustiniais ar audicininiais požymiais (analizuojami spektriniai ypatumai, priebalsių skambumas). Tai, kad nėra patikimos metodikos, nustatančios objektyvias, fonetinėmis ypatybėmis paremtas skiemenų ribas, rodo, jog kalbai svarbesnis skiemenų skaičius, o ne tikslios ribos tarp jų. Galima būtų tikėtis, kad to užtenka ir kuriant skiemenavimo įrankį kalbos kompiuterinio apdorojimo reikmėms (kalbos technologijų produktams kurti ar leidybos programoms). Tačiau tokiu principu paremtos skiemenavimo programos nebus galima pritaikyti fundamentiniams skiemens struktūros tyrimams. Tuo įsitikinta naudojant pirminį aprašomos programos variantą (apie jį kiek plačiau žr. Norkevičius, Raškinis ir Kazlauskienė, 2005, pp.235-238). Kad būtų galima analizuoti ir aprašyti skiemens modelius, reikia tiksliai nustatyti skiemenų ribas. Tam ir buvo VDU mokslininkų kuriama, tobulinama ir šiame straipsnyje plačiau aprašoma skiemenavimo programa.

Pirmąją lietuviško teksto skiemenavimo kompiuterinę programą SKIE-MUO sukūrė J. Skendelis dar XX a. 10-ojo dešimtmečio pradžioje. Programos SKIE-MUO veikimo principai nėra aprašyti mokslinėje spaudoje, tačiau žinoma, kad ji naudojo žodyną, kuriame nurodomas problemiškesnių žodžių taisyklingas skiemenavimas. Programą SKIE-MUO vėliau komerciškai platino UAB „Fotonija“. Ji buvo pritaikyta Windows terpei ir joje veikiančioms leidybos programoms.

1. Skiemenavimo principai

1.1. Skiemens centras ir jo nustatymas

Lietuvių bendrinėje kalboje skiemens pagrindas yra balsis, dvibalsis arba mišrusis dvigarsis, pvz.: *me-dus*, *sie-na*, *kalnai*. Prie skiemens centro iš vienos ir kitos pusės gali šlietis priebalsių. Jų buvimas ar nebuvimas neturi įtakos skiemeniui susidaryti, nes jį gali sudaryti ir vieni skiemens sudaromieji garsai (skiemens centras), pvz.: *ė-mė*, *uo-das*, *ar-ti*. Taigi priebalsiai yra skiemens periferija. Prieš skiemens centrą esantys priebalsiai sudaro eksplozinę priebalsių grupę, po jo esantys – implozinę grupę.

Kadangi kiekvieną skiemenį sudaro bent vienas balsis ar dvibalsis (jis yra skiemens centras, būtinoji skiemens dalis), tad suskaičiavus juos galima pasakyti, kiek skiemenų yra žodyje. Tačiau atkreiptinas dėmesys į tai, kad įprastinė rašyba netiksliai perteikia tarimą, dėl to balsė (raidė) ne visada žymi balsį (garsą). Pirmiausia minėtinas vadinamasis minkštumo ženklas, žymimas raide *i*, pvz.: žodžiuose *gėrio*, *keliu*, *važiavo* balsių-raidių ir balsių-garsų kiekis nesutampa. Iš pirmo žvilgsnio tai nesunkiai išsprendžiama problema: reikia tik garsų junginius *iu*, *ių*, *iū*, *io*, *ia*, *iq* laikyti vienu neskaidomu segmentu (kaip ir dvibalsius *ie*, *uo*, *ai*, *au*, *ei*, *ui*). Tačiau kalboje vartojama nemažai tarptautinių žodžių, kuriuose panašios samplaikos (*ia*, *iu*, *io*) žymi du balsius (vadinamasis hiatas), tarp jų tariant įterpiamas *j*, pvz.: *akvariumas*, *biologija*, *cianas*, *čempionas*. Šiuo atveju paryškintieji balsiai esti dviejų skiemenų centrai.

Kita vertus, balsė *i* ne visada žymi ir dvibalsio dėmenį, nors junginys grafiškai labai panašus į dvibalsį, pvz.: *abiturientas*, *audiencija*, *higiiena*, *orientacija*, *induiizmas*, *intuicija*, *jėzuitai*. Šiuose žodžiuose yra dviejų balsių samplaika (hiatas). Tad šie balsiai priklauso skirtingiems skiemenims.

Lietuviškuose žodžiuose tarp dviejų morfemų taip pat gali atsirasti dvibalsius ar balsio ir minkštumo ženklelio junginius (kalbininkų vadinamųjų papriešakėjusių balsių) primenančių balsinių samplaikų, pvz.: *apsiašaroti, įsiamžinti, neilgas, paimti, paupys, suiro, suošė*. Tokių dviejų balsinių samplaikų pasitaiko ir tarptautinių žodžių priešdėlio ar sangražos dalelytės ir šaknies sandūroje, pvz.: *neinicijuoti, pailiuoti, persiorientuoti, suintriguoti*. Tai nėra dvibalsiai ar papriešakėję balsiai, tad jie yra atskirų skiemenų centrai.

Dėl paminėtų atvejų negalima tik akiai suskaičiuoti balsių, dvibalsių ir pasakyti skiemenų kiekio. Reikia labai kruopščiai išrinkti tuos atvejus, kada dviejų balsių (raidžių) samplaika žymi:

- neskaidytiną dvibalsį (*ai, ei, ui, au, ie, uo, eu, oi, ou*),
- papriešakėjusį balsį (minkštumo ženklelis ir balsis, pvz., *iq, ia, iai, iau, iu, iui, iuo, io, iū, ių*),
- o kada, atvirkščiai, du gretimi balsiai priklauso skirtingiems skiemenims.

1.2. Skiemens ribų nustatymas

Kaip jau buvo minėta, kalbai svarbesnis yra skiemenų kiekis, o ne ribos tarp jų. Kuriant skiemenavimo įrankį automatinio kirčiavimo reikmėms iš principo nebūtina nustatyti tikslų ribų tarp skiemenų – svarbus tik pats skiemens centras. Vadinas, užtenka jau pateiktų skiemens centro ir skiemenų skaičiaus nustatymo taisyklių.

Tačiau tekstui redaguoti (t. y. žodžių kėlimui į kitą eilutę) nebeužtenka tik skiemenų kiekio. Čia būtina pateikti taisykles, pagal kurias raidžių samplaikos gali būti skaidomos arba jokiū būdu neskaidytinos. Tokiu atveju minkštumo ženklelis ir balsis, dvibalsiai, mišrieji dvigarsiai, dviraidžiai *ch, dž, dz* bus neskaidytini vienetai. Dabartinės žodžių kėlimo taisyklės gana liberalios: priebalsių samplaikos gali būti skaidomos įvairiai, būtina tik į kitą eilutę perkelti paskutinę priebalsę, pvz.: žodį *anksti* galima kelti *anksti, ank-sti, anks-ti*. Priešdėliniai ir sudurtiniai žodžiai gali būti skaidomi arba skiemenimis, arba pagal sudedamąsias dalis, pvz., *ant-akis* ir *an-takis*. Į kitą eilutę negalima perkelti ar palikti vienos priebalsės (žodis keliamas skiemenimis, tad atskirtojoje dalyje turi būti skiemens centras). Paprastai rekomenduojama keliamosios dalies nepradėti sunkiai perskaitomu junginiu (nederėtų kelti *an-ksti*). Manoma, kad tekstas lengviau ir greičiau skaitomas, kai ankstesnėje eilutėje paliekama daugiau priebalsių.

Norint sukurti skiemenavimo įrankį, kurį būtų galima naudoti ir taikomiesiems tikslams, ir fundamentiniams tyrimams, reikia ieškoti objektyvių skiemens ribų nustatymo taisyklių. Kadangi iki šiol nenustatyta ryškių artikuliacinių ir akustinių skiemens ribų požymių, buvo pasirinkta funkcinė arba fonologinė skiemens ribų nustatymo teorija. Pagal šią teoriją naujas skiemo pradamas maksimaliu priebalsių junginiu, kurio modelis galimas žodžio pradžioje.

Lietuvių kalbos žodžiai gali prasidėti balsiu, dvibalsiu (pvz., *audra, qžuolas, ėjo, ola, uosis*); vienu priebalsiu ar dviejų, trijų priebalsių junginiu (pvz.: *neša, ratas, kalnas, stalas, skylė, štai, slogus, švelnus, žvėris, skraidyti, sprogti, strėlė*). Jeigu žodis pradamas vienu balsiu ar priebalsiu,

nėra jokių jų išsidėstymo apribojimų (t. y. gali būti vartojami visi balsiai ar priebalsiai). Jeigu žodis pradamas dviejų ar trijų priebalsių junginiu, priebalsiai čia išsidėsto ne bet kaip, o pagal tam tikrus dėsniumus. Trinariame junginyje:

- pirmasis visada esti pučiamasis priebalsis: *s, z, š, ž, f, h, ch*, žymimas „S“ simboliu (dviraidžiu *ch* lietuvių kalboje žymimas vienas duslusis pučiamasis priebalsis [x], pasitaikantis dažniausiai tarptautiniuose žodžiuose, pvz., *chemija, choras*),
- antrasis – sprogstamasis priebalsis: *p, b, t, d, k, g*, žymimas „T“ simboliu; šiam pogrupiui priskiriamos ir afrikatos *c, č, dz, dž* (apie jų interpretaciją plačiau žr. Girdenis, 2003, pp.94-95). Tradiciškai lietuvių kalboje vienos afrikatos žymimos viena raide, kitos dviem. Į tai būtina atkreipti dėmesį rengiant automatinio skiemenavimo įrankį.
- trečiasis – sklandusis priebalsis: *l, m, n, r, v, j*, žymimas „R“ simboliu.

Vadinas, žodžio pradžioje gali būti STR tipo trinaris priebalsių junginys, pvz., *skriauda, strigti*.

Dvinariame junginyje pirmasis priebalsis gali būti tik pučiamasis arba sprogstamasis, o antrasis – sprogstamasis arba sklandusis. Taigi šie junginiai gali būti tokio tipo:

- ST, pvz., *stiklas, spausti*;
- SR, pvz., *slinkti, žvilgsnis*;
- TR, pvz., *kratyti, protas*.

Jeigu žodžio pradžioje gali būti STR, ST, SR, TR tipo priebalsių junginiai, tai tik tokio tipo junginiais gali prasidėti ir naujas skiemo. Tad žodžius *aistra, aštrus, sėkla, mįslė, akti* reikia skienuoti *ai-stra, a-štrus, sė-kla, mį-slė, ak-ti*. Jeigu šiuos žodžius skaidytume kitaip, pažeistume žodžių pradžios priebalsių jungimosi dėsnius. Tokiu būdu nustatytas skiemo yra vadinamas fonologiniu skiemeniu.

Atkreiptinas dėmesys į tai, kad žodžio su konkrečiais pradžios priebalsių junginiais galbūt nėra lietuvių kalboje, tačiau čia taikytinas apibendrintas modelis, pvz.: fonologiniais skiemenimis skaidome *iš-vai-zda, liū-dnas*, nors nėra žodžių, kurie prasidėtų *zd, dn*. Tačiau yra nemažai žodžių, kurie prasideda kitokiomis ST ir TR tipo priebalsių kombinacijomis, pvz.: *skirti, spinta, stogas* ir *greitis, priekis, traukti*.

Tarp priešdėlio ir šaknies, dviejų kamienų neretai būna vadinamoji atviroji vidinė sandūra. Šių žodžių dalių sandūra semantizuota ir garsai čia gali nepaklusti bendriesiems fonetiniams žodžio vidaus procesams (apie juos plačiau žr. Pakerys, 2003, pp.171-194). Todėl skiemenų ribas reikia tapatinti su morfemų ribomis, pvz.: *vais-krū-mis, at-rem-ti*. Kai šaknis prasideda balsiu, ankstesnės morfemos galinis priebalsis patenka į tolesnį skiemenį, pvz.: *an-takis, pe-lė-da*. Atkreiptinas dėmesys ir į kitas raidžių samplaikas tarp morfemų, pvz.: afrikatomis nelaikomos tokio tipo samplaikos: *juod-žemis*.

2. Skiemenavimo algoritmo resursai

Remiantis aptartu struktūriniu fonologinio skiemens modeliu (STRVRTS), galima automatiškai suskienuoti daugelį lietuvių kalbos žodžių. Skienuojant laikomasi nuostatos, kad visos balsių samplaikos priklauso tam pačiam skiemeniui, jei tik jos teoriškai gali sudaryti dvibalsį ar (ir) minkštumo ženklą *i* ir balsio samplaiką. Tačiau yra probleminių atvejų (apie juos jau kalbėta ir ankstesniame skyriuje), kai formaliai ta pati garsų (raidžių) samplaika gali priklausyti vienam arba dviem skiemenims (pvz., *pa-u-pys* ir *lau-kas*) ar grafema *i*, kuri vienuose žodžiuose žymi balsį, kituose yra vadinamasis minkštumo ženklelis (pvz., *bi-o-lo-gi-ja* ir *ne-šio-jo*), turi būti atskirai nurodomi. Visa tai pateikta resursų faile, toliau vadinamame skiemenavimo šablonų žodynu, kuriame eilutė aprašo vieną atvejį (šabloną) ir galimus jo požymius. Skiemenavimo šablonų žodynu naudojasi ne tik skiemenavimo algoritmas, bet ir priešdėlių nustatymo bei transkripcijos algoritmai.

Šablonai yra dvejopi: vidiniai ir pradiniai. Vidinis šablonas pradedamas ir baigiamas simboliu „*“. Jis gali atitikti bet kurią žodžio atkarpą: pradžią, vidurį ar pabaigą. Pavyzdžiui, šablonas **bi-ot** nurodo, kad tokie žodžiai, kaip

biotoksinai, *antibiotikai*, turintys raidžių seką *biot*, būtų skienuojami įterpiant skiemens skirtuką šablone pažymėtoje vietoje.

Pradinis šablonas pradedamas raide ir baigiamas trejopai: a) raide, b) simboliu „*“, c) simboliu „#“. Šis šablonas taikomas tik tada, kai atitinka nagrinėjamo žodžio pradžią. Jei pradinis šablonas baigiasi raide, jis turi tiksliai sutapti su visu nagrinėjamu žodžiu. Jei pradinis šablonas baigiasi simboliu „*“, lyginame žodyje **turi būti** viena ar daugiau raidžių po šabloną atitinkančios žodžio dalies. Simbolis „#“ apima abu aptartus atvejus: po šabloną atitinkančios žodžio dalies **gali būti** raidžių, bet gali jų ir nebūti.

Šablonų taikymo tvarka yra hierarchinė. Jei nagrinėjamam žodžiui tinka keli šablonai, pasirenkamas svarbesnis (daugeliu atveju ilgesnis) šablonas. Pavyzdžiui, žodžiui *gnai-bioti* gali būti taikomi du šablonai **bi-ot** ir **gnaibiot**, tačiau pasirenkamas tik antrasis svarbesnis (ilgesnis) šablonas ir todėl jame esanti balsių (raidžių) pora *io* nebus išskaidyta į du skiemenis. Jei žodį ar jo dalį atitinkantys šablonai yra vienodo ilgio, svarbesnis yra tas šablonas, kurio požymių sąrašė reikalaujama tam tikros lemos (pagrindinės formos).

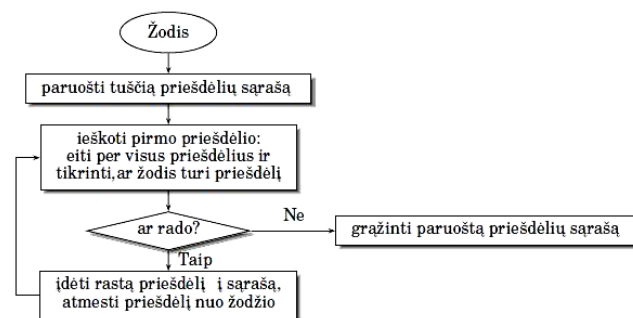
1 lentelė. Skiemenavimo šablonų žodyno įrašų pavyzdžiai

Šablonas	Pavyzdžiai	Taikymo rezultatas	Pastabos
s-dešimt	tris dešimt keturi sdešimt	tris-dešimt keturias-dešimt	
bibli-o*	bibli oteka bibli ografija	bibli-oteka bibli-ografija	
bi-o-ind*	bio indikacija	bi-o-indikacija	
bi-ot *gnaibiot*	ant ibi otikai bio toksinai nega ibio ti	antibi-otikai bi-otoksinai negaibioti	
kair* ka-ir* <kairas>	kair ės Kair o	kairės Ka-iro	Požymis <kairas> rodo, kad antrąjį šabloną atitinka tik tie žodžiai, kuriems lemuoklis nustato pagrindinę formą „kairas“.
o-inf <nj> *o-ist*	mikro in farktas ego is tas	mikro-infarktas [m'ikr:inf□.rktas] ego-istas [egɔj□stas]	Požymis <nj> rodo, kad balsio „i“ ir gretimo balsio sandūroje transkribuojant neįterpiamas priebalsis j.

3. Priešdėlių nustatymo algoritmas

Vienas iš sudėtingesnių skiemenavimo uždavinių – nustatyti skienuojamo žodžio priešdėlius ir pseudopriešdėlius. Priešdėliais šiame darbe vadinamos visos morfemos, galinčios eiti prieš žodžio šaknį: *bè-*, *tè-*, *nè-*, *àp-*, *ap□-*, *at□-*, *àt-*, *□š-*, *□-*, *nù-*, *pà-*, *pa□-*, *pér-*, *prà-*, *pr□-*, *pr□e-*, *sù-*, *ùž-*, *pó-*, *pró-*, *apý-*, *-s□-*, taip pat kai kurie šių morfemų junginiai (*nebè-*, *tebè-*). Pseudopriešdėliais vadinami tarptautiniai prefiksai ir tarptautinių (ir kai kurių lietuviškų) sudurtinių žodžių pirmieji dėmenys, pvz.: *anti-*, *avia-*, *hidro-*, *pop-*, *kilo-*, *tarp-*, *virš-* ir kiti. Žodžio priešdėlius ir pseudopriešdėlius skiemenavimo algoritmui žinoti yra svarbu dėl dviejų priežasčių. Visų pirma, daug kur skiemenų ribas reikia tapatinti su morfemų ribomis. Antra, skiemenavimo šablonų žodynas apima tik nepriešdėlinius žodžius. Tokių žodžių, kaip *išsiaiškinti*, *nebesiaiškinti*, *pasiaiškinimas*, taisyklingas skiemenavimas priklauso nuo tinkamo priešdėlių nustatymo, nes skiemenavimo šablonų žodyne nėra šablonų su visomis

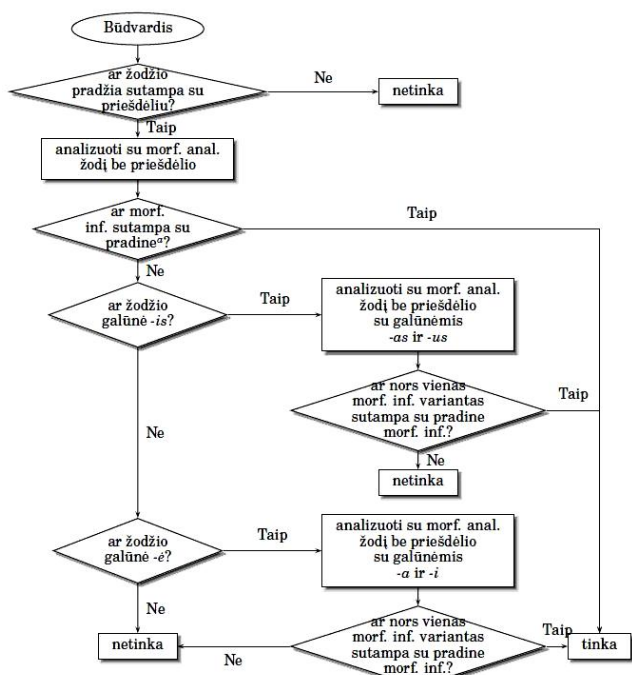
teoriškai įmanomomis priešdėlių ir sangrąžos dalelytės -si- kombinacijomis.



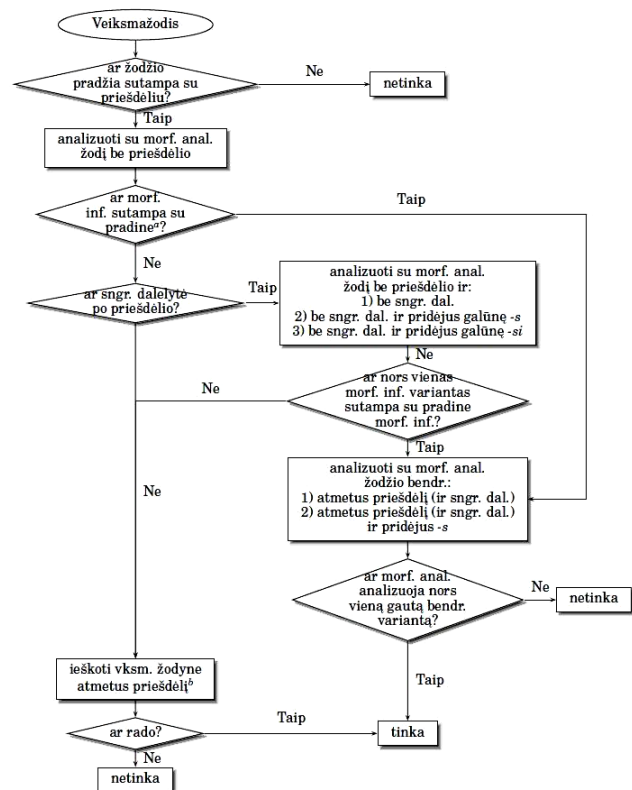
1 pav. Bendra priešdėlių nustatymo algoritmo schema

Pripažinti ar nepripažinti žodžio pradžią priešdėliu ar pseudopriešdėliu nusprendžiama pagal žodžio (ir žodžio, gauto atmetus analizuojamą dalį) morfologinės analizės rezultatus. Šiuos rezultatus priešdėlių nustatymo algoritmui pateikia morfologinis analizatorius *Lemuoklis* (plačiau apie *Lemuoklį* žr. Zinkevičius, 2000, pp.245-273). Priešdėlių ir

pseudopriešdėlių nustatymo algoritmai šiek tiek skiriasi. Būdvardžių ir veiksmažodžių priešdėlių nustatymo procedūros pateiktos 1–3 paveiksluose. Procedūros grąžina priešdėlių sąrašą, nes žodis gali turėti kelis priešdėlius.



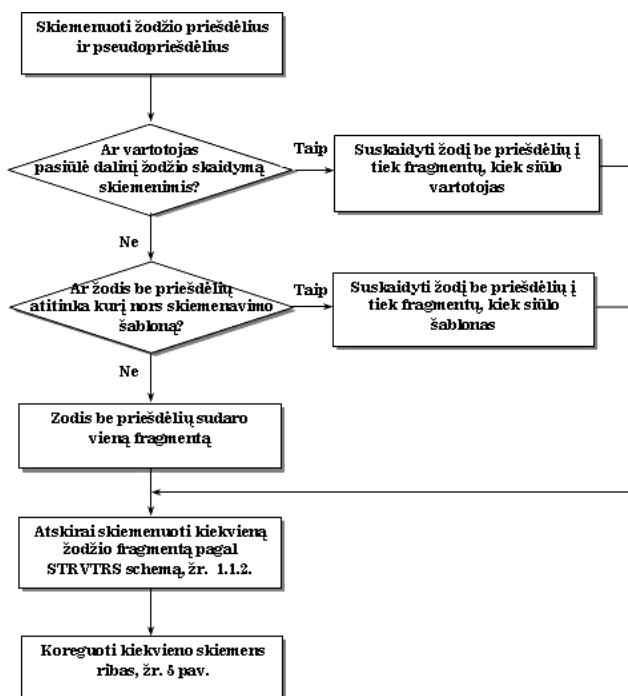
2 pav. Būdvardžio priešdėlių nustatymo schema^a Morfologinės informacijos tapatumas tikrinamas lyginant tokius parametrus: kalbos dalis, linksnis, skaičius ir t. t. Jei bent viename iš lyginamų žodžių kuris nors parametras nėra nurodytas, tas parametras nenaudojamas.



3 pav. Veiksmažodžio priešdėlių nustatymo schema^a Lyginama, ar parametrai (kalbos dalis, linksnis, skaičius ir kt.) sutampa. Jei bent viename iš lyginamų žodžių kuris nors parametras nenuro-

dytas, tas parametras nenaudojamas; ^b Pagal visas tris pagrindines veiksmažodžio formas. Tokiu būdu galima rasti priešdėlius ir tais atvejais, kai morfologinis analizatorius neatpažįsta žodžio be priešdėlio, pvz.: *supranta*.

Supaprastinta veiksmažodžio priešdėlių nustatymo schema taikoma ir daiktavardžiams. Čia priešdėliui nustatyti reikalinga tik pradinio žodžio ir žodžio, gauto atmetus prefiksus, morfologinės informacijos tapatumo. Tokiu būdu pavyksta nustatyti veiksmažodinių daiktavardžių, pvz.: *pasi-aiškinimas*, *iš-si-auklėjimas*, priešdėlius.



4 pav. Žodžio skienuojimo algoritmo schema

Lemuoklis dažnai negali atpažinti tarptautinės kilmės žodžių, turinčių priešdėlių ir ypač pseudopriešdėlių, pvz.: *poabortinis*, *antialkoholinis*, *aviaindustrija*, *hidroizoliacija*. Tokiu atveju morfologinės informacijos tapatumo principas, kuriuo grindžiami anksčiau aprašyti algoritmai, pseudopriešdėliams nustatyti negali būti taikomas. Dėl šios priežasties pseudopriešdėlių nustatymo algoritmui yra suteikta daugiau laisvės atskirti žodžio prefiksus nei priešdėlių nustatymo algoritmams. Pseudopriešdėlių nustatymo algoritmui leidžiama atskirti žodžio prefiksus pagal pateiktą sąrašą tol, kol *Lemuoklis* jau pradeda žodį be prefiksų atpažinti. Ši prefiksų atskyrimo schema kai kada „nustato“ priešdėlius, ten kur jų iš tiesų nėra (pvz.: *Popovas*, *antilopė*), bet tokie netikslumai išsprendžiami: papildomas skienuojimo šablonų žodynas (pvz.: šablonais *popov**, *antilop**).

Skienuojimo (ir transkribavimo) algoritmui svarbus šaltinis priešdėlių nustatymo algoritmo rezultatas – žodžio, gauto atskyrus priešdėlius, pagrindinė forma, kurią įvardija *Lemuoklis*. Pagrindinė forma gali būti nurodyta kai kurių skienuojimo šablonų požymių sąraše ir naudojama patikslinti šablono taikymo atvejams (t. y. skienuojamo žodžio ir šablone nurodytos pagrindinės formos turi sutapti). Tuo atveju, kai *Lemuoklis* nustato keletą galimų atsakymų, pasirenkamas vienas iš jų prioriteto mažėjimo

tvarka: veiksmazodis, būdvardis, daiktavardis ir priešveiksmai. Pavyzdžiui, žodžiui *šoko* bus pasiūlyta pagrindinė forma *šokti*, o ne *šokas*.

4. Skiemonavimo algoritmas

Remiantis skiemonavimo algoritmu skiemenuojami pavieniai žodžiai, į gretimus žodžius neatsižvelgiama. Tarp klitiko (kirčio neturinčio žodžio) ir savarankiško žodžio, jei šlijimas buvo jau pažymėtas skiemenuojamajame tekste, įterpiamas skiemens ribos ženklas, abu žodžiai skiemenuojami atskirai.

Skieonavimo algoritmo pradiniai duomenys yra:

- žodis, kurį reikia skiemenuoti;
- priešdėlių, pseudopriešdėlių ir šaknies riba žodyje;
- nepriešdėlinio žodžio pagrindinė forma.

Papildomai vartotojas gali pats pažymėti kai kurių skiemėnų ribas ar pateikti informaciją apie skiemenuojamo žodžio kirčio tipą ir vietą. Radęs vartotojo įterptus skiemens skirtukus, skieonavimo algoritmas remiasi vartotojo žymomis, o ne skieonavimo šablonų žodyno nuorodomis. Tai pravartu skiemenuojant ilgus tekstus ir siekiant

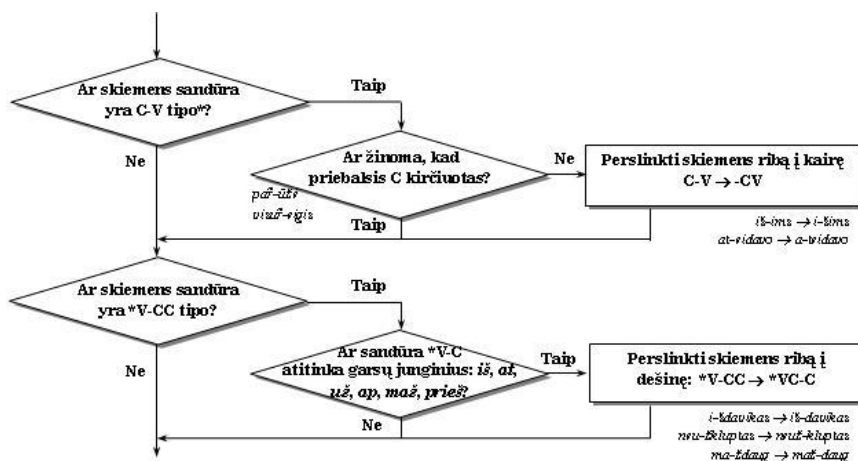
sumažinti skiemenuoklio daromas tipines klaidas (dažniausiai tai tarptautiniai žodžiai ar tikriniai daiktavardžiai specializuotuose tekstuose).

Bendra skieonavimo algoritmo schema pateikta 4 paveiksle.

Skieonavimas yra tik vienas iš automatinio kirčiavimo etapų, tad informacija apie žodžio kirčio tipą ir vietą skiemenuojant paprastai dar nėra žinoma. Dažnai ši informacija nėra labai svarbi, tačiau kai kada gali patikslinti skiemėnų ribas. Pavyzdžiui, žodis *parūžė* (kai kirtis nežinomas) būtų skiemenuojamas *pa-rūžė*, o tas pats sukirčiuotas žodis *paūžė* būtų skiemenuojamas *pa-ūžė*, nes išlaikomi abu kirčiuoto mišriojo dvigarsio dėmenys viename skiemenyje.

Galimybė skiemenuoti jau sukirčiuotą tekstą yra svarbi, nes aprašomas skieonavimo įrankis naudojamas ne tik kirčiavimo programoje, bet ir kituose VDU mokslininkų darbuose, pavyzdžiui, fonotaktikos ar ritmikos tyrimuose (vienas iš jų: Kazlauskienė ir Raškinis, 2008, pp.24-31).

Skieomėnų ribų korekcijos algoritmas pateiktas 5 paveiksle. Paveiksle minima C-V skieomėnų sandūra galima tik tarp priešdėlių arba tarp priešdėlio ir šaknies.



5 pav. Skieomėnų ribų korekcijos schema (paveiksle simboliu „V“ žymimas balsis, simboliu „C“ – bet koks priešbalsis)

Išvados

Skieonavimo algoritmas klaidų nedaro. Jų gali atsirasti tik dėl labai retų homografų, kurie rašomi vienodai, bet skiemenuojami skirtingai, pvz. *Di-e-gas* ir *die-gas*, *Tri-e-stas* ir *trie-stas*, *Ka-i-re* ir *kai-re* ir kt. Dalį homografų būtų galima vienareikšminti, jei skieonavimo algoritmas skirtingai traktuotų didžiąsias ir mažąsias raides arba gautų patikimą informaciją iš morfologinio analizatoriaus, ar skiemenuojamas žodis yra bendrinis (tikrinis) daiktavardis.

Aprašytas skieonavimo algoritmas naudojamas VDU sukurtoje automatinio transkribavimo ir kirčiavimo programose, kuria paremtas ir visiems prieinamos kirčiuoklės (<http://donelaitis.vdu.lt/main.php?id=4&nr=9>) veikimas. Šis algoritmas naudojamas ir fonotaktikos bei ritmikos tyrimuose.

Literatūra

1. Girdenis, A., 2003. *Teoriniai fonologijos pagrindai*. Vilnius: Petro ofsetas.
2. Kazlauskienė, A., Raškinis, G., 2008. Lietuvių kalbos fonologinio skieomėnų struktūrinių modelių dažnumas. *Žmogus ir žodis*, t. 10 (1), pp.24-31.
3. Norkevičius, G., Raškinis, G., Kazlauskienė, A., 2005. Knowledge-Based Grapheme-to-Phoneme Conversion of Lithuanian Words. *SPECOM 2005, 10th International Conference Speech and Computer*, pp.235-238
4. Pakerys, A., 2003. *Lietuvių bendrinės kalbos fonetika*. Vilnius: Enciklopedija.
5. Zinkevičius, V., 2000. *Lemuoklis – morfologinei analizei. Darbai ir dienos*, t. 24, pp.245-273.

Automatic Detection of Syllable Boundaries: Problems and Their Solutions

Summary

It is difficult to perceive syllable boundaries in a naturally uttered word. The number of syllables and not their exact boundaries is the primary factor affecting speech perception. However the detection of exact syllable boundaries is important to carry out fundamental investigations about the syllable structure. An automatic syllabification tool has been developed for this purpose. Syllable boundaries were defined according to the functional theory of a syllable: the syllable should begin with the maximum cluster of consonants given that such cluster (its model) is found at the beginning of other words. The paper describes some problematic cases where phonological model of a syllable cannot be directly applied. The solutions are proposed for disambiguating the juncture of a pair of consonants (diphthong, hiatus, frontalisied vowel), accurately detecting word prefixes, for syllabifying composite (two-root) words. The syllabification algorithm is quite accurate. A few errors occur because of homographs that are spelled identically but pronounced and syllabified differently. The algorithm described in this paper is being used as a component to other language processing tools created at VMU: within algorithms for automatic stress assignment and automatic phonetic transcription, and within computer tools for investigating word phonotactics and text rhythmic.

Straipsnis įteiktas 2009 11
Parengtas spaudai 2009 11

Apie autorius

Gailius Raškinis, doc. dr. (fiziniai mokslai), Vytauto Didžiojo universiteto Kompiuterinės lingvistikos centro vyresnysis mokslo darbuotojas, Sistemų analizės katedros docentas.

Mokslinės veiklos sritys: mašininio mokymo metodų kūrimas bei taikymas kalbos technologijų uždaviniams spręsti.

Adresas: Vytauto Didžiojo universitetas, K. Donelaičio g. 52, 44244 Kaunas

El. paštas: g.raskinis@if.vdu.lt

Asta Kazlauskienė, doc. dr. (humanitariniai mokslai), Vytauto Didžiojo universiteto Lietuvių kalbos katedros profesorė ir Kompiuterinės lingvistikos centro vyresnioji mokslo darbuotoja, Kauno technologijos universiteto Kalbotyros katedros profesorė.

Mokslinės veiklos sritys: fonetika, akcentologija, taikomoji kalbotyra.

Adresas: Vytauto Didžiojo universitetas, K. Donelaičio g. 52, 44244 Kaunas

El. paštas: a.kazlauskiene@hmf.vdu.lt

