

# AUTOMATIZUOTAS LIETUVIŲ KALBOS VEIKSMAŽODŽIŲ KIRČIAVIMAS: PROBLEMOS IR JŲ SPRENDIMAS

**Asta Kazlauskienė, Giedrius Norkevičius, Gailius Raškinis**

*Vytauto Didžiojo universitetas, K. Donelaičio g. 58, 3000 Kaunas*

## 1. Įvadas

Lietuvių kalbos tyrimams skirtuose skaitmeninių technologijų darbuose jau senokai pasigendama automatizuoto kirčiavimo, kuris labai aktualus ne tik lietuvių šnekos sintezei, atpažinimui, automatiniam teksto transkribavimui, teksto prozodijos tyrimams, bet ir kompiuteriniam kirčiavimo mokymui ir mokymuisi. Tai nėra atsitiktinis dalykas: daugelis kitų kalbų turi fiksuotą arba nefiksuotą pastovų kirtį, todėl šiose kalbose automatinis kirčiavimas nesukelia ypatingų problemų ir dėl to šioms kalboms nėra poreikio kurti kompiuterines programas. Lietuviai šioje srityje atsilieka dėl mūsų kalbos kirčiavimo specifikos ir dėl sudėtingos gramatinės sistemos<sup>1</sup>.

Pirmasis lietuvių kalbos automatizuoto kirčiavimo problemą bandė spręsti P. Kasparaitis (2001). Jis gana plačiai išnagrinėjo daiktavardžių ir būdvardžių kirčiavimą, veiksmažodžiams skyrė šiek tiek mažiau dėmesio. Jo sudarytos kirčiavimo taisyklės paremtos žodžio sandara, dėl to jas sunkoka suvokti, modifikuoti bei patikrinti, kaip jos atitinka tradicines kirčiavimo taisykles. Tačiau tai jokių būdu nėra darbo trūkumas. Didesnė bėda ta, kad šis algoritmas kol kas naudojamas tik paties P. Kasparaičio sukurtame sintezatoriuje. Juo negali pasinaudoti nei kiti kalbos technologijų srityje dirbantys mokslininkai, nei kalbos vartotojai, norintys pasimokyti taisyklingo kirčiavimo ar patikrinti savo žinias. Tai ir yra pagrindinė priežastis, dėl kurios VDU pradėtos tirti automatizuoto kirčiavimo galimybės<sup>2</sup>.

---

<sup>1</sup> Analizuodami kirčiavimo modelių kūrimo galimybes, pamatėme, kad tie dalykai susiję labiau, nei buvo galima tikėtis.

<sup>2</sup> Šiokia tokia problema yra ir kirčiuoti šriftai, nes viena iš kuriamos programos pritaikymo galimybių yra tekstinio dokumento automatinis kirčiavimas. Daugelio kalbininkų vartoti ankstesni rašmenų komplektai (pvz., „Fontra“) pritaikyti tik senesnėms *Microsoft Word* (pvz., 6 ar 7) sistemoms, kurios dabar kompiuteriuose jau retenybė. Tačiau tikimės, kad kuriamas naujasis šriftų rinkinys „Palemonas“ greitai išspręs šias problemas. Tad kol kas kirčio ženklus žymime kitokiais sutartiniais simboliais.

## 2. Automatizuoto kirčiavimo problemos

Kurdami kirčiavimo algoritmą, susidūrėme su daugybe problemų. Panagrinėkime tokią neveikiamosios rūšies esamojo laiko dalyvių kirčiavimo taisyklę, kuri pateikiama lietuvių kalbos vadovėliuose:

**Neveikiamieji esamojo laiko dalyviai, padaryti iš dviskiemenių o asmenuotės ir daugiaskiemenių veiksmažodžių, turi to paties laiko 3-ojo asmens kirtį ir priegaidę ir kirčiuojami pagal 1-ąją kirčiuotę**

Šis pavyzdys rodo tai, kad norėdami automatiškai sukirčiuoti konkrečią išvestinę veiksmažodžio formą iš anksto turime sugebėti atlikti tris dalykus. Pirmiausia privalome turėti algoritminiam naudojimui tinkamu būdu užrašytas specifines kirčiavimo žinias, šiuo atveju atsakančias į klausimą, kaip kirčiuojamas to paties laiko veiksmažodžio trečiasis asmuo ir ką reiškia kirčiuoti pagal 1-ąją kirčiuotę (pabraukta). Antra, turime mokėti nustatyti žodžio morfologines savybes ir pagrindinę formą (paryškinta). Trečia, privalome mokėti nustatyti reikiamų žodžio formų skiemenų kiekį (kursyvu parašyta). Taigi, analizuojant veiksmažodžių kirčiavimo taisykles ir bandant jas pritaikyti algoritmams, išryškėjo labai svarbus dalykas: automatinio kirčiavimo neįmanoma sukurti be kirčiavimo taisyklių pritaikymo kompiuteriui, reikiamos morfologinės informacijos gavimo ir skiemenavimo.

## 3. Automatizuoto kirčiavimo algoritmas

### 3.1. Specifinių kirčiavimo žinių formalizavimas

Norint sukirčiuoti veiksmažodžius, reikia žinoti, kaip kirčiuojamos trys pagrindinės jų formos, o išvestinės remiasi ta forma, iš kurios padarytos. Tačiau kompiuteris nei iš klausos, nei remdamasis kokiomis nors lengvai aprašomomis garsinėmis skiemens ypatybėmis nesukirčiuos pagrindinių formų<sup>3</sup>. Todėl pirmiausia reikėjo sudaryti pagrindinių formų sąrašą (apie 7500 žodžių). Tolesnė kirčiavimo taisyklių analizė parodė, kad veiksmažodžiai automatinio kirčiavimo požiūriu nėra lygiaverčiai

(būtina žinoti veiksmažodžio tipą: priesaginis, pirminis, mišrus), todėl visas sąrašas išskaidytas į tris dalis<sup>4</sup>.

Pagrindinėse kirčiavimo taisyklėse, kuriomis remiantis gali būti kuriami algoritmai, nurodyta informacija yra sunkiai pateikiama kompiuteriui. Be to, iš tokios taisyklių ir veiksmažodžio formų gausos sunku nustatyti prioritetus: kurios taisyklės taikytinos pirmiausia, kurios formos kirčiuotinos pagal vienodą modelį. Todėl taisykles reikia nemažai pertvarkyti, skaidyti į smulkesnes, jungti į grupes bei formalizuoti jų taikymą, t. y. suteikti taisyklėms medžio struktūrą. Medžio struktūra pasirinkta dėl kelių priežasčių: a) medis yra lengvai sudaroma ir modifikuojama žinių struktūra b) medyje aiškus operacijų eiliškumas, todėl jis lengvai paverčiamas algoritmu.

Atidžiai išnagrinėjus veiksmažodžio formų kirčiavimo dėsnius, visos išvestinės formos suskirstytos į tris grupes pagal kirčiavimo medžio šakų panašumą:

1. Formos, išlaikančios atitinkamos pagrindinės formos kirčio vietą ir priegaidę: būtasis dažninis laikas, tariamoji nuosaka, liepiamoji nuosaka, veikiamosios rūšies būtojo dažninio laiko dalyvis, veikiamosios rūšies būsimajo laiko dalyvis, būtojo dažninio laiko padalyvis, būsimajo laiko padalyvis, būtojo kartinio laiko padalyvis, veikiamosios rūšies būtojo kartinio laiko dalyvis.

2. Formos, kai priesaginiai veiksmažodžiai kirčiuojami pagal sąrašą, kiti – pagal skirtingus algoritmus: pusdalyvis, būdinys, reikiamybės dalyvis, neveikiamosios rūšies būsimajo laiko dalyvis, esamojo laiko padalyvis, neveikiamosios rūšies esamojo laiko dalyvis, neveikiamosios rūšies būtojo laiko dalyvis.

3. Formos, kirčiuojamos pagal individualius algoritmus: būsimasis laikas, veikiamosios rūšies esamojo laiko dalyvis. Šiai grupei priskirtinos ir kai kurios nepriesaginių veiksmažodžių formos: pusdalyvis, būdinys, reikiamybės dalyvis ir neveikiamosios rūšies būsimajo laiko dalyvis, esamojo laiko padalyvis, neveikiamosios rūšies esamojo laiko dalyvis, neveikiamosios rūšies būtojo laiko dalyvis.

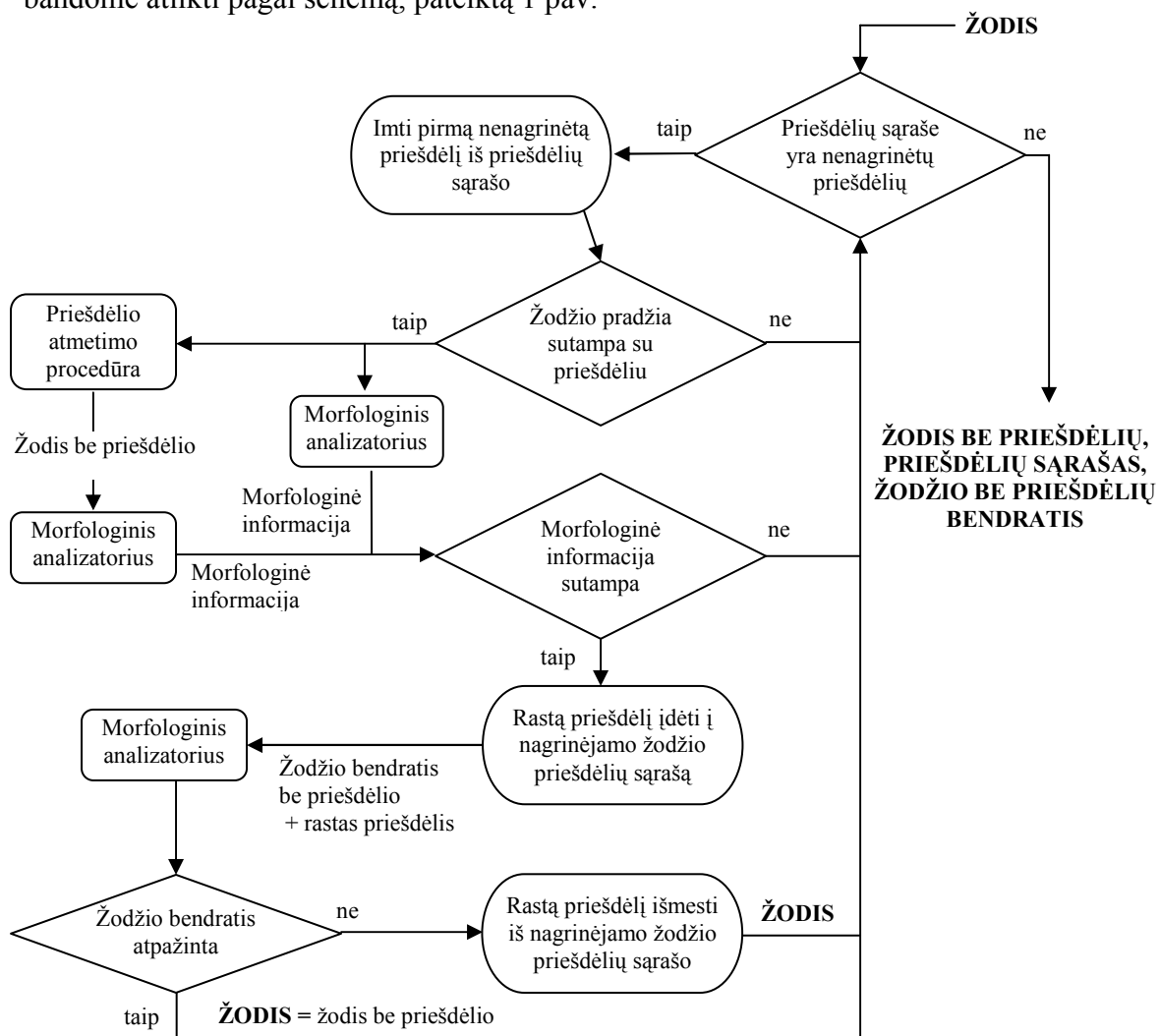
### 3.2. Morfologinė analizė

Norint sudaryti tokius kirčiavimo modelius, kuriuos būtų galima patikrinti realiomis, kalbininkams įprastomis taisyklėmis, būtina morfologinė informacija: giminė,

<sup>3</sup>Bendratis toliau vadinama 1 forma, esamojo laiko trečiasis asmuo – 2 forma, būtojo kartinio laiko trečiasis asmuo – 3 forma.

<sup>4</sup>Priesaginių veiksmažodžių užtenka turėti tik bendratį, nes esamojo ir būtojo laiko kirtį galima nustatyti automatiškai, pirminių ir mišriųjų veiksmažodžių sąrašė yra visos trys formos. Toliau A – priesaginių veiksmažodžių sąrašas, B – mišriųjų veiksmažodžių, C – pirminių veiksmažodžių.

skaičius, dalyvio linksnis, asmuo, asmenuotė, laikas, sangražiškumas, dalyvio rūšis, nuosaka. Gramatinė žodžio analizė nebuvo mūsų darbo uždavinys. Tokią morfologinę informaciją galėtų suteikti V. Zinkevičiaus morfologinis analizatorius „Lemuoklis“ (apie jį plačiau žr. Zinkevičius, 2000, 245–274). Tačiau sudarydami konkrečių formų algoritmus, įsitikinome, kad tik morfologinių duomenų apie žodį neužtenka (bent jau tokių, kokius gali pateikti minėtas morfologinis analizatorius). Norint patikrinti kai kurias sąlygas, reikia atskirų algoritmų, pavyzdžiui, reikalinga priešdėlinė žodžių analizė, kurią bandome atlikti pagal schemą, pateiktą 1 pav.



1 pav. Priešdėlinė žodžio analizė

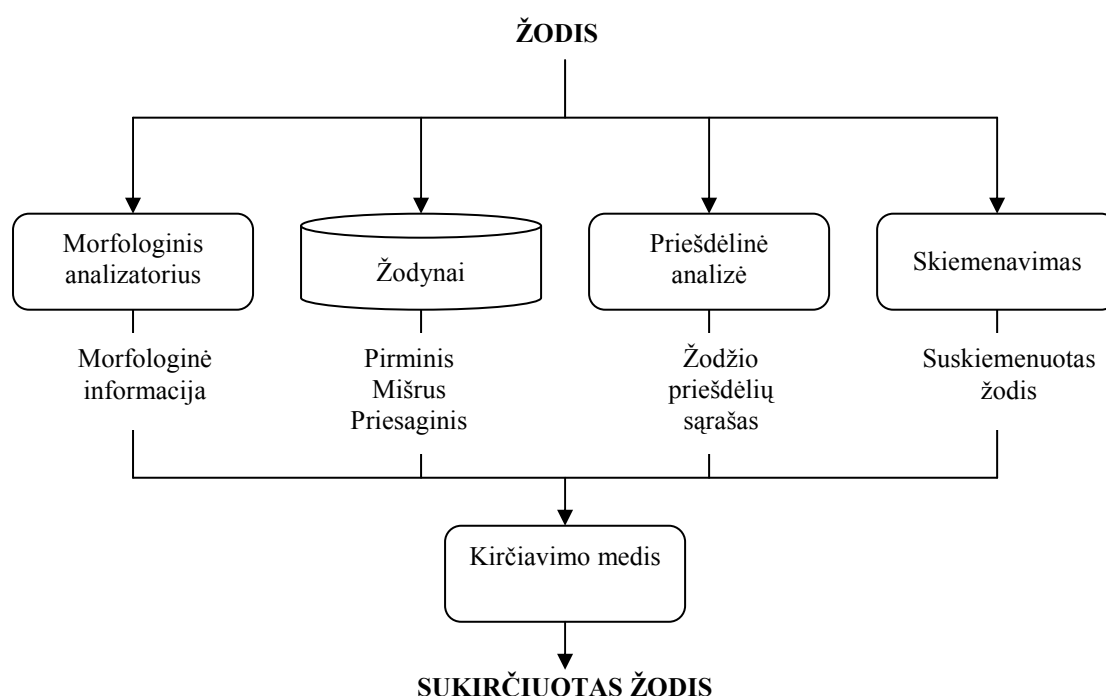
Efektyvi žodžio priešdėlių analizė leidžia sumažinti kirčiavimo bei skiemonavimo algoritmų apimtį ir sudėtingumą. Pateiktas algoritmas naudoja eilutės atitikimo (*string match*) metodą.

### 3.3. Skiemenavimas

Automatinis kirčiavimas neįmanomas ir be kirčiuojamo teksto skiemenavimo. Esantys skiemenuokliai (pvz., „Skiemuo“) nėra tobuli dėl dviejų priežasčių: a) skiemenis, kuriuos sudaro vienas balsis, jie dažnai prišlieja prie tolesnių skiemenų, b) visas dviejų balsių sandūras (ir dvigarsius, ir hiatą) laiko neskaidytiniais junginiais. Ir čia nieko keisto, nes šių skiemenuoklių paskirtis kitokia, todėl jie remiasi ne realiais skiemenimis, o žodžių kėlimo taisyklėmis. Dėl to mes negalime jais naudotis, turime sukurti tikslesnį.

### 3.4. Bendra automatizuoto kirčiavimo schema

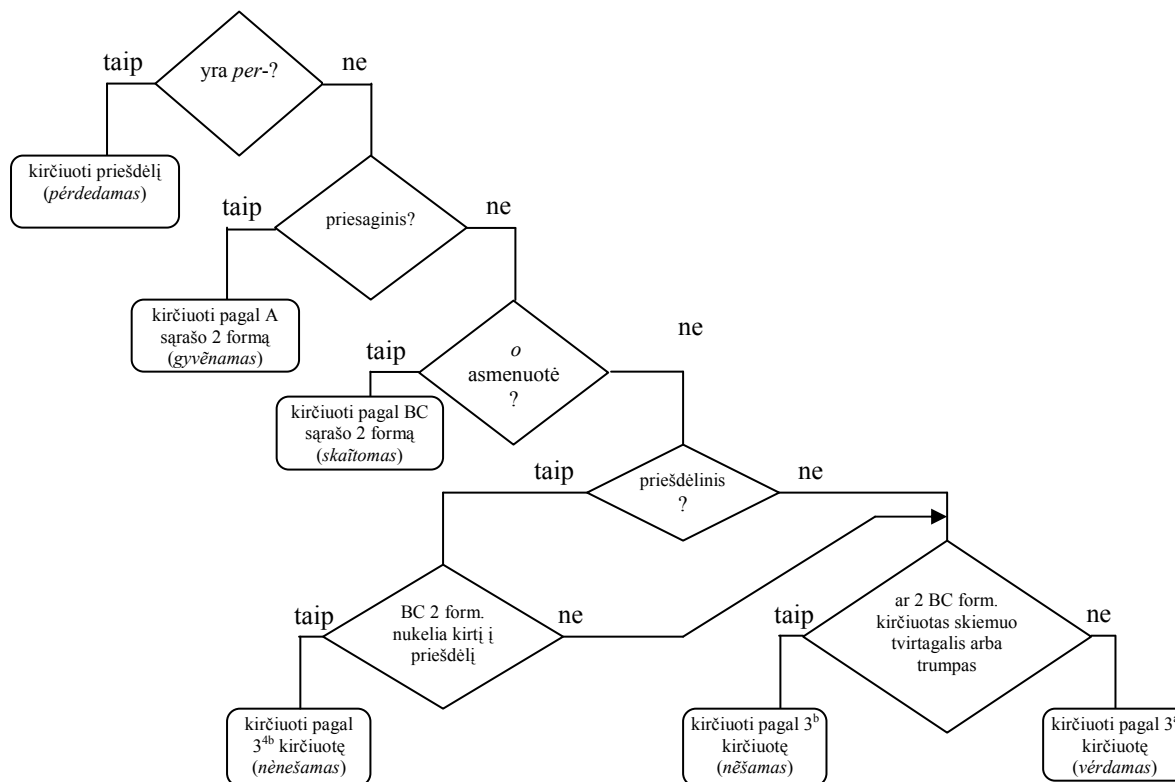
Bendroji veiksmažodžių kirčiavimo schema pavaizduota 2 pav.



2 pav. Bendroji veiksmažodžių kirčiavimo schema

Pagal šią schemą, pirmiausia atliekama morfologinė bei priešdėlinė žodžio analizė, ieškoma atitinkamame žodyne reikiamos atraminės formos, žodis suskiemenuojamas. Toliau einama į kirčiavimo medį ir ieškoma konkrečią morfologinę veiksmažodžio formą atitinkančios kirčiavimo šakos. Kaip pavyzdį pateikiame neveikiamosios rūšies esamojo laiko dalyvių kirčiavimo algoritmą medžio šakos pavidalu (žr. 3 pav.), kuris buvo sudarytas remiantis minėtais kalbos resursais bei vadovėliuose

aprašytos kirčiavimo taisyklės analize. Tarkim, reikia sukirčiuoti tokius dalyvius: *perdedamas, gyvenamas, skaitomas, nešamas, verdamas*.



3 pav. Neveikiamosios rūšies esamojo laiko dalyvių kirčiavimo šaka

#### 4. Automatizuoto kirčiavimo tikslumas

Algoritmas buvo testuotas naudojantis VDU tekstynu (jo sandarą žr. <http://donelaitis.vdu.lt>). Testavimo imtį sudarė beveik 35 tūkstančiai skirtingų dažniausiai vartojamų veiksmažodžių formų. Kirčiuodama pavienius žodžius, programa padarė 7,5 % klaidų. Toks rezultatas gautas, atmetus įvardžiuotines formas ir žodžius, kurių programa nerado mūsų turimuose sąrašuose. Taip yra todėl, kad į kirčiavimo medį įvardžiuotinės formos dar neįtrauktos, o žodyno klaidos atmestos, nes norėta patikrinti, ne bendrą programos daromų klaidų skaičių, o tiesiog algoritmo tinkamumą. Kirčiuodama rišlų tekstą, programa gali daryti ir daugiau klaidų, nes tuo atveju bus neišvengiamas žodžių leksinis ir gramatinis daugiareikšmiškumas, kuris taip pat didelė bėda ir kol kas dar neišspręsta.

## 5. Išvados

Ir algoritmų kūrimas, jų analizė, ir programos testavimas parodė, kad problemišiausias yra ne paties kirčiavimo medžio kūrimas, bet pradinių duomenų apie žodį tikslumas ir patikimumas, o tai ne visada nuo mūsų priklauso. Norėdami, kad programa kirčiuotų kuo tiksliau, turime:

- a) kartu su leksikos ir morfologijos specialistais išspręsti homografų (pvz., *girià* – *gìria*) problemą;
- b) patikslinti sudarytus sąrašus: įrašyti naujų veiksmažodžių, peržiūrėti tuos, kurie sąrašė yra kaip priešdėliniai (pvz., *pajėgti* sąrašė turi būti tik *jėgti*);
- c) sukurti ir patikrinti įvardžiuotinių formų algoritmą.

## Literatūra

*Kasparaitis P.*, 2001, Lietuvių kalbos kompiuterinės sintezė (daktaro disertacija, rankraštis, Vilniaus universitetas), Vilnius.

*Zinkevičius V.*, 2000, Lemuoklis – morfologinei analizei. – Darbai ir dienos, T. 24, 245–274.

## **The automatic accentuation of Lithuanian language verb: problems and their resolution**

### *Summary*

This paper describes automatic Lithuanian language accentuation structural model and the main part of it – tree of accentuation rules. Paper also presents principles of building such structural model, shows the feasibility of such model for creating practical automatic accentuation tool, also shows an example of one branch of accentuation rules tree and describes testing results of presented algorithm.