VILNIAUS UNIVERSITETAS

Tomas Anbinderis

KAI KURIŲ LIETUVIŲ KALBOS TEKSTO KIRČIAVIMO ASPEKTŲ MATEMATINIS MODELIAVIMAS

Daktaro disertacija
Fiziniai mokslai, informatika (09P)

Vilnius, 2010

Disertacija rengta 2005-2009 metais Vilniaus universitete.

Nuo 2005.10.01 iki 2008.04.17: mokslinis vadovas doc. dr. Algirdas Bastys (Vilniaus universitetas, fiziniai mokslai, informatika – 09P).

Mokslinis vadovas:

dr. Pijus Kasparaitis (Vilniaus universitetas, fiziniai mokslai, informatika, 09P).

Reziumė

Disertacijoje nagrinėjama viena iš kalbos sintezatoriaus sudedamųjų dalių – teksto automatinis kirčiavimas, bei su kirčiavimu susiję kiti uždaviniai: vienodai rašomų, bet skirtingai tariamų, žodžių (homografų) vienareikšminimas bei prie gretimo žodžio prišlijusių bekirčių žodžių (klitikų) paieška.

Teksto kirčiavimui pritaikytas metodas, kuris naudodamas sprendimų medžius randa raidžių sekas, vienareikšmiai nusakančias žodžio kirčiavimą. Sprendimo medžiams sudaryti buvo naudojamas didelies apimties sukirčiuotų žodžių tekstynas. Buvo sudarytos kirčiavimo taisyklės remiantis raidžių sekomis žodžių pradžioje, pabaigoje ir viduryje. Pasiūlytas kirčiavimo algoritmas pasiekia apie 95,5% tikslumą.

Homografams vienareikšminti pritaikyti iki šiol lietuvių kalbai nenaudoti metodai, pagrįsti leksemų ir morfologinių pažymų vartosenos dažniais, gautais iš vieno milijono žodžių tekstyno. Darbe parodyta, kad morfologinių pažymų dažniai yra svarbesni už leksemų dažnius. Pasiūlyti metodai leido homografus vienareikšminti 85,01% tikslumu.

Klitikų paieškai pasiūlyti metodai, kurie remiasi: 1) samplaikinių formų atpažinimu, 2) statistiniu žodžio kirčiavimo/nekirčiavimo dažniu, 3) kai kuriomis gramatikos taisyklėmis bei 4) gretimų žodžių kirčių pasiskirstymu (ritmika). Paaiškinta, kaip visus metodus sujungti į vieną algoritmą. Pritaikius šį algoritmą testavimo duomenims, klaidų ir visų žodžių santykis buvo 4,1%, o klaidų ir nekirčiuotų žodžių santykis – 18,8%.

Turinys

Įvadas

Tiriamoji problema

Šiame darbe nagrinėjamas lietuvių kalbos teksto automatinis kirčiavimas bei su tuo susiję kiti uždaviniai – homografų vienareikšminimas ir klitikų paieška.

Temos aktualumas

Šnekamoji kalba – tai vienas pagrindinių žmonių tarpusavio bendravimo (t. y. informacijos apsikeitimo) būdų. Atsiradus kompiuteriams, atsirado ir susidomėjimas, kaip realizuoti žmogaus ir kompiuterio bendravimą šnekamąja kalba. Deja, ši užduotis vis dar lieka neišspręsta.

Šnekamosios kalbos apdorojimo sistemos dažnai skirstomos į kalbos sintezės, atpažinimo ir interpretavimo sistemas. Kalbos sintezės sistemos – tai sistemos, kurios automatiškai generuoja žmogaus kalbą iš bet kokios tekstinės įvesties. Kalbos sintezė gali būti naudojama telekomunikacijose, kalboms mokyti, neįgaliems žmonėms ir pan. Šiame darbe nagrinėjama viena iš kalbos sintezės sudėtinių dalių – automatinis teksto kirčiavimas.

Tam, kad sintezuota kalba skambėtų suprantamai ir natūraliai (tiksliau – kad tekstas būtų tinkamai transkribuojamas, o intonacija ir garsų trukmė tinkamai modeliuojama), reikia nustatyti teksto žodžių kirčiavimą. Papildomų problemų atsiranda kirčiuojant žodžius, kurie yra vienodai rašomi, bet skirtingai tariami (**homografus**). Be to, šnekamajai kalbai būdingas ritmas, t. y. kirčiuotų ir nekirčiuotų skiemenų kaitaliojimasis. Siekiant išlaikyti kalbos ritmą, kai kurie žodžiai lieka nekirčiuoti (tampa **klitikais**).

Lietuvių kalba yra laisvojo kirčio kalba, be to, ji turi išplėtotą gramatinių formų sistemą (t. y. priklauso **fleksinėms** kalboms), todėl automatinio kirčiavimo uždavinys yra netrivialus. Automatinis kirčiavimas ir homografų vienareikšminimas iki šiol yra vos kelių autorių nagrinėtas, o lietuvių kalbos

klitikų paieškos algoritmai apskritai dar nebuvo analizuoti. Lietuvių kalbos teksto kirčiavimo, homografų vienareikšminimo ir klitikų paieškos problemų sprendimas svarbus siekiant sukurti aukštesnės kokybės lietuvišką balso sintezatorių.

Darbo tikslai ir uždaviniai

Pagrindinis šio darbo tikslas yra sukurti lietuvių kalbos teksto automatinio kirčiavimo algoritmus ir realizuoti juos kompiuterinėse programose. Šie algoritmai turėtų atlikti automatinį lietuvių kalbos žodžių kirčiavimą, homografų vienareikšminimą bei klitikų paiešką. Algoritmai turėtų nenusileisti tikslumu kitiems, jau egzistuojantiems, algoritmams (jei tokių yra). Siekiant šio tikslo, buvo sprendžiami šie uždaviniai:

1) Nustatyti automatinio kirčiavimo, homografų vienareikšminimo ir klitikų paieškos vietą bendroje balso sintezės schemoje, jų sąveiką su kitais moduliais, gaunamus ir perduodamus duomenis. Išnagrinėti kitose kalbose šiems uždaviniams spręsti taikytus metodus.
2) Paruošti pakankamos apimties (maždaug milijono žodžių) įvairių žanrų kirčiuotą tekstyną. Tuo tikslu sukurti programinę įrangą, reikalingą tekstynui paruošti. Šis tekstynas bus naudojamas eksperimentams ir algoritmų tikslumui įvertinti.
3) Pasiūlyti naują lietuvių kalbos homografų vienareikšminimo algoritmą.
4) Pasiūlyti naują lietuvių kalbos žodžių kirčiavimo algoritmą.
5) Pasiūlyti kalbos lietuvių kalbos ritmo nustatymo (klitikų paieškos) algoritmą.
6) Realizuoti pasiūlytus algoritmus ir eksperimentiškai įvertinti jų tikslumą.

Darbo metodai ir priemonės

Teoriniai tyrimai atlikti panaudojant metodus, sąvokas ir kitas žinias iš kalbotyros, kompiuterinės lingvistikos, atpažinimo teorijos, grafų teorijos, matematinės statistikos ir programavimo.

Eksperimentiniai tyrimai ir šiame darbe pasiūlyti algoritmai atlikti naudojantis specialiai šiuo tikslu sukurta programine įranga, parašyta *C++* programavimo kalba, naudojant *Microsoft Visual Studio 6.0* programavimo aplinką. Tai pat buvo naudojama P. Kasparaičio sukurta lietuvių kalbos žodžių kirčiavimo ir morfologinės analizės programinė įranga.

Mokslinis naujumas

Lietuvių kalbos automatiniam kirčiavimui iki šiol buvo išimtinai taikyti tik morfologine analize grįsti metodai. Tokie algoritmai yra sudėtingi, todėl sunkiai perkeliami į kitas programavimo kalbas ar operacines sistemas, juos sunku modifikuoti ar optimizuoti. Šiame darbe pasiūlytas metodas remiasi tik raidžių sekomis, nereikalauja jokių žinių apie kalbą, todėl yra itin paprastas, greitas, lengvai pritaikomas kitoms kalboms. Kirčiavimo taisyklės sudaromos automatiškai iš didelio kiekio kirčiuotų tekstų. Tokie metodai paprastai taikomi nekaitomoms arba silpnai kaitomoms (**nefleksinėms**) kalboms, pvz., anglų.

Lietuvių kalbos homografams vienareikšminti iki šiol buvo naudoti HMM, ID3 ir sintaksine analize grįsti metodai. Visi jie remiasi žodžio kontekstu. Šiame darbe pasiūlytas metodas remiasi leksemų ir morfologinių pažymų dažniais, o kontekstinės informacijos visai nenaudoja.

Taip pat šiame darbe pasiūlytas lietuvių kalbos klitikų paieškos algoritmas. Kalbotyros darbuose galima rasti aprašytas tik bendras žodžių virsmo klitikais tendencijas, o klitikų automatinio radimo algoritmai iki šiol visai nebuvo analizuoti.

Ginamieji teiginiai

1) Lietuvių kalbos homografų vienareikšminimo algoritmas, pagrįstas leksemų ir morfologinių pažymų vartosenos dažniais.
2) Lietuviško teksto kirčiavimo algoritmas, pagrįstas raidžių sekomis žodžiuose. Kirčiavimo taisyklių automatinio sudarymo iš didelio kiekio kirčiuotų tekstų algoritmas. Kirčiavimo taisyklių skaičiaus sumažinimo algoritmas.

3) Lietuvių kalbos klitikų paieškos tekste algoritmas, pagrįstas samplaikinių formų atpažinimu, žodžio kirčiavimo/nekirčiavimo statistiniu dažniu, gramatikos taisyklėmis bei gretimų žodžių kirčiavimu.

Praktinis taikymas

1) Homografų vienareikšminimo bei klitikų paieškos algoritmai yra naudojami internetiniame lietuvių kalbos sintezatoriuje [http://www.studijos.lt/sintezatorius, žiūrėta 2010.04.13].
2) Žodžių kirčiavimo algoritmas yra naudojamas *UAB „Etalinkas*“ lietuvių kalbos sintezatoriuje. Prieiga per internetą: [http://www.etalink.lt/lietuviu-kalbos-sintezatorius, žiūrėta 2010.01.05].
3) Klitikų paieškos algoritmas taip pat yra naudojamas kirčiavimo programoje *AccentTools* (žr. šio darbo 3 skyrių).

Disertacijos struktūra ir turinys

Disertacija susideda iš įvado, šešių skyrių, išvadų, literatūros sąrašo, dviejų priedų, bei sąvokų ir santrumpų sąrašo. Pagrindinę dalį sudaro 139 puslapiai įskaitant 22 paveikslėlius ir 25 lenteles. Literatūros sąraše 160 nuorodų.

Pirmame skyriuje pateikiama šnekamosios kalbos apdorojimo (t. y. sintezės, atpažinimo ir interpretavimo) sistemų apžvalga. Smulkiau išnagrinėtos kalbos sintezės sistemos, jų skaidymas į smulkesnius blokus, daugiausia dėmesio skiriant tekstinei analizei, nurodyta automatinio kirčiavimo, homografų vienareikšminimo bei klitikų paieškos uždavinių vieta šiame etape, jų sąveika su kitais moduliais, gaunami ir perduodami duomenys.

Antrame skyriuje pateikiama kirčio sąvoka, pasaulio kalbų klasifikacija pagal kirčiavimo paradigmas bei kaitymo laipsnį, apžvelgti įvairioms kalboms taikyti kirčiavimo algoritmai, jų pasirinkimas pagal kirčiavimo paradigmas ir kaitymo galimybes.

Trečiame skyriuje aprašoma duomenims ruošti naudota programinė įranga bei duomenys.

Ketvirtame skyriuje pateikiama homografo sąvoka, homografų tipai, jų vienareikšminimo problema kitose kalbose, lietuvių kalbai taikyti metodai, aprašytas autoriaus pasiūlytas homografų vienareikšminimo algoritmas bei būdai vienareikšminimo taisyklių skaičiui sumažinti, eksperimentiškai įvertintas tikslumas, rezultatai palyginti su kitų autorių sukurtų algoritmų rezultatais.

Penktame skyriuje pagrįstas naujų kirčiavimo algoritmų poreikis, aprašyti sprendimo medžiai (CART) bei jų naudojimas tekstui kirčiuoti, aprašyti autoriaus pasiūlyti lietuviško teksto automatinio kirčiavimo algoritmai, naudojantys sprendimo medžius. Išnagrinėti algoritmai kirčiavimo taisyklių skaičiui sumažinti. Eksperimentais įvertintas kirčiavimo tikslumas, rezultatai palyginti su morfologine analize grįstu metodu. Pateikta mokymo duomenų apimties plėtimo įtakos kirčiavimo tikslumui prognozė.

Šeštame skyriuje apibrėžta klitiko sąvoka, apžvelgta klitikų problema kitose kalbose, lietuvių kalbos žodžio tapsmą klitiku lemiantys veiksniai, pateikiamas autoriaus pasiūlytas lietuvių kalbos klitikų paieškos tekste algoritmas. Eksperimentais įvertintas tikslumas, numatytos tobulinimo kryptys.

Prieduose pateikti kai kurie eksperimentų rezultatai.

Autoriaus publikacijos disertacijos tema

1) Anbinderis, T., ir P. Kasparaitis (2007). Klitikų paieškos lietuviškame tekste algoritmai. *Kalbų studijos*, **10**, 30-37, (MLA).
2) Anbinderis, T., ir P. Kasparaitis (2009). Lietuvių kalbos homografų vienareikšminimas remiantis leksemų ir morfologinių pažymų vartosenos dažniais. *Kalbų studijos*, **14**, 25-31, (MLA).
3) Anbinderis, T., (2010). Automatic Stressing of Lithuanian Text Using Decision Trees. *Information Technology And Control*, **39**(1), 61-67, (INSPEC).

1 Kalbos sintezės sistemų architektūra

Dabar neįmanoma pasakyti, kada ir kaip atsirado žmonių gebėjimas kalbėti[1], tačiau šiandien šnekamoji kalba – tai vienas pagrindinių žmonių tarpusavio bendravimo, t. y. informacijos apsikeitimo, būdų. Atsiradus kompiuteriams, atsirado ir susidomėjimas, kaip realizuoti žmogaus ir kompiuterio bendravimą šnekamąja kalba. Deja, ši užduotis vis dar lieka neišspręsta ir dauguma šios dienos kompiuterinių sistemų keistis informacija dažniausiai naudoja kitus, nesusijusius su šnekamąja kalba, t. y. kalbėjimu ir klausymu, įrenginius, tokius kaip monitorius (informacijos išvedimui), klaviatūra, „pelė“ (informacijos įvedimui) ir t. t. Pvz., tam, kad būtų įvykdyta elementari užduotis *padidinti garsą*, kai kurios grafinės kompiuterinės sistemos reikalauja kelių „langų“ atidarymo bei kelių pelės paspaudimų. Kai įvedimui naudojamas balsas, šiai užduočiai atlikti užtektų pasakyti frazę: *padidinti garsą*. Kita vertus balso panaudojimas informacijos išvedimui yra patogus, kai vartotojo akys „užimtos“, pvz., vairuojant mašiną ir pan. Daugiau apie šnekamosios kalbos kompiuterinius taikymus žr. [Huang ir kt., 2001, 899-935]. Remiantis [Huang ir kt., 2001, 3], manoma, kad sistemų, kurios galės apdoroti šnekamąją kalbą taip pat, kaip žmogus, atsiradimas signalizuos „tikrai protingų mašinų“ (angl. *truly intelligent machines*) atsiradimą. 1950 metais Alan Turing [Turing, 1950] pasiūlė savo žymųjį „Turingo testą“, kur kalbos vartojimas taip, kaip ją vartoja žmogus, yra pakankama sąlyga „protingumui“ nustatyti. Turing pasiūlė žaidimą: jei mašina laimės žaidimą, ją laikysime „protingąja“. Žaidime jūs atliekate tardytojo vaidmenį. Užduodamas klausimus per teletaipą, jūs turite identifikuoti kitus du dalyvius: mašiną ir žmogų. Mašinos užduotis, atsakant į jūsų klausimus taip, kaip atsakytų žmogus,

[1] Remiantis [HistoryWorld, 2009], iš viso pasaulyje yra apie 5000 kalbų (trečdalis jų Afrikoje), tačiau mokslininkai sugrupuoja visas kalbas į maždaug dvidešimt grupių. Šiandien plačiausiai paplitusi kalbų grupė yra Indo-Europiečių (apie pusė pasaulio populiacijos). Šios kalbų grupės šaknis galima atsekti iki maždaug 3000 metų prieš mūsų erą. Jai priklauso ir lietuvių kalba.

apgauti jus ir įtikinti, kad ji (mašina) yra žmogus. Kito dalyvio (žmogaus) užduotis yra įtikinti jus, kad pirmasis dalyvis yra mašina. Nors šiandien kompiuterinės sistemos dar neįveikia Turingo testo, tyrimai šnekamosios kalbos apdorojimo (toliau **SLP**, angl. *Spoken Language Processing*) sistemų srityje vykdomi jau apie 40 metų (vieni iš pirmųjų darbų: [Flanagan, 1972a], [Flanagan, 1972b], [Hyde, 1972], [Klatt, 1977], [Reddy, 1975], [Reddy, 1976] ir kt.).

Toliau šiame skyriuje suklasifikuotos SLP sistemos, pateikta bendroji balso sintezės schema. Kadangi šiame darbe pagrindinis dėmesys skirtas lietuvių kalbos teksto automatiniam kirčiavimui bei su tuo susijusiems uždaviniams – homografų vienareikšminimui ir klitikų paieškai, tai šiame skyriuje nusakyta minėtų uždavinių vieta bendroje balso sintezės schemoje, jų sąveika su kitais moduliais, gaunamieji ir perduodamieji duomenys.

1.1 Šnekamosios kalbos apdorojimo sistemos

Pradėkim nuo SLP sistemų apibrėžimo. SLP sistemos – tai kompiuterinės sistemos, kurios turi mažiausiai vieną iš šių posistemių [Huang ir kt., 2001, 4-8]:

- **Kalbos sintezės** (toliau **TTS**, angl. *Text-to-Speech*) sistemos automatiškai generuoja žmogaus balsą iš bet kokios tekstinės įvesties (naudodamos raidžių keitimą fonemomis [Dutoit, 1997]).
- **Automatinio kalbos atpažinimo** (toliau **ASR**, angl. *Automatic Speech Recognition*) sistemos paverčia šnekamąją kalbą į rašytinį tekstą. ASR uždavinį galima įsivaizduoti kaip TTS „atvirkščiai“, tačiau ASR sistemų „kokybės lygis“ dar labiau atsilieka nuo žmogaus sugebėjimų, nei TTS sistemų: sintezuojant kalbą užtenka ją sintezuoti vieno žmogaus balsu, o atpažįstant reikia prisitaikyti ne tik prie skirtingų žmonių balsų, bet ir prie skirtingo to paties žmogaus tarimo, skirtingos kalbėjimo aplinkos ir t. t.
- **Šnekamosios kalbos suvokimo** (toliau **SLU**, angl. *Spoken Language Understanding*) sistemos susieja žodžius ir veiksmus [Huang ir kt., 2001, 4], t. y. interpretuoja frazes kontekste ir įvykdo atitinkamus veiksmus.

Pvz., net jei ASR sistema sėkmingai pavertė ištartą žodžių seką į rašytinę formą – sistema vis tiek nežino, ką daryti, nes dažnai nėra tiesioginio atitikimo tarp žodžių sekos arba sintaksinės sakinio struktūros ir sistemos funkcijų (veiksmų). Kita vertus, ASR problemos sprendimas pats gali reikalauti semantinės analizės, kalbos ar dalykinės srities žinių tam, kad sumažintų galimų atpažinimo hipotezių skaičių [Huang ir kt., 2001, 835], todėl SLU sistemų tyrimai yra tiesiogiai susiję su ASR sistemų tyrimais. Be to, dažnai SLU sistemose kalbai „įvesti“ ir „išvesti“ yra realizuojami ir ASR bei TTS blokai, pvz., „verčiančio“ telefono uždavinys [Kurematsu, 1992], leidžiantis bendravimą tarp žmonių, kalbančių skirtingomis kalbomis. SLU sistemai realizuoti reikia taikyti leksines, sintaksines ir semantines žinias. Šios sistemos yra toliausiai nuo žmogaus lygio, palyginti su ASR ir TTS, ypač bendros paskirties uždaviniams spręsti [Huang ir kt., 2001, 13].

Mus šiame darbe dominantis automatinio kirčiavimo uždavinys yra vienas iš TTS sistemų sudėtinių uždavinių, nors gali būti naudojamas ir ASR. TTS sistemas toliau ir nagrinėsime, tačiau pagrindinių technologijų ar metodų aibės, naudojamos šiose trijose sistemose, gana stipriai persidengia [Huang ir kt., 2001, 4], todėl iš pradžių trumpai apžvelkime ir ASR sistemas (1.2 skyrelis). SLU sistemos, kaip jau buvo minėta, stipriai susijusios su ASR, todėl jų atskirai nenagrinėsime. Plačiau apie SLU sistemas žr., pvz., [Allen, 1987], [Grosz ir kt. 1986], [Jurafsky, Martin, 2009], [Manning, Schutze, 1999].

1.2 Automatinis kalbos atpažinimas

Žmonių tarpusavio bendravimas balsu atrodo toks natūralus, kad kartais pamirštame kalbos sudėtingumą, o juk net tas pats tekstas, ištartas to paties kalbėtojo (diktoriaus) gali būti labai skirtingai akustiškai realizuotas [Schroeder, 2004, 42].

Tipinė ASR sistema[2] turi šiuos pagrindinius komponentus: a) signalo apdorojimo, b) dekodavimo, ir c) adaptacijos blokus, bei d) akustinį ir e) kalbos modelius. Remiantis [Huang ir kt., 2001, 4-5], [Lee, 1989, 1-9], [Schroeder, 2004, 42-44], [Waibel, Lee, 1990, 2-3], faktoriai, labiausiai įtakojantys ASR sistemos sudėtingumą ir atpažinimo kokybę, yra:

1) **Užduoties apribojimai**:
   a) Ar reikia atpažinti atskirus (izoliuotus) žodžius, ar ištisinę kalbą – izoliuotus žodžius atpažinti žymiai lengviau negu ištisinę kalbą.
   b) Ar reikia atpažinti vieną kalbėtoją, ar sistema turi atpažinti bet kokį kalbėtoją[3], t. y. sistema priklausoma ar nepriklausoma nuo kalbėtojo. Jei reikia atpažinti skirtingų kalbėtojų balsą, tai tenka atsižvelgti į tokias individualias kalbėtojų charakteristikas, kaip kalbėtojo lytis, dialektas, kalbėjimo stilius ir t. t. Dauguma šiuo metu patikimai veikiančių sistemų yra priklausomos nuo kalbėtojo. Tokiose sistemose dažniausiai reikalaujama, kad naujas kalbėtojas iš pradžių „apmokytų“ sistemą, t. y. perskaitytų tam tikrą kiekį sakinių, tačiau tai nėra patogu, nes vartotojas gaišta laiką „apmokymui“ ir t. t. [Lee, 1989, 5].
2) **Kalbos modelis**:
   a) Žodyno apribojimai, t. y. žodyno dydis – kuo didesnis leistinų žodžių kiekis, tuo daugiau klaidų galima tikėtis [Waibel, Lee, 1990, 2].
   b) Kalbos apribojimai: patikima ASR sistema turi remtis ne tik atpažįstamos kalbos gramatika (nenagrinėti gramatiškai netaisyklingų sakinių), bet, kiek įmanoma, ir semantika, t. y. potencialia reikšme. Pvz., žymaus amerikiečių lingvisto N. Chomsky sugalvotas [Chomsky, 1957] sakinio pavyzdys: „*Bespalvės žalios idėjos miega įnirtingai*“ (angl. *Colorless green*

[2] Vienas iš pirmųjų ASR sistemų apžvalgų žr. jau minėtuose [Hyde, 1972], [Reddy, 1975], [Reddy, 1976].

[3] Yra netgi atskira netriviali problema: kalbėtojo (diktoriaus) identifikacija arba atpažinimas (angl. *speaker recognition*).

ideas sleep furiously) demonstruoja gramatikos „vertę“ be semantikos. Šis gramatiškai taisyklingas sakinys neturi jokios semantinės prasmės – tokie sakiniai irgi neturi būti nagrinėjami sistemoje kaip galimas atsakymo variantas. Bendrai tariant, ASR sistemos kalbos modelio tikslas turėtų būti: kuo didesni kalbos apribojimai, tuo pačiu metu leidžiantys kuo laisvesnį įvedimą [Waibel, Lee, 1990, 2-3].

3) **Akustinis modelis**: aplinkos triukšmas ir variantiškumas. Čia galima priskirti ir mikrofono charakteristikų variantiškumą, kalbėtojo sukuriamus triukšmus (pvz., čiaudėjimą), kalbėjimo tempo pasikeitimus, kalbėtojo emocinės būklės pasikeitimus, ir t. t.

ASR labai priklauso nuo užduoties: kelių izoliuotų žodžių iš mažos apimties žodyno, kalbant vienam kalbėtojui, atpažinimas yra žymiai paprastesnis uždavinys negu rišlios skirtingų kalbėtojų kalbos su neapibrėžtu žodynu atpažinimas. Pvz., kreditinių kortelių numerio atpažinimo kokybė gali siekti 99,9 proc., tačiau „gerai“ atpažinti triukšmingą telefoninį neriboto žodyno pokalbį šiandieninės sistemos dar nesugeba [Waibel, Lee, 1990, 3]. Remiantis [Huang ir kt., 2001, 12-13], „spontaninės“ telefono kalbos atpažinimo klaida kompiuteriams yra apie 10 kartų didesnė negu žmonėms (36,7 ir 3,8 proc. atitinkamai).

Reikėtų atkreipti dėmesį, kad įmanomi įvairūs mažiausi atpažinimo vienetai. Remiantis [Lee, 1989, 12-13] jais gali būti: žodis, fonema, skiemuo, difonas ir t. t.

Remiantis [Waibel, Lee, 1990, 3-4], pagrindinius ASR metodus galima išskaidyti į:

- **Šablonais grįstus** (angl. *template-based*) metodus, kur kalbos vienetai (dažniausiai žodžiai) atvaizduojami šablonais tokia pačia forma, kaip ir įėjimo kalba. Atpažinimas vykdomas lyginant įvedimą su duomenų bazės elementais. Šablonai lyginami naudojant atstumo metrikas, o laiko variantiškumo (angl. *temporal variability*) problema sprendžiama naudojant dinaminio laiko skalės kraipymo (**DTW**, angl. *Dynamic Time*

Warping) algoritmą. Tokie metodai geriausiai tinka paprastiems taikymams. Žr., pvz., [Itakura, 1975].

- **Žiniomis grįstus** (angl. *knowledge-based*) metodus. Taisyklės gali būti sudaromos ekspertų arba automatiškai analizuojant kalbos spektrogramas, t. y. akustines kalbos signalo savybes. Lingvistinės ar fonetinės žinios apie kalbą paprastai nėra naudojamos. Žr., pvz., [Zue, 1985], [De Mori ir kt., 1987], [Waibel, Lee, 1990, 197-262].
- **Stochastinius metodus**. Stochastinės ASR sistemos yra šiuo metu pačios populiariausios. Stochastiniai balso analizės metodai naudoja tikimybinius modelius šnekamosios kalbos įvesties neapibrėžtumui modeliuoti. Dažniausiai naudojami „paslėptieji Markovo modeliai“ (toliau **HMM**, angl. *Hidden Markov Model*). HMM buvo aprašyti žymaus rusų matematiko A. A. Markovo [Markov, 1913]. Remiantis [Hain, 2001, 1], [Huang ir kt., 2001, 408], pradedant nuo 1980-ųjų metų HMM tapo pagrindine priemone kuriant ASR sistemas ir modeliuojant kalbą. Pirmieji darbai: [Baker, 1975] (*Carnegie-Mellon Universitetas*), [Jelinek, 1976] (*IBM*). Vėliau, pvz., [Rabiner ir kt., 1989] (*Bell Labs*). Kiti darbai, pvz.: [Hain, 2001], [Huang ir kt., 1990], [Jelinek, 1997, 15-38], [Lee, 1989], [Waibel, Lee, 1990, 263-370]. ASR metodų, naudojančių HMM, kritika išdėstyta: [Levinson, 1994], [Russell, 1997], [Bourlard, 1995].
- **Dirbtiniais neuroniniais tinklais** (toliau **ANN**, angl. *Artificial Neural Network*) grįstus metodus, kur ASR sistema yra apmokoma pagal įvedimo duomenis nustatant ANN neuronų svorius. Žr. pvz., [Bourlard, Morgan, 1994], [Kohonen, 1988], [Robinson, 1994], [Sakoe ir kt., 1989], [Tebelskis, 1995].

Dar 1969 metais žinomas amerikiečių komunikacijų inžinierius J. R. Pierce [Pierce, 1969] teigė, kad visos ASR kūrimo pastangos yra bevertės, kad tai yra tik laiko ir pinigų švaistymas, ir palygino ASR problemos sprendimo sudėtingumą su vėžio išgydymu, aukso iš jūros gavimu, vandens vertimu į benziną arba kelione į Mėnulį. Nors nuo to laiko įvyko nemažas progresas, tačiau galutinis tikslas – „žmogaus lygio“ ASR – vis dar

nepasiektas. Kai kurie tyrinėtojai mano, kad priartėti prie ASR problemos sprendimo neįmanoma be radikalių inovacijų [Jelinek, 1997, 12].

ASR sistemos lietuvių kalbai aprašytos, pvz., [Filipovič, Lipeika, 2004], [Kasparaitis, 2008], [Laurinčiukaitė, 2003], [Lipeika ir kt., 2002], [Raškinis, Raškinienė, 2003], [Rudžionis, Rudžionis, 1996], [Rudžionis ir kt., 2007], [Šilingas, Telksnys, 2004].

1.3 Kalbos sintezė

Kaip jau buvo minėta, kalbos sintezės arba TTS sistema[4] generuoja kalbos signalą iš bet kokio teksto. TTS sistemos šiuo metu dažniausiai naudojamos [Taylor, 2009, 2]:

- akliems žmonėms;
- skambučių centrams (angl. *call-centre*) automatizuoti (pvz., informuoti apie sąskaitas už elektrą).

Kitoms užduotims TTS sistemos dar nėra itin populiarios dėl sintezuoto balso natūralumo ar suprantamumo stokos.

TTS privalumai [Huang ir kt., 2001, 679-680]:

- užtikrina ypač aukštą suspaudimo koeficientą;
- leidžia ypač lanksčiai pasirinkti kalbėtojo lytį ir balsą, kalbėjimo stilių ir tempą, pagrindinio tono kitimo diapazoną bei kitus balso parametrus;
- galima sinchronizuoti pateikiamą tekstinę informaciją ir balsą;
- alternatyvi prieiga prie tekstinės informacijos, pvz., akliems žmonėmis, vairuojant mašiną, tamsoje ar pan.;
- sintezuojamą tekstą galima pakeisti labai greitai, daug greičiau nei įdiktuoti naują balso įrašą per mikrofoną.

TTS sistemų kokybė, kaip ir ASR bei SLU, dar toli nuo to lygio, kai jos įveiks Turingo testą (žr. 1 skyriaus įvadą), t. y. susilygins su žmogaus balso

[4] Kalbos sukūrimo akustiniai aspektai analizuojami [Fant, 1960]. Vieni iš pirmųjų TTS darbų yra jau minėti [Flanagan, 1972a], [Flanagan 1972b] bei [Allen ir kt., 1979], [Allen ir kt., 1987] ir kt. Vėlesni darbai: [Carlson, 1994], [Dutoit, 1993], [Dutoit, 1997], [Kleijn, Paliwal, 1995], [Taylor, 2009], [van Santen, 1997b] ir kt.

kokybe. Tačiau net tokio pakankamai žemo lygio sistemos jau yra naudojamos praktiškai, yra nemažai komercinių kalbos sintezatorių.

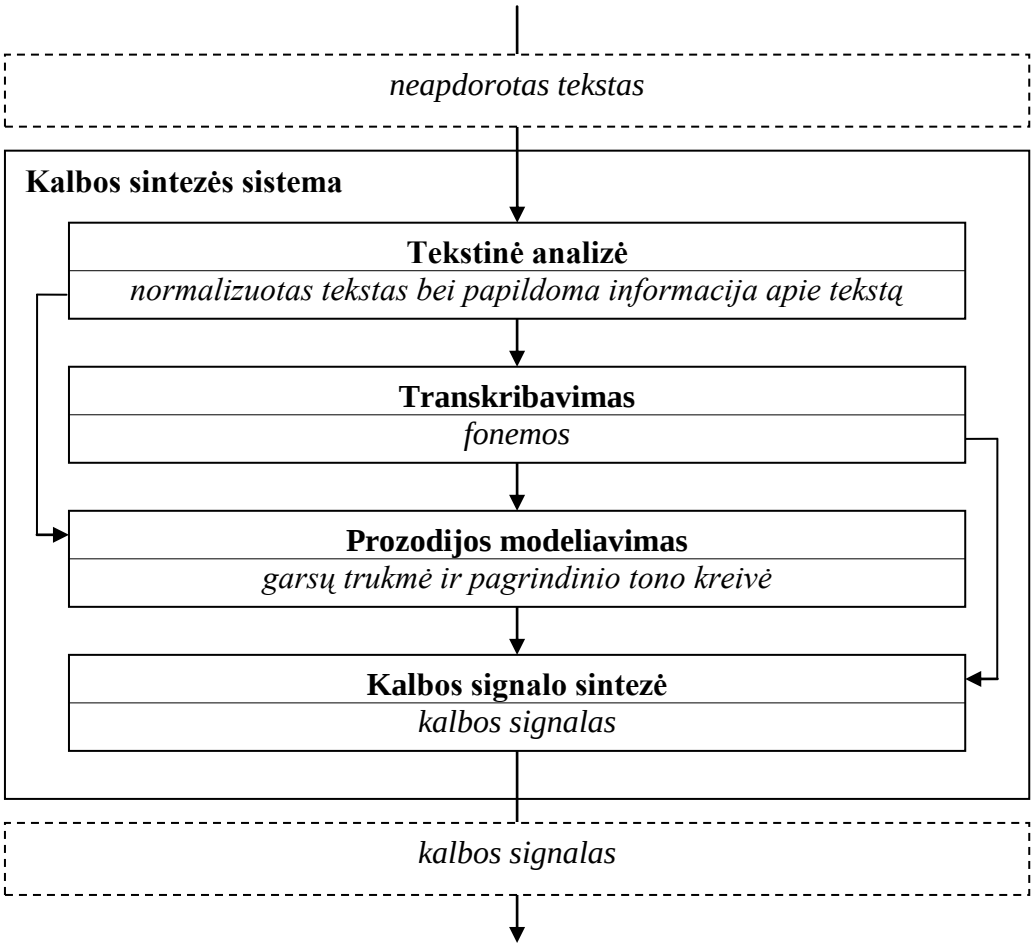
Kalbos sintezė iš teksto yra netrivialus uždavinys, jo sprendimas susiduria su įvairių nestandartinių naujų žodžių tarimu, teisingos sakinio intonacijos generavimo ir kitomis problemomis.

Tipinė TTS sistema, remiantis [Huang ir kt., 2001, 6-7], susideda iš keturių modulių[5] (žr. 1.1 pav.):

- **Tekstinės analizės** modulis a) pradinį, neapdorotą tekstą paverčia į normalizuotą formą ir b) surenka prieinamą morfologinę, sintaksinę, semantinę ir kitokią informaciją apie tekstą – jo sakinius ir žodžius.
- **Transkribavimo** (kartais vadinamas fonetinės analizės) modulis remiantis informacija iš tekstinės analizės modulio raides pakeičia fonemomis. Po šio etapo tekstinė informacija nebenaudojama, tik fonemos ir „žymės“ (angl. *tags*), pvz., kirčio ženklai.
- **Prozodijos modeliavimo** (kartais vadinamas prozodinės analizės) modulis pagal gaunamą iš tekstinės analizės ir transkribavimo modulių informaciją (fonemas, kirčių žymes ir t. t.) nustato a) pagrindinio tono parametrus ir b) segmentų (pvz., fonemų) trukmes.
- **Kalbos signalo sintezės** modulis pagal gaunamą fonetinę ir prozodinę informaciją suformuoja skaitmeninį kalbos signalą.

  Apžvelkime kiekvieną TTS sistemos modulį atskirai.

[5] Kitoks TTS „padalinimas“ pateiktas [Dutoit, 1993, 24], [Dutoit, 1997], [Šef ir kt., 1998]. Čia kalbos sintezės sistema dalinama į du pagrindinius blokus: a) lingvistinio teksto apdorojimo arba natūralios kalbos apdorojimo ir b) kalbos signalo formavimo arba apdorojimo. Tačiau lingvistinio teksto apdorojimo blokas vis tiek analogiškai skaidomas į tris etapus: i) tekstinės analizės arba teksto apdorojimo, ii) fonetinės analizės arba transkribavimo ir iii) prozodinės analizės arba prozodijos modeliavimo.



1.1 pav. Bendra TTS sistemos architektūra

1.4 Tekstinės analizės modulis

Tekstinės analizės (**TA**) modulis iš neapdoroto pradinio įvesties teksto:

1) sugeneruoja normalizuotą tekstą;
2) surenka įvairią papildomą informaciją apie tekstą, kurią gali sudaryti:
   a) sakinių ribos;
   b) teksto pastraipų, antraščių, skirsnių žymėjimai;
   c) kalbos dalys (toliau **POS**, angl. *Part of Speech*), gramatinės formos (giminė, skaičius, linksnis ir t. t.);
   d) frazių (angl. *clause*) ribos ir sintaksiniai tipai: daiktavardinė frazė (toliau **NP**, angl. *Noun Phrase*), veiksmažodinė frazė (toliau **VP**, angl. *Verb Phrase*) ir kt.;

e) sakinių funkciniai tipai (konstatuojamasis, klausiamasis, liepiamasis ir t. t.);
f) žodžių semantika (prasmė);
g) žodžių, frazių ir sakinių kirčiai;
h) rečiau – kitokia papildoma informacija, tokia kaip: i) žodžių ir frazių sinonimika (angl. *synonymy*), ii) nelaisvieji žodžių junginiai (pvz., idiomos), iii) semantinio tipo ir kalbos akto identifikavimas (prašymas, pranešimas, pasakojimas ir pan.), iv) žanro ir stiliaus analizė, ir t. t.

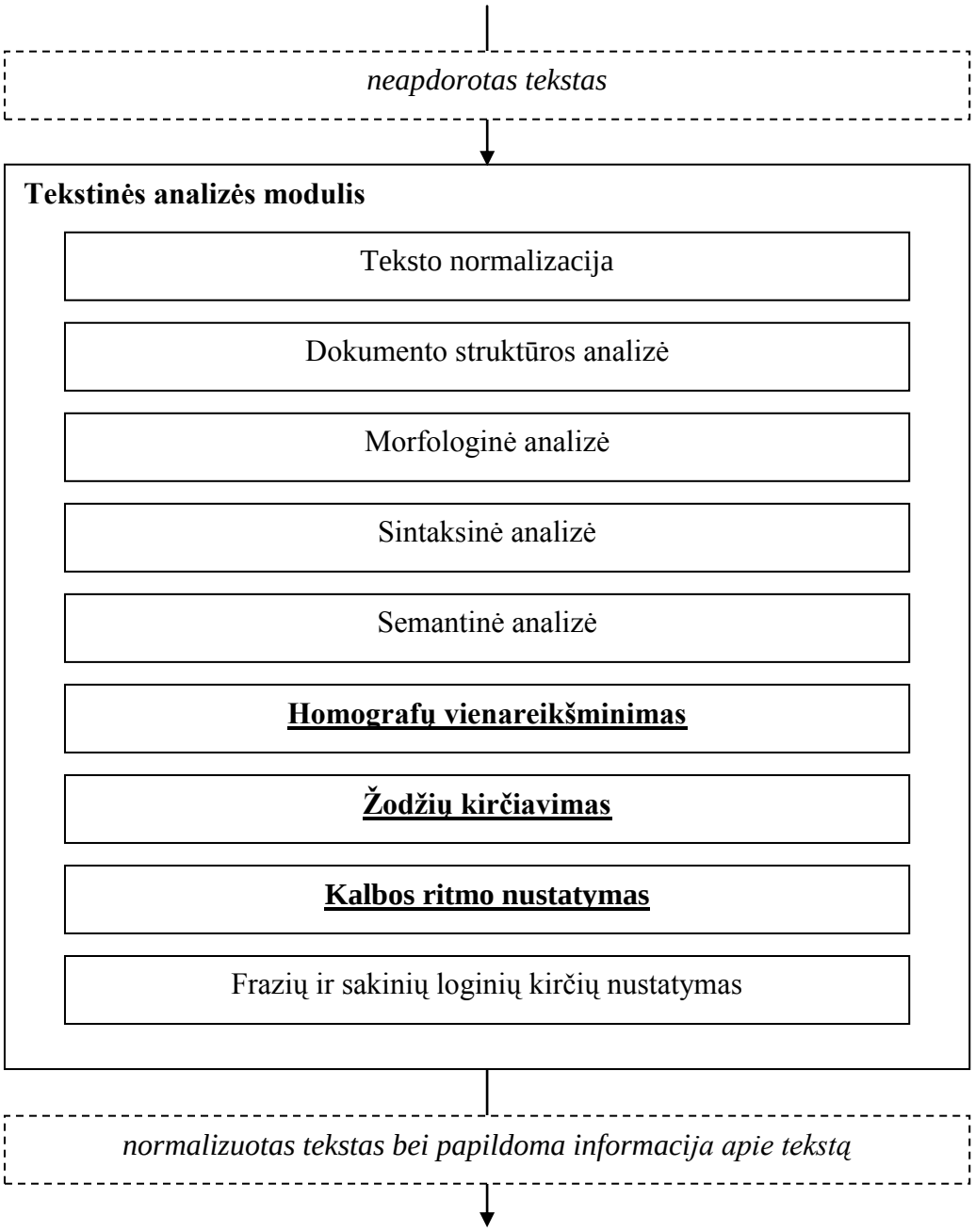
Kai kuriose sistemose TA modulis tik normalizuoja tekstą: pakeičia santrumpas ir skaičius į tekstinę formą. Sudėtingesnės sistemos surenka ir papildomą informaciją. Ši informacija gali būti naudojama transkribavimo ir prozodijos modeliavimo blokuose, žr. 1.1 pav. Taip pat ir kituose tiesiogiai su TTS nesusijusiuose uždaviniuose: informacijai gauti (angl. *information retrieval*), automatinio vertimo (angl. *machine translation*) sistemoms apmokyti ir kt.

Pagrindinės funkcijos, kurias turėtų atlikti TA modulis, pateiktos 1.2 pav. Paryškintos tos funkcijos, kurios išsamiai nagrinėjamos šiame darbe.

Visas TA modulio funkcijas jungia tai, kad visos jos naudoja tekstą ir grąžina irgi tekstą (ne fonemas ar prozodijos parametrus).

Viename paveikslėlyje atvaizduoti visus ryšius tarp funkcijų yra gana sudėtinga, todėl kiekvieno šio skyrelio poskyriuose pateiksime kiekvienos funkcijos ryšių schemą atskirai.

Pateiktoje bendroje TA modulio schemoje nėra atvaizduotos kai kurios funkcijos, pvz., skiemenavimas. Taip yra todėl, kad skiemenavimas dažnai vyksta kitų, „stambesnių“, procesų metu, pvz., kirčiuojant žodžius (jei kirčio vieta nurodoma kaip kirčiuoto skiemens numeris).



1.2 pav. Bendroji tekstinės analizės modulio funkcijų schema

Schemoje pavaizduotų TA modulio funkcijų vykdymo tvarka gali būti ir kitokia, nei siūloma 1.2 pav. pateiktoje schemoje. Pvz., dokumento struktūros analizė gali būti atliekama ir prieš teksto normalizaciją [Huang ir kt., 2001, 681], arba homografai gali būti vienareikšminami po kirčiavimo, o ne prieš jį (žr. 4 skyrių). Taigi, nors siūloma tvarka nėra vienintelė, aprašant TA modulio funkcijas bus laikomasi būtent šio eiliškumo.

Daugelyje modulių gali būti naudojami įvairūs žodynai, pvz., santrumpų žodynas teksto normalizacijos funkcijai, transkripcijų žodynas transkribavimo moduliui, kalbos signalų duomenų bazė ir pan., tačiau žodynai schemose nebus vaizduojami.

Panagrinėkime kiekvieną TA modulio funkciją atskirai.

1.4.1 Teksto normalizacija

Teksto normalizacija (**TN**) [Huang ir kt., 2001, 696-709] tai funkcija, kuri žodžiais pakeičia:

1) santrumpas (pvz., *t. t.* į *taip toliau*) ir akronimus (pvz., *LDK* į *Lietuvos Didžioji Kunigaikštystė*);
2) skaičius (telefonus, datas, laiką, valiutas, sąskaitų numerius, kelintinius ir kiekinius skaitvardžius ir t. t.);
3) specifinės srities išraiškas (pvz., matematines ar chemines formules);
4) įvairius simbolius, pvz., „~“, „©“, „™“;
5) lenteles, internetines nuorodas ir t. t.

Atrodytų, kad normalizuojant tekstą galima apsiriboti tiesiog santrumpų, skaitvardžių ir kitos ne tekstinės informacijos pakeitimu, naudojant santrumpų, skaitvardžių ir kitokius sąrašus, tačiau:

- Santrumpos gali būti nevienareikšmės, pvz., santumpa *šv.* gali reikšti ir žodį *šventas*, ir *šviesiai*, pvz., *šv. Jonas* ir *šv. žalias* [Kasparaitis, 2001b]. Tokiu atveju TN funkcija gali grąžinti ir kelias santrumpos išskleidimo hipotezes. Kelias hipotezes gali grąžinti ir kitos TA modulio funkcijos [Huang ir kt., 2001, 683-684]. Vienai hipotezei parinkti skirta homografų vienareikšminimo funkcija (žr. 1.4.6 skyrelį ir 4 skyrių). Kita vertus, TN funkcija pati gali atlikti vienareikšminimą remdamasi, pvz., kaimyninių žodžių morfologine analize. Aišku, jei sistema nesugeba pasirinkti vieno varianto arba apskritai santrumpų sąraše surasti santrumpos, galima apsiriboti perskaitymu po raidę (pvz., *LDK* perskaityti *eldėka*) arba perskaityti kaip žodį naudojant standartines tarimo taisykles (pvz., *NATO*).

- Grąžinamiems žodžiams reikia pasirinkti tinkamą gramatinę formą (giminę, skaičių linksnį ir t. t.), tai ypač aktualu fleksinei lietuvių kalbai – čia irgi galima sugeneruoti kelias gramatinių formų hipotezes ir po to vienareikšminti. Alternatyvus būdas: iš pradžių remiantis kontekstu išsiaiškinti, kokia gramatinė forma reikalinga, tik po to išskleisti santrumpą.

TN taisyklės gali būti sudaromos rankiniu būdu, pvz., [Allen ir kt., 1987] arba remiantis statistiniais metodais, pvz., [Black ir kt., 1998b].

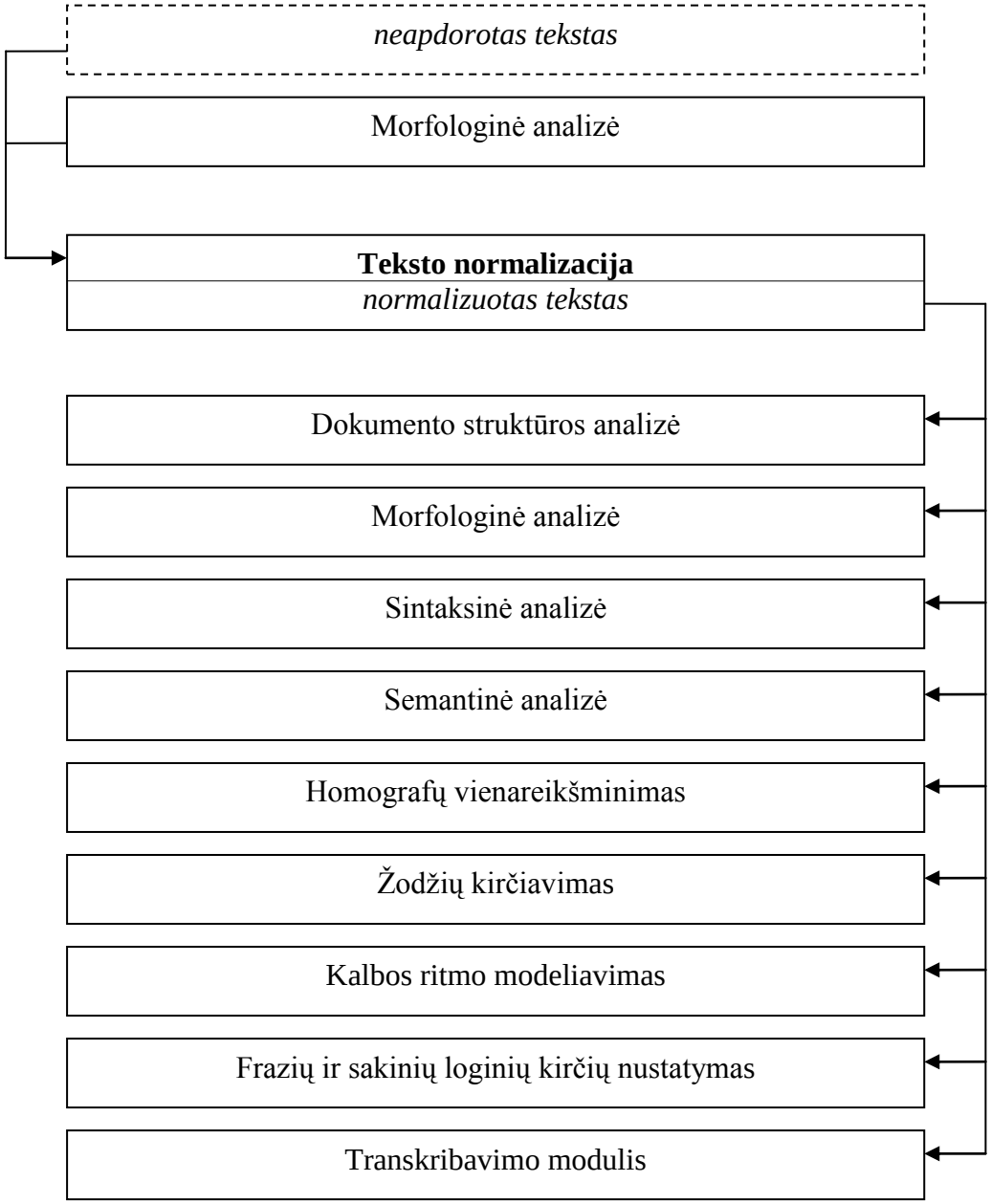
TN rezultatas (t. y. normalizuotas tekstas) toliau yra perduodamas visoms TA modulio funkcijoms bei transkribavimo moduliui, neapdorotas pradinis tekstas daugiau nebenaudojamas. TN funkcijos ryšiai su kitomis TA modulio funkcijomis ir TTS sistemos moduliais pavaizduoti 1.3 pav.

TN taip pat gali būti naudojama ir ruošiant įvairius tekstynus, arba „atvirkštinė“ teksto normalizacija automatinio kalbos atpažinimo (ASR) uždavinyje, t. y. pilni žodžiai pakeičiami santrumpomis [Huang ir kt., 2001, 696].

1.4.2 Dokumento struktūros analizė

Po teksto normalizacijos TTS sistema turėtų nustatyti dokumento struktūrą [Huang ir kt., 2001, 688-695], t. y. surasti ir pažymėti tekste tokius teksto elementus:

- sakinius (sakinių ribas);
- skyrių ir poskyrių antraštes;
- sąrašus;
- pastraipas;
- parašą elektroninio laiško pabaigoje;
- internetines nuorodas;
- skirtingo dydžio ar spalvos tekstą;
- jei tekste pateiktas dialogas – dialogo dalyvius, ir t. t.



1.3 pav. Teksto normalizacija

Kai kurių dokumento struktūros elementų išskyrimas gali būti netrivialus uždavinys, pvz., sakinio ribų nustatymas. Apie sakinių ribas dažniausiai signalizuoja skyrybos ženklai („.“, „!“, „?“) ir didžioji raidė kito sakinio pradžioje, tačiau problema ta, kad taškas tekste dažnai žymi ir santrumpas[6]. Nustatant sakinių ribas gali būti naudojami grįsti sprendimo medžiais (CART),

[6] Kai kuriose kalbose teksto išskaidymas į sakinius gali būti trivialus: pvz., kinų kalboje sakinių ribos žymimos specialiu ženklu – mažu apskritimu [Huang ir kt., 2001, 692].

įvairūs statistiniai (pvz., sakinių ilgių dažniai įvairiems žanrams) ir kiti metodai [Huang ir kt., 2001, 693-694]. Sakinio riboms nustatyti gali būti naudojama kontekstinių žodžių morfologinė, sintaksinė ir semantinė informacija.

Nuo dokumento struktūros gali priklausyti sintezuojamos kalbos prozodiniai parametrai, pvz.:

- Naujo sakinio pradžioje pagrindinis tonas dažniausiai aukštesnis negu jo viduryje ir dar žemesnis sakinio pabaigoje [Cohen ir kt., 1982].
- Analogiškai pastraipoms: naujos pastraipos pradžioje pagrindinis tonas dažniausiai aukštesnis negu jos viduryje ir dar žemesnis pastraipos pabaigoje [Sluijter, Terken, 1993].
- Elektroniniuose laiškuose galima išskirti atitinkama intonacija arba praleisti parašą laiško pabaigoje.
- Internetiniuose puslapiuose galima išskirti atitinkama intonacija internetines nuorodas, skirtingo dydžio ar spalvos tekstą.
- Jei pateiktas dialogas – rodyti kalbėtojų kaitą.

Sakinių ribos taip pat reikalingos sintaksinės analizės funkcijai.

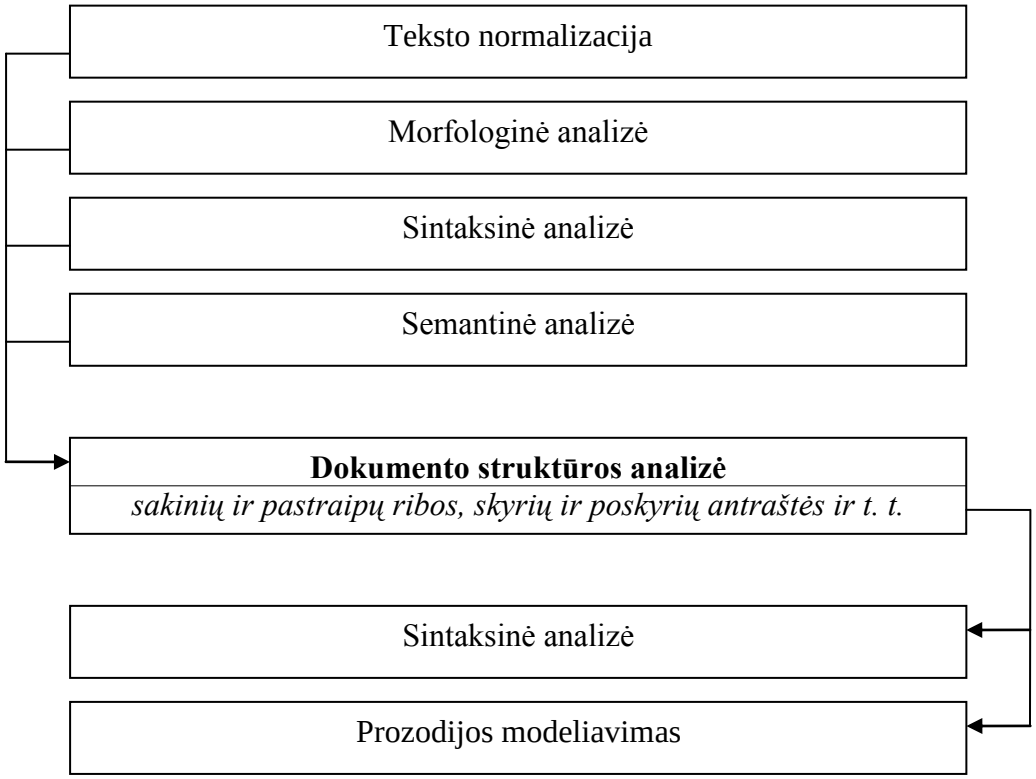
Dokumento struktūros analizės funkcijos ryšiai su kitomis TA modulio funkcijomis ir TTS sistemos moduliais pavaizduoti 1.4 pav.

1.4.3 Morfologinė analizė

Morfologinės[7] analizės (**MA**), arba dekompozicijos [Sproat, Olive, 1995], funkcija normalizuoto teksto žodžius išskaido į morfemas. Morfema – tai mažiausias gramatinis vienetas (žodžio dalis), turintis reikšmę (apibrėžimas pagal [Payne, 1997, 20-21]), pvz., lietuviškas žodis *miškelis* yra sudarytas iš trijų morfemų: šaknies *mišk*, priesagos *el* ir galūnės *is* (kamienas[8]: *miškel*).

[7] Remiantis [Huang ir kt., 2001, 55-57], yra 3 morfologijos tipai: a) kaitymo (angl. *inflectional*) morfologija i) atspindi žodžio kontekstinę situaciją ir ii) dažniausiai nepakeičia fundamentalios žodžio reikšmės (nepakeičia POS ir t. t.), b) darybinė (angl. *derivational*) morfologija gali sukurti visiškai naują žodį, dažnai su POS pakeitimu, c) sudurtinių žodžių morfologija.

[8] Formaliai kamienas nėra morfema, tačiau dažnai kompiuterinės lingvistikos srityje žodžių kamienai naudojami panašiai kaip ir žodžių šaknys.



1.4 pav. Dokumento struktūros analizė

Pasak [Šveikauskienė, 2009, 8], „Morfologinės analizės etape kiekvienam teksto žodžiui surandama antraštinė forma (bendratis, vienaskaitos vardininko linksnis ir pan.) bei nusakoma morfologinė informacija apie sakinyje pavartotą žodžio formą (giminė, skaičius, linksnis, laikas, asmuo ir kt.). Lietuvių kalbos morfologinę analizę gali atlikti Vytauto Zinkevičiaus sukurta lietuvių kalbos morfologinių modelių programinė įranga [Zinkevičius, 2000].“ Morfologinei analizei galima panaudoti ir Pijaus Kasparaičio sukurtą lietuvių kalbos žodžių kirčiavimo bei morfologinės dekompozicijos programinę įrangą [Kasparaitis, 2001b]. Žodžių šaknų ar kamienų sąrašai (duomenų bazės) paprastai naudojami nustatyti kalbos dalį, o galūnės nusako gramatinę formą.

Dažniausiai MA funkcija realizuojama sudarant:

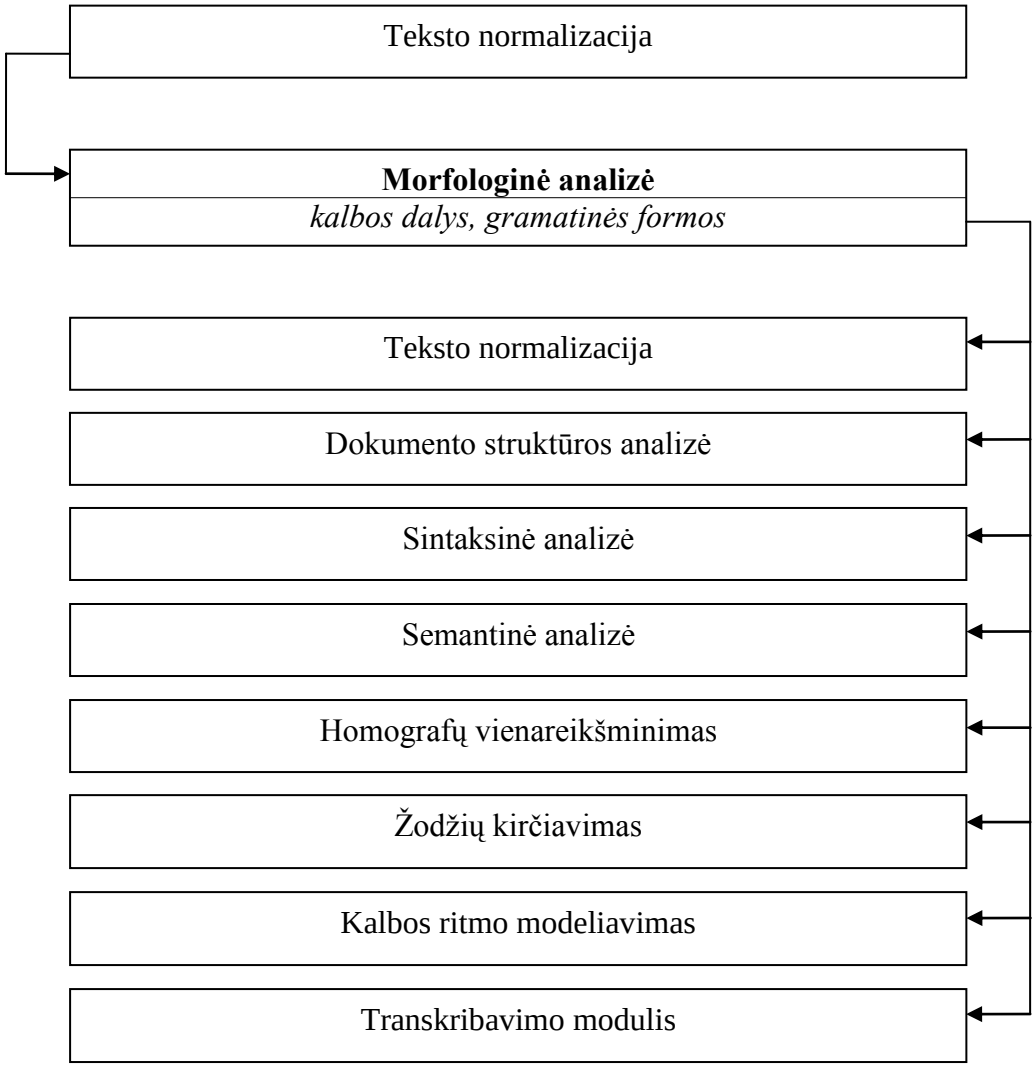
1) Morfemų žodynus, kuriuose morfemos dažniausiai grupuojamos pagal kalbos dalis.

2) Morfologines žodžio kaitymo taisykles, kurios turėtų nurodyti, kaip iš morfemų sudaryti žodį (nors iš tikrųjų reikės atlikti atvirkštinį veiksmą – pagal žodį nuspręsti, iš kokių morfemų, žodžio dalių jis sudarytas):
   a) kokios morfemos su kokiomis gali jungtis (pvz., prie veiksmažodžių ir daiktavardžių gali jungtis skirtingos priesagų aibės);
   b) kokia tvarka morfemos gali jungtis (pvz., priesaga po šaknies, galūnė po priesagos ir t. t.);
   c) ar gali būti morfema vartojama kaip žodis – atskirai, be kitų morfemų, ar gali būti sujungta su kitomis morfemomis;
   d) raidžių kitimo morfemų jungtyse taisyklės, pvz., angl. žodis *stopping* (*sustojimas*) sudarytas iš morfemų *stop* ir *ing*, tačiau jungiant šias morfemas pridedama dar viena raidė *p*, arba liet. *mesiu* sudarytas iš morfemų *mes* ir *siu*, arba *vešiu* = *vež* + *siu*;
   e) išimtys, pvz., angl. žodžių daugiskaitos formos, kurios yra sudaromos ne pagal bendras taisykles: *goose* (*žąsis*) ir *geese* (*žąsys*), bei *child* (*vaikas*) ir *children* (*vaikai*), ir t. t.

Atliekant MA gali iškilti skaidymo į morfemas (vienareišminimo) problema, pvz.:

- angl. žodis *establishment* (*įkūrimas*): atmetam priesagą *-ment*, iškyla klausimas, ar reikia toliau skaidyti *establish* į *establ* ir *-ish*;
- jau minėtas pavyzdys *stopping* turi būti išskaidytas į morfemas *stop* + *ing*, bet ne į *stopp* + *ing*. T. y. morfologinės dekompozicijos metu šiuos pakeitimus reikia atkurti;
- tų pačių raidžių sekos skaidomos į morfemas skirtingai, pvz., sekos žožių pradžioje: *neper-skaito – ne-perinti*, *prisi-rinko – pri-sirpo*, sekos žožių pabaigoje: *samuraj-ai – ger-ajai* [Kasparaitis, 2001b, 37, 43];

MA funkcija gali grąžinti kelias rezultato hipotezes (pvz., kalbos dalis ar gramatines formas), tokiu atveju kitos TA modulio funkcijos gali atlikti vienareikšminimą.



1.5 pav. Morfologinė analizė

MA funkcija kaip duomenis naudoja normalizuotą tekstą, o jos rezultatai (kalbos dalys, gramatinės formos) naudojami daugelyje TA modulio funkcijų, visą sąrašą žr. 1.5 pav. MA taip pat plačiau naudojama tiesiogiai nesusijusiose su TTS srityse, pvz., paieškos ar mašininio vertimo sistemose.

MA grįsti metodai ypač dažnai naudojami fleksinio tipo kalboms.

1.4.4 Sintaksinė analizė

Pasak [Šveikauskienė, 2009, 8], „Sakinio sintaksinė struktūra rodo, kaip žodžiai sakinyje yra susiję vienas su kitu [Allen, 1987, 9]. Labiausiai paplitęs sakinio struktūros pavaizdavimo metodas yra grafas, tiksliau, medis [Allen,

1987, 41]. Kalbininkų medžiai visada braižomi viršūne žemyn, t. y. jų šaknis yra viršuje, o lapai apačioje [Batori, Lenders, 1989, 23]. Žinomi du iš principo skirtingi medžio sudarymo būdai: frazių metodas ir priklausomybių metodas. <...> Sintaksinės analizės metu nustatomos žodžių sintaksinės funkcijos bei nurodomi ryšių tipai tarp jų.“

Sintaksinė analizė TTS sistemoje turėtų atlikti šias užduotis:

1) suskaidyti sakinius į frazes (t. y. surasti frazių ribas sakiniuose);
2) nustatyti frazių tipus (NP, VP ir t. t.) [Huang ir kt., 2001, 58-62];
3) nustatyti sakinių tipus. Standartiniai tipai, apibūdinantys anglų kalbos sakinį [Huang ir kt., 2001, 61-62]:
    a) Konstatuojamasis: *I gave her a book.* (*Aš daviau jai knygą.*);
    b) Tikrinamasis klausimas: *Did you give her a book?* (*Ar tu davei jai knygą?*);
    c) Klausimas su klausiamuoju įvardžiu: *What did you give her?* (*Ką tu jai davei?*);
    d) Klausimas su alternatyva: *Did you give her a book, a scarf, or a knife?* (*Ar tu davei jai knygą, šaliką, ar peilį?*);
    e) Klausimas prašant patvirtinimo: *You gave it to her, didn't you?* (*Ar tikrai tu jai tai davei?*);
    f) Pasyvas: *She was given a book.* (*Jai buvo duota knyga.*);
    g) Sudėtinis prijungiamasis: *It must have been a book that she got.* (*Tai, ką ji gavo, turbūt yra knyga.*);
    h) Šaukiamasis: *Hasn't this been a great birthday!* (*Ar gi tai nebuvo puikus gimtadienis!*);
    i) Liepiamasis: *Give me the book.* (*Duok man knygą.*);
4) nustatyti žodžių sintaksines funkcijas sakinyje (tarinys, veiksnys, papildinys ir t. t.).

Lietuvių kalbos sintaksinės analizės sistemos skiriasi nuo kitų kalbų sistemų. Anot [Labutis, 2002, 25], „Žodžių formų gausumas lietuvių kalboje ir nulėmė tai, kad jos tapo svarbiausia sakinio sintaksinių ryšių raiškos priemone. Kiekviena kalba turi rinkinį gramatinių ar kitokių priemonių, kurios rodo

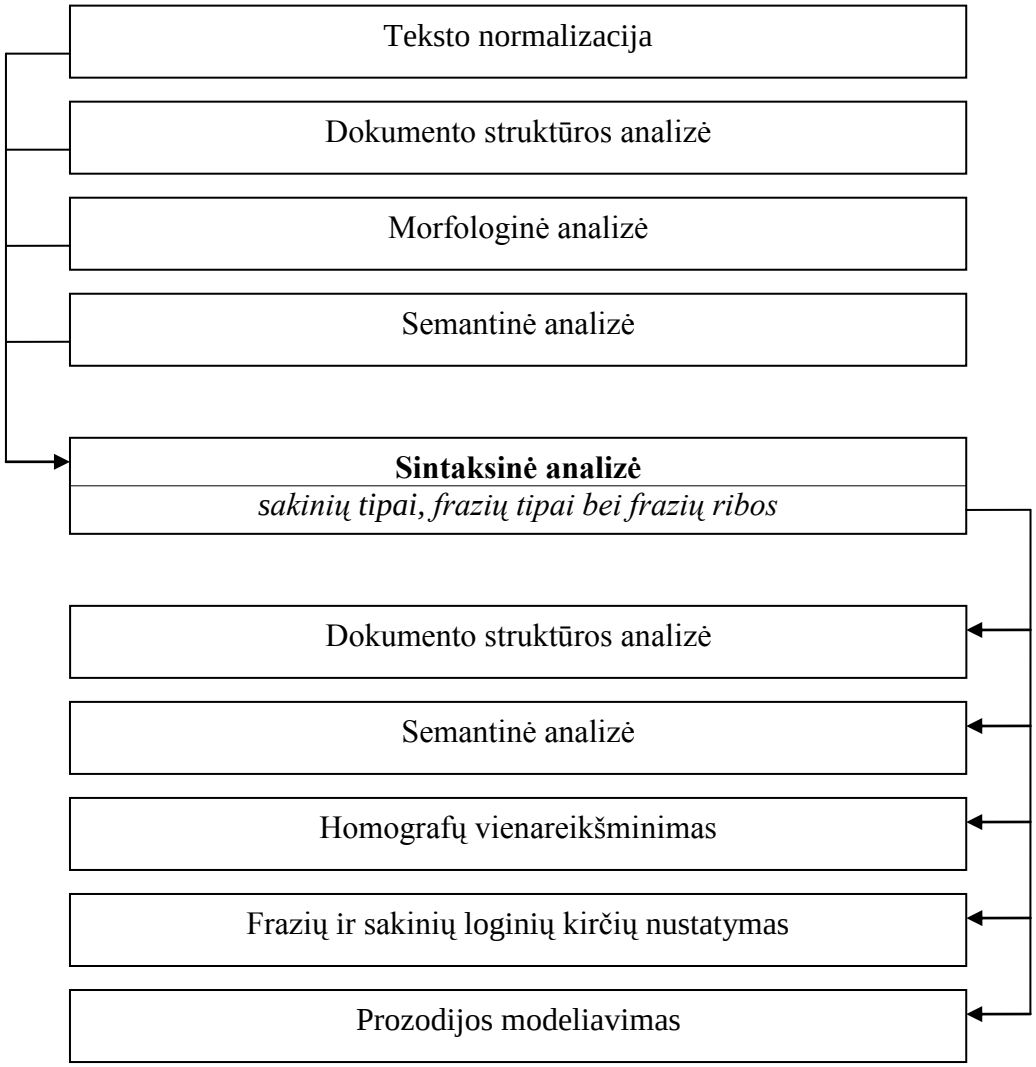
tiesioginį prasminį žodžių ryšį. Dažniausiai tai žodžių formos, tarnybiniai žodžiai (prielinksniai, jungtukai) ir žodžių tvarka sakinyje. Skirtingose kalbose pirmenybė teikiama ne toms pačioms žodžių ryšių raiškos priemonėms. Pavyzdžiui, anglų kalbai svarbiausia žodžių tvarka, po to tarnybiniai žodžiai ir tik paskiausiai žodžių formos. Lietuvių kalboje ryšiai tarp žodžių pirmiausia rodomi žodžių formomis (ypač linksniais), po to tarnybiniais žodžiais ir tik paskutinėje vietoje žodžių tvarka“. Dėl to, kad lietuvių kalbos žodžiai turi daug gramatinių formų, lietuvių kalbos sintaksinei analizei (skirtingai nuo kitų kalbų, pvz., anglų) prasminga naudoti morfologinę analizę. Lietuvių kalbos automatinė sintaksinė analizė nagrinėta [Šveikauskienė, 2009], [Grigonytė, Rimkutė, 2005].

Sintaksinės sakinio struktūros medis dažniausiai naudojamas frazių riboms surasti. Frazių ribos labai svarbios modeliuojant prozodiją. Pasak [Kasparaitis, 2001b, 15], „Frazių ribos atskiriamos pauzėmis, šalia frazių ribos esančių skiemenų pailginimu arba pagrindinio tono dažnio pakėlimu perėjus prie naujos frazės.“ Skaidymui į frazes gali būti naudojami ir paprastesni metodai, pvz., anglų kalbai [Liberman, Church, 1992] pasiūlytas *chinks and chunks* metodas; CART algoritmu grįstas metodas pateiktas [Hirschberg, 1991].

Sakinio tipo nustatymas yra svarbus viršutinio lygio (angl. *macro-level*) sakinio prozodijos modeliavimui [Huang ir kt., 2001, 712]. Frazės ir sakinio tipai taip pat gali padėti nustatyti frazės ir sakinio loginius kirčius.

Sintaksinės analizės funkcijos ryšiai su kitomis TA modulio funkcijomis ir TTS sistemos moduliais yra pavaizduoti 1.6 pav.

Kiti sintaksinės analizės taikymai, remiantis [Šveikauskienė, 2009]: a) automatinis vertimas; b) ištisinės šnekos atpažinimas; c) klaidų suradimas tekstyne.



1.6 pav. Sintaksinė analizė

1.4.5 Semantinė analizė

Remiantis [Huang ir kt., 2001, 57], semantika – lingvistikos šaka, nagrinėjanti (žodžių, sakinių, teksto) prasmę. Kitaip tariant šios funkcijos tikslas – priskirti žodžius semantinėms grupėms (arba žodžiams priskirti semantinius vaidmenis).

Remiantis [Valeckienė, 1998, 35-37], svarbiausi semantiniai vaidmenys:

1) agentas (veiksmo atlikėjas);
2) patientas (veiksmo patyrėjas bei būsenos turėtojas);
3) instrumentas (veiksmo priemonė);
4) finityvas (daikto paskirtis bei tikslas);

5) rezultatas (veiksmo padarinys);
6) lokatyvas[9] (veiksmo vieta);
7) benecifientas (turėtojas, adresatas, kieno naudai atliekamas, kam skirtas veiksmas);
8) percipientas (vidinės psichinės būsenos patyrėjas);
9) eksperientas (išorinės, fizinės būsenos patyrėjas bei suvokėjas);
10) recipientas (ko nors gavėjas, ėmėjas);
11) kontentyvas (veiksmo ar būsenos turinys);
12) proceso apimtis (veiksmo ribojimas, veikimo sfera).

Sintaksinė analizė nusako ryšius tarp žodžių, o semantiniai vaidmenys nusako reikšminius santykius.

Gali būti naudojama ir kitokia semantinė informacija, pvz., požymiai *gyvas*/*negyvas*, *suskaičiuojamas*/*nesuskaičiuojamas*. Tokia semantinė informacija dažnai saugoma žodyne. Kaip tokia informacija gali būti koduojama žodyne, žr. 1.7 pav. [Šveikauskienė, 2009, 31]:

```
vyras = (+žmogus, +vyr.gim., +suaugęs)
moteris = (+žmogus, –vyr.gim., +suaugęs)
berniukas = (+žmogus, +vyr.gim., –suaugęs)
mergaitė = (+žmogus, –vyr.gim., –suaugęs)
```

1.7 pav. Semantinės informacijos kodavimo pavyzdys

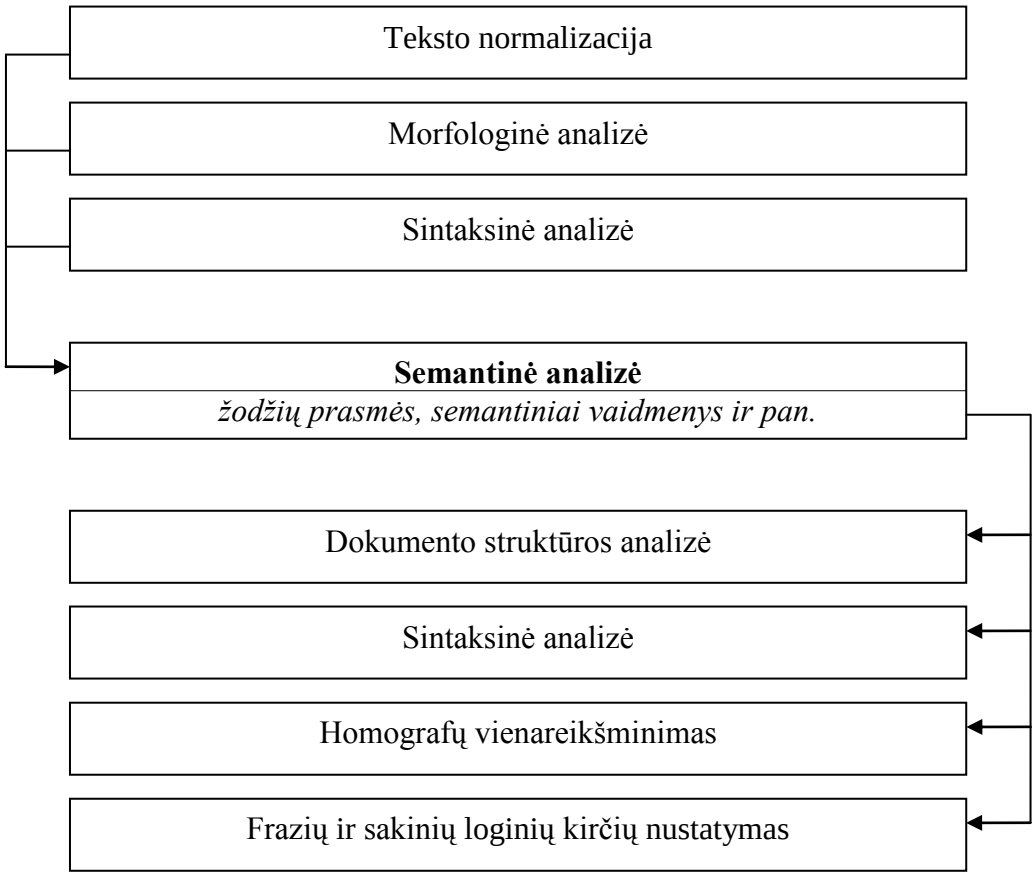
Pasak [Šveikauskienė, 2009, 9], „Semantinė analizė naudojama sintaksinės analizės rezultatams pagerinti. Jos metu žodžiams priskiriami tam tikri reikšmės požymiai. Kai kuriais atvejais, pasitelkiant tuos požymius, galima panaikinti sakino sintaksinės struktūros daugiareikšmiškumą.“ Semantinė analizė taip pat gali pagelbėti vienareikšminant homografus, nustatant frazių ar sakinių loginius kirčius (semantinius akcentus, emfazes) [Huang ir kt., 2001, 710].

[9] [Sližienė, 1994, 23-24].

Automatiniai semantinės analizės metodai dar mažai ištirti. Pasak [Šveikauskienė, 2009, 9], „Lietuvių kalbos semantinės analizės sistema taip pat dar nėra sukurta.“

Semantinė analizė nėra šio darbo tema, todėl jos detaliau nenagrinėsime.

Semantinės analizės funkcijos ryšiai su kitomis TA modulio funkcijomis ir TTS sistemos moduliais yra pavaizduoti 1.8 pav.



1.8 pav. Semantinė analizė

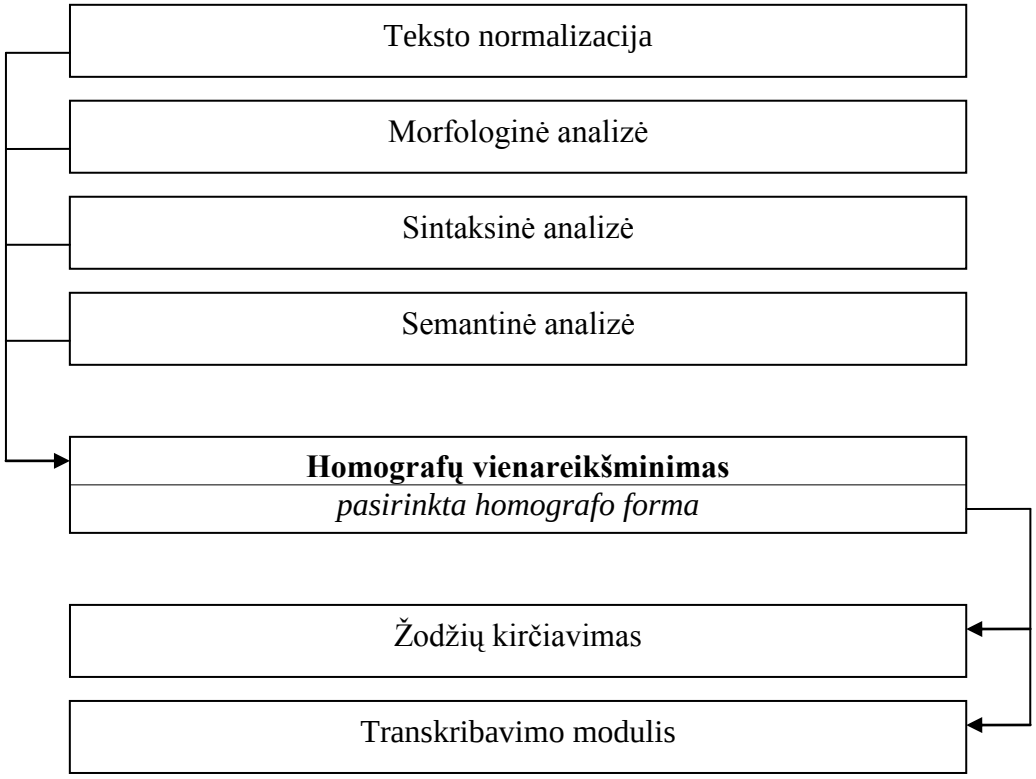
1.4.6 Homografų vienareikšminimas

Kai kurie žodžiai tekste rašomi vienodai, bet tariami (arba kirčiuojami) skirtingai. Pvz., angl. žodis *desert* (tariama /'dezət/ (*dykuma*) arba /dɪ'zə:t/ (*palikti, išvykti*)), *read* (tariama /red/ (*skaitė*) arba /ri:d/ (*skaityti*)) arba liet. *galvos* (skirtingai kirčiuojama: *gálvos* arba *galvõs*). Tokie žodžiai vadinami homografais. TA moduliui reikalingos priemonės vienai homografo tarimo formai pasirinkti. Homografams vienareikšminti gali būti naudojama

morfologinė, semantinė, sintaksinė informacija. Homografų vienareišminimo rezultatai gali būti naudojami žodžio kirčiavimui bei transkribavimui.

Homografų vienareikšminimas plačiau nagrinėtas šio darbo 4 skyriuje.

Homografų vienareikšminimo funkcijos ryšiai su kitomis TA modulio funkcijomis ir TTS sistemos moduliais pavaizduoti 1.9 pav.



1.9 pav. Homografų vienareikšminimas

1.4.7 Žodžių kirčiavimas

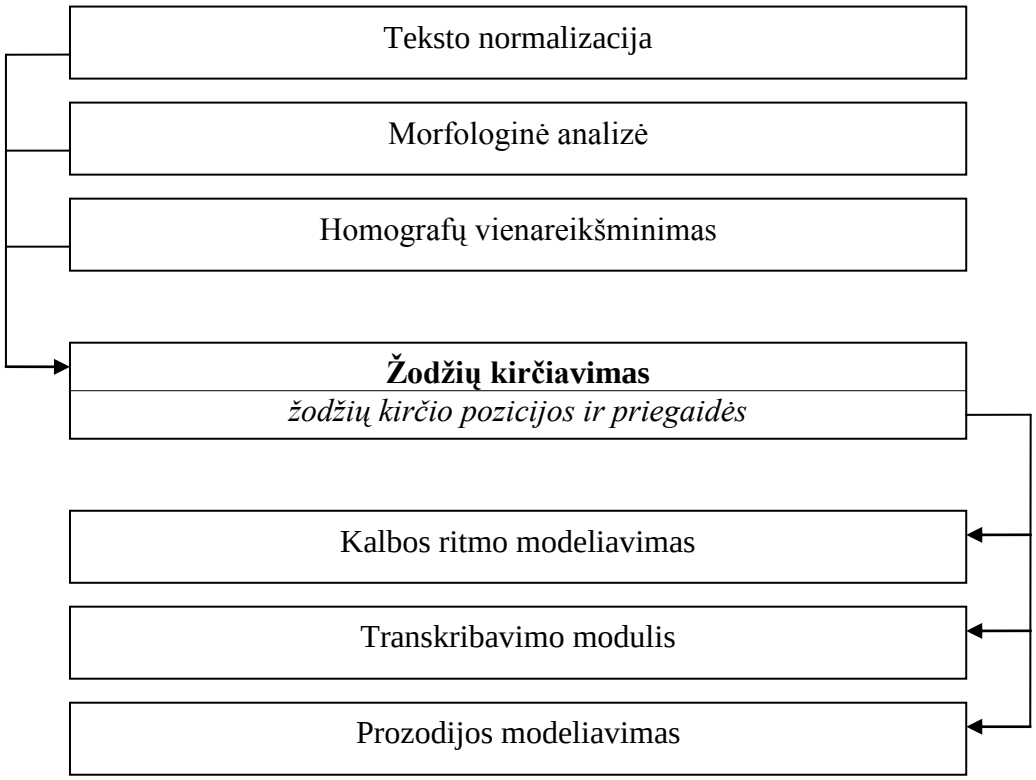
Ne visi žodžio skiemenys tariami vienodai. Vieno skiemens paryškinimas kitų atžvilgiu vadinamas kirčiavimu, o toks kirtis dar vadinamas leksiniu kirčiu [Taylor, 2009, 188]. Leksinis kirtis yra paties žodžio neatskiriama savybė (angl. *core property*) ir nesikeičia. Sąvoka „kirtis“ lingvistinėje literatūroje gali būti vartojama ir kaip sakinio ar frazės kirtis, apie tai žr. 1.4.9 skyrelyje.

Daugiau apie žodžio kirčiavimą (sąvokas, algoritmus ir t. t.) žr. šio darbo 2 skyrių, o apie lietuvių kalbos kirčiavimą naudojant sprendimo medžius žr. 5 skyrių.

Žodžių kirčiavimui gali būti naudojama morfologinė informacija, o vieno kirčiavimo varianto parinkimui – homografų vienareikšminimo informacija.

Žodžių kirčiavimas reikalingas modeliuojant kalbos ritmą, transkribuojant žodžius, modeliuojant prozodiją.

Žodžių kirčiavimo funkcijos ryšiai su kitomis TA modulio funkcijomis ir TTS sistemos moduliais pavaizduoti 1.10 pav.



1.10 pav. Žodžių kirčiavimas

1.4.8 Kalbos ritmo modeliavimas

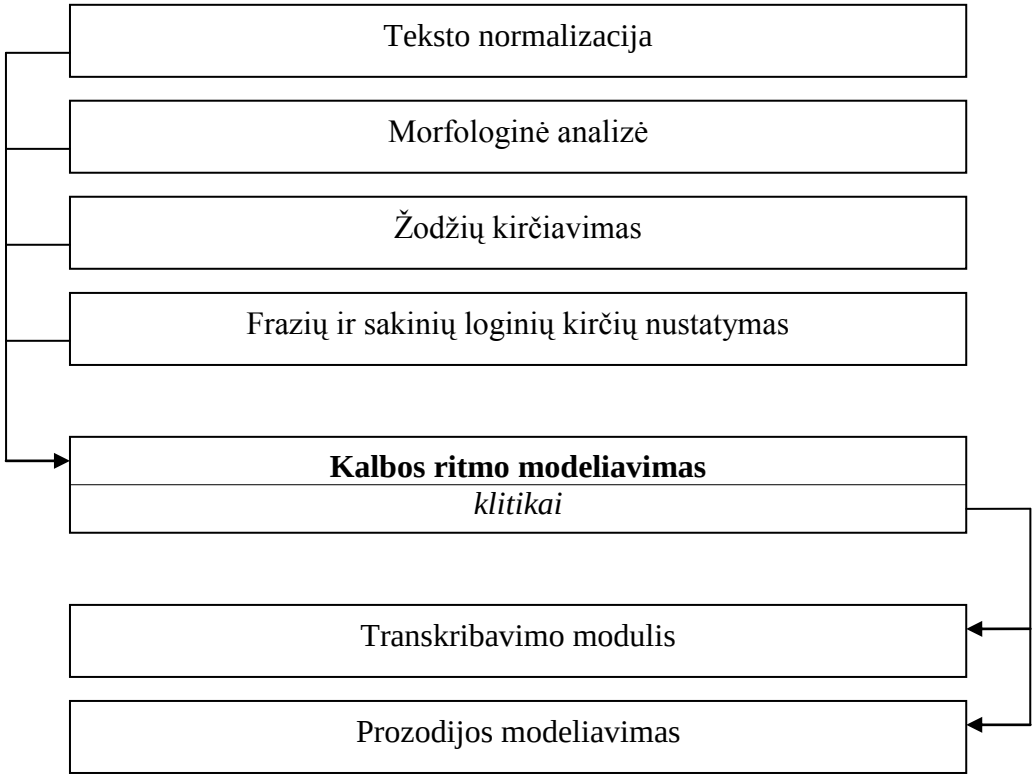
Žmonių kalbai būdingas tam tikras ritmas, kurį sukuria kirčiuotų ir nekirčiuotų skiemenų kaitaliojimasis. Siekiant išlaikyti ritmą kai kurie žodžiai paliekami nekirčiuoti. Tokie bekirčiai žodžiai vadinami klitikais [Ambrazas, 1996, 38].

Klitikų paieškai gali būti naudojama morfologinė informacija, gretimų žodžių kirčiavimas. Žodis, kuris linkęs būti klitiku, yra kirčiuojamas, jei gauna loginį frazės (sakinio) kirtį.

Klitikų paieškos rezultatai kaip kirčiavimo proceso tęsinys gali būti naudojami transkribuojant žodžius ir modeliuojant prozodiją.

Daugiau apie klitikų paiešką žr. šio darbo 6 skyrių.

Kalbos ritmo modeliavimo funkcijos ryšiai su kitomis TA modulio funkcijomis ir TTS sistemos moduliais pavaizduoti 1.11 pav.



1.11 pav. Kalbos ritmo nustatymas

1.4.9 Frazių ir sakinių loginių kirčių nustatymas

Greta žodžių kirčiavimo, pagal veikimo sritį dar galima išskirti frazės kirtį ir sakinio kirtį [Werner, Keller, 1994]. Reiškinys, kai vienas skiemuo frazėje ar sakinyje yra kirčiuojamas labiau negu kiti, vadinamas frazės ar sakinio loginiu kirčiu. Loginį kirtį lemia pragmatinės ar kitos komunikacinės priežastys, jis nėra neatskiriama žodžio savybė [Taylor, 2009, 188], t. y. loginio kirčio vietą lemia ne žodžio, bet frazės ar sakinio savybės.

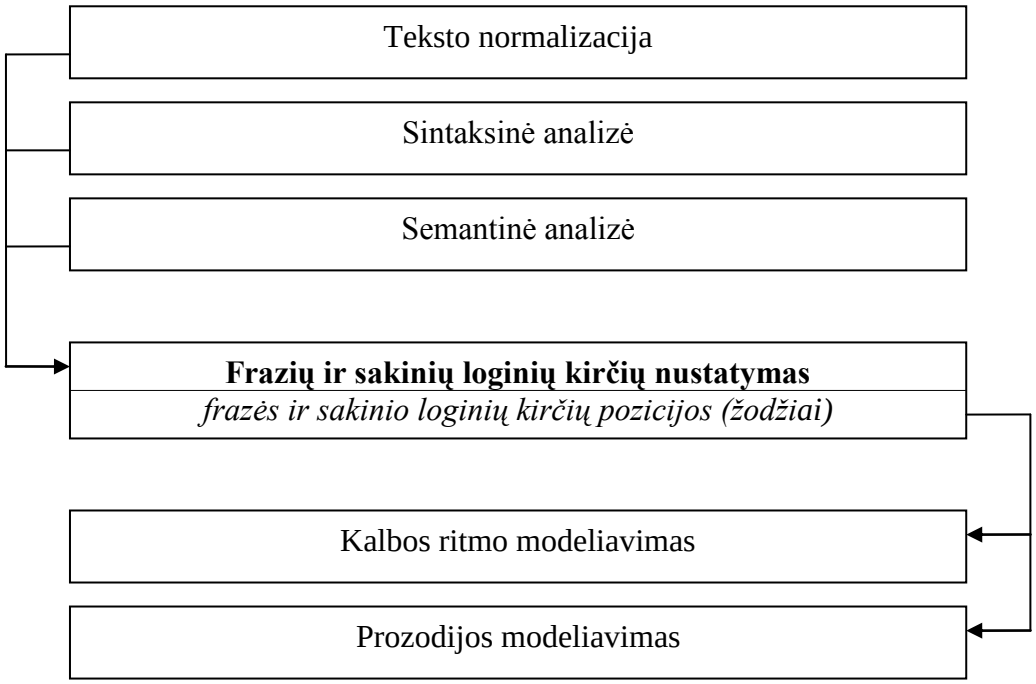
Pasak [Kasparaitis, 2001b, 14-15], „Darbe [Pfister, Traber, 1994] pateiktas sintaksine sakinio struktūra paremtas metodas, kurį naudojant iš pradžių kai kurie žodžiai pažymimi kaip nekirčiuoti, o kitiems suteikiamas

pirminis kirtis. Po to, einant nuo sintaksinio medžio lapų link šaknies, vienų kirčių svoriai padidinami, kitų – sumažinami. Proceso pabaigoje kiekviename sintaksiniame elemente, pvz., daiktavardinėje frazėje, lieka po vieną kirčiuotą žodį.“

Frazės loginio kirčio poveikis prasmei pavaizduotas 1.1 lentelėje.

1.1 lentelė. Frazės kirčiavimo pavyzdys (čia paryškinti žodžiai turintys frazės loginį kirtį)

Frazė	Reikšmė
Aš *atsisakau eiti.*	*Ne „**aš**“, bet gal „**tu**“, „**jis**“ ar „**ji**“ neatsisakys*
*Aš **atsisakau** eiti.*	*Aš „**neisiu**“*
*Aš atsisakau **eiti**.*	*Aš atsisakau „**eiti**“, bet gal „**važiuosiu**“*



1.12 pav. Frazių ir sakinių kirčiavimas

Frazių kirčiavimo funkcijos ryšiai su kitomis TA modulio funkcijomis ir TTS sistemos moduliais yra pavaizduoti 1.12 pav. Iš dokumento struktūros analizės galima sužinoti, kokie žodžiai buvo užrašyti pajuodintu, padidintu arba kitos spalvos šriftu, ir šiems žodžiams suteikti loginį kirtį. Kita vertus, paprastai loginį kirtį gauna tie žodžiai, kurie yra sintaksinio medžio viršuje.

Kartais loginiu kirčiu pabrėžtą žodį gali padėti nustatyti frazės ar sakinio sintaksinis tipas. Semantinė analizė pažymi žodžius, turinčius didesnę loginę reikšmę, jie irgi gauna loginį kirtį. Prozodijos modeliavimo blokas ima iš žodžių kirčiavimo funkcijos duomenis apie kirčiuotus skiemenis ir sustiprina loginį kirtį turinčių žodžių kirčius.

1.5 Transkribavimo modulis

Transkribavimo modulis pagal gaunamas raides bei papildomą informaciją, tokią kaip kirčio vieta, priegaidė ir t. t., grąžina teksto transkripciją (fonemas), dar vadinamas **LTS** (angl. *letter-to-sound*) moduliu. Teksto ir jo transkripcijos pavyzdys pateiktas 1.2 lentelėje (fonemų žymėjimus žr. [Kasparaitis, 2005]).

1.2 lentelė. Tekstas ir jo transkripcija

Tekstas	Po langais žydėjo didelis alyvų krūmas.
Transkripcija	_ p oo l a n g A J s _ Z' ii d' Ee j' oo d' I d' e l' i s a l' iI v uu k r Uu m a s _

Kai kuriose kalbose, pvz., ispanų, suomių, lietuvių, atitinkamybė tarp raidžių ir garsų yra paprasta. Tokios kalbos vadinamos **fonetinėmis** kalbomis. Joms paprastai taikomos raidžių perrašymo į fonemas taisyklės. Apie lietuvių kalbos transkribavimą žr. [Kasparaitis, 1999], [Skripkauskas, Telksnys, 2006].

Kalboms, kuriose priklausomybė tarp raidžių ir garsų sudėtinga, pvz., anglų, paprastai naudojami transkripcijų žodynai. Siekiant sumažinti žodynus, galima saugoti ne žodžių, o morfemų transkripcijas ir iš jų sudaryti žodžių transkripcijas, morfemoms jungti gali reikėti garsų asimiliacijos taisyklių [Huang ir kt., 2001, 714]. Tačiau naudojant žodynus reikia taisyklių naujiems žodžiams, kurių nebuvo žodyne.

Transkripcijų žodyne gali būti saugoma ir kita informacija, pvz., kirčiai, tokiu atveju kirčiavimas atliekamas kartu su transkribavimu.

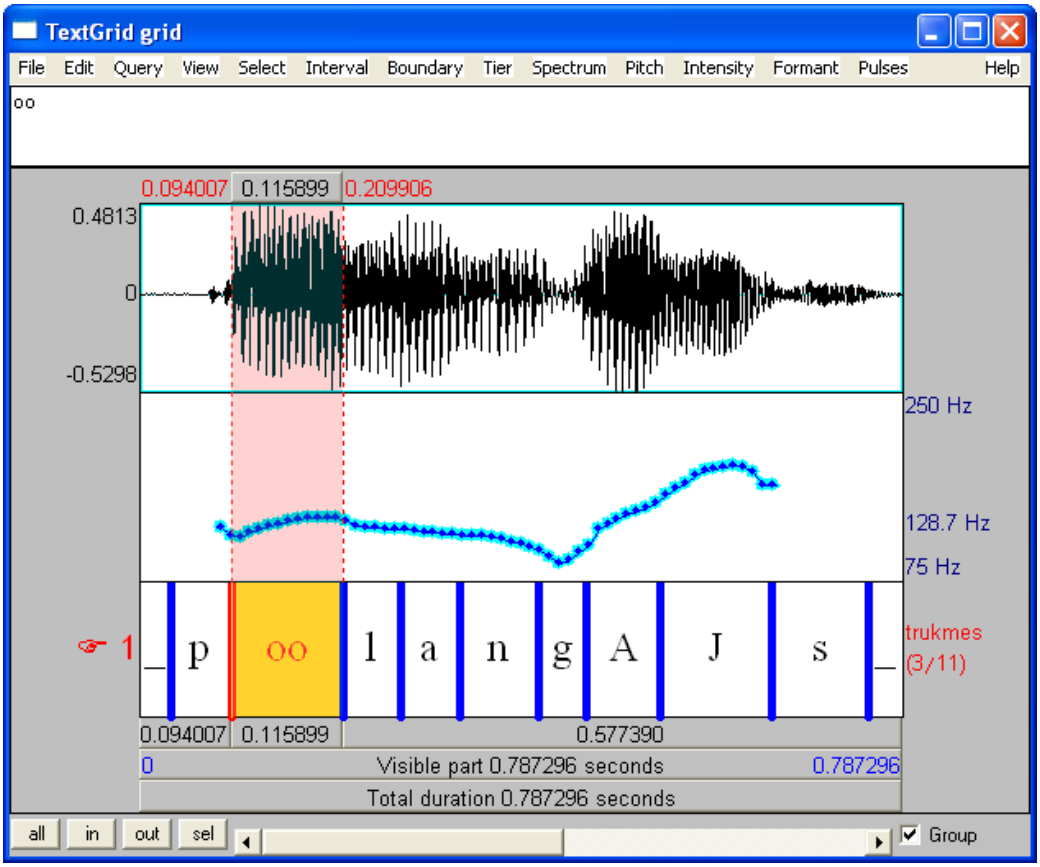
Transkribavimo modulio ryšiai su kitais TTS sistemos moduliais ir TA modulio funkcijomis yra pavaizduoti 1.13 pav.

1.13 pav. Transkribavimas

1.6 Prozodijos modeliavimas

TTS sistemoje prozodijos modeliavimo modulis apima a) segmentų (dažniausiai fonemų) trukmių ir b) pagrindinio tono nustatymą [Huang ir kt., 2001, 731]. Fonemų trukmės nusakomos kaip skaliarinis dydis milisekundėmis (ms), o pagrindinis tonas nusakomas kaip kreivė (laužtė) tam tikruose kontroliniuose taškuose. Fonemų, jų trukmių ir pagrindinio tono pavyzdys pateiktas 1.14 pav. (programos *PRAAT* [www.praat.org] langas).

Trukmės ir pagrindinis tonas apskaičiuojami remiantis tekstinės analizės ir transkribavimo modulių pateikta informacija: žodžio ir frazės kirčiavimu, frazės (sakinio) tipu, žodžio transkripcija ir t. t. Pvz., pagrindinio tono dažnis nuo frazės pradžios link pabaigos mažėja (žr. 1.4.4 skyrelį), o kirčiuotos fonemos trukmė paprastai ilgesnė negu nekirčiuotos (žr. 2.1 skyrelį).



1.14 pav. Signalas, pagrindinio tono kreivė, fonemos ir jų trukmės

Modeliuojant prozodiją iš pradžių apskaičiuojamos segmentų trukmės, po to – pagrindinio tono kreivė. **Segmentų trukmėms** modeliuoti dažniausiai naudojami dvejopi metodai:

1) Taisyklėmis grįsti metodai. Pvz., [Klatt, 1979] pasiūlytas modelis ir jo modifikacijos. Čia segmentų trukmės yra apskaičiuojamos pagal formulę:

```
DUR = MINDUR + (INHDUR - MINDUR) x PRCNT/100
```

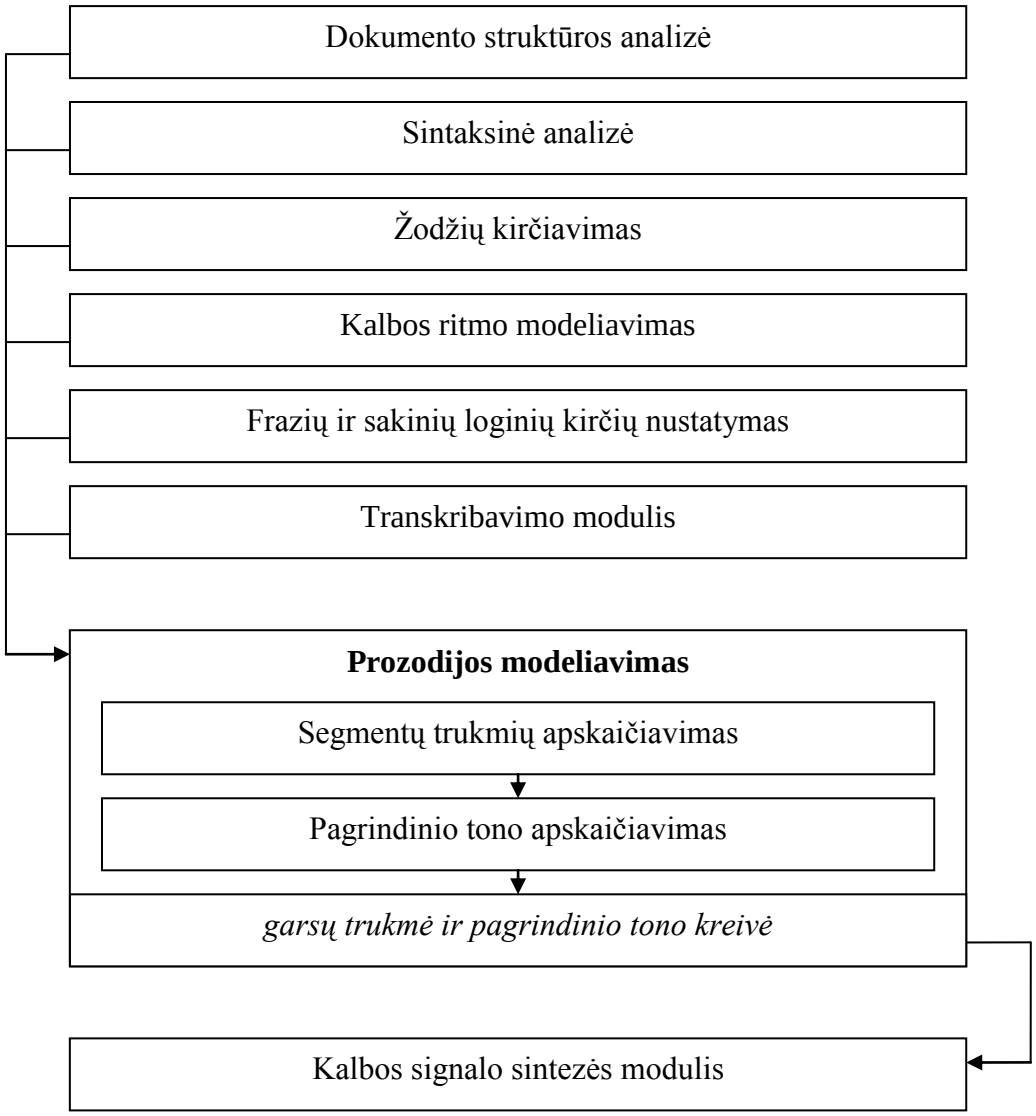
kur DUR – galutinė segmento trukmė, MINDUR – minimali segmento trukmė, INHDUR – segmento „prigimtinė“ trukmė (dažnai tai tiesiog segmento trukmių vidurkis), PRCNT – trukmės pokyčio koeficientas, kuris priklauso nuo konteksto, žodžio ilgio ir pan.

2) Kita metodų grupė – tai iš duomenų apsimokantys algoritmai, pvz., CART (išsamiau apie CART žr. 5.2 skyrelį) arba tiesinė regresija [Plumpe, Meredith, 1998], [van Santen, 1997a]. CART metodas buvo taikytas ir lietuvių kalbai [Norkevičius, Raškinis, 2008].

Pagrindinio tono modeliavimui dažniausiai taikomi:

1) Superpoziciniai modeliai, dar vadinami Fujisaki modeliais [Fujisaki, 1997], [Fujisaki, Sudo, 1971], kuriose tono kreivė gaunama sumuojant kelias skirtingų lygių (frazių, kirčiuotų skiemenų ir pan.) kreives.
2) ToBI (angl. *Tones and Break Indices*) realizacijos modelis [Anderson ir kt., 1984], [Silverman, 1987], kuris aprašo intonaciją kaip aukštų ir žemų tonų seką [Taylor, 2009, 240].

Prozodijos modeliavimo ryšiai su kitais TTS sistemos moduliais ir TA modulio funkcijomis yra pavaizduoti 1.15 pav.



1.15 pav. Prozodijos modeliavimas

1.7 Kalbos signalo sintezės modulis

Kalbos signalo sintezės arba skaitmeninio signalo apdorojimo (**DSP**, angl. *Digital Signal Processing*) [Dutoit, 1993, 39] modulis pagal a) iš transkribavimo modulio gaunamą foneminę reprezentaciją (transkripciją) ir b) iš prozodijos modeliavimo gaunamus prozodijos parametrus, sugeneruoja atitinkamą kalbos signalą (angl. *speech waveform*).

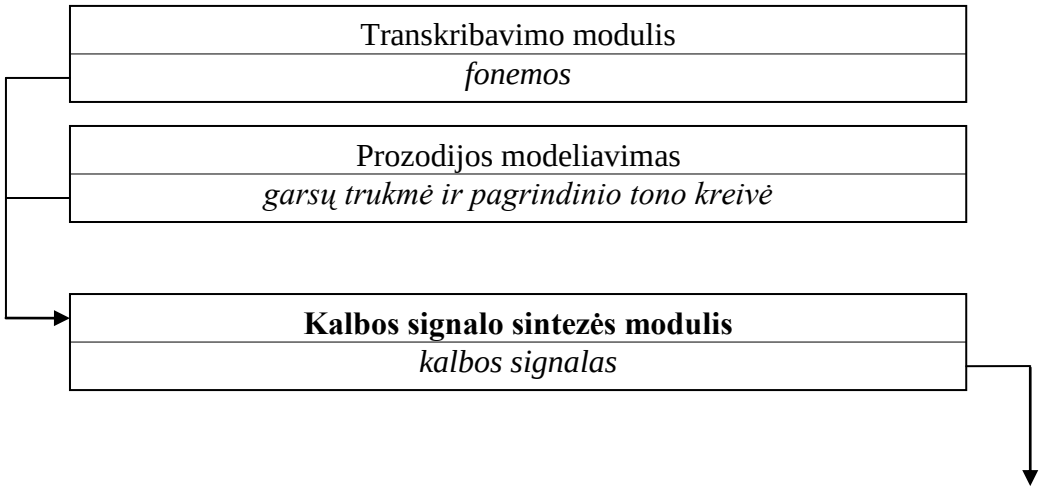
Remiantis [Huang ir kt., 2001, 777], [Pfister, Traber, 1994], [Rahim, 1994], [Syrdal, 1995], [Tatham, Morton, 2005, 23], pagal naudojamą kalbos signalo generavimo modelį kalbos sintezės sistemos klasifikuojamos į 3 pagrindinius tipus:

1) **Artikuliacinė sintezė**, kai modeliuojamas fizinis žmogaus balso sukūrimas naudojant artikuliatorius. Pastaruoju metu šis metodas nėra dažnai naudojamas, ir, nors vieni autoriai, pvz., [Pfister, Traber 1994], abejoja, ar jis yra perspektyvus, kiti autoriai, pvz., [Tatham, Morton, 2005, 23] mano, kad šis sintezės būdas turi didelį potencialą, nes jo pagalba teoriškai galima modeliuoti visus žmogaus balso aspektus.
2) **Formantinė sintezė**, kai modeliuojama balso trakto perdavimo funkcija imituojant formantinius dažnius ir jų amplitudes [Styger, Keller, 1994]. Formantė buvo apibrėžta Fant kaip balso spektro $|P(f)|$ pikas [Fant, 1960]. Šis metodas buvo gana plačiai naudojamas, pvz., sintezatoriai, Klattalk [Klatt, 1980], DECtalk, Infovox, Votrax [Syrdal, 1995], lietuvių kalbos sintezatorius Apollo II [Kasparaitis, 2001b, 70].
3) **Konkatenacinė sintezė**, kai kalba generuojama jungiant kalbos segmentus. Segmentų dydis gali būti labai įvairus: nuo fonemų, alofonų, difonų, pusskiemenių, skiemenų, trigarsių iki ištisų žodžių, frazių ar sakinių. Remiantis [Huang ir kt., 2001, 778], šiuo metu šis metodas yra geriausias pagal sintezuotos kalbos kokybę.

Taip pat, remiantis, pvz., [Dutoit, 1993, 39-40], [Huang ir kt., 2001, 777], galima ir kita metodų klasifikacija pagal rankinio „įsikišimo“ į sistemos kūrimą laipsnį: a) sintezę pagal taisykles (angl. *synthesis by rule*) – artikuliacinė,

formantinė, ir b) duomenimis grįsta sintezė (angl. *data-driven synthesis*) – konkatenacinė.

Kalbos signalo formavimo modulio ryšiai su kitais TTS sistemos moduliais yra pavaizduoti 1.16 pav.



1.16 pav. Kalbos signalo sintezė

1.8 Pirmojo skyriaus apibendrinimas

- Žmogaus ir kompiuterio bendravimas balsu yra natūraliausias būdas. Balso panaudojimas informacijos išvedimui yra patogus, kai vartotojo akys „užimtos“, pvz., vairuojant mašiną ir pan.
- Šnekamosios kalbos apdorojimo sistemas galima suskirstyti į sintezės, atpažinimo ir interpretavimo sistemas.
- Sintezės sistemas sudaro keturi pagrindiniai blokai: teksto apdorojimas, transkribavimas, prozodijos modeliavimas ir kalbos signalo sintezė.
- Šioje disertacijoje nagrinėjami uždaviniai: homografų vienareikšminimas, žodžių kirčiavimas, klitikų paieška, priklauso teksto apdorojimo blokui.
- Homografų vienareikšminimas atliekamas remiantis morfologinės, sintaksinės ir sementinės analizės rezultatais, o jo rezultatai naudojami žodžiams kirčiuoti ir transkribuoti.

- Žodžių kirčiavimas remiasi morfologine analize ir homografų vienereikšminimu, o kirčiavimo informacija naudojama transkribavimui, frazės kirčiams rasti, prozodijai ir ritmui modeliuoti.
- Ritmo modeliavimas (klitikų paieška) remiasi morfologine analize ir žodžių kirčiavimu, o rezultatai naudojami transkribuojant bei modeliuojant prozodiją.

2 Automatinio kirčiavimo algoritmai

Šiame skyriuje apibrėžta kirčio sąvoka, aprašytos įvairiose kalbose esančios kirčiavimo paradigmos, žodžių kirčiavimui taikomi metodai bei jų pasirinkimas priklausomai nuo kirčiavimo paradigmos ir atitinkamos kalbos žodžių kaitymo laipsnio.

2.1 Kirčio sąvoka

Kaip jau buvo minėta 1 skyriuje, žodžio kirčiavimas yra viena iš sudėtinių kalbos sintezės dalių. Kas yra žodžio kirtis? Panagrinėkime kurį nors lietuvių kalbos žodį, pvz., *namas*. Šio žodžio pirmo skiemens balsis tariamas kiek kitaip negu kiti balsiai žodyje, jei tiksliau, jis turi „ryškesnį, aukštesniu tonu ir didesnėmis artikuliacinėmis pastangomis tariamą balsį“ [Girdenis, 2003, 248]. Šis „ryškiau ir blankiau tariamų skiemenų kontrastas“ [Girdenis, 2003, 250] vadinamas **kirčiavimu**, o ryškiau tariamas skiemuo yra kirčiuotas arba turi **kirtį**. Kiti žodžio skiemenys yra nekirčiuoti.

Panagrinėkime kirčio fonetinę realizaciją dviem aspektais [Roach, 1991, 85]:

1) Ką daro kalbėtojas tardamas kirčiuotus skiemenis, t. y. „kirčio sukūrimas“ (angl. *stress production*).
2) Kokios garso charakteristikos iškelia kirčiuotąjį skiemenį kitų skiemenų fone, t. y. „kirčio suvokimas“ (angl. *stress perception*).

Apžvelkim kiekvieną aspektą atskirai:

1) Kaip teigiama [Roach, 1991, 85], „kirčiui sukurti“, t. y. tam, kad būtų ištartas kirčiuotas skiemuo, reikia panaudoti didesnę raumenų energiją, negu tariant nekirčiuotą (pvz., raumenys, naudojami iškvėpti orą iš plaučių, yra aktyvesni tariant kirčiuotą skiemenį, taip sukuriant didesnį subglotalinį slėgį ir pan.). Dar sakoma, kad kirčiuotas skiemuo yra tariamas su didesnėmis artikuliacinėmis pastangomis [Girdenis, 2003, 248].

2) Kaip teigiama, pvz., [Girdenis, 2003, 252], kirčiuotas skiemuo gali būti: a) tariamas tiksliau ir stipriau negu nekirčiuotieji, b) jis gali būti aukštesnis (išimtiniais atvejais – žemesnis, žr. [Allen, 1973]) arba c) ilgesnis už „foną“. Taigi „kirčiui suvokti“ būdingi šie akustiniai prozodiniai požymiai bei jų deriniai: a) balso stiprumas, b) pagrindinio tono aukštumas, c) tarimo trukmė (daugiau žr. [Girdenis, 2003, 252]). Kai kurie autoriai, pvz., [Roach, 1991, 85-86], pirmąjį požymį charakterizuoja tiesiog kaip garsumą. Ten pat [Roach, 1991, 86] tvirtinama, kad svarbesnį vaidmenį turi paskutiniai du požymiai, t. y. pagrindinis tonas ir trukmė. Vienas iš pirmųjų darbų tiriantis kirtį „suvokimo“ aspektu, buvo paskelbtas [Fry, 1958].

Kartais ne visi nekirčiuoti žodžio skiemenys tariami vienodai. Kai kuriose kalbose (pvz., lietuvių, anglų) daugiaskiemeniai sudurtiniai žodžiai (ir ne tik jie), be pagrindinio, gali turėti dar vieną ar net du šalutinius kirčius, pvz., '*dušimtàsis*, *gene*'*ralgubernãtorius*, '*mili*"*ampermètras*. Čia pirmasis šalutinis kirtis žymimas ženklu „'“ prieš kirčiuotą skiemenį, turintį šalutinį kirtį, o antrasis šalutinis kirtis žymimas ženklu „"“ [Vaitkevičiūtė, 1997, 15-18].

Kita vertus, kartais šnekamojoje kalboje kai kurie žodžiai (vadinamieji klitikai, pvz., *ant*, *be*, *dėl*) apskritai netenka kirčio, t. y. prišlyja prie gretimo žodžio. Klitikų paieška tekste gali būti sudėtingas uždavinys, nes tas pats žodis viename kontekste gali turėti kirtį, o kitame jis gali kirčio netekti (daugiau žr. šio darbo 6 skyrių).

Kai kurių kalbų kirčiavimą apsunkina ir papildomi kirčiavimo elementai, vadinami priegaidėmis arba tonais. Priegaidė – tai kirčiuoto skiemens tarimo būdas. Bendrinėje lietuvių kalboje skiriamos dvi ilgųjų skiemenų priegaidės: tvirtapradė ir tvirtagalė. Priegaidė gali skirti panašius žodžius ar jų formas, pvz, *láuk* ir *laũk*.Vienokią ar kitokią priegaidžių sistemą turi latvių, slovėnų ir kitos kalbos, daugiau žr. [Girdenis, 2003, 268-288].

2.2 Kirčiavimo paradigmos

Žodžio kirčiavimo taisyklės priklauso nuo kalbos. Pagal kirčiavimo paradigmą kalbos gali turėti:

1) fiksuotą kirtį,
2) laisvą kirtį.

Fiksuoto kirčio vietą galima nusakyti griežtomis fonetinėmis ir fonologinėmis taisyklėmis [Girdenis, 2003, 256-257], laisvo – ne. Panagrinėkim kiekvieną kirčiavimo modelį atskirai.

Fiksuotą kirtį turi dauguma kalbų [Bolinger, 1978, 481-482]. Kirčiavimo taisykles galima suskirstyti į dvi klases:

1) Visai paprastos taisyklės, nurodančios kirčiuoto skiemens poziciją, skaičiuojant nuo žodžio pradžios ar pabaigos.
2) Taisyklės, kur kirčio vieta priklauso ne tik nuo skiemens pozicijos žodyje, bet, pasak [Girdenis, 2003, 258], „ir nuo balsių bei skiemenų kiekybės“ (pvz., ilgas arba trumpas skiemuo). Taip pat, remiantis [Girdenis, 2003, 258], yra nemažai kalbų, kuriose dažniau kirčiuojamas ilgasis skiemuo.

Teigiama, kad „fiksuoto kirčio tipai yra tokie paplitę todėl, kad jie geriausiai signalizuoja žodžių ribas: pradinis ir galinis kirtis jas rodo tiesiogiai, priešpaskutinio skiemens kirtis funkcionuoja kaip savotiškas įspėjamasis ženklas.“ [Girdenis, 2003, 258].

Pagal skiemens poziciją (pirmoji taisyklių grupė), pasak [Girdenis, 2003, 256-257], „skiriami pagrindiniai trys fiksuoto kirčio modeliai“:

1) Pastoviai kirčiuojamą pirmąjį žodžio skiemenį turi, pvz., latvių, čekų, slovakų, islandų, estų, suomių, vengrų kalbos.
2) Pastoviai kirčiuojamą paskutinį skiemenį turi, pvz., dauguma tiurkų kalbų.
3) Pastoviai kirčiuojamas priešpaskutinis skiemuo. Taip kirčiuojama, pvz., lenkų kalboje.

Pasak [Girdenis, 2003, 256-257], „Visų šių trijų tipų kirčiavimo sistemose gali pasitaikyti žodžių ar tam tikrų jų grupių, turinčių kirtį, pažeidžiantį bendrąją taisyklę.“

Būna ir kitokių kirčiavimo modelių, kurie priklauso tik nuo skiemens pozicijos, pvz., makedoniečių kalboje yra kirčiuojamas trečias nuo galo skiemuo [Marinčič ir kt., 2009].

Antrosios grupės taisyklės dažnai yra analogiškos lotynų kalbos modeliui, pasak [Girdenis, 2003, 258-259], „<...> žodžiuose, turinčiuose daugiau kaip du skiemenis, kirtį gauna priešpaskutinis ilgasis skiemuo (penultima), o kai tas skiemuo būna trumpas, kirčiuojamas trečias bet kokio ilgumo skiemuo (antepenultima). <...> net lietuvių kirtis statistiškai gana stipriai susijęs su skiemens kiekybe: daugeliu atvejų jį gauna ilgasis skiemuo, nors deterministinio dėsnio, siejančio kirtį su kiekybe, ir nėra.“

Kai kuriose kalbose kirtis atlieka ir skiriamąją (distinktyvinę) funkciją, t. y. jis gali diferencijuoti žodžius ir jų formas. Šis distinktyvinis kirtis dar vadinamas **laisvuoju**, pasak [Girdenis, 2003, 255], „Tam tikra prasme galima sakyti, kad laisvasis <...> kirtis yra tipologinė anomalija“. Laisvojo kirčio atveju, „<...> nėra fonetinių nei fonologinių taisyklių, kurios nustatytų, kiek skiemenų gali eiti prieš kirčiuotąjį skiemenį (t. y. žodžio centrą) ir po jo.“ [Girdenis, 2003, 254]. Iš tikrųjų skirtinga kirčio vieta (ar priegaidė) kartais gali būti vienintelis skiriamasis požymis. Pvz., lietuvių kalbos žodžiai *kìlimas* ir *kilìmas* vienodai rašomi ir skiriasi tik kirčiu (homografai). Daugiau apie homografų vienareikšminimo problemą lietuvių kalboje žr. šio darbo 4 skyrių.

Laisvą kirtį turi, pvz., lietuvių, rusų, bulgarų, serbų-kroatų [Girdenis, 2003, 255-256], anglų, slovėnų [Marinčič ir kt., 2009], rumunų [Oancea, Badulescu, 2002] ir kitos kalbos.

Pasak [Girdenis, 2003, 259-260], „Pasitaiko ir tokių kalbų, kurios apskritai neturi pastovesnio kirčio, – tą patį jų žodį viename kontekste gali kirčiuoti vienaip, kitame – kitaip.“ Toks kirtis yra, pvz., gruzinų kalboje.

2.3 Kirčiavimo algoritmai fiksuoto kirčio kalboms

Pabandykime trumpai apžvelgti fiksuoto kirčio kalbų automatinio kirčiavimo algoritmus. Čia ir toliau žodžių automatinį kirčiavimą kompiuteriu vadinsime automatiniu kirčiavimu arba tiesiog kirčiavimu.

Remiantis anksčiau apžvelgtais fiksuoto kirčio kalbų kirčiavimo modeliais, kirčiavimo algoritmą turėtų sudaryti a) šiai kalbai skirtos paprastos kirčiavimo taisyklės ir b) išimčių sąrašas.

Paprasčiausiu atveju kalbos kirčiavimo modelio taisyklės priklauso tik nuo skiemens pozicijos žodyje. Toks paprastas kirčiavimo algoritmas yra taikomas, pvz., lenkų [Shpilewski ir kt., 2004] arba latvių kalbai [Goba, Vasiljevs, 2007]. Tiesa, kirčio pozicijos atžvilgiu paprastą latvių kalbos automatinį kirčiavimą pasunkina priegaidės nustatymas.

Kai kurių fiksuoto kirčio kalbų kirčiavimo modelių taisyklės priklauso ne tik nuo skiemenų ribų, bet ir nuo sunkiau identifikuojamų parametrų, pvz., skiemens kiekybės (ilgasis ar trumpasis). Šie parametrai gali būti svarbūs ir laisvojo kirčio kalboms. Parametrams identifikuoti svarbu, kokius įvedimo duomenis naudoja kirčiavimo algoritmas. Galimi mažiausiai du skirtingi kirčiavimo algoritmų įvedimo duomenų tipai:

1) **Fonemos**. Čia ilgi ir trumpi (t. y. skirtingos kiekybės) balsiai yra skirtingos fonemos, todėl balsio kiekybės nustatymas nesukelia problemų. Kirčiavimo algoritmo, kuris naudoja skiemenų kiekybės požymius ir fonemas kaip įvedimo duomenis, pavyzdį žr. [McPeters, Tharp, 1983]. Čia reikėtų paminėti, kad foneminė žodžio reprezentacija gali būti pasiekiama kirčiavimo algoritmui, pvz., jei kirčiavimo algoritmas ir transkribavimo modulis priklauso vienai kompiuterinei balso sintezės sistemai.
2) **Raidės**. Šiuo atveju nustatyti skiemens kiekybę gali būti sudėtinga. Rašytinėje kalboje ta pati raidė gali žymėti ir trumpą, ir ilgą balsį, pvz., žodyje *namas* pirmas *a* ilgas, antras – trumpas. Kirčiavimo algoritmo, kuris naudoja skiemenų kiekybės požymius ir raides kaip įvedimo

duomenis, pavyzdį žr. [Church, 1985]. Šiame darbe nagrinėjamas kirčiavimo nustatymas iš rašytinės formos.

Be to, kai kurių fiksuoto kirčio kalbų išimčių sąrašą gali sudaryti didelės žodžių grupės, todėl gali prireikti naudoti išimčių taisykles, o ne sąrašą, ir šios taisyklės gali dar reikalauti ir papildomos informacijos, gaunamos morfologinės analizės metu, pvz.:

1) Čekų kalboje, kur kirčiuojamas pirmas skiemuo, prielinksnis dažniausiai „<...> vertinamas kaip pirmasis fonologinio žodžio skiemuo ir todėl atitraukia į save kirtį“ [Girdenis, 2003, 256]. Todėl reikia nustatyti, ar žodis yra prielinksnis (t. y. atpažinti kalbos dalį), jį sukirčiuoti, o po jo einančio žodžio nekirčiuoti. Palyginimui, latvių kalboje, kur irgi kirčiuojamas pirmasis žodžio skiemuo, prielinksnis lieka nekirčiuotas [Girdenis, 2003, 256].
2) Lenkų kalboje dažniausiai yra kirčiuojamas priešpaskutinis žodžio skiemuo, tačiau, kaip teigiama [Bąk, 1995], pirmojo ir antrojo asmens daugiskaitos būtojo laiko veiksmažodžiuose dažniausiai yra kirčiuojamas trečias nuo galo skiemuo. Taigi šiuo atveju reikės nustatyti, ar žodis yra veiksmažodis, jei taip – jo asmenuotę ir laiką, nors [Oliver, Grice, 2003] teigiama, kad ir šiose veiksmažodžių formose, ypač šnekamojoje kalboje, yra polinkis perkelti kirtį pagal bendrą taisyklę į priešpaskutinį skiemenį.

2.4 Fleksinės ir nefleksinės kalbos

Nors, kaip matėme, kai kurių fiksuotojo kirčio kalbų kirčiavimas gali būti netrivialus uždavinys, tačiau laisvojo kirčio kalbų kirčiavimas paprastai būna daug sudėtingesnis.

Laisvojo kirčio atveju naudojami automatinio kirčiavimo metodai gali priklausyti nuo to, ar kalba yra fleksinė, ar nefleksinė. **Nefleksinėmis** kalbomis vadinsime nekaitomas arba silpnai kaitomas kalbas, pvz., anglų. Šių kalbų žodžiai turi mažai gramatinių formų.

Fleksinėmis kalbomis vadinsime kalbas su išplėtota gramatinių formų sistema, pvz., lietuvių, rusų. Šių kalbų žodžiai turi skirtingas formas

priklausomai nuo giminės, skaičiaus, linksnio, laipsnio, nuosakos, laiko, asmens ir t. t. Skirtingų to paties žodžio formų kirčio pozicija gali skirtis. Pvz., lietuvių kalbos daiktavardis *kalba* turi po 7 vienaskaitos ir daugiskaitos formas, iš kurių 5 kirtį turi šaknyje, o likusios 9 – galūnėje.

2.5 Laisvojo kirčio kalboms naudojami kirčiavimo metodai

Laisvojo kirčio kalbų automatinio kirčiavimo metodai būna grįsti:

- žodynais (pilnų žodžių, morfemų ar kitų žodžio dalių);
- taisyklėmis (čia galima priskirti daugumos fiksuoto kirčio kalbų metodus).

Žodynais grįsti metodai paprastai naudoja taisykles išimtims aprašyti, kitaip tariant galima sakyti, kad „žodynas yra taisyklė, o taisyklės – išimtys“ [Dutoit, 1993, 201]. Kita vertus, taisyklėmis grįsti metodai saugo išimtis žodyne. Norint sukirčiuoti visus teksto žodžius naudojant taisyklėmis grįstus metodus, teoriškai žodyno galima visai nenaudoti (t. y. nenaudoti išimčių), tokių metodų pavyzdžiai pateikti [Bernstein, Nessly, 1981], [McPeters, Tharp, 1983]. Tačiau remiantis žodynais grįstais metodais ir norint sukirčiuoti visus žodžius, taisyklės yra būtinos, nes joks žodynas negali apimti visų kalbos žodžių (daugiau žr. 2.5.1 skyrelį).

Analogiškai daugelyje darbų, pvz., [Dutoit, 1993, 201-202], [Pfister, Traber, 1994], [Syrdal, 1995] yra skirstomi ir transkribavimo metodai, tai dar kartą pabrėžia automatinio transkribavimo ir kirčiavimo metodų ryšį (žr. 1.5 skyrelį).

Žodynais grįsti metodai savo ruožtu gali būti skirstomi pagal tai, iš ko žodynai sudaryti:

1) iš pilnų žodžių,
2) iš morfemų ar kitų žodžių dalių.

Kita kirčiavimo metodų grupė – tai taisyklėmis grįsti metodai. Tokie metodai dažniausiai remiasi raidžių seka žodyje, skiemenų kiekybe ir struktūra (uždaras ar atviras), skiemenų skaičiumi ir t. t. Kartais taisyklės remiasi informacija apie kalbos dalį.

Pagal taisyklių sudarymo būdą šios grupės metodus galima skaidyti į dvi grupes:

1) ekspertų lingvistų arba pagal lingvistikos vadovėlius sudarytos taisyklės,
2) automatiškai iš duomenų generuotos taisyklės, naudojant ANN, CART ir kitus save mokančius algoritmus.

Čia pateikta metodų klasifikacija nėra griežta, pvz.:

- nors taisyklėmis grįsti metodai paprastai nenaudoja jokių žodynų, tačiau žodynai gali būti naudojami taisyklėms sudaryti,
- automatiškai generuojant taisykles iš žodynų ar tekstynų, ekspertų lingvistų įsikišimas paprastai nereikalingas, tačiau tekstynai ir žodynai dažnai yra sudaromi ir peržiūrimi ekspertų lingvistų,
- taikant taisyklėmis grįstus metodus, gali būti naudojami tam tikri specifiniai žodynai, pvz., priesagų, priešdėlių sąrašai,
- morfemų žodynus naudojantys metodai gali reikalauti žodžio kaitymo ir kirčiavimo taisyklių, pvz., [Kasparaitis, 2001a], [Kazlauskienė, Raškinis, 2004], [Kazlauskienė ir kt., 2004], [Norkevičius ir kt., 2004],
- taisyklėmis grįsti kirčiavimo algoritmai gali reikalauti informacijos apie žodžio kalbos dalį, o ši informacija gali būti gaunama iš žodynų,
- galimos hibridinės sistemos, kur iš pradžių naudojami žodynais grįsti metodai, o likusiems žodžiams, kurių nebuvo žodynuose – taisyklės [Church, 1985], [Church, 1986].

Pradžioje dažniau buvo naudojami ekspertų sudarytomis taisyklėmis grįsti metodai [Hunnicutt, 1976], [Bernstein, Nessly, 1981], [McPeters, Tharp, 1983], [Church, 1985]. Taip atsitiko todėl, kad šie metodai reikalauja mažiausiai kompiuterio atminties ir skaičiavimo resursų. Kita vertus, jau 1987 metais aprašytas metodas *NETtalk* [Sejnowski, Rosenberg, 1987] naudojo ANN taisyklėms automatiškai sudaryti, o vienas iš pirmųjų anglų kalbos sintezatorių *MITalk* [Allen ir kt., 1979], [Allen ir kt., 1987] naudojo morfemų žodyną (apie 10000 morfemų) kirčiavimui ir transkribavimui. Dabar dažniausiai naudojami a) morfemų žodynais grįsti metodai (žr. 2.5.2 skyrelį) ir b) automatiškai iš duomenų sudarytos taisyklės (žr. 2.5.4 skyrelį).

Toliau panagrinėkime atskirai kiekvieną iš šių metodų grupių.

2.5.1 Žodžių žodynai

Jei kirčiuojama kalba yra nefleksinė, galima bandyti tiesiog sudaryti visų jos žodžių žodyną ir kiekvienam žodžiui tame pačiame žodyne saugoti informaciją apie kirtį. Tada kirčiavimo algoritmas – tiesiog žodžio paieška žodyne. Jei kirčiuojamas žodis yra surandamas žodyne, jam yra priskiriamas saugomas šiam žodžiui kirtis, jei tokio žodžio žodyne nėra – žodis lieka nekirčiuotas, ir jį galima a) tokį ir palikti arba b) taikyti papildomas taisykles jam sukirčiuoti.

Iškyla klausimas, kokio dydžio turėtų būti žodynas, norint padengti kuo didesnę „standartinio“ teksto dalį. Pvz., [Church, 1985] teigiama, kad anglų kalbos 50000 pagrindinių žodžių žodynas padengia apie 93 proc. standartinio laikraščio teksto[10]. Norint padidinti padengimo koeficientą, reikėtų papildyti žodyną techniniais, regioniniais, medicininiais, moksliniais ir kitokiais terminais, taip pat skoliniais (pvz., lotynų kalbos žodžiais, vartojamais medicinoje, prancūzų kalbos žodžiais, vartojamais kulinarijoje, japonų kovos menų žodžiais), vardais ir pavardėmis, vietovardžiais, žodžių junginiais bei įvairioms tarmėms ir žargonui (pvz., jaunimo, kompiuteriniam) priklausančiais žodžiais. Tačiau kalboje visą laiką atsiranda naujų žodžių, todėl norint išlaikyti didesnį padengimo koeficientą, žodyną reikia nuolat pildyti ir atnaujinti. Be to, tekste gali pasitaikyti neegzistuojančių žodžių (pvz., „*abrakadabra*“, po skiemenį užrašytų žodžių, pvz., *ki–li–mas*), todėl neįmanoma padengti 100 proc. teksto.

Atskirai verta atkreipti dėmesį į daug mažesnį tikrinių daiktavardžių padengimo koeficientą. 2.1 lentelėje pateiktas atsakymas į klausimą, kokią

[10] Remiantis [http://www.askoxford.com/asktheexperts/faq/aboutenglish/?view=uk, žiūrėta 2010.04.01], anglų kalba turi didžiausią žodžių skaičių tarp pasaulio kalbų. Yra apie ¼ milijono skirtingų angliškų žodžių, neįskaičiuojant žodžių formų. Daugiau nei pusė šių žodžių yra daiktavardžiai, apie ketvirtadalį būdvardžiai, apie septintadalį veiksmažodžiai; likusi dalis sudaryta iš jaustukų, jungtukų, prielinksnių ir t. t. Jei skaičiuotume skirtingas žodžio reikšmes, tai iš viso turėtų būti apie ¾ milijono žodžių.

tekstyno dalį padengia *n* dažniausių žodžių žodynas, sudarytas iš to paties tekstyno.

2.1 lentelė. Tekstynų padengimas skirtingo dydžio žodynais

Žodyno dydis (žodžiai)	*Brown Corpus* (proc.)	*Kansas City Telephone Book* (proc.)
2000	68	46
4000	78	57
6000	83	63
8000	86	68
10000	89	72
12000	91	75
14000	92	77
16000	94	79
18000	95	81
20000	95	83
22000	96	84
24000	97	86
26000	97	87
28000	98	88
30000	98	89
32000	98	90
34000	99	91
36000	99	91
38000	99	92
40000	99	93

Yra lyginami du skirtingi tekstynai: *Brown Corpus* (*bendros paskirties anglų kalbos tekstynas*) ir *Kansas City Telephone Book* (*Kanzaso miesto telefonų knyga*). Pirma eilutė rodo, kad 2000 žodžių žodynas padengia 68 proc. *Brown Corpus* tekstyno, o 2000 pavardžių tik 46 proc. *Kansas City Telephone Book* tekstyno. Iš lentelės turėtų būti aišku, kad pavardžių žodynas turėtų būti daug didesnis. Dar blogesnė situacija, jei žodyną, sudarytą iš jau minėto *Kansas City Telephone Book* tekstyno, naudojam kitai telefonų knygai *Bell Labs Phone Book* (*Bell Labs telefonų knyga*) padengti. Maksimali padengimo reikšmė yra apie 60 proc. ir verta naudoti tik 5000-10000 dažniausių pavardžių, nes toliau žodyno didinimas beveik nepadidina tekstyno padengimo (duomenys paimti iš [Church, 1985]).

Nors anglų kalba yra nefleksinė, tačiau ir joje žodis visgi gali turėti skirtingų formų, pvz., anglų kalbos žodis *dog* (*šuo*) turi ir daugiskaitos formą *dogs* (*šunys*). Akivaizdu, kad neverta saugoti žodyne abi šio žodžio formas ir tikslinga turėti bent paprasčiausias žodžių kaitymo taisykles.

Daug sudėtingesnis yra fleksinių kalbų žodyno kūrimas. Fleksinės kalbos dažniausiai neturi didelių duomenų bazių, kurios nurodytų visų žodžių formų tarimą [Sproat, 1997]. Pvz., fleksinės lietuvių kalbos žodynas dėl gramatinių formų gausos turėtų būti žymiai didesnis nei anglų [Kasparaitis, 2000], todėl fleksinei kalbai daugiau tiktų morfemų ar kitų žodžio dalių žodynus naudojantys metodai.

Nors kompiuteriuose jau galima saugoti labai didelius žodynus, žodyno sumažinimas yra vis dar aktualus, pvz., portatyviniuose įrenginiuose. Vienas iš žodyno mažinimo būdų – pakeisti pilnų žodžių žodynus morfemų žodynais.

2.5.2 Morfemų ar kitų žodžių dalių žodynai

Šios grupės metodai taip pat remiasi žodynu, bet jame jau yra saugomi ne pilni žodžiai, bet morfemos (ar kitos žodžio dalys) – dėl to, norint padengti tą pačią teksto dalį, morfemų žodynas turėtų būti mažesnis negu pilnų žodžių. Kartais naudojamas ne morfemų, bet kamienų žodynas [Kasparaitis, 2000], [Kasparaitis, 2001a], [Kasparaitis, 2001b] arba žodžių pagrindinių formų žodynas [Kazlauskienė, Raškinis, 2004], [Kazlauskienė ir kt., 2004], [Norkevičius ir kt., 2004]. Visi šie metodai, naudojantys žodžių dalių, žodžių pagrindinių formų ar morfemų žodynus, gali būti sujungti į vieną grupę (toliau vadinsime tiesiog „morfemų žodynais grįsti metodai“).

Šios grupės metodai paprastai turi:

1) sukirčiuotų morfemų ar kitų žodžio dalių žodynus, kuriuos gali sudaryti žodžių pagrindinės formos, morfemos (priešdėliai, šaknys, prielinksniai, galūnės), kamienai ar kitos žodžių dalys,
2) taisykles, kaip iš morfemų sudaryti žodį (t. y. morfologines žodžio kaitymo taisykles, kurias jau trumpai apžvelgėme 1.4.3. skyrelyje),

3) taisykles kirčiui nustatyti; kirčiavimo taisyklių skaičius, sudėtingumas ir saugojimo būdai gali labai skirtis: nuo kelių paprastų taisyklių [Allen ir kt., 1979], [Church, 1986] iki didelės taisyklių aibės, saugomos medyje [Kazlauskienė, Raškinis, 2004], [Kazlauskienė ir kt., 2004], [Norkevičius ir kt., 2004].

Vienas iš pirmųjų kalbos sintezatorių *MITalk* [Allen ir kt., 1979], [Allen ir kt., 1987] priklauso šiai metodų grupei ir naudojo (duomenys pagal [Allen ir kt., 1979], [Dutoit, 1993, 201], [Church, 1986]):

1) maždaug 10000 anglų kalbos morfemų su kirčiais žodyną, kuris padengia 95 proc. įvesties žodžių, likę žodžiai kirčiuojami pagal taisykles,
2) paprastą morfemų jungimą su keliomis paprastomis taisyklėmis, nurodančiomis morfemų sandūrose vykstančius pakitimus:
   a) tylusis *e* panaikinamas prieš vokalizuotą sufiksą, pvz., angl. *observe* + *ance* = *observance*,
   b) priebalsis dubliuojamas prieš vokalizuotą sufiksą, pvz., angl. *red* + *est* = *reddest*,
   c) *y* pakeičiamas *i* prieš sufiksą, pvz., angl. *glory* + *ous* = *glorious*,
   d) *y* panaikinamas prieš sufiksą prasidedantį *i*, pvz., angl. *harmony* + *ize* = *harmonize*,
3) kirčiavimui – neutraliųjų afiksų principą; kirčio atžvilgiu neutralaus afikso (angl. *stress neutral affixes*) sąvoka remiasi [Chomsky, Halle, 1968], ir reiškia, kad afiksas nepaveikia kamieno kirčio, pvz., angl. *mother* (*motina*) ir *motherhood* (*motinystė*). Žodžiai, pvz., angl. *festivity* (*iškilmės*) ir *divinity* (*dievybė*), yra įvedami į žodyną kaip išimtys.

Vėliau [Church, 1986], naudodamas jau minėto *MITalk* morfologinės dekompozicijos modulį, padalino visus afiksus į 1-ojo ir 2-ojo lygio afiksus (terminologija paimta iš [Mohanan, 1982]), kur 1-ojo lygio afiksai pritraukia kirtį (pvz., prefiksas *pre-*, sufiksas *-ity*), o 2-ojo lygio (pvz., prefiksas *ultra-*, sufiksas *-hood*) nepritraukia (vadinami kirčio atžvilgiu neutraliais afiksais). Daugiau žr. [Church, 1986].

Visi žodynais grįsti metodai nagrinėja tik žodį atskirai, todėl homografų vienareikšminimo problema lieka neišspręsta. Be to, jei žodis ar jo dalys nerandamos žodyne, žodis lieka nekirčiuotas. Morfemų žodynai turi didesnį padengimo koeficientą, negu žodžių žodynai, pvz., [Kasparaitis, 2001b, 47] aprašytas metodas neranda 3,63 proc. žodžių.

2.5.3 Ekspertų sudarytos taisyklės

Šios grupės metodai [Hunnicutt, 1976], [Bernstein, Nessly, 1981], [McPeters, Tharp, 1983], [Church, 1985] remiasi įvairių kalbotyros sričių tyrimais ir taisyklėmis. Pradedant nuo plačiai žinomo darbo *Anglų kalbos garso šablonas* (angl. *Sound Pattern of English*, dažnai žymimas *SPE*) [Chomsky, Halle, 1968], vėliau buvo atlikta ir gausybė kitų fonologinių tyrimų [Liberman, Prince, 1977], [Hayes, 1980], kur buvo apibrėžtos deterministinės taisyklės kirčiavimui nustatyti naudojant „skiemenų svorius“ ir ekstrametrinių (t. y. kirčio nepritraukiančių) skiemenų skaičių.

Remiantis [McPeters, Tharp, 1983], [Church, 1985], skiemens svoris – tai binarinis parametras, kuris gali turėti dvi reikšmes: sunkusis ir lengvasis. Sunkusis skiemuo – jei baigiasi ilguoju balsiu arba dviem ar daugiau priebalsių; lengvasis – jei trumpuoju balsiu ir daugiausiai vienu priebalsiu po jo. Sunkusis skiemuo yra kirčiuojamas dažniau negu lengvasis, nors galutinis kirčiavimo rezultatas priklauso nuo skiemens konteksto (t. y. nuo kaimyninių skiemenų svorių). Pagrindinė veiksmažodžių kirčiavimo taisyklė: kirčiuojamas paskutinis skiemuo, jei jis yra sunkusis, pvz., angl. *obey* (*paklusti*), ir priešpaskutinis skiemuo – jei paskutinis skiemuo yra lengvasis, pvz., angl. *develop* (*vystyti*, *sukurti*). Panašiai kirčiuojami ir daiktavardžiai, išskyrus tai, kad paskutinis skiemuo yra ignoruojamas (dar sakoma, kad paskutinis skiemuo yra ekstrametrinis, žr. [Hayes, 1980]). Būdvardžiai kirčiuojami kaip veiksmažodžiai, jei neturi sufikso, ir kaip daiktavardžiai, jei sufiksą turi (kitaip sakant, sufiksai yra ekstrametriniai). [Liberman, Prince, 1977] pasiūlytos kirčiavimo taisyklės yra realizuotos, pvz., [McPeters, Tharp, 1983]. Apie išimtis iš šių taisyklių žr. [Liberman, Prince, 1977].

Iš pateiktų taisyklių bei apibrėžimų matome, kad jų kompiuterinis realizavimas susiduria su balsio kiekybės, kalbos dalies, būdvardžių sufikso nustatymo problemomis.

Jau nuo pat pirmųjų kompiuterinių realizacijų buvo bandoma adaptuoti minėtus lingvistinius metodus siekiant atsisakyti kai kurių žodžių, skiemenų, garsų požymių. Pvz., [Hunnicutt, 1976], [McPeters, Tharp, 1983] adaptavo minėtas taisykles taip, kad daugumą taisyklių galima taikyti nežinant kalbos dalies. [McPeters, Tharp, 1983] kirčiavimas nustatomas pagal žodžio foneminę reprezentaciją, taip išsprendžiant balsio kiekybės, kartu ir skiemens svorio nustatymo problemą.

[Bernstein, Nessly, 1981] remiasi supaprastintu [Hill, Nessly, 1973] pasiūlytu metodu. Skiemens svorio sąvoka yra pakeičiama ir naudojamas tik uždarojo/atvirojo skiemens požymis. Analogiškai minėtuose metoduose atsisakyta ir kalbos dalies požymio (tiesa, kirčiuojami tik veiksmažodžiai ir daiktavardžiai). Dėl šios taisyklės paprastumo ir stebėtinai patenkinamo rezultato – remiantis [Bernstein, Nessly, 1981], algoritmo tikslumas yra net 75 proc. – toliau pateiktas jos pseudokodas [Bernstein, Nessly, 1981]:

```
Jei (skiemenų skaičius žodyje = 1)
         Sukirčiuoti šį skiemenį
Jei (skiemenų skaičius žodyje = 2)
         Sukirčiuoti pirmą skiemenį
Priešingu atveju
         Jei (priešpaskutinis skiemuo yra uždaras)
                  Sukirčiuoti šį skiemenį
         Priešingu atveju
                  Sukirčiuoti trečia nuo galo skiemenį
         (uždėti šalutinius kirčius skiemenims iš kairės)
```

Buvo siūlomos ir kitos, sudėtingesnės minėtų lingvistinių taisyklių adaptacijos. Pvz., [Church, 1985] siūloma naudoti:

1) pseudosvorius,
2) skirtingas kirčiavimo taisykles skirtingoms kalbos dalims,
3) morfologinę analizę – kai kuriems afiksams yra saugomi svoriai,

4) žodžio etimologiją (informaciją apie žodžio kilmę), kuri nustatoma remiantis trijų raidžių sekomis (trigramais), ir pagal žodžio etimologiją naudojami skirtingi kirčiavimo modeliai.

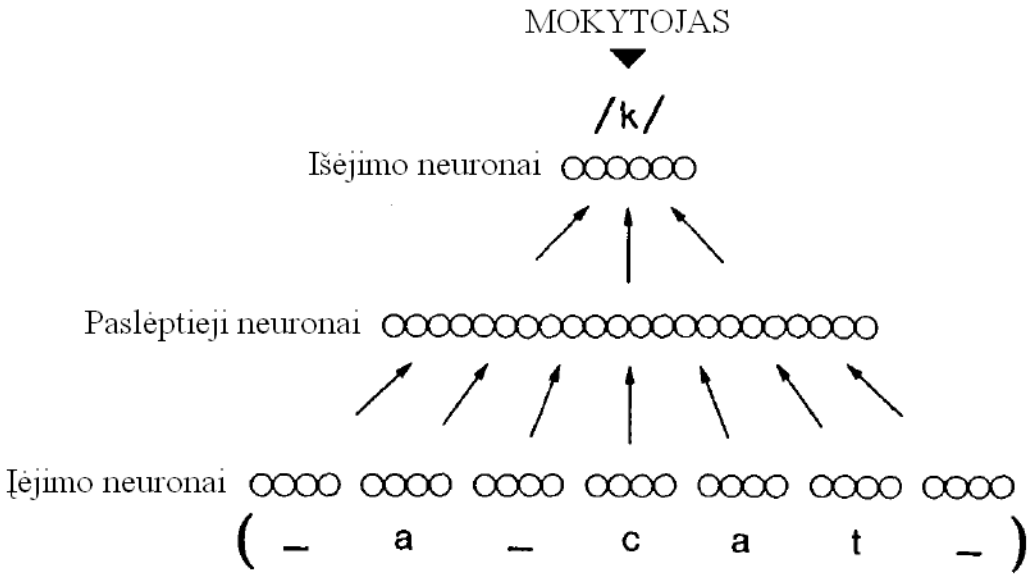
Skyrelio pabaigoje pridursime, kad, remiantis [Marinčič ir kt., 2009], lingvistų ekspertų kirčiavimo taisyklės nusileidžia tikslumu metodams grįstiems CART ir kitiems automatiškai taisykles sudarantiems metodams[11].

2.5.4 Automatiškai sudarytos taisyklės

Šios grupės taisyklės yra sudaromos automatiškai pagal mokymo duomenis. Mokymo duomenis sudaro įvairūs tekstynai ir žodynai (žodžių ar žodžių dalių). Automatinio sudarymo metodai gali būti įvairūs: dirbtiniai neuroniniai tinklai (ANN), sprendimo medžiai (CART) ir kiti metodai.

Dirbtinius neuroninius tinklus kirčiui nustatyti naudojančių sistemų atveju kirčiavimo taisyklės yra saugomos neuroninio tinklo svoriuose, t. y. neišreikštiniu (angl. *implicit*) žmogui nesuprantamu pavidalu. Viena iš pirmųjų tokių sistemų *NETtalk* [Sejnowski, Rosenberg, 1987] – tai sistema, kuri prognozuoja kirtį ir fonemą, t. y. dar atlieka ir transkribavimo uždavinį. Į įėjimo neuronų sluoksnį yra pateikiama raidė ir po 3 simbolius iš kairės ir iš dešinės (iš viso 7 simbolių langas). Kiekvienas simbolis aprašomas 29 neuronais (26 raidės ir 3 skyrybos ženklai). Iš viso yra 203 įėjimo neuronai. Išėjimo neuronų sluoksnis – tai 21 fonemos požymis (transkribavimo uždavinys) ir 5 papildomi išėjimo neuronai kirčiavimui ir skiemenų riboms (iš viso 26 išėjimo neuronai). Šio ANN schema pavaizduota 2.1 pav. [Sejnowski, Rosenberg, 1987]:

[11] Čia buvo tiriama fleksinė slovėnų kalba.



2.1 pav. ANN schema. Išėjime gaunama fonema *k*, įėjime paduodant *c* raidę ir jos kontekstą. Šiuo atveju išėjimo fonema, aišku, yra nekirčiuota

Kitas, naujesnis darbas, naudojantis ANN aprašytas [Arciuli, Thompson, 2006]. Čia kirčiuojami tik dviskiemeniai žodžiai ir kitaip negu [Sejnowski, Rosenberg, 1987] darbe, neuroninis tinklas tik kirčiuoja (netranskribuoja).

Kitas būdas automatiškai sudaryti taisykles – tai sukurti sprendimo medžius (daugiau apie kirčiavimą naudojant sprendimo medžius žr. 5.2.3 skyrelį).

2.6 Lietuvių kalbai taikyti metodai

Iš pradžių didžioji dalis tyrimų buvo skirta tik anglų kalbos kirčiavimui, ir tik vėliau atsirado tyrimų, skirtų kitoms kalboms, taip pat ir lietuvių. Reziumuojant tai, kas buvo pasakyta anksčiau šiame skyriuje, lietuvių kalba: a) turi laisvąjį kirtį, todėl nėra paprastų kirčiavimo taisyklių, b) tai fleksinė kalba ir atsižvelgiant į žodžio gramatinę formą žodžio kirtis gali pasikeisti, c) turi priegaides, todėl duotajam žodžiui reikia ne tik nustatyti kirčio vietą, bet ir priegaidę. Nors lietuvių kalbos kirčiavimas yra išsamiai aprašytas skirtinguose kalbotyros vadovėliuose (pvz., [Ambrazas, 1996]), pritaikyti šias taisykles kompiuteriui yra gana sudėtinga [Norkevičius ir kt., 2004]. Tačiau

visgi automatinio kirčiavimo uždavinys lietuvių kalbai jau pradėtas spręsti. Vienas pirmųjų algoritmų lietuvių kalbai pateiktas [Kasparaitis, 2000, 2001b, 27-48].

Algoritmas naudoja:

1) Maždaug 60000 lietuvių kalbos kamienų (~53000 daiktavardžių ir būdvardžių, ~8700 veiksmažodžių, ~2300 nekaitomų žodžių) žodyną. Taip pat yra naudojami galūnių, o veiksmažodžiams ir priešdėlių, sąrašai. Patogumo dėlei toliau morfemomis vadinsime ne tik priešdėlius ir galūnes, bet kartais ir kamienus.
2) Žodyne saugoma informacija apie tai, prie kokių kamienų kokios galūnės ir priešdėliai gali būti prijungiami, t. y. morfologinio kaitymo taisyklės. Pvz., prie žodžio *namas*, kamieno *nam* gali būti prijungtos galūnės: *-as*, *-o*, *-ui*, *-ą*, *-u*, *-e*, *-e*, *-ai*, *-ų*, *-ams*, *-us*, *-ais*, *-uose*. Dar yra naudojamos kelios papildomos raidžių kitimo kamiengaliuose taisyklės. Pvz., jei veiksmažodžio galūnės pirmoji raidė yra *s* ir kamienas baigiasi *s*, *z*, *š*, *ž*, tai galūnės *s* panaikinama ir *z*, *ž* pakeičiami *s*, *š* atitinkamai, pvz., *megz* + *siu* = *megsiu*, ir t. t. [Kasparaitis, 2001b, 42].
3) Šiuose žodynuose ir sąrašuose kiekvienam kamienui, galūnei ir priešdėliui yra saugoma informacija apie šių morfemų kirčiavimą, taip pat informacija, kuri morfema „pasiims“ kirtį. Veiksmažodžiams ir daiktavardžiams-būdvardžiams buvo sudarytos skirtingos taisyklės. Pvz., veiksmažodžių kirčiavimo taisyklė:

„Jei yra priešdėlis ir kirtis atitraukiamas į priešdėlį – kirčiuoti priešdėlį, jei kirčiuojamas paskutinis kamieno skiemuo ir jo priegaidė netvirtapradė – kirčiuoti galūnę, priešingu atveju – kirčiuoti kamieną.“ [Kasparaitis, 2001b, 39].

Visa šiai taisyklei reikalinga informacija (ar kirtis atitraukiamas į priešdėlį, kuris kamieno skiemuo yra kirčiuotas, kamieno kirčio priegaidė ir t. t.) yra saugoma pačiame žodyne.

Kituose darbuose [Kasparaitis, 2001a], [Kasparaitis, 2001b, 49-66] autorius siūlo metodus, kaip sumažinti kirčiavimui naudojamo daiktavardžių-

būdvardžių kamienų žodyno [Kasparaitis, 2000] apimtį ir kaip kirčiuoti kai kuriuos naujus žodžius, kurių kamienų nerasta žodyne. Žodynas sumažinamas pakeičiant vienodai linksniuojamų ir kirčiuojamų kamienų grupes taisyklėmis. Taisyklės remiasi kamieno pabaigos raidžių seka, dažniausiai priesaga. Pvz., vietoje žodžių *darbiniñkas, kakliniñkas* galima saugoti taisyklę *niñk-AS_2, kur „*“ reiškia bet kokią raidžių seką, AS – pridedamų galūnių sąrašą, o „2“ – kirčiuotę. Taisyklės saugomos tame pačiame žodyne, kaip ir kamienai, jos sudaromos rankiniu ir automatiniu būdu iš kamienų žodynų. Taikant šiuos metodus autorius sumažino kirčiavimui naudojamą daiktavardžių ir būdvardžių žodyną 64,4 proc. bei padidino padengimo koeficientą beveik dviem procentais [Kasparaitis, 2001b, 65]. Nors autorius vadina šį metodą taisyklėmis grįstu, tačiau žodžiai pirmiausiai skaidomi į morfemas, o taisyklės taikomos tik kamienui, todėl šis metodas artimesnis morfemų žodynais grįstiems metodams, o ne 2.5.4 skyrelyje aprašytiems taisyklėmis grįstiems metodams.

Kitas lietuvių kalbos automatinio kirčiavimo metodas aprašytas [Kazlauskienė, Raškinis, 2004], [Kazlauskienė ir kt., 2004], [Norkevičius ir kt., 2004].

Šis metodas naudoja:

1) Maždaug 50000 žodžių pagrindinių formų žodynus (45000 daiktavardžių ir 7500 veiksmažodžių).
2) Kirčiuojamo žodžio pagrindinei formai ir kitai morfologinei informacijai rasti naudotas morfologinis analizatorius *Lemuoklis* [Zinkevičius, 2000].
3) Kirčiavimo taisyklės, naudojančios šiuos žodynus buvo sudarytos rankiniu būdu ir saugomos medžio pavidalu.

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

Nors šis metodas ir naudoja žodžių pagrindinių formų žodyną (t. y. žodžių, o ne morfemų ar žodžių dalių), tačiau pagrindines formas pateikia morfologinė analizė, todėl šis metodas artimas morfemų žodynais grįstiems metodams.

2.7 Antrojo skyriaus apibendrinimas

- Kirčiavimas – svarbus faktorius, turintis įtakos šnekos (taip pat ir sintetinės) suprantamumui ir natūralumui bei realizuojamas tono, trukmės ir amplitudės modifikacijomis.
- Pagal kirčiavimo paradigmą kalbos skirstomos į fiksuotojo kirčio ir laisvojo kirčio kalbas. Pagal kaitymo laipsnį kalbas patogu skirstyti į fleksines ir nefleksines. Lietuvių kalba yra laisvojo kirčio fleksinė kalba.
- Fiksuoto kirčio kalboms kirčiuoti dažniausiai naudojamos paprastos taisyklės ir išimčių sąrašas.
- Laisvojo kirčio kalboms taikomus kirčiavimo algoritmus galima suskirstyti į grįstus žodynais ir grįstus taisyklėmis. Pirmuosius, savo ruožtu, galima skirstyti į naudojančius žodžių žodynus ir naudojančius žodžio dalių, dažniausiai morfemų, žodynus, o antruosius – į ekspertų sudarytas taisykles ir automatiškai generuotas taisykles.
- Laisvojo kirčio nefleksinėms kalboms pastaruoju metu dažniausiai naudojamos automatiškai sudarytos taisyklės.
- Lietuvių kalbai iki šiol buvo išimtinai taikyti tik morfemų žodynais grįsti metodai.

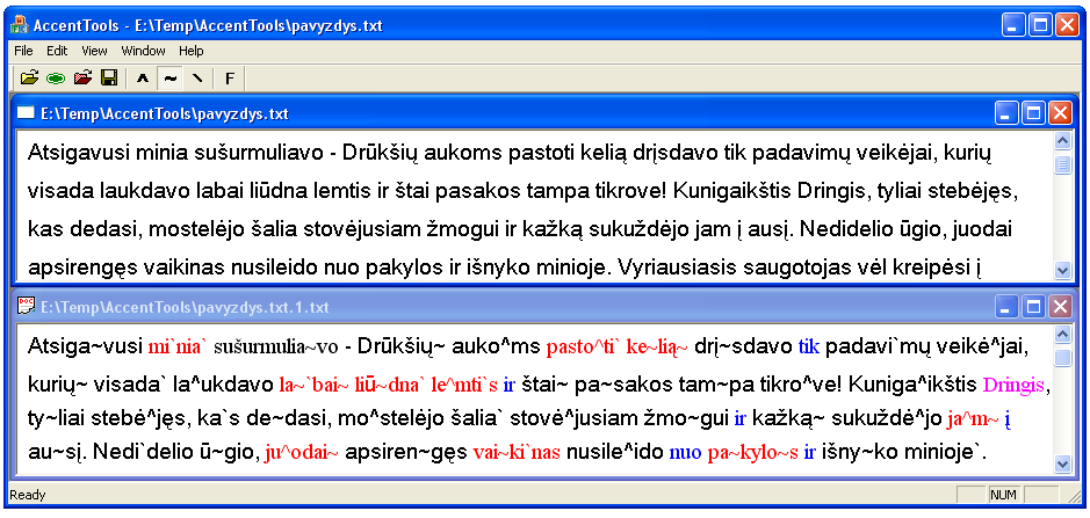
3 Duomenų paruošimas

Kirčiavimo eksperimentams reikalingas didelis kiekis kirčiuotų tekstų. Šiame skyriuje pristatoma programinė įranga, kuria naudojantis buvo paruošti duomenys 4 ir 5 skyriuose aprašytiems eksperimentams, taip pat aprašyti patys duomenys (tekstai), jų pasiskirstymas pagal žanrus.

3.1 Duomenims paruošti naudota programinė įranga

Šio darbo 4 ir 5 skyriuose aprašytiems eksperimentams atlikti reikalingas didelis kiekis kirčiuotų tekstų. Tekstams sukirčiuoti buvo panaudotas morfologine analize pagrįstas automatinio kirčiavimo algoritmas [Kasparaitis, 2001b, 27-48], aprašytas šio darbo 2.6 skyrelyje.

Minėtas algoritmas buvo įkomponuotas į specialią programą *AccentTools*, kuri a) sukirčiuoja tekstą, b) skirtingomis spalvomis išskiria žodžius, turinčius kelis kirčiavimo variantus, ir nekirčiuotus žodžius, c) leidžia vartotojui parinkti vieną kirčiavimo variantą ar pataisyti (uždėti) kirčio ženklą (žr. 3.1. pav.).



3.1 pav. Programos *AccentTools* pagrindinis langas

Raudona spalva pažymėti žodžiai, kuriems buvo grąžintos kelios kirčiavimo hipotezės, mėlyna – klitikai, violetine – neatpažinti žodžiai, juoda –

tinkamai sukirčiuoti žodžiai. Kirčio ženklas („~“ – riestinis, „`“– kairinis, „^“– dešininis) pasirenkamas meniu mygtukų pagalba. Kairysis pelės mygtukas panaikina kirtį, dešinysis – uždeda. Su šia programa profesionalus filologas sukirčiavo ir peržiūrėjo aibę tekstų, kurių bendra apimtis beveik milijonas žodžių (985.967 žodžiai).

3.2 Tekstų pasiskirstymas pagal žanrą

Tekstai buvo surinkti iš interneto ir pagal žanrą suskirstyti į šešias grupes: grožinė literatūra, mokslinė literatūra, įstatymai, respublikinė periodika, vietinė periodika, specializuota ir populiarioji periodika. Parenkant tekstus pagal žanrą buvo stengiamasi atsižvelgti į VDU tekstyne (prieiga per internetą: [http://donelaitis.vdu.lt, žiūrėta 2008.10.23]) esančias proporcijas (žr. 3.1 lentelę).

3.1 lentelė. Tekstų pasiskirstymas pagal žanrą VDU tekstyne ir autoriaus sudarytame tekstyne

VDU tekstynas			Autoriaus sudarytas tekstynas		
Teksto žanras	Žodžių skaičius	%	%	Žodžių skaičius	Teksto žanras
Respublikinė periodika	24.803.732	24,11%	23,19%	233.976	Respublikinė periodika
Vietinė periodika	16.918.103	16,45%	16,50%	166.501	Vietinė periodika
Populiarioji periodika	16.582.983	16,12%	25,16%	253.902	Populiarioji ir specializuota periodika
Specializuota periodika	9.760.811	9,49%			
Grožinė literatūra (knygos)	6.728.513	6,54%	11,06%	111.654	Grožinė literatūra
Filosofinės literatūros vertimai	2.721.990	2,65%			
Memuarai	2.226.726	2,16%			
Negrožinė literatūra (knygos)	11.472.234	11,15%	10,78%	108.762	Mokslinė literatūra
LR Valstybiniai dokumentai	8.734.236	8,49%	13,31%	134.297	Įstatymai, vyriausybės nutarimai
Seimo stenogramos	2.907.999	2,83%			
Iš viso	102.857.327			1.009.092	Iš viso

Žodžių skaičius autoriaus sudarytame tekstyne apskaičiuotas su *Microsoft Word* funkcija *Spaudos ženklų apskaita* (angl. *Word Count*). Laikantis tų pačių proporcijų, tekstai buvo padalinti į penkias maždaug vienodas dalis. Kaip šie tekstai buvo naudojami eksperimentuose, žr. 4 ir 5 skyrius.

3.3 Trečiojo skyriaus apibendrinimas

- Sukurta programinė įranga, automatiškai sukirčiuojanti pateiktą tekstą, išskirianti skirtingomis spalvomis nekirčiuotus ir keliais variantais kirčiuotus žodžius ir leidžianti vartotojui pakoreguoti kirčiavimą.
- Naudojant minėtą programinę įrangą paruoštas beveik milijono žodžių įvairių žanrų kirčiuotas tekstynas.

4 Lietuvių kalbos homografų vienareikšminimas

Lietuvių kalbos žodžiams kirčiuoti taikant morfologine analize grįstus metodus, kai kuriems žodžiams (homografams) pateikiama daugiau negu vienas kirčiavimo variantas. Šiame skyriuje aprašyti autoriaus pasiūlyti homografų vienareikšminimo metodai, pagrįsti leksemų ir morfologinių pažymų vartosenos dažniais, gautais iš vieno milijono žodžių tekstyno. Tokie metodai iki šiol lietuvių kalbai nenaudoti [Rimkutė, Grigonytė, 2006b]. Išnagrinėta metodų sąveika ir keli metodai sujungti į vieną algoritmą. Eksperimentiškai įvertintas algoritmo tikslumas, rezultatai palyginti su algoritmo ID3 rezultatais.

4.1 Homografo sąvoka

Jei žodžio kirčiavimui naudojami morfologine analize grįsti metodai, žodžio kirčiavimą patogu išskaidyti į tokius tris žingsnius: antraštinio pavidalo (lemos) suradimas, gramatinio aprašo (giminės, skaičiaus, linksnio ir pan.) suradimas ir kirčio vietos bei priegaidės nustatymas remiantis lema bei gramatiniu aprašu. Nemaža dalis lietuvių ir kitų pasaulio kalbų žodžių gali turėti kelias lemas ir kelis gramatinius aprašus. Pvz., anglų kalbos žodis *lives* gali būti: a) daiktavardžio *life* (*gyvybė* arba *gyvenimas*) daugiskaita, b) veiksmažodžio *live* (*gyventi*) esamasis laikas, III asmuo. Lietuvių kalbos žodis *galvos* gali būti: a) daiktavardžio *galva* vienaskaitos kilmininkas, b) daugiskaitos vardininkas, c) veiksmažodžio *galvoti* būsimasis laikas, III asmuo. Tokie žodžiai vadinami homoformomis, o pats reiškinys – morfologiniu daugiareikšmiškumu [Rimkutė, 2002] (žymėsime santrumpa MD). Teigiama [Rimkutė, 2002], kad homoformos sudaro 39% lietuviško teksto, o remiantis [Rimkutė, Grybinaitė, 2004] – kad net 47%.

Dalis homoformų tariamos vienodai, dalis – skirtingai. Skirtingai tariamos homoformos vadinamos **homografais**. Darbe [Kasparaitis, 2000] nagrinėtuose lietuviškuose tekstuose homografai sudarė daugiau negu 15%

visų žodžių, darbe [Rimkutė, 2002] – 8,1%. Anglų kalboje terminas „skirtingas tarimas“ reiškia skirtingą kirčio poziciją arba skirtingą transkripciją, pvz., anglų kalbos žodis (homografas) *read* (*skaityti* arba *skaitė*, transkripcijas žr. 1.4.6 skyrelyje). Lietuvių kalboje „skirtingas tarimas“ reiškia skirtingą kirtį (jo poziciją), priegaidę arba kirčiuoto skiemens ilgumą, pvz., *vỹk* (iš *vykti*) ir *výk* (iš *vyti*), *mẽs* (*aš* daugiskaita) ir *mès* (iš *mesti*).

Taigi dėl morfologinio daugiareikšmiškumo kai kurių žodžių (homografų) negalima vienareikšmiškai sukirčiuoti (ar transkribuoti), todėl reikalingas vienareikšminimas (vieno varianto išrinkimas).

Toliau šiame skyriuje (nuo 4.5 skyrelio) bus nagrinėjamos ne visos homoformos, o tik homografai ir jų vienareikšminimas, nes tik homografai sukelia sunkumų kirčiavimo algoritmą pritaikant balso sintezei pagal tekstą, tačiau, remiantis [Taylor, 2009, 110], homoformų ir homografų vienareikšminimui naudojami metodai dažnai iš esmės sutampa.

4.2 Homografų tipai

Lietuvių kalboje homografų atsiradimą gali nulemti du morfologinio daugiareikšmiškumo tipai (iš dalies pagal [Rimkutė, Grigonytė, 2006a]):

1) Lemų daugiareikšmiškumas, pvz.: *dominò* (nekaitomas daiktavardis) ir *dõmino* (veiksmažodis, lema *dominti*).
2) Morfologinių pažymų daugiareikšmiškumas (linksnių sinkretizmas), pvz., *sū́nų* (daiktavardžio *sūnus* vienaskaitos galininko linksnis) ir *sūnų̃* (daiktavardžio *sūnus* daugiskaitos kilmininko linksnis).

Kai kurios homoformos (ir homografai) gali priklausyti abiems tipams, pvz., žodis *galvos* (žr. pavyzdį aukščiau 4.1 skyrelyje). Darbe [Rimkutė, Grigonytė, 2006a] teigiama, kad „Vienokios priemonės taikomos mažinant lemų, kitokios – morfologinių pažymų daugiareikšmiškumą.“

Anglų kalboje dėl jos nefleksinio pobūdžio retai pasitaiko morfologinių pažymų daugiareikšmiškumas (nors yra ir tokių atvejų, pvz., jau minėtas veiksmažodis *read*), todėl paprastai naudojama kitokia morfologinio

daugiareikšmiškumo klasifikacija (iš dalies pagal [Taylor, 2009, 101-102], [Yarowsky, 1996]):

1) Kalbos dalių daugiareikšmiškumas (angl. *POS ambiguity*), t. y. homografai priklauso skirtingoms kalbos dalims, bet dažniausiai susiję semantiškai (pagal prasmę), pvz., *desert* (daiktavardis *dykuma* ir veiksmažodis *palikti, išvykti*, transkripcijas žr. 1.4.6 skyrelį).
2) Semantinis daugiareikšmiškumas, t. y. homografai priklauso tai pačiai kalbos daliai, bet turi skirtingą reikšmę, pvz., daiktavardžiai *bass* (/beıs/ *bosas* arba /bæs/ *ešerys*) ir *bow* (/bau/ *laivo priekis* arba /bəu/ *lankas*).

Kaip teigiama [Yarowsky, 1996], didžioji dalis anglų kalbos homografų priklauso pirmam tipui, t. y. skirtingoms kalboms dalims. Kai kurie homografai priklauso keliems tipams, pvz., *lead* (*švinas*, *pirmavimas*, *nuvesti*) turi skirtingas transkripcijas veiksmažodžio ir daiktavardžio atvejais bei skirtingas transkripcijas dviem daiktavardžio reikšmės atvejais. Kaip teigiama [Yarowsky, 1996], dažniausiai iš pradžių ribojamas POS daugiareikšmiškumas, po to (jei toks yra) semantinis daugiareikšmiškumas.

Dar galima paminėti tokius „specialius“ homografų tipus, tačiau jų toliau nenagrinėsime.

- Santrumpos: pvz.: angl. *Dr*. (*Drive* arba *Doctor*); arba liet. *šv*. (*šventas* arba *šviesiai*), plg. 1.4.1 skyrelį.
- Vardai, vietovardžiai ir kiti tikriniai daiktavardžiai, pvz., angl. *Nice* (miestas *Nica* arba *gražus, geras, šaunus*).
- Romėniški skaitvardžiai bei skaitmenys, pvz., angl. *Henry III* (*Henry the Third*), bet *Chapter III* (*Chapter Three*).
- Trupmenos ir datos, pvz., angl. *5/16* gali reikšti *penkios šešioliktosios* arba *gegužės šešioliktoji*.
- „Netikrieji homografai“ [Huang ir kt., 2001, 713-714]: tarmės arba net individualūs dialektai, kalbos greitis ir pan. Pvz., angl. žodyje *interesting* (*įdomus*) kartais gali būti praleidžiama, o kartais paliekama pirmoji *e*.

4.3 Homografų vienareikšminimo problema kitose kalbose

Homografų vienareikšminimas – tai klasifikavimo problema, kur išeities rezultatas yra nagrinėjamo žodžio (homografo) transkripcija (homoformų apskritai atveju – nagrinėjamo žodžio reikšmė), o įvesties parametrus dažniausiai sudaro pats nagrinėjamas žodis ir jo kaimynai bei įvairios jų savybės (kaip kalbos dalys ir pan.) (apibrėžimas iš dalies pagal [Yarowsky, 1996]). Kaip jau buvo minėta 4.1 skyrelyje, homoformų ir homografų vienareikšminimo metodai iš esmės sutampa, todėl toliau šiame skyrelyje juos nagrinėsime kartu ir vadinsime tiesiog „homografų vienareikšminimu“. Be to, kaip jau buvo minėta 4.2 skyrelyje, didžioji dalis anglų kalbos homografų priklauso skirtingoms kalboms dalims, todėl dažnai vienareikšminimui užtenka nustatyti, kokia kalbos dalis yra nagrinėjamas žodis [Huang ir kt., 2001, 713]. Tokie kalbos dalies nustatymo metodai vadinami POS žymėjimo metodais (angl. *POS taggers*), juos irgi šiame skyrelyje nagrinėsime kartu su kitais homografų vienareikšminimo metodais. Įvairių kalbų homografų vienareikšminimui gali būti naudojami, pvz., *n*-gramų modeliai, Bajeso klasifikatoriai, slaptieji Markovo modeliai (HMM), sprendimo medžiai (CART), sprendimo sąrašai, dirbtiniai neuroniniai tinklai (ANN), ekspertų lingvistų sukurtos vienareikšminimo taisyklės. Kai kurie autoriai ([Merialdo, 1994], [Nivre ir kt., 1996], [Orphanos, Christodoulalds, 1999]) homografų vienareikšminimo metodus dalina į dvi pagrindines grupes: a) taisyklėmis grįstus, ir b) tikimybinius (*n*-gramų modeliai, Bajeso klasifikatoriai ir kt.). Darbe [Merialdo, 1994] dar išskiriami ir dirbtiniai neuroniniai tinklai. Visi metodai, išskyrus ekspertų lingvistų sukurtas taisykles, dažniausiai kaip mokymosi duomenis naudoja anotuotus tekstynus ar žodynus, kuriuose rankiniu būdu yra sužymėtos žodžių reikšmės, POS ir t. t. Kitas būdas, pasiūlytas [Gale ir kt., 1992], naudoja dviejų kalbų vertimo tekstus, t. y. apie tai, kokia žodžio reikšmė panaudota nagrinėjamame kontekste vienoje kalboje, sprendžiama iš šio žodžio vertimo į kitą kalbą. Toliau trumpai apžvelgsime kelis populiariausius metodus.

***n*-gramų modelis** (angl. *n-gram taggers*) – tai tikimybinis modelis, kuris prognozuoja elemento reikšmę sekoje remiantis prieš tai buvusių *n* elementų reikšmėmis. Elementai gali būti tiek patys žodžiai, tiek ir jų POS žymės ar pan. Dažnai šis tikimybinis modelis yra naudojamas nustatyti, kokia kalbos dalis yra nagrinėjamas žodis. Anglų ir kai kurių kitų kalbų sakinyje yra griežta žodžių tvarka, todėl POS sekos dažnai sutampa. Anglų kalbos *n*-gramų POS žymėjimo algoritmai yra aprašyti, pvz., [Merialdo, 1994], [Church, 1988]. Metodas, aprašytas [Church, 1988], POS vienareikšminimui: a) naudoja unigramus (leksines tikimybes) $P(c_i|w_i)$, kur c_i yra hipotezė, kokia kalbos dalis (POS) yra žodis w_i. Čia žodžio POS dažnis tiesiog apskaičiuotas iš tekstyno[12]; b) trigramus (kontekstines tikimybes) $P(c_i|c_{i-1}c_{i-2}...c_1) = P(c_i|c_{i-1}c_{i-2})$, kur c_i yra POS; c) iš leksinių tikimybių (unigramų) ir kontekstinių tikimybių (POS trigramų) sudaromos įvairios sekos visiems įėjimo žodžiams ir pasirenkama seka su didžiausia tikimybe.

n-gramų POS žymėjimo modeliai remiasi tik konteksto POS, todėl nefiksuoja leksinių ryšių, pvz., „*a ribbon wound around the pole*“ (*juosta apsivyniojo aplink stulpą*) ir „*a bullet wound in the leg*“ (*kulkos žaizda kojoje*), čia žodžio „*wound*“ (*apsivyniojo* arba *žaizda*) kaimyninių žodžių POS sutampa abiejose frazėse, taigi vienareikšminant reikia atsižvelgti į kaimyninių žodžių leksinę informaciją. *n*-gramų modeliai remiasi vieno ar dviejų žodžių kontekstu (bigramai ar trigramai), ir, nors kai kuriose kalbose (pvz., anglų, vokiečių) vienareikšminant homografus dažniausiai pakanka betarpiško konteksto [Šveikauskienė, 2009, 35], tačiau, kita vertus, dėl to, kad nėra naudojami didesnio atstumo ryšiai tarp žodžių, šis metodas prastai tinka semantiniam daugiareikšmiškumui riboti [Yarowsky, 1996].

Bajeso klasifikatoriai (angl. *Bayesian classifiers*) dažnai naudojami semantiniam daugiareikšmiškumui riboti. Bajeso klasifikatoriai remiasi tikimybiniu modeliu $p(C|F_1, ..., F_n)$, kur C yra išeities (rezultato) klasės

[12] Pvz., [Francis, Kucera, 1982] teigiama, kad anglų kalbos žodžio *see* (*matyti*) POS 771 kartus buvo veiksmažodis ir tik vieną kartą jaustukas (skaičiuojant rankiniu būdu pažymėtame *Brown* tekstyne), t. y. tikimybė, kad šis žodis yra veiksmažodis, yra 771/772 arba 0,99.

kintamasis, kuris priklauso nuo parametrų F_1, ..., F_n. Šis klasifikatorius taip pat remiasi prielaida, kad kiekvienas parametras F_i yra sąlygiškai (angl. *conditionally*) nepriklausomas nuo kiekvieno kito parametro F_j, kur $j \neq i$. Ši prielaida leidžia ženkliai pagreitinti skaičiavimus. Pvz., darbe [Gale ir kt., 1992] kiekvienas homografas buvo aprašytas šimtu artimiausių žodžių, nežiūrint į jų pozicijas, tokiu būdu leidžiant fiksuoti didesnio atstumo temos lygmens (angl. *topic-level*) ryšius tarp žodžių, bet prarandant gebėjimą klasifikuoti remiantis lokaliom žodžių sekom ar sakinio struktūra.

Slaptųjų Markovo modelių analizė (HMM) dažnai yra naudojama POS vienareikšminti [Manning, 2000], [Charniak, 1993]. Šių modelių čia neaprašinėsime, tik paminėsime, kad vienareikšminant jie, panašiai kaip ir *n*-gramų modeliai, remiasi vieno ar dviejų kontekstinių žodžių POS. Remiantis [Rimkutė ir Grybinaitė, 2004], „Šis metodas buvo pritaikytas Nacionaliniame britų tekstyne. Buvo pasiektas 96–97 proc. tikslumas, priklausomai nuo analizuojamo teksto rūšies [Leech ir kt., 1994].“

Sprendimo medžiai, kurie detaliau bus aprašyti 5.2 skyrelyje, buvo naudojami homografams vienareikšminti, pvz., [Brown ir kt., 1991]. Čia prancūzų kalbos homografams vienareikšminti naudojami septyni parametrai: žodis iš kairės, žodis iš dešinės, pirmasis daiktavardis iš kairės, pirmasis daiktavardis iš dešinės, pirmasis veiksmažodis iš kairės, pirmasis veiksmažodis iš dešinės ir analizuojamojo žodžio laikas, jei tai yra veiksmažodis, arba pirmojo veiksmažodio iš kairės laikas. Anglų kalbos homografams vienareikšminti naudojami tik du parametrai: pirmasis žodis iš kairės ir antrasis žodis iš kairės. POS žymėjimo algoritmas graikų kalbai, naudojantis sprendimo medžių „mišką“, yra aprašytas [Orphanos, Christodoulalds, 1999].

Sprendimo sąrašai buvo naudojami anglų kalbos homografams vienareikšminti, pvz., [Yarowsky, 1996]. Šis klasifikatorius kiekvienam homografui ar jų tipui sudaro sprendimo sąrašus, kur didesnio patikimumo „įrodymas“ sąraše atsiduria aukščiau. „Įrodymus“ sudaro įvairūs nagrinėjamo homografo konteksto aprašymai, pvz., žodis ar jo POS iš kairės ar iš dešinės, žodis ar jo POS tarp 20 kaimyninių žodžių iš kairės ar iš dešinės ir pan. Tokiu

būdu šis metodas bando fiksuoti ir didesnio atstumo, ir lokalius ryšius tarp žodžių. Tas pats autorius pritaikė sprendimo sąrašus ir ispanų bei prancūzų kalbų leksiniam daugiareikšmiškumui panaikinti [Yarowsky, 1994].

Lingvistų ekspertų sukurtos vienareikšminimo taisyklės irgi yra naudojamos kai kurių kalbų homografams vienareikšminti. Pvz., čekų kalbai vos kelios tokios rankiniu būdu sudarytos ir šimto procentų tikslumu veikiančios taisyklės [Nepil, Popelinsky, 2001] pašalina apie pusę kalbos dalies daugiareikšmiškumo atvejų. Taisyklės remiasi kontekstiniais žodžiais, jų lemomis, kalbos dalimis ir pan. Estų kalbai buvo naudotos 1240 taisyklių, kai kurių taisyklių tikslumas buvo kiek mažesnis negu 100 proc. [Bick ir kt., 2004], [Puolakainen, 2001]. Remiantis [Rimkutė, Grigonytė, 2006a], „Anglų kalbos MD-ui riboti buvo sukurta 3300 taisyklių. Taip buvo pasiektas 77% tikslumas 1 mln. žodžių anotuotame *Browno* tekstyne“. Šių metodų pagrindinis trūkumas yra tas, kad skirtingoms kalboms reikia sudaryti skirtingas taisykles. Dažniausiai šios taisyklės taikomos kartu su kitas sudėtingesniais homografų vienareikšminimo metodais, pvz., ILP (angl. *Inductive Logic Programming*) [Nepil, Popelinsky, 2001], *n*-gramais [Rimkutė, Grigonytė, 2006a], [Dutoit, 1997, 95-96].

4.4 Lietuvių kalbai taikyti metodai

Nors, kaip jau buvo minėta 4.3 skyrelyje, kai kurių kalbų homografams vienareikšminti gali užtekti ir betarpiško konteksto, tačiau, pasak [Šveikauskienė, 2009, 35], „Lietuvių kalboje šituo pasinaudoti retai kada tegalima. *Tada mes aukštyn lipome sunkiai. Jis jau greitai mes į viršų kamuolį.* – betarpiškas kontekstas *tada mes aukštyn* ir *greitai mes į viršų* žodžio *mes* daugiareikšmiškumo nepanaikina.“ Be to, POS žymėjimas nelabai tinka lietuvių kalbai, nes: a) žodžių tvarka sakinyje nėra griežta, todėl POS sekos nėra pastovios, b) dažnas morfologinis daugiareikšmiškumas, todėl dažnai reikia nustatyti ne tik nagrinėjamo homografo POS, bet ir gramatinę formą ir pan.

Kai kurie lietuvių kalbos homoformų vienareikšminimo (ar daugiareikšmiškumo ribojimo) algoritmai: statistiniai slaptieji Markovo modeliai, loginis ID3 algoritmas, sintaksinė analizė, jau nagrinėti E. Rimkutės ir bendraautorių darbuose [Rimkutė, Grybinaitė, 2004], [Rimkutė, Grigonytė, 2006a], [Rimkutė, Grigonytė, 2006b].

Kaip pažymima [Rimkutė, Grybinaitė, 2004], „Daugiareikšmiškumo uždavinys buvo sprendžiamas iš dalies: bandyta nustatyti tik kalbos dalį, nenagrinėjant kitų morfologinių požymių“. Darbe morfologinio lietuvių kalbos daugiareikšmiškumui riboti buvo taikomi:

- **Slaptieji Markovo modeliai**: „Naudotas Viterbi algoritmas, pagrįstas prielaida, kad kalbos dalis priklauso tik nuo prieš tai buvusios kalbos dalies (bigramų atveju; trigramų atveju – nuo dviejų prieš tai buvusių kalbos dalių), t. y. nepriklauso nuo visos kalbos dalių sekos. <...> Sukurtoje vienareikšminimo sistemoje <...> randama tikėtiniausia sakinį atitinkanti kalbos dalių seka.“
- **Sprendimo medžiai**. Medžiams sudaryti buvo naudojamas **ID3 algoritmas** (detaliai apie sprendimo medžius ir ID3 algoritmą žr. šio darbo 5.2 skyrelį). „Pasirinktas algoritmas sukuria medžius, kurie gana nesudėtingai gali būti perrašomi į taisyklių aibę, suprantamą tiek informatikui, tiek lingvistui. <...> Sakiniuose ieškoma atskirų vienareikšminimo požymių kiekvienam daugiareikšmiškumo tipui. <...> Kiekvienai MD rūšiai kuriamas atskiras sprendimų medis.“ [Rimkutė, Grybinaitė, 2004].

Kaip pažymima kitame darbe [Rimkutė ir Grigonytė, 2006a]: „Automatizuotus MD-o mažinimo būdus galima skirti į tris grupes: nerealiųjų homoformų panaikinimas, nekaitomų kalbos dalių MD-o ribojimas ir kaitomų kalbos dalių MD-o ribojimas.“ a) „Paprasčiausias statistiniais duomenimis pagrįstas automatizuotas MD-o ribojimo būdas yra nerealiųjų homoformų panaikinimas.“ b) „Labiausiai riboti nekaitomų kalbos dalių MD-ą padeda sakinio skyryba; statistiniai duomenys, gauti iš Dabartinės lietuvių kalbos tekstyno; sintaksinė sakinio analizė; semantika, išsamesnė viso teksto ar kelių

gretimų sakinių analizė; pakeitimas kitais aiškiau vartojamais žodžiais; gretimų žodžių morfologinės pažymos; kaip vienas vienetas pažymėtos morfologinės samplaikos.“ c) „Kaitomų kalbos dalių MD-o ribojimas daugiausia pagrįstas **sintaksine analize**. <...> Remtasi morfologiniais ir kai kuriais semantiniais požymiais.“

4.5 Duomenų atranka ir paruošimas

Šiame skyriuje aprašytuose eksperimentuose buvo naudojami beveik milijono žodžių kirčiuoti tekstai (žr. 3 skyrių). Keturios dalys (800.000 žodžių) buvo naudojamos mokymui, o viena dalis (200.000 žodžių) – testavimui. Kadangi minėti kirčiuoti tekstai buvo sudaryti ne specialiai šiems eksperimentams, o kitais tikslais, todėl juose buvo išsaugota tik informacija apie kirčiavimą, informacijos apie lemas ir gramatines formas nėra. Dėl šios priežasties buvo modifikuotas kirčiavimo algoritmas (žr. 2.6 skyrelį ir 3 skyrių) taip, kad ne tik kirčiuotų, bet ir pateiktų informaciją apie žodžio lemą ir gramatinę formą, ir minėti tekstai buvo dar kartą apdoroti tokiu būdu: imama po vieną kirčiuoto teksto žodį, šis žodis (pašalinus kirčio ženklą) pateikiamas kirčiavimo algoritmui. Kirčiavimo algoritmas sugeneruoja visas galimas hipotezes, kokio žodžio kokia gramatinė forma tai gali būti ir kaip ji kirčiuojama. Toliau šiame darbe trumpumo dėlei vadinsime tiesiog **hipotezėmis**. Lygindami kirčiavimo algoritmo generuotų hipotezių kirčiavimą su kirčiuotu tekstu, randame hipotezes, kurių kirčiavimas sutampa su kirčiuotu tekstu (vad. **teisingomis hipotezėmis**), ir kurių nesutampa (vad. **klaidingomis hipotezėmis**). Pavyzdžiui, kirčiuotame tekste sutikus žodį *galvõs*, teisingos hipotezės bus a ir c (žr. 4.1 skyrelio pirmą pastraipą), o b – klaidinga hipotezė.

Kadangi vienas žodis gali atitikti dvi, tris ir dar daugiau gramatinių formų (hipotezių), be to, nuspėti gramatinių formų skaičių ir galimus jų derinius yra sunku, todėl nagrinėsime tik gramatinių formų (hipotezių) poras, kur viena hipotezė yra teisinga, o kita klaidinga. Pavyzdžiui, žodžiui galvos tokių hipotezių porų būtų dvi (a-b ir b-c). Kiekvienai hipotezių porai naudodami kirčiuotą tekstą galime suskaičiuoti, kiek kartų teisinga buvo pirmoji hipotezė,

ir kiek kartų – antroji. Hipotezę, kuri daugiau kartų buvo teisinga, vadinsime **dažnesne hipoteze**, o kitą poros hipotezę – **retesne hipoteze**. Rašydami hipotezių poras dažnesnę hipotezę rašysime pirmiau. Vienareikšminant homografus bus tiesiog imamos hipotezių poros ir atmetamos retesnės hipotezės, taip tikintis, kad liks tik tos, kurios kirčiuojamos vienodai.

Dar kartą verta priminti, kad kirčiavimo algoritmas naudoja tris žodynus: daiktavardžių-būdvardžių (toliau DB), veiksmažodžių (toliau Vks), nekaitomų žodžių (toliau Nek). Šių žodynų įrašus vadinsime leksemomis, kiekviename įraše saugomas žodžio kamienas, linksniuotė/asmenuotė, kirčiuotė ir kita kaitymui ir kirčiavimui reikalinga informacija. Iš kiekvienos leksemos galime nesunkiai rasti žodžio antraštinį pavidalą (lemą), todėl toliau šiame darbe kartais vietoje leksemos vartosime tik žodžio lemą, o kartais – tik kamieną, turėdami galvoje, kad pašalinus kamieną pašalinama ir visa leksema.

Dabar išsamiau pasižiūrėkime, kokią informaciją saugo kiekviena hipotezė: a) žodynas (DB, Vks, Nek); b) leksema (leksemos identifikatorius); c) gramatinė forma. Nekaitomų žodžių gramatinę formą charakterizuoja leksemos identifikatorius, todėl šie parametrai sutampa. Daiktavardžių-būdvardžių gramatinę formą charakterizuoja du parametrai: linksniuotė ir skaičius/linksnis. Taigi galime laikyti, kad kirčiavimo algoritmas užpildo hipotezių porų lentelę. Pavyzdžiui, jei visas turimas kirčiuotas tekstas sudarytas tik iš tokių dviejų žodžių *Mamà galvõs*, tuomet kirčiavimo algoritmo darbo rezultatas atrodys kaip pavaizduota 4.1 lentelėje. Stulpeliuose Ar_teis1 ir Ar_teis2 loginės reikšmės rodo atitinkamai teisingą ir klaidingą hipotezę. Veiksmažodžiams ir nekaitomiems žodžiams stulpelis Gr_f_12 arba Gr_f_22 yra tuščias.

4.1 lentelė. Hipotezių porų lentelė

Žody-nas1	Gr_f_11	Gr_f_12	Lekse-ma1	Ar_teis1	Žody-nas2	Gr_f_21	Gr_f_22	Lekse-ma2	Ar_teis2
DB	vns. V.	3 linksn.	mama	TRUE	DB	vns. Š.	3 linksn.	mama	FALSE
DB	vns. Įn.	3 linksn.	mama	TRUE	DB	vns. Š.	3 linksn.	mama	FALSE
DB	vns. K.	3 linksn.	galva	TRUE	DB	dgs. V.	3 linksn.	galva	FALSE
Vks	būs. l.	–	galvoti	TRUE	DB	dgs. V.	3 linksn.	galva	FALSE

Tolesni eksperimentai bus atliekami įvairiais pjūviais grupuojant duomenis, analogiškus pateiktiems 4.1 lentelėje.

4.6 Leksemų atmetimas

Iš pradžių panagrinėkime, kaip vienareikšminimui galima panaudoti leksemų dažnius. Galima pastebėti, kad kai kurių leksemų buvimas žodyne labiau kliudo kirčiuoti, nei padeda. Pavyzdžiui, darbe [Rimkutė, Grybinaitė, 2004] paminėtas itin retai vartojamas žodis *kokis* (žodžio *kokybė* sinonimas), kurio vienaskaitos kilmininkas sutampa su gana dažnai vartojamo įvardžio *koks* vienaskaitos kilmininku, o šie žodžiai kirčiuojami skirtingai (skiriasi priegaidė). Išmetus tokią retai vartojamą leksemą iš žodyno, daugiau žodžių būtų galima sukirčiuoti vienareikšmiškai.

Bus nagrinėjami du leksemų atmetimo būdai:

1) randamos daiktavardžių-būdvardžių žodyno leksemų poros, kurių sutampa kamienai ir linksniuotės. Tai galima užrašyti tokia užklausa:

```
SELECT Leksema1, count(Ar_teis1 = TRUE),
Leksema2, count(Ar_teis2 = TRUE),
GROUP BY Leksema1, Leksema2,
WHERE (Žodynas1 = „DB“) && (Žodynas2 = „DB“)
&& (Gr_f_11 = Gr_f_21) && (Gr_f_12 = Gr_f_22).
```

Šioje užklausoje reikalavimas, kad sutaptų ir skaičiai/linksniai, ir linksniuotės, garantuoja, kad sutampa kamienų tekstinis pavidalas. Analogiškai, kaip apibrėžėme dažnesnę hipotezę, galime apibrėžti **dažnesnę leksemą**. Dažnesne vadinsime tą, iš kurios dažniau generuojamos teisingos hipotezės. Atmetamos retesnės leksemos. Toliau pateiktuose pavyzdžiuose vietoj leksemos rašysime lemą, šalia lemos pateiksime pasikartojimų skaičių, trumpumo dėlei atskirsime dažnesnę leksemą nuo retesnės ženkliuku >. Pvz., *kláusimas* (699) > *klausìmas* (0), *Jõnas* (172) > *jònas* (17), *gė́rimas* (72) > *gėrìmas* (2), *romãnas* (49) > *Ròmanas* (5). Analogiškai galima atmesti ir nekaitomus žodžius, kurie retesni už kitus nekaitomus žodžius, pvz., *paskuĩ*

(156) > *pãskui* (11), arba už kaitomų žodžių gramatines formas, pvz., *mètro* (12) > *metrò* (9), *dõmino* (7) > *dominò* (0).

2) kiekvienai leksemai tiesiog suskaičiuojamas teisingų (klaidingų) hipotezių skaičius. Tai nusakoma užklausa:

```
SELECT Leksema1, count(Ar_teis1 = TRUE),
Leksema2, count(Ar_teis2 = TRUE),
GROUP BY Leksema1, Leksema2.
```

Kalbant apie čia pateiktas užklausas, reikia nepamiršti, kad hipotezių tvarka porose yra atsitiktinė, todėl poras sugrupavus pagal leksemas, kiekvienai grupei galima rasti simetrišką grupę, t. y. tokią, kurios dešinę pusę sukeitę su kaire gausime identišką grupę. Tokias grupes reikia sujungti sukeitus kairę pusę su dešine.

Pirmasis būdas leidžia atmesti tik nedidelį skaičių leksemų, tačiau jis garantuoja, kad nepadaugės nekirčiuotų ar klaidingai kirčiuotų žodžių. Antrasis būdas leidžia atmesti daugiau leksemų, tačiau jis gali turėti ir pašalinių efektų. Pavyzdžiui, iš veiksmažodžio kamieno *kárti* (*kãria*, *kórė*) padarytas dalyvis *kártas* sutampa su daiktavardžiu *kar̃tas*, tačiau *kar̃tas* (1911) > *kártas* (104). Analogiškai veiksmažodžio *laisvė́ti* (*laisvė́ja*, *laisvė́jo*) būsimasis laikas *laisvẽs* sutampa su daiktavardžio *láisvė* vienaskaitos kilmininku ar daugiskaitos vardininku, tačiau *láisvės* (140) > *laisvẽs* (1). Atmetus šiuos veiksmažodžių kamienus, neliks problemų kirčiuojant daiktavardžius *kartas* ir *laisvė*, tačiau liks nekirčiuoti kiti iš šių veiksmažodžių kamienų daromi žodžiai. 4.2 lentelėje pateikti duomenys, kiek kuris būdas leido atmesti leksemų ir kiek tos leksemos mokymo duomenims generavo teisingų (klaidingų) hipotezių ir teisingų hipotezių dalis procentais.

Kalbant apie nekaitomus žodžius dar verta paminėti, kad klitikai pradiniame kirčiavimo algoritmo variante ir pritaikius pirmą būdą buvo atpažįstami naudojant specialų algoritmą (žr. 6 skyrių), o naudojant antrą būdą jie tiesiog buvo pašalinti iš nekaitomų žodžių sąrašo.

4.2 lentelė. Leksemų atmetimo rezultatai mokymo duomenims

		Pradinis	Pritaikius 1 būdą		Pritaikius 1 ir 2 būdus	
			Liko	Atmesta	Liko	Atmesta
DB	Leksemų skaičius	62106	62041	65	61107	999
	Hipotezių skaičius	594344/ 102505 85,3%	593549/ 98277 85,8%	795/ 4228 15,8%	586654/ 63160 90,3%	7690/ 39345 16,3%
Vks	Leksemų skaičius	8826	8826	0	8437	389
	Hipotezių skaičius	243340/ 82934 74,6%	243340/ 82934 74,6%	0/ 0	235346/ 46919 83,4%	7994/ 36015 18,2%
Nek	Leksemų skaičius	2049	2038	11	1915	134
	Hipotezių skaičius	61211/ 1274 98,0%	61119/ 919 98,5%	92/ 355 20,6%	61081/ 715 98,8%	130/ 559 18,9%
	Iš viso	898895/ 186713 **82,8%**	898008/ 182130 83,1%	887/ 4583 16,2%	883081/ 110794 **88,9%**	15814/ 75919 17,2%

Kaip matome iš 4.2 lentelės, atmetus kai kurias leksemas teisingų hipotezių dalis tarp visų hipotezių padidėja nuo 82,8% iki 88,9%, t. y. 6,1%, tačiau, kokią įtaką šis padidėjimas turi teksto kirčiavimui, bus tiriama 4.8 skyrelyje.

4.7 Gramatinių formų dažniais grįstos taisyklės

4.7.1 Taisyklių formavimas

Šiame skyriuje panagrinėsime 4.1 lentelės duomenų grupavimą pagal gramatines formas; tam naudojama tokio tipo užklausa:

```
SELECT Žodynas1, Gr_f_11, Gr_f_12, count(Ar_teis1=TRUE),
Žodynas2, Gr_f_21, Gr_f_22, count(Ar_teis2=TRUE),
GROUP BY Žodynas1, Gr_f_11, Gr_f_12, Žodynas2, Gr_f_21,
Gr_f_22.
```

Gaunam naują hipotezių porų (dažnesnės ir retesnės) lentelę, kurios įrašus galima traktuoti kaip tam tikras vieno kirčiavimo varianto parinkimo remiantis morfologinių pažymų dažniais taisykles, turinčias tokį pavidalą:

```
(Žodynas1,Gr_f_11,Gr_f_12) > (Žodynas2,Gr_f_21,Gr_f_22).
```

Suskirstykime gautas taisykles į keturias grupes prie aukščiau minėtos užklausos pridedant papildomų sąlygų:

A. Užklausa dar papildyta sąlyga, kuri į atskirą grupę išskiria visas su nekaitomais žodžiais susijusias taisykles.

```
WHERE ((Žodynas1 = „Nek“) || (Žodynas2 = „Nek“).
```

B. Iš likusių taisyklių naudojant papildomą sąlygą išskirtos taisyklės, apimančios to paties kaitomo žodžio skirtingas gramatines formas.

```
WHERE ((Žodynas1 = Žodynas2) && (Leksema1 = Leksema2)).
```

C. Išskirtos taisyklės, gautos iš skirtingų kaitomų žodžių žodynų. Papildoma sąlyga:

```
WHERE (Žodynas1 <> Žodynas2).
```

D. Visos likusios taisyklės. Šiuo atveju žodynai sutampa, kaip ir B grupėje, tačiau skiriasi leksemos.

4.7.2 Taisyklių grupių analizė

Dabar išsamiau panagrinėkime taisyklių grupes, jų turinį dar kartą sugrupavę pagal žodynų poras (žr. 4.3 lentelę).

Iš 4.3 lentelės A grupės matome, kad labai paprastai galima išspręsti nekaitomų žodžių vienareikšminimo problemą: jei koks nors nekaitomas žodis retesnis už kitą nekaitomą žodį ar kaitomo žodžio formą, tokį žodį galima tiesiog išmesti iš žodyno (1, 3, 5 eilutės). Šį veiksmą galima perkelti į anksčiau minėtą leksemų atmetimo etapą. Likusiems atvejams kirčiuoti gauname dvi itin paprastas taisykles: (Nek, *, *) > (DB, *, *) ir (Nek, *, *) > (Vks, *, *), kurios rodo, kad nekaitomas žodis dažnesnis už bet kokią gramatinę formą, padarytą iš daiktavardžio-būdvardžio ar veiksmažodžio kamieno. Toks sprendimas leidžia teisingai parinkti kirčiavimo variantą net 98,1% tikslumu.

4.3 lentelė. Taisyklių grupių turinys.

Taisyklių grupė	Žodynų pora	Taisyklių skaičius	Teisingų hipotezių skaičius	Klaidingų hipotezių skaičius	Teisingų hipotezių procentas	Žodžių skaičius vienai taisyklei
A	DB>Nek	5	185	83	69,0%	53,6
	Nek>DB	44	2712	74	97,3%	63,3
	Vks>Nek	3	16	0	100%	5,3
	Nek>Vks	31	7885	41	99,5%	255,7
	Nek-Nek	3	315	22	93,5%	84,3
	Iš viso	87=>2	11113	220	98,1%	130,3
B	DB-DB	104	52690	12702	80,6%	628,8
	Vks-Vks	64	33911	6914	83,1%	637,9
	Iš viso	168	86601	19616	81.5%	632.2
C	DB>Vks	293	41007	4752	89,6%	156,2
	Vks>DB	209	16726	4468	78,9%	101,4
	Iš viso	502	57733	9220	86.2%	133.4
D	DB-DB	388	37074	7391	83,4%	114,6
	Vks-Vks	155	16040	2366	87,1%	118,7
	Iš viso	543	53114	9757	84,5%	115,8
Iš viso		1215	208561	38813	84,3%	

Mes norime, kad taisyklės būtų kuo labiau apibendrinančios, t. y. sudarytos iš kuo didesnio skaičiaus pavyzdžių. Kiek pavyzdžių vidutiniškai apibendrina viena taisyklė, nurodyta dešiniajame 4.3 lentelės stulpelyje. Kaip matome B grupės taisyklės yra labiau apibendrinančios (vidutiniškai 632,2 pavyzdžiai), negu kitų grupių. Tai rodo, kad homografai dažniausiai yra to paties žodžio gramatinės formos, ko ir buvo galima tikėtis. Tačiau B grupę sudaro nedaug taisyklių ir jų tikslumas mažiausias.

Kalbant apie C grupę, galima pastebėti, kad daiktavardžių ir būdvardžių gramatinės formos dažnesnės už veiksmažodžių gramatines formas, todėl 502 (293+209) taisykles pakeitus viena taisykle (DB, *, *) > (Vks, *, *) vis vien tikslumas būtų 67,9% (41007 + 4468)/ (41007 + 4468 + 16726 + 4752).

Sudarant D grupės taisykles, iš veiksmažodžių žodyno buvo gautos ir 163 taisyklės, kuriose abi gramatinės formos sutapo. Žinoma, tokios taisyklės neturi jokios prasmės, todėl jos nebuvo įtrauktos į 4.3 lentelę.

Taigi sumažinus A grupės taisyklių skaičių iki 2, gauname 1215 taisyklių rinkinį, kuris mokymo duomenims leidžia teisingą kirčiavimo variantą parinkti 84,3% tikslumu.

4.7.3 Rezultatų palyginimas

Įdomu palyginti sukurtų taisyklių tikslumą su rezultatais, gautais naudojant kitus algoritmus. Palyginimui tinkamų rezultatų pavyko rasti tik darbe [Rimkutė, Grybinaitė, 2004], tiesa, jame eksperimentai atlikti su kitais duomenimis ir buvo vienareikšminamos visos, o ne tik skirtingai kirčiuojamos, gramatinės formos. Vienareikšminimui buvo panaudotas ID3 algoritmas, kuris remiasi gretimų žodžių morfologiniais požymiais. Šiame darbe sukurtos taisyklės buvo testuojamos naudojant 200.000 žodžių tekstus, kurie nebuvo naudoti sukuriant taisykles. Rezultatų palyginimas pateiktas 4.4 lentelėje.

4.4 lentelė. ID3 algoritmo ir dažniais grįstų taisyklių palyginimas

Sutampančių gramatinių formų pora	Tikslumas taikant ID3 algoritmą	Tikslumas taikant dažniais grįstas taisykles	Taisyklių skaičius
Moteriškosios giminės vienaskaitos kilmininkas ir daugiskaitos vardininkas	73,65%	77,47%	17+2
Moteriškosios giminės vienaskaitos vardininkas ir įnagininkas	81,65%	74,73%	4+4
Bendratis ir neveikiamosios rūšies būtojo laiko neįvardžiuotinių vyriškosios giminės dalyvių daugiskaitos vardininkas	92,15%	91,98%	1

Kiekvienai linksniuotei (linksniavimo paradigmai) buvo sukurta po vieną taisyklę. Taisyklių skaičius pateiktas lentelės dešiniajame stulpelyje, pvz., pirmoje eilutėje nurodyta, kad septyniolikai linksniavimo paradigmų dažnesnis buvo vienaskaitos kilmininkas, o dviem linksniavimo paradigmoms – daugiskaitos vardininkas.

Kaip matyti iš 4.4 lentelės, rezultatai gana panašūs, gal tik nežymiai blogesni.

4.7.4 Taisyklių skaičiaus sumažinimas

Skirtumas tarp B ir D grupės taisyklių yra tas, kad B grupėje leksemų identifikatoriai sutampa, o D grupėje skiriasi. Palyginus šias grupes rasta, kad 53 taisyklės sutampa, o septyniais atvejais B ir D grupėse buvo viena kitai priešingos taisyklės (t. y. tokios, kurios sutaptų sukeitus kairę pusę su dešine). Sutampančias taisykles galima sujungti. Vietoj dviejų priešingų taisyklių palikus tik vieną, dažniau vartojamą, tikslumas sumažėtų, tačiau nežymiai (mažiau nei 0,1%). Taigi galima taisyklių skaičių sumažinti 60 taisyklių arba 4,9%.

Kadangi likusių taisyklių skaičius vis vien gana didelis (1155) ir skiriasi taisyklių naudojimo dažnis, tai įdomu panagrinėti, kaip keičiasi kirčiavimo tikslumas mažinant taisyklių skaičių, t. y. atsisakant taisyklių, kurioms skirtumas tarp teisingų ir klaidingų hipotezių skaičiaus yra mažiausias. Atsisakius kai kurių taisyklių, daliai žodžių nebus surasta nė vienos taisyklės. Laikysim, kad tokiems žodžiams vieną kirčiavimo variantą parinksim atsitiktinai ir taip parinkdami suklysim 50% atvejų (nes daugumą homografų galima kirčiuoti dviem būdais). Kai naudojamos visos taisyklės, tikslumas yra 84,3%. 4.5 lentelėje ir 4.1 pav. pavaizduota, kiek taisyklių galima atsisakyti, kad tikslumas sumažėtų tam tikru fiksuotu procentu (nuo 0,1 iki 10%).

Iš 4.5 lentelės ir 4.1 pav. matome, kad palikus 40% taisyklių, tikslumas sumažės tik 0,9%, o palikus tik 7% taisyklių – sumažės 10%.

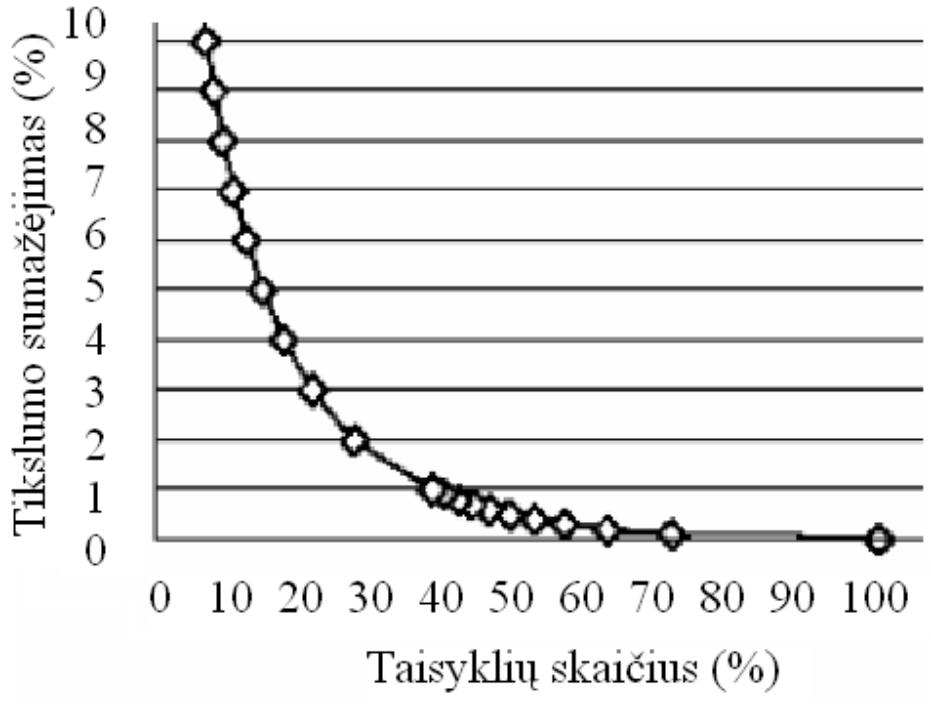
100% patikimumą turi 483 taisyklės (41,89% visų taisyklių). Toliau pavyzdžiuose pateikti žodžiai nusako tam tikrą linksniuotę ar asmenuotę ir tam tikrą gramatinę formą, o ne konkrečią leksemą, pvz.:

KARÃLIAUS (176) > KARALIAÛS (0);

VALDŽIÕS (168) > VALÕDŽIOS (0).

4.5 lentelė. Tikslumo priklausomybė nuo taisyklių skaičiaus

Tikslumo sumažėjimas procentais	Likusių taisyklių procentas	Likusių taisyklių skaičius
0	100%	1153
0,1	71%	822
0,2	62%	720
0,3	57%	652
0,4	52%	603
0,5	49%	563
0,6	46%	531
0,7	44%	504
0,8	42%	481
0,9	40%	459
1	38%	440
2	27%	315
3	22%	248
4	18%	202
5	15%	168
6	12%	142
7	10%	121
8	9%	104
9	8%	90
10	7%	77
29,6	0%	1



4.1 pav. Tikslumo priklausomybė nuo taisyklių skaičiaus

Blogiausią (50%) patikimumą turi 44 taisyklės (3,82% visų taisyklių), pvz.:

NETEISĖ́TAS (13) > NETÉISĖTAS (13).

Tam, kad būtų naudojamos tik taisyklės, kurių patikimumas yra bent 75%, reikia paimti 880 taisykles (76,32% visų taisyklių). Tokiu atveju lieka tokios taisyklės kaip, pvz.:

LÉIDŽIAMA (137) > LEIDŽIAMÀ (27), patikimumas = 137/(137+27) = 83,54%,

o nenaudojamos tokios taisyklės, kaip, pvz.:

NÚOSAVA (1689) > NUOSAVÀ (926), patikimumas = 64,59%.

Tam, kad būtų naudojamos tik taisyklės, kurių patikimumas yra bent 60%, reikia paimti 1031 taisykles (89,42% visų taisyklių). Tokiu atveju lieka tokios taisyklės kaip, pvz., jau minėta taisyklė:

NÚOSAVA (1689) > NUOSAVÀ (926), patikimumas = 64,59%,

o nenaudojamos tokios taisyklės, kaip, pvz.:

TOKIÕS (250) > TÓKIOS (172), patikimumas = 59,24%.

4.7.5 Taisyklių grupavimas remiantis lingvistiniais kriterijais

Kaip matėme 4.7.4 skyrelyje, gaunamas gana didelis gramatinėmis formomis paremtų taisyklių skaičius (daugiau kaip 1000), todėl šiame skyrelyje panagrinėsime, kaip būtų galima jį sumažinti grupuojant kai kurias taisykles pagal lingvistinius kriterijus. Sugrupavus taisykles galima gauti tokį patį arba mažesnį tikslumą, todėl po grupavimo gautoms taisyklėms įvertinti (kaip ir 4.6 skyrelyje) apsiribosime tik teisingų/klaidingų hipotezių skaičiavimu, o teksto kirčiavimo testo nebeatliksime. Kadangi šiuo atveju testavimui nebereikalingas mokymui nenaudotas tekstas, todėl taisykles sudarysime naudodami ne 800.000 žodžių, o 1.000.000 žodžių.

Taigi vėl nagrinėjame hipotezių poras iš 4.1 lentelės. Pirmiausia panagrinėkime atvejį, kai abi hipotezės gautos iš DB žodyno; be to, sudarant 4.1 lentelę buvo reikalaujama, kad kamienai sutaptų, t. y. nagrinėjame to paties žodžio (daiktavardžio-būdvardžio) gramatines formas. Akivaizdus taisyklių

grupavimas pagal linksnių kombinacijas, atsisakant informacijos apie linksniuotes. Kai kurioms linksnių kombinacijoms randame, kad vienas linksnis visada dažnesnis už kitą (nepriklausomai nuo linksniuotės), pvz., vienaskaitos vietininkas visada dažnesnis už vienaskaitos šauksmininką (*namè* vs. *nãme*), o kai kurioms linksnių kombinacijoms vienų linksniuočių dažnesnis vienas linksnis, o kitų – kitas, pvz., vienaskaitos kilmininkas dažnesnis už daugiskaitos vardininką nelyginamojo laipsnio moteriškos giminės būdvardžiams, bet retesnis aukštesnio laipsnio būdvardžiams (*gerõs* dažnesnis už *gẽros*, bet *geresnẽs* retesnis už *gerèsnės*). Yra labai paprastas būdas, kaip tokiais atvejais šiek tiek sumažinti taisyklių skaičių neprarandant tikslumo. Tarkime yra *n* taisyklių (t. y. *n* linksniuočių), kurioms linksnis *a* dažnesnis už linksnį *b*, ir *m* taisyklių, kurioms *b* dažnesnis už *a*, ir $n > m$. Tada *m* taisyklų paliekame nepakeistas ir jas taikome pirmiausia, o visiems kitiems atvejams sukuriame vieną nuo linksniuotės nebepriklausančią taisyklę, nusakančią, kad *a* dažnesnis už *b*. Taigi taisyklių skaičių galima sumažinti nuo $m + n$ iki $m + 1$. Galiausiai galima visai atsisakyti informacijos apie linksniuotes ir tam tikrai linksnių kombinacijai liks tik viena taisyklė. 4.6 lentelėje, 1 dalyje, pateiktos visos galimos linksnių kombinacijos, duodančios sutampančias gramatines formas, taisyklių skaičius, tikslumas prieš grupavimą ir sugrupavus į vieną taisyklę.

Pritaikius ką tik aprašytą taisyklių sumažinimo būdą, taisyklių skaičių galima sumažinti nuo 104 iki 24, o kiekvienai linksnių kombinacijai paliekant po vieną taisyklę liks 13 taisyklių, tačiau šiuo atveju tikslumas sumažėja nuo 80,6% iki 76,9%.

4.6 lentelė. Taisyklių grupavimo eksperimentų rezultatai

Žodynų pora	Gramatinių formų pora	Taisyklių skaičius	Teisingų/ klaidingų hipotezių skaičius	Tikslumas negrupuojant	Tikslumas sugrupavus	Pavyzdžiai
1 dalis. DB-DB	vns_V>vns_Įn	4	2517/470	74,0%	57,1%	galvà> gálva
	vns_Įn>vns_V	4	2348/1236			pìlna> pilnà
	vns_V>vns_Š	2	9536/748	92,7%	92,7%	galvà> gálva
	vns_V>dgs_G	11	3056/757	76,3%	74,7%	padėtìs> pãdėtis
	dgs_G>vns_V	2	384/310			dẽšimtis> dešimtìs
	bev.g.>vns_V	2	2753/1862	64,3%	51,8%	mãža> mažà
	vns_V>bev.g.	4	807/118			tuščià> tùščia
	vns_K>dgs_V	17	16268/4548	77,7%	77,5%	galvõs> gálvos
	dgs_V>vns_K	1	236/198			dìdelės> didelẽs
	vns_N>vns_Vt_tr	15	2805/13	99,5%	99,5%	naujám> naujam̃
	vns_G>dgs_K	2	269/130	67,4%	67,4%	sū́nų> sūnų̃
	vns_Įn>vns_Š	6	9936/533	94,9%	94,9%	ponià> põnia
	vns_Įn>bev.g.	5	1361/616	68,4%	67,7%	mažà> mãža
	bev.g.>vns_Įn	1	51/36			añtra> antrà
	vns_Vt>vns_Š	2	4223/130	97,0%	97,0%	miškè> mìške
	dgs_N>vns_Įn_tr	2	13/7	65,0%	65,0%	sūnùm> sūnum̃
	dgs_N>dgs_Įn_tr	20	539/128	80,8%	80,8%	dvíem> dviẽm
	priev.>vns_N	3	7169/3793	68,2%	59,2%	mažaĩ> mãžai
	vns_N>priev.	1	1227/125			tùščiai> tuščiaĩ
	Iš viso dažnesnių	91	60445/13735	80,6% 24 tais	76,9% 13 tais	
	Iš viso retesnių	13	5053/2023			

2 dalis. DB-Vks	V>vns_V	33	6350/810	81,8%	53,7%	móki> mokì
	vns_V>V	50	6421/2025			kalbà> kal̃ba
	vns_K>V	46	10800/839	88,4%	83,9%	tvarkõs> tvar̃kos
	V>vns_K	18	1381/764			atsãko> ãtsako
	vns_N>V	12	892/93	91,4%	51,2%	šveñtei> šventeĩ
	V>vns_N	10	821/68			manaĩ> mãnai
	vns_G>V	25	5090/564	90,1%	86,4%	kar̃tą> kártą
	V>vns_G	8	238/19			gãlinčių> galiñčių
	vns_Į̃n>V	32	5489/1261	81,9%	73,4%	šviesiù> šviẽsiu
	V>vns_Į̃n	12	772/120			jaučiù> jáučiu
	vns_Vt>V	11	541/73	91,1%	59,3%	kuriamè> kùriame
	V>vns_Vt	17	305/10			apkláusime> apklausimè
	vns_Š>V	6	2428/243	88,7%	53,1%	mótina> motinà
	V>vns_Š	29	2209/347			apkláusime> apklausìme
	dgs_V>V	46	7903/834	80,3%	73,7%	var̃tai> vartaĩ
	V>dgs_V	22	2611/1746			kartõs> kar̃tos
	V>dgs_K	8	3187/211	91,2%	61,5%	apibrė́žtų> apibrėžtų̃
	dgs_K>V	11	1966/289			kar̃tų> kártų
	dgs_N>V	6	75/2	91,1%	74,1%	mótinoms> motinóms
	V>dgs_N	12	27/8			bãram> barám
	dgs_G>V	26	3893/352	87,7%	78,1%	ausìs> aũsis
	V>dgs_G	19	827/312			skaĩčius> skaičiùs
	dgs_Į̃n>V	13	1269/51	95,7%	94,8%	purvinaĩs> pur̃vinais
	V>dgs_Į̃n	8	19/7			laĩkom> láikom
	dgs_Vt>V	6	401/32	93,5%	81,7%	júostose> juostosè
	V>dgs_Vt	5	58/0			skaičiuõs> skaĩčiuos
	bev.g.>V	8	3918/559	87,4%	86,6%	

	V> bev.g.	2	51/11			móku> mokù
	priev.>V	6	1022/139	92,0%	52,7%	palaidaĩ> paláidai
	V>priev.	7	796/20			tãriamai> tariamaĩ
	aukšt.l.>V	3	650/50	92,9%	91,8%	arčiaũ> árčiau
	V>aukšt.l.	1	8/0			mókiau> mokiaũ
	Iš viso dažnesnių	287	53908/6113	85,9% 127 tais	70,8% 16 tais	
	Iš viso retesnių	231	18510/5746			
3 dalis.Vks-Vks(dalyviai)	vns_Įn>vns_V	4	3419/1033	70,4%	69,9%	suprañtama> suprantamà
	vns_V>vns_Įn	2	752/719			būsimà> bū́sima
	vns_V>dgs_G	1	254/14	94,8%	94,8%	keĩstas> keistàs
	bev.g.>vns_V	4	3419/1033	67,1%	64,6%	suprañtama> suprantamà
	vns_V>bev.g.	2	1632/1442			baigtà> baĩgta
	dgs_V>vns_K	6	1732/663	72,4%	69,0%	ìštemptos> ištemptõs
	vns_K>dgs_V	6	137/50			būsimõs> bū́simos
	vns_Įn>bev.g.	1	880/723	54,9%	54,9%	baigtà> baĩgta
	priev&būd.>vns_N	4	395/68	85,3%	85,3%	patikimaĩ> pàtikimai
	Iš viso dažnesnių	20	10099/3534	68,7% 16 tais	67,0% 6 tais	
	Iš viso retesnių	10	2521/2211			
4 dalis. Vks-Vks	bendr.>!bendr.	3	18510/1582	92,1%	92,1%	įsteĩgti> įsteigtì
	es.l.>!es.l.	21	905/43	95,5%	95,5%	užmirštù> ùžmirštu
	es.l.IIIa>es.l.IIa	1	5252/308	94,5%	94,5%	gir̃di> girdì
	es.l.IIIa>liep.n.	1	654/60	91,6%	91,6%	žinaĩ> tesižìnai
	tar.n.>dal.dgs_K	1	4402/938	82,4%	82,4%	supìltų> supiltų̃
	likusios taisyklės	9	67/28			
	Iš viso	36	29790/2959	91,0% 14 tais	91,0% 5 tais	
Iš viso dažnesnių		434	154242/26341	83,2% 181 tais	75,8% 40 tais	
Iš viso retesnių		254	26084/9980			

Kitas atvejis, kai viena gramatinė forma daiktavardis ar būdvardis, o kita – veiksmažodis (žr. 4.6 lentelę, 2 dalį). Tokias gramatinių formų poras galima būtų grupuoti pagal daug ir įvairių tiek daiktavardžių-būdvardžių, tiek veiksmažodžių kriterijų. Panagrinėkim tik vieną – grupuoti daiktavardžius-būdvardžius pagal linksnį. Randame, kad vienaskaitos vardininko ir daugiskaitos kilmininko linksniai retesni už veiksmažodžių formas, o visi kiti linksniai dažnesni. Tokių taisyklių yra 518, sugrupavus pagal linksnius ir pritaikius minėtą metodą, taisyklių skaičių galima sumažinti iki 127, o linksniui paliekant po vieną taisyklę lieka 16 taisyklių. Tikslumas šiuo atveju krenta ženkliai – nuo 85,9% iki 70,8%.

Dar vienas atvejis, kai abi gramatinės formos padarytos iš veiksmažodžių kamienų (prie tų formų priskirtini ir dalyviai). Dalyvius galima sugrupuoti analogiškai, pagal linksnius (žr. 4.6 lentelę, 3 dalį). Gauname 30 taisyklių, jų skaičių galima sumažinti iki 16, linksniui paliekant po vieną taisyklę liks tik 6 taisyklės, tačiau šiuo atveju tikslumas krenta nežymiai – nuo 68,7% iki 67,0%. Dar galima atkreipti dėmesį, kad dvi linksnių kombinacijos elgiasi priešingai, nei daiktavardžių-būdvardžių atveju (vns_Įn>vns_V ir dgs_V>vns_K). Kitų veiksmažodžių gramatinių formų grupavimas toli gražu nėra akivaizdus, tačiau pavyko pastebėti keletą įdomių dėsningumų: a) bendratis dažnesnė už kitas gramatines formas; b) esamojo laiko asmenuojamosios formos dažnesnės už kitas gramatines formas. Šiuos dėsningumus galima būtų realizuoti kaip specialias taisykles, apimančias gramatinių formų intervalus. Lieka 12 taisyklių, iš kurių tik 3 pakankamai dažnos, ir net atmetus kitas 9 tikslumas labai nenukentėtų – sumažėtų 0,06%.

4.8 Teksto kirčiavimo eksperimentai

Kuo didesnis tam tikros taisyklės teisingų hipotezių kiekis (procentais) mokymo duomenyse, tuo tikslesnio testinio teksto kirčiavimo galime tikėtis. Priklausomybė tarp teisingų hipotezių procento ir teksto kirčiavimo tikslumo nėra tiesioginė. Tai lemia ne tik neatitikimas tarp mokymo ir testinių duomenų, bet ir tas faktas, kad kai kuriems žodžiams generuojama daugiau negu dvi

(teisinga ir klaidinga) hipotezės, bet ir sudėtingesni jų deriniai, todėl ir jų sąveika tampa sudėtingesnė. Kaip taisyklės sąveikauja ir koks realus kirčiavimo tikslumas pasiekiamas, galima sužinoti tik jas naudojant tekstui kirčiuoti. Buvo atlikti trys kirčiavimo eksperimentai su trim skirtingais taisyklių, kurios remiasi hipotezių dažniais, rinkiniais: 1) jokios taisyklės nenaudojamos, 2) grupės A, B ir C, 3) grupės A, B, C ir D. Kiekvienas iš šių eksperimentų pakartotas po du kartus: pirmuoju atveju žodžiai, turintys daug kirčiavimo variantų, laikomi klaidingai kirčiuotais, antruoju atveju imamas pirmasis kirčiavimo variantas (tai ekvivalentu atsitiktiniam vieno varianto parinkimui). Visi minėti kirčiavimo eksperimentai pakartoti po tris kartus su skirtingais žodynais: 1) naudoti pilni žodynai, 2) naudotas pirmasis leksemų atmetimo būdas, 3) naudotas pirmasis ir antrasis būdai. Hipotezių dažniais pagrįstos taisyklės visuose eksperimentuose buvo sudaromos tik iš pilnų žodynų (t. y. netaikant jokio leksemų atmetimo algoritmo), nors leksemų atmetimo poveikis jau pajuntamas kirčiavimo algoritmui grąžinant kirčiavimo variantus. Taigi iš viso atlikta 18 eksperimentų. Rezultatai pateikti 1 priede. Gauti rezultatai užima daug vietos ir nėra labai vaizdūs. Mus labiausiai domina, kiek žodžių leidžia vienareikšminti vienas ar kitas algoritmas ir kokiu tikslumu jis tai daro. Tuo tikslu iš teksto išsirinksime tik žodžius, kuriuos galima kirčiuoti keliais būdais ir bent vienas kirčiavimo variantas teisingas, suskaičiuosime, kiek žodžių iš šio sąrašo galima vienareikšminti naudojant tam tikrą algoritmą ir kokia dalis buvo vienareikšminta teisingai.

Pirmiausiai panagrinėkime, kaip veikia hipotezių dažniais pagrįstos taisyklės palyginti su atveju, kai taisyklės nenaudotos. Jos taikomos žodžiams, likusiems pritaikius leksemų atmetimo algoritmą. Atsitiktinis varianto parinkimas nenaudotas. Rezultatai 4.7 lentelėje.

4.7 lentelė. Vienareišminimas naudojant hipotezių dažniais pagrįstas taisykles

Hipotezių dažniais paremtos taisyklės	Leksemų atmetimo algoritmas nenaudotas		Leksemų atmetimo 1 algoritmas		Leksemų atmetimo 1 ir 2 algoritmai	
	Vienareikšminta žodžių	Teisingų	Vienareikšminta žodžių	Teisingų	Vienareikšminta žodžių	Teisingų
Grupės A, B ir C	24179	84,9%	24155	85,2%	19934	84,1%
Grupės A, B, C, D	29170	84,9%	29138	85,1%	20810	83,8%

Iš 4.7 lentelės matome, kad: 1) taikant leksemų atmetimo 1 algoritmą tikslumas padidėja 0,2–0,3%, o taikant 1 ir 2 – sumažėja 0,8–1,1%; 2) grupės A, B ir C paveikia žymiai daugiau žodžių nei D; 3) taikant grupę D tikslumas nežymiai krenta (iki 0,3%).

Dabar panagrinėkime, kokį tikslumą galima pasiekti, jei po visų algoritmų taikymo iš likusių variantų vieną parinksime atsitiktinai. Remdamiesi intuicija, manytume, kad tikslumas turėtų būti apie 50%. Rezultatai 4.8 lentelėje. Juos galima būtų apibendrinti taip: kuo mažiau variantų likę (išskyrus atvejį, kai naudojamos tik grupės A, B ir C), tuo didesnį tikslumą duoda atsitiktinis parinkimas. Tai rodo, kad tarp likusių hipotezių daugiau teisingų hipotezių.

4.8 lentelė. Vienareišminimas atsitiktinai parenkant vieną kirčiavimo variantą

Hipotezių dažniais paremtos taisyklės	Leksemų atmetimo algoritmas nenaudotas		Leksemų atmetimo 1 algoritmas		Leksemų atmetimo 1 ir 2 algoritmai	
	Vienareikšminta žodžių	Teisingų	Vienareikšminta žodžių	Teisingų	Vienareikšminta žodžių	Teisingų
Taisyklės nenaudotos	30197	52,5%	29517	52,3%	20967	63,5%
Grupės A, B ir C	6018	44,7%	5362	45,4%	1033	52,7%
Grupės A, B, C, D	1027	53,3%	379	67,8%	157	68,2%

Galiausiai buvo įvertinta leksemų atmetimo algoritmų įtaka. Tačiau 2 algoritmas gali paveikti ne tik kelis kirčiavimo variantus turinčius žodžius, tačiau ir kitus žodžius, pavyzdžiui, kirčiuotą žodį paversti nekirčiuotu. Žinoma, nedidelę dalį žodžių iš tikrųjų reikia palikti nekirčiuotą. Taigi 2 algoritmui

greta vienareikšminimo tikslumo dar buvo skaičiuojamas apibendrintas tikslumas, kurį galima užrašyti formule: *(tv+tn) / (tv+tn+kv+kn)*, kur

tv – teisingai vienareikšmintų žodžių skaičius,

tn – teisingai paliktų nekirčiuotais žodžių skaičius,

kv – klaidingai vienareikšmintų žodžių skaičius,

kn – klaidingai paliktų nekirčiuotais žodžių skaičius.

Rezultatai 4.9 lentelėje.

4.9 lentelė. Vienareišminimas naudojant leksemų atmetimo algoritmus

Hipotezių dažniais paremtos taisyklės	Leksemų atmetimo 1 algoritmas		Leksemų atmetimo 1 ir 2 algoritmai			
	Vienareikšminta žodžių	Teisingų	Vienareikšminta žodžių	Teisingų	Padaugėjo nekirčiuotų	Apibendrintas tikslumas
Taisyklės nenaudotos	681	91,04%	8928	95,71%	1266	84,24%
Grupės A, B ir C	740	92,03%	5569	92,48%	1213	76,56%
Grupės A, B, C, D	739	92,15%	1986	85,10%	1204	54,29%

Galima daryti išvadą, kad leksemų atmetimo 2 algoritmą verta taikyti tik tuo atveju, jei nenaudojamos taisyklės.

Turint tikslą visiems daug kirčiavimo variantų turintiems žodžiams parinkti vieną variantą geriausi rezultatai pasiekti, kai iš pradžių taikytas leksemų atmetimo 1 algoritmas, tada taisyklių grupės A, B, C ir D ir galiausiai iš likusių atsitiktinai parinktas vienas variantas. Kiek žodžių vienareikšmino kiekvienas algoritmas ir kokiu tikslumu, pateikta 4.10 lentelėje.

4.10 lentelė. Geriausius rezultatus pasiekusi vienareikšminimo algoritmų seka

Algoritmas	Vienareikšminta žodžių	Teisingų
Leksemų atmetimo 1 algoritmas	680	91,18%
Taisyklių grupės A, B, C ir D	29138	85,09%
Atsitiktinis varianto parinkimas	379	67,81%
Iš viso	30197	85,01%

Taigi iš viso testiniuose tekstuose buvo 30197 homografai (15,30% visų žodžių), teisingą variantą pavyko nustatyti 85,01% tikslumu.

4.9 Ketvirtojo skyriaus apibendrinimas

- Lietuvių kalbos tekstui kirčiuoti pritaikius morfologinėmis taisyklėmis grįstus metodus, kai kuriuos žodžius (homografus) galima sukirčiuoti keliais būdais. Norint padidinti kirčiuotų žodžių skaičių, reikalingas vienareikšminimo algoritmas.
- Šiame skyriuje pasiūlytas homografų vienareikšminimo algoritmas, pagrįstas leksemų ir morfologinių pažymų dažniais.
- Kai kurių leksemų buvimas žodyne labiau kliudo kirčiuoti, nei padeda. Jas atmetus, teisingų kirčiavimo hipotezių dalis tarp visų hipotezių (mokymo duomenims) padidėja 6,1%.
- Sudarytas morfologinių pažymų dažniais grįstų taisyklių rinkinys (1215 taisyklių), kuris mokymo duomenims leidžia teisingą kirčiavimo variantą parinkti 84,3% tikslumu.
- Jei naudojamos morfologinių pažymų dažniais grįstos taisyklės, verta atmesti tik tas leksemas, kurių sutampa ir kamienai, ir linksniuotės. Atmetus daugiau leksemų, rezultatai pablogėja – daugiau žodžių lieka nekirčiuota.
- Iš galimų kirčiavimo hipotezių atmetus rečiau vartojamas, tarp likusių hipotezių didėja teisingą kirčiavimą nusakančių hipotezių dalis (net iki 68%), todėl iš likusių hipotezių vieną variantą verta parinkti atsitiktinai.
- Nors morfologinių pažymų dažniais grįstose taisyklėse nenaudojama jokia informacija apie kontekstą, tačiau jos geba vienareikšminti kai kurias gramatines formas tikslumu, artimu kontekstinę informaciją naudojančiam ID3 algoritmui.
- Pritaikius pasiūlytus algoritmus tekstui kirčiuoti, homografus pavyko vienareikšminti 85,01% tikslumu.

5 Lietuvių kalbos žodžių kirčiavimas naudojant sprendimo medžius

Šiame skyriuje pagrįstas naujų kirčiavimo algoritmų kūrimo aktualumas, aprašytas sprendimo medžių algoritmas bei pristatomas paties autoriaus pasiūlytas lietuviško teksto kirčiavimo metodas, kuris, naudodamas sprendimų medžius, randa raidžių sekas, vienareikšmiškai nusakančias žodžio kirčiavimą. Buvo sudarytos kirčiavimo taisyklės remiantis raidžių sekomis žodžių pradžioje, pabaigoje ir viduryje. Išnagrinėti taisyklių skaičiaus sumažinimo būdai bei įvertintas kirčiavimo tikslumas, naudojant skirtingus taisyklių rinkinius. Taip pat siūlomas metodas buvo palygintas su morfologine analize grįstu metodu.

5.1 Naujų kirčiavimo algoritmų poreikis

Kaip jau buvo minėta 2 skyriuje, laisvąjį kirtį turinčioms nefleksinėms kalboms kirčiuoti paprastai naudojami kirčiuotų žodžių sąrašais grįsti metodai, laisvojo kirčio fleksinėms kalboms – morfologinėmis žodžių kaitymo taisyklėmis grįsti metodai. Tokie metodai iki šiol buvo išimtinai taikomi ir lietuvių kalbai. Tačiau lietuvių kalbos morfologijos taisyklės yra sudėtingos, todėl, jas realizavus kompiuterinės programos pavidalu, programa būna sudaryta iš daugelio kodo eilučių ir todėl sudėtinga. Tai sukelia problemų perkeliant tokį algoritmą iš vienos programavimo kalbos į kitą (pvz., iš *C++* į *Java*) ar iš vienos operacinės sistemos (**OS**) į kitą (pvz., iš *Windows* į *Symbian*). Kitas išliekantis aktualus klausimas – tai algoritmo darbo greitis. Nors egzistuojantys algoritmai sugeba kirčiuoti tekstą realiu laiku, tačiau jei kirčiavimo algoritmas įkomponuotas į serveryje esantį sintezatorių, kuriam perduodamas didelis kiekis teksto ir jis turi grąžinti sintezuotą balsą, tai savaime suprantamas noras, kad kirčiavimo algoritmas užimtų kuo mažiau procesoriaus darbo laiko.

Šiame darbe laisvojo kirčio fleksinei lietuvių kalbai pirmą kartą pritaikytas metodas, kuris nesinaudoja jokia informacija apie žodį sudarančias morfemas, jo kaitymą, priklausymą tam tikrai kalbos daliai, skiemenų ribas ir pan. Pasiūlytas metodas, naudodamas sprendimų medžius, randa raidžių sekas, kurios vienareikšmiškai nusako žodžio kirčiavimą. Šiuo atveju kirčiavimo taisyklės sudaromos automatiškai, jei yra prieinamas pakankamas kiekis kirčiuotų tekstų. O pats kirčiavimo algoritmas yra itin paprastas, greitas, gali būti nesunkiai pritaikomas kitoms pasaulio kalboms bei lengvai perkeliamas į kitas programavimo kalbas bei OS.

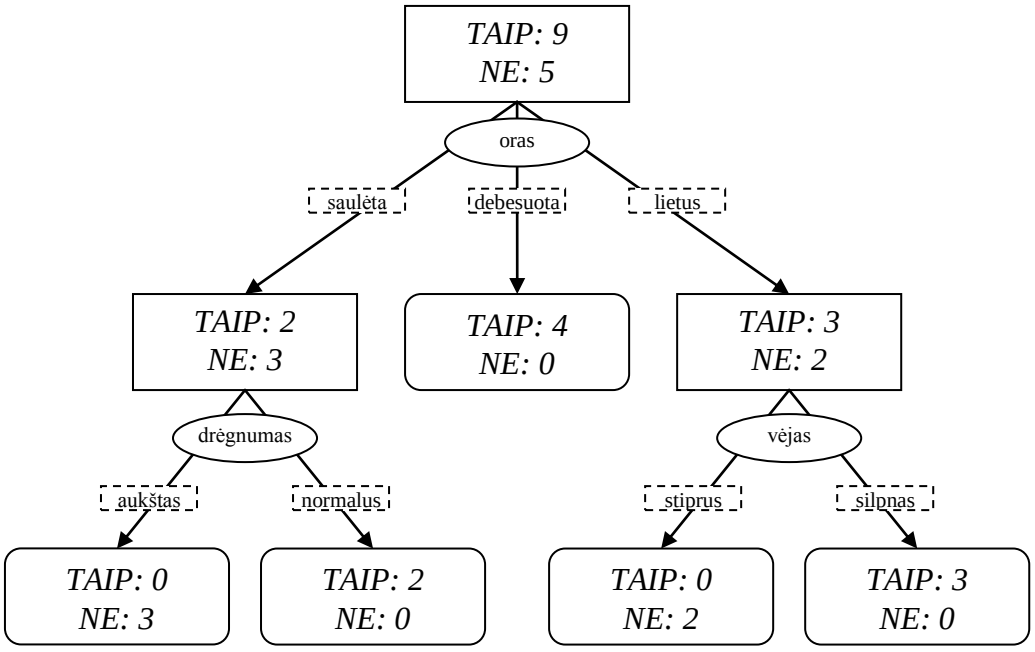
5.2 Sprendimo medžiai

5.2.1 Sprendimo medžių sąvoka

Kompiuterių moksle **medžiu** vadinama populiari duomenų struktūra, kuri realizuoja hierarchinę „tikrojo“ medžio struktūrą, naudodama aibę mazgų ir šiuos mazgus jungiančių briaunų [http://mathworld.wolfram.com/Tree.html, žiūrėta 2010.05.03]. **Mazge** gali būti saugoma informacija, pvz., reikšmė, sąlyga ir t. t. Kiekvienas medžio mazgas gali neturėti nei vieno arba turėti vieną ar daugiau **vaikų**. Mazgai sujungiamai su savo vaikais **briaunomis (**arba **šakomis**). Mazgas, turintis vaikų, vadinamas šių vaikų **tėvu**. Bet kuris medžio mazgas gali turėti tik vieną tėvą. Vienintelis mazgas, neturintis tėvo, – tai **šaknis**. Kiekvienas medis turi tik vieną šaknį. Mazgai, turintys vaikų, vadinami **vidiniais** mazgais, neturintys vaikų – **lapais**. **Kelias** medyje – tai briaunomis sujungtų mazgų seka. Medyje tarp dviejų mazgų egzistuoja vienas ir tik vienas unikalus kelias. Kelio ilgis tai mazgų kelyje skaičius. Mazgo **gylis** (arba **lygis**) – tai kelio nuo šio mazgo iki šaknies ilgis. **Eiti medžiu** reiškia judėti tam tikru apibrėžtu keliu medyje. Sąvoka **medis** dar yra naudojama ir kitose matematikos srityse, pvz., grafų teorijoje, aibių teorijoje ir kt.

Nors ši struktūra vadinama medžiu, dažniausiai grafiškai ji vaizduojama šaknimi į viršų (žr. 5.1 pav.).

Medžiai gali būti naudojami įvairių problemų sprendimams saugoti ir pavaizduoti. Medžiai, kuriuose sprendimai saugomi medžio lapuose, o kelias nuo šaknies iki lapo nusako pasirinkimų seką iki sprendimo, vadinami **sprendimo medžiais** (žr. 5.1 pav.).



5.1 pav. Sprendimo medis, skirtas prognozuoti, ar įvyks beisbolo rungtynės priklausomai nuo oro sąlygų. Medis sudarytas pagal 5.1 lentelės duomenis. Stačiakampiais pažymėti vidiniai mazgai, apvaliais stačiakampiais – sprendimo mazgai (lapai). Ovalai žymi parametrus, o stačiakampiai punktyru – parametrų reikšmes.

Duomenų gavybos (angl. *data mining*) arba mašininio mokymo (angl. *machine learning*) srityse sprendimo medžiai dar vadinami **klasifikavimo ir regresijos medžiais** (CART, angl. *Classification & Regression Tree*) ir naudojami prognozuoti kintamojo *y* reikšmei, atitinkančiai parametrų vektorių $f(p_1, p_2, ..., p_n)$. Pvz., 5.1 pav. dienos oro sąlygų stebėjimo parametrų vektorius $f(p_{oras}, p_{drėgnumas}, p_{vėjas})$ yra naudojamas prognozuoti, ar tą dieną įvyks beisbolo rungtynės (kintamasis *y*).

Jei kintamasis *y*:

1) nusako klasę arba kategoriją (t. y. priklauso natūrinių arba sveikųjų skaičių aibėms), toks medis vadinamas **klasifikavimo** medžiu (5.1 pav.

kintamasis *y* gali turėti dvi reikšmes: *TAIP* arba *NE*), o pats uždavinys – klasifikavimo uždaviniu;

2) priklauso realiųjų skaičių aibei, toks medis vadinamas **regresijos** medžiu, o pats uždavinys – regresijos uždaviniu; dažnai regresijos uždavinį galima pakeisti klasifikavimo uždaviniu padalinant kintamąjį *y* į intervalus.

Parametrų vektorių *f* gali sudaryti tiek sveikieji, tiek realieji skaičiai.

Taigi galimas kintamojo *y* reikšmes atitinka sprendimo medžio lapai. Medžio briaunos atitinka galimas parametrų vektoriaus $f(p_1, p_2, ..., p_n)$ reikšmes. Keliai nuo sprendimo medžio šaknies iki lapų atitinka parametrų reikšmių kombinacijas, kurios veda prie šių lapų sprendimų. Vidinius sprendimo medžio mazgus vadinsime **pasirinkimo** mazgais, o medžio lapus – **sprendimo mazgais**.

Toliau kalbėsime tik apie klasifikavimo (bet ne regresijos) medžius.

Sprendimo medžiai nesunkiai gali būti perrašomi taisyklių pavidalu, pvz., 5.1 pav sprendimo medis gali būti perrašytas taip:

```
JEI oras = saulėta IR drėgnumas = aukštas TAI žaisti = NE
JEI oras = lietus IR drėgnumas = aukštas TAI žaisti = NE
JEI oras = lietus IR vėjas = stiprus TAI žaisti = TAIP
JEI oras = debesuota TAI žaisti = TAIP
JEI oras = lietus IR vėjas = silpnas TAI žaisti = TAIP
```

Klasifikavimas gali būti atliekamas ne tik CART, bet ir labiau „tradiciniais“ metodais, pvz., tiesine diskriminantine analize (LDA, angl. *Linear Discriminant Analysis*).

5.2.2 Sprendimo medžių sudarymo algoritmai

Duomenų gavybos srityje sprendimo medžiai yra sudaromi automatiškai „apmokant“ medį iš duomenų. Pirmieji sprendimo medžių sudarymo algoritmai: C&RT [Breiman ir kt., 1984], CHAID (angl. *Chi-square Automatic Interaction Detector*) [Kass, 1980], ID3 [Quinlan, 1986]. Šį iki šiol populiarų algoritmą (ID3) apžvelgsime detaliau. Parametrų vektorių $f(p_1, p_2, ..., p_n)$ ir

šiam vektoriui atitinkantį rezultato (sprendimo) kintamąjį *y* sujungsime į vieną vektorių ir vadinsime apmokymo pavyzdžiu *s*, kur $s(f(p_1, p_2, ..., p_n), y) = s(p_1, p_2, ..., p_n, y)$, čia *n* žymi įvesties parametrų skaičių. Visą apmokymo duomenų aibę žymėsime *S*, kur $S = s_k(p_{1k}, p_{2k}, ..., p_{nk}, y_k)$, čia *k* žymi apmokymo pavyzdžių skaičių aibėje *S*. Šis algoritmas reikalauja, kad kiekvienas parametras p_i (kur $i = 1..n$) turėtų tik diskrečias reikšmes. Parametro p_i skirtingų reikšmių (aibėje *S*) aibę žymėsime $V_i = \{v_{i1}, v_{i2}, ..., v_{im}\}$, visų reikšmių aibę: $P_i = \{p_{i1}, p_{i2}, ..., p_{ik}\}$, čia $i = \{1, 2, ..., n\}$, skaičius *m* gali būti skirtingas kiekvienam p_i. T. y. visoms *i* ir *j* reikšmėms nuo 1 iki *n* yra teisinga, kad $|P_i|=|P_j|=k$, bet nebūtinai $|V_i|=|V_j|$ (čia $|P_i|$, $|V_i|$ ir t. t. žymi elementų skaičių aibėje P_i, V_i ir t. t.).

Toliau pateikiamas ID3 algoritmo pseudokodas:

1. Jei visi pavyzdžiai aibėje *S* priklauso vienai klasei (t. y. turi vienodas kintamojo *y* reikšmes), tai sukurti medžio lapą (sprendimo mazgą), įrašyti į jį kintamojo *y* reikšmę ir sustoti.

2. Priešingu atveju pasirinkti parametrą p_i, kuris turi **didžiausią informacinę naudą**, ir sukurti pasirinkimo mazgą. Tegul p_i reikšmių aibė $V_i = v_{i1}, v_{i2}, ..., v_{im}$.

3. Padalinti apmokymo pavyzdžių aibę *S* į poaibius $S_1, S_2, ..., S_m$ pagal reikšmes $v_{i1}, v_{i2}, ..., v_{im}$.

4. Taikyti algoritmą rekursyviai kiekvienai S_u aibei (čia $u = \{1, 2, …, m\}$).

Antrame šio algoritmo žingsnyje nurodyta, kad reikia pasirinkti **didžiausią informacinę naudą** turintį parametrą. Parametro p_i informacinės naudos (angl. *information gain*) požymis nusako, kaip gerai pasirinktas parametras p_i padalina apmokymo pavyzdžius į tikslines (angl. *targeted*) klases. Informacinės naudos sąvoka apibrėžiama remiantis **entropija**, kuri nurodo informacijos kiekį aibėje *S*. Entropija skaičiuojama ne parametrui, bet

visai apmokymo pavyzdžių aibei *S*. Jei kintamojo *y* skirtingų reikšmių aibė $\{y_1, y_2, ..., y_r\}$, tai:

$$Entropija(S) = \sum (-p(y_i) \log_2 p(y_i))$$

kur suma skaičiuojama visoms skirtingoms kintamojo *y* reikšmėms (t .y. nuo $i = 1$ iki r); $p(y_i)$ yra pavyzdžių, turinčių reikšmę y_i, dalis tarp visų aibės *S* pavyzdžių; $\log_2$ žymi logaritmą pagrindu 2.

Panagrinėkime dažnai literatūroje randamą pavyzdį. Tegul aibę *S* sudaro 5.1 lentelėje pateikti duomenys, t. y. 14 pavyzdžių, iš kurių 9 su rezultatu *TAIP* (y_1) ir 5 su rezultatu *NE* (y_2), vadinasi $r = 2$. Tada aibės *S* entropija:

$$Entropija(S) = - (9/14) \log_2(9/14) - (5/14) \log_2(5/14) = 0{,}940$$

Jei visi aibės *S* elementai priklauso tai pačiai klasei (duomenys idealiai klasifikuoti), tai entropijos reikšmė yra 0 (čia laikome, kad $\log_2 0 = 0$). Jei elementai aibėje *S* yra po lygiai pasiskirstę į klases, tai entropija lygi 1.

5.1 lentelė. Duomenys apie tai, ar įvyko beisbolo rungtynės priklausomai nuo oro sąlygų

Diena	**Oras**	**Tempe-ratūra**	**Drėgnumas**	**Vėjas (stiprumas)**	**Ar įvyko rungtynės?**
	p_1	p_2	p_3	p_4	y
s_1	Saulėta	Aukšta	Aukštas	Silpnas	NE
s_2	Saulėta	Aukšta	Aukštas	Stiprus	NE
s_3	Debesuota	Aukšta	Aukštas	Silpnas	TAIP
s_4	Lietus	Vidutinė	Aukštas	Silpnas	TAIP
s_5	Lietus	Žema	Normalus	Silpnas	TAIP
s_6	Lietus	Žema	Normalus	Stiprus	NE
s_7	Debesuota	Žema	Normalus	Stiprus	TAIP
s_8	Saulėta	Vidutinė	Aukštas	Silpnas	NE
s_9	Saulėta	Žema	Normalus	Silpnas	TAIP
s_{10}	Lietus	Vidutinė	Normalus	Silpnas	TAIP
s_{11}	Saulėta	Vidutinė	Normalus	Stiprus	TAIP
s_{12}	Debesuota	Vidutinė	Aukštas	Stiprus	TAIP
s_{13}	Debesuota	Aukšta	Normalus	Silpnas	TAIP
s_{14}	Lietus	Vidutinė	Aukštas	Stiprus	NE

5.1 lentelės stulpeliai nuo antro iki penkto saugo 14-os dienų oro sąlygų stebėjimo rezultatus (parametrai $p_1, p_2, ..., p_k$, $k = 4$). Paskutinis (šeštas) lentelės

stulpelis nurodo, ar tą dieną įvyko beisbolo rungtynės (kintamasis *y*). Oro sąlygų stebėjimo parametrai p_1, p_2, ..., p_k yra oras, temperatūra, drėgnumas ir vėjo stiprumas. Šie parametrai gali turėti tokias reikšmes: oras = {*saulėta, debesuota, lietus*}, temperatūra = {*aukšta, vidutinė, žema*}, drėgnumas = {*aukštas, normalus*}, vėjas = {*silpnas, stiprus*}. Paskutinis stulpelis (kintamasis *y*) gali turėti dvi reikšmes {*TAIP, NE*}.

Parametro p_i informacinė nauda aibėje *S* apibrėžiama kaip (čia parametras p_i kaip ir anksčiau gali turėti *m* skirtingų reikšmių):

$$Nauda(S, p_i) = Entropija(S) - \sum ((|S_u| / |S|) * Entropija(S_u))$$

kur suma skaičiuojama visoms skirtingoms parametro p_i reikšmėms (t .y. nuo $u = 1$ iki m); S_u – aibės *S* poaibis, kurio elementų parametro p_i reikšmė lygi u; $|S_u|$ – elementų skaičius aibėje S_u; $|S|$ – elementų skaičius aibėje *S*.

Grįžkime prie 5.1 lentelės duomenų: parametro *vėjas* reikšmė 8 kartus buvo *silpnas* ir 6 kartus *stiprus*. Kai vėjas silpnas, 6 kartus rungtynės įvyko ir 2 kartus neįvyko; kai vėjas stiprus, 3 kartus įvyko ir 3 kartus neįvyko. Taigi:

$$Nauda(S,p_{vėjas})=Entropija(S)-(8/14)*Entropija(S_{silpnas})-(6/14)*Entropija(S_{stiprus}) = 0,940 - (8/14)*0,811 - (6/14)*1,00 = 0,048$$

čia:

$$Entropija(S_{silpnas}) = - (6/8)*\log_2(6/8) - (2/8)*\log_2(2/8) = 0,811$$
$$Entropija(S_{stiprus}) = - (3/6)*\log_2(3/6) - (3/6)*\log_2(3/6) = 1,00$$

Analogiškai apskaičiavę likusių parametrų informacinę naudą gauname:

$$Nauda(S, p_{oras}) = 0,246$$
$$Nauda(S, p_{temperatūra}) = 0,029$$
$$Nauda(S, p_{drėgnumas}) = 0,151$$

Taigi iš 4 parametrų didžiausią informacinę naudą aibėje *S* turi parametras p_{oras}. Grįžtame prie ID3 algortimo: būtent parametras p_{oras} yra naudojamas „dalinant“ šakninį mazgą. Šis parametras gali turėti 3 skirtingas reikšmes: $\{v_{saulėta}, v_{debesuota}, v_{lietus}\}$, todėl šakninis mazgas turi 3 vaikus, atitinkančius 3 aibės *S* poaibius: $S_{saulėta}$, $S_{debesuota}$, S_{lietus}. Toliau skaičiuojama

likusių parametrų {$p_{temperatūra}$, $p_{drėgnumas}$, $p_{vėjas}$} informacinė nauda kiekvename iš šių poaibių. Parametrą p_{oras} jau naudojome šakniniame mazge, todėl poaibiuose nagrinėjami tik likę 3 parametrai. Aibėje $S_{debesuota}$ visi 4 elementai turi reikšmę *TAIP* (kitamasis *y*), t. y. klasifikuojami vienodai, todėl galima sukurti medžio lapą. Toliau, pvz., poaibiui $S_{saulėta}$ = {*s1, s2, s8, s9, s11*}:

$Nauda(S_{saulėta}, p_{drėgnumas}) = 0,970$
$Nauda(S_{saulėta}, p_{temperatūra}) = 0,570$
$Nauda(S_{saulėta}, p_{vėjas}) = 0,019$

Dabar parametras $p_{drėgnumas}$ turi didžiausią informacinę naudą, todėl jis naudojamas „dalinant“ mazgą. Šis procesas tęsiamas tol, kol visi duomenys klasifikuoti idealiai arba nėra daugiau parametrų. Galutinis sprendimo medis (rezultatas) pavaizduotas 5.1 pav. Reikėtų atkreipti dėmesį, kad galutiniame medyje nėra naudojamas temperatūros parametras (algoritmas pats atmetė nereikalingą parametrą, neturintį jokios informacijos).

Darbuose [Quinlan, 1993], [Quinlan, 1996] pateiktas C4.5 algoritmas, kuris yra patobulinta ID3 algoritmo versija. Keli patobulinimai:

1) gali apdoroti parametrus, turinčius tiek diskrečias, tiek tolydžias reikšmes (kitaip nei ID3 algoritmas, kuris dirba tik su diskrečiomis parametrų reikšmėmis); tam, kad apdorotų tolydžias reikšmes turinčius parametrus, C4.5 apskaičiuoja parametro reikšmių slenkstį ir naudoja jį dalindamas parametro reikšmių aibę į poaibius;
2) gali apdoroti mokymo duomenis su trūkstamomis parametrų reikšmėmis; trūkstamos parametrų reikšmės tiesiog nenaudojamos skaičiuojant entropiją ir informacinę naudą;
3) gali apdoroti parametrus su skirtingais svoriais;
4) atlieka medžio genėjimą: po medžio sukūrimo C4.5 algoritmas pereina atgal per medį ir bando atmesti briaunas, kurios nepadeda priimti sprendimo.

Dabar trumpai apžvelkime, kaip sprendimų medžiai yra naudojami automatiniškai kirčiuojant įvairių pasaulio kalbų tekstus.

5.2.3 Sprendimo medžiai teksto kirčiavimo algoritmuose

Sprendimo medžiai gana dažnai naudojami kuriant automatinio žodžių kirčiavimo algoritmus. Kaip buvo minėta 5.2.1 skyrelyje, iš medžio galima nesudėtingai gauti taisykles [Šef, 2005]. Medžiai gali būti sudaromi automatiškai – generuojant iš duomenų – arba rankiniu būdu ([Kazlauskienė, Raškinis, 2004], [Kazlauskienė ir kt., 2004], [Norkevičius ir kt., 2004], dar žr. 2.6 skyrelį).

Sprendimo medžius naudojantys fleksinės slovėnų kalbos automatinio kirčiavimo algoritmai aprašyti [Šef, 2005], [Marinčič ir kt., 2009]. Darbe [Šef, 2005] sprendimo medžiams sudaryti naudojamas [Quinlan, 1986] metodas (ID3). Kirtis nustatomas dviem etapais, sprendimo medžiai naudojami tik pirmajame. Antrame žodžio kirtis nustatomas pagal žodžio ilgį ir kirčiuotų raidžių skaičių. Kiekvienam balsiui (ir priebalsiui *r*) yra sukuriamas atskiras medis. Medžio lapai (kintamasis *y*) nusako, ar nagrinėjama raidė kirčiuota, ir kokia priegaidė. Naudojami 66 atributai (parametrų vektorius *f*): skiemenų skaičius žodyje, nagrinėjamos raidės skiemens pozicija žodyje, priešdėlių ir priesagų buvimas žodyje ir jų klasė, nagrinėjamos raidės kontekstas (po 3 raides iš kairės ir iš dešinės) ir kt. Kaip jau buvo minėta 2.5.3 skyrelyje, darbe [Marinčič ir kt., 2009] buvo įrodyta, kad slovėnų kalbos automatinio kirčiavimo metodai, kurie remiasi sprendimo medžiais, yra tikslesni negu lingvistų ekspertų sudarytos kirčiavimo taisyklės.

Rumunų fleksinei kalbai skirtas algoritmas aprašytas [Oancea, Badulescu, 2002]. Sprendimo medžiai sudaromi automatiškai iš 4500 rumunų kalbos žodžių žodyno. Medžio lapai (kintamasis *y*) nusako kirčiuoto skiemens poziciją. Parametrų vektorių *f* sudaro: žodžio ilgis (raidžių skaičius), skiemenų žodyje skaičius, fonetinis kodas, raidžių seka žodžio pradžioje, pabaigoje ir viduryje. Medis buvo sudaromas pradedant nuo pirmo parametro ir einant nuosekliai iki paskutinio parametro, taigi nereikia ieškoti, pagal kurį parametrą geriausia „dalinti“ mazgą. Gautos taisyklės pavyzdys: „Žodžiai, kurių ilgis yra

n raidžių, *r* skiemenų, kurių fonetinis kodas yra *u* ir pradžios raidžių seka yra s_i, turi kirtį ant *p* skiemens.“

Sprendimo medžiai naudojami kituose TTS uždaviniuose, pvz.: a) transkribuojant [Black ir kt., 1998a], [Pagel ir kt., 1998], [Webster, 2004]; b) modeliuojant garsų trukmes [Norkevičius, Raškinis, 2008], ir kt.

Sprendimo medžiai dar gali būti naudojami ir srityse, nesusijusiose su kalbos apdorojimu, pvz., medicininėje diagnostikoje, augalų klasifikacijai, vertinant kreditų riziką, aparatūros gedimų paieškai pagal priežastis, paieškai internete ir kt.

5.3 Duomenys autoriaus pasiūlytam algoritmui

Prieš sudarant sprendimų medžius kirčiavimo taisyklėms gauti, iš kirčiuotų tekstų (žr. 3 skyrių) sudaromi du žodžių sąrašai: nekirčiuotų žodžių ir kirčiuotų žodžių. **Nekirčiuotų žodžių sąrašo** didžiąją dalį sudaro klitikai – žodžiai, kurie lietuvių kalboje linkę būti nekirčiuojami, pavyzdžiui, žodžiai *be*, *ant*, *bei*, *nuo*, *ir* ir kt. Į šį sąrašą taip pat pakliūna ir nelietuviški žodžiai bei santrumpos (daugiau apie klitikų paiešką lietuvių kalboje žr. 6 skyrių). Jei koks nors žodis tekste būna ir kirčiuotas, ir nekirčiuotas, į nekirčiuotų žodžių sąrašą jis pakliūva tuo atveju, jei dažniau nekirčiuojamas. Toliau nekirčiuotų žodžių (klitikų) sąrašas visuomet naudojamas prieš taikant kirčiavimo taisykles, t. y. jei kirčiuojamas žodis priklauso šiam sąrašui, jis paliekamas nekirčiuotas ir jokios kirčiavimo taisyklės jam nėra taikomos.

Iš kirčiuotų teksto žodžių sudaromas **kirčiuotų žodžių sąrašas**. Jei vienodi žodžiai skirtingai kirčiuojami (homografai), tai į sąrašą įtraukiamas tas kirčiavimo variantas, kuris statistiškai pasitaiko dažniau. Plačiau homografų kirčiavimo problema nagrinėta 4 skyriuje. Kirčiuotų žodžių sąrašas toliau naudojamas sprendimo medžiams sudaryti ir kirčiavimo taisyklėms gauti. Kirčiuotų žodžių sąrašą galima taikyti ir kaip žodžių kirčiavimo taisykles, tačiau šiuo atveju sunku tikėtis gerų rezultatų, nes žodžiai, kurių nebuvo pradiniame tekstyne, nebus sukirčiuoti (žr. 5.6 skyrelį).

5.4 Žodžio pradžios ir pabaigos taisyklių algoritmas

Šiame skyriuje pasiūlyti metodai kirčiavimo taisyklėms sudaryti naudoja sprendimo medžius, kuriuose parametrų vektorius *f* atitinka žodžio raidžių seką, o kintamasis *y* – kirčiuotą raidę (indeksą) ir kirčio tipą (priegaidę). Buvo išbandyti trys skirtingi metodai: raidžių sekos žodžio pradžioje, žodžio pabaigoje ir bet kurioje žodžio vietoje.

Iš pradžių panagrinėkime medžio sudarymą imant raides nuo žodžio pradžios. Medžio mazguose saugomas galimas kirčiavimas (kirčio vieta ir priegaidė), o mazgus jungiančiose briaunose – raidės (žr. 5.2 pav.). Pradedame nuo tuščio medžio, turinčio tik vieną mazgą – šaknį. Kiekvieną žodį iš kirčiuotų žodžių sąrašo pridedame į medį nuo šaknies, pradėdami nuo pirmos raidės ir iki žodžio pabaigos[13]. Pridedant žodį į medį, visi kelyje pasitaikantys mazgai (t. y. tiek, kiek žodyje yra raidžių) papildomi ta pačia informacija apie kirčiavimą. Po to, kai visi žodžiai iš sąrašo jau pridėti į medį, šakninis mazgas saugo visus įmanomus kirčiavimo variantus. Toliau algoritmas aprašytas pseudokodu:

```
Kiekvienam žodžiui iš kirčiuotų žodžių sąrašo
        Einamuoju mazgu tampa šaknis
        Kiekvienai raidei pradedant nuo žodžio pradžios
                Pridėti raidę į medį
                Pakeisti einamąjį mazgą ir papildyti jį
                informacija apie kirčiavimą
```

Jei vienas žodis yra kito žodžio poaibis ir šie žodžiai kirčiuojami skirtingai, reikia sudaryti dvi skirtingas taisykles, tačiau pagal aprašytą algoritmą bus sudaryta tik ilgesnio žodžio kirčiavimą atitinkanti taisyklė. Tam, kad būtų išspręsta ši problema, prie visų žodžių pabaigų buvo pridėti specialūs žodžio pabaigos simboliai („#“). Po šios operacijos pirmasis žodis jau nebėra kito žodžio poaibis.

[13] Dėl to, kad žodžiai buvo pridedami į medį nuosekliai nuo pirmos raidės ir iki žodžio pabaigos, kiekviename medžio mazge dalinimo parametras yra žinomas iš anksto, todėl mūsų atveju neverta naudoti sudėtingesnių sprendimo medžio sudarymo algoritmų (ID3 ir pan.).

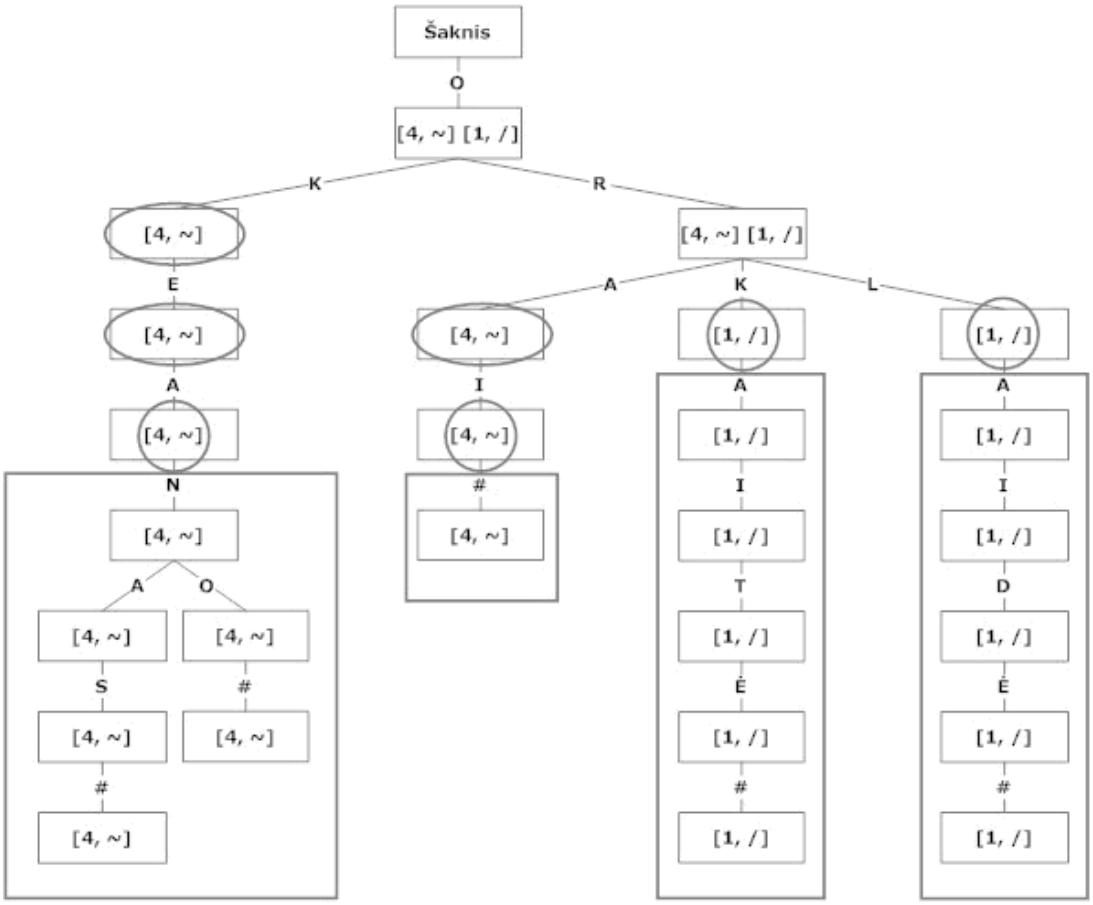
Po to, kai visi žodžiai yra pridėti į medį, iš medžio galima gauti **taisykles**, t. y. tokias raidžių sekas, kurios vienareikšmiškai apibrėžia žodžio kirčiavimą. Einant nuo šaknies visomis medžio šakomis, ieškoma mazgų, kuriuose įrašytas vienareikšmis kirčiavimas, ir kur kirčiuotos raidės indeksas yra mažesnis arba lygus einamojo mazgo lygiui (sprendimo mazgai). Kirčiavimo taisyklė gaunama nuo medžio šaknies iki sprendimo mazgo surinkus ant briaunų esančias raides į seką. Reikia paminėti, kad sprendimo mazgo vaikus galima atmesti netikrinus, nes iš jų gautume tik tą patį kirčiavimą nusakančias, tačiau ilgesnes taisykles. Į taisyklių sąrašą įtraukiama tik trumpiausioji, pvz., **ÓRK**, ÓRKA, ÓRKAI, ÓRKAIT, ÓRKAITĖ, ÓRKAITĖ#. Čia ir toliau taisyklės tekstinis pavidalas (pvz., „ORK“), kirčiuotos raidės indeksas (1) ir priegaidė (‚/‘) yra jungiami į vieną formą („ÓRK“).

Toliau pavyzdžiais pailiustruosime sprendimų medžio ir taisyklių sudarymą. 5.2 lentelėje kairėje pateiktas žodžių sąrašas, o 5.2 pav. – iš tų žodžių sudarytas medis (žodžiai imti nuo pradžios). Gautos taisyklės pateiktos 5.2 lentelėje dešinėje.

5.2 lentelė. Žodžių sąrašas ir iš jo sudarytos žodžio pradžios taisyklės

Įvesti žodžiai	Gautos pradžios taisyklės
OKEÃNAS# OKEÃNO# ORAĨ# ÓRKAITĖ# ÓRLAIDĖ#	OKEÃ ORAĨ ÓRK ÓRL

Norint sukirčiuoti žodį naudojant žodžio pradžios taisykles, reikia tiesiog taisyklių aibėje rasti tą, kuri sutaptų su žodžio pradžia. Paieškai paspartinti galima taisykles surūšiuoti ir atlikti binarinę paiešką. Jei tinkamos taisyklės nerandama, žodis paliekamas nekirčiuotas. Paprastai sprendimų medžiais grįstuose algoritmuose tokiais atvejais imamas medyje vienu lygiu aukščiau esantis sprendimas, tačiau šiame darbe taip nėra daroma. Galime manyti, kad egzistuoja dar viena briauna atitinkanti visas kitas raides, kurios mazge įrašytas sprendimas – „nekirčiuoti duotojo žodžio“.



5.2 pav. Žodžio pradžios taisyklių medis. Apskritimais pažymėti sprendimo mazgai; stačiakampiais – sprendimo mazgų vaikai, kurie yra atmetami; elipsėmis – tie mazgai, kurie negali būti sprendimo mazgais, nes nepasiekta kirčiuota raidė

Žodžio pabaigos taisyklių medis yra formuojamas taip pat, kaip ir žodžio pradžios taisyklių medis, tik prieš pridedant į medį kiekvienas žodis yra apverčiamas. Be to, pridedami ne žodžio pabaigos, bet pradžios simboliai („#“). Šiuo atveju medžio šaknis atitinka žodžio pabaigą, todėl ir kirčiavimo metu taisyklės lyginamos su žodžio pabaiga.

5.5 Žodžio vidurio taisyklių algoritmas

Toliau nagrinėsime raidžių sekas, kurios gali būti žodyje bet kur: tiek pradžioje, tiek viduryje, tiek pabaigoje. Paprastumo dėlei jas vadinsime žodžio vidurio taisyklėmis. Algoritmas yra panašus į žodžio pradžios taisyklių algoritmą, tik kiekvienas žodis į medį yra pridedamas kelis kartus (tiek, kiek

žodyje yra raidžių) atmetant po vieną raidę nuo žodžio pradžios. Dar reikėtų paminėti, kad, kitaip nei sudarant žodžio pradžios bei pabaigos taisyklių medžius, papildomu simboliu „#“ reikia pažymėti ir žodžio pradžią, ir pabaigą. Kirčiavimo taisyklės iš sprendimų medžio gaunamos visiškai taip pat, kaip ir žodžio pradžios taisyklių atveju. Norint sukirčiuoti žodį, su taisyklėmis reikia lyginti ne tik žodžio pradžią, bet ir ieškoti bet kurios žodžio vietos sutapimo su taisykle. Tai sulėtina taisyklių paiešką, todėl aktualus tampa taisyklių skaičiaus mažinimo klausimas.

Kaip jau buvo parodyta anksčiau (5.4 skyriuje), sprendimo mazgo vaikų atmetimas leidžia atmesti taisyklę, kurios pradžia sutampa su kita taisykle. Tačiau, sudarius žodžio vidurio taisyklių aibę, tarp jų gali būti tokių taisyklių, kur vienos pabaiga sutampa su kita taisykle, pvz.: #ORAĨ, ORAĨ, **RAĨ.** Pradedant paiešką nuo ilgiausių taisyklių, surandamos tokių taisyklių poros ir ilgesnės taisyklės išmetamos. Toliau šį taisyklių sumažinimą vadinsime **pirmuoju sumažinimu.**

Net ir sumažinus taisyklių skaičių, vis vien pasitaiko atvejų, kai tam pačiam žodžiui tinka kelios taisyklės (nusakančios tą patį kirčiavimą). Be to, taisyklės gali skirtis pagal jų taikymo statistinį dažnumą. Mažinant taisyklių skaičių, į tai taip pat reikia atsižvelgti. Taisyklių atmetimo algoritmo esmė: imami visi žodžiai, iš kurių buvo sudarytos taisyklės, ir taisyklės taikomos tol, kol yra nors vienas žodis, kuriam nebuvo pritaikyta nei viena taisyklė. Kai taisyklės surastos visiems žodžiams , visas likusias taisykles galima išbraukti iš taisyklių sąrašo. Pradedama nuo taisyklių, kurios tinka didžiausiam žodžių skaičiui. Toliau šį taisyklių sumažinimą vadinsime **antruoju sumažinimu.** Antrasis sumažinimas pseudokodu:

```
Kiekvienam žodžiui nustatyti, kokios taisyklės gali būti
pritaikytos
Kiekvienai taisyklei apskaičiuoti, keliems žodžiams gali
būti pritaikyta
Kol nors vienas žodis yra aktyvus
        Pasirinkti taisyklę, turinčią maksimalų taikomų
        žodžių skaičių
        Deaktyvuoti žodžius, kuriems šį taisyklė gali
        būti taikoma
        Kiekvienai taisyklei, kuri gali būti taikoma
        deaktyvuotiems žodžiams, sumažinti taikomų
        žodžių skaitliuką
```

5.6 Eksperimentų rezultatai

Kaip buvo minėta 3 skyriuje, turimi kirčiuoti tekstai buvo padalinti į penkias maždaug lygias dalis po 200.000 žodžių. Pirmiausia atlikti eksperimentai, kuriuose taisyklėms sudaryti ir testuoti buvo naudoti tie patys tekstai. Visų pirma, tai leidžia įvertinti, kokią įtaką kirčiavimo tikslumui daro homografai. Eksperimentų metu žodžių skaičių didinant nuo 200.000 iki 1.000.000, klaida monotoniškai didėjo nuo 1,02% iki 1,22%. Be to, klaidos dydis nepriklausė nuo kirčiavimo taisyklių sudarymo būdo. Tai rodo, kad visi pasiūlyti kirčiavimo taisyklių rinkiniai turi ne mažiau informacijos, nei kirčiuotų žodžių sąrašas.

Tolesniuose eksperimentuose vieni duomenys buvo naudojami taisyklėms sudaryti, o kiti testuoti. Taisyklėms sudaryti naudojamus duomenis toliau vadinsime **mokymo duomenimis**, o taisyklių sudarymą iš šių duomenų – **mokymu**. Taisyklėms sudaryti naudotos keturių dydžių žodžių aibės: po 200.000, 400.000, 600.000 ir 800.000 žodžių. Šios aibės gautos visais įmanomais būdais kombinuojant 200.000 žodžių dydžio tekstyno dalis (iš viso 75 tokios kombinacijos). Testavimas buvo atliekamas su visais žodžiais, kurie nebuvo naudoti taisyklių sudaryme.

Kiekvienam mokymo duomenų kiekiui buvo skaičiuojamas klaidos vidurkis ir taisyklių skaičiaus vidurkis. T. y. kiekvienam metodui skaičiuojami

4 klaidų vidurkiai: klaidos vidurkis su 200.000 žodžių, su 400.000 žodžių, su 600.000 žodžių, su 800.000 žodžių, taip pat 4 taisyklių skaičiaus vidurkiai.

Atlikti eksperimentai naudojant 7 kirčiavimo taisyklių sudarymo metodus:

1) Kaip kirčiavimo taisyklės naudotas kirčiuotų žodžių sąrašas (žymėsime santrumpa **žod**).
2) Iš pradžių taikomos žodžio pradžios taisyklės, likusiems nekirčiuotiems žodžiams – žodžio pabaigos taisyklės (**pr-pab**).
3) Iš pradžių taikomos žodžio pabaigos taisyklės, likusiems nekirčiuotiems žodžiams – žodžio pradžios taisyklės (**pab-pr**).
4) Taikomos tik žodžio pradžios taisyklės (**prad**).
5) Taikomos tik žodžio pabaigos taisyklės (**pab**).
6) Taikoma visa žodžio vidurio taisyklių aibė (**vid**) arba ši aibė po pirmojo sumažinimo (**vid1**).
7) Taikoma žodžio vidurio taisyklių aibė po antrojo sumažinimo (**vid2**).

Šeštame metode sujungti du metodai, nes jų klaidos vienodos, skiriasi tik gautų taisyklių skaičius. Dar verta priminti, kad pirmiausia patikrinama, ar kirčiuojamas žodis nepriklauso nekirčiuotų žodžių (klitikų) sąrašui. Jei priklauso – jis paliekamas nekirčiuotas ir jokios taisyklės jam netaikomos. Šis metodas taip pat gali tam tikrą skaičių žodžių klaidingai priskirti nekirčiuotiems žodžiams.

Teksto kirčiavimo tikslumo vidurkiai įvairiems kirčiavimo taisyklių sudarymo metodams, kai taisyklėms sudaryti naudota 800.000 žodžių, o testuoti – 200.000, pateikti 5.3 lentelėje. Rezultatai, kai taisyklėms sudaryti buvo naudota 200.000, 400.000, 600.000 ir 800.000 žodžių yra pateikti 2 priede.

Kaip matyti iš 5.3 lentelės, geriausią rezultatą duoda 3 metodas (**pab-pr**) – 4,47% klaidų. Nedaug (mažiau nei 0,3%) nusileidžia 2 metodas (**pr-pab**), 6 metodas (**vid, vid1**) ir 7 metodas (**vid2**). 4 (**prad**) ir 5 (**pab**) metodai, kuriuose naudojamos tik pradžios arba tik pabaigos taisyklės, nusileidžia

daugiau kaip 1,5%, o kaip kirčiavimo taisykles naudojant kirčiuotų žodžių žodyną (**žod**) klaida viršija 10%.

5.3 lentelė. Klaidos priklausomybė nuo naudojamo metodo. Lentelės stulpeliai: A – Klitikai sukirčiuoti (klaidingi); B – Klitikai nekirčiuoti (teisingi); C – Nesukirčiuoti žodžiai (klaidingi); D – Nežinomi (pvz., užsienio kalbų) žodžiai nekirčiuoti (teisingi); E – Nežinomi žodžiai sukirčiuoti (klaidingi); F – Ne ten kirtis arba ne ta priegaidė (klaidingi); G – Sukirčiuoti teisingai; H – Klaidingai iš viso (A+C+E+F); I – Teisingai iš viso (B+D+G)

Metodas	A	B	C	D	E	F	G	H	I
1 žod	0,19	15,82	8,81	0,69	0,10	1,10	73,28	10,21	89,79
2 pr-pab	0,19	15,82	1,54	0,46	0,33	2,53	79,13	4,59	95,41
3 pab-pr	0,19	15,82	1,54	0,46	0,33	2,41	79,25	**4,47**	**95,53**
4 prad	0,19	15,82	3,51	0,57	0,22	2,15	77,53	6,08	93,92
5 pab	0,19	15,82	3,64	0,55	0,24	2,00	77,56	6,07	93,93
6 vid, vid1	0,19	15,82	1,04	0,35	0,44	2,99	79,17	4,66	95,34
7 vid2	0,19	15,82	1,93	0,44	0,35	2,29	78,98	4,76	95,24

Kiekvieno metodo taisyklių skaičiaus vidurkiai pateikti 5.4 lentelėje.

5.4 lentelė. Taisyklių skaičiaus priklausomybė nuo naudojamo metodo

Metodas	200.000	400.000	600.000	800.000
tik klitikai	1381	2441	3407	4309
žod	43215	67564	86633	102760
prad	28293	42608	53433	62424
pab	29790	44688	56009	65404
vid	118442	175615	218280	253379
vid1	39545	57165	70047	80510
vid2	19627	28840	35676	41291

Kaip matome iš 5.4 lentelės, mažiausiai taisyklių reikalauja metodas **vid2**. Apie pusantro karto daugiau reikalauja metodai **prad** ir **pab**. Taigi mažiausią klaidą pasiekusiam metodui (**pab-pr**) reikia apie 3 kartus daugiau taisyklių negu metodui **vid2**. Tačiau žodžio vidurio taisyklių metodai veikia kiek lėčiau, nes taisykles reikia lyginti ne tik su žodžio pradžia ar pabaiga, tačiau ir pradedant kiekviena žodžio raide. Antrasis žodžio vidurio taisyklių sumažinimas leidžia jų skaičių sumažinti apie du kartus, o tikslumas sumažėja

tik apie 0,1%. Kita išvada: žodžio pabaigos taisyklių (**pab**) visada yra šiek tiek daugiau negu žodžio pradžios taisyklių (**prad**).

5.7 Rezultatų palyginimas su morfologiniu metodu

5.3 lentelėje geriausias metodas (**pab-pr**) yra lyginamas su [Kasparaitis, 2000], [Kasparaitis, 2001b] darbuose pasiūlytu metodu, besiremiančiu morfologinėmis taisyklėmis, kuris dar buvo papildytas klitikų (žr. 6 skyrių) ir homografų (žr. 4 skyrių) kirčiavimo taisyklėmis. Abu metodai yra testuojami su identiškais duomenimis. Šiuo atveju disertacijoje pasiūlytas algoritmas yra mokomas tik su viena kombinacija iš 800.000 žodžių, todėl nėra skaičiuojamas vidurkis, taigi algoritmo rezultatai skiriasi nuo pateiktųjų 5.3 lentelėje. Nors pasiūlyto metodo rezultatai yra šiek tiek blogesni už morfologinio metodo rezultatus (apie 0,8%), tačiau pasiūlytas metodas yra žymiai paprastesnis tiek gavimo tiek ir taikymo atžvilgiu.

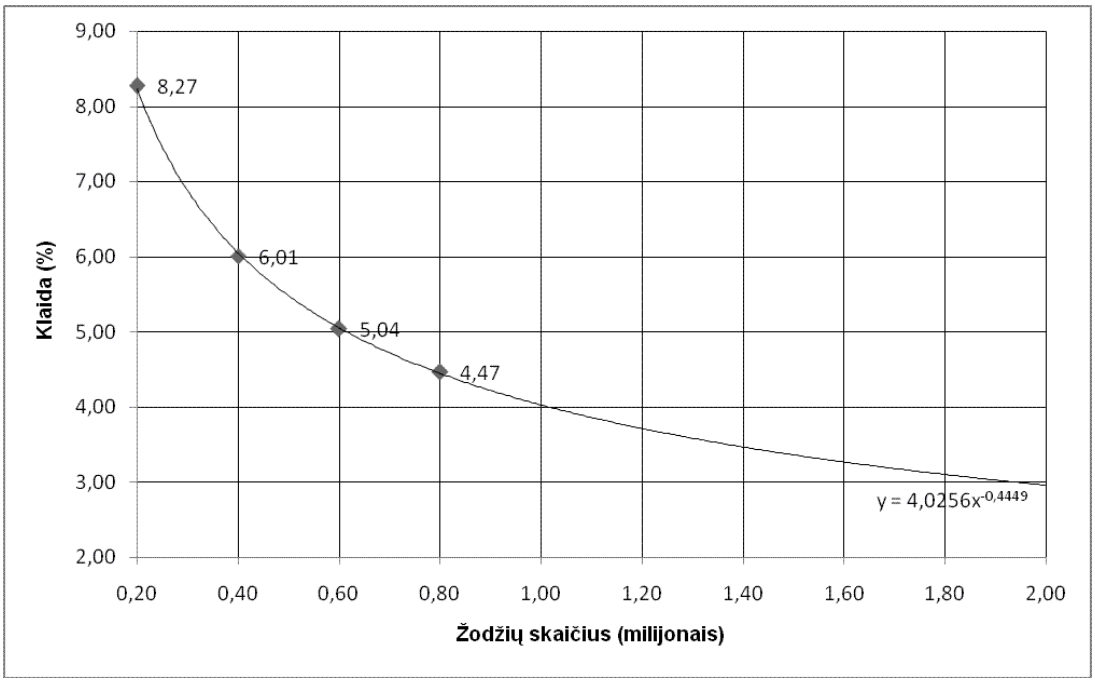
5.5 lentelė. Geriausio metodo (**pab-pr**) lyginimas su morfologiniu metodu (**morfolog**) testuojant tik su viena kombinacija iš 200.000 žodžių. Stulpelių reikšmes žr. 5.3 lentelėje

Metodas	A	B	C	D	E	F	G	H	I
3 pab-pr	0,17	15,80	1,32	0,51	0,36	2,37	79,47	4,22	95,78
morfolog	0,07	13,49	1,54	3,07	0,11	1,67	80,05	3,40	96,61

5.8 Tekstyno didinimo įtakos kirčiavimo tikslumui prognozė

Kaip ir buvo galima tikėtis, eksperimentai parodė, kad kuo daugiau žodžių naudojama taisyklėms sudaryti, tuo didesnis naujo teksto kirčiavimo tikslumas pasiekiamas. Remiantis klaidos reikšmėmis, gautomis apmokius tiksliausią metodą **pab-pr** su 200.000, 400.000, 600.000 ir 800.000 žodžių, buvo bandoma prognozuoti (ekstrapoliuoti) klaidą didesniam žodžių kiekiui. Ekstrapoliacijai pasirinktas mažiausių kvadratų metodas (prieiga per internetą: [http://mathworld.wolfram.com/LeastSquaresFittingPowerLaw.html, žiūrėta 2009.07.01]). Rezultatai pateikti 5.3 pav. Gauta ekstrapoliacijos funkcija:

$y = 4{,}0256*x^{-0{,}4449}$, kur x – žodžių skaičius, y – klaidos procentas. Klaidų skaičius, analogiškas pasiektam taikant morfologinį metodą (3,40%), būtų pasiektas mokant su maždaug 1.500.000 žodžių, o apmokius su 2.000.000 prognozuojama klaida būtų 2,96%.



5.3 pav. Tekstyno didinimo įtakos kirčiavimo tikslumui prognozė

5.9 Penktojo skyriaus apibendrinimas

- Šiame skyriuje pasiūlyti metodai, kurie kirčiavimo taisykles sudaro iš kirčiuotų žodžių sąrašo. Tokie metodai paprastai taikomi nefleksinėms kalboms.
- Taisyklėms sudaryti naudotas sprendimo medžių algoritmas. Nagrinėti keli taisyklių sudarymo būdai. Parodyta, kad didžiausią tikslumą duoda pabaigos-pradžios taisyklių metodas, o mažiausiai taisyklių gaunama taikant žodžio vidurio taisyklių metodą ir taisyklių aibė sumažinama, naudojant šiame skyriuje aprašytą algoritmą.
- Savo tikslumu geriausias pasiūlytas metodas tik 0,8% nusileidžia morfologinę analizę naudojančiam metodui, tačiau parodyta, kad,

didinant taisyklėms sudaryti naudojamų tekstų kiekį, galima tikėtis šį tikslumą pasiekti ir aplenkti.

- Pasiūlyti metodai remiasi tik raidžių sekomis ir nereikalauja jokių žinių apie kalbą: morfemas, kalbos dalis, žodžių kaitymą, skiemenavimą ir pan. Reikalingas tik pakankamas kiekis kirčiuotų tekstų, iš kurių taisyklės sudaromos automatiškai. Todėl šie metodai gali lengvai būti pritaikyti ir kitoms kalboms.
- Pats kirčiavimo algoritmas – tai iš esmės binarinė paieška surūšiuotame taisyklių sąraše, todėl algoritmas labai greitas, be to, jį lengva perkelti į kitą programavimo kalbą ar kitą operacinę sistemą.

6 Lietuvių kalbos ritmo modeliavimas (klitikų paieška)

Sintezuojant balsą iš teksto, svarbu išlaikyti kalbos ritmą. Kalbai ritmą suteikia tarp kirčiuotų skiemenų esantys nekirčiuoti skiemenys. Tačiau jei vieno žodžio kirčiuotas paskutinis skiemuo, o kito žodžio – pirmasis, turėsime greta du kirčiuotus skiemenis. Tai ypač dažnai pasitaiko, jei bent vienas iš dviejų žodžių yra vienskiemenis. Siekiant išlaikyti kalbos ritmą, kai kurie žodžiai lieka nekirčiuoti (prišlyja prie kirčiuotų gretimų žodžių ir lyg virsta pastarųjų skiemenimis, t. y. tampa klitikais). Tačiau iki šiol sukurti lietuvių kalbos žodžių kirčiavimo algoritmai [Kasparaitis, 2000], [Kasparaitis, 2001], [Kazlauskienė, Raškinis, 2004], [Kazlauskienė ir kt., 2004], [Norkevičius ir kt., 2004] kirčiuodavo visus žodžius. Šiame skyriuje apžvelgiami šlijimą lemiantys faktoriai ir autoriaus pasiūlyti keturi metodai, kaip tekste rasti žodžius, kurie gali būti bekirčiai. Kiekvienam metodui apibrėžtos žodžių klasės, kurioms jis geriausiai tinka, paaiškinta, kaip visus metodus sujungti į vieną algoritmą. Eksperimentais įvertintas pasiūlytų metodų tikslumas.

6.1 Klitiko sąvoka

Jei į tekstą žiūrėsime ne kaip į pavienius žodžius, o kaip į visumą, tai nemažai žodžių yra nekirčiuojami. Kalbos ritmo modeliavimą galima įsivaizduoti kaip kirčių „pašalinimą“ nuo kai kurių žodžių. Bekirčiai žodžiai vadinami **klitikais** [Ambrazas, 1996, 38]. Žodžio tapimą klitiku vadinsime **akcentiniu šlijimu**. Kalbotyros darbuose (pvz., [Mauricaitė, 1985] [Stundžia, 1991]) galima rasti tik bendras žodžių akcentinio šlijimo tendencijas, o klitikų paieškos lietuviškame tekste algoritmai – dar visai nenagrinėta tema. Be to, tai gana sudėtinga problema, nes tas pats žodis viename kontekste gali turėti kirtį, o kitame jis gali netekti kirčio. Remiantis [Zwicky, 1995], klitikais vadinami kalbos elementai su kai kuriomis nepriklausomo (atskiro) žodžio ir kartu afikso savybėmis: kaip ir afiksai, šie elementai vienaip ar kitaip priklauso nuo gretimo žodžio. Kalbant apie prozodijos modeliavimą, klitikai – tai žodžiai, kuriems

nepriskiriamas pagrindinio tono akcentas (angl. *pitch accent*) [Goldsmith, 1999]. Jeigu klitikas prišlyja prie kirčiuoto žodžio pradžios, jis vadinami proklitiku, jei prie galo – enklitiku. Savarankiško žodžio ir klitiko junginys, vienijamas bendro kirčio, laikomas vienu fonetiniu žodžiu[14].

Skirtingų kalbų klitikų sistemos turi bendrų bruožų, pvz., kalbos dalys, linkusios būti klitikais, dažnai yra tos pačios (įvardžiai, prielinksniai, jungtukai ir pan.), tačiau skirtingų kalbų klitikų sistemos gali turėti ir ypatumų. Pvz., anglų kalboje prie klitikų priskiriamas ir kilmininko linksnio žymėjimas *'s*, pvz., *Paul's book* (*Pauliaus knyga*) [Taylor, 2009, 61]. Taip pat klitikais laikomi ir veiksmažodžio *be* (*būti*) sutrumpintos formos [Zwicky, 1995], pvz., esamasis laikas: *he's walking* (pilna forma *he is walking* (*jis eina*)), arba būsimasis laikas: *he'll walk* (pilna forma *he will walk* (*jis eis*)). Tokių klitikų yra baigtinis skaičius, jie yra atskiriami apostrofu, todėl nesukelia problemų realizuojant TTS sistemą ir yra interpretuojami kaip vienas žodis: skaitomi taip, kaip rašomi, jų nereikia išskleisti kaip santrumpų [Taylor, 2009, 60-61].

Taip pat kai kuriose kalbose kai kurie įvardžiai vartojami sutrumpinta forma, pvz., prancūzų kalbos įvardžiai: *me* (*aš*) ir *le* (*jis*); ir atitinkami „pilni“ įvardžiai: *moi* ir *lui*, arba serbų-kroatų kalbos įvardžiai (enklitikai): *im* (*jiems*) ir *ti* (*tau*) ir atitinkami „pilni“ įvardžiai *njima* ir *tebi* [Zwicky, 1977, 3-4]. Tokių klitikų paieška neturėtų būti sudėtinga.

Kai kuriose kalbose klitikas gali būti rašomas kartu su žodžiu, t. y. neatskiriamas nuo jo tarpeliu (pvz., lenkų kalboje). Jų paieška skiriasi nuo atskirų (ar apostrofais pažymėtų) žodžių paieškos, tačiau tokių klitikų nustatymas irgi gali būti reikalingas, nes dažnai fiksuoto kirčio kalbose prie žodžių prišliję enklitikai „griauna“ bendrąsias šios kalbos kirčiavimo taisykles: „jie pailgina fonologinį žodį vienu ar keliais skiemenimis, o kirtis gali likti senoje vietoje“ [Girdenis, 2003, 258].

[14] Tačiau, pasak [Balčiūnaitė, Pakerys, 2006], lietuvių ir užsienio lingvistų yra „svarstoma, ar klitikai labiau susiję su žodžiu, prie kurio šlyja ir su juo sudaro vieną fonologinį žodį, ar tai yra frazės (sakinio) sudedamoji dalis (frazės morfema) [Zwicky, 1985, 283; ...]“.

6.2 Klitikų paieškos algoritmai kitoms kalboms

Tam, kad automatiškai būtų galima surasti klitiką anglų kalbos tekste, dažniausiai apsiribojama keliomis paprastomis taisyklėmis [Huang ir kt., 2001, 757]. Pirma taisyklė: reikšminių (angl. *content*) žodžių kategorijos (POS), tokios kaip daiktavardžiai, veiksmažodžiai, būdvardžiai ir prieveiksmiai, turi būti kirčiuojamos, o funkcinių žodžių[15] kategorijos (visos likusios kalbos dalys, pvz., įvardžiai, prielinksniai, jungtukai, artikeliai (*a*, *an*, *the*), pagalbiniai (angl. *auxiliary*) veiksmažodžiai, veiksmažodis *be* (*būti*) ir pan.) paliekamos be kirčio. Ši taisyklė gali būti patobulinta, nurodant kontekstinių žodžių POS (naudojant *n*-gramus ar pan.). Kitos taisyklės pavyzdys: vienskiemeniai bendri veiksmažodžiai (angl. *common verbs*) nekirčiuojami. Be abejo, šių taisyklių tikslumas nėra šimtaprocentinis. Tam, kad būtų galima atskirti funkcinius ir reikšminius žodžius, galima tiesiog sudaryti funkcinių žodžių (t. y. klitikų) sąrąšą, nes jų kalboje yra baigtinis skaičius. Kitas būdas – pasinaudoti POS žymėjimo algoritmais (žr. 4.3 skyrelį), kurie dažnai naudojami ir kitiems tikslams (pvz., homografams vienareikšminti). Galimas ir tiesioginis POS žymėjimo algoritmo naudojimas klitikams nustatyti, pvz., jau minėtame 4.3 skyrelyje darbe [Orphanos, Christodoulalds, 1999] graikų kalbos klitikų nustatymas yra POS žymėjimo algoritmo dalis – čia klitikas apskritai yra interpretuojamas kaip savarankiška kalbos dalis. Funkciniams ir reikšminiams žodžiams atskirti gali būti papildomai naudojamas ir šio darbo 1.4.4 skyrelyje minėtas metodas *chinks and chunks* [Liberman, Church 1992], kur jį buvo siūloma naudoti anglų kalbos sakiniams skaidyti į frazes. Šio metodo esmė: frazė susideda iš funkcinių žodžių sekos ir po jos einančios reikšminių žodžių sekos.

Kitoks požiūris į anglų kalbos ritmo modeliavimą aprašytas [McPeters, Tharp, 1983]. Čia iš pradžių pažymimi žodžių kirčiai, o vėliau taikoma kirčio

[15] Remiantis [Zwicky, 1995], dar vadinami „gramatiniais žodžiais“, „dalelytėmis“ (angl. *particles*), „markeriais“ (angl. *markers*), arba (dėl to, kad jų skaičius keičiasi labai lėtai) „uždaros kategorijos žodžiais“.

pašalinimo taisyklė EDR (angl. *English Destressing Rule*). EDR remiasi „metrinių medžių“ (angl. *metrical trees*) sąvoka.

Remiantis [Huang ir kt., 2001, 757], idealiu atveju kalbos ritmui nustatyti TTS sistema turėtų naudoti visavertę semantinę bei kitokią analizę. Pvz., anglų kalbos artikeliai nekirčiuojami, tačiau artikelis bus kirčiuotas, jei jis turi loginį frazės kirtį, pvz.: „*English articles are: a, an and the.*“ (*Anglų kalbos artikeliai yra: a, an ir the.*). Čia pabraukti žodžiai turintys loginį kirtį. Loginis kirtis paveikia klitikus ir lietuvių kalboje [Mauricaitė, 1985]. Tačiau loginio kirčio nustatymas yra sudėtinga ir dar neišspręsta problema (žr. 1.4.9 skyrelį).

6.3 Akcentinį šlijimą lietuvių kalboje lemiantys veiksniai

Ar žodis yra klitikas, priklauso nuo įvairių faktorių: nagrinėjamo žodžio skiemenų skaičiaus, jo funkcinio svorio tekste, atstumo tarp kirčiuotų skiemenų, nekirčiuotų skiemenų skaičiaus frazės pradžioje ir pabaigoje, frazės ilgio, loginio kirčio [Mauricaitė, 1985], [Stundžia, 1991].

6.3.1 Skiemenų skaičiaus ir funkcinio reikšmingumo įtaka

Skiemenų skaičius turi tiesioginę įtaką žodžių akcentiniam šlijimui, t. y. dažniausiai šlyja vienskiemeniai žodžiai, rečiau dviskiemeniai, dar rečiau triskiemeniai ir t. t. Šio darbo 6.5 skyrelyje nagrinėtuose tekstuose 95,3% klitikų buvo vienskiemeniai. Todėl toliau kalbėsime tik apie vienskiemenius ir dviskiemenius žodžius.

Darbe [Mauricaitė, 1994] teigiama, kad dažniausiai šlyja vienskiemeniai tarnybiniai žodžiai (prielinksniai, jungtukai, dalelytės), funkciškai reikšmingesni vienskiemeniai žodžiai (įvardžiai, prieveiksmiai, skaitvardžiai) dažniau yra kirčiuojami. Kartais gali šlyti ir dviskiemeniai tarnybiniai žodžiai bei veiksmažodžiai – tarinio jungtys.

Darbe [Mauricaitė, 1986] pateikiami tikrųjų (jie neturėtų būti kirčiuojami) ir potencialiųjų klitikų pavyzdžiai. Tikraisiais klitikais laikomi:

1) Vienskiemeniai prielinksniai: *ant, be, dėl, į, iš, lig, link, nuo, pas, per, po, prie, pro, su, tarp, už.*

2) Vienskiemeniai jungtukai: *ar, bei, bet, ir, jog, kad, kai, kaip, kiek, lyg, nes, o*.
3) Vienskiemenės dalelytės: *ar, be, bet, dar, gal, gi, ir, jau, juk, kai, kaip, ne, tik*.

Potencialieji klitikai:

1) Vienskiemeniai įvardžiai: *aš, jis, ji, kas, šis, ši, tas, ta, tu*.
2) Vienskiemeniai prieveiksmiai: *čia, kur, ten, tuoj*.
3) Vienskiemeniai skaitvardžiai: *du, dvi, trys*.
4) Veiksmažodžiai – tarinio jungtys: *bus*.
5) Dviskiemeniai prielinksniai: *apie, iki*.
6) Dviskiemenė dalelytė *nebe*.

Reikia atkreipti dėmesį į tai, kad čia laikomasi nuostatos, jog iš anksto žinoma, kokia kalbos dalis yra tam tikras žodis, pvz., *o* – jungtukas, tačiau apdorojant tekstą kompiuteriu kalbos dalis nėra žinoma, ją dar reikia nustatyti.

6.3.2 Gretimų skiemenų kirčių įtaka

Darbe [Mauricaitė, 1994] tyrinėta, kokią įtaką šlijimui turi gretimo skiemens kirčiavimas. Potencialus klitikas greta kirčiuoto skiemens dažnai bus nekirčiuotas, o greta nekirčiuoto skiemens arba tarp nekirčiuotų skiemenų – kirčiuotas. Nagrinėti tie žodžiai, kurių kirčiuotų ir nekirčiuotų rasta maždaug po lygiai. Minėta hipotezė pasiteisino, gretimo skiemens kirčiuotumas labiausiai veikia įvardžių, kiek mažiau dalelyčių ir prieveiksmių kirčiavimą. Kalbant apie konkrečius žodžius, minėta taisyklė dažniausiai tinka (tikimybės mažėjimo tvarka) įvardžiams *tas, ta, aš, tu, jis, ji*, prieveiksmiui *čia*, dalelytei *vis*.

Siekiant dar geriau įsivaizduoti kirčių išsidėstymą lietuvių kalbos frazėse galima pasinaudoti [Stundžia, 1991] pateiktais statistiniais duomenimis apie atstumus tarp kirčiuotų skiemenų ir kirčio padėtį frazės pradžioje ir pabaigoje. Dažniausiai tarp kirčiuotų skiemenų būna 2 nekirčiuoti skiemenys, dažniausiai kirčiuojamas antras skiemuo nuo frazės pradžios ir pabaigos. Siekiant išlaikyti atstumus tarp kirčiuotų skiemenų, kai kurie žodžiai paliekami nekirčiuoti.

Šiomis idėjomis remiantis ir bus bandoma sukurti klitikų atpažinimo algoritmą. Frazės ilgio, loginio kirčio ir kalbėjimo tempo įtaka nebus nagrinėjama.

6.4 Klitikai samplaikinėse formose

Verta atskirai pakalbėti apie kai kurias kalbos dalis (įvardžius, prieveiksmius, jungtukus, dalelytes), kurios sudarytos iš kelių žodžių ir vadinamos samplaikinėmis formomis. Samplaikinių formų radimas ir priskyrimas morfologinėms kategorijoms buvo nagrinėtas [Rimkutė ir kt., 2005], šiame darbe nagrinėjamas jų kirčiavimas. Daugelyje samplaikinių formų vienas žodis yra nekirčiuojamas ir prišlyja prie kito samplaikinės formos žodžio [Mauricaitė, 1987]. Daugumos šių samplaikinių formų kirčiavimas yra nusistovėjęs, t. y. žodynuose galima rasti, kurį samplaikinės formos žodį kirčiuoti, o kurio ne.

6.4.1 Samplaikinių formų atpažinimas

Siekiant sukurti samplaikinių formų atpažinimo algoritmą, pirmiausiai remiantis šiais leidiniais [Keinys, 1993] ir [Mauricaitė, 1987] buvo sudarytas samplaikinių formų sąrašas. Jame 95 samplaikinės formos priklauso nekaitomoms kalbos dalims (prieveiksmiai, jungtukai, dalelytės) ir 35 kaitomoms (įvardžiai). Įvardžiai turi 6 linksnių galūnes, be to, kai kurie iš jų turi vienaskaitą ir daugiskaitą, vyriškąją ir moteriškąją giminę. Taigi iš viso gautos 604 įvardžių formos. Trys samplaikiniai įvardžiai nebuvo nagrinėjami, nes abu juos sudarantys žodžiai yra kirčiuojami. Taigi gautas 699 samplaikinių formų sąrašas.

Dauguma į samplaikines formas įeinančių žodžių gali būti vartojami ir atskirai, todėl ne visada lengva atpažinti samplaikines formas. Toliau šiame darbe buvo bandoma išsiaiškinti, kaip dažnai žodžių grupė, kuri gali sudaryti samplaikinę formą, tėra tik atskiri žodžiai. Šiam tikslui buvo atliekama žodžių grupės paieška lietuvių kalbos tekstyne [http://donelaitis.vdu.lt] ir internete [http://www.google.lt], o rasti rezultatai rankiniu būdu išnagrinėti bei

užsirašyta statistika, kiek kartų žodžių grupė sudaro samplaikinę formą, o kiek kartų tai yra atskiri žodžiai. Radus daugiau nei 100 pavyzdžių, buvo nagrinėti tik 100 pirmųjų pavyzdžių. Taip pat reikia pastebėti, kad daugeliu atvejų (564 iš 699) žodžių grupė kirčiuojama vienodai, nesvarbu, ar ji sudaro samplaikinę formą, ar ne. Tokios žodžių grupės nebuvo nagrinėjamos, o paprastumo dėlei laikoma, kad grupė sudaro samplaikinę formą. Likusioms žodžių grupėms išnagrinėjus apie 10000 pavyzdžių nustatyti tokie dėsningumai, kurie bus panaudoti kuriant algoritmą:

1) žodžių grupė, galinti sudaryti samplaikinę formą, dažniausiai ją ir sudaro, išskyrus žodžių grupes *kaip ir*, *kas sykį*, *kam kita*, kurios dažniau samplaikinių formų nesudaro. Šias žodžių grupes paprasčiausia tiesiog pašalinti iš samplaikinių formų sąrašo;
2) žodžių grupės, kurios baigiasi žodžiu *nors*, nesudaro samplaikinių formų eidamos prieš žodį *kiek*, prieš kitus žodžius samplaikines formas sudaro. Paprasčiausias problemos sprendimo būdas, papildyti sąrašą nauja samplaikine forma *nors kiek* (nors tokia forma nebuvo minima nei žodyne [Keinys, 1993], nei straipsnyje [Mauricaitė, 1987]), o samplaikinių formų paiešką pradėti nuo sakinio galo (kodėl paieška atliekama nuo sakinio pabaigos, paaiškinta 6.4.2 skyrelyje).

Kiek išsamiau panagrinėta žodžių grupė *ką tik*. Ji samplaikinę formą sudaro apie 86% atvejų. Tarp likusių apie 10000 nagrinėtų pavyzdžių, žodžių grupės nesudarė samplaikinių formų tik 10 atvejų (0,1%).

6.4.2 Žodžių grupių skaidymas į samplaikines formas

Tarkime, kad turime trijų žodžių grupę *bet kas nors*. Samplaikinę formą gali sudaryti tiek žodžiai *bet kas*, tiek žodžiai *kas nors*. Siekiant išsiaiškinti, kaip spręsti tokias skaidymo problemas, buvo surastos samplaikinių formų poros, kur vienos formos paskutinis žodis sutampa su kitos formos pirmuoju, ir sudarytos visos galimos iš dalies persidengiančių samplaikinių formų žodžių grupės. Su jomis atlikta analogiška paieška tekstyne arba internete ir rasti rezultatai rankiniu būdu išnagrinėti bei užsirašyta vienokio ar kitokio skaidymo

žodžių grupėmis statistika. Įvardžių, kurie gali turėti vienaskaitą ir daugiskaitą, buvo nagrinėjama tik vienaskaita. Iš viso buvo nagrinėtos 328 grupės, 9 iš jų rasta daugiau nei po 100 pavyzdžių, 48 rasta nuo 1 iki 99 pavyzdžių, kitų grupių pavyzdžių nerasta. Iš viso buvo išanalizuota apie 1500 pavyzdžių ir sudarytos tokios taisyklės, kurios bus panaudotos kuriant algoritmą:

1) Pirmiausiai tikrinama, ar grupėje yra samplaikinių formų (visi linksniai) *kas ne kas, koks ne koks, kuris ne kuris, kada ne kada, kame ne kame, kur ne kur, vos ne vos*. Šios formos niekada neskaidomos.
2) Jei grupė prasideda žodžiu *bet* po skyrybos ženklo, samplaikinę formą sudaro paskutiniai du žodžiai, pvz., *bet | kas nors*, priešingu atveju pirmieji du, pvz., *bet kas | nors*.
3) Toliau išvardintas grupes skaidyti taip: *vis vien | tik, vos ne | tik, daug kas | kita*.
4) Kitais atvejais samplaikinė forma yra grupės pabaigoje, t. y. nuo grupės pabaigos atskeliama samplaikinė forma, o likusiai grupei vėl gali būti taikoma ši taisyklė.

Buvo rasti tik 5 pavyzdžiai, kurie netenkina šių taisyklių. Kadangi daugelyje atvejų tenka taikyti 4 taisyklę, todėl samplaikinių formų paiešką patogiausia atlikti nuo sakinio galo.

6.5 Klitikų paieška remiantis kirčiavimo / nekirčiavimo dažniu

Tolimesniuose eksperimentuose buvo naudojami lietuvių kalbos tarties mokymuisi skirti kirčiuoti tekstai, paimti iš [Pakerys, Pupkis, 1976]. Šiuose tekstuose klitikai palikti nekirčiuoti. Buvo naudojami tik prozos tekstai, nes poezijoje kartais kirtis paslenkamas siekiant išlaikyti ritmą. Tolimesniuose eksperimentuose buvo naudojama tekstų dalis sudaryta iš 8397 žodžių, iš jų 1842 (21,9%) nekirčiuoti.

Toliau visuose eksperimentuose pirmiausiai surandamos ir sukirčiuojamos samplaikinės formos. Samplaikinėms formoms priklausantiems žodžiams kiti algoritmai nebetaikomi. Tokių formų rasta 49 (98 žodžiai), kirčiuojant padaryta viena klaida (2 žodžiai).

Toliau buvo bandoma išsiaiškinti, kaip sėkmingai galima aptikti nekirčiuojamus žodžius remiantis vien tik statistika, t. y. remiantis tekstais statistiškai nustatyti, kad tam tikras žodis dažniau yra nekirčiuotas, nei kirčiuotas, ir iš tokių žodžių sudaryti sąrašą, kurio žodžius laikysime nekirčiuojamais. Analogiškai galima būtų sudaryti ir kirčiuojamų žodžių sąrašą, bet galima tiesiog visus likusius žodžius laikyti kirčiuojamais. Tekstai buvo padalinti į dvi maždaug lygias dalis, kurios pakaitomis buvo naudojamos sąrašui sudaryti ir testuoti. Eksperimentų metu gali būti padaromos dviejų tipų klaidos: a) žodis, kurį reikėtų kirčiuoti, paliekamas nekirčiuotas; b) sukirčiuojamas žodis, kurio galima būtų nekirčiuoti. Visuose tolimesniuose eksperimentuose bus stengiamasi minimizuoti abiejų tipų klaidų sumą. Rezultatai pateikti 6.1 lentelėje.

6.1 lentelė. Klitikų paieškos, remiantis nekirčiuojamų žodžių sąrašu, rezultatai

Sąrašas sudarytas iš	Kirčiuotų žodžių + nekirčiuotų žodžių= suma, (**dalis tarp visų žodžių**)	Sąrašo ilgis	Pirmojo tipo klaidų + antrojo tipo klaidų = suma (*klaidų ir nekirčiuotų žodžių santykis*)	
			1 dalis testavimui	2 dalis testavimui
1 dalies	3225 + 910 = 4135 **(49,8%)**	69	60 + 56 = 116 *(12,7%)*	83 + 96 = 179 *(19,7%)*
2 dalies	3281 + 883 = 4164 **(50,2%)**	74	76 + 90 = 166 *(18,8%)*	45 + 67 = 112 *(12,7%)*

Taigi sąrašui sudaryti ir testuoti naudojant tuos pačius duomenis pasiektas 12,7% tikslumas, o naudojant skirtingus duomenis – 19,2% tikslumas.

6.6 Klitikų radimas remiantis gramatika

Išanalizavus 6.5 skyrelyje gautus rezultatus bus bandoma remiantis lietuvių kalbos gramatika sukurti taisykles, taikomas visai žodžių grupei, o gauti rezultatai bus lyginami su kiekvienam žodžiui atskirai pasiektais statistinio metodo rezultatais.

6.6.1 Kalbos dalys – potencialūs klitikai

Akcentinis žodžio šlijimas labai priklauso nuo to, kokia tai kalbos dalis [Mauricaitė, 1994], todėl pirmiausiai apibrėžkime, tarp kokių kalbos dalių ieškosime klitikų. Dažniausi klitikai yra prielinksniai, jungtukai, dalelytės, rečiau įvardžiai, prieveiksmiai, skaitvardžiai, veiksmažodžiai (sakinyje atliekantys tarinio jungties funkciją). Kitas kalbos dalis laikysime visada kirčiuojamomis. Turimuose tekstuose buvo kirčiuoti visi skaitvardžiai, todėl juos taip pat laikysime visada kirčiuojamais, nors darbe [Mauricaitė, 1986] yra pateikta pavyzdžių su nekirčiuotais skaitvardžiais.

Be to, turimuose tekstuose buvo rasti 6 žodžiai, kurie nepriklauso kalbos dalims, tarp kurių ieškosime klitikų. Dažniausiai tai žodžių skiemenys, parašyti atskirai siekiant parodyti tarimą skiemenimis.

6.6.2 Nekaitomos kalbos dalys

Remiantis 6.5 skyrelio rezultatais galima tikėtis, kad kai kuriems nekaitomiems žodžiams (prielinksniams, jungtukams, dalelytėms) galima sėkmingai taikyti statistinį metodą. Iš ankstesniame skyrelyje aprašyto nekirčiuojamų žodžių sąrašo buvo išrinkti tie nekaitomi žodžiai, kurie daugiau nei du kartus dažniau buvo nekirčiuoti, nei kirčiuoti. Gautas toks 41 žodžio sąrašas:

- prielinksniai: *ant*, *apie*, *be*, *dėl*, *į*, *iki*, *iš*, *ligi*, *nuo*, *pas*, *per*, *po*, *prie*, *pro*, *su*, *tarp*, *ties*, *už*;
- jungtukai: *ar*, *bet*, *ir*, *jog*, *kad*, *kai*, *lyg*, *lygu*, *negu*, *nei*, *nes*, *nors*, *o*, *tad*, *tai*;
- dalelytės: *gi*, *jau*, *juk*, *kaži*, *nė*, *ne*, *nebe*;
- prieveiksmis *tik*.

Čia nurodyta tik ta kalbos dalis, kuria atitinkamas žodis dažniausiai vartojamas. Lyginant su 6.3.1 skyrelyje pateiktu sąrašu, sąrašas ne tik papildytas kai kuriais vienskiemeniais ir dviskiemeniais žodžiais (*apie*, *bei*, *iki*, *kaži*, *ligi*, *lygu*, *nė*, *nebe*, *negu*, *nei*, *nors*, *tad*, *tai*, *ties*), tačiau buvo pastebėta,

kad kai kurie žodžiai gan dažnai buvo kirčiuoti (*kaip, kiek, link*), todėl reikės sudėtingesnių taisyklių, o kai kuriuos geriau iš viso laikyti visada kirčiuojamais (pvz., žodis *dar* 36 kartus aptiktas kirčiuotas ir nė karto bekirtis, žodis *gal* – 11 kartų kirčiuotas ir 1 nekirčiuotas).

Kai kurie žodžiai, pvz., *ligi*, gali būti laikomi keliomis kalbos dalimis, tačiau tiek būdamas prielinksniu, tiek jungtuku, tiek dalelyte, šis žodis linkęs netekti kirčio. Tačiau yra žodžių, kurie būdami viena kalbos dalimi linkę netekti kirčio, o kita kalbos dalimi – kirčiuojami. Pvz., žodis *ant* dažniausiai būna nekirčiuojamas kaip prielinksnis, tačiau gali būti vartojamas ir kaip dalelytė (santrumpa nuo *antai*), kuri linkusi būti kirčiuojama. Analogiškai žodis *ir* dažniausiai būna nekirčiuojamu jungtuku, bet gali būti ir kirčiuotu prieveiksmiu (santrumpa nuo *irgi*). Tačiau pastarieji vartojimo atvejai yra itin reti, todėl laikysime, kad *ant* ir *ir* visada nekirčiuojami.

Turimuose tekstuose anksčiau minėtam 41 nekirčiuojamo žodžio sąrašui priklausantys žodžiai buvo rasti 1354 kartus, 16 kartų tekstuose šie žodžiai buvo kirčiuoti. Taigi minėtas sąrašas padengia (1354-16)/1842=72,6% nekirčiuojamų žodžių, o klaidos sudaro tik 16/1354=1,2%.

6.6.3 Skyrybos ženklo panaudojimas

Kai kirčiuoto ir nekirčiuoto varianto dažniai nedaug skiriasi, kalbos daliai nustatyti tenka taikyti kiek sudėtingesnius algoritmus. Pvz., žodis *o* būdamas jungtukas yra nekirčiuojamas, o jaustukas – kirčiuojamas. Tačiau jaustukas paprastai eina prieš skyrybos ženklą, o jungtukas – ne. Analogiškai dalelytės *ne* ir *nebe* paprastai kirčiuojamos prieš skyrybos ženklą (neigiamai atsakant į klausimą) ir nekirčiuojamos kitais atvejais. Apibrėžkime aibę O&NE = {*o*, *ne*, *nebe*} ir šios aibės kirčiavimui suformuluokim tokią taisyklę:

1 taisyklė. Aibės O&NE žodį kirčiuoti prieš skyrybos ženklą, priešingu atveju nekirčiuoti.

Toliau aibės O&NE žodžiai buvo išimti iš nekirčiuojamų žodžių sąrašo, kuriame lieka 1266 žodžiai (padaroma 14 klaidų), ir kirčiuojami naudojant 1

taisyklę. Rezultatai pateikti 6.5 lentelėje 1 eilutėje. Be to, 6.5 lentelėje palyginimui pateikti ir duomenys, kiek klaidų būtų padaroma kiekvienam grupės žodžiui atskirai taikant statistinį metodą.

6.6.4 Prielinksnių atpažinimas

Kai kurie žodžiai, pvz., *prieš*, *viršum*, *šalia*, gali būti ir prielinksniai (paprastai nekirčiuojami), ir prieveiksmiai (paprastai kirčiuojami). Be to, kai kurie žodžiai, pvz., *link*, *dėlei*, gali būti prielinksniai ir polinksniai (paprastai kirčiuojami). Šiais atvejais prielinksnis atpažįstamas pagal tai, kad po jo eina atitinkamo linksnio žodis. Paprastas, nors nepakankamai patikimas žodžio linksnio atpažinimo būdas – tai pasinaudoti kiekvieno linksnio galūnių sąrašais. Apibrėžkime tokias aibes PRIEL_K= {*viršum*, *šalia*, *link*, *dėlei*} ir PRIEL_G={*prieš*}, ir šioms aibėms suformuluokime tokią taisyklę:

2 taisyklė. Aibės PRIEL_K žodį nekirčiuoti, jei po jo eina žodis kilmininko linksniu, o aibės PRIEL_G žodį nekirčiuoti, jei po jo eina žodis galininko linksniu, priešingu atveju kirčiuoti.

Rezultatai pateikti 6.5 lentelėje 2 eilutėje.

6.6.5 Parodomieji įvardžiai

Dabartinės lietuvių kalbos gramatikoje [Ambrazas, 1996, 252] teigiama, kad įvardžiai *tas, ta, šis, ši* gali būti vartojami daiktavardiškai (pvz., *Tàs ir sãko.*) arba būdvardiškai (pvz., *Tas põnas ir sãko.*). Buvo pastebėta, kad daiktavardiškai vartojami šie įvardžiai paprastai yra kirčiuojami, o būdvardiškai – nekirčiuojami. Kai žodis vartojamas būdvardiškai, po jo eina to paties linksnio žodis. Apibrėžkime parodomųjų įvardžių aibę PAROD_IV = {*tas, to, tam, tą, tuo, tame, tie, tų, tiems, tuos, tais, tuose, ta, tos, toje, toms, tomis, tose, šis, šio, šiam, šį, šiuo, šiame, šie, šių, šiems, šiuos, šiais, šiuose, ši, šios, šiai, šią, šia, šioje, šioms, šias, šiomis, šiose*}. Aibės PAROD_IV žodžius galima sugrupuoti pagal linksnius. Aibėje nėra įvardžio *tai*, nes šis žodis daug

dažniau vartojamas kaip jungtukas. Aibės PAROD_IV žodžiams kirčiuoti naudosime taisyklę:

3 taisyklė. Aibės PAROD_IV žodį nekirčiuoti, jei po jo eina to paties linksnio žodis, priešingu atveju kirčiuoti.

Rezultatai pateikti 6.5 lentelėje 3 eilutė. Naudojant paprastą linksnių galūnių paiešką padaroma 14 klaidų, jei pavyktų visur teisingai atpažinti linksnius, klaidų liktų 11.

6.7 Klitikų nustatymas remiantis konteksto kirčiavimu

Kai kurių žodžių kirčiavimas labai priklauso nuo to, ar kirčiuoti greta esantys žodžiai. Darbe [Mauricaitė, 1994] tarp tokių žodžių minimi įvardžiai *aš, tu, jis, ji,* prieveiksmis *čia*, dalelytė *vis*. Peržvelgus turimus pavyzdžius, dar buvo nuspręsta panagrinėti veiksmažodžio *būti* formas, įvardžių *koks, kokia, kuris, kuri* grupę, įvardžių *šis, ši, tas, ta* grupę, įvardžių ir prieveiksmių *kas, kada, kaip, kiek* grupę.

Buvo atlikti tokie eksperimentai: pasirenkama tam tikra žodžių aibė, kuriai bus bandoma sukurti vieną taisyklę. Surenkama statistika (turimuose daugiau kaip 8000 žodžių tekstuose), kiek kartų šios grupės žodžiai buvo kirčiuoti / nekirčiuoti tam tikrame kontekste. Buvo nagrinėti tokie kontekstai: a) po skyrybos ženklo, b) po nekirčiuoto žodžio, c) po kirčiuoto skiemens, d) po vieno nekirčiuoto skiemens kirčiuotame žodyje, e) po dviejų ir daugiau nekirčiuotų skiemenų kirčiuotame žodyje, f) prieš skyrybos ženklą, g) prieš nekirčiuotą žodį, h) prieš kirčiuotą skiemenį, i) prieš vieną nekirčiuotą skiemenį kirčiuotame žodyje, j) prieš du nekirčiuotus skiemenis kirčiuotame žodyje, k) prieš tris ir daugiau nekirčiuotus skiemenis kirčiuotame žodyje. Nekirčiuotas vienskiemenis žodis ir vienas nekirčiuotas kirčiuoto žodžio skiemuo gali turėti visai skirtingą įtaką, todėl tai buvo laikoma skirtingais kontekstais. Statistika surašoma į lentelę. Remiantis lentele sukuriama taisyklė,

kiekvienai kontekstų „po“ ir „prieš“ porai nusakanti, kirčiuoti ar nekirčiuoti aibei priklausančius žodžius.

6.7.1 Asmeniniai įvardžiai

Pirmiausiai buvo atlikti eksperimentai su įvardžių *aš, tu, jis, ji* įvairiomis gramatinėmis formomis (iš viso 61 skirtingas žodis). Iš karto buvo pastebėta, kad kai kurios gramatinės formos elgiasi visiškai skirtingai, pvz., *aš*, *tu* lyginant su *man, tau*, todėl buvo bandoma jas skaidyti į dvi grupes ir sukurti dvi taisykles. Siekiant, kad taisyklės būtų kiek galima bendresnės, mažiau priklausytų nuo konkrečių duomenų, vienai grupei buvo priskiriami visi tam tikrą linksnį turintys ar tam tikra savybe pasižymintys žodžiai.

6.2 lentelė. Asmeninių įvardžių kirčiavimo priklausomybė nuo konteksto kirčiavimo

prieš / po	skyrybos ženklą	nekirčiuotą žodį	kirčiuotą skiemenį	1 nekirčiuotą skiemenį	2 nekirčiuotus skiemenis	3 ir daugiau nekirčiuotų skiemenų
skyrybos ženklo	**1/3**	*9/2*	**0/30**	**3/14**	*3/2*	*1/0*
nekirčiuoto žodžio	*4/0*	*9/0*	*11/9*	*17/9*	*7/1*	–
kirčiuoto skiemens	**0/3**	**2/6**	**2/16**	**2/17**	**1/2**	**0/1**
1 nekirčiuoto skiemens	**1/2**	4/4	**1/13**	**1/10**	**1/3**	–
2 ir daugiau nekirčiuotų skiemenų	–	**1/2**	**0/1**	*2/1*	1/1	–

Perrinkus įvairius grupavimo variantus nustatyta, kad mažiausia klaidų padaroma, kai į vieną grupę priskiriami visi vienskiemeniai žodžiai, išskyrus naudininko linksnį. Apibrėžkim asmeninių įvardžių aibę iš 20 žodžių: ASM_IV = {*aš*, *mes*, *mūs*, *mus*, *tu*, *jūs*, *jus*, *jis*, *jo*, *jį*, *juo*, *jie*, *jų*, *juos*, *jais*, *ji*, *jos*, *ją*, *ja*, *jas*}. Turimuose duomenyse rasta 84 kirčiuoti ir 152 nekirčiuoti šios grupės žodžiai. Jų kirčiavimo statistika įvairiuose kontekstuose pateikta 6.2 lentelėje.

6.2 lentelėje pirmas skaičius rodo, kiek kirčiuotų, antras – kiek nekirčiuotų įvardžių. Jei daugiau nekirčiuotų, skaičiai paryškinti, jei daugiau

kirčiuotų – parašyta kursyvu. Siekdami kiek supaprastinti taisyklę, laikykim, kad antroje eilutėje visus reikia kirčiuoti, o paskutinėse trijose eilutėse – visus nekirčiuoti. Dabar taisyklė atrodo taip:

4 taisyklė. Aibės ASM_IV žodžius kirčiuoti po skyrybos ženklo prieš nekirčiuotą žodį arba prieš du ar daugiau nekirčiuotų skiemenų kirčiuotame žodyje, taip pat kirčiuoti po nekirčiuoto žodžio, kitais atvejais nekirčiuoti.

Rezultatus žr. 6.5 lentelėje 4 eilutė. Likusius asmeninius įvardžius (41 gramatinę formą) geriausia visada kirčiuoti. Jų rasta 88 ir padaromos 9 klaidos.

6.7.2 Veiksmažodis būti

Kita nagrinėta grupė buvo žodžio *būti* gramatinės formos: BUTI = {*buvo, bus, buvau, buvai, yra, esu, esi, nesu, nesi, nėra*}. Rasta 64 kirčiuoti ir 30 nekirčiuotų grupės žodžių, jiems gauta statistika pateikta 6.3 lentelėje.

6.3 lentelė. Veiksmažodžio *būti* formų kirčiavimo priklausomybė nuo konteksto kirčiavimo

prieš / po	skyrybos ženklą	nekirčiuotą žodį	kirčiuotą skiemenį	1 nekirčiuotą skiemenį	2 nekirčiuotus skiemenis	3 ir daugiau nekirčiuotų skiemenų
skyrybos ženklo	*1/0*	*1/0*	*3/0*	*3/1*	–	–
nekirčiuoto žodžio	*5/0*	*5/0*	*3/1*	*5/0*	*1/0*	*2/0*
kirčiuoto skiemens	*6/0*	*3/1*	**3/5**	**2/3**	**2/3**	–
1 nekirčiuoto skiemens	*4/0*	*1/0*	**3/8**	**5/6**	**1/2**	*1/0*
2 ir daugiau nekirčiuotų skiemenų	*1/0*	–	*2/0*	*1/0*	–	–

Pagal ją galima užrašyti taisyklę:

5 taisyklė. Aibės BUTI žodžius nekirčiuoti po kirčiuoto skiemens arba po vieno nekirčiuoto skiemens kirčiuotame žodyje prieš kirčiuotą skiemenį, 1 arba 2 nekirčiuotus skiemenis, kitais atvejais kirčiuoti.

Rezultatai 6.5 lentelėje 5 eilutė.

6.7.3 Kitos žodžių grupės

Analogiški eksperimentai atlikti su klausiamųjų įvardžių grupe KLAUS_IV = {*koks, kokio, kokiam, kokį, kokiu, kokie, kokių, kokiems, kokius, kokiais, kokia, kokios, kokiai, kokią, kokioms, kokias, kurs, kuris, kurio, kuriam, kurį, kuriuo, kurie, kurių, kuriems, kuriuos, kuriais, kuri, kurios, kuriai, kurią, kuria, kurioms, kurias*}, apimančia 34 skirtingus vienskiemenius ir dviskiemenius žodžius, taip pat prieveiksmiu *čia*, dalelyte *vis*. Gautos analogiškos statistikos lentelės, jomis remiantis užrašytos tokios trys taisyklės:

6 taisyklė. Aibės KLAUS_IV žodžius nekirčiuoti prieš kirčiuotą skiemenį po skyrybos ženklo arba kirčiuoto skiemens, kitais atvejais kirčiuoti.

7 taisyklė. Žodį *čia* kirčiuoti po skyrybos ženklo prieš nekirčiuotą žodį arba kirčiuotą skiemenį, po nekirčiuoto žodžio prieš skyrybos ženklą, nekirčiuotą žodį, vieną ar daugiau nekirčiuotų skiemenų, kitais atvejais nekirčiuoti.

8 taisyklė. Žodį *vis* nekirčiuoti prieš nekirčiuotą žodį po kirčiuoto žodžio, prieš kirčiuotą skiemenį po vieno ar daugiau nekirčiuotų skiemenų.

Rezultatai 6.5 lentelėje atitinkamai 6, 7 ir 8 eilutėse. Analogiški eksperimentai atlikti ir su įvardžių *šis, ši, tas, ta* gramatinėmis formomis, tačiau jiems šis metodas nepasiteisino, nes buvo gauta 21 klaida, o tai daugiau, nei taikant 5.5. skyrelyje aprašytą metodą (kur buvo padaryta 14 klaidų).

6.7.4 Klausiamieji įvardžiai ir prieveiksmiai

Dar sudėtingesnės taisyklės reikia klausiamųjų įvardžių ir prieveiksmių grupei KI&P = {*kas, ko, kieno, kam, ką, kuo, kur, kame, kada, kaip, kiek, kodėl*}. Pagal parengtą statistikos lentelę užrašius taisyklę padaroma 40 klaidų. Buvo pastebėta, kad minėti žodžiai gali būti vartojami klausimuose, ir šiuo

atveju jie paprastai būna kirčiuoti, todėl daliai KI&P aibės žodžių suformuluota tokia taisyklė:

9a taisyklė. Jei KI&P aibės žodis eina po skyrybos ženklo, o sakinys baigiasi klaustuku, žodis kirčiuojamas.

Ši taisyklė apima 32 žodžius, padarytos 2 klaidos (2 žodžiai turėjo likti nekirčiuoti). Likusiems (9a taisyklei nepriklausantiems) 138 žodžiams buvo sudaryta 6.4 lentelė ir pagal ją užrašyta tokia taisyklė:

9b taisyklė. Aibės KI&P žodis kirčiuojamas prieš skyrybos ženklą ir nekirčiuotą žodį, taip pat po nekirčiuoto žodžio prieš du ar daugiau nekirčiuotų skiemenų.

6.4 lentelė. Klausiamųjų įvardžių ir prieveiksmių kirčiavimo priklausomybė nuo konteksto kirčiavimo

po \ prieš	skyrybos ženklą	nekirčiuotą žodį	kirčiuotą skiemenį	1 nekirčiuotą skiemenį	2 nekirčiuotus skiemenis	3 ir daugiau nekirčiuotų skiemenų
skyrybos ženklo	–	*15/9*	**5/27**	**5/19**	***0/3***	–
nekirčiuoto žodžio	*2/0*	*2/0*	**1/5**	**2/4**	*2/0*	–
kirčiuoto skiemens	–	*2/1*	**0/7**	**0/12**	***0/1***	–
1 nekirčiuoto skiemens	–	–	**0/5**	**0/4**	***0/1***	–
2 ir daugiau nekirčiuotų skiemenų	–	–	***0/2***	–	***0/2***	–

Taikant 9b taisyklę padaromos 26 klaidos. Taisyklės 9a ir 9b buvo sujungtos į vieną 9 taisyklę. Rezultatus žr. 6.5 lentelėje 9 eilutė.

Taigi taikant specializuotas taisykles žodžių grupėms gauta 15,7% klaidų, o tai 8,7% mažiau, nei remiantis atskirai kiekvienam žodžiui apskaičiuotu kirčiavimo / nekirčiavimo dažniu.

6.5 lentelė. Sudarytų taisyklių ir statistinio metodo palyginimas

Taisyklės Nr.	Žodžių aibė (*pavyzdžiai*)	Kirčiuotų žodžių + nekirčiuotų žodžių = suma, (**dalis tarp visų žodžių**)	Pirmo tipo klaidų + antro tipo klaidų = suma *(klaidų ir žodžių santykis)*	
			Taikyta taisyklė žodžių grupei	Taikyta statistika atskiriems žodžiams
1	O&NE (*o, ne, nebe*)	2 + 86 = 88 **(11,3%)**	0 + 0 = 0 *(0,0%)*	2 + 0 = 2 *(2,3%)*
2	PRIEL_K (*link*), PRIEL_G (*prieš*)	2 + 6 = 8 **(1,0%)**	1 + 0 = 1 *(12,5%)*	0 + 2 = 2 *(25,0%)*
3	PAROD_IV (*tas, šis*)	35 + 56 = 91 **(11,7%)**	9 + 5 = 14 *(15,4%)*	19 + 3 = 22 *(24,2%)*
4	ASM_IV (*aš, tu, jis*)	84 + 152 = 236 **(30,3%)**	23 + 23 = 46 *(19,5%)*	68 + 1 = 69 *(29,2%)*
5	BUTI (*yra, esu*)	64 + 30 = 94 **(12,1%)**	16 + 3 = 19 *(20,2%)*	0 + 30 = 30 *(31,9%)*
6	KLAUS_IV (*koks, kuris*)	19 + 20 = 39 **(5,0%)**	2 + 7 = 9 *(23,1%)*	3 + 3 = 6 *(15,4%)*
7	*čia*	10 + 23 = 33 **(4,2%)**	2 + 2 = 4 *(12,1%)*	10 + 0 = 10 *(30,3%)*
8	*vis*	6 + 13 = 19 **(2,4%)**	3 + 0 = 3 *(15,8%)*	6 + 0 = 6 *(31,6%)*
9	KI&P (*kas, kaip*)	65 + 105 = 170 **(21,9%)**	13 + 13 = 26 *(15,3%)*	16 + 26 = 42 *(24,7%)*
Iš viso		287 + 491 = 778 **(100%)**	69 + 53 = 122 *(15,7%)*	124 + 65 = 189 *(24,3%)*

6.7.5 Nuo konteksto priklausomų taisyklių sąveika

Sudarydami šiame skyriuje aprašytas taisykles laikėmės nuostatos, kad žodžio konteksto kirčiavimas jau yra žinomas, ir tereikia nustatyti tik vieno žodžio kirčiavimą. Iš tikrųjų gali greta eiti keli žodžiai, kirčiuojami pagal čia aprašytas taisykles, ir tokiu atveju konteksto kirčiavimas dar nėra žinomas. Taigi paskutinis klausimas, kurį liko išnagrinėti, tai šiame skyriuje aprašytų taisyklių sąveiką. Turimuose daugiau kaip 8000 žodžių tekstuose buvo rasti 52 atvejai (apimantys 104 žodžius), kai greta eina pagal anksčiau aprašytas taisykles kirčiuojami žodžiai. Su jais išbandžius žodžių peržiūrėjimą iš kairės ir iš dešinės, o greta esantį nežinomo kirčiavimo žodį laikant kirčiuotu, po to nekirčiuotu, buvo nustatyta, kad geriausia tokias žodžių sekas peržiūrėti iš kairės, aibių ASM_IV, KI&P žodžius bei žodžius *čia* ir *vis* laikyti

nekirčiuotais, o kitus kirčiuotais. Yra dvi išimtys: 1) aibės ASM_IV žodžius po KI&P žodžių laikyti kirčiuotais. 2) radus aibės BUTI žodį, po kurio eina aibės KI&P žodis, pirmiau sukirčiuoti KI&P žodį (aibės BUTI žodį laikyti kirčiuotu), tada grįžti prie aibės BUTI žodžio.

Panagrinėkim tokį pavyzdį: *Aš jį čia buvau matęs.* Pateiktame pavyzdyje *Aš* yra po skyrybos ženklo prieš nekirčiuotą žodį (*jį* laikom nekirčiuotu), todėl *Aš* kirčiuojam (4 taisyklė), dabar žodis *jį* yra po kirčiuoto skiemens prieš nekirčiuotą žodį (*čia* laikom nekirčiuotu), todėl nekirčiuojam (4 taisyklė). Žodis *čia* yra po nekirčiuoto žodžio prieš kirčiuotą žodį (žodį *buvau* laikom kirčiuotu, bet pirmas skiemuo nekirčiuotas, o antras kirčiuotas), todėl kirčiuojam (7 taisyklė). Pagaliau *buvau* eina po kirčiuoto skiemens prieš kirčiuotą skiemenį, todėl nekirčiuojam (5 taisyklė). Taigi visą sakinį siūloma tarti taip: *Àš jį čià buvau mãtęs.*

6.8 Bendras algoritmas, testavimo rezultatai, tobulinimas

Dabar bendrą klitikų paieškos algoritmą galima nusakyti tokiais 4 žingsniais:

1) Remiantis samplaikinių formų sąrašu rasti samplaikines formas ir jas sukirčiuoti.
2) Remiantis nekirčiuojamų žodžių sąrašu rasti visus nekirčiuojamus žodžius. Sukirčiuoti žodžius remiantis 1-3 taisyklėmis.
3) Pasižymėti žodžius, kuriems bus taikomos 4-9 taisyklės, visus kitus žodžius kirčiuoti.
4) Pritaikyti 4-9 taisykles.

Remiantis anksčiau aprašytomis taisyklėmis bei jų sąveika buvo sukurtos kompiuterinės programos. Testavimas buvo atliktas su anksčiau taisyklėms sudaryti naudotais tekstais (daugiau kaip 8000 žodžių) bei tekstais (beveik 1000 žodžių), kurie iki šiol dar nebuvo naudoti. Testavimo su abiem tekstais rezultatai pateikti 6.6 lentelėje.

6.6 lentelė. Testavimo rezultatai.

Taisyklės Nr.	Žodžių aibė (*pavyzdžiai*)	Tie patys tekstai taisyklių sudarymui ir testavimui		Skirtingi tekstai taisyklių sudarymui ir testavimui	
		Kirčiuotų žodžių + nekirčiuotų žodžių = suma, (**dalis tarp visų žodžių**)	Pirmo tipo klaidų + antro tipo klaidų = suma (*klaidų ir žodžių santykis*)	Kirčiuotų žodžių + nekirčiuotų žodžių = suma, (**dalis tarp visų žodžių**)	Pirmo tipo klaidų + antro tipo klaidų = suma (*klaidų ir žodžių santykis*)
1	O&NE (*o, ne, nebe*)	2 + 86 = 88	0 + 0 = 0 *(0,0%)*	0 + 12 = 12	0 + 0 = 0 *(0,0%)*
2	PRIEL_K (*link*), PRIEL_G (*prieš*)	2 + 6 = 8	1 + 0 = 1 *(12,5%)*	0 + 3 = 3	0 + 1 = 1 *(33,3%)*
3	PAROD_IV (*tas, šis*)	35 + 56 = 91	9 + 5 = 14 *(15,4%)*	5 + 2 = 7	1 + 0 = 1 *(14,3%)*
4	ASM_IV (*aš, tu, jis*)	84 + 152 = 236	23 + 26 = 49 *(20,8%)*	11 + 27 = 38	7 + 7 = 14 *(36,8%)*
5	BUTI (*yra, esu*)	64 + 30 = 94	15 + 4 = 19 *(20,2%)*	8 + 2 = 10	2 + 1 = 3 *(30,0%)*
6	KLAUS_IV (*koks, kuris*)	19 + 20 = 39	2 + 7 = 9 *(23,1%)*	3 + 1 = 4	1 + 0 = 1 *(25,0%)*
7	*čia*	10 + 23 = 33	2 + 0 = 2 *(6,1%)*	2 + 4 = 6	0 + 0 = 0 *(0,0%)*
8	*vis*	6 + 13 = 19	3 + 0 = 3 *(15,8%)*	0 + 1 = 1	0 + 0 = 0 *(0,0%)*
9	KI&P (*kas, kaip*)	65 + 105 = 170	17 + 14 = 31 *(18,2%)*	5 + 12 = 17	4 + 2 = 6 *(35,3%)*
Iš viso žodžių taisyklėse		287 + 491 = 778 **(9,3%)**	72+56 = 128 *(16,5%)*	34 + 64 = 98 **(10,2%)**	15 + 11 = 26 *(26,5%)*
Samplaikinės formos		49 + 49 = 98 **(1,2%)**	1 + 1 = 2 *(2,0%)*	9 + 9 = 18 **(1,9%)**	0 + 0 = 0 *(0,0%)*
Nekirč. žodžių sąrašas		14 + 1252 = 1266 **(15,1%)**	14 + 0 = 14 *(1,1%)*	3 + 124 = 127 **(13,3%)**	3 + 0 = 3 *(2,4%)*
Kiti žodžiai		6205 + 50 = 6255 **(74,5%)**	0 + 50 = 50 *(0,8%)*	702 + 12 = 714 **(74,6%)**	0 + 12 = 12 *(1,7%)*
Iš viso		6555 + 1842 = 8397	87+107 = 194 *(2,3%)*	745+207 = 957 **(100%)**	18 + 23 = 41 *(4,3%)*

Remiantis 6.6 lentele galima pastebėti, kad dėl taisyklių sąveikos atsirado 6 papildomos klaidos. Atliekant testavimą su tais pačiais duomenimis, kuriais remiantis buvo kuriamos taisyklės, gauta 2,3% klaidų, o klaidų santykis su nekirčiuojamais žodžiais yra 10,5%. Atlikus testavimą su dar nenaudotais duomenimis gauta 4,3% klaidų, o klaidų santykis su nekirčiuojamais žodžiais yra 19,8%. Geriausius rezultatus davė 1 taisyklė, samplaikinės formos ir

nekirčiuojamų žodžių sąrašas. Be to, atrinkus žodžius visoms taisyklėms, lieka labai nedaug nekirčiuojamų žodžių.

6.6 lentelėje pateiktas žodžių, kuriems buvo pritaikytos tam tikros grupės taisyklės, ir visų žodžių santykis rodo taisyklių grupės naudojimo dažnį. Nekirčiuoti žodžiai dažniausiai randami remiantis nekirčiuojamų žodžių sąrašu.

Kalbant apie sukurto algoritmo tobulinimą, pirmiausia reikia atkreipti dėmesį į tai, kad sudarant taisykles ir nekaitomų žodžių sąrašą naudojami tik tie žodžiai, kurie pasitaikė tekstuose. Svarbu yra tai, kad nekaitomų tarnybinių žodžių skaičius lietuvių kalboje yra baigtinis, todėl [Keinys, 1993] galima rasti visus nekaitomus tarnybinius lietuvių kalbos žodžius. Toliau apsiribosime tik vienskiemeniais ir dviskiemeniais žodžiais. Kadangi remiantis naudotais tekstais visi prielinksniai pakliuvo į nekirčiuojamų žodžių sąrašą, todėl ir kitus prielinksnius laikysime nekirčiuojamais. Remiantis [Keinys, 1993] nekirčiuojamų žodžių sąrašą galima papildyti tokiais 14 prielinksnių: *anot, apsuk, aukščiau, lig, pagal, palei, pasak, pirm, sulig, užu, vidur, virš, viršuj, žemiau*. Jungtukai taip pat dažniausiai nekirčiuojami (išskyrus jungtuką *jei* ir jam sinonimišką jungtuką *jeigu*), todėl nekirčiuojamų žodžių sąrašą papildome 19 jungtukų: *arba, begu, bei, betgi, būtent, idant, ik, nebent, neg, norint, pakol, tačiau, tartum, tegu, tegul, tiktai, užuot, vienok, visgi*. Taip pat galima papildyti ir 2 taisykle apibūdinamus žodžius, kurie gali būti ir prielinksniai, ir kitos kalbos dalys. Aibę PRIEL_K papildome tokiais 6 žodžiais: *arti, greta, paskiau, pusiau, šiapus, vietoj*, o aibę PRIEL_G tokiais 4 žodžiais: *aplink, paskui, paskum, priešais*. Sudėtingesnė situacija su likusiomis (maždaug 65) dalelytėmis. Be išsamesnių tyrimų negalima nustatyti, kurios dalelytės linkusios netekti kirčio, o kurios jį išlaiko.

Anksčiau aprašytu būdu papildžius sukurto algoritmo žodžių sąrašus ir patikrinus su testavimo duomenimis klaidų skaičius sumažėjo iki 39, t. y. 4,1% klaidų, o klaidų santykis su nekirčiuojamais žodžiais 18,8%.

6.9 Šeštojo skyriaus apibendrinimas

- Šnekamajai kalbai būdinga ritmika, t. y. kirčiuotų ir nekirčiuotų skiemenų kaitaliojimas. Siekiant išvengti greta esančių kirčiuotų skiemenų, kai kurie žodžiai lieka nekirčiuoti (tampa klitikais).
- Šiame skyriuje nekirčiuojamų žodžių paieškai lietuvių kalbos tekste pasiūlyti keturių tipų metodai: pagrįsti samplaikinių formų atpažinimu, žodžio kirčiavimo/nekirčiavimo statistiniu dažniu, gramatikos taisyklėmis bei gretimų žodžių kirčiavimu.
- Kiekvienam metodui apibrėžtos žodžių klasės, kurioms jis geriausiai tinka. Paaiškinta, kaip visus metodus sujungti į vieną algoritmą.
- Kuriant algoritmą buvo stengiamasi minimizuoti pirmosios ir antrosios rūšies klaidų sumą. Taikant šį algoritmą testavimo duomenims gauta 4,1% klaidų tarp visų žodžių, o klaidų ir nekirčiuotų žodžių santykis yra 18,8%.

Rezultatai ir išvados

Šiame darbe nagrinėjamas lietuvių kalbos teksto automatinis kirčiavimas bei su tuo susiję kiti du uždaviniai – homografų vienareikšminimas ir klitikų paieška.

1) Darbe nusakyta automatinio kirčiavimo, homografų vienareikšminimo ir klitikų paieškos vieta bendroje balso sintezės schemoje, jų sąveika su kitais moduliais, gaunami ir perduodami duomenys. Išnagrinėti metodai, taikyti šiems uždaviniams spręsti kitose kalbose, jų pasirinkimas atsižvelgiant į kalbos kaitymo laipsnį ir kirčiavimo paradigmą.
2) Lietuvių kalbos tekstui kirčiuoti pritaikius morfologinėmis taisyklėmis grįstus metodus, kai kuriuos žodžius (homografus) galima sukirčiuoti keliais būdais. Norint padidinti kirčiuotų žodžių skaičių, reikalingas vienareikšminimo algoritmas. Autoriaus pasiūlytas homografų vienareikšminimo algoritmas, pagrįstas leksemų ir morfologinių pažymų dažniais. Kai kurių leksemų buvimas žodyne labiau kliudo kirčiuoti, nei padeda. Jas atmetus, teisingų kirčiavimo hipotezių dalis tarp visų hipotezių (mokymo duomenims) padidėja 6,1%. Sudarytas morfologinių pažymų dažniais grįstų taisyklių rinkinys (1215 taisyklių), kuris mokymo duomenims leidžia teisingą kirčiavimo variantą parinkti 84,3% tikslumu. Pritaikius pasiūlytus algoritmus tekstui kirčiuoti, homografus pavyko vienareikšminti 85,01% tikslumu. Nors pasiūlytas algoritmas nenaudoja jokios informacija apie kontekstą, tačiau gauti rezultatai artimi kontekstinę informaciją naudojančiam ID3 algoritmui.
3) Morfologinėmis taisyklėmis grįsti lietuvių kalbos teksto kirčiavimo metodai pasižymi sudėtingumu, todėl juos problematiška perkelti į kitas sistemas ar programavimo kalbas. Šiame darbe autoriaus pasiūlyti metodai, kurie remiasi tik raidžių sekomis ir nereikalauja jokių žinių apie kalbą: morfemas, kalbos dalis, žodžių kaitymą, skiemenavimą ir pan., o kirčiavimo taisyklės yra itin paprastos ir sudaromos tiesiog iš kirčiuotų

žodžių sąrašo. Tokie metodai paprastai taikomi nefleksinėms kalboms. Taisyklėms sudaryti naudotas sprendimo medžių algoritmas. Nagrinėti keli taisyklių sudarymo būdai. Parodyta, kad didžiausią tikslumą (95,53%) duoda pabaigos-pradžios taisyklių metodas, o mažiausiai taisyklių gaunama taikant žodžio vidurio taisyklių metodą ir kai taisyklių aibė sumažinama naudojant vieną iš autoriaus pasiūlytų algoritmų. Savo tikslumu pasiūlytas metodas tik 0,8% nusileidžia morfologinėmis taisyklėmis grįstam metodui, tačiau parodyta, kad, didinant taisyklėms sudaryti naudojamų tekstų kiekį, galima tikėtis šį tikslumą pasiekti ir aplenkti.

4) Šnekamajai kalbai būdinga ritmika, t. y. kirčiuotų ir nekirčiuotų skiemenų kaitaliojimas. Siekiant išvengti greta esančių kirčiuotų skiemenų, kai kurie žodžiai lieka nekirčiuoti (tampa klitikais). Nekirčiuojamų žodžių paieškai lietuvių kalbos tekste autoriaus pasiūlyti keturių tipų metodai: pagrįsti samplaikinių formų atpažinimu, žodžio kirčiavimo/nekirčiavimo statistiniu dažniu, gramatikos taisyklėmis bei gretimų žodžių kirčiavimu. Kiekvienam metodui apibrėžtos žodžių klasės, kurioms jis geriausiai tinka, bei paaiškinta, kaip visus metodus sujungti į vieną algoritmą. Minimizuojant pirmosios ir antrosios rūšies klaidų sumą, testavimo duomenims gauta 4,1% klaidų tarp visų žodžių, o klaidų ir nekirčiuotų žodžių santykis yra 18,8%.

Priedai

1 priedas. Teksto kirčiavimo eksperimentų rezultatai (iš viso 197362 žodžiai)

Gramatinių formų dažniais paremtos taisyklės	*Kamienų atmetimo algoritmas*	*Atsitiktinis kirčiavimo varianto parinkimas*	*Žodžiai su vienu kirčiu*			*Žodžiai su*
			klaidingai nesukirčiuoti	teisingai sukirčiuoti	klaidingai sukirčiuoti	klaidingai nesukirčiuoti
Taisyklės nenaudotos	-	-	156 (0,08%)	***130072 (65,91%)***	98 (0,05%)	39 (0,02%)
		+	195 (0,10%)	145930 (73,94%)	14455 (7,32%)	0 (0,00%)
	I	-	156 (0,08%)	***130692 (66,22%)***	149 (0,08%)	39 (0,02%)
		+	195 (0,10%)	146117 (74,04%)	14269 (7,23%)	0 (0,00%)
	I ir II	-	126 (0,06%)	137696 (69,77%)	428 (0,22%)	27 (0,01%)
		+	153 (0,08%)	151016 (76,52%)	8131 (4,12%)	0 (0,00%)
Grupės A, B ir C	-	-	194 (0,10)%	150606 (76,31%)	3685 (1,87%)	1 (0,00%)
		+	195 (0,10)%	153299 (77,67%)	7086 (3,59%)	0 (0,00%)
	I	-	194 (0,10)%	151267 (76,64%)	3694 (1,87%)	1 (0,00%)
		+	195 (0,10)%	153703 (77,88%)	6683 (3,39%)	0 (0,00%)
	I ir II	-	153 (0,08)%	154459 (78,26%)	3653 (1,85%)	0 (0,00%)
		+	153 (0,08)%	155003 (78,54%)	4144 (2,10%)	0 (0,00%)
Grupės A, B, C ir D	-	-	195 (0,10)%	154825 (78,45%)	4505 (2,28%)	0 (0,00%)
		+	195 (0,10)%	155372 (78,72%)	5013 (2,54%)	0 (0,00%)
	I	-	195 (0,10)%	***155486 (78,78%)***	4494 (2,28%)	0 (0,00%)
		+	195 (0,10)%	***155743 (78,91%)***	4643 (2,35%)	0 (0,00%)
	I ir II	-	153 (0,08)%	155142 (78,61%)	3848 (1,95%)	0 (0,00%)
		+	153 (0,08)%	155249 (78,66%)	3898 (1,98%)	0 (0,00%)

Geriausius rezultatus davusi vienareikšminimo algoritmų seka (žr. 4.10 lentelę):

(130692 – 130072)/(30197 – 29517) = ***91,18%*** *30197 – 29517* = ***680***
(155486 – 130692)/(29517 – 379) = ***85,09%*** *29517 – 379* = ***29138***
(155743 – 155486)/(379 – 0) = ***67,81%*** *379 – 0* = ***379***
(155743 – 130072)/(30197 – 0) = ***85,01%*** *30197 – 0* = ***30197***

daug kirčiavimo variantų		*Nekirčiuoti (pvz., užsienio k.) žodžiai*		*Klitikai*		*Klaidos (iš viso)*	
teisingai sukirčiuoti (vienas kirt. iš buvusių)	klaidingai sukirčiuoti	teisingai nesukirčiuoti	klaidingai sukirčiuoti	teisingai nesukirčiuoti	klaidingai sukirčiuoti	klaidingai	klaidingai + daug kirčiavimo variantų
30197 (15,30%)	18 (0,01%)	6084 (3,08%)	3938 (2,00%)	26622 (13,49%)	138 (0,07%)	4387 (2,22%)	34584 (17,52%)
0 (0,00%)	0 (0,00%)	6084 (3,08%)	3938 (2,00%)	26622 (13,49%)	138 (0,07%)	18726 (9,49%)	18726 (9,49%)
29517 (14,96%)	28 (0,01%)	6084 (3,08%)	3937 (1,99%)	26622 (13,49%)	138 (0,07%)	4447 (2,25%)	33964 (17,21%)
0 (0,00%)	0 (0,00%)	6084 (3,08%)	3937 (1,99%)	26622 (13,49%)	138 (0,07%)	18539 (9,39%)	18539 (9,39%)
20967 (10,62%)	56 (0,03%)	6126 (3,10%)	5176 (2,62%)	26622 (13,49%)	138 (0,07%)	5951 (3,02%)	26918 (13,64%)
0 (0,00%)	0 (0,00%)	6126 (3,10%)	5176 (2,62%)	26622 (13,49%)	138 (0,07%)	13598 (6,89%)	13598 (6,89%)
6018 (3,05%)	76 (0,04%)	6084 (3,08%)	3938 (2,00%)	26622 (13,49%)	138 (0,07%)	8032 (4,07%)	14050 (7,12%)
0 (0,00%)	0 (0,00%)	6084 (3,08%)	3938 (2,00%)	26622 (13,49%)	138 (0,07%)	11357 (5,75%)	11357 (5,75%)
5362 (2,72%)	63 (0,03%)	6084 (3,08%)	3937 (1,99%)	26622 (13,49%)	138 (0,07%)	8027 (4,07%)	13389 (6,78%)
0 (0,00%)	0 (0,00%)	6084 (3,08%)	3937 (1,99%)	26622 (13,49%)	138 (0,07%)	10953 (5,55%)	10953 (5,55%)
1033 (0,52%)	2 (0,00%)	6126 (3,10%)	5176 (2,62%)	26622 (13,49%)	138 (0,07%)	9122 (4,62%)	10155 (5,15%)
0 (0,00%)	0 (0,00%)	6126 (3,10%)	5176 (2,62%)	26622 (13,49%)	138 (0,07%)	9611 (4,87%)	9611 (4,87%)
1027 (0,52%)	28 (0,01%)	6084 (3,08%)	3938 (2,00%)	26622 (13,49%)	138 (0,07%)	8804 (4,46%)	9831 (4,98%)
0 (0,00%)	0 (0,00%)	6084 (3,08%)	3938 (2,00%)	26622 (13,49%)	138 (0,07%)	9284 (4,70%)	9284 (4,70%)
379 (0,19%)	27 (0,01%)	6084 (3,08%)	3937 (1,99%)	26622 (13,49%)	138 (0,07%)	8791 (4,45%)	9170 (4,65%)
0 (0,00%)	0 (0,00%)	6084 (3,08%)	3937 (1,99%)	26622 (13,49%)	138 (0,07%)	8913 (4,52%)	8913 (4,52%)
157 (0,08%)	0 (0,00%)	6126 (3,10%)	5176 (2,62%)	26622 (13,49%)	138 (0,07%)	9315 (4,72%)	9472 (4,80%)
0 (0,00%)	0 (0,00%)	6126 (3,10%)	5176 (2,62%)	26622 (13,49%)	138 (0,07%)	9365 (4,75%)	9365 (4,75%)

2 priedas. Teksto kirčiavimo eksperimentų rezultatų vidurkiai (iš viso 197193,4 žodžių (vidurkis))

	Taisyklių skaičius			*Kirčiuoti žodžiai*					
Metodas	pilnų žodžių taisyklių	žodžio pradžios taisyklių	žodžio pabaigos taisyklių	klaidingai nesu-kirčiuoti	%	teisingai sukirčiuoti	%	klaidingai sukirčiuoti	%
apmokymas su 200.000 žodžių									
1	44596,4	0,0	0,0	34710,5	*17,60*	127354,5	*64,58*	2027,1	*1,03*
2	1381,4	28293,4	29790,2	7560,7	*3,83*	148331,0	*75,22*	8200,3	*4,16*
3		28293,4	29790,2	7560,7	*3,83*	148835,5	*75,48*	7695,8	*3,90*
4		28293,4	0,0	15522,8	*7,87*	142065,3	*72,04*	6503,9	*3,30*
5		0,0	29790,2	16293,7	*8,26*	142063,8	*72,04*	5734,5	*2,91*
6		39545,2	0,0	4841,6	*2,46*	148717,3	*75,42*	10533,1	*5,34*
7		19627,6	0,0	9123,4	*4,63*	147497,6	*74,80*	7471,0	*3,79*
8		560644,2	0,0	4841,6	*2,46*	148717,3	*75,42*	10533,1	*5,34*
apmokymas su 400.000 žodžių									
1	70005,9	0,0	0,0	24957,1	*12,66*	137005,7	*69,48*	2112,3	*1,07*
2	2441,6	42608,8	44688,2	4837,7	*2,45*	152917,7	*77,55*	6319,7	*3,20*
3		42608,8	44688,2	4837,7	*2,45*	153266,7	*77,72*	5970,7	*3,03*
4		42608,8	0,0	10454,1	*5,30*	148420,3	*75,27*	5200,8	*2,64*
5		0,0	44688,2	11011,6	*5,58*	148363,3	*75,24*	4700,2	*2,38*
6		57165,6	0,0	3179,3	*1,61*	153130,4	*77,65*	7765,4	*3,94*
7		28840,3	0,0	6000,2	*3,04*	152382,5	*77,28*	5692,4	*2,89*
8		843841,3	0,0	3179,3	*1,61*	153130,4	*77,65*	7765,4	*3,94*
apmokymas su 600.000 žodžių									
1	90041,2	0,0	0,0	20267,8	*10,28*	141652,1	*71,83*	2149,9	*1,09*
2	3407,3	53433,9	56009,1	3711,5	*1,88*	154898,2	*78,55*	5460,1	*2,77*
3		53433,9	56009,1	3711,5	*1,88*	155163,0	*78,69*	5195,4	*2,63*
4		53433,9	0,0	8241,2	*4,18*	151241,2	*76,70*	4587,4	*2,33*
5		0,0	56009,1	8627,0	*4,37*	151222,5	*76,69*	4220,3	*2,14*
6		70047,6	0,0	2476,3	*1,26*	155033,5	*78,62*	6560,0	*3,33*
7		35676,5	0,0	4632,2	*2,35*	154504,4	*78,35*	4933,3	*2,50*
8		1059317,0	0,0	2476,3	*1,26*	155033,5	*78,62*	6560,0	*3,33*
apmokymas su 800.000 žodžių									
1	107069,2	0,0	0,0	17380,2	*8,81*	144504,2	*73,28*	2176,6	*1,10*
2	4309,2	62424,2	65404,4	3045,4	*1,54*	156029,4	*79,13*	4986,2	*2,53*
3		62424,2	65404,4	3045,4	*1,54*	156266,6	*79,25*	4749,0	*2,41*
4		62424,2	0,0	6923,6	*3,51*	152888,2	*77,53*	4249,2	*2,15*
5		0,0	65404,4	7183,0	*3,64*	152938,6	*77,56*	3939,4	*2,00*
6		80510,6	0,0	2045,8	*1,04*	156123,0	*79,17*	5892,2	*2,99*
7		41291,8	0,0	3812,2	*1,93*	155734,4	*78,98*	4514,4	*2,29*
8		1238408,2	0,0	2045,8	*1,04*	156123,0	*79,17*	5892,2	*2,99*

Metodas: **1** – tik pilni; **2** – pradžios pabaigos; **3** – pabaigos pradžios; **4** – pradžios; **5** – pabaigos; **6** – vidurio (1-as sumaž.); **7** – vidurio (2-as sumaž.); **8** –vidurio (visi);

Nekirčiuoti (pvz., užsienio k.) žodžiai				*Klitikai*				*Klaidos (iš viso)*	
teisingai nesu-kirčiuoti	%	klaidingai sukirčiuoti	%	teisingai nesu-kirčiuoti	%	klaidingai sukirčiuoti	%	klaidingai	%
apmokymas su 200.000 žodžių									
1908,8	*0,97*	191,0	*0,10*	30656,1	*15,55*	345,7	*0,18*	37274,1	*18,90*
1395,2	*0,71*	704,6	*0,36*					16811,2	*8,53*
1395,2	*0,71*	704,6	*0,36*					16306,7	*8,27*
1658,0	*0,84*	441,8	*0,22*					22814,1	*11,57*
1595,2	*0,81*	504,6	*0,26*					22878,4	*11,60*
1120,2	*0,57*	979,6	*0,50*					16699,9	*8,47*
1352,1	*0,69*	747,7	*0,38*					17687,8	*8,97*
1120,2	*0,57*	979,6	*0,50*					16699,9	*8,47*
apmokymas su 400.000 žodžių									
1642,6	*0,83*	197,0	*0,10*	30916,2	*15,68*	362,5	*0,18*	27628,9	*14,01*
1163,2	*0,59*	676,4	*0,34*					12196,3	*6,18*
1163,2	*0,59*	676,4	*0,34*					11847,3	*6,01*
1402,8	*0,71*	436,8	*0,22*					16454,1	*8,34*
1349,5	*0,68*	490,1	*0,25*					16564,4	*8,40*
903,5	*0,46*	936,1	*0,47*					12243,3	*6,21*
1109,1	*0,56*	730,5	*0,37*					12785,6	*6,48*
903,5	*0,46*	936,1	*0,47*					12243,3	*6,21*
apmokymas su 600.000 žodžių									
1475,0	*0,75*	202,4	*0,10*	31078,4	*15,76*	367,9	*0,19*	22987,9	*11,66*
1019,2	*0,52*	658,2	*0,33*					10197,6	*5,17*
1019,2	*0,52*	658,2	*0,33*					9932,9	*5,04*
1243,7	*0,63*	433,8	*0,22*					13630,2	*6,91*
1195,3	*0,61*	482,2	*0,24*					13697,3	*6,95*
774,8	*0,39*	902,6	*0,46*					10306,8	*5,23*
965,5	*0,49*	712,0	*0,36*					10645,2	*5,40*
774,8	*0,39*	902,6	*0,46*					10306,8	*5,23*
apmokymas su 800.000 žodžių									
1356,0	*0,69*	202,8	*0,10*	31197,0	*15,82*	376,6	*0,19*	20136,2	*10,21*
910,0	*0,46*	648,8	*0,33*					9057,0	*4,59*
910,0	*0,46*	648,8	*0,33*					8819,8	*4,47*
1127,6	*0,57*	431,2	*0,22*					11980,6	*6,08*
1081,2	*0,55*	477,6	*0,24*					11976,6	*6,07*
681,6	*0,35*	877,2	*0,44*					9191,8	*4,66*
874,2	*0,44*	684,6	*0,35*					9387,8	*4,76*
681,6	*0,35*	877,2	*0,44*					9191,8	*4,66*

Literatūros sąrašas

[Allen ir kt., 1979] Allen, J., S. Hunnicutt, R. Carlson, B. Granstrom (1979). MITalk-79: The 1979 MIT text-to-speech system. *Proceedings of the 97th Meeting of the Acoustical Society of America*, USA, pp. 507-510.

[Allen ir kt., 1987] Allen, J., S. Hunnicutt, D. Klatt (1987). *From Text to Speech: The MITTALK System*. Cambridge University Press, 213 pp.

[Allen, 1973] Allen, W. S. (1973). *Accent and Rhytm: Prosodic Features of Latin and Greek: A Study in Theory and Reconstruction*. Cambridge University Press, Cambridge.

[Allen, 1987] Allen, J. (1987). *Natural Language Understanding*. The Benjamin/Cummings Publishing Company, Amsterdam.

[Ambrazas, 1996] Ambrazas, V. (red.) (1996). *Dabartinės lietuvių kalbos gramatika* (*DLKG*). Mokslo ir enciklopedijų leidykla, Vilnius.

[Anderson ir kt., 1984] Anderson, M. D., J. B. Pierrehumbert, M. Y. Liberman (1984). Synthesis by Rule of English Intonation Patterns. *Proceedings of the International Conference on Acoustics Speech and Signal Processing 1984*, pp. 2.8.1-2.8.4.

[Arciuli, Thompson, 2006] Arciuli, J., J. Thompson (2006). Improving the Assignment of Lexical Stress in Text-to-Speech Systems. *Proceedings of the 11th Australian International Conference on Speech Science & Technology*, pp. 296-300.

[Bąk, 1995] Bąk, P. (1995). *Gramatyka Języka Polskiego. Zarys popularny*. Wiedza Powszechna, Warszawa.

[Baker, 1975] Baker, J. K. (1975). The DRAGON System – An Overview. *IEEE Transactions on Acoustics, Speech and Signal Processing*, **23**(1), pp. 24-29.

[Balčiūnaitė, Pakerys, 2006] Balčiūnaitė, A., A. Pakerys (2006). Klitikų problema Alfonso Maldonio poezijoje. *Žmogus ir žodis: Didaktinė lingvistika*, **8**(1), pp. 10-14.

[Batori, Lenders, 1989] Batori, I. S., W. Lenders (1989). *Computerlinguistik: Ein internationales Handbuch zur computergestützten Sprachforschung und ihrer Anwendung*. Walter de Gruyter, Berlin.

[Bernstein, Nessly, 1981] Bernstein, J., L. Nessly (1981). Performance comparison of component algorithms for the phonemicization of orthography. *Proceedings of 19th Annual Meeting of the Association for Computational Linguistics*, pp. 19-22.

[Bick ir kt., 2004] Bick, E., H. Uibo, K. Müürisep (2004). Arborest – a Growing treebank of Estonian. *Nordic Language Technology*, Museum Tusculanums Forlag Københavns Universitet, København, pp. 125-142.

[Black ir kt., 1998a] Black, A. W., K. Lenzo, V. Pagel (1998). Issues in building general letter to sound rules. *The 3rd ESCA Workshop on Speech Synthesis*, pp. 77-80.

[Black ir kt., 1998b] Black, A. W., P. Taylor, R. Caley (1998). The architecture of the Festival Speech Synthesis System. *The 3rd ESCA Workshop on Speech Synthesis*, Jenolan Caves, Australia, pp. 147-151.

[Bolinger, 1978] Bolinger, D. L. (1978). Intonation across Languages. *Universals of Human Language*, **2**, Stanford University Press, Stanford, pp. 471-524.

[Bourlard, 1995] Bourlard, H. (1995). Towards increasing speech recognition error rates. *Proceeding of EUROSPEECH'95*, pp. 883-894.

[Bourlard, Morgan, 1994] Bourlard, H., N. Morgan (1994). *Connectionist speech recognition*. Kluwer Academic Publishers.

[Breiman ir kt., 1984] Breiman, L., J. H. Freidman, R. A. Olshen, C. J. Stone (1984). *Classification and Regression Trees*. Wadsworth.

[Brown ir kt., 1991] Brown, P. F., S. A. Della Pietra, V. J. Della Pietra, R. L. Mercer (1991). Word-Sense Disambiguation Using Statistical Methods. *Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics*, pp. 264-270.

[Carlson, 1994] Carlson, R. (1994). Models of Speech Synthesis. In D. B. Roe, J. G. Wilpon (eds.), *Voice Communications Between Humans and*

Machines. National Academy of Sciences, National Academy of Sciences, Washington, DC.

[Charniak, 1993] Charniak, E. (1993). *Statistical Language Learning*. A Bradford Book The MIT Press, Cambridge, Massachusetts, London.

[Chomsky, 1957] Chomsky, N. (1957). *Syntactic structures*. Mouton, The Hague.

[Chomsky, Halle, 1968] Chomsky, N., M. Halle (1968). *The Sound Pattern of English (SPE)*. Harper and Row, New York etc.

[Church, 1985] Church, K. (1985). Stress Assignment in Letter to Sound Rules for Speech Synthesis. *Proceedings of the 23rd Annual Meeting on Association for Computational Linguistics*, pp. 246-253.

[Church, 1986] Church, K. (1986). Morphological Decomposition and Stress Assignment for Speech Synthesis. *Proceedings of the 24th Annual Meeting on Association for Computational Linguistics*, pp. 156-164.

[Church, 1988] Church, K. (1988). A Stochastic Parts Program and Noun Phrase Parser for Unrestricted Text. *Proceedings of the Second Conference on Applied Natural Language Processing*, Austin, Texas, pp. 136-143.

[Cohen ir kt., 1982] Cohen, A., R. Collier, J. Hart (1982). Declination: Construct or intrinsic feature of speech pitch? *Phonetica*, **39**, pp. 254-273.

[De Mori ir kt., 1987] De Mori, R., L. Lam, M. Gilloux (1987). Learning and plan refinement in a knowledge-based system for automatic speech recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **9**, issue 2, pp. 289-305.

[Dutoit, 1993] Dutoit, T. (1993). *High Quality Text-To-Speech Synthesis of the French Language*. PhD dissertation, Faculte Polytechnique de Mons, TCTS Lab, Mons, Belgium.

[Dutoit, 1997] Dutoit, T. (1997). *A Short Introduction to Text-to-Speech Synthesis*. Kluwer Academic Publishers, Dortrecht, Netherlands.

[Fant, 1960] Fant, G. (1960). *Acoustic Theory of Speech Production*. Mouton & Co, The Hague, Netherlands.

[Filipovič, Lipeika, 2004] Filipovič, M., A. Lipeika (2004). Development of HMM/neural network-based medium-vocabulary isolated-word lithuanian speech recognition system. *Informatica*, **15**(4), pp. 465–474.

[Flanagan, 1972a] Flanagan, J. (1972). *Speech Analysis Synthesis and Perception*. Springer-Verlag, New York.

[Flanagan, 1972b] Flanagan, J. (1972). Voices Of Men And Machines. *Journal of Acoustical Society of America*, **51**, pp. 1375.

[Francis, Kucera, 1982] Francis, W., H. Kucera (1982). *Frequency Analysis of English Usage*. Houghton Mifflin, New York, NY.

[Fry, 1958] Fry, D. B. (1958). Experiments in the perception of stress. *Language and Speech*, **1**, pp. 126-52, reprinted in D. B. Fry (ed.) *Acoustic Phonetics*, Cambridge University Press.

[Fujisaki, 1997] Fujisaki, H. (1997). Prosody, Models, and Spontaneous Speech. In Y. Sagisaka, N. Campbell, N. Higuchi (eds.), *Computing Prosody*, Springer, New York.

[Fujisaki, Sudo, 1971] Fujisaki, H., H. Sudo (1971). A Generative Model of the Prosody of Connected Speech in Japanese. *Annual Report of Eng. Research Institute*, **30**, pp. 75-80.

[Gale ir kt., 1992] Gale, W., K. Church, D. Yarowsky (1992). A Method for Disambiguating Word Senses in a Large Corpus. *Computers and Humanities*, **26**, pp. 415-439.

[Girdenis, 2003] Girdenis, A. (2003). *Teoriniai lietuvių fonologijos pagrindai*. 2-asis leid., Mokslo ir enciklopedijų leidybos institutas, Vilnius.

[Goba, Vasiljevs, 2007] Goba, K., A. Vasiljevs (2007). Development of Text-To-Speech system for Latvian. J. Nivre, H.-J. Kaalep, K. Muischnek, M. Koit (eds.), *NODALIDA 2007 Conference Proceedings*, pp. 67-72.

[Goldsmith, 1999] Goldsmith, J. (1999). *Dealing with Prosody in a Text-to-Speech System*. http://humanities.uchicago.edu/faculty/goldsmith/IJST/index.html

[Grigonytė, Rimkutė, 2005] Grigonytė, G., E. Rimkutė (2005). Formal Specifications for a Dependency Grammar of the Lithuanian Language.

Tarptautinės konferencijos The Second Baltic Conference on Human Language Technologies pranešimų medžiaga, Tallinn, pp. 237-242.

[Grosz ir kt., 1986] Grosz, B., F. S. Jones, B. L. Webber (1986). *Readings in Natural Language Processing*. Morgan Kaufmann, Los Altos, CA.

[Hain, 2001] Hain, T. (2001). *Hidden Model Sequence Models for Automatic Speech Recognition*. Ph. D. thesis, Cambridge University.

[Hayes, 1980] Hayes, B. (1980). *A Metrical Theory of Stress Rules*. Ph. D. Thesis, MIT.

[Hill, Nessly, 1973] Hill, K., L. Nessly (1973). Review of The Sound Pattern of English. *Linguistics*, 106, pp. 57-101.

[Hirschberg, 1991] Hirschberg, J. (1991). Using text analysis to predict intonational boundaries. *Proceedings EUROSPEECH 91*, Genova, pp. 1275-1278.

[HistoryWorld, 2009]. HistoryWorld. „History of Language“. Žiūrėta 2009.12.30. http://www.historyworld.net/wrldhis/PlainTextHistories.asp?historyid=ab13

[Hyde, 1972] Hyde, S. R. (1972). Automatic Speech Recognition: Literature, Survey, And Discussion. In E. E. David, P. B. Denes (eds.), *Human Communication, A Unified Approach*, McGraw Hill, New York.

[Huang ir kt., 1990] Huang, X., Y. Ariki, M. A. Jack (1990). *Hidden Markov Models for Speech Recognition*. Edinburgh University Press, Edinburgh, U.K.

[Huang ir kt., 2001] Huang, X., A. Acero, H.-W. Hon (2001). *Spoken Language Processing. A Guide to Theory, Algorithm, and System Development*. Prentice-Hall, Upper Saddle River, New Jersey.

[Hunnicutt, 1976] Hunnicutt, S. (1976). Phonological rules for a text-to-speech system. *AJCL Microfiche 57*.

[Itakura, 1975] Itakura, F. (1975). Minimum prediction residual principle applied to speech recognition. *Acoustics, Speech and Signal Processing, IEEE Transactions on*, **23**, issue 1, pp. 67-72.

[Yarowsky, 1994] Yarowsky, D. (1994). Decision Lists For Lexical Ambiguity Resolution: Application to Accent Restoration in Spanish and French. *Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics*, Las Cruces, NM, pp. 88-95.

[Yarowsky, 1996] Yarowsky, D. (1996). Homograph Disambiguation in Text-to-Speech Synthesis. In J. van Santen, R. Sproat, J. Olive, J. Hirschberg (eds.), *Progress in Speech Synthesis*, Springer, New York, pp. 159-175.

[Jelinek, 1976] Jelinek, F. (1976). Continuous Speech Recognition by Statistical Methods. *Proceedings of the IEEE*, **64**(4), pp. 532-556.

[Jelinek, 1997] Jelinek, F. (1997). *Statistical methods for speech recognition*. MIT Press.

[Jurafsky, Martin, 2009] Jurafsky D., J. H. Martin (2009). *Speech and language processing: an introduction to natural language processing, computational linguistics, and speech recognition*. Edition 2, Prentice Hall.

[Kasparaitis, 1999] Kasparaitis, P. (1999). Transcribing of the Lithuanian Text Using Formal Rules. *Informatica*, **10**(4), pp. 367-376.

[Kasparaitis, 2000] Kasparaitis, P. (2000). Automatic Stressing of the Lithuanian Text on the Basis of a Dictionary. *Informatica*, **11**(1), pp. 19-40.

[Kasparaitis, 2001a] Kasparaitis, P. (2001). Automatic Stressing of the Lithuanian Nouns and Adjectives on the Basis of Rules. *Informatica*, **12**(2), pp. 315-336.

[Kasparaitis, 2001b] Kasparaitis, P. (2001). *Lietuvių kalbos kompiuterinė sintezė*. Daktaro disertacija, Vilniaus universitetas, Vilnius.

[Kasparaitis, 2005] Kasparaitis, P. (2005). Diphone Databases for Lithuanian Text-to-Speech Synthesis. *Informatica*, **16**(2), pp. 193-202.

[Kasparaitis, 2008] Kasparaitis, P. (2008). Lithuanian Speech Recognition Using the English Recognizer. *Informatica*, **19**(4), pp. 505-516.

[Kass, 1980] Kass, G. V. (1980). An exploratory technique for investigating large quantities of categorical data. *Applied Statistics*, **29**, pp. 119-127.

[Kazlauskienė ir kt., 2004] Kazlauskienė, A., G. Norkevičius, G. Raškinis (2004). Automatizuotas lietuvių kalbos veiksmažodžių kirčiavimas: problemos ir jų sprendimas. *Baltų ir kitų kalbų fonetikos ir akcentologijos problemos*, pp. 166-173.

[Kazlauskienė, Raškinis, 2004] Kazlauskienė, A., G. Raškinis (2004). Veiksmažodžių automatinio kirčiavimo galimybės. *Kalbos teorija ir praktika*, pp. 80-82.

[Keinys, 1993] Keinys, S. (vyr. red.) (1993). *Dabartinės lietuvių kalbos žodynas (DLKŽ)*. Mokslo ir enciklopedijų leidykla, Vilnius.

[Klatt, 1977] Klatt, D. (1977). Review of the ARPA Speech Understanding Project. *Journal of Acoustical Society of America*, **62**(6), pp. 1324-1366.

[Klatt, 1979] Klatt, D. (1979). Synthesis by Rule of Segmental Durations in English Sentences. In B. Lindblom, S. Öhman (eds.), *Frontiers of Speech Communication Research*, Academic, New York, pp. 287-300.

[Klatt, 1980] Klatt, D. (1980). Software for a cascade/parallel formant synthesizer. *The Journal of the Acoustical Society of America*, **67**, pp. 971-995.

[Kleijn, Paliwal, 1995] Kleijn, W. B., K. K. Paliwal (1995). *Speech Coding and Synthesis*. Elsevier, Amsterdam, Netherlands.

[Kohonen, 1988] Kohonen, T. (1988). The 'Neural' Phonetic Typewriter. *IEEE Computer Society Press*, **21**, issue 3, Los Alamitos, CA, USA, pp. 11-22.

[Kurematsu, 1992] Kurematsu, A. (1992). Future Perspective of Automatic Telephone Interpretation. *IEICE TRANSACTIONS on Communications*, vol. E75-B, no.1, pp. 14-19.

[Labutis, 2002] Labutis, V. (2002). *Lietuvių kalbos sintaksė*. Vilniaus universiteto leidykla, Vilnius.

[Laurinčiukaitė, 2003] Laurinčiukaitė, S. (2003). Atskirai pasakytų lietuvių kalbos žodžių atpažinimas, remiantis Paslėptaisiais Markovo Modeliais. *Proceedings of Information Technologies 2003*, vol. IX, KTU, Kaunas, pp. 21-24.

[Lee, 1989] Lee, K.-F. (1989). *Automatic speech recognition: the development of the SPHINX system*. Springer.

[Leech ir kt., 1994] Leech, G., R. Garside, M. Bryant (1994). The large-scale grammatical tagging of text: Experience with the British National Corpus. In N. Oostdijk, P. de Haan (eds.), *Corpus-based research into Language*, Radopi, Amsterdam, pp. 47-63.

[Levinson, 1994] Levinson, S. E. (1994). Speech recognition technology: A critique. In D. B. Roe, J. G. Wilpon (eds.), *Voice communication between humans and machines*, National Academic Press, pp. 159-164.

[Liberman, Prince, 1977] Liberman, M., A. Prince (1977). On Stress and Linguistic Rhythm, *Linguistic, Inquiry*, **8**(2), pp. 249-336.

[Liberman, Church, 1992] Liberman, M., K. Church (1992). Text Analysis and Word Pronunciation in Text-to-Speech Synthesis. In S. Furuy, M. M. Sondhi (eds.), *Advances in Speech Signal Processing*, Dekker, New York, pp. 791-831.

[Lipeika ir kt., 2002] Lipeika, A., J. Lipeikienė, L. Telksnys (2002). Development of isolated word speech recognition system. *Informatica*, **13**(1), pp. 37–46.

[Manning, 2000] Manning, C. (2000). *Probabilistic Models in Computational Linguistics*. http://nlp.stanford.edu/~manning/talks/ima2000.pdf

[Manning, Schutze, 1999] Manning, C., H. Schutze (1999). *Foundations of Statistical Natural Language Processing*. MIT Press, Cambridge, USA.

[Marinčič ir kt., 2009] Marinčič, D., T. Tušar, M. Gams, T. Šef (2009). Analysis of Automatic Stress Assignment in Slovene. *Informatica*, **20**(1), pp. 35-50.

[Markov, 1913] Markov, A. A. (1913). An Example of Statistical Investigation in the Text of 'Eugene Onyegin', Illustrating Coupling of Tests in Chains. *Proceedings of the Academy of Sciences of St. Petersburg*, Russia, pp. 153-162.

[Mauricaitė, 1985] Mauricaitė, V. (1985). Kai kurių frazės faktorių įtaka žodžių šlijimui. *Kalbotyra*, **36**(1), pp. 38-43.

[Mauricaitė, 1986] Mauricaitė, V. (1986). Kiek žodžių, tiek kirčių? *Lietuvių kalba mokykloje*, **2**, Šviesa, Kaunas, pp. 95-103.

[Mauricaitė, 1987] Mauricaitė, V. (1987). Samplaikinių formų dėmenų akcentinis šlijimas. *Mūsų kalba*, **2**, pp. 3-6.

[Mauricaitė, 1994] Mauricaitė, V. (1994). Akcentinis šlijimas frazės viduje (statistinis tyrimas). *Kalbotyra*, **43**(1), pp. 61-64.

[McPeters, Tharp, 1983] McPeters, D. L., A. L. Tharp (1983). Application of the Liberman-Prince Stress Rules to Computer Synthesized Speech. *Proceedings of the First Conference on Applied Natural Language Processing*, pp. 192-197.

[Merialdo, 1994] Merialdo, B. (1994). Tagging English Text with a Probabilistic Model. *Computational Linguistics*, **20**(2), pp. 155–171.

[Mohanan, 1982] Mohanan, K. P. (1982). *Lexical phonology*. Massachusetts Institute of Technology. http://dspace.mit.edu/bitstream/handle/1721.1/15652/10583970.pdf?sequence=1

[Nepil, Popelinsky, 2001] Nepil, M., L. Popelinsky (2001). Part-of-speech tagging by means of ILP and active learning. *Proceedings of the workshop on instance selection at ECML/PKDD 2001*, Freiburg, pp. 25-31.

[Nivre ir kt., 1996] Nivre, J., L. Gronqvist, M. Gustafsson, T. Lager, S. Sofková (1996). Tagging Spoken Language Using Written Language Statistics. *COLING 1996*, pp. 1078-1081.

[Norkevičius ir kt., 2004] Norkevičius, G., A. Kazlauskienė, G. Raškinis (2004). Bendrinės lietuvių kalbos daiktavardžių ir būdvardžių kirčiavimo struktūrinis modelis, algoritmas ir realizacija. *Kalbų studijos*, **6**, pp. 72-76.

[Norkevičius, Raškinis, 2008] Norkevičius, G., G. Raškinis (2008). Modeling Phone Duration of Lithuanian by Classification and Regression Trees, using Very Large Speech Corpus. *Informatica*, **19**(2), pp. 271-284.

[Oancea, Badulescu, 2002] Oancea, E., A. Badulescu (2002). Stressed Syllable Determination for Romanian Words within Speech Synthesis Applications. *International Journal of Speech Technology*, **5**(3), pp. 237-246.

[Oliver, Grice, 2003] Oliver, D., M. Grice (2003). Phonetics and Phonology of lexical stress in Polish verbs. *Proceedings of the 15th International Congress of Phonetic Science*, Barcelona, Spain. http://www.coli.uni-saarland.de/~dominika/icphs_1002.pdf

[Orphanos, Christodoulalds, 1999] Orphanos, G. S., D. N. Christodoulalds (1999). POS Disambiguation and Unknown Word Guessing with Decision Trees. *Proceedings of EACL '99*, pp. 134-141.

[Pagel ir kt., 1998] Pagel, V., K. Lenzo, A. Black (1998). Letter to sound rules for accented lexicon compression. *ICSLP98*, Sydney, Australia.

[Payne, 1997] Payne, T. E. (1997). *Describing morphosyntax: A guide for field linguists*. Cambridge University Press, Cambridge, New York, 413 p.

[Pakerys, Pupkis, 1976] Pakerys, A., A. Pupkis (1976). *Lietuvių kalbos bendrinė tartis*. Plokštelių tekstai, Vilnius.

[Pfister, Traber, 1994] Pfister, B., C. Traber (1994). Text-to-Speech Synthesis: An Introduction and A Case Study. In E. Keller (ed.), *Fundamentals of Speech Synthesis and Speech Recognition: Basic Concepts, State of the Art and Future Challenges*, John Willey & Sons, Chichester, New York, Brisbane, Toronto, Singapore, pp. 87-108.

[Pierce, 1969] Pierce, J. R. (1969). Whither Speech Recognition? *Journal of the Acoustical Society of America*, 46, pp. 1049-1051.

[Plumpe, Meredith, 1998] Plumpe, M., S. Meredith (1998). Which is More Important in a Concatenative Text-to-Speech System: Pitch, Duration, or Spectral Discontinuity. *Third ESCA/COCOSDA Int. Workshop on Speech Synthesis*, Jenolan Caves, Australia, pp. 231-235.

[Puolakainen, 2001] Puolakainen, T. (2001). *Computer Grammar of Estonian: Morphological Disambiguation*. Dissertationes Mathematicae Universitatis Tartuensis, Tartu.

[Quinlan, 1986] Quinlan, J. R. (1986). Induction of decision trees. *Machine Learning*, **1**, pp. 81-106.

[Quinlan, 1993] Quinlan, J. R. (1993). *C4.5 Programs for Machine Learning*. Morgan Kaufman, San Mateo, CA.

[Quinlan, 1996] Quinlan, J. R. (1996). Improved use of continuous attributes in C4.5. *Journal of Artificial Intelligence Research*, **4**, pp. 77-90.

[Rabiner ir kt., 1989] Rabiner, L. R., J. G. Wilpon, F. K. Soong (1989). High performance connected digit recognition using hidden Markov models. *IEEE Transactions on Acoustics, Speech and Signal Processing*, **37**, issue 8, pp. 1214-1225.

[Rahim, 1994] Rahim, M. G. (1994). *Artificial Neural Networks for Speech Analysis/Synthesis*. Chapman & Hall, London, Glasgow, Weinheim, New York, Tokyo, Melbourne, Madras.

[Raškinis, Raškinienė, 2003] Raškinis, G., D. Raškinienė (2003). Building medium-vocabulary isolated-word Lithuanian HMM speech recognition system. *Informatica*, **14**(1), pp. 75-84.

[Reddy, 1975] Reddy, D. R. (ed.) (1975). *Speech Recognition*. Academic Press, New York.

[Reddy, 1976] Reddy, D. R. (1976). Speech Recognition by Machine: A Review. *IEEE Proceedings*, **64**(4), pp. 502-531.

[Rimkutė ir kt., 2005] Rimkutė, E., G. Jarašiūnaitė, P. Homola (2005). Morfologinių samplaikų atpažinimas ir klasifikavimas. *Lituanistica*, **62**(2), pp. 58-75.

[Rimkutė, 2002] Rimkutė, E. (2002). Homoformos dabartinės lietuvių kalbos tekstyne. *Lituanistika*, **2**(50), pp. 86-101.

[Rimkutė, Grigonytė, 2006a] Rimkutė, E., G. Grigonytė (2006). Automatizuotas lietuvių kalbos morfologinio daugiareikšmiškumo ribojimas. *Kalbų studijos*, **9**, pp. 30-37.

[Rimkutė, Grigonytė, 2006b] Rimkutė, E., G. Grigonytė (2006). Statistiniai, loginiai ir kompiuterių mokymosi metodai lietuvių kalbos morfologiniam

daugiareikšmiškumui riboti. *Konferencijos Informacinės technologijos 2006, pranešimų medžiaga*, Technologija, Kaunas, pp. 104-108.

[Rimkutė, Grybinaitė, 2004] Rimkutė, E., A. Grybinaitė, (2004). Dažniausios lietuvių kalbos morfologinio daugiareikšmiškumo rūšys ir jų automatinis vienareikšminimas. *Kalbų studijos*, **5**, pp. 74-78.

[Roach, 1991] Roach, P. (1991). *English Phonetics and Phonology. A practical course.* Second edition, Cambridge University Press, Cambridge.

[Robinson, 1994] Robinson, A. J. (1994). An application of recurrent nets to phone probability estimation. *IEEE Trans. Neural Networks*, **5**(2), pp. 298-305.

[Rudžionis ir kt., 2007] Rudžionis, A., K. Ratkevičius, R. Maskeliūnas (2007). Adaptation of English speech recognition engines for Lithuanian speech recognition. *Proceedings of 3rd Baltic Conference on Human Language Technologies*, Kaunas.

[Rudžionis, Rudžionis, 1996] Rudžionis, A., V. Rudžionis (1996). Izoliuotų žodžių atpažinimas vidurkinant fonetiškai segmentuotus kalbinių signalų parametrus. *Informacinės technologijos-96*, Technologija, Kaunas, pp. 168-174.

[Russell, 1997] Russell, M. J. (1997). Progress towards speech models that model speech. *Proceedings 1997 IEEE Workshop on Automatic Speech Recognition and Understanding*, pp. 114-115.

[Sakoe ir kt., 1989] Sakoe, H., R. Isotani, K. Yoshida, K.-I. Iso, T. Watanabe (1989). Speaker-independent word recognition using dynamic programming neural networks. *International Conference on Acoustics, Speech, and Signal Processing 1989, ICASSP-89*, **1**, Glasgow, UK, pp. 29-32.

[Schroeder, 2004] Schroeder, M. R. (2004). *Computer speech: recognition, compression, synthesis*. Springer.

[Sejnowski, Rosenberg, 1987] Sejnowski, T., C. R. Rosenberg (1987). Parallel Networks that Learn to Pronounce English Text. *Complex Systems*, **1**, pp. 145-168.

[Shpilewski ir kt., 2004] Shpilewski, E., B. Piurkowska, J. Rafalko, B. M. Lobanov, V. Kiselov, L. I. Tsirulnik (2004). Polish TTS in Multi-Voice Slavonic Languages Speech Synthesis System. *SPECOM-2004*, pp. 565-570.

[Silverman, 1987] Silverman, K. (1987). *The Structure and Processing of Fundamental Frequency Contours*. Ph.D. Thesis, University of Cambridge, Cambridge, UK.

[Syrdal, 1995] Syrdal, A. K. (1995). Text-to-Speech Systems. In A. Syrdal, R. Bennett, S. Greenspan (eds.), *Applied Speech Technology*, CRC Press, Boca Raton, Ann Arbor, London, Tokyo, pp. 99-126.

[Skripkauskas, Telksnys, 2006] Skripkauskas, M., L. Telksnys (2006). Automatic Transcription of Lithuanian Text Using Dictionary. *Informatica*, **17**(4), pp.1-14.

[Sližienė, 1994] Sližienė, N. (1994). *Lietuvių kalbos veiksmažodžių junglumo žodynas*. Mokslo ir enciklopedijų leidykla, Vilnius.

[Sluijter, Terken, 1993] Sluijter, A. M. C., J. M. B. Terken (1993). Beyond Sentence Prosody: Paragraph Intonation in Dutch. *Phonetica*, **50**, pp. 180-188.

[Sproat, 1997] Sproat, R. (1997). *Multilingual Text-to-Speech Synthesis*. Kluwer Academic Publishers, Norwell, MA.

[Sproat, Olive, 1995] Sproat, R., J. Olive (1995). An Approach to Text-to-Speech Synthesis. In W. B. Kleijn, K. K. Paliwal (eds.), *Speech Coding and Synthesis*, Elsevier Science, Amsterdam, pp. 611-634.

[Styger, Keller, 1994] Styger, T., E. Keller (1994). Formant Synthesis. In E. Keller (ed.), *Fundamentals of Speech Synthesis and Speech Recognition: Basic Concepts, State of the Art and Future Challenges*, John Willey & Sons, Chichester, New York, Brisbane, Toronto, Singapore, pp. 109-128.

[Stundžia, 1991] Stundžia, B. (1991). Kirtis tekste. *Žodžiai ir prasmės*, **1**, Mokslas, Vilnius, pp. 86-92.

[Šef ir kt., 1998] Šef, T., A. Dobnikar, M. Gams (1998). Text-to-Speech Synthesis in Slovenian Language. *EUSIPCO 98*, **2**, Rhodes, Greece, pp. 1157-1160.

[Šef, 2005] Šef, T. (2005). A Two Level Lexical Stress Assignment Model for Highly Inflected Slovenian Language. *Proceedings of the Third International Conference on Information Technology and Applications*, pp. 347-351.

[Šilingas, Telksnys, 2004] Šilingas D., L. Telksnys (2004). Specifics of Hidden Markov Model Modifications for Large Vocabulary Continuous Speech Recognition. *Informatica*, **15**(1), pp. 93-110.

[Šveikauskienė, 2009] Šveikauskienė, D. (2009). *Lietuvių kalbos vientisinių sakinių automatinė sintaksinė analizė*. Daktaro disertacija, VGTU leidykla Technika, Vilnius.

[Taylor, 2009] Taylor, P. (2009). *Text-to-Speech Synthesis*. Cambridge University Press.

[Tatham, Morton, 2005] Tatham, M., K. Morton (2005). *Developments in Speech Synthesis*. John Wiley Sons.

[Tebelskis, 1995] Tebelskis, J. (1995). *Speech Recognition using Neural Networks*. Ph.D. thesis, School of Computer Science, Carnegie Mellon University, Pittsburgh, Pennsylvania.

[Turing, 1950] Turing, A. M. (1950). Computing Machinery and Intelligence. *Mind*, **LIX**(236), pp. 433-460.

[Vaitkevičiūtė, 1997] Vaitkevičiūtė, V. (1997). *Bendrinės lietuvių kalbos kirčiavimas*. Šviesa, Kaunas.

[Valeckienė, 1998] Valeckienė, A. (1998). *Funkcinė lietuvių kalbos gramatika*. Mokslo ir enciklopedijų leidyklos institutas, Vilnius.

[van Santen, 1997a] van Santen, J. (1997). Segmental Duration and Speech Timing. In Y. Sagisaka, N. Campbell, N. Higuchi (eds.), *Computing Prosody*, Springer, New York, USA, pp. 225-250.

[van Santen, 1997b] van Santen, J. (ed.) (1997). *Progress in Speech Synthesis*. Springer-Verlag, New York.

[Waibel, Lee, 1990] Waibel, A. H., K. F. Lee (1990). *Readings in Speech Recognition*. Morgan Kaufman Publishers, San Mateo, CA.

[Webster, 2004] Webster, G. (2004). Improving letter-to-pronunciation accuracy with automatic morphologically-based stress prediction. *ICSLP*, pp. 2573-2576.

[Werner, Keller, 1994] Werner, S., E. Keller (1994). Prosodic Aspects of Speech. In E. Keller (ed.), *Fundamentals of Speech Synthesis and Speech Recognition: Basic Concepts, State of the Art and Future Challenges*, John Willey & Sons, Chichester, New York, Brisbane, Toronto, Singapore, pp. 23-40.

[Zinkevičius, 2000] Zinkevičius, V. (2000). Lemuoklis – morfologinei analizei. *Darbai ir dienos*, **24**, pp. 245-274.

[Zue, 1985] Zue, V. (1985). The Use of Speech Knowledge in Speech Recognition. *Invited paper, Special issue on Man-Machine Communication, Proceedings IEEE*, **73**(11), pp. 1602-1615.

[Zwicky, 1977] Zwicky, A. (1977). *On Clitics*. Bloomington, Indiana University Linguistics Club. http://www.stanford.edu/~zwicky/on_clitics.pdf

[Zwicky, 1985] Zwicky, A. (1985). Clitics and particles. *Language*, **61**(2), pp. 283-305. http://www.stanford.edu/~zwicky/cliticsparticles.pdf

[Zwicky, 1995] Zwicky, A. (1995). What is a clitic? In J. Nevis, B. Joseph, D. Wanner, A. Zwicky (eds.), *Clitics Bibliography*. http://www.stanford.edu/~zwicky/what-is-a-clitic.pdf

Sąvokos

fleksinė kalba – kalba su išplėtota gramatinių formų sistema.

fonema – pagrindinis garsinis kalbos vienetas.

frazė – intonaciškai ir reikšmiškai išbaigtas žodžių junginys.

homografai – vienodai rašomi, bet skirtingai tariami žodžiai.

intonacija – pagrindinio tono kitimas, leidžiantis išskirti atskirus žodžius ar frazes.

kirčiavimas – ryškiau ir blankiau tariamų skiemenų kontrastas.

kirtis – vieno skiemens išskyrimas iš kitų.

klitikai – teksto bekirčiai žodžiai.

linksniavimas – žodžio kaitymas skaičiais ir linksniais.

morfema – mažiausia reikšminė žodžio dalis.

morfologija – kalbotyros sritis, tirianti žodžių struktūrą, jų formas ir tomis formomis žymimas gramatines reikšmes.

nefleksinė kalba – nekaitoma arba silpnai kaitoma kalba.

pagrindinis tonas – vyraujantis garso šaltinio sukuriamas dažnis.

priegaidė – skiemens požymis, diferencijuojantis kitais atžvilgiais vienodus skiemenis.

sakinys – iš žodžių sudarytas ir gramatikos dėsnių sąlygotas kalbos vienetas.

semantika – kalbotyros sritis, nagrinėjanti žodžių reikšmes.

sintaksė – žodžių (ir jų formų) jungimo į žodžių junginius ir sakinius, o taip pat sakinių jungimo į sudėtinius sakinius, būdai.

sintezatorius – įrenginys, generuojantis žmogaus balsą imituojantį garsą.

skiemuo – vienu kartu ištariama garsinė žodžio dalis, kurios pagrindą sudaro balsis, dvibalsis ar mišrusis dvigarsis.

transkribavimas – raidžių sekos keitimas fonetinių vienetų seka.

Santrumpos

ANN – dirbtinis neuroninis tinklas.

ASR sistema – automatinio kalbos atpažinimo sistema.

CART – klasifikavimo ir regresijos medis.

DSP modulis – kalbos signalo sintezės modulis.

HMM – paslėptasis Markovo modelis.

LTS modulis – transkribavimo modulis.

MA – morfologinė analizė.

MD – morfologinis daugiareikšmiškumas.

NP – daiktavardinė frazė.

OS – operacinė sistema.

POS – kalbos dalis.

SLP sistema – šnekamosios kalbos apdorojimo sistema.

SLU sistema – šnekamosios kalbos suvokimo sistema.

TA modulis – tekstinės analizės modulis.

TN – teksto normalizacija.

TTS sistema – kalbos sintezės sistema.

VP – veiksmažodinė frazė.