VILNIAUS UNIVERSITETAS

Karolina Piaseckienė

STATISTINIAI METODAI LIETUVIŲ KALBOS SUDĖTINGUMO ANALIZĖJE

Daktaro disertacija

Fiziniai mokslai, matematika (01 P)

Vilnius, 2014

Disertacija parengta 2008–2013 metais Vilniaus universiteto Matematikos ir informatikos institute.

Mokslinis vadovas

prof. dr. Marijus Radavičius (Vilniaus universitetas, fiziniai mokslai, matematika – 01 P).

VILNIUS UNIVERSITY

Karolina Piaseckienė

THE STATISTICAL METHODS IN THE ANALYSIS OF THE LITHUANIAN LANGUAGE COMPLEXITY

Doctoral Dissertation

Physical Sciences, Mathematics (01 P)

Vilnius, 2014

Doctoral dissertation was prepared at the Institute of Mathematics and Informatics of Vilnius University in 2008–2013.

Scientific Supervisor

Prof. Dr. Marijus Radavičius (Vilnius University, Physical sciences, Mathematics – 01 P).

Žymėjimai

Simboliai

i, j, k, l, n, N, s, t – natūralieji skaičiai

$\mathbf{N}$ – natūraliųjų skaičių aibė

$\mathbf{R}$ – realiųjų skaičių aibė

$\mathbf{R}^n$ – n-matė realiųjų skaičių erdvė, $\mathbf{R}^n = \underbrace{\mathbf{R} \times \ldots \times \mathbf{R}}_{n}$

$\mathbf{Z}$ – sveikųjų skaičių aibė

$\mathbf{Z}_+$ – sveikųjų neneigiamų skaičių aibė, $\mathbf{Z}_+ = \{0, 1, \ldots\}$

$\mathbf{Z}_+^n$ – n-matė sveikųjų neneigiamų skaičių erdvė

$\mathbf{E}X$ – a. d. X vidurkis

$N(\mu, \sigma^2)$ – normalusis skirstinys su parametrais μ ir σ^2

$X \sim N(\mu, \sigma^2)$ – a. d. X, pasiskirstęs pagal normalųjį dėsnį su parametrais μ ir σ^2

$\mathbf{P}(A)$ – įvykio A tikimybė

$\mathbb{I}(E)$ – įvykio (aibės, ryšio) E indikatorius

Santrumpos

a. d. – atsitiktinis dydis

AL modelis – apibendrintasis logit modelis

AT modelis – apibendrintasis tiesinis modelis

edf – empirinė pasiskirstymo funkciją (angl. *empirical distribution function*)

LNRE – didelis kiekis retų įvykių (angl. *large number of rare events*)

MAL – Markovo atsitiktiniai laukai (angl. *Markov random fields*)

MK – metakalbiniai komentarai

Padėka

Nuoširdžiai dėkoju darbo vadovui prof. dr. Marijui Radavičiui už nuoširdų vadovavimą disertaciniam darbui, skirtą laiką ir energiją, vertingas konsultacijas ir nuolatinį skatinimą tobulėti.

Dėkoju recenzentams prof. dr. Kęstučiui Dučinskui ir doc. dr. Rimantui Eidukevičiui už pastabas, padėjusias patobulinti disertaciją.

Taip pat dėkoju skaitmeninės bibliotekos „Literatūros kūriniai 5–8 klasėms“, sukurtos įgyvendinant projektą „Pagrindinio ugdymo pirmojo koncentro (5–8 kl.) mokinių esminių kompetencijų ugdymas“, kūrėjams, kad leido naudotis projekto rezultatais.

Turinys

Įvadas

Tyrimų sritis ir darbo aktualumas

XX amžiaus antrojoje pusėje ypač spartus mokslo ir kompiuterių technikos vystymasis bei kompiuterių technikos virtimas teksto apdorojimo priemone, taip pat moderniosios kalbotyros ryšiai su semiotika bei kitomis „kibernetikos šeimos“ disciplinomis sąlygojo tiksliųjų mokslų skverbimąsi į kalbotyrą.

Matematinė lingvistika, anot G. K. Pullum ir A. Kornai, yra matematinių struktūrų ir metodų, kurie yra svarbūs lingvistikai, analizė. [67]

Visame pasaulyje, taip pat ir Lietuvoje, pastaruoju metu sparčiai vystosi kalbos kompiuterizavimo procesai: kuriamos programos tekstui koreguoti, klaidoms tikrinti, žodžiams bei garsams atpažinti, elektroniniams žodynams sudaryti ir t.t. Kalbinės technologijos panaudojamos ir pačiai kalbai tyrinėti kiekybiniu bei struktūriniu aspektais.

Deja, Lietuvoje, kaip ir daugelyje Europos šalių, kalbos technologijų erdvė yra plėtojama netolygiai. Sukurta gana kokybiška programinė įranga bazinei teksto analizei, pavyzdžiui, įrankiai morfologinei ir sintaksinei analizei, tačiau pažangesnių technologijų, kurioms reikia nuodugnesnio lingvistinio apdorojimo ir semantinių žinių, kol kas tėra tik užuomazgos. Yra parengta nemažai pirminių skaitmeninių kalbos išteklių (elektroninių žodynų, tekstynų, terminynų) ir pagrindinių kalbos analizės priemonių (morfologinių požymių nustatymo ir generavimo, rašybos tikrinimo įrankių), taip pat sukurtas lietuviškas sintezatorius, automatinio vertimo sistemos ir t.t., tačiau menkai išplėtoti semantikos tyrimai lėmė mažesnę kalbos generavimo, teksto interpretavimo ir teksto analizės pažangą. [91]

Užsienio šalyse naudojamus metodus pritaikyti lietuvių kalbai ne visada pavyksta dėl mūsų kalbos specifiškumo. Lietuvių kalba yra sudėtinga fleksinė kalba, pasižyminti gramatinių formų įvairove, morfologiniu daugiareikšmiškumu, dideliu kaitomumu, laisva žodžių tvarka sakinyje ir pan., todėl tiesiogiai negalima pasinaudoti kitose šalyse jau sukurta programine įranga, pavyzdžiui, automatinei sintaksinei analizei, ir tai gerokai apsunkina efektyvių algoritmų kūrimą automatiniam lietuviškų tekstų apdorojimui. Paprastai klasifikuojant tekstus remiamasi raktiniais žodžiais, tačiau lietuvių kalboje dėl linksniavimo, asmenavimo ir kitos kaitos gali keistis tiek žodžio galūnė, tiek ir kamienas. Tai labai apsunkina raktinių žodžių parinkimo uždavinį.

Šiuo metu, plėtojantis struktūrinei lingvistikai, tampa itin svarbūs ir kalbos modeliavimo klausimai. Lingvistikoje yra skiriamos dvi modelių rūšys: nestatistiniai (arba baziniai) ir statistiniai (arba stochastiniai). Tai susiję su dvipusiu kalbos traktavimu jos funkcionavimo metu. Pirmiausia, kalbą galima tyrinėti jos žodžių junginių identifikavimo požiūriu. Iš kitos pusės, kalbą galima traktuoti kaip tikimybinį procesą, susijusį su kalbos elementų panaudojimo dažnumu kalbos aktuose. Sudarant šiuos modelius, taikomi įvairūs matematiniai metodai. [55]

Nestatistiniai modeliai sudaromi remiantis matematine logika, grafų teorija. Šie modeliai turi ypač didelę reikšmę sintaksinio nagrinėjimo metodikai, kadangi jie leidžia sudaryti neribotą skaičių realiai egzistuojančių sakinių. Kaip teigiama D. Šveikauskienės darbuose (žr. [81], [82]), dėl sudėtingų, palyginti su kitomis kalbomis, pvz., anglų ar vokiečių, sintaksinių santykių lietuvių kalboje (nėra griežtos žodžių tvarkos sakiniuose ir žodis gali priklausyti ne nuo vieno žodžio) sakinio sintaksinės struktūros negalima pavaizduoti medžiu. Visą sintaksinę informaciją gali atskleisti tik grafas, turintis ciklų.

Statistiniams modeliams sukurti taikomi matematinės statistikos, informacijos teorijos ir tikimybių teorijos metodai (Zipfo dėsnis, paslėptieji Markovo modeliai ir t.t.). Nors N. Chomskio (N. Chomsky) pasekėjai ir atmeta tikimybinius kalbos vartosenos modelius, jie yra labai vertingi, ypač kompiuterinei lingvistikai. [52]

Statistiniai metodai yra dažnai naudojami kiekybinėje lingvistinėje analizėje (žr. [1], [10], [79]). Tekstynų lingvistikoje pagrindinė prielaida yra tekstynų atsitiktinumas, kuris suprantamas kaip vienodas tikėtinumas (žr. [9], [27], [13]). Tačiau dideliame tekstyne duomenys yra labai heterogeniški, nes tekstynas yra skirtingų rūšių, įvairiau-

sių žanrų ir tipų tekstų mišinys, kuris yra skirtas skirtingiems tikslams, skirtingiems skaitytojams, jau nekalbant apie skirtingų autorių skirtingas žodžių vartojimų preferencijas („mental lexicon“ [6]), taip pat leistinas ir mėgstamas natūralios kalbos struktūras. Norint atskirti pačiai kalbai būdingas ypatybes nuo autorių preferencijų labai svarbus yra tikslus tiriamos (baigtinės) populiacijos apibrėžimas; jis didele dalimi nulemia ir statistinės analizės rezultatus. Tai leidžia išrinkti reprezentatyvią imtį su atitinkamomis statistinėmis savybėmis ir identifikuoti skirtingus (idealiu atveju – nepriklausomus) statistinio kintamumo šaltinius. Tačiau tekstynų lingvistikoje tai yra problematiška. Atrodo, kad (tiesiogiai nesuformuluotame) populiacijos apibrėžime, kuris paprastai naudojamas lingvistiniuose tyrimuose, laikoma, kad bazinis tekstyno elementas, tenkinantis atsitiktinumo hipotezę, yra „einamasis“ žodis (angl. *running word*). Yra ir kitoks požiūris, pavyzdžiui, pateikiamas M. Baroni ir S. Evert [13], kur nagrinėjamos atsitiktinių dokumentų statistinės imtys (tiriamas elementas yra dokumentas, kūrinys). Tačiau ir šiuo atveju imtyse bei populiacijose autorių, kurių teksto dokumentai yra dideli, preferencijos ir tų tekstų lingvistinės savybės yra atstovaujamos labiau negu tų, kurių teksto dokumentai yra mažesni.

Struktūrinis skirstinys yra vienas iš pagrindinių statistinės lingvistikos tyrimo objektų, glaudžiai susijęs su tikimybiniu požiūriu (tikimybinių modelių taikymu) lingvistikoje, atskiru atveju su Zipfo-Mandelbroto dėsniu (žr. [97], [54]), Julo-Saimono (Yule-Simon) dėsniu (žr. [95], [78]) ir t.t. (žr. [43], [10], [44], [26] ir ten esančias nuorodas). Pagrindinė struktūrinio skirstinio ir Zipfo-Mandelbroto dėsnio tyrimų dalis nagrinėja anglų kalbos tekstynus. Išimčių pavyzdžiai [31], [88]. A. Utka [88] pateikia lietuviško tekstyno (102 mln. žodžių) stebėtų dažnių struktūrinį skirstinį. Didelis lietuvių kalbos kaitomumas, laisva žodžių tvarka ir kitos savybės skiria ją nuo kitų kalbų, ypač anglų. D. Šveikauskienės darbuose [80], [81], [82] pateiktas išsamus šios temos aptarimas. Visa tai iškelia uždavinį patikrinti teiginių ir faktų, nustatytų kitoms kalboms, pagrįstumą lietuvių kalbai.

Peržiūrėjus pagrindinius lituanistų žurnalus („Lietuvių kalbotyros klausimai“ ir kt.) ir išstudijavus, kokie statistiniai metodai yra taikomi, galima teigti, kad lituanistiniuose darbuose vyrauja aprašomosios statistikos metodai, tiriamoji populiacija dažniausiai

nėra apibrėžiama, neaiškiai išrenkama imtis tyrimui ir pan. Tokiu atveju iškyla imties reprezentatyvumo ir statistinių išvadų pagrįstumo klausimas, kuris yra aktualus ir patiems lituanistams (žr. [47], [53]).

Tyrimų objektas

Disertacijos tyrimų objektas – kokybinių požymių tarpusavio priklausomybių struktūriniai (grafiniai) modeliai ir įvairios lietuvių kalbos struktūros.

Matematinis tyrimo objektas – Zipfo-Mandelbroto dėsnis ir su juo susijęs struktūrinis skirstinys, stebėtų dažnių lentelės ir skirstinių modeliai.

Lingvistiniai tyrimo objektai – garsai, raidės, žodžiai, metakalbiniai komentarai, sakiniai (morfologijos ir sintaksės aspektais) ir jų tarpusavio ryšiai; funkciniai stiliai.

Tirti realūs duomenys iš „Dabartinės lietuvių kalbos tekstyno“, ŠU bibliotekoje esančios vaikiškos literatūros, laisvai prieinamos skaitmeninės bibliotekos. Tyrimo metu tyrimo objektai – tekstynai – keitėsi.

Darbo tikslas ir uždaviniai

Pagrindinis darbo tikslas – pritaikyti matematinius ir statistinius metodus lietuvių kalbos analizėje, identifikuojant ir atsižvelgiant į lietuvių kalbos ypatumus, jos heterogeniškumą, sudėtingumą ir variabilumą.

Siekiant numatyto tikslo, buvo sprendžiami tokie uždaviniai:

- Atlikti statistinių taikymų lietuvių kalbos tyrimuose apžvalgą.
- Taikant (grafinę) logtiesinę analizę bei kitus statistinius ir grafinius metodus ištirti lietuvių kalbos savybes, struktūras ir jų sudėtingumą; statistiniais metodais ištirti, ar kitoms kalboms pastebėti ypatumai tinka lietuvių kalbai.
- Aprašyti ir įvertinti kalbos heterogeniškumą ir variabilumą (kintamumą), kurį sąlygoja jos autoriaus pasirinkimai, siekiant identifikuoti nuo autoriaus santykinai mažai priklausančias, vadinasi, potencialiai pačiai kalbai būdingas savybes,

sąryšius ir struktūras. Sudaryti atitinkamą metodiką, ją pritaikyti konkretiems lietuvių kalbos tyrimams.

- Pademonstruoti matematinių metodų galimybes sprendžiant konkrečius lietuvių kalbos uždavinius: skirtingų funkcinių stilių identifikavimas remiantis raidžių / garsų proporcijomis, metakalbinių komentarų konstravimo ypatumai, sakinio struktūros sudėtingumo matavimas (charakteristikos ir jų pasiskirstymas).
- Pritaikyti empirinį Bajeso metodą aprašant kalbos variabilumą ir vertinant struktūrinį skirstinį.

Tyrimų metodai

Darbe taikomi šie tyrimų metodai: mokslinės literatūros disertacijos tema analizė; imčių metodai; grafų teorijos elementai; matematinis modeliavimas ir statistinė duomenų analizė, naudojant logtiesinius ir grafinius logtiesinius modelius, taip pat logistinė regresija bei neigiama binominė regresija su pertekliniu nulių skaičiumi; Bajeso metodologija ir empirinis Bajeso metodas; asimptotiniai metodai; lingvistikos mokslo pagrindai.

Naudotos statistinės analizės programos: SPSS, SAS, R.

Mokslinis darbo naujumas ir praktinė vertė

Atlikto darbo rezultatai papildo ir praplečia kitų šioje bei giminiškose srityse atliktų tyrimų rezultatus.

Lingvistiniuose tyrimuose, kurie remiasi tekstynais, pirminis tyrimo elementas yra žodis, žodžių junginys, kartais – sakinys. Šiame darbe pirminis tyrimo elementas yra autorius, jo pasirinkimas yra kalbos heterogeniškumo ir variabilumo šaltinis. Šis požiūris, drauge su atitinkamais imčių bei statistinės analizės metodais, sudaro siūlomos naujos metodologijos, kuri leidžia nustatyti nuo autoriaus santykinai mažai priklausančias, vadinasi, potencialiai pačiai kalbai būdingas savybes, sąryšius ir struktūras, pagrindą. Ši metodologija pritaikyta dviem konkretiems lietuvių kalbos analizės

uždaviniams.

Sprendžiant automatinio kalbos apdorojimo, pvz., automatinio sintaksinio anotavimo (šį uždavinį savo darbuose [81], [82] iškėlė D. Šveikauskienė) arba vertimo, uždavinius labai svarbu įvertinti tiriamų lingvistinių struktūrų sudėtingumą, nes jis gali nulemti ne tik naudojamų metodų pasirinkimą, bet ir bazinio kalbos modelio sudarymo principus bei analizės metodiką. Šiame darbe atlikta pradinė sakinio (grafinės) sintaksinės struktūros sudėtingumo statistinė analizė. Nors paprastos tekstų lietuvių kalba sudėtingumo charakteristikos, matyt, jau skaičiuotos ne kartą, vis dėlto darbai, skirti nuoseklesnei jų sudėtingumo analizei, autorei nežinomi.

Remiantis neigiamos binominės regresijos su pertekliniu nulių kiekiu modeliu ir empiriniu Bajeso metodu buvo sukonstruotas žodžių formų tekste struktūrinio skirstinio įvertinys, kuris panaudoja turimą papildomą informaciją apie teksto autorius ir žodžio formas ir tokiu būdu leidžia atsižvelgti į tekstų nehomogeniškumą bei nestebėtų žodžio formų efektą. Struktūrinis skirstinys yra žymiai subtilesnis kalbos tyrimo įrankis negu metodai, kurie remiasi parametriniais Zipfo-Mandelbroto tipo modeliais.

Šis darbas parodo imčių metodų taikymo svarbą ir galimybes, o kaip tekstyno statistinės analizės bazinį elementą išskiria tekstų autorius.

Ginamieji disertacijos teiginiai

1. Statistiniai metodai plačiai taikomi lietuvių kalbos analizėje, bet pastaruoju metu vyrauja aprašomoji statistika ir informatikų sukurtos procedūros, pritaikytos daugiau anglų kalbai ir orientuotos į konkrečių praktinių uždavinių sprendimą.

2. Apskritai Herdano ir Zipfo dėsniai gana tiksliai aproksimuoja žodžių formų kiekį ir pasiskirstymą lietuvių kalbos tekstuose. Tačiau tuos dėsnius aprašančių parametrų reikšmės tarp autorių ženkliai skiriasi, dar labiau tarp lietuvių ir užsienio autorių.

3. Lietuvių kalboje betarpiškai susiję žodžiai gali būti gerokai nutolę vienas nuo kito, todėl modeliai ir automatinės taisyklės, sudarytos remiantis trigramų statistika, turi gana ribotas galimybes tinkamai modeliuoti ir prognozuoti lietuvių kalbos

struktūras.

4. Tam, kad sudėtingesni statistiniai kalbos tyrimai būtų atlikti ir interpretuoti korektiškai, pastoviai palaikomi tekstynai turėtų suteikti galimybę tyrimo duomenų sudarymui taikyti imčių metodus ir išrinkti tekstus pagal įvairius požymius, taip pat ir pagal autorius. Tekstynai, kurie neleidžia kontroliuoti imties sudarymo taisyklių, turi labai ribotas statistinės analizės galimybes.

5. Tinkamai taikant imčių metodus surinkti duomenys, (grafiniai) logtiesiniai modeliai, taip pat ir empirinis Bajeso metodas leidžia išnaudoti turimą papildomą informaciją ir modeliuoti sudėtingas kalbos struktūras, jų heterogeniškumą bei individualų kintamumą ir tokiu būdu sudaro pagrindą nustatyti ir tyrinėti pačiai kalbai būdingas savybes bei sąryšius.

Darbo rezultatų aprobavimas

Disertacijos rezultatai paskelbti 3-uose recenzuojamuose moksliniuose leidiniuose bei konferencijų pranešimų medžiagoje. Publikacijų sąrašas pateiktas disertacinio darbo pabaigoje.

Tarpiniai disertacijos rezultatai pristatyti šiose mokslinėse konferencijose:

- K. Piaseckienė, M. Radavičius. *Lietuviškų tekstų stilių palyginimas remiantis universalių kiekybinių charakteristikų statistine analize.* LMD LI konferencija, Šiaulių universitetas, 2010 m. birželio 17–18 d.;

- K. Piaseckienė. *Mokslinio ir grožinio stilių palyginimas remiantis raidžių statistine analize.* VI Tarptautinė mokslinė konferencija „Pasaulio vaizdas kalboje“, Šiaulių universitetas, 2010 m. spalio 21–22 d.;

- K. Piaseckienė, M. Radavičius. *Lietuvių kalbos vaizdingumo raiškos priemonių analizė.* LMD LII konferencija, Gen. J. Žemaičio Lietuvos karo akademija, 2011 m. birželio 16–17 d.;

- K. Piaseckienė, M. Radavičius. *Metakalbinių komentarų funkcijų ir vartosenos palyginamoji analizė.* LMD LIII konferencija, Klaipėdos universitetas, 2012 m. birželio 11–12 d.;

- T. Rekašius, K. Piaseckienė, M. Radavičius. *Apie Zipfo dėsnį kai kurioms lietuvių kalbos sakinio struktūroms.* LMD LIV konferencija, Lietuvos edukologijos universitetas, 2013 m. birželio 19–20 d.;

- K. Piaseckienė. *Grožinės literatūros tekstų struktūrinių skirstinių palyginimas.* LMD LIV konferencija, Lietuvos edukologijos universitetas, 2013 m. birželio 19–20 d.;

- K. Piaseckienė. *Structural distributions of words in Lithuanian texts.* 2-oji tarptautinė Skaičių teorijos konferencija, skirta prof. habil. dr. A. Laurinčiko 65-erių metų sukakčiai paminėti, Šiaulių universitetas, 2013 m. rugsėjo 9–12 d.;

- K. Piaseckienė, T. Rekašius. *Statistinių metodų taikymai sakinio struktūros ir jos sudėtingumo analizėje.* LMD LV konferencija, Mykolo Romerio universitetas, 2014 m. birželio 26–27 d.

Taip pat skaityti pranešimai Matematikos ir informatikos instituto Tikimybių teorijos ir statistikos skyriaus seminare bei Šiaulių universiteto Infomatikos, matematikos, e. studijų instituto seminare.

Disertacijos struktūra

Disertaciją sudaro įvadas, 3 skyriai, išvados, naudotos literatūros sąrašas, autorės publikacijų disertacijos tema sąrašas ir 2 priedai.

Bendra disertacijos apimtis yra 124 puslapiai, kuriuose pateikta 45 numeruotos formulės, 17 paveikslų, 21 lentelė. Rašant disertaciją remtasi 99 literatūros šaltiniais.

1

Statistikos metodų taikymai lingvistikoje

1.1 Kiekybinė lingvistika užsienyje

Bene pirmieji artimesnio ryšio tarp kalbotyros ir matematikos reikalingumą aiškiai suprato du žymūs XX amžiaus struktūrinės kalbos analizės pradininkai – F. de Sosiūras (F. de Saussure) ir B. de Kurtenė (B. de Courtenay).

Dar 1894 metais F. de Sosiūras išsakė tokią mintį, kad kalba ir jos sąryšiai gali būti užrašomi matematinėmis formulėmis, o 1911 metais, rašydamas paskutinį bendrosios kalbotyros skyrių, pabrėžė, kad šis mokslas jam pasirodė kaip geometrijos sistema: „Siekiama įrodyti kaip teoremas“. [35]

B. de Kurtenė studijuodamas kalbą bandė panaudoti dalį pagrindinio tuometinės matematikos supratimo. 1909 metais publikuotos kalbotyros apžvalgoje jis išreiškė įsitikinimą, kad kada nors kalbotyra taps artimesnė tiksliesiems mokslams, t.y., viena vertus, matematikoje kada nors bus galima išdėstyti daugiau „kiekybinės minties“, kita vertus, kalbotyroje bus vystomi nauji „dedukcinės minties“ metodai. [35]

Prancūzų matematikas J. Hadamard 1943 metais pripažino struktūrizacijos kalbos moksle tendenciją, skelbdamas, kad kalbotyra yra tiltas tarp matematikos ir humanitarinių mokslų. [35]

Kalbos struktūros ir jos matematinių aspektų simpoziumą inicijavo Amerikos ma-

tematinė visuomenė, kuri suprato, kad kalbininkų, logikų ir matematikų dėmesys yra sutelktas į bendrų interesų problemas. Iki vėlyvų 1960-ųjų metų, kai A. V. Gladkij ir I. A. Mel'čuk išleido knygą „Matematinės lingvistikos elementai“ („Elements of Mathematical Linguistics“, 1969), matematinė lingvistika buvo mažai žinoma, ypač JAV, kur ji buvo dar tik neseniai pradėjusi kurtis.

Vienas iš pirmųjų darbų, kurie šiuolaikiniu požiūriu galėtų būti vadinami matematinės lingvistikos darbais, yra A. A. Markovo (A. A. Markov, 1912) silpnojo didžiųjų skaičių dėsnio apibendrinimas 1-os eilės Markovo grandinėms su baigtiniu būsenų skaičiumi (žr. [45]):

Teorema. *Bet kokiems pasirinktinai mažiems teigiamiems skaičiams ε, δ egzistuoja ilgis N toks, kad jei m_i žymi, kiek kartų N ilgio bandymo metu procesas (tiksliau – 1-os eilės tranzityvioji Markovo grandinė su baigtiniu būsenų skaičiumi) buvo būsenoje a_i, tai turime*

$$P\left(|m_i/N - T_i| > \delta\right) < \varepsilon.$$

Čia T_i žymi stacionariąją būsenos a_i tikimybę.

XX amžiaus antrojoje pusėje ypač spartus mokslo ir kompiuterių technikos vystymasis bei kompiuterių technikos virtimas teksto apdorojimo priemone, taip pat moderniosios kalbotyros ryšiai su semiotika bei kitomis „kibernetikos šeimos“ disciplinomis sąlygojo tiksliųjų mokslų skverbimąsi į kalbotyrą. Kalbos tyrinėjimų, paremtų tiksliųjų mokslų metodais, sfera iki šiol vadinama skirtingais terminais: mašininė lingvistika, statistinė lingvistika, kompiuterinė lingvistika, nors bene populiariausias – matematinė lingvistika.

Pradedant L. Blumfildo (L. Bloomfield) postulatais pagrindinis matematinės lingvistikos sampratos aparatas, ypač hierarchinių struktūrų sudarymas iš palyginti stabilių pasikartojančių elementų, formavosi pirmiausia fonologijos ir morfologijos pagrindu. N. Chomskis suformulavo tris teorinius modelius lingvistinės struktūros apibūdinimui. Vienas paremtas automatais su baigtiniu būsenų skaičiumi (*Finite-State Automata*), kitas – nepriklausančiomis nuo konteksto gramatikomis (*Context-Free Grammars*) ir trečiasis – priklausančiomis nuo konteksto gramatikomis (*Context-Sensitive Grammars*) ir/ar tik efektyvesnėmis „neapribotomis perrašymo sistemomis“ (*Unrestricted Rewriting Systems*). Ryšys tarp jų yra ištirtas ir tapo tolimesnių formalios kalbos teorijos

kompiuterių moksle darbų pagrindu. [67]

Yra keletas logikų ir lingvistų, įskaitant T. Batóg, F. H. H. Kortlantą (F. H. H. Kortlandt), J. Mulder ir A. Wedberg, kurie, taikydami matematinius metodus, nagrinėjo fonemų teoriją nuo 1960–1970-ųjų metų, bet šis darbas turėjo nedaug įtakos lingvistinei praktikai. Galutinis teorinės fonologijos ir morfologijos formalizavimas buvo pasiūlytas N. Chomskio ir M. Halle 1968 m. Pirmas svarbus N. Chomskio techninis indėlis į lingvistiką buvo gretimų sudėtinių dalių analizės formalizavimas pasinaudojant nekontekstinėmis gramatikomis. [67]

Vienas iš žymiausių Vokietijos mokslininkų – M. Krachtas (M. Kracht), parašęs svarbių modalinės logikos, matematikos bei pagrindinių lingvistikos sričių (fonologijos, morfologijos, sintaksės, semantikos) darbų. „Kalbos matematika“ („The Mathematics of Language“) yra viena iš autoritetingiausių knygų, kurioje didelis dėmesys skiriamas svarbioms šiuolaikinių kalbininkų problemoms. Knygoje M. Krachtas aprašo teiginių logiką, kategorinę gramatiką, Montague semantiką, supažindina su intensyvumu, cilindrine algebra, pateikia teoriniu modeliu pagrįstą požiūrį į fonemų, morfemų, medžių, nurodomųjų santykių apibūdinimą, taip pat į naujoviškų žodžių junginių struktūrinę ir transformacinę gramatiką. [48]

Informacijos teorija pagrįstas požiūris į semantiką vis dar yra didžia dalimi apribotas leksinės semantikos, nors daug uždavinių, tokių kaip mašininis vertimas, kurie iš pradžių buvo priskiriami sudėtingai semantinei analizei, dabar dažnai yra atliekami tiesiog naudojantis statistiniais modeliais. [67]

Šiuo metu vis daugiau žmonių dirba su kompiuteriais ar kitais informacijos apdorojimo įrenginiais. Didelė dalis laiko skiriama informacijai surinkti kompiuteriu. Galimybė šią užduotį atlikti balsu labai sumažintų darbo laiką, kurį žmogus praleidžia surinkdamas tekstą klaviatūra. Todėl pasaulyje, spręsdami šią problemą, labai intensyviai darbuojasi šnekos atpažinimo specialistai. Plačiai paplitusioms kalboms (anglų, ispanų ir kt.) jau yra sukurtos komercinės atpažinimo sistemos.

Dar vienas šiuo metu, gausėjant medžiagos internete ir didėjant poreikiui tobulinti paieškos sistemas, ne mažiau svarbus kompiuterinės lingvistikos uždavinys – automatinis tekstų klasifikavimas į žanrus, temas ar kitas kategorijas. Automatinis teksto

rūšies ir pan. atpažinimas paieškos sistemose leidžia vartotojui tiksliau atlikti paiešką ir gauti geresnius rezultatus. Automatinio klasifikavimo uždavinys sprendžiamas naudojant skirtingus klasifikavimo požymius, statistinius ir nestatistinius metodus. [89]

Bene labiausiai ištyrinėtas (žr. [25], [18], [94]) ir įdiegtas daugelyje šiuolaikinių paieškos sistemų yra klasifikavimas pagal tekstų temą. Klasifikavimas pagal tekstų žanrą remiasi ne reikšminiais žodžiais, o stilistiniais kalbos požymiais: struktūriniais, statistiniais, skyrybos arba labai dažnomis žodžių formomis. J. Karlgren ir D. Katingas (D. Cutting) (žr. [38]) klasifikavimui naudoja tik struktūrinius požymius ir taiko diskriminantinę analizę, B. Kesleris (B. Kessler) ir kiti (žr. [42]) be struktūrinių dar atsižvelgia į leksinius, skyrybos bei statistinius požymius ir taiko logistinę regresiją.

Kai kuriuose anglų kalbos tekstų klasifikavimo darbuose (pvz., [83]), kurių uždavinys – nustatyti grožinės literatūros tekstų žanrą arba autorystę, įrodyta, kad labai dažni žodžiai ir žodžių formos yra geri tekstų klasifikavimo požymiai.

Naudodamas faktorių analizę žanrams tirti, D. Baiberis (D. Biber) nustatė ir teoriškai apibrėžė aštuonis anglų kalbos teksto tipus, kurie pagrįsti kalbinių elementų pasiskirstymu tekstuose. D. Baiberio paskelbta faktorių analizės metodologija pradėta dažniau taikyti ir kituose lingvistiniuose tekstų tipologijos tyrinėjimuose (R. Sigley, E. Csomay, C. Geisler ir kt.). [89]

Europos ir JAV lingvistai ne kartą nagrinėjo dažniausių žodžių savybes ir jų svarbą kalbai. Buvo pastebėta, kad dažniausi kalbos žodžiai arba jų formos turi savybių, nebūdingų retesniems žodžiams.

Pirmasis labai dažnų žodžių ir jų formų savybes aprašė amerikiečių lingvistas G.K. Zipfas (G.K. Zipf) (žr. [97]). Jis nustatė, kad egzistuoja matematinė priklausomybė tarp žodžio dažnio ir jo vietos dažniniame sąraše bei tarp žodžio dažnio ir žodžių, turinčių tą dažnį, skaičiaus (vadinamasis Zipfo dėsnis – *Zipf's Law*). Ši matematinė priklausomybė gali būti išreikšta formule

$$f_r = \frac{K}{r^\gamma}, \tag{1.1}$$

kuri yra atskiras Zipfo-Mandelbroto dėsnio atvejis. Tarkime, kad f_r yra r-ąjį rangą

žodžių dažnių lentelėje turintis dažnis, išreiškiamas kaip rango r funkcija:

$$f_r = \frac{K}{(r+B)^\gamma}.$$

Čia rangas $r = 1, 2, \ldots, N$, γ – žodžių gausumo parametras, $\gamma > 0$, B – parametras, rodantis nukrypimą nuo Zipfo dėsnio, $B \geq 0$ (kai $B = 0$, gauname klasikinį Zipfo dėsnį), K – normuojanti konstanta. [8]

Atrodo, kad dėsnis tinka duomenims, kadangi trumpi žodžiai paprastai yra dažnesni negu ilgi žodžiai. Kaip pabrėžia G. K. Zipfas, ekonominiai faktoriai reikalauja trumpesnių žodžių, kadangi ilgus žodžius parašyti ir perskaityti, ištarti ir suprasti užtrunka daugiau laiko. [77]

Teigiama, kad Zipfo dėsnis yra universalus, t.y. tinka bet kurios kalbos dažniniam sąrašui. Vis dėlto G. K. Zipfas negalėjo paaiškinti, kodėl beveik visiems žodžiams taikytina matematinė priklausomybė negalioja 2% pačių dažniausių, trumpesnių negu 3 raidės, anglų kalbos žodžių. Tiek anglų, tiek švedų kalbose vienos ir dviejų raidžių žodžiai yra retesni negu trijų raidžių žodžiai, prieštaraujant Zipfo dėsniui. [88], [77]

G. K. Zipfas nesiūlė jokios funkcijos dažnio gavimui pagal žodžio ilgį. Šiuo klausimu XX a. pabaigoje–XXI a. pradžioje domėjosi keletas tyrinėtojų (pvz., K.-H. Best, G. Wimmer, A. Wilson, T. McEnery ir kiti). [77]

Bengt Sigurd su bendraautoriais, ištyrę, kad anglų ir švedų kalbose paprastas Zipfo dėsnis galioja tik ilgesniems negu 3 raidžių arba fonemų žodžiams, nustatė sudėtingesnę formulę (panašią į gama skirstinį), kurios bendra išraiška $f_{exp} = aL^b c^L$ (čia L yra žodžio ilgis raidėmis) ir kuri aprašo abu faktus: ir tai, kad galimų žodžių skaičius didėja didėjant žodžių ilgiui (iki 3 raidžių ilgio), ir tai, kad ilgesni žodžiai yra linkę būti vartojami rečiau galimai dėl to, kad yra neekonomiški. [77]

Įvairiems kalbininkų, kurie nori remtis realiais duomenimis, tyrimams labai naudingi ištekliai yra tekstynai, kurie, tinkamai sudaryti, reprezentatyviai atspindi rašytinę / šnekamąją ar pan. kalbą. Taikomosios lingvistikos profesoriaus Ch. F. Mejerio (Ch. F. Meyer) knyga „Anglų kalbos tekstynų lingvistika“ („English Corpus Linguistics“, [60]) yra nuoseklus kalbinių tekstynų kūrimo ir analizavimo vadovas. Knygoje rašoma apie tai, kaip planuoti tekstyno kūrimą, kaip rinkti ir kompiuterizuoti duomenų įtraukimą

į tekstyną, kaip anotuoti surinktus duomenis ir kaip atlikti užbaigto tekstyno analizę.

Ne mažiau svarbus yra ir natūraliosios kalbos reiškinių apdorojimas (kompiuteriu) (*natural language processing*), neretai naudojamas darbui su tekstynais, taip pat – statistinis natūraliosios kalbos reiškinių apdorojimas, kuris nuo tradicinio skiriasi tuo, kad kalbininkui nereikia pagal turimus lingvistinius duomenis sudaryti modelio ranka, nes modelis yra sudaromas (pusiau) automatiškai. Naudojami maksimalaus tikėtinumo įvertinimai (*maximum likelihood estimation*), maksimalios entropijos modeliai (*maximum entropy models*). Gramatinio anotavimo modeliams kurti naudojami paslėptieji Markovo modeliai (*Hidden Markov Models*). [17]

Iš kitos pusės, gramatiniai anotatoriai (*taggers*) gali būti kuriami ir ranka. Pavyzdžiui, Ch. Samuelsson ir A. Voutilainen (žr. [75]) teigia, kad rankiniu būdu sudarytas gramatinis anotatorius gali konkuruoti su savo stochastiniais pusbroliais. Tačiau rankiniu būdu sukurti gramatinį anotatorių reikia daug daugiau pastangų nei jį kurti automatiškai, darant prielaidą, kad tekstynas egzistuoja. [17]

G. Vilkoko (G. Wilcock) knygoje [93] siekiama parodyti, kad geros lingvistinės anotacijos yra esminis geros teksto analizės pagrindas.

N. A. Smito (N. A. Smith) knygoje „Lingvistinių struktūrų prognozavimas“ („Linguistic Structure Prediction“, [79]) teigiama, kad į daugelį pagrindinių natūraliosios kalbos reiškinių apdorojimo problemų galima žiūrėti kaip į struktūros prognozavimo problemas, t.y. samprotavimą apie daugelį tarpusavyje susijusių įvykių. Knygoje pateikiamas bendras supratimas apie tikimybes ir statistiką, algoritmų ir duomenų struktūras.

1.2 Kiekybinė lingvistika Lietuvoje

Kaip ir visame pasaulyje, taip ir Lietuvoje pastaruoju metu sparčiai vystosi kalbos kompiuterizavimo procesai: kuriamos programos tekstui koreguoti, klaidoms tikrinti, žodžiams bei garsams atpažinti, elektroniniams žodynams sudaryti ir t.t. Kalbinės technologijos naudojamos ir pačiai kalbai tyrinėti kiekybiniu bei struktūriniu aspektais.

Dabartinę (rašytinę) lietuvių kalbą, įvairius jos stilius, anot kūrėjų, reprezentatyviai atspindi Vytauto Didžiojo universiteto Kompiuterinės lingvistikos centro (toliau KLC) „Dabartinės lietuvių kalbos tekstynas“, kurį sudaro daugiau kaip 140 mln. žodžių ir kuris yra plačiai Lietuvoje naudojama duomenų bazė (plačiau žr. *http://tekstynas.vdu.lt/tekstynas*).

Lietuvių kalbai yra sukurta gerai veikianti automatinė morfologinė analizės programa *Lemuoklis*, žodžio formai pateikianti antraštinį pavidalą (lemą) ir gramatines pažymas. Programa sukurta 2000 metais ir skirta VDU KLC tekstyno moksliniams lingvistiniams tyrinėjimams automatizuoti. Visus programavimo darbus atliko V. Zinkevičius, kuris taip pat sukūrė kompiuterinę žinių apie lietuvių kalbos leksiką ir gramatiką bazę, lingvistinės informacijos paieškos šioje bazėje ir jos panaudojimo automatiško lemavimo procesui metodiką bei programinę įrangą. Taigi *Lemuoklis* automatiškai lemuoja lietuviškas žodžių formas iš pradinių tekstinių failų, įrašydamas lemavimo rezultatus į tekstinius rezultatų failus. [96]

Didėjant informacinių technologijų plėtrai bei spartėjant kalbos kompiuterizavimui, iškilo būtinybė kurti didelius anotuotus tekstynus tam, kad būtų galima pasinaudoti jų duomenimis pereinant į aukštesnius kalbos kompiuterizavimo lygmenis, pvz., automatinę sintaksę ir semantinę analizę, mašininį vertimą.

Lietuvių kalboje ypač aktuali morfologinio daugiareikšmiškumo problema. Atsižvelgiant į sėkmingą pasaulinę patirtį sprendžiant morfologinio daugiareikšmiškumo problemą statistiniais metodais, šiuos metodus pabandyta pritaikyti ir lietuvių kalbai. Anotuojant tekstyną su *Lemuokliu* nepavyko išspręsti morfologinio daugiareikšmiškumo, nes programa gali pateikti tik lemas ir morfologines pažymas, kurios dažnai yra daugiareikšmės. [70]

Morfologinis daugiareikšmiškumas atsiranda dėl įvairių priežasčių: dėl *Lemuok-*

lio specifikos; dėl to, kad analizuojamos pavienės formos, o ne kontekstas; dėl žodynuose pateiktų leksinių ir gramatinių duomenų netikslumo ar trūkumo. Taigi reikia ieškoti būdų morfologiniam daugiareikšmiškumui riboti. Gana išsamiai lietuvių kalbos morfologinis daugiareikšmiškumas ir metodai bei priemonės, kaip jį riboti, aprašyti E. Rimkutės daktaro disertacijoje „Morfologinio daugiareikšmiškumo ribojimas kompiuteriniame tekstyne“ (2006).

Lietuvių kalbos morfologinis daugiareikšmiškumas pradėtas riboti gana neseniai. Pirmieji morfologinio daugiareikšmiškumo ribojimo būdai buvo automatiniai – tai dažniausiai informatikų pritaikyti statistiniai ir loginiai metodai. Nuo 2001 m. pradėti taikyti statistiniai lietuvių kalbos morfologinio daugiareikšmiškumo ribojimo metodai: paslėptieji Markovo modeliai, Viterbi algoritmas. Šiais metodais buvo pasiektas apytiksliai 85% efektyvumas. Nuo 2002 m. lietuvių kalbos morfologiniam daugiareikšmiškumui riboti imti taikyti ne tik statistiniai, bet ir kompiuterių loginio mokymosi metodai, kuriuos pritaikius pasiektas 90,69% morfologinio daugiareikšmiškumo ribojimo tikslumas. [71]

Automatizuotai ribojant morfologinį daugiareikšmiškumą yra kuriamos taisyklės. Vienos iš jų pagrįstos statistika, kitos – morfologine ar sintaksine analize. Kai kuriuos daugiareikšmius atvejus galima panaikinti gana greitai ir lengvai, o kitos morfologiškai daugiareikšmės formos lieka net ir pritaikius kelis metodus. Paprasčiausias statistiniais duomenimis pagrįstas automatizuotas morfologinio daugiareikšmiškumo ribojimo būdas yra nerealiųjų homoformų panaikinimas. Analizuojant nekaitomas kalbos dalis taikoma įvairialypė analizė, o kaitomų kalbos dalių morfologinio daugiareikšmiškumo ribojimas dažniausiai pagrįstas sintaksine analize. [71]

2005–2006 m. trukusio Lietuvos valstybinio mokslo ir studijų fondo finansuojamo projekto metu buvo nagrinėti statistiniai lietuvių kalbos modeliai ir jų taikymas automatinio morfologinio vienareikšminimo problemai spręsti.

Kuriant automatinio lietuvių kalbos morfologinio anotavimo priemonę buvo pasitelkta čekų patirtis morfologinio anotavimo srityje. Čekų darbuose atskleidžiamas paslėptųjų Markovo modelių ir formaliųjų (taisyklių) metodų taikymas čekų ir anglų kalboms. Buvo atlikti įvairūs eksperimentai, kuriuose derinami, redukuojami įvairūs čekų kalbos morfologiniai požymiai, ir buvo pasiektas anglų kalbai artimas lietuvių kal-

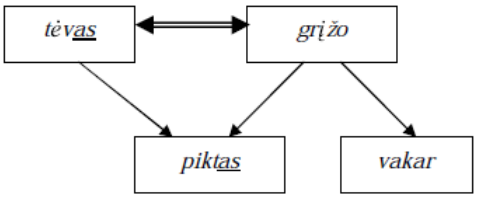
bos morfologinio anotavimo tikslumas – 96%. Taip pat buvo pasiektas 99% tikslumas nustatant antraštines lietuvių kalbos žodžių formas (lemas). Tikslumas skaičiuojamas įvertinant vienareikšminimo klaidas, neatpažinti žodžiai nėra įskaičiuojami. [70]

2007–2008 m. Valstybinio mokslo ir studijų fondo finansuotame projekte „Interneteniai ištekliai: anotuotas lietuvių kalbos tekstynas ir anotavimo priemonės (ALKA2)“ vienas iš numatytų ir įgyvendintų darbų – parengtas morfologiškai anotuotas tekstynas, internete pateikta visiems prieinama morfologinio anotavimo programa, padidintas morfologinio anotatoriaus atpažįstamų žodžių kiekis ir papildyta leksinė duomenų bazė. [74]

Taip pat tiesiogiai negalima pasinaudoti ir kitose šalyse jau sukurta automatinės sintaksinės analizės programine įranga, nes lietuvių kalbai būdingas didelis kaitomumas ir laisva žodžių tvarka sakinyje. Ši problema sprendžiama lietuvių kalbos automatinėje sintaksinėje analizėje į vieną visumą sujungiant visas tris gramatikos sritis – morfologiją, sintaksę ir semantiką. [81]

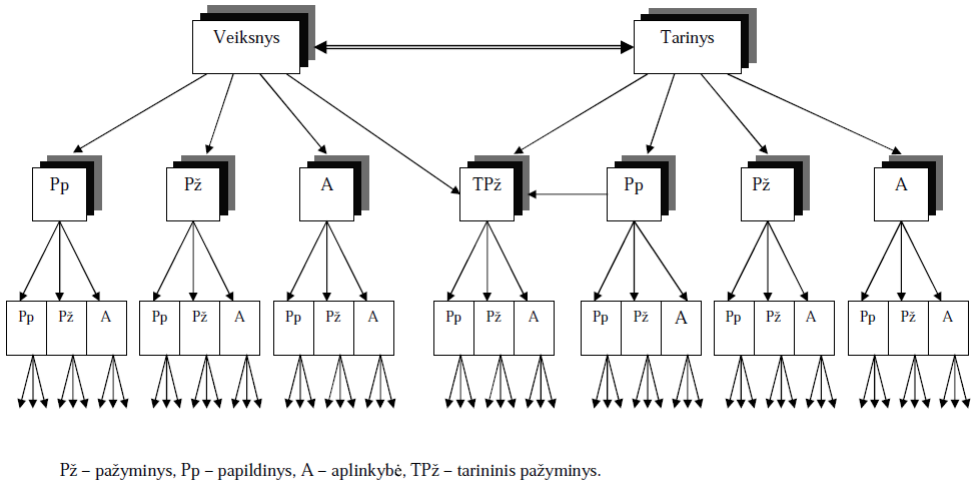
Kompiuterizuojant kalbas, skiriami du pagrindiniai etapai: automatinė morfologinė analizė ir automatinė sintaksinė analizė. Po šių etapų seka semantinė analizė. [30]

Kitų kalbų sintaksinei struktūrai naudojamas priklausomybių medis negali atspindėti visų sintaksinių ryšių, esančių lietuviškame sakinyje, todėl, norint neprarasti dalies sintaksinės informacijos, lietuvių kalbos sintaksinės struktūros gali būti vaizduojamos tik priklausomybių grafu (plačiau žr. [82]). Pvz., sakinio *Tėvas vakar grįžo piktas* (žr. 1.1 pav.) sintaksinė struktūra, pavaizduota medžiu, neatskleistų visos informacijos, t.y. tarininio pažyminio dvigubų sintaksinių ryšių.



Pav. 1.1: Sakinio *Tėvas vakar grįžo piktas* sintaksinė struktūra (paimta iš [82])

D. Šveikauskienės daktaro disertacijoje [81] pateikiama lietuvių kalbos vientisinių sakinių apibendrinta sintaksinė struktūra, išreikšta grafu (žr. 1.2 pav.). V. Karaciejūtė



Pav. 1.2: Apibendrinta lietuvių kalbos vientisinio sakinio struktūrinė schema (paimta iš [81])

[37] sakinių sintaksinę struktūrą naudojo mėgindama nustatyti, kokiam funkciniam stiliui priklauso tekstas.

XX amžiaus pabaigoje programinės įrangos pramonėje bei akademiniuose sluoksniuose susidomėta statistinių metodų taikymu mašininiame vertime. [24]

Kalbų apdorojimo priemonių plėtrą bei įvairovę skatina ir mašininio vertimo sistemos, nes siekiant gerinti mašininio vertimo kokybę reikia tobulinti esamas automatinės kalbos analizės priemones ir kurti naujas, pvz., automatinės morfologinės, sintaksinės, semantinės analizės programas, terminų atpažinimo priemones, vienareikšminimo įrankius ir pan. Kiekvienoje kalboje realiai veikianti mašininio vertimo sistema yra svarbus įvykis, turintis įtakos pačiai kalbai, tos kalbos vartotojams ir tyrėjams. [72]

Iki šiol nėra nė vienos realiai veikiančios „tobulos“ mašininio vertimo programos, kuri verstų į lietuvių kalbą ir iš jos. Anot vokiečių lingvisto V. Toiberto (W. Teubert, 1997), niekada nebus sukurta tokia mašininio vertimo sistema, kuri galės pateikti teisingus ir galutinius vertimus 'atviriems tekstams', priklausantiems tam tikram kultūriniam ar socialiniam diskursui. [87]

Vienas iš Lietuvoje siūlomų produktų, palengvinančių vertimą, yra nuo 2001 m. leidžiamas „Tildės biuras“ (žr. *http://www.tilde.lt*). 2006 m. išleistame „Tildės biure 2006“ įdiegta daugiapakopė paieškos sistema ir automatizuotas vertimo įrankis –

„Vertimo vedlys“. Tai pirmoji Lietuvoje sukurta automatizuota vertimo priemonė, padedanti skaityti ar versti elektroninius tekstus iš anglų kalbos į lietuvių kalbą. Ši automatizuota vertimo priemonė analizuoja sakinių struktūrą ir automatiškai siūlo sakinio, jo dalies ar atskirų žodžių vertimą. 2009 m. buvo išleista nauja programinės įrangos versija „Tildės biuras 2009“, kurioje žodynų rinkinys papildytas naujais kalbų ir terminų žodynais, todėl atsirado galimybė skaityti tekstus ne tik anglų, vokiečių, rusų, bet ir prancūzų kalbomis. Nauja versija „Tildės biuras 2012“ leidžia vartotojams naudotis ne tik gausiausiais Lietuvoje įvairių kalbų kompiuteriniais žodynais, atnaujintu rašybos tikrintuvu, bet ir mašininiu lietuvių-anglų-lietuvių kalbų vertimu.

Šiuo metu Lietuvoje geriausiai žinoma iš anglų ir rusų kalbų į lietuvių kalbą ir atvirkščiai verčianti ir laisvai prieinama kompiuterinio vertimo sistema – *Google Vertėjas*. Yra ir daugiau nemokamų vertimo prieigų, pavyzdžiui, Vytauto Didžiojo universiteto „Internetinė informacijos vertimo priemonė“, verčianti iš anglų kalbos į lietuvių kalbą. Tačiau nė viena automatizuota vertimo priemonė nepajėgia idealiai išversti teksto dėl kalbų savitumo, daugiareikšmiškumo ir kitų priežasčių. I. Petkevičiūtė ir B. Tamulynas savo straipsnyje [66] aptaria, kokios vertimo klaidos dažniausiai daromos ir kaip būtų galima tobulinti kompiuterinio vertimo sistemas.

Tobulinant mašininio vertimo programas yra aktualus ir leksinio vieneto atpažinimas. J. Kovalevskaitės daktaro disertacijoje „Lietuvių kalbos samplaikos“ (2012) tiriamas lietuvių kalbos samplaikų, kaip leksinio vieneto, pasižyminčio formos ir turinio stabilumu, autonomiškumas.

R. Marcinkevičienės monografijoje [52] aptariamas lietuvių kalbos kolokacijų (t.y. daiktavardinių frazių – stabiliai vartojamų junginių su daiktavardžiais) žodynas, jo sudarymo principai bei paskirtis. Šis žodynas „naudingas kuriant statistinius, tikimybinius lietuvių kalbos modelius ir jais pagrįstus programinius įrankius“ (197 psl.), jį taip pat galima panaudoti ir kuriant mašininio vertimo sistemas.

Kalbos kompiuterinio apdorojimo reikmėms, taip pat tekstui redaguoti yra svarbus ir skiemenavimo įrankis. Pirmąją lietuviško teksto skiemenavimo kompiuterinę programą SKIE-MUO sukūrė J. Skendelis XX a. 10-ojo dešimtmečio pradžioje.

Ši programa naudojo žodyną, kuriame nurodomas problemiškesnių žodžių taisyklingas skiemenavimas. Programa buvo pritaikyta Windows terpei ir joje veikiančioms leidybos programoms. Tačiau modelis, besiremiantis funkcine skiemens teorija, ne visada gali būti taikomas be išlygų, kadangi lietuvių kalbai būdingas didelis žodžių ir jų formų daugiareikšmiškumas, atsiranda naujadarų, esama nemažai kirčiavimo gretimybių. Kad būtų galima analizuoti ir aprašyti skiemens modelius, reikia tiksliai nustatyti skiemenų ribas. Todėl VDU mokslininkai tobulino skiemenavimo programą. G. Raškinio ir A. Kazlauskienės straipsnyje [69] aprašytas skiemenavimo algoritmas naudojamas VDU sukurtoje automatinio transkribavimo ir kirčiavimo programoje, kuria paremtas ir visiems prieinamos kirčiuoklės (*http://donelaitis.vdu.lt/main.php?id=4&nr=9*) veikimas. Šis skiemenavimo įrankis naudojamas ir fonotaktikos bei ritmikos tyrimuose. [69]

Skiemenavimo, kirčiavimo ir transkribavimo principai bei algoritmai išsamiai aprašyti A. Kazlauskienės, G. Raškinio ir A. Vaičiūno studijoje [40]. Skiemenavimo, kirčiavimo ir transkribavimo algoritmus taip pat kūrė ir Vilniaus universiteto bei Matematikos ir informatikos instituto mokslininkai (plačiau žr. [40] 7–8 psl.).

Šiuolaikiniame pasaulyje vis aktualesnis automatinio kalbos / šnekos atpažinimo klausimas – kuriama vis daugiau diktavimo, balso operatorių, balsu valdomos paieškos ir navigacijos sistemų. Nors lietuvių šnekos atpažinimo sistemos pradėtos kurti palyginti neseniai, tačiau šiuo klausimu jau nemažai nuveikta.

A. Lipeika ir kt. [51] aprašo atskirai tariamų žodžių šnekos atpažinimo sistemą, kuri remiasi dinaminio laiko suspaudimo metodika. Sistemos naudojami požymiai – tiesinės prognozės koeficientai.

G. ir D. Raškinių straipsnyje [68] pristatomas sukurtõs lietuvių šnekos atpažinimo sistemos parametrų ir jų įtakos atpažinimo tikslumui tyrimas. Nagrinėjama atpažinimo tikslumo priklausomybė nuo naudojamo fonemų rinkinio, nuo naudojamų paslėptųjų Markovo modelių prototipų ir kontekstinių fonemų klasterizacijos algoritmų.

I. Dabašinskienė [22] nagrinėja šnekamosios lietuvių kalbos morfologines ypatybes. Autorė išskiria tokius kriterijus, kurie galėtų būti tinkami šnekamajai kalbai atpažinti: labai dažnas veiksmažodžio, įvardžio, dalelytės ir kitų tarnybinių kalbos dalių vartoji-

mas, itin retas būdvardžio vartojimas ir t.t. Gautieji rezultatai buvo pirmasis bandymas statistiškai pagrįsti dažnai tik stebėjimu užfiksuotus šnekamosios ir rašytinės kalbos skirtumus.

Pastaruoju metu parašyta nemažai daktaro disertacijų, kuriose nagrinėjamas automatinis šnekos atpažinimas.

A. Vaičiūno daktaro disertacijoje [90] nagrinėjami statistiniai lietuvių kalbos modeliai ir jų taikymas automatinio šnekos atpažinimo problemai spręsti. Sukurtas labai didelio žodyno, daugelio diktorių, lietuvių rišlios šnekos atpažinimo sistemos prototipas, paremtas paslėptųjų Markovo modelių metodologija. Įvertinta atpažinimo sistemos žodyno dydžio ir įvairių statistinių kalbos modelių įtaka tokios sistemos atpažinimo tikslumui.

G. Tamulevičiaus daktaro disertacijoje „Pavienių žodžių atpažinimo sistemų kūrimas“ (2008) analizuojamos pavienių žodžių atpažinimo problemos, lyginami atpažinimo metodai, sprendžiami pavienių žodžių atpažinimo sistemos kūrimo klausimai – sukurta pavienių žodžių atpažinimo (sukurtas automatinis žodžio ribų nustatymo metodas, pasižymintis stabilumu bei atsparumu signalo kokybės kitimui) ir segmentavimo sistema (remiantis garsų ribų nustatymo metodika sukurti du metodai žodžiams segmentuoti – tikėtinumo funkcijos maksimizavimo ir prognozės klaidos minimizavimo).

S. Laurinčiukaitės daktaro disertacijoje „Lietuvių šnekos atpažinimo akustinis modeliavimas“ (2008) tirtas žodžiais, skiemenimis, kontekstiniais skiemenimis, fonemomis ir kontekstinėmis fonemomis grįstas šnekos atpažinimas, atlikti lyginamieji atpažinimo tyrimai. Modeliuojamos automatinio šnekos atpažinimo sistemos naudoja statistinį metodą, susijusį su paslėptaisiais Markovo modeliais.

G. Norkevičiaus disertacijoje [64] nagrinėjama anksčiau netyrinėta problema – nuo kalbos nepriklausomų garsų trukmių prognozavimo modelių kūrimas. Kaip teigiama disertacijoje, kalbininkų atliekami lietuvių kalbos garsų trukmės tyrimai „yra daugiau aprašomosios statistikos pobūdžio ir apsiriboja pavienių požymių įtakos garso trukmei analize“ (3 psl.). Disertacijoje požymių įtaka garsų trukmei mašininio mokymo algoritmo pagalba išmokstama iš duomenų ir užrašoma sprendimo medžio pavidalu.

Šiuo metu kalbos atpažinimo klausimai sprendžiami Matematikos ir informatikos institute, Vytauto Didžiojo universitete bei Kauno technologijos universitete. Pagrindinis dėmesys yra skiriamas ištisinės kalbos atpažinimui taikant paslėptuosius Markovo

modelius (kuriami kalbos modeliai, atliekami eksperimentai), kalbos duomenų bazėms (kaupiamos pavienių žodžių ir ištisinės kalbos duomenų bazės) bei bazių kaupimo automatizavimui. [84]

Lietuvių kalbos analizėje buvo taikomi ir matematinės statistikos metodai. R. Merkytė, taip pat ir kartu su V. Kalinka, nagrinėjo komponentų, sudarančių lingvistinį vienetą, kiekio pasiskirstymo dėsnius ir jų statistinio vertinimo uždavinius (žr. [57], [56]).

Kadangi žodis susideda tarsi iš dviejų dalių: pagrindinės (šaknis) ir papildomos (priešdėliai, priesagos, galūnės), tai natūralu tarti, kad žodžių susidarymą aprašo skirtingi tikimybiniai dėsniai. Šia mintimi remiasi V. Fukso (W. Fuchs, 1957) pasiūlytas skiemenų skaičiaus žodyje (komponentų kiekio lingvistiniame vienete) tikimybinis modelis, kurį R. Merkytė ir V. Kalinka išvystė ir pritaikė lietuvių kalbos tyrimuose. Skiemenų žodyje skaičius Z aprašomas kaip diskretaus atsitiktinio dydžio η ir nepriklausomo Puasono atsitiktinio dydžio ξ suma. [56] darbe buvo išvestos skaičiavimui patogios formulės, susiejančios atsitiktinio dydžio Z skirstinio nežinomus parametrus, $n-1$ atsitiktinio dydžio η tikimybę ir ξ vidurkį su to dydžio Z pirmaisiai momentais, ir aprašyta tų parametrų statistinio vertinimo procedūra, kuri remiasi momentų metodu. Autoriai ištyrė teorinio modelio suderinamumą su lietuvių kalbos žodyno duomenimis: su $n = 4$ buvo pasiekta gera atitiktis. Skiemenų skaičiaus skirstinys žodyne skiriasi nuo jo skirstinio žodžiuose iš lietuvių kalbos teksto, nes pastarajame tas pats žodis gali kartotis keletą kartų, todėl skiemenų skaičiaus skirstinio tekstuose patikslinimui buvo įvestos atitinkamos pataisos, nusakytos žodžių iki 3 skiemenų imtinai dažnumu nagrinėjamame tekste.

Kitame R. Merkytės straipsnyje (žr. [58]) pastebėta, kad ilgesniuose žodžiuose skiemenys turi tendenciją trumpėti, ir, naudojant regresinę analizę, išvesta apytikslė formulė, siejanti skiemenų žodyje skaičių Z su fonemų jame kiekiu k: $Z \approx 2,2k + 0,9$. Taip pat R. Merkytė tyrė raidžių lietuvių kalbos tekstuose entropiją ir informacijos kiekį, jų pasiskirstymą žodžiuose (žr. [59]).

Dar viena matematinės lingvistikos sritis, kurioje gana nemažai nuveikta ir Lietuvoje – fonemų, raidžių, žodžių ir jų formų dažnumo tyrinėjimas.

A. Girdenio ir V. Karosienės straipsnyje [29] aptariami skiemens ir žodžio pirmųjų bei paskutiniųjų fonemų dažnumo tyrimo rezultatai. Kompiuterio programos iš rišlių tekstų masyvo automatiškai atrinko pirmąsias ir paskutiniąsias žodžių bei skiemenų fonemas, apskaičiavo jų absoliutų bei santykinį dažnumą ir rezultatus sugrupavo dažnumų mažėjimo tvarka. Paaiškėjo, kad skiemens bei žodžio pìrmosios ir paskutinės fonemos dažnumo atžvilgiu visiškai skiriasi – pradžiose (ypač skiemens) ryškiai dominuoja priebalsiai, o pabaigose – balsiai (išskyrus didelio morfologinio krūvio fonemą /s/, kuri žodžio gale pasižymi dideliu dažnumu).

Išsamesnę lietuvių kalbos garsų dažnumo analizę pateikia A. Kazlauskienė ir G. Raškinis [39].

Be žinomų L. Grumadienės ir V. Žilinskienės „Dažninio dabartinės rašomosios lietuvių kalbos žodyno (mažėjančio dažnio tvarka)“ (1997) bei „Dažninio dabartinės rašomosios lietuvių kalbos žodyno (abėcėlės tvarka)“ (1998), 2009 metais paskelbta ir elektroninė dažninio žodyno versija – A. Utkos sudarytas „Dažninis rašytinės lietuvių kalbos žodynas“, kuriame pateikiami ne tik žodžių, bet ir kaitomų žodžių formų dažniai (žr. [86]), o 2011 metais paskelbtas E. Rimkutės, A. Kazlauskienės ir G. Raškinio parengtas „Dažninis lietuvių kalbos morfemikos žodynas“ (I–III dalys).

Savo disertacijoje ir straipsniuose (žr. [89], [88]) A. Utka nagrinėja dažniausių lietuvių kalbos žodžių ir žodžių formų savybes ir jų svarbą teksto analizei. Anot autoriaus, dažniausios žodžių formos nuo retesnių skiriasi ne tik ypač dideliu dažnumu, bet ir kitomis tik joms būdingomis savybėmis. Šių žodžių ir žodžių formų pasiskirstymas tekstuose nėra atsitiktinis. Būdami dažniausi struktūriniai teksto vienetai, jie yra tiesiogiai susiję su teksto funkcijomis, todėl gali būti laikomi reikšmingais teksto funkcinių ypatybių rodikliais.

Lietuvos mokslo tarybos 2010–2011 m. finansuoto projekto „Morfeminė lietuvių kalbos žodžių struktūra“ metu išsiaiškinta, kad žodžių morfeminės struktūros modelių gali būti labai daug. Remiantis tyrimo medžiaga sudaryti elektroniniai internete laisvai prieinami morfemų žodynai: abėcėlinis, atgalinis ir jau minėtas dažninis. [41]

Žodynus galima rasti VDU puslapyje, adresu *http://donelaitis.vdu.lt/lkk/index.php?item=6&subid=2*.

2

Naudojamos sąvokos, modeliai ir metodai

2.1 Pagrindinės darbe vartojamos lingvistikos sąvokos

Šiame poskyryje pateikiamos pagrindinės lingvistikos sąvokos (plačiau žr. [7]), vartojamos disertacijoje ir cituojamoje literatūroje.

Afiksas – kiekviena reikšminė žodžio dalis (morfema), išskyrus šaknį (pvz., priešdėlis, priesaga, galūnė, intarpas). Skiriami žodžių darybos ir kaitybos afiksai. Lietuvių kalboje labiausiai paplitę kaitybos afiksai yra galūnės, skiriančios įvairias to paties žodžio formas (pvz., linksnių formas *rakt-as, rakt-o, rakt-ui* ir pan.).

Balsės – raidės, žyminčios balsius.

Balsiai – kalbos garsai (žr. *kalbos garsai*), kuriuos tariant nesusidaro aiškaus tarimo židinio ir kurių suvokimą lemia muzikinis tonas. Pagal liežuvio horizontalų pasislinkimą dantų atžvilgiu ir jo pakilimo vietą balsiai gali būti *priešakinės eilės* (žymimi raidėmis e, ę, ė, i, į, y) arba *užpakalinės eilės* (žymimi raidėmis a, ą, o, u, ų, ū). Pagal lūpų veiklą skiriami *lūpiniai*, kuriuos tariant lūpos atkišamos į priekį ir sudedamos ratuku (žymimi raidėmis o, u, ų, ū), ir *nelūpiniai* balsiai (visi kiti balsiai: a, ą, e, ę, ė, i, į, y).

Daugiareikšmiškumas, polisemija – žodžio (rečiau kurio nors kito reikšminio kalbos vieneto) daugiareikšmiškumas, t.y. dviejų arba daugiau reikšmių turėjimas. Pvz., *žalias, -ia*: 1) 'žolės spalvos', 2) 'neprinokęs, nesubrendęs', 3) 'neišdžiūvęs', 4) 'neapdorotas, neišdirbtas (neišviręs, neiškepęs, nedegtas)', 5) 'jaunas, sveikas, stiprus', 6) 'nepatyręs, neišpruses'.

Fonema – skiriamasis (fonologinis) garsinis kalbos elementas, nebesuskaidomas į mažesnius vienas po kito einančius vienetus. Keičiant vieną fonemą kita, kinta žodžių dalykinė reikšmė (plg. *kãras – gãras, takù – tekù, màno – mãno*, kadangi /*k*/ ir /*g*/, /a/ ir /e/, /*a*/ ir /*a:*/ yra skirtingos fonemos).

Fonetika: **1.** Kurios nors kalbos garsų (žr. *kalbos garsai*) akustinių ir artikuliacinių savybių visuma. **2.** Mokslas, tiriantis kalbos garsų funkcijas ir jų akustines bei artikuliacines ypatybes. Plačiau suprantama fonetika apima ir fonologiją (žr. *fonologija*).

Fonologija – mokslas, tiriantis kalbos garsus (žr. *kalbos garsai*), jų derinius bei požymius žodžių ir jų formų skiriamosios funkcijos atžvilgiu.

Funkcinis stilius – istoriškai susidariusi kalbos atmaina, kurios ypatybes lemia kalbos vartojimo sritis ir funkcijos. Dabartinėje lietuvių kalboje skiriami šie svarbiausi funkciniai stiliai: buitinis (šnekamasis), publicistinis, kanceliarinis (administracinis), mokslinis ir grožinis (meninis).

Gramatinė forma – kalbos forma, reiškianti gramatinius ryšius ir turinti apibrėžtą funkciją gramatinėje kalbos sandaroje. Skiriamos sintaksinės formos (sakinių ir žodžių tarpusavio ryšio modeliai bei sandaros tipai) ir morfologinės bei darybos formos (žodžiai su tam tikromis kaitybos bei darybos galūnėmis, priesagomis, priešdėliais). Lietuvių kalboje tas pats kaitomas žodis gali turėti įvairias gramatines formas. Pvz., tas pats daiktavardis turi įvairias linksnių, skaičių gramatines formas (*namas, namo, namui, ..., namai, namų,...*).

Homoformos – skirtingų žodžių vienodai tariamos ãtskiros kaitybos formos. Pvz., *sakaĩ* (daiktavardžio daugiskaitos vardininkas) ir *sakaĩ* (veiksmažodžio *sakyti* esamojo laiko 2 asmuo).

Kalbos garsai – mažiausios kalbos srauto atkarpos, maždaug atitinkančios vieną savarankišką fonemą (žr. *fonema*).

Kontekstas – sakytinė arba rašytinė žodžio, sakinio ar didesnio kalbos vieneto aplinka.

Leksema – dvipusis (reikšminis) kalbos vienetas, tarpinis tarp morfemos (žr. *morfema*) ir sintaksinio junginio. Terminas reiškia tą patį, ką ir *žodis*, kai pastaruoju suprantamos visos paradigminės formos ir reikšmės (pvz., leksema *naujas* apima *naujas, -a, naujesnis, -ė, naujausias, -ia, naujasis, -oji* su visomis jų linksnių formomis ir visomis reikšmėmis). Jei vartojami abu terminai, skirtumas tik tas, kad leksema reiškia abstraktų kalbos (jos leksikos sistemos) vienetą, o žodis – konkrečią to vieneto realizaciją (su visais formos ir reikšmės variantais) kalbėjime.

Leksika – kalbos žodžių visuma. Taip dažnai vadinami ne tik visi kurios nors kalbos žodžiai (paprastai kartu su frazeologizmais – pastoviais, džn. vaizdingais pasakymais, turinčiais vientisinę reikšmę, neišvedamą iš juos sudarančių žodžių įprastinių reikšmių, pvz., *į akį dėti* – 'miegoti'), bet ir atskirų jos atmainų bei variantų (tarmių, socialinių dialektų, funkcinių stilių ir pan.) žodžių visuma.

Leksikologija – kalbotyros šaka, tirianti leksiką (žr. *leksika*).

Lema – tai žodyno straipsnio antraštinis žodis; tekstyne pavartoto žodžio antraštinė forma (pvz., daiktavardžiams, būdvardžiams, skaitvardžiams – (vyriškosios giminės – jeigu kaitoma giminėmis) vienaskaitos vardininkas, veiksmažodžiams – bendratis). [70], [73], [81]

Matematinė kalbotyra – sąlyginis pavadinimas dabartinės kalbotyros krypčių, kurios kalbai modeliuoti bei tirti naudojasi matematine logika, algebra, aibių, tikimybių bei automatų teorijomis ir statistine analize.

Metakalbinis komentaras – kalbos arba teksto kalbinės raiškos aptarimas: vertinimas, vartojimo motyvavimas (t.y. kalbėtojas vertina žodį arba jo tinkamumą šnekoje, motyvuoja savo sprendimo pasirinkimą, pvz., *kitaip nepasakysi, nelinksmai sakant*), raiškos paieška (pvz., *kaip čia pasakius*) ir kt. [99]

Morfema – mažiausias dvipusis (turintis formą ir reikšmę) kalbos sistemos vienetas. Skiriamos šakninės morfemos ir afiksai (žr. *afiksas*). Pvz., žodį *užsienietis* sudaro šakninė morfema *-sien-* ir afiksai: priešdėlis *už-*, priesaga *-iet-* ir galūnė *-is*.

Morfologija: **1.** Žodžių formų ir tomis formomis žymimų gramatinių reikšmių sistema. **2.** Gramatikos sritis, tirianti žodžių formas, tų formų santykius bei sistemas

ir kaitybos paradigmas.

Padalyvis – neasmenuojamoji ir nelinksniuojamoji veiksmažodžio forma, turinti veiksmažodžio ir prieveiksmio ypatybių. Žymi pašalinio veikėjo atliekamą arba savaime vykstantį šalutinį veiksmą. Pvz., *dirbant, dirbus, dirbdavus, dirbsiant* ir pan.

Priebalsės – raidės, žyminčios priebalsius (žr. *priebalsiai*).

Priebalsiai – kalbos garsai (žr. *kalbos garsai*), skiemenyje užimantys periferinę padėtį, turintys aiškų tarimo židinį. Pagal balso stygų veiklą priebalsiai skirstomi į *skardžiuosius*, kuriuos tariant virpa balso stygos, oro srovė būna silpnesnė (žymimi raidėmis b, d, g, z, ž, h), *dusliuosius*, kuriuos tariant balso stygos būna prasiskleidusios ir oras, pučiamas iš plaučių, jų nevirpina (žymimi raidėmis p, t, k, s, š, ch, f), ir *pusbalsius*, kurie tuo pačiu yra ir skardieji priebalsiai (žymimi raidėmis l, m, n, r, j, v).

Prieveiksmis – kalbos dalis, kurią sudaro nelinksniuojami ir neasmenuojami žodžiai, reiškiantys veiksmų, būsenų ir ypatybių požymius arba aplinkybes. Pvz., *tinkamai, aukščiau, toliausiai, platyn, veltui, tyčia, dabar* ir t.t.

Semantika: **1.** Kalbotyros šaka, tirianti kalbos vienetų turinį (reikšmę). **2.** Žodžio, jo formos ar žodžių junginio reikšminė sandara.

Semema – tai reikšminis morfemos (žr. *morfema*), žodžio ar tam tikro žodžių junginio elementas, nesusijęs su konkrečia šių vienetų forma (pvz., žodžių *duoti, duoklė* semema susijusi su „davimo“ sąvoka, žodžių junginio (frazeologizmo) *pūsti į akį* semema – su „miegojimo“ sąvoka ir pan.). [65]

Semiotika – mokslas apie ženklus ir jų sistemas.

Sintaksė: **1.** Kurios nors kalbos sakinių ir kitų didesnių už žodį vienetų sandara. **2.** Gramatikos šaka, tirianti sakinių ir kitų didesnių už žodį kalbos vienetų sandarą bei vartoseną. Pagrindinis sintaksės objektas yra sakinys kaip žodžių tarpusavio ryšių ir prasminių santykių visuma. **3.** Semiotikos (žr. *semiotika*) šaka, tirianti linijinius ženklų tarpusavio santykius.

Stilistika, lingvistinė stilistika, – kalbotyros šaka, kurios atstovai tiria kalbos priemones, atsižvelgdami į jų ekspresinę, emocinę bei estetinę paskirtį ir bendrinės kalbos funkcinius stilius (žr. *funkcinis stilius*).

Stilius: **1.** Kalbos atmaina, kurios ypatybes lemia vartojimo sritis ir funkcijos visuomenėje. Beveik visose rašytinę tradiciją turinčiose kalbose galima išskirti neutralų,

knyginį („aukštąjį“) ir buitinį stilių. Pagal ypatybes, susijusias su kalbos funkcijomis visuomenėje, skiriami funkciniai stiliai (žr. *funkcinis stilius*). **2.** Tam tikram autoriui, jų grupei ar srovei būdingas kalbos priemonių parinkimas ar jų vartosenos savitumas. Tai literatūros stilius, literatūros tyrimo objektas.

Tekstas: **1.** Užrašyta kalbos atkarpa. **2.** Savo prasme susijusių sakinių visuma.

Tekstynas – kompiuterinių (elektroninių, suskaitmenintų) tekstų rinkinys, reprezentuojantis kurią nors kalbą ar kalbos atmainą (pvz., tam tikrą funkcinį stilių).

Tekstynų lingvistika (angl. *corpus linguistics*) – kalbotyros šaka, tirianti tekstynų sudarymo ir panaudojimo principus bei metodus.

Teksto lingvistika – kalbotyros sritis, tirianti didesnių už sakinį kalbos vienetų sandarą ir sakinių tarpusavio ryšius.

Žodžio forma, kaitybos forma, – (kaitomo) žodžio viena iš galimų gramatinių atmainų (žr. *gramatinė forma*).

2.2 Imčių metodų elementai

Šiame poskyryje, remiantis D. Krapavickaitės ir A. Plikuso knyga [49], aprašoma keletas imčių teorijos sąvokų.

Daugelio sričių tyrimų pagrindą sudaro netikimybinės imtys iš baigtinių populiacijų, pvz., proginė (patogioji) arba tikslinė (ekspertinė) imtis, kai daromos išvados iš tyrėjams lengviausiai prieinamų arba ekspertų parinktų „tipinių“ populiacijos elementų. „Tokiais atvejais būna nagrinėjama tiktai pati imtis ir nekeliamas klausimas apie tai, koks yra jos ryšys su visa populiacija“ ([49], p. 33). Be abejo, tinkamai taikoma bet kuri netikimybinė imtis gali duoti gerus rezultatus, tačiau iš netikimybinių imčių gautų įverčių paklaidos negali būti statistiškai įvertintos.

Imtis laikoma reprezentatyvia, jeigu ji gerai atspindi tiriamą populiaciją. Lenkų kilmės amerikiečių statistikas J. Neimanas (J. Neyman) 1934 m. teoriškai įrodė ir iliustravo praktiniais pavyzdžiais, kad, skirtingai nuo tikslinių imčių, atsitiktinės (reprezentatyviosios) imtys leidžia statistiškai pagrįsti rezultatus, kokybiškai įvertinti gaunamų rezultatų tikslumą (žr. [63]).

Tokie tyrimai, kai renkami ir nagrinėjami imties elementų duomenys, siekiant padaryti išvadas apie visą populiaciją, vadinami imčių tyrimais. Imčių tyrimai taikomi daugelyje mokslo ir praktikos sričių.

Tarp imčių metodų teorijos specialistų išskiriamų pagrindinių tyrimo planavimo ir vykdymo dalių minimos ir tokios: tyrimo tikslo nusakymas, tikslo populiacijos apibrėžimas, reikalingų tyrimo rezultatų nusakymas, informacijos rinkimo būdo pasirinkimas, imties išrinkimo būdo pasirinkimas ir imties dydžio nustatymas, tikimybinės imties išrinkimas.

D. Krapavickaitės ir A. Plikuso knygoje [49] teigiama, kad atlikdami imčių tyrimą, imčių teorijos specialistai pirmiausia sprendžia tokį uždavinį: kaip geriausiai išrinkti imtį ir surinkti duomenis, t.y. koks turėtų būti imties dydis, kaip ją išrinkti, kokius reikėtų taikyti duomenų rinkimo metodus, kokių kintamųjų reikšmes matuoti.

Vienas paprasčiausių ir dažniausiai taikomų imties išrinkimo būdų yra paprastasis atsitiktinis ėmimas.

Paprastoji atsitiktinė (negrąžintinė) imtis – tai tokia n skirtingų elementų imtis iš N dydžio baigtinės populiacijos, kai bet kuris n skirtingų elementų rinkinys turi vienodą tikimybę būti išrinktas.

Iš N dydžio populiacijos $\mathcal{U} = \{1, 2, \ldots, N\}$ galima išrinkti C_N^n tokių imčių. Tikimybė, kad kiekvienas n skirtingų elementų rinkinys $\mathbf{i} = \{i_1, \ldots, i_n\}$ bus išrinktas iš N dydžio populiacijos, $p(\mathbf{i}) = 1/C_N^n$ visoms galimoms $\mathbf{i}$. Tikimybė, kad k-asis populiacijos elementas priklauso kuriai nors n dydžio paprastajai atsitiktinei imčiai,

$$\pi_k = \mathbf{P}(\mathbf{i} : k \in \mathbf{i}) = \frac{n}{N}, \quad k = 1, 2, \ldots, N.$$

Tikimybė, kad populiacijos elementai k ir l priklauso paprastajai atsitiktinei imčiai,

$$\pi_{kl} = \mathbf{P}(k \in \mathbf{i}, l \in \mathbf{i}) = \mathbf{P}(k \in \mathbf{i})\mathbf{P}(l \in \mathbf{i} | k \in \mathbf{i}) = \frac{n}{N}\frac{n-1}{N-1},$$

$k, l = 1, \ldots, N, k \neq l$. Taigi, įvykiai $\{k \in \mathbf{i}\}$ ir $\{l \in \mathbf{i}\}$ yra priklausomi, nes $\pi_{kl} \neq \pi_k \pi_l = (n/N)^2$, tačiau esant didelei populiacijai galima laikyti, kad $\pi_{kl} \approx (n/N)^2$.

Paprastosios atsitiktinės imties $\mathbf{i}$ iš baigtinės populiacijos $\mathcal{U} = \{1, 2, \ldots, N\}$ tyrimo kintamojo y vidurkis μ vertinamas imties vidurkiu

$$\overline{y} = \frac{1}{n} \sum_{k \in \mathbf{i}} y_k.$$

Teiginys 1. *Turint paprastąją atsitiktinę n dydžio imtį iš N dydžio baigtinės populiacijos,*

a) *populiacijos vidurkio μ įvertinys $\overline{y}$ yra nepaslinktasis;*

b) *šio įvertinio dispersija yra*

$$\mathbf{D}\overline{y} = \left(1 - \frac{n}{N}\right)\frac{s^2}{n}, \quad \text{čia} \quad s^2 = \frac{1}{N-1}\sum_{i=1}^{N}(y_i - \mu)^2;$$

c) *dispersijos įvertinys*

$$\widehat{\mathbf{D}}\overline{y} = \left(1 - \frac{n}{N}\right)\frac{\hat{s}^2}{n}, \quad \text{čia} \quad \hat{s}^2 = \frac{1}{n-1}\sum_{k \in \mathbf{i}}(y_k - \overline{y})^2,$$

yra nepaslinktasis.

Šio ir kitų 2.2 poskyrio teiginių įrodymai yra [49] vadovėlyje.

Kadangi populiacijos suma t nuo populiacijos vidurkio μ skiriasi tik pastoviuoju daugikliu N, t.y.

$$t = \sum_{k=1}^{N} y_k = N\mu,$$

todėl sumos t (nepaslinktojo) įvertinio $\hat{t} = N\overline{y}$ dispersijos įvertinys

$$\widehat{\mathbf{D}}\hat{t} = N^2\widehat{\mathbf{D}}\overline{y}$$

taip pat neturės poslinkio.

Vertinant populiacijos dalį sukuriamas fiktyvus kintamasis

$$y_k = \begin{cases} 1, & \text{jei } k\text{-asis elementas turi tiriamą požymį,} \\ 0 & \text{priešingu atveju,} \end{cases}$$

čia $k = 1, 2, \ldots, N$. Tuomet išplaukia, kad, turint paprastąją atsitiktinę n dydžio imtį iš N dydžio baigtinės populiacijos, populiacijos dalies p (nepaslinktasis) įvertinys yra $\hat{p} = \overline{y}$, o jo dispersijos (nepaslinktasis) įvertinys

$$\widehat{\mathbf{D}}\hat{p} = \left(1 - \frac{n}{N}\right)\frac{\hat{p}(1-\hat{p})}{n-1}. \tag{2.1}$$

Intervaliniai įvertiniai, arba pasikliautinieji intervalai, yra patogus būdas apibendrinti imčių tyrimo rezultatus. Paprastai jie remiasi parametrų ir jų dispersijų įvertiniais bei normaliąja aproksimacija. Tarkime, kad statistika

$$\xi = \frac{\hat{\theta} - \theta}{\sqrt{\widehat{\mathbf{D}}\hat{\theta}}}$$

yra apytiksliai pasiskirsčiusi pagal standartinį normalųjį skirstinį. Čia $\widehat{\mathbf{D}}\hat{\theta}$ yra parametro įvertinio $\hat{\theta}$ dispersijos įvertinys. Tada parametro θ įvertinio apytikslis pasikliautinasis intervalas yra toks:

$$\left(\hat{\theta} - d, \hat{\theta} + d\right) = \left(\hat{\theta} - z_{\alpha/2}\sqrt{\widehat{\mathbf{D}}\hat{\theta}}, \hat{\theta} + z_{\alpha/2}\sqrt{\widehat{\mathbf{D}}\hat{\theta}}\right), \tag{2.2}$$

čia $z_{\alpha/2}$ yra standartinio normaliojo skirstinio $1 - \alpha/2$ lygmens kvantilis.

Paprastoji atsitiktinė grąžintinė imtis – tai tokia n elementų imtis iš N dydžio baigtinės populiacijos, kai bet kuris n elementų rinkinys, besiskiriantis nuo kitų elementų rinkinių ir kuriame galbūt yra pasikartojimų, turi vienodą tikimybę būti išrinktas.

Negrąžintinės tokio pat dydžio imties atveju vidurkio įvertinys yra tikslesnis, kadangi grąžintinėje imtyje galimas informacijos dubliavimasis.

Vienu iš svarbiausių imčių teorijos rezultatų laikomas 1952 m. D. G. Horvico (D. G. Horvitz) ir D. J. A. Tompsono (D. J. A. Thompson) pasiūlytas (žr. [34]) universalus sumos įvertinys, tinkamas bet kokiam imties planui.

Tegu $\pi_k, k = 1, 2, \ldots, N$, žymi k-ojo populiacijos elemento priklausymo imčiai tikimybę

$$\pi_k = \mathbf{P}(k \in \mathbf{i}) = 1 - (1 - p_k)^n;$$

$$\pi_{kl} = \mathbf{P}(k \in \mathbf{i}, l \in \mathbf{i}) = \pi_k + \pi_l - (1 - (1 - p_k - p_l)^n),$$

$k, l = 1, \ldots, N$, – k-ojo ir l-ojo populiacijos elementų tikimybę kartu priklausyti imčiai arba antrosios eilės priklausymo imčiai tikimybę; ν – efektyvųjį imties dydį, t.y. skirtingų elementų skaičių n dydžio imtyje.

Teiginys 2 (Horvico ir Tompsono įvertinys)**.** *Tegu* $\mathbf{i}$ *yra atsitiktinė imtis, išrinkta pagal bet kokį imties planą. Tuomet*

a) *populiacijos sumos įvertinys*

$$\hat{t}_\pi = \sum_{i=1}^{\nu} \frac{y_i}{\pi_i}$$

yra nepaslinktasis;

b) *šio įvertinio dispersija yra*

$$\mathbf{D}\hat{t}_\pi = \sum_{k=1}^{N} \frac{1 - \pi_k}{\pi_k} y_k^2 + \sum_{k=1}^{N} \sum_{\substack{l=1 \\ l \neq k}}^{N} \frac{\pi_{kl} - \pi_k \pi_l}{\pi_k \pi_l} y_k y_l;$$

c) *įvertinio* $\hat{t}_\pi$ *dispersijos įvertinys*

$$\widehat{\mathbf{D}}\hat{t}_\pi = \sum_{i=1}^{\nu} \frac{1 - \pi_i}{\pi_i^2} y_i^2 + \sum_{i=1}^{\nu} \sum_{\substack{j=1 \\ j \neq i}}^{\nu} \frac{\pi_{ij} - \pi_i \pi_j}{\pi_i \pi_j} \frac{y_i y_j}{\pi_{ij}},$$

jei $\pi_{ij} > 0$*, yra nepaslinktasis.*

Sluoksnine imtimi, arba sluoksniniu ėmimu, vadinami tokie imties planai, kai, suskaidžius populiaciją į kelias nesikertančias dalis – sluoksnius,– imtys renkamos iš kiekvieno sluoksnio atskirai, nepriklausomai nuo kitų sluoksnių. Skirtingiems sluoksniams gali būti taikomi skirtingi imčių planai ir parametrų vertinimo būdai.

Sluoksninis ėmimas taikomas, jei:

• populiacijos skaidymas į sluoksnius, kuriuos sudaro panašūs elementai, leidžia išvengti visiškai blogų imčių ir tikėtis tikslesnių įverčių;

• reikia gauti įverčius populiacijos srityse – skaidant populiaciją į sluoksnius galima sritis laikyti sluoksniais, taip užtikrinant jose reikiamo dydžio imtį ir reikiamą įverčių tikslumą.

Tinkamai parinkus sluoksnius galima sumažinti tyrimo išlaidas, supaprastėja tyrimo organizavimas ir vykdymas.

Dažnai taikomas toks sluoksninis ėmimas, kai iš kiekvieno sluoksnio renkama paprastoji atsitiktinė imtis. Tokiu būdu gaunama imtis vadinama paprastąja atsitiktine sluoksnine imtimi.

Tegu N dydžio baigtinė populiacija $\mathcal{U}$ suskaidyta į H bendrų elementų neturinčių sluoksnių $\mathcal{U}_h$, kurių dydžiai yra N_h, $h = 1, 2, \ldots, H$, $N_1 + \ldots + N_H = N$.

Sakykime, kad n_h žymi imties, išrinktos iš h-ojo sluoksnio, elementų skaičių. Tada bendras imties dydis $n = \sum_{h=1}^{H} n_h$. Tyrimo kintamojo y i-ojo elemento reikšmę h-ajame sluoksnyje pažymėkime y_{hi}, $i = 1, 2, \ldots, N_h$, $h = 1, 2, \ldots, H$. Tada $t_h = \sum_{i=1}^{N_h} y_{hi}$, sluoksnio vidurkis $\mu_h = t_h/N_h$, o visos populiacijos suma $t = \sum_{h=1}^{H} t_h$.

Tegu $\hat{\mu}_h$ žymi vidurkio μ_h, $\hat{t}_h$ – sumos t_h, $h = 1, 2, \ldots, H$ įvertinius, o sumos t įvertinys $\hat{t}_{sl} = \hat{t}_1 + \ldots + \hat{t}_H$.

Teiginys 3. *Paprastosios atsitiktinės sluoksninės imties atveju*

a) *populiacijos vidurkio įvertinys*

$$\hat{\mu}_{sl} = \sum_{h=1}^{H} \frac{N_h}{N} \overline{y}_h, \quad \check{c}ia \quad \overline{y}_h = \frac{1}{n_h} \sum_{i=1}^{n_h} y_{hi},$$

yra nepaslinktasis;

b) *šio įvertinio dispersija yra*

$$\mathbf{D}\hat{\mu}_{sl} = \sum_{h=1}^{H} \left(\frac{N_h}{N}\right)^2 \left(1 - \frac{n_h}{N_h}\right) \frac{s_h^2}{n_h}, \quad \check{c}ia \quad s_h^2 = \frac{1}{N_h - 1} \sum_{j=1}^{N_h} (y_{hj} - \mu_h)^2;$$

c) *dispersijos $\mathbf{D}\hat{\mu}_{sl}$ įvertinys*

$$\widehat{\mathbf{D}}\hat{\mu}_{sl} = \sum_{h=1}^{H} \left(\frac{N_h}{N}\right)^2 \left(1 - \frac{n_h}{N_h}\right) \frac{\hat{s}_h^2}{n_h}, \quad \check{c}ia \quad \hat{s}_h^2 = \frac{1}{n_h - 1} \sum_{j=1}^{n_h} (y_{hj} - \overline{y}_h)^2,$$

yra nepaslinktasis.

Racionaliai sudarant sluoksnius ir optimaliai paskirstant imties dydį beveik visada galima gauti tikslesnį sumos įvertinį, negu turint tokio paties dydžio paprastąją atsitiktinę imtį. Norint gauti kuo tikslesnį sluoksninės imties įvertinį, reikia, kad sluoksnių dispersijos būtų kuo mažesnės, t.y. sluoksniai turi būti homogeniški. Kad sumos įvertinys turėtų kuo mažesnę dispersiją, populiaciją į sluoksnius reikėtų skaidyti taip, kad kintamojo reikšmės sluoksnių viduje būtų kuo panašesnės, arba kad sluoksnių vidurkiai kuo labiau skirtųsi. Jeigu, pavyzdžiui, sluoksnių vidurkiai bus lygūs: $\mu_h = \mu$, $h = 1, \ldots, H$, tada toks sluoksniavimas bus neefektyvus, nes paprastosios atsitiktinės imties planas duos tikslesnį sumos (taip pat ir vidurkio) įvertinį.

Lizdine imtimi, arba lizdiniu ėmimu, vadinamas toks imties sudarymo būdas, kai renkama lizdų (bendrų elementų neturinčių populiacijos dalių) imtis ir apklausiami visi į imtį išrinktų lizdų (pirminių elementų) elementai (antriniai elementai).

Sluoksninis ir lizdinis ėmimas skiriasi tuo, kad sluoksninio ėmimo atveju imtis renkama iš kiekvieno sluoksnio, o lizdinio ėmimo atveju renkama lizdų imtis ir imties elementais laikomi visi išrinktų populiacijos dalių elementai.

Lygiai taip pat, kaip renkama paprastoji atsitiktinė elementų imtis, galima išrinkti ir paprastąją atsitiktinę lizdinę imtį arba grąžintinę lizdinę imtį, renkamą su tikimybėmis p_i, proporcingomis lizdų dydžiui.

Jei lizdai yra homogeniški ir jų vidurkių įvairovė didelė, tada lizdinė imtis nėra efektyvi. Lizdų homogeniškumui matuoti naudojamas pataisytasis lizdų vidinės koreliacijos koeficientas

$$\rho_a = 1 - \frac{SSW}{(N-L)s^2},$$

čia

$$SSW = \sum_{i=1}^{L}\sum_{j=1}^{M_i}(y_{ij} - \mu_i)^2, \quad s^2 = \frac{1}{LM_i - 1}\sum_{i=1}^{L}\sum_{j=1}^{M_i}(y_{ij} - \mu)^2,$$

L – lizdų (pirminių elementų) skaičius populiacijoje, M_i – (antrinių) elementų skaičius i-ajame populiacijos lizde ($i = 1, \ldots, L$), šiuo atveju $M_1 = \ldots = M_L$, N – populiacijos (antrinių) elementų skaičius, y_{ij} – i-ojo lizdo j-ojo elemento kintamojo y reikšmė ($j = 1, \ldots, M_i, i = 1, \ldots, L$), $\mu_i = t_i/M_i$ – kintajomo y vidurkis i-ajame lizde, $\mu = t/N$ – populiacijos vidurkis. Kuo ρ_a reikšmės mažesnės, tuo lizdinė imtis bus efektyvesnė.

Kai lizdai yra dideli ir juos sudarantys populiacijos elementai yra panašūs, gali būti renkama dviejų pakopų lizdinė imtis, kai į imtį įtraukiami ne visi lizdo elementai, o renkama antrinių elementų tikimybinė imtis. Skirtinguose lizduose gali būti taikomi skirtingi imčių planai.

2.3 Logtiesiniai modeliai

Paprasčiausiu atveju tiesinio ryšio tarp aiškinamojo (atsako) kintamojo Y ir aiškinančiųjų (nepriklausomų) kintamųjų $X_1, X_2, ..., X_m$ įvertinimui gali būti naudojamas tiesinis regresijos modelis (žr. [12], [76], [21]), kuriame laikoma, kad Y (sąlyginis) skirstinys yra normalusis su vidurkiu, tiesiškai priklausančiu nuo X-ų. Tačiau, kai aiškinamojo kintamojo Y (sąlyginis) skirstinys nėra normalusis arba priklausomybė tarp Y ir $X_1, X_2, ..., X_m$ nėra tiesinė (abi šios savybės būdingos kokybiniams duomenims), tokių ryšių, naudojant paprastą tiesinį modelį, adekvačiai aprašyti ir įvertinti negalima (žr. [3]). Todėl J. A. Nelder ir R. W. M. Wedderburn (1972) praplėtė tiesinių modelių klasę ir ją pavadino *apibendrintaisiais tiesiniais modeliais* (angl. *Generalized linear models*) ([62], taip pat žr. [4]).

Logtiesiniai modeliai yra atskiras apibendrintųjų tiesinių modelių atvejis, kai jungties funkcija yra *log* arba *logit* funkcija.

2.3.1 Apibendrintasis tiesinis modelis

Apibendrintąjį tiesinį modelį nusako 3 komponentės:

1. Atsitiktinė komponentė apibrėžia aiškinamojo kintamojo skirstinį iš eksponentinės skirstinių šeimos.

2. Sisteminė komponentė yra nusakoma tiesiniu prediktoriumi, kuris aprašo aiškinančiųjų kintamųjų įtaką aiškinamojo kintamojo skirstiniui.

3. Jungties funkcija, kuri susieja aiškinamojo kintamojo vidurkį su tiesiniu prediktoriumi.

Apibrėžimas. *Atsitiktinis vektorius Y arba jo skirstinys priklauso eksponentinei skirstinių šeimai su parametru $\theta \in \Theta \subset \mathbf{R}^d$, jeigu jo skirstinio tankis atžvilgiu σ-baigtinio mato ν turi tokį pavidalą:*

$$p(y \mid \theta) = h(y) \exp\left\{(\psi(\theta))^\top T(y) - b(\theta)\right\}, \quad y \in \mathbf{R}^n, \tag{2.3}$$

čia $\psi\colon \mathbf{R}^d \to \mathbf{R}^d$, $h\colon \mathbf{R}^n \to [0, \infty)$, $T\colon \mathbf{R}^n \to \mathbf{R}^d$ yra žinomos funkcijos, o $b\colon \mathbf{R}^d \to \mathbf{R}$ yra normuojantis narys ir yra vadinama kumuliantine (angl. cumulantive) arba

logaritmine dalinimo (angl. logpartition) funkcija:

$$\exp\{b(\theta)\} = \int_{\mathbf{R}^n} h(y) \exp\left\{(\eta(\theta))^\top T(y)\right\} \nu(\mathrm{d}y).$$

Ji apibrėžta ir baigtinė, kai $\theta \in \Theta := \{u \in \mathbf{R}^n \colon b(u) < \infty\}$. *Funkcijos* ψ, T *ir* h *kartu su matu* ν *nusako konkrečios eksponentinės šeimos pavidalą, jos skirstinių „formą“.*

Funkcija h yra perteklinė, ją galima eliminuoti pakeitus pradinį matą ν matu $\nu_h(dy) = h(y)\nu(dy)$. Tačiau ji leidžia eksponentinės šeimos apibrėžime kaip bazinį matą ν naudoti standartinius matus (Lebego, skaičiuojantįjį matą). Funkcija T yra eksponentinės šeimos (minimali) pakankama statistika.

Kai θ yra kanoninis parametras (t.y., kai $\psi(\theta) \equiv \theta$), eksponentinės skirstinių šeimos tankis turi tokį pavidalą:

$$p(y \mid \theta) = h(y) \exp\left\{\theta^\top T(y) - b(\theta)\right\}, \quad y \in \mathbf{R}^n,$$

kuris vadinamas *kanoniniu*.

Kanoninį parametrą θ ir $T(Y)$ vidurkį bei kovariacijų matricą sieja lygtys

$$\mathbf{E}_\theta T(Y) := \mu(\theta) = \frac{\partial b(\theta)}{\partial \theta},$$

$$\mathrm{Var}(\theta) := \mathbf{E}_\theta T(Y)(T(Y))^\top - \mu(\theta)\,(\mu(\theta))^\top = \frac{\partial^2 b(\theta)}{\partial\theta\partial^\top\theta}.$$

Diskrečių (kategorinių) duomenų atveju galima laikyti, kad $Y \in \mathbf{Z}_+^n$ (čia $\mathbf{Z}_+$ žymi sveikųjų neneigiamų skaičių aibę, $\mathbf{Z}_+ := \{0, 1, \ldots\}$). Kai ν yra skaičiuojantysis matas, apibrėžtas aibėje $\mathbf{Z}_+^n$, lygybė (2.3) suvedama į

$$\mathbf{P}(Y = y) = p(y \mid \theta) = h(y) \exp\left\{(\psi(\theta))^\top T(y) - b(\theta)\right\}, \quad y \in \mathbf{Z}_+^n.$$

Išsamesnį eksponentinės šeimos ir jos savybių aprašymą galima rasti [62], [11].

Tarkime, turime stebinius $\{X, Y\}^N := \{(X(t), Y(t)),\, t = 1, \ldots, N\}$, čia $Y(t) \in \mathbf{R}^n$ yra pagrindinis tyrimo kintamasis, o $X(t) \in \mathbf{R}^k$ yra aiškinamųjų kintamųjų vektorius $(t = 1, \ldots, N)$, $X^N = (X(1), \ldots, X(N))$, $Y^N = (Y(1), \ldots, Y(N))$.

Apibrėžimas. *Sakoma, kad stebiniai* $\{X, Y\}^N$ *tenkina apibendrintąjį tiesinį (AT) modelį, jeigu:*

1) $\{Y(t),\ t = 1, \dots, N\}$ *yra sąlyginai nepriklausomi, kai žinomas* X^N*, ir* $Y(t)$ *sąlyginis skirstinys priklauso eksponentinei skirstinių šeimai su kanoniniu parametru* $\theta(t) = \theta(t \mid X(t)) \in \Theta$ $(t = 1, \dots, N)$*;*

2) apibrėžtas tiesinis prediktorius $\eta(t) := BX(t)$*, čia* B *yra* $n \times k$ *nežinomų AT modelio parametrų matrica;*

3) nusakytos jungties funkcijos $g_t \colon \mathbf{R}^n \to \mathbf{R}^n$*, susiejančios sąlyginį* $T(Y(t))$ *vidurkį* (*kai žinomos* $X(t)$ *reikšmės*)

$$\mu(t) = \mu(t \mid X(t)) := \mathbf{E}[T(Y(t)) \mid X(t)]$$

su tiesiniu prediktoriumi $\eta(t)$ *lygybe:*

$$g_t(\mu(t)) = \eta(t), \quad t = 1, \dots, N.$$

AT modelių statistinė analizė remiasi didžiausio tikėtinumo įvertinių teorija. Šiuo atveju didžiausio tikėtinumo įvertiniai sutampa su momentų metodu gautais įvertiniais.

Logistinės, binominės logistinės, Puasono, neigiamos binominės regresijos modeliai, taip pat polinominė regresija ir apibendrintasis logit modelis yra atskiri AT modelio atvejai. Savo ruožtu, visi jie yra *logtiesiniai* modeliai (žr. 2.3.2 paragrafą).

Detalesnį (apibendrintojo) tiesinio modelio ir jo statistinės analizės aprašymą galima rasti [62], jo taikymai kategorinių duomenų analizėje aptariami [4].

2.3.2 Logtiesinio modelio apibrėžimas

Diskretieji logtiesiniai modeliai yra taikomi stebėtų dažnių daugiamačių lentelių statistinei analizei ir leidžia aprašyti sudėtingus didelio matavimo kategorinių požymių tarpusavio sąryšius (žr. [14]). Juose aiškinamuoju kintamuoju laikomos ne pačios tiriamų požymių stebėtos reikšmės, o tų reikšmių įvairių kombinacijų pasitaikymo duomenyse dažniai.

Tegul $y = (y_1, y_2, \dots, y_n)^\top$ yra dažnių lentelės ląstelių, sunumeruotų kokia nors tvarka, dažnių vektorius. Pažymėkime $\mu = (\mu_1, \mu_2, \dots, \mu_n)^\top := \mathbf{E}y$ jo tikėtinų dažnių vektorių.

Logtiesinis modelis remiasi prielaida $\mu > 0$ (visi vidurkiai μ_i yra teigiami) ir apibrėžia μ natūrinį logaritmą, kaip tam tikro aiškinančiųjų kintamųjų z ir (naujų) neži-

nomų parametrų β rinkinio tiesinę kombinaciją:

$$\ln \mu_i = \sum_{j=1}^{k} z_{ij}\beta_j, \quad \forall i. \tag{2.4}$$

Matricinėje formoje

$$\ln \mu = \mathbf{Z}\beta, \quad \mathbf{Z} := (z_{ij}),\ \beta := (\beta_1, \ldots, \beta_k)^\top. \tag{2.5}$$

Sudarant ir interpretuojant logtiesinius modelius patogu naudotis *Mobiuso* formule.

Mobiuso formulė. Duotai aibei $J \subset [n]$ apibrėžkime atvaizdavimą $w = H_J(z)$, $H_J \colon \mathcal{Z} \to \mathcal{Z}$, tokiu būdu: $w_j = z_j,\ \forall j \in J$ ir $w_j = b_j,\ \forall j \in [n] \setminus J$, čia $b_j \in \mathcal{A}_j$ yra pasirinkta bazinė (arba atskaitos, angl. *reference*) būsena, bazinė alfabeto $\mathcal{A}_j$ raidė.

Tegu $\psi \colon \mathcal{Z} \to \mathbf{R}$ yra bet kokia funkcija. Tada Mobiuso formulė (tapatybė) jai užrašoma taip (žr. [50], [98])

$$\psi(z) = \sum_{K\colon K\subset[n]} \sum_{J\colon J\subset K} (-1)^{|K|-|J|}\, \psi(H_K(z)) =: \sum_{K\colon K\subset[n]} u_K(z_K), \tag{2.6}$$

$$u_K(z_K) := \sum_{J\colon J\subset K} (-1)^{|K|-|J|}\, \psi(H_K(z)), \quad z \in \mathcal{Z}, \tag{2.7}$$

čia $|J|$ žymi aibės J elementų skaičių (galią), o funkcijos $u_K(z_K)$ priklauso tik nuo $z_K := (z_j,\ j \in K)$. Funkcijos u_K virsta 0, jeigu kuriam nors $j \in K$ funkcija ψ nepriklauso nuo z_j. Pavyzdžiui, kai $n = 3$,

$$\begin{aligned}
\psi(z_1, z_2, z_3) &= \psi(b_1, b_2, b_3) + (\psi(z_1, b_2, b_3) - \psi(b_1, b_2, b_3)) + \\
&+ (\psi(b_1, z_2, b_3) - \psi(b_1, b_2, b_3)) + (\psi(b_1, b_2, z_3) - \psi(b_1, b_2, b_3)) + \\
&+ (\psi(z_1, z_2, b_3) - \psi(z_1, b_2, b_3) - \psi(b_1, z_2, b_3) + \psi(b_1, b_2, b_3)) + \\
&+ (\psi(z_1, b_2, z_3) - \psi(z_1, b_2, b_3) - \psi(b_1, b_2, z_3) + \psi(b_1, b_2, b_3)) + \\
&+ (\psi(b_1, z_2, z_3) - \psi(b_1, z_2, b_3) - \psi(b_1, b_2, z_3) + \psi(b_1, b_2, b_3)) + \\
&+ (\psi(z_1, z_2, z_3) - \psi(z_1, z_2, b_3) - \psi(z_1, b_2, z_3) - \psi(b_1, z_2, z_3) + \\
&+ \psi(z_1, b_2, b_3) + \psi(b_1, z_2, b_3) + \psi(b_1, b_2, z_3) - \psi(b_1, b_2, b_3)) =: \\
&=: u^\circ + u_1(z_1) + u_2(z_2) + u_3(z_3) + u_{12}(z_1, z_2) + u_{13}(z_1, z_3) + \\
&+ u_{23}(z_2, z_3) + u_{123}(z_1, z_2, z_3),
\end{aligned} \tag{2.8}$$

(čia trumpumo dėlei vietoje $u_{\{i,j,\ldots\}}$ rašoma $u_{ij\ldots}$), ir jeigu funkcija ψ nepriklauso,

tarkim, nuo pirmojo argumento, t.y. $\psi(z_1, z_2, z_3) \equiv \psi(b_1, z_2, z_3)$, tai

$$\begin{aligned}
\psi(b_1, z_2, z_3) &= \psi(b_1, b_2, b_3) + (\psi(b_1, b_2, b_3) - \psi(b_1, b_2, b_3)) + \\
&+ (\psi(b_1, z_2, b_3) - \psi(b_1, b_2, b_3)) + (\psi(b_1, b_2, z_3) - \psi(b_1, b_2, b_3)) + \\
&+ (\psi(b_1, z_2, b_3) - \psi(b_1, b_2, b_3) - \psi(b_1, z_2, b_3) + \psi(b_1, b_2, b_3)) + \\
&+ (\psi(b_1, b_2, z_3) - \psi(b_1, b_2, b_3) - \psi(b_1, b_2, z_3) + \psi(b_1, b_2, b_3)) + \\
&+ (\psi(b_1, z_2, z_3) - \psi(b_1, z_2, b_3) - \psi(b_1, b_2, z_3) + \psi(b_1, b_2, b_3)) + \\
&+ (\psi(b_1, z_2, z_3) - \psi(b_1, z_2, b_3) - \psi(b_1, b_2, z_3) - \psi(b_1, z_2, z_3) + \\
&+ \psi(b_1, b_2, b_3) + \psi(b_1, z_2, b_3) + \psi(b_1, b_2, z_3) - \psi(b_1, b_2, b_3)) = \\
&= \psi(b_1, z_2, z_3) = u^\circ + u_2(z_2) + u_3(z_3) + u_{23}(z_2, z_3). \qquad (2.9)
\end{aligned}$$

Pateiksime paprastą logtiesinio modelio pavyzdį dviejų požymių dažnių lentelei (žr. 2.1 lentelę).

Lentelė 2.1: Dviejų požymių A ir B kryžminė dažnių $y_{(i,j)}$ lentelė

		B				
	$y_{(i,j)}$	1	2	...	s	
A	1	$y_{(1,1)}$	$y_{(1,2)}$	...	$y_{(1,s)}$	$y_{(1,+)}$
	2	$y_{(2,1)}$	$y_{(2,2)}$	...	$y_{(2,s)}$	$y_{(2,+)}$
	·	·	·	...	·	·
	·	·	·	...	·	·
	·	·	·	...	·	·
	r	$y_{(r,1)}$	$y_{(r,2)}$	...	$y_{(r,s)}$	$y_{(r+)}$
		$y_{(+,1)}$	$y_{(+,2)}$	...	$y_{(+,s)}$	$y_{(+,+)} = N$

Pavyzdys. Tegu tiriami požymiai A ir B, $A \in \{1, 2\}$, $B \in \{1, 2, 3\}$. Numeravimo tvarka τ požymių A ir B dažnių lentelėje yra iš kairės į dešinę ir iš viršaus į apačią. y_ℓ yra stebėtas dažnis ląstelėje $\ell \in \mathcal{L}$, čia $\mathcal{L} := \{1, 2\} \times \{1, 2, 3\} = \{(i, j), i = 1, 2, j = 1, 2, 3\}$ su $N := |\mathcal{L}| = 6$.

Tada

$$y := \begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \\ y_5 \\ y_6 \end{pmatrix} = \begin{pmatrix} y_{(1,1)} \\ y_{(1,2)} \\ y_{(1,3)} \\ y_{(2,1)} \\ y_{(2,2)} \\ y_{(2,3)} \end{pmatrix}.$$

Tokiu pat būdu susiejami ir y vidurkiai μ. Kadangi $\mu_{(i,j)}$ yra dviejų argumentų funkcija, tai pritaikę Mobiuso tapatybę (2.6) gauname išdėstymą

$$\ln \mu_{(i,j)} = u^{\circ} + u^{A}(i) + u^{B}(j) + u^{AB}(i,j),$$

kuris vadinamas pilnuoju (angl. *saturated*) logtiesiniu modeliu. Bazinėmis reikšmėmis laikysime paskutiniąsias kategorijų reikšmes: 2 – požymiui A ir 3 – požymiui B. Todėl, kai $i = 2$, $u^{A}(i) = 0$, o kai $j = 3$, $u^{B}(j) = 0$. Be to, funkcijos $u^{AB}(2, j) = 0 \quad \forall j$, o $u^{AB}(i, 3) = 0 \quad \forall i$, t.y. nelygūs nuliui bus tik $u^{AB}(1, 1)$ ir $u^{AB}(1, 2)$. Tokiu būdu, pirmosioms dviems funkcijoms parametrizuoti užtenka po vieną parametrą, o likusioms dviems funkcijoms reikia poros parametrų. Atitinkama plano matrica formulėje (2.5) yra

$$\mathbf{Z} = \left(Z^{\circ}, Z_1^{A}, Z_1^{B}, Z_2^{B}, Z_{1,1}^{AB}, Z_{1,2}^{AB}\right).$$

Čia Z°, Z_1^{A}, Z_1^{B}, Z_2^{B}, $Z_{1,1}^{AB}$ ir $Z_{1,2}^{AB}$ – koeficientai atitinkamai prie parametrų u°, $u^{A}(1)$, $u^{B}(1)$, $u^{B}(2)$, $u^{AB}(1, 1)$ ir $u^{AB}(1, 2)$. Taigi, turime

$$Z_1^{A} = \begin{pmatrix} 1 \\ 1 \\ 1 \\ 0 \\ 0 \\ 0 \end{pmatrix}, \; Z_1^{B} = \begin{pmatrix} 1 \\ 0 \\ 0 \\ 1 \\ 0 \\ 0 \end{pmatrix}, \; Z_2^{B} = \begin{pmatrix} 0 \\ 1 \\ 0 \\ 0 \\ 1 \\ 0 \end{pmatrix} \quad \text{ir t.t.}$$

Vadinasi,

$$\mathbf{Z} = \begin{pmatrix} 1 & 1 & 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 1 & 0 & 1 \\ 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}.$$

Logtiesinis modelis, gaunamas iš pilnojo modelio eliminavus paskutinįjį narį, apibrėžia nepriklausomų požymių A ir B modelį:

$$\ln \mu_{(i,j)} = u^{\circ} + u^{A}(i) + u^{B}(j), \quad i = 1, 2,\ j = 1, 2, 3,$$

jo matrica

$$\mathbf{Z} = \left(Z^{\circ}, Z_1^A, Z_1^B, Z_2^B\right) = \begin{pmatrix} 1 & 1 & 1 & 0 \\ 1 & 1 & 0 & 1 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \end{pmatrix}.$$

Pateiksime formalų logtiesinio modelio apibrėžimą.

Tegul $\{y_{(\ell)}, \ell \in \mathcal{L}\}$ yra neneigiamų atsitiktinių dydžių rinkinys, $\mathcal{L}$ yra baigtinė indeksų aibė, turinti n elementų. Aibės $\mathcal{L}$ elementų numeravimo tvarką apibrėšime abipus vienareikšmiu atvaizdavimu $\tau\colon [n] \to \mathcal{L}$. Tada $y := (y_i, i \in [n])$ yra n-matis atsitiktinių dydžių $\{y_{(\ell)}, \ell \in \mathcal{L}\}$, surikiuotų tvarka τ, vektorius. Čia $y_i := y_{(\tau(i))}, i \in [n]$. Tarkime, kad $\mu := \mathbf{E}y > 0$, čia laikoma, kad visos operacijos, tiek palyginimo, tiek ir logaritmavimo, atliekamos pakomponenčiui.

Apibrėžimas. *Stebiniai y tenkina logtiesinį modelį su tiesiniu poerdviu $\mathcal{M} \subset \mathbf{R}^n$, jeigu* $\ln(\mu) \in \mathcal{M}$.

Logtiesinio modelio apibrėžime stebinių vektoriaus y skirstinys nėra pilnai nusakytas, nes apribojimai įvedami tik jo vidurkiui. Kokybinių požymių statistinėje analizėje y skirstinys paprastai būna Puasono skirstinių sandauga, polinominis skirstinys arba polinominių skirstinių sandauga. Visi 3 atvejai priklauso AT modelių klasei.

Logtiesinių modelių parinkimas ir statistinės išvados, kaip ir AT modelių, remiasi tikėtinumo funkcijų ir didžiausio tikėtinumo įvertinių (asimptotine) teorija ([4], taip pat žr. [98]). Logtiesinių modelių logtikėtinumo funkcija yra iškila, tai suteikia tam tikrų privalumų apskaičiuojant nežinomų parametrų įvertinius. Sukurti efektyvūs parametrų algoritmai yra realizuoti statistiniuose paketuose.

Viena iš pagrindinių kategorinių duomenų, taip pat ir logtiesinės analizės problemų yra gautų rezultatų suvokimas ir interpretavimas. Šiai problemai spręsti naudojami grafai, kurie leidžia sudaryti kokybinį sąryšių tarp kategorinių kintamųjų vaizdą. Interpretavimo problemas jau gana gerai iliustruoja trimačių kategorinių duomenų analizė.

2.3.3 Trimatės dažnių lentelės logtiesiniai modeliai

Tarkime, kad turime trijų kokybinių požymių, $A \in \{1, \ldots, m_A\}$, $B \in \{1, \ldots, m_B\}$ ir $C \in \{1, \ldots, m_C\}$, dažnių lentelę, sudarytą iš stebėtų tų požymių dažnumų. Pažymėkime y_{ijs} stebėtą požymių (A, B, C) kombinacijos (i, j, s) dažnį, o μ_{ijs} tegu žymi jo vidurkį: $\mu_{ijs} = \mathbf{E}y_{ijs}$. Pasinaudoję Mobiuso formulėmis (2.6)–(2.7) (taip pat žr. (2.8) bei (2.9)) gauname pilnąjį („prisotintą“, angl. *saturated*) logtiesinį modelį

$$\ln \mu_{ijs} = u^{\circ} + u_i^A + u_j^B + u_s^C + u_{ij}^{AB} + u_{is}^{AC} + u_{js}^{BC} + u_{ijs}^{ABC}. \tag{2.10}$$

Simboliškai jis žymimas [ABC]. Čia ir toliau interpretacijos patogumui bei taupumo sumetimais Mobiuso išskaidymo funkcijose u argumentų numerių aibes, nurodytas apatiniuose indeksuose, pakeitėme atitinkamų kintamųjų vardais ir perkėlėme iš apatinių indeksų į viršutinius, o argumentus nukėlėme į indeksus.

Siekiant dešinėje (2.6) formulės pusėje gauti parametrinę formą (2.4), funkcijos u parametrizuojamos. Joms parametrizuoti reikia tiek parametrų, kiek jos gali įgyti skirtingų nenulinių narių. Vadinasi, funkcijai u° parametrizuoti reikia vieno parametro, funkcijai u^A reikia $m_A - 1$, u^{AB} užtenka $(m_A - 1)(m_B - 1)$ parametrų ir t.t. Nesunku patikrinti, kad parametrų skaičius dešinėje (2.6) lygybės pusėje bus lygus $n_A n_B n_C$, t.y. parametrų skaičiui kairėje pusėje.

Įvedus papildomus apribojimus funkcijoms u, gaunami pilnojo modelio daliniai atvejai, kurie turi atitinkamą interpretaciją. Tokiu būdu gautų trimačių logtiesinių modelių specifikacijos ir simboliniai pažymėjimai pateikti 2.2 lentelėje.

Lentelė 2.2: Galimi logtiesiniai modeliai trijų požymių dažnių lentelėje

Nr.	Specifikacija	Žymėjimas
1	$\ln \mu_{ijs} = u^{\circ} + u_i^A + u_j^B + u_s^C$	$[A][B][C]$
2	$\ln \mu_{ijs} = u^{\circ} + u_i^A + u_j^B + u_s^C + u_{js}^{BC}$	$[A][BC]$
3	$\ln \mu_{ijs} = u^{\circ} + u_i^A + u_j^B + u_s^C + u_{is}^{AC}$	$[AC][B]$
4	$\ln \mu_{ijs} = u^{\circ} + u_i^A + u_j^B + u_s^C + u_{ij}^{AB}$	$[AB][C]$
5	$\ln \mu_{ijs} = u^{\circ} + u_i^A + u_j^B + u_s^C + u_{is}^{AC} + u_{js}^{BC}$	$[AC][BC]$
6	$\ln \mu_{ijs} = u^{\circ} + u_i^A + u_j^B + u_s^C + u_{ij}^{AB} + u_{js}^{BC}$	$[AB][BC]$
7	$\ln \mu_{ijs} = u^{\circ} + u_i^A + u_j^B + u_s^C + u_{ij}^{AB} + u_{is}^{AC}$	$[AB][AC]$
8	$\ln \mu_{ijs} = u^{\circ} + u_i^A + u_j^B + u_s^C + u_{ij}^{AB} + u_{is}^{AC} + u_{js}^{BC}$	$[AB][AC][BC]$
9	$\ln \mu_{ijs} = u^{\circ} + u_i^A + u_j^B + u_s^C + u_{ij}^{AB} + u_{is}^{AC} + u_{js}^{BC} + u_{ijs}^{ABC}$	$[ABC]$

Pateiksime kai kurių modelių interpretaciją:

1 modelis: požymiai A, B ir C yra tarpusavyje (statistiškai) nepriklausomi.

5 modelis: A, B ir C yra tarpusavyje priklausomi. Tačiau požymiai A ir B yra *sąlyginai nepriklausomi*, kai yra žinomos požymio C reikšmės. Kitaip tariant, A ir B tarpusavio priklausomumas pasireiškia tik per C.

8 modelis neturi paprastos interpretacijos nepriklausomumo ar sąlyginio nepriklausomumo terminais. Šiuo atveju šansų (galimybių) santykiai dvimatėse požymių A ir B lentelėse, kai duota likusio požymio C reikšmė, nuo tos reikšmės nepriklauso. Kadangi modelis simetrinis atžvilgiu A, B ir C, tai minėta savybė galios ir poroms A ir C, bei B ir C. Šis modelis nėra grafinis, jo grafas sutampa su 8 modelio grafu. Apie grafinius modelius rašoma 2.4 poskyryje.

9 modelis yra pilnasis modelis. Jis laikomas grafiniu.

Sudėtingesnių modelių pavyzdys ir aptarimas pateiktas 3.1.3 paragrafe.

Išsamus trimatės dažnių lentelės logtiesinių modelių aprašymas pateiktas [4] monografijoje (318–320 psl.).

2.3.4 Apibendrintasis logit modelis

Kai aiškinamasis kintamasis yra kategorinis ir gali įgyti daugiau kaip dvi skirtingas reikšmes, jo sąlyginiam skirstiniui modeliuoti taikomas *apibendrintasis*, arba *daugia-*

naris, logit (angl. *generalized logits*) *modelis*. Logistinę regresiją galima laikyti apibendrintojo logit modelio atskiru atveju, kai aiškinamasis kategorinis kintamasis gali įgyti tik dvi skirtingas reikšmes.

Tarkime, kad aiškinamojo kintamojo galimų reikšmių aibė yra $[m]$. Tegu $x(j) \in \mathbf{R}^k, j \in [n]$, yra stebėtos skirtingos aiškinamųjų kintamųjų X reikšmės, o $y_i(j)$ yra aiškinamojo kintamojo reikšmės i pasitaikymo duomenyse su $X = x(j)$ dažnis, $i \in [m], j \in [n]$. Taigi, turime duomenis $(x(j), y_i(j)), i \in [m], j \in [n]$.

Tegu $p_i(x)$ žymi sąlyginę tikimybę, kad aiškinamasis kintamasis įgis reikšmę i, kai $X = x$. Tuomet apibendintasis logit (AL) modelis apibrėžiamas lygtimis

$$\ln \frac{p_i(x)}{p_m(x)} = \alpha_i + x^\top \beta_i, \quad i = 1, \ldots, m-1,$$

kurios nusako aiškinamojo kintamojo reikšmės i *šansus* lyginant su bazine reikšme m (šiuo atveju bazine reikšme pasirinkta paskutinė kategorija).

AL modelyje kiekvienai kategorinio aiškinamojo kintamojo reikšmei, išskyrus bazinę, sudaroma atskira logit funkcija su savais nepriklausomais parametrais. Taigi, turime $m-1$ logit funkciją ir $(m-1)(k+1)$ nežinomų skaliarinių parametrų. Dalykinė parametrų interpretacija remiasi *šansų*, arba *galimybių, santykiu*. Iš (2.11) išplaukia, kad aiškinamojo kintamojo reikšmių i^* ir i šansų santykis, kai X reikšmė x keičiama reikšme x^*, yra lygus

$$\frac{p_{i^*}(x^*)}{p_i(x^*)} : \frac{p_{i^*}(x)}{p_i(x)} = \exp\{(x^* - x)^\top (\beta_{i^*} - \beta_i)\}.$$

AL modelis modeliuoja sąlyginį aiškinamojo kintamojo Y skirstinį, kai yra žinomos aiškinančiųjų kintamųjų X reikšmės. Apie patį X skirstinį jokių prielaidų nedaroma. Taigi, AL modelis yra klasikinių regresinės analizės modelių analogas.

Jeigu X taip pat yra kategoriniai, tai galima nagrinėti bendrą aiškinamojo ir aiškinančiųjų kintamųjų logtiesinį modelį. Kadangi AL modelyje X skirstinys nėra nusakytas, tai jis faktiškai yra aprašomas pilnuoju logtiesiniu modeliu. Papildžius AL modelį pilnuoju logtiesiniu X modeliu, gaunamas bendras logtiesinis modelis, ekvivalentus pradiniam AL modeliui tiek kokybiškai, tiek ir statistinių išvadų prasme.

Pavyzdys. Tegu X yra fiktyvusis aiškinantysis kintamasis, atitinkantis kokybinius

požymius B ir C, o tiriame požymius A, B ir C. Tada AL modelį galima užrašyti taip:

$$\ln \frac{p_a(b,c)}{p_m(b,c)} = \alpha_a + \beta^B_{ab} + \beta^C_{ac} + \beta^{BC}_{abc},$$

čia $X = x := (b,c)$, b yra B galima reikšmė, o c yra C galima reikšmė. Jeigu A gali įgyti tik dvi reikšmes, tai indeksą a galima praleisti, ir gauname:

$$\lambda_x := \ln \frac{p_1(x)}{1-p_1(x)} = \alpha_1 + \beta^B_b + \beta^C_c + \beta^{BC}_{bc}. \tag{2.11}$$

Galimos (2.11) modelio specifikacijos ir apibendrinto logit modelio atitinkamybė logtiesiniam modeliui pateikta 2.3 lentelėje (taip pat žr. [4]).

Lentelė 2.3: Atitinkamybė tarp apibendrinto logit ir logtiesinio modelio

ALm			Logtiesinis modelis
(0) $\lambda_x = \alpha_1$	$(-)$	$A = \emptyset$	$[A][BC]$
(1) $\lambda_x = \alpha_1 + \beta^B_b$	(B)	$A = B$	$[AB][BC]$
(2) $\lambda_x = \alpha_1 + \beta^C_c$	(C)	$A = C$	$[AC][BC]$
(3) $\lambda_x = \alpha_1 + \beta^B_b + \beta^C_c$	$(B + C)$	$A = B + C$	$[AB][AC][BC]$
(4) $\lambda_x = \alpha_1 + \beta^B_b + \beta^C_c + \beta^{BC}_{bc}$	$(B * C)$	$A = B + C + BC$	$[ABC]$

Modelis $[A][BC]$ reiškia, kad požymis A ir požymiai B, C yra tarpusavyje (statistiškai) nepriklausomi. Savo ruožtu, $[AB][BC]$ reiškia, kad požymiai A ir C yra tarpusavyje nepriklausomi, kai yra žinomos požymio B reikšmės. Sąveika tarp A ir C pasireiškia tik per B.

Apibendrinant, logtiesinis modelis aprašo *visų požymių tarpusavio sąveiką*, o AL modelis modeliuoja aiškinamojo kintamojo *sąlygines tikimybes*, kai duotos aiškinančiųjų kintamųjų reikšmės.

2.4 Grafiniai modeliai

2.4.1 Pagrindinės grafų teorijos sąvokos

Išskiriami du grafinių modelių tipai: orientuotieji (kryptiniai) (angl. *directed*) ir neorientuotieji (nekryptiniai) (angl. *undirected*).

Grafas G yra sutvarkyta pora (V, U), tokia, kad V yra baigtinė aibė, *viršūnių aibė*, o U yra visų nesutvarkytų porų iš V aibės $V^{(2)}$ poaibis, *briaunų* (neorientuotojo grafo atveju) arba *lankų aibė* (orientuotojo grafo atveju). [16], [50]

Dvi viršūnės x ir y yra *gretimos* (kaimyninės) grafo G viršūnės, jeigu $(x, y) \in U$, o viršūnės x ir y yra *incidentiškos* briaunai (x, y). [16]

Poaibis $A \subseteq V$, kurio visos viršūnės yra gretimos su visomis likusiomis, yra vadinamas pilnuoju poaibiu, t.y.

$$A \text{ pilnasis } \Leftrightarrow \forall \alpha, \beta \in A : (\alpha, \beta) \in U.$$

Jei pilnasis poaibis yra maksimalus su šita savybe, tai jis vadinamas *klika*:

$$C \text{ yra klika } \Leftrightarrow [C \text{ yra pilnasis ir } C^* \supset C \Rightarrow C^* \text{ nėra pilnasis}].$$

Grafo klikų aibė žymima $\mathcal{C}$.

V visada galima išskaidyti klikų sąjunga $V = C_1 \cup \ldots \cup C_k$. Tada $\mathcal{C}= (C_1, \ldots, C_k)$.

Grafas, turintis daugiausiai vieną kliką, yra *pilnasis grafas*.

Grafas (W, F) yra grafo (V, U) *pografis*, jeigu $W \subseteq V$ ir $F \subseteq U$. Jeigu $A \subseteq V$, jis indukuoja pografį

$$(A, U_A) = (A, \{(\alpha, \beta) \in U : \alpha \in A \wedge \beta \in A\}),$$

t.y. viršūnių aibė sutampa su aibe A, o briaunų (lankų) aibę U_A sudaro tos grafo G briaunos (lankai), kurių abu galai priklauso aibei A. [50]

Bet koks pilnasis grafo G pografis yra grafo G klika.

Gretimų briaunų (lankų) seka (v_1, v_2) (v_2, v_3) $(v_3, v_4) \ldots (v_{k-1}, v_k)$ vadinama *grandine* (keliu). Jei grandinės pirmoji ir paskutinė viršūnės sutampa, tai tokia grandinė vadinama *ciklu*. Kelio atveju – kontūru.

Grafas yra *jungusis*, jeigu bet kuri skirtingų viršūnių pora $\{x, y\}$ yra sujungta grandine. Maksimaliai jungus pografis yra grafo *jungioji komponentė*. Kitaip sakant, grafo $G = (V, U)$ jungioji komponentė – tai pografis, kurį indukuoja aibė A, sudaryta iš bet kurios grafo G viršūnės v ir visų tų viršūnių, į kurias galima nukeliauti iš viršūnės v. Jeigu grafas sudarytas iš vienos jungiosios komponentės, tai jis yra jungusis.

Jungusis grafas, neturintis ciklų, vadinamas *medžiu*.

Teiginys 4. *Šie tvirtinimai grafui G yra ekvivalentūs:*

a) G yra medis;

b) G yra minimalus jungusis grafas, t.y. G yra jungusis ir jeigu $(x, y) \in U$, tai $G - (x, y)$ yra nejungusis;

c) G yra maksimalus beciklis grafas, t.y. G yra beciklis ir jeigu x ir y yra negretimos grafo G viršūnės, tai $G + (x, y)$ gaunamas ciklas.

Du grafai yra *izomorfiniai*, jeigu yra jų viršūnių aibių atitiktis, kuri išlaiko gretimumą. Taigi $G = (V, U)$ yra izomorfinis $G' = (V', U')$, jeigu yra tokia bijekcija $\varphi : V \to V'$, kad, jei $(x, y) \in U$, tai $(\varphi(x), \varphi(y)) \in U'$. Jei grafai G ir H yra izomorfiniai, tai rašome $G \cong H$. [16]

Orientuotojo grafo atveju bijekcija φ turi išlaikyti ir orientaciją (kryptį).

Grafų sąjungą ir sankirtą galima apibrėžti taip:

$$\begin{aligned}(V_1, U_1) \cup (V_2, U_2) &= (V_1 \cup V_2, U_1 \cup U_2),\\ (V_1, U_1) \cap (V_2, U_2) &= (V_1 \cap V_2, U_1 \cap U_2).\end{aligned}$$ [50]

2.4.2 Tikimybiniai grafiniai modeliai

Tikimybiniai grafiniai modeliai apjungia tikimybių ir grafų teoriją. Tikimybių teorija yra reikalinga susieti atskiras sudėtingos sistemos dalis ir parinkti jai pagal stebėtus duomenis tikimybinį modelį. Grafai geriausiai tinka struktūros vidinių sąryšių pavaizdavimui ir yra ypač naudingi, modeliuojant aukštos eilės kintamųjų sąveikas. [61]

Sistemos (struktūros) elementai arba kintamieji tapatinami su grafo viršūnėmis. Dviejų viršūnių sujungimas reiškia tiesioginę priklausomybę tarp atitinkamų sistemos

elementų. Nesujungtos grafo viršūnės yra nepriklausomos arba sąlyginai nepriklausomos, kai yra žinomos kitų viršūnių būsenos. [4]

Suformuluosime tai tiksliau. Tegu duotas grafas ir jo visos viršūnės suskirstytos į 3 aibes $\mathcal{A}$, $\mathcal{B}$, $\mathcal{C}$. Jeigu bet kuris kelias iš viršūnės, priklausančios aibei $\mathcal{A}$, į viršūnę, priklausančią aibei $\mathcal{B}$, eina per viršūnę, priklausančią aibei $\mathcal{C}$, tai požymiai, atitinkantys viršūnes iš $\mathcal{A}$, ir požymiai, atitinkantys viršūnes iš $\mathcal{B}$, yra *sąlyginai nepriklausomi*, kai žinomos požymių, atitinkančių viršūnes iš $\mathcal{C}$, reikšmės.

Natūralu, kad grafiniais modeliais vadinami tokie modeliai, kuriuos galima pavaizduoti grafiškai. Deja, tik dalis logtiesinių modelių yra grafiniai. Pavyzdžiui, trimatėje 2.2 lentelėje 8-as modelis nėra grafinis. Tokiu atveju ieškomas mažiausias (mažiausiai parametrų turintis) grafinis modelis, apimantis parinktą logtiesinį modelį.

Mažiausio grafinio modelio algoritmas: įtraukiami visi kintamieji (kiekvieną jų atitinka grafo viršūnė); jei parinktame logtiesiniame modelyje yra įtraukta n-tos eilės sąveika, tai automatiškai įtraukiamos visos žemesnės eilės sąveikos, ir priešingai, jei yra įtrauktos visos žemesnės eilės sąveikos, tai įtraukiama ir aukštesnės eilės sąveika (plačiau žr. [92], taip pat [50]).

Priklausomybė tarp kintamųjų gali būti nusakoma parametriniu sąlyginiu skirstiniu arba dar kitaip vadinama potencialine funkcija (angl. *potential function*). Grafo jungčių aibė ir sąlyginiai skirstiniai drauge apibrėžia bendrą visų grafo kintamųjų tikimybinį skirstinį (angl. *joint probability distribution*), nusakantį visos sistemos funkcionavimą. Paprastai grafo jungčių aibė vadinama grafo struktūra, arba topologija, o sąlyginių skirstinių parametrai – tiesiog grafo parametrais.

Daugelis statistikoje naudojamų modelių yra specialūs grafinių modelių atvejai. Kryptiniams grafiniams modeliams priskiriami Bajeso tinklai, priežastiniai modeliai (angl. *causal models*) ir kt. Nekryptiniams modeliams, kurie dar vadinami Markovo tinklais (angl. *Markov Networks*) arba Markovo atsitiktiniais laukais (angl. *Markov random fields*), priskiriami logtiesiniai modeliai. Disertacijoje yra naudojami nekryptiniai statistiniai grafiniai modeliai.

Logtiesinių modelių ryšį su Markovo atsitiktiniais laukais (MAL) pastebėjo ir aprašė J. N. Darroch, S. L. Lauritzen ir T. P. Speed (1980; [23]). Kadangi diskretieji MAL yra

glaudžiai susiję su grafiniais modeliais bei logtiesinių modelių interpretacija nepriklausomumo ir sąlyginio nepriklausomumo terminais, čia pateiksime trumpą įvadą.

Diskretieji Markovo atsitiktiniai laukai

Tegul $Y := \{Y(v),\ v \in V\}$ yra duotas atsitiktinių dydžių $Y(v)$ su reikšmėmis baigtinėje būsenų aibėje $\mathcal{A}_v$ rinkinys ($v \in V$). Čia V žymi baigtinę indeksų aibę, $n := |V|$ yra aibės V elementų skaičius, $\mathcal{A}^V := \otimes_{v\in V}\mathcal{A}_v$.

Indeksų aibėje V įvesime ekvivalentumo santykį, kurį nusako neorientuotas grafas $(V, \mathcal{B})$ be kilpų. Ekvivalentumo santykis yra susietas su aibės V poaibių, sudarytų iš kaimyninių elementų, rinkiniu $\mathcal{U} := \{U_\ell,\ \ell \in V\}$. Aibėje $U_\ell \subset V$ yra išvardintos visos viršūnei ℓ kaimyninės grafo viršūnės, t.y. viršūnės, turinčios bendrą briauną su ℓ.

Pažymėkime:

$$Y(U) := \{Y(u),\ u \in U\}, \quad W_{-v} := W \setminus \{v\}, \quad v \in W \subset V.$$

Apibrėžimas. *Atsitiktinis laukas $Y = Y(V)$ vadinamas (diskrečiuoju) Markovo atsitiktiniu lauku (MAL) atžvilgiu grafo $(V, \mathcal{B})$ (arba Markovo atsitiktiniu lauku su kaimynų sistema $\mathcal{U}$), jeigu su visais $v \in V$ atsitiktinis dydis $Y(v)$ ir atsitiktiniai dydžiai $Y(V_{-v} \setminus U_v)$ yra sąlyginai nepriklausomi, kai yra žinomos dydžių $Y(U_v)$ reikšmės:*

$$Y(v) \perp Y\left(V_{-v} \setminus U_v\right) \mid Y\left(U_v\right).$$

Vadinasi, Markovo atsitiktiniam laukui Y su visais $a \in \mathcal{A}^V$ galioja lygybė

$$\mathbf{P}\left\{Y(v) = a(v) \mid Y\left(V_{-v}\right) = a\left(V_{-v}\right)\right\} = \mathbf{P}\left\{Y(v) = a(v) \mid Y\left(U_v\right) = a\left(U_v\right)\right\}. \quad (2.12)$$

Iš (2.12) matome, kad lokalūs sąlyginiai skirstiniai

$$p_v(a(v) \mid a(U_v)) := \mathbf{P}\left\{Y(v) = a(v) \mid Y\left(U_v\right) = a\left(U_v\right)\right\}, \quad a \in \mathcal{A}^V,$$

nusako pilnąsias sąlygines tikimybes (angl. *full conditional distributions*)

$$p_v\left(a(v) \mid a(V_{-v})\right) := \mathbf{P}\left\{Y(v) = a(v) \mid Y\left(V_{-v}\right) = a\left(V_{-v}\right)\right\}.$$

Iš Brooko lemos (žr. [98]) išplaukia, kad lokalieji sąlyginiai skirstiniai visiškai nusako ir bendrą $Y(V)$ skirstinį, jeigu tik išpildyta teigiamų tikimybių sąlyga:

$$\mathbf{P}\{Y = a\} > 0 \quad \forall\ a \in \mathcal{A}^V. \quad (2.13)$$

Gibso skirstinys

Gibso skirstinys (fizikoje) yra apibrėžiamas per potencialų rinkinį $\{u_K(a),\ K \in \mathcal{K}\}$

$$P_G(a) := Z^{-1} \exp\{-\sum_{K \in \mathcal{K}} u_K(a)\}, \quad a \in \mathcal{A}^V, \tag{2.14}$$

čia Z yra normuojantis daugiklis (*dalinimo funkcija*).

Tarkime, kad potencialai tenkina sąlygą

$$u_K(a) = u_K(z) \quad \forall\ a, z \in \mathcal{A}^V,\ a(K) = z(K). \tag{2.15}$$

Tegu $\mathcal{K}$ yra rinkinys V poaibių, sudarytų iš grafo $(V, \mathcal{B})$ klikų viršūnių.

Apibrėžimas. *Skirstinį, nusakytą (2.14) su potencialų funkcijomis u_K, $K \in \mathcal{K}$, kurios tenkina (2.15) sąlygą, vadinsime Gibso skirstiniu su lokalios sąveikos aplinkomis $\mathcal{U}$ arba grafu $(V, \mathcal{B})$.*

Teiginys 5 (Hammersley-Clifford, 1971; žr. [50], [92])**.** *Tarkime, kad atsitiktinio lauko Y skirstinys P_Y tenkina teigiamų tikimybių sąlygą (2.13). Tuomet P_Y yra Markovo atsitiktinis laukas su aplinkų sistema $\mathcal{U}$ tada ir tik tada, kai jis yra Gibso skirstinys su lokalios sąveikos aplinkomis $\mathcal{U}$ ir kokiu nors potencialo funkcijų $\{u_K(a),\ K \in \mathcal{K}\}$, tenkinančių (2.15) sąlygą, rinkiniu.*

2.5 Struktūriniai skirstiniai

Tegu S – fiksuota objektų arba tekstinės informacijos šaltinių, kurie yra laikomi statistiškai nepriklausomais, populiacija. Mus domina ne tekstų turinys ar semantika, o tik juos sudarantys žodžiai, tiksliau – skirtingos tų tekstų žodžių formos, ir jų pasitaikymo juose dažnis. Šiame darbe laikoma, kad visi žodžiai, kurie rašomi vienodai, turi tą pačią žodžio formą (angl. *word type*).

Tegu W_s žymi visų (skirtingų) žodžių formų aibę šaltinyje $s \in S$ (to šaltinio žodyną) ir tegu $V_s := |W_s|$ yra to žodyno dydis (žodyno apimtis). Tariama, kad visi žodynai W_s yra bendro žodyno $\mathcal{W}$ poaibiai. Taigi duomenys, kuriuos mes nagrinėjame, yra $\{(y_w(s), x_w(s)), w \in \mathcal{W}, s \in S\}$, čia $y_w(s)$ yra žodžių formų $w \in \mathcal{W}$ dažnis šaltinyje $s \in S$, $x_w(s)$ yra atitinkamų aiškinamųjų kintamųjų vektorius, kuris suteikia papildomos informacijos tiek apie žodžio formą $w \in \mathcal{W}$, tiek ir apie tos formos šaltinį $s \in S$.

Kiekybinėje lingvistikoje teigiama, kad žodynas $\mathcal{W}$ iš esmės yra neaprėžtas (begalinis; žr., pvz., [10], [46]). Kad galėtume tai aprašyti formaliai, įveskime asimptotinį parametrą $M \to \infty$, kuris nusako nagrinėjamų tekstinių dokumentų bendrą apimtį. Pavyzdžiui, jeigu $\mathcal{W} = \cup_{s \in S} W_s$, parametras M gali būti apibrėžtas kaip $|S|$, t.y. šaltinių rinkinyje S kiekis (kita alternatyva yra nagrinėjama žemiau). Tada (teoriniuose išvedžiojimuose) reikalaujama, kad

$$V = V^{(M)} := V(\mathcal{W}_M) \to \infty, \quad M \to \infty. \tag{2.16}$$

Aibės W_s yra laikomos V_s dydžio imtimis iš kokios nors (begalinės) žodžių, generuotų tam tikru stochastiniu mechanizmu, superpopuliacijos (plg. [27], [13]). Lingvistinėje literatūroje yra pasiūlyti ir nagrinėjami (aptariami) įvairūs žodžių superpopuliacijos (šitos sąvokos neminint) modeliai. Faktiškai bet kuris iš lingvistikoje taikomų tikimybinių modelių (pvz., Zipfo-Mandelbroto, Julo-Saimono ir kiti; žr. [8], [10], [44]) gali būti laikomas superpopuliacijos modeliu.

Nemažai empirinių tyrimų, pradedant nuo M. Estoupo (1916) (nuoroda paimta iš [44]), rodo, kad beveik pusė žodžių yra sutinkami tekstyne tik po vieną kartą (tokie žodžiai vadinami *hapax legomena*). Pagal klasikinę rekomendaciją (žr., pvz., [4], p. 396) daugumoje dažnių lentelės ląstelių stebėtas dažnis turėtų būti mažiausiai 5. Jei šis reikalavimas yra neišpildytas, klasikinių testų statistikų skirstinių χ^2 aproksimacijos

tikslumas gali būti nepakankamas standartinei statistinei analizei atlikti. Tada sakoma, kad kategoriniai duomenys yra išretinti (arba stebėtų dažnių lentelė yra išretinta). Taigi, tekstyno žodžių kiekio duomenys yra išretinti. Jeigu galioja (2.16) prielaida, sakoma, kad turime išretinimo asimptotikos schemą (angl. *sparse asymptotics*) (žr. [15]) arba didelio kiekio retų įvykių (angl. *large number of rare events*, LNRE) asimptotikos schemą (žr. [43]).

Toliau yra aprašomi du pagrindiniai išretintų kategorinių duomenų modeliai.

2.5.1 Latentinis skirstinys

Paprastumo dėlei tariame, kad yra tik vienas tekstinės informacijos šaltinis s, todėl toliau nuorodą į šaltinį galime praleisti. Tegu žodžio formos bendrajame žodyne $\mathcal{W}$, kurio dydis $V = V^{(M)}$, yra išdėstytos tam tikra tvarka r. Gauname žodžių formų vektorių $\underline{w} = \underline{w}_V(r) := (w_1, \ldots, w_V)$. Ta pačia tvarka išdėstomi stebėtieji ir vidutiniai (tikėtini) žodžio formų dažniai, atitinkamai y ir $\mu := \mathbf{E}y$, bei aiškinantieji kintamieji x (taip pat ir kiti susiję objektai): $\underline{y} := (y_1, \ldots, y_V)$, $\underline{\mu} := (\mu_1, \ldots, \mu_V)$ ir $\underline{x} := (x_1, \ldots, x_V)$.

Vienas iš paprasčiausių būdų apeiti išretintų dažnių problemą yra tarti, kad $\underline{\mu}$ yra nusakomas latentine pasiskirstymo funkcija F intervale $[0, 1]$ ir tokiomis lygybėmis:

$$\mu_i = \mu_+ \left(F(t_i) - F(t_{i-1})\right), \quad \mu_+ := \sum_{i=1}^{V} \mu_i,$$

čia $t_i := i/V$, $i = 0, 1, \ldots, V$ (plg. [15], [43]). Ši schema yra dažnai naudojama ekonometriniuose ranginių kintamųjų tyrimuose arba rangavimo uždaviniuose (žr. [2]). Paprastai daroma prielaida, kad egzistuoja aprėžtas ir glodus latentinio skirstinio tankis f, $f(u) = dF(u)/du$. Pastaroji prielaida reiškia, kad tikėtini dažniai $\mu_i = O\left(\mu_+/V\right)$, $V \to \infty$. Tegu

$$\begin{aligned} \widehat{V}_m &= \widehat{V}_m(s) = \sum_{i=1}^{V} \mathbb{I}\{y_i = m\}, \quad m = 0, 1, \ldots, \\ \widehat{V}_+ &= \widehat{V}_+(s) := \sum_{i=1}^{V} \mathbb{I}\{y_i > 0\} = \sum_{m=1}^{\infty} \widehat{V}_m \end{aligned}$$

yra atitinkamai žodžių formų, stebėtų lygiai m kartų, skaičius ir bendras visų realiai stebėtų žodžių formų (nagrinėjamame šaltinyje s) skaičius. Čia ir toliau $\mathbb{I}(E)$ yra įvykio (aibės, ryšio) E indikatorius. Vadinasi, nagrinėjamame (šaltinio s) tekstyne

yra $\widehat{V}_+$ skirtingų žodžių formų ir y_+ teksto žodžių, $y_+ := \sum_{i=1}^{V} y_i$. Pastarąjį skaičių lingvistinėje literatūroje įprasta žymėti N. Taigi, yra patogu asimptotiniu parametru imti $M := \mu_+ = \mathbf{E}y_+$, kadangi jis gali būti naudojamas taip pat ir tuo atveju, kai šaltinių skaičius yra fiksuotas, bet didėja patys šaltinių tekstai. Tada $V = V^{(M)} \to \infty \Longleftrightarrow M = M^{(V)} \to \infty$.

Duomenų išretinimą galima apibūdinti įvairiais dydžiais, pavyzdžiui,

$$\rho_A = \rho_A(M) := \frac{M}{V^{(M)}}, \quad \rho_1 = \rho_1(M) := \frac{\mathbf{E}\widehat{V}_1}{\mathbf{E}\widehat{V}_+} \tag{2.17}$$

yra, atitinkamai, vidutinis tikėtinas žodžių formų dažnis ir santykinis tikėtinas tik kartą pasitaikančių žodžių formų (t.y. *hapax legomena*) skaičius tekstiniuose duomenyse. Dydį ρ_1 įvedė E. V. Khmaladze [43], siekdamas apibrėžti LNRE modelius (schemas).

Apibrėžimas (plg. [43])**.** *Sakoma, kad stebėti dažniai* $\underline{y}$ *tenkina LNRE modelį, jeigu*

$$\liminf_{M\to\infty} \rho_1(M) > 0, \quad \mathbf{E}\widehat{V}_+ \to \infty. \tag{2.18}$$

LNRE modelį, taip pat ir sąlygą $\rho_A(N) = O(1)$, galima laikyti formaliu išretintų kategorinių duomenų (modelio) apibrėžimu.

Pastaba. Latentinio skirstinio modelis remiasi prielaida, kad egzistuoja (latentinis) ranginis atsitiktinis kintamasis r (t.y. kintamasis, kuriam yra prasmingos palyginimo operacijos). Vadinasi, jis tiesiogiai nepritaikomas nominaliesiems duomenims, ypač žodžių formoms tekstiniuose dokumentuose. Siekiant išspręsti šią problemą galima įvesti tam tikrą fiktyviai sutvarkytą kintamąjį, kokiu nors būdu susijusį su nagrinėjamais nominaliaisiais duomenimis. Žodžių kiekių duomenims natūralūs (tradiciniai) fiktyvūs ranginiai kintamieji yra stebėtų žodžių formų dažnių tam tikrame tekstyne arba žodžių formų, sutvarkytų didėjimo (mažėjimo) tvarka pagal jų dažnius tekstyne, rangai. Tokiu būdu sudaryto ranginio kintamojo r (galbūt jį atitinkamai normavus) latentinis skirstinys vadinamas *struktūriniu skirstiniu*. Formalus struktūrinio skirstinio apibrėžimas yra pateiktas ir aptariamas kitame paragrafe. Deja, žodžių formas pakeičiant jų rangais prarandamas žodžių formų vienetinumas, nes tą patį dažnį tekstyne gali turėti ne viena žodžių forma.

2.5.2 Struktūrinis skirstinys

Kartais yra natūralu manyti, kad statistinės išvados turi turėti tam tikras simetrijos savybes, kitaip tariant, yra invariantiškos atžvilgiu tam tikrų transformacijų. Nagrinėjant žodžių kiekių duomenis, galima tarti, kad patys žodžių formų pavadinimai (indentifikatoriai) w, pagal kuriuos jos buvo surikiuotos žodyne, yra neinformatyvūs ir tyrėjo nedomina. Vadinasi, šiuo atveju, taikomos statistinės procedūros turėtų būti invariantiškos atžvilgiu žodžių formų tvarkos žodyne bet kokių keitinių.

Tai reiškia, kad bet kuri „populiacijos“ $\mathcal{W}$ kiekybinė charakteristika gali būti pilnai aprašoma jos empiriniu skirstiniu. Atskiru atveju, žodžių formų w iš $\mathcal{W}$ tikėtini dažniai yra pilnai aprašyti jų empirine pasiskirstymo funkcija (edf)

$$\hat{F}(u) = \frac{1}{V}\sum_{i=1}^{V} \mathbb{I}\{\mu_i \le u\}, \quad u \ge 0.$$

Empirinė pasiskirstymo funkcija $\hat{F}$ yra vadinama *empiriniu struktūriniu skirstiniu.* Galima tikėtis, kad $\hat{F}$, galbūt atitinkamai pakeitus mastelį, konverguoja (kai $V = V^{(M)} \to \infty$) į pasiskirstymo funkciją F.

Apibrėžimas (plg. [28])**.** *Tarkime, kad edf $\hat{F}(\rho t)$ su mastelio daugikliu $\rho = \rho(M)$ silpnai konverguoja į paskirstymo funkciją F, kai $M \to \infty$. Tada F vadinama tikėtinų dažnių $\underline{\mu}$ (arba tiesiog žodyno $\mathcal{W}$) struktūriniu skirstiniu su mastelio daugikliu ρ.*

Puasono ėmimo schemoje y reikšmė, atsitiktinai su vienodomis tikimybėmis pasirinkta iš stebėtų dažnių $\{y_1, \dots, y_V\}$, tenkina

$$[y \mid \lambda] \stackrel{\mathcal{L}}{=} Poisson(\lambda), \quad \lambda \stackrel{\mathcal{L}}{=} \hat{F} \quad (\stackrel{\mathcal{L}}{=} \text{ apibrėžia skirstinio dėsnį}), \tag{2.19}$$

čia $Poisson(\lambda)$ žymi Puasono skirstinio dėsnį su intensyvumo parametru arba vidurkiu λ, t.y.

$$\mathbf{P}(y_i = k \mid \lambda = \mu_i) = \Pi_k(\mu_i) := \mu_i^k \mathrm{e}^{-\mu_i}/(k!),\ k = 0, 1, \dots.$$

Jei seka $\rho^{-1}\underline{\mu}_V := (\rho^{-1}\mu_1, \dots, \rho^{-1}\mu_V)$ yra nepriklausomų ir vienodai pasiskirsčiusių atsitiktinių dydžių su bendru skirstiniu F seka, tai F yra $\mathcal{W}$ su mastelio daugikliu ρ struktūrinis skirstinys. Tada žodžių skaičiaus skirstinys, apibrėžtas (2.19), gali būti aproksimuotas Puasono skirstinių mišinio modeliu

$$[y \mid \lambda] \stackrel{\mathcal{L}}{=} Poisson(\rho\lambda), \quad \lambda \stackrel{\mathcal{L}}{=} F.$$

Puasono skirstinių mišinių taikymus lingvistikoje nagrinėjo K. W. Church ir W. A. Gale [19].

Empiriniai tyrimai (žr., pvz., [10], [26]) rodo, kad realius duomenis geriau aprašo netikriniai struktūriniai skirstiniai. S. Evertas [26] nagrinėjo Zipfo dėsnį LNRE modelyje. Šio skirstinio netikrinis struktūrinis tankis f yra

$$f(z) = z^{-\tau-1}, \quad 0 < z \le b < \infty,\ \tau \in (0,1).$$

Parametras τ nusako Zipfo eksponentę $\alpha = 1/\tau$.

Siekiant išvengti netikrinių skirstinių, empirinį struktūrinį skirstinį galima aproksimuoti baigtiniu Zipfo-Mandelbroto (žr. [26]) arba nukirstuoju Pareto skirstiniais $F_\varrho(\cdot \mid \tau, \delta)$ su tankiu

$$f_\varrho(z \mid \tau, \delta) := \varrho^{-1} f_1(\varrho^{-1} z \mid \tau, \delta), \quad 0 < \delta < 1,\ \tau \in (0, \infty), \tag{2.20}$$

$$f_1(z) := c_1 z^{-\tau-1} \mathbb{I}\{\delta < z < 1\}, \tag{2.21}$$

čia mastelio parametras ϱ, apatinis f_1 atramos rėžis δ, vadinasi, ir normuojantis daugiklis $c_1 = c_1(\tau, \delta) := \tau(\delta^{-\tau} - 1)^{-1}$ gali priklausyti nuo asimptotinio parametro M. Čia daroma prielaida, kad $\tau > 0$ yra fiksuotas,

$$\lim_{M\to\infty} \varrho(M) = b_0 \in (0, \infty], \quad \lim_{M\to\infty} \delta(M)\varrho(M) = a_0 \in [0, b_0). \tag{2.22}$$

Taikymuose prielaida, kad aproksimuojančio struktūrinio skirstinio atrama turi teigiamą apatinį rėžį, iš tikrųjų nėra apribojimas, kadangi neįmanoma įvertinti labai retų žodžių formų, kurių pasitaikymo tikimybė yra žemiau tam tikros teigiamos ribos (kuri priklauso nuo M), tikėtinumo.

Toliau šiame paragrafe remsimės S. Everto ([26], taip pat žr. [46]) argumentais apie struktūrinio skirstinio aproksimacijų (2.20), (2.21) ryšį su Herdano-Hipso ir Zipfo dėsniais. Skirtingai negu darbe [26], kuriame $N = y_*$ buvo laikomas fiksuotu, čia N yra atsitiktinis.

Tegul $\mathbf{E}_\rho$ ($\mathbf{P}_\rho$) žymi teorinį vidurkį (atitinkamai, tikimybę) atžvilgiu tankio f_ϱ. Tada žodžių formų, pasitaikiusių tekstyne m kartų, vidutinis skaičius

$$\mathbf{E}_\varrho \widehat{V}_m = V \mathbf{P}_\varrho\{y = m\} = V \int_0^\infty \Pi_m(u) \mathrm{d}F_\varrho(u) = V \frac{c_1 \varrho^\tau}{m!} \int_{\delta\varrho}^{\varrho} u^{m-\tau-1} \mathrm{e}^{-u} \mathrm{d}u. \tag{2.23}$$

Pažymėkime nepilną gama funkciją $\Gamma(u;\beta)$ ir tarkime, kad $m \geq 1$ ir $\tau \in (0,1)$. Tada iš (2.22) ir (2.23) išplaukia, kad

$$\frac{\mathbf{E}_\varrho \widehat{V}_m}{V} \sim \frac{c_1 \rho^\tau (\Gamma(b_0; m-\tau) - \Gamma(a_0; m-\tau))}{\Gamma(m+1)}. \tag{2.24}$$

Analogiškai, vidutiniam žodyno dydžiui, t.y. vidutiniam žodžių formų, pasitaikiusių tekstyne bent kartą, skaičiui, turime

$$\frac{\mathbf{E}_\varrho \widehat{V}_+}{V} = \mathbf{P}_\varrho\{y > 0\} = \int_0^\infty \left(1 - \mathrm{e}^{-u}\right) \mathrm{d}F_\varrho(u) \sim c_1 c_0 \rho^\tau \frac{c_0 \tau (\delta\rho)^\tau}{1-\delta^\tau} \tag{2.25}$$

su

$$c_0 = c_0(a_0, b_0, \tau) := \int_{a_0}^{b_0} u^{-(\tau+1)} \left(1 - \mathrm{e}^{-u}\right) \mathrm{d}u \sim \int_{\delta\varrho}^{\varrho} \frac{1 - \mathrm{e}^{-u}}{u^{\tau+1}} \mathrm{d}u.$$

Reikia pastebėti, kad $a_0 = 0, b_0 = \infty$ duoda $c_0 = \Gamma(1-\tau)/\tau$. Žodyno $\mathcal{W}$, generuoto pagal aproksimuojantį struktūrinį skirstinį $F_\varrho(\cdot \mid \tau, \delta)$, apibrėžtą (2.20), (2.21), vidutinis tikėtinas dažnis (2.17) yra

$$\rho_A = \frac{\mathbf{E}_\varrho y_+}{V} = \mathbf{E}_\varrho y = \varrho \int_0^\infty u \mathrm{d}F_1(u) = c_1 \varrho \int_\delta^1 \frac{\mathrm{d}u}{u^\tau} = \frac{c_1 \varrho \,(1 - \delta^{1-\tau})}{1-\tau}. \tag{2.26}$$

Iš (2.17), (2.25) ir (2.26) seka, kad

$$\mathbf{E}_\varrho \widehat{V}_+ \sim c_0 \frac{1-\tau}{(1-\delta^{1-\tau})} M \varrho^{\tau-1}. \tag{2.27}$$

Tarkime, kad $\varrho \sim c_\varrho M^\beta,\ c_\varrho > 0,\ \beta \in [0,1]$. Tada iš (2.27) išvedamas laipsninis dėsnis (angl. *power law*)

$$\log\left(\mathbf{E}_\varrho \widehat{V}_+\right) = \log\left(\frac{c_0(1-\tau)\, c_\varrho^{\tau-1}}{1-\delta^{1-\tau}}\right) + (1 - \beta + \beta\tau) \log M + o(1). \tag{2.28}$$

Kai $M \to \infty$, neformaliai pritaikę didžiųjų skaičių dėsnį gauname, kad $\widehat{V}_+ \sim \mathbf{E}_\varrho \widehat{V}_+$ ir $N = y_+ \sim \mathbf{E}_\varrho y_+ = M$. Vadinasi, (2.28) išraiška gali būti interpretuojama kaip apytikslis ryšys tarp stebėtų žodžių formų skaičiaus $\widehat{V}_+$ ir stebėtų teksto žodžių skaičiaus N dideliame tekstyne

$$\log\left(\widehat{V}_+\right) \approx \text{const} + (1 - \beta + \beta\tau) \log N.$$

Ši išraiška yra žinoma kaip Herdano dėsnis (žr. [33]) kiekybinėje lingvistikoje ir kaip Hipso dėsnis (žr. [32]) informacijos paieškoje.

Atsižvelgiant į (2.24) ir (2.25), žodyno $\mathcal{W}$, generuoto pagal (2.20), (2.21), santykinių dažnių spektras

$$\frac{\mathbf{E}_\varrho \widehat{V}_m}{\mathbf{E}_\varrho \widehat{V}_+} \sim \frac{\Gamma(b_0; m-\tau) - \Gamma(a_0; m-\tau)}{c_0(a_0, b_0, \tau)\Gamma(m+1)} \tag{2.29}$$

yra (asimptotiškai, kai $M \to \infty$) nepriklausomas nuo vidutinio žodyno dydžio M. Taigi, išretinimo sąlyga (2.18) yra išpildyta.

Tarkime, kad $a_0 = 0, b_0 = \infty$. Žinomos aproksimacijos $\log\left(\Gamma(t+h)/\Gamma(t)\right) = h\log(t+h) + O\left(t^{-1}\right)$, $t \to \infty$, dėka (2.29) reiškia (labai dideliems M)

$$\log\left(\mathbf{E}_\varrho \widehat{V}_m\right) = \log\left(\mathbf{E}_\varrho \widehat{V}_+\right) - \log\left(\Gamma(1-\tau)/\tau\right) - (1+\tau)\log(m) + O\left(m^{-1}\right), \quad m \to \infty.$$

Ši išraiška gali būti laikoma teoriniu Zipfo antrojo dėsnio variantu:

$$\log\left(\widehat{V}_m\right) \approx \log(\widehat{V}_+) - \log\left(\Gamma(1-\tau)/\tau\right) - (1+\tau)\log(m). \tag{2.30}$$

Šioje išraiškoje žodžių formų rangai r, naudojami Zipfo pirmajame dėsnyje, yra pakeisti tų formų stebėtu dažniu m (žr. [46], p. 63).

Kiti struktūrinių skirstinių parametriniai modeliai aprašyti [8], [10], taip pat žr. [44], [26], [13] ir ten esančius literatūros šaltinius. Baigtinis Zipfo-Maldebroto modelis yra vienas tinkamiausių modelių realiuose taikymuose.

2.6 Bajeso metodologija

Bajeso statistinės analizės metodologija yra paremta aprioriniu skirstiniu ir tikimybių perskaičiavimu pagal Bajeso formulę. Bajeso statistikai skirta daug vadovėlių ir straipsnių, paminėsime tik monografiją [20], skirtą Bajeso metodams kategorinių duomenų analizėje, bei apžvalgą [5].

Tegu $Y_N = \{y_1, \ldots, y_N\}$ yra dydžio N paprastoji atsitiktinė imtis, a. d. $y := y_1$ skirstinys turi tankį $f(t;\theta)$ atžvilgiu duoto σ-baigtinio mato (paprastai Lebego arba skaičiuojančiojo mato). Čia $\theta \in \Theta$ yra k-matis nežinomas parametras. Bajeso metodologija remiasi prielaida, kad nežinomas parametras θ yra atsitiktinis dydis su reikšmėmis srityje Θ ir jų skirstiniu Π. Jis vadinamas *aprioriniu* skirstiniu ir atspindi tyrėjo apriorinę informaciją apie θ. Paprastai laikoma, kad Π yra absoliučiai tolydus. Tegu π žymi apriorinio skirstinio tankį.

Taigi, atskiro stebinio Bajeso modelis nusakomas formulėmis $[y \mid \theta] \stackrel{\mathcal{L}}{=} f(\cdot \mid \theta), \quad \theta \stackrel{\mathcal{L}}{=} \Pi$.

Aposterioriniu skirstiniu vadinamas a. d. θ sąlyginis pasiskirstymas, kai žinomos a. d. y stebinių Y_N reikšmės. Aposteriorinio skirstinio tankis surandamas remiantis Bajeso formule:

$$\pi(\theta \mid Y_N) = \frac{L_N(\theta \mid Y_N)\pi(\theta)}{f_0(y \mid \pi)}, \quad f_k(y \mid \pi) := \int_\Theta v^k \, f(y \mid v) \, \pi(v) \mathrm{d}v.$$

Čia $L_N(\theta \mid Y_N) = \prod_{j=1}^N f(y_j \mid \theta)$ yra parametro θ tikėtinumo funkcija duomenims Y_N.

Pagrindinė Bajeso statistikos problema – apriorinio skirstinio Π parinkimas. Jis svarbus galutiniam tyrimo rezultatui, ypač turint mažą imtį.

Empirinio Bajeso metodo esmė: apriorinis skirstinys laikomas nežinomu parametru ir yra įvertinamas iš turimų duomenų. T.y. tariama, kad duomenų Y_N skirstinys yra nusakytas (marginalinio) atskiro stebinio y skirstinio tankiu $f_0(\cdot \mid \pi)$ su nežinomu hiperparametru π, kuris įvertinamas pagal Y_N. Paprastai taikomas didžiausio tikėtinumo metodas. Jeigu tariama, kad π priklauso parametrinei skirstinių klasei, tai empirinis Bajeso metodas vadinamas *parametriniu*, kitais atvejais – *neparametriniu*.

Empirinis Bajeso metodas labai praverčia incidentinių parametrų schemoje, kai kiekvieną stebinį y_j atitinka sava nežinomo atsitiktinio parametro θ reikšmė θ_j ($j =$

$1, \ldots, N$). Reikšmės θ_j, $j = 1, \ldots, N$, laikomos nepriklausomomis ir vienodai pasiskirsčiusiomis pagal skirstinį Π.

Žinant Π, nežinomus parametrus $\theta_j, j = 1, \ldots, N$, galima įvertinti *aposteriориniu vidurkiu*

$$\hat{\theta}_j(y_j \mid \pi) := \frac{f_1(y_j \mid \pi)}{f_0(y_j \mid \pi)}, \quad j = 1, \ldots, N. \tag{2.31}$$

Pritaikius empirinį Bajeso metodą gaunami įvertiniai $\hat{\theta}_j(y_j \mid \hat{\pi})$. Čia nežinomas marginalinio y skirstinio tankio $f_0(\cdot \mid \pi)$ hiperparametras π (2.31) formulėje yra pakeičiamas jo įvertiniu $\hat{\pi}$, gautu remiantis imtimi Y_N.

Šiame darbe empirinis Bajeso metodas taikomas konstruojant struktūrinio skirstinio įvertinį.

3

Statistinių metodų taikymai lietuvių kalbos analizėje

3.1 Lietuviškų tekstų palyginimas pagal raidinę ir fonetinę žodžių struktūrą

Lietuvių kalba yra gana sudėtinga ir lanksti, ir tai gerokai apsunkina efektyvių algoritmų kūrimą automatiniam lietuviškų tekstų apdorojimui. Paprastai klasifikuojant tekstus remiamasi kokiais nors raktiniais žodžiais, tačiau lietuvių kalboje dėl linksniavimo, asmenavimo ir kitos kaitos gali keistis tiek žodžio galūnė, tiek ir kamienas. Tai labai apsunkina raktinių žodžių parinkimo uždavinį.

Kadangi raktiniai žodžiai atspindi teksto turinį, tai jie patys ir jų dažnumai tekste labai priklauso nuo to teksto autoriaus, temos ir net nuo paties kūrinio. Vadinasi, aktualus uždavinys – pamatuoti įvairias tiriamų tekstų formos ir/ar stiliaus ypatybes, išreikšti jas per kiekybines charakteristikas, kurias būtų galima iš tų tekstų suskaičiuoti. Tokias charakteristikas, kurios nesusijusios su teksto turiniu ir gali būti suskaičiuotos bet kuriam tekstui, vadinsime *universaliomis kiekybinėmis charakteristikomis*.

Buvo iškelti tokie uždaviniai:

a) Nustatyti stabilias, invariantiškas proporcijas tarp įvairių raidžių tipų lietuviškuose tekstuose. Klausimas: Kuo visi tekstai panašūs?

b) Nustatyti tose proporcijose pasireiškiančius skirtumus tarp grožinės ir mokslinės

literatūros. Klausimas: Kuo skiriasi grožinės ir mokslinės literatūros tekstai?

Pasirinktas logtiesinių modelių tyrimas, kadangi naudojantis logtiesiniais modeliais galima aprašyti ir nagrinėti sudėtingus, taip pat ir aukštesnės (negu 3-ios) eilės tiriamų požymių tarpusavio sąryšius. Kitų alternatyvų tokiems tyrimams praktiškai nėra.

3.1.1 Pirminė statistinė duomenų analizė

Analizuojami lietuviškų tekstų duomenys buvo gauti iš atsitiktinai paimtų skirtingų tekstų (kūrinių) lietuvių kalba, laisvai prieinamų vartotojams elektroninėje formoje. Kadangi šis tyrimas buvo labiau eksperimentinis, siekiant iliustruoti logtiesinių modelių panaudojimo galimybes tekstų apdorojime, nėra tiksliai apibrėžiama, kas yra lietuviškas tekstas ir kas yra tiriamoji populiacija.

Šiam tyrimui duomenys buvo paimti iš 17 skirtingų grožinio ir mokslinio stilių tekstų, atsitiktinai iš kiekvieno atrenkant po 3000 žodžių. Grožinio stiliaus tekstų (pvz., Maironio „Pavasario balsai“, E. Hemingvėjaus „Senis ir jūra“) buvo paimta 11 (t.y. juos sudarė 33000 žodžių, 228229 raidės), mokslinio stiliaus tekstų (3 filosofijos, 2 lingvistikos, 1 politologijos) buvo paimta 6 (t.y. juos sudarė 18000 žodžių, 137939 raidės).

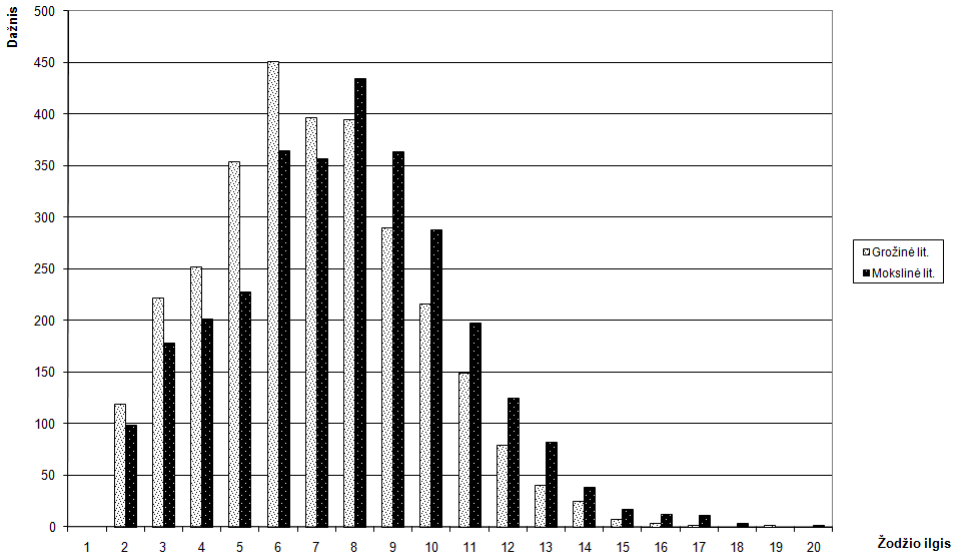
Palyginus pradinius duomenis matome, kad mokslinio stiliaus tekstuose žodžiai vidutiniškai yra šiek tiek ilgesni. 3.1 lentelėje parodyta, kiek kokio ilgio žodžių yra grožinio ir mokslinio stilių tiriamose imtyse.

Lentelė 3.1: Žodžių ilgis skirtingų stilių tekstuose

Žodžių ilgis	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Grožinis stilius	0	1306	2438	2772	3891	4960	4367	4337	3187	2372	1642	866	447	277	75	34	14	3	12	0
Mokslinis stilius	0	589	1071	1208	1365	2186	2139	2606	2179	1726	1185	749	491	233	104	72	66	19	1	11

Kadangi skiriasi žodžių skaičius grožinio ir mokslinio stilių imtyse, lyginti skirtingo ilgio žodžių dažnumą galima tik „suvienodinus“ tiriamų imčių apimtis, t.y. grožinio stiliaus dažnius padalijus iš 11, o mokslinio – iš 6. 3.1 diagramoje pavaizduotas vidutiniškas skirtingų ilgių atitinkamo stiliaus žodžių dažnis.

Matome, kad trumpesni žodžiai (iki 8 raidžių) yra dažnesni grožinio stiliaus tekstuose, o ilgesni (nuo 8 raidžių) – mokslinio stiliaus tekstuose. Išimtis – 19 raidžių žodžiai,



Pav. 3.1: Žodžių ilgis skirtingų stilių tekstuose

kurių nagrinėjamuose grožinio stiliaus tekstuose buvo rasta daugiau negu mokslinio stiliaus tekstuose. Taigi, galima daryti prielaidą, kad grožinio ir mokslinio stiliaus tekstus galima atskirti pagal žodžių ilgius.

3.2 lentelėje pateikiama, kiek procentų visų atitinkamo stiliaus imties raidžių / garsų sudaro balsiai, priebalsiai apskritai, bei pasižymintys tam tikromis savybėmis.

Lentelė 3.2: Garsų ir raidžių dažnumas (procentais) skirtingų stilių tekstuose

	Priebalsiai	Duslieji prieb.	Skardieji prieb.	Pusbalsiai	Skard. prieb. su pusb.
Grožinis stilius	52,67	24,24	7,81	20,62	28,43
Mokslinis stilius	52,78	24,51	6,90	21,37	28,27
Skirtumas	$-0,11$	$-0,27$	$0,91$	$-0,75$	$0,16$
	Balsiai	Lūpiniai b.	Nelūpiniai b.	Priešakinės eilės b.	Užpakalinės eilės b.
Grožinis stilius	47,33	11,56	35,77	22,95	24,38
Mokslinis stilius	47,22	11,77	35,45	23,30	23,92
Skirtumas	$0,11$	$-0,21$	$0,32$	$-0,35$	$0,46$
		Ilgosios balsės	Nosinės balsės		
Grožinis stilius		1,65	2,42		
Mokslinis stilius		1,93	2,19		
Skirtumas		$-0,28$	$0,23$		

Pagal rezultatus, pateikiamus 3.2 lentelėje, galima pastebėti tokius ryškesnius grožinio ir mokslinio stilių skirtumus: grožinio stiliaus tekstuose daugiau skardžiųjų priebalsių (beveik 1%) ir užpakalinės eilės balsių (beveik 0,5%) negu mokslinio stiliaus tekstuose, o pastaruosiuose daugiau pusbalsių negu grožinio stiliaus tekstuose.

Ar pastebėti skirtumai yra statistiškai reikšmingi? Norint gauti nuoseklų atsakymą

buvo parinktas logtiesinis modelis, kuris parodė, kad šitie skirtumai yra sąlygoti autorių.

3.1.2 Kintamieji

Logtiesinių modelių tyrimui buvo apibrėžti tokie kintamieji: „*teksto numeris*“ (žymima *nr*), „*teksto stilius*“ (nurodo – grožinė ar mokslinė literatūra; žym. *st*), „*žodžio ilgis 5*“ (žodžiai skirstomi į 5 ilgio grupes: žodžiai, kurių ilgis < 4 raides, 4 arba 5 raidžių, 6 raidžių (žodžių, kurių ilgis $= 6$, daugiausia), 7 arba 8 raidžių žodžiai ir žodžiai, kurių ilgis > 8 raides; žym. *i5*).

Visos raidės buvo suskirstytos į balses (kintamasis *b*) bei priebalses ir joms priskirti atitinkami požymiai, susiję su rašyba arba tarimu. Apibrėžti tokie kintamieji: „*ilgosios*“ ir „*nosinės balsės*“ (žymima atitinkamai *by* ir *bn*), „*lūpiniai balsiai*“ (*bl*), „*priešakinės eilės balsiai*“ (*bp*), „*duslieji priebalsiai*“ (*prd*) ir „*pusbalsiai*“ (*psb*) (apie balsių ir priebalsių klasifikavimą žr. 2.1 poskyryje). Raidžių kodavimas, t.y. priskyrimas atitinkamoms savybėms, pateiktas 3.3 lentelėje. Iš lentelės galima pastebėti, kad yra vienintelė balsė, nepasižyminti išskirtomis savybėmis, t.y. nėra nei ilgoji, nei nosinė raidė ir nežymi lūpinės balsės. Kitaip sakant, kintamasis *b* žymi vienintelę raidę – *a*.

Lentelė 3.3: Balsių ir priebalsių kodavimas

<table>
<tr><td colspan="4">Balsiai
$b = 1$</td><td colspan="3">Priebalsiai
$b = 0$</td></tr>
<tr><td colspan="2" rowspan="2"></td><td rowspan="2">$by = 1$</td><td colspan="2">$by = 0$</td><td rowspan="2">$prd = 1$</td><td colspan="2">$prd = 0$</td></tr>
<tr><td>$bn = 1$</td><td>$bn = 0$</td><td>$psb = 1$</td><td>$psb = 0$</td></tr>
<tr><td colspan="2">$bl = 1$, $bp = 0$</td><td>ū</td><td>ų</td><td>o, u</td><td>c, č, f, k, p, s, š, t</td><td>j, l, m, n, r, v</td><td>b, d, g, h, z, ž</td></tr>
<tr><td rowspan="2">$bl = 0$</td><td>$bp = 1$</td><td>y</td><td>ę, į</td><td>e, ė, i</td><td rowspan="2"></td><td rowspan="2"></td><td rowspan="2"></td></tr>
<tr><td>$bp = 0$</td><td></td><td>ą</td><td>a</td></tr>
</table>

Taip pat buvo apskaičiuoti visų tipų raidžių kiekiai kiekviename žodyje.

3.1.3 Empirinio tyrimo rezultatai

Naudojantis „SAS“ sistemos procedūra CATMOD buvo tirtas raidinės ir garsinės struktūros sąryšis su moksliniu ir grožiniu stiliumi. Gauti rezultatai pateikti 3.4 lentelėje.

Atitinkamų logtiesinio modelio narių (veiksnių) statistinis reikšmingumas pateiktas stulpelyje $Pr > ChiSq$, kuriame nurodytos į modelį įtrauktų požymių bei jų sąveikų (kombinacijų) atitinkamos p reikšmės.

Lentelė 3.4: Didžiausio tikėtinumo dispersinė analizė

Maximum Likelihood Analysis of Variance					
Source	**DF**	**Pr > ChiSq**	**Source**	**DF**	**Pr > ChiSq**
bl	1	0,0035	psb*nr	16	0,0128
bn	1	<,0001	i5*psb*nr	64	<,0001
bl*bn	1	0,0508	prd	1	<,0001
i5	4	<,0001	i5*prd	4	0,0003
i5*bn	4	<,0001	i5*prd*nr	64	<,0001
b	1	<,0001	i5*b*nr	64	<,0001
i5*b	4	0,0107	bl*nr	16	0,0002
bp	1	<,0001	i5*bl*nr	64	0,0005
by	1	<,0001	bp*nr	16	0,0328
i5*bl	4	0,0176	bp*bn*nr	16	<,0001
i5*by	4	<,0001	i5*bp*nr	64	<,0001
i5*bp*bn	4	0,0144	bl*bn*nr	16	<,0001
i5*bp*by	4	<,0001	st*i5*by	4	<,0001
nr	16	<,0001	by*nr	16	<,0001
i5*nr	64	<,0001	bp*by	1	0,0314
psb	1	<,0001	**Likelihood Ratio**	**361**	**0,0685**

Tradiciškai veiksnys laikomas statistiškai reikšmingu, jeigu atitinkama p reikšmė yra mažesnė už reikšmingumo lygmenį $\alpha = 0,05$. Paskutinėje lentelės eilutėje pateikta tikėtinumo santykio kriterijaus p reikšmė, kuri šiuo atveju rodo, kad parinktas modelis pakankamai gerai atitinka turimus duomenis, todėl nulinę hipotezę apie parinkto modelio adekvatumą atmesti nėra pagrindo.

Gautasis modelis nėra grafinis, todėl jo negalima aprašyti klikomis.

Pamėginsime šiek tiek formaliau pateikti šito modelio aprašymą. Stengdamiesi išlaikyti simbolinius kintamųjų žymėjimus apibrėšime indeksų (žr. 2.3.3 paragrafo (2.10) formulę) aibę. Statistinėje analizėje naudoti kintamųjų pavadinimai ir naujai įvesti matematiniai jų analogai yra pateikti ir susieti 3.5 lentelėje.

Lentelė 3.5: Modelio kintamųjų ir indeksų sąsajos

Nr.	1	2	3	4	5	6	7	8	9	10
Kintamieji	*st*	*nr*	*i5*	*b*	*by*	*bn*	*bl*	*bp*	*prd*	*psb*
Kategorijų skaičius	2	17	5	2	2	2	2	2	2	2
j	j_1	j_2	j_3	j_4	j_5	j_6	j_7	j_8	j_9	j_{10}

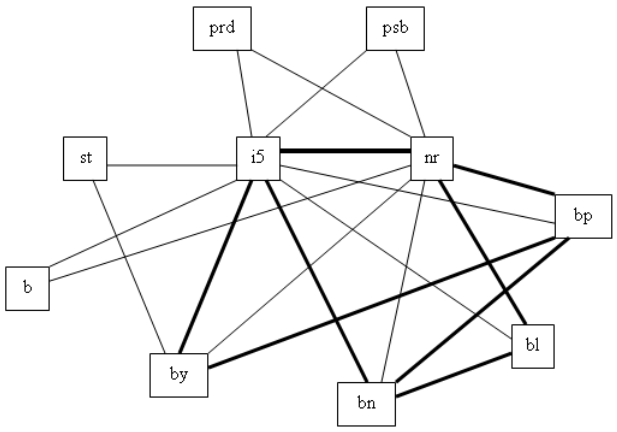
$J'_n = \{1, \ldots, n\}$. Bendra būsenų aibė yra $J = J'_2 \times J'_{17} \times J'_5 \times \underbrace{J'_2 \times \ldots \times J'_2}_{7}$. $j = (j_1, \ldots, j_{10}) \in J$.

Taigi, aptariamą logtiesinį modelį galima būtų užrašyti taip:

$$\ln \mu_j = u^\circ + u^{nr}(j_2) + u^{i5}(j_3) + u^b(j_4) + u^{by}(j_5) + u^{bn}(j_6) + u^{bl}(j_7) + u^{bp}(j_8) +$$

$$
\begin{aligned}
&+u^{prd}(j_9) + u^{psb}(j_{10})+\\
&+u^{nr,i5}(j_2,j_3) + u^{nr,by}(j_2,j_5) + u^{nr,bl}(j_2,j_7) + u^{nr,bp}(j_2,j_8) + u^{nr,psb}(j_2,j_{10})+\\
&+u^{i5,b}(j_3,j_4) + u^{i5,by}(j_3,j_5) + u^{i5,bn}(j_3,j_6) + u^{i5,bl}(j_3,j_7) + u^{i5,prd}(j_3,j_9)+\\
&+u^{by,bp}(j_5,j_8) + u^{bn,bl}(j_6,j_7)+\\
&+u^{st,i5,by}(j_1,j_3,j_5) + u^{nr,i5,b}(j_2,j_3,j_4) + u^{nr,i5,bl}(j_2,j_3,j_7) + u^{nr,i5,bp}(j_2,j_3,j_8)+\\
&+u^{nr,i5,prd}(j_2,j_3,j_9) + u^{nr,i5,psb}(j_2,j_3,j_{10}) + u^{nr,bn,bl}(j_2,j_6,j_7)+\\
&+u^{nr,bn,bp}(j_2,j_6,j_8) + u^{i5,by,bp}(j_3,j_5,j_8) + u^{i5,bn,bp}(j_3,j_6,j_8).
\end{aligned}
$$

Gautą logtiesinį modelį papildžius tam tikromis sąsajomis, prie kurių koeficientai lygūs 0, galime gauti bendresnį – grafinį – modelį, kurio grafas pavaizduotas 3.2 pav.



Pav. 3.2: Žodžių raidinės ir fonetinės struktūros sąryšių grafas

Ryškesnės briaunos žymi, kurie elementai sudaro maksimalias klikas: $[i5, nr, bl, bn]$, $[i5, nr, bp, bn]$ ir $[i5, nr, bp, by]$. Visose klikose, kurių ilgis $= 4$, yra kintamieji *i5* ir *nr*. Tai reiškia, kad į klikas įeinančių atitinkamų savybių balsių pavartojimas priklauso nuo žodžio ilgio ir teksto numerio, t.y. ir autoriaus.

Aptarsime parinkto modelio ypatumus, jo interpretaciją. Ji nusakoma tais veiksniais ir tomis jų sąveikomis, kurios nebuvo statistiškai reikšmingos ir todėl į modelį nebuvo įtrauktos.

Atsakymui į a) klausimą svarbios yra raidžių / garsų grupės, kurios nėra susijusios (neturi sąveikos nario) su teksto numeriu (kintamuoju *nr*). Tai reiškia, kad tų raidžių / garsų grupių proporcijos buvo maždaug tokios pačios visuose nagrinėtuose tekstuose. Atsakymui į b) klausimą – priešingai, yra svarbios tos raidžių / garsų grupės, kurios yra susijusios (turi sąveikos narį) su kintamuoju „*teksto stilius*“ (*st*). Tai reiškia, kad

tų raidžių / garsų grupių proporcijos statistiškai reikšmingai skyrėsi moksliniuose ir grožinės literatūros tekstuose.

a) Lietuviškų tekstų invariantai yra aprašomi, pavyzdžiui, veiksniais *i5*bn*, *i5*by* ir *i5*bp*by*. Jie interpretuojami taip. Nosinių balsių (kint. *bn*) šansų santykis atžvilgiu žodžių ilgių grupių (*i5*) nepriklauso nuo autoriaus (*nr*). Analogiškai, ilgųjų balsių (*by*) šansų santykis atžvilgiu *bp* nepriklauso nuo *nr*. Lygiai taip pat, kaip *bp* atžvilgiu *by*.

b) Mokslinės ir grožinės literatūros tekstų skirtumus parodo vienintelis veiksnys *st*i5*by*, kuris reiškia, kad ilgųjų balsių (*by*) šansų santykis atžvilgiu žodžių ilgių grupių (*i5*) nepriklauso nuo *nr*, t.y. ir nuo autoriaus, bet priklauso nuo stiliaus (*st*).

Tuo nesunku įsitikinti. Pažymėkime (žr. 3.5 lent.)

$$\begin{aligned} j &= (j_1, j_2, b_3, j_4, 0, j_6, \ldots, j_{10}), \quad b_3 \in \{1, \ldots, 5\}, \\ j' &= (j_1, j_2, b_3, j_4, 1, j_6, \ldots, j_{10}), \\ \tilde{j} &= (j_1, j_2, g, j_4, 0, j_6, \ldots, j_{10}), \quad g \in \{1, \ldots, 5\}, g \neq b_3, \\ \tilde{j}' &= (j_1, j_2, g, j_4, 1, j_6, \ldots, j_{10}). \end{aligned}$$

Tuomet *by* šansų santykis atžvilgiu $i5$

$$\begin{aligned} \ln\left(\frac{\mu_{\tilde{j}'}\mu_j}{\mu_{\tilde{j}}\mu_{j'}}\right) &= u^{by}(1) + u^{nr,by}(j_2, 1) + u^{by,bp}(j_5, j_8) + \\ &\quad + u^{i5,by}(g, 1) + u^{st,i5,by}(j_1, g, 1) + u^{i5,by,bp}(g, 1, j_8) - \\ &\quad - \left(u^{by}(1) + u^{nr,by}(j_2, 1) + u^{by,bp}(j_5, j_8)\right) = \\ &= u^{i5,by}(g, 1) + u^{st,i5,by}(j_1, g, 1) + u^{i5,by,bp}(g, 1, j_8). \end{aligned} \tag{3.1}$$

Čia naudojomės tuo, kad funkcijos u virsta 0, kai jose bet kuris iš argumentų įgyja bazinę reikšmę. Grupės kintamajam $i5$ ta reikšmė yra b_3, o kintamajam *by* ta reikšmė yra 0. Matome, kad dešinė (3.1) lygybės pusė nepriklauso nuo kintamojo *nr* (t.y. nuo argumento j_2, žr. 3.5 lent.), bet priklauso nuo kintamojo *st* (t.y. nuo argumento j_1, žr. 3.5 lent.).

Analogiškai patikrinami ir punkto a) teiginiai apie šansų santykius. Galima pastebėti, kad visi jie negalioja išplėstiniam grafiniam modeliui (žr. 3.2 pav.).

Statistiškai labai reikšmingo veiksnio *st*i5*by* sąveikaujančių elementų reikšmingumo analizė parodė, kad statistiškai reikšmingas skirtumas yra tarp trumpiausių žodžių.

Grožinėje literatūroje ilgųjų balsių labai trumpuose žodžiuose yra mažiau negu mokslinėje literatūroje, o ilgesniuose žodžiuose – priešingai. Galima daryti prielaidą, kad šie mokslinio ir grožinio stiliaus skirtumai yra susiję su dažnesniu žodžio „yra“ vartojimu mokslinėje literatūroje.

Sudarytame modelyje 13 veiksnių iš 31 sąveikauja su kintamuoju *nr*. Norint atskleisti grožinės ir mokslinės literatūros tekstų skirtumus buvo stengiamasi visus modelio kintamuosius „*teksto numeris*“ (*nr*) pakeisti kintamuoju „*teksto stilius*“ (*st*). Deja, pavyko gauti tik vieną tokią sąveiką, o visuose kituose veiksniuose šis pakeitimas pažeidė modelio ir duomenų suderinamumą. Tai reiškia, kad tie veiksniai atspindi specifines kiekvieno autoriaus arba teksto (raidinės ir fonetinės žodžių struktūros) charakteristikas.

Apibendrinant atliktą tyrimą ir jo rezultatus galima būtų pasiūlyti tokią kalbos statistinių dėsningumų tyrimo metodiką. Ji leistų aprašyti ir įvertinti kalbos heterogeniškumą ir variabilumą (kintamumą), kurį sąlygoja jos autoriaus pasirinkimai, identifikuoti nuo autoriaus santykinai mažai priklausančias kalbos elementų savybes bei jų sąryšius ir tokiu būdu atskleisti potencialiai pačiai kalbai būdingus (statistinius) dėsningumus.

1. Tarkime, kad turime tekstyną, kuriame yra galimybė išrinkti tekstus pagal autorius. Tuomet renkama paprastoji atsitiktinė autorių imtis. Jeigu autorių mažai – imami visi autoriai.

2. Iš kiekvieno autoriaus tekstų nepriklausomai nuo likusių autorių išrenkama visiems autoriams vienodo dydžio paprastoji (grąžintinė) atsitiktinė tiriamų kalbos elementų ar struktūrų imtis. Kartais tokį imties planą realizuoti sudėtinga. Tada siekiama, kad sudaryta imtis kuo mažiau iškreiptų minėto imties plano savybes.

3. Imties duomenims sudaroma autoriaus ir tiriamų elementų požymių pasitaikymo imtyje dažnių lentelė, kuriai parenkamas (grafinis) logtiesinis modelis arba apibendrintasis logit modelis, kuriuose yra visos tiriamų požymių sąveikos su požymiu „autorius“.

4. Kiekviename parinkto logtiesinio modelio naryje, kuriame yra sąveika su požymiu

„autorius“, bandoma tą požymį pakeisti jo apibendrinimu arba (net) iš viso jį eliminuoti. Jeigu tai pavyksta padaryti nepažeidžiant modelio adekvatumo, t.y. taip, kad „apibendrintas“ modelis statistiškai reikšmingai nesiskiria nuo pradinio, gaunamas narys, kuris aprašo nuo individualių autoriaus savybių santykinai mažai priklausančias bendras ar (net) universalias tiriamų požymių sąveikas ir priklausomybes.

Analogišką metodiką galima pritaikyti ne vien autorių bet ir kitų pagrindinių tyrimo požymių analizei ir apibendrinimui. Ši metodika iš dalies buvo taikoma ir metakalbinių komentarų statistinei analizei (žr. 3.2 posk.).

3.1.4 Dalinės išvados

Atlikus raidžių ir garsų logtiesinių modelių analizę galima daryti tokias išvadas.

1. Nagrinėjant turimus duomenis buvo nustatyta, kad visiems nagrinėjamiems tekstams yra būdingos tam tikros, nuo teksto nepriklausančios, ilgųjų ir priešakinės eilės balsių proporcijos, kurios priklauso ir nuo žodžio ilgio.

2. Skirtumus tarp mokslinio ir grožinio stilių parodo nevienodas ilgųjų balsių pasiskirstymas įvairaus ilgio žodžiuose, ypač trumpiausiuose žodžiuose. Galima daryti prielaidą, kad šie stiliaus skirtumai yra susiję su dažnesniu žodžio „yra“ vartojimu mokslinėje literatūroje.

3. Kadangi šiame tyrime analizuojami tekstai buvo paimti atsitiktinai, tiksliai neapibrėžus, kas yra lietuviškas tekstas ir kas yra tiriamoji populiacija, tai gautų rezultatų negalima apibendrintai taikyti visiems lietuviškiems tekstams. Lietuviško teksto ir atitinkamos tyrimo populiacijos apibrėžimas yra labai aktualus ir sudėtingas uždavinys. Vienas iš galimų sprendimų – naudoti Lingvistikos centre sudarytą ir jo palaikomą tekstyną (žr. *http://tekstynas.vdu.lt/*). Deja, vartotojas turi labai ribotas galimybes (sudėtinga surinkti informaciją apie autorius ir pan.), todėl dabartinis tekstynas nėra tinkamas sudėtingesniems tyrimams atlikti (tai iš dalies atsiskleidžia kitame poskyryje).

3.2 Metakalbinių komentarų ir jų funkcijų statistinė analizė

Lietuvių kalba pasižymi vaizdingumu, kuris gali būti perteikiamas įvairiomis raiškos priemonėmis. Viena iš jų – metakalbiniai komentarai (MK).

Metakalbinius komentarus, kaip kalbos konkurencijos priemonę, nemažai nagrinėjo K. Župerka (žr. [99]) ir kiti lituanistai, taip pat ir kitų šalių (latvių, čekų) kalbininkai. Dauguma jų MK-us nagrinėja kaip stilistikos objektą, tačiau nėra nagrinėjami kitokio pobūdžio klausimai, pavyzdžiui, kokia tikimybė, kad pavartotas žodis yra MK-o dalis. Šis darbas turėtų bent iš dalies užpildyti minėtą spragą.

Dalis MK-ų apibūdina kalbos, pasakymo toną: malonus (*švelniai, meiliai kalbant*), iškilmingas (*iškilmingai, skambiai tariant*), juokaujamas (*juokaujamai, šmaikščiai kalbant*), piktas (*piktai, šiurkščiai tariant*). Iš MK-ų įvairovės galima išskirti apibūdinančius vartojamo pasakymo (emocinį) toną ir tuo pačiu išreiškiančius emocinę bei estetinę priešpriešą: *gražiai*, *švelniai*, *skambiai*, *aiškiai*, *atvirai*, *vaizdingai*, *vaizdžiai* ir *negražiai*, *grubiai*, *šiurkščiai*, *vulgariai*, *griežtai*, *ironiškai*, *banaliai kalbant*, *sakant* ir pan., kurie ir bus nagrinėjami šiame poskyryje.

Nors yra manoma, kad neigiamas vertinimas kalboje išreikštas dažniau negu teigiamas, mūsų išrinkti MK rodo, kad kalbos vartotojai nevengia pabrėžti ir teigiamo kalbos bei kalbėjimo (šnekos) vertinimo.

Komunikacines šnekos ypatybes įvardijantys MK rodo antrojo konkurento orientavimą į kalbamąjį daiktą, turinį (*geriau pasakius, tiksliau sakant*), į adresatą (*aiškiau pasakius, paprasčiau tariant*), į sąmonę (*vaizdingai, metaforiškai kalbant*), į glaustumo bei apibendrinimo siekimą (*trumpai tariant, apibendrintai kalbant*) (plačiau žr. [99]). Tai reiškia, kad MK, kaip ir paprasti žodžiai, komunikacijos procese atlieka skirtingas funkcijas.

Kalbos analizėje taikant statistinius metodus labai svarbu tiksliai apibrėžti tiriamą populiaciją. Nuo to priklauso gauti rezultatai, jų interpretacija ir patikimumas. Tai iliustruoja ir šiame darbe gauti rezultatai.

Šiame darbe, remiantis VDU sudarytu lietuvių kalbos tekstynu, nagrinėjama rašytinė lietuvių kalba. Tokiu būdu tiriamoji populiacija yra nusakyta to tekstyno sudarymo

taisyklėmis. Deja, konkreti metodika, pagal kurią yra sudarytas tekstynas, autorei nėra žinoma. Neaišku, ar sudarant tekstyną buvo remtasi imčių teorija (žr. 2.2 poskyrį). Vadinasi, šiuo atveju negalime tvirtinti, kad tiriamoji populiacija adekvačiai atspindi rašytinę lietuvių kalbą, juolab kad ir pats terminas „rašytinė lietuvių kalba“ nėra tiksliai apibrėžtas.

MK-ų nagrinėjimui taikoma Pirsono koreliacija ir logistinė regresija, naudojama SPSS ir SAS programinė įranga.

3.2.1 Pirminė statistinė duomenų analizė

Tyrimui duomenys buvo išrinkti iš VDU KLC viešai prieinamo internete „Dabartinės lietuvių kalbos tekstyno“ (žr. *http://tekstynas.vdu.lt/*). Šiame internetiniame puslapyje yra dvi tekstyno versijos – senoji ir naujoji. 3.6 lentelėje pateiktas abiejų tekstyno versijų sandaros bei jose rastų, šiame poskyryje analizuojamų MK-ų kiekio palyginimas.

Lentelė 3.6: Palyginimas

Tekstyno senoji versija		**MK**	**Tekstyno naujoji versija**		**MK1**	**MK2**
Respublikinė periodika	23%	23,83%				
Vietinė periodika	17%	23,10%				
Populiarioji periodika	18%	19,41%				
Specializuota periodika	8%	6,6%	Publicistika	63,8%	66,47%	55,24%
Grožinė literatūra	7%	8,71%	Grožinė lit.	11,6%	17,57%	15,84%
Negrožinė literatūra	11%	7,72%	Negrožinė lit.	14,2%	10,84%	25,87%
Seimo stenogramos	2%	5,94%	Administrac. lit.	10%	4,66%	3,05%
LR valstyb. dokumentai	8%	0%	Sakytinė kalba	0,3%	0,46%	
Filosofinės lit. vertimai	3%	2,77%				
Memuarai	3%	1,92%				
≈102 mln. žodžių		1515MK	≈141 mln. žodžių		2186MK	5778MK

Su MK-ais buvo atliekami du tyrimai, todėl iš pradžių buvo išrinkta mažesnė imtis, pirmiausia iš senosios tekstyno versijos (1515 MK), o pasirodžius naujajai versijai – tie patys MK (2186 MK) buvo išrinkti ir iš jos. Detalesniam (MK-ų funkcijų) tyrimui tik iš naujosios tekstyno versijos buvo išrinkti visi komunikacines šnekos ypatybes įvardijantys MK, kurių struktūra: padalyvis + prieveiksmis (žr. 2.1 poskyrį).

Kadangi kai kurių tiriamų MK-ų dažniai tekstyne yra labai maži, sudarant logistinės regresijos modelį pagal pirmuosius duomenis buvo susidurta su retų ir išsklaidytų

dažnių lentelių problema (plačiau žr. [4]) ir norint parinkti patikimą ir stabilų modelį tuos labai retus MK-us iš tolesnio tyrimo teko eliminuoti, todėl kitam tyrimui MK iš sakytinės kalbos dalies iš viso neberinkti.

Kadangi naujojo tekstyno apimtis 40 mln. žodžių didesnė, visiškai suprantama, kad jame ir MK-ų rasta kur kas daugiau. Be to, galima būtų tikėtis, kad MK-ų pasiskirstymas skirtingose tekstyno srityse yra tiesiogiai proporcingas atitinkamos srities dydžiui. Tačiau iš lentelės matome (čia lyginama senosios tekstyno versijos imtis su pirmąja imtimi iš naujosios tekstyno versijos, t.y. tas atvejis, kai abiejose tekstyno versijose buvo renkami tie patys MK), kad, tarkim, naujajame tekstyne grožinės literatūros srityje rasta daugiau MK-ų, negrožinėje mažiau nei tikėtasi, o administracinėje literatūroje MK-ų rasta net dvigubai mažiau. Tai rodo, kad MK yra ne paprasti žodžiai, o vaizdingumo raiškos priemonė, ir jų atliekamos specifinės – kalbos vaizdingumo, pasakymo vertinimo – funkcijos ne kiekvienoje srityje vienodai reikalingos. Pavyzdžiui, grožinėje literatūroje vaizdingumo reikia daugiau, mokslinėje ar administracinėje – mažiau, o valstybiniuose dokumentuose (ši sritis buvo išskirta senojoje tekstyno versijoje) vaizdingumas visai nepriimtinas, todėl nagrinėjamų MK-ų šioje srityje nėra.

Apskaičiavus naujojo tekstyno srityse rastų pirmos imties MK-ų koreliacijas paaiškėjo, kad stipriausiai koreliuoja sakytinės kalbos ir administracinės literatūros MK, nors tiek MK-ų bazinių žodžių, tiek visų žodžių, esančių šaltiniuose, kuriuose buvo bent vienas nagrinėjamas MK, šios dvi sritys koreliuoja silpnai.

Silpniausiai koreliuoja grožinės ir negrožinės literatūros MK, nors jau minėtose kitose imtyse šių dviejų sričių koreliacija yra vidutiniška ar net stipri.

Tai dar kartą patvirtina, kad MK yra išskirtinė kalbos raiškos priemonė.

3.3 paveikslėlyje–lentelėje pateikiama SPSS programa apskaičiuota antrosios (papildytos, bet be sakytinės kalbos) MK-ų imties Pirsono koreliacija. Matome, kad stipriausiai koreliuoja publicistikos (*Publ_MK*) ir administracinės literatūros (*AdmL_MK*) MK. Silpniausiai koreliuoja administracinės (*AdmL_MK*) ir negrožinės literatūros (*NegL_MK*) MK.

Kadangi išrinkti iš tekstyno visus MK-us yra praktiškai neįmanoma, apsiribota tokia MK-ų struktūra: padalyvis + prieveiksmis (bazinis žodis). Pagal tai, kokią

Correlations

		GrL_MK	NegL_MK	AdmL_MK	Publ_MK
GrL_MK	Pearson Correlation	1	,617**	,851**	,789**
	Sig. (2-tailed)		,000	,000	,000
	N	48	48	48	48
NegL_MK	Pearson Correlation	,617**	1	,400**	,653**
	Sig. (2-tailed)	,000		,005	,000
	N	48	48	48	48
AdmL_MK	Pearson Correlation	,851**	,400**	1	,871**
	Sig. (2-tailed)	,000	,005		,000
	N	48	48	48	48
Publ_MK	Pearson Correlation	,789**	,653**	,871**	1
	Sig. (2-tailed)	,000	,000	,000	
	N	48	48	48	48

**. Correlation is significant at the 0.01 level (2-tailed).

Pav. 3.3: Antros imties MK-ų koreliacija

būseną nurodo padalyvis, MK-us sąlyginai galima skirstyti į tokias grupes: „kalbėjimo“, „galvojimo“ ir „žiūrėjimo“ (žr. 3.7 lent.)

Lentelė 3.7: Metakomentarų grupės

„Kalbėjimas“		„Galvojimas“		„Žiūrėjimas“	
kalbant šnekant sakant pasakius tariant tarus	abstrakčiai, konkrečiai, tiksliai, tiesiai, primityviai, nuoširdžiai, paprastai, trumpai, rimtai, bendrai, logiškai, gerai, apibendrintai, teisingai, tikriau, metaforiškai, vaizdžiai, vaizdingai, švelniai, skambiai, aiškiau, atvirai, vulgariai, šiurkščiai, grubiai, griežtai, banaliai	galvojant pagalvojus mąstant pamąsčius svarstant pasvarsčius	abstrakčiai, konkrečiai, nuoširdžiai, paprastai, rimtai, nuosekliai, logiškai, blaiviai, gerai	žiūrint žvelgiant pažvelgus	abstrakčiai, primityviai, paprastai, rimtai, bendrai, nuosekliai, logiškai, blaiviai, gerai
94,86%		4,57%		0,57%	

Matome, kad didžiąją dalį visų išrinktų MK-ų sudaro „kalbėjimo“ grupei priklausantys MK. Būtent šiai grupei priklauso ir 10 dažniausių MK-ų, kurie sudaro net 60% visų MK-ų (skliausteliuose nurodytas MK-o dažnis): *trumpai tariant* (934), *švelniai tariant* (674), *tiksliau sakant* (549), *atvirai kalbant* (270), *tiksliau pasakius* (227), *atvirai sakant* (182), *atvirai pasakius* (180), *tiksliau tariant* (169), *vaizdžiai tariant* (165), *trumpiau tariant* (119).

3.8 lentelėje mažėjimo tvarka surašytos tekstyno dalys pagal atitinkamos būsenos MK-ų santykius su visais tam tikros tekstyno dalies MK-ais.

Lentelė 3.8: Santykinės MK-ų būsenos tekstyno dalyse

„Kalbėjimas“	„Galvojimas“	„Žiūrėjimas“
NegL – 0,983	GrL – 0,082	GrL – 0,012
AdmL – 0,977	Publ – 0,052	NegL – 0,005
Publ – 0,943	AdmL – 0,023	Publ – 0,004
GrL – 0,906	NegL – 0,012	AdmL – 0,000

Palyginę šiuos rezultatus matome, kad grožinėje literatūroje (*GrL*) mažiau nei kitose srityse „kalbama“, o daugiau „galvojama“ ir „žiūrima“, t.y. grožinėje literatūroje būdinga stebėti, apmąstyti. Negrožinėje literatūroje (*NegL*) daugiau nei kitose srityse „kalbama“ ir mažiau „galvojama“, nes šioje srityje paprastai yra pateikiami faktai, o ne samprotavimai, išvedžiojimai.

3.2.2 Empirinio tyrimo rezultatai

Pagal pirmuosius duomenis (2186 MK) buvo sudarytas logistinės regresijos modelis, kuriuo siekiama aprašyti, nuo ko priklauso MK-ų proporcija tarp visų bazinio žodžio (baziniu žodžiu laikomas MK-o konstrukcijoje esantis prieveiksmis) pavartojimų, t.y. tikimybę, kad pavartotas bazinis žodis yra MK-o dalis. Kadangi kai kurių tiriamų MK-ų dažniai tekstyne buvo labai maži, tai parenkant patikimą ir stabilų modelį tuos labai retus MK-us teko eliminuoti iš tolesnio tyrimo.

Šiam tyrimui buvo apibrėžti tokie kintamieji: *bz_lg* – bazinio žodžio dažnio dešimtainis logaritmas; *bz_salt_lg* – bazinio žodžio santykinio dažnio tarp visų jo šaltinių (t.y. nors kartą jį paminėjusių šaltinių) žodžių dešimtainis logaritmas; *bz_srit_lg* – bazinio žodžio santykinio dažnio tarp visų srities žodžių dešimtainis logaritmas; *tipas* – MK-ą apibūdinanti savybė (vaizdingumas, atvirumas, švelnumas ir t.t., priklausomai nuo bazinio žodžio); *laipsnis* – bazinio žodžio nelyginamasis (0) arba aukštesnysis (1) laipsnis; *sritis* – keturios naujojo tekstyno sritys (dėl jau minėtos problemos sakytinės kalbos sritis buvo nenaudojama), grožinę literatūrą laikant bazine kategorija; *tn_atspalv* – teigiamas / neigiamas bazinio žodžio (tuo pačiu ir MK-o) atspalvis (*vaizdžiai, vaizdingai, švelniai, aiškiau, atvirai* turi teigiamą atspalvį (1), o *šiurkščiai,*

grubiai, griežtai, banaliai turi neigiamą atspalvį (-1)).

Remiantis informaciniais kriterijais ir eliminavus statistiškai nereikšmingus veiksnius buvo parinktas toks (žr. 3.9 lent.) logistinės regresijos modelis.

Lentelė 3.9: Logistinės regresijos modelis 1

Type 3 Analysis of Effects				
Effect	**DF**	**Estimate**	**Wald Chi-Square**	**Pr>ChiSq**
bz_lg	1	$-1,2027$	$12,4734$	$0,0004$
bz_salt_lg	1	$3,8537$	$20,2096$	$<,0001$
$tn\_atspalv * tipas$	8		$131,2744$	$<,0001$
$tn\_atspalv * laipsnis$	1	$0,8033$	$15,7000$	$<,0001$
$tn\_atspalv * tipas * sritis$	24		$479,1275$	$<,0001$

Atitinkamų logistinės regresijos modelio narių statistinis reikšmingumas pateiktas paskutiniame lentelės stulpelyje, kuriame yra visų į modelį įtrauktų veiksnių bei jų sąveikų atitinkamos p reikšmės.

Tradiciškai veiksnys laikomas statistiškai reikšmingu, jeigu atitinkama p reikšmė yra mažesnė už reikšmingumo lygmenį $\alpha = 0,05$. Galima pastebėti, kad šį modelį sudarantys veiksniai ir jų sąveikos yra statistiškai reikšmingi.

Hosmerio ir Lemešou kriterijus, kurio statistikos p reikšmė $p = 0,9792 > 0,05 = \alpha$, taip pat rodo, kad parinktas modelis yra tinkamas, t.y. gana gerai suderintas su duomenimis.

Sudarytame logistinės regresijos modelyje aiškinančiųjų kintamųjų $X = X(t)$ įtaką tikimybėms

$$p(X) = \frac{\exp\{\beta^\top X\}}{1 + \exp\{\beta^\top X\}}$$

aprašo toks tiesinis prediktorius (žr. 2.3.1 paragrafą):

$$\begin{aligned}
\beta^\top X \;=\; & 21,28 - 1,2027 \times bz_lg + 3,8537 \times bz_salt_lg - 4,3116 \times tn_atspalv * tipas[1] - \\
& -4,949 \times tn_atspalv * tipas[2] - 5,3402 \times tn_atspalv * tipas[3] - \\
& -12,0316 \times tn_atspalv * tipas[5] + 7,2443 \times tn_atspalv * tipas[10] + \\
& +5,9315 \times tn_atspalv * tipas[11] + 10,2918 \times tn_atspalv * tipas[12] + \\
& +9,3509 \times tn_atspalv * tipas[14] + 0,8033 \times tn_atspalv * laipsnis - \\
& -0,6289 \times tn_atspalv * tipas * sritis[1][2] - 0,5112 \times tn_atspalv * tipas * sritis[1][3] + \\
& +1,442 \times tn_atspalv * tipas * sritis[1][4] - 0,2009 \times tn_atspalv * tipas * sritis[2][2] + \\
& +1,3976 \times tn_atspalv * tipas * sritis[2][3] - 0,1853 \times tn_atspalv * tipas * sritis[2][4] - \\
& -0,5491 \times tn_atspalv * tipas * sritis[3][2] + 1,2788 \times tn_atspalv * tipas * sritis[3][3] + \\
& +1,7994 \times tn_atspalv * tipas * sritis[3][4] + 4,7411 \times tn_atspalv * tipas * sritis[5][2] - \\
& -16,8797 \times tn_atspalv * tipas * sritis[5][3] + 6,1705 \times tn_atspalv * tipas * sritis[5][4] - \\
& -0,41 \times tn_atspalv * tipas * sritis[10][2] + 0,6035 \times tn_atspalv * tipas * sritis[10][3] - \\
& -1,6192 \times tn_atspalv * tipas * sritis[10][4] + 0,5648 \times tn_atspalv * tipas * sritis[11][2] -
\end{aligned}$$

$$-0,3962 \times tn\_atspalv * tipas * sritis[11][3] - 1,3124 \times tn\_atspalv * tipas * sritis[11][4] -$$
$$-2,7794 \times tn\_atspalv * tipas * sritis[12][2] + 7,3357 \times tn\_atspalv * tipas * sritis[12][3] -$$
$$-2,5229 \times tn\_atspalv * tipas * sritis[12][4] + 0,5282 \times tn\_atspalv * tipas * sritis[14][2] +$$
$$+7,3927 \times tn\_atspalv * tipas * sritis[14][3] - 4,5713 \times tn\_atspalv * tipas * sritis[14][4].$$

Čia t žymi tiriamų MK-ų bazinių žodžių pavartojimo atvejį. X – vektorius, kurio matavimas lygus visų laisvės laipsnių 3.9 lentelėje sumai, o komponentės susideda iš lentelėje pateiktų kintamųjų ir kokybinius veiksnius atitinkančių fiktyvių kintamųjų: $X = (bz\_lg, bz\_salt\_lg, tn\_atspalv * laipsnis, tn\_atspalv * tipas[1], tn\_atspalv * tipas[2], \ldots)$.

Pastaba. Kadangi turime tik agreguotus duomenis, neturime konkrečių atvejų, t.y. t, negalime suskaičiuoti tikimybių ir įvertinti testo kokybės.

Gautus rezultatus galima būtų interpretuoti taip.

Neigiamas parametro prie *bz_lg* įverčio ženklas rodo, kad kuo daugiau kartų tiriamos srities tekste yra pavartotas pats bazinis žodis, tuo mažesnė tikimybė, kad jis bus pavartotas ne kaip paprastas žodis, o kaip MK. Tačiau, kadangi bazinio žodžio dažnis priklauso nuo srities dydžio, kurį lemia tekstyno sudarymo taisyklės, šio rezultato dalykinė interpretacija yra neaiški. Informatyvesnis yra santykinis dydis – bazinio žodžio proporcija tekstuose, kur jis nors kartą pavartotas, – nes jis mažiau įtakojamas tekstyno sudarymo taisyklių. Kaip matyti iš parametro prie *bz_salt_lg* įverčio ženklo, ši proporcija yra teigiamai susijusi su tikimybe baziniam žodžiui būti MK-o dalimi. Vaizdžiai kalbant, kuo dažniau autorius, „linkęs vartoti“ vieną ar kitą bazinį žodį, jį vartoja, tuo didesnė tikimybė, kad tas bazinis žodis bus MK-o dalimi.

MK-ų proporcija tarp visų bazinio žodžio pavartojimų reikšmingai priklauso ir nuo teigiamo / neigiamo bazinio žodžio atspalvio, tačiau ši priklausomybė nėra tiesioginė, bet pasireiškia per sąveikas su MK-us apibūdinančia savybe (*tipas*), laipsniu ar savybe ir tekstyno sritimi drauge.

Pirmoji sąveika ($tn\_atspalv * tipas$) nėra informatyvi, nes kintamasis *tn_atspalv* yra sudarytas iš kintamojo *tipas*. Vis dėlto įdomu pastebėti, kad teigiamą atspalvį turinčių parametrų didžiausio tikėtinumo įverčių ženklas yra neigiamas, o neigiamą atspalvį turinčių – teigiamas.

Sąveika $tn\_atspalv * laipsnis$ rodo, kad MK-o emocinis atspalvis priklauso nuo bazinio žodžio laipsnio, t.y. aukštesnysis žodžio laipsnis tarsi paryškina teigiamą arba

neigiamą atspalvį ir tuo pačiu įtakoja bazinio žodžio buvimą MK-o dalimi (kaip matyti iš atitinkamo įverčio ženklo, teigiamas atspalvis didina buvimo MK-o dalimi tikimybę).

Iš trečiosios sąveikos $tn\_atspalv * tipas * sritis$ galima spręsti, kad teigiamą / neigiamą atspalvį turinčio metakomentaro bazinio žodžio pavartojimas priklauso nuo tekstyno srities (grožinė, negrožinė ir t.t.), nes skirtingose srityse santykinis teigiamą ir neigiamą atspalvį turinčių žodžių vartojimas yra skirtingas.

Kadangi neturėjome informacijos apie autorius, apie MK-ų vartojimą teko spręsti pagal kitus parametrus. Gali būti, kad turint papildomą informaciją apie autorius parinktasis modelis būtų netinkamas.

Pagal antrą MK-ų imtį (5778 MK, dažnesnių MK-ų duomenų lentelė pateikta 2 priede) buvo sudarytas logistinės regresijos modelis, kuriuo siekiama nustatyti, ar atliekantys vienodas funkcijas MK atskiroje srityje vartojami vienodu dažnumu.

Šiam tyrimui buvo apibrėžti tokie kintamieji: *MK* – metakomentaro bazinis žodis; *bz_lg* – bazinio žodžio dažnio dešimtainis logaritmas; *sritis* – keturios tekstyno dalys, bazine kategorija laikoma grožinė literatūra, kaip lietuvių literatūrinės kalbos etalonas. Skirtingose srityse įvairių MK-ų pavartojimo poreikis yra skirtingas. Natūralu manyti, kad tai yra apsprendžiama MK-ų atliekamomis funkcijomis ir atitinkamų funkcijų poreikiu tam tikrose srityse. Buvo išskirtos tokios šešios MK-ų funkcijos: *detal* – „detalumas“ (detalus / bendras, konkretus / abstraktus); *tiksl* – „tikslumas“ (tikslus / vaizdus); *pagr* – „pagrįstumas“ (pagrįstas / nepagrįstas); *svarb* – „svarba“ (pabrėžiama asmeninė vertybė ar emocinė motyvacija); *skamb* – „skambesys“ (akcentuojama išsakomos minties svarba kitiems skambumo atžvilgiu); *krit* – „kritiškumas“ (akcentuojamas neigiamas autoriaus požiūris į pateikiamą informaciją) (žr. 2 priedas).

Buvo parinktas toks (žr. 3.10 lent.) logistinės regresijos modelis.

Galima pastebėti, kad šį modelį sudarantys veiksniai ir jų sąveikos yra statistiškai reikšmingi – visos p reikšmės yra $< 0,05$.

Hosmerio ir Lemešou kriterijus, kurio statistikos p reikšmė yra didesnė už pasirinktą reikšmingumo lygmenį, t.y. $p = 0,8074 > 0,05 = \alpha$, taip pat rodo, kad parinktas modelis yra tinkamas, t.y. gana gerai suderintas su duomenimis.

Gautus rezultatus galima būtų interpretuoti taip.

Lentelė 3.10: Logistinės regresijos modelis 2

Type 3 Analysis of Effects				Type 3 Analysis of Effects			
Effect	**DF**	**Wald ChiSq**	**Pr> ChiSq**	**Effect**	**DF**	**Wald ChiSq**	**Pr> ChiSq**
MK	46	2247,9413	<,0001	$mk28 * sritis$	3	10,3710	0,0157
bz_lg	1	27,3058	<,0001	$mk38 * sritis$	3	11,5805	0,0090
$sritis * pagr$	6	24,9458	0,0003	$sritis1 * svarb$	2	11,1794	0,0037
$sritis * detal * tiksl$	33	244,6549	<,0001	$sritis3 * svarb$	2	39,2513	<,0001
$mk14 * sritis$	3	113,2982	<,0001	$sritis3 * skamb$	4	22,0648	0,0002
$mk34 * sritis$	3	24,2803	<,0001	$bz\_lg * sritis$	3	24,8285	<,0001
$mk39 * sritis$	3	29,9476	<,0001				

Kadangi išsikeltas uždavinys buvo nustatyti, ar atliekantys vienodas funkcijas MK atskiroje srityje vartojami vienodu dažnumu, tai galima manyti, kad MK-o vartojimo (santykinis) dažnumas gali būti susijęs su bazinio žodžio vartojimo dažnumu. Pavyzdžiui, jeigu dažniau vartojamas bazinis žodis, tai proporcingai dažniau vartojamas ir MK su tuo baziniu žodžiu. Tokiu atveju, tikimybė, kad atsitiktinai paimtas bazinis žodis yra MK-o dalis, būtų pastovi, ji nepriklausytų nuo to – dažniau ar rečiau vartojamas pats bazinis žodis. Tada logistiniame modelyje, kuris skirtas tai tikimybei modeliuoti, neturėtų būti narių, kurie (tiesiogiai) priklauso nuo bazinio žodžio vartojimo dažnumo. Vadinasi, prie narių su atitinkamu veiksniu *bz_lg* regresijos koeficientai turėtų būti statistiškai nereikšmingi. Bet parinktame modelyje priešingai – jie yra statistiškai reikšmingi.

Jeigu skirtingus įvairių MK-ų (santykinius) dažnius galima būtų paaiškinti vien tik išskirtomis šešiomis MK-ų funkcijomis („detalumas“, „tikslumas“, „pagrįstumas“, „svarba“, „skambesys“, „kritiškumas“), tai sudarytame modelyje neturėtų būti statistiškai reikšmingų sąveikų tarp požymio *sritis* ir *MK* identifikatoriaus, o būtų tik sąveikos tarp požymio *sritis* ir kintamųjų, nusakančių įvairias MK-ų funkcijas.

Iš modelio (3.10) lentelės matome, kad funkcijomis „detalumas“, „tikslumas“, „pagrįstumas“, „svarba“, „skambesys“ paaiškinami daugumos MK-ų vartojimo skirtumai tarp sričių, bet nepaaiškinami penkių MK-ų (su baziniais žodžiais *trumpai, tikriau, švelniai, aiškiau, atvirai*) vartojimo srityse ypatumai. Todėl akivaizdu, kad MK atlieka daugiau ir įvairesnių funkcijų, negu nagrinėtos šešios.

3.2.3 Dalinės išvados

1. Atliktas tyrimas rodo, kad ne visose tekstyno srityse vaizdingumo raiškos priemonių vartojimas yra vienodas – kintamasis, kuris koduoja srities numerį, yra statistiškai reikšmingas. Jis statistiškai reikšmingai priklauso nuo kitų faktorių, t.y. nuo teigiamo / neigiamo atspalvio ir nuo *tipo*.

2. Kuo didesnis tekstuose bazinio žodžio santykinis dažnis šaltinyje, tuo didesnė tikimybė, kad jis bus pavartotas kaip MK-o dalis.

3. Nepavyko paaiškinti MK-ų vartojimo dažnumo srityse skirtumų jų atliekamų funkcijų vartojimo tose srityse dažnumu.

4. Dėl duomenų didelio agregavimo lygio, dėl to, kad buvo paimta tik dalis MK-ų ir kad jie gali atlikti daug skirtingų funkcijų, šios išvados yra greičiau hipotezės negu galutiniai teiginiai. Tam reikia papildomų tyrimų su kruopščiai, pagal tikimybines imtis surinktais duomenimis.

Bendros išvados apie lietuvių kalbos ypatybes ir statistinius dėsningumus labai priklauso nuo to, kaip apsibrėšime, kas yra rašytinė lietuvių kalba, t.y. kuri (tekstyno) sritis ją geriausiai atspindi. Jeigu konstruosime modelius, imdami skirtingas nagrinėtų keturių sričių proporcijas, gausime skirtingas išvadas.

3.3 Statistinių metodų taikymai sakinio struktūros ir jos sudėtingumo analizėje

Sakinio ar teksto sudėtingumo įvertinimas – labai aktualus uždavinys lingvistikoje. L. Tanguy ir N. Tulechki [85] pateikia trumpą įvairių požiūrių į tekstų (kalbos) sudėtingumą apžvalgą. Pirmiausia, teksto, sakinių ar šnekamosios kalbos sudėtingumą būtina įvertinti atsižvelgiant į skaitymo ar kitų kalbos komponentų įsisavinimo lygį, sudarant supaprastintus, adaptuotus tekstus, kuriant kalbos mokymosi metodikas ir pan. Šiuo atveju teksto ar kalbos fragmento sudėtingumo įvertinimo (matavimo) tikslas yra parodyti, kiek sunku besimokančiajam jį bus įsisavinti. Analogiškai, į sakinių sudėtingumą reikia atsižvelgti ir modeliuojant kalbą, sudarant ir pagrindžiant įvairius kalbos generavimo modelius, kuriant automatinius vertėjus. Šiuo atveju teksto ar kalbos fragmento sudėtingumo įvertinimo (matavimo) tikslas yra parodyti, kiek sudėtingas yra adekvatus automatinis to fragmento apdorojimas, fragmente pateiktos aktualios informacijos išgavimas ir pateikimas reikalinga forma.

Yra pasiūlyta įvairių sudėtingumo indeksų, atspindinčių sudėtingesnių gramatinių ir sintaksinių darinių procentinę dalį tekste, bet dažnai naudojami patys paprasčiausi: vidutinis žodžio ar sakinio ilgis, kablelių procentinė dalis ir pan. Įdomesnės sudėtingumo charakteristikos yra vidutinis sakinio sintaksinis gylis ir vidutinė sintaksinių ryšių apimtis bei vidutinis tarpinių žodžių skaičius tarp sintaksiškai susietos žodžių poros. [85]

Šio tyrimo tikslas – ištirti sakinių struktūras, išreikštas kalbos dalimis, ir sakinių struktūrų, išreikštų sakinio dalimis, sudėtingumą.

3.3.1 Duomenys

Šiame darbe tiriami tekstai yra lietuvių rašytojų 1995–2011 m. išleistos vaikams skirtos prozos knygos (kurių apimtis ne mažesnė kaip 44 psl.), kurios yra saugomos Šiaulių universiteto bibliotekoje (žr. Šaltinių sąrašas 1 priede). Galutinis (smulkiausias) tyrimo elementas yra sakinys. Iš minėtų tekstų buvo renkama 3-jų pakopų atsitiktinė imtis, bet dėl naudotų specialių imties ėmimo planų 1-ojoje ir 3-iojoje pakopose realiai atsitiktinė buvo tik 2-osios pakopos imtis. Pirminius populiacijos elementus sudaro

autoriai. Laikoma, kad jie yra pagrindiniai tyrimo elementai, kad būtent jų populiacija ir yra tiriama. Kadangi ši populiacija yra maža, 36 autoriai, tai jie visi buvo paimti kaip pirminiai elementai 1-ojoje imties sudarymo pakopoje. Jeigu populiacija būtų kur kas didesnė, iš jos būtų renkama paprastoji negrąžintinė atsitiktinė imtis. Kiekvieno autoriaus parašytos knygos buvo apjungtos tarsi į vieną ištisą tekstą ir tas tekstas suskirstytas į lizdus: kiekvienas puslapis atitiko atskirą sakinių lizdą, sudarytą iš tame puslapyje prasidedančių sakinių. Kiekvienam autoriui buvo imama paprastoji grąžintinė 20 lizdų imtis (2-oji pakopa), o kiekviename į imtį įtrauktame lizde (puslapyje) buvo imamas 1-asis sakinys (3-ioji pakopa). Jeigu tas pats autoriaus tekstų puslapis pasitaikė kelis kartus, tai jame imamas kitas (gretimas) sakinys. Jeigu pagal ėmimo taisyklę imamas puslapis buvo netinkamas (jame nebuvo teksto; pvz., jis buvo tuščias arba jame buvo paveiksliukas), tada buvo imamas kitas (gretimas) puslapis. Kadangi pirmoje pakopoje į imtį yra įtraukti visi pirminiai elementai, tai formaliai tą pakopą galima ignoruoti ir laikyti, kad 2-ojoje pakopoje buvo renkama sluoksninė paprastoji grąžintinė lizdų imtis su visuose sluoksniuose vienodu lizdų imties dydžiu.

Tariant, kad 1-asis puslapio sakinys reprezentuoja likusius puslapio sakinius taip pat gerai kaip ir bet kuris kitas to puslapio sakinys ir kad autoriaus tekstai turi pakankamai daug puslapių (į tyrimą įtrauktos tik knygos, turinčios daugiau kaip 44 psl.), o tie puslapiai turi pakankamai daug sakinių, galima laikyti, kad kiekvieno autoriaus sakinių imtis apytiksliai yra (paprastoji) tarpusavyje nepriklausomų elementų imtis iš sakinių, kuriuos tas autorius galėtų parašyti, visumos (superpopuliacijos). Elementų nepriklausomumas grindžiamas tuo, kad galiojant minėtoms prielaidoms į imtį paimtus sakinius, išskyrus labai retus atvejus, skirs daugiau kaip puslapis teksto. Šis argumentas negaliotų tradicinei lizdų imčiai, kai į imtį įtraukiami visi lizdo elementai (žr. [49], 188 psl.).

Taigi, imtį („tekstyną“) sudaro 720 sakinių, kurie buvo morfologiškai ir sintaksiškai anotuoti rankiniu būdu, t.y. nurodyta kiekvieno žodžio kalbos dalis su atitinkamomis savybėmis, bei nurodyti veiksniai ir tariniai su (vientisiniuose sakiniuose arba vientisiniuose sakinio dėmenyse) jiems pavaldžiomis kitomis sakinio dalimis.

Kadangi sakinių anotavimas (net ir naudojant *Lemuoklį*) reikalauja daug kvalifikuoto darbo sąnaudų, šiame tyrime naudojama nedidelė imtis, todėl ir pats tyrimas

nėra išsamus – tai kol kas labiau žvalgomasis darbas.

3.3.2 Empirinio tyrimo rezultatai

Sakinių struktūrų analizė

Bendru atveju sakinio struktūrą galima būtų aprašyti grafu. Lietuvių kalbos sakinių grafines struktūras išsamiai aprašė D. Šveikauskienė (žr. [82], [81]).

Dėl mažo duomenų kiekio minėtų sakinio struktūrų statistinė analizė negalima – susiduriama su išretintų duomenų (angl. *sparse data*) problema, su kuria susiduriama lingvistikoje nagrinėjant vien tik žodžius (kurių bet kurioje kalboje yra labai daug, bet didelė jų dalis vartojama labai retai), jau nekalbant apie žodžių poras, o juo labiau – sakinius.

Atliekant sakinių struktūrų analizę išretintų duomenų problema buvo sprendžiama perkoduojant anotuotus sakinius ir nagrinėjant iš veiksmažodžio ir daiktavardžio sudarytą sakinio „karkasą“, kuris sąlyginai vadinamas *kodu*. Kodas aprašo redukuotą (supaprastintą) sakinio struktūrą. Redukcijos (supaprastinimo) lygis priklauso nuo to, kokios savybės ir kaip detaliai yra koduojamos.

Sakinio kodas sudaromas kiekvieną sakinio žodį pakeičiant simboliu (raide arba skaičiumi), kuris *koduoja* to žodžio, kaip sakinio sudėtinio elemento, vienokią ar kitokią savybę. Tokiu būdu sakinys tampa tarsi „žodžiu“, kurio „alfabetas“ yra sudarytas iš nagrinėjamas savybes koduojančių simbolių. Žemiau yra pateikti 4 tokių „alfabetų“ pavyzdžiai ir jiems paskaičiuotos statistikos.

Sudaryti tokių tipų sakinių struktūrų kodai:

I — išlaikant anotuoto sakinio tvarką, paliekami tik daiktavardžiai ir veiksmažodžiai, o visos kitos kalbos dalys pakeičiamos simboliu „–“, keli iš eilės einantys „–“ simboliai apjungiami;

Ia — gaunamas iš I tipo kodo, apjungiant kelis iš eilės einančius daiktavardžius ar veiksmažodžius;

II — sudaromas panašiai kaip ir I tipo, tik išsaugoma informacija apie daiktavardžio linksnį, t.y. vietoje daiktavardžio įrašomas linksnio numeris (vardininkas – 1, kilmininkas – 2 ir t.t.);

IIa — gaunamas iš II tipo kodo, apjungiant kelis iš eilės einančius vienodus simbolius.

3.11 lentelėje pateikti koduočių pavyzdžiai. Čia D – daiktavardis, V – veiksmažodis, B – būdvardis, I – įvardis, n – kita kalbos dalis.

Lentelė 3.11: Sakinių struktūrų koduočių pavyzdžiai

	Sakinio struktūra (be linksnių)	Koduotė I	Koduotė Ia	Sakinio struktūra (su linksniais)	Koduotė II	Koduotė IIa
1	nVDnD	-VD-D	-VD-D	nV5n5	-V5-5	-V5-5
2	DDVBnnBD	DDV-D	DV-D	21VBnnB1	21V-1	21V-1
3	nIDDnnDVVnDnn	-DD-DVV-D-	-D-DV-D-	nI11nn1VVn2nn	-11-1VV-2-	-1-1V-2-
4	IVVnV	-VV-V	-V-V	IVVnV	-VV-V	-V-V
5	DVD	DVD	DVD	1V4	1V4	1V4
6	VD	VD	VD	V7	V7	V7
7	BDVnn	-DV-	-DV-	B1Vnn	-1V-	-1V-
8	nIVnnn	-V-	-V-	nIVnnn	-V-	-V-
9	IDDDnDVnDnD	-DDD-DV-D-D	-D-DV-D-D	I411n1Vn4n4	-411-1V-4-4	-41-1V-4-4
10	DVnDDnnnV	DV-DD-V	DV-D-V	1Vn24nnnV	1V-24-V	1V-24-V

3.12 lentelėje pateikti kodų pasikartojimų dažniai. Iš šios lentelės matome, kad 257 I tipo sakinio kodai pasitaikė tik po vieną kartą, o II tipo, kai yra atsižvelgiama į daiktavardžio linksnį, po vieną kartą sutinkamos net 407 struktūros (daugiau nei pusė visų sakinių). Visais atvejais yra struktūrų, sutinkamų po 30 ir daugiau kartų.

Lentelė 3.12: Sakinių struktūrų kodų pasikartojimo dažniai

Kodų pasikartojimo dažnis	1	2	3	4	5	6	7	8	9	10	11	12	13	14
I tipo kodų skaičius	257	31	9	11	3	6	4	1		1	1	2	1	1
Ia tipo kodų skaičius	120	36	8	6	5	3	3	2	2	1	3	1		1
II tipo kodų skaičius	407	30	9	6	2	2	5		2		1			
IIa tipo kodų skaičius	355	37	12	4	4	3	3		4	1	1			

Kodų pasikartojimo dažnis	15	16	17	18	19	21	23	26	27	30	32	34	38	40
I tipo kodų skaičius		1		2			1			1	1	1		
Ia tipo kodų skaičius	1		3		1	1	2	1	1	1			1	1
II tipo kodų skaičius		1		2						1		1		
IIa tipo kodų skaičius			1		2					1			1	

Vienas iš kertinių klausimų: ar aukščiau aptartu būdu sudarytiems „žodžiams“ galioja Zipfo-Mandelbroto dėsnis? Šiuo atveju jis reikštų, kad didelė dalis tekste

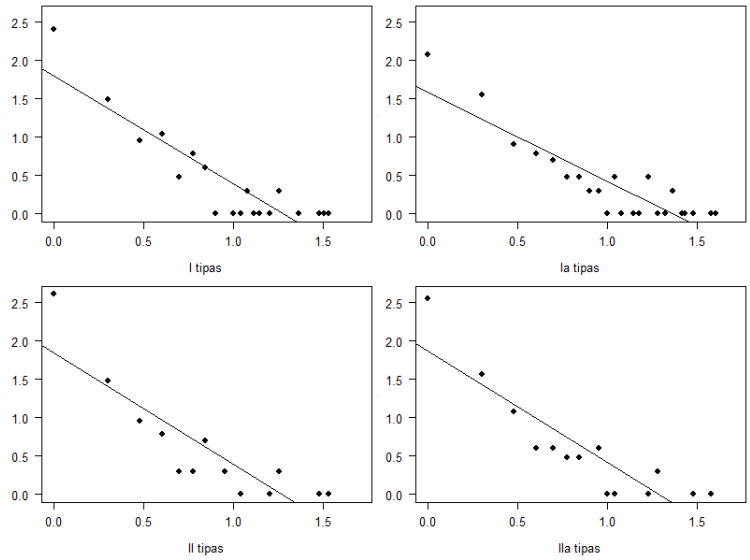
sutinkamų sakinių turi beveik standartinę struktūrą (nagrinėjamu požiūriu, kuri apibrėžia naudojamą kodavimo taisyklę), iš kitos pusės – dauguma sakinio struktūrų yra (beveik) unikalios, jos tekstuose pavartotos tik vieną kartą (1 ar 2 kartus).

Zipfo dėsnio parametrai (žr. 3.13 lent.) sakinio struktūros kodams skaičiuojami pagal (1.1) formulę. Šia formule užrašytą dėsnį paprasčiau interpretuoti log-log skalėje. Tada parametrų vertinimui galima taikyti mažiausių kvadratų metodą.

Lentelė 3.13: Zipfo dėsnio parametrai

Kodas	$\log K$	γ
I	1,795	-1,405
Ia	1,585	-1,166
II	1,845	-1,457
IIa	1,863	-1,453

3.4 paveikslėlyje pavaizduoti I, Ia, II ir IIa tipų sakinių struktūrų grafikai; čia x ašyje yra kodų dažnių logaritmas, y ašyje – kodų dažnių pasikartojimo dažnių logaritmas.



Pav. 3.4: Sakinio struktūros kodų dažnių log-log grafikai

Pastaba. Galima rasti analogijų su 2.5.2 paragrafe išdėstyta medžiaga. Ten minimas antrasis Zipfo dėsnis (2.30), susiejantis žodžio formų, pasitaikiusių tekste m kartų, kiekį $\widehat{V}_m$ su m. Pats Zipfo dėsnis susieja stebėtų žodžių su rangu r kiekį f_r su rangu r (žr. (1.1) formulę).

3.4 pav. matome, kad tinkamai (mažiausių kvadratų metodu) parinktos tiesės gana gerai aprašo logaritminėje skalėje (t.y. log-log skalėje) atidėtų porų (r, f_r) duome-

nis. Čia r yra „žodžio“ formos rangas, o f_r yra „žodžio“ formų, kurių rangas yra r, kiekis. Visais atvejais, išskyrus pačią paprasčiausią (labiausiai redukuojančią) koduotę Ia, Zipfo tiesės parametrai labai panašūs (žr. 3.13 lent.).

Vadinasi, gerai „išmokus“ identifikuoti ir analizuoti (anotuoti, versti ir t.t.) paprasčiausios struktūros sakinius, galima automatiškai apdoroti nemažą dalį teksto sakinių. Paprastomis struktūromis laikant tas, kurios pasitaikė, tarkime, ne mažiau kaip 10 kartų, pagal II koduotę galima identifikuoti 17,64% sakinių (apytikslis proporcijos 95% pasikliautinasis intervalas (žr. (2.1) ir (2.2)) yra nuo $14,86\%$ iki $20,42\%$), o pagal I koduotę – net 33,75% sakinių (apytikslis proporcijos 95% pasikliautinasis intervalas yra nuo $30,3\%$ iki $37,2\%$).

Sakinio struktūros sudėtingumo analizė

Remiantis užsienio patirtimi, ir Lietuvoje, modeliuojant kalbą, populiariausiu kalbos modeliu išlieka 2-os eilės paslėptasis Markovo modelis (angl. *Hidden Markov Model*), kuris remiasi trigramomis, t.y. trijų iš eilės einančių žodžių statistika tekste (žr. [70], [90]). Dėl ankstesniuose skyriuose minėto lietuvių kalbos specifiškumo ir sudėtingumo vargu ar trigramomis pagrįsti algoritmai, tinkantys, pvz., anglų kalbai, yra perspektyvūs, taikant lietuvių kalbos tyrimuose ir taikomojoje lingvistikoje.

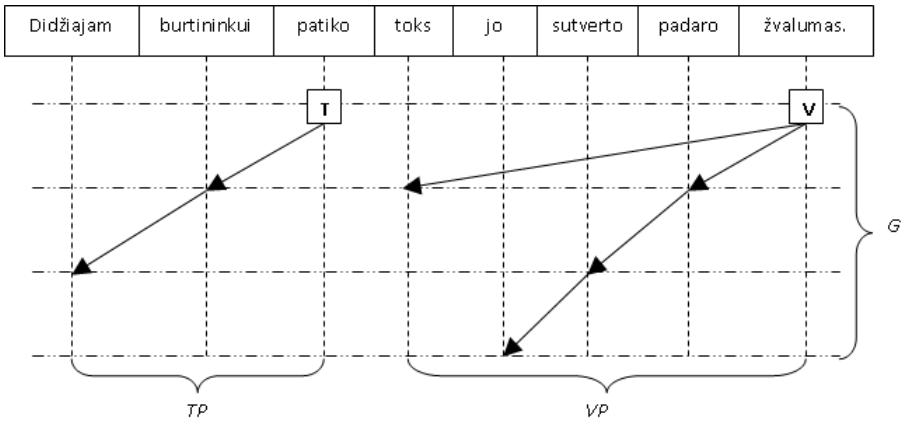
Sakinio struktūros sudėtingumą galima suvokti kaip sakinio gylį ir sakinio plotį pagal pagrindinėms sakinio dalims (veiksniui ir tariniui) pavaldžius žodžius. Ypač svarbi sakinio struktūros sudėtingumo charakteristika – sakinio gylis, kuris parodo, kiek yra betarpiškai susijusių žodžių (aktualu – ar jų nėra daugiau nei 3).

Žodis, iš kurio keliamas klausimas kitam žodžiui, gali būti laikomas pastarojo žodžio „tėvu“. Pastarasis žodis taip pat gali būti kurio nors kito žodžio „tėvu“ ir t.t. Taip kiekvienam sakinio žodžiui galima sudaryti jo „geneologiją“ – jo „tėvų“ seką. Tokio „geneologinio“ medžio šaknis bus veiksnys arba tarinys, o maksimalus atstumas nuo medžio šaknies iki (k-osios) kabančios viršūnės parodo sakinio gylį. Kitaip tariant, kiekvienam sakinio žodžiui z_i galima priskirti jo „geneologijos“ seką sudarančių žodžių skaičių n_i. Tada sakinio gylis yra $\max(n_i)$.

Sakinio plotis suprantamas kaip maksimalus skirtumas tarp vieno „tėvo“ „vaikų“ eilės numerių, įskaitant ir „tėvą“, sakinyje: $N_{[\max]} - N_{[\min]} + 1$, t.y. žodžių skaičius

atitinkamoje struktūroje. Pavyzdžiui, jei žodis yra pats sau „tėvas“ ir neturi jam pavaldžių žodžių, laikoma, kad tokios grupės plotis yra vienetinis.

Atotrūkis tarp pagrindinių sakinio dalių suprantamas kaip tarpinių žodžių skaičius tarp veiksnio ir tarinio: $|tarinio.numeris - veiksnio.numeris| - 1$. Kai veiksnys ir tarinys yra greta, tada atotrūkis lygus 0.



Pav. 3.5: Sakinio struktūros gylis ir plotis

3.5 schemoje matome, kad (maksimalus) atstumas tarp pagrindinių sakinio dalių, t.y. atotrūkis tarp veiksnio ir tarinio lygus 4. Tarinio struktūros plotis $TP = 3$. Veiksnio struktūros plotis $VP = 5$. Maksimalus sakinio gylis $G = 3$ (tarinio struktūros gylis lygus 2, maksimalus veiksnio struktūros gylis lygus 3).

Šiame tyrime nagrinėti tik tie sakiniai, kurių struktūrą galima pavaizduoti medžiu, t.y. 327 (iš turimos 720 sakinių imties) vientisiniai sakiniai, kuriems priskirti „tėvai“, t.y. pažymėta, iš kurio žodžio keliamas klausimas. Šiems sakiniams apskaičiuotas jų gylis ir plotis.

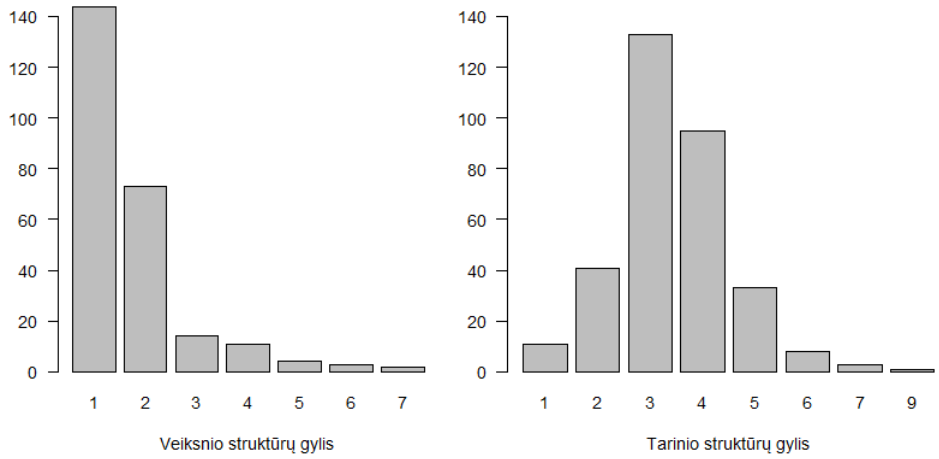
Sakinio gylis yra maksimalus žodžių grandinės, sudarytos iš vienas kitam betarpiškai pavaldžių žodžių, ilgis sakinyje. Tirtuose duomenyse maksimalus sakinio gylis buvo lygus 9.

Sakinio gylio dažnių lentelėje 3.14 pateiktas bendras sakinio gylis (pagal aukščiau esantį apibūdinimą) ir veiksnio bei tarinio struktūrų gyliai.

Veiksnio ir tarinio struktūrų gylio dažniai pavaizduoti 3.6 pav. Matome, kad daugumą tarinio struktūrų sudaro 3–4 žodžiai, o veiksnio struktūros dažniausiai yra iš 1–2 žodžių. Tai rodo, kad tariniai yra linkę prisijungti ilgesnes jiems pavaldžių žodžių grandines. Pastebėtina, kad apie 2/3 sakinių struktūros yra kompaktiškos, sudaro

Lentelė 3.14: Sakinio gylio dažniai

Gylis	1	2	3	4	5	6	7	9
Sakinio gylio (bendro) dažniai		41	129	102	38	11	5	1
Veiksnio struktūros gylio dažniai	144	73	14	11	4	3	2	
Tarinio struktūros gylio dažniai	11	41	133	95	33	8	3	1



Pav. 3.6: Veiksnio ir tarinio struktūrų gylis

tarpusavyje nepersidengiančius vieno „tėvo“ pavaldžių žodžių blokus.

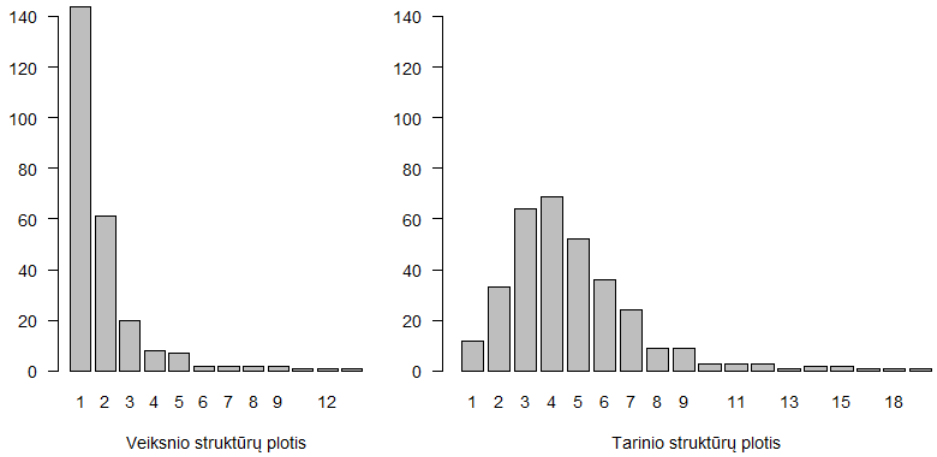
3.15 lentelėje pateikiamas bendras sakinio plotis ir sakinio plotis, kai jis skaičiuojamas atskirai tarp veiksniui priklausančių ir tariniui priklausančių žodžių (taip pat žr. 3.7 pav.).

Lentelė 3.15: Sakinio pločio dažniai

Plotis	1	2	3	4	5	6	7	8	9	10	11
Sakinio pločio (bendro) dažniai		33	50	68	42	50	29	13	14	7	5
Veiksnio struktūros pločio dažniai	144	61	20	8	7	2	2	2	2		1
Tarinio struktūros pločio dažniai	12	33	64	69	52	36	24	9	9	3	3

Plotis	12	13	14	15	16	17	18	20	21	23	
Sakinio pločio (bendro) dažniai	2	4	2	2	2	2		1		1	
Veiksnio struktūros pločio dažniai	1		1								
Tarinio struktūros pločio dažniai	3	1	2	2		1	1		1		

Matome, kad veiksnio struktūrai priklausantys žodžiai gali būti nuo veiksnio nutolę net per 14 žodžių, nors dauguma jų yra greta veiksnio (plotis lygus 1), o tarinio struktūrai priklausantys žodžiai gali būti nutolę nuo tarinio net per 21 žodį ir daugumos tarinio struktūrų plotis yra 2–4 žodžiai.



Pav. 3.7: Veiksnio ir tarinio struktūrų plotis

Turint omenyje aukščiau minėtą trigramų taikymą lietuvių kalbos tyrimuose, svarbu įvertinti, kokia lietuviškuose tekstuose (šiuo atveju, vaikams skirtoje literatūroje) yra proporcija sakinių, kuriuose visi betarpiškai sintaksiškai susiję žodžiai yra nutolę vienas nuo kito ne didesniu kaip 3 atstumu, t.y. per vieną tarpinį žodį.

Nagrinėjant bendrą sakinio plotį gauta, kad proporcija sakinių, kurių plotis yra ne mažesnis nei 4, tarp visų anotuotų (šio tyrimo) sakinių sudaro net 74,62% (apytikslis proporcijos 95% pasikliautinasis intervalas (žr. (2.1), (2.2)) yra nuo 70% iki 79, 3%). Ypač tai yra būdinga tarinio struktūroms, kurių tiek plotis, tiek gylis daugeliu atvejų yra didesnis, negu reikia, norint aprašyti trigramomis. Akivaizdu, kad lietuvių kalbos sintaksinės (sakinių) struktūros yra per daug sudėtingos, kad jas būtų galima aprašyti trigramomis, nebent visi sakiniai būtų sudaryti tik iš veiksnio struktūrų.

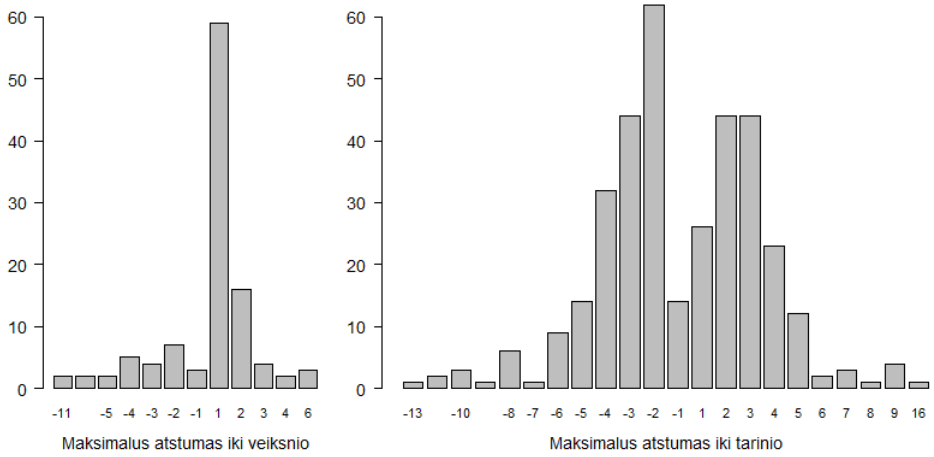
Taip pat svarbi kita sudėtingumo charakteristika – maksimalus žodžio atstumas iki jo „tėvo“, atsižvelgiant į tai, kurioje „tėvo“ pusėje tas žodis yra: $\max(tevo.numeris - vaiko.numeris)$. Rezultatai pateikti 3.16 lentelėje (bendras atstumas bei atstumas iki veiksnio ir iki tarinio) ir 3.8 pav.

Galima pastebėti, kad dauguma veiksniui pavaldžių žodžių būna jo kairėje, dažniausiai šalia. Tarinio atžvilgiu jam pavaldūs žodžiai dažniau būna išsidėstę dešinėje (54%), dažniausiai per vieną žodį nutolę nuo tarinio.

Proporcija sakinių, kurių maksimalus atstumas absoliutine reikšme tarp žodžio ir jo „tėvo“ yra ne mažesnis už 3, tarp visų nagrinėtų sakinių sudaro 66,03% (apytikslis proporcijos 95% pasikliautinasis intervalas yra nuo 61, 2% iki 70, 9%).

Lentelė 3.16: Maksimalus žodžio atstumas iki jo „tėvo“

Maks. atstumas	−14	−13	−12	−11	−10	−9	−8	−7	−6	−5	−4	−3	−2
Dažniai													
Bendrai	2	1	1	4		3	7	3	9	22	35	46	50
Iki veiksnio				2					2	2	5	4	7
Iki tarinio		1	2		3	1	6	1	9	14	32	44	62
Maks. atstumas	−1	1	2	3	4	5	6	7	8	9	10	16	
Dažniai													
Bendrai	8	31	36	44	28	15	11	5	1	2	3	1	
Iki veiksnio	3	59	16	4	2		3						
Iki tarinio	14	26	44	44	23	12	2	3	1	4		1	



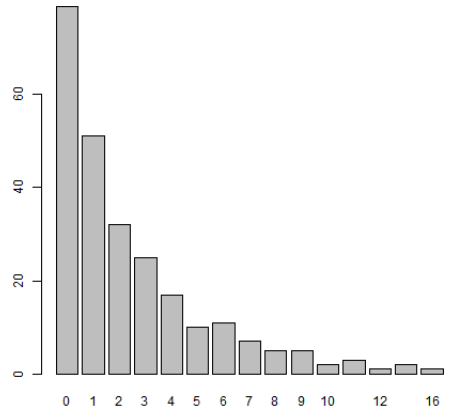
Pav. 3.8: Maksimalus atstumas iki veiksnio ir iki tarinio

Maksimalus atotrūkis tarp veiksnio ir tarinio, jei sakinyje yra keli veiksniai arba keli tariniai, – tai maksimali visų veiksnio–tarinio porų tarpusavio atotrūkių reikšmė. Jei sakinyje yra tik veiksnys arba tik tarinys, tokiu atveju atotrūkis neskaičiuojamas. Dėl pastarosios priežasties atotrūkis tarp veiksnio ir tarinio buvo skaičiuotas tik 251 sakinyje.

3.17 lentelėje ir 3.9 pav. pateikti maksimalaus atotrūkio tarp veiksnio ir tarinio nagrinėjamuose sakiniuose rezultatai.

Lentelė 3.17: Maksimalus atstumas tarp veiksnio ir tarinio

Atstumas	0	1	2	3	4	5	6	7	8	9	10	11	12	13	16
Dažnis	79	51	32	25	17	10	11	7	5	5	2	3	1	2	1



Pav. 3.9: Maksimalus atstumas tarp veiksnio ir tarinio

Matome, kad dažniausiai veiksnys ir tarinys yra šalia ($31,47\% \pm 5,75\%$) arba tarp jų yra vienas žodis ($20,32\% \pm 4,98\%$), o didėjant atstumui tarp veiksnio ir tarinio, dažnis beveik tolygiai mažėja. Įdomu pastebėti, kad vieną kartą pasitaikė ir toks sakinys, kuriame tarp veiksnio ir tarinio buvo net 16 žodžių.

3.3.3 Dalinės išvados

Ištyrus sakinių struktūras ir jų sudėtingumą galima daryti tokias išvadas.

1. Sakinių struktūrose žodžius užkodavus atitinkamais simboliais, gautiems „žodžiams“ galioja Zipfo dėsnis.

2. Gerai „išmokus“ identifikuoti ir analizuoti (anotuoti, versti ir t.t.) paprasčiausios struktūros sakinius, galima automatiškai apdoroti reikšmingą dalį (daugiau negu 17%) teksto sakinių.

3. Apskaičiuotos sudėtingumo charakteristikos rodo, kad sakinio struktūros grafas dažnai būna gana sudėtingas ir norint tinkamai jį aprašyti naudojant tik trijų gretimų sakinyje žodžių (trigramų statistiką) informaciją nepakanka, kadangi, pavyzdžiui, tik 25,38% sakinių plotis yra ne didesnis negu 3.

3.4 Lietuvių kalbos tekstų struktūrinių skirstinių analizė

Šiame tyrime pagrindinis tyrimo objektas yra teksto autorius – tiriama grožinės literatūros vaikams autorių populiacija. Būtent tekstinių dokumentų autoriai yra baziniai populiacijos, kuria domimės, elementai. Tariama, kad tekstinių dokumentų rinkinio (tekstyno) heterogeniškumą sąlygoja autorių preferencijų ir pasirinkimo heterogeniškumas. Sudarant tiriamos autorių populiacijos statistinį modelį taikoma Bajeso metodologija.

Tyrimo tikslas – ištirti įvairių autorių, rašančių įvairaus žanro kūrinius, sukurtų tekstų („tekstyno“) ir vartojamų žodžių (žodyno) ypatumus, panašumus ir variabilumą.

3.4.1 Duomenys ir kintamieji

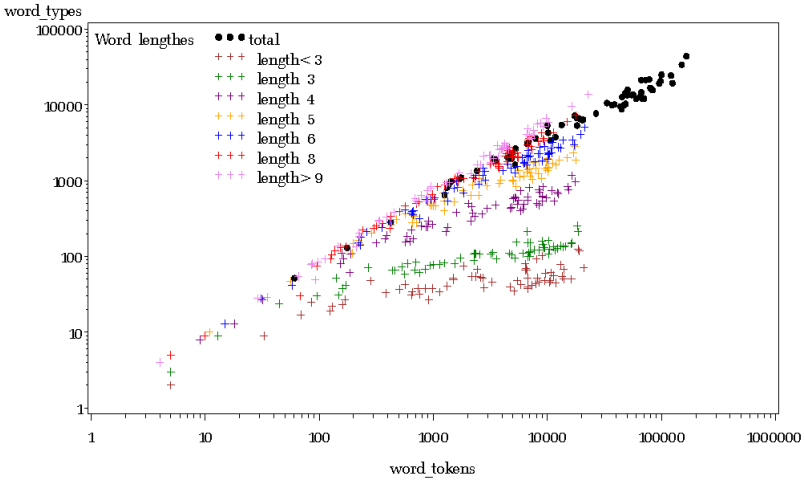
Tyrime naudojama tekstų, parašytų lietuvių kalba, duomenų aibė. Duomenų rinkinį sudaro mokyklai (5–8 kl.) rekomenduojami lietuvių ir užsienio autorių grožinės literatūros kūriniai (romanai, apsakymai, poemos, eilėraščiai, pjesės), paimti iš laisvai prieinamos skaitmeninės bibliotekos (*http://ebiblioteka.mkp.emokykla.lt/*). Susidarytą „tekstyną“ sudaro 63-ų autorių parašytų 80 tekstinių dokumentų (knygų, kūrinių); iš viso tekstuose buvo 2567290 žodžių, iš jų 206453 skirtingi žodžiai (žodžių formos).

Šiame tyrime autoriai laikomi skirtingais ir nepriklausomais tekstinių duomenų šaltiniais. Todėl skirtingi $s \in S$ reiškia skirtingus autorius. Tegul W_s žymi autorius s žodyną. Laikoma, kad bendrasis žodynas $\mathcal{W} = \cup_{s \in S} W_s$.

Aiškinamųjų kintamųjų $x_w(s)$ vektorius susideda iš dviejų kategorinių kintamųjų s ir ℓ_w bei jų sąveikų. Kategorinis kintamasis $s \in S$ turi $|S| = 63$ kategorijas, kategorinis kintamasis $\ell_w \in \{2, \ldots, 10\}$ yra žodžių formos w ilgio grupė. Grupė, kurios $\ell_w = 2$, susideda iš žodžių formų, kurių ilgis 1 arba 2, žodžių formos grupėje, kurios $\ell_w = 10$, turi 10 ir daugiau raidžių. Likusiose grupėse žodžių formų ilgis ir grupės numeris sutampa.

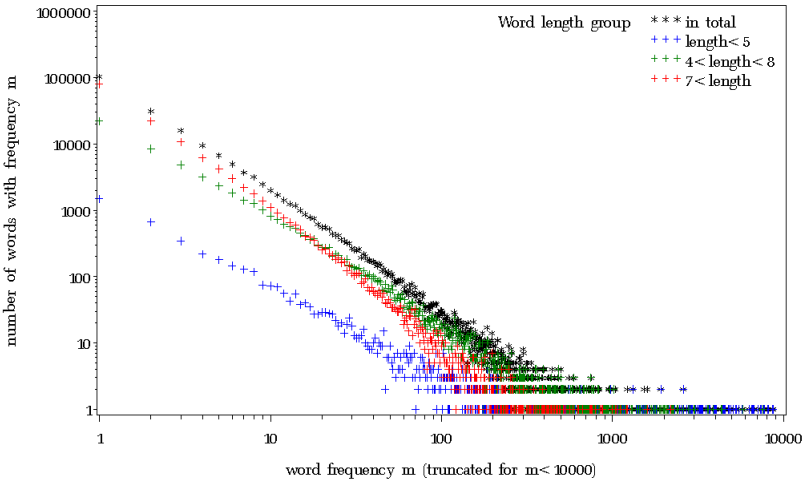
Taip pat naudojamas išvestinis požymis „gimtasis“ (*native*), nurodantis, ar autorius rašo gimtąja lietuvių kalba, ar jis/ji yra užsienio rašytojai, t.y. kūrinys yra verstinis.

3.4.2 Empirinio tyrimo rezultatai



Pav. 3.10: Herdano-Hipso dėsnis skirtingo ilgio ir visiems žodžiams

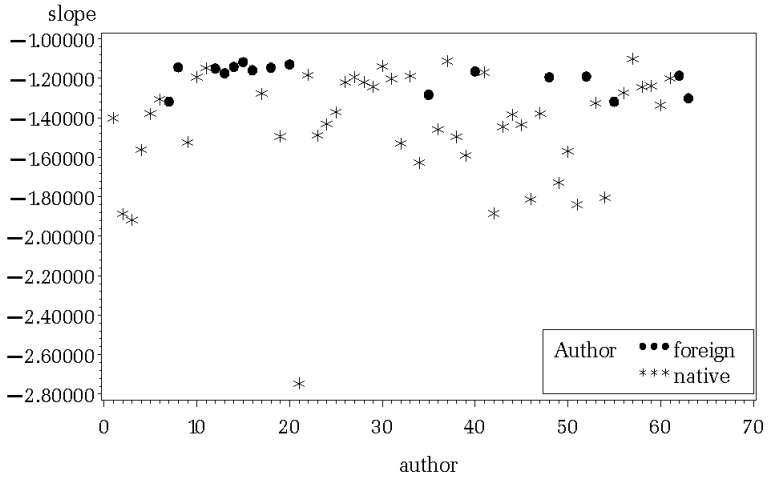
3.10 pav. pateikta Herdano-Hipso dėsnio (žr. 2.5.2 paragrafas) iliustracija, kuri taip pat parodo ir teksto bei žodyno didumų pasiskirstymą tarp autorių. Duomenys log-log skalėje pakankamai gerai atitinka tiesią liniją įvairaus ilgio (net ir gana trumpiems) tekstiniams dokumentams. Tačiau matome, kad trumpesni žodžiai (kai $\ell_w < 6$) nukrypsta nuo šio dėsnio, jo netenkina. Dėsnis galioja tik labai ilgiems žodžiams. Dažniausi – 6 raidžių ilgio žodžiai, kurie jau beveik tenkina šį dėsnį.



Pav. 3.11: Zipfo antrasis dėsnis skirtingo ilgio ir visiems žodžiams

Zipfo antrojo dėsnio (žr. (2.30) 2.5.2 paragrafe) grafinės iliustracijos pateikiamos

3.11 pav. Vėlgi galima pastebėti, kad ir Zipfo dėsnį geriausiai tenkina ilgesni (kai $\ell_w > 7$) žodžiai.



Pav. 3.12: Zipfo antrojo dėsnio (2.30) posvyrio įverčių sklaidos diagrama

Tekstinių duomenų heterogeniškumą iliustruoja ir antrąjį Zipfo dėsnį atitinkančios tiesės posvyrio įverčių, gautų formaliai pritaikius mažiausių kvadratų metodą, sklaidos diagrama pagal autorius, pateikta 3.12 pav., kurioje skirtingai pažymėti lietuvių ir užsienio autoriai. Kaip matome, gaunami gana skirtingi posvyrio įverčiai. Užsienio autorių posvyro įvertinys, savaime suprantama, yra savo absoliutine reikšme gerokai mažesnis už lietuvių autorių posvyro įvertinius. Tai leidžia daryti natūralią prielaidą, kad verstinėje literatūroje yra kur kas mažiau retų žodžių negu lietuviškoje.

3.4.3 Empirinis Bajeso metodas

E. V. Khmaladze [43] pastebėjo, kad struktūrinis skirstinys gali būti traktuojamas kaip latentinis mišinio skirstinys empiriniame Bajeso metode.

Šiame paragrafe pateikiamas paprastas ir patogus skaičiavimams, bet gana informatyvus Bajeso statistinis modelis ir pasitelkus parametrinį empirinį Bajeso metodą jis pritaikomas žodžių formų ir jų struktūrinių skirstinių statistiniam vertinimui.

Latentinio ir struktūrinio skirstinių modeliai, pristatyti 2.5.1 ir 2.5.2 paragrafuose, ignoruoja turimą papildomą informaciją $\{x_w(s), w \in \mathcal{W}, s \in S\}$. Čia pateikiamas Bajeso modelis išnaudoja papildomą informaciją, darant prielaidą, kad žodžių formos w šaltinyje s $y_w(s)$ dažnio sąlyginis skirstinys, kai žinoma aiškinančiųjų kintamųjų $x_w(s)$

reikšmė x, priklauso nuo x per skaliarines funkcijas $p = p(x)$, $\mu = \mu(x)$ ir $\kappa = \kappa(x)$. Tiksliau,

$$\begin{aligned}
[y_w(s) \mid z_w(s) = 0] &= 0, && (3.2)\\
[z_w(s) \mid x_w(s) = x] &\stackrel{\mathcal{L}}{=} Binomial(1, 1 - p(x)), && (3.3)\\
[y_w(s) \mid z_w(s) = 1, \lambda_w(s) = \lambda] &\stackrel{\mathcal{L}}{=} Poisson(\lambda), && (3.4)\\
[\lambda_w(s) \mid x_w(s) = x] &\stackrel{\mathcal{L}}{=} Gamma(\kappa(x), \mu(x)). && (3.5)
\end{aligned}$$

Čia $\{z_w(s), w \in \mathcal{W}, s \in S\}$ yra latentiniai binariniai atsitiktiniai kintamieji, sąlyginai (tarpusavyje) nepriklausomi, kai yra duotos aiškinančiųjų kintamųjų $\{x_w(s), w \in \mathcal{W}, s \in S\}$ reikšmės. Savo ruožtu, $\{y_w(s), w \in \mathcal{W}, s \in S\}$ yra atsitiktiniai kintamieji, sąlyginai (tarpusavyje) nepriklausomi, kai latentinių teigiamų atsitiktinių kintamųjų $\{\lambda_{sw}, w \in \mathcal{W}, s \in S\}$ reikšmės ir aiškinančiųjų kintamųjų reikšmės yra laikomos fiksuotomis. $Gamma(\kappa, \mu)$ žymi gama skirstinio dėsnį su vidurkiu $\mu > 0$, dispersija $\kappa\mu^2$ ir skirstinio tankiu

$$g\left(u \mid \kappa, \mu\right) := \frac{u^{1/\kappa - 1} \exp\{-u/(\mu\kappa)\}}{\Gamma(1/\kappa)\ (\mu\kappa)^{1/\kappa}}, \quad u > 0.$$

Latentinio kintamojo $z_w(s)$ reikšmė 0 parodo, kad žodžių forma w yra nevartotina šaltinyje s, $p(s)$ yra tikimybė, kad žodžių forma w su $x_w(s) = x$, bus nevartotina. Latentinis kintamasis $\lambda_w(s)$ yra žodžių formos w šaltinyje s vidutinis (tikėtinas) dažnis.

Prielaidą apie sąlyginį y'ų nepriklausomumą nėra realistiška. Tačiau asimptotinės statistikos požiūriu ši prielaida prilygsta y'ų silpnos (sąlyginės) priklausomybės sąlygai, kuri atrodo gana pagrįsta prielaida, kai nagrinėjami žodžių skaičiaus skirstiniai pakankamai dideliuose tekstiniuose dokumentuose.

Marginalinis (ir sąlyginis, kai yra duotos x'ų reikšmės) y'o skirstinys gaunamas išintegruojant nestebimus atsitiktinius kintamuosius z ir λ

$$\begin{aligned}
Q_k(x) &:= \mathbf{P}(y_w(s) = k \mid x_w(s) = x)\\
&= p(x)\mathbb{I}\{k = 0\} + (1 - p(x)) \int_0^\infty \Pi_k(u)\ g\left(u \mid \kappa(x), \mu(x)\right) \mathrm{d}u, \qquad (3.6)
\end{aligned}$$

kuris iš tikrųjų yra išsigimusio taške 0 skirstinio ir neigiamo binominio skirstinio q, su vidurkio parametru $\mu = \mu(x)$ ir dispersijos parametru $\kappa = \kappa(x)$, mišinys, su

atitinkamomis apriorinėmis tikimybėmis $p(x)$ ir $1-p(x)$:

$$q(k \mid \mu, \kappa) := \frac{\Gamma(1/\kappa + k)}{\Gamma(1/\kappa)k!} \left(\frac{\mu}{1+\mu}\right)^k \left(\frac{1}{1+\mu}\right)^{1/\kappa}, \quad k = 0, 1, \ldots.$$

(3.2)–(3.5) lygtys apibrėžia jungtinį hierarchinį Bajeso modelį su apriioriniu skirstiniu, kuris yra dviejų komponenčių Gama-Puasono mišinių sandauga. Apriorinis skirstinys yra nusakomas tarpusavyje nepriklausomomis poromis nežinomų parametrų $(z_w(s), \lambda_w(s))$, $w \in \mathcal{W}, s \in S$, kurie turi skirstinius

$$\begin{aligned} [z_w(s) \mid p_w(s) = p] &\stackrel{\mathcal{L}}{=} Binomial(1, 1-p), \\ [\lambda_w(s) \mid z_w(s) = 1, \mu_{ws} = \mu,\, \kappa_{ws} = \kappa] &\stackrel{\mathcal{L}}{=} Gamma(\kappa, \mu), \\ [\lambda_w(s) \mid z_w(s) = 0, \mu_{ws} = \mu,\, \kappa_{ws} = \kappa] &= 0, \end{aligned}$$

priklausančius nuo hiperparametrų $p_{ws} := p(x_w(s))$, $\mu_{sw} := \mu(x_w(s))$, $\kappa_{sw} := \kappa(x_w(s))$, $w \in \mathcal{W}, s \in S$. Vadinasi, nežinomų parametrų aposteriorinis skirstinys, kuris remiasi imtimi $y\{D\} := \{y_w(s), w \in \mathcal{W}, s \in D\}$, $D \subset S$, vėlgi yra dviejų komponenčių Gama-Puasono mišinys su atnaujintais hiperparametrais

$$\hat{p}_{sw} = \hat{p}_{sw}(y\{D\}) := \frac{p_{sw}\mathbb{I}\{y_w(s) = 0\}}{p_{sw}\mathbb{I}\{y_w(s) = 0\} + (1 - p_{sw})\, q(y_w(s) \mid \mu_{sw}, \kappa_{sw})}, \tag{3.7}$$

$$\hat{\mu}_{sw} = \hat{\mu}_{sw}(y\{D\}) := \frac{\mu_{sw}(1 + \kappa_{sw} y_w(s))}{1 + \kappa_{sw}\mu_{sw}}, \tag{3.8}$$

$$\hat{\kappa}_{sw} = \hat{\kappa}_{sw}(y\{D\}) := \frac{\kappa_{sw}}{1 + \kappa_{sw} y_w(s)}, \quad s \in D. \tag{3.9}$$

Pagrindinė Bajeso statistikos problema yra apriorinio skirstinio parinkimas. Mūsų atveju tai reiškia hiperparametrų $p_{sw}, \mu_{sw}, \kappa_{sw}$, $w \in \mathcal{W}, s \in S$, arba funkcijų $p(\cdot), \mu(\cdot)$ ir $\kappa(\cdot)$ parinkimą. Pagal empirinį Bajeso metodą, hiperparametrai įvertinami priderinant marginalinius y'ų skirstinius (3.6) prie turimų duomenų $y\{D\}$. Tarus, kad funkcijos $p(\cdot), \kappa(\cdot)$ ir $\mu(\cdot)$ turi tam tikrą parametrinę formą, šį uždavinį pavyksta išspręsti. Pavyzdžiui, jeigu $p(\cdot)$ ir $\mu(\cdot)$ priklauso nuo tiesinių prediktorių su, atitinkamai, *logit* ir logaritmine jungties funkcijomis, o $\kappa(\cdot)$ yra konstanta, tada (3.2)–(3.5) lygtys apibrėžia K-mišinio skirstinių regresijos modelį (žr. [19]), taip pat gerai žinomą ir ekonometrijoje kaip neigiamas binominis skirstinys su pertekliniais nuliais (angl. *Zero-Inflated Negative Binomial*) regresijos modelis. M. Jansche [36] siūlė modelius su pertekliniais nuliais naudoti modeliuojant žodžių skaičiaus tekstuose skirstinius. Šio modelio parinkimui

gali būti taikoma standartinė statistikos programinė įranga (R, SAS (nuo 9.2 versijos), STATA). Nežinomų parametrų įvertiniai yra gaunami taikant didžiausiojo tikėtinumo metodą ir apskaičiuojami pasinaudojant iteratyviai persveriamu mažiausiųjų kvadratų arba/ir EM algoritmu.

Turint atnaujintus hiperparametrus (3.7)–(3.9), žodžių formų šaltiniui $s \in S$ struktūrinis skirstinys gali būti įvertintas tiesiogiai, kaip

$$\hat{F}_s(u) = \frac{1}{V} \sum_{w \in \mathcal{W}} \left(\mathbb{I}\{\hat{\mu}_{ws} \leq u, y_w(s) > 0\} + (1 - \hat{p}_{sw}) \mathbb{I}\{\hat{\mu}_{ws} \leq u, y_w(s) = 0\} \right). \quad (3.10)$$

Antrasis dėmuo šiame reiškinyje įvertina *nematomų žodžių formų* indėlį šaltiniui $s \in S$. Norint gauti bendrojo žodyno $\mathcal{W}$ žodžių formų struktūrinio skirstinio statistinį įvertį, reikia paimti struktūrinių skirstinių įvertinių (3.10) svertinį vidurkį

$$\hat{F}_*(u) := \frac{1}{\omega_+} \sum_{s \in S} \omega_s \hat{F}_s \left(\frac{u N^*}{\hat{\mu}_{*s}} \right), \quad (3.11)$$

prieš tai atitinkamai normuojant (pakeičiant mastelį) su

$$\hat{\mu}_{*s} := \widehat{\mathbf{E} N_s} = \int_0^\infty u \hat{F}_s(\mathrm{d}u),$$

taip, kad šaltinių struktūriniai skirstiniai turėtų tokius pačius įvertintus vidutinius tekstų dydžius N^*. Atliktame empiriniame tyrime buvo naudojami vienodi svoriai ir svoriai, proporcingi šaltinio teksto (ar žodyno) dydžiui. Galbūt paprastas mastelio pakeitimas nėra geriausias būdas prisiderinti prie nevienodų tekstų dydžių, tačiau šis uždavinys šiame darbe nenagrinėjamas.

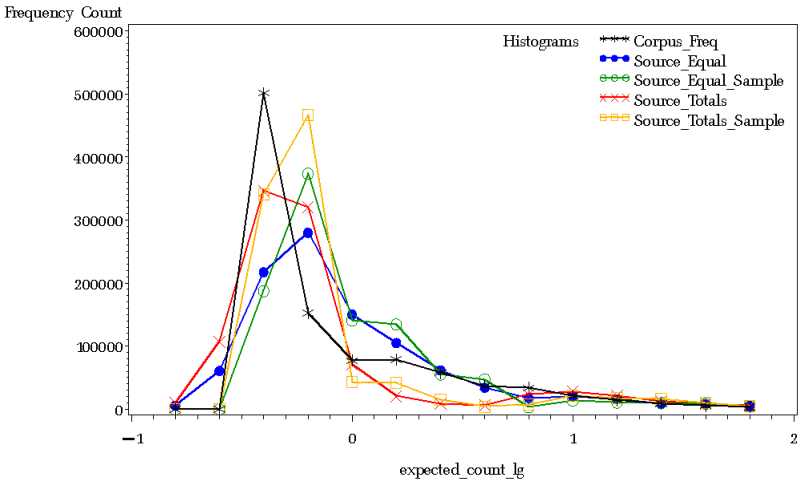
Bendrojo žodyno $\mathcal{W}$ struktūrinis skirstinys taip pat gali būti įvertintas ir tiesiogiai, nenaudojant tarpinių skaičiavimų vertinant atskirų šaltinių struktūrinius skirstinius:

$$\hat{F}_W(u) = \frac{1}{V} \sum_{w \in \mathcal{W}} \mathbb{I}\{\hat{\mu}_{w+} \leq u\}, \quad (3.12)$$

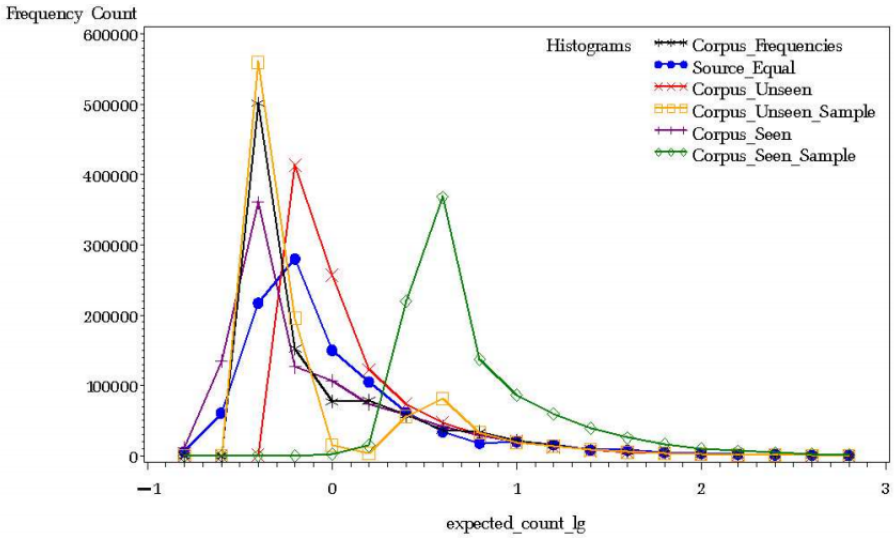
$$\hat{\mu}_{w+} := \sum_{s \in S} \left(\hat{\mu}_{ws} \mathbb{I}\{y_w(s) > 0\} + (1 - \hat{p}_{sw}) \hat{\mu}_{ws} \mathbb{I}\{y_w(s) = 0\} \right). \quad (3.13)$$

Norint šį įvertinį palyginti su (3.11), reikėtų atitinkamai pakeisti jo mastelį.

Empirinis Bajeso metodas, pritaikytas turimiems duomenims, leidžia įvertinti nematomų (angl. *unseen*) žodžių formų skaičių kiekviename šaltinyje, todėl galima atitinkamai pataisyti struktūrinių skirstinių įvertinius. Kaip pasireiškia toks pataisymas,



Pav. 3.13: Struktūrinių skirstinių įverčių, gautų įvairiais metodais, palyginimas 1

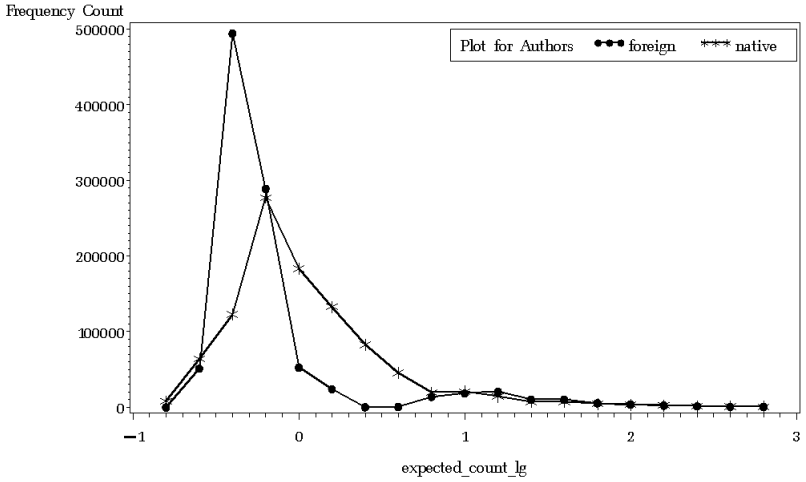


Pav. 3.14: Struktūrinių skirstinių įverčių, gautų įvairiais metodais, palyginimas 2

matome 3.13 ir 3.14 pav., kuriuose pavaizduotos struktūrinių skirstinių įverčių histogramos, gautos naudojant skirtingus metodus. Struktūrinio skirstinio įverčiai yra atitinkamai transformuoti pakeičiant mastelį taip, kad atitiktų suminį teksto žodžių kiekį $N^* = 10^6$, ir jų histogramos yra standartizuotos, kad atitiktų tekstinius dokumentus, kurių žodyno dydis lygus 10^6 žodžių formų. Histogramų grafikai yra nukirsti ties tam tikru tikėtino dažnio dydžiu, arba 10^3, arba 10^2. *Source_Equal* ir *Source_Totals* žymi atitinkamai įverčius, gautus iš (3.11) su vienodais svoriais $\{\omega_s\}$ ir svoriais, proporcingais teksto dydžiui N_s šaltinyje $s \in S$. Įverčiai su svoriais, proporcingais šaltinio žodyno dydžiui, yra labai artimi įverčiams su vienodais svoriais, todėl čia nepatei-

kiami. Įvertis, skaičiuojamas pagal (3.12), (3.13), yra žymimas *Corpus_Unseen*, o *Corpus_Seen* reiškia įvertį, gautą iš (3.12) be dėmens su įvertintu nematomų žodžių formų indėliu. Be to, yra nupieštos analogiškos struktūrinio skirstinio įverčių, tik dabar apskaičiuotos iš pradinių duomenų atsitiktinės dydžio 9 (dalinės) imties, histogramos. Tai nurodoma prie pradinės žymos pridedant žodį *Sample*. Palyginimo patogumui abiejuose paveikslėliuose pridedama įverčio *Source_Equal* histograma ir atitinkamai transformuota (keičant mastelį ir standartizuojant) žodžių formų stebėtų dažnių histograma, pavadinta *Corpus_Frequencies*.

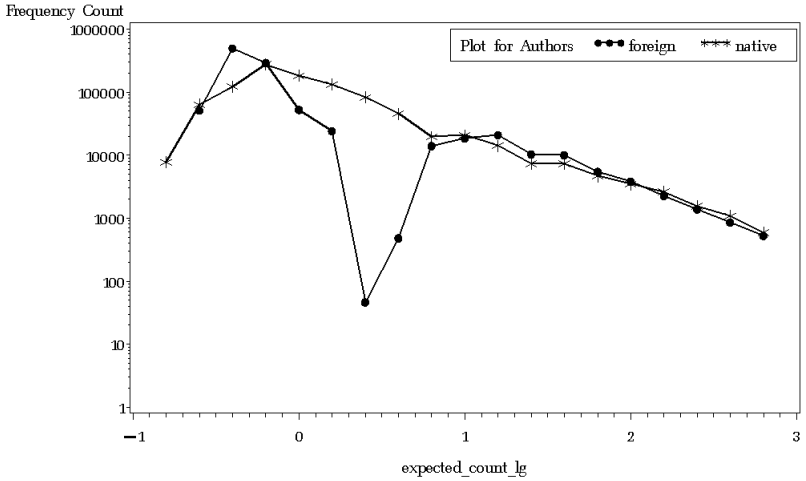
Svertiniai struktūriniai skirstinių įverčiai *Source_Equal_Sample* ir *Source_Totals_Sample* duoda priimtinas atitinkamų įverčių *Source_Equal* ir *Source_Totals*, gautų remiantis visa imtimi S, tuo pačiu galbūt ir tikrojo struktūrinio skirstinio F, prognozes. Įverčiai *Corpus_Unseen* ir *Corpus_Unseen_Sample* šiuo atveju, atrodo, yra paslinkti.



Pav. 3.15: Struktūrinio skirstinio įverčiai lietuvių ir užsienio rašytojams

Dar kartą iliustruojant tekstinių duomenų heterogeniškumą, 3.15 pav. pateikiamos empiriniu Bajeso metodu įvertintų struktūrinių skirstinių histogramos užsienio ir lietuvių autoriams. Galima įžvelgti subtilesnę užsienio rašytojų žodžių formų dažnių struktūrą, palyginus su įvertintų posvyrių Zipfo antrame dėsnyje sklaidos diagrama 3.12 pav. Viena vertus, verstiniuose tekstuose paprastai naudojamas labiau standartinis žodynas (tai atsispindi tikėtinų dažnių sumažėjimu intervale $(0; 0,8)$ $\log_{10}$ skalėje), kita vertus, juose yra žodžių, susijusių su kitų tautų geografijos, kultūros ir buities ypatumais, vadinasi, – retų originaliuose lietuvių tekstuose (tai atsispindi piku ties

$-0,4$). Dar akivaizdžiau minėti atvejai išsiskiria histogramas pateikus logaritminėje skalėje (3.16 pav.). Zipfo dėsnio atveju šios kreivės logaritminėje skalėje būtų tiesės (išskyrus galbūt kraštus, kur laikoma, kad Zipfo dėsnis gali negalioti). Lietuvių autorius atitinkanti kreivė ir yra beveik tiesė. Tačiau užsienio autorius vaizduojanti kreivė išsiskiria jau minėtomis savybėmis – didelė „duobė“, atitinkanti, kaip galima spėti, „retus“ lietuviškus žodžius, bei pikas ties „itin retais“ lietuvių kalboje žodžiais (aiškiau matomas nelogaritmuotoje skalėje). Naudojant standartinį Zipfo-Mandelbroto parametrinį struktūrinio skirstinio modelį to nesimatytų, nes vietoj kreivės būtų tiesė.



Pav. 3.16: Struktūrinio skirstinio įverčiai (logaritminėje skalėje) lietuvių ir užsienio rašytojams

3.4.4 Dalinės išvados

1. Bajeso modelis, kuris remiasi neigiama binomine regresija su pertekliniu nulių kiekiu, ir empirinis Bajeso metodas leido sukonstruoti struktūrinio skirstinio įvertinius, kurie atsižvelgia į tekstynų nehomogeniškumą ir poslinkį, atsirandantį dėl juose nestebimų žodžių formų.

2. Mokyklinės literatūros tekstinių duomenų pagrindu gautų autorių struktūrinių skirstinių įverčių palyginimas rodo, kad lietuvių kalbos tekstai yra gana heterogeniški duomenys. Lyginant lietuvių ir užsienio autorių struktūrinių skirstinių įverčius pademonstruota, kad jie įgalina atskleisti subtilesnius tiriamų teks-

tų ypatumus lyginant su bendromis jų charakteristikomis, nusakomomis Zipfo-Mandelbroto dėsniu.

3. Herdano ir Zipfo dėsniai gana neblogai aprašo žodžių ir žodžių formų lietuviškuose tekstuose dažnumų pasiskirstymo bendras charakteristikas net ir gana trumpiems tekstams. Tačiau jos gana skirtingos skirtingiems autoriams ir faktiškai sutampa su ilgų žodžių formų, kurios sudaro pagrindinę tekstų dalį, dažnumų charakteristikomis. Trumpoms žodžių formoms Herdano dėsnis negalioja.

Bendrosios išvados

1. Taikant statistinius metodus labai svarbu tiksliai apibrėžti tyrimo objektą, o tuo pačiu ir tiriamąją populiaciją, kadangi nuo to priklauso gautų rezultatų interpretacija ir patikimumas.

2. Kaip rodo Zipfo-Mandelbroto ir Herdano-Hipso dėsnių analogai bei struktūrinio skirstinio analizė, tekstynų duomenys yra nehomogeniški, priklauso ne tik nuo stiliaus, žanro ir pan., bet ir nuo paties autoriaus. Stebimi ryškūs skirtumai tarp stilių ir tarp pačių tekstų autorių. Darant išvadas apie statistinius dėsningumus lingvistikoje reikėtų į tą nehomogeniškumą atsižvelgti.

3. Bajeso modelis, kuris remiasi neigiama binomine regresija su pertekliniu nulių kiekiu, ir empirinis Bajeso metodas leido sukonstruoti struktūrinio skirstinio įvertinius, kurie atsižvelgia į tekstynų nehomogeniškumą ir poslinkį, atsirandantį dėl juose nestebimų žodžių formų.

4. Logtiesiniai ir grafiniai logtiesiniai modeliai yra gana lankstūs ir leidžia aprašyti sudėtingas struktūras, jas pavaizduoti grafiškai. Taigi, tie modeliai yra patogus įrankis formuluoti ir tikrinti įvairias hipotezes apie lingvistinių objektų struktūras ir priklausomybes, taip pat ir pačiai kalbai būdingas savybes.

5. Specialiu būdu užkodavus žodžius, į sakinį galima žiūrėti kaip į naują žodį ir traktuojant sakinius kaip žodžius galima tirti Zipfo-Mandelbroto dėsnio analogus. Pirminė analizė rodo, kad nemaža dalis (daugiau negu 17%) sakinių turi paprasčiausią struktūrą.

6. Lietuvių kalbos sintaksinės (sakinių) struktūros yra per daug sudėtingos, kad jas būtų galima aprašyti modeliais, besiremiančiais trigramų statistika. Pavyzdžiui, net 74,62% sakinių plotis yra didesnis negu 3.

Literatūros sąrašas

1. S. Abney. Statistical Methods and Linguistics. *The Balancing Act: Combining Symbolic and Statistical Approaches to Language*, 1996, 1–26.

2. M. Aerts, I. Augustyns, P. Janssen. Sparse consistency and smoothing for multinomial data. *Statistics & Probability Letters*, 1997, **33**: 41–48.

3. A. Agresti. *An Introduction to Categorical Data Analysis*. Second Edition. New York: Wiley & Sons, 2007.

4. A. Agresti. *Categorical Data Analysis*. New York: Wiley & Sons, 2002.

5. A. Agresti, B. D. Hitchcock. Bayes inference for categorical data analysis. *Statistical Methods and Applications*, 2005, **14**: 297–330.

6. A. E. Allahverdyan, W. Deng, Q. A. Wang. Explaining Zipf's law via mental lexicon. *arxiv.org: 1302.4383*, 2013. http://arxiv.org/pdf/1302.4383.pdf.

7. V. Ambrazas, A. Girdenis, K. Morkūnas [et al.]. *Lietuvių kalbos enciklopedija*, Vilnius: Mokslo ir enciklopedijų leidybos institutas, 2008.

8. R. H. Baayen. Statistical Models for Word Frequency Distributions: A Linguistic Evaluation. *Computers and the Humanities*, 1993, **26**: 347–363.

9. R. H. Baayen. The randomness assumption in word frequency statistics. *Research in Humanities Computing*, Oxford: Oxford University Press, 1996, **5**: 17–31.

10. R. H. Baayen. *Word Frequency Distributions*. Kluwer Academic Publishers, 2001.

11. V. Bagdonavičius, J. Kruopis. *Matematinė statistika*, I dalis. Vilnius: TEV, 2007.

12. V. Bagdonavičius, J. Kruopis. *Tiesiniai modeliai*. Vilnius, 2011. http://www.duomenuanalize.lt/zinynas/892_275.

13. M. Baroni, S. Evert. Words and echoes: Assessing and mitigating the non-randomness problem in word frequency distribution modeling. *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics*, Prague, 2007, 904–911.

14. W. P. Bergsma. *Marginal Models for Categorical Data*. Tilburg University Press, 1997.

15. Y. M. M. Bishop, S. E. Fienberg, P. W. Holland. *Discrete Multivariate Analysis: Theory and Practice*. Cambridge: MIT Press, 1975.

16. B. Bollobás. Modern Graph Theory. *Graduate Texts in Mathematics*, **184**, New York: Springer, 1998.

17. Ch. Callison-Burch, M. Osborne. Statistical Natural Language Processing. *A Handbook For Language Engineers*, editor Ali Farghaly, 2003, 1–29.

18. A. McCallum, K. Nigam. A Comparison of Event Models for Naive Bayes Text Classification. *Proceedings of AAAI–98 Workshop on Learning for Text Categorization*, 1998.

19. K. W. Church, W. A. Gale. Poisson mixtures. *Journal of Natural Language Engineering*, 1995, **1**: 163–190.

20. P. Congdon. *Bayesian Models for Categorical Data*, New York: Wiley & Sons, 2005.

21. V. Čekanavičius, G. Murauskas. *Statistika ir jos taikymai*. I–III dalys. Vilnius: TEV, 2006–2009.

22. I. Dabašinskienė. Šnekamosios lietuvių kalbos morfologinės ypatybės. *Acta Linguistica Lithuanica*, 2009, **LX**: 1–15.

23. J. N. Darroch, S. L. Lauritzen, T. P. Speed. Markov Fields and Log-Linear Interaction Models for Contingency Tables. *The Annals of Statistics*, 1980, **8 (3)**: 522—539.

24. V. Daudaravičius. Pradžia į begalybę: Mašininis vertimas ir lietuvių kalba. *Darbai ir dienos*, 2006, **45**: 7–18.

25. S. Dumais, J. Platt, D. Heckerman, M. Sahami. Inductive Learning Algorithms and Representations for Text Categorization. *Proceedings of ACM Conference on Information and Knowledge Management (CIKM'98)*, 1998, November, 148–155.

26. S. Evert. A simple LNRE model for random character sequences. *Proceedings of the 7èmes Journées Internationales d'Analyse Statistique des Données Textuelles (JADT2004)*, Belgium: Louvain-la-Neuve, 2004, 411–422.

27. S. Evert. How random is a corpus? the library metaphor. *Zeitschrift für Anglistik und Amerikanistik*, 2006, **54(2)**: 177–190.

28. B. van Es, C. A. J. Klaassen, R. M. Mnatsakanov. Estimating the structural distribution function of cell probabilities. *Austrian Journal of Statistics*, 2003, **32**: 85–98.

29. A. Girdenis, V. Karosienė. Skiemens ir žodžio pirmųjų bei paskutinių fonemų dažnumas bendrinėje lietuvių kalboje. *Baltistica*, 2004, **39(2)**: 213–231.

30. G. Grigonytė, E. Rimkutė. Priklausomybių gramatika pagrįstų lietuvių kalbos sintaksinių taisyklių išgavimas iš *Dabartinės lietuvių kalbos tekstyno. 10-osios tarpuniversitetinės magistrantų doktorantų konferencijos „Informacinės technologijos" pranešimų medžiaga*, Kaunas, 2005, 65–67.

31. L. Q. Ha, D. W. Stewart, P. Hanna, F. J. Smith. Zipf and type-token rules for the English, Spanish, Irish and Latin languages. *Web Journal of Formal, Computational and Cognitive Linguistics*, 2006, **1(8)**: 1-12.

32. H. S. Heaps. *Information Retrieval – Computational and Theoretical Aspects.* Academic Press, 1978.

33. G. Herdan. *Quantitative Linguistics*. London: Buttersworths, 1964.

34. D. G. Horvitz, D. J. Thompson. A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 1952, **47**: 663—685.

35. R. Jakobson. Introduction. *Structure of Language and its Mathematical Aspects. Proceedings of Symposia in Applied Mathematics*, 1961, **XII**: v–vi.

36. M. Jansche. Parametric Models of Linguistic Count Data. *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, Stroudsburg, PA, USA, 2003, **1**:288–295.

37. V. Karaciejūtė. Individualiųjų tekstų sakinio ilgis ir jo parametrai. *Žmogus ir žodis*, 2007, **8 (1)**: 105–110.

38. J. Karlgren, D. Cutting. Recognizing Text Genres with Simple Metrics Using Discriminant Analysis. *Proceedings of the 15th International Conference on Computational Linguistics (COLING '94)*, Kyoto, 1994, 1071–1075.

39. A. Kazlauskienė, G. Raškinis. Bendrinės lietuvių kalbos garsų dažnumas. *Respectus Philologicus*, 2009, **16(21)**: 169–182.

40. A. Kazlauskienė, G. Raškinis, A. Vaičiūnas. *Automatinis lietuvių kalbos žodžių skiemenavimas, kirčiavimas, transkribavimas*. Kaunas: VDU l-kla, 2010.

41. A. Kazlauskienė, E. Rimkutė, A. Utka. Kiekybiniai tyrimai kalbotyroje (I). *Gimtasis žodis*, 2011, **10**: 2–7.

42. B. Kessler, G. Nunberg, H. Shütze. Automatic Detection of Text Genre. *Proceedings of the 35th conference on Association for Computational Linguistics*, 1997, 32–38.

43. E. V. Khmaladze. The statistical analysis of large number of rare events. *CWI Report MS-R8804*, Amsterdam: Dept.Math.Statist., 1988.

44. E. V. Khmaladze. Zipf's Law. *Encyclopaedia of Mathematics, Supplement III*. Dordrecht: Kluwer Academic Publishers, 2002.

45. A. Kornai. *Mathematical Linguistics*. 2007.

46. A. Kornai. How many words are there? *Glottometrics*, 2002, **4**: 61–86.

47. J. Kovalevskaitė. *Dabartinės lietuvių kalbos tekstynas* – 10 metų kaupimo ir naudojimo patirtis. *Prace Bałtystyczne: język, literatura, kultura*, 2006, **3**: 231–239.

48. M. Kracht. *The Mathematics of Language*. Berlin, 2003 (reviewed by Geoffrey K. Pullum).

49. D. Krapavickaitė, A. Plikusas. *Imčių teorijos pagrindai*. Vilnius: Technika, 2005.

50. S. L. Lauritzen. Lectures on Contingency Tables. 2002. *http://www.stats.ox.ac.uk/∼steffen/papers/cont.pdf*.

51. A. Lipeika, J. Lipeikienė, L. Telksnys. Development of Isolated Word Speech Recognition System. *Informatica*, 2002, **13(1)**: 37–46.

52. R. Marcinkevičienė. *Lietuvių kalbos kolokacijos*. Monografija. Kaunas, 2010.

53. R. Marcinkevičienė. Tekstynų lingvistika (teorija ir praktika). *Darbai ir dienos*, 2000, **24**: 7–64.

54. B. Mandelbrot. An informational theory of the structure of language based upon the theory of the statistical matching of messages and coding. *Communication Theory*, London: Acad. Press, 1953, 503–512.

55. L. Mauzienė. Lingvistiniai ir psichologiniai lingvodidaktikos pagrindai (teorinė interpretacija). *Santalka. Filologija. Edukologija*, 2009, **17(2)**: 61–67.

56. R. Merkytė, V. Kalinka. Apie V. Fukso lingvistinių elementų susidarymo dėsnį. *Lietuvos matematikos rinkinys*, 1968, **8(2)**: 279–287. (Rusų kalba)

57. R. Merkytė. Kai kurios žodžių iš skiemenų ir skiemenų iš raidžių susidarymo statistinės charakteristikos. *Lietuvos matematikos rinkinys*, 1962, **2(1)**: 91–105. (Rusų kalba)

58. R. Merkytė. Skiemenų ir fonemų skaičiaus lietuvių kalbos žodžiuose savitarpio priklausomybės tyrimas. *Eksperimentinė ir praktinė fonetika.* Vilnius, 1974, 73–84.

59. R. Merkytė. The information content of the Lithuanian languane. *Lithuanian mathematical journal*, 1978, **18(3)**: 384–389.

60. Charles F. Meyer. *English Corpus Linguistics. An Introduction.* Cambridge: Cambridge University Press, 2002.

61. K. P. Murphy. An introduction to graphical models. 2001. *http://www.cs.ubc.ca/∼murphyk/Papers/intro_gm.pdf.*

62. J. A. Nelder, R. W. M. Wedderburn. Generalized Linear Models. *Journal of the Royal Statistical Society.* Series A (General), 1972, **135 (3)**: 370–384.

63. J. Neyman. On the two different aspects of the representative method: the method of stratified sampling and the method of purposive selection. *Journal of the Royal Statistical Society*, 1934, **97 (4)**: 558–606.

64. G. Norkevičius. *Garsų trukmių modelių kūrimo metodas, naudojant didelės apimties daugelio kalbėtojų garsyną*: daktaro disertacija. Kaunas, 2011.

65. J. Palionis. *Kalbos mokslo pradmenys*, Vilnius: Jandrija, 1999.

66. I. Petkevičiūtė, B. Tamulynas. Kompiuterinis vertimas į lietuvių kalbą: alternatyvos ir jų lingvistinis vertinimas. *Kalbų studijos*, 2011, **18**: 38–45. http://www.kalbos.lt/archyvas4.html.

67. G.K. Pullum, A. Kornai. Mathematical Linguistics. *http://www.metacarta.com/Collateral/Documents/English-US/Mathematical-linguistics-Kornai.pdf.*

68. G. Raškinis, D. Raškinienė. Lietuvių šnekos atpažinimo sistemos, pagrįstos paslėptaisiais Markovo modeliais, parametrų tyrimas ir optimizacija. *Informacinės technologijos*, 2003, **IX**: 41–48.

69. G. Raškinis, A. Kazlauskienė. Automatinis skiemenavimas: problemos ir jų sprendimas. *Kalbų studijos*, 2009, **15**: 71–76. http://fcim.vdu.lt/∼asta_kazlauskiene/publikacijos/Automatinis%20skiemenavimas-%20problemos%20ir%20ju %20sprendimas.pdf.

70. E. Rimkutė, V. Daudaravičius. Morfologinis dabartinės lietuvių kalbos tekstyno anotavimas. *Kalbų studijos*, 2007, **11**: 30–35. http://donelaitis.vdu.lt/publications/Rimkute_2007.pdf.

71. E. Rimkutė, G. Grigonytė. Automatizuotas lietuvių kalbos morfologinio daugiareikšmiškumo ribojimas. *Kalbų studijos*, 2006, **9**: 30–37. http://fcim.vdu.lt/∼erika_rimkute/straipsniai/automatizuotas_MD_ribojimas_2006.pdf.

72. E. Rimkutė, J. Kovalevskaitė. Mašininis vertimas – greitoji pagalba globalėjančiam pasauliui. *Gimtoji kalba*, 2007, **9**: 3–10.

73. E. Rimkutė. *Morfologinio daugiareikšmiškumo ribojimas kompiuteriniame tekstyne*: daktaro disertacija. Kaunas, 2006.

74. E. Rimkutė, V. Valskys, J. Vaskelienė. Lietuvių kalbos leksemų morfologinis anotavimas: ypatumai ir sunkumai. *Kalbų studijos*, 2009, **15**: 63–70. http://fcim.vdu.lt/∼erika_rimkute/straipsniai/leksemu_morfologinis_anotavimas.pdf.

75. Ch. Samuelsson, A. Voutilainen. Comparing a Linguistic and a Stochastic Tagger. *Proceedings of the Thirty-Fifth Annual Meeting of the Association for Computational Linguistics and Eighth Conference of the European Chapter of the Association for Computational Linguistics*, Somerset, New Jersey: Association for Computational Linguistics, 1997, 246–253.

76. D. Sengupta, S. R. Jammalamadaka. *Linear Models: An Integrated Approach.* World Scientific, 2003.

77. Bengt Sigurd, Mats Eeg-Olofsson & Joost van de Weijer. Word length, sentence length and frequency – Zipf revisited. *Studia Linguistica*, 2004, **58(1)**: 37–52.

78. H. A. Simon. On a class of skew distribution functions. *Biometrika*, 1955, **42**: 425–440.

79. N. A. Smith. Linguistic Structure Prediction. *Synthesis Lectures on Human Language Technologies*, Morgan & Claypool Publishers, 2011.

80. D. Šveikauskienė. Formal description of the syntax of the Lithuanian language. *Information Technology and Control*, 2005, **34(3)**: 245–256.

81. D. Šveikauskienė. *Lietuvių kalbos vientisinių sakinių automatinė sintaksinė analizė*: daktaro disertacija. Vilnius, 2009.

82. D. Šveikauskienė. Lietuvių kalbos sintaksinė analizė. *Lietuvių kalba*, 2013, **7**. http://www.lietuviukalba.lt/index.php?id=231.

83. T. Tabata. Narrative Style and the Frequencies of Very Common Words: A Corpus Based Approach to Dickens' First Person and Third Person Narratives. *English Corpus Studies 2*, 1995, 91–109.

84. G. Tamulevičius. *Pavienių žodžių atpažinimo sistemų kūrimas*: daktaro disertacija. Vilnius, 2008.

85. L. Tanguy, N. Tulechki. Sentence Complexity in French: a Corpus-Based Approach. *Intelligent Information Systems*, 2009, 1–14.

86. A. Utka. Dažninis rašytinės lietuvių kalbos žodynas: 1 milijono žodžių morfologiškai anotuoto tekstyno pagrindu. *http://donelaitis.vdu.lt/publikacijos/Dazninis_zodynas.pdf*, 2009.

87. A. Utka. Kalbinė įranga ir jos galimybės. *Darbai ir dienos*, 2000, **24**: 275–285.

88. A. Utka. Labai dažnų lietuvių kalbos žodžių ir žodžių formų ypatybės. *Lituanistica*, 2005, **1(61)**: 48–55.

89. A. Utka. *Statistinis tekstų funkcijų nustatymas*: daktaro disertacija. Kaunas, 2004.

90. A. Vaičiūnas. *Lietuvių kalbos statistinių modelių ir jų taikymo šnekos atpažinimui tyrimas, kai naudojami labai dideli žodynai*: daktaro disertacija. Kaunas, 2006.

91. D. Vaišnienė, J. Zabarskaitė. Lietuvių kalba skaitmeniniame amžiuje. *Baltųjų knygų serija*, 2012.

92. M. J. Wainwright, M. I. Jordan. Graphical Models, Exponential Families, and Variational Inference. *Foundations and Trends in Machine Learning*, 2008, **1(1–2)**: 1–305.

93. G. Wilcock. Introduction to Linguistic Annotation and Text Analytics. *Synthesis Lectures on Human Language Technologies*, Morgan & Claypool Publishers, 2009.

94. Y. Yang. An Evaluation of Statistical Approaches to Text Categorization. *Information Retrieval*, 1999, **1(1/2)**: 67–88.

95. G. U. Yule. A mathematical theory of evolution. *Philosophical Transactions of the Royal Society*, 1924, **B 213**: 21–87.

96. V. Zinkevičius. Lemuoklis – morfologinei analizei. *Darbai ir dienos*, 2000, **24**: 245–273.

97. G. K. Zipf. *The Psycho-Biology of Language.* New York: Houghton Mifflin, 1935.

98. J. Židanavičiūtė. *Kategorinių požymių priklausomybių struktūros statistinė analizė ir jos taikymas genetikoje*: daktaro disertacija. Vilnius, 2010.

99. K. Župerka. *Kalbos priemonių konkurencija kaip lietuvių kalbos stilistikos objektas.* Šiauliai, 1995.

Autorės publikacijų disertacijos tema sąrašas

1. K. Piaseckienė, M. Radavičius, R. Stiklius. Lietuviškų tekstų stilių palyginimas remiantis universalių kiekybinių charakteristikų statistine analize. *Lietuvos matematikos rinkinys. LMD darbai*, 2010, **51**: 307–312.

2. K. Piaseckienė, M. Radavičius. Lietuvių kalbos vaizdingumo raiškos priemonių analizė. *Lietuvos matematikos rinkinys. LMD darbai*, 2011, **52**: 220–224.

3. K. Piaseckienė, M. Radavičius. Empirical Bayes estimators of structural distribution of words in Lithuanian texts. *Nonlinear Analysis: Modelling and Control*, 2014, **19(4)**. (Priimtas spausdinti)

Priedai

1 priedas. Šaltinių sąrašas

S. Ach. Ambrozijus pradeda pirmas: [pasaka]. 2006.

G. Adomaitytė. Debesėlis ieško vardo: (Gintarės pasakos). 1999.

G. Adomaitytė. Vėjų miesto pasakos. 2003.

J. Avyžius. Stebuklingas miestas: pasakojimai mažiems ir dideliems. 1997.

J. Avyžius. Aštuonetas iš Trepsės namų: apysaka. 1999.

J. Avyžius. Bardo nuotykiai ir žygiai: apysaka. 2002.

B. Beniušienė. Ruginė varna: novelės vaikams. 1998.

B. Beniušienė. Pabėgėlė: vaikystės vaizdeliai ir prisiminimai. 1999.

A. Butkus. Grįšim, nepražūsim: šiuolaikinė nuotykių apysakaitė. 1995.

A. Butkus. Penkiolikos metų kapitonė: šiuolaikinė nuotykių apysakaitė. 2001.

A. Butkus. Marių vaikai: šiuolaikinė nuotykių apysakaitė. 2001.

R. Černiauskas. Pasakėlės vaikams, vanagams ir sliekams: [apysakaitė vaikams]. 2003.

L. Dovydėnas. Jaujos pasakos. 1996.

L. Dovydėnas. Katino sodas. 2006.

A. Gricienė. Bičiuliai ir šunė Iglė: apsakymėliai vaikams. 1997.

A. Gricienė. Kelionė aplink mėlyną..: išmonė ir tikrovė vyraičiams ir mergaitėms. 1998.

K. Gudonytė. Gėlių dvaras: romanas vaikams, kurie jau šį tą išmano apie gyvenimą. 2002.

K. Gudonytė. Blogos mergaitės dienoraštis: [romanas]. 2009.

A. Gustaitis. Algio Trakio ir Taksiuko Šleivio nutikimai: apysaka-pasaka. 2001.

J. Jankus. Auksinis vabaliukas: pasakos. 1997.

M. Jonutis. Kirminas paukštis: tavo pirmoji knyga apie skraidymo meną: [eseistika].

2009.

P. Juodišius. Amarėlio pyragas: [pasaka]. 2006.

A. Kandroškaitė. Pasakos vaikams, tėvams ir seneliams apie puošnų pelėsį, vilko šaltinį, šuns uodegą ir dar kai ką. 1995.

A. Kandroškaitė. Prie tėviškės verdenių: padavimai. 1999.

A. Kandroškaitė. Vilke, nebūk pikčius!: užklasiniam skaitymui. 2001.

N. Kepenienė. Kopūstų riteris ir kiti. 1996.

N. Kepenienė. Džiovintas debesėlis: pasakos. 1999.

N. Kepenienė. Skriek vaivorykštės karusele: apysaka-pasaka. 2000.

N. Kepenienė. Po riestainio saule: pasakos. 2001.

N. Kepenienė. Tititatos užburti. 2002.

N. Kepenienė. Baltosios žąsytės pasakos. 2006.

R. Kundrotienė. Karti katinėlio dalia: apysaka-pasaka. 1998.

V. V. Landsbergis. Rudnosiuko istorijos: (atsiminimai, esė, nesąmonės). 1995.

V. V. Landsbergis. Obuolių pasakos: 26 dalykėliai. 1999.

V. V. Landsbergis. Angelų pasakos. 2003.

V. V. Landsbergis. Rudnosiuko istorijos. 2004.

V. V. Landsbergis. Arklio Dominyko meilė: [pasaka]. 2004.

V. V. Landsbergis. Obuolių pasakos. 2005.

V. V. Landsbergis. Pelytė Zita: [apysaka-pasaka]. 2005.

V. V. Landsbergis. Berniukas ir žuvėdros: (pasaka paaugusiems vaikams ir jų tėveliams). 2005.

V. V. Landsbergis. Arklio Dominyko kelionė į žvaigždes: antroji knyga apie arklį Dominyką ir jo svajonių rugiagėlę. 2007.

V. V. Landsbergis. Rudnosiuko istorijos: [pasakos ir pasakojimai]. 2007.

V. V. Landsbergis. Gediminas ir keturi seneliai. 2007.

V. V. Landsbergis. Briedis Eugenijus: pasakos apie meilę ir kitus nesusipratimus. 2007.

V. V. Landsbergis. Obuolių pasakos ir kriaušių. 2008.

V. V. Landsbergis. Stebuklingas Dominyko brangakmenis: trečioji knyga apie arklį Dominyką ir jo svajonių rugiagėlę. 2010.

J. Laucius. Vaivos nuotykiai Žvilgsnių šalyse. 1995.

A. Laurinčiukas. Tigro gyvenimas ir mirtis: apsakymai. 1998.

A. Laurinčiukas. Tigro gyvenimas ir mirtis: kelionų apsakymai . 2002.

E. Liegutė. Dramblienė: apysaka. 1997.

E. Liegutė. Šuo Džimas – geras: apysaka. 2000.

E. Liegutė-Balionienė. Rudis, kuris tapo Džimu: apysaka. 1995.

R. Liutkutė. Mažieji įnamiai. 1998.

R. Lužytė. Septynios šypsenos: apysaka. 1999.

L. Mykolaitis. Šešiakojis auksarankis: trylika pasakų. 1999.

S. Paltanavičius. Tai mes, pelės!: apysaka vaikams, pelėms, kitiems žmonėms. 2004.

V. Petkevičius. Kodėlčius. 1999.

V. Petkevičius. Sieksnis, Sprindžio vaikas: apysaka. 2000.

V. Petkevičius. Didysis medžiotojas Mikas Pupkus: [apysaka]. 2001.

V. Petkevičius. Molio Motiejus – žmonių karalius: [apysaka]. 2005.

V. Račickas. Šlepetė: [apysaka vaikams ir tėvams]. 1996.

V. Račickas. Zuika padūkėlis: [apysaka]. 1997.

V. Račickas. Zuika dar gyvas: apysaka. 1997.

V. Račickas. Kita šlepetės istorija: apysaka vaikams ir tėvams. 1998.

V. Račickas. Mano vaikystės ledai: apsakymai. 2000.

V. Račickas. Nauji Zuikos nuotykiai, arba Tikrasis džiaugsmas. 2001.

V. Račickas. Mikė ir Juozapėlis. 2002.

V. Račickas. Jos vardas Nippė: mergaitė, kuri mylėjo tėtį. 2002.

V. Račickas. Šlepetė-3. 2002.

V. Račickas. Šlepetė: [premijuota apysaka]. 2003.

V. Račickas. Geriausias draugas: apsakymai. 2004.

V. Račickas. Nippė nori namo: mergaitė, kuri mylėjo tėtį. 2004.

V. Račickas. Seni pažįstami, arba Ketvirtoji šlepetė: apysaka. 2004.

V. Račickas. Sunku būti mokiniu: apsakymai. 2005.

V. Račickas. Stebuklingas portfelis: apsakymai. 2005.

V. Račickas. Nebaigtas dienoraštis: [apysaka]. 2006.

V. Račickas. Berniukai šoka breiką: novelių apysaka. 2007.

V. Račickas. Patricija, Antanas, jo tėtis ir mama: romanas vaikams. 2009.

V. Račickas. Nippė namie: mergaitė, kuri mylėjo tėtį. 2011.

R. Repšienė. Lietuvos piliakalnių legendos. 2007.

R. Sadauskas. Saulės laikrodžių meistras: [apsakymai vaikams]. 1995.

K. Saja. Ei, slėpkitės!: kam pasaka, o kam teisybė, arba dviejų dalių apysaka su pagražinimais. 1999.

K. Saja. Stulpininkas: fantastiniai apsakymai. 1999.

K. Saja. ...kurio nieks nemylėjo: romanas jaunajam skaitytojui. 2005.

K. Saja. Septyni miegantys broliai: [mažas romanas nebemažam skaitytojui]. 2009.

M. Sluckis. Milžinai nenorėjo karaliais būti: [pasaka-apysaka]. 2004.

B. Steckienė. Seku seku pasakas... 1998.

R. Šerelytė. Krakatukų pievelė: [literatūrinė pasaka]. 2007.

R. Šerelytė. Krakatukų jūra: [pasaka]. 2011.

D. Vaitkevičiūtė. Trise prieš mafiją: [apysaka]. 2006.

D. Vaitkevičiūtė. Marius Pietaris ir Juodasis Bokštas: [apysaka]. 2006.

A. Zurba. Raganaitė Džilda: pasaka-apysaka. 2007.

V. Žilinskaitė. Paršiuko puota. 1996.

V. Žilinskaitė. Tiputapė: humoristinė apysaka. 1996.

V. Žilinskaitė. Kelionė į Tandadriką: apysaka-pasaka. 2003.

J. Žilinskas. Gugis – girių kaukas ir žmonių draugas: romanas. 2006.

2 priedas

MK	MK_txt	detal	tikslum	argum	svarb	skamb	kritik	Groz_MK	Negr_MK	Adm_MK	Pub_MK	viso_MK	Baze_txt	Groz_B	Negr_B	Adm_B	Pub_B
10	abstrakčiai kalbant	-2	0	0	0	0	0	1	7	0	6	14	abstrakčiai	42	208	13	234
17	abstrakčiau kalbant	-2	0	0	0	0	0	0	4	0	1	5	abstrakčiau	3	18	1	14
20	konkrečiai kalbant	2	1	0	0	0	0	2	14	4	18	38	konkrečiai	191	807	856	2678
23	konkrečiau kalbant	1	1	0	0	0	0	10	25	3	14	52	konkrečiau	59	137	45	419
30	tiksliai kalbant	2	2	1	0	0	0	1	44	0	5	50	tiksliai	1272	2260	528	5794
34	tiksliau kalbant	1	1	1	0	0	0	157	427	25	416	1025	tiksliau	782	1625	362	4321
40	tiesiai kalbant	0	0	0	1	0	0	0	4	2	35	41	tiesiai	4589	1026	172	4736
44	tiesiau sakant	0	0	0	1	0	0	0	1	0	2	3	tiesiau	35	12	0	34
50	primityviai kalbant	0	-1	0	1	-2	0	1	0	0	8	8	primityviai	30	25	10	180
60	nuoširdžiai kalbant	0	0	0	2	0	0	16	85	5	188	294	paprastai	2937	7855	1104	15786
70	paprastai kalbant	-1	-1	0	0	0	0	22	47	3	184	256	paprasčiau	272	292	67	1712
80	paprasčiau kalbant	-1	-1	0	0	0	0	1	2	0	3	6	paprasčiausiai	1268	1103	417	4415
86	paprasčiausiai sakant	-1	-1	0	0	0	0	155	497	13	464	1129	trumpai	1567	1688	948	4732
90	trumpai kalbant	-1	-1	0	0	0	0	35	39	0	107	181	trumpiau	135	208	167	858
100	trumpiau kalbant	-1	-1	0	0	0	0	8	4	1	20	33	rimtai	1824	927	311	5273
110	rimtai kalbant	0	0	0	1	1	0	0	9	0	6	15	bendrai	146	643	469	1922
130	bendrai kalbant	-2	0	0	0	0	0	1	5	0	7	13	bendriau	3	19	4	19
136	bendriau kalbant	-2	0	0	0	0	0	1	9	0	26	36	apibendrintai	21	137	17	239
140	apibendrintai kalbant	1	1	2	0	0	0	2	2	0	4	8	nuosekliai	206	1003	188	2057
150	nuosekliai maštant	0	1	2	0	0	0	25	7	4	107	143	logiškai	144	570	60	792
160	logiškai snekant	0	1	2	0	0	0	6	2	0	14	22	blaiviai	139	108	12	421
170	blaiviai pamąsčius	0	1	1	0	0	0	0	1	0	0	1	blaiviau	20	8	1	81
176	blaiviau pagalvojus	0	1	1	0	0	0	23	2	0	26	51	gerai	15954	9095	6329	42831
180	gerai pamąsčius	0	1	1	1	0	0	42	12	1	30	85	geriau	6931	6019	1117	24520
182	geriau sakant	0	1	1	1	0	0	1	0	0	0	1	teisingai	1362	1701	1020	927
190	teisingai pasakius	1	1	2	0	0	0	12	2	3	7	24	teisingiau	198	163	62	536
191	teisingiau kalbant	1	1	2	0	0	0	6	0	9	8	23	tikriau	212	111	18	259
200	tikriau sakant	1	1	2	0	0	0	2	9	1	21	33	metaforiškai	23	127	2	181
210	metaforiškai kalbant	0	-2	0	0	0	0	14	41	6	218	279	vaizdžiai	64	242	18	656
220	vaizdžiai kalbant	0	-1	0	0	0	0	1	4	1	10	16	vaizdžiau	5	13	3	37
230	vaizdžiau kalbant	0	-1	0	0	0	0	4	24	2	9	39	vaizdingai	53	134	4	242
240	vaizdingai sakant	0	-2	0	0	1	0	1	2	0	1	4	vaizdingiau	8	14	0	19
245	vaizdingiau kalbant	0	-2	0	0	0	1	66	41	37	617	761	šviesiniai	1980	379	59	1989
250	šviesiniai kalbant	0	-2	0	0	0	1	2	2	2	16	22	šviesniau	98	40	12	164
256	šviesniau pasakius	0	-2	0	0	0	1	2	1	0	1	2	šviesniausiai	8	3	0	9
258	šviesniausiai kalbant	0	-2	0	0	0	1	0	0	0	6	6	skambiai	103	12	2	143
260	skambiai tariant	0	-2	0	0	2	0	0	0	0	16	33	aiškiau	380	623	177	1179
270	aiškiau kalbant	1	1	0	0	0	0	10	7	0	424	772	atviriai	1322	970	230	4588
280	atviriai kalbant	0	0	0	1	0	0	271	32	45	0	2	atviriau	46	48	0	180
286	atviriau kalbant	0	0	0	1	0	0	2	0	0	0	0	vulgariai	31	12	2	54
290	vulgariai sakant	0	-1	0	0	0	2	1	1	1	3	6	vulgariai	31	12	2	54
300	šiurkščiai kalbant	0	-2	0	0	0	2	0	4	3	17	24	šiurkščiai	306	94	148	461
310	grubiai kalbant	-2	-2	0	0	0	1	1	5	5	38	49	grubiai	133	79	42	419
315	grubiau pasakius	-2	-2	0	0	0	1	0	0	0	3	3	grubiau	2	2	1	9
320	griežtai kalbant	1	2	2	0	0	0	8	67	0	47	122	griežtai	979	1642	324	5388
324	griežčiau kalbant	1	1	2	0	0	0	2	3	0	8	13	griežčiau	58	129	37	648
330	banaliai kalbant	0	0	0	0	-1	0	1	1	0	10	12	banaliai	48	14	1	111