ŠIAULIŲ UNIVERSITETAS
MATEMATIKOS IR INFORMATIKOS FAKULTETAS
MATEMATIKOS KATEDRA

Raimundas Stiklius

Lietuviškų tekstų stilių palyginimas naudojant universalias kiekybines charakteristikas

Magistro darbas

Darbo vadovas
doc. dr. Marijus Radavičius

Šiauliai, 2010

TURINYS

ĮVADAS

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, pvz., informacijos teorija, 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.

Matematinėje lingvistikoje domėjimosi objektas yra kalbos žodžių, sakinių, reikšmių ir pan. rinkimas. Tobulėjant kompiuterinei technikai vis daugiau dėmesio skiriama kuo efektyvesniam kompiuterinių programų panaudojimui kalbos bei šnekos analizei, pavyzdžiui, mašininiam vertimui arba šnekos atpažinimui.

Matematinė lingvistika, kaip rašo Geoffrey K. Pullum ir András Kornai, yra matematinių struktūrų ir metodų, kurie yra svarbūs lingvistikai, analizė [13].

Realūs procesai (įskaitant ir kalbą) gamtoje paprastai yra charakterizuojami kaip signalai. Šie signalai gali būti diskretūs (raidės abėcėlėje) arba tolydūs (kalba). Signalo šaltinis gali būti stacionarus (kai jo statistinės savybės nekinta laike) arba nestacionarus (kai signalo savybės kinta laike) [12].

Charakterizuojant tokius realius signalus svarbu jų savybes aprašyti modeliais. Lingvistikoje skiriamos dvi modelių rūšys: nestatistiniai (arba baziniai) ir statistiniai (arba stochastiniai). Tai susiję su dvipusiu kalbos traktavimu jos funkcionavimo metu. Pirma, kalbą galima tyrinėti jos žodžių junginių identifikavimo požiūriu. Antra, kalbą galima traktuoti kaip tikimybinį procesą, susijusį su kalbos elementų panaudojimo dažnumu kalbos aktuose. Sudarant šiuos modelius taikomi įvairūs matematiniai metodai. Statistiniams modeliams sukurti taikomi matematinės statistikos, informacijos teorijos ir tikimybių teorijos metodai [11]. Statistinių modelių pagrindinė prielaida yra ta, kad signalas gali būti gerai aprašomas parametriniu atsitiktiniu procesu, kurio parametrai gali būti tiksliai surasti gerai apibrėžtu būdu [12].

Nestatistiniai modeliai sudaromi remiantis matematine logika, aibių grafų teorija [11].

Apdorojant kalbos signalus sėkmingai naudojamos abi modelių klasės.

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

Kultūros srityje ypač svarbu stiprinti lietuvių kalbos atpažinimo, sintezės ir vertimo mokslinius tyrimus, nes tik taip padėsime išlikti lietuvių kalbai modernioje skaitmeninėje terpėje. Kaip pabrėžia F. Čermakas, čekų nacionalinio tekstyno instituto direktorius, „kompiuteriu nevaldomoms kalboms, neradusioms vietos Europos informacinėje visuomenėje, gresia sumenkėjimas ir išnykimas“ [19]. Kad būtų galima plėtoti informacinių technologijų krypties mokslinius tyrimus ir eksperimentinės plėtros darbus siekiant sukurti lietuvių kalbos vertimo priemones, būtina sukaupti didelės apimties, autentiškos kalbinės medžiagos išteklių bazę.

Paminėsime keletą svarbiausių matematinės lingvistikos darbų, atliktų arba atliekamų Lietuvoje.

2000-iais metais lietuvių kalbai buvo sukurta gerai veikianti automatinė morfologinė analizės programa *Lemuoklis*, žodžio formai pateikianti antraštinį pavidalą (lemą) ir gramatines pažymas. Ši programa skirta Vytauto Didžiojo universiteto Kompiuterinės lingvistikos centro tekstyno moksliniams lingvistiniams tyrinėjimams automatizuoti [22].

Didėjant informacinių technologijų plėtrai, 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.

Dabartinės lietuvių kalbos tekstynas, kurį sudaro 100 mln. žodžių, – plačiai Lietuvoje naudojama duomenų bazė, reprezentatyviai atspindinti dabartinę lietuvių kalbą, įvairius jos stilius (plačiau žr. http://donelaitis.vdu.lt/).

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 [15].

Nuo 2001 m. pradėti taikyti statistiniai lietuvių kalbos morfologinio daugiareikšmiškumo ribojimo metodai: paslėptieji Markovo modeliai, Viterbi algoritmas. 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 [16].

Kalbų apdorojimo priemonių plėtrą ir 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.

Iki šiol nėra nė vienos realiai veikiančios „tobulos“ mašininio vertimo programos, kuri verstų į lietuvių kalbą ir iš jos. Vienas iš Lietuvoje siūlomų produktų, palengvinančių vertimą, yra „Tildės biuro 2006“ dalis „Vertimo vedlys“ (žr. http://www.tilde.lt). Ši automatizuota vertimo priemonė analizuoja sakinių struktūrą ir automatiškai siūlo sakinio, jo dalies ar atskirų žodžių vertimą [17].

Šiuolaikiniame pasaulyje vis aktualesnis automatinis kalbos–šnekos atpažinimo klausimas – kuriama vis daugiau diktavimo, balso operatorių, balsu valdomos paieškos ir navigacijos sistemų. Nors lietuvių šnekos atpažinimo sistemos pradėtos kurti palyginus neseniai, tačiau šiuo klausimu jau nemažai nuveikta, parašyta nemažai straipsnių, apginta ne viena daktaro disertacija.

Šiuo metu kalbos atpažinimo klausimai sprendžiami Matematikos ir informatikos institute, Vytauto Didžiojo universitete bei Kauno technologijos universitete. Pagrindinis dėmesys 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 [18].

Atliekami ir kitokio pobūdžio – konkrečių literatūros sričių – tyrinėjimai. Šiaulių universitete buvo parašyti keli bakalauro darbai, kuriuose nagrinėtas tam tikrų garsų bei kalbos dalių pasiskirstymas lietuvių tautosakoje.

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

Rašomosios lietuvių kalbos žodžių dažnumo tyrinėjimus atspindi tokios knygos kaip L. Grumadienės ir V. Žilinskienės „Dažninis dabartinės rašomosios lietuvių kalbos žodynas (mažėjančio dažnio tvarka)“ (1997) bei „Dažninis dabartinės rašomosios lietuvių kalbos žodynas (abėcėlės tvarka)“ (1998). 2009 metais paskelbta ir elektroninė dažninio žodyno versija. Tai – A. Utkos sudarytas „Dažninis rašytinės lietuvių kalbos žodynas: 1 milijono žodžių morfologiškai anotuoto tekstyno pagrindu“, kuriame pateikiami ne tik žodžių, bet ir kaitomų žodžių formų dažniai [21].

Savo straipsniuose 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 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 [20].

Kaip savo straipsnyje teigia A. Utka, tiek tradicinės kalbotyros, tiek tekstynų lingvistikos darbuose dažniausių kalbos žodžių reikšmė tekstams nėra pakankamai nagrinėta [20].

Siekiant statistiškai išnagrinėti ir įvertinti tam tikrų (trumpų) žodžių ir žodžių formų dažnumą bei sąsajas su tam tikrais funkciniais stiliais, buvo pasirinkta tokia magistro darbo tema – „Lietuviškų tekstų stilių palyginimas naudojant universalias kiekybines charakteristikas“.

Tekstai analizuojami ne vien tik lingvistine prasme, kalbiniu ar moksliniu požiūriu, siekiant nustatyti svarbias kalbos ypatybes ir jas pritaikyti praktiniams uždaviniams. Labai aktualūs ir plačiai taikomi yra tekstų rūšiavimo, klasifikavimo bei atpažinimo ir identifikavimo metodai. Paminėsime tik 2 aktualiausias tokių metodų taikymų sritis: informacijos paieška internete bei plagiato nustatymas.

Kaip jau minėjome, lietuvių kalba yra gana sudėtinga ir lanksti, 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 šaknis. Tai labai apsunkina ir be to sudėtingą raktinių žodžių parinkimo uždavinį.

Darbe aptariami 2 būdai kaip tą problemą supaprastinti arba apeiti. Siūloma skaičiuojant įvairias tiriamų tekstų skaitines charakteristikas

1) naudoti edit (redagavimo) tipo atstumus (pvz., Levenšteino atstumą) tarp žodžių, juos sutapatinant, jeigu atitinkamu būdu apibrėžtas atstumas tarp jų yra mažesnis už parinktą slenkstį;

2) naudoti nekintančius ar mažai kintančius (stabilius) lietuvių kalbos objektus ar darinius, pavyzdžiui, skyrybos ženklus, raides,jungtukus, prielinksnius ir pan.).

Levenšteino (redagavimo) atstumo apibrėžimas ir jo skaičiavimo algoritmas trumpai aptariami teoriniame skyrelyje. Tipinė Levenšteino atstumo realizacija nėra pakankamai lanksti, jos adaptavimas lietuvių kalbai, kaip išaiškėjo, yra sudėtingas (programavimo ir lituanisto) uždavinys. Todėl perspektyvesniu pasirodė antrasis būdas, kuris ir nagrinėjamas šiame darbe.

Kadangi raktiniai žodžiai atspindi teksto turinį, tai jie patys ir jų dažnumai tekste labai priklauso nuo to teksto autoriaus, temos, ir net pačio kūrinio, iš kurio tas tekstas paimtas. Todėl raktinių žodžių pagrindu keblu palyginti, pavyzdžiui, grožinę ir mokslinę literatūrą. Tokiais atvejais skirtumai išryškėja teksto stiliuje, pateikimo formoje. Vadinasi, aktualus uždavinys – pamatuoti įvairias tiriamų tekstų formos 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, ir vadinsime *universaliomis kiekybinėmis charakteristikomis*.

Magistro darbo tikslas – išanalizuoti kelis lietuvių kalbos stilius taikant statistinius metodus.

Darbo uždaviniai:

1. Statistiškai išanalizuoti grožinį ir mokslinį stilius.
2. Ištirti žodžių ilgio ir dažnio priklausomybę pritaikant koreliacinę ir regresinę analizę.
3. Ištirti daugiamačių lentelių duomenis.
4. Pritaikyti logtiesinius modelius.

Problematika ir temos aktualumas. Norint pasiekti gerų rezultatų kuriant tobulą lietuvių kalbos lingvistiką informacinėse technologijose visų pirma reikia ištirti esamą terpę, atlikti skaičiavimus, atrasti priklausomybes bei pasikartojimus. Statistiškai apdorojus turimus duomenis galima kurti programines struktūras. Tai galima padaryti remiantis statistikos duomenimis ir jų apdorojimo metodais. Ši globali problema paaiškina atliekamo tyrimo aktualumą.

1. TEORINĖ DALIS

1.1. PRADINIŲ DUOMENŲ ANALIZĖ

Statistika yra mokslas, kuris tiria, kaip rinkti duomenis, juos tvarkyti ir analizuoti [3: 7].

Empirinio skirstinio statistinės charakteristikos (vidurkis, dispersija, pradiniai ir antriniai momentai ir t.t.) vadinamos **empirinėmis imties charakteristikomis** [10: 84].

Stulpelinė diagrama – bet kuri iš stačiakampių stulpelių sudaryta diagrama [8].

Histograma **–** tai stulpelinė diagrama, sudaryta iš besiliečiančių stačiakampių, kurių aukščiai lygūs dažniams, o duomenų reikšmės yra stačiakampio pločių vidurio taškai [8].

Mediana yra toks dydis, kuris variacinę seka padalija į dvi vienodo dydžio dalis:

$$Me = \begin{cases} X_{\left(\frac{n+1}{2}\right)}, & \text{jei } n = 2k+1, \\ \frac{1}{2}\left(X_{\left(\frac{n}{2}\right)} + X_{\left(\frac{n}{2}+1\right)} \right), & \text{jei } n = 2k. \end{cases} \qquad \text{[3: 38]}$$

1.2. LOGTIESINIAI MODELIAI

Dažnių lentelės

Statistinėje eilutėje kintamojo *X* reikšmės gali kartotis. Tegul statistinėje eilutėje yra *k* skirtingų reikšmių. Tarkime, kad skirtingosios reikšmės $x_1, x_2, \ldots, x_k$.

Galima suskaičiuoti, kiek kartų statistinėje eilutėje pasikartojo kiekviena reikšmė, ir rasti, kurią visų stebėjimų dalį ji sudarė. Sakykime, kad stebima reikšmė x_j pasikartojo f_j kartų. Tuomet $f_1 + f_2 + \ldots + f_k = n$, o x_j statistinėje eilutėje sudaro f_j/n dalį visų stebėjimų.

Kintamojo reikšmės dažnis f_j – tai skaičius, nusakantis, kiek kartų reikšmė x_j pasikartojo statistinėje eilutėje.

Kintamojo reikšmės santykinis dažnis f_j/n – tai skaičius, nusakantis, kurią statistinės eilutės dalį sudaro x_j.

Skaičiuojami kiekybinių ir kokybinių kintamųjų dažniai ir santykiniai dažniai. Duomenims sisteminti naudojami ir sukauptieji bei santykiniai sukauptieji dažniai. Kaip skaičiuojami dažniai, matome žemiau pateiktoje lentelėje esančiose formulėse [9].

Reikšmė	X_1	X_2	X_3	…	X_k
Santykinis dažnis	f_1/n	f_2/n	f_3/n	…	f_k/n
Sukauptasis santykinis dažnis	f_1/n	$(f_1+f_2)/n$	$(f_1+f_2+f_3)/n$	…	$(f_1+f_2+\ldots+f_k)/n=1$

Logtiesinių modelių dažnių lentelės interpretacija

Panagrinėkime 3-jų požymių, *A*, *B* ir *C*, dažnių lentelės $L=L(A,B,C)$ atvejį. Tuomet

$$\lambda_{ijs}=\ln\mu_{ijs}=u_0+u_i^A+u_j^B+u_s^C+u_{ij}^{AB}+u_{is}^{AC}+u_{js}^{BC}+u_{ijs}^{ABC}$$

yra pilnas (prisotintas) logtiesinis modelis $M=M(A,B,C)$.

Nr.	Specifikacija	Žymėjimas
0	$\ln\mu_{ijs}=u_0+u_i^A+u_j^B+u_s^C$	$[A]+[B]+[C]$
1	$\ln\mu_{ijs}=u_0+u_i^A+u_j^B+u_s^C+u_{js}^{BC}$	$[A][BC]$
2	$\ln\mu_{ijs}=u_0+u_i^A+u_j^B+u_s^C+u_{is}^{AC}$	$[AC][B]$
3	$\ln\mu_{ijs}=u_0+u_i^A+u_j^B+u_s^C+u_{ij}^{AB}$	$[AB][C]$
4	$\ln\mu_{ijs}=u_0+u_i^A+u_j^B+u_s^C+u_{is}^{AC}+u_{js}^{BC}$	$[AC][BC]$
5	$\ln\mu_{ijs}=u_0+u_i^A+u_j^B+u_s^C+u_{ij}^{AB}+u_{js}^{BC}$	$[AB][BC]$
6	$\ln\mu_{ijs}=u_0+u_i^A+u_j^B+u_s^C+u_{ij}^{AB}+u_{is}^{AC}$	$[AB][AC]$
7	$\ln\mu_{ijs}=u_0+u_i^A+u_j^B+u_s^C+u_{ij}^{AB}+u_{is}^{AC}+u_{js}^{BC}$	$[AB][AC][BC]$
8	$\ln\mu_{ijs}=u_0+u_i^A+u_j^B+u_s^C+u_{ij}^{AB}+u_{is}^{AC}+u_{js}^{BC}+u_{ijs}^{ABC}$	$[ABC]$

0 modelis: *A*, *B* ir *C* yra nepriklausomi.

1 modelis: tarp *A* ir *C* yra sąveika požymis *A* nepriklauso nuo poros *B* ir *C*, bet *B* ir *C* yra priklausomi.

2 ir 3 modeliai turi analogišką interpretaciją kaip ir 1 modelis.

4 modelis: šiuo atveju *A* ir *C* bei *B* ir *C* priklausomi.

• Kaip interpretuoti šį modelį? Šis modelis nusako požymių *A* ir *B* *sąlyginį nepriklausomumą*, kai žinomos *C* reikšmės.

A ir *B* tarpusavio priklausomumas pasireiškia tik per *C*.

5 ir 6 modeliai turi analogišką interpretaciją kaip ir 4 modelis.

7 modelis neturi paprastos interpretacijos: galimybių santykiai dvimatėse lentelėse *L*(*AB*), kai duota *C* reikšmė, nuo *C* reikšmės nepriklauso.

8 modelis yra pilnasis modelis.

Laikoma, kad požymių A ir B stebėti dažniai turi vieną iš trijų tikimybinių skirstinių.

Polinominis (P). Atsitiktiniai dydžiai $\{\mu_{ij}\}$ turi polinominį skirstinį su parametrais $(N, \{p_{ij}\})$. Čia $\{\mu_{ij}\}$ (trumpumo dėlei) žymi rinkinį $\{\mu_{ij}, i \in I; j \in J\}$. Toliau tai žymėsime taip: $\{\mu_{ij}\} \sim Multinomial(N, \{p_{ij}\})$.

Polinominių skirstinių sandauga (PS) (angliškai *Product of Multinomials, Multinomial products*). Jis atsiranda tada, kai vienas iš požymių, pavyzdžiui, B, $B = 1, \ldots, q_B$, reiškia numerį pakartotinos požymio A imties (eksperimento) iš nepriklausomų imčių serijos (pakartojimų skaičius yra q_B). Tokiu būdu turime q_B kartų nepriklausomai stebėtas požymio A reikšmių dažnių lenteles.

A	1	2	...	q_A
1-oji imtis			...	
2-oji imtis			...	
	...	...	...	...
q_B-oji imtis			...	

Pastebėsime, kad šiuo atveju n_{+j} jau yra *neatsitiktiniai*.

Puasono skirstinys (Po, Poisson). $\{n_{ij}\}$ yra nepriklausomi Puasono atsitiktiniai dydžiai su atitinkamais vidurkiais $\{\mu_{ij}\}$:

$$P\{n_{ik} = k\} = \frac{\mu_{ij}^k}{k!} e^{-\mu_{ij}}, i \in I, j \in J.$$

Šiuo atveju ir $n_{++} = N$ yra atsitiktinis.

Remiantis Puasono skirstinio savybėmis galima parašyti:

$$N \sim Poisson(\mu_{++}), \qquad \mu_{++} = \sum_{i \in I} \sum_{j \in J} \mu_{ij}.$$

Pastaba. Visi 3 modeliai yra atskiri *apibendrintojo tiesinio modelio (ATm)* atvejai.

Statistikos. Aptarsime pagrindines statistikas, kurios naudojamos dažnių lentelių statistinėje analizėje.

Klasikinė statistika yra (Pirsono) χ^2 statistika. Jos bendras pavidalas yra toks:

$$\chi^2 = \sum_{i \in I} \sum_{j \in J} \frac{(O_{ij} - E_{ij})^2}{E_{ij}},$$

kur $\{O_{ij}\}$ yra stebėti dydžiai ("*O*" yra paimtas nuo angliško žodžio *observed*), $\{E_{ij}\}$ yra atitinkamos tų stebėtų dydžių prognozės (vidurkiai, tipinės reikšmės), remiantis duotu modeliu, t.y. nuline hipoteze H_0 ("*E*" yra paimtas nuo angliško žodžio *expected*).

Tikėtinumo (santykio) statistika apibrėžiama taip:

$$G^2 := 2 \sum_{i \in I} \sum_{j \in J} O_{ij} \ln \frac{O_{ij}}{E_{ij}}.$$

Skleidžiant Teiloro eilute nesunku įsitikinti, kad

$$G^2 := X^2 + o((O_{ij} - E_{ij})^2), O_{ij} \approx E_{ij}.$$

1.3. KORELIACINĖ IR REGRESINĖ ANALIZĖ

Apibrėžimas. Atsitiktinių dydžių *X* ir *Y* ***kovariacija*** vadiname dydį

$$\operatorname{cov}(X,Y) = M(X - MX)(Y - MY).$$

Taikydami vidurkio savybes gauname, kad

$$\operatorname{cov}(X,Y) = MXY - MX \cdot MY.$$

Tuomet:

$$\operatorname{cov}(X,X) = DX,$$

$$\operatorname{cov}(X,Y) = \operatorname{cov}(Y,X),$$

$$\operatorname{cov}(cX,Y) = c \operatorname{cov}(X,Y), \text{ kur } c = const.$$

Galima užrašyti ir taip:

$$D(X+Y) = DX + DY + 2\operatorname{cov}(X,Y).$$

Priklausomiems atsitiktiniams dydžiams *X* ir *Y*

$$\operatorname{cov}(X,Y) \neq 0,$$

o nepriklausomiems

$$\operatorname{cov}(X,Y) = 0.$$

Todėl $\operatorname{cov}(X,Y)$ tam tikra prasme charakterizuoja priklausomybę tarp *X* ir *Y*, nors tam tikslui geriau naudoti normuotą kovariaciją (kai $DX \neq 0$, $DY \neq 0$)

$$\rho(X,Y) = \frac{\operatorname{cov}(X,Y)}{\sqrt{DX \cdot DY}},$$

vadinamą koreliacijos koeficientu [10: 65].

Tegu dvimačio atsitiktinio dydžio (X,Y) stebėjimų duomenys duoti koreliacine lentele. Tuomet empiriniu koreliacijos koeficientu vadinamas skaičius:

$$r = \frac{n}{n-1} \cdot \frac{\overline{xy} - \bar{x} \cdot \bar{y}}{s(x) \cdot s(y)},$$

kuris yra koreliacijos koeficiento $\rho(X,Y)$ nepaslinktasis įvertis. Čia

$$\overline{xy} = \frac{1}{n} \sum_{i=1}^{k} \sum_{j=1}^{l} x_i y_j m_{ij}, \qquad \bar{x} = \frac{1}{n} \sum_{i=1}^{k} x_i m_i, \qquad \bar{y} = \frac{1}{n} \sum_{j=1}^{l} y_j m_j,$$

$$s^2(x) = \frac{1}{n-1} \cdot \left(\overline{x^2} - (\bar{x})^2 \right), \qquad s^2(y) = \frac{1}{n-1} \cdot \left(\overline{y^2} - (\bar{y})^2 \right),$$

$$\overline{x^2} = \frac{1}{n} \sum_{i=1}^{k} x_i^2 m_i, \qquad \overline{y^2} = \frac{1}{n} \sum_{j=1}^{l} y_j^2 m_j.$$

Žinoma, kad didelių imčių koreliacija $r \approx \rho$ [1: 324].

Teorema. Koreliacijos koeficientui $\rho(X,Y)$ galioja savybės:

1) $|\rho(X,Y)| \le 1$.

2) Jei X ir Y yra nepriklausomi, tai $\rho(X,Y) = 0$.

3) $|\rho(X,Y)| = 1$ tada ir tik tada, kai egzistuoja realūs skaičiai a ir b, $a \neq 0$, tokie, kad $Y = aX + b$. Be to $|\rho(X,Y)| = 1$, kai $a > 0$, ir $\rho(X,Y) = -1$, kai $a < 0$.

Galima apskaičiuoti koreliacijos koeficiento ρ pasikliautinąjį intervalą su pasikliovimo tikimybe P:

$$\left[r - \frac{v(1-r^2)}{\sqrt{n}}; r + \frac{v(1-r^2)}{\sqrt{n}} \right],$$

čia skaičius v randamas iš [1: 352] priedo, $\Phi(v) = \frac{P}{2}$.

Dvimačio atsitiktinio dydžio (X,Y) imtį pavaizdavus xOy plokštumoje taškais, gaunama tam tikra plokštumos taškų aibė, kuri vadinama sklaidos diagrama. Sklaidos diagrama atskleidžia statistinį požymio (X,Y) komponenčių X ir Y priklausomumą [1. 325].

1.4. LEVENŠTEINO ATSTUMAS

Duomenų apdorojime yra svarbi tam tikro eiliškumo duomenų paieška, pvz., žodžio paieška tekste. Kuriant paieškos algoritmus svarbus užklausos apdorojimas, t.y. ne visuomet vartotojo užklausa yra korektiška.

Levenšteino *atstumo skaičiavimas* – vienas iš būdų kaip palyginti dvi eilutes (galima įsivaizduoti, kad pirmoji yra užklausa, o antroji – įrašas duomenų bazėje).

Levenšteino arba pakeitimo (angl. *edit distance*) ***atstumu*** vadinsime mažiausią žingsnių skaičių, reikalingą norint vieną eilutę pakeisti kita. Yra keli eilučių pertvarkymo būdai: ***kaita***, ***įterpimas***, ***trynimas***. Kiekvienas veiksmas yra įvertinamas *kaina*, kuri gali būti parenkama skirtinga konkrečiu atveju.

- *Kaita* gali kainuoti 0, jei simboliai sutampa, arba 1, jei nesutampa.
- *Įterpimas* ir *pašalinamas* kainuoja po 1.
- Visų šių veiksmų kainų suma ir nulemia Levenšteino atstumą.

Sprendimo procesą labai patogu pavaizduoti naudojant $m \times n$ matricą *D*, kur *m* ir *n* yra eilučių ilgiai.

D[*i, j*] – pati mažiausia veiksmų kaina tarp eilučių $A = (a_1\, a_2\, a_3 \ldots a_i)$ ir $B = (b_1\, b_2\, b_3 \ldots b_j)$ *i*-tųjų ir *j*-tųjų simbolių.

Į *D* matricos [*i, j*] poziciją yra įrašomas mažiausias skaičius iš šių matricos elementų:
D[*i* – 1, *j* – 1]+*kaina.* Jei *a*[*i*]=*b*[*j*], tai *kaina*=0, kitu atveju *kaina*=1,
D[*i* – 1, *j*]+1. Eilutė turi vienu simboliu per daug – sumokama ištrynimo kaina,
D[*i, j* – 1]+1. Eilutė turi vienu simboliu per mažai – sumokama įterpimo kaina.

Visi duomenys surašomi į matricą.

Pvz., iš žodžio KIAMAS norint gauti KAIMAS, sudaroma lentelė

		K	A	I	M	A	S
	0	1	2	3	4	5	6
K	1	0	1	2	3	4	5
I	2	1	1	1	2	3	4
A	3	2	1	2	2	2	3
M	4	3	2	2	2	3	3
A	5	4	3	3	3	2	3
S	6	5	4	4	4	3	**2**

Paskutinis matricos *D*[*m, n*] narys parodo Levenšteino atstumą.

Pagrindinė skaičiavimo taisyklė:

$$S(a_1 \ldots a_i, b_1 \ldots b_j) = \begin{cases} S(a_1 \ldots a_{i-1}, b_1 \ldots b_j) + \\ \quad + kaina(I, T(a_i)), \text{ kai } i, j > 0 \\ S(a_1 \ldots a_i, b_1 \ldots b_{j-1}) + \\ \quad + kaina(I, T(b_j)), \text{ kai } i, j > 0 \\ S(a_1 \ldots a_{i-1}, b_1 \ldots b_{j-1}) + \\ \quad + kaina(K(a_i, b_j)), \text{ kai } i, j > 0 \\ i, \text{kai } j = 0 \\ j, \text{kai } i = 0 \\ 0, \text{kai } i,j = 0 \end{cases}$$

2. PRAKTINĖ DALIS

Visi skaičiavimai buvo atlikti naudojant R-packet (žr. priedai 7 ir 8), Excel, SAS programas su statistiniais paketais. Tyrimai buvo atliekami su daugybe skirtingų autorių grožinio ir mokslinio stilių tekstais. Pvz., grožinio stiliaus: Abe Kobas – „Moteris smėlynuose“, Maironis – „Pavasario balsai“, Ernestas Hemingvėjus – „Senis ir jūra“ ir kiti. Mokslinio stiliaus: „Filosofijos įvadas“ (Karl Jaspers), „Politologijos įvadas 2“, „Ekonomika“ ir kiti. Kaip matome, tiriamųjų tekstų aspektas yra gana platus, kad būtų pasiekti optimalesni rezultatai. Taip pat tyrimams atlikti buvo imami tekstai iš Lietuvos tekstynų archyvų, pvz.: www.donelaitis.vdu.lt.

2.1. PRADINIŲ DUOMENŲ ANALIZĖ

Prieš pradėdami tirti tekstus ir nežinodami, kokius gausim rezultatus, iškėlėme tokią hipotezę: ar statistiniais tyrimais galime atskirti grožinį stilių nuo mokslinio. H_0 – statistiškai mokslinis ir grožinis stiliai vienodi. H_1 – statistiškai mokslinis ir grožinis stiliai skiriasi.

1 lentelė. Statistinės charakteristikos

Lyginamoji grožinio ir mokslinio stilių analizė		
	Grožinis stilius	Mokslinis stilius
Tiriamų žodžių kiekis	5000	5000
Panaudotų raidžių kiekis	28258	32632
Minimalus žodžių ilgis	1	1
Maksimalus žodžių ilgis	19	20
Žodžių ilgio mediana	5	6
Žodžių ilgio vidurkis	5,65	6,51
Žodžių ilgio dispersija	7,41	9,68
Žodžių ilgio standartinis nuokrypis	2,72	3,11
Sakinių kiekis	468	317
Sakinių vidurkis	93,6	63,4
Sakinių dispersija	259	131

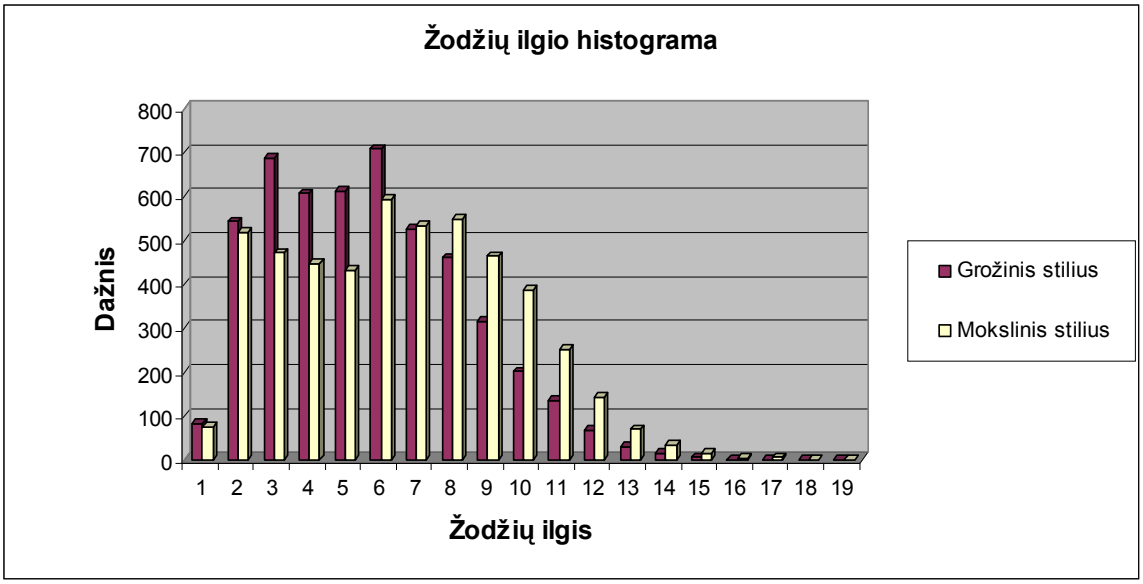
Pažvelgus į šiuos pradinius rezultatus galima pastebėti, kad grožinis ir mokslinis stiliai šiek tiek skiriasi. Matome, kad buvo paimtas vienodas kiekis žodžių, tačiau panaudotų raidžių kiekis ganėtinai skiriasi: moksliniame jų net 4400 daugiau. Mokslinio stiliaus žodžių ilgių mediana ir vidurkis didesni. Tai reiškia, kad moksliniuose tekstuose yra daugiau ilgesnių žodžių, nei grožiniuose tekstuose. Didesnė dispersija rodo, kad grožiniame stiliuje yra daugiau žodžių, kurių ilgis artimesnis vidurkiui, o moksliniame yra daugiau ilgų žodžių. Tačiau pažvelgus į sakinių struktūrą akivaizdžiai matyti, kad grožiniuose tekstuose sakinių yra žymiai daugiau. Suprantama, tai parodo, kad grožiniame stiliuje sakiniai yra daug trumpesni, o moksliniame – ilgesni.

Žinant tokius duomenis jau galėtume preliminariai nuspėti, kokį tekstą skaitome.

2 lentelė. Žodžių ilgių dažnių lentelė

Grožinio stiliaus																			
Žodžių ilgis	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
Dažnis	83	542	687	606	611	707	526	460	316	202	136	68	31	15	6	2	1	0	1

Mokslinio stiliaus																			
Žodžių ilgis	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	20
Dažnis	76	517	471	447	431	594	533	549	464	388	252	143	70	34	17	5	6	2	1



1 pav. Žodžių ilgių histograma

Iš 2 lentelėje ir 1 paveiksle pateiktų duomenų matome, kad grožinio stiliaus tekstuose iki 7 raidžių žodžių yra daugiau nei moksliniame. O moksliniuose tekstuose ilgesnių nei 7 raidžių žodžių yra kur kas daugiau. Taigi matome, kad grožinį ir mokslinį stilius galima atskirti pagal žodžių ilgius.

Dabar patyrinėkime lenteles (1 ir 2 priedai), kuriose yra ištirtas kai kurių žodžių paplitimas tekstuose. Buvo atlikti tyrimai su šiais dažniau vartojamais trumpais žodeliais bei tų žodžių formomis: **o, į, ir, ar, su, yra, kad, bet, tai, kaip, ji, jis, jie, jos, jų**. Grožiniuose tekstuose yra plačiau vartojami šie žodžiai: **o, į, bet, tai, jis, jie**. Moksliniuose šie: **ir, su, yra, ji, jų**. Abiejų stilių tekstuose vartojamas panašus kiekis šių žodžių: **ar, kad, kaip, jos.**

Tiriami žodžiai padeda atskirti mokslinį stilių nuo grožinio. Grožiniuose tekstuose yra 2 kartus daugiau prielinksnių **į**, šiek tiek mažiau jungtukų **ir**, 6 kartus mažiau veiksmažodžių **yra**, 3 kartus daugiau jungtukų **bet**, beveik 3 kartus daugiau įvardžių **tai**, o žodis **jis** vartojamas 8 kartus dažniau. Be to, moksliniuose tekstuose, galima sakyti, visiškai nenaudojami **!** ir **?,** o tai dar vienas iš būdų atskirti šiuos tekstus.

Apskritai, jungtuko **ir** vartojimas sudaro apie 4% visų tekstuose vartojamų žodžių. Žinant, kad lietuvių kalbos žodynas yra labai platus, o šie 15 žodelių (tarp kurių yra ir šių žodžių formų) sudaro apie 12% paplitimą tarp visų žodžių.

Taigi, po pradinių duomenų analizės galima teigti, kad hipotezė H_1 yra patenkinta ir statistinių tyrimų pagalba galime atskirti grožinius tekstus nuo mokslinių.

2.2. ŽODŽIŲ ILGIO IR DAŽNIO PRIKLAUSOMUMAS

2.2.1. Žodžių ilgio ir dažnio priklausomumas grožinio stiliaus tekstuose

Ištirsime priklausomybę tarp žodžių ilgio ir dažnio grožiniame stiliuje.

Turime imtį, jos $N = 19$.

Randame imties XY vidurkį:

$$\overline{XY} = \frac{1}{19}\sum_{i=1}^{19} X_i Y_i = 1487 .$$

Randame imties X vidurkį:

$$\overline{X} = \frac{1}{19}\sum_{i=1}^{19} X_i = 10 .$$

Randame imties Y vidurkį:

$$\overline{Y} = \frac{1}{19}\sum_{i=1}^{19} Y_i = 263.$$

Randame imties *X* standartinį nuokrypį:

$$\sigma(X) = \sqrt{S^2} = \sqrt{\frac{1}{19}\sum_{i=1}^{19} X_i^2 - \left(\overline{X}\right)^2} = 5{,}9.$$

Randame imties *Y* standartinį nuokrypį:

$$\sigma(Y) = \sqrt{S^2} = \sqrt{\frac{1}{19}\sum_{i=1}^{19} Y_i^2 - \left(\overline{Y}\right)^2} = 274.$$

Gauname koreliacijos koeficientą:

$$\rho(X,Y) = \frac{\overline{XY} - \overline{X}\cdot\overline{Y}}{\sigma(X)\cdot\sigma(Y)} = \frac{1487 - 10\cdot 263}{5{,}9\cdot 274} \approx -0{,}784.$$

Gautas koreliacijos koeficientas parodo stiprią neigiamą priklausomybę (žr. 4 priedas).

Apskaičiuojame koreliacijos koeficiento ρ pasikliautinąjį intervalą su tikimybe $P = 0{,}95$. $v = 1{,}96$ žiūrėti V. Čekanavičiaus ir G. Murausko knygoje [4: 352]. (Mūsų atveju *v* visada bus konstanta.)

$$\left[\rho - \frac{v(1-\rho^2)}{\sqrt{n}}; \rho + \frac{v(1-\rho^2)}{\sqrt{n}}\right] = \left[-0{,}784 - \frac{1{,}96(1-(-0{,}784)^2)}{\sqrt{19}}; -0{,}784 + \frac{1{,}96(1-(-0{,}784)^2)}{\sqrt{19}}\right] =$$
$$= \left[-0{,}957; -0{,}61\right]$$

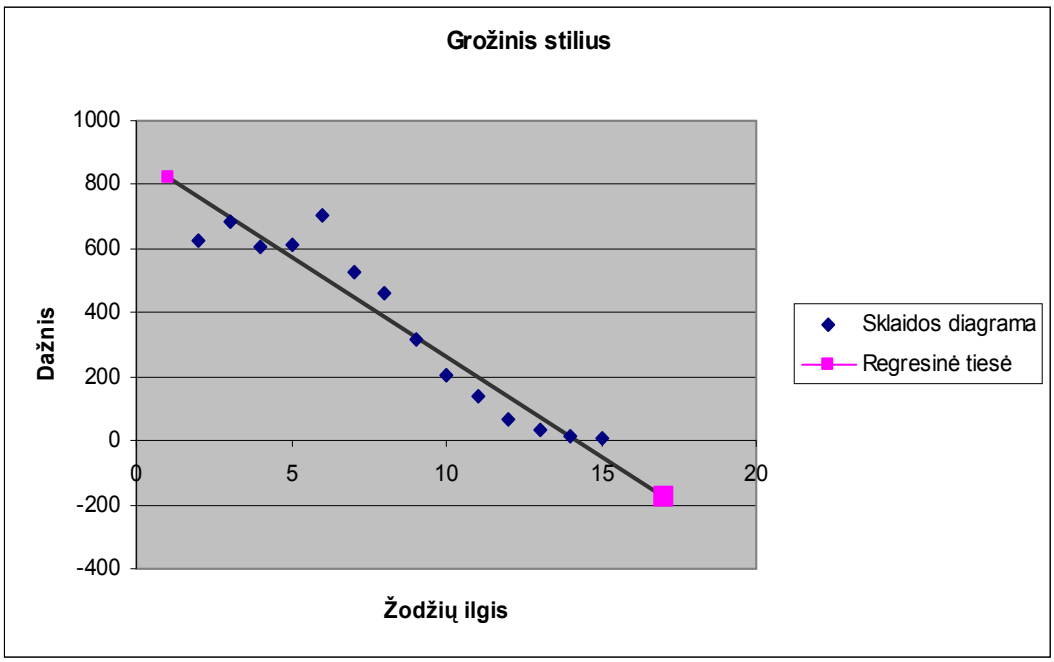
Sudarome regresijos tiesės lygtį ir surandame jos koeficientus:

$$Y = aX + b, \quad a = \frac{\overline{xy} - \overline{x}\cdot\overline{y}}{\overline{x^2} - \overline{x}^2} \approx -38{,}5, \quad b = \overline{y} - a\overline{x} \approx 644{,}6, \quad Y = -38{,}5X + 644{,}6.$$

Ši tiesė apibūdina žodžių dažnio kitimą kintant žodžių ilgiui grožiniuose tekstuose.

Turėdami visus duomenis sudarome sklaidos diagramą su regresijos tiese:

Kad būtų gerai atspindėta regresijos tiesė, tam tikslui buvo apjunkti duomenys. 1 ir 2 raidžių ilgių žodžiai bei imami žodžių ilgiai iki 16 raidžių, nes ilgesnių yra labai nedaug, todėl jie nėra reikšmingi.



2 pav. Žodžių ilgio sklaida grožiniuose tekstuose

2.2.2. Žodžių ilgio ir dažnio priklausomumas mokslinio stiliaus tekstuose

Skaičiavimai su moksliniu stiliumi yra analogiški, todėl trumpumo dėlei jie nebus rodomi.

Gaunamas koreliacijos koeficientas:

$$\rho(X,Y) \approx -0{,}763\,.$$

Gautas koreliacijos koeficientas parodo stiprią neigiamą priklausomybę (žr. 4 priedas).

Apskaičiuojame koreliacijos koeficiento ρ pasikliautinąjį intervalą su tikimybe $P = 0{,}95$:

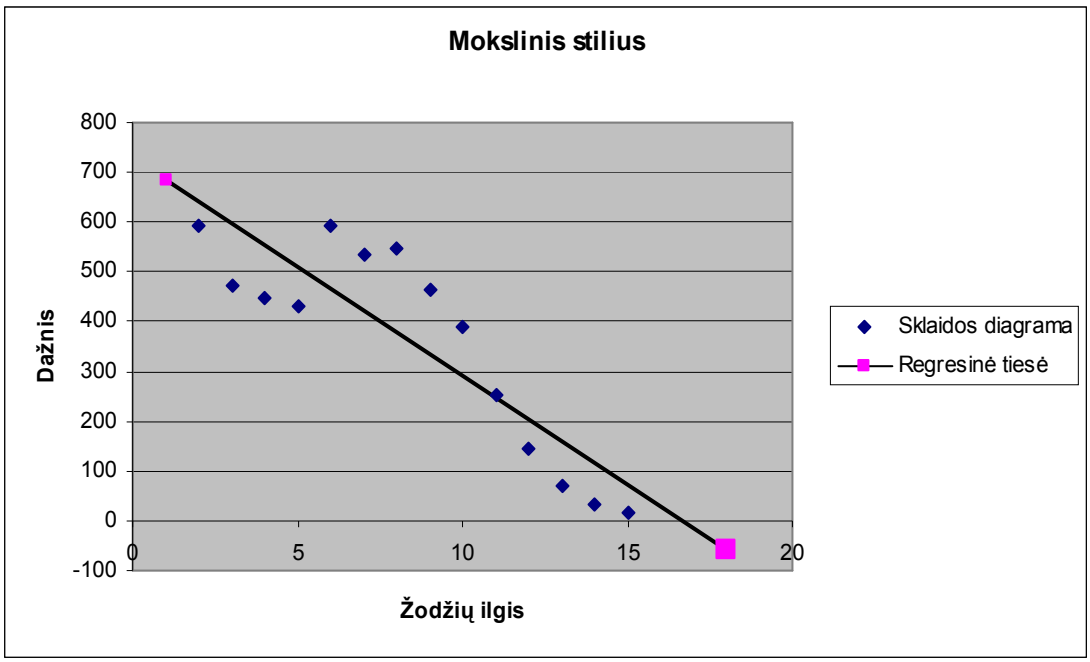
$$[-0{,}951;\ -0{,}575].$$

Sudarome regresijos tiesės lygtį ir surandame jos koeficientus:

$$Y = aX + b, \qquad a \approx -29{,}96, \qquad b \approx 564{,}6, \qquad Y = -29{,}96X + 564{,}6.$$

Ši tiesė apibūdina žodžių dažnio kitimą kintant žodžių ilgiui moksliniuose tekstuose.

Turėdami visus duomenis sudarome sklaidos diagramą su regresijos tiese:



3 pav. Žodžių ilgio sklaida moksliniuose tekstuose

Buvo tikrinama priklausomybė tarp žodžių ilgio ir jų dažnio. Tiek grožiniame, tiek moksliniame stiliuose buvo gauta stipri neigiama priklausomybė. Iš regresijos tiesės matome, kaip didėjant žodžio ilgiui mažėja jų dažnis. Matome, kad lietuvių kalboje labai ilgų žodžių nėra daug. Taip pat buvo patikrintas jų reikšmingumas pasikliautinaisiais intervalais.

2.3. LENTELIŲ ANALIZĖ

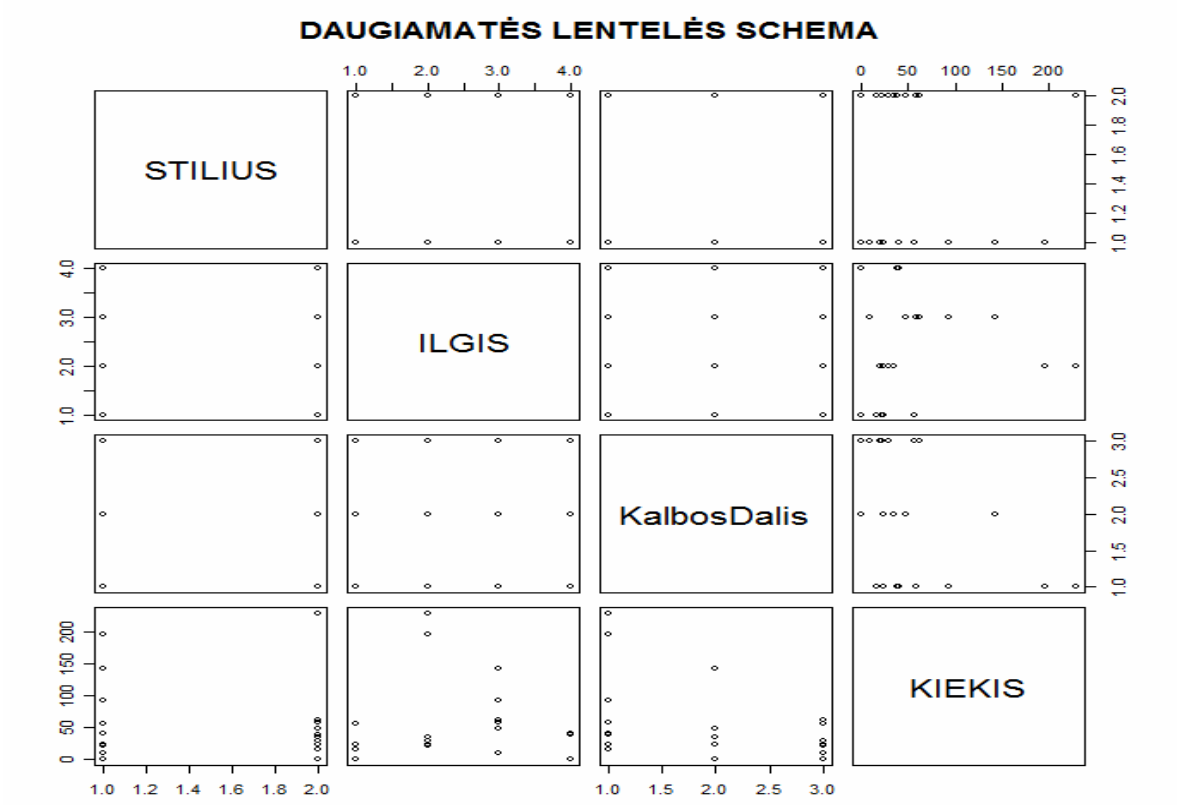
Prieš pradėdami analizuoti lentelių (1 ir 2 priedai) duomenis, pirma jas apjungsime į vieną daugiamatę lentelę. Šiai analizei naudosime „r-packet“ programą. Gauname naują daugiamatę lentelę, ją sugrupuojame į faktorius. *STILIUS* yra 2 lygių faktorius, *ILGIS* – 4 lygių faktorius, *KALBOS DALIS* – 3 lygių faktorius. Duomenims surinkti buvo ištirti mokslinio ir grožinio stilių tekstai, kurių bendra apimtis – 10000 žodžių. 3 lentelėje yra pateikta 1147 tyrimui atrinktų žodžių kiekybinė analizė.

3 lentelė. Daugiamatė lentelė

	STILIUS	ILGIS	KALBOS DALIS	KIEKIS
1	Grožinis	1	Jungtukas	24
2	Mokslinis	1	Jungtukas	16
3	Grožinis	2	Jungtukas	196
4	Mokslinis	2	Jungtukas	229
5	Grožinis	3	Jungtukas	93
6	Mokslinis	3	Jungtukas	59
7	Grožinis	4	Jungtukas	40
8	Mokslinis	4	Jungtukas	38
9	Grožinis	1	Įvardis	0

10	Mokslinis	1	Įvardis	0
11	Grožinis	2	Įvardis	24
12	Mokslinis	2	Įvardis	35
13	Grožinis	3	Įvardis	142
14	Mokslinis	3	Įvardis	48
15	Grožinis	4	Įvardis	0
16	Mokslinis	4	Įvardis	0
17	Grožinis	1	Kita	57
18	Mokslinis	1	Kita	23
19	Grožinis	2	Kita	21
20	Mokslinis	2	Kita	30
21	Grožinis	3	Kita	10
22	Mokslinis	3	Kita	62
23	Grožinis	4	Kita	0
24	Mokslinis	4	Kita	0

4 paveiksle pavaizduotos daugiamačių lentelių grafinės schemos.



4 pav. Brėžinys, sudarytas iš daugiamatės lentelės

Turint daugiamatę lentelę galima atlikti įvairius tyrimus. Daugiamatę lentelę bet kada galima susiaurinti, t.y. panaikinti kokį nors faktorių tariant, kad jis yra nereikšmingas. Atliekant tokius pertvarkymus reikėtų nepamiršti, jog galima prarasti kai kurios duomenis arba juos „iškraipyti“ (žr. 3 priedas, **Simpsono paradoksas**).

Tarkime, kad norime tyrinėti kalbos dalis, tada mums yra nereikalingas faktorius *stilius*. Jį tiesiog apjungiame. Arba atvirkščiai, norime tyrinėti stilius, bet mums trukdo faktorius *kalbos dalis*, todėl apjungiame jį.

4 lentelė. Be skilties „Kalbos dalis“

	Stilius	
Ilgis	Grožinis	Mokslinis
1	81	39
2	241	294
3	245	169
4	40	38

5 lentelė. Be skilties „Stilius“

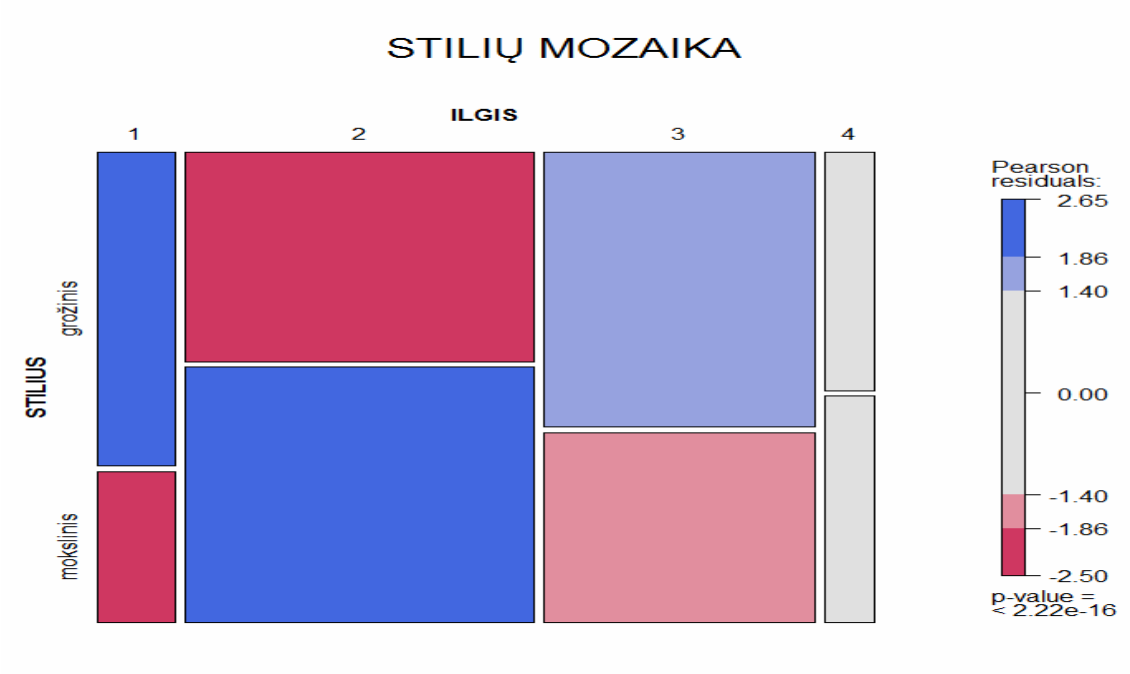
	Kalbos dalis		
Ilgis	Jungtukas	Įvardis	Kita
1	40	0	80
2	425	59	51
3	152	190	72
4	78	0	0

Štai apjungus faktorius buvo gautos dvi skirtingos lentelės. Remiantis 4 lentelės duomenimis galima nagrinėti žodžių ilgį pagal stilių, o pagal 5 lentelės duomenis galima tyrinėti kalbos dalis pagal žodžių ilgį. Dabar išanalizuosime šias gautas lenteles.

4 lentelėje 1147 žodžiai yra apjungti į du faktorius – tai *ilgis* ir *stilius*. Sudarome šios lentelės grafinę schemą (žr. 6 priedas), iš kurios matome, kurio stiliaus (mūsų tiriamų) žodžių pagal jų ilgius yra daugiau. Ši schema parodo tik kiekių palyginimus. Norėdami išsiaiškinti daugiau, sudarome mozaiką.

MOZAIKA skirta dažnių lentelių analizės rezultatams pavaizduoti vizualiai. Kiekvieno stačiakampio plotas yra proporcingas dažnių lentelės atitinkamos ląstelės reikšmei, o spalva parodo, kiek tas dažnis nukrypsta į vieną ar kitą pusę nuo tikėtinos (*expected* = „tipinės“, „vidutinės“, ...) to dažnio reikšmės, kai teisinga nulinė hipotezė, t.y. bazinis modelis. Šiuo atveju bazinis modelis tvirtina, kad tiriamieji požymiai yra nepriklausomi.

2.3.1. Tekstų stilių lentelės analizė



5 pav. Mozaika, sudaryta iš 4 lentelės

Mozaika parodo, kaip duomenys yra pasiskirstę. Čia galime matyti, kaip atskiri duomenys yra priklausomi arba nepriklausomi.

Matome, kad moksliniame tekste dažniausiai pasitaiko *ilgis* = 2 žodžiai, rečiau *ilgis* = 1, o grožiniame tekste rečiau *ilgis* = 2, o dažniau *ilgis* = 1 ir *ilgis* = 3. *Ilgis* = 1 žodžiai (beveik) statistiškai reikšmingai vyrauja grožiniuose tekstuose („beveik“ būtų galima praleisti, jeigu pasirinktas reikšmingumo lygmuo α būtų lygus 0.1, kas atitinka kritinę reikšmę apie 1.64, >1.64 arba < –1.64; arba jeigu α būtu 0.05, bet mozaikoje vietoje didesnių +1.86 ir mažesniu –1.86 reikšmių būtų atitinkamai naudojamos šios reikšmės reikšmes +1.96 ir –1.96). Priešingas teiginys galioja *ilgis* = 2 žodžiams, kurie (beveik) statistiškai reikšmingai vyrauja moksliniuose tekstuose. *Ilgis* = 3 žodžiai dažnesni grožiniuose tekstuose, bet statistiškai reikšmingas skirtumas nenustatytas, yra nustatyta tik „tendencija“, „polinkis“. *Ilgis* = 4 žodžių dažniai moksliniame ir grožiniame tekste (beveik) nesiskiria.

Ir šiems duomenims iškeliama hipotezė apie požymių statistinę priklausomybę:

$$\begin{cases} H_0 : p > \alpha = 0{,}05, \\ H_1 : p < \alpha = 0{,}05. \end{cases}$$

Hipotezė H_0 teigia, kad stilius ir žodžių ilgis yra nesusiję, H_1 – stilius ir žodžių ilgis yra susiję. Šiai hipotezei ištirti naudosime χ^2 testą (Chi-Square) ir tikslųjį Fišerio testą.

6 lentelė. χ^2 testas

Pearson's Chi-squared test
data: GSStabs
X-squared = 30.1426, *df* = 3, *p*-value = 1.288e-06

Grafoje „Pearson's Chi-squared test“ randame $\chi^2 = 30{,}1426$ reikšmę, laisvės laipsnių $df = 3$ ir p reikšmę, lygią $1{,}288 \cdot 10^{-6}$. Kadangi p reikšmė mažesnė už 0,05, hipotezę apie požymių nepriklausomumą atmetame – požymiai statistiškai priklausomi.

7 lentelė. Fišerio testas

Fisher's Exact Test for Count Data
data: GSStabs
p-value = 1.147e-06
alternative hypothesis: two.sided

Fišerio testas yra tikslesnis. Jis hipotezę apie požymių nepriklausomumą atmetė – požymiai statistiškai priklausomi, nes *p* reikšmė mažesnė už 0,05. Tačiau testas perspėja, kad (two.sided) hipotezė gali būti dviprasmiška, reikėtų tirti daugiau duomenų.

Dabar ištirsime kiekvieno laukelio reikšmingumą visai lentelei.

8 lentelė. Laukelių analizės paaiškinimas

Cell Contents	Laukelio turinys
N	*N* – gardelės duomenų kiekis
Chi-square contribution	χ^2 – reikšmė gardelėje
N / Row Total	Gardelės duomenų dalis eilutėje
N / Col Total	Gardelės duomenų dalis stulpelyje
N / Table Total	Gardelės duomenų dalis lentelėje

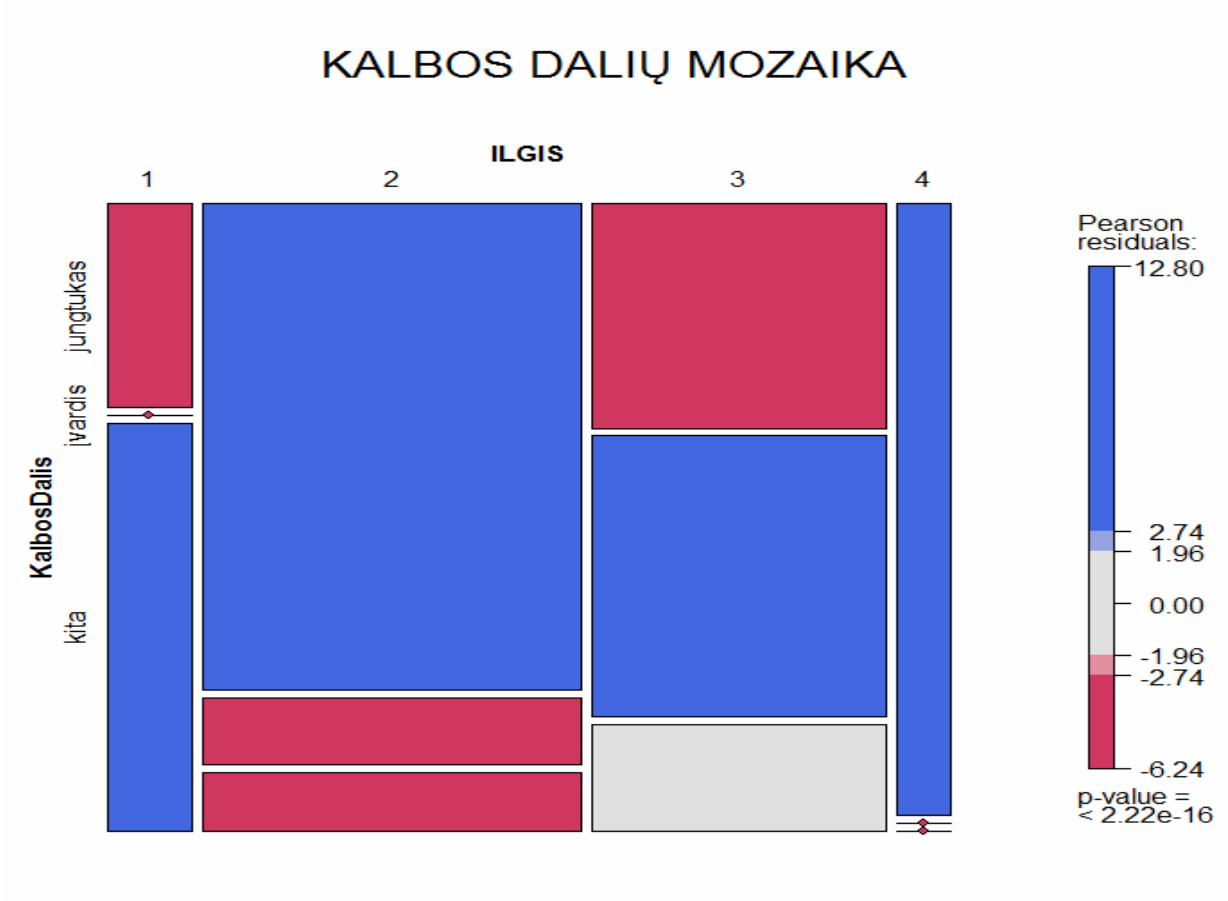
9 lentelė. Laukelių analizė

	STILIUS		
ŽODŽIŲ ILGIS	**GROŽINIS**	**MOKSLINIS**	**VISA EILUTĖ**
1	81 4,82 0,675 0,133 0,071	39 5,418 0,325 0,072 0,034	120 0,105
2	241 6,268 0,45 0,397 0,21	294 7,045 0,55 0,544 0,256	535 0,466
3	245 3,064 0,592 0,404 0,214	169 3,444 0,408 0,313 0,147	414 0,361
4	40 0,04 0,513 0,066 0,035	38 0,044 0,487 0,07 0,033	78 0,068
VISAS STULPELIS	607 0,529	540 0,471	1147

Šioje lentelėje matome, kiek yra duomenų kiekviename laukelyje. Taip pat yra suskaičiuoti χ^2 kiekvienam laukeliui bei kokią dalį laukelis sudaro eilutėje, stulpelyje ir visoje lentelėje. Iš šios lentelės sužinome, kuris laukelis yra reikšmingas, o kuris nedaro įtakos visai lentelei.

2.3.2. Kalbos dalių lentelės analizė

5 lentelėje 1147 žodžiai yra apjungti į du faktorius, tai *ilgis* ir *kalbos dalis*. Sudarome šios lentelės grafinę schemą, (žr. 5 priedas), iš kurios matome, kurių kalbos dalių (iš mūsų tiriamų žodžių) pagal žodžių ilgius yra daugiausia. Ši schema parodo tik kiekių palyginimus. Norėdami išsiaiškinti daugiau, sudarome mozaiką.



6 pav. Mozaika, sudaryta iš 5 lentelės

Kaip buvo minėta, lentelių analizė atliekama naudojant „r-packet“ programą, kurios pateikiami rezultatai nėra susiję su lietuvių kalba. Taip susiduriama su viena iš anksčiau minėtų problemų, kad dauguma programų skirta anglų kalbai. Todėl rezultatai bus pateikiami anglų kalba su lietuviškais paaiškinimais.

Matome, kad jungtukai dažniausiai pasitaiko 2 ir 4 raidžių, įvardžių dažniau yra 3 raidžių ilgio. „*Kita*“ kalbos dalis (beveik) statistiškai reikšmingai vyrauja, kai *ilgis* = 1 Statistiškai reikšmingai vyrauja jungtukai, kurių *ilgis* = 2 ir 4. Raudona spalva parodo, kurios kalbos dalys rečiau pasitaiko tarp nagrinėjamų žodžių ilgių. Labiausiai tikėtiną reikšmę atitiko kita kalbos dalis, kai *ilgis* = 3.

Toliau iškelsime hipotezę apie požymių statistinę priklausomybę:

$$\begin{cases} H_0 : p > \alpha = 0{,}05, \\ H_1 : p < \alpha = 0{,}05. \end{cases}$$

Hipotezė H_0 teigia, kad kalbos dalis ir žodžio ilgis yra nesusiję, H_1 – kalbos dalis ir žodžio ilgis yra susiję. Šiai hipotezei ištirti naudosime χ^2 testą (Chi-Square).

10 lentelė. χ^2 testas

Pearson's Chi-squared test
data: GSStab
X-squared = 484.2337, *df* = 6, *p*-value < 2.2e-16

Grafoje „Pearson's Chi-squared test“ randame $\chi^2 = 484.2337$ reikšmę, laisvės laipsnių $df = 6$ ir $p = 2.2 \cdot 10^{-16}$. Kadangi p reikšmė mažesnė už 0,05, hipotezę apie požymių nepriklausomumą atmetame – požymiai statistiškai priklausomi.

Dabar ištirsime kiekvieno laukelio reikšmingumą visai lentelei.

11 lentelė. Laukelių analizė

	KALBOS DALIS			
ILGIS	**JUNGTUKAS**	**ĮVARDIS**	**KITA**	**VISA EILUTĖ**
1	40 14,72 0,333 0,058 0,035	0 26.05 0 0 0	80 162,59 0,667 0,394 0,07	120 0,105
2	425 31,36 0,794 0,612 0,371	59 28,114 0,11 0,237 0,051	51 20,16 0,095 0,251 0,044	535 0,466
3	152 38,96 0,367 0,219 0,133	190 111,55 0,459 0,763 0,166	72 0,022 0,174 0,355 0,063	414 0,361
4	78 19,99 1 0,112 0,068	0 16,93 0 0 0	0 13,81 0 0 0	78 0,068
VISAS STULPELIS	695 0.606	249 0,217	203 0,177	1147

Šioje lentelėje matome, kiek yra duomenų kiekviename laukelyje. Taip pat yra suskaičiuoti χ^2 kiekvienam laukeliui bei kokią dalį laukelis sudaro eilutėje, stulpelyje ir visoje lentelėje. Iš šios lentelės sužinome, kuris laukelis yra reikšmingas, o kuris nedaro įtakos

visai lentelei. Matome, kad vienos raidės ilgio įvardžių nėra, o dviejų raidžių ilgio jungtukų daugiausiai, taigi šie duomenys visiškai skirtingai įtakoja visą lentelę.

2.4. LOGTIESINIAI MODELIAI

2.4.1. Logtiesiniai raidžių ir garsų modeliai

Atliekant tyrimus su logtiesiniais modeliais buvo naudojama „SAS“ programa. Buvo gauti rezultatai anglų kalba, todėl jie bus pateikiami su lietuviškais paaiškinimais. Šiam tyrimui duomenys buvo imami iš 15 skirtingų tekstų, atrenkant atsitiktinai po 1000 žodžių gabaliuką. Po to buvo tirta raidinės ir garsinės struktūros sąryšis su moksliniu ir grožiniu stiliumi. Gauti rezultatai pateikti 12 lentelėje.

12 lentelė. Logtiesinis raidžių modelis

Data Summary			
Response	il*ba*sk*be*by*bn*te*tip	**Response Levels**	525
Weight Variable	None	**Populations**	1
Data Set	RAIDES	**Total Frequency**	51919
Frequency Missing	0	**Observations**	51919

Maximum Likelihood Analysis of Variance			
Source	**DF**	**Chi-Square**	**Pr > ChiSq**
ilgis5	4	170.94	<.0001
balse1	1	222.52	<.0001
ilgis5*balse1	4	101.43	<.0001
by1	2	4765.84	<.0001
ilgis5*by1	8	115.10	<.0001
ilgis5*bei1*tipas2	4	27.11	<.0001
ilgis5*tipas2	4	36.52	<.0001
ilgis5*by1*tipas2	4*	26.24	<.0001
bnosine1	1	4595.25	<.0001
ilgis5*bnosine1	4	37.33	<.0001

Maximum Likelihood Analysis of Variance			
Source	**DF**	**Chi-Square**	**Pr > ChiSq**
ilgis5*bnosine1*tipas2	4	14.36	0.0062
skardus1	2	2306.33	<.0001
ilgis5*skardus1	8	179.53	<.0001
teksto_nr	14	61.59	<.0001
ilgis5*teksto_nr	52*	525.53	<.0001
skardus1*teksto_nr	28	37.50	0.1083
ilgis5*skardus1*teksto_n	112	142.37	0.0278
balse1*bei1*tipas2	1	3.27	0.0704
balse1*teksto_nr	14	22.12	0.0762
ilgis5*balse1*teksto_nr	56	76.42	0.0363
by1*teksto_nr	27*	83.22	<.0001
balse1*bnosine1*teksto_n	14	39.36	0.0003
balse1*by1*tipas2	0*	.	.
Likelihood Ratio	156	176.35	0.1265

Logtiesinių modelių tyrimui buvo apibrėžti tokie kintamieji: „teksto numeris“ (kintamasis „*teksto_nr*“), „teksto pavadinimas“, „teksto tipas“ (kintamasis „*tipas2“*), kuris nurodo grožinę ar mokslinę literatūrą. Visos raidės suskirstytos į grupes ir joms priskirti atitinkami požymiai. Kintamuosius: balsės (kintamasis „*balse1“*), jų rūšys: nosinės (kintamasis „*bnosine1“*), ilgosios (kintamasis „*by1*“), „švelniosios“ balsės **e** ir **i** (kintamasis „*bei1*“), o taip pat priebalsių rūšies požymį, pusbalsiai, skardžiosios priebalsės ir dusliosios priebalsės (kintamasis „*skardus1*“). Taip pat buvo apskaičiuoti visų žodžių ilgiai (kintamasis „*ilgis*“) bei visų tipų raidžių kiekiai kiekviename žodyje. Kintamasis „*ilgis*“ sugrupuotas („sustambinus“), apjungus į 3 arba 5 didesnes grupes. Pavyzdžiui, kintamasis „*ilgis5*“ yra lygus 1, jeigu žodžio ilgis < 4.

Atitinkamų logtiesinio modelio narių statistinis reikšmingumas pateiktas paskutiniame 12 lentelės stulpelyje, kuriame pateiktos visų į modelį įtrauktų veiksnių bei jų sąveikų (kombinacijų) atitinkamos *p* reikšmės. Tradiciškai veiksnys (faktorius, požymis) laikomas statistiškai reikšmingu, jeigu atitinkama *p* reikšmė yra mažesnė už reikšmingumo lygmenį

$\alpha = 0.05$. Iš lentelės matome, kad šios sąlygos neišpildo požymiai „*skardus1*teksto_nr“*, „*balse1*bei1*tipas2“*, „*balse1*teksto_nr“*. Kadangi atitinkama *p* reikšmė yra nedaug didesnė už 0,05, tai ši sąveika būtų beveik reikšminga, jeigu būtume pasirinkę „tolerantiškesnį“ reikšmingumo lygmenį $\alpha = 0.1$ vietoje tradicinio $\alpha = 0.05$. Galima tikėtis, kad ši sąveika būtų reikšminga, jeigu tyrime būtų naudojama didesnė imtis (daugiau duomenų).

Reikšmingumo lygmenį $\alpha = 0.05$ viršija *p* reikšmė 0.1265 paskutinėje lentelės eilutėje. Šiuo atveju tai – geras požymis. Jis reiškia, kad parinktas modelis pakankamai gerai atspindi turimus duomenis, todėl nulinę hipotezę apie parinkto modelio adekvatumą atmesti nėra pagrindo.

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

a) Nustatyti stabilias, invariantiškas proporcijas tarp įvairių raidžių tipų lietuviškuose tekstuose; klausimas: kuo visi tekstai panašūs?

b) Nustatyti tose proporcijose (galbūt) pasireiškiančius (ženklius) skirtumus tarp grožinės ir mokslinės literatūros; klausimas: kuo skiriasi grožinės ir mokslinės literatūros tekstai?

Atsakymui į a) klausimą svarbios yra raidžių grupės, kurios nėra susijusios (neturi sąveikos nario) su teksto numeriu (kintamuoju „*teksto_nr*“). Tai reiškia, kad tų raidžių grupių proporcijos buvo maždaug tokios pačios visuose (nagrinėtuose) tekstuose.

Atsakymui į b) klausimą – priešingai, yra svarbios tos raidžių grupės, kurios yra susijusios (turi sąveikos narį) su kintamuoju „*tipas2*“. Tai reiškia, kad tų raidžių grupių proporcijos statistiškai reikšmingai skyrėsi moksliniuose ir grožinės literatūros tekstuose.

a) Lietuviškų tekstų invariantas yra ilgųjų balsių (jų kiekį aprašo kintamasis „*by1“*) proporcija (proporcija tarp balsių), kuri, žinoma, priklauso nuo žodžio ilgio. Tas pats galioja ir balsėms su nosinėmis (jų kiekį aprašo kintamasis „*bnosine1“*). Priebalsių ir balsių proporcija bei priebalsių tipų proporcijos gali labai skirtis įvairiuose tekstuose, jie nėra „invariantas“. Tai atspindi statistiškai labai reikšminga kintamųjų „*balse1“* ir „*skardus1“* sąveika su „*teksto_nr“*. Priebalsių tipų proporcijos žodyje taip pat labai priklauso ir nuo žodžio ilgio.

b) Mokslinių ir grožinės literatūros tekstų skirtumai pasireiškia ilgųjų ir nosinių balsių proporcijų skirtumuose, taip pat jų skirtingu pasiskirstymu įvairaus ilgio žodžiuose. Tai parodo statistiškai reikšmingos sąveikos *balse”*, „*bnosine1“* ir „*tipas2“*, „*ilgis5“*, „*by1“* ir „*tipas2“*. Mokslinėje ir grožinėje literatūroje, matyt, yra skirtingas ir balsių proporcijų pasiskirstymas žodžio ilgio atžvilgiu, nors jau minėta atitinkama sąveika „*skardus1*teksto_nr“, „balse1*bei1*tipas2“, „balse1*teksto_nr“* ir nėra statistiškai reikšminga su $\alpha = 0.05$.

2.4.2. Logtiesiniai skyrybos ženklų modeliai

Taip pat buvo atliktas analogiškas tyrimas su skyrybos ženklais. Atsitiktine tvarka buvo imama po 500 žodžių.

13 lentelė. Logtiesinis skyrybos ženklų modelis

Data Summary			
Response	sak*sky*kab*tek*rep*tipa	**Response Levels**	271
Weight Variable	None	**Populations**	1
Data Set	TXT_PJ	**Total Frequency**	22500
Frequency Missing	0	**Observations**	22500

Maximum Likelihood Analysis of Variance			
Source	**DF**	**Chi-Square**	**Pr > ChiSq**
teksto_nr	14	138.59	<.0001
kablelis*teksto_nr	14	174.81	<.0001
skyris*kablelis*tipas2	1	11.60	0.0007
sakinys*skyris*tipas2	1	11.48	0.0007
sakinys*teksto_nr	14	207.58	<.0001
skyris	1	17.57	<.0001
kablelis	1	39.59	<.0001
skyris*kablelis	1	36.49	<.0001
sakinys	1	57.26	<.0001
sakinys*skyris	1	36.44	<.0001
sakinys*kablelis	1	64.22	<.0001
sakinys*skyris*kablelis	1	17.78	<.0001
Likelihood Ratio	219	224.41	0.3866

Šiame modelyje atsiranda naujų požymių: „*kablelis*“, „*skyris*“, „*sakinys*“.

Atitinkamų logtiesinio modelio narių statistinis reikšmingumas pateiktas paskutiniame 13 lentelės stulpelyje, kuriame pateiktos 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 $\alpha = 0.05$. Iš lentelės matome, kad ši sąlyga buvo išpildyta visais atvejais.

Reikšmingumo lygmenį $\alpha = 0.05$ ženkliai viršija p reikšmė 0.3866 paskutinėje lentelės eilutėje. Šiuo atveju tai – geras požymis. Jis reiškia, kad parinktas modelis pakankamai gerai atspindi turimus duomenis, todėl nulinę hipotezę apie parinkto modelio adekvatumą atmesti nėra pagrindo.

Kaip matome, tekstuose esantys skyrybos ženklai ir sakiniai yra statistiškai priklausomi nuo tekstų stilių.

IŠVADOS

Kalbos elementų skirstymas pagal kiekybinius požymius nėra įprastas tradicinėje kalbotyroje. Įvairūs kalbos elementai paprastai tiriami kokybiškai, nes atsižvelgiama į jų priklausomumą tam tikroms teorinėms klasėms, pvz., morfologinėms, semantinėms, sintaksinėms. Kadangi kiekybiniai požymiai dažnai gali paaiškinti, patvirtinti, ar paneigti kokybinės analizės išvadas, todėl pirmiausiai būna atliekami kiekybiniai tyrimai.

Atlikus visus skaičiavimus, ištyrus priklausomybes, išnagrinėjus lenteles bei sudarius logtiesinius modelius, buvo gauta nemažai statistiškai reikšmingų rezultatų. Išanalizavus duomenis gauta nemažai būdų, kaip galima atskirti mokslinį stilių nuo grožinio.

- Grožinis stilius nuo mokslinio skiriasi:

žodžių ilgiu ir jų dažniu;
sakinių kiekiu tekstuose;
naudojamais simboliais;
vartojamų žodelių, pvz. **ir**, **ar**, **į**, **yra**..., kiekiu.

Skirtumai pasireiškia ilgųjų ir nosinių balsių proporcingume.

Taip pat buvo gauta ir kitokių rezultatų.

- Jungtuko **ir** vartojimas sudaro apie 4% visų tekstuose vartojamų žodžių.
- Žodžių ilgio ir jų dažnio priklausomybė yra stipri neigiama.
- Kalbos dalys ir žodžių ilgiai yra statistiškai priklausomi.
- Tekstų stiliai ir žodžių ilgiai yra statistiškai priklausomi.
- Mozaika vizualiai parodo, kaip lentelėje esantys dažniai nukrypsta į viena ar į kitą pusę nuo tikėtinos to dažnio reikšmės.
- Statistiškai reikšmingi yra sąryšiai tarp balsių, balsių ir priebalsių, raidžių ir tekstų, raidžių ir žodžių bei jų ilgių.

LITERATŪROS SĄRAŠAS

1. APYNIS A.; Stankus E., *Matematika*. Vilnius, (2001).
2. BAKŠTYS A. *Skaičiavimo statistika*. Šiauliai, (2007).
3. BAKŠTYS A. *Statistika ir tikimybė*. Vilnius, (2006).
4. ČEKANAVIČIUS V.; MURAUSKAS G. *Statistika ir jos taikymai II dalis*. Vilnius, (2006).
5. ČEKANAVIČIUS, V.; MURAUSKAS G. *Statistika ir jos taikymai I dalis*. Vilnius. (2004).
6. http://coralit.lt/ Prieiga per internetą [žiūrėta 2010-05-18]
7. http://donelaitis.vdu.lt/ Prieiga per internetą [žiūrėta 2010-05-18]
8. http://protas.pypt.lt/matematika/statistika Prieiga per internetą [žiūrėta 2010-05-18]
9. http://staneikaite.wordpress.com/2009/10/29/dazniu-lenteles/ Prieiga per internetą [žiūrėta 2010-05-18]
10. KANIŠAUSKAS V. *Tikimybių teorijos ir matematinės statistikos pagrindai*. Šiauliai, (2000)
11. MAUZIENĖ, L. Lingvistiniai ir psichologiniai lingvodidaktikos pagrindai, *Santalka*, Filologija, Edukologija, **17(2)**, 61–67 (2009).
12. NAVAKAUSKAS, D.; et al. *Šiuolaikinės SSA priemonės*, in: http://www2.el.vgtu.lt/ssa/node3.html (2002).
13. PULLUM, G. K.; KORNAI, A. *Mathematical Linguistics*, in http://www.metacarta.com/Collateral/ Documents/English-US/Mathematical-linguistics-Kornai.pdf.
14. RAŠKINIS, G.; RAŠKIENĖ, D. Lietuvių šnekos atpažinimo sistemos, pagrįstos paslėptaisiais Markovo modeliais, parametrų tyrimas ir optimizacija, *Informacinės technologijos*, IX 41–48 (2003).
15. RIMKUTĖ, E.; DAUDARAVIČIUS, V. Morfologinis dabartinės lietuvių kalbos tekstyno anotavimas, *Kalbų studijos*, **11**, 30–35 (2007).
16. RIMKUTĖ, E.; GRIGONYTĖ, G. Automatizuotas lietuvių kalbos morfologinio daugiareikšmiškumo ribojimas, *Kalbų studijos*, **9**, 30–37 (2006). [13]
17. RIMKUTĖ, E.; KOLEVSKAITĖ, J. Mašininis vertimas – greitoji pagalba globalėjančiam pasauliui, *Gimtoji kalba*, **9** (2007).
18. TAMULEVIČIUS, G. *Pavienių žodžių atpažinimo sistemų kūrimas*: daktaro disertacija, Vilnius (2008).

19. USONIENĖ, A. GRIGALIŪNIENĖ, J.; et al. http://www.leidykla.eu/fileadmin/Baltistika/43-1/09_Usoniene_ir_kt.pdf
20. UTKA, A. *Dažninis rašytinės lietuvių kalbos žodynas*: 1 milijono žodžių morfologiškai anotuoto tekstyno pagrindu, 2009. // Donelaitis.vdu.lt/publikacijos/Dazninis_zodynas.pdf
21. UTKA, A. Labai dažnų lietuvių kalbos žodžių ir žodžių formų ypatybės, *Lituanistica*, **1(61)**, 48–55 (2005).
22. ZINKEVIČIUS, V. *Lemuoklis* – morfologinei analizei, *Darbai ir dienos*, **24**, 245–273 (2000).

The comparison of Lithuanian texts‘ styles by using the universal quantitative characteristics

SUMMARY

The processes of the language computerization is developing all over the world, and Lithuania is not an exception. There are lots of new programs developed on purpose to correct texts, to check spelling mistakes, to recognize words, to create e-dictionaries and etc. A lack of information and many not discovered and not analyzed linguistic fields arise because of these fast development. In the cultural approach it is necessary to strengthen the scientific research in pursuance to recognize the Lithuanian language and its synthesis and to translate it into other languages. This would help to keep the use of the Lithuanian language in this modern digital environment. This final master‘s thesis is based on research according these aspects.

The recency of the thesis: the statistical researches and adaptation of the Lithuanian language‘s styles in computer programs.

The topicality of the thesis: the fostering of the Lithuanian language. The utilization of the research results for texts‘ interpretations.

The goal: to analyze the styles of the Lithuanian language by using the statistical methods. Tasks:

1. To analyze statistically the fiction and scientific literature styles.
2. To investigate dependence of word frequencies on its‘ length by using the correlation and regression analysis.
3. To visualize and analyze multivariate frequency tables.
4. To apply the loglinear models.

The object of research: multivariate statistical methods in the Lithuanian language processing. Problems: Lithuanian language is quite complex, the alternation of genders, declension. The programs and programming are not familiar to the Lithuanian language.

In conclusion, the results of the research revealed that the fiction and scientific literature styles are different in: the lengthes and frequencies of the words, the amount of sentences, proportions of symbols, the amounts of certain auxiliary words that are used (for instance: “and”, “if”, “to” “is” and etc.). The dependence of word frequency on its length is strong but negative. Parts of speech and length of words are statistically related. The same is valid for styles of texts and the length of words – they have strong relevance. For visualization of frequency tables and their deviation from the expected frequencies under model the software “MOSAIC” is applied.

PRIEDAI

1 priedas

Grožinio stiliaus lentelė																
Ištraukos po 1000 žodžių	**sakiniai**	**o**	**į**	**ir**	**ar**	**su**	**yra**	**kad**	**bet**	**tai**	**kaip**	**ji**	**jis**	**jie**	**jos**	**jų**
	80	4	11	46	2	2	8	10	12	12	6	3	14	2	4	3
	77	6	17	49	1	4	2	5	2	1	10	0	29	8	4	1
	122	2	6	22	11	7	0	16	10	12	6	1	12	9	3	2
	91	8	6	32	9	1	0	4	8	5	13	1	3	0	5	6
	98	4	17	23	1	7	0	15	11	2	5	4	9	5	3	3
Vidurkis	**93,6**	**4,8**	**11,4**	**34,4**	**4,8**	**4,2**	**2**	**10**	**8,6**	**6,4**	**8**	**1,8**	**13,4**	**4,8**	**3,8**	**3**
Dispersija	**258,64**	**4,16**	**24,24**	**127**	**18,6**	**6,16**	**9,6**	**24**	**12,6**	**22,6**	**9,2**	**2,16**	**74,6**	**11,8**	**0,56**	**2,8**
Pasiskirstymas %		**0,48**	**1,14**	**3,44**	**0,48**	**0,42**	**0,2**	**1**	**0,86**	**0,64**	**0,8**	**0,18**	**1,34**	**0,48**	**0,38**	**0,3**
Suma	**468**	**24**	**57**	**172**	**24**	**21**	**10**	**50**	**43**	**32**	**40**	**9**	**67**	**24**	**19**	**15**

2 priedas

Mokslinio stiliaus lentelė																
Ištraukos po 1000 žodžių	**sakiniai**	**o**	**į**	**ir**	**ar**	**su**	**yra**	**kad**	**bet**	**tai**	**kaip**	**ji**	**jis**	**jie**	**jos**	**jų**
	70	3	5	39	3	9	6	2	1	7	7	0	2	2	6	11
	70	9	6	27	8	1	10	13	3	4	7	11	4	1	4	1
	59	2	3	37	4	8	18	18	2	3	3	0	0	1	1	4
	75	1	2	44	3	10	8	7	4	2	11	1	2	0	2	3
	43	1	7	57	7	2	20	7	2	2	10	3	0	0	5	1
VIDURKIS	**63,4**	**3,2**	**4,6**	**41**	**5**	**6**	**12**	**9,4**	**2,4**	**3,6**	**7,6**	**3**	**1,6**	**0,8**	**3,6**	**4**
Dispersija	**131,44**	**8,96**	**3,44**	**96**	**4,4**	**14**	**31**	**31**	**1**	**3,4**	**7,8**	**17,2**	**2,24**	**0,56**	**3,44**	**13,6**
Pasiskirstymas %		**0,32**	**0,46**	**4,1**	**0,5**	**0,6**	**1**	**0,9**	**0,2**	**0,4**	**0,8**	**0,3**	**0,16**	**0,08**	**0,36**	**0,4**
SUMA	**317**	**16**	**23**	**204**	**25**	**30**	**62**	**47**	**12**	**18**	**38**	**15**	**8**	**4**	**18**	**20**

3 priedas

Simpsono paradoksas

Turime tokius duomenis:

	Lytis			
	V		M	
Rezultatas	S	N	S	N
Gydymas 1	60	20	40	80
Gydymas 2	100	50	10	30

Lentelėje naudojami sutrumpinimai:

S – sėkmė, *N* – nesėkmė, *V* – vyras, *M* – moteris.

Suskaičiuojame:

Rezultatas	S	N
Gydymas 1	100	100
Gydymas 2	110	80

Pastebėkime, kad:

Vyrams sėkmingas	
Gydymas 1	60/80=75%
Gydymas 2	100/150=67%

Moterims sėkmingas	
Gydymas 1	40/120=33%
Gydymas 2	10/40=25%

Pirmasis gydymo būdas akivaizdžiai geresnis tiek vyrams, tiek moterims.

Dabar suskaičiuokime rezultatus lentelei, agreguotai pagal lytį.

Turint dideles lenteles su daug požymių, norisi kai kurias požymių reikšmes apjungti arba net agreguoti pagal dalį požymių.

Gydymas 1	100/200=50%
Gydymas 2	110/190=58%

2-asis gydymo būdas yra geresnis -> PARADOKSAS

Kada tai galima daryti ?

Būtinos sąlygos yra gana sudėtingos. Todėl jų neminėsime.

Pakankamos sąlygos.

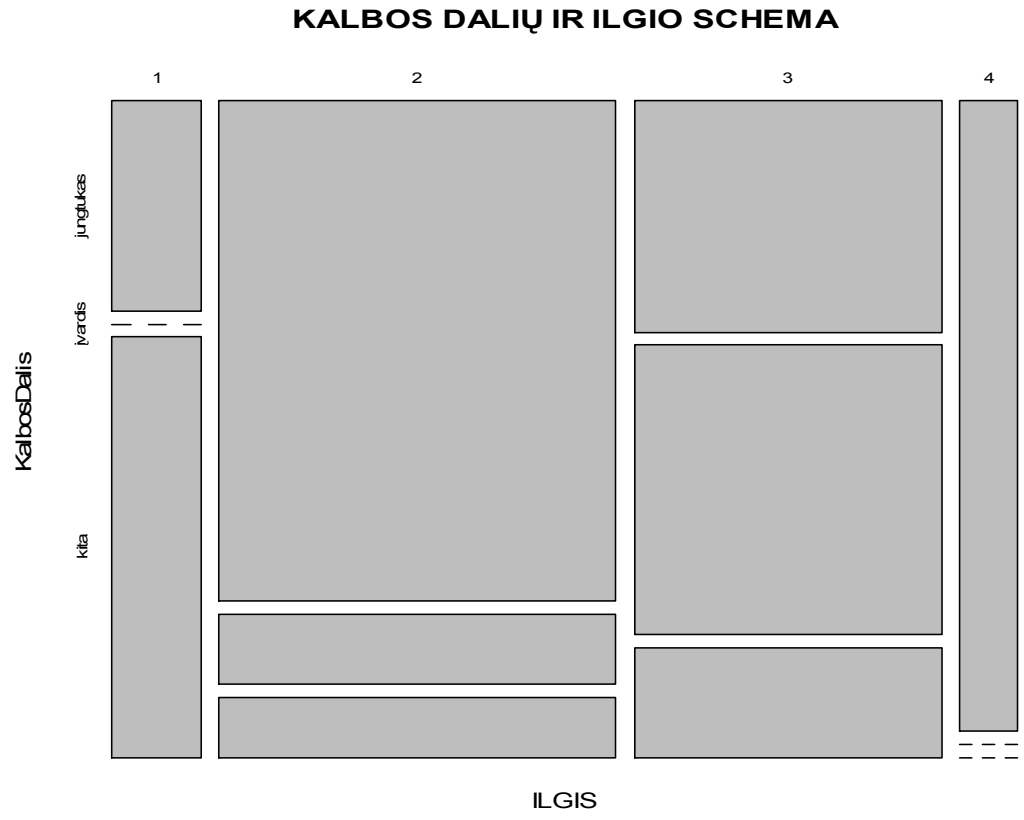
Imant požymius iš A ir B galima agreguoti pagal C, jeigu A ir C sąlyginai nepriklausomi, kai duotos požymių iš B reikšmės.

4 priedas

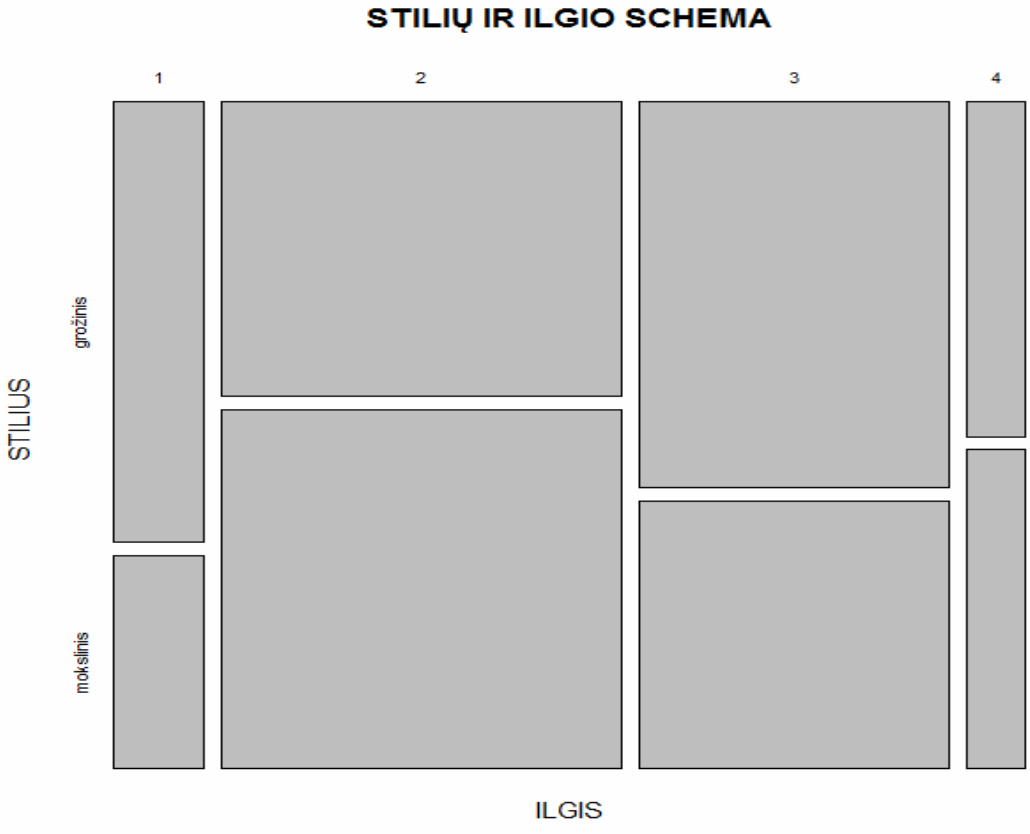
Koreliacijos koeficiento priklausomumo lentelė

ρ reikšmė	Koreliacija
$\rho = 0$	Atsitiktiniai dydžiai nekoreliuoti
$0 < \lvert\rho\rvert < 0.3$	labai silpna (*teigiama, neigiama*)
$0.3 \le \lvert\rho\rvert < 0.5$	silpna *(teigiama, neigiama*)
$0.5 \le \lvert\rho\rvert < 0.7$	vidutinė (*teigiama, neigiama*)
$0.7 \le \lvert\rho\rvert < 0.9$	stipri (*teigiama, neigiama*)
$0.9 \le \lvert\rho\rvert \le 1$	labai stipri (*teigiama, neigiama*)

5 priedas



6 priedas



Programos „R-packet“ kodas, su kuria buvo atlikta dauguma skaičiavimų.

7 priedas

Pradinių duomenų analizei.

```
#TEKSTU ANALZE
setwd("d:/tekstai")
#scan("grozinis.txt", nlines=497, what="")
zodziaiG= scan("grozinis.txt", nlines=497, what="")
ilgisG=nchar(zodziaiG)
 ilgisG[1:10]
#scan("mokslinis.txt", nlines=523, what="")
zodziaiM= scan("mokslinis.txt", nlines=523, what="")
 ilgisM=nchar(zodziaiM)
 ilgisM[1:10]
#SKAICIAVIMAI
sum(ilgisG)   #raidziu kiekis
sum(ilgisM)    #raidziu kiekis
 max(ilgisG)   #maximumas
  max(ilgisM)   #maximumas
 min(ilgisG)    #minimumas
 min(ilgisM)    #minimumas
 median(ilgisG)   #mediana
 median(ilgisM)   #mediana
 mean(ilgisG)      #vidurkis
 mean(ilgisM)      #vidurkis
 var(ilgisG)       #dispersija
 var(ilgisM)       #dispersija
 sd(ilgisG)     #standartinis nuokrypis
 sd(ilgisM)     #standartinis nuokrypis
table(ilgisG)   #dazniu lentele
table(ilgisM)   #dazniu lentele
op <-par(mfrow=c(2,1))  #SUJUNGE 2 GRAFIKUS
 hist(ilgisG, breaks=c(seq(from=0.5, to=19.5)))
 abline(v=mean(ilgisG),col=2,lty=2,lwd=2)   #vidurkio linija
 abline(v=median(ilgisG),col=3,lty=3,lwd=2)   #medianos linija
 hist(ilgisM, breaks=c(seq(from=0.5, to=20.5)))
 abline(v=mean(ilgisM),col=2,lty=2,lwd=2)      #vidurkio linija
abline(v=median(ilgisM),col=3,lty=3,lwd=2)      #medianos linija
```

8 priedas

Lentelių analizei.

```
library(gmodels)
library(MASS)
library(vcd)
library(colorspace)
library(grid)
library(gridBase)
#BENDRA LENTELE: tyrimas is 10000 žodzių.
GSS <- data.frame(expand.grid(STILIUS=c("grožinis" , "mokslinis") , ILGIS=c("1" , "2", "3"
, "4"),
```

```
 KalbosDalis=c("jungtukas" , "įvardis" , "kita")),
KIEKIS=c(24,16,196,229,93,59,40,38,0,0,24,35,142,48,0,0,57,23,21,30,10,62,0,0))
GSS              # lentele
str(GSS)         # informacija apie lenteles struktura
sum(GSS$KIEKIS)    # suma
plot(GSS, main="DAUGIAMATĖS LENTELĖS SCHEMA")
summary(GSS)
(GSStab <- xtabs(KIEKIS ~ ILGIS + KalbosDalis, data=GSS))  #matrica
summary(GSStab)          #informacija
as.data.frame(GSStab)      #lentele
plot(GSStab, main="KALBOS DALIŲ IR ILGIO SCHEMA")
mosaic(GSStab, gp = shading_max, split_vertical = TRUE, main="KALBOS DALIŲ
MOZAIKA")
CrossTable(GSStab)
chisq.test(GSStab)
fisher.test(GSStab)
assocstats(GSStab)
(GSStabs <- xtabs(KIEKIS ~ ILGIS + STILIUS, data=GSS))  #matrica
summary(GSStabs)          #informacija
as.data.frame(GSStabs)      #lentele
plot(GSStabs, main="STILIŲ IR ILGIO SCHEMA")
mosaic(GSStabs, gp = shading_max, split_vertical = TRUE, main="STILIŲ MOZAIKA")
CrossTable(GSStabs)
chisq.test(GSStabs)
fisher.test(GSStabs)
assocstats(GSStabs)
```