VILNIAUS GEDIMINO TECHNIKOS UNIVERSITETAS
FUNDAMENTINIŲ MOKSLŲ FAKULTETAS
MATEMATINĖS STATISTIKOS KATEDRA

Monika Lapėnaitė-Gedvilė

KALBOS DALIŲ PASISKIRSTYMO LIETUVIŠKUOSE TEKSTUOSE ANALIZĖ IR PROGNOZĖ

THE ANALYSIS AND PREDICTION OF THE PART OF SPEECH DISTRIBUTIONS IN LITHUANIAN TEXTS

Baigiamasis magistro darbas

Taikomosios statistikos studijų programa, valstybinis kodas 621G31001
Statistinių metodų finansuose ir ekonomikoje specializacija
Statistikos studijų kryptis

Vilnius, 2014

VILNIAUS GEDIMINO TECHNIKOS UNIVERSITETAS
FUNDAMENTINIŲ MOKSLŲ FAKULTETAS
MATEMATINĖS STATISTIKOS KATEDRA

TVIRTINU
Katedros vedėjas

_____
(Parašas)

_____
(Vardas, pavardė)

_____
(Data)

Monika Lapėnaitė-Gedvilė

KALBOS DALIŲ PASISKIRSTYMO LIETUVIŠKUOSE TEKSTUOSE ANALIZĖ IR PROGNOZĖ

THE ANALYSIS AND PREDICTION OF THE PART OF SPEECH DISTRIBUTIONS IN LITHUANIAN TEXTS

Baigiamasis magistro darbas

Taikomosios statistikos studijų programa, valstybinis kodas 621G31001
Statistinių metodų finansuose ir ekonomikoje specializacija
Statistikos studijų kryptis

Vadovas doc. dr. Marijus Radavičius _____ _____
(Parašas) (Data)

Lietuvių kalbos konsultantas doc. dr. Angelė Kaulakienė _____ _____
(Parašas) (Data)

Vilnius,
2014

Vilniaus Gedimino technikos universitetas
Fundamentinių mokslų fakultetas
Matematinės statistikos katedra

ISBN ISSN
Egz. sk. .....
Data .....-.....-.....

Antrosios pakopos studijų **Taikomosios statistikos** programos magistro baigiamasis darbas 4

Pavadinimas **Kalbos dalių pasiskirstymo lietuviškuose tekstuose analizė ir prognozė**

Autorius **Monika Lapėnaitė-Gedvilė**

Vadovas **doc. dr. Marijus Radavičius**

Kalba: lietuvių

Anotacija

Baigiamajame magistro darbe yra nagrinėjami tokie klausimai: ar naudojantis statistine informacija apie dažnas žodžio formas galima prognozuoti žodžio formų su tam tikromis savybėmis pasitaikymo (lietuviškame) tekste dažnius, kokiu tikslumu, kaip tai priklauso nuo teksto autoriaus? Darbe apžvelgti ankstesnių mokslinių tyrimų rezultatai. Siekiant išsiaiškinti atrinktų žymeklių tinkamumą linksniuojamų kalbos dalių prognozavimui bei jų ryšį su autoriais, atlikta pirminė statistinė ir koreliacinė analizė bei remtasi apibendrintų tiesinių modelių teorija. Sudaryti logistinės ir Puasono regresijų modeliai ir įvertintas jų tinkamumas trims reprezentatyviausioms kalbos dalių grupėms. Išnagrinėjus teorinius ir praktinius baigiamojo darbo aspektus, pateikiamos išvados ir rekomendacijos.

Tyrime naudojami mokykloms skirti suskaitmeninti lietuvių grožinės literatūros kūriniai. Skaičiavimai atlikti su paketu R.

Darbą sudaro 6 dalys: įvadas, ankstesnių mokslinių tyrimų apžvalga, analitinė – metodinė dalis, eksperimentinė – tiriamoji dalis, išvados ir rekomendacijos, literatūros sąrašas.

Darbo apimtis – 71 p. teksto be priedų, 9 pav., 33 lent., 27 bibliografiniai šaltiniai.

Atskirai pridedami darbo priedai.

Prasminiai žodžiai: apibendrintasis tiesinis modelis; funkciniai žodžiai; logistinė regresija; prognozė; Puasono regresija; ROC kreivė;

Vilnius Gediminas Technical University
Faculty of Fundamental Sciences
Department of Mathematical Statistics

ISBN ISSN
Copies No. .....
Date .....-.....-.....

Master Degree Studies **Applied Statistics** study programme Master Graduation Thesis 4

Title **The analysis and prediction of the part of speech distributions in Lithuanian texts**

Author **Monika Lapėnaitė-Gedvilė**

Academic supervisor **Assoc Prof Dr Marijus Radavičius**

Thesis language:
Lithuanian

Annotation

In the master thesis the following problems are considered: if it is possible to predict the frequency of occurrences of word forms with specific properties in Lithuanian texts using statistical information about frequent word forms, to what occuracy and how it depends on authors? Results of previous studies are outlined. In order to ascertain the suitability of the selected markers for prediction of inflective parts of speech and relations of the markers with authors, primary statistical analysis and correlation analysis have been performed and generalized linear models have been applied. Logistic and Poisson regressions models are composed for three the most representative groups of parts of speech and suitability of these models are assessed. After the examination of the practical and theoretical aspects, the conclusions and recommendations have been presented.

Lithuanian digitized literary works for schools are used in the study. Calculations are performed with R.

Thesis consists of 6 parts: introduction, review of previous studies, analytical - methodical part, experimental - research part, conclusions and suggestions, references.

Thesis consists of: 71 p. text without appendixes, 9 pictures, 33 tables, 27 bibliographical entries. Appendixes included.

Keywords: generalized linear model; functional words; logistic regression; prediction; Poisson regression; ROC curve;

Vilniaus Gedimino technikos universiteto egzaminų, sesijų ir baigiamųjų darbų rengimo bei gynimo organizavimo tvarkos aprašo 2011-2012 m. m.

1 priedas

(Baigiamojo darbo sąžiningumo deklaracijos forma)

VILNIAUS GEDIMINO TECHNIKOS UNIVERSITETAS

Monika Lapėnaitė-Gedvilė, 20121713

(Studento vardas ir pavardė, studento pažymėjimo Nr.)

Fundamentinių mokslų fakultetas

(Fakultetas)

Taikomoji statistika, TSMfm-12

(Studijų programa, akademinė grupė)

BAIGIAMOJO DARBO (PROJEKTO) SĄŽININGUMO DEKLARACIJA

2014 m. birželio 4 d.

Patvirtinu, kad mano baigiamasis darbas tema „Kalbos dalių pasiskirstymo lietuviškuose tekstuose analizė ir prognozė" patvirtintas 2013 m. spalio 24 d. dekano potvarkiu Nr. 360fm, yra savarankiškai parašytas. Šiame darbe pateikta medžiaga nėra plagijuota. Tiesiogiai ar netiesiogiai panaudotos kitų šaltinių citatos pažymėtos literatūros nuorodose.

Mano darbo vadovas doc. dr. Marijus Radavičius.

Kitų asmenų indėlio į parengtą baigiamąjį darbą nėra. Jokių įstatymų nenumatytų piniginių sumų už šį darbą niekam nesu mokėjęs (-usi).

	Monika Lapėnaitė-Gedvilė
(Parašas)	(Vardas ir pavardė)

VILNIAUS GEDIMINO TECHNIKOS UNIVERSITETAS
FUNDAMENTINIŲ MOKSLŲ FAKULTETAS
MATEMATINĖS STATISTIKOS KATEDRA

TVIRTINU
Katedros vedėjas

Statistikos studijų kryptis

Taikomosios statistikos studijų programa, valstybinis kodas 621G31001

Statistinių metodų finansuose ir ekonomikoje specializacija

_____
(Parašas)

_____
(Vardas, pavardė)

_____
(Data)

BAIGIAMOJO MAGISTRO DARBO UŽDUOTIS

…….....Nr. .....
Vilnius

Studentui (ei) .....Monikai Lapėnaitei-Gedvilei.....….....….....
(Vardas, pavardė)

Baigiamojo darbo tema: ..... Kalbos dalių pasiskirstymo lietuviškuose tekstuose analizė ir prognozė …..……

patvirtinta 2013m. spalio mėn. 24 d. dekano potvarkiu Nr. 360 fm

Baigiamojo darbo užbaigimo terminas 2014m. birželio 04 d.

BAIGIAMOJO DARBO UŽDUOTIS:

Ištirti, ar galima, panaudojant tinkamai sudarytą žymeklių rinkinį pakankamai tiksliai įvertinti kitų žodžių linksnių vartojimo tekste dažnumus ir proporcijas. Žymekliais vadinami tekste dažnai pasitaikantys (funkciniai) žodžiai, glaudžiai susiję, kaip manoma, su tiriama savybe. Tyrimui naudoti kūrinius iš suskaitmenintos lietuvių grožinės literatūros, skirtos mokykloms.

.....

.....

.....

Baigiamojo darbo rengimo konsultantai:

doc. dr. Angelė Kaulakienė .....
(Moksl. laipsnis/pedag. vardas, vardas, pavardė)

Vadovas ....doc. dr. Marijus Radavičius.....
(Paraša) (Moksl. laipsnis/pedag.vardas, vardas, pavardė)

Užduotį gavau

……..
(Parašas)
...Monika Lapėnaitė-Gedvilė...
(Vardas, pavardė)
…...……2012-10-01……....
(Data)

Turinys

PAVEIKSLŲ SĄRAŠAS

LENTELIŲ SĄRAŠAS

1. ĮVADAS

Pastaraisiais dešimtmečiais vyksta ypač spartus informacinių technologijų vystymasis, kuris paveikia įvairias mokslo sritis ir kasdieninį gyvenimą. Tai palietė ir labai seną mokslo sritį - kalbotyrą (lingvistiką). Galingesni kompiuteriai leidžia kaupti didelius duomenų kiekius, kurie gali būti saugomi tiek tekstinio, tiek audio formatu. Yra kuriami įvairūs tekstų redaktoriai, vertėjai, elektroniniai žodynai, programos įvairiems žodžiams, kalboms atpažinti. Pažengus technologijoms, jau yra norima ne tik atpažinti atskirus žodžius, bet ir sakinius, jų struktūrą. Šiam tikslui pasiekti reikalingi išsamūs kalbos tyrimai.

Kalba, kaip ir bet koks kitas gamtos, visuomenės ar fizikinis reiškinys, gali būti tyrinėjama įvairiais metodais. Paskutiniais dešimtmečiais ypač suaktyvėjo statistinių metodų taikymas lingvistikoje, tačiau lietuvių kalboje atliktų tyrimų yra mažai.

Lietuvių kalba yra sintetinė kalba, kuri turi žodžių kaitymą giminėmis, skaičiais, linksniais, nuosakomis, laikais, asmenimis, laipsniais, laisvą žodžių tvarką sakiniuose, nėra apibrėžtų griežtų sakinio struktūros taisyklių bei pasižymi morfologinio daugiareikšmiškumo savybe, todėl atsiranda įvairių problemų su automatiniu teksto apdorojimu. Taip pat tik labai maža dalis žodžių tekstuose yra vartojami dažnai, didelė dalis vos vieną kartą bei lieka labai daug žodžių, kurie iš viso nepasitaiko tekstuose.

Nagrinėjant šias problemas kyla klausimas: ar remiantis dažniau pasitaikančiomis žodžio formomis ar kitais dariniais, jų pasiskirstymu tekste, galima nuspėti kitų (retesnių) žodžio formų (ar kitų lingvistinių objektų) savybių skirstinius. Viena iš kalbos savybių, kuri galėtų padėti nuspėti nežinomą sakinio struktūrą ar padėti spręsti kitas aktualias problemas yra linksnis. Tai yra viena iš pagrindinių kalbos savybių, apimanti didelę dalį žodžių bei sutinkama praktiškai kiekviename sakinyje. Būtent dėl šių priežasčių ji buvo pasirinkta tyrimams. Taip pat darbe akcentuojama dažnų žodžio formų ypatybė: kokiu laipsniu jos reprezentuoja nepriklausomas nuo autorių, kitaip tariant, būdingas pačiai kalbai lietuviškų tekstų savybes, sąryšius tarp įvairių lingvistinių objektų.

Nustatyti visų linksniuojamų kalbos dalių linksnius yra didelių laiko ir kompiuterio resursų reikalaujantis uždavinys. Tačiau jeigu uždavinys yra apytiksliai įvertinti įvairių linksnių ir jų kombinacijų pasitaikymo tekste dažnius, tai galbūt nėra būtina nustatyti visų linksniuojamų žodžių linksnius, nes pakankamai tiksliai (statistinių paklaidų tikslumu) tuos dažnius pavyks prognozuoti pasirinktų žymeklių pagrindu. Tačiau šios hipotezės tikrinimas reikalauja daug techninio darbo, kadangi ekspertas turi nustatyti (suskaičiuoti) dominančio požymio pasireiškimo lygį visame duomenų rinkinyje. Be to, net su pakankamai galingais kompiuteriais apdoroti tokio dydžio duomenų rinkinį užtrunka labai daug laiko. Dar vienas tyrimo etapas reikalaujantis itin didelių darbo ir laiko resursų yra žymeklių paieška - tai sudėtingas uždavinys ir kol jis nėra išspręstas, net neaišku, ar jie egzistuoja. Todėl prieš tikrinant minėtą hipotezę yra būtina atlikti bandomuosius tyrimus, kurie padėtų įvertinti

galimybę, gauti gerus rezultatus, atliekant pagrindinį tyrimą. Tai ir bus pagrindinis šio baigiamojo magistro darbo tikslas.

Jeigu dviejų grupių žymekliai turi daugmaž vienodą ir reikšmingą informaciją apie tam tikro linksnio vartojimą tekste, tai tikėtina, kad vienos grupės žymeklių savybės, susijusios su tuo linksniu, bus pakankamai gerai prognozuojamos per atitinkamas kitos žymeklių grupės savybes. Norint patikrinti šią hipotezę reikia išspręsti kelis uždavinius:

1) nustatyti, ar galima nuspėti pasirinktos grupės pasirodymą bloke, remiantis likusių grupių pasiskirstymu;

2) patikrinti, ar galima prognozuoti pasirinktos grupės žymeklių skaičių bloke, remiantis likusių grupių pasiskirstymu;

3) patikrinti, ar galima įvertinti keliuose bloko sakiniuose pasirodys žymekliai iš pasirinktos grupės, remiantis kitomis grupėmis.

Sprendžiant šiuos uždavinius buvo įvykdytos tokios užduotys:

1) atlikta žymeklių paieška,

2) tekstai suskaidyti į blokus, parenkant tinkamą blokų dydį,

3) naudojant atsitiktinę imtį, duomenys suskaidyti į dvi dalis: modelio sudarymui ir prognozavimui,

4) sukurti reikiami kintamieji, atlikta jų pirminė statistinė analizė bei nustatyti tarpusavio ryšiai,

5) naudojant apmokymo duomenis tiriamam požymiui parinkti statistiniai modeliai,

6) sukonstruota prognozavimo taisyklė ir taikant ją testiniams duomenims įvertinta prognozavimo paklaida.

Visų uždavinių sprendimui buvo atrinkta 30 panašaus žanro lietuviškų kūrinių, suskaitmenintų vykdant ES struktūrinių fondų remiamą projektą „Pagrindinio ugdymo pirmojo koncentro (5–8 kl.) mokinių esminių kompetencijų ugdymas“, 2012 [17].

Uždavinių sprendimui darbe taikomi atskiri apibendrintų tiesinių modelių atvejai: logistinės ir Puasono regresijos modeliai, o parenkant tinkamą modelį ir informatyviausius aiškinamuosius kintamuosius vadovautasi Akaike informaciniu kriterijumi (AIC), deviacijos statistika, χ^2 statistika bei į modelį įtrauktų kintamųjų statistiniu reikšmingumu (Voldo testu).

Šiame magistro darbe atlikti tyrimai parodė, kad linksniuojamos kalbos dalys gali būti pakankamai tiksliai prognozuojamos, remiantis funkcinių (dažnų, vienareikšmiškai

apibrėžiamų) žodžių statistikomis. Aprašyti reprezentatyviausi modeliai leidžia prognozuoti trijų grupių $G1$, $V1$, $Įn1$ pasirodymus tekstuose. Atlikta žymeklių paieška, analizė ir jų pritaikymas modeliuose patvirtino, kad įvardžiai ir prielinksniai yra tinkami žymekliai tekstinių duomenų charakteristikų nagrinėjimui. Gauti darbo rezultatai gali būti naudojami tolimesniuose statistiniuose lietuvių kalbos tyrinėjimuose bei pritaikomi įvairių lingvistinių problemų sprendimui.

Dalis magistriniame darbe pateiktų rezultatų buvo pristatyti konferencijoje IVUS 2014 ir paskelbti straipsnyje "Dažnos žodžių formos ir linksnių vartojimas", kuris išspausdintas el. leidinyje "Informacinės technologijos" ISSN 2029–4832 (pateiktas priede A).

Baigiamojo magistro darbo 2 skyriuje yra pateikta mokslinės literatūros apžvalga, 3 skyriuje aprašyti darbe naudojami statistiniai ir duomenų analizės metodai, 4 skyriuje pateikta duomenų apdorojimo ir kintamųjų kūrimo metodika, pirminė statistinė kintamųjų analizė bei bandomųjų tyrimų rezultatai, 5 skyriuje - išvados ir rekomendacijos.

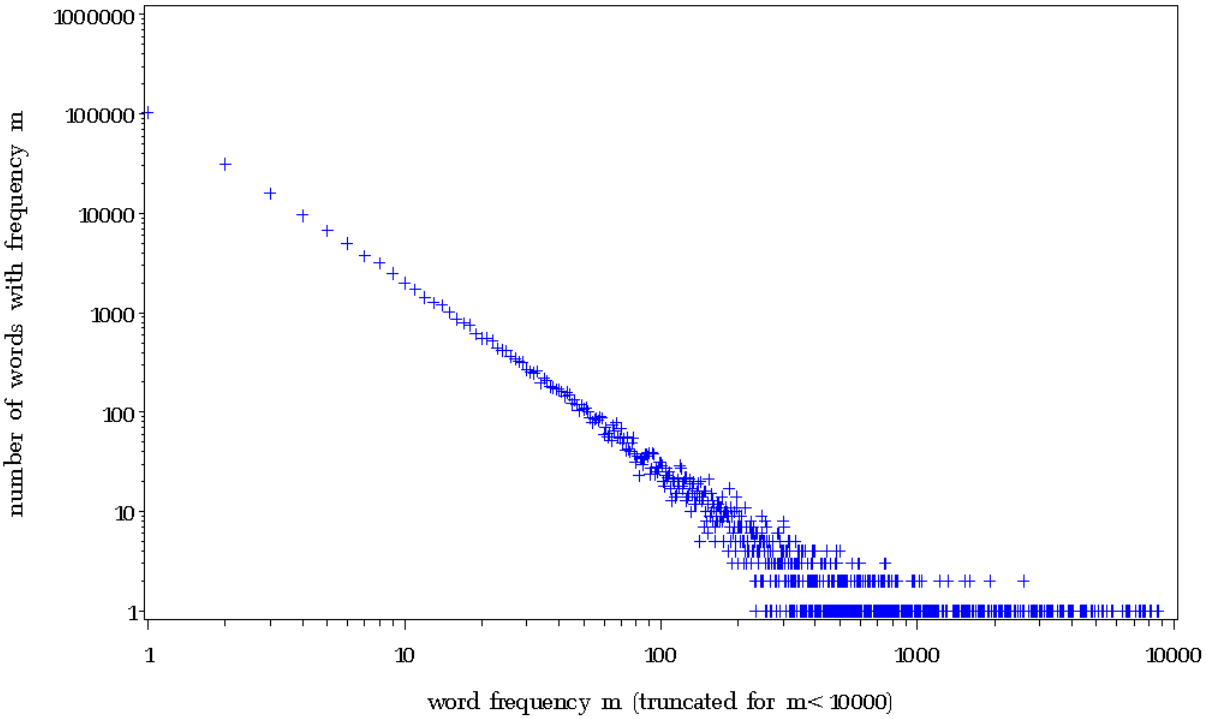
2. ANKSTESNIŲ MOKSLINIŲ TYRIMŲ APŽVALGA

Kalbos tyrinėjimų, paremtų tikimybių teorija, matematinės statistikos ir duomenų analizės teorija, sfera iki šiol neturi nusistovėjusio pavadinimo. Mokslinėje literatūroje galime aptikti įvairius jos pavadinimus: matematinė lingvistika, statistinė lingvistika, statistika lingvistikoje, kalbos statistika, lingvostatistika, tekstometriniai tyrimai, inžinerinė lingvistika, mašininė lingvistika. Pastarieji du dažniau naudojami kalbant apie automatinę teksto analizę ar garsinės kalbos automatinį apdorojimą.

Statistiniai metodai jau seniai įprasti eksperimentinėje fonetikoje matuojant, analizuojant, lyginant garsinio signalo akustinius požymius, vertinant paklaidas. A. Girdenio, Vidos Karosienės, Irenos Kruopienės, Astos Kazlauskienės ir Gailiaus Raškinio bei kt. darbuose statistiniai metodai taikomi fonotaktikos (garsų junginių, skiemens struktūros, jų vartosenos dėsningumų) tyrimams.

Kaip teigia Asta Kazlauskienė, Erika Rimkutė, Andrius Utka XX a. pabaigoje atsirado nemažai ir vadinamųjų tekstometrinių tyrimų, kuriais siekiama išsiaiškinti kalbos elementų vartojimo dėsningumus įvairiuose tekstuose. Gana dažnai tokie tyrimai atliekami su labai didelėmis duomenų sankaupomis, kurių niekaip rankomis neapdorosi. Tai iš dalies lėmė ir rimtų tekstometrinių tyrimų pasirodymą tik tada, kai atsirado galimybė pasinaudoti techninėmis priemonėmis (pirmosiomis skaičiavimo mašinomis ar moderniais kompiuteriais). Tekstometriniais tyrimais naudojasi ir kalbos ekspertai, norėdami nustatyti autoriaus stiliaus ypatybes ar net autorystę.

Šiame darbe iškeltų uždavinių sprendimui yra taikomi duomenų analizės ir statistikos metodai [1, 4, 18]. Vienas iš statistinių metodų taikymo lingvistikoje (kalbotyroje) ypatumų yra susijęs su Zipfo-Mandelbroto dėsniu [4, 9, 14, 27], kuriuo nusakomas žodžių formų, išrikiuotų jų pasitaikymo tekstyne dažnio mažėjimo (didėjimo) tvarka, pasiskirstymo dėsnis (skirstinys). Jis teigia, kad nedidelė žodžių formų dalis pasitaiko labai dažnai, o ženkli žodžių formų dalis, kartais virš 50%, tekstynuose yra pavartota tik kartą, o dar didesnė jų dalis yra iš viso nestebima. Šis dėsnis pastebimas ir magistro darbe naudojamame tekstų rinkinyje (žiūrėti 1 pav.). Formaliai iš Zipfo-Mandelbroto dėsnio išplaukia, kad potencialusis (teorinis) kalbos žodynas yra begalinis [4, 12]. Kompiuterių moksle (mašininio mokymosi taikymas apdorojant tekstus) tai kartais įvardijama kaip protrūkis, „sprogimas“ (angl. burstiness, [13]).



1 pav. Zipfo-Mandelbroto dėsnis

Nors šiame darbe yra sprendžiamas prognozavimo uždavinys, kuris yra (labiau) būdingas natūraliosios kalbos apdorojimo (NPA, angl. Natural Language Processing, NLP) metodologijai, jis turi sąsajų su lingvistikos mokslo problemomis. Jau seniai pastebėta, kad dažniausi kalbos žodžiai arba jų formos turi savybių, nebūdingų retesniems žodžiams, ir paprastai yra funkciniai žodžiai, t.y. atlieka tam tikras funkcijas [24]. Kai kuriuose anglų kalbos tekstų klasifikavimo darbuose (pvz., [20]), kurių uždavinys – nustatyti grožinės literatūros tekstų žanrą arba autorystę, tvirtinama, kad labai dažni žodžiai ir žodžių formos yra geri tekstų klasifikavimo požymiai.

Paprastai empiriniai-statistiniai tyrimai lingvistikoje remiasi tekstynų duomenimis. Didelės elektroninės tekstų (tiek sakytinės, tiek rašytinės kalbos) sankaupos vadinamos tekstynais. Tekstynai paprastai naudojami natūralios kalbos vartosenai tirti, leidžia daryti objektyvias išvadas. Jais remiantis tiriamas žodžių dažnumas, sudaromi dažniniai žodynai, analizuojama žodžių kontekstinė aplinka [10]. Dabartinę (rašytinę) lietuvių kalbą, įvairius jos stilius, anot kūrėjų, reprezentatyviai atspindi Vytauto Didžiojo universiteto Kompiuterines lingvistikos centro (toliau - KLC) „Dabartinės lietuvių kalbos tekstynas“, kurį sudaro daugiau kaip 140 mln. žodžių ir kuris yra plačiai Lietuvoje naudojama duomenų bazė [7]. Tačiau jis nėra pritaikytas nuodugnesnei statistinei analizei, kai imtį reikia sudaryti pagal tyrimui svarbius tekstų požymius, visų pirma tekstų autorius, taip pat kitas svarbias charakteristikas: tekstų sukūrimo datas, įvairius stilius, žanrus, ar pan.

Lietuvių kalbai yra sukurta automatinė morfologinės analizės programa Lemuoklis [26], žodžio formai pateikianti antraštinį pavidalą (lemą) ir gramatines pažymas. Ji skirta VDU KLC tekstyno moksliniams lingvistiniams tyrinėjimams automatizuoti. Bet net ir naudojant šį įrankį didesnio teksto morfologinis anotavimas ir linksnių nustatymas reikalauja daug darbo sąnaudų.

3. ANALITINĖ - METODINĖ DALIS

Šiame skyriuje bus išdėstyta baigiamajame magistro darbe naudojama metodika, susijusi su hipotezių tikrinimu, kintamųjų tarpusavio ryšių analize - koreliacine analize, apibendrintais tiesiniais modeliais bei modelių ir aiškinamųjų kintamųjų parinkimo - atmetimo kriterijais.

3.1. Kintamieji ir jų matavimo skalės

Parenkant statistinį modelį yra būtina atsižvelgti į naudojamų kintamųjų matavimo skales. Kintamuoju vadiname požymį, kurį matuojame pasirinktu matavimo instrumentu. Jie yra skirstomi į kiekybinius ir kokybinius. Kokybiniams kintamiesiems matuoti naudojama pavadinimų arba rangų skalė, kiekybiniams kintamiesiems – intervalų arba santykių skalė. Skaitinės operacijos modelyje, atliekamos su išmatuotais tiriamo objekto ar reiškinio parametrais turi derintis su operacijomis realybėje, turi būti prasmingos ir tikrovėje. Matavimo skalę charakterizuoja dvi (tarpusavyje susijusios) savybės:

a) prasmingos (leistinos) duotoje skalėje operacijos,

b) operacijos, kurios nekeičia turimos informacijos, t.y. transformacijos, atžvilgiu kurių matavimų duomenyse esanti informacija yra invariantiška.

Vardų skalėje matuojamas požymis aprašomas tik lygybės/nelygybės savybe. Vadinasi,

a) nominaliojoje skalėje prasminga tik lyginimo operacija: lygu/nelygu;

b) bet kuris abipus vienareikšmis atvaizdavimas nekeičia nominaliųjų duomenų (reikšminės) informacijos.

Ranginis požymis nusako tam tikrą tvarką tarp objektų, juos galima surikiuoti. Vadinasi,

a) prasmingas tik objektų palyginimas pagal duoto požymio išreikštumą: daugiau/lygu/mažiau; apibrėžtos santykio operacijos ">,<,=";

b) informacijos nekeičia duotojo požymio (išreikštumo) bet kokios griežtai monotoniškai (didėjančios) transformacijos.

Intervalinėje skalėje:

a) prasmingi visi palyginimai, sumos ir skirtumai bei daugyba ir dalyba iš teigiamo skaičiaus, bet neturi prasmės intervalinio požymio reikšmių dalyba ar daugyba;

b) nenusakyta koordinačių pradžia.

Santykinė skalė. Tai pati turtingiausia skalė, kurioje prasmingos visos aritmetinės operacijos ir palyginimai. Nuo intervalinės skalės ji skiriasi tuo, kad nusakyta reali prasminga (interpretuojama) koordinačių pradžia.

3.2. Hipotezių tikrinimas

Šiame poskyryje bus supažindinama su hipotezių savokomis [19] bei jų tikrinimo metodika [5].

Hipoteze vadinamas bet kuris nepatvirtintas teiginys. Bet koks tvirtinimas apie atsitiktinio dydžio (kelių atsitiktinių dydžių) pasiskirstymo formą vadinamas *statistine hipoteze*, o bet koks tvirtinimas apie jo (jų) pasiskirstymo parametrus vadinamas *parametrine hipoteze*. Pradinę hipotezę vadiname *nuline* hipoteze ir žymime H_0. Alternatyva vadiname hipotezę H_1, priešingą nulinei. Alternatyvos skirstomos į *dvipuses* $\Theta \neq \Theta_0$ ir *vienpuses* $\Theta < \Theta_0$ (arba $\Theta > \Theta_0$). Hipotezę, kuri visiškai nusakoma viena parametro reikšme, vadiname *paprastąja*. Priešingu atveju hipotezė vadinama *sudėtinga*.

Taisyklė, pagal kurią iš imties rezultatų darome išvadą apie hipotezės teisingumą ar klaidingumą, vadinama statistiniu kriterijumi. Kriterijai, sudaryti remiantis duomenų skirstinio parametriniu modeliu, vadinami *parametriniais*, priešingu atveju - *neparametriniais*.

Priimdami arba atmesdami hipotezę H_0, galime padaryti dviejų rūšių klaidas. Pirmosios rūšies klaida: H_0 atmetame, kai ji teisinga. Antros rūšies klaida: H_0 priimame, kai ji klaidinga. Dėl imties atsitiktinumo praktiškai neįmanoma sudaryti kriterijaus, kad abiejų klaidų tikimybės būtų lygios 0, todėl parenkamas kriterijaus reikšmingumo lygmuo $\alpha = P(H_0\ atmetame\ |H_0\ teisinga)$.

Tarkime, tikriname hipotezę su vienpuse alternatyva:

$$\begin{cases} H_0 : \Theta = \Theta_0, \\ H_1 : \Theta > \Theta_0. \end{cases}$$

Tegul sprendimui naudojama statistika $T(X_1,X_2,...,X_n)$ yra tolydi. Tarkime, t^* yra statistikos T realizacija. Tikimybė, kad kriterijaus statistika T (tuo atveju, kai H_0 teisinga) nemažesnė už stebimą realizaciją t^*, vadinama *p reikšme*. Tuomet:

$$p = P(T \geqslant t^*), \text{ kai } H_0 \text{ teisinga.}$$

Jeigu skaičius t^* patenka į kritinę sritį W (hipotezę H_0 atmetame), tai $p < \alpha$. Ir atvirkščiai, jeigu t^* nepatektų į W, tai $p \geqslant \alpha$.

Žinoma, jeigu alternatyva $\Theta < \Theta_0$ arba $\Theta \neq \Theta_0$, tai keičiasi ir p reikšmės apibrėžimas. Galima suformuluoti bendrą taisyklę, tinkančią visų rūšių nulinėms hipotezėms H_0 bei alternatyvoms H_1.

Tegul α yra reikšmingumo lygmuo, o p - p reikšmė.

Jeigu $p < \alpha$, tai hipotezė H_0 atmetama.
Jeigu $p \geqslant \alpha$, tai hipotezė H_0 neatmetama.

3.3. Apibendrintasis tiesinis modelis

Šiame darbe atliekant statistinę analizę taikomi atskiri apibendrintojo tiesinio modelio (toliau - ATM) atvejai. Išsamų ATM aprašymą galima rasti McCullagho ir Nelderio (1989) monografijoje [15].

Prieš pateikiant ATM apibrėžimą, reikia susipažinti su eksponentinės skirstinių šeimos apibrėžimu. Taip pat šiame poskyryje bus pateikiama atskiri ATM atvejai: logistinės, Puasono ir neigiamos binominės regresijos modeliai.

3.3.1. Eksponentinė skirstinių šeima

Tarkime, turime atsitiktinį kintamąjį y, kurio tikimybinis skirstinys priklauso nuo vieno parametro μ. Skirstinys priklauso eksponentinei skirstinių šeimai, jei jį galima užrašyti pavidalu:

$$f(y, \mu) = s(y)t(\mu)e^{a(y)b(\mu)}; \tag{1}$$

čia a, b, s, t - žinomos funkcijos. Parametras $b(\mu)$ vadinamas natūraliuoju parametru.

Lygtį (1) galima perrašyti taip:

$$f(y, \mu) = e^{a(y)b(\mu)+c(\mu)+d(y)}; \tag{2}$$

čia $s(y) = e^{d(y)}$, $t(\mu) = e^{c(\mu)}$. Iš (2) lygties galima pastebėti simetriškumą tarp y ir μ.

Eksponentinei skirstinių šeimai priklauso daug gerai žinomų ir plačiai naudojamų skirstinių, pvz.: normalusis, Puasono, binominis ir t.t.

3.3.2. Apibendrinto tiesinio modelio apibrėžimas

Tarkime, turime tyrimo (priklausomojo) kintamojo Y stebinius (y_1, y_2,..., y_n) ir aiškinamųjų kintamųjų $X = (x_1, .., x_k)$, $x_i = (x_{1i},...,x_{ni})^T$ stebinius

$$\begin{pmatrix} x_{11} & ... & x_{1k} \\ ... & ... & ... \\ x_{n1} & ... & x_{nk} \end{pmatrix}.$$

Laikoma, kad turimuose duomenyse stebiniai y_i, $i = 1,...,n$, yra tarpusavyje sąlygiškai nepriklausomi atsitiktiniai dydžiai, kai yra žinomos aiškinamųjų kintamųjų reikšmės.

Apibendrintą tiesinį modelį apibrėžia 3 komponentės:

1) atsitiktinė komponentė $f(y, \mu)$, kuri nusakoma tyrimo kintamojo y skirstiniu iš eksponentinės skirstinių šeimos;

2) tiesinis prediktorius η, kuris aprašo bendrą aiškinamųjų kintamųjų $x_1, ..., x_k$ įtakos tyrimo kintamojo y skirstiniui kiekybinę išraišką kaip tiesinę jų kombinaciją

$$\eta = \beta_0 + \beta_1 x_1 + ... + \beta_k x_k \tag{3}$$

su nežinomais koeficientais β_i , kuriuos reikės įvertinti iš duomenų;

3) jungties funkcija g, kuri susieja y skirstinio vidurkį μ su tiesiniu prediktoriumi η ,

$$g(\mu) = \eta. \tag{4}$$

Atsitiktinė komponentė $f(y, \mu)$ dažniausiai nusakoma skirstiniu iš eksponentinės skirstinių šeimos su parametru $\mu = E(y)$. Jungties funkcija g yra bet kokia griežtai monotoninė diferencijuojama funkcija. Nežinomų modelio parametrų β_i įvertinimui paprastai taikomas didžiausio tikėtinumo metodas.

Kituose skyreliuose aprašysime atskirus ATM atvejus: logistinės, Puasono, neigiamos binominės regresijų modelius.

3.3.3. Logistinės regresijos modelis

Dvireikšmės logistinės regresijos modelis – toks modelis, kai vienam (priklausomam) dvireikšmiui kintamajam Y daro įtaką vienas ar keletas (nepriklausomų, aiškinamųjų) kintamųjų X.

Tarkime, turime

$$y_i = \begin{cases} 1, \ jei \ rezultatas \ teigiamas, \\ 0, \ jei \ rezultatas \ neigiamas \end{cases}$$

ir aiškinamuosius kintamuosius

$$X = (x_1, ..., x_k) = \begin{pmatrix} x_{11} & ... & x_{1k} \\ ... & ... & ... \\ x_{n1} & ... & x_{nk} \end{pmatrix},$$

čia $x_i = (x_{1i},...,x_{ni})^T$.

Tikimybės $P\{y_i = 1|X = x\}$ kitimą, kintant x, patogu aprašyti logistine funkcija:

$$\frac{e^{\beta x}}{1 + e^{\beta x}}; \tag{5}$$

čia $\beta = (\beta_0, \beta_1, ..., \beta_k)^T$ modelio parametrai.

Logistinės regresijos modelis yra:

$$p_i = \frac{e^{\eta_i}}{1 + e^{\eta_i}}, \tag{6}$$

$$\eta_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_k x_{ik}. \quad (7)$$

Logistinės regresijos modelį galima užrašyti ir taip:

$$\eta_i = ln\frac{p_i}{1-p_i}. \quad (8)$$

Ši funkcija yra logistinės regresijos jungties funkcija. Santykis $\frac{p_i}{1-p_i}$ vadinamas galimybe (šansu).

Parametrų β_0, β_1,...,β_k įverčius $\widehat{\beta}_0$, $\widehat{\beta}_1$,...,$\widehat{\beta}_k$ reikia parinkti taip, kad (6) modelis su turimais duomenimis būtų kuo geriau suderintas. Šiuo atveju taikomas didžiausiojo tikėtinumo metodas. Parametrų įverčiai parenkami taip, kad tikėtinumo funkcija

$$L = \prod_{i:y_i=1} \widehat{p}_i \prod_{i:y_i=0} (1-\widehat{p}_i); \quad (9)$$

būtų maksimali. Čia:

$$\widehat{p}_i = \frac{e^{\widehat{\eta}_i}}{1+e^{\widehat{\eta}_i}}, \quad (10)$$

$$\widehat{\eta}_i = \widehat{\beta}_0 + \widehat{\beta}_1 x_{i1} + \widehat{\beta}_2 x_{i2} + ... + \widehat{\beta}_k x_{ik}; \quad (11)$$

Parametrų įverčiai skaičiuojami taip: pasirenkamos pradinės jų reikšmės ir keičiamos tol, kol stabilizuojasi, t.y. atliekamas iteracinis procesas. Gauti koeficientai naudojami apskaičiuoti $\widehat{\eta}_i$. Taip pat galima apskaičiuoti ir tikimybės įvertį apibrėžtą formule (10). Žinodami šį įvertį, galime prognozuoti y_i reikšmę. Jeigu $\widehat{p}_i > c$, kur c pasirinkta kritinė reikšmė, tai prognozuojama y_i reikšmė yra 1. Jeigu $\widehat{p}_i < c$, tai prognozuojama y_i reikšmė yra 0.

3.3.4. Puasono regresijos modelis

Tarkime, kad priklausomojo kintamojo y_i skirstinys yra Puasono ($y_i \sim Poisson(\mu_i)$, $f(y_i|\mu_i) = \frac{{\mu_i}^{y_i}}{y_i!}e^{-\mu_i}$). Puasono regresijos modelis yra tiesinis modelis sudarinėjamas priklausomojo kintamojo y_i vidurkiui.

Modelio lygtis:

$$\mu_i = e^{\eta_i}, \quad (12)$$

$$\eta_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_k x_{ik}; \quad (13)$$

Puasono regresijos jungties funkcija yra:

$$\ln\mu_i = \eta_i, \quad (14)$$

$$\eta_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_k x_{ik}. \quad (15)$$

Puasono skirstinys unikalus tuo, kad jo vidurkis sutampa su dispersija. Pagrindinė problema - priklausomojo kintamojo reikšmės dažnai būna labiau "išsibarsčiusios" nei tikimasi pagal modelį, todėl dispersija išauga. Tai vadinama dispersijos padidėjimu, o pats modelis

vadinamas Puasono regresijos modelis su padidėjusia dispersija [25]. Tokiu atveju į Puasono skirstinį įvedamas papildomas dispersijos padidėjimo parametras, kuris įvertinmas iš duomenų. Kita alternatyva šiuo atveju gali būti Puasono skirstinys su pertekliniu nulių kiekiu (zero-inflated Poisson distribution), nusakomas parametrais μ_i ir π_i, $0 < \pi_i < 1$:

$$P\{y_i = 0\} = \pi_i e^{-\mu_i} + 1 - \pi_i, \tag{16}$$

$$P\{y_i = k\} = \pi_i \mu_i^k \frac{e^{-\mu_i}}{k!}. \tag{17}$$

Nežinomi šio skirstinio parametrai μ_i ir π_i vertinami didžiausio tikėtinumo metodu. Dar viena alternatyva yra neigiamas binominis skirstinys, aprašytas kitame skyrelyje.

3.3.5. Neigiamos binominės regresijos modelis

Kai duomenims netinka Puasono regresijos modelis dėl didesnės dispersijos negu ji turėtų būti pagal Puasono dėsnį, dar vienas iš būdų tai išspręsti yra neigiamos binominės regresijos modelis, kuris taip pat sudaromas kintamojo vidurkiui.

Tarkime, kad kintamasis y pasiskirstęs pagal Puasono dėsnį, bet μ yra atsitiktinis dydis pasiskirstęs pagal Gama skirstinį, t.y.

$$y|\mu \sim Poisson(\mu), \tag{18}$$

$$\mu \sim Gamma(\alpha, \beta); \tag{19}$$

kur $Gamma(\alpha, \beta)$ yra Gama skirstinys su vidurkiu $\alpha\beta$ ir dispersija $\alpha\beta^2$, kurio tankio funkcija:

$$f(\mu) = \frac{1}{\beta^\alpha \Gamma(\alpha)} \mu^{\alpha-1} e^{-\alpha/\beta}, \tag{20}$$

kai $\mu > 0$ ir $f(\mu) = 0$ kitu atveju. Tuomet kintamojo y besąlyginis skirstinys yra neigiamas binominis,

$$f(y) = \frac{\Gamma(\alpha + y)}{\Gamma(\alpha) y!} \left(\frac{\beta}{1+\beta}\right)^y \left(\frac{1}{1+\beta}\right)^\alpha. \tag{21}$$

Šio skirstinio vidurkis $E(y) = \alpha\beta$, o dispersija $D(y) = \alpha\beta + \alpha\beta^2$. Taigi, sudarant modelį, parametrai imami: $\mu = \alpha\beta$ ir $k = \frac{1}{\alpha}$, tuomet $E(y) = \mu$, o $D(y) = \mu + k\mu^2$. Tada

$$y \sim Negbin(\mu, k),\ f(y) = \frac{\Gamma(k^{-1} + y)}{\Gamma(k^{-1}) y!} \left(\frac{k\mu}{1+k\mu}\right)^y \left(\frac{1}{1+k\mu}\right)^{1/k}. \tag{22}$$

Indeksas k^{-1} vadinamas dispersijos parametru. Kai $k^{-1} \to 0$, tai $D(Y) \to \mu$ ir neigiamas binominis skirstinys konverguoja į Puasono skirstinį. Paprastai k^{-1} yra nežinomas, o jo įvertinimas padeda nustatyti dispersijos padidėjimo koeficientą.

Modelio lygtis atsitiktiniam kintamajam y_i:

$$\mu_i = e^{\eta_i}, \tag{23}$$

$$\eta_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_k x_{ik}. \tag{24}$$

3.4. Modelio tinkamumo nustatymo ir aiškinamųjų kintamųjų parinkimo kriterijai

Tinkamas modelis tenkina šias savybes: jame atspindėti visi tiriami sąryšiai, kurie yra reikšmingi arba statistiškai arba remiantis nagrinėjamos srities dėsniais ar teoroniais modeliais, kuo tiksliau prognozuoti naujų duomenų reikšmes, ir maksimaliai paprastas išlaikant anksčiau paminėtas savybes. Šiame darbe parenkant tinkamą modelį ir informatyviausius aiškinamuosius kintamuosius vadovautasi deviacijos statistika ir Akaike informaciniu kriterijumi (AIC), bei į modelį įtrauktų kintamųjų statistiniu reikšmingumu. Prognozių tikslumo vertinimui buvo taikomas kryžminio patikrinimo metodas, χ^2 statistika, logistinės regresijos atvejui buvo naudojama ROC kreivių analizė.

Deviacijos (deviation) statistika. Prisotintasis (ATM) modelis yra toks modelis, į kurį yra įtraukti visi potencialiai informatyvūs kintamieji ar veiksniai, aprašomi tų kintamųjų sąveikomis. Daroma prielaida, kad prisotintojo modelio parametrus galima "tinkamai" įvertinti iš turimų duomenų. Kitus modelius kartais vadina *apribotaisiais*, kadangi jie gaunami iš prisotintojo modelio uždedant apribojimus prisotintojo modelio parametrams, o dažnai ir tiesiog dalį tų parametrų užnulinant.

Tarkime, β_{max} prisotinto modelio parametrų vektorius, o b_{max} - šio vektoriaus didžiausiojo tikėtinumo įvertis. Prisotinto modelio tikėtinumo funkcijos didžiausia reikšmė $L(b_{max}; y)$ bus didesnė už už bet kokios kitos tikėtinumo funkcijos reikšmes, nes ji pateikia išsamiausią duomenų aprašymą. Pažymėkime $L(b,y)$ tikrinamo modelio tikėtinumo funkcijos maksimalią reikšmę. Tuomet tikėtinumo santykis apskaičiuojamas pagal formulę [8]:

$$\lambda = \frac{L(b_{max}; y)}{L(b; y)}. \tag{25}$$

Praktikoje naudojamas tikėtinumo santykio logaritmas, kuris lygus logaritmuotų tikėtinumo funkcijų skirtumui:

$$ln\lambda = l(b_{max}; y) - l(b; y); \tag{26}$$

čia $l(b_{max}; y) = lnL(b_{max}; y)$, $l(b; y) = lnL(b; y)$. Nelderis ir Wedderburnas (1972) deviacija pavadino dydį:

$$Dev = 2[l(b_{max}; y) - l(b; y)]. \tag{27}$$

Deviacija, kai teisingas nulinėje hipotezėje suformuluotas modelis, turi χ^2 skirstinį. Šio teiginio įrodymą pateikė Dobsonas [8].

Taigi deviacija parodo, kiek tiriamasis modelis skiriasi nuo prisotinto modelio. Gerai duomenis aprašančio apibendrintojo tiesinio modelio deviacijos ir *liekamųjų* (vadinamų paklaidų) laisvės laipsnių santykis turi būti artimas vienetui. Jeigu dydis r yra aiškinamųjų kintamųjų X matricos rangas, o iš viso turime n stebinių, tai paklaidų laisvės laipsniai bus $(n - r)$.

Akaike informacinis kriterijus (toliau - AIC). AIC apibrėžiamas taip [2]:

$$AIC(M) = -2(l(b_M; y)) - k_M); \tag{28}$$

čia k_M - modelio M parametrų skaičius, $l(b_M; y)$ - modelio M tikėtinumo funkcijos didžiausios reikšmės (duomenims y) natūrinis logaritmas. Geriausiai duomenis y aprašantis modelis M iš duoto alternatyvių modelių rinkinio yra tas, kurio AIC kriterijaus reikšmė AIC(M) yra mažiausia.

Voldo kriterijai. Sudarant modelį labai svarbu, kad jis būtų kuo paprastesnis, t.y. kad jame nebūtų statistiškai nereikšmingų kintamųjų, kurie neturi didelės įtakos modelio tikslumui. Nesvarbių kintamųjų paieškai naudojami Voldo testai [6]. Jie padeda nuspręsti, ar aiškinamasis kintamasis šalintinas iš modelio. Tikrinama hipotezė:

$$\begin{cases} H_0 : \beta_i = 0, \\ H_1 : \beta_i \neq 0. \end{cases} \tag{29}$$

Valdo kriterijus atsako į klausimą ar konkretus koeficientas $\beta_j \neq 0$. Žinant Voldo kriterijaus p reikšmę hipotezė H_0 atmetama, jeigu $p < \alpha$, kur α pasirinktas reikšmingumo lygmuo. Jeigu $p \geqslant \alpha$, hipotezė H_0 neatmetama.

Kai j-tasis kintamasis x_j įgyja skaitines reikšmes, hipotezei (29) tikrinti naudojamas Valdo kriterijus su statistika $W = \widehat{\beta}_j / se(\widehat{\beta}_j)$. Čia $se(\widehat{\beta}_j)$ yra $\widehat{\beta}_j$ standartinio nuokrypio įvertis, paprastai apskaičiuojamas per kvadratinę šaknį iš Fisherio informacinės matricos, apskaičiuotos didžiausiojo tikėtinumo įverčio taške. Jeigu p reikšmė $< \alpha$, kur α pasirinktas reikšmingumo lygmuo (dažniausiai 0,05), tai kintamasis yra statistiškai reikšmingas ir jį modelyje paliekame. Jeigu $p \geqslant \alpha$, tai kintamasis yra statistiškai nereikšmingas ir modelyje jis paliekamas tik ypatingais atvejais. Tokiu atveju reikėtų sudaryti logistinės regresijos modelį be j-ojo kintamojo ir pažiūrėti, kaip pasikeičia koeficientų reikšmės ir klasifikacinė lentelė.

ROC kreivių analizė. Tarkime, atliekamas testas, kurio rezultatas gali būti teigiamas arba neigiamas. Pažymėkime: TP - nustatytas teisingas teigiamas rezultatas, NP - neteisingai teigiamas rezultatas, NN - neteisingai neigiamas rezultatas ir TN - teisingai neigiamas rezultatas.

Jautrumas - tai tikimybė, kad turinčio požymį objekto testo rezultatas yra teigiamas. Jis apskaičiuojamas $TP/(TP + NN)$.

Specifiškumas - tai tikimybė, kad neturinčio požymio objekto testo rezultatas yra neigiamas. Jis apskaičiuojamas $TN/(NP + TN)$.

ROC kreivė - grafikas, rodantis klasifikatoriaus jautrumo ir specifiškumo sąryšį [25], t.y. rodo teisingai nustatytų teigiamų rezultatų skaičiaus priklausomybę nuo neteisingai nustatytų neigiamų rezultatų skaičiaus. Ji dažniausiai naudojama rezultatų tikslumui įvertinti, sprendžiant klasifikavimo į dvi grupes uždavinius. ROC kreivė konstruojama taip:

1) kiekviena rodiklio reikšmė x_i naudojama kaip kritinis taškas ir procentais skaičiuojamas jo jautrumas j_i ir specifiškumas s_i;

2) koordinačių plokštumoje atidedamas taškas kiekvienam i su koordinatėmis X ašyje - $(1 - s_i)$, Y ašyje - j_i;

3) gretimi taškai sujungiami, o gauta tiesė vadinama ROC kreive.

Kuo plotas po ROC kreive didesnis, tuo geresnis modelis. Galima lyginti ir kelis modelius tarpusavyje. Tuomet lyginamas plotas po ROC kreive - kuo plotas po ROC kreive artimesnis vienetui, tuo modelis geriau prognozuoja duomenis. Teoriškai jis gali įgyti bet kokias reikšmes iš intervalo [0, 1], tačiau prasmingu modeliu laikomas tas, kurio kreivė yra aukščiau tiesės y=x. Idealaus modelio atveju plotas po ROC kreive lygus 1, trivialaus modelio - 0,5.

3.5. Tyrimo atlikimo metodika

Šiame poskyryje aprašoma darbe naudojama tyrimo atlikimo metodika. Ši metodika taikoma visiems bandomiesiems tyrimams atlikti, o jos etapai yra:

1) Tiriamai savybei (požymiui) modeliuoti sudaromas žymeklių rinkinys.

2) Visi duomenys suskaidomi į optimalaus dydžio blokus. Blokų dydis parenkamas derinant priešingus siekius: maksimaliai atsižvelgti į galimą (tikėtiną) duomenų nehomogeniškumą, bet tuo pačiu nagrinėti pakankamai didelius duomenų blokus, kad išryškėtų statistiniai dėsningumai.

3) Tekstų blokai atsitiktinai su vienodomis tikimybėmis priskiriami vienai iš 2-jų grupių. Pirmoji yra naudojama statistinio modelio parinkimui (apmokymui), o antroji yra skirta prognozavimo tikslumui įvertinti (testavimui).

4) Naudojantis kiekvieno bloko žymeklių statistika sukuriami aiškinantieji kintamieji statistiniam modeliui sudaryti.

5) Naudojant apmokymo duomenis tiriamam požymiui parenkamas statistinis modelis, taikant logistinę, Puasono ar neigiamos binominės regresijos modelius, sukonstruojama prognozavimo taisyklė ir taikant ją testiniams duomenims įvertinama prognozavimo paklaida. Parenkant tinkamą modelį ir informatyviausius aiškinančiuosius kintamuosius vadovaujamasi deviacijos statistika ir

Akaike informaciniu kriterijumi (AIC), bei į modelį įtrauktų kintamųjų statistiniu reikšmingumu (p reikšmė $< 0{,}01$).

Pirmieji du tyrimo etapai yra specifiniai, 3) - 5) etapai yra standartiniai, dažnai naudojami panašaus pobūdžio tyrimuose.

3.6. Pirminių tekstinių duomenų aprašymas

Šiame poskyryje bus aprašytas pasirinktas duomenų rinkinys, pateikti tyrime naudojamų kalbos savybių apibrėžimai ir aprašymai.

Baigiamajam magistro darbui buvo pasirinktas duomenų rinkinys, kurį sudaro 80 kūrinių, suskaitmenintų vykdant ES struktūrinių fondų remiamą projektą „Pagrindinio ugdymo pirmojo koncentro (5–8 kl.) mokinių esminių kompetencijų ugdymas“ [17]. Skaitmeninėje bibliotekoje pateikiama 80 literatūros kūrinių 5–8 klasių mokiniams. Kūriniai suskaitmeninti TXT, PDF, EPUB ir MP3 formatais. Nacionalinį ES struktūrinių fondų remiamą projektą „Pagrindinio ugdymo pirmojo koncentro (5–8 kl.) mokinių esminių kompetencijų ugdymas“ 2009–2012 m. įgyvendino Ugdymo plėtotės centras. Projekto veiklose dalyvavo dešimties konkurso būdu atrinktų Lietuvos mokyklų pedagogai (po 6 iš kiekvienos mokyklos).

Bandomiesiems tyrimams buvo paimta 30 panašaus žanro lietuviškų kūrinių iš minėto tekstų rinkinio. Kiekvienas tekstas pateiktas atskirame tekstiniame faile, kuriame nurodyta autorius, pavadinimas, ISBN kodas bei surinktas visas kūrinys. Šis duomenų rinkinys ypač tinkamas mano tyrimui, kadangi kalbos vartojimo ypatumai formuojami nuo mažens šeimoje, buityje (šiuos duomenis surinkti sudėtinga) ir mokykloje. Todėl galima manyti, jog tam tikros savybės randamos mokyklose skaitomuose, nagrinėjamuose tekstuose egzistuoja ir kitoje aplinkoje naudojamoje lietuvių kalboje.

Šiame darbe yra nagrinėjama linksniuojamų kalbos dalių gramatinė kategorija - linksnis. Savo turiniu ir reikšme ši kategorija yra sintaksinė kategorija, nagrinėjanti sakinių ir žodžių junginių sandarą. Lietuvių kalboje yra septyni linksniai: vardininkas, kilmininkas, naudininkas, galininkas, įnagininkas, vietininkas ir šauksmininkas. Pastarasis linksnis tyrime nenagrinėjamas, nes jis neturi reikšmių, juo įvardijame tik kreipimąsi.

Vardininkas - linksnis, kuris atsako į klausimą „kas?“ [11]. Nors vardininkas dažniausias linksnis lietuvių kalboje, reikšmių jis neturi daug, dažniausiai reiškia sakinio veikėją, arba subjektą - tai, apie ką sakinyje kalbama, neretai būvį arba koks tai daiktas (išreiškiamas būdvardiškaisiais žodžiais).

Kilmininkas - linksnis, kuriuo atsakoma į klausimą „ko?“, „kieno?“ [11]. Šis linksnis yra labai dažnas, turi daug reikšmių bei aiškų ryšį su vardininku ir galininku. Kilmininkas gali rodyti į ką krypsta veiksmas (su veiksmažodžiais), patikslinti tam tikrą kiekybę reiškiančius žodžius, atskleisti jų konkretų turinį. Nemažai kilmininkų yra pažymimieji (apibūdina

daiktą), požymio arba turi priklausymo reikšmę. Taip pat kilmininkai gali nusakyti rūšį, gentį, veislę, tipą bei turėti daug kitų reikšmių.

Naudininkas - linksnis, kuriuo atsakoma į klausimą „kam?“ [11]. Naudininko reikšmėms būdinga bendra paskirties ir naudos reikšmė.

Galininkas - linksnis, kuriuo atsakoma į klausimą „ką?“ [11]. Šio linksnio reikalauja daug veiksmažodžių, tokiu atveju galininkas reiškia objektą, į kurį krypsta veiksmas. Kartais galininkas reiškia laiką, parodo kada, kuriuo metu, kiek ilgai, kaip dažnai vyksta veiksmas arba objekto ar būsenos kiekį.

Įnagininkas - linksnis, kuriuo atsakoma į klausimą „kuo?“ [11]. Įnagininkas gali reikšti objektą, būvį, požymį, įvairias aplinkybes.

Vietininkas - linksnis, kuris atsako į klausimą „kame, kur?“ [11]. Jis dažniausiai eina aplinkybėmis, ypač vietos, rečiau – laiko, o vienu kitu išskirtiniu atveju reiškia veiksmo atlikimo būdą ar būvį.

Lietuvių kalboje egzistuoja daiktavardžio, būdvardžio, skaitvardžio ir įvardžio linksnio kategorijos.

Magistro darbo tyrimui buvo pasirinktos dvi kalbos dalys: įvardis ir prielinksnis. *Įvardis* - tai yra kalbos dalis, kuri nurodo daiktą, ypatybę arba skaičių, bet jų nepavadina [22]. Įvardis gali pakeisti daiktavardžius, būdvardžius, ar net nurodyti kiekybę. Taigi tikrai nemaža dalis žodžių gali būti pakeista įvardžiais, kurie dažnai pasirodo lietuvių kalbos tekstuose ir dauguma jų linksnių yra vienareikšmiškai nustatomi. Todėl tikėtina, kad įvardis yra viena iš kalbos dalių, gana gerai atspindinti žodžių ir linksnių padėtį sakiniuose.

Dar viena kalbos dalis, padedanti nustatyti žodžio linksnį, yra *prielinksniai* - nekaitoma kalbos dalis, kuri eina su linksniu ir parodo linksniuojamojo žodžio ryšį su kitais žodžiais [23]. Yra nemažai prielinksnių, kurie eina tik kartu su tam tikru linksniu, pavyzdžiui, po prielinksnio „su“ visada eis žodis, esantis įnagininko linksnio.

Šiame skyriuje išdėstyta teorija buvo naudojama baigiamojo magistro darbo tyrimuose, kurių eiga ir gauti rezultatai aprašomi kitame skyriuje.

4. EKSPERIMENTINĖ - TIRIAMOJI DALIS

Šiame skyriuje bus pateiktas trumpas kintamųjų kūrimo ir apdorojimo aprašymas, atlikta pirminė statistinė analizė, koreliacinė analizė, nagrinėjami kintamųjų tarpusavio ryšiai, bus pateikiami apibendrintų tiesinių modelių taikymo rezultatai, taip pat kintamųjų prognozavimo rezultatai bei tikslumo įvertinimas, padarytos išvados.

Bandomųjų tyrimų tikslas yra nustatyti ar naudojami žymekliai tinkami linksniuojamų kalbos dalių prognozavimui bei patikrinti hipotezę: jeigu dviejų grupių žymekliai turi daugmaž vienodą ir reikšmingą informaciją apie tam tikro linksnio vartojimą tekste, tai tikėtina, kad vienos grupės žymeklių savybės, susijusios su tuo linksniu, bus pakankamai gerai prognozuojamos per atitinkamas kitos žymeklių grupės savybes. Tyrimas atliekamas remiantis 3.5. skyrelyje aprašytos tyrimo metodikos 5-6 punktais.

Atliekant tyrimą prognozuojamas kintamasis y buvo apskaičiuojamas kaip vienos grupės, pasirinktos iš 12 grupių (žr. 4.1. poskyrį), žymeklių dažnis bloke, jų pasirodymo bloke indikatorius arba sakinių skaičius, kuriame pasirodė žymeklis, o likusių grupių analogiška statistika blokuose buvo (pradiniai) aiškinantieji kintamieji. Kiekvienas blokas yra atskiras apmokymo imties elementas, j yra jo numeris.

Aptarsime tik trijų grupių, $V1$, $G1$ ir $Įn1$, rezultatus. Jie gana gerai reprezentuoja kitus atvejus. Vardininkas dažniausias linksnis, net 87% blokų turėjo žymeklių iš $V1$ grupės. Galininkas yra gana dažnai vartojamas linksnis. Įnagininkas vartojamas gana retai, virš 81 % blokų neturėjo grupės $Įn1$ žymeklių. Remiantis AIC kriterijumi minėtiems kintamiesiems buvo parinkti modeliai su pradinių aiškinančiųjų kintamųjų sąveikomis (2-os eilės) ir be jų, taip pat lakoniškieji šių modelių variantai, parinkti naudojant reikšmingumo lygmenį 0,01 (modelyje palikti tik nariai, kurių p reikšmė $< 0{,}01$).

Modelius pateiksime simbolinėje formoje, kuri naudojama R pakete. Pvz., užrašas

$$y \sim x_1 + x_2 : x_3, \tag{30}$$

kur x_i yra tolydieji arba binariniai kintamieji, reiškia, kad kintamajam y sudaryto modelio prediktorius yra

$$\eta = \beta_0 + \beta_1 x_1 + \beta_2 x_2 x_3, \tag{31}$$

o simbolinis užrašas $x_1 * x_2$ reškia tą patį kaip ir $x_1 + x_2 + x_1 : x_2$.

4.1. Tekstinių duomenų apdorojimas, žymeklių parinkimas ir kintamųjų kūrimas

Šiame poskyryje trumpai aprašysime optimalaus teksto blokų dydžio ir žymeklių parinkimo principus bei kintamųjų kūrimo metodiką. Kaip jau buvo minėta anksčiau, tyrimams naudojamas 30-ies tekstų rinkinys. Tekstai apdorojami ir analizuojami naudojant

statistinį paketą R [21]. Duomenų apdorojimo ir kintamųjų kūrimo programa pateikta D priede.

Tekstai buvo nagrinėjami skirtingais pjūviais: pagal žodžius, sakinius ir sakinių blokus, nes skirtingais tyrimo etapais reikėjo vis kitaip pateikiamų duomenų. Atliekant tinkamų žymeklių paiešką tekstas turėjo būti skaidomas pagal žodžius, nes šiame tyrimo etape nagrinėjami objektai yra žodžiai. Nustatant optimalų teksto dalies dydį, jis turėjo būti skaidomas pagal sakinius. Pasirinkus tinkamą sakinių blokų dydį, tekstai buvo suskaidomi pagal blokus. Tekstus išskaidžius šiais trimis būdais buvo galima vykdyti tolimesnes tyrimo užduotis.

Optimalaus teksto dalies dydžio parinkimas. Blokų dydis parinktas derinant priešingus siekius: maksimaliai atsižvelgti į galimą (tikėtiną) tekstų nehomogeniškumą, bet tuo pačiu nagrinėti pakankamai didelius tekstų blokus, kad išryškėtų statistiniai dėsningumai. Tekstas yra gana didelis elementas ir nagrinėti įvairias savybes tokiuose plačiuose rėžiuose yra netikslinga. Tuo tarpu vienas sakinys yra per mažas elementas ir tyrimui naudingų rezultatų nepavyks gauti. Dėl šios priežasties tekstai buvo suskaidyti į blokus po 10 sakinių. Iš viso 12167 blokai. Tokiu atveju buvo galima tikėtis, jog viename bloke bus šnekama apie vieną veiksmą ir tame bloke bus pakankamai tipinių žodžių, t. y. bent 10 % žodžių tame bloke bus iš pasirinktų žymeklių.

Apmokymo ir testavimo duomenys. Kadangi baigiamojo darbo tikslas ne tik sudaryti modelius, bet ir patikrinti jų prognozių tikslumą, duomenys buvo suskaidyti į dvi grupes: duomenys modelio sudarymui (6085 blokai) ir duomenys prognozavimui (6082 blokai), kurie buvo suskaidyti dar į 10 grupių. Tuomet prognozių tikslumas galėjo būti tikrinamas, lyginamas ir įvertinamas. Šis duomenų skaidymas buvo atliekamas naudojant funkciją "sample". Ši funkcija išrenka atsitiktinę imtį (šiuo atveju su pasikartojimais). Turint sudarytus reikiamus duomenų rinkinius, buvo kuriami tyrimo kintamieji.

Žymekliai. Priklausomais ir aiškinamaisiais kintamaisiais šiame tyrime laikoma žymeklių grupės. Žymekliai tai pasirinkti žodžiai, kurie yra vienareikšmiškai apibrėžiami ir gana dažni lietuvių kalboje. Dėl gana gero reprezentatyvumo buvo pasirinktos dvi kalbos dalys: įvardis ir prielinksnis. Pradžioje buvo atrinkti 87 įvardžiai:

aš, manęs, man, mane, manimi, manyje, tu, tavęs, tau, tave, tavimi, tavyje, mes, mūsų, mums, mus, mumis, mumyse, jūs, jūsų, jums, jus, jumis, jumyse, jis, jo, jam, jį, juo, jame, ji, jos, jai, ją, ja, joje, jie, jų, jiems, juos, jais, juose, joms, jas, jomis, jose, tas, to, tam, tą, tuo, tame, ta, tos, tai, toje, šis, šio, šiam, šį, šiuo, šiame, ši, šios, šiai, šią, šia, šioje, tie, tų, tiems, tuos, tais, tuose, toms, tomis, tose, šie, šių, šiems, šiuos, šiais, šiuose, šioms, šias, šiomis, šiose.

Tačiau vien jų buvo per maža, kad galima būtų atlikti tolesnį tyrimą. Vidutiniškai įvardžiai sudarė 8,62 % visų žodžių viename bloke. Todėl dar buvo atrinka 24 prielinksniai:

anapus, anot, antrapus, arčiau, dėl, iš, link, linkui, nuo, pasak, prie, pusiau, tarp, žemiau, apie, į, pagal, palei, pas, per, prieš, pro, su, ties.

Šie prielinksniai visada eina tik su tam tikrais linksniais. Taigi iš viso atrinkta 111 žymeklių, suskaičiuota, kiek jų yra kiekviename sakinių bloke. Vidutiniškai žymekliai sudarė 13,007 % visų žodžių viename bloke. Šis santykis buvo pakankamas tolesniam tyrimui pradėti.

Kintamųjų kūrimas. Kadangi baigiamajame magistro darbe buvo pasirinkta nagrinėti kalbos dalis ir jų linksnius, tai buvo sudaryta 12 grupių pagal kalbos dalį ir linksnį:

$G1$, $Įn1$, $K1$, $N1$, V_G1, $V\_Įn1$, V_K1, $V1$, $Vt1$, $G2$, $Įn2$, $K2$.

Raidės nusako, kurio linksnio yra žodžiai, o skaitmuo kalbos dalį, t.y. ar tai įvardžiai, ar prielinksniai (1 – įvardžiai, 2 – prielinksniai). Priede B pateikiama lentelė, kurioje parodoma, koks žodis kuriai grupei priklauso. Pažymėtina, kad tarp sudarytų grupių yra tokių, kurios turi dvi raides atskirtas "_" brūkšneliu. Tyrimo metu buvo susidurta su problema, jog kai kurie įvardžiai gali būti poros linksnių, pvz: įvardis „jos“ – šiuo atveju, nežinant konteksto, negalima pasakyti, ar šis žodis yra vardininko ar kilmininko linksnio. Tai vadinama morfologiniu daugiareikšmiškumu [16]. Kadangi negalima vienareikšmiškai nusakyti, kuriai grupei turėtų žodis priklausyti, buvo sukurtos jungtinės grupės, pvz.: V_K1 į kurią įeina anksčiau minėtas įvardis „jos“.

Pirmajam bandomajam tyrimui buvo sukurtas kintamasis - kiekvienos grupės indikatorius, lygus atitinkamai 1 ir 0, kai grupės žymekliai bloke pasitaikė ir kai nepasitaikė. Antrajam bandomajam tyrimui, naudojant surinktą statistiką apie žymeklius kiekviename bloke kiekvienai minėtai grupei, buvo sudarytas kintamasis, rodantis, kiek kartų bloke pasitaikė žymekliai iš pasirinktos grupės. Trečiajam bandomajam tyrimui buvo sukurti kintamieji, kurie nurodė keliuose bloko sakiniuose pasitaikė žodžiai iš pasirinktos grupės.

Taip pat buvo sukurtas kintamasis „*autoriai*“, nurodantis kūrinio autorių. Šis kintamasis leidžia nustatyti, ar statistiniai sąryšiai tarp tiriamųjų kintamųjų priklauso nuo autoriaus. Tokiu būdu galima atskirti modelio komponentes, priklausančias nuo autorių stiliaus, ir komponentes nuo jų nepriklausančias, vadinasi, galbūt atspindinčias pačiai lietuvių kalbai būdingas savybes. Kintamasis „*autoriai*“ naudojamas visuose tyrimuose.

Taigi toliau tyrimuose naudojama 12 kintamųjų, nurodančių žymeklių iš pasirinktos grupės skaičių kiekviename bloke, 12 dvinarių kintamųjų, 12 kintamųjų, nurodančių keliuose sakiniuose pasirodė žymekliai iš duotosios grupės, bei kintamasis „*autoriai*“. Ki-

tuose skyreliuose bus atlikta šių kintamųjų pirminė statistinė analizė, bei kuriami modeliai naudojant 3. skyriuje išdėstytą teoriją.

4.2. Pirminė statistinė kintamųjų analizė

Šiame poskyryje bus atlikta 4.1. poskyryje aprašytų kintamųjų pirminė statistinė analizė, tačiau dėl kintamųjų gausos aprašytos bus ne visos grupės. Pirmiausia apžvelgsime dvinarius kintamuosius, tuomet kintamuosius, nurodančius žymeklių skaičių sakinių bloke (toliau vadinami dažnių kintamaisiais), ir kintamuosius, nurodančius keliuose sakiniuose pasirodė žymekliai pasirinktame bloke (toliau vadinami trečio tipo kintamaisiais). Duomenų rinkinys buvo suskaidytas į dvi dalis: modelio sudarymo ir prognozavimo duomenys, dėl to pirminė analizė bus atliekama kiekvienam rinkiniui atskirai, kad būtų galima apžvelgti panašumus ir skirtumus.

4.2.1. Dvinarių kintamųjų statistinė analizė

Dvinarių kintamųjų pirminei statistinei analizei yra naudojamos dažnių lentelės, kurios suteikia visą reikiamą ir įmanomą informaciją apie šio tipo kintamuosius. 1 lentelėje pateikiama visų grupių dažniai modelio sudarymo duomenyse.

1 lentelė. Dvinarių kintamųjų dažnių lentelė (apmokymo duomenys)

	0	1
V1	724	5361
K1	2364	3721
N1	1481	4604
G1	2212	3873
Įn1	4948	1137
Vt1	5609	476
V_K1	4667	1418
V_G1	5236	849
V_Įn1	5806	279
K2	1494	4591
G2	1298	4787
Įn2	3493	2592

Dažniausiai tekstuose pasitaiko žymekliai iš $V1$ grupės, rečiausiai - iš $V\_Įn1$. Taip pat dažnas yra naudininko linksnis ir prielinksniai, kurie pasirodo didelėje dalyje blokų. Tokią pačią situaciją matoma ir tarp prognozavimo duomenų (2 lentelė).

2 lentelė. Dvinarių kintamųjų dažnių lentelė (prognozavimo duomenys)

	0	1
V1	816	5266
K1	2388	3694
N1	1441	4641
G1	2253	3829
Įn1	4928	1154
Vt1	5549	533
V_K1	4610	1472
V_G1	5221	861
V_Įn1	5826	256
K2	1542	4540
G2	1330	4752
Įn2	3547	2535

Žymeklius galima suskirstyti į tris kategorijas: dažni ($V1$, $G2$, $N1$, $K2$), vidutinio dažnumo ($G1$, $K1$, $Įn2$, V_K1) ir reti ($V\_Įn1$, $Vt1$, V_G1, $Įn1$). Šiame darbe bus nagrinėjama po vieną žymeklių grupę iš kiekvienos grupės.

4.2.2. Dažnio tipo kintamųjų analizė

Šiame skyrelyje bus atliekama dažnio tipo, kurie parodo, kiek kartų žymekliai iš duotosios grupės pasitaikė sakinių bloke, kintamųjų pirminė statistinė analizė, pateiktos kintamųjų dažnių lentelės, histogramos.

Modelio sudarymo duomenys. Į šią dalį pateko 6085 blokai, kuriuose iš viso pasirodė 76303 žymekliai. Bendros žymeklių sumos pagal grupes pateiktos 3 lentelėje.

3 lentelė. Bendras žymeklių skaičius pagal grupes modelio sudarymo duomenyse

G1	Įn1	K1	N1	V_G1	V_Įn1	V_K1	V1	Vt1	G2	Įn2	K2
7601	1386	7119	10871	986	308	1930	18392	540	12412	4043	10715

Lietuvių kalboje dažniausiai vartojamas linksnis yra vardininkas. Tai atsispindi lentelėje bei C priede pateiktose dažnių lentelėse. Dažnai vartojami ir naudininko, kilmininko bei galininko linksniai. Rečiausias yra vietininkas, neskaitant jungtinės grupės $V\_\acute{I}n1$, į kurią patenka tik vienas įvardis "ta". Tokio rezultato buvo galima tikėtis, nes lietuvių kalboje labai retai yra (taisyklingai) vartojami įvardžiai vietininko linksnyje.

Grupės $V1$ žymekliai sudaro 24,10% visų žymeklių pasirodymų. Prielinksnių grupės $G2$ ir $K2$ bendrai sudaro 30,31%. Taigi vien šios trys grupės sudaro apie 54,41% visų žymeklių pasirodymų, tačiau tai neleidžia daryti prielaidos, kad jos yra svarbesnės nei likusios grupės.

Prognozavimo duomenys. Šiems duomenims buvo priskirta 6082 sakinių blokai, kuriuose pasirodė 75957 žymekliai. Šie duomenys atsitiktinai buvo suskirstyti į 10 dalių, o žymeklių skaičius juose pasiskirstė taip:

4 lentelė. Žymeklių pasiskirstymas pagal grupes prognozavimo duomenų grupėse

	G1	Įn1	K1	N1	V_G1	V_Įn1	V_K1	V1	Vt1	G2	Įn2	K2
1	784	151	731	1138	108	26	229	1857	47	1213	382	1062
2	812	135	729	1117	101	26	234	1887	69	1236	408	1072
3	814	142	773	1117	101	29	195	1890	53	1287	403	1068
4	745	158	753	1071	95	26	198	1828	67	1212	414	1053
5	736	137	710	1074	85	31	207	1712	56	1240	395	1058
6	748	140	661	1031	94	35	199	1751	72	1158	353	1006
7	771	124	698	1121	111	23	203	1814	60	1295	388	1032
8	702	123	610	1074	94	30	194	1751	59	1232	433	1024
9	764	131	694	1066	103	25	194	1792	49	1175	400	1033
10	709	145	738	1132	111	27	204	1855	69	1333	394	1113

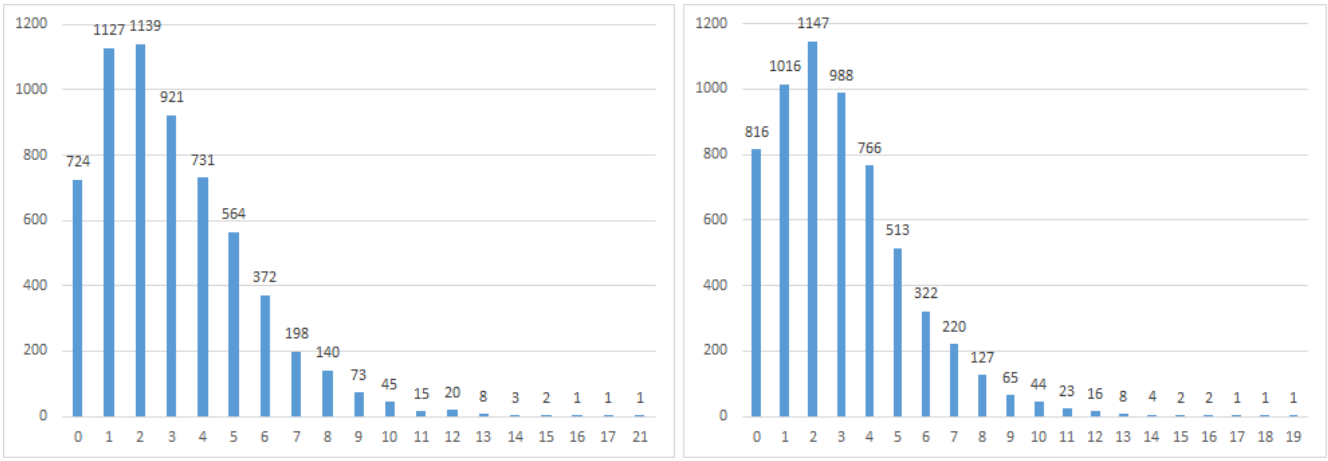
Ši lentelė, mums vėl parodo, kad dažniausias linksnis yra vardininkas. Pažymėtina, jog atsitiktinai skaidant prognozavimo duomenis į 10 dalių, žymeklių skaičius kiekvienoje

dalyje yra panašus.

Priklausomam kintamajam bus sudaroma prognozė, imant kiekvienos dalies duomenis, taigi iš viso turėsime 10 prognozių, kurias galėsime palyginti, išskirti tam tikras tendencijas, savybes, bei tiksliau įvertinti paklaidą.

Histogramos. 2 - 4 paveikslėliuose pateikta $V1$, $G1$ ir $Įn1$ grupių dažnio kintamųjų histogramos. Būtent šioms grupėms sudaryti modeliai ir bus aptarti tolesniuose poskyriuose.

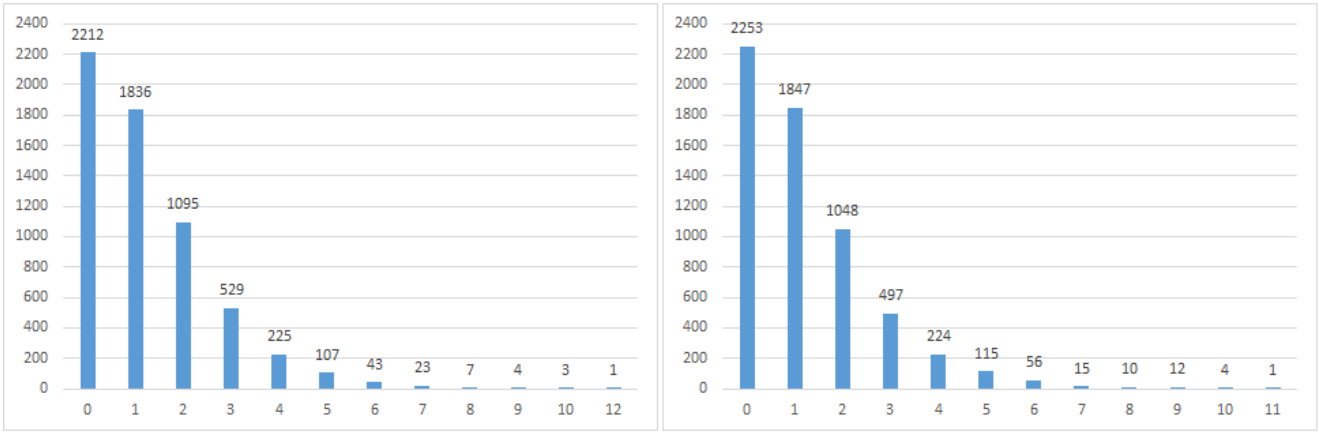
Pagal $V1$ histogramą 2, matome, kad įvardžiai esantys vardininko linksnyje yra panašiai pasiskirstę kaip Puasono dėsnis. Taip pat pastebėtina, jog pasiskirstymas abiejuose duomenų rinkiniuose yra gana panašus. Apmokymo duomenų vidurkis yra 2,6, o testavimo duomenų 2,64.



2 pav. V1 grupės dažnio kintamojo modelio sudarymo (kairė) ir testavimo (dešinė) duomenų histogramos

Be to, tai jog iš viso buvo tik 724 sudarymo duomenyse ir 816 testavimo duomenyse blokai, kuriuose nepasitaikė žymekliai iš šios grupės dar kartelį patvirtina, jog vardininkas yra dažnas linksnis.

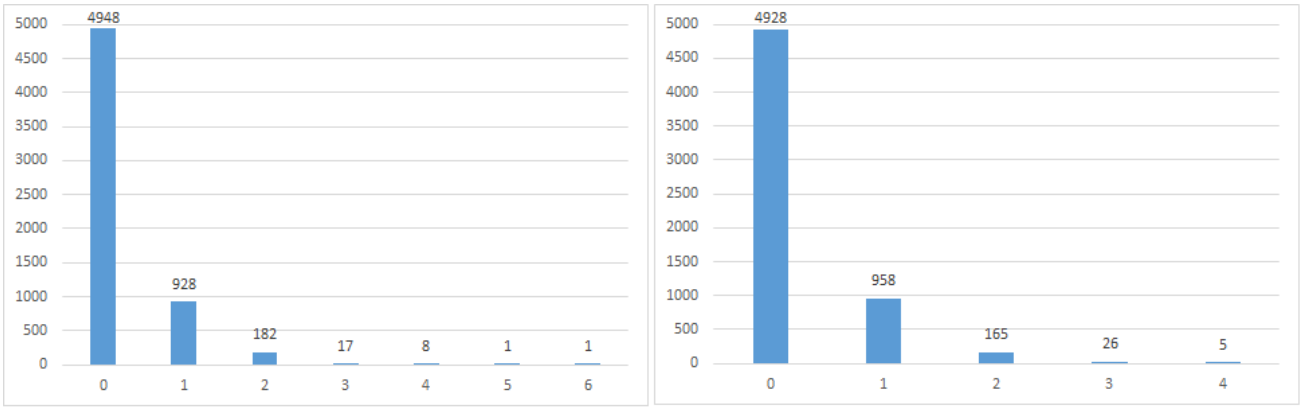
$G1$ grupės histogramos (3 pav.) primena neigiamo binominio skirstinio grafikus. Šios grupės žymekliai yra retesni nei prieš tai minėtosios. Vėl gi, abiejuose duomenų rinkiniuose stebimas panašus pasiskirstymas.



3 pav. G1 grupės dažnio kintamojo modelio sudarymo (kairė) ir testavimo (dešinė) duomenų histogramos

Apmokymo duomenų vidurkis $G1$ grupės lygus 1,14, o testavimo duomenų - 1,15. Matyti, jog tikimybė, kad žymeklis iš $G1$ grupės pasirodys sakinių bloke 7 ar daugiau kartų yra labai maža.

Pažvelgus į $Įn1$ histogramas (4 pav), akivaizdu, jog tai labai reta grupė.



4 pav. Įn1 grupės dažnio kintamojo modelio sudarymo (kairė) ir testavimo (dešinė) duomenų histogramos

Įvardžiai esantys įnagininko linksnyje pasitaiko po vieną kartą tik 928 (modelio sudarymo duomenyse) ir 958 (testavimo duomenyse) blokuose, o tikimybė, kad žymekliai iš šios grupės pasirodys du ar daugiau kartų yra maža.

4.2.3. Trečio tipo kintamųjų analizė

Šiame skyrelyje pateiktos lentelės parodo, keliuose sakiniuose pasirinktame bloke pasitaikė atrinktos grupės žymekliai.

5 lentelė. Sakinių, kuriuose pasirodė žymekliai, skaičius pagal grupes modelio sudarymo duomenyse

V1	K1	N1	G1	Įn1	Vt1	V_K1	V_G1	V_Įn1	K2	G2	Įn2
16066	6490	9606	6990	1338	533	1857	970	302	9323	10440	3741

Remiantis 5 ir 3 lentelėmis, galima daryti išvadas, kad tik retais atvejais pasitaiko, kad viename sakinyje būtų du žymekliai iš tos pačios grupės. Tai natūralu, nes negalima viename sakinyje naudoti daug įvardžių, o tuo labiau dar vienodo linksnio, nes tampa nebeaišku, kas ir ką veikia ar su kuo veikia ar pan. Panašiai yra ir su prielinksniais: dažnai yra išvardinami vienas po kito žodžiai esantys to pačio linksnio, tačiau tokiu atveju, prielinksnis dažniausiai pavartojamas tik vieną kartą, kad kalba nebūtų perkrauta žodžiais, kuriuos galime nutuokti. Pavyzdžiui: lietuvių kalboje sakoma: "Jis kalbėjo su Jonu, Petru ir Antanu", bet ne "Jis kalbėjo su Jonu, su Petru ir su Antanu". Taigi, šiame pavyzdyje yra tik vienas prielinksnis "su" ir net trys žodžiai esantys įnagininko linksnyje.

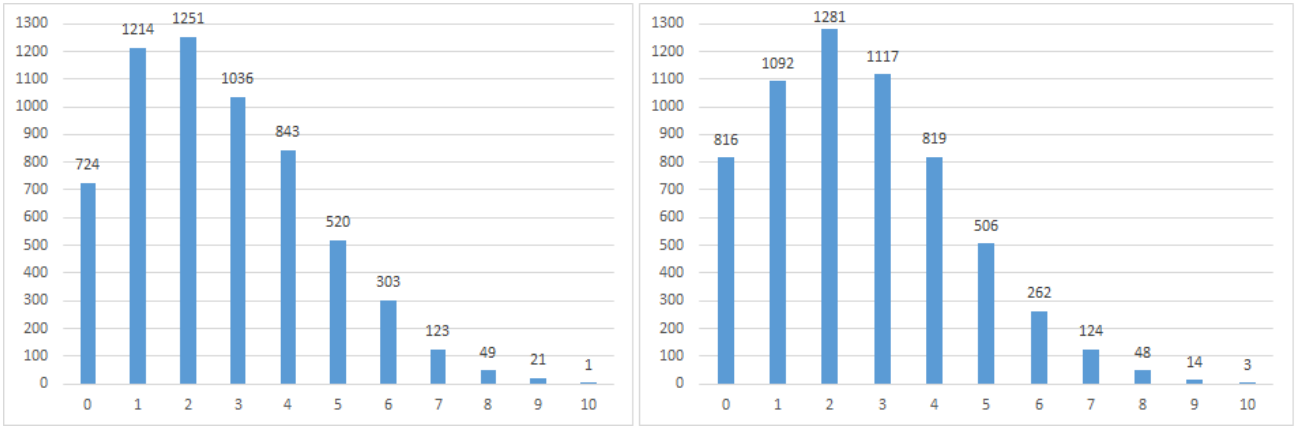
6 lentelėje pateikiama keliuose sakiniuose iš viso pasitaikė pasirinkti žymekliai prognozavimo duomenyse.

6 lentelė. Sakinių, kuriuose pasirodė žymekliai, skaičius pagal grupes prognozavimo duomenyse

V1	K1	N1	G1	Įn1	Vt1	V_K1	V_G1	V_Įn1	K2	G2	Įn2
15791	6471	9691	6929	1334	594	1965	989	272	9140	10473	3661

Prognozavimo duomenyse vyrauja panašios tendencijos kaip ir modelio sudarymo duomenyse. Vėl pastebėtina, jog dažniausiai į vieną sakinį neįeina daugiau nei vienas žymeklis.

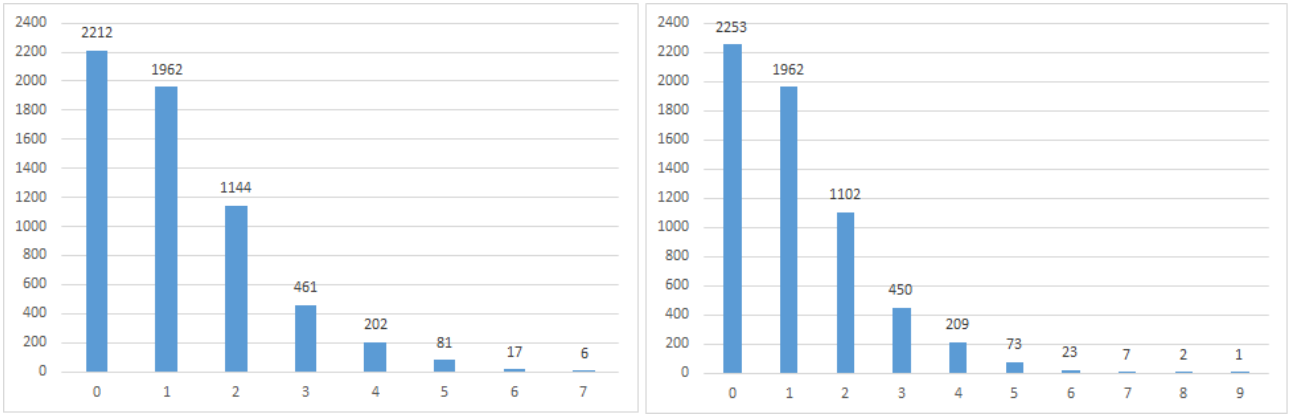
Histogramos. 5 - 7 paveikslėliuose pateikta $V1$, $G1$ ir $Įn1$ trečio tipo kintamųjų histogramos. Pagal $V1$ histogramą (5 pav.), vėl gi matyti, kad įvardžių vardininko linksnyje pasiskirstymas panašus į Puasono skirstinį.



5 pav. V1 grupės trečio tipo kintamojo modelio sudarymo (kairė) ir testavimo (dešinė) duomenų histogramos

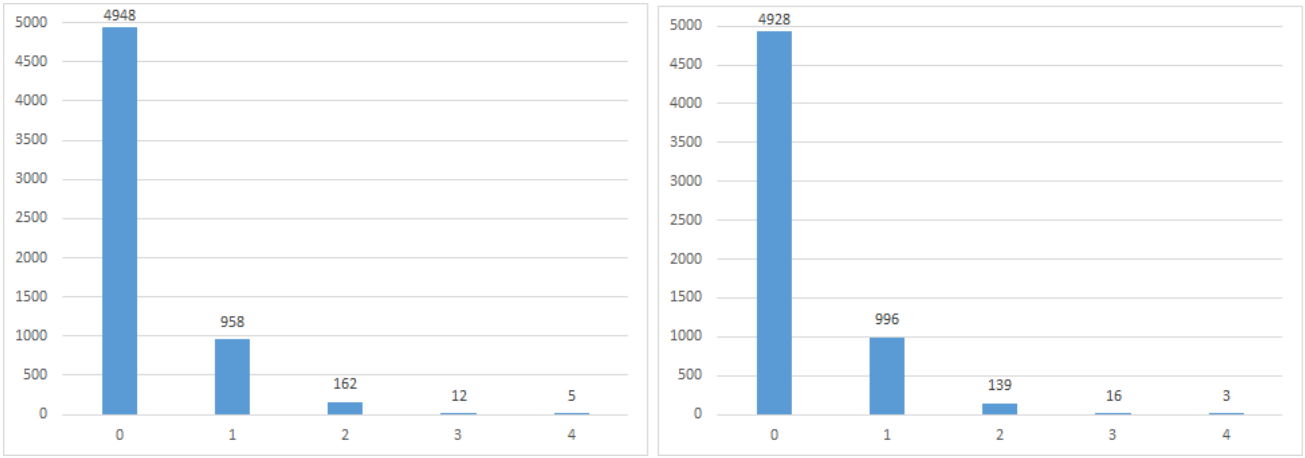
Taip pat vėl pastebėtina, jog pasiskirstymai abiejuose duomenų rinkiniuose yra gana panašūs. $V1$ grupės vidurkiai apmokymo duomenyse ir prognozavimo duomenyse taip pat skiriasi labai nedaug, atitinkamai 3,02 ir 2,98.

$G1$ grupės histogramos (6 pav.) primena neigiamo binominio skirstinio grafikus. Vėl gi, abiejuose duomenų rinkiniuose stebimas panašus pasiskirstymas. $G1$ grupės vidurkis apytiksliai vienodas abiejuose duomenų rinkiniuose apie 1,25.



6 pav. G1 grupės trečio tipo kintamojo modelio sudarymo (kairė) ir testavimo (dešinė) duomenų histogramos

Pažvelgus į Į$n1$ histogramas (7 pav), akivaizdu, jog tai labai reta grupė. Žymekliai iš šios grupės pasirodo tik kas penktame sakinių bloke. Į$n1$ grupės vidurkiai abiejuose duomenų rinkiniuose taip pat sutampa ir yra lygūs 0,228.



7 pav. Įn1 grupės trečio tipo kintamojo modelio sudarymo (kairė) ir testavimo (dešinė) duomenų histogramos

Nors duomenys buvo skaidomi atsitiktinai, ši grupė ir vėl yra panašiai pasiskirsčiusi abiejuose rinkiniuose. Remiantis šiuo faktu, galima daryti prielaidą, kad linksniai tekste nėra pasiskirstę visiškai atsitiktinai, o turi tendencijas išsidėlioti tam tikru dažnumu sakiniuose. Be to, nors trečiojo tipo kintamųjų interpretacija ir sudarymo principai skiriasi nuo dažnio kintamųjų, pastebėtina, jog skirstiniai yra panašūs.

Be aukščiau minėtų grupių, dar vienas svarbus kintamasis yra „*autoriai*“, kuris bus aptariamas sekančiame skyrelyje.

4.2.4. Kintamasis „*autoriai*“

Prieš atliekant tolimesnį tyrimą svarbu įsitikinti, jog kiekvienas autorius patektų į modelių sudarymui skirtus duomenis, nes tik tokiu atveju, mes galėsime įvertinti kiekvieno autoriaus įtaką nagrinėjamoms grupėms. Šiam tikslui, patogu pažiūrėti į dažnių lentelę 7.

7 lentelė. Autorių dažnių lentelė (modelio sudarymo duomenys)

Autoriaus nr.	1	2	3	4	5	6	7	8	9	10	11	12
Blokų skaičius	638	124	371	30	232	472	348	22	505	80	369	24
Autoriaus nr.	13	14	15	16	17	18	19	20	21	22	23	24
Blokų skaičius	500	236	388	327	226	212	203	302	24	195	41	216

Dažnių lentelės (7, 8) mums parodo, kad iš viso yra 24 autoriai (priminsiu, jog nagrinėjama 30 tekstų) ir į modelio sudarymo duomenis įeina bent keli blokai iš kiekvieno autoriaus. Daugiausiai sakinių blokų paimta iš pirmojo autoriaus Vinco Pietario kūrybos, mažiausiai- iš aštuntojo autoriaus Jono Biliūno - žinant jo kūrybą, to buvo galima tikėtis,

8 lentelė. Autorių dažnių lentelė (prognozavimo duomenys)

Autoriaus nr.	1	2	3	4	5	6	7	8	9	10	11	12
Blokų skaičius	668	117	365	20	193	516	370	13	504	54	375	28
Autoriaus nr.	13	14	15	16	17	18	19	20	21	22	23	24
Blokų skaičius	482	241	386	327	226	242	228	259	26	220	41	181

nes net ir turint palyginus daug jo kūrinių, jų trumpumas neleidžia sudaryti daug sakinių blokų.

9 lentelėje pateikiama, kiek vidutiniškai duotos grupės žymeklių autorius pavartoja viename sakinių bloke.

9 lentelė. Žymeklių pasiskirstymas pagal autorius

	G1	Įn1	K1	N1	V_G1	V_Įn1	V_K1	V1	Vt1	G2	Įn2	K2
1	0.909	0.071	1.525	2.281	0.215	0.078	0.235	2.616	0.143	2.182	1.103	2.509
2	0.726	0.145	0.685	1.218	0.040	0.024	0.274	2.460	0.008	1.581	0.315	1.081
3	1.647	0.199	1.035	2.402	0.146	0.040	0.604	4.094	0.051	2.216	0.736	1.819
4	1.567	0.100	1.133	2.067	0.033	0.000	0.567	5.100	0.000	3.333	0.767	2.100
5	1.228	0.323	1.134	1.664	0.289	0.121	0.388	2.026	0.017	2.909	1.375	2.302
6	1.646	0.326	1.710	2.076	0.197	0.042	0.303	3.839	0.087	2.233	0.663	1.909
7	0.851	0.126	0.830	1.282	0.043	0.023	0.195	2.718	0.009	0.954	0.351	0.865
8	1.636	0.045	0.864	2.773	0.182	0.136	0.864	2.091	0.000	1.591	0.682	2.591
9	1.727	0.356	1.362	2.026	0.091	0.036	0.560	4.434	0.063	1.772	0.648	1.618
10	1.150	0.112	0.838	1.812	0.025	0.000	0.150	1.875	0.050	1.137	0.375	1.087
11	0.547	0.092	0.531	0.919	0.119	0.008	0.100	1.501	0.024	1.453	0.268	1.192
12	1.458	0.542	0.792	1.417	0.125	0.042	0.750	2.917	0.083	2.833	0.875	0.875
13	0.560	0.298	0.868	0.994	0.086	0.022	0.246	1.376	0.120	1.494	0.524	1.494
14	0.949	0.258	1.127	1.360	0.114	0.017	0.403	1.504	0.229	2.708	1.042	2.513
15	1.747	0.206	0.773	2.157	0.415	0.111	0.222	2.755	0.082	2.428	0.660	1.222
16	1.599	0.153	1.034	1.979	0.187	0.073	0.370	5.287	0.067	2.064	0.581	1.425
17	0.757	0.146	0.801	1.195	0.155	0.066	0.177	2.558	0.031	2.009	0.553	1.602
18	1.170	0.311	1.330	2.203	0.104	0.014	0.453	3.608	0.066	2.250	0.642	2.509
19	3.862	0.754	3.177	3.931	0.108	0.089	0.468	5.517	0.374	4.414	1.064	3.453
20	1.325	0.238	1.291	1.974	0.199	0.033	0.285	4.162	0.070	2.007	0.477	1.593
21	1.125	0.167	1.500	1.000	0.042	0.000	0.042	3.167	0.042	2.583	0.708	1.792
22	1.036	0.215	1.169	1.108	0.154	0.082	0.251	2.174	0.149	1.974	0.379	1.415
23	0.537	0.024	1.415	0.707	0.171	0.073	0.415	1.561	0.195	1.780	0.317	1.561
24	0.551	0.116	0.634	0.907	0.213	0.056	0.120	1.546	0.046	1.204	0.366	1.597

Labiausiai iš visų išsiskiria 19 autorius - Daniel Defoe, kuri savo kūriniuose naudoja palyginti daug įvardžių ir prielinksnių. Nagrinėjant, tarkime, 24 autorių - Vladą Dautartą, pastebėtina, jog dažniau jis naudoja įvardžius, esančius vardininko ir naudininko linksniuose, ir prielinksnius, einančius su kilmininko ir galininko linksniais. Tačiau įvardžiai esantys kilmininko ar galininko linksnyje jo kūryboje vartojami palyginus retai, todėl galima daryti išvadą, kad vis gi autorius linkęs dažnai vartoti šiuos du linksnius su kitomis kalbos dalimis. Toks samprotavimas veda prie to, kad prielinksniai gali būti naudingi ir perspektyvūs pagrindinio tyrimo, nustatant visų linksniuojamų kalbos dalių pasireiškimo lygį, metu.

Kitame skyrelyje pateiksime dažnio ir trečio tipo kintamųjų koreliacinės analizės rezultatus.

4.2.5. Koreliacinė analizė

Koreliacinė analizė naudojama norint nustatyti, kokio stiprumo tiesinis ryšys veikia kintamuosius. Šiame skyrelyje pateiksime koreliacinę matricą modelio sudarymo duomenims (10 lentelė).

10 lentelė. Dažnio kintamųjų apmokymo duomenų koreliacinė matrica

	G1	Įn1	K1	N1	V_G1	V_Įn1
G1	1.00000000	0.145184919	0.22323196	0.27777011	0.071329192	0.031731001
Įn1	0.14518492	1.000000000	0.17064443	0.11195081	0.035218908	0.008625833
K1	0.22323196	0.170644430	1.00000000	0.21642634	0.042584092	0.017117366
N1	0.27777011	0.111950808	0.21642634	1.00000000	0.075015744	0.036447990
V_G1	0.07132919	0.035218908	0.04258409	0.07501574	1.000000000	0.019779181
V_Įn1	0.03173100	0.008625833	0.01711737	0.03644799	0.019779181	1.000000000
V_K1	0.12106163	0.074792462	0.06104662	0.06035315	-0.011591445	0.055658123
V1	0.35182170	0.140901692	0.24548283	0.35179232	0.007813335	-0.001646042
Vt1	0.03984798	0.034803603	0.07775956	0.01877598	0.006374396	0.005465484
G2	0.29820321	0.126176961	0.17879197	0.14825170	0.085419234	0.044495648
Įn2	0.08495044	0.221737446	0.16876999	0.08877902	0.046683389	0.055092205
K2	0.13171501	0.078112856	0.29744511	0.09411327	0.051878221	0.047347796

	V_K1	V1	Vt1	G2	Įn2	K2
G1	0.12106163	0.351821698	0.039847980	0.29820321	0.08495044	0.13171501
Įn1	0.07479246	0.140901692	0.034803603	0.12617696	0.22173745	0.07811286
K1	0.06104662	0.245482834	0.077759558	0.17879197	0.16876999	0.29744511
N1	0.06035315	0.351792315	0.018775978	0.14825170	0.08877902	0.09411327
V_G1	-0.01159145	0.007813335	0.006374396	0.08541923	0.04668339	0.05187822
V_Įn1	0.05565812	-0.001646042	0.005465484	0.04449565	0.05509220	0.04734780
V_K1	1.00000000	0.114815958	0.021783998	0.11238372	0.08261341	0.14578771
V1	0.11481596	1.000000000	-0.020462776	0.11793981	0.04823975	0.05407354
Vt1	0.02178400	-0.020462776	1.000000000	0.10275870	0.06557064	0.10478246
G2	0.11238372	0.117939811	0.102758699	1.00000000	0.20348160	0.28073846
Įn2	0.08261341	0.048239749	0.065570643	0.20348160	1.00000000	0.21427979
K2	0.14578771	0.054073543	0.104782464	0.28073846	0.21427979	1.00000000

Natūralu, kad didesnis koreliacijos koeficientas yra tarp $G1$ ir $G2$, $Įn1$ ir $Įn2$, $K1$ ir $K2$ grupių, nes parinkti prielinksniai eina kartu su tam tikro linksnio žodžiais. Stipriausia koreliacija pastebima tarp $V1$ - $G1$ ir $V1$ - $N1$ kintamųjų. Reikėtų priminti, jog šiame darbe yra nagrinėjama lietuvių kalba, kuri turi daug specifinių savybių, o jos elementus (žodžius, sakinius) sieja labai sudėtingi ryšiai. Dėl šių priežasčių, koreliacijos koeficientas įgyjantis reikšmę 0,3 laikomas pakankamai dideliu.

Trečiojo tipo kintamųjų koreliacinė matrica pateikta 11 lentelėje.

11 lentelė. Trečiojo tipo kintamųjų apmokymo duomenų koreliacinė matrica

	V1	K1	N1	G1	Įn1	Vt1
V1	1.000000000	0.21027845	0.30728242	0.30703661	0.11386347	-0.030951800
K1	0.210278446	1.00000000	0.19663177	0.20016112	0.15478271	0.072103911
N1	0.307282418	0.19663177	1.00000000	0.24669042	0.10285599	0.014554275
G1	0.307036612	0.20016112	0.24669042	1.00000000	0.12619919	0.028713264
Įn1	0.113863472	0.15478271	0.10285599	0.12619919	1.00000000	0.036315584
Vt1	-0.030951800	0.07210391	0.01455428	0.02871326	0.03631558	1.000000000
V_K1	0.106328002	0.05690419	0.06229067	0.12176115	0.06803354	0.024913111
V_G1	0.007554081	0.04171141	0.07236872	0.06469413	0.03150516	0.002465063
V_Įn1	-0.003488715	0.01949279	0.03582321	0.02913603	0.01212515	0.005623754
K2	0.043743446	0.27850501	0.07936044	0.11533523	0.07365921	0.100915465
G2	0.089360371	0.14057132	0.11602951	0.28393076	0.10709336	0.085753824
Įn2	0.037016256	0.15627947	0.08477082	0.08190875	0.23302932	0.067926117

	V_K1	V_G1	V_Įn1	K2	G2	Įn2
V1	0.10632800	0.007554081	-0.003488715	0.04374345	0.08936037	0.03701626
K1	0.05690419	0.041711407	0.019492794	0.27850501	0.14057132	0.15627947
N1	0.06229067	0.072368716	0.035823209	0.07936044	0.11602951	0.08477082
G1	0.12176115	0.064694134	0.029136027	0.11533523	0.28393076	0.08190875
Įn1	0.06803354	0.031505162	0.012125148	0.07365921	0.10709336	0.23302932
Vt1	0.02491311	0.002465063	0.005623754	0.10091547	0.08575382	0.06792612
V_K1	1.00000000	-0.014627306	0.058653306	0.14328537	0.10459198	0.08870693
V_G1	-0.01462731	1.000000000	0.017779769	0.05432707	0.07775993	0.04607785
V_Įn1	0.05865331	0.017779769	1.000000000	0.04150437	0.04639646	0.05699814
K2	0.14328537	0.054327073	0.041504367	1.00000000	0.24101508	0.19835960
G2	0.10459198	0.077759928	0.046396458	0.24101508	1.00000000	0.20393336
Įn2	0.08870693	0.046077845	0.056998137	0.19835960	0.20393336	1.00000000

Naujų ryšių iš 11 lentelės nepastebėta, tačiau visi ryšiai yra truputėlį silpnesni nei

tarp dažnio kintamųjų. Taip pat pastebėtina, jog tik tarp kelių kintamųjų yra atvirkštinė koreliacija (labai silpna). Įdomu tai, jog atvirkštinė koreliacija yra tik tuo atveju, jei bent į vieną iš dviejų grupių įeina įvardžių esančių vardininko linksnyje.

4.3. Pirmasis bandomasis tyrimas - logistinės regresijos modeliai

Pirmasis iš tyrimo klausimų: ar galima nuspėti pasirinktos grupės pasirodymą pagal kitas grupes. Šiam uždaviniui spręsti naudojama logistinė regresija ir ROC kreivių analizė. Šiame poskyryje pateikta trijų grupių - $V1$, $G1$, $Įn1$ - rezultatai. Priklausomas kintamasis yra dvinaris, o aiškinantieji kintamieji nurodo žymeklių iš duotos grupės pasirodymų skaičių bloke.

Tyrimo eiga:

1) priklausomas kintamasis yra paverčiamas dvinariu, t.y. jeigu žymekliai iš duotosios grupės pasirodė bloke, tuomet įgyjama reikšmė 1, priešingu atveju - 0.

2) Sudaromi logistinės regresijos modeliai, naudojant apmokymo duomenis.

3) Naudojant apmokymo duomenis yra prognozuojamos priklausomojo kintamojo reikšmės. Šiame etape prognozavimas vyksta su visomis c reikšmėmis intervale [0,1], parenkant žingsnį kas 0,001. Apskaičiuojamas padarytų klaidų skaičius kiekvienai c reikšmei ir išrenkama ta c, su kuria buvo suklysta mažiausiai kartų.

4) Atliekamas prognozavimas naudojant testinius duomenis ir pasirinktą c reikšmę.

5) Remiantis gautais rezultatai, nustatomas modelio tinkamumas prognozėms.

Tyrimo metu kiekvienai grupei sudaroma keturių tipų modeliai. Viena iš grupių laikoma priklausomu kintamuoju, o likusios grupės bei kintamasis „*autoriai*“ laikomi aiškinamaisiais kintamaisiais. Nagrinėjamos visos įmanomos aiškinamųjų kintamųjų kombinacijos. Pirmo tipo tinkamiausias modelis atrenkamas remiantis AIC kriterijumi, t.y. pasirenkamas tas modelis, kurio AIC kriterijaus reikšmė yra mažiausia. R pakete tai galima atlikti su funkcija „stepwise“. Antro tipo aiškinamaisiais kintamaisiais laikomos grupės bei jų 2-os eilės sąveikos ir kintamasis „*autoriai*“. Trečio ir ketvirto tipo modeliai sudaromi imant tuos pačius kintamųjų rinkinius, tačiau modelyje paliekami tik statistiškai reikšmingiausi kintamieji, t.y. kurių p reikšmė <0,01. Šie modeliai bus vadinami lakoniškaisiais.

Toliau šiame skyrelyje bus pateikiami logistinės regresijos modelių tarpiniai ir galutiniai rezultatai.

4.3.1. V1 grupė

Dažnas grupes reprezentatyviai atspindi $V1$ grupė - įvardžiai esantys vardininko linksnyje, kurios žymekliai pasitaiko 87% blokų. Šiame skyrelyje bus pateikti 4 modeliai, sudaryti šiai grupei, aptartos jų prognozavimo charakteristikos, išrinktas geriausias modelis bei įvertintos prognozavimo paklaidos.

Pirmasis modelis, sudarytas atsižvelgiant į AIC kriterijų.

$$V1 \sim K1 + N1 + G1 + Įn1 + autoriai. \tag{32}$$

Kaip jau buvo minėta anksčiau, tai nėra tikrasis modelio pavidalas, o tik simbolinis užrašas. Koeficientai ir p reikšmės pateikiamos 12 lentelėje.

12 lentelė. Logistinės regresijos modelio $V1$ grupei koeficientų ir p reikšmių lentelė

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	0.99764	0.14022	7.115	1.12e-12
K1	0.10955	0.04477	2.447	0.014396
N1	0.30576	0.04048	7.554	4.21e-14
G1	0.25851	0.04885	5.292	1.21e-07
Įn1	0.28994	0.10616	2.731	0.006310
autoriai2	-1.16797	0.44147	-2.646	0.008154
autoriai3	-0.24621	0.17612	-1.398	0.162115
autoriai4	1.41507	0.34476	4.105	4.05e-05
autoriai5	-0.85927	0.15662	-5.486	4.11e-08
autoriai6	-0.38623	0.16904	-2.285	0.022320
autoriai7	16.30991	3956.18033	0.004	0.996711
autoriai8	0.79160	0.25236	3.137	0.001708
autoriai9	15.06397	3956.18033	0.004	0.996962
autoriai10	-0.28800	0.16265	-1.771	0.076611
autoriai11	15.62549	874.55782	0.018	0.985745
autoriai12	0.56626	0.31913	1.774	0.076004
autoriai13	15.58898	214.65797	0.073	0.942107
autoriai14	0.52516	0.24607	2.134	0.032823
autoriai15	1.30403	0.31934	4.084	4.44e-05
autoriai16	1.42764	1.03080	1.385	0.166057
autoriai17	0.61840	0.27713	2.231	0.025650
autoriai18	-0.30722	0.38627	-0.795	0.426408
autoriai19	1.12331	0.29197	3.847	0.000119
autoriai20	15.24891	1237.90331	0.012	0.990172
autoriai21	1.45912	0.47153	3.094	0.001972
autoriai22	0.78441	0.22948	3.418	0.000630
autoriai23	-2.47861	0.94317	-2.628	0.008590
autoriai24	1.87461	0.34317	5.463	4.69e-08

Autorius modelyje yra faktorinis kintamasis, todėl kiekvienas autorius yra įvertinamas atskirai. Pastebėtina, jog ne visi autoriai yra statistiškai reikšmingi $V1$ grupei. Statistiškai reikšmingiausios $N1$ ir $G1$ grupės, kurios atitinkamai padidina tiesinio prediktoriaus reikšmę po 0,30576 ir 0,25851 už kiekvieną grupių vienetą. Iš tiesų, lietuvių kalboje dažnai yra naudojamos struktūros „kas kam“ arba „kas ką“. Pastebėtina, jog prognozuojant $V1$ grupę labai reikšminga yra ir laisvoji konstanta, kurios koeficientas yra 0,99764, t.y. jeigu

bloke nepasirodytų nei vienas žymeklis nei iš vienos grupės, tiesinio prediktoriaus reikšmė būtų 0,99764, o prognozuojama $V1$ grupės tikimybė pagal (6) formulę būtų 0,7305943.

Pritaikant sudarytą modelį apmokymo duomenims gautos prognozavimo charakteristikos pateikiamos 13 lentelėje.

13 lentelė. Logistinės regresijos modelio $V1$ grupei prognozių charakteristikos apmokymo duomenims

c	TP	Jautrumas	TN	Specifiškumas	VT
0.731	4881	0.9104645	249	0.3439227	0.8430567

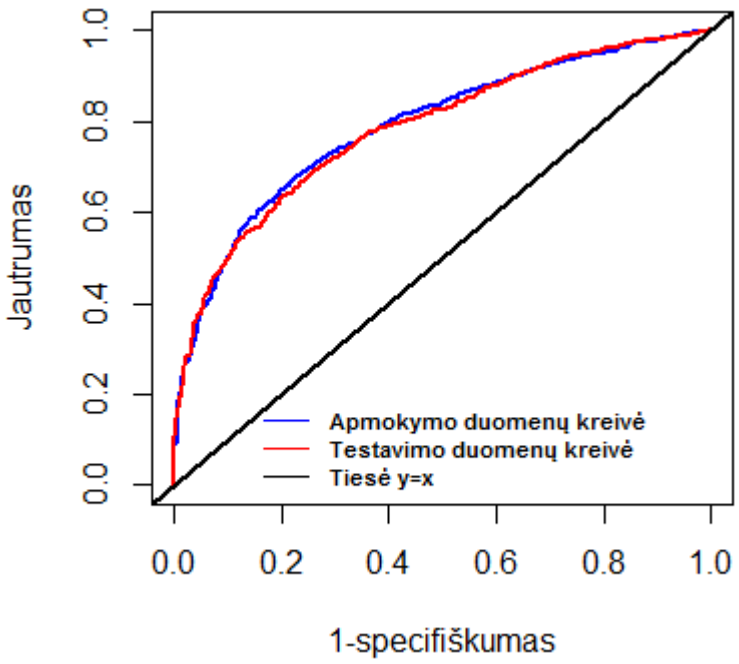
Šioje lentelėje TP - teisingai atspėti pasirodymai, TN - teisingai atspėti nepasirodymai, VT - santykis tarp teisingai atspėtų atvejų ($TP + TN$) ir imties dydžio (visų atvejų). Kritinė reikšmė c=0,731 buvo parinkta pagal apmokymo duomenis minimizuojant klaidų kiekį. Modelis pritaikytas apmokymo duomenims, teisingai atspėjo 84,31% atvejų. Teisingai atspėjo 4881 pasirodymus, t.y. apie 91% ir 249 nepasirodymus, t.y. tik apie 34%.

Pritaikius modelį testavimo duomenims, gautos prognozės charakteristikos pateiktos 14 lentelėje.

14 lentelė. Logistinės regresijos modelio $V1$ grupei prognozių charakteristikos testavimo duomenims

c	TP	Jautrumas	TN	Specifiškumas	VT
0.731	4856	0.922142	251	0.307598	0.8396909

Modelis pritaikytas testavimo duomenims, teisingai atspėjo 83,97% atvejų. Teisingai atspėjo 4856 pasirodymus, t.y. apie 92% ir 251 nepasirodymus, t.y. tik apie 31%. Natūralu, jog modelis pritaikytas testavimo duomenims prognozavo truputėlį prasčiau, tačiau šis skirtumas toks mažas, jog galima teikti, kad šis modelis yra tinkamas prognozėms. 8 paveikslėlyje pateiktos ROC kreivės, vaizdžiai iliustruoja panašų modelio prognozavimą, pritaikius jį apmokymo ir testavimo duomenims.



8 pav. V1 grupės logistinės regresijos modelio be sąveikų ROC kreivės

Tarp kintamųjų gali būti ir sudėtingesnių ryšių, todėl į modelius taip pat buvo įtrauktos ir jų 2-os eilės sąveikos.

Antrasis modelis - V1 grupės modelis su sąveikomis.

$$\begin{aligned} V1 \sim {} & \text{K1} + \text{N1} + \text{G1} + \text{Įn1} + \text{Vt1} + \text{V\_K1} + \text{V\_Įn1} + \text{K2} + \text{G2} + \text{Įn2} + \\ & + \text{autoriai} + \text{K1:G2} + \text{K1:Įn2} + \text{N1:Vt1} + \text{G1:Įn1} + \text{G1:V\_Įn1} + \text{G1:G2} + \\ & + \text{Vt1:V\_K1} + \text{Vt1:Įn2} + \text{V\_K1:K2} + \text{V\_Įn1:Įn2} + \text{K2:Įn2}. \end{aligned} \tag{33}$$

Šio ir kitų modelių koeficientų įverčių ir p reikšmių lentelės darbe nebus pateiktos, nes šio tyrimo tikslas yra analizuoti modelių struktūras (kokie kintamieji ir jų saveikos patenka į modelį) bei įvertinti jų tinkamumą prognozavimui, o ne stebėti konkrečius įvertintus koeficientus prie kintamųjų, ypač nominaliųjų kintamųjų su dideliu kategorijų skaičiumi.

Pasirinkta c kritinė reikšmė - 0,726. Prognozavimo charakteristikos apmokymo duomenims pateiktos 15 lentelėje.

15 lentelė. Logistinės regresijos modelio su saveikomis $V1$ grupei prognozių charakteristikos apmokymo duomenims

c	TP	Jautrumas	TN	Specifiškumas	VT
0.726	4887	0.9115837	251	0.3466851	0.8443714

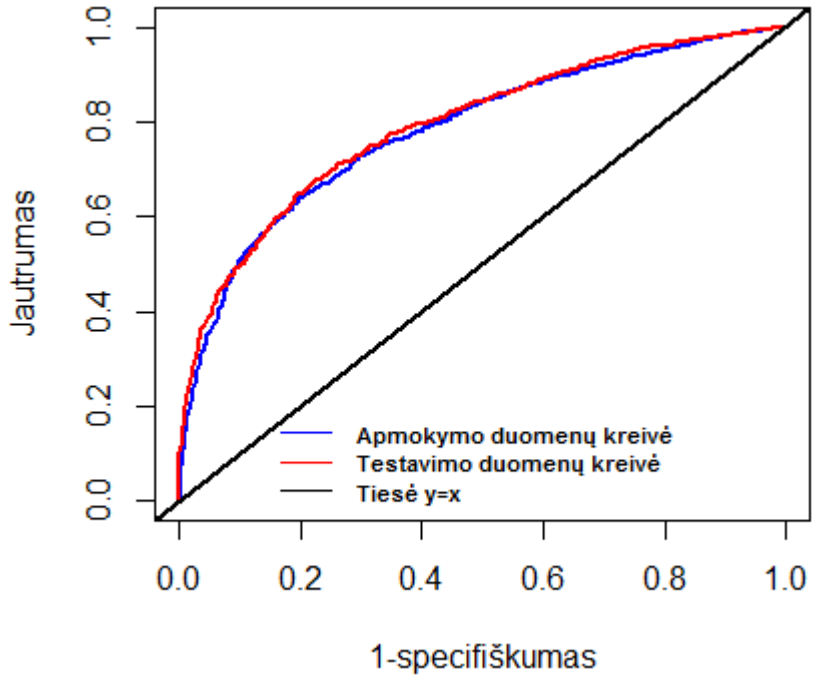
Modelis pritaikytas apmokymo duomenims, teisingai atspėjo 84,44% atvejų. Teisingai atspėjo 4887 pasirodymus, t.y. apie 91% ir 251 nepasirodymus, t.y. tik apie 35%.

Pritaikius modelį testavimo duomenims, gautos prognozės charakteristikos pateiktos 16 lentelėje.

16 lentelė. Logistinės regresijos modelio su saveikomis $V1$ grupės prognozių charakteristikos testavimo duomenims

c	TP	Sensitivity	TN	Specifity	Corect
0.726	4838	0.9187239	249	0.3051471	0.8364025

Modelis pritaikytas testavimo duomenims, teisingai atspėjo 83,64% atvejų. Teisingai atspėjo 4838 pasirodymus, t.y. apie 92% ir 249 nepasirodymus, t.y. tik apie 31%. 9 paveikslėlyje pateiktos ROC kreivės vėl parodo, jog pasirinkus bet kokią c reikšmę modelis pritaikytas apmokymo duomenims prognozuos labai panašiai kaip ir pritaikytas prognozavimo duomenims.



9 pav. $V1$ grupės logistinės regresijos modelio su sąveikomis ROC kreivės

Pastebėtina, jog yra tam tikrų c reikšmių su kuriomis modelis, pritaikytas testavimo duomenims, gautų net tikslesnes prognozes, nei pritaikytas apmokymo duomenims.

Tai, kad modelyje su sąveikomis nebuvo statistiškai reikšmingų kintamojo „*autoriai*“ sąveikų su kitais kintamaisiais, leidžia daryti prielaidą, kad sąryšiai tarp įvairių linksnių var-

tojimo yra gana universalūs, mažai priklauso nuo autoriaus preferencijų. Sąveikų įtraukimas į modelį nepagerino parinkto modelio tinkamumo matų, taip pat ir prognozės tikslumo.

Trečiasis modelis, sudarytas remiantis aiškinamųjų kintamųjų statistiniu reikšmingumu (p reikšmė < 0,01). Lakoniškasis V1 modelis:

$$V1 \sim N1 + G1 + Įn1 + autoriai. \tag{34}$$

Šis modelis skiriasi nuo pirmojo, aprašyto (32) formule, tuom, kad į jį neįeina kintamasis $K1$. Lakoniškasis modelis pritaikytas modelio sudarymo duomenims teisingai atspėjo 84,26 % atvejų, pasirinkus kritinę reikšmę c=0,734. Išmetus iš pirmojo modelio kintamąjį $K1$ atspėjimo procentas buvo sumažintas 0,05 procentinio punkto, t.y. buvo padaryta 3 klaidomis daugiau nei su pirmuoju modeliu. Lakoniškasis modelis teisingai prognozavo 4895 pasirodymus (jautrumas = 0,9131) ir 232 nepasirodymus (specifiškumas = 0,3204).

Modelis aprašytas (34) formule ir pritaikytas testavimo duomenims, teisingai prognozavo 84,03% atvejų, o tai net geriau nei pirmasis modelis, kuris prognozavo 83,97% atvejų teisingai. Teisingai atspėta buvo 4855 pasirodymai (jautrumas = 0,922) ir 256 (specifiškumas = 0,3137) nepasirodymai.

Ketvirtasis modelis - V1 grupės lakoniškasis modelis su sąveikomis.

$$V1 \sim N1 + G1 + autoriai. \tag{35}$$

Kaip matome, lakoniškajame modelyje neliko sąveikų, o tai parodo, kad sudarinėjant modelius $V1$ grupei sąveikos nėra statistiškai reikšmingos ir šiai grupei nusakyti pakanka vien pagrindiniais veiksniais aprašomų ryšių. Taip pat, šis modelis prognozuoja tiksliau nei modelis su sąveikomis (antrasis). Šis modelis pritaikytas testiniams duomenims teisingai atspėja 83,95% atvejų. Teisingai prognozavo 4850 pasirodymus, t.y. apie 92,1% ir 256 nepasirodymus, t.y. tik apie 31,4%.

Šis modelis dar supaprastėjo, lyginant su trečiuoju modeliu, nereikšmingas tapo ir kintamasis $Įn1$. Tačiau jo atsisakymas lėmė, kad prognozuojant buvo padaryta dar 5 vienetais daugiau klaidų. Kadangi tai nėra labai didelis klaidų skaičius, priklausomai nuo uždavinio nustatoma, ar svarbiau tiksliau atspėti keliuose blokuose pasirodė įvardis esantis vardininko linksnyje, ar sutaupyti darbo ir laiko resursų, neskaičiuojant $Įn1$ grupės žymeklių, tačiau padarant papildomas kelias klaidas.

Geriausias modelis - trečiasis. Geriausiai prognozuojantis $V1$ grupę modelis yra trečiasis, aprašytas (34) formule. Šis modelis atitinka pagrindines tinkamo modelio savybes: yra paprastas - naudojama palyginus nedidelis skaičius žymeklių, taip pat atitinka teoriją bei realią situaciją, t.y. logiška, kad į modelį įeina $G1$ ir $N1$ linksniai, nes jie yra labai

dažnai vartojami kartu su vardininku, bei kintamasis „*autoriai*“ parodo, kad autoriai iš šio duomenų rinkinio linkę vartoti skirtingą kiekį $V1$ grupės žymeklių. Deviacijos statistikos reikšmė yra 3745,8 su 6058 laisvės laipsniais, o AIC kriterijaus reikšmė yra 3799,8. 17 lentelėje pateikiama, kaip šis modelis prognozavo kiekvienoje prognozavimo duomenų dalyje, sumuojant prognozuojamas reikšmes (1 arba 0).

17 lentelė. V1 grupės sumos prognozės, sumuojant prognozuojamas reikšmes

Dalies nr.	1	2	3	4	5
Prognozuojama suma	544	547	566	544	523
Tikra suma	527	530	543	538	506
Skirtumas	-17	-17	-23	-6	-17
Santykinė paklaida	0.032	0.032	0.042	0.011	0.034
Dalies nr.	6	7	8	9	10
Prognozuojama suma	534	544	527	527	559
Tikra suma	509	529	506	526	552
Skirtumas	-25	-15	-21	-1	-7
Santykinė paklaida	0.049	0.028	0.042	0.002	0.013

Lentelė parodo, jog kiekvienoje prognozavimo dalyje modelis prognozavo truputį per daug pasirodymų, vadinasi reikėjo pasirinkti didesnę kritinę c reikšmę, tačiau padaromos paklaidos yra nedidelės. Vidutinė paklaida yra 0,0285, didžiausia padaroma paklaida yra 0,049 t.y. 4,9%, šiuo atveju pasirodymų skaičius buvo per daug padidintas 25 vienetais.

Prognozuojant bendras (sumines, vidutines ir pan.) charakteristikų reikšmes, galima naudoti ir prognozių tikimybes, t.y. dydžius, kuriuos betarpiškai prognozuoja parinktas modelis. Rezultatai pateikiami 18 lentelėje.

18 lentelė. V1 grupės sumos prognozės, sumuojant prognozių tikimybes

Dalies nr.	1	2	3	4	5
Prognozuojama suma	535,5	540,1	554,4	548,3	516,7
Tikra suma	527	530	543	538	506
Skirtumas	8,5	10,1	11,4	10,3	10,7
Santykinė paklaida	0,016	0,019	0,021	0,019	0.021
Dalies nr.	6	7	8	9	10
Prognozuojama suma	529,3	546,2	518,2	528,1	556,5
Tikra suma	509	529	506	526	552
Skirtumas	20,3	17,2	12,2	2,1	4,5
Santykinė paklaida	0,040	0,033	0,024	0,004	0,008

Sumuojant prognozių tikimybes gaunami tikslesni rezultatai nei sumuojant prognozuojamas reikšmes, tačiau ir šiuo atveju prognozės kiekvienoje dalyje yra didesnės už tikrąsias sumas. Tai gali nutikti dėl kelių priežasčių:

1) sudarant apmokymo ir testinius duomenis, $V1$ pastaruosiuose pasitaikė ženkliai dažniau;

2) vardininkas sunkiai prognozuojamas linksnis, kuris ženkliai priklauso nuo autoriaus. Kai kurie autoriai linkę naudoti "menamus veiksnius", todėl jų kūriniuose vardininkų bus mažiau;

3) naudojant logistinės regresijos modelį prarandama nemažai informacijos apie vardininkų grupę. Vardininkas yra labai dažnas linksnis, todėl svarbesnė informacija yra ne jo pasirodymas ar nepasirodymas sakinių bloke, bet dažnis, t.y. kiek kartų jis pasirodė bloke.

4.3.2. G1 grupė

$G1$ grupė reprezentuoja vidutinio dažnumo linksnius. Šiai grupei vėl gi buvo sudaryti 4 modeliai:

1) modelis be sąveikų (pagal AIC):

$$G1 \sim K1 + N1 + V\_K1 + V1 + G2 + autoriai; \tag{36}$$

2) modelis su sąveikomis (pagal AIC):

$$
\begin{aligned}
G1 \sim\ & \text{Įn1} + \text{K1} + \text{N1} + \text{V\_G1} + \text{V\_Įn1} + \text{V\_K1} + \text{V1} + \text{Vt1} + \\
& + \ \text{G2} + \text{Įn2} + \text{K2} + \textit{autoriai} + \text{Įn1:Vt1} + \text{Įn1:K2} + \text{K1:V\_Įn1} + \\
& + \ \text{K1:K2} + \text{N1:V\_G1} + \text{N1:V\_K1} + \text{N1:V1} + \text{V\_G1:V\_K1} + \\
& + \ \text{V\_G1:K2} + \text{V\_Įn1:Įn2} + \text{V\_Įn1:K2} + \text{V\_K1:Vt1} + \\
& + \text{V\_K1:G2} + \text{V1:Vt1} + \text{Vt1:Įn2} + \text{Įn1:V\_G1};
\end{aligned}
\tag{37}
$$

3) lakoniškasis modelis be sąveikų:

$$G1 \sim N1 + V\_K1 + V1 + G2 + autoriai; \tag{38}$$

4) lakoniškasis modelis su sąveikomis:

$$G1 \sim N1 + V\_K1 + V1 + G2 + autoriai; \tag{39}$$

Parinktus modelius, ypač su sąveikomis, interpretuoti sudėtinga. Jų sudėtingumas dalinai gali būti paaiškintas gana siauru naudojamų žymeklių rinkiniu, susietu tik su dviejų tipų funkciniais žodžiais (įvardžiais ir prielinksniais). Tų tipų funkcinių žodžių vartojimo ypatybės gali būti viena iš sudėtingos parinktų modelių struktūros priežasčių.

Tačiau galima pastebėti tam tikras ypatybes. Į visus modelius vėl patenka kintasis „*autoriai*“, tai dar kartą parodo jo svarbą tyrime. Palyginus pirmąjį modelį (36) su trečiuoju (38), matoma, jog kintamasis $K1$ tampa statistiškai nereikšmingu įvedus apribojimą, kad p reikšmė turi būti mažesnė už 0,01. Svarbiausias rezultatas šiai grupei yra tai, jog lakoniškajame modelyje (39) nelieka sąveikų, dėl to galima daryti prielaidą, kad sąveikos nėra statistiškai reikšmingos prognozuojant įvardžių esančių galininko linksnyje pasirodymą. Taip pat, lakoniškieji modeliai (38 ir 39) sutampa.

Visų 4 modelių prognozavimo rezultatai pateikiami 19 lentelėje.

19 lentelė. G1 grupės prognozių charakteristikos apmokymo duomenims

	c	TP	Jautrumas	TN	Specifiškumas	VT
1 modelis	0,571	2905	0,750065	1247	0,563743	0,682334
2 modelis	0,572	2906	0,750323	1243	0,561935	0,681841
3 - 4 modeliai	0,572	2913	0,75213	1252	0,566004	0,68447

Daugiausiai atvejų teisingai atspėjo (apie 69%) 3 – 4 modeliai. $G1$ grupės modeliai skiriasi nuo $V1$ grupės modelių tuom, kad geriau prognozuoja nepasirodymus, bet prasčiau pasirodymus. Tačiau tiek pasirodymus, tiek ir nepasirodymus modeliai atspėja daugiau nei 50%. Likusių modelių (1 ir 2), pritaikytų apmokymo duomenims prognozavimo rezultatai panašūs.

Iš 20 lentelės matyti, kad mažiausia paklaida gaunama taikant 3 – 4 modelius. Šie modeliai nepadaro nei vienos klaidos, taigi duomenys šiais modeliai aprašomi puikiai. Tačiau likę modeliai irgi pakankamai tiksliai prognozuoja.

20 lentelė. G1 grupės prognozės ir paklaidos (apmokymo duomenys), sumuojant prognozuojamas reikšmes

	Prognozuojama suma	Tikra suma	Skirtumas	Santykinė paklaida
1 modelis	3870	3783	3	0,000775
2 modelis	3875	3783	-2	0,00052
3 - 4 modeliai	3873	3873	0	0

Gauti rezultatai, pritaikius sudarytus modelius testavimo duomenims, pateikti 21 – 22 lentelėse. Matyti, kad 3 – 4 modeliai ir testavimo duomenis geriausiai aprašo. Šiuo atveju, jie teisingai atspėjo 65,97% atvejų.

21 lentelė. G1 grupės prognozių charakteristikos testavimo duomenims

	c	TP	Jautrumas	TN	Specifiškumas	VT
1 modelis	0,571	2831	0,739358	1190	0,528185	0,661131
2 modelis	0,572	2818	0,735962	1191	0,528629	0,659158
3 - 4 modeliai	0,572	2812	0,734395	1200	0,532623	0,659651

*G*1 grupės atveju, prognozuojant sumines charakteristikas geriau sumuoti prognozuojamas reikšmes, nes tokiu atveju padaromos mažesnės paklaidos. Kaip matome iš 22 lentelės, sumuojant reikšmes, prognozuotas naudojant 3 – 4 modelius, santykinė paklaida 0,0094, t.y. apie 1%.

22 lentelė. G1 grupės prognozės ir paklaidos (testavimo duomenys), sumuojant prognozuojamas reikšmes

	Prognozuojama suma	Tikra suma	Skirtumas	Santykinė paklaida
1 modelis	3894	3829	-65	0,01698
2 modelis	3880	3829	-51	0,01332
3 - 4 modeliai	3865	3829	-36	0,0094

Sumuojant abiem atvejais prognozuojama suma gaunama truputį didesnė, nei tikroji suma. Pastebėtina, kad sumuojant prognozuojamas tikimybes (23 lentelė) gaunamos daug stabilesnės paklaidos nei sumuojant prognozuojamas reikšmes. Jeigu sumuojant prognozuojamas reikšmes skirtumas svyruoja nuo 36 iki 65, tai prognozuojamų tikimybių sumavimo atveju, skirtumas svyruoja tik nuo 42,1 iki 43,7. Taigi, sumuojant tikimybes padaroma didesnė paklaida, tačiau ji mažai skiriasi tarp skirtingų modelių.

23 lentelė. G1 grupės prognozės ir paklaidos (testavimo duomenys), sumuojant prognozuojamas tikimybes

	Prognozuojama suma	Tikra suma	Skirtumas	Santykinė paklaida
1 modelis	3871,8	3829	-42,8	0,01118
2 modelis	3871,1	3829	-42,1	0,011
3 - 4 modeliai	3872,7	3829	-43,7	0,01141

Plačiau nagrinėjama 3 – 4 modeliai, nes jie tikslesni ir paprastesni. 24 lentelėje pateikiama prognozavimo rezultatai kiekvienoje testavimo duomenų dalyje. Smarkiai išsiskiria 8 ir 10 dalys: 8 dalyje žymekliai iš $G1$ grupės pasirodė ženkliai mažiau kartų, tuo tarpu 10 dalyje modelis nepagrįstai padidino prognozuojamą sumą.

24 lentelė. G1 grupės suminės prognozės, sumuojant prognozių tikimybes

Dalies nr.	1	2	3	4	5
Prognozuojama suma	387,6	391,1	399,4	392	373,2
Tikra suma	391	387	405	384	370
Skirtumas	-3,4	4,1	-5,6	8	3,2
Santykinė paklaida	0,009	0,011	0,014	0,021	0,009
Dalies nr.	6	7	8	9	10
Prognozuojama suma	377,4	394,9	374,6	380,1	402,4
Tikra suma	383	393	355	381	380
Skirtumas	-5,6	1,9	19,6	-0,9	22,4
Santykinė paklaida	0,015	0,005	0,055	0,002	0,059

Tačiau apžvelgus 24 lentelę matome, jog modeliai ne visur prognozuoja didesnes sumas nei yra tikrosios, o tas skirtumas atsiranda būtent iš anksčiau minėtų dviejų išsiski-

riančių dalių.

4.3.3. Įn1 grupė

Įvardžių esančių įnagininko linksnyje grupė reprezentuoja labai retus atvejus. Sudaryti 4 modeliai:

1) modelis be sąveikų (pagal AIC):

$$\text{Įn1} \sim \text{K1} + \text{N1} + \text{V1} + \text{G1} + \text{G2} + \text{Įn2} + \textit{autoriai}; \tag{40}$$

2) modelis su sąveikomis (pagal AIC):

$$\begin{aligned}\text{Įn1} \sim\ & \text{G1} + \text{K1} + \text{N1} + \text{V\_G1} + \text{V\_Įn1} + \text{V\_K1} + \text{V1} + \text{Vt1} + \\ & + \text{K2} + \textit{autoriai} + \text{G1:K1} + \text{G1:N1} + \text{N1:G2} + \text{V\_G1:V\_Įn1} + \\ & + \text{V\_G1:V\_K1} + \text{G2} + \text{Įn2} + \text{V\_Įn1:V\_K1} + \text{V1:Vt1} + \\ & + \text{V1:Įn2} + \text{Vt1:G2} + \text{Vt1:Įn2} + \text{G2:Įn2} + \text{Įn2:K2};\end{aligned} \tag{41}$$

3) lakoniškasis modelis be sąveikų:

$$\text{Įn1} \sim \text{K1} + \text{V1} + \text{Įn2} + \textit{autoriai}; \tag{42}$$

4) lakoniškasis modelis su sąveikomis:

$$\begin{aligned}\text{Įn1} \sim\ & \text{K1} + \text{Įn2} + \textit{autoriai} + \text{G1:N1} + \text{V\_G1:V\_K1} + \\ & + \ \text{V1:Įn2} + \text{Įn2:K2};\end{aligned} \tag{43}$$

Į kiekvieną logistinės regresijos modelį sudarytą $Įn1$ grupei įeina autorių nurodantis kintamasis. Taip pat visuose modeliuose yra įtrauktas kintamasis $Įn2$. Modeliuose su sąveikomis nėra sąveikų tarp grupių ir kintamojo „*autoriai*“, tai vėl gi parodo, kad įnagininko linksnio vartojimas nepriklauso nuo autorių polinkių.

Šiuos modelius interpretuoti yra labai sudėtinga dėl didelio įtrauktų kintamųjų skaičiaus. Paprastesnis modelis yra trečiasis. Natūralu, kad į modelį įeina kintamieji $V1$, $Įn2$, „*autoriai*“, tačiau kombinacija $Įn1$ su $K1$ yra retesnė. Žinoma, reikia nepamiršti, kad šiame darbe yra nagrinėjami ne sakiniai, o jų blokai, todėl tikėtina, kad grupė $K1$ patenka į modelį ne dėl ryšio su įnagininko linksniu sakinio rėžiuose, bet dėl sakinų tarpusavio ryšių.

Kiekvieno modelio prognozių charakteristikos, naudojant testavimo duomenis 25 lentelėje.

25 lentelė. Įn1 grupės prognozių charakteristikos apmokymo duomenims

	c	TP	Jautrumas	TN	Specifiškumas	VT
1 modelis	0,277	465	0,408971	4279	0,864794	0,779622
2 modelis	0,281	466	0,409851	4280	0,864996	0,779951
3 modelis	0,276	456	0,401055	4268	0,862571	0,776335
4 modelis	0,273	461	0,405453	4278	0,864592	0,7788

Prognozuojant pasirodymus atskiruose blokuose tiksliausias yra antrasis modelis, kuris teisingai prognozuoja apie 80% atvejų. Likę modeliai prognozuoja panašiai - prasčiausiai prognozuojantis trečiasis modelis teisingai atspėja 77,63% atvejų.

Tačiau suminių prognozių lentelė (26 lentelė) parodo, kad geriausiai prognozuojantis modelis yra trečiasis. Taigi, kurį modelį pasirinkti reikėtų spręsti pagal tai koks yra tyrimo tikslas: ar nustatyti atskirų pasirodymų atvejus, ar prognozuoti bendrą pasirodymų sumą.

26 lentelė. Įn1 grupės prognozės ir paklaidos (apmokymo duomenys), sumuojant prognozuojamas reikšmes

	Prognozuojama suma	Tikra suma	Skirtumas	Santykinė paklaida
1 modelis	1134	1137	3	0,002639
2 modelis	1134	1137	3	0,002639
3 modelis	1136	1137	1	0,00088
4 modelis	1131	1137	6	0,005277

Prognozuojant trečiuoju modeliu bendrą blokų, kuriuose pasirodė žymekliai, skaičių apsirinkama tik vienu vienetu, santykinė paklaida yra 0,00088.

Pritaikius modelius testavimo duomenims gaunamas truputį prastesnis rezultatas (27 lentelė), tačiau vis tiek teisingai atspėtų atvejų procentas viršija 77,5 %.

27 lentelė. Įn1 grupės prognozių charakteristikos testavimo duomenims

	c	TP	Jautrumas	TN	Specifiškumas	VT
1 modelis	0,277	459	0,397747	4263	0,865057	0,776389
2 modelis	0,281	476	0,404679	4265	0,865463	0,778034
3 modelis	0,276	463	0,401213	4255	0,863433	0,775732
4 modelis	0,273	468	0,405546	4267	0,865869	0,778527

Kaip ir apmokymo duomenyse (25 lentelė), taip ir testavimo duomenyse (27 lentelė) konkrečius blokus geriau prognozuoja modeliai į kuriuos yra įtrauktos sąveikos, tuo tarpu bendrą sumą (28 lentelė) geriau prognozuoja tyrimo paprasčiausias modelis - trečiasis modelis. Jis testavimo duomenyse apsirinka 18 vienetų, santykinė paklaida - 0,016.

28 lentelė. Įn1 grupės prognozės ir paklaidos (testavimo duomenys), sumuojant prognozuojamas reikšmes

	Prognozuojama suma	Tikra suma	Skirtumas	Santykinė paklaida
1 modelis	1124	1154	30	0,025997
2 modelis	1130	1154	24	0,020797
3 modelis	1136	1154	18	0,015598
4 modelis	1129	1154	25	0,021664

Į$n1$ grupės atveju, geriau prognozuojant bendrą sumą yra sumuoti prognozuojamas reikšmes, tokiu atveju padaroma mažesnė paklaida. Tačiau kaip ir $G1$ grupei, tai ir šiai grupei, sumuojant prognozuojamas reikšmes parinktas modelis paklaidos dydžiui turi didesnę įtaką nei sumuojant prognozuojamas tikimybes (29 lentelė).

29 lentelė. Įn1 grupės prognozės ir paklaidos (testavimo duomenys), sumuojant prognozuojamas tikimybes

	Prognozuojama suma	Tikra suma	Skirtumas	Santykinė paklaida
1 modelis	1124,5	1154	29,5	0,025563
2 modelis	1130,2	1154	23,8	0,020624
3 modelis	1123,1	1154	30,9	0,026776
4 modelis	1128	1154	26	0,02253

30 lentelėje pateikiama trečiojo modelio pritaikyto testinių duomenų atskiroms dalims prognozavimo rezultatai. Skirtumai tarp prognozuojamų ir tikrųjų sumų yra mažesni nei $V1$ ar $G1$ grupių, tačiau paklaidos yra didesnė. Priežastis yra ta, kad $Įn1$ grupė yra labai reta ir kiekviena padaryta klaida smarkiai didina paklaidą, kuri 8 bloke išaugo net iki 16,2%.

30 lentelė. Įn1 grupės suminės prognozės, sumuojant prognozių tikimybes

Dalies nr.	1	2	3	4	5
Prognozuojama suma	111	123	112	121	113
Tikra suma	121	123	121	123	118
Skirtumas	10	0	9	2	5
Santykinė paklaida	0,083	0	0,074	0,016	0,042
Dalies nr.	6	7	8	9	10
Prognozuojama suma	101	112	115	115	113
Tikra suma	114	106	99	108	121
Skirtumas	13	-6	-16	-7	8
Santykinė paklaida	0,114	0,057	0,162	0,065	0,066

4.3.4. Pirmojo bandomojo tyrimo išvados

Atlikus tyrimą buvo pastebėta, kad nėra vieno tipo modelio, kuris tiktų visiems atvejams. Jei norima prognozuoti bendras (sumines, vidurkines ar pan.) charakteristikas geriau yra naudoti lakoniškuosius modelius be sąveikų. Jie ne taip tiksliai prognozuoja pavienius atvejus, tačiau gana neblogai įvertina tikėtiną sumą. Tuo tarpu, jei tyrimo tikslas

yra prognozuoti atskirus blokus, tuomet vertėtų naudoti išsamesnius modelius, įtraukiant ir kintamųjų sąveikas. Tačiau reikėtų nepamiršti, jog šių modelių naudojimas atima nemažai laiko bei reikalauja pakankamai galingų kompiuterių, be to juos labai sudėtinga interpretuoti.

Taip pat buvo pastebėta, kad logistinės regresijos modeliai tinkamiausi vidutinio dažnumo grupėms prognozuoti. Labai retų ar labai dažnų grupių prognozavimas naudojant logistinės regresijos modelius yra ne toks tikslus. $V1$ grupės atveju, didesnę paklaidą taip pat įtakojo padidintas žymeklių skaičius testiniuose duomenyse.

Vienas svarbiausių pastebėjimų: visuose sudarytuose modeliuose buvo įtrauktas kintamasis „*autoriai*“, tai parodo, kad skirtingi autoriai yra linkę naudoti skirtingą kiekį įvairių linksnių.

Paverčiant dažnas grupes, tokias kaip $V1$ ar $G1$ binariniais kintamaisiais prarandama daug informacijos, todėl svarbu nagrinėti ne tik funkcinių žodžių pasirodymą - nepasirodymą, bet ir jų dažnius. Tai bus atliekama kitame poskyryje, naudojant Puasono regresijos modelius.

4.4. Antrasis bandomasis tyrimas - Puasono regresijos modeliai

Dar vienas labai svarbus klausimas: ar galima nuspėti vienų žymeklių dažnius per kitus žymeklius. Šiam uždaviniui spręsti naudojami Puasono regresijos modeliai. Priklausomu kintamuoju laikoma pasirinkta grupė, kurios reikšmės yra žymeklių skaičius sakinių bloke.

Šio tyrimo etapai:

1) suskaičiuojamos priklausomojo ir aiškinamųjų kintamųjų reikšmės.

2) Sudaromi Puasono ir neigiamos binominės regresijos modeliai, naudojant apmokymo duomenis.

3) Naudojant apmokymo duomenis yra prognozuojamos priklausomojo kintamojo reikšmės.

4) Atliekamas prognozavimas naudojant testinius duomenis.

5) Remiantis gautais rezultatai, nustatomas modelio tinkamumas prognozėms.

Šiame bandomajame tyrime taip pat bus sudaromi keturių tipų modeliai, kurie buvo aprašyti 4.3. poskyryje. Toliau bus apžvelgiami modeliai sudaryti toms pačioms grupėms $V1$, $G1$, $Įn1$.

4.4.1. V1 grupė

Žemiau išvardinti keturi Puasono regresijos modeliai, sudaryti V1 grupei.

1) modelis be sąveikų (pagal AIC):

$$V1 \sim \text{G1} + \text{Įn1} + \text{K1} + \text{N1} + \text{V\_G1} + \text{V\_Įn1} + \text{V\_K1} + \\ + \ \text{Vt1} + \text{K2} + \textit{autoriai}; \tag{44}$$

2) modelis su sąveikomis (pagal AIC):

$$V1 \sim \text{G1} + \text{Įn1} + \text{K1} + \text{N1} + \text{V\_G1} + \text{V\_Įn1} + \text{V\_K1} + \text{Vt1} + \\ + \text{G2} + \text{Įn2} + \text{K2} + \textit{autoriai} + \text{G1:V\_G1} + \text{G1:Įn2} + \text{Įn1:Įn2} + \\ + \text{K1:Įn2} + \text{N1:V\_Įn1} + \text{N1:G2} + \text{N1:}\textit{autoriai} + \text{V\_K1:Įn2} + \\ + \text{Vt1:G2} + \text{Vt1:K2} + \text{G2:}\textit{autoriai} + \text{Įn2:K2}; \tag{45}$$

3) lakoniškasis modelis be sąveikų:

$$V1 \sim \text{G1} + \text{Įn1} + \text{K1} + \text{N1} + \text{Vt1} + \textit{autoriai}; \tag{46}$$

4) lakoniškasis modelis su sąveikomis:

$$V1 \sim \text{G1} + \text{K1} + \text{N1} + \text{Vt1} + \text{G2} + \textit{autoriai} + \text{G1:V\_G1} + \\ + \ \text{Įn1:Įn2} + \text{K1:Įn2} + \text{N1:}\textit{autoriai} + \text{V\_K1:Įn2} + \text{G2:}\textit{autoriai}; \tag{47}$$

Dažnių kintamiesiems sudaryti modeliai yra labai sunkiai interpretuojami, nes jie sudaryti iš didelio kiekio kintamųjų ir jų sąveikų. Šie modeliai yra daug sudėtingesni nei logistinės regresijos modelių atveju. Viena iš priežasčių gali būti, kad logistinės regresijos atveju prarastai informacijai aprašyti reikia papildomų kintamųjų. Tai reikštų, kad modeliuose atsiradę nauji kintamieji atneša informaciją apie priklausomo kintamojo kiekį blokuose. Taip pat, Puasono regresijos modeliuose pasirodė sąveikos tarp funkcinių žodžių grupių ir kintamojo „*autoriai*“, kurios išlieka net lakoniškuosiuose modeliuose. Kaip ir logistinės regresijos atveju, taip ir šiuose modeliuose vienas iš pagrindinių kintamųjų yra „*autoriai*“.

Šių modelių suminės prognozės ir jų charakteristikos kiekvienai testavimo duomenų daliai pateiktos 31 lentelėje.

31 lentelė. Puasono regresijos modelio V1 grupei prognozių charakteristikos testavimo duomenims

Dalies nr.	1	2	3	4	5
Tikra suma	1857	1887	1890	1828	1712
1 modelio prognozuojama suma	1869,605	1891,149	1934,868	1885,855	1802,67
Skirtumas 1	-12,605	-4,149	-44,868	-57,855	-90,67
Santykine paklaida 1	0,0068	0,0022	0,0237	0,0316	0,0530
2 modelio prognozuojama suma	1883,244	1898,153	1923,37	1881,927	1774,476
Skirtumas 2	-26,244	-11,153	-33,37	-53,927	-62,476
Santykine paklaida 2	0,0141	0,0059	0,0177	0,0295	0,0365
3 modelio prognozuojama suma	1864,52	1884,585	1932,773	1886,831	1801,015
Skirtumas 3	-7,52	2,415	-42,773	-58,831	-89,015
Santykine paklaida 3	0,0040	0,0013	0,0226	0,0322	0,0520
4 modelio prognozuojama suma	1874,084	1895,626	1923,266	1877,957	1770,935
Skirtumas 4	-17,084	-8,626	-33,266	-49,957	-58,935
Santykine paklaida 4	0,0092	0,0046	0,0176	0,0273	0,0344

Dalies nr.	6	7	8	9	10
Tikra suma	1751	1814	1751	1792	1855
1 modelio prognozuojama suma	1839,142	1855,576	1752,337	1790,732	1864,942
Skirtumas 1	-88,142	-41,576	-1,337	1,268	-9,942
Santykine paklaida 1	0,0503	0,0229	0,0008	0,0007	0,0054
2 modelio prognozuojama suma	1836,998	1856,268	1742,848	1778,534	1863,277
Skirtumas 2	-85,998	-42,268	8,152	13,466	-8,277
Santykine paklaida 2	0,0491	0,0233	0,0047	0,0075	0,0045
3 modelio prognozuojama suma	1839,745	1851,968	1752,272	1789,569	1866,007
Skirtumas 3	-88,745	-37,968	-1,272	2,431	-11,007
Santykine paklaida 3	0,0507	0,0209	0,0007	0,0014	0,0059
4 modelio prognozuojama suma	1837,398	1856,554	1744,753	1775,95	1860,293
Skirtumas 4	-86,398	-42,554	6,247	16,05	-5,293
Santykine paklaida 4	0,0493	0,0235	0,0036	0,0090	0,0029

Remiantis santykinėmis paklaidomis geriausiai prognozavo ketvirtasis modelis, kurio vidutinė santykinė paklaida buvo 0,01813, o maksimali - 0,0493. Skirtumai tarp prognozuojamų ir tikrųjų sumų keliose dalyse yra gana dideli, tačiau įvertinus tikrųjų reikšmių išsibarstymą toks rezultatas natūralus. Testavimo duomenyse tikrosios sumos svyravo nuo 1712 iki 1890, t.y. per 178 vienetus, o didžiausias skirtumas buvo tik apie 86 vienetus.

Puasono regresijos modeliams taip pat buvo apskaičiuotos simetrizuotos χ^2 statistikos. χ^2 statistikos prognozių atžvilgiu: $\chi^2_1 = 1.009732$, $\chi^2_2 = 1.295286$, $\chi^2_3 = 1.260754$, $\chi^2_4 = 0.944646$. Remiantis šiomis statistikomis galime teigti, kad geriausias modelis yra ketvirtasis, nes jo χ^2 reikšmė yra mažiausia.

Kad įvertinti sudarytų modelių naudą, reikia pasirinkti geriausią taisyklę, kuri nesiremtų jokia papildoma informacija tik duomenimis apie patį tiriamąjį kintamąjį. Tuomet modelių sudarytą taisyklę, kuri nenaudoja jokios informacijos apie patį priklausomąjį kintamąjį, o tik aiškinamųjų kintamųjų informaciją, galima palyginti su geriausia taisykle. Vienas iš tokių tradicinių geriausių taisyklių yra vidurkis, kuris gali būti įvertinamas iš duomenų. Jeigu iš anksto būtų žinomas testavimo duomenų vidurkis ir nebūtų jokios kitos papildomos informacijos, prognozuodami pagal šį vidurkį, prognozių χ^2 būtų lygi 1,865087893. Akivaizdu, jog bet kuris šiame tyrime sudarytas Puasono regresijos modelis prognozuoja geriau už testinių duomenų vidurkį, todėl galima daryti prielaidą, kad aiškinamųjų kintamųjų turimos informacijos panaudojimas leidžia prognozuoti tokiu tikslumu, kurio negalima pasiekti remiantis vien informacija apie priklausomą kintamąjį.

4.4.2. G1 grupė

$G1$ grupei sudaryti keturi Puasono regresijos modeliai:

1) modelis be sąveikų (pagal AIC):

$$G1 \sim K1 + N1 + V\_G1 + V\_K1 + V1 + Vt1 + G2 + K2 + autoriai; \quad (48)$$

2) modelis su sąveikomis (pagal AIC):

$$\begin{aligned} G1 \sim {} & \text{Įn1} + \text{K1} + \text{N1} + \text{V\_G1} + \text{V\_K1} + \text{V1} + \text{Vt1} + \text{G2} + \text{Įn2} + \\ & + \text{K2} + autoriai + \text{Įn1:K1} + \text{Įn1:N1} + \text{Įn1:G2} + \text{K1:Įn2} + \text{N1:V1} + \\ & + \text{V\_G1:K2} + \text{V1:G2} + \text{Vt1:K2} + \text{G2:K2} + \text{Įn1:V\_G1}; \end{aligned} \quad (49)$$

3) lakoniškasis modelis be sąveikų:

$$G1 \sim K1 + N1 + V\_G1 + V\_K1 + V1 + G2 + autoriai; \quad (50)$$

4) lakoniškasis modelis su sąveikomis:

$$\begin{aligned} G1 \sim {} & \text{K1} + \text{N1} + \text{V\_K1} + \text{V1} + \text{Vt1} + \text{G2} + autoriai + \\ & + \text{N1:V1} + \text{V\_G1:K2} + \text{V1:G2} + \text{Vt1:K2} + \text{G2:K2}; \end{aligned} \quad (51)$$

32 lentelė. Puasono regresijos modelio G1 grupei prognozių charakteristikos testavimo duomenims

Dalies nr.	1	2	3	4	5
Tikra suma	784	812	814	745	736
1 modelio prognozuojama suma	773,1374	766,9636	812,0775	765,9409	762,4604
Skirtumas 1	10,8626	45,0364	1,9225	-20,9409	-26.4604
Santykinė paklaida 1	0,0139	0,0555	0,0024	0,0281	0,0360
2 modelio prognozuojama suma	777,9538	768,9525	820,1086	770,8597	751,6783
Skirtumas 2	6,0462	43,0475	-6,1086	-25,8597	-15,6783
Santykinė paklaida 2	0,0077	0,0530	0,0075	0,0347	0,0213
3 modelio prognozuojama suma	772,1357	769,2811	811,9759	767,365	762,228
Skirtumas 3	11,8643	42,7189	2,0241	-22,365	-26,228
Santykinė paklaida 3	0,0151	0,0526	0,0025	0,03002	0,0356
4 modelio prognozuojama suma	775,3429	768,1607	812,4723	770,1178	744,0659
Skirtumas 4	8,6571	43,8393	1,5277	-25,1178	-8,0659
Santykinė paklaida 4	0,0110	0,0540	0,0019	0,0337	0,0110
Dalies nr.	6	7	8	9	10
Tikra suma	748	771	702	764	709
1 modelio prognozuojama suma	735,9643	778,7939	710,6225	741,8234	796,8149
Skirtumas 1	12,0357	-7,7939	-8,6225	22,1766	-87,8149
Santykinė paklaida 1	0,0161	0,0101	0,0123	0,0290	0,1239
2 modelio prognozuojama suma	730,325	772,4708	710,7488	744,1461	786,8719
Skirtumas 2	17,675	-1,4708	-8,7488	19,8539	-77,8719
Santykinė paklaida 2	0,0236	0,0019	0,0125	0,0260	0,1098
3 modelio prognozuojama suma	737,1919	779,0585	711,0603	742,5485	797,1214
Skirtumas 3	10,8081	-8,0585	-9,0603	21,4515	-88,1214
Santykinė paklaida 3	0,0144	0,0105	0,0129	0,0281	0,1243
4 modelio prognozuojama suma	730,5907	770,6247	709,8258	742,299	788,9818
Skirtumas 4	17,4093	0,3753	-7,8258	21,701	-79,9818
Santykinė paklaida 4	0,0233	0,0005	0,0111	0,0284	0,1128

Puasono regresijos modeliai sudaryti $G1$ grupei, vėl gi yra labai sudėtingi ir sunkiai interpretuojami.Taip pat kiekviename modelyje yra kintamasis „*autoriai*“, tačiau nėra sąveikų tarp šio ir likusių kintamųjų. Remiantis santykine paklaida (32 lentelė), tiksliausias modelis yra ketvirtasis. Šio modelio prognozių vidutinė santykinė paklaida yra 0.028770542, o maksimali santykinė paklaida - 0.112809309.

$G1$ grupės Puasono regresijos modeliams taip pat buvo apskaičiuotos simetrizuotos χ^2 statistikos. χ^2 statistikos prognozių atžvilgiu: $\chi^2_1 = 1.551450383$, $\chi^2_2 = 1.281373955$, $\chi^2_3 = 1.535378319$, $\chi^2_4 = 1.311536771$. Remiantis šiomis statistikomis galime teigti, kad geriausias modelis yra antrasis, nes jo χ^2 reikšmė yra mažiausia. Jeigu būtų prognozuojama pagal testinių duomenų vidurkį, nanaudojant jokios papildomos informacijos χ^2 statistikos reikšmė būtų $\chi^2 = 1.742691376$. Ir šiuo atveju, bet kuris iš keturių sudarytų modelių yra tinkamesnis prognozavimui nei tikrasis vidurkis.

4.4.3. Įn1 grupė

Į$n1$ grupei buvo sudaryti keturi Puasono regresijos modeliai:

1) modelis be sąveikų (pagal AIC):

$$\text{Įn1} \sim \text{K1} + \text{N1} + \text{V\_G1} + \text{V1} + \text{G2} + \text{Įn2} + \textit{autoriai}; \tag{52}$$

2) modelis su sąveikomis (pagal AIC):

$$\begin{aligned}\text{Įn1} \sim\ & \text{G1} + \text{K1} + \text{N1} + \text{V\_G1} + \text{V\_Įn1} + \text{V\_K1} + \text{V1} + \text{Vt1} + \\ & + \text{G2} + \text{Įn2} + \text{K2} + \textit{autoriai} + \text{G1:K1} + \text{G1:N1} + \text{N1:V\_G1} + \\ & + \text{V\_G1:V\_Įn1} + \text{V\_G1:V\_K1} + \text{V\_Įn1:V\_K1} + \text{V1:Vt1} + \\ & + \text{Vt1:G2} + \text{Vt1:Įn2} + \text{G2:Įn2} + \text{Įn2:K2};\end{aligned} \tag{53}$$

3) lakoniškasis modelis be sąveikų:

$$\text{Įn1} \sim \text{K1} + \text{V\_G1} + \text{V1} + \text{G2} + \text{Įn2} + \textit{autoriai}; \tag{54}$$

4) lakoniškasis modelis su sąveikomis:

$$\begin{aligned}\text{Įn1} \sim\ & \text{K1} + \text{V1} + \text{G2} + \text{Įn2} + \textit{autoriai} + \text{G1:K1} + \text{G1:N1} + \\ & + \text{V\_G1:V\_K1} + \text{Įn2:K2};\end{aligned} \tag{55}$$

33 lentelė. Puasono regresijos modelio Įn1 grupei prognozių charakteristikos testavimo duomenims

Dalies nr.	1	2	3	4	5
Tikra suma	151	135	142	158	137
1 modelio prognozuojama suma	139,4674	139,8939	150,8376	144,3729	138,4417
Skirtumas 1	11,5326	-4,8939	-8,8376	13,6271	-1,4417
Santykine paklaida 1	0,0764	0,0363	0,0622	0,0862	0,0105
2 modelio prognozuojama suma	141,6874	139,2878	153,3936	146,0632	140,527
Skirtumas 2	9,3126	-4,2878	-11,3936	11,9368	-3,527
Santykine paklaida 2	0,0617	0,0318	0,0802	0,0755	0,0257
3 modelio prognozuojama suma	139,1446	139,8031	150,754	144,354	136,0549
Skirtumas 3	11,8554	-4,8031	-8,754	13,646	0,9451
Santykine paklaida 3	0,0785	0,0356	0,0616	0,0864	0,0069
4 modelio prognozuojama suma	141,9612	140,4567	152,3116	146,6743	138,2521
Skirtumas 4	9,0388	-5,4567	-10,3116	11,3257	-1,2521
Santykine paklaida 4	0,0599	0,0404	0,0726	0,0717	0,0091

Dalies nr.	6	7	8	9	10
Tikra suma	140	124	123	131	145
1 modelio prognozuojama suma	127,8045	141,7957	126,9018	137,002	140,168
Skirtumas 1	12,1955	-17,7957	-3,9018	-6,002	4,832
Santykine paklaida 1	0,0871	0,1435	0,0317	0,0458	0,0333
2 modelio prognozuojama suma	145,2111	144,4746	127,3062	138,8709	139,5006
Skirtumas 2	-5,2111	-20,4746	-4,3062	-7,8709	5,4994
Santykine paklaida 2	0,0372	0,1651	0,03501	0,0601	0,0379
3 modelio prognozuojama suma	127,805	141,6168	126,1357	136,783	140,3109
Skirtumas 3	12,195	-17,6168	-3,1357	-5,783	4,6891
Santykine paklaida 3	0,0871	0,1421	0,0255	0,0441	0,0323
4 modelio prognozuojama suma	128,9107	144,5221	127,2211	136,9394	139,6817
Skirtumas 4	11,0893	-20,5221	-4,2211	-5,9394	5,3183
Santykine paklaida 4	0,0792	0,1655	0,0343	0,0453	0,0367

Lakoniškasis modelis be sąveikų skiriasi nuo modelio be sąveikų, sudaryto remiantis AIC kriterijumi, tik $N1$ kintamuoju. Bet lakoniškasis modelis su sąveikomis yra daug paprastesnis už modelį su sąveikomis (pagal AIC). Į antrąjį modelį buvo įtraukti kintamieji $G1$, $G1 : K1$ ir $G1 : N1$, o ketvirtajame modelyje liko tik $G1 : K1$ ir $G1 : N1$. Galima daryti prielaidą, kad kintajam $Įn1$ $G1$ įtaka statistiškai reikšmingai pasireiškia per jo sąveikas su $K1$ ir $N1$ kintamaisiais, t.y. kintamasis $G1$ daro poveikį $Įn1$ grupei tik dalyvaudamas kartu su $K1$ ir $N1$ grupėmis. Taip pat visuose modeliuose buvo įtrauktas kintamasis „*autoriai*“.

Lentelėje 33 pateikiama kievieno iš keturių modelių prognozavimo charakteristikos. Pagal lentelės duomenis mažiausia padaroma vidutinė santykinė paklaida lygi 0.060015988 yra trečiojo modelio, kurio prognozių didžiausia santykinė paklaida yra 0.142070968.

$Įn1$ grupės Puasono regresijos modeliams taip pat buvo apskaičiuotos simetrizuotos χ^2 statistikos. χ^2 statistikos prognozių atžvilgiu: $\chi^2_1 = 0.691490499$, $\chi^2_2 = 0.67656672$, $\chi^2_3 = 0.682888296$, $\chi^2_4 = 0.700580653$. Remiantis šiomis statistikomis galime teigti, kad geriausias modelis yra antrasis, nes jo χ^2 reikšmė yra mažiausia. Jeigu būtų prognozuojama pagal testinių duomenų vidurkį, nanaudojant jokios papildomos informacijos χ^2 statistikos reikšmė būtų $\chi^2 = 0.801083927$. Ir šiuo atveju, bet kuris iš keturių sudarytų modelių yra tinkamesnis prognozavimui nei tikrasis vidurkis.

Antrojo bandomojo tyrimo metu taip pat buvo sudarinėjami ir neigiamos binominės regresijos modeliai toms žymeklių grupėms, kuriose buvo padidinta dispersija. Tačiau tiek jų struktūra, tiek ir prognozavimo rezultatai buvo labai panašūs, dėl šios priežasties jie šiame darbe nėra plačiau aprašomi.

4.4.4. Antrojo bandomojo tyrimo išvados

Antrojo bandomojo tyrimo metu buvo pastebėta ir išsiaiškinta, kad prognozuojant bet kurį linksnį, tyrime sudaryti modeliai, naudojantys tik informaciją apie aiškinamuosius kintamuosius, sugeba tiksliau prognozuoti nei geriausia taisyklė, kuri naudoja informaciją apie priklausomąjį kintamąjį.

Taip pat į visus modelius buvo įtrauktas kintamasis „*autoriai*“ ir šis faktas dar kartą įrodo, kad linksnių vartojimas kūriniuose labai priklauso nuo autorių polinkių ir stiliaus.

Pastebėta, kad dažnas ir labai dažnas grupes Puasono regresijos modeliai prognozuoja geriau nei labai retas grupes.

4.5. Trečiasis bandomasis tyrimas

Šiame tyrime priklausomu kintamuoju laikomas sakinių skaičius bloke, kuriuose buvo žymeklių iš pasirinktos grupės, o aiškinamaisiais kintamaisiais - likusių grupių žymeklių skaičius bloke (dažnio kintamieji). Tyrimo tikslas prognozuoti keliuose sakiniuose bloke

pasirodo žymekliai iš pasirinktos grupės.

Šio tyrimo etapai:

1) suskaičiuojamos priklausomojo ir aiškinamųjų kintamųjų reikšmės.

2) Sudaromi Puasono ir binominės regresijos modeliai, naudojant apmokymo duomenis.

3) Naudojant apmokymo duomenis yra prognozuojamos priklausomojo kintamojo reikšmės.

4) Atliekamas prognozavimas naudojant testinius duomenis.

5) Remiantis gautais rezultatai, nustatomas modelio tinkamumas prognozėms.

Šiame bandomajame tyrime taip pat buvo sudaromi keturių tipų modeliai, kurie buvo aprašyti 4.3. poskyryje. Atlikto trečiojo bandomojo tyrimo su trečiojo tipo kintamaisiais rezultatai buvo labai panašūs į pirmojo ir antrojo bandomųjų tyrimų rezultatus, todėl jų aprašymo šiame baigiamojo magistro darbe nebus pateikta.

5. IŠVADOS IR REKOMENDACIJOS

Baigiamojo magistro darbe atlikus mokslinių tyrimų apžvalgą pastebėta, kad lietuvių kalba, o ypač jos linksniuojamos kalbos dalys, yra mažai išnagrinėta statistinių metodų pagrindu. Dėl tyrimų trūkumo yra reikalingi įvairūs bandomieji (pasiruošiamieji) tyrimai, kurie ir buvo atlikti šiame darbe.

Atlikus dviejų kalbos dalių, įvardžių ir prielinksnių, pirminę statistinę ir koreliacinę analizes buvo pastebėtos tam tikros linksnių vartojimo tendencijos priklausančios nuo autorių ir kalbos dalies.

Sudaryti logistinės regresijos modeliai parodė, kad remiantis funkcinių žodžių pasisikirstymu galima pakankamai tiksliai prognozuoti priklausomojo kintamojo (pasirinktos grupės) pasirodymą tekste (indikatorių). Apmokymo duomenims sudaryti modeliai panašiai tiksliai prognozavo ir testinius duomenis. Taip pat, buvo pastebėta, kad modelio pasirinkimas priklauso nuo siekiamų tikslų. Jei tikslas yra prognozuoti bendras (sumines, vidurkines ar pan.) charakteristikas, tuomet verta rinktis lakoniškuosius modelius. Šie modeliai sugebėjo pakankamai tiksliai (su mažesne nei 3% vidutine paklaida) atspėti blokų, kuriuose pasirodė žymekliai, sumas prognozavimo duomenų dalyse. Tačiau jei prognozavimo tikslas yra atspėti žymeklių pasirodymą atskiruose blokuose, tuomet vertėtų rinktis išsamesnius modelius, į kuriuos yra įtraukta daugiau kintamųjų bei jų saveikos. Taip pat, buvo pastebėta, kad logistinės regresijos modeliai taikomi tekstiniams duomenims geriau prognozuoja vidutinio dažnumo grupes, labai retų ar labai dažnų grupių pasirodymų prognozės buvo ne tokios tikslios, tačiau vidutinė santykinė paklaida vis tiek neviršijo 3%.

Tyrimų metu pavyko gana tiksliai prognozuoti ir grupių dažnius. Dažnių prognozavimui buvo naudojamos Puasono ir neigiamos binominės regresijų modeliai, kurie panašiai tiksliai prognozavo tiriamojo kintamojo vidurkį. Pastebėta, kad modeliai geriau prognozavo dažnas arba labai dažnas grupes nei labai retas. Antrojo bandomojo tyrimo metu sudaryti modeliai buvo lyginami su geriausia prognozavimo taisykle - priklausomojo kintamojo vidurkiu testavimo duomenyse. Buvo padaryta išvada, kad be aiškinamųjų kintamųjų (funkcinių žodžių grupių) tokio tikslumo nebūtų įmanoma pasiekti, nes visais atvejais sudaryti modeliai prognozavo geriau už geriausią taisyklę.

Trečiojo bandomojo tyrimo, kurio metu buvo prognozuojama keliuose bloko sakiniuose pasirodė žymekliai iš pasirinktos grupės, rezultatai stipriai nesiskyrė nuo pirmųjų dviejų bandomųjų tyrimų rezultatų.

Atlikus visus tyrimus buvo padaryta išvada, jog linksnių vartojimas labai smarkiai priklauso nuo autoriaus polinkių, kadangi į kiekvieną sudarytą modelį buvo įtrauktas kintamasis „*autoriai*“. Tačiau tik $V1$ grupės modeliuose pasitaikė sąveikos tarp aiškinamųjų kintamųjų ir autorių.

Šiame darbe parinktus modelius, ypač su sąveikomis, interpretuoti sudėtinga. Jų

sudėtingumas dalinai gali būti paaiškintas gana siauru naudojamų žymeklių rinkiniu, susietu tik su dviejų tipų funkciniais žodžiais. Tų funkcinių žodžių vartojimo ypatybės gali būti viena iš sudėtingos parinktų modelių struktūros priežasčių. Todėl atliekant tolimesnius tyrimus reikėtų ženkliai praplėsti naudojamų žymeklių rinkinį. Tai galbūt leistų ne tik supaprastinti modelių struktūras, bet ir pasiekti didesnį prognozavimo tikslumą. Taip pat reikėtų į modelius įtraukti kintamuosius, kurie turėtų informacijos apie patį tekstą, pvz.: kintamasis, nurodantis blokų dydį - žodžių skaičių bloke, arba kūrinio žanrą nurodantis kintamasis.

LITERATŪRA

[1] Abney, S. Statistical methods and linguistics. In: The Balancing Act: Combining Symbolic and Statistical Approaches to Language, MIT Press, Cambridge, 1996, pp. 1–26.

[2] Agresti, A. Categorical Data Analysis, second edition, Canada, 2002.

[3] Aksomaitis, A. Tikimybių teorija ir statistika. Kaunas: Technologija, 2000.

[4] Baayen, R., H. Word Frequency Distributions, Kluwer Academic Publishers, 2001.

[5] Čekanavičius, V.; Murauskas, G. Statistika ir jos taikymai, knyga I. Vilnius: TEV, 2000.

[6] Čekanavičius, V.; Murauskas, G. Taikomoji regresinė analizė socialiniuose tyrimuose, Vilnius: Vilniaus universiteto leidykla, 2014.

[7] Dabartinės lietuvių kalbos tekstynas [žiūrėta 2014 m. balandžio 29 d.]. Prieiga per internetą:<`http://tekstynas.vdu.lt/tekstynas`>.

[8] Dobson, A., J. Introduction to generalized linear models, second edition, Boca Raton, London, New York, Washington A CRC Press Company 2001

[9] Evert, S. A simple LNRE model for random character sequences. In: Proceedings of the 7emes Journees Internationales d'Analyse Statistique des Donnees Textuelles, Louvain-la-Neuve, Belgium, 2004, pp. 411—422.

[10] Kazlauskienė, A.; Rimkutė, E.; Utka, A. Kiekybiniai tyrimai kalbotyroje [žiūrėta 2014 m. balandžio 29 d.]. Prieiga per internetą:<`http://donelaitis.vdu.lt/lkk/pdf/I_dalis.pdf`>.

[11] Keinys, S.; Bilkis, L.; Paulauskas, J.; Vitkauskas, V. Dabartinės lietuvių kalbos žodynas: šeštas (trečias elektroninis) leidimas, Vilnius: Lietuvių kalbos institutas, 2011.[žiūrėta 2014 m. balandžio 29 d.]. Prieiga per internetą: <`http://dz.lki.lt/`>.

[12] Kornai, A. How many words are there?, Glottometrics, 4, 61-86, 2002.

[13] Madsen, R., E.; Kauchak, E.; Elkan, C. Modeling Word Burstiness using the Dirichlet distribution, In: Proceedings of the 22nd International Conference on Machine Learning (ICML), pp. 545–552, 2005.

[14] Mandelbrot, B. An informational theory of the structure of language based upon the theory of the statistical matching of messages and coding. In: Communication Theory, Acad. Press, London, 1953, pp. 503—513.

[15] McCullagh, P.; Nelder, J., A. Generalized linear models. 2nd ed., London, Chapman and Hall, 1989.

[16] Rimkutė, E.; Daudaravičius, V. Morfologinis dabartinės lietuvių kalbos tekstyno anotavimas. Kalbų studijos, 2007,11: 30–35. [žiūrėta 2014 m. balandžio 29 d.]. Prieiga per internetą: <http://donelaitis.vdu.lt/publications/Rimkute_2007.pdf>.

[17] Skaitmeninė biblioteka „Literatūros kūriniai 5–8 klasėms“ [žiūrėta 2014 m. balandžio 29 d.]. Prieiga per internetą: <http://mkp.emokykla.lt/ebiblioteka/>.

[18] Smith, N., A. Linguistic structure prediction, Morgan Claypool Publishers, 2011.

[19] Svetulevičienė, V. Matematinės statistikos elementai, I dalis, Vilnius, 1976.

[20] Tabata, T. 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.

[21] The R Project for Statistical Computing [žiūrėta 2014 m. balandžio 29 d.]. Prieiga per internetą: <http://www.r-project.org/>.

[22] Urbanavičius, A. Įvardis [žiūrėta 2014 m. balandžio 29 d.]. Prieiga per internetą: <http://ualgiman.dtiltas.lt/ivardis.html>.

[23] Urbanavičius, A. Prielinksnis [žiūrėta 2014 m. balandžio 29 d.]. Prieiga per internetą: <http://ualgiman.dtiltas.lt/prielinksnis.html>.

[24] Utka, A. Labai Dažnų Lietuvių Kalbos Žodžių ir Žodžio Formų Ypatybės. Kaunas: Vytauto Didžiojo universitetas, 2005.

[25] Venclovienė, J. Statistiniai metodai medicinoje. Kaunas: Vytauto Didžiojo universitetas, 2010.

[26] Zinkevičius, V. Lemuoklis – morfologinei analizei. Darbai ir dienos, 24: 245–273, 2000.

[27] Zipf, G., K. The psycho-biology of language; an introduction to dynamic philology, Houghton Mifflin, Boston, 1935.

A PRIEDAS

Dažnos žodžių formos ir linksnių vartojimas

Monika Lapėnaitė-Gedvilė
Matematinės statistikos katedra
VGTU
Vilnius, Lietuva
lapenaite.monika@gmail.com

Marijus Radavičius
Matematikos ir informatikos institutas, VU
Vilnius, Lietuva
marijus.radavicius@mii.vu.lt

Karolina Piaseckienė
Matematikos ir informatikos institutas, VU
Vilnius, Lietuva
karol@delfi.lt

***Santrauka* — Darbe nagrinėjami tokie klausimai: ar naudojantis statistine informacija apie dažnas žodžio formas galima prognozuoti žodžio formų su tam tikromis savybėmis pasitaikymo (lietuviškame) tekste dažnius, kokiu tikslumu, kaip tai priklauso nuo teksto autoriaus? Aptariami preliminarūs žodžio linksnio prognozavimo rezultatai. Skaičiavimai atlikti paketu R.**

Raktiniai žodžiai — apibendrintasis tiesinis modelis; funkciniai žodžiai; linksnis; logistinė regresija; natūraliosios kalbos apdorojimas; prognozė; Puasono regresija; ROC kreivė.

I. ĮVADAS

Visame pasaulyje šiuo metu viskas yra kompiuterizuojama, ne išimtis ir kalba. Yra kuriami įvairūs tekstų redaktoriai, vertėjai, elektroniniai žodynai, programos įvairiems žodžiams, kalboms atpažinti. Pažengus technologijoms jau yra norima ne tik atpažinti atskirus žodžius, bet ir sakinius, jų struktūrą.

Sprendžiant šiuos uždavinius (plačiai) taikomi duomenų analizės ir statistikos metodai [1, 2, 14].

Vienas iš statistinių metodų taikymo lingvistikoje (kalbotyroje) ypatumų yra susijęs su Zipfo-Mandelbroto dėsniu [2, 3, 10, 18], kuriuo nusakomas žodžių formų, išrikiuotų jų pasitaikymo tekstyne dažnumų mažėjimo (didėjimo) tvarka, pasiskirstymo dėsnis (skirstinys). Jis teigia, kad nedidelė žodžių formų dalis pasitaiko labai dažnai, o ženkli žodžių formų dalis, kartais virš 50%, tekstynuose yra pavartota tik kartą (šios formos vadinamos *hapax legomena*), o dar didesnė jų dalis yra iš viso nestebima. Formaliai iš šio dėsnio išplaukia, kad potencialusis (teorinis) kalbos žodynas yra begalinis [2, 8]. Informatikos ir kompiuterių moksle tai kartais įvardinama kaip protrūkis, „sprogimas“ (angl. burstiness, [9]). Taigi, aktualus uždavinys – remiantis dažniau pasitaikančiomis žodžio formomis ar kitais dariniais (toliau juos vadinsime markeriais), jų pasiskirstymu tekste, nuspėti kitų (retesnių) žodžio formų (ar kitų lingvistinių objektų) savybių pasiskirstymus. Markeriai turi būti lengvai identifikuojami ir veikiami tų pačių veiksnių, kurie yra susiję su tiriamomis savybėmis. Be to, tekstuose jie turi pasitaikyti pakankamai dažnai. Tinkamų markerių paieška yra sudėtingas uždavinys, ir kol jis neišspręstas, net neaišku, ar jie egzistuoja.

Nors šiame darbe yra sprendžiamas prognozavimo uždavinys, kuris yra (labiau) būdingas Natūraliosios Kalbos Apdorojimo (NPA, angl. Natural Language Processing, NLP) metodologijai, jis turi sąsajų su lingvistikos mokslo problemomis.

Jau seniai pastebėta, kad dažniausi kalbos žodžiai arba jų formos turi savybių, nebūdingų retesniems žodžiams, ir paprastai yra funkciniai žodžiai, t.y., atlieka tam tikras funkcijas [16]. Kai kuriuose anglų kalbos tekstų klasifikavimo darbuose (pvz., [15]), kurių uždavinys – nustatyti grožinės literatūros tekstų žanrą arba autorystę, tvirtinama, kad labai dažni žodžiai ir žodžių formos yra geri tekstų klasifikavimo požymiai.

Šiame darbe akcentuojama kita dažnų žodžio formų ypatybė: kokiu laipsniu jos reprezentuoja nepriklausomas nuo autorių, kitaip tariant, būdingas pačiai kalbai lietuviškų tekstų savybes, sąryšius tarp įvairių lingvistinių objektų.

Pasirinkta tyrimui savybė yra linksnis. Linksnio kategorija, jo svarba, įvairūs jo vartojimo ir įsisavinimo aspektai yra aptarti Savickienės straipsnyje [13]. Lietuvių kalboje žodžių tvarka

sakinyje yra gana laisva, nėra apibrėžtų griežtų sakinio struktūros taisyklių. Žodžių linksniai yra viena iš kalbos savybių, kuri galėtų padėti nuspėti nežinomą sakinio struktūrą. Tai specifinė savybė, kurios kartais negalime vienareikšmiškai nusakyti. Tarkime žodis „jos“ – šiuo atveju nežinant konteksto negalima pasakyti, ar tai įvardis ir kokio – vardininko ar kilmininko – linksnio, ar veiksmažodžio „joti“ būsimasis laikas. Tai vadinama morfologiniu daugiareikšmiškumu [12]. Lietuvių kalbai yra sukurta automatinė morfologinės analizės programa Lemuoklis [17], žodžio formai pateikianti antraštinį pavidalą (lemą) ir gramatines pažymas. Ji skirta Vytauto Didžiojo universiteto Kompiuterinės lingvistikos centro (toliau KLC) sukurto tekstyno moksliniams lingvistiniams tyrinėjimams automatizuoti. Bet net ir naudojant šį įrankį didesnio teksto morfologinis anotavimas, o tuo pačiu ir linksnių nustatymas, reikalauja daug darbo sąnaudų. Tačiau, jeigu uždavinys yra (apytiksliai) įvertinti įvairių linksnių ir jų kombinacijų pasitaikymo tekste dažnius, tai galbūt nėra būtina nustatyti visų linksniuojamų žodžių linksnius, nes pakankamai tiksliai (statistinių paklaidų tikslumu) tuos dažnius pavyks prognozuoti pasirinktų markerių pagrindu.

Šiai hipotezei patikrinti šiame darbe atliktas žvalgybinis tyrimas. Jis remiasi tokiu samprotavimu. Jeigu dviejų grupių markeriai turi daugmaž vienodą ir reikšmingą informaciją apie tam tikro linksnio vartojimą tekste, tai tikėtina, kad vienos grupės markerių savybės, susijusios su tuo linksniu, bus pakankamai gerai prognozuojamos per atitinkamas kitos markerių grupės savybes.

Paprastai empiriniai-statistiniai tyrimai lingvistikoje remiasi tekstynų duomenimis. Dabartinę (rašytinę) lietuvių kalbą, įvairius jos stilius, anot kūrėjų, reprezentatyviai atspindi VDU 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). Tačiau jis nėra pritaikytas nuodugnesnei statistinei analizei, kai imtį reikia sudaryti pagal tyrimui svarbius tekstų požymius, visų pirma, tekstų autorius, bei kitas svarbias charakteristikas: tekstų sukūrimo datas, įvairius stilius, žanrus, ar pan. Todėl tyrimui buvo paimta 30 panašaus žanro lietuviškų kūrinių, suskaitmenintų vykdant ES struktūrinių fondų remiamą projektą „Pagrindinio ugdymo pirmojo koncentro (5–8 kl.) mokinių esminių kompetencijų ugdymas“, 2012 [4]. Duomenų apdorojimui ir analizei buvo naudojamas paketas R [7].

Kitame skyrelyje trumpai aprašyta tyrimo metodika, 3 skyrelyje pateikti naudoti statistiniai modeliai, 4 skyrelyje aptarti rezultatai ir 5 skyrelyje suformuluotos išvados.

II. TYRIMO METODIKA

Žemiau pateiktos metodikos pagrindą sudaro 1 ir 2 žingsniai, likę žingsniai yra standartiniai.

1. Tiriamai savybei (požymiui) modeliuoti sudaromas markerių rinkinys. Markeriai – tai dažnai (dažniau) tekste pasitaikančios žodžių formos (ar kiti lingvistiniai objektai), glaudžiai, kaip manoma, susijusios su tiriamu požymiu.

2. Visi tekstai skaidomi į sakinių blokus po 10 sakinių kiekviename bloke (iš viso 12167 blokai). Blokų dydis parinktas derinant priešingus siekius: maksimaliai atsižvelgti į galimą (tikėtiną) tekstų nehomogeniškumą, bet tuo pačiu nagrinėti pakankamai didelius tekstų blokus, kad išryškėtų statistiniai dėsningumai. Tikimasi, kad blokų dydis iš 10 sakinių (pasirinktam markerių rinkiniui) yra būtent toks.

3. Tekstų blokai atsitiktinai su vienodom tikimybėm priskiriami vienai iš 2-jų grupių. Pirmoji yra naudojama statistinio modelio parinkimui (apmokymui), o antroji yra skirta prognozavimo tikslumui įvertinti (testavimui).

4. Naudojantis kiekvieno bloko markerių statistika sukuriami aiškinantieji kintamieji statistiniam modeliui sudaryti.

5. Naudojant apmokymo duomenis tiriamam požymiui parenkamas statistinis modelis, sukonstruojama prognozavimo taisyklė ir taikant ją testiniams duomenims įvertinama prognozavimo paklaida.

Markerių parinkimas. Viena iš kalbos dalių, gana gerai atspindinti žodžių ir linksnių padėtį sakiniuose, yra įvardis. Tai yra kalbos dalis, kuri nurodo daiktą, ypatybę arba skaičių, bet jų nepavadina [5]. Įvardis gali pakeisti daiktavardžius, būdvardžius, ar net nurodyti kiekybę. Taigi, tikrai nemaža dalis žodžių gali būti pakeista įvardžiais, kurie dažnai pasirodo lietuvių kalbos tekstuose ir dauguma jų linksnių yra vienareikšmiškai nustatomi.

Dar viena kalbos dalis, padedanti nustatyti žodžio linksnį, yra prielinksniai – nekaitoma kalbos dalis, kuri eina su linksniu ir parodo linksniuojamojo žodžio ryšį su kitais žodžiais [6]. Yra nemažai prielinksnių, kurie eina tik kartu su tam tikru linksniu, pavyzdžiui, po prielinksnio „su“ visada eis žodis, esantis įnagininko linksnio.

Šiam tyrimui naudojamame duomenų rinkinyje dažniausiai pasirodančių žodžių 50-uke, pateko net 27 įvardžiai ir 10 prielinksnių. Taigi, tyrimui buvo atrinkti 87 įvardžiai ir 24 prielinksniai (iš viso 111 tipinių žodžių), kurie buvo apjungti į grupes pagal kalbos dalį ir linksnį. Sudaryta 12 grupių: *G*1, *Įn*1, *K*1, *N*1, *V_G*1, *V_Įn*1, *V_K*1, *V*1, *Vt*1, *G*2, *Įn*2, *K*2 (raidė nusako linksnį, o skaičiai – kalbos dalį: 1 – įvardžiai, 2 – prielinksniai). Į grupes *V_G*1, *V_Įn*1, *V_K*1 patenka tie įvardžiai, kurie atitinka porą linksnių, pvz.: įvardis „šios“ gali būti daugiskaitos vardininko arba vienaskaitos kilmininko linksnio, todėl šis įvardis įtraukiamas į grupę *V_K*1. Apskaičiuotas tipinių žodžių iš kiekvienos grupės pasirodymų skaičius kiekviename sakinių bloke. Kaip pavyzdį, pateiksime Vt1 grupės dažnių lentelę:

LENTELĖ I. VT1 GRUPĖS DAŽNIŲ LENTELĖ

Reikšmė	0	1	2	3	4
Dažnis	11158	893	102	12	2

čia: reikšmė reiškia, kiek kartų tipinis žodis iš šios grupės pasirodė sakinio bloke,

dažnis – keliuose blokuose buvo toks tipinių žodžių kiekis.

Matome, kad markeris iš Vt1 grupės pasitaikė gana retai; maždaug 10% blokų.

Aiškinantieji kintamieji. Naudojant surinktą statistiką apie markerius kiekviename bloke kiekvienai minėtai grupei buvo sudarytas kintamasis, rodantis, kiek kartų bloke pasitaikė markeris iš duotos grupės, ir kiekvienos grupės indikatorius, lygus atitinkamai 1 ir 0, kai grupės markeriai bloke pasitaikė ir kai nepasitaikė. Taip pat, buvo sukurtas kintamasis „*autoriai*“, nurodantis kūrinio autorių. Šis kintamasis leidžia nustatyti, ar statistiniai sąryšiai tarp tiriamų kintamųjų priklauso nuo autoriaus. Tokiu būdu galima atskirti modelio komponentes, priklausančias nuo autorių stiliaus, ir komponentes nuo jų nepriklausančias, vadinasi, galbūt atspindinčias pačiai lietuvių kalbai būdingas savybes. Kadangi tiriamas požymis ir pradiniai aiškinantieji kintamieji gali būti tarpusavyje susiję sudėtingais ryšiais, tai į parenkamą modelį be pradinių aiškinančiųjų kintamųjų buvo bandoma įtraukti ir jų sąveikas.

III. STATISTINIAI MODELIAI

Šiame darbe atliekant statistinę analizę taikoma logistinė ir Puasono regresija. Kadangi abu modeliai yra atskiri Apibendrintojo Tiesinio Modelio (ATM) atvejai, tai jo trumpą aprašymą ir pateiksime, nurodydami minėtų regresijos modelių ypatumus. Išsamų ATM modelio aprašymą galima rasti McCullagh ir Nelder monografijoje [11].

ATM apibrėžia 3 komponentės:

1) atsitiktinė komponentė, kuri nusakoma tyrimo kintamojo y skirstiniu iš eksponentinės skirstinių šeimos;

2) tiesinis prediktorius η, kuris aprašo bendrą aiškinamųjų kintamųjų $x_1, \dots, x_k$ įtakos tyrimo kintamojo y skirstiniui kiekybinę išraišką kaip tiesinę jų kombinaciją

$$\eta = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k$$

su nežinomais koeficientais β_i, kuriuos reikės įvertinti iš duomenų;

3) jungties funkcija g, kuri susieja y skirstinio vidurkį μ su tiesiniu prediktoriumi η,

$$g(\mu) = \eta.$$

Laikoma, kad turimuose duomenyse $\{(x_1(j),\dots,x_k(j),y(j)),\ j = 1,\dots,n\}$, stebiniai $y(j)$, $j = 1,\dots,n$, yra tarpusavyje sąlyginai nepriklausomi atsitiktiniai dydžiai, kai yra žinomos aiškinančiųjų kintamųjų reikšmės.

Nežinomų modelio parametrų $\beta_1,\dots,\beta_k$ įvertinimui paprastai taikomas didžiausio tikėtinumo metodas. Logistinėje regresijoje y įgyja tik 2 reikšmes, 0 ir 1, ir turi binominį skirstinį, jungties funkcija g yra logit funkcija

$$g(u) = ln\big(u/(1-u)\big).$$

Puasono regresijoje y įgyja sveikas neneigiamas reikšmes ir turi Puasono skirstinį, jungties funkcija

$$g(u) = ln(u).$$

Parenkant tinkamą modelį ir informatyviausius aiškinančiuosius kintamuosius vadovautasi deviacijos statistika ir Akaike informaciniu kriterijumi (AIC), bei į modelį įtrauktų kintamųjų statistiniu reikšmingumu. Logistinės regresijos modelyje y reikšmės prognozė, kai žinomas $x = a$, $\hat{y}(a) = 1$, jeigu $\hat{p}(a) > c$, ir $\hat{y}(a) = 0$ priešingu atveju. Čia $\hat{p}(a)$ yra pagal parinktą logistinės regresijos modelį įvertinta įvykio $y = 1$ sąlyginė tikimybė, kai $x = a$, o c yra parenkama kritinė reikšmė.

IV. Rezultatų apžvalga

Atliekant tyrimą prognozuojamas kintamasis y buvo apskaičiuojamas kaip vienos grupės, pasirinktos iš 12 grupių (žr. punktą Markerių parinkimas), markerių dažnis bloke arba jų pasirodymo bloke indikatorius, o likusių grupių analogiška statistika blokuose buvo (pradiniai) aiškinantieji kintamieji x. Kiekvienas blokas yra atskiras mokymo imties elementas, j yra jo numeris.

Aptarsime tik dviejų grupių, $G1$ ir Įn1, rezultatus. Jie gana gerai reprezentuoja kitus atvejus. Galininkas yra labai dažnai vartojamas linksnis. Įnagininkas vartojamas gana retai, virš 81% blokų neturėjo grupės Įn1 markerių. Remiantis AIC minėtiems kintamiesiems buvo parinkti modeliai su pradinių aiškinančiųjų kintamųjų sąveikomis (iki 2-os eilės) ir be jų, taip pat lakoniškieji šių modelių variantai, parinkti naudojant reikšmingumo lygmenį 0.01 (modelyje palikti tik nariai, kurių p-reikšmė < 0.01).

Modelius pateiksime simbolinėje formoje, kuri naudojama R pakete. Pvz., užrašas

$$y \sim x_1 + x_2{:}x_3,$$

kur x_i yra tolydieji arba binariniai kintamieji, reiškia, kad kintamajam y sudaryto modelio prediktorius yra

$$\eta = \beta_0 + \beta_1 x_1 + \beta_2 x_2 x_3,$$

o simbolinis užrašas $x_1 * x_2$ reškia tą patį kaip ir $x_1 + x_2 + x_1{:}x_2$.

A. Grupė G1, Puasono regresija

Modelis be sąveikų grupės $G1$ markerių dažniams:

$$G1 \sim K1 + N1 + V\_G1 + V\_K1 + V1 + Vt1 + G2 + K2 + autoriai$$

Prognozuojamas $G1$ grupės markerių dažnumas testiniams duomenims yra 7645 vietoje tikrojo dažnio 7585, santykinė paklaida yra 0,79%.

Tai, kad modelyje su sąveikomis nebuvo statistiškai reikšmingų kintamojo „*autoriai*“ sąveikų su kitais kintamaisiais, leidžia daryti prielaidą, kad sąryšiai tarp įvairių linksnių vartojimo yra gana universalūs, mažai priklauso nuo autoriaus preferencijų. Sąveikų įtraukimas į modelį pagerino parinkto modelio tinkamumo matus, taip pat ir prognozės tikslumą.

Lakoniškasis modelis su sąveikomis:

$$G1 \sim K1 + N1 + V\_K1 + V1 + Vt1 + G2 + autoriai + N1{:}V1 + V\_G1{:}K2 + V1{:}G2 + Vt1{:}K2 + G2{:}K2$$

Prognozuojamas dažnumas 7612, santykinė paklaida 0,36%. Ir šiuo atveju nėra statistiškai reikšmingų kintamojo „*autoriai*“ sąveikų su kitais kintamaisiais.

Logistinės regresijos rezultatai grupės $G1$ indikatoriui yra panašūs, bet kadangi galininkas yra gana dažnas linksnis, tai naudojant indikatorių

prarandama reikšminga informacija, ir todėl logistinės regresijos modeliai turi mažiau informatyvių kintamųjų ir jų sąveikų negu Puasono regresijos modeliai.

B. Grupė Įn1*, logistinė regresija*

Kadangi įnagininkas vartojamas gana retai, tai šiuo atveju prasminga vietoje Puasono naudoti logistinę regresiją. Lakoniškasis modelis be sąveikų grupės Įn1 indikatoriui:

$$\text{Įn1} \sim K1 + V1 + \text{Įn2} + autoriai.$$

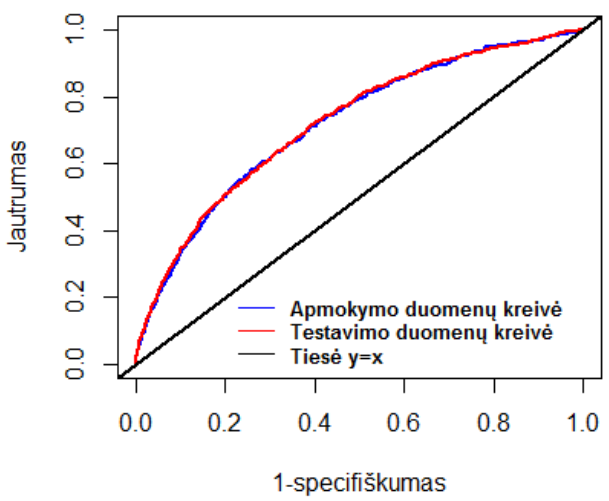
Prognozavimo testiniams duomenims charakteristikos: teisingai prognozuotų indikatoriaus reikšmių proporcija yra 77,4 %, jautrumas (teisingai prognozuotų reikšmių proporcija tarp blokų su grupės Įn1 markeriais) 39,8 %, specifiškumas (teisingai prognozuotų reikšmių proporcija tarp blokų be grupės Įn1 markerių) 86,1 %. Kritinė reikšmė $c = 0{,}272$ buvo parinkta pagal apmokymo duomenis minimizuojant klaidų kiekį.

Prognozavimo rezultatai:

LENTELĖ II. Įn1 PROGNOZAVIMO REZULTAI

c	TP	Sensitivity	TN	Specifity	Corect
0,272	460	0,3986	4246	0,8616	0,7738

ROC kreivė – grafikas, rodantis jautrumo ir specifiškumo sąryšį. Jautrumas aprašomas kaip funkcija atžvilgiu parametro 1-specifiškumas. Grafikas naudojama rezultatų tikslumui vertinti. Kuo plotas po kreive didesnis, tuo modelis tinkamesnis prognozei.



1 pav. ROC kreivės apmokymo ir testiniams duomenims

Natūralu, kad apmokymo duomenims plotas po ROC kreive truputį didesnis, nei testiniams duomenims, bet tas skirtumas toks mažas, jog galime teigti, kad parinktas modelis, pritaikytas testiniams duomenis, prognozuoja taip pat gerai, kaip ir apmokymo duomenims.

Lakoniškasis modelis su sąveikomis grupės Įn1 indikatoriui:

$$\text{Įn1} \sim K1 + \text{Įn2} + autoriai + G1{:}N1 + V\_G1{:}V\_K1 + V1{:}\text{Įn2} + \text{Įn2}{:}K2.$$

Modelių su sąveikomis prognozių charakteristikos testiniams duomenims labai panašios. Parinktas pagal AIC modelis be sąveikų prognozuoja truputį prasčiau. Vėlgi, ir grupei Įn1 pagal AIC parinktuose modeliuose nėra statistiškai reikšmingų kintamojo „*autoriai*“ sąveikų su kitais kintamaisiais.

Pastaba: Parinktus modelius, ypač su sąveikomis, interpretuoti sudėtinga. Jų sudėtingumas dalinai gali būti paaiškintas gana siauru naudojamų markerių rinkiniu, susietu tik su dviejų tipų funkciniais žodžiais (įvardžiais ir prielinksniais). Tų tipų funkcinių žodžių vartojimo ypatybės gali būti viena iš sudėtingos parinktų modelių struktūros priežasčių.

V. IŠVADOS

1. Preliminarus žvalgybinis tyrimas parodė, kad tinkamai parinkto markerių rinkinio ir statistinio modelio pagrindu galima gauti gana tikslius dominančios savybės (šiuo atveju – linksnių vartojimo dažnumo) pasiskirstymo įverčius.
2. Nors vieno ar kito linksnio vartojimo dažnis tarp autorių gali ženkliai skirtis, tačiau jo sąryšiai su markerių grupėmis, matyt, nuo autorių priklauso silpnai. Taigi, tie sąryšiai gali būti susieti su universaliomis lietuvių kalbos savybėmis.

3. Tai, kad buvo naudoti tik 2-jų tipų funkcinių žodžių markeriai, gali sąlygoti tiek parinkto prediktoriaus struktūrą, tiek ir jo pagrindu sukonstruotos prognozės ypatumus. Vertėtų ženkliai praplėsti markerių rinkinį kitų tipų funkciniais žodžiais arba kitomis dažnomis žodžių formomis ar struktūromis.

PADĖKA

Dėkojame Vilniaus Gedimino technikos universiteto dr. Tomui Rekašiui už vertingus patarimus ir pagalbą, dirbant su paketu R.

LITERATŪROS SĄRAŠAS

[1] Abney, S. Statistical methods and linguistics. In: The Balancing Act: Combining Symbolic and Statistical Approaches to Language, MIT Press, Cambridge, 1–23, 1996.

[2] Baayen, R. H. Word Frequency Distributions, Kluwer Academic Publishers, 2001.

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

[4] http://mkp.emokykla.lt/ebiblioteka/

[5] http://ualgiman.dtiltas.lt/ivardis.html

[6] http://ualgiman.dtiltas.lt/prielinksnis.html

[7] http://www.r-project.org/

[8] Kornai, A. How many words are there?, Glottometrics, 4, 61–86, 2002.

[9] Madsen, R.E., Kauchak, E., Elkan, C. Modeling Word Burstiness using the Dirichlet distribution, In: Proceedings of the 22nd International Conference on Machine Learning (ICML), 545–552, 2005.

[10] Mandelbrot, B. B. An information theory of the statistical structure of language. In: Communication Theory, W. Jackson, ed., London, 486–502, 1953.

[11] McCullagh, P., Nelder, J. A. Generalized linear models. 2nd ed., London, Chapman and Hall, 1989.

[12] Rimkutė, E., Daudaravičius, V. Morfologinis dabartinės lietuvių kalbos tekstyno anotavimas. Kalbų studijos, 30–35, 2007.11. http://donelaitis.vdu.lt/publications/Rimkute_2007.pdf.

[13] Savickienė, I. Linksnio kategorijos įsisavinimas: lietuvių kalba kaip gimtoji ir svetimoji ISSN 1392–1517. Kalbotyra. 122–129, 2006. 56(3).

[14] Smith, N. A. Linguistic structure prediction, Morgan & Claypool Publishers, 2011.

[15] Tabata, T. 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, 91–109, 1995.

[16] Utka, A. Labai dažnų lietuvių kalbos žodžių ir žodžio formų ypatybės. Vytauto Didžiojo universitetas, Kaunas, 2005.

[17] Zinkevičius, V. Lemuoklis – morfologinei analizei. Darbai ir dienos, 24: 245–273, 2000.

[18] Zipf, G. K. The psycho-biology of language; an introduction to dynamic philology, Houghton Mifflin, Boston, 1935.

B PRIEDAS

Nr.	Markeris	Grupė
1	aš	V1
2	manęs	K1
3	man	N1
4	mane	G1
5	manimi	Įn1
6	manyje	Vt1
7	tu	V1
8	tavęs	K1
9	tau	N1
10	tave	G1
11	tavimi	Įn1
12	tavyje	Vt1
13	mes	V1
14	mūsų	K1
15	mums	N1
16	mus	G1
17	mumis	Įn1
18	mumyse	Vt1
19	jūs	V1
20	jūsų	K1
21	jums	N1
22	jus	G1
23	jumis	Įn1
24	jumyse	Vt1
25	jis	V1
26	jo	K1
27	jam	N1
28	jį	G1
29	juo	Įn1
30	jame	Vt1
31	ji	V1
32	jos	V_K1
33	jai	N1
34	ją	G1
35	ja	Įn1
36	joje	Vt1
37	jie	V1

Nr.	Markeris	Grupė
38	jų	K1
39	jiems	N1
40	juos	G1
41	jais	Įn1
42	juose	Vt1
43	joms	N1
44	jas	G1
45	jomis	Įn1
46	jose	Vt1
47	tas	V_G1
48	to	K1
49	tam	N1
50	tą	G1
51	tuo	Įn1
52	tame	Vt1
53	ta	V_Įn1
54	tos	V_K1
55	tai	N1
56	toje	Vt1
57	šis	V1
58	šio	K1
59	šiam	N1
60	šį	G1
61	šiuo	Įn1
62	šiame	Vt1
63	ši	V1
64	šios	V_K1
65	šiai	N1
66	šią	G1
67	šia	Įn1
68	šioje	Vt1
69	tie	V1
70	tų	K1
71	tiems	N1
72	tuos	G1
73	tais	Įn1
74	tuose	Vt1

Nr.	Markeris	Grupė
75	toms	N1
76	tomis	Įn1
77	tose	Vt1
78	šie	V1
79	šių	K1
80	šiems	N1
81	šiuos	G1
82	šiais	Įn1
83	šiuose	Vt1
84	šioms	N1
85	šias	G1
86	šiomis	Įn1
87	šiose	Vt1
88	anapus	K2
89	anot	K2
90	antrapus	K2
91	arčiau	K2
92	dėl	K2
93	iš	K2
94	link	K2
95	linkui	K2
96	nuo	K2
97	pasak	K2
98	prie	K2
99	pusiau	K2
100	tarp	K2
101	žemiau	K2
102	apie	G2
103	į	G2
104	pagal	G2
105	palei	G2
106	pas	G2
107	per	G2
108	prieš	G2
109	pro	G2
110	su	Įn2
111	ties	Įn2

C PRIEDAS

G1

0	1	2	3	4	5	6	7	8	9	10	12
2212	1836	1095	529	225	107	43	23	7	4	3	1

Įn1

0	1	2	3	4	5	6
4948	928	182	17	8	1	1

K1

0	1	2	3	4	5	6	7	8	9	10	11
2364	1850	1006	494	198	106	37	17	9	1	2	1

N1

0	1	2	3	4	5	6	7	8	9	10	11	12	13	14
1481	1680	1255	816	438	207	112	50	21	8	10	3	2	1	1

V_G1

0	1	2	3	4	6
5236	739	88	19	2	1

V_Įn1

0	1	2	3	6
5806	256	20	2	1

V_K1

0	1	2	3	4	5	6
4667	1053	255	83	19	6	2

V1

0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
724	1127	1139	921	731	564	372	198	140	73	45	15	20	8	3	2
16	17	21													
1	1	1													

Vt1

0	1	2	3	4
5609	419	51	5	1

G2

0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1298	1545	1290	808	518	313	151	77	30	33	12	4	2	1	2	1

Įn2

0	1	2	3	4	5	6	7	9
3493	1644	624	210	75	20	14	4	1

K2

0	1	2	3	4	5	6	7	8	9	10	12	15	16	17
1494	1716	1247	785	450	203	99	38	32	11	6	1	1	1	1

Autoriai

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
638	124	371	30	232	472	348	22	505	80	369	24	500	236	388	327	226	212	203
20	21	22	23	24														
302	24	195	41	216														

D PRIEDAS

```
rm(list = ls())
setwd("C:/Users/Monika/Desktop/Straipsniui/30tekstu")
Sys.setlocale(locale = "Lithuanian")
txt.dir <- file.path(getwd())
txt.path <- list.files(txt.dir, full.names = TRUE)
######################################################################
######################NUSKAITAU TEKSTUS###############################
######################################################################
read.tekstas <- function(failas) {
tekstas <- scan(failas, what="char", sep=".", blank.lines.skip = TRUE, encoding="UTF-8",
    quiet = TRUE)
tekstas<- tekstas[-(1:20)]
rep_rez <- gsub("[!?]", ".", tekstas, perl=TRUE)
rep_rez <- gsub("[,:-]", "", rep_rez, perl=TRUE)
rep_rez <- gsub("[?*?]", "", rep_rez)
rep_rez <- gsub("[0-9]", "", rep_rez)
rep_rez <- gsub("[][]", "", rep_rez)
rep_rez<-tolower(rep_rez)
spl_rez<-unlist(strsplit(rep_rez, ".", fixed=T))
spl_rez<- gsub(" +", " ", spl_rez)
spl_rez<- gsub("?", "", spl_rez)
spl_rez<- gsub("\t", "", spl_rez)
spl_rez<- gsub("[);(]", "", spl_rez)
spl_rez<- gsub("[???]", "", spl_rez) #(lietuvi?kos kabut?s)
spl_rez<-spl_rez[lapply(spl_rez, nchar)>1]
return(spl_rez)
}
tekstai<- lapply(txt.path , read.tekstas)
l<-length(tekstai)
#############################################################################
######Autoriu priskyrimas#####################################################
#############################################################################
read.tekstas1 <- function(failas) {
tekstas1 <- scan(failas, what="char", sep=".", blank.lines.skip = TRUE, encoding="UTF-8",
     quiet = TRUE)
tekstas1 <-tekstas1[lapply(tekstas1, nchar)>8]
tekstas1<- gsub("? ", "", tekstas1)
tekstas1<- gsub("[\t]", "", tekstas1)
tekstas1<-tolower(tekstas1)
tekstas1<-tekstas1[1]
tekstas1<-unlist(strsplit(tekstas1, " ", fixed=T))
tekstas1<-tekstas1[tekstas1!=""]
tekstas1<-paste(tekstas1, collapse = " ")
return(tekstas1)
```

```
}
autoriai<-unlist(lapply(txt.path , read.tekstas1))
autkukodas<-c()
autk<-gsub("[^[:alpha:]]", "", autoriai)
autku<-unique(autk)
for (i in 1:length(autku)){
autkukodas[i]<-i }
autlent<-cbind(autku,autkukodas)
autoriukodas<-NULL
for (i in 1:length(autoriai)){
autoriukodas[i]<- autlent[,2][autlent[,1]==autk[i]]}
autoriulent<-cbind(autoriai, autk, autoriukodas)
#########################################################################
######################SUBLOKUOJU TEKSTUS PO 10 SAKINIU###################
#########################################################################
sakiniai<-NULL
bbb<-NULL
tekst<-NULL
tekstblok<-NULL
nr<-NULL
kurinionr<-NULL
for (j in 1:l) {
tekst<-unlist(tekstai[j])
bbb<-NULL
for (k in 1:round(length(tekstai[[j]])/9)){
aaa<-list(tekst[1:10])
bbb<-c(bbb, aaa)
tekst<-tekst[-(1:10)]
}
bbb<-lapply(bbb, function (x) x[!is.na(x)])
bbb<-bbb[lapply(bbb,length)==10]
tekstblok<-c(tekstblok, length(bbb))
sakiniai<-c(sakiniai,bbb)
nr<-rep(j,length(bbb))
kurinionr<-c(kurinionr,nr)
}
sakiniailent<-as.data.frame(unlist(sakiniai))
setwd("C:/Users/Monika/Desktop/Straipsniui/papildomi failai")
write.table(sakiniailent, file = "sakiniailent1.txt", sep = "\t", row.names = FALSE)
head(sakiniai)
#########################################################################
######NUSKAITOM POZYMIUS#################################################
#########################################################################
ivardis = scan("C:/Users/Monika/Desktop/Baigiamasis01-23/papildomi failai/ivardis.txt",
    what = "char", sep = "\t", blank.lines.skip = TRUE, encoding="UTF-8", quiet = TRUE)
```

```
linksnis= scan("C:/Users/Monika/Desktop/Baigiamasis01-23/papildomi failai/linksnis.txt",
    what = "char", sep = "\t", blank.lines.skip = TRUE, encoding="UTF-8", quiet = TRUE)
skaicius = scan("C:/Users/Monika/Desktop/Baigiamasis01-23/papildomi failai/skaicius.txt",
     what = "char", sep = "\t", blank.lines.skip = TRUE, encoding="UTF-8", quiet = TRUE)
Prielinksnis = scan("C:/Users/Monika/Desktop/Baigiamasis01-23/papildomi failai/priel.txt
    ", what = "char", sep = "\t", blank.lines.skip = TRUE, encoding="UTF-8", quiet = TRUE
    )
linksnispriel= scan("C:/Users/Monika/Desktop/Baigiamasis01-23/papildomi failai/
    linksnispriel.txt", what = "char", sep = "\t", blank.lines.skip = TRUE, encoding="UTF
    -8", quiet = TRUE)
poz<-c(ivardis, Prielinksnis)
linksn<-c(linksnis, linksnispriel)
markeriai<-cbind(poz, linksn)
write.table(markeriai, file = "markeriai.txt", sep = "\t", row.names = FALSE)
write.table(as.data.frame(poz), file = "poz.txt", sep = "\t", row.names = FALSE,col.names
     = FALSE )
##########################################################################
###SUSKAICIUOJU KIEK YRA KIEKVIENO POZYMIO KIEKVIENAME SAKINIU BLOKE ####
##########################################################################
s<-length(sakiniai)
lentele<- data.frame(poz, linksn)
santykiulent<-data.frame(poz, linksn)
laik<-data.frame(poz)
sii<-NULL
for (j in 1:s) {
sak<-unlist(strsplit(unlist(sakiniai[j]), " "))
blokas<-sak[sak!=""]
laik<-data.frame(poz)
for (k in 1:length(blokas)) {
sii<-NULL
sii <- sapply(poz, function(i) sum(blokas[k] == i))
laik[k+1]<-sii
ss<-rowSums((laik[-1]))
sant<-ss/(length(blokas))
}
lentele[j+2] <-ss
santykiulent[j+2]<-sant
}
lentele[, 1:10]
##########################################################################
##########SUDARAU POZYMIU GRUPES IR SUSKAICIUOJU#########################
##########################################################################
grupiulentele1<-lentele[1:(length(ivardis)),]
grupiulentele2<-lentele[(length(ivardis)+1):(length(poz)),]
grupsantykiulent1<-santykiulent[1:(length(ivardis)),]
```

```
grupsantykiulent2<-santykiulent[(length(ivardis)+1):(length(poz)),]
suma1<-rowsum(grupiulentele1[,-(1:2)], grupiulentele1$linksn)
suma2<-rowsum(grupiulentele2[,-(1:2)], grupiulentele2$linksn)
suma1T<-t(suma1)
suma2T<-t(suma2)
sumasant1<-rowsum(grupsantykiulent1[,-(1:2)], grupsantykiulent1$linksn)
sumasant2<-rowsum(grupsantykiulent2[,-(1:2)], grupsantykiulent2$linksn)
sumasant1T<-t(sumasant1)
sumasant2T<-t(sumasant2)
bendrasant<-cbind(sumasant1T, sumasant2T)
sumavisupozymiu<-rowSums(bendrasant)
summary(sumavisupozymiu)
################################################################################
###########DUOMENYS (teksto nr, pozymiu skaicius grupeje)#######################
################################################################################
duomenys<-as.data.frame(cbind(kurinionr,suma1T, suma2T))
summary(duomenys)
library(pastecs)
stat.desc(as.matrix(duomenys))
autoriaiduom<-c()
for(i in 1:length(duomenys[,1])){
autoriaiduom[i]<-autoriulent[duomenys[i,1],3] }
autoriaiduom<-as.factor(autoriaiduom)
duomenys<-cbind(duomenys, autoriaiduom)
#imtis<-sample(0:1,length(duomenys[,1]),replace=T)
imtis = read.table("C:/Users/Monika/Desktop/Straipsniui/papildomi failai/imtis30.txt",
    header = FALSE, sep = "\t", blank.lines.skip = TRUE)
table(imtis)
#write.table(imtis, file = "imtis30.txt", sep = "\t", row.names = FALSE, col.names =FALSE
     )
duomenys<-cbind(duomenys,imtis)
#write.table(duomenys, file = "duomenys.txt", sep = "\t", row.names = FALSE,col.names =
    TRUE )
duomenys = read.table("C:/Users/Monika/Desktop/Straipsniui/papildomi failai/duomenys.txt
    ", header = TRUE, sep = "\t", blank.lines.skip = TRUE)
head(duomenys)
duomenysmodeliui<-duomenys[duomenys$imtis==0,]
duomenysprognozavimui<-duomenys[duomenys$imtis==1,]
#write.table(duomenysmodeliui, file = "duomenysmodeliui.txt", sep = "\t", row.names =
    FALSE,col.names =TRUE )
n<-10
#imtisprog<-sample(1:n,length(duomenysprognozavimui[,1]),replace=T)
#write.table(imtisprog, file = "imtisprog.txt", sep = "\t", row.names = FALSE,col.names =
    FALSE)
imtisprog = read.table("C:/Users/Monika/Desktop/Straipsniui/papildomi failai/imtisprog.
```

```
    txt", header =FALSE, sep = "\t", blank.lines.skip = TRUE)
table(imtisprog)
colnames(imtisprog)<-"imtisprog"
duomenysprognozavimui$imtis<-NULL
duomenysprognozavimui<-cbind(duomenysprognozavimui, imtisprog)
#write.table(duomenysprognozavimui, file = "duomenysprognozavimui.txt", sep = "\t", row.
    names = FALSE,col.names =TRUE )
duomenysprognozavimui= read.table("C:/Users/Monika/Desktop/Straipsniui/papildomi failai/
    duomenysprognozavimui.txt", header = TRUE, sep = "\t", blank.lines.skip = TRUE)
head(duomenysprognozavimui)
```