Lithuanian Journal of Statistics *2015, vol. 54, No 1, pp. 92–100*
Lietuvos statistikos darbai *2015, 54 t., Nr. 1, 92–100 p.*
www.statisticsjournal.lt

TEKSTŲ NEHOMOGENIŠKUMO TYRIMAS NAUDOJANT ŽYMEKLIUS

Monika Lapėnaitė-Gedvilė[1], Karolina Piaseckienė[2], Marijus Radavičius[3]

[1, 3] Vilniaus universitetas, Matematikos ir informatikos institutas
[1]Vilniaus Gedimino technikos universitetas
[2] Šiaulių universitetas, Informatikos, matematikos, e. studijų institutas
Adresas: [1, 3] Akademijos g. 4, LT-08663 Vilnius, Lietuva
[1] Saulėtekio al. 11, LT-10223 Vilnius, Lietuva
[2] Višinskio g. 19, LT-77156 Šiauliai, Lietuva
El. paštas: [1]*lapenaite.monika@gmail.com*, [2]*k.piaseckiene@gmail.com*, [3]*marijus.radavicius@mii.vu.lt*

Gauta: 2015 m. rugpjūtis Pataisyta: 2015 m. rugsėjis Paskelbta: 2015 m. lapkritis

Santrauka. Straipsnio tikslas – įvertinti tekstų statistinį nehomogeniškumą pagal funkcinių žodžių ir kitų lingvistinių elementų vartoseną. Atliktas empirinis tyrimas remiasi mokykloms rekomenduojamų suskaitmenintų grožinės literatūros kūrinių biblioteka http://ebiblioteka.mkp.emokykla.lt. Apskaičiuojami sudarytų dažnų žodžių formų ar kitų kalbinių struktūrų rinkinių, juos vadinsime žymekliais, dažnumai tekstų blokuose, jungiančiuose 50 iš eilės einančių sakinių. Pastebėta, kad žymeklių dažnumai blokuose turi ženklią perteklinę sklaidą, palyginti su lingvistikoje įprastu homogeniškumo modeliu. Pasirinktoms žymeklių grupėms parinkti kelių tipų hierarchiniai binominės logistinės regresijos modeliai, naudojantys autoriaus identifikatorių, bloko ilgį ir likusių žymeklių dažnius blokuose kaip aiškinančiuosius kintamuosius, leido paaiškinti didelę dalį pasirinktųjų žymeklių perteklinės sklaidos.

Reikšminiai žodžiai: statistinė lingvistika, perteklinė sklaida, tikėtinumo nuokrypis, binominė logistinė regresija, funkciniai žodžiai.

1. Įvadas

Spartus kompiuterių mokslo ir technikos vystymasis ir vis platesnis panaudojimas sąlygojo kiekybinių metodų, tuo pat metu ir duomenų analizės bei statistikos, skverbimąsi į kalbotyrą. Vienas iš statistikos (duomenų analizės) taikymo kalbotyroje ypatumų yra susijęs su apibendrintuoju *Zipfo-Mandelbroto dėsniu* [25, 12], kuriuo nusakomas žodžių, išrikiuotų jų pasitaikymo tekstyne dažnumų mažėjimo tvarka, skirstinys. Praktiškai šis dėsnis pasireiškia tuo, kad nedidelė žodžių dalis pasitaiko labai dažnai, o ženkli jų dalis, kartais viršijanti 50 %, tekstynuose yra pavartota tik kartą. Dažnai tekstuose sutinkami žodžiai (arba žodžių formos) turi savybių, nebūdingų retesniems žodžiams, ir paprastai yra funkciniai žodžiai, t. y. atlieka tekstuose tam tikras funkcijas. Labai dažnų žodžių ar žodžių formų reikšmę ir panaudojimo galimybes lietuviškų tekstų analizėje nagrinėjo Utka (2005a, 2005b, 2006) [21–23]. Straipsnyje [22] pastebėta, kad labai dažnų žodžių proporcijos skirtinguose tekstuose (kūriniuose) yra gana skirtingos. Dažniausių žodžių rinkiniui, sudarytam iš 98 žodžių, aprašomosios faktorinės analizės metodu išskirti 7 faktoriai, kurių kokybinė interpretacija leido įvardyti ir aprašyti 7 tekstų funkcijas ir paradigmas. Dabašinskienė (2009) [5] tyrė šnekamąją kalbą ir lygino ją su rašytine kalba. Jos pateikta statistika rodo, kad įvardžių, prielinksnių ir prieveiksmių (jų pagrindinę dalį sudaro labai dažnai vartojami žodžiai) vartosenos lietuviškuose tekstuose proporcijos šnekamojoje ir rašytinėje kalboje pastebimai skiriasi. Vadinasi, jos turėtų skirtis ir tekstuose, kuriuose skiriasi tiesioginės ir netiesioginės kalbos proporcija. Tiesioginės kalbos ypatumus, taip pat ir labai dažnų žodžių pasiskirstymą tyrė Leonavičienė (2006, 2007) [10, 11]. Pacituoti pastebėjimai rodo, kad lietuviški tekstai yra gana *nehomogeniški* pagal labai dažnų, funkcinių žodžių vartojimo dažnį.

Kalbos (tekstų) nehomogeniškumas yra suprantamas įvairiai. Kalbos variantiškumas (angl. *variation*) ir kitimas (angl. *change*) yra pagrindinės sociolingvistikos sąvokos [13, 28]. Sociolingvistika tyrinėja tarmes, kalbos skirtumus ir panašumus tarp skirtingų socialinių grupių, skirtingų geografinių regionų, laikmečių ir pan. Tačiau skiriasi ne tik

socialinių grupių, bet ir atskirų autorių kalba. Tekstų heterogeniškumas apibrėžiamas kaip intertekstualumo apibendrinimas, svetimų tekstų, posakių tiesioginis (žymėtasis heterogeniškumas) ir netiesioginis (nežymėtasis heterogeniškumas) panaudojimas tekste [9, 3].

Tekstų (statistinis) nehomogeniškumas yra seniai pastebėtas ir taikomas natūralios kalbos apdorojime (angl. *natural language processing*) bei informacijos paieškoje ir išrinkime (angl. *information retrieval*), pavyzdžiui, siekiant padidinti užklausų (angl. *query*) efektyvumą [15] ir sprendžiant tekstų klasifikavimo (pagal žanrą, stilių ir pan.), taip pat ir autoriaus identifikavimo uždavinius [19, 7, 27]. Tačiau kalbos mokslo požiūriu autorystės nustatymas nėra įdomus uždavinys, nebent siekiama nustatyti to paties teksto, kurį reikėjo kuo tiksliau atpasakoti, atpasakojimo autorių (plg. [27, 26]). Kitaip autoriai atskiriami pagal mintis ir nuostatas, kurios atsispindi kalboje, o ne pagal pačią kalbą.

Šiame straipsnyje tiriamas tekstų (kalbos) nehomogeniškumas skiriasi nuo aukščiau minėtų jo variantų. Jis pasireiškia per labai dažnų, funkcinių žodžių ir kalbos elementų vartosenos (statistinius) skirtumus, jos kintamumą tekste. Jis būdingas ir tam pačiam autoriui, ir net tam pačiam jo kūriniui, nes skirtingoms mintims išdėstyti reikia skirtingų kalbos priemonių bei struktūrų. Tekstų nehomogeniškumas apsunkina (analitinės) statistikos metodų taikymą kalbotyros moksle, kadangi yra sudėtinga sukonstruoti adekvatų, atsižvelgiantį į nehomogeniškumą, galimų tekstų statistinį modelį. Jeigu autorius galima atskirti pagal jų tekstų statistines savybes, vadinasi, kiekvienam jų reikia sudaryti skirtingus jų kalbos statistinius modelius. Taip formuluojant, tai – neperspektyvus uždavinys. Callison-Burch ir Osborne (2003) ([4], 25 psl.; plg. [1], p. 19–21, kur Abney polemizuoja su Chomskiu) vadina tai natūraliosios kalbos nestacionarumo problema ir mano, kad artimiausiu metu ji, matyt, nebus išspręsta. Kol kas modeliuojant kalbą įprasta laikyti, kad ji yra generuojama to paties stacionaraus šaltinio. Tariama, kad sakiniai tekste yra generuojami nepriklausomai vienas nuo kito, o sakinio elementų skirstinys yra aprašomas n-tos eilės Markovo grandine ($n \leq 3$) arba pasitelkiant tikimybines kontekstines gramatikas. Nežinomi modelio parametrai, sąlyginės tikimybės, įvertinamos remiantis n-gramų (n iš eilės einančių žodžių) dažnumų statistika tekstyne [4, 17, 7].

Labai dažniems žodžiams pritaikyta faktorinė analizė [22] liudija, kad tarp jų dažnumų tekstuose yra koreliacija. Remiantis tuo, kad dauguma labai dažnų žodžių yra funkciniai ir susiję su tam tikromis kalbos struktūromis, galima iškelti hipotezę, kad minėta koreliacija atspindi statistinius sąryšius tarp atitinkamų kalbos elementų ir struktūrų. Jeigu pasirodytų, kad tie sąryšiai yra nepriklausomi arba silpnai priklausomi nuo teksto autoriaus, temos ir pan., tai būtų vienas iš svarbių požymių, kad jie atspindi ne atskiram konkrečiam tekstui, o pačiai kalbai būdingas savybes [14].

Šiame darbe atliktas eksperimentas, kurio tikslas – įvertinti lietuviškų tekstų nehomogeniškumą pagal vartojamų funkcinių žodžių ir kitų kalbos elementų tikėtinus dažnumus ir proporcijas. Tariama, kad tuos dažnumus galima įvertinti pasitelkiant *žymeklių* statistiką [8]. Žymekliai – tai tekstuose pakankamai dažnai pasitaikantys ir gana paprastu algoritmu identifikuojami žodžiai, žodžių formos ar kiti teksto elementai, glaudžiai, kaip manoma, susiję su tiriamąja kalbos funkcija, struktūra ar savybe. Tinkamų žymeklių (grupių) parinkimas yra sudėtingas uždavinys, palyginamas ar net sudėtingesnis negu informacijos paieškos ir išrinkimo efektyvios sistemos sukūrimas (plg. [18]). Taigi, žymekliai yra terminų *labai dažnas žodis*, *labai dažna žodžio forma* [21], taip pat natūraliosios kalbos apdorojime taikomų terminų *žodžių krepšelis* (angl. *bag-of-words*), *rodyklės–terminai* (angl. *indexing terms*) analogas (žr., pvz., [4, 15, 20]), bet bendrą žymeklių idėją ir paskirtį geriausiai atitinka lingvistinėje tipologijoje įvestas terminas *klasterinė kalbos dalis*. Tai – laisvas vertimas iš anglų kalbos termino *distributional part of speach*, kuris apibrėžiamas kaip žodžių formų rinkinys, gautas pritaikius tekstynui klasterizaciją (klasterinę analizę) [24]. Klasterinėms kalbos dalims nekeliami jokie išankstiniai kokių nors kalbos struktūrų reikalavimai, jos apibrėžiamos remiantis vien tik jų pasiskirstymu tiriamajame tekstyne (tai paaiškina angliško termino kilmę). Tačiau žymekliai skiriasi nuo visų aukščiau išvardytų terminų vienu ar kitu aspektu. Pavyzdžiui, skirtingai nuo klasterinių kalbos dalių, žymeklių rinkinių sudarymui gali būti taikomi kiti metodai, ne vien tik klasterinė analizė. Nors šiame darbe, kaip ir [8], tinkamų žymeklių, glaudžiai susietų su konkrečiomis tiriamomis kalbos struktūromis, parinkimo uždavinys nėra sprendžiamas, ši tema paliečiama aptariant tyrimo rezultatus.

Konkretizuosime sprendžiamus uždavinius:

(1) Parodyti, kad tekstuose (kalboje) ženkli statistinės informacijos apie kalbą dalis, vaizdžiai tariant, yra už žodžių trigramų ir sakinio ribų, o taip pat šalia bendros ištisinės tekstynų statistikos, nes, skirtingai nuo pastarosios, ta informacija remiasi lokalia (sakykim, penkiasdešimties sakinių ilgio) nuoseklaus teksto fragmentų statistika.

(2) Pademonstruoti, kad tam tikri dažnų žodžių ar kitų kalbos elementų rinkiniai (juos vadiname žymekliais) gali tikti aukščiau minėtai statistinei informacijai kaupti.

(3) Sudaryti kelis potencialių žymeklių rinkinius ir ištirti jų tarpusavio lokalius, bet išeinančius už sakinio ribų statistinius ryšius. Patogus būdas tokiems ryšiams aprašyti yra regresiniai (prognozavimo) modeliai, kurie, remiantis informacija apie vienus žymeklius, leidžia prognozuoti kitų žymeklių dažnumą ir proporciją tekste bei įvertinti jų skirstinį. Sociometrijoje jau įprasta taikyti logistinę regresiją [13, 6], šis modelis taikomas ir šiame darbe.

Tyrime naudojami duomenys yra dvidešimt keturių autorių trisdešimties panašaus žanro lietuviškų kūrinių, skirtų 5–8 kl. mokiniams, tekstai (http://mkp.emokykla.lt/ebiblioteka/). Duomenų apdorojimui ir statistinei analizei buvo naudojamas paketas *R* [http://www.r-project.org/] ir sistema SAS [19] (procedūra *LOGISTIC*).

Kitame skyrelyje trumpai aprašyta tyrimo metodika, 3 skyrelyje aptarti rezultatai ir 4 skyrelyje suformuluotos išvados.

2. Tyrimo metodika

Darbe nagrinėjami 3 žymeklių rinkiniai: (I) įvardžiai, (L) prielinksniai ir (B) įvardžiuotiniai būdvardžiai bei būdvardžių aukštesnysis ir aukščiausiasis laipsniai. Kai kuriems įvardžiams, prielinksniams ir išvardytoms būdvardžio formoms būdingas morfologinis nevienareikšmiškumas [16, 17]. Pavyzdžiui, įvardis „mano“ yra ir veiksmažodžio „manyti“ forma, o įvardis „mes“ yra ir veiksmažodžio „mesti“ forma. Todėl šie rinkinių pavadinimai yra sąlyginiai, faktiškai jie yra tik *žodžių formų* rinkiniai. Žymeklių morfologinis vienareikšmiškumas kartais gali būti pageidautinas, bet apskritai jis nėra būtinas. O remiantis analogija tarp žymeklių ir klasterinių kalbos dalių, morfologinis vienareikšmiškumas visai nereikalingas [24]. Tyrime naudojami neanotuoti tekstai ir žymeklių morfologinio nevienareikšmiškumo problema nesprendžiama.

Žymeklių rinkiniai dar yra skirstomi į grupes pagal linksnį, paskirtį ar kitas ypatybes. Kiekvienos grupės pavadinimas susideda iš 3 simbolių. Pirmasis simbolis rodo pavadinimą vieno iš jau minėtų rinkinių (I – įvardis, L – prielinksnis, B – būdvardžio formos), antrasis – papildomas grupės ypatybes, savitas kiekvienam rinkiniui (žr. žemiau), o trečiasis – su žymekliu susijusį linksnį (1 – vardininkas, 2 – kilmininkas, 3 – naudininkas, 4 – galininkas, 5 – įnagininkas, 6 – vietininkas, o simbolis 0 žymi, kad linksnis nėra vienareikšmiškai nusakytas arba jis ignoruojamas).

Antruoju simboliu pažymėtos ypatybės kiekvienam rinkiniui yra skirtingos.

Įvardžiams tokių ypatybių yra 10: raidė A žymi įvardžio „aš“ formas, T – įvardžio „tu“ formas, J – įvardžių „jis“ ir „ji“ formas, R – parodomųjų įvardžių formas (pvz., „šis“, „ši“, „tas“, „toks“ ir pan.), B – apibendrinamųjų įvardžių formas (pvz., „viskas“, „niekas“, „visas“, „visa“ ir pan.), Q – klausiamųjų įvardžių formas (pvz., „kas“, „kuris“, „kuri“, „katras“, „katra“ ir pan.), K – atskiriamųjų įvardžių formas (pvz., „kitas“, „kita“, „kitoks“, „kitokia“, „vienoks“, vienokia“ ir pan.) ir t.t.

Prielinksniams išskirti 7 papildomi požymiai. Raidė V žymi prielinksnius, kurie dalyvauja tik vietos aplinkybėse (pvz., „anapus“, „ant“, „greta“, „pas“, „paskui“, „prie“, „palei“, „pro“, „šiapus“, „ties“, „viduj“), T – prielinksnius, kurie yra susiję tik su vieta ir su laiku (pvz., „apie“, „arti“, „iki“, „ligi“, „nuo“, „tarp“), C – prielinksnius, kurie yra susiję su išrikiavimu ir palyginimu („aukščiau“, „pirma“, „pirmiau“, „sulig“, „žemiau“), A – prielinksnius, kurie yra susiję su tikslu, priežastimi ar nurodo „centrą“, į kurį nukreiptas arba iš kurio kyla veiksmas (pvz., „anot“, „dėka“, „dėl“, „į“, „iš“, „pagal“, „pasak“); E – prielinksnius, kurie rodo išskyrimą, atmetimą ar pakeitimą („be“, „vietoj“), S žymi prielinksnį „su“. Visi kiti prielinksniai sudaro grupę, kuri žymima simboliu „_“ (apatiniu brūkšniu).

Būdvardžių formoms yra sudarytos 3 grupės: skaičius „2“ žymi aukštesnįjį būdvardžio laipsnį, 3 – aukščiausiąjį būdvardžio laipsnį, o raidė „i“ – įvardžiuotinę būdvardžio formą. Šio rinkinio žodžiai identifikuojami pagal minėtoms būdvardžių formoms būdingas galūnes (bet tokias galūnes turi ne vien tik būdvardžiai!).

Sudarytų žymeklių grupių statistinis tyrimas atliekamas pagal tokią metodiką.

1. Pasirenkama žymeklių grupė, kurios narių pasitaikymo tekste dažnumui prognozuoti bus sudaromas regresinis modelis, naudojantis informaciją apie kitų žymeklių grupių pasitaikymo tekste dažnumus.

2. Kiekvieno autoriaus tekstai suskaidomi į nesikertančius blokus po 50 sakinių ir kiekviename bloke suskaičiuojami visų žymeklių grupių pasitaikymo juose dažnumai. Prognozuojamas pasirinktos žymeklių grupės dažnumas laikomas *aiškinamuoju kintamuoju*, o žymeklių grupių pasitaikymo tekste dažnumai yra laikomi *aiškinančiaisiais kintamaisiais*. Aiškinančiųjų kintamųjų sąrašas papildomas bloko ilgiu (bloko žodžių kiekiu), tiksliau – jo (dešimtainiu) logaritmu, bei autoriaus identifikatoriumi (numeriu).

3. Ankstesniame punkte sudarytiems blokų duomenims parenkamas binominės logistinės regresijos modelis ir kryžminio patikrinimo metodu įvertinami pasirinktos žymeklių grupės pasirodymo kiekviename bloke dažnumai ir tikimybės.

4. Apskaičiuojami Pearsono (normuotas chi-kvadrato) atstumas

$$\mathbf{C^2} = \frac{1}{N}\sum_{j=1}^{N} \frac{(y_j - \hat{y}_j)^2}{\hat{y}_j}$$

ir tikėtinumo nuokrypis (arba normuota Kullbacko–Leiblerio divergencija, angl. *deviance*)

$$\mathbf{T^2} = \frac{2}{N}\sum_{j=1}^{N} y_j \log\left(\frac{y_j}{\hat{y}_j}\right),$$

kurie matuoja neatitikimo dydį tarp stebėtųjų reikšmių ir jų prognozių, gautų naudojant įvairius statistinius modelius (žr., pavyzdžiui, [2, 13]). Čia y_j žymi j-ąją stebėtąją aiškinamojo kintamojo reikšmę, o $\hat{y}_j$ žymi jos prognozę, naudojant atitinkamą modelį su įvertintais nežinomais to modelio parametrais, $j = 1, \dots, N$.

1 pastaba. Žymeklių grupė prognozavimui pasirenkama remiantis keliomis taisyklėmis. Visų pirma, ji turi būti pakankamai skaitlinga: yra nagrinėjamos tik tos grupės, kurių vidutinis dažnumas bloke viršijo 1. Be to, buvo atrenkamos grupės su didesniu tarpblokiniu nehomogeniškumu (statistiniu kintamumu). Ir, žinoma, prioritetą turi tos grupės, kurias gana paprasta susieti su kokia nors kalbos struktūra ar funkcija. Tokiu būdu buvo pasirinktos trys grupės: IT0, LV2 ir Inag. Grupė IT0 – visos įvardžio „tu“ formos, jos siejamos su tiesiogine kalba. Grupė LV2 – prielinksniai, reikalaujantys kilmininko linksnio ir susieti su vietos aplinkybe (pvz., „anapus“, „ant“, „greta“, „kitapus“, „link“, „netoli“, „prie“, „vidur“, „vidury“, „virš“, „viršuj“, „viršum“). Kadangi įnagininko linksnis vartojamas santykinai gana retai, tai buvo sudaryta nauja sudėtinė grupė Inag, į kurią apjungtos visos nagrinėjamos žymeklių grupės, susietos vien tik su įnagininko linksniu. Tai yra, į grupę Inag sujungtos visos žymeklių grupės, kurių pavadinimas baigiasi simboliu „5“.

2 pastaba. Straipsnyje [8] bloko dydis buvo 10 sakinių. Šiame darbe jis padidintas iki 50 sakinių, nes į tyrimą įtrauktos ir retesnės žymeklių grupės. Idealiu atveju parinktas blokų dydis turėtų suderinti du prieštaringus siekius: sudaryti pakankamai didelius tekstų blokus, kad juose išryškėtų statistiniai dėsningumai, bet nesujungti į vieną bloką teksto fragmentų, kurie reikšmingai skiriasi naudojamomis kalbos struktūromis ir funkcijomis. Surinkus pakankamai statistinės informacijos apie lietuvių kalbą (kalbos elementus ir struktūras) galima būtų bandyti kurti adaptyvias teksto skaidymo į blokus procedūras.

3 pastaba. Buvo sudaryti 5 tipų hierarchiniai binominės logistinės regresijos modeliai. Prieš juos aptardami, pateiksime trumpą binominės logistinės regresijos įvadą. Išsamų logistinės regresijos ir jos modelio parinkimo metodų aprašymą galima rasti [2], žr. taip pat [19]. Logistinė regresija dažnai taikoma sociolingvistikoje [13, 6].

Logistinės regresijos modelyje tariama, kad tiriamojo įvykio pasitaikymo tikimybė p priklauso nuo aiškinančiųjų kintamųjų vektoriaus $x := (x_1,\dots,x_k)' \in R^k$, $p = p(x)$, ir ta priklausomybė yra nusakoma logit transformacija, vadinama jungties funkcija, bei tiesiniu prediktoriumi η su nežinomais parametrais $\beta := (\beta_1, \dots, \beta_k)' \in R^k$:

$$\text{logit}\left(p(x;\beta)\right) := \log\left(\frac{p(x;\beta)}{1-p(x;\beta)}\right) = \eta(x;\beta) = \beta' x. \tag{1}$$

Tegu y_j žymi tiriamojo įvykio stebėtą dažnumą j-oje imtyje, sudarytoje iš n_j nepriklausomų ėmimų, $j = 1, \dots, N$. Laikoma, kad tie stebėti dažnumai $y_1, \dots, y_N$ yra sąlyginai nepriklausomi, kai žinomos visos aiškinančiųjų kintamųjų reikšmės $x_{*j} := \left(x_{1j}, \dots, x_{kj}\right)', j = 1, \dots, N$. Taigi, y_j, kai žinomas x_{*j}, turi binominį skirstinį su parametrais n_j ir $p_j := p\left(x_{*j}; \beta\right)$. Nežinomi logistinės regresijos modelio parametrai $\beta = (\beta_1, \dots, \beta_k)'$ įvertinami naudojant didžiausio tikėtinumo metodą. Tegu $\hat{\beta} = \left(\hat{\beta}_1, \dots, \hat{\beta}_k\right)'$ žymi gautą didžiausio tikėtinumo įvertinį. Tuomet tikimybių p_j didžiausio tikėtinumo įvertiniai (prognozės) yra $\hat{p}_j := p\left(x_{*j}; \hat{\beta}\right)$, o stebėtų dažnumų y_j prognozės yra $\hat{y}_j = \hat{p}_j n_j, j = 1, \dots, N$.

Statistiniuose paketuose naudojamos simbolinės modelių užrašymo (specifikavimo) formos. Jeigu visi aiškinantieji kintamieji (1) modelyje yra kiekybiniai, o x_1 yra daugiklis prie laisvojo nario, tai simbolinėje formoje šis modelis atrodytų taip: $y \sim x_2 + \ldots + x_k$. Kairėje šios išraiškos pusėje nurodomas aiškinamasis kintamasis, o dešinėje – aiškinantieji kintamieji, kitaip tariant, adityvūs veiksniai, įtraukti į tiesinio prediktoriaus apibrėžimą. Šiuo atveju yra paprasta susieti simbolinį modelio užrašą su matematiniu modeliu. Tačiau, jeigu simboliniame modelio aprašyme yra naudojami faktoriai, t. y. kokybiniai kintamieji, ir jų sąveikos (angl. *interaction*), tai detalizuoti, kaip tas aprašymas siejasi su skaliariniais kintamaisiais $x_1, \ldots, x_k$ yra keblu, nes sukuriama gana daug fiktyvių skaliarinių kintamųjų (statistiniuose paketuose tai atliekama automatiškai) priklausomybėms nuo faktorių ir sąveikų tinkamai aprašyti. Taikymuose ta sąsaja paprastai nėra svarbi, ypač tada, kai naudojami hierarchiniai modeliai. Hierarchiniuose modeliuose reikalaujama į modelį įtraukti visus kintamuosius ir visas jų žemesnės kaip *m* eilės sąveikas, jeigu pagal kokį nors modelio parinkimo kriterijų į modelį jau įtrauktos visų jų *m*-tos eilės sąveikos. Pavyzdžiui, jeigu į modelį jau įtraukta kintamųjų a, b ir c sąveika (ji žymima a*b*c), tai į modelį būtinai įtraukiami ir veiksniai a, b, c, a*b, a*c, b*c.

Šiame darbe logistinės regresijos modelio parinkimas atliktas su SAS procedūra LOGISTIC [19], kuri turi dideles galimybes, yra lanksti ir leidžia efektyviai spręsti didelės apimties uždavinius. Kai kurių darbe tirtų modelių pradinis kintamųjų sąrašas gerokai viršijo 1000.

Aptarsime nagrinėtus binominės logistinės regresijos modelių tipus. Modelių aprašymuose, siekiant pažymėjimų paprastumo, stebėti blokuose žymeklių grupių dažnumai yra žymimi taip pat, kaip ir atitinkamos žymeklių grupės.

Pirmajame modelyje (M1) tariama, kad tiriamos žymeklių grupės pasitaikymo tekste tikimybė yra vienoda visuose blokuose, nepriklausomai nuo bloko teksto autoriaus. Ši tikimybė yra vienintelis šio modelio parametras. Modelis (M1) atitiktų įprastas (nors tiesiogiai ir neformuluojamas) tekstynų lingvistikos prielaidas.

Antrajame modelyje (M2) laikoma, jog ta tikimybė gali būti skirtinga skirtingų autorių tekstuose. Modelis (M2) turi 24 nežinomus parametrus.

Trečiasis modelis (M3) atsižvelgia ir į bendrą kiekvieno bloko žodžių kiekį, kitaip tariant, į vidutinį sakinio ilgį blokuose, kadangi sakinių kiekis visuose blokuose yra vienodas. Be to, šiame modelyje laikoma, jog vidutinio sakinio ilgio įtaka skirtingų autorių tekstuose gali skirtis. Simboliškai jį užrašyti galima taip:

$$\text{IT0} \sim \text{autor} + \text{ilgis\_lg} + \text{autor*ilgis\_lg}. \qquad \text{(M3)}$$

Kairėje šios išraiškos pusėje yra nurodytas aiškinamasis kintamasis, šiuo atveju IT0. Kaip jau buvo minėta, kintamasis IT0 žymi stebėtus blokuose grupės IT0 dažnius. Dešinėje išraiškos pusėje yra išvardyti tiesinio prediktoriaus adityvieji veiksniai (angl. *effects*), išreiškiami per pradinius aiškinančiuosius kintamuosius, šiuo atveju kintamuosius autor ir ilgis_lg. Kintamasis autor žymi faktorių, kurį nusako autoriaus identifikatorius (numeris), ilgis_lg yra bloko ilgio dešimtainis logaritmas. Narys autor*ilgis_lg aprašo faktoriaus autor ir kintamojo ilgis_lg sąveikos poveikį aiškinamojo kintamojo skirstiniui. Modelis (M3) turi 48 nežinomus parametrus.

Ketvirtasis modelis (M4) sudaromas nuosekliai papildant trečiąjį modelį (statistiškai reikšmingais arba tenkinančiais Akaike informacinį kriterijų AIC) žymeklių grupių dažnumų, stebėtų blokuose, kintamaisiais bei jų antrosios eilės sąveikomis (sandaugomis) tol, kol Hosmerio-Lemeshow'o testas pagaliau neatmeta nulinės hipotezės apie logistinės regresijos modelio suderinamumą (atitiktį duomenims) su reikšmingumo lygmeniu 0,05. Tam naudojama SAS procedūra Logistic su automatiniu, einančiu pirmyn (angl. *forward*) į modelį įtraukiamų kintamųjų parinkimu [19, 2].

Penktasis modelis (M5) sudaromas analogiškai kaip ir modelis (M4), bet nenaudojama informacija apie tekstų autorius (faktorius autor).

3. Rezultatų apžvalga

Bendra nagrinėjamų tekstų apimtis yra 1 176 822 žodžiai, 2 434 blokai, tarp jų IT0 grupės žymekliai (žodžiai) pasitaikė 10 091 kartą, tai sudaro apie 0,857 % visų žodžių, vidutiniškai 4,146 atvejo bloke; LV2 grupės žodžiai pasitaikė 9 736 kartus, tai yra maždaug 0,827 % visų žodžių, vidutiniškai 4,000 atvejai bloke, Inag grupės žodžiai pasitaikė 13 621 kartą, tai yra apie 1,157 % visų žodžių, vidutiniškai 5,596 atvejo bloke.

1 lentelė. Parinktų modelių Pearsono atstumas ir tikėtinumo nuokrypis

Žymeklių grupė	Suderinamumo matas	Modeliai					
		M1	M2	M3	M4	M5	Max
IT0	Pearsono atstumas	4,64	3,41	2,77	1,84	1,79	1
		0 %	27 %	40 %	60 %	61 %	78 %
	Tikėtinumo nuokrypis	4,45	3,13	2,596	1,81	1,91	1
		0 %	30 %	42 %	59 %	57 %	78 %
LV2	Pearsono atstumas	2,46	1,65	1,62	1,16	1,42	1
		0 %	33 %	34 %	53 %	42 %	59 %
	Tikėtinumo nuokrypis	2,4	1,71	1,69	1,24	1,51	1
		0 %	29 %	30 %	48 %	37 %	58 %
Inag	Pearsono atstumas	1,9	1,62	1,6	1,51	1,61	1
		0 %	15 %	16 %	21 %	15 %	47 %
	Tikėtinumo nuokrypis	2,02	1,7	1,66	1,58	1,69	1
		0 %	16 %	18 %	22 %	16 %	50 %

Pagrindiniai rezultatai apibendrinti 1-oje lentelėje. Joje pateiktas normuotas Pearsono atstumas ir tikėtinumo nuokrypis tarp atitinkamo aiškinamojo kintamojo (grupių IT0, LV2 ar Inag dažnumų) blokuose stebėtų reikšmių ir jų prognozių, sudarytų modelių (M1)–(M5) pagrindu. Normuotas Pearsono atstumas ir tikėtinumo nuokrypis yra įvertinti kryžminio patikrinimo metodu (angl. *cross-validation*) [19]. Gerai parinktam modeliui minėti suderinamumo matai turėtų būti artimi 1. Matome, kad nė vienas modelis šiuo atveju nėra geras. Grupės IT0 žymeklių dažnumai blokuose labai varijuoja, Pearsono atstumo įvertis modeliui (M1), suderintam su homogeniškumo hipoteze, viršija 4,64. Tai rodo didelį tekstų heterogeniškumą įvardžio „tu“ formų vartojimo atžvilgiu. Bet atsižvelgus tik į teksto autorius ir bloko ilgį, pavyksta paaiškinti apie 40 % jų vartojimo sklaidos (statistinio kintamumo). Papildomai į modelį įtraukus ir informaciją apie kitus žymeklius, pavyksta aprašyti jau apie 60 % sklaidos. Iki viršutinės ribos – 78,46 % $\approx$ (1 – 1/ 4,64) 100 % (lentelėje ji pateikta stulpelyje Max) trūksta ne tiek daug. Svarbu paminėti, kad beveik tokio paties prognozės tikslumo, kaip su parinktu modeliu (M4), pavyksta pasiekti ir su modeliu (M5), kuriame visai nenaudojama informacija apie tekstų autorius.

Grupių LV2 ir Inag žymeklių dažnumų kintamumas blokuose gerokai mažesnis: Pearsono atstumo įvertis homogeniškajam modeliui (M1) yra atitinkamai lygus 2,46 ir 1,9.

Detaliau aptarsime IT0 grupės rezultatus. Įvardžio „tu“ formos dažniausiai naudojamos tiesioginėje kalboje, kreipiantis į vieną iš pašnekovų. Todėl natūralu IT0 grupės žymeklius laikyti šios kalbinės „struktūros“ skiriamuoju požymiu. Parinktas modelis (M4) (modelių parinkimo procedūra aprašyta 3-ioje pastaboje) turi 143 parametrus, o modelis (M5) turi 228 parametrus. Žemiau pateiktose simbolinėse modelių išraiškose apsiribojome tik svarbesniais kintamaisiais (visi jie statistiškai reikšmingi su reikšmingumo lygmeniu $< 0{,}0001$), paaiškinančiais ženklią IT0 perteklinės sklaidos dalį:

$$\begin{aligned} IT0 \sim {} & autor - ilgis\_lg + autor*ilgis\_lg + IA0 + IB2 - IB1 - IJ1 + LA4 - Bi6 + IQ0 + IB0*IB1 - IK4*IB2 \\ & + L\_4*IJ1 + LE2*IJ6 + LV4*IZ1 - LT2 + \ldots . \end{aligned} \qquad (M4)$$

$$\begin{aligned} IT0 \sim {} & IA0 - ilgis\_lg + IJ4 + IQ0 - LA4 + IJ3*LC2 - Bi6 - B22 + IA0*IJ1 - L\_4*IB4 - LT2 - IA0*IJ4 \\ & - Bi2*LC5 - LV4*IJ4 + B24*L\_2 - IQ3*B26 + Bi2*B26 + Bi5*LV4 + \ldots . \end{aligned} \qquad (M5)$$

Vaizdumo dėlei prieš aiškinantįjį kintamąjį rašome minuso ženklą, jeigu pastarajam didėjant aiškinamasis kintamasis mažėja. Nors dauguma veiksnių modeliuose (M4) ir (M5) yra neinterpretuojami, kai kuriems veiksniams galima rasti natūralų paaiškinimą. Didžiausią dalį IT0 sklaidos aprašo įvardžių grupės IA0, kuri yra jungtinė grupei IT0, dažnumai.

IT0 žymekliai retai naudojami sakinyje dažniau negu 1-ą kartą, tad augant vidutiniam sakinio ilgiui jų proporcija natūraliai mažėja. Be to, tiesioginės kalbos sakiniai yra vidutiniškai trumpesni už netiesioginės kalbos sakinius [11]. Tai paaiškina neigiamą ženklą prieš kintamąjį ilgis_lg. Kategorinis kintamasis autor aprašo skirtingą autorių polinkį savo kūriniuose (tekstuose) naudoti IT0 žymeklius (tuo pačiu ir tiesioginę kalbą), o sąveikos veiksnį autor*ilgis_lg kiek supaprastintai galima interpretuoti kaip autorių individualius skirtumus tarp sakinių su žymekliu IT0 ir be jo ilgių. Narys IQ0 atspindi tai, kad tiesioginėje kalboje daug dažniau negu netiesioginėje pasitaiko klausiamieji sakiniai. Veiksnys (– Bi6) (jis išlieka ir (M5) modelyje), t. y. įvardžiuotinės būdvardžio formos vartojimas, gali reikšti, kad tekste aptariamas svarbesnis objektas negu pašnekovai. Panašią prasmę galima būtų suteikti ir nariui (– IK4*IB2), rodančiam, kad be apibendrinamųjų įvardžių bloke pasitaiko ir (daug) atskiriamųjų įvardžių. Bet šie bandymai interpretuoti parinkto modelio veiksnius yra tik spėjimai. Apskritai, net paprasčiausių kalbos elementų adekvataus, pagrįsto ir prasmingai interpretuojamo prognozavimo modelio parinkimas yra sudėtingas uždavinys. Tai nėra šio darbo tikslas. Pateikta euristinė modelio (M4) veiksnių interpretacija tik iliustruoja kol kas kalbotyroje neišnaudotas statistinio modeliavimo galimybes ir leidžia iškelti hipotezę, kad skirtumai tarp šnekamosios (tiesioginės) ir rašytinės (netiesioginės) kalbos yra subtilesni negu rodo straipsnyje [5] pateikta bendra įvairių kalbos elementų (tekstynuose) stebėtų proporcijų statistika (žr. taip pat [9, 10]).

Modelio (M5) pagrindiniai veiksniai reikšmingai skiriasi nuo modelio (M4) pagrindinių veiksnių, veiksnys LA4 net keičia savo ženklą. Tai suprantama, nes modelio (M5), kuriame nenaudojama informacija apie autorius, dalis veiksnių tampa pagrindiniais todėl, kad padeda atskirti autorius, kurie dažniau už kitus autorius vartoja tiesioginę kalbą ir IT0 grupės žymeklius. Bet keli pagrindiniai veiksniai išlieka tie patys: IA0, (– ilgis_lg), IQ0, (– Bi6), (– LT2). Tai, kad modelis (M5) pasiekia beveik tokį patį prognozavimo tikslumą kaip ir (M4), rodo, kad nepanaudotą informaciją apie autorius praktiškai pavyksta kompensuoti tam tikrų žymeklių grupių ir jų tarpusavio sąveikų statistine informacija.

Prielinksnių, reikalaujančių kilmininko linksnio ir susietų su vietos aplinkybe, grupei, t. y. žymeklių grupei LV2, parinktų modelių (M4) (193 parametrai) ir (M5) (180 parametrų) pagrindiniai veiksniai yra tokie:

$$\begin{aligned} LV2 \sim {} & autor + ilgis\_lg + autor*ilgis\_lg - IR3 + LT2 - IR4*LT2 - IR2 - IR5 + B34*IR5 \\ & + IR1*IB2 - IB4 - B34 - IA0 + IQ2 - IJ5 + IA0*LV4 + \dots ; \end{aligned} \quad (M4)$$

$$\begin{aligned} LV2 \sim {} & ilgis\_lg - IR5 + LT2 - IJ5 + LS5 - IR4*LT2 + LV5 - B34 + IR6 \\ & + LV4*B30 + B34*IR5 + IB3 - IQ4 - IQ1 + \dots . \end{aligned} \quad (M5)$$

Visi išvardyti pagrindiniai veiksniai yra statistiškai reikšmingi su reikšmingumo lygmeniu < 0,0001. Šiuo atveju apie pusė pagrindinių veiksnių modeliuose (M4) ir (M5) sutampa: ilgis_lg, (– IR5), LT2, (– IJ5), (– IR4*LT2), (– B34), B34*IR5, IR1*IB2. Gal dėl to, kad tiriami autorių tekstai yra gana panašūs pagal grupės LV2 žymeklių vartojimą (žr. 1 lentelę).

Žemiau pateikti žymeklių grupei Inag, sudarytoje iš žymeklių grupių, reikalaujančių vien tik įnagininko linksnio, parinktų modelių (M4) (74 parametrai) ir (M5) (77 parametrai) pagrindiniai veiksniai:

$$Inag \sim autor + ilgis\_lg + autor*ilgis\_lg - IQ2 - IR4 + IB4 - IB4*L\_4 - Y34 - LA2*IZ3 + IJ3*Y34 + \dots ; \quad (M4)$$

$$Inag \sim ilgis\_lg + Yi4 - Y34 + IB4 + LT4 + IR3*IK6 + IJ3*IJ2 + IZ1 - IB4*IR1 + \dots . \quad (M5)$$

Visi išvardyti modelio (M4) pagrindiniai veiksniai yra statistiškai reikšmingi su reikšmingumo lygmeniu < 0,001, o modelio (M5) – su lygmeniu < 0,0001.

4. Išvados

Lietuviški dvidešimt keturių autorių neanotuoti tekstai buvo padalyti į blokus po 50 iš eilės einančių sakinių ir tirtas šių blokų homogeniškumas pagal tam tikrų žodžių, vadinamų žymekliais, vartoseną. Žymekliai – tai tekstuose pakankamai dažnai pasitaikantys ir gana paprastu algoritmu identifikuojami žodžiai, žodžių formos ar kiti teksto elementai, glaudžiai, kaip manoma, susiję su tiriamąja kalbos funkcija, struktūra ar savybe. Pagal bendrą idėją ir paskirtį žymekliai geriausiai atitinka lingvistinėje tipologijoje įvestą terminą klasterinė kalbos dalis [24]. Klasterinėms kalbos dalims nekeliami jokie išankstiniai kokių nors kalbos struktūrų reikalavimai, jos apibrėžiamos remiantis vien tik jų pasiskirstymu tiriamajame tekstyne. Buvo pasirinktos trys žymeklių grupės: IT0 – visos įvardžio „tu“ formos, jos siejamos su tiesiogine kalba, LV2 – prielinksniai, reikalaujantys kilmininko linksnio ir susieti su vietos aplinkybe, ir jungtinė grupė Inag, jungianti visas žymeklių grupes, susietas su įnagininko linksniu. Šioms žymeklių grupėms atliktas jų pasiskirstymo blokuose tyrimas rodo, kad tekstai tų grupių atžvilgiu yra gana nehomogeniški. Perteklinės (angl. *excessive*) sklaidos dalis skirtingoms žymeklių grupėms skiriasi ir svyruoja nuo 47 % (žymeklių grupė Inag) iki 78 % (žymeklių grupė IT0). Kadangi žymeklių grupes sudaro labai dažni, funkciniai žodžiai, galima daryti prielaidą, kad tekstai yra gana nehomogeniški ir pagal kalbos struktūrų vartoseną, kitaip tariant, yra funkciškai nehomogeniški.

Ženklią tiriamų žymeklių grupių sklaidos blokuose dalį pavyksta paaiškinti naudojant tekstų autorystės požymį (modelis (M2)) bei papildomą informaciją apie blokų ilgį (modelis (M3)), o taip pat informaciją apie kitų žymeklių grupių pasitaikymo blokuose dažnumus (modelis (M4)). Kai kuriais atvejais beveik tokį pat, o kartais ir geresnį prognozavimo tikslumą kaip modelyje (M3) pavyksta pasiekti naudojant vien informaciją apie blokų ilgį ir kitų žymeklių grupių dažnumus (modelis (M5)). Vadinasi, žymeklių grupės yra tarpusavyje statistiškai susijusios, ir tos sąsajos bent iš dalies sietinos su autoriaus preferencijomis ir pasirinkimu. Nors dauguma veiksnių, įtrauktų į modelius (M4) ir (M5), yra neinterpretuojami, kai kuriems jų galima rasti natūralų paaiškinimą.

Kadangi šiame tyrime yra naudojami neanotuoti tekstai ir skaičiuojant pasitaikymo dažnius blokuose nenaudojama jokia informacija apie žodžių juose tvarką, tai gauti rezultatai liudija, kad tyrimai, kurie remiasi vien bendra trigramų (*n*-gramų) ir kita sakinių apribota informacija arba ištisine, tekstų gretimumą ignoruojančia tekstynų statistika, neišnaudoja reikšmingos dalies statistinės informacijos. Į statistinį tekstų nehomogeniškumą, perteklinę sklaidą reikėtų atsižvelgti darant statistines išvadas.

Aktualus ir sudėtingas tinkamų žymeklių (grupių) parinkimo ir jų susiejimo su svarbiais kalbos elementais uždavinys šiame darbe nespendžiamas. Jis turėtų remtis anotuotų tekstų statistine analize.

Literatūra

1. Abney S. 1996: Statistical methods and linguistics. In: *The Balancing Act: Combining Symbolic and Statistical Approaches to Language*, MIT Press, Cambridge, p. 1–26.
2. Agresti A. 2002: *Categorical Data Analysis*, Wiley & Sons.
3. Brasiūnaitė J. 2010: Straipsnių antraščių tekstų heterogeniškumas, *Žurnalistikos tyrimai,* 3, p. 79–98.
4. Callison-Burch Ch. and Osborne M. 2003: Statistical Natural Language Processing. In: *A Handbook For Language Engineers*, (ed. Ali Farghaly), CSLI Publications, p. 1–29.
5. Dabašinskienė I. 2009: Šnekamosios lietuvių kalbos morfologinės ypatybės, *Acta Linguistica Lithuanica*, LX, p. 1–15.
6. Johnson D.E. 2009: Getting off the GoldVarb Standard: Introducing Rbrul for Mixed-Effects Variable Rule Analysis, *Language and Linguistics Compass*, 3/1, p. 359–383.
7. Kapočiūtė-Dzikienė J., Utka A., Šarkutė L. 2014: Seimo posėdžių stenogramų tekstynas autorystės nustatymo bei autoriaus profilio sudarymo tyrimams, *Kalbotyra*, 66, p. 27–45.
8. Lapėnaitė-Gedvilė M., Radavičius M., Piaseckienė K. 2014: Dažnos žodžių formos ir linksnių vartojimas. In: *Informacinės technologijos*, KTU, Kaunas, p. 156–160.
9. Leonavičienė A. 2005: Spaudos tekstų heterogeniškumas funkcinių stilių sandūros aspektu, *Kalbotyra*, 55 (3), p. 38–46.
10. Leonavičienė A. 2006: Tiesioginė šnekamojo stiliaus kalba – spaudos tekstų konversacionalumo požymis, *Filologija*, 11, p. 48–56.
11. Leonavičienė A. 2010: Sakinių ilgis – publicistinio ir šnekamojo stiliaus sandūros tekstuose požymis, *Kalbotyra*, 62 (3), p. 95–107.
12. Mandelbrot B. B. 1953: An information theory of the statistical structure of language. In: *Communication Theory*, (ed. W. Jackson), London, p. 486–502.
13. Paolillo J.C. 2002: *Analyzing Linguistic Variation*, CSLI Publications, Stanford.
14. Piaseckienė K. 2014: *Statistiniai metodai lietuvių kalbos sudėtingumo analizėje*. Fizinių mokslų (matematikos) daktaro disertacija. Vilnius, Vilniaus universitetas.
15. Riloff E. 1995: Little Words Can Make a Big Difference for Text Classification. In: *Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Association for Computing Machinery, New York, p. 130–136.

16. Rimkutė E., Grigonytė G. 2006: Automatizuotas lietuvių kalbos morfologinio daugiareikšmiškumo ribojimas, *Kalbų studijos*, 9, p. 30–37.
17. Rimkutė E., Daudaravičius V. 2007: Morfologinis dabartinės lietuvių kalbos tekstyno anotavimas, *Kalbų studijos*, 11, p. 30–35. http://donelaitis.vdu.lt/publications/Rimkute_2007.pdf.
18. Saracevic T. 1995: Evaluation of evaluation in information retrieval. In: *SIGIR '95 Proceedings of the 18th annual international ACM SIGIR conference on Research and development in information retrieval*, Association for Computing Machinery, New York, p. 138–146.
19. SAS Institute Inc. 2013: *SAS/STAT* ® 13.1 *User's Guide*. Cary, NC: SAS Institute Inc.
20. Stamatatos E., Fakotakis N., Kokkinakis G. 2000: Text Genre Detection Using Common Word Frequencies. In: *COLING '00 Proceedings of the 18th conference on Computational linguistics,* Vol. 2, Association for Computational Linguistics, Stroudsburg, p. 808–814.
21. Utka, A. 2005a: Labai dažnų lietuvių kalbos žodžių ir žodžių formų ypatybės, *Lituanistica,* 1(61), p. 48–55.
22. Utka A. 2005b: Statistinis tekstų funkcijų nustatymas, *Žmogus ir žodis: didaktinė lingvistika*, 7(1), p. 22–25.
23. Utka A. 2006: Common words as indicators of text functions. *Prace Bałtystyczne*, 3, p. 213–224.
24. Wälchli B. 2009: Distributional Parts of Speech. In: *Transalpine Typology Meeting*, University of Bern, Bern.http://attach.matita.net/caterinamauri/sitovecchio/1389216084_Waelchli-parts-of-speech-TTM.pdf.
25. Zipf G. K. 1935: *The psycho-biology of language: an introduction to dynamic philology*, Houghton Mifflin, Boston.
26. Žalkauskaitė G. 2011: Idiolekto požymiai elektroninių laiškų leksikoje, *Kalbotyra*, 63(3), p. 149–164.
27. Žalkauskaitė G. 2012: *Idiolekto požymiai elektroniniuose laiškuose*. Humanitarinių mokslų daktaro disertacija. Vilnius: Vilniaus universitetas.
28. Žilinskienė V. 2005: Daiktavardžio ir jo gramatinių formų vartojimo dažnis lietuvių kalbos stiliuose, *Kalbos istorijos ir dialektologijos problemos,* 1, p. 379–393.

ANALYSIS OF TEXT NON-HOMOGENEITY USING MARKERS

Monika Lapėnaitė-Gedvilė, Karolina Piaseckienė, Marijus Radavičius

Abstract. The aim of the paper is to assess the distributional non-homogeneity of texts in the usage of functional words and other linguistic units. Our empirical study is based on recommended school fiction works taken from a digital library at http://ebiblioteka.mkp.emokykla.lt. Sets of frequent word forms, called markers, are made, and their frequency counts in blocks of 50 successive sentences are calculated. The frequency counts of the markers show significant excess variability (overdispersion) with respect to a text homogeneity model usually assumed in linguistics. For chosen markers, different kinds of hierarchical binomial logistic regression models with the author's identifier, the block length and the frequency counts of the remaining markers as explanatory variables are fitted to the block data in order to explain the observed overdispersion of the markers chosen.

Keywords: statistical linguistics, over-dispersion, deviance, binomial logistic regression, functional words.