ŽODŽIŲ DAŽNIŲ PASISKIRSTYMO ANALIZĖ SKIRTINGŲ ŽANRŲ LIETUVIŲ KALBOS TEKSTUOSE

Neringa Bružaitė[1], Tomas Rekašius[2]

Vilniaus Gedimino technikos universitetas. Adresas: Saulėtekio al. 11, 10223, Vilnius, Lietuva
El. paštas: [1]ner.bruzaite@gmail.com, [2]tomas.rekasius@vgtu.lt

Gauta: 2016 m. liepa Pataisyta: 2016 m. rugpjūtis Paskelbta: 2016 m. lapkritis

Santrauka. Darbe nagrinėjami skirtingų autorių ir skirtingų žanrų tekstai, parašyti lietuvių kalba. Pagrindinės mus dominančios tekstų savybės – žodžių skaičius, teksto žodyną sudarančių skirtingų žodžių skaičius ir žodžių dažniai. Žodžių dažnių pasiskirstymui tekste aprašyti taikomas struktūrinis skirstinys ir Zipfo dėsnis. Akivaizdu, kad bet kokio teksto leksinę įvairovę nusako jame vartojamų žodžių žodynas. Pademonstruota, kad redukuotame žodyne esančios informacijos užtenka darbe nagrinėtiems tekstams suskirstyti į grupes pagal žanrus ir autorius naudojant hierarchinio klasterizavimo metodą. Šiuo atveju atstumai tarp klasterių matuojami naudojant Jaccardo atstumo matą, o klasteriai apjungiami naudojant Wardo metodą.

Reikšminiai žodžiai: žodžių dažniai, struktūrinis skirstinys, Zipfo dėsnis, hierarchinis klasterizavimas, Jaccardo atstumas, Wardo metodas.

1. Įvadas

Kiekvienas tekstas gali būti suvokiamas kaip žodžių (angl. *word tokens*) seka, sudaryta iš N, $N \in \mathbb{N}$ elementų. Tada skirtingų žodžių tekste, dar vadinamų žodžių formomis (angl. *word types*), skaičius $V(N)$ įprastai daug mažesnis už bendrą žodžių skaičių N, kadangi dalis žodžių pasikartoja daugiau nei vieną kartą; bendru atveju $V(N) \leqslant N$. Šio darbo pagrindinis tyrimo objektas – dažninis žodžių sąrašas, kurį sudaro visos žodžių formos ir jų pasitaikymo dažnis tekste. Vieno autoriaus kūrinių dažniniai žodžių sąrašai dažniausiai naudojami norint nustatyti autoriaus leksikos įvairovę, vyraujančią kūrinio tematiką. Taip pat didelio tekstyno dažninis sąrašas yra nepamainomas įrankis žodynininkams. Juk dėl riboto žodyno dydžio jų kūrėjai stengiasi pirmiausia užtikrinti, kad į žodynus patektų dažniausiai vartojami žodžiai [5].

Andrius Utka savo darbe [10] aprašė dažniausiai pasikartojančių žodžių tekstuose savybes tokias kaip vidutinis žodžių ilgis, žodžių semantinė reikšmė, ir jų svarbą teksto funkcijoms identifikuoti. Priešingai nei A. Utkos darbe, čia nėra atsižvelgiama į žodžių, sudarančių dažninį sąrašą, semantinę reikšmę ar sutelkiamas dėmesys į vieną žodžių grupę. Vienas iš pagrindinių šio darbo uždavinių – aprašyti žodžių dažnių pasiskirstymą taikant *empirinį struktūrinį skirstinį* ir *klasikinį Zipfo dėsnį*.

Kitas uždavinys – tekstų klasterizavimas, kuris žinomas kaip vienas iš būdų greitai surasti informaciją. Pirmieji siūlymai panaudoti klasterizavimą informacijos gavybai pagerinti buvo pateikti 1971 metais [11]. Žinoma, kad klasterizuojami dokumentai į grupes patenka ne pagal ieškomų žodžių ar frazių buvimą/nebuvimą juose, bet pagal dokumentų turinio panašumą [11], todėl taikant hierarchinį klasterizavimą tikimasi, kad skirtingų autorių kūrinių leksika skiriasi. Šiuo tyrimu siekiama išsiaiškinti, ar įmanoma suklasterizuoti kūrinius į grupes pagal jų autorių ir (ar) žanrą, turint tik tekstus sudarančių žodžių formų sąrašą; tekstams grupuoti naudojamas hierarchinis klasterizavimas, klasteriams apjungti pritaikytas modifikuotas Wardo metodas su Jaccardo atstumo matu.

Tyrime naudojamas nedidelių tekstų duomenų rinkinys, susidedantis iš trijų autorių (Jono Biliūno, Jurgio Savickio, Bitės Vilimaitės) kūrinių. Visų pasirinktų kūrinių žanras – novelė. Novelės buvo paimtos iš klasikinės lietuvių literatūros portalo (http://antologija.lt/), skaitmeninės bibliotekos (http://ebiblioteka.mkp.emokykla.lt/), lietuvių literatūros antologijos (http://www.tekstai.lt). Taip pat į duomenų rinkinį buvo įtraukti 8 populiariosios žurnalistikos straipsniai (http://technologijos.lt/) bei baudžiamojo kodekso (BK) II–IX skyriai (http://www.baudziamasiskodeksas.lt/). Iš viso nagrinėjami 40 tekstų, kurių dydis kinta nuo 362 iki 2 683 žodžių.

2. Empirinis struktūrinis skirstinys

Prieš pradedant nagrinėti tekstus, apibrėžiama, kas yra tyrimo populiacija ir imtis. Vieną kūrinį galima laikyti statistine populiacija tik tada, kai sutelkiamas dėmesys į jį kaip literatūros vienetą. Šiame darbe populiacija apibrėžiama kaip fiksuota visų žodžių formų, kurias autorius gali panaudoti tekste, aibė S. Tada tikimybė žodžio formai w_i pasirodyti tekste lygi π_i, $\pi_i \in [0,1]$, $i = 1,2,\ldots,S$. Imtį šiame darbe atitinka vieno autoriaus kūrinys, sudarytas iš N žodžių ir $V(N)$ žodžių formų. Tokių imčių iš viso turima 40. Žodžio formos w_i dažnis N dydžio tekste žymimas $f(i,N)$, $i = 1,2,\ldots,V(N)$. Laikomasi prielaidų, kad: a) tekstų ir juos sudarančių žodžių eiliškumas nėra svarbus ir b) visi žodžiai, kurie rašomi vienodai, turi tą pačią žodžio formą.

Nemažai empirinių tyrimų rodo, kad beveik pusė žodžių yra sutinkami tekstyne tik po vieną kartą [9]. 1 lentelėje parodyta vidutinė vieną kartą pasikartojančių žodžių dalis skirtingų autorių kūriniuose.

1 lentelė. Vidutinė vieną kartą pasitaikančių žodžių dalis tekstuose

J. Biliūno	J. Savickio	B. Vilimaitės	Straipsnių	BK skyrių
44,19 %	56,51 %	54,19 %	54,04 %	22,63 %

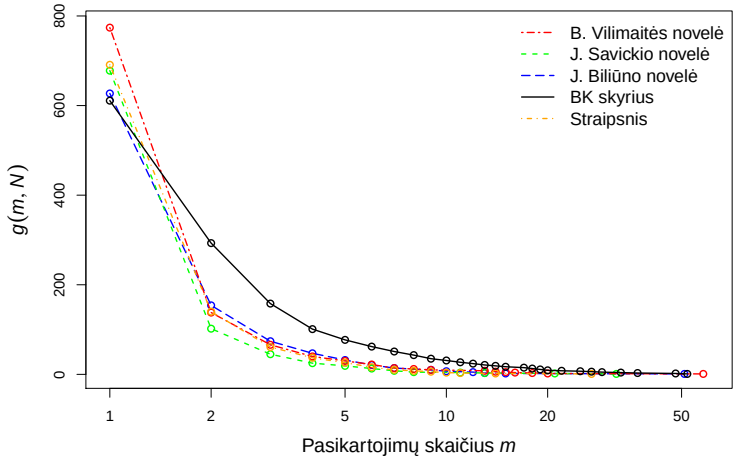
Kadangi labai didelė dalis žodžių tekste pasirodo tik po vieną kartą (iki 56,51 %), žodžių dažnių lentelės yra išretintos, todėl tekstų analizei netinka klasikiniai statistiniai metodai (pvz., χ^2 kriterijus). **Struktūrinis skirstinys** (angl. *structural type distribution*) – tai svarbi išretintų kategorinių duomenų modelių charakteristika, kurios išraiška:

$$G(\pi) = \sum_{i=1}^{S} \mathbf{1}_{[\pi_i \geqslant \pi]}. \tag{1}$$

Struktūrinis skirstinys apibrėžia žodžių formų skaičių populiacijoje su tikimybe lygia arba didesne už π [1]. Analogiškas struktūriniam skirstiniui yra **empirinis struktūrinis skirstinys** $g(m,N)$, kuris apibūdinamas kaip vienas iš neparametrinių metodų dažnių lentelėms aprašyti. Sakoma, kad tą patį pasikartojimų skaičių m turintys žodžiai sudaro klasę. Empirinio struktūrinio skirstinio reikšmės nurodo skirtingų žodžių formų, kurios pasirodo m ir daugiau kartų, skaičių N dydžio tekste, $m \geqslant 0$ [1]. Jo išraiška:

$$g(m,N) = \sum_{i=1}^{V(N)} \mathbf{1}_{[f(i,N) \geqslant m]}. \tag{2}$$

1 pav. pavaizduotas empirinio struktūrinio skirstinio reikšmių, apskaičiuotų pagal (2) formulę, grafikas skirtingų autorių ir žanrų kūriniams. Kad būtų galima palyginti, buvo parinkti tekstai, turintys panašių žodžių formų skaičių.



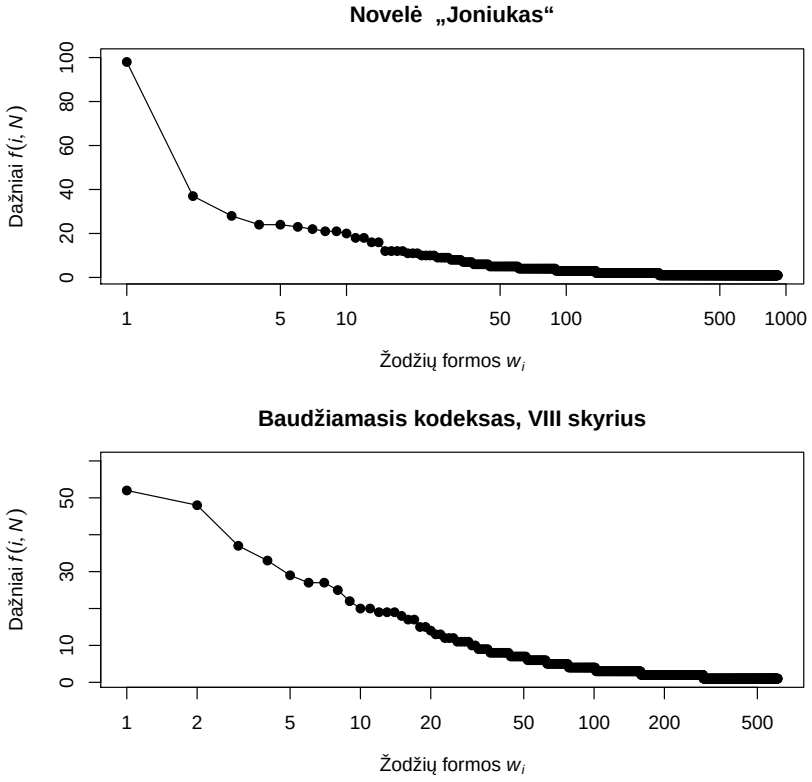
1 pav. Kai kurių tekstų empirinių struktūrinių skirstinių grafikai

Horizontali pasikartojimų klasės m ašis buvo logaritmuota dėl vaizdinio patogumo. Iš grafiko matyti, kad struktūrinis skirstinys yra asimetriškas: 1 pav. matomas labai didelis skirtingų žodžių

skaičiaus kritimas žemyn, kai iš pasikartojimų vieną ir daugiau kartų pereinama prie pasikartojimų du ir daugiau kartų. Taip pat matoma, kad struktūrinio skirstinio reikšmės yra „ištemptos“ į dešinę. Vadinasi, augant žodžių pasikartojimų skaičiui m, struktūrinio skirstinio reikšmių $g(m,N)$ mažėjimas lėtėja. Atliekant tyrimą nustatyta, kad baudžiamojo kodekso skyrių struktūrinio skirstinio reikšmės, lyginant su kitais kūriniais, gęsta lėčiau, todėl pateiktas grafikas atspindi bendrą tendenciją. Iš to matyti, kad baudžiamojo kodekso skyrių žodžių dažnių pasiskirstymas yra tolygesnis. Tai gali būti susiję su administraciniu stiliumi parašytų tekstų žodynų lakoniškumu ir ribotu žodžių pavartojimu.

3. Žodžių dažnių modeliavimas Zipfo dėsniu

Vienas iš mįslingiausių faktų apie žmonių kalbą yra ir vienas paprasčiausių: žodžiai atsiranda pagal sistemingą dažnių pasiskirstymą – yra keli aukšto dažnio žodžiai, kurie sudaro didžiąją dalį visų žodžių tekste, ir yra daug žemo dažnio žodžių [8]. Pavyzdžiui, 2 pav. kiekvienas taškas pagal horizontalią ašį atitinka atskirą žodžio formą, o vertikalioje ašyje atidėtas tos žodžio formos dažnis. Čia žodžių formos išrikiuotos dažnių nedidėjimo tvarka, o horizontali ašis logaritmuota.



2 pav. Žodžių formų dažniai dviejų skirtingų žanrų tekstuose

Pirmasis dažniausių kalbos žodžių formų neįprastas ypatybes užrašė amerikiečių lingvistas George Kingsley Zipf. Jis nustatė, kad egzistuoja matematinė priklausomybė tarp žodžio dažnio ir jo vietos dažniniame sąraše bei tarp žodžio dažnio ir žodžių, turinčių tą dažnį, skaičiaus; teigiama, kad Zipfo dėsnis yra universalus, t. y. tinka bet kurios kalbos dažniniam sąrašui [9, 10]. Šiame darbe patikrinsime, ar Zipfo dėsnį galima pritaikyti trumpiems lietuvių kalbos tekstams.

2 lentelėje pateikiamas 3 kūrinių išrikiuotas penkiolikos dažniausiai pasikartojančių žodžių dažnių sąrašas. Kaip pavyzdys parenkama J. Biliūno novelė „Joniukas“ ($N = 1891$), baudžiamojo kodekso VIII skyrius ($N = 1771$), kuriame kalbama apie bausmės skyrimą, ir straipsnis apie IT technologijas ($N = 1025$). 2 lentelėje kūrinyje „Joniukas“ dažniausiai pasikartojantis žodis *ir* turi rangą $z = 1$, antras pagal dažnumą žodis – $z = 2$ ir taip toliau iki žodžio, turinčio rangą $z = V(N)$. Jei keli žodžiai pasikartoja tiek pat kartų, tai tokiems žodžiams suteikiama vienoda rango reikšmė – rangų aritmetinis vidurkis.

2 lentelė. Dažniausiai pasikartojantys žodžiai tekstuose

„Joniukas“			Baudž. kodeksas			Straipsnis apie IT		
z	Žodis	$f_z(z,N)$	z	Žodis	$f_z(z,N)$	z	Žodis	$f_z(z,N)$
1	ir	98	1	bausmės	52	1	ir	27
2	joniukas	37	2	ar	48	2	iš	14
3	kaip	28	3	ir	37	3,5	di	11
4,5	iš	24	4	veiką	33	3,5	go	11
4,5	tik	24	5	už	29	5	kad	10
6	bet	23	6,5	bausmę	27	6,5	su	9
7	kad	22	6,5	teismas	27	6,5	tai	9
8,5	ant	21	8	nusikalstamą	25	8,5	kaip	8
8,5	jis	21	9	arba	22	8,5	žmogaus	8
10	jam	20	10,5	atsakomybę	20	10	į	7
11,5	int	18	10,5	veika	20	13,5	apie	6
11,5	motutė	18	13	bausmė	19	13,5	kompiuteriai	6
13,5	jį	16	13	laisvės	19	13,5	metais	6
13,5	savo	16	13	straipsnio	19	13,5	o	6
16,5	jo	12	15	padaryta	18	13,5	savo	6

Pastaba. Harald Baayen knygoje [1] žodžiai, turintys tą patį dažnį, turi skirtingas rango reikšmes, t. y. kiekvienam tolesniam žodžiui priskiriama vis didesnė rango reikšmė.

Zipfo rango-dažnio pasiskirstymas (angl. *Zipf rank-frequency distribution*) yra ekvivalentus empiriniam struktūriniam skirstiniui [1]:

$$g(m,N) = z \Leftrightarrow f_z(z,N) = m,$$

čia $f_z(z,N)$ – žodžio, turinčio rangą z, pasikartojimų skaičius. Pavyzdžiui, novelėje „Joniukas“ žodžio didžiausias dažnis yra $m = 98$. Šį dažnį turinčio žodžio empirinio struktūrinio skirstinio reikšmė lygi vienam, nes nėra jokio kito žodžio, kuris pasikartotų 98 ir daugiau kartų:

$$g(98,1891) = 1.$$

Tada atitinkamai Zipfo rangą $z = 1$ turintis žodis tekste pasikartoja $m = 98$ kartus:

$$f_z(1,1891) = 98.$$

Zipfo dėsnis įprastai aiškinamas „mažiausių pastangų“ (angl. *least effort*) kalboje principu [3]. A. Utka teigia, kad „<…> dažniausių žodžių formų bruožas – trumpumas. Žodžiai turi tendenciją ilgėti leidžiantis dažniniu sąrašu žemyn, todėl pačios trumpiausios žodžių formos yra dažninio sąrašo viršuje. <…> Šis reiškinys susijęs su kalbos ekonomija ir efektyvumu, nes priešingu atveju, jei ypač dažnai vartojamos žodžių formos būtų ilgos, kalba būtų labai neefektyvi bendravimo priemonė laiko ir energijos eikvojimo požiūriu“ [10]. Iš 2 lentelės galima pastebėti, kad J. Biliūno novelėje „Joniukas“ ir straipsnyje apie IT technologijas dažniausiai pasikartojantys žodžiai yra labai trumpi, sudaryti iš 4 ar mažiau simbolių. Tačiau nagrinėtame baudžiamojo kodekso skyriuje didžiąją daugumą dažniausiai vartojamų žodžių sudaro ilgesni nei 4 simbolių žodžiai. Baudžiamojo kodekso skyrių tekstai gali pažeisti „mažiausių pastangų“ kalboje principą, todėl, kad tokių tekstų kalba šabloniška, joje vartojama daug vienodų standartinių frazių, kurios padeda visiems skaitytojams vienodai suprasti tekstą. Taigi, administracinio stiliaus viena pagrindinių funkcijų yra *aiškaus pranešimo*, o ne efektyvaus bendravimo, kuri prieštarauja Zipfo dėsnio formuluotei.

Karolinos Piaseckienės disertacijoje [9] buvo nagrinėjamas Zipfo dėsnis sakinių struktūroms, t. y. skaičiuojami ne žodžių, sudarančių tekstą, dažniai, o pasikartojančių sakinių struktūrų dažniai. Buvo nustatyta, kad gautosioms sakinių struktūroms galioja Zipfo dėsnis. Šiame darbe taikomas Zipfo dėsnis tekstą sudarančių žodžių dažniams.

Ryšys tarp dažnio rango z ir žodžio dažnio $f_z(z,N)$, turinčio rangą z, N dydžio tekste yra vadinamas **Zipfo dėsniu**:

$$f_z(z,N) = \frac{C}{z^{\alpha}}, \tag{3}$$

kuris yra atskiras atvejis *Zipfo-Mandelbroto* dėsnio:

$$f_z(z,N) = \frac{C}{(z+\beta)^{\alpha}}, \tag{4}$$

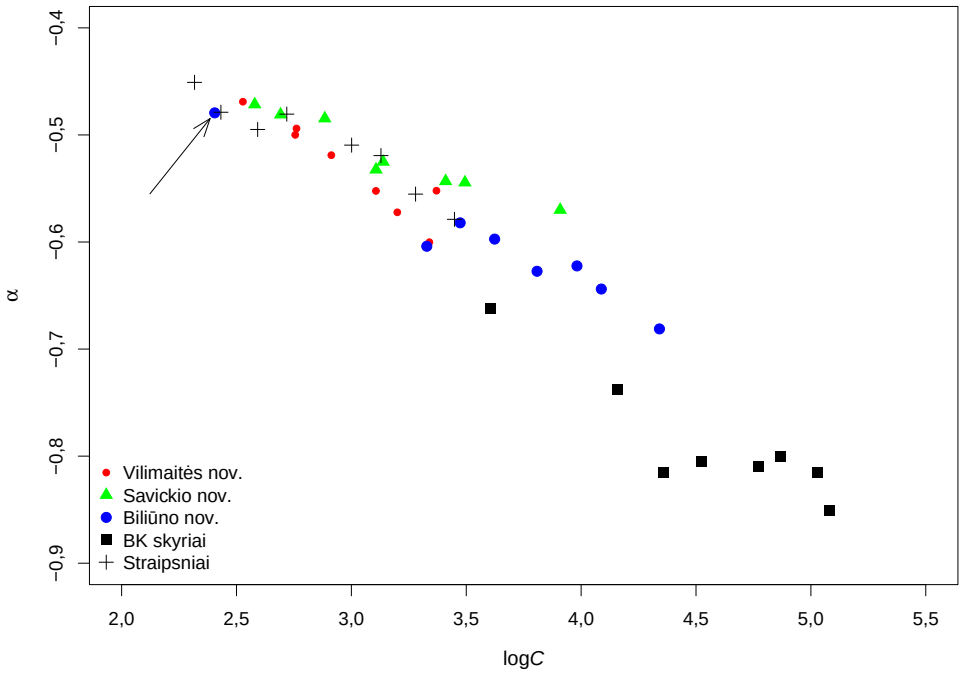
čia $C > 0$ – normuojanti konstanta, β – parametras, rodantis nukrypimą nuo Zipfo dėsnio, $\alpha > 0$ – žodžių gausumo parametras [9]. Kai $\beta = 0$ gaunamas klasikinis Zipfo dėsnis. (3) formulėje aprašytas modelis prognozuoja spartų dažnių mažėjimą tarp dažniausių žodžių tekste ir dažnių mažėjimo lėtėjimą, kai rangas didėja t. y. tada, kai žodžių pasikartojimų skaičius mažėja. Taigi, Zipfo dėsnis prognozuoja, kad labai didelis skaičius žodžių turi labai panašius, nebūtinai sveikaisiais skaičiais išreikštus, dažnius [3].

Nagrinėjamiems tekstams bus taikomas klasikinis Zipfo dėsnis. Zipfo dėsnio patogi savybė yra ta, kad logaritmuojant abi (3) formulės puses, gaunama tiesinė funkcija:

$$\log f_z(z,N) = \log C - \alpha \log z, \tag{5}$$

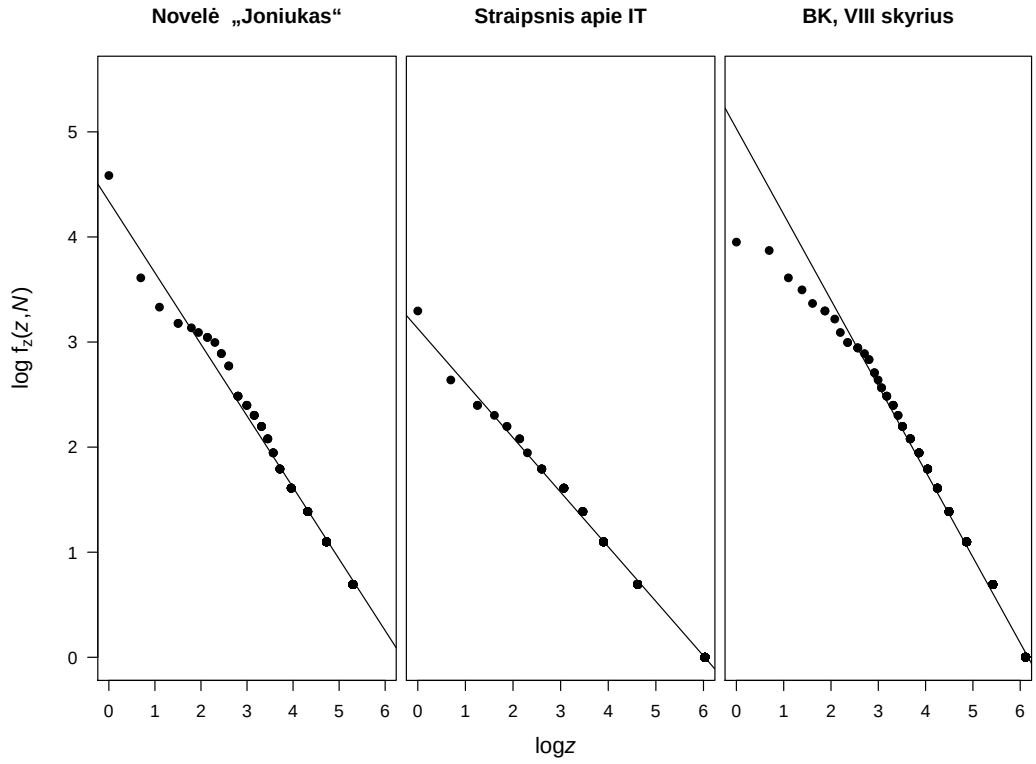
čia $\log C$ – tiesinės funkcijos laisvasis narys, α – modelio parametras, kuris nusako tiesinės regresijos krypties koeficientą [1]. Parametrai α ir $\log C$ randami *mažiausių kvadratų metodu.*

Buvo įvertinti visų nagrinėjamų tekstų Zipfo dėsnio parametrai. Parametrų α ir $\log C$ įverčiai vaizduojami 3 pav. Rodykle pažymėtas taškas atitinka J. Biliūno novelę „Kliudžiau“, kuri pagal Zipfo dėsnio parametrus išsiskiria iš kitų J. Biliūno novelių. Kadangi tiesinės regresijos modelio parametrų kovariacija neigiama, pats savaime taškų išsibarstymas sklaidos diagramoje jokios ypatingos prasmės neturi. Kuo mažesnė parametro α reikšmė (ir didesnė $\log C$ reikšmė), tuo mažesnis retai pasitaikančių žodžių tekste skaičius. Matyti, kad Zipfo dėsnio parametrai novelėms ir straipsniams yra panašūs, tačiau baudžiamojo kodekso yra akivaizdžiai nutolę nuo kitų kūrinių parametrų. Tai galima paaiškinti tuo, kad baudžiamojo kodekso skyriuose unikalių žodžių skaičius, lyginant su kitų žanrų tekstais, yra mažesnis.



3 pav. Visų nagrinėjamų tekstų Zipfo dėsnio parametrų sklaidos diagrama

Galima pasižiūrėti, kaip modelio prognozuojami logaritmuoti dažniai skiriasi nuo tikrųjų dažnių logaritmo. 4 pav. atidėtos logaritmuotos $(z,\ f_z(z,N))$ taškų poros. Taškai grafike – stebėti žodžių dažniai tekstuose, tiesė – prognozuoti dažniai pagal Zipfo dėsnį. Matyti, kad pritaikytas modelis gana gerai prognozuoja mažus logaritmuotus dažnius, tačiau keliems didžiausiems dažniams daro paklaidą.



4 pav. Žodžių dažnių log-log grafikas

4. Redukuotų žodynų sudarymas ir jų pritaikymas tekstams klasterizuoti

Iš 2 lentelės matyti, kad dažniausiai pasikartojantys žodžiai tekstuose – funkcinės paskirties nekaitomi žodžiai (jungtukai, įvardžiai ir pan.). Paprastai tokių didelį dažnį turinčių žodžių tekste nėra daug ir labai daug žodžių pasikartoja tik po vieną kartą (žr. 2 pav.). Vieno teksto skirtingų žodžių aibę pavadinkime *žodynu*. Žodynas nusako to teksto leksinę įvairovę. Akivaizdu, kad funkciniai žodžiai tekste neperteikia daug informacijos, o vieną kartą pasikartojantys žodžiai yra netipiniai. Galima sudaryti **redukuotą žodyną** – tokį žodyną, kuriame nėra žodžių, tenkinančių sąlygas:

1. žodis turi didžiausią pasikartojimų skaičių, $\max\limits_{i=1,\ldots,V(N)} f(i,N)$;
2. žodis tekste pasirodo tik vieną kartą, $f(i,N)=1$, $i=1,2,\ldots,V(N)$.

Daroma prielaida, kad redukuotas žodynas atspindi autoriui būdingiausius žodžius. Taip sukuriama nauja charakteristika, naudojama tekstams apibūdinti. Gautus redukuotus žodynus galima panaudoti tekstams palyginti: jų panašumui ir skirtumui nustatyti. Pavyzdžiui, Vaido Balio disertacijoje [2] buvo naudojamas redukuoto žodyno analogas – teksto raktažodžiai – moksliniams kūriniams klasterizuoti. 3 lentelėje pateikiami vidutiniai žodynų ir vidutiniai redukuotų žodynų dydžiai. Matyti, kad redukuoti žodynai vidutiniškai 4,1–5,9 kartų mažesni už novelių ir publicistinių straipsnių žodynus bei vidutiniškai 2,3 kartų mažesni už baudžiamojo kodekso skyrių žodynus.

3 lentelė. Vidutiniai nagrinėjamų tekstų žodynų ir redukuotų žodynų dydžiai

	J. Biliūno	J. Savickio	B. Vilimaitės	Straipsnių	BK skyrių
Vidutinis žodyno žodžių skaičius	690,63	808,50	482,38	480,13	441,13
Vid. redukuoto žodyno žodžių skaičius	168,13	138,13	89,25	102,50	188,75
Žodyno sumažėjimas, kartais	4,11	5,85	5,41	4,68	2,34

Klasifikavimas be apmokymo imties (klasterizavimas) naudojamas tuo atveju, kai turimas objektų sąrašas, bet nežinomas kintamasis, kuris aprašytų, kuriai klasei kiekvienas objektas priklauso.

Vienas iš klasifikavimo metodų yra hierarchinis klasterizavimas, kuris atliekamas dviem būdais: aglomeratyviu (angl. *agglomerative*) ir skaidančiu (angl. *divisive*). Aglomeratyvus hierarchinis klasterizavimas pirmiausia atskiria kiekvieną objektą į jo individualų klasterį, todėl klasterių skaičius pradžioje lygus objektų skaičiui. Toliau artimiausi klasteriai jungiami tol, kol gaunamas vienas klasteris su visais objektais [4, 11]. Ką reiškia „artimiausias klasteris“ priklauso nuo pasirinkto atstumo mato: euklidinio, Manhatano ir t. t.

Šiuo atveju klasterizuojamus tekstus atitinka žodynai, t. y. žodžių aibės, todėl jų panašumui įvertinti galima naudoti Jaccardo panašumo indeksą

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|},$$

čia A ir B – aibės, o $|A|$ – aibės A elementų skaičius [7]. Naudojant Jaccardo panašumo indeksą, apibrėžiamas Jaccardo atstumo matas:

$$J_D(A,B) = 1 - J(A,B) = 1 - \frac{|A \cap B|}{|A \cup B|}. \tag{6}$$

Jaccardo atstumas įgyja reikšmes iš intervalo [0,1]. Jei atstumas lygus 1, tai dvi aibės neturi bendrų elementų, jei atstumas lygus 0, tai dvi aibės yra vienodos. Šiuo atveju aibės A ir B atitinka dviejų skirtingų tekstų žodynus.

Egzistuoja kelios klasterių jungimo taisyklės, pavyzdžiui, *artimiausio kaimyno metodas* (angl. *single linkage*), *tolimiausio kaimyno metodas* (angl. *complete linkage*). Šiame darbe naudojamas Wardo minimalios dispersijos metodas. Pagal Fionn Murtagh [6], Wardo metodu atstumas tarp dviejų klasterių q ir q', kuriuos sudaro atitinkamai m_q ir $m_{q'}$ skaičius elementų, apskaičiuojamas kaip

$$d(q,q') = \frac{m_q m_{q'}}{m_q + m_{q'}} \|q - q'\|^2, \tag{7}$$

čia $\|q-q'\|^2$ – euklidinio atstumo tarp q ir q' kvadratas. Šiame darbe vietoje euklidinio atstumo naudojamas formulėje (6) apibrėžtas Jaccardo atstumas, o q ir q' žymi klasterius, sudarytus iš žodynų.

Kadangi po kiekvieno klasių apjungimo žingsnio reikia perskaičiuoti atstumų matricą, labai patogi formuluotė, kuri apima visus prieš tai išvardytus klasių apjungimo metodus, yra Lance-Williams atstumų atnaujinimo (angl. *dissimilarity update*) formulė, kuri parodo kaip keičiasi atstumų matricos reikšmės po kiekvieno klasterių apjungimo [6]. Jei objektai i ir j yra apjungiami į klasterį $i \cup j$, tada atstumai tarp naujai gauto klasterio ir kitų klasterių gaunami pagal tokią formulę:

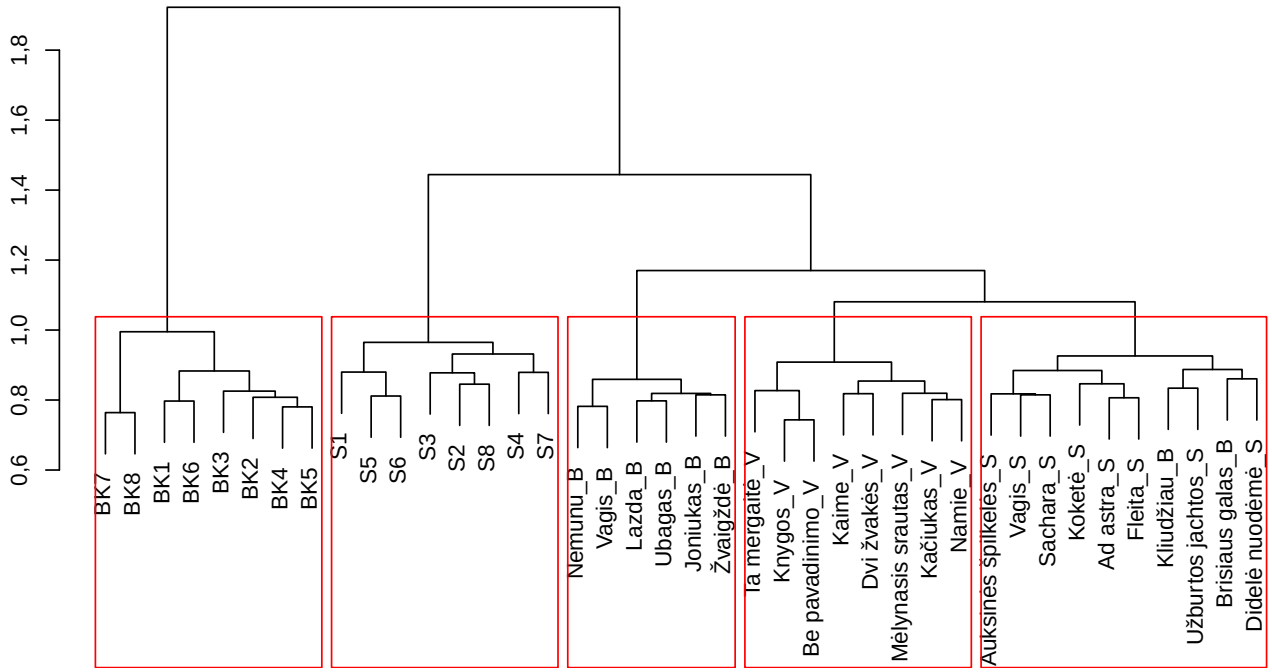
$$d(i \cup j, k) = \alpha_i d(i,k) + \alpha_j d(j,k) + \beta d(i,j) + \gamma |d(i,k) - d(j,k)|. \tag{8}$$

Wardo metodo atveju

$$\alpha_i = \frac{|i| + |k|}{|i| + |j| + |k|}, \quad \alpha_j = \frac{|j| + |k|}{|i| + |j| + |k|}, \quad \beta = -\frac{|k|}{|i| + |j| + |k|}, \quad \gamma = 0,$$

čia $|i|$ – objektų skaičius klasteryje i. Lance-Williams išraiškas kitiems klasių apjungimo metodams galima rasti [6].

5 pav. pateikta dendrograma, gauta klasterizavimui naudojant redukuotus tekstų žodynus. BK1–BK8 žymi baudžiamojo kodekso skyrius, o S1–S8 – populiariosios žurnalistikos straipsnius. Atstumai tarp skirtingų kūrinių apskaičiuojami naudojant Jaccardo atstumo matą, o klasteriai jungiami naudojant Wardo taisyklę. Galima pastebėti, jei tekstai būtų klasterizuojami į dvi grupes, tai grožinės literatūros J. Biliūno, B. Vilimaitės, J. Savickio kūriniai bei populiariosios žurnalistikos straipsniai būtų vienoje grupėje, o baudžiamojo kodekso skyriai – kitoje. Klasterizuojant tekstus į tris grupes, į atskirą grupę atsiskiria visi publicistiniai straipsniai. Grožinės literatūros klasteryje galima išskirti tris mažesnius klasterius, kuriuose kūriniai beveik tiksliai sugrupuoti pagal jų autorių. Išimtį sudaro J. Biliūno novelės „Kliudžiau“ ir „Brisiaus galas“, kurios priskiriamos prie J. Savickio novelių. Kūriniams klasterizuoti naudojant pradinius žodynus gaunamas panašus rezultatas: tik viena J. Biliūno novelė „Kliudžiau“ buvo priskirta prie J. Savickio novelių. Tai, kad šis kūrinys išsiskiria iš kitų J. Biliūno novelių, buvo galima pastebėti ir iš Zipfo dėsnio parametrų (žr. 3 pav.).



5 pav. Visų nagrinėtų tekstų dendrograma

5. Rezultatų apibendrinimas

Buvo nagrinėjami nedidelės apimties skirtingų žanrų lietuvių kalbos tekstai. Gauta, kad Zipfo dėsnis visiems nagrinėtiems tekstams gana gerai prognozuoja mažus žodžių dažnius, o didžiausios paklaidos daromos prognozuojant didelius dažnius. Tačiau žodžių dažnių prognozės tikslumui įtaką daro rangavimo taisyklė. Čia naudojama rangavimo taisyklė, pagal kurią tą patį pasikartojimų skaičių turintiems žodžiams priskiriama vienoda rango reikšmė – rangų aritmetinis vidurkis. Taikant rangavimo taisyklę, pagal kurią žodžių dažnių rangai nevidurkinami, Zipfo dėsnio parametrų įverčiams didesnę įtaką turėtų retų žodžių dažniai.

Taip pat pastebėta, kad pagal Zipfo dėsnio parametrų reikšmes baudžiamojo kodekso skyriai išsiskiria iš kitų nagrinėtų tekstų, o taikant empirinį struktūrinį skirstinį baudžiamojo kodekso skyrių struktūrinio skirstinio kreivės gęsta lėčiau nei kitų nagrinėjamų tekstų. Todėl baudžiamojo kodekso skyrių žodžių dažnių pasiskirstymas yra tolygesnis. Tai galima paaiškinti tuo, kad teisinė kalba dėl savo specifikos turi kitiems žanrams nebūdingų savybių (dažnas santrumpų naudojimas, pasikartojantys terminai ir t. t.), ir dėl to didelė dalis žodžių turi didesnį dažnį nei kitų žanrų tekstuose.

Klasterizavimui naudojant redukuotus žodynus, visi tekstai tiksliai suskirstomi į grupes pagal žanrą, o tik 2 iš 24 grožinės literatūros kūrinių neteisingai priskiriami kitam autoriui. Tai rodo, kad turint net ir nevisą informaciją apie teksto žodyną, galima gauti gana gerus klasterizavimo rezultatus.

Literatūra

[1] Baayen R. H. *Word Frequency Distributions*. Nijmegen universitetas. Kluwer Academic Publishers. Nyderlandai, 2001. p. 10, 13–14, 47.

[2] Balys V., *Mokslinės terminijos matematiniai modeliai ir jų taikymas leidinių klasifikavime* [interaktyvus]. Daktaro disertacija, 2009. Žiūrėta [2016-06-29]. <http://www.mii.lt/files/mii_dis_2009_balys.pdf>

[3] Baroni M., *39 Distributions in text* [interaktyvus]. Straipsnis, 2006. Žiūrėta [2016-06-24]. p. 12, 17. <http://sslmit.unibo.it/ baroni/termsett/05_1/hsk_39_dist_rev1.pdf>

[4] Yim O. ir Ramdeen K. T. *Hierarchical Cluster Analysis: Comparison of Three Linkage Measures and Application to Psychological Data* [interaktyvus]. Prancūzija, 2015. Žiūrėta [2016-06-25]. p. 9-11, 13. <http://www.tqmp.org/RegularArticles/vol11-1/p008/p008.pdf>

[5] Kazlauskienė A., Rimkutė E., Utka A. *Kiekybiniai tyrimai kalbotyroje (III)* [interaktyvus]. Straipsnis. Žiūrėta [2016-07-05]. p. 3. <http://donelaitis.vdu.lt/lkk/pdf/III_dalis.pdf>

[6] Murtagh F. *Correspondence Analysis and Data Coding with Java and R.* Leidykla Chapman & hall/crc. Boca Raton, Florida, 2005. p. 49, 51.

[7] Niwattanakul S. ir kt. *Using of Jaccard Coefficient for Keywords Similarity* [interaktyvus]. Honkongas, 2013. Žiūrėta [2016-07-04]. p. 2. <http://www.iaeng.org/publication/IMECS2013/IMECS2013_pp380-384.pdf>

[8] Piantadosi S. T. *Zipf's word frequency law in natural language: a critical review and future directions* [interaktyvus]. Apžvalga. 2015. Žiūrėta [2016-06-25]. p. 1. <https://colala.bcs.rochester.edu/papers/piantadosi2014zipfs.pdf>

[9] Piaseckienė K. *Statistiniai metodai lietuvių kalbos sudėtingumo analizėje* [interaktyvus]. Daktaro disertacija. Vilnius, 2014. Žiūrėta [2016-06-27]. p. 20–21, 60–61, 90. <http://www.mii.lt/files/mii_dis_2014_piaseckiene.pdf>

[10] Utka A. *Statistinis tekstų funkcijų nustatymas* [interaktyvus]. Daktaro disertacija. Kaunas, 2004. Žiūrėta [2016-06-27]. p. 37, 39. <http://donelaitis.vdu.lt/~andrius/sites/default/files/files/A_Utka-disertacija.pdf>

[11] Žalinauskas M. *Individualiai klasifikuotų dokumentų klasterizavimo metodas* [interaktyvus]. Magistro darbas. Kaunas, 2006. Žiūrėta [2016-06-25]. p. 8–9, 23. http://vddb.library.lt/fedora/get/LT-eLABa-0001:E.02~2006~D_20060522_143851-15319/DS.005.0.02.ETD>

STATISTICAL ANALYSIS OF WORD FREQUENCY DISTRIBUTION IN LITHUANIAN TEXTS OF DIFFERENT GENRES

Neringa Bružaitė, Tomas Rekašius

Abstract. The paper examines Lithuanian texts of different authors and genres. The main points of interest – the number of words, the number of different words and word frequencies. Structural type distribution and Zipf's law are applied for describing the frequency distribution of words in the text. It is obvious that the lexical diversity of any text can be defined by different words that are used in the text, also called vocabulary. It is shown that the information contained in a reduced vocabulary is enough for dividing the texts analyzed in this article into groups by genre and author using a hierarchical clustering method. In this case, distances between clusters are measured using the Jaccard distance measure, and clusters are aggregated using the Ward method.

Keywords: word frequencies, structural distribution, Zipf's law, hierarchical clustering, Jaccard distance, Ward method.