Lietuvos matematikos rinkinys. LMD darbai
51 tomas, 2010, 307–312

ISSN 0132-2818
www.mii.lt/LMR/

Lietuviškų tekstų stilių palyginimas remiantis universalių kiekybinių charakteristikų statistine analize

Karolina Piaseckienė[1,2], Marijus Radavičius[1,2], Raimundas Stiklius[1]

[1] *Šiaulių universitetas*
P. Višinskio g. 19, 77156 Šiauliai
[2] *Matematikos ir informatikos institutas*
Akademijos g. 4, LT-08663 Vilnius
E. paštas: karol@delfi.lt; mrad@ktl.mii.lt; raimundas19@gmail.com

Santrauka. Lietuvių kalba yra gana sudėtinga ir lanksti, ir tai gerokai apsunkina efektyvių algoritmų kūrimą automatiniam lietuviškų tekstų apdorojimui. Tekstų stiliaus ypatybių nagrinėjimui buvo pasirinktos *universalios kiekybinės charakteristikos*, kurios yra nesusijusios su teksto turiniu ir gali būti suskaičiuotos bet kuriam tekstui. Šiame straipsnyje parodoma, kaip matematinės statistikos pagalba galima atskirti ir interpretuoti lietuvių kalbos stilius. Atlikti logtiesinių modelių tyrimai parodo raidinės ir garsinės struktūros sąryšį su moksliniu ir grožiniu stiliumi.

Raktiniai žodžiai: universalios kiekybinės charakteristikos, logtiesiniai modeliai, statistiškai reikšmingi sąryšiai.

Įvadas

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

Matematinė lingvistika, anot Geoffrey K. Pullum ir András Kornai, yra matematinių struktūrų ir metodų, kurie yra svarbūs lingvistikai, analizė [3].

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

Dabartinę lietuvių kalbą, įvairius jos stilius reprezentatyviai atspindi „Dabartinės lietuvių kalbos tekstynas“, kurį sudaro 100 mln. žodžių ir kuris yra plačiai Lietuvoje naudojama duomenų bazė (plačiau žr. http://donelaitis.vdu.lt/).

Lietuvių kalboje ypač aktuali morfologinio daugiareikšmiškumo problema [4]. Taip pat tiesiogiai negalima pasinaudoti ir kitose šalyse jau sukurta automatinės sintaksinės analizės programine įranga, nes lietuvių kalbai būdingas didelis kaitomumas ir laisva žodžių tvarka sakinyje. Ši problema sprendžiama lietuvių kalbos automatinėje

sintaksinėje analizėje į vieną visumą sujungiant visas tris gramatikos sritis – morfologiją, sintaksę ir semantiką [7].

XX amžiaus pabaigoje programinės įrangos pramonėje bei akademiniuose sluoksniuose susidomėta statistinių metodų taikymu mašininiame vertime [2].

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

Be žinomų 1997, 1998 metais išleistų L. Grumadienės ir V. Žilinskienės „Dažninių dabartinės rašomosios lietuvių kalbos žodynų“, 2009 metais paskelbta ir elektroninė dažninio žodyno versija – A. Utkos sudarytas „Dažninis rašytinės lietuvių kalbos žodynas“, kuriame pateikiami ne tik žodžių, bet ir kaitomų žodžių formų dažniai [6].

Savo straipsniuose A. Utka nagrinėja dažniausių lietuvių kalbos žodžių ir žodžių formų savybes ir jų svarbą teksto analizei. Anot autoriaus, dažniausi struktūriniai teksto vienetai yra tiesiogiai susiję su teksto funkcijomis, todėl gali būti laikomi reikšmingais teksto funkcinių ypatybių rodikliais [5].

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

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

Pirmame skyrelyje trumpai supažindinsime su logtiesinių modelių dažnių lentelės interpretacija. Antrame skyrelyje pateikiami raidžių ir garsų logtiesinių modelių tyrimo rezultatai.

1 Trimatės dažnių lentelės logtiesiniai modeliai

Kadangi šis darbas yra taikomasis, o nuoseklus logtiesinių modelių matematiniu pagrindu išdėstymas yra gana ilgas, tai apsiribosime tik trimačio kategorinio požymio logtiesinių modelių ir jų interpretacijos trumpu aprašymu. Išsamus aprašymas yra pateiktas klasikinėje monografijoje [1].

Tarkime, kad turime trijų kategorinių (vardinių) požymių, $A \in \{1, \dots, k\}$, $B \in \{1, \dots, m\}$ ir $C \in \{1, \dots, n\}$, dažnių lentelę, sudarytą iš tiriamuose duomenyse stebėtų tų požymių dažnumų. Pažymėkime Y_{ijs} stebėtą požymių (A, B, C) kombinacijos (i, j, s) dažnį, o μ_{ijs} tegu žymi jo vidurkį: $\mu_{ijs} = \mathbf{E}Y_{ijs}$.

Tarkime, kad $\mu_{ijs} > 0$ su visais (i, j, s). Tada galima apibrėžti

$$\lambda_{ijs} = \ln \mu_{ijs} = u_0 + u_i^A + u_j^B + u_s^C + u_{ij}^{AB} + u_{is}^{AC} + u_{js}^{BC} + u_{ijs}^{ABC}. \tag{1}$$

Lygybės (1) kartu su parametru u identifikuojamumo sąlyga: su visais i, j, s

$$0 = u_k^A = u_m^B = u_n^C = u_{kjs}^{ABC} = u_{ims}^{ABC} = u_{ijn}^{ABC},$$
$$0 = u_{kj}^{AB} = u_{im}^{AB} = u_{ks}^{AC} = u_{in}^{AC} = u_{ms}^{BC} = u_{jn}^{BC},$$

1 lentelė.

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

apibrėžia pilnąjį („prisotintą“, angl. *saturated*) logtiesinį modelį (simboliškai jis žymimas [ABC]). Įvedus papildomus apribojimus parametrams u, gaunami šio modelio daliniai atvejai, kurie turi atitinkamą interpretaciją. Tokiu būdu gautų trimačių logtiesinių modelių specifikacijos ir simboliniai pažymėjimai pateikti 1 lentelėje.

0 modelis: požymiai A, B ir C yra tarpusavyje nepriklausomi.

1 modelis: požymis A nepriklauso nuo požymių poros (B, C), bet požymiai B ir C yra priklausomi. 2 ir 3 modeliai turi analogišką interpretaciją kaip ir 1 modelis su akivaizdžiais pakeitimais.

4 modelis: A ir C bei B ir C yra priklausomi. Šis modelis nusako požymių A ir B *sąlyginį nepriklausomumą*, kai žinomos C reikšmės, t.y., A ir B tarpusavio priklausomumas pasireiškia tik per C. 5 ir 6 modeliai turi analogišką interpretaciją kaip ir 4 modelis su akivaizdžiais pakeitimais.

7 modelis neturi paprastos interpretacijos nepriklausomumo terminais. Šiuo atveju galimybių santykiai dvimatėse požymių A ir B lentelėse, kai duota C reikšmė, nuo C reikšmės nepriklauso.

Svarbu pabrėžti, kad logtiesinio modelio lygtis (1) ir jos daliniai atvejai, aprašyti 1 lentelėje, nusako tik stebėtų dažnių vidurkio pavidalą. Kategorinių požymių analizėje paprastai laikoma, kad stebėtų dažnių skirstinys yra polinominis arba polinominių skirstinių sandauga, arba stebėti dažniai yra tarpusavyje nepriklausomi Puasono atsitiktiniai dydžiai. Vieno iš šių skirstinių pasirinkimas paprastai gana silpnai įtakoja statistines išvadas.

2 Logtiesiniai raidžių ir garsų modeliai

Darbe analizuojami lietuviškų tekstų duomenys buvo gauti iš atsitiktinai paimtų skirtingų tekstų (kūrinių) lietuvių kalba, laisvai prieinamų vartotojams elektroninėje formoje. Taigi, kas yra lietuviškas tekstas ir kas yra atitinkama tiriamoji populiacija, kol kas nėra tiksliai apibrėžta. Reikia pažymėti, kad tai – sudėtingas uždavinys. Vienas iš galimų sprendimų – naudoti Lingvistikos centre sudarytą ir jo palaikomą tekstyną (žr. http://donelaitis.vdu.lt/). Deja, vartotojas turi labai ribotas galimybes, ir todėl dabartinis tekstynas yra netinkamas sudėtingesniems tyrimams atlikti.

Buvo iškelti tokie uždaviniai:

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

2 lentelė.

Maximum likelihood analysis of variance					
Source	DF	Pr > ChiSq	Source	DF	Pr > ChiSq
bl	1	0.0035	psb*nr	16	0.0128
bn	1	<.0001	i5*psb*nr	64	<.0001
bl*bn	1	0.0508	prd	1	<.0001
i5	4	<.0001	i5*prd	4	0.0003
i5*bn	4	<.0001	i5*prd*nr	64	<.0001
b	1	<.0001	i5*b*nr	64	<.0001
i5*b	4	0.0107	bl*nr	16	0.0002
bp	1	<.0001	i5*bl*nr	64	0.0005
by	1	<.0001	bp*nr	16	0.0328
i5*bl	4	0.0176	bp*bn*nr	16	<.0001
i5*by	4	<.0001	i5*bp*nr	64	<.0001
i5*bp*bn	4	0.0144	bl*bn*nr	16	<.0001
i5*bp*by	4	<.0001	st*i5*by	4	<.0001
nr	16	<.0001	by*nr	16	<.0001
i5*nr	64	<.0001	bp*by	1	0.0314
psb	1	<.0001	**Likelihood Ratio**	**361**	**0.0685**

b) Nustatyti tose proporcijose pasireiškiančius skirtumus tarp grožinės ir mokslinės literatūros. Klausimas: Kuo skiriasi grožinės ir mokslinės literatūros tekstai?

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

Visos raidės buvo suskirstytos į balses (kintamasis *b*) bei priebalses ir joms priskirti atitinkami požymiai, susiję su rašyba arba tarimu. Apibrėžti tokie kintamieji: „*ilgosios*“ ir „*nosinės balsės*“ (žymima atitinkamai *by* ir *bn*), „*lūpiniai balsiai*“ (*bl*), „*priešakinės eilės balsiai*“ (*bp*), „*duslieji priebalsiai*“ (*prd*) ir „*pusbalsiai*“ (*psb*) (plačiau apie garsų klasifikaciją žr. http://ualgiman.dtiltas.lt/fonetika.html). Taip pat buvo apskaičiuoti visų tipų raidžių kiekiai kiekviename žodyje.

Logtiesinių modelių tyrimui duomenys buvo imami iš 17 skirtingų grožinio ir mokslinio stilių tekstų, atsitiktinai iš kiekvieno atrenkant po 3000 žodžių. Naudojantis „SAS“ programa buvo tirtas raidinės ir garsinės struktūros sąryšis su moksliniu ir grožiniu stiliumi. Gauti rezultatai pateikti 2 lentelėje.

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

Tradiciškai veiksnys laikomas statistiškai reikšmingu, jeigu atitinkama p reikšmė yra mažesnė už reikšmingumo lygmenį $\alpha = 0.05$. Paskutinėje lentelės eilutėje pateikta tikėtinumo santykio kriterijaus p reikšmė, kuri šiuo atveju rodo, kad parinktas modelis pakankamai gerai atitinka turimus duomenis, todėl nulinę hipotezę apie parinkto modelio adekvatumą atmesti nėra pagrindo.

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

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

a) Lietuviškų tekstų invariantai yra nosinių ir ilgųjų balsių proporcijos, kurios, žinoma, priklauso nuo žodžio ilgio (sąveikos *i5*bn* ir *i5*by*). Statistiškai labai reikšminga sąveika *i5*bp*by* parodo, kad ilgųjų priešakinės eilės balsių, t.y. balsių *y*, kiekis priklauso tik nuo žodžio ilgio, bet nepriklauso nei nuo konkretaus teksto, nei nuo jo stiliaus. Šios sąveikos struktūrinių elementų analizė (čia nepateikiama) parodė, kad balsių *y* yra tuo mažiau, kuo trumpesni žodžiai.

b) Mokslinės ir grožinės literatūros tekstų skirtumus parodo skirtingos ilgųjų balsių proporcijos skirtingo ilgio žodžiuose (sąveika *st*i5*by*). Šios statistiškai labai reikšmingos sąveikos struktūrinių elementų reikšmingumo analizė parodė, kad statistiškai reikšmingas skirtumas yra tarp trumpiausių žodžių. Grožinėje literatūroje ilgųjų balsių labai trumpuose žodžiuose yra mažiau negu mokslinėje literatūroje, o ilgesniuose žodžiuose – priešingai. Galima spėti, kad šie mokslinio ir grožinio stiliaus skirtumai yra susiję su dažnesniu žodžio „yra“ vartojimu mokslinėje literatūroje.

Išvados

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

1. Nagrinėjant turimus duomenis buvo nustatyta, kad visiems nagrinėjamiems tekstams yra būdingos tam tikros ilgųjų ir priešakinės eilės balsių proporcijos, kurios priklauso ir nuo žodžio ilgio.
2. Skirtumus tarp mokslinio ir grožinio stilių parodo nevienodas ilgųjų balsių pasiskirstymas įvairaus ilgio žodžiuose, ypač trumpiausiuose žodžiuose. Galima spėti, kad šie stiliaus skirtumai yra susiję su dažnesniu žodžio „yra“ vartojimu mokslinėje literatūroje.
3. Kadangi šiame darbe analizuojami tekstai buvo paimti atsitiktinai, tiksliai neapibrėžus, kas yra lietuviškas tekstas ir kas yra tiriamoji populiacija, tai gautų rezultatų negalima apibendrintai taikyti visiems lietuviškiems tekstams.

Aliktas tyrimas iliustruoja logtiesinių modelių panaudojimo galimybes tekstų apdorojime bei supažindina su jų pagrindu gautų rezultatų interpretacija. Naudodamiesi logtiesiniais modeliais galime aprašyti ir nagrinėti sudėtingus, taip pat ir aukštesnės (negu 3-ios) eilės tiriamų požymių tarpusavio sąryšius. Kitų alternatyvų tokiems tyrimams praktiškai nėra.

Literatūra

[1] A. Agresti. *Categorical Data Analysis*. Wiley-Interscience, New York, 2002.

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

[3] G.K. Pullum and A. Kornai. *Mathematical Linguistics*. Available from Internet: http://www.metacarta.com/Collateral/Documents/English-US/Mathematical-linguistics-Kornai.pdf.

[4] E. Rimkutė, V. Daudaravičius. Morfologinis dabartinės lietuvių kalbos tekstyno anotavimas. *Kalbų studijos*, **11**:30–35, 2007.

[5] A. Utka. Labai dažnų lietuvių kalbos žodžių ir žodžių formų ypatybės. *Lituanistica*, **1**(61):48–55, 2005.

[6] A. Utka. *Dažninis rašytinės lietuvių kalbos žodynas: 1 milijono žodžių morfologiškai anotuoto tekstyno pagrindu*. 2009. Adresas internete: http://donelaitis.vdu.lt/publikacijos/Dazninis_zodynas.pdf.

[7] D. Šveikauskienė. *Lietuvių kalbos vientisinių sakinių automatinė sintaksinė analizė*. Adresas internete: http://www.mii.lt/files/mii_dis_2010_sveikauskiene.pdf.

SUMMARY

The comparison of Lithuanian texts' styles by using the statistical analysis of the universal quantitative characteristics

K. Piaseckienė, M. Radavičius, R. Stiklius

Lithuanian language is quite complex and flexible, and its significantly complicates the development of efficient algorithms for the automatic processing of Lithuanian texts. For studying text-styles features were selected the universal quantitative characteristics that are unrelated to the text content and can be calculated for any text. This article shows how mathematical Statistics can help to distinguish and interpret the Lithuanian language styles. Studies of the log-linear models show the connection between the letters and sounds structure and the scientific and fiction.

Keywords: universal quantitative characteristics, log-linear models, statistically significant connections.