Lietuvos matematikos rinkinys. LMD darbai
52 tomas, 2011, 220–224

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

Lietuvių kalbos vaizdingumo raiškos priemonių analizė

Karolina Piaseckienė[1,2], **Marijus Radavičius**[1,2]

[1] *Šiaulių universitetas, Matematikos ir informatikos fakultetas*
P. Višinskio 19, 77156 Šiauliai

[2] *Vilniaus universitetas, Matematikos ir informatikos institutas*
Akademijos 4, LT-08663 Vilnius

E. paštas: karol@delfi.lt; mrad@ktl.mii.lt

Santrauka. Lietuvių kalba pasižymi vaizdingumu, kuris gali būti perteikiamas įvairiomis raiškos priemonėmis. Viena iš jų – metakalbiniai komentarai. Šiame straipsnyje, remiantis VDU sudarytu Lietuvių kalbos tekstynu, statistiškai nagrinėjami vartojamo pasakymo (emocinį) toną apibūdinantys metakalbiniai komentarai rašytinėje lietuvių kalboje. Sudarytas logistinės regresijos modelis, kuriuo siekiama aprašyti, nuo ko priklauso metakalbinio komentaro proporcija tarp visų bazinio žodžio pavartojimų.

Raktiniai žodžiai: lietuvių kalba, metakalbiniai komentarai, koreliacija, tekstynas, logistinė regresija.

Įvadas

Matematinė lingvistika, anot Geoffrey K. Pullum ir András Kornai, yra matematinių struktūrų ir metodų, kurie yra svarbūs lingvistikai, analizė [4].

Šiuo metu, plėtojantis struktūrinei lingvistikai, tampa itin svarbūs kalbos modeliavimo klausimai. Lingvistikoje skiriamos dvi modelių rūšys: nestatistiniai (arba baziniai) ir statistiniai (arba stochastiniai). Tai susiję su dvipusiu kalbos traktavimu jos funkcionavimo metu. Pirma, kalbą galima tyrinėti jos žodžių junginių identifikavimo požiūriu. Antra, kalbą galima traktuoti kaip tikimybinį procesą, susijusį su kalbos elementų panaudojimo dažnumu kalbos aktuose. Sudarant šiuos modelius, taikomi įvairūs matematiniai metodai [6].

Lietuvių kalba pasižymi vaizdingumu, kuris gali būti perteikiamas įvairiomis raiškos priemonėmis. Viena iš jų – metakalbiniai komentarai.

Metakalbiniu komentaru (MK) vadinamas kalbos arba teksto kalbinės raiškos aptarimas: vertinimas, vartojimo motyvavimas ir kt. [7].

Metakalbinius komentarus, kaip kalbos konkurencijos priemonę, nemažai nagrinėjo K. Župerka ir kiti lituanistai, taip pat ir kitų šalių (latvių, čekų) kalbininkai. Dauguma jų MK nagrinėja kaip stilistikos objektą, tačiau nėra nagrinėjami kitokio pobūdžio klausimai, pavyzdžiui, kokia tikimybė, kad pavartotas žodis yra MK dalis. Šis darbas turėtų bent iš dalies užpildyti minėtą spragą.

Iš MK įvairovės galima išskirti apibūdinančius vartojamo pasakymo (emocinį) toną ir tuo pačiu išreiškiančius emocinę bei estetinę priešpriešą: *gražiai*, *švelniai*, *skambiai*,

aiškiai, *atvirai*, *vaizdingai*, *vaizdžiai* ir *negražiai*, *grubiai*, *šiurkščiai*, *vulgariai*, *griežtai*, *ironiškai*, *banaliai kalbant*, *sakant* ir pan., kurie ir bus nagrinėjami šiame straipsnyje.

Nors yra manoma, kad neigiamas vertinimas kalboje išreikštas dažniau negu teigiamas, mūsų išrinkti MK rodo, kad kalbos vartotojai nevengia pabrėžti ir teigiamo kalbos bei kalbėjimo (šnekos) vertinimo.

Kalbos analizėje taikant statistinius metodus labai svarbu tiksliai apibrėžti tiriamą populiaciją. Nuo to priklauso gauti rezultatai, jų interpretacija ir patikimumas. Tai iliustruoja ir šiame darbe gauti rezultatai.

Šiame darbe, remiantis VDU sudarytu lietuvių kalbos tekstynu, nagrinėjama rašytinė lietuvių kalba. Tokiu būdu tiriamoji populiacija yra nusakyta to tekstyno sudarymo taisyklėmis. Deja, konkreti metodika, pagal kurią yra sudarytas tekstynas, autoriams nėra žinoma. Neaišku, ar sudarant tekstyną buvo remtasi imčių teorija (žr. [3]). Vadinasi, šiuo atveju negalime tvirtinti, kad tiriamoji populiacija adekvačiai atspindi rašytinę lietuvių kalbą, juolab kad ir pats terminas „rašytinė lietuvių kalba" nėra tiksliai apibrėžtas.

MK nagrinėjimui taikoma koreliacija ir logistinė regresija, naudojama SPSS ir SAS programinė įranga.

Pirmame skyrelyje trumpai supažindinama su logistinės regresijos modeliu. Antrame skyrelyje pateikiama metakalbinių komentarų apžvalginė analizė. Trečiame skyrelyje pateikiami logistinės regresijos modelio tyrimo rezultatai.

1 Logistinės regresijos modelis

Binarinė logistinė regresija yra naudojama tyrinėti binarinio atsitiktinio dydžio Y statistiniams sąryšiams su bet kurios skalės aiškinančiaisiais kintamaisiais (kategoriniai kintamieji yra abipus vienareikšmiškai perkoduojami į atitinkamą binarinių kintamųjų rinkinį). Kadangi binarinis kintamasis paprastai reiškia įvykį, kuris gali įvykti arba neįvykti, naudojant binarinę logistinę regresiją yra modeliuojama šio įvykio tikimybė, sąlygojama aiškinančiųjų kintamųjų.

Tikimybė p_i, kad i-ame stebėjime tiriamas atsitiktinis dydis Y_i įgys reikšmę 1, kai yra žinomos aiškinančiųjų kintamųjų $X_1, \ldots, X_k$ reikšmės $x_{1i}, \ldots, x_{ki}$, yra aprašomos formule

$$p_i = \frac{\exp\{z(\mathbf{x}_i)\}}{1 + \exp\{z(\mathbf{x}_i)\}}, \quad z(\mathbf{x}_i) = a + b_1 x_{1i} + b_2 x_{2i} + \cdots + b_k x_{ki},$$

čia $\mathbf{x}_i = (x_{1i}, \ldots, x_{ki})$ [2].

Kadangi šis darbas yra taikomasis, detalesnio logistinės regresijos modelio aprašymo nepateikiame – jį galima rasti [1]. Tik pastebėsime, kad tikimybė p_i, augant reikšmei x_{ji}, didėja (mažėja), jeigu $b_j > 0$ ($b_j < 0$).

2 MK apžvalginė analizė

Tyrimui duomenys buvo išrinkti iš VDU kompiuterinės lingvistikos centro viešai prieinamo internete „Dabartinės lietuvių kalbos tekstyno" (žr. http://donelaitis.vdu.lt/), kurį sudaro daugiau kaip 100 mln. žodžių ir kuris yra viena iš didžiausių lietuvių kalbos kalbinių duomenų bazių, kuria naudojasi įvairių sričių specialistai, atlikdami

1 lentelė. Palyginimas.

Tekstyno senoji versija		MK	Tekstyno naujoji versija		MK
Respublikinė periodika	23%	23,83%			
Vietinė periodika	17%	23,10%			
Populiarioji periodika	18%	19,41%			
Specializuota periodika	8%	6,6%	Publicistika	63,8%	66,47%
Grožinė literatūra	7%	8,71%	Grožinė literatūra	11,6%	17,57%
Negrožinė literatūra	11%	7,72%	Negrožinė lit.	14,2%	10,84%
Seimo stenogramos	2%	5,94%	Administracinė lit.	10%	4,66%
LR valstybiniai dokumentai	8%	0%	Sakytinė kalba	0,3%	0,46%
Filosofinės lit. vertimai	3%	2,77%			
Memuarai	3%	1,92%			
≈102 mln. žodžių		1515 MK	≈141 mln. žodžių		2186 MK

tyrimus Lietuvoje ir užsienyje. Tekstynas gana reprezentatyviai atspindi dabartinę lietuvių kalbą, įvairius jos stilius [5].

Aukščiau nurodytame internetiniame puslapyje yra dvi tekstyno versijos – senoji ir naujoji. 1 lentelėje pateiktas abiejų tekstyno versijų sandaros bei jose rastų MK kiekio palyginimas.

Kadangi naujojo tekstyno apimtis 40 mln. žodžių didesnė, visiškai suprantama, kad jame ir MK rasta kur kas daugiau. Be to, galima būtų tikėtis, kad MK pasiskirstymas skirtingose tekstyno srityse yra tiesiogiai proporcingas atitinkamos srities dydžiui. Tačiau iš lentelės matome, kad, tarkim, naujajame tekstyne grožinės literatūros srityje rasta daugiau MK, negrožinėje mažiau nei tikėtasi, o administracinėje literatūroje MK rasta net dvigubai mažiau. Tai rodo, kad MK yra ne paprasti žodžiai, o vaizdingumo raiškos priemonė, nevienodai vartojama skirtingose srityse. Grožinėje literatūroje vaizdingumo reikia daugiau, mokslinėje ar administracinėje – mažiau, o valstybiniuose dokumentuose (ši sritis buvo išskirta senojoje tekstyno versijoje) vaizdingumas visai nepriimtinas, todėl nagrinėjamų MK šioje srityje nėra.

Apskaičiavus naujojo tekstyno srityse rastų MK koreliacijas paaiškėjo, kad stipriausiai koreliuoja sakytinės kalbos ir administracinės literatūros MK, nors tiek MK bazinių žodžių, tiek visų žodžių, esančių šaltiniuose, kuriuose buvo bent vienas nagrinėjamas MK, šios dvi sritys koreliuoja silpnai.

Silpniausiai koreliuoja grožinės ir negrožinės literatūros MK, nors jau minėtose kitose imtyse šių dviejų sričių koreliacija yra vidutiniška ar net stipri.

Tai dar kartą patvirtina, kad MK yra išskirtinė kalbos raiškos priemonė.

3 MK logistinės regresijos modelis

Šiame tyrime logistinės regresijos modeliu siekiama aprašyti, nuo ko priklauso MK proporcija tarp visų bazinio žodžio (baziniu žodžiu laikomas MK konstrukcijoje esantis prieveiksmis, t. y. *gražiai*, *švelniai* ir t. t.) pavartojimų, t. y. tikimybę, kad pavartotas bazinis žodis yra MK dalis. Kadangi kai kurių tiriamų MK dažniai tekstyne buvo labai maži, tai parenkant patikimą ir stabilų modelį tuos labai retus MK teko eliminuoti iš tolesnio tyrimo.

Šiam tyrimui buvo apibrėžti tokie kintamieji: *zb_lg* – bazinio žodžio dažnio dešimtainis logaritmas; *zb_salt_lg* – bazinio žodžio santykinio dažnio tarp visų jo šaltinių (t. y., nors kartą jį paminėjusių šaltinių) žodžių dešimtainis logaritmas; *zb_srit_lg* –

2 lentelė. Logistinės regresijos modelis.

Logistinės regresijos modelis				
Veiksnys	L.L.	Įvertis	Wald'o statistika	p-reikšmė
zb_lg	1	−1,202	12,4734	0,0004
zb_salt_lg	1	3,853	20,2096	<,0001
*tn_atspalv*tipas*	8		131,2744	<,0001
*tn_atspalv*laipsnis*	1	0,803	15,7000	<,0001
*tn_atspalv*tipas*sritis*	24		479,1275	<,0001

bazinio žodžio santykinio dažnio tarp visų srities žodžių dešimtainis logaritmas; *tipas* – MK apibūdinanti savybė (vaizdingumas, atvirumas, švelnumas ir t. t.); *laipsnis* – bazinio žodžio nelyginamasis (0) arba aukštesnysis (1) laipsnis; *sritis* – keturios naujojo tekstyno sritys (dėl jau minėtos problemos sakytinės kalbos sritis buvo nenaudojama), grožinę literatūrą laikant bazine kategorija; *tn_atspalv* – teigiamas/neigiamas bazinio žodžio (tuo pačiu ir MK) atspalvis (vaizdžiai, vaizdingai, švelniai, aiškiau, atvirai turi teigiamą atspalvį (1), o šiurkščiai, grubiai, griežtai, banaliai turi neigiamą atspalvį (-1)).

Remiantis informaciniais kriterijais ir eliminavus statistiškai nereikšmingus veiksnius buvo parinktas toks (žr. 2 lentelę) logistinės regresijos modelis.

Atitinkamų logistinės regresijos modelio narių statistinis reikšmingumas pateiktas paskutiniame lentelės stulpelyje, kuriame yra visų į modelį įtrauktų veiksnių bei jų sąveikų atitinkamos p reikšmės.

Tradiciškai veiksnys laikomas statistiškai reikšmingu, jeigu atitinkama p reikšmė yra mažesnė už reikšmingumo lygmenį $\alpha = 0,05$. Galima pastebėti, kad šį modelį sudarantys veiksniai ir jų sąveikos yra statistiškai reikšmingi.

Hosmerio ir Lemešou kriterijus, kurio statistikos p reikšmė $p = 0,9792 > 0,05 = \alpha$, taip pat rodo, kad parinktas modelis yra tinkamas, t. y., gana gerai suderintas su duomenimis.

Gautus rezultatus galima būtų interpretuoti taip.

Neigiamas parametro prie *zb_lg* įverčio ženklas rodo, kad kuo daugiau kartų tiriamos srities tekste yra pavartotas pats bazinis žodis, tuo mažesnė tikimybė, kad jis bus pavartotas ne kaip paprastas žodis, o kaip MK. Tačiau, kadangi bazinio žodžio dažnis priklauso nuo srities dydžio, kurį lemia tekstyno sudarymo taisyklės, šio rezultato dalykinė interpretacija yra neaiški. Informatyvesnis yra santykinis dydis – bazinio žodžio proporcija tekstuose, kur jis nors kartą pavartotas, – nes jis mažiau įtakojamas tekstyno sudarymo taisyklių. Kaip matyti iš parametro prie *zb_salt_lg* įverčio ženklo, ši proporcija yra teigiamai susijusi su tikimybe baziniam žodžiui būti MK dalimi. Vaizdžiai kalbant, kuo dažniau autorius, „linkęs vartoti" vieną ar kitą bazinį žodį, jį vartoja, tuo didesnė tikimybė, kad tas bazinis žodis bus MK dalimi.

MK proporcija tarp visų bazinio žodžio pavartojimų reikšmingai priklauso ir nuo teigiamo/neigiamo bazinio žodžio atspalvio, tačiau ši priklausomybė nėra tiesioginė, bet pasireiškia per sąveikas su MK apibūdinančia savybe (*tipas*), laipsniu ar savybe ir tekstyno sritimi drauge.

Pirmoji sąveika nėra informatyvi, nes kintamasis *tn_atspalv* yra sudarytas iš kintamojo *tipas*. Vis dėlto įdomu pastebėti, kad teigiamą atspalvį turinčių parametrų didžiausio tikėtinumo įverčių ženklas yra neigiamas, o neigiamą atspalvį turinčių – teigiamas (taupant vietą, konkrečios tų įverčių reikšmės čia nepateikiamos).

Sąveika *tn_atspalv*laipsnis* rodo, kad MK emocinis atspalvis priklauso nuo bazinio žodžio laipsnio, t. y., aukštesnysis žodžio laipsnis tarsi paryškina teigiamą arba neigiamą atspalvį ir tuo pačiu įtakoja bazinio žodžio buvimą MK dalimi (kaip matyti iš atitinkamo įverčio ženklo, teigiamas atspalvis didina buvimo MK dalimi tikimybę).

Iš trečiosios sąveikos *tn_atspalv*tipas*sritis* galima spręsti, kad teigiamą/neigiamą atspalvį turinčio metakomentaro bazinio žodžio pavartojimas priklauso nuo tekstyno srities (grožinė, negrožinė ir t. t.), nes skirtingose srityse santykinis teigiamą ir neigiamą atspalvį turinčių žodžių vartojimas yra skirtingas.

Išvados

1. Atliktas tyrimas rodo, kad ne visose tekstyno srityse vaizdingumo raiškos priemonių vartojimas yra vienodas.
2. Kuo didesnis tekstuose bazinio žodžio santykinis dažnis, tuo didesnė tikimybė, kad jis bus pavartotas kaip MK dalis.
3. Tikimybė pavartotam teigiamą/neigiamą atspalvį turinčiam baziniam žodžiui būti MK dalimi priklauso nuo tekstyno srities.

Bendros išvados apie lietuvių kalbos ypatybes ir statistinius dėsningumus labai priklauso nuo to, kaip apsibrėšime, kas yra rašytinė lietuvių kalba, t. y. kuri (tekstyno) sritis ją geriausiai atspindi. Jeigu konstruosime modelius, imdami skirtingas nagrinėtų keturių sričių proporcijas, gausime skirtingas išvadas.

Literatūra

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

[2] V. Čekanavičius ir G. Murauskas. *Statistika ir jos taikymai II*. TEV, Vilnius, 2008.

[3] D. Krapavickaitė ir A. Plikusas. *Imčių teorijos pagrindai*. Technika, Vilnius, 2005.

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

[5] J. Kovalevskaitė. Dabartinės lietuvių kalbos tekstynas – 10 metų kaupimo ir naudojimo patirtis. *Prace Baltystyczne*, (3):231–241, 2006.

[6] L. Mauzienė. Lingvistiniai ir psichologiniai lingvodidaktikos pagrindai (teorinė interpretacija). *Santalka. Filologija. Edukologija*, **17**(2):61–67, 2009.

[7] K. Župerka. *Kalbos priemonių konkurencija kaip lietuvių kalbos stilistikos objektas*. Šiauliai, 1995.

SUMMARY

The analysis of the imagery expression devices in the Lithuanian language

K. Piaseckienė, M. Radavičius

Lithuanian language is characterized by imagery, which can be conveyed in a variety of expression means. One of them is metalanguage comments. In this article, metalanguage comments that describe the (emotional) tone of saying are analyzed statistically using corpus of contemporary Lithuanian language. A fitted logistic regression model describes effects which influence the proportion of the metalanguage comments among all occurrences of their basic word in a text.

Keywords: Lithuanian language, metalanguage comment, correlation, corpus, logistic regression.