



Asta Kazlauskienė, Erika Rimkutė,
Andrius Utkā

Kiekybiniai tyrimai kalbotyroje (III)



Dažniniai žodžių sąrašai

Ar susimąstėte, kokie yra dažniausi lietuvių kalbos žodžiai? Gal manote, kad tai gražūs, prasmingi žodžiai, pvz., *Lietuva, meilė, geras, dirbti, gyventi*? Deja, tarp dešimties dažniausių žodžių formų vien tik nekaitomi žodžiai, įvardžių ir veiksmažodžio *būti* formos. Lietuvių kalbos dažniniuose žodžių sąrašuose neabejotinas lyderis yra jungtukas *ir*: paprastai jis sudaro apie 3 proc. visų teksto žodžių. Taigi tekste vidutiniškai vienas iš 34 žodžių yra *ir*, jis pavartojamas kas antroje eilutėje. Beje, kiekviena kalba turi savų lyderių, pavyzdžiui, anglų kalboje dažniausias yra žymimasis artikelis *the*, vokiečių kalboje – *der*, rusų kalboje – jungtukas *u* (ir). Dešimt pačių dažniausių lietuvių kalbos žodžių formų 1 mln. žodžių morfologiškai anotuotame tekстыne yra: *ir, j, kad, yra, iš, o, buvo, kaip, ar, savo*. Beje, tarp dažniausių daiktavardžio formų yra *darbo, Lietuvos*.



Nors esame linkę manyti, kad intuityviai jaučiame, kurie žodžiai dažnesni, vis dėlto neretai klystame. Žmonės atkreipia dėmesį į rūpimus žodžius arba į keistus ir neįprastus kalbos reiškinius, o kasdieniai,

dažnai kartojami kalbos elementai praslysta nepastebėti. Kompiuteriu sugeneruoti dažniniai žodžių sąrašai suteikia objektyvią informaciją, kuri leidžia daryti išvadas apie žodžių vartojimą tiriamame tekste.

Lietuvių kalbininkai jau nemažai rašė apie dažniausiai vartojamus žodžius, sudarė ne vieną dažninį žodyną: V. Žilinskienė *Lietuvių kalbos dažninis žodynas*, 1990; *Dabartinės lietuvių kalbos atgalinis žodynas* (sud. V. Žilinskienė), 1995; V. Žilinskienė, L. Grumadienė *Dažninis dabartinės rašomosios lietuvių kalbos žodynas: mažėjančio dažnio tvarka*, 1997; V. Žilinskienė, L. Grumadienė *Dažninis dabartinės rašomosios lietuvių kalbos žodynas: abėcėlės tvarka*, 1998; *Bendriniai XX a. spaudos žodžiai: elektroninis dažninis žodynas* (sud. ir red. V. Mauricaitė, M. Norkaitienė, A. Pakerys, R. Petrokienė), 2004; *Tikriniai XX a. spaudos žodžiai: elektroninis dažninis žodynas* (sud. ir red. V. Mauricaitė, M. Norkaitienė, A. Pakerys, D. Sviderskienė), 2005; A. Utkā *Dažninis rašytinės lietuvių kalbos žodynas*, 2009.

Reikia paminėti, kad žodžių dažnis gali būti nustatomas ne tik didelėse tekstų sancaupose (tekstynuose), bet ir atskiruose tekstuose ar tekstų rinkiniuose (pvz., kurio nors rašytojo romanuose). Dažniniams žodžių sąrašams generuoti naudoja-

mos specialios kompiuterinės programos – dažninių žodžių generatoriai. Savaimė suprantama, kad dažninio sąrašo turinys ir dažniai priklauso nuo analizuojamų tekstų ar tekstyno sudėties ir dydžio. Tam tikros srities tekstų rinkinyje arba tekstyne dažninio sąrašo viršuje rasite tai sričiai būdingus žodžius, o dideliame, įvairias sritis apimančiame tekstyne galėsite nustatyti bendras tiriamos kalbos tendencijas.

Didelio tekstyno dažninis sąrašas yra nepamainomas įrankis žodynininkams. Juk dėl riboto žodyno dydžio jų kūrėjai stengiasi pirmiausia užtikrinti, kad į žodynus patektų dažniausi žodžiai. Todėl dažninis žodžių sąrašas, surūšiuotas mažėjančio dažnio tvarka, yra labai pravartus. Šie sąrašai gali būti panaudoti ir mokant(is) kalbų, nes tiek dėstytojas, tiek studentai turėtų žinoti, kuriuos dažniausius žodžius ir žodžių formas reikėtų pirmiausia išmokti.

Galima sudaryti ir vieno autoriaus visos kūrybos ar pavienių kūrinių dažninius sąrašus. Tokie dažniniai sąrašai atskleidžia autoriaus leksikos įvairovę, vyraujančią tematiką ir leidžia autoriaus kalbos tyrėjams pažvelgti į kūrybą kitu aspektu ir taip galbūt atrasti įdomių dėsningumų.

Tekstynų lingvistikoje (tai lingvistikos sritis, kurios tyrimai remiasi elektroniniais tekstynais) dažniniai sąrašai neretai yra tam tikro tyrimo pradžia, nes juose galime pastebėti neįprastai dažnai arba retai vartojamų žodžių ar jų grupių, kurios vėliau gali būti nuodugniau analizuojamos.

Žodis ir žodžio forma

Skaičiuojant teksto žodžius, labai svarbu atskirti *žodžio* ir *žodžio formos* sąvokas. Juk tas pats žodis gali pasirodyti įvairiomis formomis, pavyzdžiui, žodis *mintis* gali turėti 12 formų: *mintis*, *minties*, *minčiai*, *mintį*, *mintimi*, *mintyje*, *mintie*, *mintys*, *minčių*, *mintims*, *mintimis*, *mintyse*. Atkreipkite dėmesį, kad sutampa vienaskaitos vardininko ir daugiskaitos galininko, daugiskaitos vardininko ir šauksmininko formos. Paprastame (morfologiškai neanotuotame) žodžių formų sąrašė rasite visų pavartotų žodžių formų dažnius, pavyzdžiui, bus pateikiamas minėto žodžio *mintis* 12 skirtingai grafiškai užrašomų formų pavartojimo kiekis (vienaskaitos vardininko ir daugiskaitos galininko, daugiskaitos vardininko ir šauksmininko duomenys bus susumuoti). Tokius dažninius sąrašus galima sugeneruoti palyginti lengvai, nes nereikia jokios papildomos analizės.

Morfologiškai anotuoti dažniniai sąrašai yra sudėtingesni nei paprasti dažniniai sąrašai, nes juose visos pavartotos konkretaus žodžio formos yra apibendrinamos antraštinėmis formomis (lemomis), pavyzdžiui, pateikiamas visų žodžio *mintis* formų pavartojimo kiekis ir gali būti pateikiamos gramatinės kategorijos. Šiems sąrašams sudaryti reikalingos sudėtingesnės kalbos analizės priemonės – lemuokliai ir morfologiniai analizatoriai. Šiuo atveju gaunami konkrečių formų pavartojimo duomenys, pavyzdžiui, jau žinotume, kiek kartų pavartotas žodžio *mintis* vienaskaitos vardininkas, o kiek – daugiskaitos galininkas.

1 lentelėje pateiktos dažniausios žodžių formos ir žodžiai iš 1 mln. žodžių morfologiškai anotuoto tekstyno. Čia matome, kad lemuotame sąrašė į sąrašo viršų šalia nekaitomų kalbos dalių pakyla kaitomi įvardžiai – *jis*, *tas*, *aš*, *kuris*, *šis*, *kitas*,

visas, kas, savo; veiksmažodžiai – *būti, galėti, turėti*; daiktavardžiai – *darbas, žmogus, Lietuva*. Tiesa, reiktų nepamiršti, kad dažniniuose sąrašuose, skirtingai nei žodynuose, nėra sprendžiamos homonimų (pvz., *klanas, kasa*), homoformų (pvz., *laužo* – sutampa daiktavardžio *laužas* ir veiksmažodžio *laužyti* formos) ir homografų (pvz., *šalis* ir *šālįs*) problemos. Darant išvadas apie tokių žodžių dažnius reikalinga detalesnė jų analizė.

1 lentelė. 25 dažniausios žodžių formos ir žodžiai 1 mln. žodžių lietuvių kalbos tekstyne

Žodžių formų dažninis sąrašas			Lemuotas dažninis sąrašas		
Žodžio forma	Dažnis		Žodis	Kalbos dalis	Dažnis
1. ir	32713		1. ir	jng	28174
2. į	8369		2. jis	įv	18117
3. kad	7487		3. būti	vksm	17128
4. yra	5621		4. tas	įv	8819
5. iš	5015		5. į	prl	8369
6. o	4816		6. kad	jng	7487
7. buvo	3952		7. aš	įv	6522
8. kaip	3868		8. kuris	įv	5831
9. ar	3148		9. su	prl	5533
10. savo	3127		10. šis	įv	5309
11. nuo	2878		11. galėti	vksm	5181
12. jis	2836		12. iš	prl	5015
13. bet	2810		13. o	jng	4816
14. ne	2791		14. ir	dll	4539
15. apie	2759		15. darbas	dkt	4243
16. jų	2742		16. kitas	įv	3923
17. jo	2549		17. kaip	jng	3868
18. jau	2500		18. turėti	vksm	3863
19. dar	2451		19. visas	įv	3797
20. darbo	2352		20. žmogus	dkt	3647
21. Lietuvos	2267		21. Lietuva	tikr. dkt	3394
22. gali	2238		22. kas	įv	3204
23. būti	2187		23. ar	jng	3148
24. aš	2130		24. savo	įv	3127
25. bei	2079		25. nuo	prl	2878

Dažniausi kalbos žodžiai

Gali atrodyti, kad dažniausi kalbos žodžiai yra dažni visuose tekstuose ir dėl to pagal jų vartojimo kiekį negalima pasakyti nieko svarbaus apie teksto pobūdį. Pastebėta, kad dažniausi kalbos žodžiai, be dažnio, turi ir kitų išskirtinių savybių, pavyzdžiui, paprastai jie yra labai trumpi ir užima neproporcingai didelę teksto dalį

(pvz., 20 dažniausių lietuvių kalbos žodžių užima 12,5 proc. visų teksto žodžių, anglų kalboje – 10 proc.). Be to, buvo nustatyta, kad jų pasiskirstymas atspindi svarbias teksto funkcines savybes. Išsiaiškinta, kad skirtingų žanrų tekstams būdingi ne tie patys dažniausi žodžiai, pavyzdžiui, grožinės literatūros tekstams būdingi trečiojo asmens įvardžiai ir erdvės santykius reiškiantys prielinksniai (*ant, į, prie*), o administracinio stiliaus tekstams būdinga modalumo raiška (*būti, gali, turi, galima, reikia*) ir alternatyvų bei sąlygų vardijimą rodantys nekaitomi žodžiai (*arba, ar, tam, jei, pagal, jeigu*) (plačiau apie tai žr. Utkė 2005).

Rūšiavimas

Paprastai dažniniai sąrašai yra rūšiuojami didėjančio ar mažėjančio dažnio tvarka arba pagal abėcėlę. Pagal dažnį surūšiuotuose sąrašuose išsryškėja dažniausi žodžiai ir jų formos.

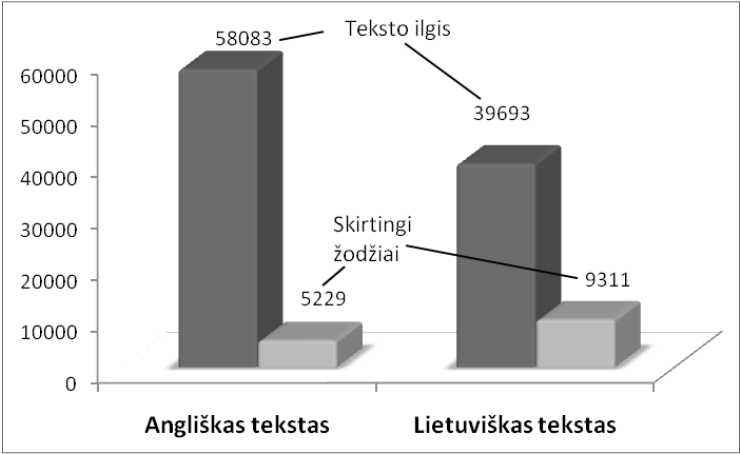
Pagal abėcėlę surūšiuotuose sąrašuose patogų grupuoti visus bendrašaknius ir giminiškus žodžius. Šiuose sąrašuose taip pat lengviau rasti konkretų žodį. 2 lentelėje matome žodžių *komendantūra, komentaras* ir *komentatorius* dažnių pasiskirstymą pagal linksnius abėcėlės tvarka išrikiuotame dažniniame sąrašė iš 1 mln. žodžių lietuvių kalbos tekstyno.

2 lentelė. Žodžiai *komendantūra, komentaras* ir *komentatoriai* pagal abėcėlę surūšiuotame dažniniame sąrašė

Žodžio forma	Dažnis
komendantūros	2
komendantūrų	1
komentariai	13
komentariais	2
komentaras	2
komentaro	2
komentaruose	2
komentarus	16
komentarą	12
komentarų	2
komentatoriai	4
komentatorius	2
komentatorių	2

Skirtingų kalbų dažniniai sąrašai

Sintetinių kalbų (pvz., lietuvių, rusų) žodžių formų dažniniai sąrašai yra gero- kai ilgesni nei analitinių kalbų (pvz., anglų, vokiečių, prancūzų). Taip yra dėl to, kad sintetinėse kalbose gramatiniai žodžių santykiai reiškiami žodžių formomis, tad šios kalbos pasižymi žodžių formų gausa. Palyginkime tos pačios knygos lietuviško vertimo ir angliško originalo žodžių formų dažninius sąrašus. 1 paveiksle



1 pav. Agatos Kristi „Paslaptingas atsitikimas Stailze“ (1992 m., leidykla UAB „Židinyš“; vertėjas Vladas Bubliauskas) originalo ir vertimo teksto ilgis ir skirtingų žodžių kiekis

matyti, kad nors lietuviškos knygos teksto ilgis matuojant žodžiais (39 693 žodžių) yra trumpesnis nei originalo (58 083), bet skirtingų žodžių sąrašas (arba dažninis sąrašas) yra 1,8 karto ilgesnis. Tai rodo daug didesnę lietuviško teksto žodžių formų įvairovę.

Žodingumas

Skirtingų ir visų teksto žodžių santykis leidžia nustatyti vadinamąjį *iteracijos indeksą*, arba teksto žodingumą. Iteracijos indeksas parodo, kokią dalį visų teksto žodžių sudaro skirtingi žodžiai, todėl žodingesnis tekstas turės didesnę iteracijos indeksą, o tekstas, kuriame yra daugiau pasikartojančių žodžių, turės mažesnę iteracijos indeksą. Dviejų vienodo ilgio sakinių iteracijos indekso skaičiavimo pavyzdys pateiktas 3 lentelėje. Matome, kad pirmame sakinyje skirtingi yra keturi žodžiai (67 proc.), o antrajame visi žodžiai skirtingi, todėl antrasis sakinys yra žodingesnis. Pavyzdžiui, suskaičiavus skirtingo stiliaus tekstų žodingumą, pastebėta, kad administracinio stiliaus tekstai yra ne tokie žodingi kaip grožinės literatūros tekstai.

3 lentelė. Iteracijos indekso skaičiavimo iliustracija

Sakinys	Skaičiavimas	Iteracijos indeksas
<i>Koksai gražus pasaulis, mylimoji, koksai gražus!</i>	Iš viso žodžių – 6, skirtingų žodžių – 4	$4 / 6 \times 100 \% = 67 \%$
<i>Štai Raseiniai, gražus miestas prie Dubysos.</i>	Iš viso žodžių – 6, skirtingų žodžių – 6	$6 / 6 \times 100 \% = 100 \%$

O gal patys norėtumėte suskaičiuoti, kokie yra dažniausi žodžiai jūsų tekste arba koks jūsų teksto iteracijos indeksas? Tam nereikia gaišti daug laiko – pasinaudokite VDU Kompiuterinės lingvistikos centre esančia teksto funkcijų nustatymo programa (žr. <http://tekstynas.vdu.lt>, programą rasite Įrankių skyrelyje). Nors ši programa pirmiausia skirta tam, kad nustatytumėte, kokio pobūdžio yra analizuojamas tekstas, ją galima naudoti ir kitiems tikslams. Jei pasirinksite minėtos programos funkciją „Pateikti dažnumų sąrašą“, sužinosite, kokie dažniausiai jūsų vartojami žodžiai. Naudojantis funkcija „Statistika“ galima gauti duomenis apie jūsų teksto žodžių ir sakinių skaičių, vidutinį sakinio, žodžio ilgį ir iteracijos indeksą.

Baigiamoji pastaba

Šiame straipsnių cikle apžvelgtos kiekybinių tyrimų kryptys yra tik maža dalelė lingvistikoje naudojamų statistinių metodų. Norime įspėti, kad, analizuojant kalbą skaitmeninėmis priemonėmis, jas reikia taikyti atsargiai, o interpretuoti kritiškai. Netinkamai naudojant statistinius metodus, galima prieiti prie nepagrįstų išvadų ir rezultatų. Be to, kartais statistika ir skaičiai gali tapti trukdžiu, neleidžiančiu išsiaiškinti prasminių kalbos ypatybių. Tačiau tie, kurie sugebės šiuos sunkumus nugalėti ir suprasti skaičių kalbą, išmoks naudoti svarbų tyrimo įrankį, kuris leis atrasti įdomių dėsningumų, pasiūlys naujų idėjų ir toliau plėtoti kalbos tyrimus.

Literatūra

1. *Utka A.*, 2000, Kalbinė įranga ir jos galimybės // *Darbai ir dienos*, t. 24, 275–285.
2. *Utka A.*, 2005, Labai dažnų lietuvių kalbos žodžių ir žodžių formų ypatybės // *Lituanistica*, t 1(61), 48–55.