RAIDŽIŲ DAŽNUMAS BENDRINĖS LIETUVIŲ KALBOS RIŠLIUOSE TEKSTUOSE

Petras SKIRMANTAS

Beveik visi, kam reikalinga informacija apie lietuviškų raidžių dažnumą, dabar paprastai remiasi V. Žilinskienės (1978) duomenimis, paskelbtais prieš 20 metų. Kadangi nežinoma, ar V. Žilinskienės rezultatai buvo tikrinami kitais analogiško išsamumo tyrimais, buvo sumanyta lietuvių k. raidžių dažnumą patyrinėti papildomai. Be kita ko, papildomai jį patyrinėti pasirodė racionalu ir dėl poros bendresnio pobūdžio aplinkybių. Viena, V. Žilinskienės pateiktieji raidžių dažnumai yra skaičiuoti ne iš „natūralių" lietuvių bendrinės kalbos tekstų, bet iš publicistikos dažnumų žodyno, – jam buvę imami įvairios tematikos laikraščių ir žurnalų straipsnių fragmentai po 1000 žodžių, iš kurių eliminuoti tikriniai daiktavardžiai (Žilinskienė, 1978, 83). Antra, nesunku pastebėti, kad V. Žilinskienės darbe kiek nutolta ir nuo tradicinio raidžių supratimo: lyginamosiose lentelėse „raide" taip pat laikomas ir žodžių tarpo simbolis – visiškai neaišku, kodėl jis ten šitaip „pagerbtas" ir išskirtas iš visų kitų rašmenų, neįeinančių į lietuvių kalbos raidyną (abėcėlę), bet neretai pasitaikančių tekstuose (tokie rašmenys būtų visi skaitmenys, skyrybos ženklai, kitų raidynų simboliai ir t. t.).

Duomenys apie raidžių dažnumus bei tarpusavio proporcijas eksperimentiniams kalbos tyrimams svarbūs ir reikalingi. Pavyzdžiui, jie tiesiog būtini, norint optimizuoti lietuviškus tekstus apdorojančias kompiuterių programas: jeigu jos, identifikuodamos rašmenis ar jų derinius, jų atitikmenų ieškos etaloninėse simbolių eilutėse, kuriose raidės išrikiuotos dažnumų mažėjimo tvarka, tai atpažinimui reikės žymiai mažiau laiko, negu ieškant jų, tarkim, abėcėlės tvarka išdėstytose eilutėse. Kuo tikslesni dažnumo duomenys, tuo greičiau tokios programos veiks. Taip pat nėra abejonės, kad į raidžių dažnumus atsižvelgti būtina kuriant įvairias tekstų šifravimo (įslaptinimo) bei šifruotų tekstų dešifravimo sistemas, mėginant automatizuoti įvairių specifinių teksto požymių – tokių kaip žanrinis pobūdis, stilistiniai ypatumai, tikėtiniausia autorystė ir pan. identifikaciją. Be šių duomenų taip pat sunkiai įsivaizduojami kalbos sintezės darbai, rašytinės ir sakytinės kalbos tarpusavio santykio tyrinėjimai, pagaliau – ergonominiu požiūriu optimalus simbolių išdėstymas kompiuterių ir lietuviškų rašomųjų mašinėlių klaviatūrose, racionalus jų kodavimas kompiuterinių rašmenų lentelėse ir kiti panašūs dalykai.

Šis tyrinėjimas nėra ir jokiu būdu negali būti galutinis. Juo tik siekta palyginti jau esamą, V. Žilinskienės darbuose pateiktą informaciją apie raidžių dažnumus su duomenimis, gautais iš kitų, naujų tekstų, o jei pasirodytų tikslinga, tai tą informaciją ir papildyti bei pakoreguoti.

Tyrimui buvo sukurta speciali kompiuterio programa, nustatanti lietuvių kalbos abėcėlės raidžių kiekius bei apskaičiuojanti jų santykinius dažnumus tekste, kuris tekstinio failo pavidalu laikomas išorinėje kompiuterio atmintyje – kuriame nors iš jo magnetinių diskų. Visų kitų rašmenų ir simbolių, t. y. tų, kurie lietuviškajai abėcėlei nepriklauso, – nežiūrima ir į jų kiekį neatsižvelgiama. Tad ir konkrečių raidžių santykiniai dažnumai nustatomi tiktai visų lietuviškosios abėcėlės raidžių atžvilgiu, o ne apskritai tekste pasitaikančių rašmenų (simbolių) atžvilgiu.

Duomenys apie lietuvių raidžių dažnumus rinkti iš palyginti stambių, didelės apimties rišlių bendrinės lietuvių kalbos tekstų: personaliniais kompiuteriais surinktų, jais redaguotų ir jau anksčiau spaudai parengtų knygų bei žurnalų – ačiū visiems, kurie leido naudotis savo kompiuteriniais failais. Itin stambūs tekstai su minėta programa buvo apdorojami dalimis, paskui duomenys susumuoti. Programa, suskaičiavusi lietuvių abėcėlės raides ir apskaičiavusi kiekvienos jų santykinius dažnumus nurodytame faile (t. y. tekste ar jo dalyje), visą šią informaciją užrašydavo į diską taip pat tekstiniu failu – specialia lentele. Susikaupė keliasdešimt tokių lentelių (kartu su tomis, kurios gautos išmėginant ir testuojant programą ar tiriant šiaip įvairius atsitiktinius tekstus), bet tolesniam apdorojimui imtos tik tos, kuriose atsispindi duomenys apie raidžių dažnumą tyrimui atrinktuose tekstuose.

Tekstai, iš kurių rinkti čia apibendrinami duomenys, yra šie:

1. Moksleivio fiziologija: Šviečiamojo pobūdžio knyga moksleiviams ir pedagogams.
2. *Gudaitis L.* Knyga apie Tadą Lomsargį: Biografijos ir kūrybos apybraiža.
3. *Stundžia B.* Lietuvių bendrinės kalbos kirčiavimo sistema, Vilnius, 1995.
4. *Malūkas E.* Juodieji želmenys: Romanas.
5. Gimtoji kalba: 1993 m. komplektas.
6. *Tomonis M.* Žinia: Esė, filosofiniai apmąstymai, poezija, dienoraščiai, laiškai.
7. *Almonaitis V.* Ką šniokščia Jūros rėvos, Vilnius–Kaunas, 1994.
8. Baltistica: T. 29, sąs.1 (lietuviškieji straipsniai).
9. *Strozzi E. di*. Šiaurės rožė: Romanas.

Toliau šie tekstai dažniausiai nurodomi čia pažymėtu eilės numeriu.

Tolesniam apdorojimui duomenys apie raidžių dažnumą išvardintuose tekstuose iš pirminių lentelių buvo perkelti į elektroninę skaičiuoklę *MS Excel for Windows* (v. 7.0) – jos priemonėmis ir atlikta didžiuma lyginamųjų skaičiavimų, nubraižyti grafikai. Tiktai dažnio priklausomybės nuo rango aproksimacijoms skaičiuoti panaudota labiau specializuota skaičiuoklė *MicroCal Origin* (v. 2.8), o klasterinei analizei – spacializuotas programų paketas *Statistica* (v. 4.3).

1 lentelė. **Raidžių dažnumai tekstuose**

Raidė:	Teksto numeris									IŠ VISO:
	1	2	3	4	5	6	7	8	9	
a	0,112943	0,114574	0,117313	0,125078	0,118989	0,11135	0,120151	0,110213	0,131942	0,117971
ą	0,006642	0,007097	0,005188	0,007998	0,006001	0,008055	0,006051	0,004338	0,00849	0,006916
b	0,010611	0,015392	0,015144	0,01361	0,017252	0,015491	0,012356	0,015495	0,016775	0,014622
c	0,005748	0,003631	0,006419	0,001147	0,003258	0,003544	0,00186	0,00515	0,001325	0,003419
č	0,003354	0,00465	0,006976	0,004148	0,004343	0,004606	0,008036	0,004407	0,004208	0,004694
d	0,025768	0,026533	0,030567	0,025452	0,02758	0,024331	0,024423	0,027308	0,026941	0,026276
e	0,055082	0,050655	0,051485	0,054763	0,053076	0,060064	0,054801	0,057547	0,052601	0,054915
ę	0,001699	0,001993	0,001866	0,003482	0,001611	0,002584	0,003447	0,0019	0,002859	0,002408
ė	0,013307	0,014813	0,017253	0,023319	0,014826	0,016591	0,01854	0,013807	0,021242	0,017331
f	0,004044	0,002738	0,005302	0,000868	0,002105	0,002241	0,000633	0,004219	0,001943	0,002469
g	0,024282	0,020235	0,019878	0,021221	0,016998	0,020958	0,017524	0,01797	0,020789	0,020267
h	0,00182	0,002177	0,001195	0,000544	0,000916	0,001179	0,000695	0,002869	0,001554	0,001248
i	0,150658	0,130824	0,152788	0,127812	0,139844	0,143809	0,136917	0,140796	0,129609	0,139422
į	0,004652	0,006679	0,004454	0,006	0,00478	0,006499	0,006329	0,004282	0,006064	0,005559
y	0,014717	0,016303	0,012751	0,01184	0,014914	0,014749	0,014647	0,016132	0,010175	0,013937
j	0,017835	0,018897	0,014506	0,018067	0,018563	0,021616	0,020971	0,016732	0,024052	0,019014
k	0,043206	0,0471	0,044726	0,052053	0,049919	0,04173	0,046545	0,047078	0,046531	0,04675
l	0,034159	0,031174	0,027367	0,030494	0,034126	0,030497	0,035388	0,033583	0,035138	0,032149
m	0,041574	0,037225	0,033709	0,02976	0,032972	0,037157	0,030627	0,031933	0,028558	0,034148
n	0,052966	0,04815	0,05338	0,052244	0,050239	0,048049	0,047365	0,052691	0,050438	0,050659
o	0,05671	0,062196	0,049863	0,053587	0,056093	0,055356	0,054614	0,051609	0,054831	0,055075
p	0,025759	0,030743	0,030792	0,031809	0,02572	0,026081	0,032612	0,028471	0,029026	0,028493
r	0,054135	0,057101	0,055073	0,046078	0,053465	0,048488	0,048602	0,057279	0,05336	0,05153
s	0,082434	0,0773	0,075946	0,080829	0,076816	0,086204	0,074464	0,075667	0,072902	0,07951
š	0,009109	0,018592	0,011578	0,017523	0,013778	0,014034	0,01703	0,015563	0,014014	0,014358
t	0,047805	0,055229	0,056946	0,057618	0,061101	0,055325	0,053406	0,065104	0,057627	0,056353
u	0,045235	0,049861	0,036452	0,052916	0,041556	0,043009	0,050419	0,039915	0,052518	0,045774
ų	0,015148	0,011312	0,01313	0,006422	0,014575	0,010735	0,012101	0,014401	0,007034	0,011442
ū	0,00473	0,004112	0,0058	0,005372	0,004668	0,006594	0,005322	0,004338	0,00491	0,005251
v	0,021733	0,022291	0,024151	0,025048	0,026399	0,026252	0,029501	0,023489	0,022109	0,024804
z	0,004148	0,002101	0,00351	0,001009	0,002564	0,002885	0,001673	0,004507	0,001179	0,002541
ž	0,007988	0,008322	0,014491	0,011886	0,010951	0,00994	0,012945	0,011207	0,009255	0,010696
Raidžių:	436166	222781	271205	608509	546881	579302	208571	159990	205302	3238707

Tyrimu nustatyti raidžių santykiniai dažnumai bei bendras raidžių skaičius kiekviename iš minėtųjų tekstų, taip pat jų dažnumas visame ištirtų 9 tekstų masyve pateikta 1 lentelėje.

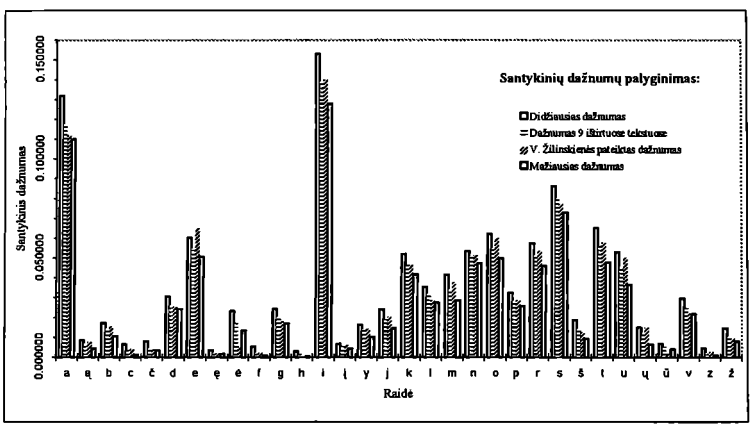
2 lentelė. **Raidžių santykinių dažnumų palyginimas**

Raidė	Santykinis dažnumas			
	Bendras 9 tekstuose	Didžiausias	Mažiausias	V.Žilinskienės pateiktas
a	0,117971	0,131942	0,110213	0,112
ą	0,006916	0,00849	0,004338	0,0076
b	0,014622	0,017252	0,010611	0,0155
c	0,003419	0,006419	0,001147	0,0041
č	0,004694	0,008036	0,003354	0,004
d	0,026276	0,030567	0,024331	0,0255
e	0,054915	0,060064	0,050655	0,0651
ę	0,002408	0,003482	0,001611	0,0021
ė	0,017331	0,023319	0,013307	0,0057
f	0,002469	0,005302	0,000633	0,0021
g	0,020267	0,024282	0,016998	0,0182
h	0,001248	0,002869	0,000544	0,0001
i	0,139422	0,152788	0,127812	0,1401
į	0,005559	0,006679	0,004282	0,0062
y	0,013937	0,016303	0,010175	0,0143
j	0,019014	0,024052	0,014506	0,0202
k	0,04675	0,052053	0,04173	0,0466
l	0,032149	0,035388	0,027367	0,0288
m	0,034148	0,041574	0,028558	0,0378
n	0,050659	0,05338	0,047365	0,0515
o	0,055075	0,062196	0,049863	0,0599
p	0,028493	0,032612	0,02572	0,0285
r	0,05153	0,057279	0,046078	0,0537
s	0,07951	0,086204	0,072902	0,0772
š	0,014358	0,018592	0,009109	0,0127
t	0,056353	0,065104	0,047805	0,0578
u	0,045774	0,052916	0,036452	0,0504
ų	0,011442	0,015148	0,006422	0,0151
ū	0,005251	0,006594	0,004112	0,0018
v	0,024804	0,029501	0,021733	0,0224
z	0,002541	0,004507	0,001009	0,0028
ž	0,010696	0,014491	0,007988	0,0094

Iš lentelės nesunku matyti, kad nors bendras lietuviškų raidžių kiekis konkrečiuose tekstuose (t. y. imčių tūris) palyginti didelis (šimtai tūkstančių), vis dėlto įvairių tekstų raidžių santykiniai dažnumai gerokai skiriasi – pastebimos gana žymios santykinių dažnumų variacijos. Tos variacijos kone tiesiogiai proporcingos dažnumui: kuo didesnis raidės santykinis dažnumas, tuo dažniausiai didesnė ir jo variacija. Pavyzdžiui, pačios dažniausios lietuvių tekstų raidės *i* santykinio dažnumo varijavimo mastas (specifiniu statistikos terminu sakant – variacijos žingsnis) didesnis kaip 3,7% (mažiausias jis 4, o didžiausias – 3 tekste – atitinkamai 12,8% ir 16,5%), o palyginti nedažnų raidžių santykinis dažnumas svyruoja daug mažiau.

Be abejo, įdomu palyginti šiuos raidžių dažnumus su tais, kuriuos minėtame straipsnyje yra pateikusi V. Žilinskienė. Jos nustatytus dažnumus (imami iš viso žodyno apskaičiuoti raidžių dažnumai, pateikti min. str. 1 lentelėje) logiška lyginti su raidžių dažnumais visų 9 dabar ištirtų tekstų masyve bei su didžiausiu ir mažiausiu jų dažnumais, rastais šiuose tekstuose. Tikslūs („skaitmeniniai“) lyginimo duomenys pateikti 2 lentelėje, o bendras grafinis jų vaizdas – 1 pav.

Nors šiuo tyrimu nustatyti ir V. Žilinskienės pateikti raidžių dažnumai šiek tiek skiriasi, tie skirtumai nėra dideli ir greičiausiai bus atsiradę dėl „natūralaus" santykinių dažnumų varijavimo. Ypač pabrėžtina, kad V. Žilinskienės apskaičiuoti didžiumos raidžių dažnumai yra didesni už mažiausią, bet mažesni už didžiausią jų dažnumą, rastą dabar ištyrus 9 tekstus, kitaip tariant, jie patenka į tą patį intervalą, kuriame svyruoja ir dabartinių tekstų raidžių dažnumai. Tiktai 4 raidžių – *e*, *ė*, *h* ir *ū* – santykiniai dažnumai šį intervalą pražengia, bet jau pačios V. Žilinskienės buvo paaiškintos (žr. min. str. 84 p.) priežastys, dėl kurių kaip tik tų raidžių dažnumas jos skaičiavimuose galįs būti kiek iškreiptas.



1 pav. **Raidžių santykinių dažnumų diagramos**

Aiškiai pastebima raidžių santykinių dažnumų variacija iš teksto į tekstą, kai imtys yra po kelis šimtus tūkstančių, leistų daryti prielaidą, kad raidžių santykiniams dažnumams ryškus statistinis stabilumas (plg. Kubilius, 1980, 14–15; Tutubalin, 1972, 5–7, 144–145) nebūdingas – kad tos ar kitos raidės proporcija tekste, jos santykinis kiekis greičiausiai yra susijęs su kokiais nors specifiniais teksto ypatumais ar požymiais ir galbūt jų sąlygojamas. Apskritai imant, tai gana natūralu, nes, pirma, dažnio statistinis stabilumas paprastai siejamas su atsitiktinių dydžių nepriklausomumu, t. y. su tokiais atsitiktiniais rezultatais, kurie vienas kito tiesiogiai neveikia. O rišlaus teksto raidžių jokiu būdu neišeitų laikyti tiesiogiai nesusijusiomis, nes beveik kiekvienõs raidės buvimo ar nebuvimo konkrečioje teksto pozicijoje galimybė labai smarkiai priklauso nuo ankstesnių, prieš ją einančių raidžių. Liktų tik pridurti, kad patys tarpusavy susijusių dydžių tikimybiniai modeliai – vadinamosios Markovo grandinės – buvo

grindžiami remiantis ir raidžių derinių A. Puškino „Eugenijaus Onegino" tekste analize (plg. Fišas, 1968, 28t.; Markov, 1913). Iš čia – praktinė pastaba: tiriant raidžių, taip pat ir visų kitų tarpusavy susijusių teksto segmentų pasiskirstymus, negalima aklai remtis tais tikimybiniais modeliais, kurie taikytini šiaipjau nepriklausomų dydžių pasiskirstymams. O antra, gal verta būtų taip pat paabejoti, ar rišlūs, „natūralūs" tekstai, ištisos knygos ar žurnalai, laikytini korektiškomis raidžių imtimis: tą sąvoką matematinė statistika supranta kiek kitaip ir paprastai labiausiai pabrėžia vadinamąjį homogeniškumą (plg. Kruopis, 1993, 270tt. Ajvazjan ir kt., 1983, 30) – apriorinį reikalavimą, kad „tikroji" tiriamojo įvykio tikimybė per visą imtį išliktų pastovi. Šiuo gi atveju homogeniškumo klausimas išvirstų į klausimą, ar faktinės, nors tiksliai ir nežinomos, raidžių tikimybės per visą tekstą bei iš teksto į tekstą išlieka pastovios, ar atvirkščiai, jos kinta dar ir priklausomai nuo kokių nors kitų, papildomų sąlygų. Deja, į jį kaip reikiant neatsakyta. Iš 1 ir 2 lentelių matoma palyginti nemenka kai kurių raidžių santykinių dažnumų variacija leistų spėti, kad homogeniškumo požiūriu santykiai tarp įvairių tekstų gali būti gana nevienodi: vienur labiau tikėtina vienokia, o kitur – kitokia, priešinga išvada. Tuo tarpu ką nors pasakyti apie homogeniškumą atskirų tekstų ribose iš šio tyrimo būtų dar kebliau, nes tiesioginių duomenų tam jis neduoda.

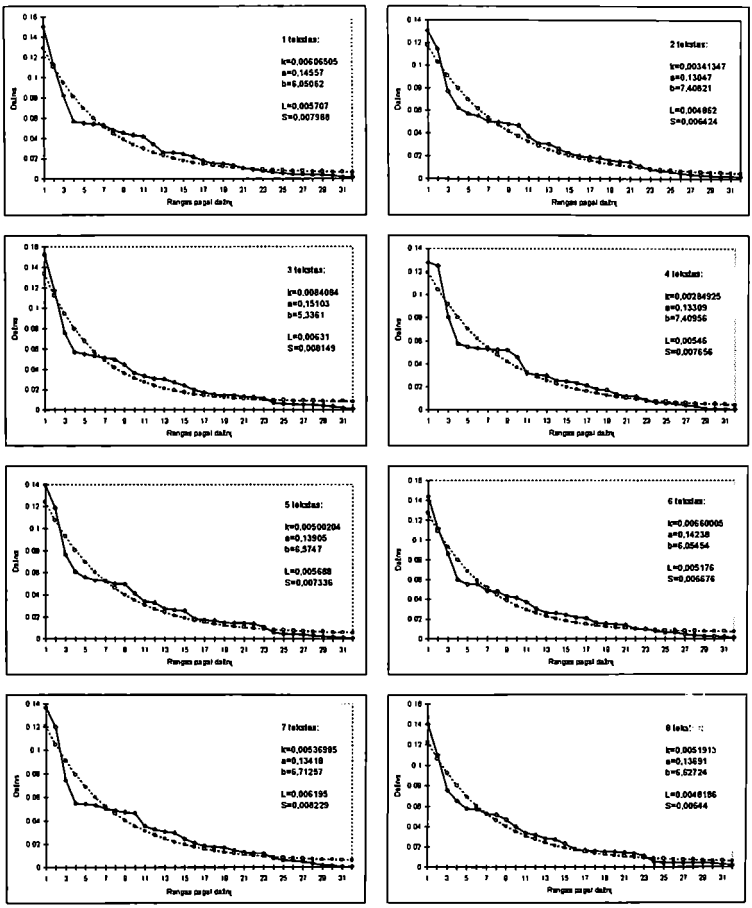
Buvo mėginta pasižiūrėti, kaip raidžių dažnumų dydis (reikšmė) priklauso nuo jų rango. Raidžių dažnumai, tiek atskirų tekstų, tiek ir sumariniai, buvo rikiuojami mažėjimo tvarka ir numeruojami nuo 1 iki 32; šie numeriai ir yra dažnumų rangai: pirmojo rango dažnumas būna pats didžiausias, antrojo – antras pagal dydį, trečiojo – trečias ir t. t., o paskutiniojo, trisdešimt antrojo, – pats mažiausias. Šitaip sutvarkytos dažnumų sekos, ypač – pavaizduotos grafikais, būna gana informatyvios. Iš jų nesunku yra susekti bendrą dažnumų mažėjimo pobūdį: nustatyti, kur jie mažėja sparčiau ir kur lėčiau, pažiūrėti, ar jų mažėjimo sparta kinta visur tolygiai, ar ne, pagaliau – sugretinti tokią mažėjančių dažnumų seką su kokiais nors žinomais teoriniais mažėjančių dydžių modeliais.

Kita gi vertus, mėginimai analitiškai pavaizduoti dažnumų priklausomybę nuo jų rangų paprastai vienaip ar kitaip siejami su garsiuoju Zipfo dėsniu, tiksliau pasakius, yra ne kas kita, kaip įvairios to dėsnio modifikacijos, taikomos dažniausiai žodžių dažnumams (plg. Jaglom, 1973, 265t.; Piotrovskij ir kt., 1977, 19), bet iš principo taikytinos ir smulkesnių kalbos elementų – skiemenų, fonemų bei raidžių – dažnumams (plg. Piotrovskij, 1975, 98tt., ypač – 109). Šį kartą irgi mėginta rasti analitinę raidžių dažnių priklausomybės nuo jų rangų išraišką, tačiau nueita kiek kitu keliu: nebesistengta įrodyti, jog Zipfo dėsnį galima modifikuoti taip, kad jis tiktų ir raidžių dažnumams, o tiesiog pagal bendrą dažnių mažėjimo pobūdį tai priklausomybei aproksimuoti parinkta mažėjanti laipsninė funkcija $p_r = k + ae^{-(rb)}$; čia r – rangas (t. y. bet kuris sveikasis skaičius nuo 1 iki 32 imtinai), p_r – atitinkamo rango teorinė tikimybė (kitaip – prognozuojamas santykinis dažnumas), o k, a ir b – koeficientai. Pasitelkus kiek specializuotą elektroninę skaičiuoklę *MicroCal Origin* nėra sunku tuos koefi-

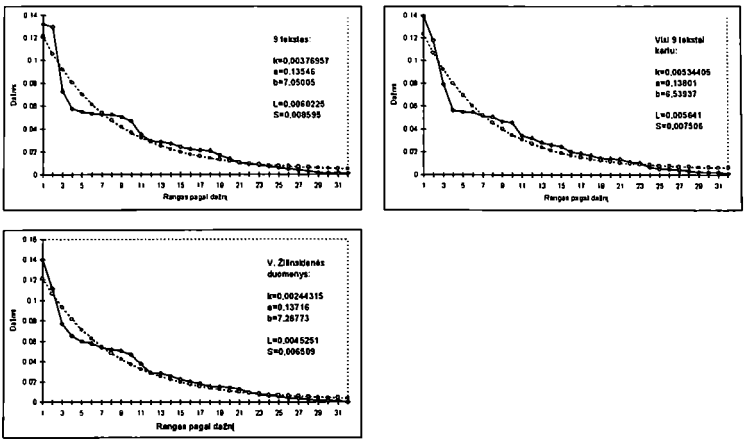
cientus parinkti taip, kad visų 32 p_r suma būtų lygi 1, o skirtumų tarp realių ir atitinkamų prognozuojamų santykinių dažnumų visuma taptų minimali. Kiekvienam tekstui, taip pat bendram visų 9 tekstų masyvui ir V. Žilinskienės pateiktiems duomenims tie koeficientai apskaičiuoti atskirai: mat rūpėjo ne rasti kokias nors „universalias", visiems tekstams neblogai tinkančias tos lygties koeficientų reikšmes, bet pažiūrėti, kiek ir kaip įvairiuose tekstuose svyruoja minėtieji minimalūs skirtumai, susidarą tarp faktinių, empiriškai nustatytų ir „optimalių" teoriškai apskaičiuotų raidžių dažnumų. Svyravimo matai imti patys paprasčiausi: vidutinis linijinis (L) ir vidutinis kvadratinis (S) skirtumai, abu reiškiantys iš esmės tą patį – vienai raidei tenkantį vidutinį neatitikimą tarp empiriškai nustatytos ir teoriškai apskaičiuotos dažnumo reikšmės (prisimintina, kad ir patys dažnumai, tiek empirinis, tiek ir teorinis, ir jų skirtumas, ir, tuo pačiu, dydžiai L bei S yra išreikšti vieneto dalimis). Vidutinis linijinis skirtumas L gaunamas sudėjus visų 32 raidžių empirinių ir teorinių dažnumų skirtumų modulius ir gautąją sumą padalinus iš 32, o vidutinis kvadratinis skirtumas S – analogiškai sudėjus tų pačių skirtumų kvadratus, padalinus sumą iš 32 ir iš dalmens ištraukus kvadratinę šaknį (užrašant „formaliai", $L=(\text{abs}(d_1-p_1)+\text{abs}(d_2-p_2)+...+\text{abs}(d_{32}-p_{32}))/32$; $S=\text{sqrt}(((d_1-p_1)^2+(d_2-p_2)^2+...+(d_{32}-p_{32})^2)/32)$; čia d_r – empirinis, o p_r – teorinis r rango raidės dažnumas). Tad vidutinis kvadratinis skirtumas S didesnius, žymesnius faktiškojo ir prognozuojamo dažnumų neatitikimus išryškina (bei mažesnius – neutralizuoja) labiau, negu vidutinis linijinis skirtumas L.

Raidžių dažnumų empirinių pasiskirstymų ir jų teorinių modelių grafikai pavaizduoti 2 pav. brėžiniuose; kartu nurodomos koeficientų k, a ir b bei vidutinio linijinio skirtumo L ir vidutinio kvadratinio skirtumo S reikšmės.

Iš 2 pav. grafikų gerai matyti, kad bendrasis dažnio priklausomybės nuo rango pobūdis visur išlieka maždaug panašus: pats didžiausias, pirmojo rango faktiškasis dažnumas visuomet viršija teoriškai apskaičiuotą reikšmę, bet rangų skalės pradžioje, iki ketvirtojo rango imtinai, jis mažta labai sparčiai, žymiai labiau, negu prognozuojamas tų pačių rangų dažnumas (empirinių dažnumų kreivė šioje atkarpoje žemyn smunka daug staigiau už teorinę kreivę, ir 4 rango faktiškasis dažnumas niekur nebesiekia nė pusės didžiausio, t. y. pirmojo rango dažnumo). Toliau iki pat paskutiniojo rango dažnumai mažėja jau daug lėčiau, – vienuose tekstuose (pvz., 2, 6 ir 8) palyginti visai tolygiai, kituose kiek labiau paįvairuodami. Kartu derėtų pabrėžti, kad dažnumų empirinio ir teorinio pasiskirstymų „sutapties matai" L ir S kokios nors didelės naudos, matyt, neduoda: iš teksto į tekstą jie kinta nedaug, ir nei to kitimo, nei jų absoliučių dydžių nėra kaip sieti nei su tekstų pobūdžiu ar turiniu, nei su kokiais kitais veiksniais. O galimà padaryti išvada, kad L ir S reikšmės mažesnės tuose tekstuose, kur dažnumai, didėjant rangų numeriams, mažėja tolygiau, būtų perdaug banali: taip ir turi būti, tokia yra šių matų „prigimtis". Vadinasi, tiek grafinis dažnumų kreivių vaizdas, regimasis jų santykis, tiek ir L bei S duoda iš esmės tą pačią informaciją, tik vienu atveju ji yra analoginio pobūdžio, apibendrinta ir apytikslė, o antru – skaitmeninė,



2 pav. **Raidžių dažnumų priklausomybė nuo rango**: empiriniai pasiskirstymai (ištisinė linija) ir jų teoriniai modeliai (punktyrinė linija; aproksimacijos lygtis $p_r = k + ae^{-(r/b)}$).



2 pav. **Raidžių dažnumų priklausomybė nuo rango**: empiriniai pasiskirstymai (ištisinė linija) ir jų teoriniai modeliai (punktyrinė linija; aproksimacijos lygtis $p_r = k + ae^{-(r/b)}$).

diskretizuota pasirinktu tikslumu. O tų kreivių forma? Įdomesnė, be abejo, yra empirinių dažnumų kreivių forma (teorinių kreivių forma visiškai priklauso nuo aproksimacijos lygties koeficientų), nes ji labai įvairuoja, kiekviename tekste yra savita ir tam tikru mastu unikali.

Nesileidžiant į apibendrinimus šįkart tik norėtųsi atkreipti dėmesį į dviejų tekstų – ketvirtojo ir devintojo – empirinių dažnumų kreivių vizualinį panašumą. Jų abiejų pradžia kiek primena kėdę ar terasą: 1 ir 2 rango dažnumų skirtumas palyginti mažas ir kreivė pačioje pradžioje gana gulsčia, nuo 2 iki 4 rango ji labai staigiai, kone statmenai smunka žemyn, o nuo 4 iki 9 – tampa beveik horizontali, bet tuoj pat, nuo 9 iki 11 rango vėl gana staigiai pasuka žemyn, išryškindama „laiptelį“ ir tik po to jau visiškai tolygiai žemėja iki pat galo, iki paskutiniojo rango. Tiesa, panašaus „laiptelio“ žymių esama ir kitų tekstų (septintojo, iš dalies – antrojo bei penktojo) kreivėse, tačiau 4 ir 9 tekstuose jis yra visų ryškiausias bei „figūriškiausias“ Čia pat reiktų pastebėti, kad abu šie tekstai yra stambiosios grožinės prozos kūriniai, romanai, tad natūraliai kyla klausimas, ar akivaizdus raidžių rangų dažnumų kreivių formos panašumas yra tik gryna atsitiktinybė, ar ne. Gal jis gali būti kaip nors susijęs su tų tekstų vidinės struktūros ypatumais, sąlygotais žanro ar kitų literatūrinių ar šiaip aukštesnio lygmens veiksnių? Taip pakreiptą klausimą kol kas tegalima tik iškelti, nes atsakymui į :štirtų gerokai atsitiktinių tekstų nė

iš tolo neužtenka: tejmanoma tik labai nedrąsi prielaida apie tokios sąsajos galimybę, kuri dar bus prisiminta vėliau.

3 l e n t e l ė. **Raidžių išsidėstymas pagal dažnumo rangus**

	Tekstas									Būna
Rangas	1	2	3	4	5	6	7	8	9	raidės
1	i	i	i	i	i	i	i	i	a	i,a
2	a	a	a	a	a	a	a	a	i	a, i
3	s	s	s	s	s	s	s	s	s	s
4	o	o	t	t	t	e	e	t	t	t,e,o
5	e	r	r	e	o	o	o	e	o	o,e,r
6	r	t	n	o	r	t	t	r	r	r,t,n,o
7	n	e	e	u	e	r	u	n	e	e,u,n,r
8	t	u	o	n	n	n	r	o	u	n,o,u,t,r
9	u	n	k	k	k	u	n	k	n	k,n,u,
10	k	k	u	r	u	k	k	u	k	k,u,r,
11	m	m	m	p	l	m	l	l	l	m,l,p
12	l	l	p	l	m	l	p	m	p	l,p,m
13	d	p	d	m	d	v	m	p	m	d,m,p,v
14	p	d	l	d	v	p	v	d	d	d,p,v,l
15	g	v	v	v	p	d	d	v	j	v,d,g,p,j
16	v	g	g	ė	j	j	j	g	v	g,j,v,ė
17	j	j	ė	g	b	g	ė	j	ė	j,ė,g,b
18	ų	š	b	j	g	ė	g	y	g	g,ų,š,b,j,y,ė
19	y	y	j	š	y	b	š	š	b	y,š,b,j
20	ė	b	ž	b	ė	y	y	b	š	b,ė,y,š,ž
21	b	ė	ų	ž	ų	š	ž	ų	y	ų,ž,b,ė,š,y
22	š	ų	y	y	š	ų	b	ė	ž	ų,y,š,b,ė,ž
23	ž	ž	š	ą	ž	ž	ų	ž	ą	ž,ą,š,ų
24	ą	ą	č	ų	ą	ą	č	c	ų	ą,č,ų,c
25	c	j	c	į	į	ū	į	z	į	į,c,ū,z
26	ū	č	ū	ū	ū	į	ą	č	ū	ū,č,į,a
27	į	ū	f	č	č	č	ū	ą	č	č,ū,ą,f,į
28	z	c	ą	ę	c	c	ę	ū	ę	c,ę,ą,ū,z
29	f	f	į	c	z	z	c	į	f	f,c,z,į
30	č	h	z	z	f	ę	z	f	h	z,f,h,č,ę
31	h	z	ę	f	ę	f	h	h	c	h,ę,f,c,z
32	ę	ę	h	h	h	h	f	ę	z	h,ę,f,z

Įdomu taip pat, kokia raidė kurį rangą kuriame tekste užima, kitaip tariant, – kokios gaunamos raidžių sekos, kai raidės išdėstomos jų dažnumų mažėjimo tvarka. Jos pateikiamos 3 lentelėje.

Akivaizdu, kad įvairiuose tekstuose raidžių dažnumo rangai gerokai įvairuoja: sakysim, *n* raidės rangas svyruoja nuo 6 iki 9, *v* – nuo 13 iki 16, *b* – nuo 17 iki 22, *š* – nuo 18 iki 23 ir t. t. Pamėginta suskaičiuoti, kelių raidžių rangai sutampa ir kelių skiriasi, įvairus tekstus lyginant vienus su kitais. Buvo sudarytos visos įmanomos skirtingų tekstų poros, kurių iš viso susidaro 36 (bendras informatyvių porų skaičius yra ((9×9)–9)/ 2=36, nes tekstai negali sudaryti porų patys su savimi, o be to, nesvarbu yra, kurią vietą poroje, pirmąją ar antrąją, kuris tekstas užima). Skaičiavimo rezultatai pateikti 4 lentelėje.

Išvada ir čia kiek netikėta: išdėsčius raides pagal jų dažnumo rangus, jų eilės skirtinguose tekstuose kur kas labiau skiriasi, negu sutampa (palyginus visus tekstus vienus su kitais, nustatyta, jog skiriasi vidutiniškai 23, o sutampa tik 9 raidžių rangai; tikslesni vidurkiai yra atitinkamai 23,22 ir 8,78). Labiausiai panašūs pasirodė beesą 1 ir 2 tekstų raidžių rangai – net 13 sutapimų. Po 12 kartų sutampa 2 ir 8, 3 ir 5 bei 5 ir 8 tekstų raidžių rangai. O labiausiai skirtingi yra 6 ir 8

(vos 4 sutapimai), taip pat 1 ir 7, 3 ir 6, 6 ir 9 bei 8 ir 9 (rasta po 5 sutapimus) tekstų raidžių rangai. Pastebėtina, kad kai kurie panašumai (pvz., 3 ir 5 bei 5 ir 8 tekstų) ir praktiškai visi didžiausieji skirtumai vėlgi palyginti gerai atliepia lyginamųjų tekstų turiniui.

4 l e n t e l ė. **Raidžių dažnumo eilių (rangų) sutapimas įvairiuose tekstuose**

Lyginami tekstai	Skiriasi	Sutampa	Lyginami tekstai	Skiriasi	Sutampa	Lyginami tekstai	Skiriasi	Sutampa
1 ir 2	19	13	2 ir 7	24	8	4 ir 8	22	10
1 ir 3	25	7	2 ir 8	20	12	4 ir 9	22	10
1 ir 4	26	6	2 ir 9	23	9	5 ir 6	22	10
1 ir 5	21	11	3 ir 4	22	10	5 ir 7	23	9
1 ir 6	22	10	3 ir 5	20	12	5 ir 8	20	12
1 ir 7	27	5	3 ir 6	27	5	5 ir 9	22	10
1 ir 8	22	10	3 ir 7	25	7	6 ir 7	22	10
1 ir 9	26	6	3 ir 8	21	11	6 ir 8	28	4
2 ir 3	24	8	3 ir 9	26	6	6 ir 9	27	5
2 ir 4	24	8	4 ir 5	22	10	7 ir 8	26	6
2 ir 5	23	9	4 ir 6	23	9	7 ir 9	21	11
2 ir 6	21	11	4 ir 7	21	11	8 ir 9	27	5

Kadangi prielaida apie raidžių dažnumų ryšio su tekstų turiniu bei pobūdžiu galimumą vis iškyla, buvo atliktas dar vienas eksperimentas, kuriuo siekta nustatyti, kuriuose iš tiriamųjų tekstų raidžių santykiniai dažnumai apskritai skiriasi mažiausiai ir kuriuose – labiausiai. Vėl buvo imamos tos pačios 36 (t. y. visos įmanomos) tiriamųjų tekstų poros, bet dabar jose lyginti visų raidžių santykiniai dažnumai: poroms apskaičiuoti jau anksčiau aptarti dydžiai S (vidutinis kvadratinis skirtumas) ir L (vidutinis linijinis skirtumas). Tik, jei anksčiau buvo žiūrima empiriškai nustatytų ir teoriškai prognozuojamų dažnumų skirtumo, tai dabar imti tiktai empirinių santykinių dažnumų abiejuose poros tekstuose skirtumai, tiksliau – jų kvadratai bei moduliai. Tokiu būdu ir S, ir L – tik kiekvienas savaip, L išlikdamas visiškai neutralus skirtumų dydžiui, o S labiau išryškindamas didesniuosius ir neutralizuodamas mažesniuosius skirtumus – atspindi raidžių dažnumų visumos panašumo laipsnį suporuotuose tekstuose: kur jų reikšmės mažesnės, ten raidžių santykiniai dažnumai porą sudarančiuose tekstuose panašesni, artimesni vienas kitam, o kur jos didesnės – ten atitinkamai didesnis ir raidžių santykinių dažnumų visumos neatitikimas. Kartu rūpi pažiūrėti, kuris rodiklis, S ar L, apskritai imant, yra efektyvesnis, kuris iš jų geriau, adekvačiau atspindi (jeigu, žinoma, iš viso atspindi) tekstų pobūdžio panašumą. Pirminiai lyginimo rezultatai pateikti 5 lentelėje.

5 l e n t e l ė. **Raidžių dažnumų panašumo įvairiuose tekstuose įvertinimas**

Lyginami tekstai	S	L	Lyginami tekstai	S	L	Lyginami tekstai	S	L
1 ir 2	0,005050	0,003470	2 ir 7	0,003696	0,002933	4 ir 8	0,005842	0,004351
1 ir 3	0,004309	0,003438	2 ir 8	0,004334	0,003015	4 ir 9	0,003181	0,002259
1 ir 4	0,006894	0,004958	2 ir 9	0,004561	0,003034	5 ir 6	0,003726	0,002707
1 ir 5	0,004601	0,003148	3 ir 4	0,006638	0,004503	5 ir 7	0,003378	0,002668
1 ir 6	0,003511	0,002952	3 ir 5	0,003835	0,002934	5 ir 8	0,002599	0,001933
1 ir 7	0,005350	0,004305	3 ir 6	0,004558	0,003686	5 ir 9	0,004567	0,003141
1 ir 8	0,004903	0,003256	3 ir 7	0,005232	0,003898	6 ir 7	0,004199	0,003095
1 ir 9	0,006980	0,004823	3 ir 8	0,003784	0,002758	6 ir 8	0,004068	0,003235
2 ir 3	0,005793	0,003727	3 ir 9	0,006663	0,004488	6 ir 9	0,006151	0,004008
2 ir 4	0,004403	0,003323	4 ir 5	0,004646	0,003535	7 ir 8	0,004585	0,003594
2 ir 5	0,003569	0,002721	4 ir 6	0,005440	0,003789	7 ir 9	0,003840	0,002922
2 ir 6	0,004447	0,002944	4 ir 7	0,003480	0,002619	8 ir 9	0,006021	0,004124

Tačiau tekstų porų panašumo pagal raidžių santykinius dažnumus laipsnis dar geriau išryškėtų 5 lentelės duomenis kiek pertvarkius. Reikėtų sudaryti dvi tekstų porų eiles: vieną – išrikiavus tas poras *S* didėjimo, o kitą – *L* didėjimo tvarka. Eilės numeris ir atitiktų tekstų, sudarančių poras, panašumo mastą, nes eilės pradžioje atsidurtų pačių panašiausių, jos gale – pačių skirtingiausių tekstų poros, o palaipsniui einant nuo tokios eilės pradžios į pabaigą panašumas taip pat laipsniškai mažėtų. Šitaip sutvarkyti duomenys ir pateikti 6 lentelėje.

Iš jos aiškiai matyti, kad tiek pagal vidutinį kvadratinį (*S*), tiek ir pagal vidutinį linijinį (*L*) raidžių santykinių dažnumų skirtumą patys panašiausi yra 5 ir 8, t. y. žurnalų „Gimtoji kalba“ ir „Baltistica“ tekstai. Savaime suprantama, kad jie išties daug kuo artimi ir savo pobūdžiu bei turiniu. Taip pat abu šie rodikliai antruosius pagal panašumą iškelia 4 ir 9, t. y. romanų „Juodieji želmenys“ ir „Šiaurės rožė“ tekstus, be abejo priklausančius ir tam pačiam funkciniam stiliui, ir literatūros žanrui. Vadinasi, pagal raidžių dažnumą bent jau patys panašiausi tekstai iš tikro yra panašūs ir savo „aukštaisiais lygmenimis“ (funkciniu pobūdžiu, žanru, iš dalies – tematika ir pan.). Tačiau toliau tasai ryšys ima darytis nebe toks aiškus. Išsiskiria porų, išrikiuotų pagal *S* ir *L* eilės. Sakysim, jau trečioji vieta pagal *S* atitenka 5 ir 7, o pagal *L* – 4 ir 7 tekstų poroms. Šiuo atveju bene adekvatesnis būtų rodiklis *L*: 7 tekstas, V. Almonaičio knyga „Ką šniokščia Jūros rėvos“, kurion sudėti vaizdingi kelionių Žemaitijos upėmis aprašymai, istorinio bei kraštotyrinio pobūdžio pasakojimai, iš tiesų artimesnis yra taip pat pasakojamajam tekstui – romanui (5), negu kalbos žurnalui (4).

6 l e n t e l ė. **Tekstų panašumas pagal raidžių santykinių dažnumų visumą**

Panašumo	Tekstų pora pagal		Panašumo	Tekstų pora pagal		Panašumo	Tekstų pora pagal	
laipsnis	S	L	laipsnis	S	L	laipsnis	S	L
1	5 ir 8	5 ir 8	13	6 ir 7	2 ir 8	25	1 ir 2	3 ir 6
2	4 ir 9	4 ir 9	14	1 ir 3	2 ir 9	26	3 ir 7	2 ir 3
3	5 ir 7	4 ir 7	15	2 ir 8	6 ir 7	27	1 ir 7	4 ir 6
4	4 ir 7	5 ir 7	16	2 ir 4	5 ir 9	28	4 ir 6	3 ir 7
5	1 ir 6	5 ir 6	17	2 ir 6	1 ir 5	29	2 ir 3	6 ir 9
6	2 ir 5	2 ir 5	18	3 ir 6	6 ir 8	30	4 ir 8	8 ir 9
7	2 ir 7	3 ir 8	19	2 ir 9	1 ir 8	31	8 ir 9	1 ir 7
8	5 ir 6	7 ir 9	20	5 ir 9	2 ir 4	32	6 ir 9	4 ir 8
9	3 ir 8	2 ir 7	21	7 ir 8	1 ir 3	33	3 ir 4	3 ir 9
10	3 ir 5	3 ir 5	22	1 ir 5	1 ir 2	34	3 ir 9	3 ir 4
11	7 ir 9	2 ir 6	23	4 ir 5	4 ir 5	35	1 ir 4	1 ir 9
12	6 ir 8	1 ir 6	24	1 ir 8	7 ir 8	36	1 ir 9	1 ir 4

Apie tekstų, sudarančių dar tolesnes vietas šiose eilėse užimančias poras, giminingumą ar skirtingumą taip pat būtų galima mėginti samprotauti, tačiau šįkart labiau norėtųsi atkreipti dėmesį į eilių pabaigą, į kelias paskutiniąsias poras, kurias ir vienoje, ir kitoje eilėje sudaro iš tikro labai skirtingi, savo pobūdžiu tolimi tekstai (1 ir 9, 1 ir 4, 3 ir 9, 3 ir 4 ir t. t.). Todėl apskritai galima būtų sakyti, kad, nors raidžių, ypač – dažnesniųjų, santykiniai dažnumai tekstuose fliuktuoja gana žymiai, vis dėlto jų santykinių dažnumų skirtumų visuma gerokai priklauso nuo to, kokie tekstai lyginami: ji aiškiai mažesnė tarp žanriniu atžvilgiu artimų ir didesnė – tarp tuo pačiu atžvilgiu tolimų tekstų.

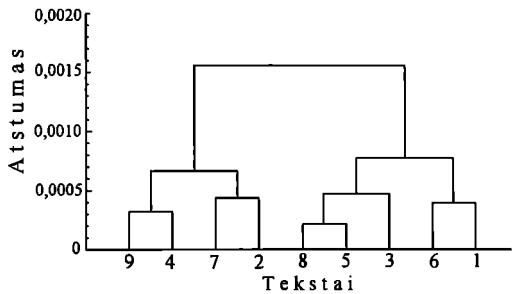
Vidutinis kvadratinis ir vidutinis linijinis skirtumai leidžia lyginti tekstus poromis, išryškinti panašiausių ir skirtingiausių tekstų poras, vienaip ar kitaip klasifikuoti visą įmanomų skirtingų jų porų visumą ir pan., bet dar patrauklesnė yra mintis paieškoti kokios nors universalesnės, visus tekstus iš karto aprėpiančios formalaus jų klasifikavimo, paremto raidžių dažnumais, procedūros ir pažiūrėti, ar tokios klasifikacijos rezultatai gerai atitinka natūralią tų tekstų skaidą pagal žanrus, pobūdį, turinį ir pan., ar ne. Sudėtingesnės statistinių duomenų apdorojimo procedūros iš esmės leidžia atlikti panašias taksonomijos operacijas (apie jas žr., pvz., Bitinas 1974, 135tt.; Ajvazjan ir kt. 1989, 249tt.). Iš jų mūsų kalbininkai bene labiausiai mėgsta ir plačiausiai taiko vadinamąjį hierarchinio grupavimo metodą, kartais dar vadinamą, nors, atrodo, ir ne visai preciziškai, klasterine analize (angl. *cluster analysis*): juo remiantis, daug kartų yra klasifikuoti garsai bei fonemos (pirmiausia minėtini A. Pakerio bei A. Girdenio ir jo mokinių, ypač G. Kačiuškienės ir I. Remenytės-Mažiulienės darbai), taip pat kalbos dalys (Žilinskienė, 1981; Linkevičienė, 1994) ir netgi mėginti klasifikuoti

taip pat tekstai, tiesa, tik vieno, mokslinio, funkcinio stiliaus ir charakterizuoti tik trimis požymiais (Bitinienė, 1993). Kita vertus, šiuolaikinės programinės priemonės leidžia gana lanksčiai parinkti reikiamą klasterinės analizės atmainą ir laisvai kaitalioti įvairius jos parametrus, tad atsiranda geros galimybės eksperimentuoti bei ieškoti kuo tikslesnio formalinės klasifikacijos atitikimo realiai tiriamųjų objektų sklaidai. Tad pasinaudojant specializuotu programų paketu *STATISTICA* (v. 4.3) ir buvo pamėginta klasterinės analizės būdu tiriamuosius tekstus sugrupuoti į kokias nors logiškai paaiškinamas hierarchiškas grupes (vadinamuosius taksonus arba klasterius).

Pradiniai grupavimo duomenys yra iš 1 lentelės paimti raidžių dažnumai. Kiekvienas tekstas traktuojamas kaip atskiras objektas, o visos raidės – kaip 32 skirtingi tų objektų požymiai. Tad kiekvienos raidės dažnumas kiekviename tekste yra tarsi konkretus to objekto atitinkamo požymio įvertis (reikšmė). Todėl galima vaizduotis, kad 1 lentelės duomenys išsamiai „aprašo" 9 skirtingų objektų (tekstų) padėtį (lokalizaciją) 32 matavimų erdvėje. Klasterinės analizės tikslas ir būtų aptikti bei išryškinti toje erdvėje susiformuojančias tų objektų sankaupas (taksonus). Šį kartą buvo pasinaudota jungiančiąja klasterizacija (angl. *joining* arba *tree clustering*), nes ji grupavimo rezultatus – vis sudėtingesnės struktūros taksonus – vaizduoja lingvistiniuose darbuose itin įprastu būdu – dendrograma. „Nuotolio" tarp klasifikuojamų objektų (tekstų) matu imtas vadinamasis kvadratinis euklidinis atstumas (angl. *squared Euclidean distances*), kuris yra labai artimas ne kartą minėtam vidutiniam kvadratiniam skirtumui *S* ir šiuo atveju lygus tiesiog dažnumų skirtumų kvadratų sumai. O formuojamų taksonų padėtis vienas kito atžvilgiu nustatoma (angl. *amalgamation* arba *linkage rule*) pagal vadinamąjį „tolimiausios kaimynystės" principą (angl. *complete linkage* arba *furthest neighbor*), t. y. pagal visų didžiausią tiems taksonams priskirtų objektų (tekstų) „nuotolį". Tekstai buvo sugrupuoti taip, kaip parodyta 3 pav.

Grupavimo rezultatai iš tikrųjų pakankamai logiški ir įtikinantys. Bent jau 3 iš 4 mažiausiųjų taksonų sudaro žanro ir turinio atžvilgiu itin artimi tekstai – 8 ir 5 (lingvistikos žurnalai), 9 ir 4 (romanai) bei 7 ir 2 (apybraižų tipo knygos, pasakojamoji publicistika). Sunkiau būtų paaiškinti ketvirtą minimalųjį taksoną, suformuotą iš 6 (M. Tomonio raštai: filosofiniai apmąstymai, eseistiniai rašiniai, poezija, dienoraščiai, laiškai) ir 1 „Moksleivio fiziologija" tekstų, tačiau juos kaip tik ir vienija tai, kad abu jie yra ir savo turiniu, ir žanru nepanašūs nė į vieną kitą tirtąjį tekstą, individualūs, tad neturėdami vidinio, jie vis dėlto turi tam tikro išorinio, pasakytum, „situacinio" panašumo. Aukštesnių lygmenų taksonai taip pat visiškai logiški: prie lingvistinių žurnalų toliau prisijungia B. Stundžios studija „Lietuvių bendrinės kalbos kirčiavimo sistema" (3 tekstas) ir susiformuoja lingvistinių tekstų grupę atitinkantis taksonas, o prie šio prisijungus minėtam abejonių kiek keliančiam taksonui iš 6 ir 1 tekstų, susidaro vienas iš dviejų stambių taksonų, kuris atitinka nesiužetinių, mokslinės pakraipos tekstų grupę. Tuo tarpu romanų ir pasakojamosios publicistikos taksonai, susijungę tarpusavy, sudaro antrąjį stambų taksoną, atitinkantį siužetu ir daugiau ar mažiau nuosekliu pasakojimu besiremiančius teks-

tus. Vadinasi, raidžių dažnumais paremta formali tekstų klasifikacija šiuo atveju pakankamai vykusiai atspindi natūralią jų skaidą, tik reikia atitinkamai parinkti klasifikavimo procedūrų tipus ir parametrus.



3 pav. **Tekstų grupavimo pagal raidžių dažnumus klasterinės analizės būdu dendrograma**

Pabaigai – glaustos išvados.

1. Šiuo tyrimu nustatyti raidžių santykiniai dažnumai iš esmės panašūs į tuos, kuriuos buvo nustačiusi V. Žilinskienė (1978), o palyginti nedideli skirtumai gali būti aiškinami natūraliais raidžių santykinio dažnumo svyravimais, kurie išryškėjo ir tarp atskirų šiam tyrimui paimtų tekstų.

2. Raidžių dažnumų (p) priklausomybę nuo jų rangų (r) galima aproksimuoti lygtimi $p_r=k+ae^{-(r/b)}$, atitinkamai parenkant koeficientus k, a ir b.

3. Raidžių dažnumų mažėjimo kreivė visuose tirtuose tekstuose yra iš esmės panaši, nors kai kurie jos formos ypatumai galbūt ir galėtų būti atsargiai siejami su teksto pobūdžiu.

4. Raidžių, išdėstytų dažnumų mažėjimo tvarka, sekos skirtinguose tekstuose smarkiai įvairuoja.

5. Raidžių dažnumų atžvilgiu patys panašiausi tekstai yra panašūs ir savo žanriniu pobūdžiu, o patys skirtingiausi – ryškiai skiriasi.

6. Klasterinės analizės būdu tekstus galima pagal raidžių dažnumus sugrupuoti į pakankamai natūralias, jų žanrinę specifiką bei kalbos tipą (funkcinį stilių) gerai atitinkančias struktūrines grupes.

ЧАСТОТА БУКВ В ТЕКСТАХ ЛИТОВСКОГО ЛИТЕРАТУРНОГО ЯЗЫКА

Резюме

В статье рассматриваются частоты букв литовского алфавита в 9 сравнительно больших текстах на современном литовском литературном языке. Заново определенные значения частот букв сравниваются с соответствующими значениями, ранее предоставленными В. Жилинскене; исследуется общий характер убывания частот и зависимость значения частоты от ее ранга; для отражения этой зависимости предлагается вариант экспоненциально убывающей функции. Установлено, что в рядах букв, ранжированных по частоте, позиции конкретных букв расходятся в весьма значительной мере: при сравнении всех текстов попарно обнаружено, что в таких рядах совпадают в среднем лишь позиции 9 букв. Делается попытка установить соответствие между распределением частот букв и жанровым или языковым характером текста. Выявлено, что наиболее близкие по частоте букв тексты являются в данном отношении сходными, а самые разные – весьма различными. Возможность такого соответствия подтверждает и группировка текстов методом кластерного анализа.

LITERATŪRA

Ajvazjan ir kt., 1983 – Айвазян С. А., Енюков И. С., Мешалкин Л. Д. Прикладная статистика: Основы моделирования и первичная обработка данных. Москва: Финансы и статистика.

Ajvazjan ir kt., 1989 – Айвазян С. А., Бухштабер В. М., Енюков И. С., Мешалкин Л. Д. Прикладная статистика: Классификация и снижение размерности. Москва: Финансы и статистика.

Bitinas B., 1974, Statistiniai metodai pedagogikoje ir psichologijoje. Kaunas: Šviesa

Bitinienė A., 1993, Sakinio ilgis – statistinis stiliaus parametras. – Kalbotyra, t. 42 (1), 4–16.

Fišas M., 1968, Tikimybių teorija ir matematinė statistika. Vilnius: Mintis.

Jaglom, 1973 – Яглом А. М., Яглом И. М., Вероятность и информация. Москва: Наука.

Kruopis J., 1993, Matematinė statistika. Vilnius: Mokslas.

Kubilius J., 1980, Tikimybių teorija ir matematinė statistika. Vilnius: Mokslas.

Linkevičienė N., 1994, Kalbos dalių dažnumas šiaurės žemaičių tarmėje. – Kalbotyra, t. 43(1), 53–60.

Markov, 1913 – Марков А. А., Пример статистического исследования над текстом „Евгения Онегина" иллюстрирующий связь испытании в цепь. – Известия Академии наук, т. 7 (6).

Piotrovskij, 1975 – Пиотровский Р. Г., Текст, машина, человек. Ленинград: Наука.

Piotrovskij ir kt., 1977 – Пиотровский Р. Г., Бектаев К. Б., Пиотровская А. А., Математическая лингвистика. Москва: Высшая школа.

Tutubalin, 1972 – Тутубалин В. Н., Теория вероятностей. Москва: МГУ.

Žilinskienė V., 1978, Lietuvių kalbos raidžių dažnumas publicistikos tekstuose. – Kalbotyra, t. 29(1), 83–95.

Žilinskienė V., 1981, Lietuvių publicistikos tekstų kalbos dalių koreliacinė ir klasterinė analizė. – Kalbotyra, t. 32(1), 121–133.

Vilniaus universitetas
Bendrosios kalbotyros katedra

Įteikta
1997 m. spalio mėn.