

RAIDŽIŲ DAŽNIŲ LIETUVIŲ IR KITOSE KALBOSE, VARTOJANČIOSE LOTYNIŠKUS RAŠMENIS, ANALIZĖ

Gintautas GRIGAS¹, Anita JUŠKEVIČIENĖ²

Vilniaus universitetas, Matematikos ir informatikos institutas, Akademijos g. 4, 08663 Vilnius, Lietuva
El. paštas: ¹gintautas.grigas@mii.vu.lt; ²anita.juskeviciene@mii.vu.lt

Įteikta 2015-08-31; priimta 2015-11-23

Abėcėlių savitumus, ypač raidžių dažnius, svarbu įvertinti projektuojant klaviatūras, analizuojant tekstus, projektuojant abėcėlinius žaidimus, atliekant kitus darbus su tekstais. Todėl svarbu nustatyti, kas gali būti naudinga tekstų įvedimo priemonių ir kitokių įrenginių, susijusių su raidžių aibėmis, projektuotojams. Vis daugiau tekstų publikuojama internete. Siekiant adekvačiai palyginti įvairių kalbų raidžių dažnius, vartojamus internetinėje erdvėje, jie buvo paimti iš visoms kalboms giminingo šaltinio – Vikipedijos. Pasiūlytas ir taikytas dažnių santykių ir pokyčių metodas padėjo išryškinti panašumus ir skirtumus tarp lietuvių ir kitų kalbų. Pagrindinis dėmesys skiriamas raidžių, esančių takoskyroje tarp rečiausiai vartojamų savos abėcėlės ir dažniausiai vartojamų užsieninių, dažnių analizei.

Reikšminiai žodžiai: abėcėlė, diakritinis ženklas, klaviatūra, lietuvių kalba, lotyniški rašmenys, mobilusis telefonas, raidė, raidžių dažniai, teksto įvedimas.

Įvadas

Lotyniški rašmenys vartojami daugelyje kalbų, bet jų abėcėlės skirtingos. Tam tikras bendrumas atsiranda dėl pagrindinės lotynų abėcėlės, turinčios 26 raides¹. Ją galima vadinti kitų kalbų abėcėlių stuburu. Tačiau ne visos lotynų abėcėlės raidės įtrauktos į daugelio kitų kalbų abėcėles. Kita vertus, kiekviena kalba turi

savitųjų raidžių. Todėl lotyniško pagrindo abėcėlių įvairovė didelė, kur kas didesnė negu kitų rašmenų grupių (pvz., graikų, kirilicos) abėcėlių.

Abėcėlių savitumus, ypač raidžių dažnį, svarbu įvertinti projektuojant klaviatūras, analizuojant tekstus, projektuojant abėcėlėmis pagrįstus žaidimus, atliekant kitus darbus su tekstais. Vis daugiau tekstų publikuojama internete. Todėl svarbu pasitikrinti, kokia yra dažninė internetinių tekstų charakteristika.

Lietuvių kalba – viena iš Europos Sąjungos oficialiųjų kalbų. Todėl svarbu matyti lietuvių

¹ Pagrindine lotyniška abėcėle vadinsime klasikinę lotynų kalbos abėcėlę, turinčią 23 raides, papildytą raidėmis J, U ir W. Faktiškai tai anglų kalbos abėcėlė.

kalbos abėcėlės vietą tarp kitų kalbų abėcėlių. Mūsų tikslas – ištirti lietuvių kalbos abėcėlės raidžių vartoseną skaitmeninėje, ypač internetinėje, erdvėje, palyginti įvairių kalbų raidžių dažnines charakteristikas, pagrindinį dėmesį skiriant savitosioms raidėms ir lietuvių kalbos abėcėlėje nesančioms raidėms q , w ir x .

Savitųjų raidžių paprastai būna mažiau negu perimtų iš pagrindinės lotyniškos abėcėlės ir jų dažnis mažesnis. Stengiantis sumažinti mobiliųjų telefonų, planšetinių kompiuterių ir kitų mažų įrenginių klaviatūras, kartais atsisakoma dalies raidžių rinkimo tiesiogiai, t. y. vienu klavišo paspaudimu (palietimu). Į atsisakomų raidžių sąrašą dažnai patenka savitosios raidės ar jų dalis (Bi *et al.* 2012). Kita vertus, didėjant globalizacijai dažniau prireikia svetimų kalbų raidžių. Norint nubrėžti ribą, po kurios raidžių rinkimas galėtų būti lėtesnis, pavyzdžiui, keliais klavišų paspaudimais, reikia žinoti rečiau vartojamų raidžių dažnius.

Raidės dažnis – tai tos raidės pasikartojimų skaičiaus santykis su visų raidžių skaičiumi tiriamame tekste, išreikštas procentais. Skaičiuosime tik raides, esančias tiriamų 25 kalbų abėcėlėse. Kitų rašto sistemų (pvz., graikų, kirilicos) raidžių ir kitų ženklų (pvz., skaitmenų, skyrybos ženklų, tarpų) į skaičiavimus neįtrauksime. Neskirsime didžiųjų raidžių nuo jas atitinkančių mažųjų.

Lietuvių kalbos raidžių dažnių statistiką pateikia Vida Žilinskienė knygoje „Lietuvių kalbos dažninis žodynas“ (Žilinskienė 1990). Į skaičiavimus įtraukta arti dviejų milijonų raidžių. Nebuvo įtraukti tikriniai daiktavardžiai, nes, pasak autorės, jie didesnės reikšmės lyginamiesiems leksikografiniams tyrimams neturi. Nesant tikrinių daiktavardžių, į statistiką nepateko ir užsienietiškuose asmenvardžiuose pasitaikančios raidės, kurių nėra lietuvių kalbos abėcėlėje (pvz.: q , w , x).

Lietuvių kalbos žodžių dažniai pateikti dar dviejuose žodynuose (Utkā 2009; Grumadienė, Žilinskienė 1998). Jų imtis didesnė (per penkis milijonus raidžių). Iš juose pateiktos

medžiagos galima suskaičiuoti ir raidžių dažnius. Tačiau ir šiuose vengiama tikrinių daiktavardžių.

Analogiški statistiniai duomenys yra ir apie kitų kalbų raidžių dažnius, bet tie šaltiniai labai įvairūs. Raidžių dažniai priklauso nuo teksto stiliaus, paskirties, autoriaus. Kriptografijoje tokie duomenys naudojami tekstų autorystei nustatyti. Dėl to lyginamų kalbų statistika turėtų būti gauta iš tokios pat rūšies tekstų. Todėl mūsų analizei reikia ieškoti stambesnių daugiakalbių šaltinių.

Praktinės kriptografijos interneto svetainėje (Practical Cryptography 2015) pateikiami dešimties kalbų (anglų, danų, islandų, ispanų, lenkų, prancūzų, rusų, suomių, švedų, vokiečių) raidžių dažniai. Imtys didelės, ne mažesnės kaip po 90 milijonų kiekvienos kalbos rašto ženklų. Deja, ištekluose nėra nei lietuvių, nei mums giminingos latvių kalbos.

Tematikos įvairovė ir straipsnių gausa pasižymi Vikipedija (Wikipedia 2015). Ją kuria daugelis autorių. Joje aprašomi globalūs reiškiniai, jos tekstuose dažniau vartojami svetimų kalbų rašto ženklai ir todėl išryškėja ryšiai tarp kalbų. Todėl šis šaltinis gerai tiktų mūsų tyrimams. Juolab kad pradžiai jau padaryta. Rašto ženklų statistiką 262 kalbų Vikipedijų tekstuose pateikia Denis Vrandečicas (Denny Vrandečić) interneto svetainėje (Vrandečić 2012) ir analizuoja straipsnyje (Vrandečić *et al.* 2011). Tai unigramų (rašto ženklų), diagramų ir trigramų skaičiai. Todėl analizei pasirinkome minėtos svetainės (Vrandečić 2012) išteklis. Toliau juos vadinsime Vikipedijų ženklų baze, trumpiau – Baze.

Kalbų ir raidžių rinkiniai

Europos Sąjungoje šiuo metu yra 24 oficialiosios kalbos. Iš jų 22-jose vartojami lotyniški rašmenys. Jas visas įtraukėme į analizę. Taip pat įtraukėme dar tris Europoje dažniau vartojamas kalbas: islandų, norvegų ir turkų. Taigi iš viso analizuosime 25 kalbas. Visų šių kalbų statistika yra pateikta Bazėje.

Dažniausiai domins santykis tarp gretimų raidžių, t. y. kai $j = i + 1$. Santykinį dažnių pokytį $d(i, j) = (f_i - f_j)/f_j = f_i/f_j - 1$ vadinsime tiesiog pokyčiu.

Pokytis rodo, kiek kartų raidės *i* dažnis f_i didesnis už raidės *j* dažnis f_j (arba f_j mažesnis už f_i). Jeigu raidės išdėstytos jų dažnių mažėjimo eile, tai dažnių pokytis tarp gretimų raidžių bus lygus nuliui (kai abiejų raidžių dažnis vienodas) arba didesnis už nulį. Tuo pokytį kartais patogiau naudoti negu santykį.

Didelius pokyčius tarp gretimų raidžių vadiname dažnių lūžiais. Identifikuosime dažnio lūžio taškus – dažnio pokyčius, ne mažesnius už vienetą, t. y. tokius, kai gretimos raidės dažnis sumažėja du ar daugiau kartų.

Lietuvių kalbos raidžių dažniai

Vikipedijos raidžių dažnius palyginome su dažniais, pateiktais žodynuose (2 lentelė).

2 lentelė. Lietuvių kalbos raidžių dažniai

Raidė	V. Žilinskienės žodyne	A. Utkos žodyne	Vikipedijoje	
	%	%	%	Pokytis
i	14,0	13,7081	12,9593	
a	11,1	11,4997	11,1912	0,1580
s	7,7	7,8496	7,8811	0,4200
o	6,0	5,9718	6,7429	0,1688
r	5,4	5,4813	5,6689	0,1895
e	6,4	5,4946	5,6205	0,0086
t	5,8	5,9735	5,3323	0,0540
n	5,2	5,0750	5,1441	0,0366
u	5,1	4,6032	4,5860	0,1217
k	4,7	4,4993	4,1708	0,0995
m	3,8	3,6232	3,5790	0,1654
l	2,9	3,0893	3,5019	0,0220
p	2,9	2,9297	2,7344	0,2806
v	2,2	2,3537	2,6557	0,0296
d	2,5	2,5225	2,5802	0,0293
j	2,1	1,9802	2,3380	0,1036
g	1,8	1,9203	1,7946	0,3028
ė	0,6	1,5486	1,6643	0,0783
b	1,6	1,5162	1,4842	0,1213
y	1,4	1,3795	1,4325	0,0361

Raidė	V. Žilinskienės žodyne	A. Utkos žodyne	Vikipedijoje	
	%	%	%	Pokytis
ų	1,5	1,2424	1,2602	0,1367
š	1,3	1,2720	1,1339	0,1114
ž	0,9	0,8521	0,7965	0,4236
c	0,5	0,4359	0,6019	0,3230
ą	0,8	0,7287	0,5418	0,1109
į	0,6	0,6282	0,4848	0,1176
č	0,3	0,4269	0,4303	0,1267
ū	0,2	0,4894	0,4079	0,0549
f	0,2	0,2471	0,3468	0,1762
z	0,2	0,2808	0,3454	0,0041
h	0,1	0,1240	0,2760	0,2514
ę	0,2	0,2293	0,1721	0,6037
x		0,0170	0,0733	1,3479
w		0,0052	0,0368	0,9918
q		0,0008	0,0054	5,8148
é			0,0038	0,4211
á			0,0024	0,5833
ó			0,0018	0,3333
í			0,0015	0,2000
ö			0,0014	0,0714
ü			0,0013	0,0769
ä			0,0012	0,0833
ł			0,0010	0,2000
ä			0,0010	0
Kitos	0	0,0009	0,0096	
Iš viso	100	100	100	

Lentelėje raidžių dažniai pateikiami procentais keturių ženklų po kablelio tikslumu iki tol, kol dažnis užrašytas ne mažiau kaip dviem reikšminiais skaitmenimis. Suminiai duomenys visų kitų, dar rečiau vartojamų, raidžių pateikiami eilutėje *Kitos*. Jų procentai Vikipedijoje:

- ī – 0,0007,
- ñ, ú – 0,0005,
- ç, ē, è – 0,0004,
- ā, à, á, é, í, ã, ø – 0,0003,
- ë, â, ß, ž, š, ı, â, ä, æ – 0,0002,
- ì, ô, õ, ê, î, ý, ñ, ò, õ, ù, ğ – 0,0001.

Utkos žodyne (2009):

ž – 0,0007,

ā, ī – 0,0001.

Didžiausi dažnių pokyčiai (šuočiai) yra abėcėlės pabaigoje – ten, kur prasideda svetimų kalbų raidės.

Žilinskienės žodyne (1990) dažnių procentai apskaičiuoti įvertinant ir tarpo ženklą, kurio dažnis yra 13,1 proc. Kituose šaltiniuose tarpai neskaičiuojami. Todėl procentus perskaičiavome pagal formulę $f'_i = f_i \times 100 / (100 - 13,1)$, kai f_i – i -tosios raidės dažnis.

Koreliacijos koeficientai tarp trijų šaltinių dažnių artimi vienetui: tarp Vikipedijos ir Žilinskienės žodyno – 0,993, tarp Vikipedijos ir Utkos žodyno – 0,996, tarp abiejų žodynų – 0,996.

Todėl galima tvirtinti, kad duomenys patikimi. Vikipedijos dažnių pokyčiai tarp gretimų raidžių, užrašytų ties tarpais tarp stulpelių, grafiškai pavaizduoti 1 pav.

Didelių šuočių zona išsiskiria ties raidėmis x , w , q . Pirmasis lūžis tarp raidžių ξ – x . Jis sutampa su abėcėlės pabaiga. Ne ką mažesnis

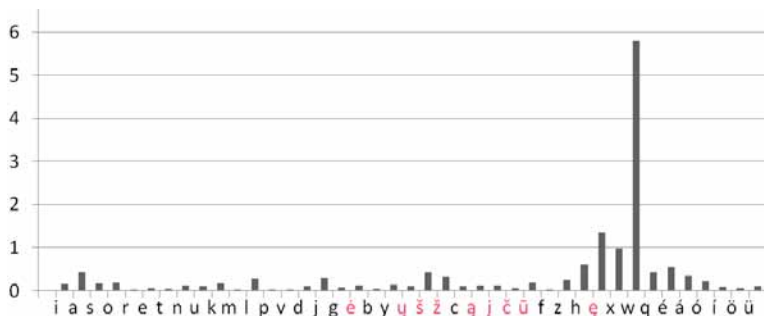
ir antrasis lūžis tarp raidžių x – w , o tarp w – q – labai didelis.

Pirmoji dažniausia raidė, nesanti abėcėlėje, yra x . Ji vartojama ne tik svetimžodžiuose, bet atlieka ir romėniško skaitmens X funkciją. Rašytinės kalbos dažnių žodyne (Utko 2009) radome 0,002 proc. ženklų x , turinčių raidės reikšmę, ir 0,015 proc. – romėniško dešimtuko. Tuo galima paaiškinti, kad ženklas x lietuviškoje klaviatūroje yra skaitmenų eilėje – ten, kur ir kiti specialieji ženklai. Pirmųjų rašomųjų mašinelių klaviatūrose, kuriose buvo striuka su klavišų skaičiumi, iš minėtų trijų raidžių buvo tik x . Tik vėliau, kai klaviatūros padidėjo dviem klavišais, į ją buvo įtrauktos raidės w ir q (Grigas 2014). Raidei q buvo paskirtas klavišas, neturintis pastovios vietos, taigi pats prasčiausias.

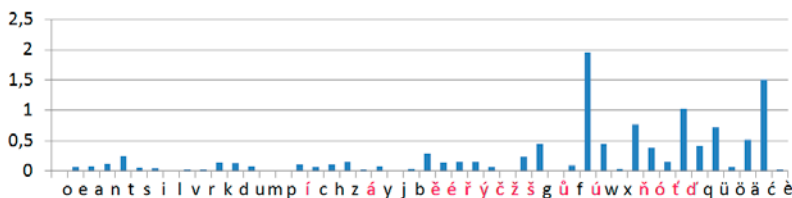
Reziumuojant galima pasakyti, kad turėjome tris lietuviškų klaviatūrų tipus su:

- 32 raidėmis,
- 32 + 1 raidžių (papildyta raide x),
- 32 + 3 raidžių (papildyta raidėmis q , w ir x).

Palyginimui pateiksime čekų kalbos dažnių pokyčių diagramą (2 pav.).



1 pav. Lietuvių kalbos raidžių dažnių pokyčiai



2 pav. Čekų kalbos raidžių dažnių pokyčiai

Diagramoje matome tris didelius lūžius. Po pirmojo, didžiausio, lūžio, esančio tarp raidžių *f* ir *ú*, patenka keletas čekų abėcėlės raidžių (*ú, ř, ó, t, d*), tarp kurių yra taip pat nemažų lūžių. Tuo čekų kalbos raidžių dažnių diagramos pavidalas gerokai skiriasi nuo lietuvių kalbos raidžių dažnių diagramos.

Dažnių pokyčiai ir lūžių taškai tarp gretimų raidžių Europos kalbose

Visų kalbų pokyčių diagramos užimtų daug vietos. Todėl pokyčius pateiksime supaprastintu pavidalu (3 lentelė). Raidės suskirstytos

į keturias dažnių grupes ir išdėstytos jų dažnių mažėjimo eile.

Raidės, esančios standarto ETSI grupėje A, paryškintos. Pagrindines lotyniškos abėcėlės raidės, kurios standarte ETSI priskirtos grupei B, pabrauktos. Į lūžių vietas įterpti balti skaitmenys juodame fone, reiškiantys sveikąsias pokyčių dalis.

Pokyčių šuoliai nevienodi. Mažiausi (po vienetą) čekų, vokiečių, estų, suomių ir norvegų kalbose. Vadinasi, šiose kalbose sunku nubrėžti ribą tarp dažnai ir retai vartojamų raidžių.

Didžiausias šuolis (11), pažymėtas kvadratu, priklauso ispanų kalbai. Ne ką mažesni (8) – anglų ir italų kalboms. Toliau eina lietuvių kalba (5).

3 lentelė. 25 kalbų raidės, išdėstytos jų dažnių mažėjimo eile ir lūžiai tarp jų

	$f \geq 0,1$	$0,1 > f \geq 0,01$	$0,01 > f \geq 0,001$	$0,001 > f \geq 0,0001$
cs	oeantsilvrkdumpíchzávyjběérýčžšgúf <u>1</u> úw	xñ	ót <u>1</u> dq	üöä <u>1</u> čèl'èl'
da	erntaisldlogmkfvubhpâyöæcj <u>2</u> w	z	xéqöü	ä <u>1</u> áoíèðásßlç
de	enirstadhulgocmbfkwpvüäöjyß	<u>1</u> x	<u>1</u> q <u>1</u> é	<u>1</u> áoíèšçčäløácúäüö
en	eationsrhldcumfpgwybv <u>1</u> xjzq	<u>8</u> é	<u>1</u> áoüíóääèøç	ñšæ...èà...è...í...œö
es	eaonsrildtcumpbgvfyóhqjzáexñkú	w	ü <u>1</u> öèçäää	äaüèšcôö
et	aiestlunokrdmvgpjhäböü <u>1</u> fcö <u>1</u> y	wz	šx <u>1</u> žqé	ääöóíä
fi	aintesloukämrvjhpyd <u>1</u> ögcbfw	z	x <u>1</u> šéq	<u>1</u> äüäóözi <u>1</u> èøç
fr	easnitrluodcpmévqfbhqàxèyjkèz	wçóáiüœüi	<u>2</u> éäüíóíöäüñ	šcúa
ga	ainhrestclodgmubíáéíóúp <u>4</u> vyk	wjzxq	<u>1</u> äöü <u>1</u> äl'èèçñòšäç	äðö...æ...è...qí...œ
hr	aioenrjstuklvdmpgzbçhšžčfđy <u>w</u>	<u>2</u> x <u>1</u> qé	áuüíóáüüääçñäè	
hu	eatlsnkriozáégmbyvdhupjőfcóóíüüü	<u>1</u> wx	<u>3</u> q	<u>2</u> šäöçčäèšçžlô
is	arniestulgmðkfovhdíájböpyðæüécý	<u>1</u> wx	z <u>2</u> q	üöäääèšçláíäüä
it	eiaonltrscdupmgvfbzh <u>1</u> qèàkyò	wùjxié	<u>3</u> áoüíóíäçšäçšäüñä	üö
lt	iasoretntukmlpvdjgèbyqšžcąčüfzhe	<u>1</u> x	<u>w</u>	<u>5</u> qéáoíöüäläñ
lv	aisetrunäoklmdpvjzjgbcš <u>1</u> ñflühžkğčl	<u>1</u> ywx	<u>2</u> q <u>1</u> óäüö	óí...ö...ī
mt	aaltenrumsojkdflbpgwhzğzxqvcç <u>2</u> y	à <u>4</u> éü	òèäüçíóöçíšçä	
nl	enairtodslghvmukcpbwjzf <u>2</u> y	<u>1</u> xèé	<u>1</u> qèüüö	óáäflçšçöçääßä
no	erntasioldgkmvfpubháyjoc <u>1</u> æw	zxéqöä	üáoíèäðäüásçlçíüçö	ñè
pl	aioenrzwscdkpdmuljlgbhąęóźśf <u>1</u> ńcý	žx <u>2</u> éq	áuö	íäšèçø
pt	aeosirdntmuclp <u>1</u> gvbfhâqéçázíjxókyêwóú	ääó	<u>4</u> ü <u>1</u> öèänlş	ääšè
ro	eiarntulocsdpmäfvígbsťzhájxky <u>w</u>	<u>4</u> éq	áuöèóíäöäşçşlúçç	ßääöè
sk	oaienrvltskdmpuchjbzáýýíčěšžúgflťóôw <u>x</u>	ñďä	<u>4</u> lq	řěöřèü <u>1</u> lúí
sl	eaionrstjvkdpmuzbghčcsž <u>1</u> ý	<u>w</u> <u>2</u> xéqčá	üöíóíèäüðçäöóöł	èääiè
sv	eanrtsildomkgvfhupäçböäýj <u>3</u> xw	zé <u>1</u> q	<u>1</u> üáoèóíäöäçšçéçü	ðñlā
tr	aeinrlhkdstmyuobüşvzghçpçğöf <u>4</u> j	<u>w</u> <u>2</u> xâiq	é <u>1</u> üáoíäääš	ñcúéäč

Kuo didesnis šuolis, tuo ryškiau atsiskiria retai vartojamos raidės.

Jeigu pirmojo šuolio kairėje yra visos kalbos abėcėlės raidės, o dešinėje visos kalboje pasitaikančios svetimų kalbų raidės, tai galima tvirtinti, kad abėcėlės raidžių vartojimas yra tolygus ir abėcėlė atitinka realių raidžių vartojimą internetinėje erdvėje. Iš lentelės matome, kad lietuvių kalba šį kriterijų tenkina. Po didžiausio anglų kalbos lūžio matome dar 12 paryškintų raidžių: *é, ô, è, ç, ñ, æ, ë, a, ê, î, œ, õ*, kurios įtrauktos į ETSI standarto raidžių grupę A. Vadinasi, pagal šį standartą jos priivalomos ir turėtų turėti anglų kalbos abėcėlei priklausančių raidžių statusą, analogišką kitų kalbų savitosioms raidėms.

Yra nuomonių, kad anglų kalboje yra ne vien 26 pagrindinės lotynų abėcėlės raidės. Michaelis Eversonas rašo (Everson 2004), kad dažnai pasitaiko iš kitų kalbų perimtų svetimžodžių, kuriuose yra raidžių su diakritiniais ženklais, pavyzdžiui: *à la carte, abbé, Ægean, archaeology, beloved, café*. Jis pateikia anglų kalbos abėcėlę, papildytą net 16 raidžių. Tačiau, kaip matome, šios raidės patenka už didelio lūžio (8). Taigi ir jų vartoseną menka.

Papildomų raidžių įtraukimas į standartą ETSI sukėlė prieštaravimų pačiame standarte. Jos, kaip esančios pagrindinėje raidžių grupėje, turėtų būti traktuojamos kaip ir kitos šios grupės raidės, užrašytos ant klavišų abėcėlės tvarka (pvz., *2aàæbc*) ir tokia pat eile (*aàæbc*) renkamos. Tačiau į klaviatūrą T12 jos įtrauktos kaip antraeilės – nukeltos po skaitmenų ir neužrašomos ant klavišų. Pavyzdžiui, klavišo *2abc* raidės išdėstytos tokia rinkimo eile *abc2aæ*. Renkant po skaitmens einančią raidę reikalingas papildomas klavišo paspaudimas skaitmeniui peršokti. Anglų kalbai tai tinka, nes jos raidžių dažnių lūžis šioje vietoje didelis.

Standarte ETSI pagal tą pačią taisyklę dėstomos ir visų kitų kalbų raidės – savitosios raidės nukeliamos po skaitmenų, nors jos priklauso grupei A ir daugelio jų dažnis didesnis negu prieš skaitmenis einančių raidžių. Dėl to padidėja paspaudimų skaičius ir sumažėja rinkimo sparta.

Iš viso to galima daryti prielaidą, kad dėl dažniais nepagrįsto raidžių dėstymo standarte ETSI atsirado prieštaravimas (dėstymas neatitinka raidžių skirstymo į grupes A ir B) ir dėl to sumažėjo visų kalbų, išskyrus anglų, SMS žinučių rinkimo sparta.

Standarte ETSI pateikto raidžių skirstymo į grupes A ir B neatitiktį jų vartojimo dažniams, kai prieš grupės A raidę eina grupės B raidė, matome ir kitose kalbose.

Nutolusių raidžių dažnių santykiai

Dažnių santykius ir pokyčius galima analizuoti ne tik tarp gretimų raidžių dažnių juostoje, bet ir tarp nutolusių raidžių. Dažnių santykių gali prireikti įvertinant raidžių išdėstymo klaviatūroje alternatyvas, pavyzdžiui, kuriai raidei ar raidžių grupei skirti patogesni rinkimo būdai. Savitųjų ir rečiau vartojamų pagrindinės lotynų abėcėlės raidžių dažnių santykiai pateikti 4 lentelėje. Naudosime tokius žymenis:

f_{s1} – dažniausios (1-osios sąrašė) savitosios raidės dažnis,

f_{sn} – rečiausios (paskutinės sąrašė, nesuskliaustos) savitosios raidės dažnis,

f_{svisu} – visų savitųjų raidžių dažnis,

f_{pn} – rečiausios pagrindinės lotynų abėcėlės raidės dažnis,

f_{s1}/f_{pn} – dažniausios savitosios raidės dažnio santykis su rečiausios pagrindinės lotynų abėcėlės raidės dažniu,

f_{sn}/f_{pn} – rečiausios savitosios raidės dažnio santykis su rečiausios pagrindinės lotynų abėcėlės raidės dažniu.

Savitųjų raidžių sąrašė suskliaustos raidės tos, kurios įtrauktos į standarto ETSI raidžių grupę A, bet pagal jų dažnius (žr. 3 lentelę) šiai grupei neturėtų priklausyti.

Savitąsias raides dažniausiai (per 10 %) vartoja čekai, vengrai, islandai, latviai, turkai. Lietuviai – netoli vidurkio.

Dažnių santykiai rodo savitųjų kalbos raidžių svarbą ir tos svarbos ribinius įvertinimus dažniausiai (f_{s1}/f_{pn}) ir rečiausiai (f_{sn}/f_{pn}) vartojamos raidės atžvilgiu. Dažniausios savitosios

4 lentelė. Dažnių santykiai

Kalba	Savitosios raidės				Paskutinė en		Dažnių santykis	
	Sąrašas	f_{s1}	f_{sn}	$f_{svisų}$	Raidė	f_{pn}	f_{s1}/f_{pn}	f_{sn}/f_{pn}
cs	íáééřýčžšúúńóťď	2,7952	0,0170	12,5090	q	0,0120	233	1
da	åøæ	0,7926	0,7320	2,2866	q	0,0154	51	48
de	üäöß	0,5672	0,1567	1,5488	q	0,0344	16	5
en	(éôèçñæëäëïceô)	0	0	0	q	0,1099	0	0
es	óíáéñú(ü)	0,7998	0,1252	2,2796	w	0,0831	10	2
et	äõüö(šž)	1,0694	0,2330	2,9515	q	0,0117	91	20
fi	äö(ä)	3,3433	0,4363	3,7796	q	0,0136	246	32
fr	éàèèçóáîùceüï(ë)	2,4438	0,0151	3,6890	w	0,0881	28	0
ga	íáéóú	1,8851	0,8751	7,1006	q	0,0124	152	71
hr	čšžćđ	0,8837	0,1981	2,7016	q	0,0154	57	13
hu	áéőóóíúúú	3,5430	0,2253	11,3774	q	0,0122	290	18
is	ðíáóþöæúéý	3,6600	0,2497	12,0487	q	0,0109	336	23
it	èàòùìé(ó)	0,2381	0,0411	0,7084	x	0,0585	4	1
lt	ėųšžąjčūę	1,6643	0,1721	6,8919	q	0,0054	308	32
lv	āīēšņļūžķģċ(ōŗ)	4,0671	0,1169	10,2346	q	0,0064	635	18
mt	ħġżċ	2,0288	0,5814	4,1295	q	0,6844	3	1
nl	(ëïüöä)	0	0	0	q	0,0233	0	0
pl	łąęóźśńćż	1,7848	0,0660	6,0278	q	0,0121	148	5
pt	ãéçáíóêôüääô(ü)	0,6669	0,0412	3,2523	w	0,1235	5	0
ro	ăîșțâ	2,2527	0,4485	5,7356	q	0,0156	144	29
sk	áýíčěšžúľťóňďäľŕ	1,7484	0,0067	9,0332	q	0,0126	139	1
sl	čšž	1,1240	0,5336	2,5372	q	0,0132	85	40
sv	äöä	1,6661	1,2724	4,2160	q	0,0194	86	66
no	åøæ(é)	1,0216	0,1678	1,9692	q	0,0176	58	10
tr	üüşçğö	8,5689	0,8869	10,5980	q	0,0124	691	72
Vidurkis		2,4592	0,3039	5,1042		0,0569	153	6

lietuviškos raidės santykis (308) maždaug du kartus didesnis negu nagrinėtų kalbų vidurkis (153), rečiausios (32) – penkis kartus didesnis už vidurkį (6). Tai reiškia, kad lietuvių kalboje savitosios raidės atlieka svarbesnį vaidmenį negu daugelyje kitų kalbų.

Santykis f_{s1}/f_{pn} gali būti panaudotas nustatyti, kuriai raidėi teikti pirmenybę, kai tenka riboti tiesiogiai pasiekiamų raidžių skaičių. Pavyzdžiui, kai kurių planšetinių kompiuterių lietuvių kalbai skirtų klaviatūrų pagrindiniame plane išdėstytos tik pagrindinės lotynų raidės, tarp jų ir rečiausiai vartojama raidė *q*, o savitosios lietuviškos renka-

mos keliais klavišų paspaudimais – kelis kartus lėčiau. Taigi pirmenybė teikiama raidėi *q* prieš raidę *é*, vartojamai 308 kartus dažniau. Netgi rečiausia lietuvių kalbos abėcėlės raidė *ę* už raidę *q* vartojama 32 kartus dažniau.

Lietuvių kalbos abėcėlė ir kitos kalbos

Daugelis savitųjų lietuvių kalbos abėcėlės raidžių vartojamos ir kitose kalbose (5 lentelė). Pusjuodžiu šriftu pažymėti dažniai, ne mažesni kaip 0,1 %. Kai kurios savitosios lietuvių kalbos

raidės kitose kalbose netgi dažnesnės negu lietuvių kalboje: *q* – lenkų, *č* – čekų, vengrų, slovėnų ir slovakų, *ę* – lenkų, *ž* – čekų. Kitos taip

pat gausiai vartojamos, tik mažiau negu lietuvių kalboje ir yra tų kalbų abėcėlėse: *č* – latvių, *š* – čekų, vengrų, latvių, slovėnų ir slovakų, *ū* – latvių.

5 lentelė. Lietuvių kalbos savitosios raidės ir lietuvių kalbos abėcėlėje nesančios raidės *q*, *w* ir *x*

Kalba	Savitosios lietuvių kalbos raidės									Kitos raidės		
	ą	č	ę	ė	į	š	ų	ū	ž	q	w	x
cs	0,0003	0,8566	0,0004	0,0011	0,0000	0,6801	0,0002	0,0004	0,8448	0,0120	0,1007	0,0971
da	0,0001	0,0006	0,0001	0,0004	0,0000	0,0014	0,0000	0,0007	0,0006	0,0154	0,1566	0,0498
de	0,0000	0,0011	0,0002	0,0000	0,0000	0,0018	0,0000	0,0007	0,0006	0,0344	1,4132	0,0769
en	0,0001	0,0005	0,0001	0,0001	0,0000	0,0008	0,0000	0,0007	0,0003	0,1099	1,6729	0,2019
es	0,0000	0,0003	0,0001	0,0000	0,0000	0,0006	0,0000	0,0007	0,0002	0,5928	0,0831	0,2037
et	0,0002	0,0014	0,0002	0,0005	0,0000	0,0425	0,0001	0,0017	0,0184	0,0117	0,0947	0,0382
fi	0,0001	0,0011	0,0001	0,0002	0,0000	0,0161	0,0000	0,0011	0,0036	0,0136	0,1177	0,0330
fr	0,0001	0,0005	0,0001	0,0000	0,0000	0,0008	0,0000	0,0007	0,0003	0,8278	0,0881	0,4178
ga	0,0003	0,0010	0,0002	0,0001	0,0000	0,0011	0,0001	0,0006	0,0004	0,0124	0,0982	0,0214
hr	0,0001	0,8837	0,0001	0,0001	0,0000	0,7053	0,0000	0,0005	0,4692	0,0154	0,1017	0,0324
hu	0,0001	0,0024	0,0002	0,0000	0,0000	0,0033	0,0000	0,0002	0,0011	0,0122	0,0898	0,0604
it	0,0001	0,0021	0,0001	0,0001	0,0000	0,0018	0,0000	0,0008	0,0007	0,3384	0,0934	0,0585
lv	0,0002	0,1169	0,0002	0,0004	0,0000	0,9113	0,0002	0,3721	0,2326	0,0064	0,0378	0,0249
mt	0,0001	0,0011	0,0002	0,0001	0,0000	0,0013	0,0000	0,0003	0,0005	0,6844	1,4289	0,7059
nl	0,0006	0,0021	0,0009	0,0001	0,0000	0,0027	0,0000	0,0003	0,0011	0,0233	1,4806	0,0888
pl	0,8744	0,0022	0,8666	0,0002	0,0000	0,0029	0,0001	0,0007	0,0013	0,0121	4,9081	0,0639
pt	0,0002	0,0005	0,0002	0,0000	0,0000	0,0007	0,0000	0,0004	0,0002	0,5925	0,1235	0,2748
ro	0,0002	0,0010	0,0002	0,0001	0,0000	0,0014	0,0000	0,0005	0,0006	0,0156	0,1052	0,2060
sk	0,0005	0,8815	0,0007	0,0001	0,0000	0,7725	0,0000	0,0001	0,7566	0,0126	0,1021	0,1009
sl	0,0001	1,1240	0,0001	0,0000	0,0000	0,8795	0,0000	0,0002	0,5336	0,0132	0,0939	0,0271
sv	0,0001	0,0010	0,0001	0,0001	0,0000	0,0016	0,0000	0,0004	0,0006	0,0194	0,1416	0,1466
is	0,0000	0,0003	0,0001	0,0000	0,0000	0,0006	0,0000	0,0004	0,0002	0,0109	0,0866	0,0731
no	0,0001	0,0008	0,0002	0,0001	0,0000	0,0015	0,0000	0,0011	0,0004	0,0176	0,1454	0,0443
tr	0,0001	0,0007	0,0001	0,0001	0,0000	0,0010	0,0000	0,0006	0,0004	0,0124	0,0935	0,0296
Σ24	0,8781	3,8834	0,8714	0,0039	0,0000	4,0326	0,0007	0,3859	2,8683	3,4164	12,8573	3,0770
%24	0,0366	0,1618	0,0363	0,0002	0,0000	0,1680	0,0000	0,0161	0,1195	0,1424	0,5357	0,1282
lt	0,5418	0,4303	0,1721	1,6643	0,4848	1,1339	1,2602	0,4079	0,7965	0,0054	0,0368	0,0733
Σ25	1,4199	4,3137	1,0435	1,6682	0,4848	5,1655	1,2609	0,7938	3,6648	3,4218	12,8941	3,1503
%25	0,0568	0,1725	0,0417	0,0667	0,0019	0,0205	0,0504	0,0318	0,1466	0,1364	0,5158	0,1260

Pastabos:

Kitos raidės – pagrindinės lotynų raidės, nesančios lietuvių kalbos abėcėlėje,

Σ24 – dažnių procentų suma 24-iose kalbose (visose, išskyrus lietuvių),

%24 = Σ24/24 – dažnių vidurkio procentas 24-iose kalbose,

lt – dažnis lietuvių kalboje,

Σ25 = Σ24+lt – dažnių procentų suma visose 25 kalbose,

%25 = Σ25/25 – dažnių vidurkio procentas 25-iose kalbose.

Viena eile mažesni, bet ne mažesni kaip 0,01 proc. dažniai pažymėti pusjuodžiu kursyvu: š – estų ir suomių, ž – estų. Kai kurie šaltiniai (Everson 2004) mini jas esant šių dviejų kalbų abėcėlėse. Visų kitų raidžių (ė, į, ū) pavartojimą kitų kalbų Vikipedijose reikėtų laikyti epizodišku, pavyzdžiui, kai reikia paminėti lietuvių kalbos žodį, užrašyti lietuvišką asmenvardį arba vietovardį. Iš šių trijų dažniau pavartota raidė ė.

Pagrindinės lotynų abėcėlės raidės *q*, *w* ir *x*, kurių nėra lietuvių abėcėlėje, retai vartojamos ir daugelyje kitų kalbų. Mažesnis negu 0,1 proc. dažnis yra: *q* – 19, *w* – 11, *x* – 17 kalbų iš 25. Mažiausias dažnis yra šiose kalbose: *q* – lietuvių, *w* – lietuvių, *x* – airių. Didžiausias: *q* – prancūzų, *w* – lenkų, *x* – maltiečių kalbose.

Šių raidžių dažnius lietuviškoje Vikipedijoje palyginome su dažniais, apskaičiuotais iš Utkos žodyne pateiktų duomenų. Jie yra tokie: *q* – 0,0008 %, *w* – 0,0052 %, *x* (raidė) – 0,0020 %, *x* (romėniškas dešimtukas) – 0,0150 %. Iš viso 0,0230 %. Iš viso Vikipedijoje 0,1154 %, arba 3,8 karto, daugiau negu žodyne. Skirtumą galima paaiškinti tuo, kad šis žodynas sudarytas iš spausdintų tekstų, daugiausia literatūrinių ir publicistinių. Juose asmenvardžių, vietovardžių ir kitų tikrinių daiktavardžių nėra daug. Kompiuteriniuose tekstuose, ypač enciklopediniuose, jų gerokai daugiau. Vienas iš mūsų tikslų ir buvo paanalizuoti kraštinį atvejį, kuriame svetimų kalbų rašmenų vartojimas didžiausias.

Vikipedijoje pagal ženklą *x* dažnį lietuviškai lenkia 11 kitų kalbų. A. Utkos žodyne išskirtos ženklų *x* funkcijos: raidė ir romėniškas dešimtukas. Ženklas *x* skaitmenį žymi dažniau negu raidė. Čia dar reikia pridėti Vikipedijoje neišskirtus atvejus, kai juo žymimas nežinomas dydis matematikoje, kai vartojamas kaip daugybos ženklų pakaitalas. Šie atvejai lemia, kad lietuviai raidę *x* vartoja dažniau negu *q* ir *w*.

Apibendrinimai ir išvados

Raidžių dažniai, paimti iš visoms kalboms giminingo šaltinio (Vikipedijos) sudarė sąlygas

adekvačiai palyginti įvairių kalbų raidžių, vartojamų internetinėje erdvėje, dažnius. Panaudotas dažnių santykių ir pokyčių metodas padėjo išryškinti panašumus ir skirtumus tarp kalbų.

Lyginant lietuvių kalbą su kitomis Europos kalbomis, vartojančiomis lotyniškus rašmenis, paaiškėjo, kad lietuvių kalbos abėcėlės raidžių dažniai jų eilėje išsidėstę tolygiau (tarp jų nėra šuolių). Ryškesnė takoskyra tarp lietuvių kalbos abėcėlės raidžių ir užsienietiškų (tarp jų yra šuolių), ko pasigendama kai kuriose kitose kalbose. Šios savybės suteikia stabilumo lietuvių kalbos abėcėlei.

Šešios savitosios lietuvių kalbos raidės vartojamos ir kitose kalbose. Keturių raidžių (*q*, *č*, *ę*, *ž*) dažnis kai kuriose kalbose netgi didesnis negu lietuvių kalboje.

Dažniausios savitosios lietuviškos raidės *ė* dažnio santykis su rečiausios pagrindinės lotynų abėcėlės raide *q* yra 308, maždaug du kartus didesnis negu analogiškas santykių vidurkis kitose nagrinėtose kalbose, rečiausios raidės *ę* santykis yra 32, penkis kartus didesnis už analogišką kitų kalbų vidurkį. Tai reiškia, kad savitųjų raidžių vaidmuo lietuvių kalboje svarbesnis negu daugelyje kitų kalbų.

Tikimės, kad pateikti duomenys ir analizės rezultatai galės būti naudingi tekstų įvedimo priemonių ir kitokių įrenginių, susijusių su raidžių aibėmis, projektuotojams.

Literatūra

Bi, X.; Smith, B. A.; Zhai, S. 2012. Multilingual touchscreen keyboard design and optimization, *Human-Computer Interaction* 27(4): 352–382.

ETSI ES 202 130 v.2.1.2: 2007. *Character repertoires, orderings and assignments to the 12-key telephone keypad (for European languages and other languages used in Europe)*. European Telecommunications Standards Institute, 2007.

Everson, M. 2004. *The alphabets of Europe. Everytype* [interaktyvus], [žiūrėta 2015 m. balandžio 20 d.]. Prieiga per internetą: <http://www.everytype.com/alphabets/>

Grigas, G. 2014. Klaviatūros: nuo rašomųjų mašinėlių iki planšetinių kompiuterių, *Santalka*:

Filologija, Edukologija 22(1): 25–38.

<http://dx.doi.org/10.3846/cpe.2014.03>

Grumadienė, L.; Žilinskienė, V. 1998. *Dažninis dabartinės rašomosios lietuvių kalbos žodynas*. LKI, MII. 476 p.

ISO 12199:2000. *Alphabetical ordering of multi-lingual terminological and lexicographical data represented in the Latin alphabet*. International Organization for Standardization, 2000.

Practical Cryptography. 2015. *Letter frequencies for various languages* [interaktyvus], [žiūrėta 2015 m. balandžio 20 d.]. Prieiga per internetą: <http://practicalcryptography.com/cryptanalysis/letter-frequencies-various-languages/>

Utka, A. 2009. *Dažninis rašytinė lietuvių kalbos žodynas*. VDU. 549 p.

Vrandečić, D. 2012. *Letter frequency* [interaktyvus], [žiūrėta 2015 m. balandžio 20 d.]. Prieiga per internetą: <http://simia.net/letters/unigrams.zip>

Vrandečić, D.; Sorg, P.; Studer, R. 2011. Language resources extracted from Wikipedia, in *Proceedings of the Sixth International Conference on Knowledge Capture*, 26–29 June 2011, Baniff, Alberta, Canada, 153–160. <http://dx.doi.org/10.1145/1999676.1999703>

Wikipedia [interaktyvus]. 2015 [žiūrėta 2015 m. balandžio 20 d.]. Prieiga per internetą: <https://www.wikipedia.org/>

Žilinskienė, V. 1990. *Lietuvių kalbos dažninis žodynas*. Vilnius: „Mokslas“. 177 p.

LETTER FREQUENCY ANALYSIS OF LITHUANIAN AND OTHER LANGUAGES USING THE LATIN ALPHABET

Gintautas GRIGAS¹, Anita JUŠKEVIČIENĖ²

Institute of Mathematics and Informatics, Vilnius University, Akademijos g. 4, LT-08663 Vilnius, Lithuania

E-mails: ¹gintautas.grigas@mii.vu.lt; ²anita.juskeviciene@mii.vu.lt

It is important to evaluate specificities of alphabets, particularly the letter frequencies while designing keyboards, analyzing texts, designing games based on alphabets, and doing some text mining. In order to adequately compare letter frequencies of Lithuanian language to other languages in the Internet space, Wikipedia source was selected which content is common to different languages. The method of letter frequency jumps is used. The main attention is paid to the analysis of letter frequencies at the boundary between native letters and foreign letters used in Lithuanian and other languages.

Keywords: alphabet, diacritic, keyboard, Latin script, letter, letter frequency, Lithuanian language, mobile phone, text input.