EESTI NSV TEADUSTE AKADEEMIA TOIMETISED, 26. KÖIDE
FÜÜSIKA * MATEMAATIKA. 1977, NR. 4

ИЗВЕСТИЯ АКАДЕМИИ НАУК ЭСТОНСКОЙ ССР. ТОМ 26
ФИЗИКА * МАТЕМАТИКА. 1977, № 4

https://doi.org/10.3176/phys.math.1977.4.10

УДК 537.782:311.21:801.41

Э. КЮННАП

ЧАСТОТА БУКВ В ПЕЧАТНОМ ТЕКСТЕ НА ЭСТОНСКОМ ЯЗЫКЕ

Введение

При синтезе речи по печатному тексту буквы преобразуются в акустические волны речевых сигналов.

В синтезе слитной речи участвуют сочетания различных фонем, их составляющие характеризуются своими определенными значениями параметров (частотой, амплитудой и шириной формант, частотой основного тона, спектром и амплитудой шума и т. д.). При произнесении гласных и сонорных согласных различаются три этапа — появление, квазистационарная часть и исчезновение. Слитность произношения вызывает взаимное влияние смежных фонем друг на друга. Поэтому при разработке законов управления параметрами синтезаторов речевых сигналов необходимо располагать данными о существующих сочетаниях фонем, выраженных в печатном тексте буквами.

Статистическим анализом текстов на эстонском языке занимались несколько авторов. В [1, 2] были определены частота появления в словах отдельных букв и комбинаций слогов из двух и трех букв. Однако в разговоре отдельные слова сливаются и поэтому воспринимаются слухом как последовательность звуковых единств, т. е. синтагмами, паузы между которыми на письме выделяются, как правило, знаками препинания. Поэтому при вводе печатного текста для обработки в ЭВМ знаки препинания (запятые, точки, двоеточия, точки с запятой, вопросительные и восклицательные знаки, кавычки, скобки, тильды и тире, а также обрывы слитности при чтении) должны кодироваться как паузы.

Материалом для анализа служили произвольно выбранные отрезки текстов из книг, журналов и газет в соотношении

научно-техническая литература	47,1 %
художественная литература	22,1 %
общественно-публицистическая литература	30,8 %

Изучено было 102 307 букв. Исследование по синтагмам давало больше комбинаций букв по сравнению с исследованием по словам, поскольку в речевом потоке слова сливаются и произносятся с паузами только после нескольких слов.

В алфавите эстонского языка 23 буквы. В написаниях имен и в заимствованиях встречаются, кроме того, буквы *c*, *f*, *q*, *š*, *z*, *ž*, *w*, *x* и *y*. Из них мы учитывали только *f*, как наиболее употребительную.

Таблица 1

Буква	Частота	Относительная частота, %	Доверительные границы, %	
a	12148	11,87	11,61	12,13
e	12086	11,81	11,55	12,07
i	10278	10,05	9,81	10,29
s	9389	9,18	8,95	9,42
t	8012	7,83	7,62	8,05
l	5966	5,83	5,64	6,02
u	5855	5,72	5,54	5,91
m	4599	4,50	4,67	4,34
k	4518	4,42	4,26	4,59
n	4513	4,41	4,25	4,58
d	3868	3,78	3,63	3,94
o	3852	3,76	3,61	3,92
r	3414	3,34	3,20	3,49
v	2434	2,38	2,26	2,51
g	1863	1,82	1,72	1,93
j	1738	1,70	1,60	1,81
h	1623	1,59	1,49	1,69
p	1611	1,57	1,47	1,67
õ	1244	1,22	1,13	1,31
ä	1240	1,21	1,12	1,30
b	818	0,80	0,73	0,88
ü	781	0,76	0,69	0,83
ö	374	0,37	0,32	0,42
f	83	0,08	0,06	0,11
Всего	102307	100,00		

Результаты исследования

Данные изучения частотности отдельных букв в печатном тексте приведены в табл. 1. Как видно, относительная частота встречаемости всех девяти гласных составляет 46,8%, причем наиболее употребительны *a*, *e* и *i* — 33,7% всех букв и 71,4% всех гласных. Среди согласных наиболее распространена буква *s*. Взрывная *t* встречается почти в пять раз чаще, чем *p*, и почти в два раза чаще, чем *k*. Общий процент повторяемости *t*, *p* и *k* в два раза выше такового *b*, *d* и *g*. Гласная *i* встречается в шесть раз чаще полугласного *j*, а если считать, что они описывают одну и ту же фонему, то вероятность их появления вместе взятых приближается к вероятности появления *a* и *e*.

В табл. 2 представлены данные о частотности диад, т. е. двухбуквенных сочетаний. По горизонтали выписаны первые и по вертикали вторые буквы диад. Всего выявлено 94 010 сочетаний. В отдельных словах эстонского языка используются 25 комбинаций гласных: *ae, ai, ao, au, ea, ei, eo, io, iu, oa, oe, oi, ou, ui, õa, õe, õi, õo, õu, äe, äi, äo, äu, öe* и *öi*. В слитной же речи их количество увеличивается до 60, т. е. по сравнению с отдельными словами становится больше, чем в два раза. Теоретически из девяти гласных можно образовать 72 варианта. Значит, в синтагмах встречается 83,2% возможных сочетаний, причем гласные *a*, *e*, *i* и *u* имеют все комбинации. Из 24 букв (23 эстонских плюс *f*) теоретически можно сформировать 552 двухбуквенные комбинации. В синтагмах обнаружено 504 разные диады, т. е. 90,9% теоретически возможных. Из них чаще десяти раз появилось 397 диад и чаще пяти раз 442 диады. Последние составляют 78,3% их общего количества.

Исследованы частоты встречаемости триад, т. е. комбинации из трех

Таблица 2

	a	*b*	*d*	*e*	*f*	*g*	*h*	*i*	*j*	*k*	*l*	*m*	*n*	*o*	*p*	*r*	*s*	*t*	*u*	*v*	*õ*	*ä*	*ö*	*ü*	Σ
a	902*	209	726	426	8	309	332	496	477	659	898	765	708	144	229	591	1285	1413	288	377	32	14	6	53	10445
b	140	5	9	94		5	7	105	37	46	58	26	21	21	24	45	47	39	36	14	2	4	3	13	796
d	818	2		675	3	15	27	257	66	123	140	161	81	89	76	69	226	140	328	78	16	5	7	39	3441
e	581	96	383	922	17	334	220	756	265	943	1251	637	681	150	289	560	1566	954	125	385	19	10	4	52	10278
f	2			7	4		5	14		1	3		2	6		77	1	3	10	2		3	1	14	81
g	441			187			7	347	1	23	45	24	92	20	17	144	63	61	301	9	1	1	1	2	1787
h	184	1	14	342			17	134	37	73	56	13	66	64	3	15	32	295	142	49	7	58	1	9	1595
i	586	79	772	417	5	226	97	629	80	723	631	691	954	292	170	262	2011	793	120	509	25	9	7	42	9501
j	962			145				4						26					428		32	137	3	1	1738
k	695	4	91	359	1	19	14	295	14	134	80	37	68	491	28	53	748	320	656	22	141	88	7	89	4320
l	684	19	175	1387	1	100	44	923	196	150	489	152	101	222	56	23	193	489	257	72	70	59	8	60	5441
m	1130	48	48	416	2	9	3	1101	10	55	67	180	86	94	109	5	135	67	314	16	154	125	9	17	4020
n	494	15	618	840	9	221	19	799	16	92	80	66	102	101	28	28	145	236	308	48	60	136	3	28	4390
o	84	58	67	184	4	176	101	127	22	54	496	328	737	463	102	267	210	135	36	16	1	4	1	2	33212
p	216	4	6	230			7	203	2	27	87	3	2	139	34	231	56	22	101	3	111	78	11	36	1575
r	484	14	78	396	1	157	9	768	160	111	46	84	40	318	7	56	123	109	174	187	19	17	6	6	3314
s	993	21	18	2624	1	14	23	1430	198	314	195	143	149	374	165	57	349	1751	519	171	95	23	10	172	9460
t	1359	36	86	1708	4	45	53	747	56	191	248	164	102	349	82	279	620	283	933	129	77	169	95	52	7584
u	141	136	748	301	1	147	217	278	77	187	485	173	257	38	75	275	1117	589	418	103	6	6	2	9	5388
v	956	2	25	162		2	8	254	12	20	48	19	28	64	8	25	37	36	172	6	332	201	1	5	2417
õ	19	8	12	26		13	120	361	6	13	66	15	88	3	64	74	89	113	100	10	42				1200
ä	41	30	16	72		40	122	166		28	146	8	17	4	8	302	85	42	15	1		95			1143
ö	12	4	3	32		10	2	5	2	8	24	3	2	4	8	15	21	30	5	3			181		193
ü	6	7	40	17		7	114	26	4	59	171	69	6	2	14	19	118	20	3	3				81	691
Σ	11028	793	3921	11047	57	1849	1551	9596	1738	3900	5321	3581	4288	3015	1562	3346	8948	7657	5371	2207	1200	1147	186	701	94010

* Подчеркнутые цифры в общей сумме не учтены.

Таблица 3

ü	ö	ä	õ	v	u	t	s	r	p	o	n	m	l	k	j	i	h	g	f	e	d	b	a	
4	1	4	2	443	48	588	241	214	115	10	231	461	260	309	421	126	94	190	3	263	299	51		*a*
3	2	11	4	2	57	3	2	3		29	1	20	2			58	1			51	5		131	*b*
23	8	7	9	6	293	24	19	41	6	35	278	16	83	11		357	6			229		1	310	*d*
3	14	14	14	74	68	641	901	140	87	36	317	278	555	150	40	140	182	96	3		294	28	201	*e*
					1	4	2	1		3	9	3	2	1		6				15	1		5	*f*
5	2	22	6	2	78	10	2		3	55	114	1	54			134				204	13	3	153	*g*
57	1	70	57	6	88	14	30	68	1	53	6	8	24	4		58		2		125	8	4	173	*h*
3	2	85	149	86	147	267	459	269	79	35	277	407	458	156			82	170	11	290	152	54	170	*i*
	4	2	9	6	36	78	49	61		18	2	2	90	5		44	13	1		99	21	28	173	*j*
32	10	16	5	2	95	119	184	69	7	25	46	19	39			371	24	12	3	419	82	26	297	*k*
68	4	74	31	19	249	76	75	12	17	263	30	16		31		259	9	18		576	57	19	519	*l*
16	3	3	14	8	94	82	83	45	1	149	40		60	10		342	8	25		240	81	16	291	*m*
13	1	19	41	6	136	46	64	21	8	245		41	46	21		446	15	85		284	40	15	353	*n*
		3		18	15	132	114	78	68		22	23	48	252		89	29	6	14	84	37	9	88	*o*
5		5	21	1	32	31	72	6		32	9	28	36	3		61		2		119	33	11	105	*p*
7	6	153	33	6	97	135	32		64	143	16	5	8	35		111	2	14		253	26	21	287	*r*
61	8	27	13	41	628	274		51	31	94	56	83	100	277		770	28	30	2	572	84	19	604	*s*
11	23	19	43	28	223		892	62	12	72	69	33	205	104		383	157	25		408	85	22	536	*t*
	3	2	37	35		531	164	64	41	18	243	119	123	368	167	48	42	155	9	46	166	18	143	*u*
1	5	2	9		61	77	70	61	1	24	20	7	42	12		142	19	5		192	51	15	220	*v*
				143	2	46	41	3	46		25	52	57	78	16	13	4	2	1	15	10		15	*õ*
				93	1	77	11	3	52		71	29	31	35	61	5	13			6	4	3	6	*ä*
					3	3	1	1				1				2				1	1		5	*ö*
					2	24	68	2	13		1	6	15	47		18	2	1	5	12	12	6	27	*ü*

Таблица 4

	a	*b*	*d*	*e*	*f*	*g*	*h*	*i*	*j*	*k*	*l*	*m*	*n*	*o*	*p*	*r*	*s*	*t*	*u*	*v*	*õ*	*ä*	*ö*	*ü*
a		121	323	179	4	193	143	163	114	271	513	248	305	91	103	253	525	481	109	160	13	8	7	26
b	53		2	25		4	4	65	28	26	16	16	20	10	11	27	16	24	20	12		1	2	5
d	406	2		323	1	17	7	156	57	82	62	76	41	42	36	28	77	91	152	62	10	6	6	12
e	253	58	259		14	191	127	286	84	389	656	200	217	85	105	178	576	364	47	170	14	7	4	12
f	3			3				12		2				11			2	2	4			2	1	11
g	271	1	3	109			2	184	5	12	22	25	85	4	2	13	35	26	179	5	4		1	1
h	120	1	7	189		1		77	19	30	12	7	12	22		1	35	171	38	23	5	26		6
i	142	61	394	133	6	104	60		47	432	247	263	390	117	60	76	739	349	51	92	13	5	7	14
j	472			131										11					135		15	89		
k	283		9	150	1		10	152	6		30	10	21	233	4	27	415	161	294	9	48	31	2	42
l	269	2	89	575	3	54	15	405	97	40		58	37	91	30	7	121	273	130	42	38	26	1	19
m	506	20	14	351	1	1	7	272	8	18	15		35	30	28	5	71	38	92	6	40	37	5	6
n	202	1	291	418	9	130	5	345	5	42	27	42		25	11	15	64	74	135	20	18	57		10
o	18	24	19	35	3	35	14	35	19	13	189	131	226		28	162	70	65	8	5				
p	104		2	71		1		79		8	17		7	78		38	30	10	49	1	34	53	5	24
r	241	3	40	142	1	60	3	343	73	72	10	56	22	81	6		78	169	78	61	4	7	4	2
s	274	4	15	966	2	2	26	458	49	189	73	95	64	107	72	34		1054	186	79	35	12	9	42
t	634	2	35	798	5	10	13	342	50	126	99	83	47	153	30	128	234		460	60	35	55	32	29
u	57	67	462	57		77	87	164	39	95	254	75	128	14	41	43	579	231		62	3	1	5	2
v	452		3	100		3	5	220	9	5	6	8	6	24	1	7	38	36	31		90	89		
õ	5	2	11	14		6	62	163	11	8	35	13	56		29	39	13	56	36	9				
ä	5	10	9	32		20	64	91		17	69	2	16	2	7	112	20	23	1	1				
ö			2	9			1	2								2			1					
ü		1	2	1		4	35	3		25	61	13	14		4	6	47	8		1				

букв (табл. 3 и 4). Теоретически из 24 букв можно образовать 12 696 триад, в которых рядом друг с другом одна и та же буква не повторяется. В нашем материале зафиксировано только 4065 комбинаций, т. е. 32,0% теоретически возможных. Более 100 раз появились 29 триад, от 50 до 99 раз — 104, от 10 до 49 раз — 936, от 5 до 9 раз — 740 и от 1 до 4 раз — 2256, причем 1023 триады встретились только единожды. Десять наиболее употребительных триад и частота их появления: *ste* — 266, *ise* — 240, *rel* — 230, *use* — 224, *est* — 216, *ast* — 210, *mis* — 197, *ist* — 189, *ust* — 185 и *sta* — 179 раз.

В табл. 3 приведены левые соседи и в табл. 4 правые соседи центральных букв в триадах. Последние указаны на вертикальной линии таблиц.

Статистико-вероятностный анализ

Частота употребительности отдельных букв z_i в процентном выражении ко всем зафиксированным буквам вычисляется по формуле

$$z_i = x_i/n \cdot 100,$$

где x — частотность данной буквы i среди всех букв n, вероятность появления p которой неизвестна. Располагая большим количеством n, для приближенного значения вероятности можем записать $p_i \approx z_i$. Однако это равенство обязательно содержит некоторую ошибку. Поскольку z_i является лишь одним значением случайной величины, желательно знать доверительные границы для искомой вероятности.

Из теории вероятности известно, что вероятность отклонения случайной величины от математического ожидания более чем на утроенное квадратичное отклонение крайне мала. В нашем случае на основе проверки большого материала можно предположить, что разность между появлением и вероятностью появления букв не превышает квадратичное отклонение более чем на $1-\alpha$:

$$|z-p| \leqslant A[p(1-p)/n]^{1/2}. \qquad (1)$$

Коэффициент A, выражающий отклонение случайной величины от нормального распределения в зависимости от доверительного уровня двухсторонней границы, был определен несколькими авторами. Например, в [3] при $\alpha = 0{,}002$ получено $A = 3{,}09$, а при $\alpha = 0{,}01$ соответственно $A = 2{,}578$.

Чтобы найти доверительные границы для искомой вероятности, уравнение (1) было решено относительно p:

$$z - p = A[p(1-p)/n]^{1/2}, \qquad (2)$$

$$p_{1,2} = \{nz + A/2 \pm A[nz(1-z) + 1/4A^2]^{1/2}\}/(n + A^2). \qquad (3)$$

Результаты расчета приведены в табл. 1.

Например, буква *a* встретилась 12 148 раз на 102 307 букв, т. е. частота z ее составляет $0{,}1187 = 11{,}87\%$. Значит, доверительные границы для 1%-ного уровня значимости ($\alpha = 0{,}01$) будут равны

$$p_1 = 0{,}1161 = 11{,}61\%, \quad p_2 = 0{,}1213 = 12{,}13\%,$$

$$\Delta p = (12{,}13 - 11{,}61)/2 = 0{,}26\%.$$

Следовательно, в печатном тексте на эстонском языке в 99 случаях из 100 на каждые 100 000 букв *a* может появиться в пределах $11{,}87 \pm 0{,}26\%$, т. е. от 11 610 до 12 130 раз. Частотность диад значи-

тельно меньше. Из всех букв 94 010 диад наиболее употребительна диада *se* — 2624 раза. Доверительные границы для 1%-ного уровня значимости ($\alpha = 0{,}01$) равны $p_1 = 2{,}93\%$ и $p_2 = 2{,}66\%$, т. е. диада *se* может встретиться в тексте из 100 000 букв от 2496 до 2757 раз.

Заключение

Проведен статистический анализ частотности появления отдельных букв, двух- и трехбуквенных сочетаний в произвольно выбранных фрагментах печатного текста на эстонском языке. Установлено, что наиболее часто встречаются буква *a*, диада *se* и триада *ste*. Слово состоит в среднем из 7 букв и синтагма из 35 букв. Статистика показала, что в слитной речи появляются сочетания букв, отсутствующие в отдельных словах. Результаты исследования будут использованы при разработке законов управления параметрами синтезатора речевых сигналов по печатному тексту.

ЛИТЕРАТУРА

1. Kaasik, Ü., Laugaste, E., Keel ja kirjandus, **12**, 600 (1969).
2. Kaasik, Ü., Laugaste, E., Ääremaa, K., Keel ja kirjandus, **18**, 21 (1975).
3. Fisher, R. A., Statistische Methoden für die Wissenschaft, Edinburgh, 1956.

Институт кибернетики
Академии наук Эстонской ССР

Поступила в редакцию
10/XI 1976

E. KÜNNAP

TÄHTEDE ESINEMISSAGEDUS EESTIKEELSES TEKSTIS

Tuuakse andmeid tähtede ning kahe- ja kolmetäheliste kombinatsioonide esinemissageduse kohta eestikeelsetes tekstides. Uurimise tulemusi kasutatakse trükiteksti-kõnesüntesaatori juhtimissüsteemide väljatöötamisel.

E. KÜNNAP

FREQUENCY OF OCCURRENCE OF LETTERS IN WRITTEN ESTONIAN

Statistical information was obtained about written Estonian. The frequency of occurrence of letters, dyads and triads is presented. The investigation results will be used for working out rules for control of the synthesizers when synthesizing speech from a written text.