МРНТИ 16.21.21

А.Қ.Жұбанов[1], Жанабекова А.Ә.[2], Тоқмырзаев Д.О.[3]

[1]А.Байтұрсынұлы атындағы Тіл білімі институтының бас ғылыми қызметкері, филология ғылымдарының докторы, профессор. Алматы қаласы, Қазақстан

[2]А.Байтұрсынұлы атындағы Тіл білімі институтының бас ғылыми қызметкері, филология ғылымдарының докторы. Алматы қаласы, Қазақстан

[3]А.Байтұрсынұлы атындағы Тіл білімі институтының ғылыми қызметкері,бағдарламашы. Алматы қаласы, Қазақстан

ҚАЗАҚ ТІЛІ ӘРІПТЕРІ МЕН ӘРІП ТІРКЕСТЕРІНІҢ ЖИІЛІКТЕРІН АНЫҚТАУ ЛАТЫН ӘЛІПБИІНЕ КӨШУДЕ ӨЗ СЕПТІГІН ТИГІЗЕДІ

Аннотация: Әріптер мен әріп тіркестерінің жиіліктерін анықтау қазақ әліпбиін латын қарпіне көшірудегі бірнеше теориялық-практикалық бағыттағы мәселелерді шешуге септігін тигізеді.

Біріншіден, әріп, әріп тіркестерінің жиілігі мәтінді қолмен теруде жылдамдықты арттыру үшін латын қаріпті қазақ пернетақтасына әріптерді тиімді орналастыруда қажет болса, *екіншіден,* қазақ жазуын реформалаудағы орфографиялық емле ережелерін шешу барысында жасалатын зерттеулер мен лингвистикалық әзірленімдерге фактологиялық сандық деректер үшін, *үшіншіден,* кирилл қаріпті мәтіндерді латын қарпіне автоматты аударма жасаудағы программаға қажетті емле ережелерін жасау барысында жиілік бойынша реттеу, сонымен қатар нысанға алынған әріп тіркестері туралы сандық мәліметтер беру және әріп тіркестерін мысалмен дәйектейтін сөздерді қазақ тілінің мәтіндер корпусынан іздестіріп, көптеген мысал сөздерді табу сияқты мәселелерді шешуде аса қажет.

Мақалада қазақ тілінің жазба мәтінін әріптік (графемалық) деңгейде және қазақ тілі сөзформаларының құрылымына, яғни әріптерге, әріп тіркестеріне, буындарға статистикалық әдіспен зерттеу нәтижелері баяндалады. Сөз құрылымындағы әріп пен буынды табиғи тілдің жазба мүмкіндігі арқылы бейнелейтін ең бір қарапайым түрдегі көрнекі тілдік бірліктері ретінде санау негізгі ұстаным болып табылады. Графемдік деңгейдегі статистикалық зерттеудің негізгі мақсаты – компьютер пернетақтасында қазақ әріптерін ұтымды орналастыру.

Тірек сөздер: Жиілік сөздік, сөз, сөзформа, буын, әріп, әріп тіркесі, екіәріптік тіркес, жиілік, абсолютті жиілік, қатынастық жиілік, жоғары жиілікті, төмен жи-ілікті әріптер және әріп тіркестері, клавиатура, қазақ әліпбиінің клавиатурада ұтым-ды орналасуы

А.К. Жубанов[1], А.А. Жанабекова[2], Д. Токмырзаев[3]

[1]главный научный сотрудник Института языкознания имени А. Байтурсынулы, д-р.филол.наук, профессор, Алматы, Казахстан

[2] главный научный сотрудник Института языкознания имени А. Байтурсынулы, д-р.филол.наук, Алматы, Казахстан

[3]научный сотрудник Института языкознания имени А. Байтурсынулы, программист, Алматы, Казахстан

ОПРЕДЕЛЕНИЕ ЧАСТОТ КАЗАХСКИХ БУКВ И БУКВОСОЧЕТАНИЙ ДАЕТ ВОЗМОЖНОСТЬ К ПЕРЕХОДУ НА ЛАТИНСКУЮ ГРАФИКУ

Аннотация. Определение частоты букв и их сочетаний поможет решить несколько теоретических и практических задач при переводе казахского алфавита на латинский алфавит.

Во-первых, частота букв и словосочетаний крайне необходима для эффективного расположения букв на казахской клавиатуре для увеличения скорости ручного набора, во-вторых, для фактологических числовых данных исследований и лингвистических разработок, которые будут проводиться в ходе решения орфографических правил в реформировании казахской письменности, в-третьих, в-третьих, упорядочение по частоте в ходе разработки правилы правописания, необходимой для программы автоматического перевода текстов на кириллице – на латиницу, также в решении таких проблем, как предоставление количественных данных о целевых сочетаниях букв и поиск слов, доказывающих примером сочетаний букв, из корпуса текстов казахского языка, нахождение многих слов для примера. Регулировка частоты важна при создании правил правописания для программы автоматического перевода, а также при предоставлении числовой информации о целевых сочетаниях букв и поиске множества примеров слов путем поиска слов, следующих за комбинациями букв с примерами.

В статье рассматриваются результаты статистических исследований букв и буквосочетаний, а также слогов казахского языка, с учетом их позиционных положений в словоформах. Данное статистическое исследование проведено с целью поиска удобного расположения букв казахского алфавита на клавиатуре компьютера.

Ключевые слова: Частотный словарь, слова, словоформы, буквы, сочетания букв, графемы, сочетания графем, абсолютная частота, относительная частота, высокочастотные и низкочастоные буквы и их сочетания, процент покрываемости текста, клавиатура, расположение букв на клавиатуре, позиционное расположение букв, статистика букв и буквосочетаний, статистика слогов.

A.K. Zhubanov[1], A.A. Zhanabekova[2], D. Tokmyrzaev[3]

[1]Chief Researcher of the A. Baitursynuly Institute of Linguistics, Doctor of Philology, Professor, Almaty, Kazakhstan

[2]Chief Researcher of the A. Baitursynuly Institute of Linguistics, Doctor of Philology, Almaty, Kazakhstan

[3]Researcher at the A. Baitursynuly Institute of Linguistics, programmer, Almaty, Kazakhstan

DETERMINATION THE FREQUENCIES OF KAZAKH LETTERS AND LETTER COMBINATIONS MAKES IT POSSIBLE TO SWITCH TO THE LATIN SCRIPT

Annotation. Determining the frequency of letters and their combinations will help to

solve several theoretical and practical problems when translating the Kazakh alphabet into the Latin alphabet.

Firstly, the frequency of letters and phrases is necessary for the effective placement of letters on the Kazakh keyboard to increase the speed of manual typing, and secondly, for actual numerical data on research and linguistic developments in solving spelling rules in the reform of Kazakh writing, and thirdly, for Cyrillic texts in Latin. Adjusting the frequency is important when creating spelling rules for an automatic translation program, as well as when providing numerical information about target letter combinations and searching for many example words by looking for words that follow the example letter combinations.

The article describes the results of a statistical study of the written text of the Kazakh language at the alphabetical (grapheme) level and the structure of word forms of the Kazakh language: letters, letter combinations, syllables. The basic principle is to consider the letters and syllables in the structure of the word as the simplest visual linguistic units, reflecting the writing abilities of a natural language. The main goal of statistical studies at the graphochemical level is the rational placement of Kazakh letters on the computer keyboard.

The article deal with the results of statistical studies of letters and letter combinations, as well as the syllables of the Kazakh language, taking into account their positional positions in word forms. This statistical study was conducted to find the convenient location of the letters of the Kazakh alphabet on a computer keyboard.

Keywords: Frequency dictionary, words, word forms, letters, combinations of letters, graphemes, combinations of graphemes, absolute frequency, relative frequency, high-frequency and low-frequency letters and their combinations, percentage of text coverage, keyboard, arrangement of letters on the keyboard, positional arrangement of letters, statistics of letters and letter combinations, statistics of syllables

2016 жылы А.Байтұрсынұлы атындағы Тіл білімі институты Қазақстан Республикасы Білім және ғылым министрі Е.Сағадиевтің тапсырмасы бойынша «Қазақ тілінің ұлттық корпусын әзірлеу және жасау» зерттеу жобасы аясында «Жалпы білім берудегі қазақ тілінің жиілік сөздігін» шығарды. Бұл сөздікте 7 миллионнан астам сөзқолданыстан тұратын қазақ тілінің 5 стилінен мәтіндер қамтылып, 61 түрленім нұсқасы дайындалды. Сөзтізбе саны – 36262 сөз.

Аталған 7 миллионнан астам сөзқолданыстан тұратын қазақ тілінің мәтінінен «Қазақ тілінің сөзформалар жиілік сөздігін» де түзілетінін ескерсек, бұл сөздікті де қазақ графемаларының мәтіндерде кездесу мүмкіндіктерін толық қамтыған жиілік сөздік деп те санауға болады.

Әріптер мен әріп тіркестерінің жиіліктерін анықтау қазақ әліпбиін латын қарпіне көшірудегі бірнеше теориялық-практикалық бағыттағы мәселелерді шешуге септігін тигізеді.

Біріншіден, әріп, әріп тіркестерінің жиілігі мәтінді қолмен теруде жылдамдықты арттыру үшін латын қаріпті қазақ пернетақтасына әріптерді тиімді орналастыруда қажет болса, *екіншіден,* қазақ жазуын реформалаудағы орфографиялық емле ережелерін шешу барысында жасалатын зерттеулер мен лингвистикалық әзірленімдерге фактологиялық сандық деректер үшін, *үшіншіден,* кирилл қаріпті мәтіндерді латын қарпіне автоматты аударма жасаудағы программаға қажетті емле ережелерін жасау барысында жиілік бойынша реттеу, сонымен қатар нысанға алынған әріп тіркестері туралы сандық мәліметтер беру және әріп тіркестерін мысалмен дәйектейтін сөздерді қазақ тілінің мәтіндер корпусынан іздестіріп, көптеген мысал сөздерді табу сияқты мәселелерді шешуде аса қажет.

Төмендегі *1.1-кестеде* Қазақ тіліндегі әріптердің жиілікті-әліпбилі жиілік сөздігі, яғни әріптердің қайталану жиіліктерінің ең көп кездесуінен бастап, біртіндеп кему тәртібімен орналасқан әріптер тізімі көрініс тапты:

1.1-кестедегі мәліметтерді 3 топқпа бөліп қарастыратын болсақ, 1-топқа ең жиі қолданатын 19 әріптерді: *а, е, ы, н, і, т, р, л, д, с, м, қ, о, к, ғ, б, й, у, з* жатқызуға болады. Бұл әріптердің ішінде ең жиі қолданатыныс тауып, 1-орынға ие *«а»* әріпі *8.630862* рет қолданып, мәтіннің *12,796* пайызын қамтиды екен, ал жиілік жағынан 2-орынға ие болатын «*е*» әріпі *5.660025* рет қолданып, мәтіннің *8.391* пайызын қамтиды. Енді осы *«а», «е»* әріптерінің бірігіп қолданысы барлық мәтіннің *21,187* пайызын қамтып жатқанына *1.1-кесте* бойынша көз жеткізуге болады. Сол сияқты, 1-топқа жататын ең жиі қолданатын 19 әріп *(а, е, ы, н, і, т, р, л, д, с, м, қ, о, к, ғ, б, й, у, з)* барлық мәтіннің *87,630* пайызын қамтып жатқанын айтуға болады.

1.1-кесте. Қазақ тіліндегі әріптердің жиілікті-әліпбилі жиілік сөздігі

№	Әріп	Абсолютті жиілігі	Бір топ әріптің мәтінді қамту пайызы	№	Әріп	Абсолютті жиілігі	Бір топ әріптің мәтінді қамту пайызы
1	2	3	4	1	2	3	4
1	а	8630862	12.796	22	г	903479	91.763
2	е	5660025	21.187	23	ң	872812	93.057
3	ы	5378832	29.161	24	и	866761	94.342
4	н	4192570	35.377	25	ж	738252	95.437
5	і	4154893	41.537	26	ұ	673484	96.435
6	т	4066134	47.565	27	ө	655787	97.408
7	р	4028709	53.537	28	ү	598755	98.295
8	л	3824239	59.207	29	ә	524073	99.072
9	д	3413312	64.267	30	я	231433	99.415
10	с	2708334	68.283	31	х	105688	99.572
11	м	2040371	71.308	32	ц	81165	99.692
12	қ	1876617	74.090	33	ф	58378	99.779
13	о	1639224	76.520	34	в	55978	99.862
14	к	1633360	78.941	35	э	30250	99.907
15	ғ	1256146	80.804	36	ь	25076	99.944
16	б	1229938	82.627	37	ю	20877	99.975
17	й	1175543	84.370	38	h	5406	99.983
18	у	1103283	86.006	39	ъ	5148	99.991
19	з	1095869	87.630	40	ч	3702	99.996
20	п	978676	89.081	41	щ	1679	99.998
21	ш	905698	90.424	42	ё	1018	100

Енді 2-топқа, қолдану жиілігі жағынан абсолютті жиіліктері *978676-524073* аралығында жататын сан жағынан 10-ға тең *(п, ш, г, ң, и, ж, ұ, ө, ү, ә)* әріптер екен. Аталған 10 әріп қайталана қолдана келіп, барлық мәтіннің 36,443 пайызын қамтып жатыр. Сол сияқты, 3-топқа сирек қолданатын басқа тілдерден енген келесі 13 әріпті жатқыздық: *я, х, ц, ф, в, э, ь, ю, h, ъ, ч, щ, ё* .

3-топтағы кірме дыбыстарға жататын бұл әріптер 7 миллионнан астам сөзқолданыстан тұратын қазақ тілінің 5 стилі бойынша алынған мәтіндердің бар болғаны 0,928 пайызын ғана қамтиды екен, яғни ол 1 пайызға да жетпейтін мәтінді қамту пайызы.

Жоғарыда сөз болған қазақ әріптерінің статистикалық мәліметтері қазақтың ұлттық компьютерлік пернетақсына қазақ әліпбиін латынға көшіру кезінде пайдалануға мүмкіндік береді.

Енді қазақ әріптерінің екіәріптік тіркестерінің статистикалық сипаттамасына қысқаша тоқталайық.

1.2-кесте. Қазақ тілінің екі әріптік тіркестерінің жиілік сөздігінен үзінді

Рет саны №	Әріп тіркесі	Абсолютті жиілік	Бір топ әріп тіркестерінің мәтінді қамту пайызы
1	ар	688045	2.035495
2	ан	582201	3.757865
3	ал	565089	5.429610
4	ын	562671	7.094202
5	да	560939	8.753670
6	та	511161	10.265877
7	ен	501533	11.749600
8	ер	481450	13.173909
9	қа	460416	14.535993
10	де	458037	15.891038
11	ің	451434	17.226550
12	ла	422562	18.476647
13	ды	417831	19.712748
14	ты	404750	20.910150
15	нд	399330	22.091518
16	га	396123	23.263399
17	ке	371958	24.363790
18	ай	369359	25.456493
19	лы	342672	26.470245
20	ті	333046	27.455520
21	ді	331481	28.436166
22	ас	330440	29.413731
23	сы	327308	30.382031
24	ме	323281	31.338418
25	ат	284740	32.180786
26	ба	281240	33.012799
27	ны	276945	33.832107
28	ын	275833	34.648124
29	ол	273816	35.458175
30	те	273269	36.266607
31	ет	272744	37.073486
32	ры	263800	37.853906
33	ық	261904	38.628716
34	ра	257203	39.389619
35	на	250705	40.131299
36	жа	248531	40.866547
37	ақ	243641	41.587328
38	ыл	241848	42.302805
39	ле	241186	43.016324
40	не	239816	43.725790
41	ст	239516	44.434368
42	ір	237178	45.136030
43	рі	230151	45.816903
44	ма	229464	46.495744
45	са	226934	47.167100
46	ге	226772	47.837976
47	ел	209952	48.459093
48	ек	208945	49.077231
49	гы	207327	49.690582
50	лд	198170	50.276843
51	ыр	197109	50.859966
52	аз	196589	51.441550
53	рд	195039	52.018548
54	ес	189725	52.579826
55	іл	189148	53.139397
56	тт	184683	53.685759
57	бі	184178	54.230626
58	ің	181594	54.767850
59	лі	181226	55.303984
60	гі	178293	55.831442
61	ау	177800	56.357441
62	қы	172081	56.866521
63	ре	169334	57.367475
64	ыс	168863	57.867035
65	сі	168801	58.366412
66	ып	166290	58.858360
67	ші	163289	59.341430
68	ам	161755	59.819962
69	ік	154050	60.275700
70	ым	147729	60.712738
71	ім	147386	61.148761
72	бо	146149	61.581124
73	кт	144973	62.010009
74	ад	144878	62.438613
75	бе	143874	62.864246
76	шы	140779	63.280723
77	се	137879	63.688621
78	ем	133338	64.083085
79	із	131926	64.473371
80	кі	130520	64.859499
81	ег	128229	65.238848
82	ко	127963	65.617411
83	ор	126556	65.991811
84	ит	120653	66.348748
85	мы	120124	66.704120
86	ыз	119848	67.058675
87	ұр	118474	67.409166
88	ша	113697	67.745524
89	ей	113265	68.080604
90	зд	113149	68.415342
91	аз	113080	68.749875
92	өз	112899	69.083872
93	ко	112540	69.416808
94	рт	107298	69.734236
95	ін	105436	70.046155
96	ші	105243	70.357503
97	он	104735	70.667349
98	рл	104106	70.975334
99	же	103331	71.281025
100	ед	102174	71.583294
101	йы	98085	71.873467
102	лг	96179	72.158000
103	кт	95018	72.439099
104	па	92954	72.714092
105	за	91058	72.983476
106	мі	90191	73.250295
107	ыг	89655	73.515528
108	зі	89577	73.780530
109	ұы	89453	74.045166
110	мд	87990	74.305473
111	іс	86953	74.562713
112	со	86002	74.817139
113	йт	85321	75.069551
114	йд	84808	75.320445
115	ец	83836	75.568463
116	ту	82158	75.811517
117	ту	81604	76.052932
118	ой	81473	76.293960
119	ру	78168	76.525210
120	ұл	78080	76.756200

Рет саны №	Әріп тіркесі	Абсолютті жиілік	Бір топ әріп тіркестерінің мәтінді қамту пайызы
121	*ку*	77632	76.985865
122	*рм*	77476	77.215068
123	*еп*	77425	77.444120
124	*ия*	76718	77.671081
125	*ше*	75995	77.895903
126	*ац*	69420	78.101273
127	*ос*	68825	78.304883
128	*ән*	67263	78.503873
129	*ур*	67191	78.702649
130	*рг*	66537	78.899490
131	*пт*	64861	79.091373
132	*рг*	64652	79.282638
133	*ор*	63611	79.470824
134	*ок*	62543	79.655849
135	*аб*	62542	79.840872
136	*то*	61576	80.023037
137	*әр*	61565	80.205169
138	*бұ*	60962	80.385518
139	*йл*	60354	80.564068
140	*уд*	60316	80.742505
141	*жсы*	59382	80.918179
142	*кұ*	59169	81.093223
143	*жо*	58694	81.266862
144	*тұ*	57334	81.436478
145	*ер*	54592	81.597981
146	*ңш*	54403	81.758926
147	*жә*	53682	81.917737
148	*ка*	53591	82.076280
149	*ңд*	53060	82.233251
150	*жұ*	52798	82.389447
151	*аш*	52156	82.543744
152	*лм*	51437	82.695914
153	*от*	50604	82.845620
154	*ти*	50538	82.995130
155	*лг*	49924	83.142824
156	*өл*	49481	83.289208
157	*ез*	48885	83.433828
158	*ду*	48642	83.577729
159	*үй*	47945	83.719568
160	*ха*	47869	83.861183
161	*іг*	47318	84.001167
162	*со*	45850	84.136809
163	*уа*	45677	84.271939
164	*сқ*	45262	84.405841
165	*тк*	44909	84.538698
166	*ик*	44264	84.669648
167	*бы*	44121	84.800174
168	*рс*	43707	84.929476
169	*үш*	42635	85.055606
170	*әс*	42547	85.181476
171	*ус*	42402	85.306917
172	*йі*	40912	85.427950
173	*ни*	40860	85.548830
174	*зы*	40487	85.668605
175	*ур*	40401	85.788127
176	*йн*	40271	85.907264
177	*лт*	40245	86.026323
178	*су*	40052	86.144812
179	*ұл*	39988	86.263112
180	*бі*	39980	86.381388
181	*ри*	39419	86.498004
182	*еш*	39268	86.614173
183	*пе*	38952	86.729408
184	*уш*	38833	86.844290
185	*ие*	38602	86.958490
186	*өт*	38302	87.071801
187	*мә*	37581	87.182980
188	*ко*	37456	87.293789
189	*зе*	37193	87.403820
190	*тқ*	37160	87.513753
191	*си*	36662	87.622213
192	*ыт*	35854	87.728283
193	*уг*	35743	87.834024
194	*әл*	35728	87.939721
195	*ми*	35603	88.045048
196	*ул*	35192	88.149159
197	*сп*	34919	88.252462
198	*ом*	33837	88.352565
199	*ри*	33105	88.450502
200	*жұ*	33023	88.548196
201	*ші*	32705	88.644950
202	*ог*	32191	88.740183
203	*ял*	32129	88.835233
204	*нт*	31942	88.929729
205	*ум*	31057	89.021608
206	*йц*	30876	89.112950
207	*уі*	30113	89.202036
208	*іш*	29962	89.290675
209	*ен*	29755	89.378701
210	*лш*	29713	89.466603
211	*ұл*	29290	89.553254
212	*ск*	29049	89.639192
213	*мұ*	29004	89.724997
214	*ая*	28939	89.810609
215	*лс*	28186	89.893994
216	*нг*	27874	89.976456
217	*цы*	27579	90,058045

Қазақ тілінің екі әріптік тіркестерінің жиілік сөздігін құрастыру үшін жоғарыда аталған 7 миллионнан астам сөзқолданыстан тұратын қазақ тілінің мәтінінен құрастырылған «Қазақ тілінің сөзформалар жиілік сөздігі» негіз болды.

Компьютер көмегімен құрастырылған екіәріптік жиілік сөздікте әртүрлі тіркестердің реестрлік рет саны 1160-қа тең, ал осы екіәріптік тіркестер қайталанып қолдана келіп, олардың абсолютті жиіліктерінің қосындысы 67.604670 мәтін бойындағы екіәріптік тіркестердің 100 пайызын құрайды.

Енді осы сөздіктің жоғары жиілікті бөлігінен үзінді ретінде барлық мәтіннің 90

пайызын ғана қамтыйтын жоғары жиілікті зонаға жататын 217 екіәріптік тіркестер бөлігін *1.2-кесте* бойынша қарастыруға болады.

Қазақ тілінің ұлттық компьютерлік пернетақтасында *1.1-кестедегі* жиі кездесетін 1-топтағы 19 әріп екі бөлікке бөлініп, пернетақтаның ортаңғы бөлігіне орналасады. Сол сияқты, *1.2-кестедегі* екіәріптік тіркестердің ең жиі қолданатын, мәселен мәтіннің 90 пайызын қамтитын рет сандары 1-217 аралығын қамтитын тіркестердің екінші әріптерін *1.1-кестедегі* рет сандары 1-29 аралығындағы әріптерге тіркесуге пернетақтаның жақын аралығына, яғни сол қол мен оң қол саусақтарына сәйкестендіріп орналастыру мүмкіндіктері қарастырылды. Әрине, бұл әрекетке үлкен мән берілуі қажет деп білеміз.

Статистикалық мәліметтер толық болу үшін біз өз зерттеуімізде сөз басында және сөз соңында кездесетін екіәріптік тіркестерінің де жиілік сөздіктерін құрастырдық. Енді ол сөздіктердің үзінділерімен, яғни жиі қолдана отырып, мәтіннің 90 пайызын қамтитыны жайлы *1.3-кесте* және *1.4-кесте* арқылы танысуға болады.

1.3-кесте бойынша сөз басындағы екіәріптік тіркестердің статистикалық сипатын айқындауға болады.

1.3-кесте. Қазақ тілінің сөз басындағы екі әріптік тіркестерінің жиілік сөздігі (үзінді)

Рет саны №	Әріп тіркесі	Абсолютті жиілік	Бір топ әріп тіркестерінің мәтінді қамту пайызы
1	2	3	4
1	*ка*	314514	5.33657
2	*жа*	228241	9,20929
3	*ба*	222845	12.99045
4	*ке*	177014	15.99396
5	*бі*	149781	18.53540
6	*бо*	144030	20.97925
7	*та*	142349	23.39458
8	*кө*	126340	25.53828
9	*ме*	123027	27.62576
10	*са*	116450	29.60164
11	*де*	116241	31.57398
12	*ко*	108225	33,41031
13	*ал*	99351	35.09607
14	*бе*	91785	36.65344
15	*со*	81582	38.03770
16	*же*	75181	39.31335
17	*құ*	73538	40.56111
18	*ту*	70582	41.75873
19	*ай*	65318	42.86702
20	*бұ*	59954	43.88430
21	*ор*	58427	44.87567
22	*қы*	58085	45.86124
23	*ку*	57775	46.84154
24	*жо*	57314	47.81403
25	*ту*	55624	48.75784
26	*ма*	55004	49.69113
27	*ар*	54779	50.62060
28	*жә*	53475	51.52795
29	*шы*	51825	52.40730
30	*өз*	50541	53.26486
31	*те*	50297	54.11828

Рет саны №	Әріп тіркесі	Абсолютті жиілік	Бір топ әріп тіркестерінің мәтінді қамту пайызы
1	2	3	4
30	*өз*	50541	53.26486
31	*те*	50297	54.11828
32	*жү*	49984	54.96639
38	*ша*	43897	59.68356
39	*се*	42067	60.39734
40	*то*	38897	61.05733
41	*ха*	37809	61.69886
42	*не*	36931	62.32549
43	*от*	36633	62.94707
44	*ау*	36611	63.56827
45	*он*	36401	64.18591
46	*ре*	36121	64.79880
47	*мә*	35588	65.40265
48	*ос*	34590	65.98956
49	*да*	34176	66.56945
50	*жү*	32232	67.11635
51	*кі*	31591	67.65238
52	*ес*	31011	68.17856
53	*ок*	30805	68.70125
54	*ту*	30296	69.21530
55	*өт*	28845	69.70473
56	*ер*	28770	70.19289
57	*па*	28510	70.67664
58	*ад*	27917	71.15033
59	*ше*	27320	71.61389
60	*ко*	27211	72.07559
61	*ак*	27077	72.53503
62	*үш*	25670	72.97059
63	*сы*	25579	73.40460
64	*му*	25496	73.83721
65	*ас*	25250	74.26564

1	2	3	4
66	**то**	24990	74,68967
67	**ет**	24806	75,11057
68	**ол**	24071	75,51899
69	**за**	24004	75,92629
70	**тә**	23876	76,33141
71	**га**	23492	76,73001
72	**мы**	23137	77,12259
73	**ой**	22516	77,50464
74	**әр**	22221	77,88167
75	**ел**	21756	78,25082
76	**ұл**	20966	78,60657
77	**әл**	20143	78,94835
78	**іс**	19666	79,28203
79	**бө**	19398	79,61117
80	**ем**	18988	79,93335
81	**ед**	18731	80,25118
82	**му**	18613	80,56700
83	**сұ**	18011	80,87260
84	**дә**	17976	81,17761
85	**үй**	17527	81,47500
86	**ан**	17500	81,77194
87	**би**	17496	82,06880
88	**на**	17402	82,36407
89	**ен**	17090	82,65405
90	**әд**	16843	82,93984
91	**ты**	16671	83,22271
92	**су**	16160	83,49690
93	**өл**	15674	83,76286
94	**өн**	15137	84,01970

1	2	3	4
95	**си**	15010	84,27438
96	**іш**	14993	84,52878
97	**ұл**	14866	84,78102
98	**ұс**	14847	85,03294
99	**пр**	14819	85,28438
100	**кә**	14494	85,53031
101	**әк**	13795	85,76438
102	**ти**	13599	85,99512
103	**өм**	13112	86,21760
104	**сә**	12973	86,43772
105	**бұ**	12953	86,65751
106	**бә**	12880	86,87605
107	**жи**	12741	87,09224
108	**ки**	12615	87,30628
109	**мо**	12530	87,51889
110	**аш**	12060	87,72352
111	**жі**	12013	87,92735
112	**ми**	11877	88,12887
113	**үй**	11858	88,33008
114	**аг**	11826	88,53074
115	**аз**	11363	88,72354
116	**ең**	10523	88,90209
117	**ая**	10468	89,07971
118	**сұ**	10463	89,25724
119	**дұ**	10427	89,43416
120	**ин**	10291	89,60878
121	**әс**	9983	89,77817
122	**до**	9936	89,94676
123	**жө**	9872	90,11426

Мысалы, ең жиі қолданатын екіәріптік тіркестердің алғашқы әріпіне қарай олардың екінші әріптерінің қандай екенін анықтауға болады. *1.3-кестеде* рет саны 1-ші болып тұрған *«қа»* екіәріптік тіркес мәтінде ең жиі кездесіп, оның абсолютті жиілігі 314514-ке тең және барлық мәтінді қамту пайызы – 5,33%. Енді, мысал үшін, осы *«қ»* әріпінің мәтіннің 90 пайызын қамтып жатқан 123 тіркестер ішінде қандай әріптермен тіркесіп жиі қолданатынын анықтайық. Мысалы, жиілік сөздіктің жиі қолданған аясында орналасқан *қа, қо, құ, қы* тіркестерін, сол сияқты *«ж»* әріпінің 11 түрлі тіркестері: *жа, же, жо, жә, жү, жы, жы, жұ, жи, жі, жө* жиі қолданып кездескен екен. Сол сияқты, *«т»* әріпінің 8 түрлі жиі тіркесін келтірсек, олар: *та, тү, тұ, те, то, ту, ты, ти.* Міне мұндай мысалдарды қазақ әліпбиінің басқа да әріптері бойынша статистикалық мәліметтер мақсатында келтіруге болады.

Екіәріптік тіркестердің жиілік сөздіктерінің үзінділері *1.2, 1.3, 1.4-кестелерде* көрініс тапты, оларда тіркестердің сөз ішінде орналасу орнына қарай бөліп қарастырылды. Мысалы, *1.2-кестеде* әріп тіркесі сөзформа ішінде алатын орнына қатыссыз, *1.3-кестеде* – сөз басындағы тіркестер, *1.4-кестеде* – сөз соңындағы тіркестердің статистикасы жайлы мәліметтер қамтылды. Енді оларды қазақ тілінің компьютерлік ұлттық пернетақтасын жасауда қалай пайдаланамыз деген сұрақ туындауы мүмкін.

1.4-кесте. Қазақ тілінің сөз соңындағы екіәріптік тіркестерінің жиілік сөздігі (үзінді)

Рет саны №	Әріп тіркесі	Абсолютті жиілік	Бір топ әріп тіркестерінің мәтінді қамту пайызы	Рет саны №	Әріп тіркесі	Абсолютті жиілік	Бір топ әріп тіркестерінің мәтінді қамту пайызы
1	2	3	4	1	2	3	4
1	ен	319537	5,41253	37	ек	46525	72,87324
2	ан	272130	10,02205	38	ім	44391	73,62517
3	ың	240700	14,09919	39	ау	43245	74,35768
4	ды	240100	18,16617	40	қа	43056	75,08699
5	ын	195229	21,47309	41	ыс	42655	75,80951
6	ді	187867	24,65530	42	ым	39675	76,48155
7	ін	176123	27,63859	43	ша	37192	77,11154
8	ық	165604	30,44370	44	ет	36345	77,72717
9	ің	157597	33,11319	45	лі	35812	78,33378
10	ар	149557	35,64648	46	та	35423	78,93380
11	ып	146130	38,12173	47	ұл	32593	79,48588
12	да	130213	40,32737	48	сі	32030	80,02843
13	ай	113729	42,25378	49	ру	31677	80,56499
14	не	109065	44,10120	50	ес	30999	81,09007
15	де	102506	45,83752	51	ні	30949	81,61431
16	га	101192	47,55157	52	ке	30465	82,13035
17	ты	98951	49,22767	53	са	29680	82,63309
18	ік	95044	50,83759	54	ол	27248	83,09463
19	ер	91362	52,38514	55	ші	26643	83,54593
20	на	89806	53,90634	56	ей	26551	83,99567
21	сы	89508	55,42248	57	се	25855	84,43362
22	гы	80934	56,79340	58	ыр	24862	84,85474
23	ті	80568	58,15811	59	ат	24729	85,27362
24	ге	80523	59,52207	60	ра	24671	85,69152
25	ін	78056	60,84423	61	ас	23516	86,08985
26	ақ	75244	62,11876	62	ам	23501	86,48792
27	гі	69045	63,28829	63	те	23009	86,87766
28	лы	68298	64,44517	64	ла	22852	87,26475
29	рі	64818	65,54310	65	ма	22668	87,64871
30	ір	64774	66,64029	66	шы	22100	88,02306
31	ап	64127	67,72651	67	кі	22094	88,39730
32	еп	56991	68,69187	68	ту	21449	88,76062
33	ыз	53348	69,59551	69	уы	20591	89,10940
34	ры	51962	70,47568	70	ше	19712	89,44330
35	із	47858	71,28633	71	ыл	18768	89,76120
36	ны	47161	72,08517	72	ок	17887	90,06418

Осы сауалға қысқаша жауап берсек, біздің ұстанымыз бойынша ең жиі қолданатын қазақ әріптері пернетақтаның ортаңғы жағында орналасуы қажет. Олар екі топқа бөлініп, бірінші, сол жақ қол саусақтарына және екінші, оң жақ қол саусақтарына деп ажыратылуы қажет. Сонымен бірге, екіәріптік тіркестердің статистикасы бойынша тіркестің екінші әріпі, тіркестің бірінші әріпінің оң жағына жақын орналасуы қажет деп білеміз. Ондай дәрежеге бір емес, бірнеше әріп «үміткер» болуы мүмкін. Осындай жағдайда екіәріптік тіркестің позициялық жағдайына да мән берілуі керек.

Келесі айтайын дегеніміз, қазақ әліпбиін латынға көшіру мәселесінде тек әріп және әріп тіркестерінің статистикасынан басқа қазақ сөзінің құрылымдық бірліктің алғашқысы ретінде «буындық» бірлікті жатқызуға болады. Сол себепті біз «Қазақ тілі буындарының жиілік сөздігін» компьютер көмегімен жасап, қазақ тілі буындарына статистикалық сипаттама беруді де мақсат етіп қойдық.

Қазақ тілінің сөзформаларындағы буын жігін автоматты түрде ажырату мәселесі әлі толық шешімін таппағандықтан бұл процесті жартылай қолмен жүзеге асыруға тура келді.

Төменде көрініс тапқан «Қазақ тілі буындарының жиілік сөздігі» негізінде қазақ тіліндегі буындардың реттік саны 1225 тең болса, олардың қайталана қолдануы негізінде буындардың қосынды жиілігі 124063 тең екендігі дәлелденді.

Қазақ тіліндегі буындарды дыбыстық жағынан қарастыратын болсақ, оларды 3 топқа бөліп қарастыруға болатыны мәлім: 1) Ашық буын; 2) Тұйық буын; 3) Бітеу буын. Жиілік сөздікте буын түрлері шартты белгілермен таңбаланды: *ашық буын – АБ, тұйық буын – ТБ, бітеу буын – ББ* арқылы, ал *дауысты дыбыс – Ды, дауыссыз дыбыс – Дз* арқылы және олардың халықаралық шартты белгілері: *дауысты дыбыс – V, дауыссыз дыбыс – С* арқылытаңбаланды.

Енді қазақ тілі буындарының кері-әліпбилі жиілік сөздігінен шартты белгіленімдер бойынша буындардың түрлерін, яғни *ашық буын – АБ, тұйық буын – ТБ, бітеу буын – ББ* бойынша жеке-жеке қарастыруға болады. Ашық буынға қатысты кері-әліпбилі жиілік сөздігінен үзінді төменде *2.1-кестеде* көрініс тапты. Жиілік сөздіктің реттік саны бойынша барлық *ашық буынға (АБ)* жататын бірліктер саны 173-ке тең, ал олар қайталана қолдана келіп мәтіннің 54,59 пайызын қамтитынына көз жеткізуге болады.

Кесте деректеріне сүйенетін болсақ, *ашық буындардың (АБ)* 11-і бір дауысты дыбыстан тұратын буындар, яғни *а, ә, е, и, о, ө, у, ұ, ү, ы, і* дауысты дыбыстар. Аталған буындар мәтін ішінде қайталана қолданып, барлық буын түрлерінің 5,7 пайызын қамтиды. Аталған 11 дауысты дыбыстардың ішінен ең жиі қолданғандары: *«а»* – 1944 рет, *«о»* – 1603 рет, *«е»* – 1083 рет.

Сол сияқты, *бітеу буын (ББ)* жайлы сөз ететін болсақ, олар кері-әліпбилі жиілік сөздікте ашық буынға қарағанда сандық жағынан айтарлықтай жиі қолданады деуге болады. Мәселен, рет саны бойынша бітеу буынның реестрдегі саны 952-ге тең де, ал барлық буын бірліктерінің мәтін ішінде қайталана қолданып, қазақ тілінің 5 стилі бойынша алынған мәтіннің 41 пайызын қамтиды екен. Қазақ тіліндегі бітеу буындардың кері-әліпбилі жиілік сөздігінің үзіндісі *2.2-кестеде* берілді.

2.1-кесте. Қазақ тіліндегі ашық буындардың кері-әліпбилі жиілік сөздігі

Рет саны №	Буын	Абсолютті жиілік	Бір топ буынның мәтінді қамту пайызы
1	*а/V/A*	1944	1,5669
2	*ә/V/A*	639	2,0820
3	*е/V/A*	1083	2,9550
4	*и/V/A*	50	2,9953
5	*о/V/A*	1603	4,2873
6	*ө/V/A*	802	4,9338
7	*у/V/A*	15	4,9459
8	*ұ/V/A*	403	5,2707
9	*ү/V/A*	375	5,5730
10	*ы/V/A*	30	5,5972
11	*і/V/A*	138	5,7084
12	*ба/C+V/A*	997	6,5120
13	*ға/C+V/A*	1	6,5128
14	*ға/C+V/A*	1538	7,7525
15	*да/C+V/A*	2522	9,7854
16	*жа/C+V/A*	1020	10,6075
17	*за/C+V/A*	256	10,8139
18	*йа/C+V/A*	307	11,0613
19	*қа/C+V/A*	2520	13,0925
20	*ла/C+V/A*	1793	14,5378

1	2	3	4
21	*ма/С+V/А*	876	15,2439
22	*на/С+V/А*	1354	16,3352
23	*ңа/С+V/А*	107	16,4215
24	*па/С+V/А*	230	16,6069
25	*ра/С+V/А*	1129	17,5169
26	*са/С+V/А*	1080	18,3874
27	*та/С+V/А*	1852	19,8802
28	*ұа/С+V/А*	213	20,0519
29	*ха/С+V/А*	278	20,2760
30	*һа/С+V/А*	4	20,2792
31	*ша/С+V/А*	726	20,8644
32	*бә/С+V/А*	88	20,9353
33	*дә/С+V/А*	70	20,9918
34	*жә/С+V/А*	357	21,2795
35	*зә/С+V/А*	5	21,2835
36	*иә/С+V/А*	6	21,2884
37	*кә/С+V/А*	103	21,3714
38	*лә/С+V/А*	1	21,3722
39	*мә/С+V/А*	216	21,5463
40	*нә/С+V/А*	37	21,5761
41	*пә/С+V/А*	38	21,6068
42	*рә/С+V/А*	13	21,6172
43	*сә/С+V/А*	31	21,6422
44	*тә/С+V/А*	100	21,7228
45	*ұә/С+V/А*	60	21,7712
46	*шә/С+V/А*	5	21,7752
47	*бе/С+V/А*	581	22,2435
48	*ге/С+V/А*	901	22,9698
49	*де/С+V/А*	2226	24,7640
50	*же/С+V/А*	364	25,0574
51	*зе/С+V/А*	97	25,1356
52	*ие/С+V/А*	35	25,1638
53	*йе/С+V/А*	243	25,3597
54	*ке/С+V/А*	1366	26,4607
55	*ле/С+V/А*	1058	27,3135
56	*ме/С+V/А*	705	27,8818

1	2	3	4
57	*не/С+V/А*	1366	28,9829
58	*ңе/С+V/А*	30	29,0070
59	*пе/С+V/А*	56	29,0522
60	*ре/С+V/А*	454	29,4181
61	*се/С+V/А*	638	29,9324
62	*те/С+V/А*	774	30,5562
63	*уе/С+V/А*	20	30,5724
64	*ше/С+V/А*	436	30,9238
65	*би/С+V/А*	142	31,0383
66	*ги/С+V/А*	34	31,0657
67	*ди/С+V/А*	12	31,0753
68	*жи/С+V/А*	93	31,1503
69	*ки/С+V/А*	10	31,1584
70	*қи/С+V/А*	91	31,2317
71	*ли/С+V/А*	23	31,2503
72	*ми/С+V/А*	68	31,3051
73	*ни/С+V/А*	61	31,3542
74	*пи/С+V/А*	6	31,3591
75	*ри/С+V/А*	118	31,4542
76	*си/С+V/А*	73	31,5130
77	*ти/С+V/А*	48	31,5517
78	*хи/С+V/А*	10	31,5598
79	*ши/С+V/А*	16	31,5727
80	*бо/С+V/А*	579	32,0394
81	*до/С+V/А*	18	32,0539
82	*жо/С+V/А*	214	32,2264
83	*қо/С+V/А*	407	32,5544
84	*мо/С+V/А*	23	32,5730
85	*но/С+V/А*	1	32,5738
86	*ро/С+V/А*	6	32,5786
87	*со/С+V/А*	224	32,7592
88	*то/С+V/А*	103	32,8422
89	*шо/С+V/А*	17	32,8559
90	*бө/С+V/А*	103	32,9389
91	*гө/С+V/А*	16	32,9518
------	----------	-------	---------

2.2-кесте. Қазақ тіліндегі бітеу буындардың кері-әліпбилі жиілік сөздігінен үзінді

Рет саны №	Буын	Абсолютті жиілік	Бір топ буын-ның мәтінді қамту пайызы
174	*желк/С+V+С+С/Б*	2	54.5908
175	*булк/С+V+С+С/Б*	5	54.5948
176	*сұлк/С+V+С+С/Б*	2	54.5965

Рет саны №	Буын	Абсолютті жиілік	Бір топ буын-ның мәтінді қамту пайызы
177	*сылқ/С+V+С+С/Б*	3	54.5989
178	*данк/С+V+С+С/Б*	4	54.6021
179	*жарк/С+V+С+С/Б*	2	54.6037

1	2	3	4
180	*жүрк/С+V+С+С/Б*	1	54.6045
181	*шырк/С+V+С+С/Б*	5	54.6085
182	*балп/С+V+С+С/Б*	2	54.6102
183	*тырп/С+V+С+С/Б*	1	54.6110
184	*дүрс/С+V+С+С/Б*	2	54.6126
185	*мырс/С+V+С+С/Б*	5	54.6174
186	*ғайт/С+V+С+С/Б*	1	54.6182
187	*жайт/С+V+С+С/Б*	11	54.6271
188	*кайт/С+V+С+С/Б*	12	54.6368
189	*тайт/С+V+С+С/Б*	1	54.6376
190	*жәйт/С+V+С+С/Б*	1	54.6384
191	*сөйт/С+V+С+С/Б*	2	54.6400
192	*жүйт/С+V+С+С/Б*	1	54.6408
193	*туйт/С+V+С+С/Б*	4	54.6440
194	*жалт/С+V+С+С/Б*	2	54.6456
195	*салт/С+V+С+С/Б*	12	54.6553
196	*шалт/С+V+С+С/Б*	1	54.6561
197	*зелт/С+V+С+С/Б*	1	54.6569
198	*селт/С+V+С+С/Б*	1	54.6577
199	*бұлт/С+V+С+С/Б*	3	54.6601
200	*жүлт/С+V+С+С/Б*	1	54.6609
201	*ғылт/С+V+С+С/Б*	3	54.6634
202	*кілт/С+V+С+С/Б*	4	54.6666
203	*лант/С+V+С+С/Б*	3	54.6690
204	*уант/С+V+С+С/Б*	1	54.6698
205	*зент/С+V+С+С/Б*	1	54.6706
206	*карт/С+V+С+С/Б*	2	54.6722
207	*нарт/С+V+С+С/Б*	3	54.6746
208	*тарт/С+V+С+С/Б*	11	54.6835
209	*шарт/С+V+С+С/Б*	9	54.6908
210	*берт/С+V+С+С/Б*	1	54.6916
211	*дерт/С+V+С+С/Б*	4	54.6948
212	*зерт/С+V+С+С/Б*	28	54.7174
213	*керт/С+V+С+С/Б*	3	54.7198
214	*серт/С+V+С+С/Б*	2	54.7214
215	*жорт/С+V+С+С/Б*	1	54.7222
216	*төрт/С+V+С+С/Б*	17	54.7359
217	*жүрт/С+V+С+С/Б*	55	54.7802
218	*құрт/С+V+С+С/Б*	6	54.7851
219	*мүрт/С+V+С+С/Б*	2	54.7867
220	*күрт/С+V+С+С/Б*	4	54.7899

1	2	3	4
221	*түрт/С+V+С+С/Б*	3	54.7923
222	*ғырт/С+V+С+С/Б*	1	54.7931
223	*дырт/С+V+С+С/Б*	2	54.7947
224	*жырт/С+V+С+С/Б*	3	54.7972
225	*йырт/С+V+С+С/Б*	1	54.7980
226	*кырт/С+V+С+С/Б*	3	54.8004
227	*пырт/С+V+С+С/Б*	2	54.8020
228	*сырт/С+V+С+С/Б*	16	54.8149
229	*тырт/С+V+С+С/Б*	1	54.8157
230	*уырт/С+V+С+С/Б*	3	54.8181
231	*шырт/С+V+С+С/Б*	1	54.8189
232	*гірт/С+V+С+С/Б*	1	54.8197
233	*дірт/С+V+С+С/Б*	1	54.8205
234	*кірт/С+V+С+С/Б*	3	54.8230
235	*сірт/С+V+С+С/Б*	1	54.8238
236	*тірт/С+V+С+С/Б*	1	54.8246
237	*карш/С+V+С+С/Б*	4	54.8278
238	*кірш/С+V+С+С/Б*	2	54.8294
239	*жаб/С+V+С/Б*	4	54.8326
240	*раб/С+V+С/Б*	20	54.8487
241	*хаб/С+V+С/Б*	1	54.8496
242	*шаб/С+V+С/Б*	1	54.8504
243	*гиб/С+V+С/Б*	1	54.8512
244	*коб/С+V+С/Б*	1	54.8520
245	*сыб/С+V+С/Б*	4	54.8552
246	*мег/С+V+С/Б*	1	54.8560
247	*бөг/С+V+С/Б*	4	54.8592
248	*баг/С+V+С/Б*	87	54.9294
249	*даг/С+V+С/Б*	34	54.9568
250	*жаг/С+V+С/Б*	109	55.0446
– – –	– – – – – – – – – – –	– – –	– – – – –
1030	*быт/С+V+С/Б*	2	94.2279
1031	*ғыт/С+V+С/Б*	35	94.2561
1032	*зыт/С+V+С/Б*	1	94.2570
1033	*йыт/С+V+С/Б*	4	94.2602
1034	*кыт/С+V+С/Б*	55	94.3045
1035	*мыт/С+V+С/Б*	8	94.3110
1036	*ныт/С+V+С/Б*	11	94.3198
1037	*ныт/С+V+С/Б*	1	94.3206
1038	*рыт/С+V+С/Б*	2	94.3222
1039	*уыт/С+V+С/Б*	4	94.3255

Қазақ тіліндегі тұйық буындардың кері-әліпбилі жиілік сөздігінің статистикалық мәліметтері бойынша барлық тұйық буындардың рет саны 100-ге тең және олар қайталана қолдана келіп барлық мәтіннің 4,44 пайызын ғана қамтиды екен. Ал енді олардың құрылымдық жағына келетін болсақ, оны екі түрге (вариантқа) бөліп қарастыруға болады. Біріншісі, 3 дыбыстан тұрады, оның алғашқысы дауысты (V), ал келесі екеуі дауыссыз дыбыстар (С/С), яғни моделі – V/C/C. Мұндай тұйық буынды бірліктердің жиілік сөздіктегі саны 9-ға ғана тең (*айт, әйт, өйт, ұйт, ұлт, ант, арт, ерт, өрт)* және олардың қайталана қолдану нәтижесі барлық мәтіннің 0,2 пайызын ғана қамтыйды екен. Осы 9 тұйық буындарының ішінде ең жоғары қолданатындар *айт* – 97рет, *ұлт* – 67 рет және *өйт* – 24 рет қолданыс тапқан екен.

Үш дыбысты тұйық буындарға қарағанда е*кі дыбысты тұйық буындардың* са-ны айтарлықтай жоғары, яғни ол 91 санына тең және қайталана қолдану нәтижесін-де мәтінді қамту пайызы – 4,24%. Олардың ішінен ең жоғары қолдастағылар (100-ден көп) деп мына 16 тұйық буынды атауға болады: *ал* – 587, *ай* – 360, *өз* – 264, *ар* – 240, *ол* – 221, *іс* – 211, *ел* – 210, *ау* – 183, *өт* – 170, *ор* – 166, *ақ* – 153, *үй* – 149, *ең* – 143, *ат* – 138, *ас* – 121, *әл* – 106.

Ал қазір біз, ең жиі қолданылатын буын түрлеріне статистикалық сипаттама беру үшін «Қазақ тілі буындарының жиілікті-әліпбилі жиілік сөздігі» жайлы қысқаша

сөз етпекпіз. Өздеріңізге белгілі, жиілікті-әліпбилі жиілік сөздіктің анықтамасы бойынша, бұл сөздікте буын бірліктерінің ең жиі қолданылатынан бастау алып, әрі қарай олардың біртіндеп кему тәртібімен орналасатындығы жайлы айтылған. Сол себепті, жиілік сөздіктің бас жағында – буын түрлерінің ең жиі қолданылатын түрлері орналасқандықтан, сол буындардың статистикалық сипаттамасына қысқаша тоқталайық.

Жоғары жиілікті зонаға абсолютті жиіліктері 1000-нан жоғары, реттік саны 1 – 26 аралығындағы буындардың қайталана қолдана келіп, мәтіннің 30 пайызын қамтитын жиілік сөздік бөлігін жатқызуды ұйғардық және оны *бірінші топ* деп атауды ұйғардық. Аталған аралықта орналасқан 26 буынның 25-і ашық буын, ал тек 13-ші рет санында орналасқан «***ған***» бітеу буын ғана 1430 жиілікпен барлық мәтіннің 1,15 пайызын қамтып, жоғары жиілікті зонаға еніп тұр.

Енді буын жиіліктері 1000-нан төмен және 100-ден жоғары, яғни 100-реттен 1000 ретке дейін қолданған ашық буындарды *екінші топқа* жатқызатын болсақ, олардың түрлерінің саны 73-ке тең және олар қайталана қолдану нәтижесінде барлық мәтіннің 22,45 пайызын қамтып жатыр.

Қорыта айтқанда, жазба тілдің сөйлеу тілі сияқты практикалық және теориялық тұрғыдан аса маңызды зерттеу нысаны екені мәлім. Мысалы, ақпаратты автомат-ты түрде өңдеу мен шартты белгілермен таңбалау (кодтау) кезеңінде сол тілдің графемаларының статистикалық сипаттамаларын білу аса маңызды. Графемаларды статистикалық әдіспен талдау арқылы алынған нәтижелер баспа ісін жетілдіруде және компьютерді қазақ тілін кирилден латынға ұтымды көшіру шараларында да аса қажет. Жазба тіліне тән ерекшеліктер, яғни жазба мәтіннің сандық және сапалық қатынастары қазақ тілінің әртүрлі әдеби жанрларынан орын алады.

Мақалада қазақ тілінің жазба мәтінін әріптік (графемалық) деңгейде және қазақ тілі сөзформаларының құрылымына статистикалық әдіспен зерттеуге әрекет жасалған. Себебі, әріп пен сөз құрылымындағы буын табиғи тілді жазба мүмкін-дігі арқылы бейнелейтін ең бір қарапайым түрдегі көрнекі тілдік бірліктер болып саналады.

Зерттеу нәтижелері қазақ тілін фонологиялық және фонетикалық деңгейде тілдің морфемдік құрылымын талдау мен тілдік бірліктерді синтагматикалық тұрғыда қарастыру кезінде аса маңызды. Әріптердің, әріп тіркестерінің жалпы және шептік орналасу мүмкіндіктерінің статистикасын анықтауда профессор А.Қ.Жұбановтың «Задача получения на ЭВМ частотных списков лингвистических единиц» [1] атты мақаласында келтірілген алгоритмдік сызбасы компьютерлік программа жазуға негіз болды. Сонымен бірге, аталған автордың «К вопросу о графемной статистике казахского текста» [2] атты еңбегіндегі әріптердің статистикасы жайлы мәліметтер де осы мақаланы жазуға ықпал етті.

ӘДЕБИЕТТЕР ТІЗІМІ:

[1] Жубанов А.К. Задача получения на ЭВМ частотных списков лингвистических единиц. // Статистика казахского текста. – Алма-Ата, 1973. – С. 263-299.

[2] Жубанов А.К. К вопросу о графемной статистике казахского текста. В кн.: Вопросы казахской фонетики и фонологии. – Алма-Ата: Наука, 1979. – С.49-52.

[3] Исенгельдина А.А. Факторы, определяющие относительную частотность фонем. // Статистика казахского текста. – Алма-Ата, 1973. – С. 659-662.

[4] Пиотровский Р.Г. Моделирование фонологических систем и методы их сравнения. – М.Л.: Наука, 1966. – 299 с.