МРНТИ 16.21.37

А.Қ. Жұбанов

А. Байтұрсынұлы атындағы Тіл білімі институтының бас ғылыми қызметкері, филология ғылымдарының докторы, профессор
Алматы, Казахстан

ЛАТЫН ГРАФИКАСЫНА КӨШУДЕГІ ҚАЗАҚ ТІЛІ ӘРІПТЕРІНІҢ СТАТИСТИКАЛЫҚ ДЕРЕКТЕРІ

Аннотация: Жазу мен әліпби мәселесі бір-бірімен тығыз байланысты болатындықтан, ең алдымен, әр тілдің дыбыс қорын анықтап, содан кейін әр дыбысты әріптермен таңбалау қажет болады. Жазбаша тілді ауызша тілдің графикалық бейнесі деуге болады.

Мақалада қазақ әліпбиінің кирилден латын әліпбиіне көшу қажеттілігі және көшу барысында кездесетін қиындықтар, сондай-ақ латын графикасына көшу жолдары туралы айтылады. Қазақ әліпбиі графемаларының кездесу жиілігіне және олардың мәтінде қамтылу пайызына қазақ тіліндегі үш мәтіннің мысалында статистикалық талдау жүргізіледі. Қазақ әліпбиінің кирилден латын графикасына көшуі кезінде статистикалық деректер ескерілуі тиіс.

Тірек сөздер: мәтін, латын, әліпби, әріп, тіркес, графема, фонема, орфоэпия, орфография, дыбыс қоры, генерация, жиілік, абсолютті жиілік, қатынастық жиілік, компьютер, компьютерлік бағдарлама, бағдарламалық жасақтама, ақпарат.

А.К.Жубанов

главный научный сотрудник Института языкознания имени А. Байтурсынова, доктор филологических наук, профессор, Алматы, Казахстан

СТАТИСТИЧЕСКИЕ ДАННЫЕ ОТНОСИТЕЛЬНО ПЕРЕХОДА КАЗАХСКОГО АЛФАВИТА НА ЛАТИНСКУЮ ГРАФИКУ

Аннотация. Поскольку проблема письма и алфавита тесно связана между собой, прежде всего, необходимо будет определить звуковой фонд каждого языка, а затем маркировать каждый звук буквами. Письменный язык можно назвать графическим изображением устного языка.

В статье говорится о необходимости перехода казахского алфавита с кириллицы на латиницу и о возможных трудностях такого перехода, а также о путях перехода на латинскую графику. Проводится статистический анализ частоты встречаемости графем казахского алфавита и процентов их покрываемости на примере 3-х текстов казахского языка. Статистические данные должны быть учтены при переходе казахского алфавита с кириллицы на латинскую графику.

Ключевые слова: текст, латиница, алфавит, буква, фраза, графема, фонема, орфоэпия, правописание, звуковой фонд, генерация, частота, абсолютная частота, частота связи, компьютер, компьютерная программа, программное обеспечение, информация.

A.K.Zhubanov

Chief Researcher of the Institute of Linguistics named after A. Baitursynov,
doctor of philological sciences, professor

STATISTICAL DATA ON THE TRANSITION OF THE KAZAKH ALPHABET TO LATIN GRAPHICS

Annotation. Since the problem of writing and the alphabet is closely related, first of all, it will be necessary to determine the sound foundation of each language, and then label each sound with letters. Written language can be called a graphic representation of the spoken language.

The article discusses the necessity of the transition of the Kazakh alphabet from Cyrillic script to Latin graphics and possible difficulties of such transition, as well as possible ways of the transition to Latin graphics. The statistical analysis of frequency of occurrence of graphemes of the Kazakh alphabet and the percent of their coverage by example of 3 written texts of the Kazakh language has been conducted. Statistical data should be taken into account in the process of the transition of the Kazakh alphabet from Cyrillic script to Latin graphics

Keywords: text, Latin script, alphabet, letter, phrase, graphics, phoneme, orthoepy, spelling, sound fund, generation, frequency, absolute frequency, communication frequency, computer, computer program, software, information.

Латын әліпбиіне көшу, сайып келгенде, ана тіліміздің болашағын ойлап, қолданыс аясын одан әрі кеңейте түсуге мүмкіндік жасап, тіліміздің ішкі табиғи әліпбиіміз арқылы жазудың қазақы айтылым (орфоэпия) мен жазылым (орфография) талаптарын жүйеге түсіру деп түсіну керек.

Шындығында, бір жазудан екіншісіне көшу халықтың осы рухани байлықтан сусындауына қосымша қиындық келтіруі де мүмкін. Сондықтан әліпби мен жазу мәселесіне әлеуметтік лингвистика тұрғысынан жете назар аударған жөн. Жазу мен әліпби мәселесі бір-бірімен тығыз байланысты болатындықтан, ең алдымен, әр тілдің дыбыс қорын анықтап, содан кейін әр дыбысты әріптермен таңбалау қажет болады. Жазбаша тілді ауызша тілдің графикалық бейнесі деуге болады. Жазудың қандай түрі болмасын олар белгілі бір таңбалар арқылы жасалады. Әліпби жасау үшін алдымен графикалық лингвистика теориясын терең танып, тілдің дыбыс жүйесіндегі фонемалардың өзіндік фонологиялық ерекшеліктеріне жете назар аударып, бөгде тілдерден енген сөздерге қатысты фонемаларды қалай таңбалау керек деген мәселелерді шешіп алу керек, одан кейін тілдің болмысын танытатын негізгі заңдылықтарын біліп, тілдің табиғатына сай келетін, өзіндік дыбысталу ерекшеліктеріне кері әсер етпейтіндей мәселелер ескерілуі керек. Жазу барысында әр әріптің өзара кездесу жиілігі де назарға алынуы тиіс деп білеміз.

Латын графикасына (таңбасына) негізделген қазақ әліпбиін жасау ісінде тіліміз дыбыс жүйесіндегі төл дыбыстар мен өзге тілден енген кірме дыбыстарды жеке-жеке қарастырып, әрқайсысының өзіндік ерекшеліктері мен жазу барысында туындайтын әртүрлі заңдылықтары ескерілуі тиіс.

Әліпби өзгерту мәселесі тек кірме сөздерге ғана қатысты емес, өз тіліміздегі төл сөздерімізді де жазуда кеткен олқылықтарды жөнге келтіруге көмектеседі. Адам жазу арқылы білім алады, жазу арқылы бір-бірімен байланыс жасайды. Осы жазулар бүгінгі

ұрпақтың орфоэпия (айтылым) заңдылықтарын ұмытып, айтуын да, жазуын да орфография (жазылым) заңдылықтарымен жүруіне итермеледі. Осы орайда, тіліміздегі әріптерді қысқартып, компьютерге икемдеу, яғни тіліміздегі бар дыбыстарды орынсыз қысқартып жіберуге де болмайтынын ескеруіміз қажет. Компьютерді тілімізге икемдеудің орнына, тілімізді компьютерге икемдегеніміз дұрыс болмайды. Олай болса, қазақ тіліндегі 28 фонеманы компьютер пернетақтасына орналастыруға әбден болады демекпіз.

Тілдің қоғаммен бірге дамып, ғылым мен техника дамыған сайын көршілес елдерден және басқа да шетелдерден жаңа сөздер кіруі арқылы да толығып отыратыны табиғи заңдылық болып табылады. Сондықтан тіліміздің табиғи айтылымына қайшы келетін кемшіліктерімізден арылатын кез келді.

Қазіргі таңда латын әліпбиі үлкен беделге ие болып, қолданыс аясы да, мүмкіндігі де зор екендігін танытып отыр. Жер бетінде латын әліпбиі барлық салада қолданылатыны анық. Барлық дәрі-дәрмек атаулары, математика, физика, химия формулалары, көптеген терминдер, мамандықтарға қатысты ғылыми әдебиеттер – барлығы да латын әліпбиімен байланысты екенін байқауға болады. Латын графикасын қолданатын барлық елдердің әліпбиіндегі әріп саны тілдегі фонемалар санынан әлдеқайда аз болуы да оның жетістігі болып табылады.

Ғылыми қауым латын әліпбиіне көшкен түркі халықтарының жағдайына да жеке зерттеу жүргізіп келеді. Осы бағытта арнайы компьютерлік бағдарлама да бар.

Жазба тілдің сөйлеу тілі сияқты практикалық және теориялық тұрғыдан аса маңызды зерттеу нысаны екені мәлім. Мысалы, ақпаратты автоматты түрде өңдеу мен шартты белгілермен таңбалау (кодтау) кезеңінде сол тілдің графемаларының статистикалық сипаттамаларын білу керек болады. Графемаларды статистикалық әдіспен талдау арқылы алынған нәтижелер баспа ісін жетілдіруде және компьютерді қазақ тілін кирилден латынға ұтымды көшу 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 статистикалық әдіспен зерттеуге әрекет жасалған. Себебі, әріп – табиғи тілді жазба мүмкіндігі арқылы бейнелейтін ең бір қарапайым түрдегі көрнекі тілдік бірлік болып саналады. Зерттеу нәтижелері қазақ тілін фонологиялық және фонетикалық деңгейде тілдің морфемдік құрылымын талдау мен тілдік бірліктерді синтагматикалық тұрғыда қарастыру жағдайларында 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.

Әріптердің жиілік сөздіктерін құрастыруда қазақ тілінің үш түрлі стильдер мәтіндері зерттеу нысаны ретінде алынды. Олар:

1) М.О.Әуезовтің «Абай жолы» романы мәтіні (*51290* сөзқолданыс немесе *280812* әріп):

2) математика пәніне арналған мектеп оқулықтарының мәтіні (*19467* сөзқолданыс немесе *130388* әріп);

3) қазақ тілінің екі томдық түсіндірме сөздігінің мәтіні (*17585* сөзқолданыс немесе *116317* әріп).

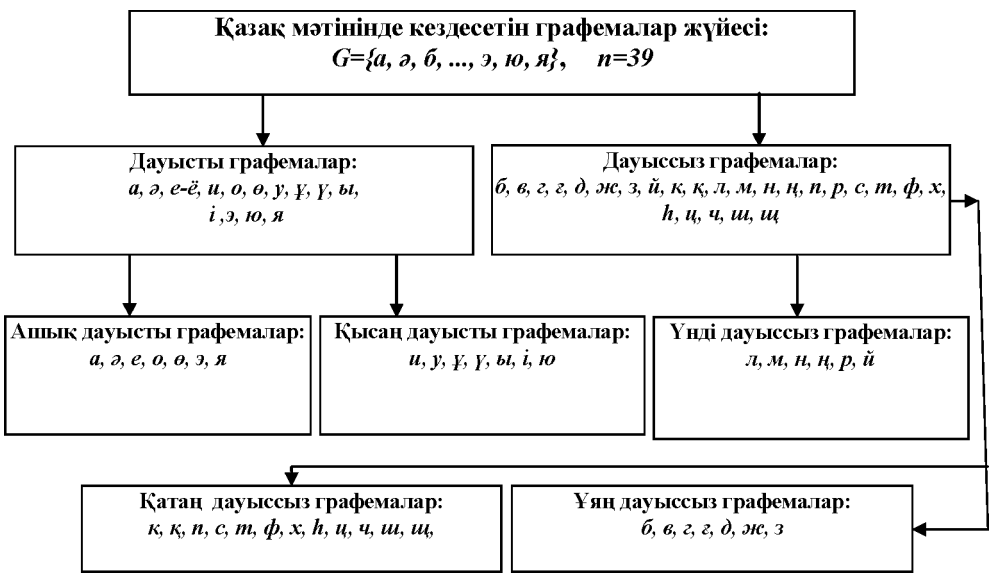
Тәжірибе түрінде алынған стильдер мәтіндерінің тең көлемді болмауы, олардың толық түрдегі шығармалар болуымен түсіндіріледі. Сондықтан мәтіндер ішіндегі тілдік бірліктердің (әріптердің) сандық сипаттары жайлы мәлімет алу үшін олардың қатынастық жиіліктері (немесе пайыздық салмақтары) пайымдалды.

Мәтін әріптерін статистикалық әдіспен талдаудан бұрын олардың графемалық құрамын анықтап алған жөн. Мысалы, егер қазақ тілі әліпбиіндегі барлық әріптерді (графемаларды) *G* жиыны деп белгілесек, жиын элементтерінің саны *n*-ге тең болады.

Бұл жиынға (*G*) мәтіндегі сөз бен сөздің арасын ажыратып тұратын «бос орын» белгісін және қос сөздер арасындағы «дефисті», «тыныс белгілерін» есепке алмағанда, әліпбидегі жіңішкелік белгі (*ь*) мен қатаңдық белгіні (*ъ*) бір графема деп есептесек, онда *n=39* деуге болады. Яғни *G={а, ә, б, ..., э, ю, я}* және жиын элементтерінің саны *n=39*.

Зерттеу барысында мәтінде кездесетін әріптер жиынтығы топ-тобымен ірілі-ұсақты бөліктерге бөлініп (шартты түрде – дауысты, дауыссыз), сызба-топтама түрінде көрініс тапты.

Сызба-топтама



Болашақ зерттеу жоспарымызда сызбада көрсетілген графемалар (әріптер) мен олардың тіркестерінің мәтінде қолданылу мүмкіншіліктерінің статистикасы, әртүрлі стильге қатысты мәтіндер бойынша және олардың сөздегі шептік орналасу тәртібіне қарай да (сөз басында, сөз ішінде, сөз соңында) зерделенеді. Ал қазіргі ұсынып отырған мақаламызда біз тек қазақ әріптерінің белгілі көлемдегі мәтін бойын қамту мүмкіндіктерінің пайызына байланысты ғана жайттардың статистикасына тоқталамыз.

Төменде ***1-кестеде*** қазақ тілінің үш стилі бойынша компьютер көмегімен түзілген қазақ әріптерінің (графемаларының) жиілік сөздігінен үзінді келтірілді. Қысқаша пайымдағанда, аталған кестеде ең жиі кездесетін графемалар олардың жиіліктерінің кему тәртібімен орналасқан. Әрі қарайғы пайымдауымызда графемалардың үш түрлі стиль бойынша түзілген жиілік сөздіктеріндегі алғашқы он орынына қатысты статистикалық мәліметтер, яғни әрбір әріптің және олардың жиынтықтарының мәтінді қамту мүмкіндіктерінің пайыздық салмақтары сөз болады. Кестедегі мәліметтерге

сүйенсек, қазақ әліпбиіндегі *39* графеманың тек 10-ы ғана әр стиль бойынша алынған мәтіндердің *64-66* пайызын қамтитынын байқаймыз. Ал көркем әдебиет (роман) стилі мен ғылыми-техникалық (математика) стильде ең жиі қолданатын төрт графема ғана (***а, е-ё, ы, н***) барлық мәтіннің *35-36* пайызын қамтиды екен. Бұл жерде әмбебап дауыстылар қатарына жататын *4*-ші "*і*" дауысты графема аталған стильдерде 5-ші және 6-шы орындарға (іргелес орындарға) ие болып тұр.

1-кестедегі сөздік мәтіні (қазақ тілінің екітомдық түсіндірме сөздігі) бойынша алынған әріптердің жиілік сөздігіндегі ең жиі қолданыстағы әріптердің алатын орны мен статистикасы жоғарыда қысқаша қарастырылған екі стиль бойынша алынған мәліметтерден белгілі дәрежеде айырым табады. Әсіресе, сөздіктегі тұйық раймен берілген етістіктердің әсерінен «*у*» әрпінің 3-орынға ие болуы және «*қ*» әрпінен басталатын қазақ сөздерінің сөздіктегі басымдылығы бірден көзге түседі. Ал жиілік сөздіктегі алғашқы ең жиі кездесетін «*у*» мен «*қ*»-дан басқа 8 графема орналасу орындарымен ғана айырым тапқанымен, мәтінді қамтудағы пайыздық салмағы жағынан айтарлықтай айырым таппайды.

Біз бұл мақалада жиілік сөздік үзіндісіндегі мәліметтермен бірге, қазақ тілінің дауысты және дауыссыз графемалар топтарының қысқаша статистикасын да беруді жөн көрдік (2-кесте және 3-кестені қара).

1-кесте

Стиль-дердегі әріптің алатын орны	Әріптердің мәтінді қамту пайызы (%)								
	Роман			Математика			Сөздік		
	Әріп аты	Әріптің %	Жиынтық %	Әріп аты	Әріптің %	Жиынтық %	Әріп аты	Әріптің %	Жиынтық %
1	а	13,2	13,2	а	11,0	11,0	а	13,0	13,0
2	е-ё	8,1	21,3	е- ё	8,3	19,3	т	7,1	20,1
3	ы	7,5	28,8	н	8,2	27,5	у	7,0	27,1
4	н	7,1	35,9	ы	7,4	34,9	е-ё	6,3	33,4
5	і	6,0	41,9	д	5,4	40,3	ы	6,1	39,5
6	т	5,0	46,9	і	5,3	45,6	р	6,0	45,5
7	р	4,8	51,7	р	5,3	50,9	л	6,0	51,5
8	д	4,6	56,3	т	5,2	56,1	с	4,5	56
9	л	4,2	60,5	л	5,1	61,2	қ	4,1	60,1
10	с	4,1	64,6	с	4,4%	65,6	н	4,1	64,2
Қосындысы:			65%			66%			64%

Сөз басында да, соңында да және ішінде де қолданыла беретін әріптерді «сөз ішіндегі» қолданыс деп шартты түрде атауды ұйғардық. Міне, осындай дауысты және дауыссыз дыбыстарды таңбалайтын графемалардың статистикасын анықтауда, олардың қатынастық жиіліктерінің стильдік (жанрлық) айырым-белгі ретінде жүрмейтіндігін ***2-ші*** және ***3-ші кестедегі*** деректерден аңғаруға болады. Әрине, сөздік мәтініндегі реестр сөздердің дыбыс құрамының статистикасы қиындасқан сөздер тізбегіндегі статистикадан аздап болса да айырым табады.

Жеке графемалардың жиілігін талдай келе, кейбір дауыстылардың және дауыссыздардың басқа графемаларға қарағанда жиі қолданатынын байқау қиын емес. Оған мысал ретінде «әмбебап» дауыстылар деп аталатын ***(а, е-ё, ы, і)*** графемаларды

айтуға болар еді. Мұндай графемалардың мәтін ішінде қолданылуының қосынды нәтижесі: романда – *34,8%;* математикада – *32,0%;* сөздікте – *29,0%.* Егер біз қарастырған мәтіндердегі барлық дауысты графемалардың қолданылу жиілігінің пайыздық қосындысы – *43,14%; 43,19%* және *44,60%* екенін ескерсек, әмбебап аталатын ***а, е-ё, ы, і*** төрт дауысты графеманың қолданылу дәрежесін өте жоғары деп санауға әбден болады.

2-кесте

Дауыссыз графемалардың топтары	Дауыссыз графемалардың мәтінді қамту пайызы (%)		
	Роман	Математика	Сөздік
Үнді	22,60%	24,60%	21,20%
Қатаң	19,95%	17,71%	23,20%
Ұяң	14,31%	14,50%	11,00%
Қосындысы:	56,86%	56,81%	55,40%

3-кесте

Дауысты графемалардың топтары	Дауысты графемалардың мәтінді қамту пайызы (%)		
	Роман	Математика	Сөздік
Ашық	25,93%	25,27%	24,40%
Қысаң	17,21%	17,92%	20,20%
Қосындысы:	43,14%	43,19%	44,60%

Дауыссыз графемалардың статистикасына байланысты айтатынымыз, қазақтың жазба және сөйлеу мәтіні сипатына үнді және жабысыңқы-шұғыл графемалардың тән болуы. Ал шұғыл (үзілмелі) графемалар көркем әдебиет мәтінінде *23,5%* тең, математика мәтінінде – *22,19* және сөздік мәтінінде *22,2* пайызға тең екен және бұндай пайыздық шамалар барлық графемалардың төрттен біріне ғана жуық деуге болады. Кейбір ***в, х, ч, һ, щ, ц, э, ю, я*** – графемалар және ***ь-ъ*** белгілері басқа тілдерден (орыс, араб-иран) енген кірме сөздерде ғана кездесуіне байланысты, олардың пайыздық салмағы бір пайызға да жетпейтін дәрежеде қолданылған.

1, 2, 3-кестелерде келтірілген жеке графемалардың статистикалық мәліметтері мен келесі мақаламызда сөз болатын қазақ тілі әліпбиінің әріп тіркестерінің статистикасы латын әліпбиіне көшу шаралары кезінде оң әсерін тигізеді деген ниеттеміз.

ӘДЕБИЕТТЕР ТІЗІМІ:

[1] Джубанов А.Х. Задача получения на ЭВМ частотных списков лингвистических единиц // В кн.: Статистика казахского текста. – Алма-Ата, 1973. – С. 263-299.

[2] Джубанов А.Х. 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 // В кн.: 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.