А.Қ. ЖҰБАНОВ, Е.Б. ЖҰБАНОВА

ҚАЗАҚ ЖАЗБА МӘТІНДЕРІНДЕГІ ГРАФЕМА ҚОЛДАНЫСЫНЫҢ СТАТИСТИКАСЫ

Зерттеу нысаны бойынша алынатын таңдама бөліктің ең аз мөлшердегі көлемін анықтау. Фонетика мен фонология түркі тілдер аясында ең көп зерттелген салалар екені белгілі. Мәселен, И.А.Батмановтың, А.Н.Кононовтың, Н.А.Баскаковтың, Қ.Қ.Жұбановтың, І.К.Кеңесбаевтың, Ж.А.Аралбаевтың және т.б. ғалымдардың еңбектерінде түркі тілдерінің фонемалық құрамы жеткілікті түрде зерттелген деуге болады. Түркі тілдерінің салыстырмалы және салыстырмалы-тарихи фонетикасы жайлы ғылыми еңбектерді де көптеп атап кетуге болады. Бірақ бұл зерттеулерде статистикалық әдіс өте сирек қолданыс тапқан. Ал шындығына келсек, қай тіл болмасын, оның фонемалар жиынтығы статистикалық әдіспен талдауға оңай көнетін тілдік бірліктердің қатарына жатады. Статистика-ықтималдықтар пәнінің әдістері арқылы алынған сандық нәтижелер лингвистикалық зерттеулерге сапалы қорытындылар жасауға мүмкіндік тудырады [1; 2; 4, 62-63; 6 және т.б.]. Бірақ кейбір зерттеулер шектеулі көлемдегі мәтін үзінділері бойынша жүргізілгендіктен, олардың нәтижелерінің сенімділік дәрежесі төмен. Сол себепті ондай зерттеулердегі алынған мәліметтерді әр уақытта репрезентативті (өкілетті) деп санай беруге болмайды. Міне, осыған байланысты нысан ретінде алынған мәтін көлемін дұрыс анықтау – аса маңызды мәселеге айналып отыр.

Математикалық статистиканың негізгі міндеттерінің бірі – нысандар жиынтығының қайсы-бір қасиеттерін бақылау негізінде **бас жиынтықтың** (генеральная совокупность) табиғаты жайлы тұжырым жасау. Егер **бас жиынтық** ретінде қазақ тілінің мәтіндері алынды десек, онда оның көлемі шексіздікке ұмтылатыны дау туғызбайтын жайт.

Математикалық статистика саласында жиынтық бірліктерінің қасиеттерін зерттеу мәселесі үш түрлі тәсіл бойынша жүзеге асады:

1) жиынтық бірліктерінің барлығын бірдей қамту арқылы (жаппай зерттеу);

2) субъективті көзқарас негізінде барлық жиынтық ішінен бірліктердің белгілі бір бөлігін таңдап алу арқылы;

3) барлық жиынтық ішінен белгілі бір бөлікгі (таңдама бөлік немесе таңдама) кездейсоқ түрде бөліп алу арқылы барлық жиынтық (**бас жиын**) бойынша қорытынды жасалады. Мұндай қорытындыға келу нақты ережелер негізінде ғана жүзеге асады.

Жоғарыда аталған бірінші тәсілді қолдану, яғни жаппай зерттеу жүргізу мәтін көлемінің шексіздігіне байланысты мүмкін бола бермейді. Ал екінші тәсілді, яғни арнайы жолмен іріктеп, жеке көзқарастағы мәтін көлемін таңдап алу әдісін қолданған жағдайда тек шектелген сипаттағы нәтижеге ғана қол жетеді. Мұндай нәтижелер барлық жиынтыққа қатысты статистикалық қорытынды жасауға мүмкіндік тудырмайды.

Сонымен, ең дұрысы, үшінші таңдама әдісті қолдану. Мұндай таңдама бөліктер (таңдамалар) бас жиынның ішкі жиыны деп те аталады және бас жиынның барлық бірліктерінің таңдама ішіне ену ықтималдығы бірдей дәрежеде болады.

Таңдамалар екі түрлі тәсілмен қалыптасады:

а) кездейсоқ сандар кестесі бойынша;

ә) жүйелі түрде іріктеп алу әдісі арқылы.

Мәселен, қазіргі қазақ жазушыларының шығармаларынан үзінділер таңдауда арнайы түзілген "кездейсоқ сандар кестесі" әдісін қолдануға болады. Ол үшін, мысалы, шығарманың бет санын көрсететін нөмірі мен әр беттегі бірінші (немесе басқа) абзац мәтінін таңдамаға енгізу үшін аталған арнайы кестені қолдану керек. Шығарма бетін таңдау арнайы кестедегі сандарға сәйкес келу шартымен жүзеге асады.

Екінші тәсіл бойынша, қажетті зерттеу нысаны ретіндегі материалды іріктеу, яғни әрбір келесі бет саны мен тиісті абзацты таңдап алу жүйелі түрде жүзеге асады. Мұндай іріктеу әдісі, көбінде, газет пен журнал бетіндегі мәтін үзінділерін таңдама бөлікке енгізу кезінде қолданылады. Нысан ретіндегі жеке мақаладағы сөйлемдер осы мақсатпен ретретімен нөмірленеді. Содан кейін барып, $K_n=K_{n+1}+L$ формуласы бойынша сөйлемдер реті таңдама бөлік ретінде алынады. Бұл өрнектегі K_n таңдамаға енетін сөйлемнің бас-тапқы нөмірі ($K_n=1,2,3\ldots$), ал L интервал (арақашықтық) шамасы, әдетте, ол *3* пен *5* сандарына тең алынады. Жүйелі түрде іріктеу кезінде K_n-ның бастапқы шамасы мен интервал мөлшері

зерт-теушінің өз еркінше алынады.

Статистикалық зерттеудегі таңдама әдісті қолданудың мақсаты – **бас жиынтықты** сипаттауға мүмкіндік тудыратын деректерді алу. Осыған орай, таңдаманың өкілділігі (репрезентативтілігі) жайында сұрақ (ақиқаттықтың қандай дәреже-сімен зерттеу нәтижесін барлық жиынтыққа қатысты сөз етуге болады? – деген сұрақ) туындауы мүмкін.

Математикалық статистика, таңдама бөліктер бойынша жүргізілген зерттеу деректерінің, **бас жиынтықтың** осындай қасиеттеріне сәйкес келуін бейнелеудің дәлдік дәрежесін анықтауға мүмкін-дік береді. Статистикалық талдаудың дәлдігі таңдаманың көлеміне, яғни зерттеуге түсетін бірлік санына айтарлықтай тәуелді болатындықтан, ең алдымен таңдаманың көлемін анықтау мәселесін шешіп алу қажет болады. Аталған мәселе лингвистикалық материалды статистикалық әдіспен зерттеуде аса маңызды деп саналады [8, 165; 9, 82; 3, 34-35; 7, 76]. Таңдама бөліктің ең кіші көлемдік мөлшерін анықтауға байланысты бірне-ше тәсілдер орын алуда [8, 165-166; 3, 33 және т.б.].

Біз өз зерттеуімізде таңдаманың ең кіші көлемін анықтау үшін мынадай өрнекті қолдандық:

$$N = \frac{z_p^2 (1 - f_{гр})}{\delta^2 \cdot f_{гр}},$$

мұндағы Z_p – алдын ала берілген ықтималдықтың сенімділік мәні бойынша алынатын тұрақты сан; $f_{гр}$ – лингвистикалық бірліктің қатынастық жиілігі; **d** – қатынастық қате [3, 34-35]. Әдетте, лингвистикалық статистикада ықтималдық мәні ретінде ***P=0,95*** және ***P=0,99*** алынады да, ал оған сәкес келетін Z_p-ның кестелік мәні ***1,64*** және ***1,96*** немесе ***2,58*** сандарына тең болып келеді [3, 33]. Ал Z_p-ның кестелік мәндері ықтималдықтар тео-риясы мен математикалық статистикаға арналған оқулықтардың қосымшасында беріледі.

Қазақ жазба мәтініндегі графемалардың статистикасы. 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 әдіс-пен талдау арқылы алынған нәтижелер баспа ісін жетілдіруде және компьютерді ұтымды пайдалану 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 бірлік-терді синтагматикалық тұрғыда қарастыру жағдайларында аса маңызды. Әріптердің, әріп тіркестерінің жалпы және шептік орналасу мүм-кіндіктерінің статистикасын анықтауда А.Х. Джу-бановтың (қазір – А.Қ.Жұбанов) "Задача получения на ЭВМ частотных списков лингвистических единиц" [5] атты мақаласында келтірілген алгоритмдік сызбасы компьютерлік программа жазуға негіз болды. Сонымен бірге, аталған автордың "К вопросу о графемной статистике казахского текста" [6] атты еңбегіндегі әріптердің статистикасы жайлы мәліметтер де осы мақаланы жазуға ықпал етті.

Жиілік сөздіктерді құрастыруда қазақ тілінің үш түрлі стильдер материалдары зерттеу нысаны ретінде алынды. Олар:

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

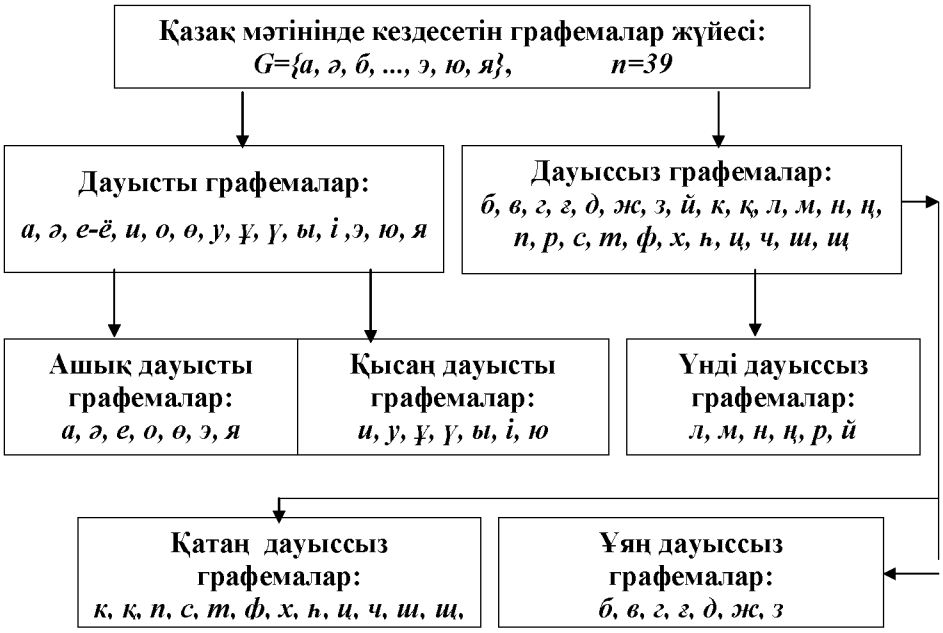
2) математика пәніне арналған мектеп оқулықтарының мәтіні (***19467*** сөзқолданыс немесе ***130388*** әріп);

3) қазақ тілінің екі томдық түсіндірме сөзді-гінің мәтіні (***17585*** сөзқолданыс немесе ***116317*** әріп)

Әрі қарайғы пайымдауымызда, бұл стильдерді қысқаша түрде – ***роман, математика, сөздік*** деп атауды жөн көрдік.

Тәжірибе түрінде алынған стильдер мәтінде-рінің 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 әліп-биіндегі барлық әріптерді (графемаларды) ***G*** жиы-ны 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.



Сызба-топтама

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), төмендегі сызба-топтама түрінде көрініс тапты.

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 үзін-ді 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 сөзде-рінің 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.

Әрине, қазақ графемаларының толық жиілік сөздігіндегі мәліметтер жеке пайымдауды қажет етеді. Ал біз бұл мақалада жиілік сөздік үзіндісіндегі мәліметтермен бірге, қазақ тілінің дауысты және дауыссыз графемалар топтарының қысқаша статистикасын да беруді жөн көріп отырмыз (2-кесте және 3-кестені қара).

Сөздің басында да, соңында да және ішінде де

1-кесте

Стильдердегі 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%

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%

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 графемалардың қолданылу дәрежесін өте жоғары деп санауға әбден болады.

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 пайыз-ға да жетпейтін дәрежеде қолданылған.

1, 2, 3-кестелер бойынша басқа да мәліметтерді сөз етуге болатыны белгілі, бірақ олар келесі мақаламыздың нысаны болмақ.

ӘДЕБИЕТ

1. *Байтанаева Д.А.* Информационные характеристики казахского текста: Автореф. дисс. ... канд. филол. наук. Алма-Ата, 1985. 17 с.

2. *Бектаев К.Б.* Статистико-инженерные методы в тюркологии // СОПИЯЛ. Чимкент, 1980. С. 7-10.

3. *Бектаев К.Б.* Статистико-информационная типология тюркского текста. Алма-Ата: Наука, 1978. 183 с.

4. *Головин Б.Н.* Язык и статистика. М.: Просвещение, 1971. 191 с.

5. *Джубанов А.Х.* Задача получения на ЭВМ частотных списков лингвистических единиц. В кн.: Статистика казахского текста. Алма-Ата, 1973. С. 263-299.

6. *Джубанов А.Х.* К вопросу о графемной статистике казахского текста. В кн.: Вопросы казахской фонетики и фонологии. Алма-Ата: Наука, 1979. С. 49-52.

Исенгельдина А.А. Факторы, определяющие относительную частотность фонем. В кн.: Статистика казахского текста. Алма-Ата, 1973. С. 659-662.

7. *Нелюбин Л.Л.* Перевод и прикладная лингвистика. М.: Высшая школа, 1983. 207 с.

8. *Пиотровский Р.Г.* Моделирование фонологических систем и методы их сравнения. М.Л.: Наука, 1966. 299 с.

9. *Пиотровский Р.Г., Бектаев К.Б., Пиотровская А.С.* Математическая лингвистика. М.: Высшая школа, 1977. 384 с.

10. *Фрумкина Р.М.* Статистические методы изучения лексики. М.: Наука, 1964. 116 с.