

А.Қ. Жұбанов

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

**ҚАЗАҚ ӘДЕБИ ТІЛІНІҢ ЖИЛІК СӨЗДІГІ БІРЛІКТЕРІНІҢ
СТАТИСТИКАЛЫҚ АРАҚАТЫНАСЫ**

Аннотация. Мақалада М.Әуезовтің 20 томдық шығармалар жинағының жиілік сөздіктері құрастырылып жатқан «Қазақ әдеби тілінің жиілік сөздігі» бірліктерінің статистикалық арақатынасы қарастырылады.

Тірек сөздер: лексикалық қор, жиілік сөздік.

А.К. Жубанов

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

**СТАТИСТИЧЕСКОЕ СООТНОШЕНИЕ ЧАСТОТНЫХ СЛОВАРНЫХ
ЕДИНИЦ КАЗАХСКОГО ЛИТЕРАТУРНОГО ЯЗЫКА**

Аннотация. В статье рассматривается статистическое соотношение единиц, входящих в состав «Частотного словаря казахского литературного языка», на основе 20-томного издания произведений М.О.Ауэзова.

Ключевые слова: лексическая база, частотный словарь.

A.K. Zhubanov

Chief researcher of the Institute of Linguistics named after A. Baitursynov,
Doctor of Philological sciences, professor
Almaty, Kazakhstan

**STATISTICAL CORRELATION OF FREQUENCY DICTIONARY UNITS
OF THE KAZAKH LITERARY LANGUAGE**

Annotation. The article considers the statistical ratio of the units included in the «Frequency Dictionary of the Kazakh literary language», on the basis of the 20-volume edition of the works of M. O. Auezov.

Keywords: lexical base, frequency dictionary.

Ішкі және сыртқы тілдік факторлардың әсерінен тіліміздегі сөздік қор бірліктері бір-бірімен өзара тығыз және күрделі түрде әрекеттеседі. Егер сөздердің тілімізде қолданыс табуы кездейсоқтық оқиға деп есептесек, оның заңды не заңсыз құбылыс екендігі тек ықтималдықтар теориясы мен математикалық статистика әдістері арқылы ғана анықталады. Сөздік қор мәселесін зерттеу барысында тілге әсер ететін қалам иесінің (автордың) әлемге деген көзқарасының барлық жақтары – әлеуметтік, тұрмыстық жағдай ерекшеліктері

және т.б. бірдей ескерілуі тиіс. Себебі олар лексикалық қорға өз таңбасын қалдырумен бірге, тілді статистикалық жолмен зерттеудің қол жеткізер нысанына айналады [1, 47].

Қазақ тілінің кейбір сандық заңдылықтар сапасының сырын ашуға көмектесетін деректерін, негізінен, зерттеуші-ғалымдар М.Әуезовтің 20 томдық шығармалар жинағының жиілік сөздігі негізінде қарастыратын болсақ, алынған деректер қазақ мәтінінің құрылымы жайлы дәстүрлі көзқарасқа да және жаңаша тұрғыдағы – статистикалық жолмен зерттеуге де көптеген мәліметтер беруде [2].

Жазушы лексиконының байлығын сөз етудің ең қарапайым жолына қысқаша тоқталайық. Егер мәтін көлемі – « N » деп белгіленсе, ал сөздік бойындағы сөзтұлға (не сөз) саны (сөздік ұзындығы) – « L » деп белгіленсе, онда $K_a = L:N$ – қатынас шамасы арқылы автордың сөздік қорының көп не аз екендігін қарапайым түрде, яғни қатынас шамасының « L » санына жақын не алшақ болуына қарай пайымдауға болады [3, 42-43]. Мәселен, $K_a = L:N$ қатынасы 1-ге тең болса, онда бөлшектің алымы бөліміне тең деген сөз, ал мұндай теңдік орындалу үшін мәтін тек қайталанбайтын сөздер тізбегінен ғана тұруы қажет. Бұл жағдай, әрине, көркем шығармалар мәтініне тән құбылыс емес. Егер де қатынастың алымы бөлімінен аз айырмашылықта болса қатынас шамасы да « 1 » санынан жуық болады, яғни автордың сөздік қоры бай, ал керісінше болса, яғни бөлшектің алымы бөлімінен көп кіші, онда шығарманың (қалам иесінің) сөздік қоры «кедей» деп пайымдалады.

Қазақ тілінің көркем әдебиет стилі («Абай жолы») романы жиілік сөздігі [4]) және ағылшын, орыс, латын тілдерінің публицистика стильдерінің сөздіктері негізінде есептелген $K_a = L:N$ қатынас шамалары *1-кестедегідей* көрініс тапты [1, 50].

Көптеген зерттеу деректері бойынша байқалатын жайт, $K_a = L:N$ қатынасының мәтін көлемінің құбылуына қарай өзгеріске ұшырауы және жуық көлемдегі мәтіндерде ондай қатынас мәндерінің де шамалас болуы. Сондықтан ескерту ретінде айтарымыз, мәселен, екі не үш жазушының шығармалар тілін зерттеу қажет болатын болса, олардан алынатын мәтіндер көлемі де бір-бірімен жуық шамада болуы шарт. Сол сияқты, олардан түзілетін сөздіктер бірліктері де бірдей дәрежеде болуы да шарт, яғни олар не сөз, не сөзтұлға түрінде болуы керек.

1-кесте

Тілдердің аналитикалық не синтетикалық сипатын

K_a коэффициенті арқылы анықтау

Тіл	Стиль	Сөздер саны	Сөзтұлға саны	K_a
Ағылшын тілі	публиц.	5197	8300	0, 62
Орыс тілі	публиц.	5225	14117	0, 37
Латын тілі	публиц.	1331916983	45211	0, 30
Қазақ тілі: «Абай жолы»	көркем әдебиет		61424	0, 28

Жоғарыда сөз болған $K_a = L:N$ коэффициенті тілдің аналитикалық не синтетикалық сипатын анықтау үшін де қолданылады. Бұл коэффициент мәні неғұрлым «бірге» жуық болса, соншалықты ол аналитикалық тілге жақын, ал егер ол «нөлге» жуық шама болса, онда ол тілдің синтетикалық сипаты арта түседі деп пайымдалады. Түрлі тілдерге тән мұндай қасиетті *1-кестеден* байқауға болады. Кесте деректері бойынша аналитикалық-аморфты ағылшын тілінде K_a коэффициентінің мәні ең үлкен шама – 0,62 пайыз, агглютинативті қазақ тілінде ең аз шама – 0,28 пайыз. Ал флективті-синтетикалық орыс, латын тілдерінде K_a мәні ағылшын тілінен аз да, ал агглютинативті қазақ тілінен көп. Сонымен, бұл коэффициентті ($K_a = L:N$) тілдердің аналитикалық (синтетикалық) сандық сипаттамасы ретінде де алуға болады деп қорытындылауға болады [1, 49-50].

Сөздік пен мәтін арақатынасын анықтаудың ең бір қолданбалылық маңызы зор түрі – жиілік сөздіктегі сөздің не сөздер тобының мәтін бойын қамту дәрежесін айқындау. Мәселен, шет тілін не ана тілін үйренуде не үйретуде минимум сөздіктер жасау қажеттігі

туатыны белгілі. Минимум сөздікке енетін сөздер сан жағынан аз болғанымен, мәтін ішінде көп қайталанып барып, оның өн бойын қамту мүмкіндігі мол. Немесе тілімізде кейбір сөздер сирек қолданылғанымен, ондай сөздерді ескермеу, мәтін мазмұнының (сөйлем мағынасының) дұрыс ашылмауына әкеліп соғуы ықтимал. Міне, осындай және басқа да қолданбалы мәні жоғары мәтін мен сөздік арасындағы қатынастар туралы кейбір зерттеулерден мәліметтер келтіре отырып, сөздердің мәтінді қамту сипатына тоқталайық.

Ол үшін М.Әуезовтің 20 томдық шығармалар жинағында қанша сөз немесе сөзтұлға бар және олар қандай жиілікте қолданылған, мәтін мен жиілік сөздіктер бірліктерінің арақатынасы қандай деген мәселелердің басын ашу мәселесін қарастырайық. Жиілік сөздіктердің статистикалық сипаттамалары тілдің теориялық және қолданбалы мәнін неғұрлым тереңірек танып-білуге мүмкіндік беретіні анық. Жиілік сөздіктерді олардың бірлігіне (сөзтұлға және сөз) қарай екі жағдайда: сөзтұлға негізіндегі жиілік сөздік және реестрлік сөз (лексема) негізіндегі жиілік сөздік ретінде қарастыруға болады. Мәтін бойындағы сөзқолданыстардан компьютер арқылы жасалатын әліпби-жиілік сөздіктің бірлігі – сөзтұлға.

Енді М.О.Әуезовтің 20 томдық шығармалар жинағындағы 2,1 млн сөзқолданыстан тұратын мәтін мен 29483 реестрлік сөзден тұратын (жалқы есімді қоспағанда) жиілік сөздігі арасындағы сандық қатынастарға қысқаша тоқталайық [2]. Мұндай сандық қатынастар **2-кестеде** көрсетілді.

2-кесте

М.О.Әуезовтің 20 томдық шығармалар жинағы
мәтін мен жиілік сөздігі арасындағы сандық қатынастар

Реестрлік сөз жиілігі (аралығы)	Жиілікке сәйкес сөз саны	Жиынтық абсолютті жиілік	Сөздікті қамту пайызы	Мәтінді қамту пайызы
1	2	3	1	2
10000- жоғары	17	311559	0,0006	18,260
5001-10000	30	208325	0,001	0,122
1001-5000	217	443576	0,007	0,260
501-1000	244	171572	0,008	0,101
401-500	120	53964	0,004	0,032
301-400	172	60588	0,006	0,035
201-300	401	97836	0,014	0,057
101-200	839	118999	0,028	0,070
51-100	1091	78077	0,037	0,046
41-50	454	20516	0,015	0,012
31-40	712	25014	0,024	0,015
21-30	1056	26429	0,036	0,896
11-20	2370	33602	0,080	0,097
10	378	3780	0,013	0,002
9	465	4185	0,016	0,002
8	512	4096	0,017	0,002
7	659	4613	0,022	0,003
6	762	4572	0,026	0,003
5	992	4960	0,034	0,003
4	1364	5456	0,046	0,003
3	2028	6084	0,069	0,004
2	3712	7424	0,126	0,004
1	10988	10988	0,373	0,006
Қосындысы:	29483	1706195		

Жиілік сөздіктегі сөздер жиілігіне сәйкес топ-топқа бөлініп, олардың тізбедегі сөздердің қанша пайызын және қайталану қабілетін ескеріп, мәтіннің (жалқы есімсіз) қанша пайызын қамтитынын аталған кестеден байқауға болады. Мысалы, жиілігі 10000-нан астам 17 сөз тізбедегі (реестрдегі) барлық сөздердің 0,0006% қамтығанымен, олар қайтала-

на келе бүкіл мәтіндегі сөзқолданыстың 18,26% құрайды. Сол сияқты, жиілігі 5001-ден 10000-ға дейінгі сөздердің саны 30-ға тең. Тізбедегі 29483 сөздің бұл 0,001% пайызы, ал барлық мәтіннің (қайталануын ескергенде) 0,122%-ы екен. Жиілігіне сәйкес басқа топтардың қамту пайыздарын осылайша 2-кестеден анықтауға болады.

Осы кестеден барлық тілдерге (қазақ тілінде де) тән заңдылық байқалады. Ол – сөздердің жиілігі кеміген сайын реестрдегі сөз санының ұлғаюы, бірақ мәтіндегі сөзқолданысты қамту пайызы барынша аз болуы. Бұл заңдылықты кестедегі жиілігі 10-нан 1-ге дейін кеміп отыратын сөздердің санынан анық байқаймыз. Мысалы, жиілігі 10-ға тең сөздердің саны сөздікте 378 (0,013%) болса, жиілігі 9-ға тең сөздердікі – 465 (0,016%). Осылай, жиілігі кеміген сайын олардың қатары арта түсетінін аңғарамыз. Мәселен, жиілігі 2-ге тең сөз саны 3712 (0,126%), ал жиілігі 1-ге тең болғанда – 10 988 (0,373%) дейін өсетіні заңды құбылыс деп есептеуіміз қажет. Мұндай төменгі жиілікте қолданған сөздер жазушы тілінің мәйегі деуге болады.

Көбіне жазушының сөз байлығын осындай сирек кездесетін сөздердің санымен мөлшерлейді. Шынында да, М.Әуезовтің сөз қолдану шеберлігі оқырмандарға белгілі, сонымен қатар бір қолданған сөзін екінші рет қайталамай, баламасын іздеп синоним, дублет сөздерді көп қолданатыны да сөз байлығын көрсететін бірден-бір дерек.

Енді 2,1 млн сөзқолданыстан тұратын М.Әуезовтің 20 томдық шығармалар жинағындағы сөз таптарының сөздік пен мәтін арасындағы салыстырмалы статистикалық деректеріне қысқаша тоқталамыз [1, 58-59].

3-кестедегі сандық көрсеткіштен мәтінде ең көп қайталанып қолданылатын сөз таптары алдымен зат есім, екінші орында – етістік, ал ең аз кездесетіндер одағай мен еліктеуіш сөздер.

Мәтіндегі сөзқолданыстың 18%-ын құрайтын жалқы есімдерді қоспағанда М.Әуезовтің 20 томдық шығармалар жинағында 29483 сөз қолданылғаны анықталды. Оның 5743-і кірме сөздер. Мұндағы кірме сөздер – негізінен орыс тіліндегі орфограммасын сақтап енген сөздер және солардан өрбіген туынды түбірлер. Бұлардың қатарында қазіргі орфографиялық нормаға сәйкес келмейтін араб-парсы сөздері де бар. Мысалы, *фар-сы, фиғыл, фарасат, фақыр, фәле* т.б. Орыс сөздері де әртүрлі фонетикалық нұсқада ұшырайды. Мысалы, *прокурор – пыркорал, протокол – пұртөкөл, генерал – жанарал, приговор – пригоуор – пір-ғауар* т.б. Кейіпкерлердің тіліндегі айтылу ыңғайына орай жазылған әртүрлі фонетикалық нұсқадағы сөздерге де тізбеде (реестрде) орын берілді. Мысалы, *ажуа* (55 рет қайталанып қолданылған), *ажуа – 22, ажусала – 7, ажым – 65, әжім – 14, ажымды – 11, әжімді – 4* т.б.

Бұл кестеден М.Әуезовтің сөз байлығының үштен екісі зат есім мен етістік сөздер екенін байқаймыз. Оның есесіне есімдік пен шылау сөздер бірігіп, жазушы тілінің жарым пайызын құрайды, алайда олардың мәтінде қайталану жиілігі өте жоғары, барлық сөзқолданыстың 18%-ына жуықтайды. Мәселен, зат есім тудыратын ең өнімді *-шы, -ші* жұрнағы 343 сөздің құрамында кездесе, етістік тудыратын *-ла// -ле, -да// -де, -та// -те* 1409 сөз жасап тұр.

Жазушы тілінде кездесетін сын есім сөздердің жартысынан көбі (3054 сөз) *-дай// -дей, -тай// -тей, -лы// -лі, -ды// -ді, -ты// -ті, -сыз// -сіз* жұрнақтары арқылы жасалған.

М.Әуезовтің шығармалар жинағында жалқы есімдер 386 729 рет қайталанып қолданылған. Бұл сөздер сөздіктің реестр қатарында жоқ, бірақ әр том бойынша және түгел 20 том ішіндегі жалқы есім жайлы мәліметтер келтірілген. Жалқы есімдер өзара сегіз түрге бөлініп, мәтінде шартты белгілермен таңбаланды. 20 томдық шығармалар жинағы бірнеше жанрда, әртүрлі стильмен жазылғаны белгілі. Осыған байланысты жалқы есімнің түрлері барлық томдарда бірдей емес. Мысалы, пьеса мен драмалық шығармаларда адам есімдері көп кездесе, әңгіме, повесть, роман мәтіндерінде ру, тайпа және жан-жануар аттары көбірек кездеседі.

3-кесте

М.Әуезовтің 20 томдық шығармалар жинағындағы
сөз таптарының сөздік пен мәтін арасындағы сандық қатынасы

Рет саны	Сөзтабы	Сөз саны	Сөздік бойындағы үлесі (%)	Сөзқолданыс саны	Мәтін бойындағы үлесі (%)
1	Зат есім (зт)	8843	37, 25	520521	30, 50
2	Етістік (ет)	6853	28, 87	534836	31, 35
3	Сын есім (сн)	5885	24, 79	189224	11, 00
4	Есімдік (ес)	209	0, 88	173495	10, 17
5	Сан есім (са)	94	0, 40	28009	1, 64
6	Үстеу (үс)	1023	4, 31	74472	4, 36
7	Шылау (шл)	162	0, 68	129536	7, 60
8	Одағай (од)	327	1, 38	8408	0, 50
9	Елктеуіш (ел)	344	1, 45	2029	0, 12
	Кірме сөздер (бн)	5743	19, 40	45265	2, 65
	Қосындысы:	29483	<i>Жалқы есім мен кірме сөздерсіз</i>	1,70619	<i>Жалқы есімсіз</i>

386729 рет қайталанған 20 томдық шығармадағы жалқы есімдер жайлы мәліметтер төмендегідей: адам есімдері – 173300 рет, жер-су аттары – 115396 рет, ругайпа аттары – 89705 рет, жан-жануар аттары – 3813 рет, баспа аттары – 3345 рет, мекеме аттары – 592 рет, қысқарған сөздер – 526 рет және атақ-данқ аттары – 12 рет қайталанып қолданылған.

Бұл мәліметтерден бірден көзге түсетіні – мәтіннің өн бойында адам аттары мен есімдік сөздердің қайталану санының (173300; 173495) жуық болуы және атақданқ, лауазым атауларының сирек қолданылып, тек мақалалар мен зерттеулерде ғана аздап кездесетіндігі.

Сонымен, М.Әуезовтің 20 томдық шығармалар жинағында қайталанып, әртүрлі сөз таптары мен атаулардың сан мөлшерінің өзіндік қыры-сыры, тілдік таным-түсінігі бар. М.Әуезовтің 20 томдық шығармалар жинағының үш түрлі жиілік сөздіктері жазушының тіл байлығы мен стильдік ерекшеліктерін тереңірек зерттеуге аса құнды-құнды материал болатындығы сөзсіз. Жеке қалам шеберлерінің сөздігімен қатар, қазақ тілінің сөз байлығын, яғни қазіргі қазақ тілінің жалпы лексикасын түгел қамтитын жиілік сөздік құрастыру да күн тәртібінде тұрған мәселе – заман қажеттілігі. М.Әуезовтің 20 томдық шығармалар жинағының жиілік сөздіктері құрастырылып жатқан «Қазақ әдеби тілінің жиілік сөздігіне» қосылатын сүбелі үлес болатыны анық.

ӘДЕБИЕТТЕР ТІЗІМІ:

- [1] Жұбанов А.Қ. Қолданбалы лингвистика: қазақ тілінің статистикасы: Оқу құралы. Алматы: Қазақ университеті, 2007. 210 б.
- [2] Бектаев Қ.Б., Жұбанов А.Қ., Мырзабеков С., Белботаев А.Б. М.О. Әуезовтің 20 томдық шығармалар текстерінің жиілік сөздіктері. Алматы-Түркістан, 1995. 346-б.
- [3] Фрумкина Р.М. Статистические методы изучения лексики. М., 1964. 115 с.
- [4] Бектаев Қ.Б., Жұбанов А.Қ., Мырзабеков С., Белботаев А.Б. М.Әуезовтің «Абай жолы» романының жиілік сөздігі. Алматы, 1979. 334-б.