Оскар Құдайбергенұлы
ЖҰБАНОВ

Қолданбалы лингвистика:
ҚАЗАҚ ТІЛІНІҢ СТАТИСТИКАСЫ

ӘЛ-ФАРАБИ атындағы
ҚАЗАҚ ҰЛТТЫҚ УНИВЕРСИТЕТІ

Асқар Құдайбергенұлы
ЖҰБАНОВ

Қолданбалы лингвистика: ҚАЗАҚ ТІЛІНІҢ СТАТИСТИКАСЫ

Оқу құралы

Алматы
"Қазақ университеті"
2004

ББК 81.2
Ж 80

Баспаға әл-Фараби атындағы Қазақ ұлттық университеті филология факультетінің Ғылыми кеңесі және Редакциялық-баспа кеңесі ұсынған

Пікір жазғандар:

филология ғылымдарының докторы ***Р.Ә. Авакова;***
филология ғылымдарының докторы ***М. Малбақов;***
физика-математика ғылымдарының докторы ***М.Ы. Тілеубергенов***

Жұбанов А.Қ.

Ж 83 Қолданбалы лингвистика: қазақ тілінің статистикасы: Оқу құралы. – Алматы: Қазақ университеті, 2004. – 209 бет.

ISBN 9965-12-753-0

Кітапта қазақ тілін зерттеуге қажетті қолданбалы лингвистика саласының статистикалық әдіс-тәсілдері мен математикалық әрнектері филолог-мамандар үшін түсінікті түрде баяндалады.

Бірінші тарауда аталған сала бойынша қазақ тілін зерттеу нәтижелеріне қысқаша ғылыми шолу жасалып, жиілік сөздіктер мәселесі және олардың тілді зерттеу ісінде пайдаланылуы, автоматты түрдегі жасалу жолдары көрсетілді. Екінші тарауда тіл зерттеудегі статистикалық әдістің бастама түсітары сипатталса, үшінші тарауда негізгі сөз таптарының қазақ мәтініндегі үлестірілуінің ықтималды-статистикалық үлгісін (моделін) құрудың проблемалық мәселелері қарастырылды. Төртінші тарауда тілдің лексика-морфологиялық құрылымына статистикалық әдісті қолданудың алғышарттары берілген. Қосымшада осы кітапта пайдаланылған салалық терминдердің қазақша-орысша сөздігі, қолданбалы тіл білімінің негізгі терминдерінің анықтамалары мен қысқаша орысша-қазақша сөздікшесі және М.Әуезовтің «Абай жолы» роман-эпопеясы бойынша түзілген жиілік сөздіктің ең жиі қолданылған 500 сөзі көрініс тапты.

Кітап қолданбалы лингвистика мамандығын даярлайтын жоғары оқу орындарының студенттеріне және математикалық, құрылымдық тіл білімі мамандары мен барша филолог-ғалымдарға арналған.

ЖК $\frac{4603000000\text{-}060}{460(05)\text{-}04}$

ББК 81.2

ISBN 9965-12-753-0

АЛҒЫ СӨЗ

Оқырмандар назарына ұсынылып отырған бұл еңбек қолданбалы лингвистиканың бір тармағы болып саналатын статистикалық лингвистиканың кейбір өзекті мәселелерін сөз етеді. Автор аса көрнекті ғалым Бодуэн де Куртенэнің: «Тіл білімінде сандық, математикалық ойлауды жиі қолданып, оны мейлінше нақты ғылымдарға жақындату керек» деген ұстанымын басшылыққа алады [22]. Сондай-ақ Р.М.Фрумкинаның «Статистические методы при заведомо неполной информации об объектах исследования предоставляют возможность делать выводы об этих объектах с заданной точностью и надежностью» [91] – деген пікірімен толық келісе отырып, тілге қатысты көптеген тұжырымға статистикалық ғылыми баға беріп, олардың дұрыс-бұрыстығын тексеруге болатындығына арнайы тоқталады.

Ықтималды-статистикалық заңдылықтар табиғат пен қоғамда кездесетін әр алуан құбылыстардың өзара қарым-қатынастарында көрініс табатыны белгілі. Осындай көрініс «жиілікпен» өлшенеді десек, артық айтқандық болмас. Ал тілдік құбылыстың жиілігі – әмбебап тілдік категория. Бұл жерде тіл зерттеу барысында тек математикалық, дәлірек айтсақ, статистикалық әдіс-тәсіл мен математикалық орнекті (формуланы)

---

*Бұл оқу құралына автордың орыс тілінде жарық көрген «Квантитативная структура казахского текста» атты монографиясы негіз болды [2]. Бірақ аталған құрал осы еңбектің тікелей аудармасы емес. Өйткені оқу құралы мен ғылыми монографияның құрылымдары мен мақсаты әр түрлі және 1987–2004 жылдар аралығындағы ғылыми жаңалықтар ескеріліп, оқу құралына тиісті толықтырулар енгізілді.

қолдану туралы ғана сөз болып отырған жоқ, сонымен бірге сандық сипат белгілі бір тілдің ішкі табиғатына объективті түрде тән деп түсінген жөн. Шынында да, бүгінгі таңда тіл құрылымындағы сапалық және сандық сипаттар іштей тығыз байланыста болатыны ешқандай күмән туғызбайды. Мәселен, кейбір беделді, ірі ғалымдар өз тұжырымының дұрыстығына толықканды дәлелдер келтіруі мүмкін. Осыдан тіл білімінде не басқа ғылым салаларында бірін-бірі терістейтін теориялар мен болжамдар (гипотезалар), тіпті ғылыми мектептер пайда болады. Міне, сондықтан тілді статистикалық тәсілдермен зерттеу тілші-мамандардың субъективті тұжырымдарын не объективті шындыққа айналдыруға көмектеседі, немесе оның теріс екендігін дәлелдеп береді.

Кіған тілдің сөздік қорын зерттеуді мақсат еткен филолог-студент, филолог-мұғалім, филолог-зерттеуші оқырмандарды ең қарапайым түрдегі статистикалық әдістермен және оның қарапайым есеп-қисап жүргізу жолдарымен таныстыруды көздейді.

Сөздік қор – аса күрделі тілдік нысан, бірақ көптеген теориялық және практикалық мәселелер үшін оның кейбір қасиеттерін, яғни ең негізгі деген сандық сипаттарын білу жеткілікті. Мәселен, кейбір жағдайда белгілі көлемдегі мәтін мен оның сөздігі арасындағы жоғары ықтималдыққа ие қатынастарды білсек те жеткілікті екен. Мәтін көлемінің өзгеруіне қарай зерттеу нысанына алынған тілдік бірлік (сөз не сөз тіркес) жиіліктерінің мәтін мен сөздік бойында таралу (үлестірілу) сипаты қандай, алынған сандық нәтижелерді ана тілін, не шет тілін үйретуде қалай қолдануға болады – деген сауалдарға жауап алуға мүмкіндік береді. Сонымен қатар тілдік ақпаратты тарату, ғылыми ақпараттарды автоматты түрде оңдеу, ұзақ мерзімге сақтау, оларды қажеттігіне қарай іздеп, тауып алу мүмкіндігін анықтау да сөздік қордың статистикасын зерттеу арқылы іске асады.

Қолыңыздағы оқу құралына қолданбалы лингвистика саласының «статистикалық тіл білімі» негізгі нысан болуының өзіндік себептері де бар.

Кезінде Бүкілодақ бойынша «Тіл статистикасы» 1957 жылдан зерттеле бастапса, филология ғылымдарының докторы,

математик Қ.Б.Бектаев жетекшілік еткен қазақстандық ғылыми шығармашылық топ зерттеу жұмыстарын 1970 жылдан бастап жүргізе бастағаны ғылыми көпшілікке мәлім. ҚазССР ҒА Тіл білімі институтындағы «Статолингвистикалық зерттеу және автоматтандыру» деп аталатын ғылыми шығармашылық топ қазақ әдебиетіндегі белгілі ақын-жазушылар шығармалары мен басқа да стильдерге қатысты мәтіндер тілін статистикалық әдіспен зерттеді. Содан бері осы сала бойынша едәуір ғылыми еңбек жарық көрді. Қазақстандық статолингвист ғалымдардың жұмысына қысқаша шолу жасаумен қатар біз өз зерттеу тәжірибемізді де оқырмандарға таныстыруды мақсат етеміз.

Тілдегі заңдылық жүйеге ғана тән болғандықтан, статистикалық заңдылықтарды ашуға байланысты қолданылатын әдістер кезінде тілдің өзінше бір бөлек жүйе құратындығы туралы сөз ете бермейміз. Басқаша айтсақ, тіл – біріне-бірі қатыссыз құбылыстар жиынтығы емес, керісінше, заңды түрде үйымдасқан, көпөлшемді және тәуелсіз сипаттағы ақиқат.

Тілдік бірліктердің (дыбыс, морфема, сөз тіркестері, сөйлем) мәтін ішінде (жазба тіл ағымында) кездесуі кездейсоқ оқиға деп есептесек, онда белгілі бір жүйеден туындайтын тілдің өзі де (яғни мәтін) кездейсоқ процеске қатысты болады. Осының салдарынан зерттеу нысанына алынатын нақты бір мәтін белгілі дәрежеде ақиқат сөйлеу тіліне жақын келетін модельденуші мәтінге айналады. Мұндай жағдайда мәтін қасиеттері тұрақты және элементтері бір-біріне тәуелсіз деген қағидатты ұстану қажет болады.

Лингвистикалық статистикада тілдік заңдылықтарды анықтау барысы тәжірибелік байқау мен сол байқаудың негізінде туындайтын нәтижелерді математикалық жолмен өңдеу арқылы іске асады. Сондықтан лингвостатистикалық зерттеулер жаратылыстану ғылымдарындағыдай табиғи және ғылыми түрғыда жүргізіледі.

Статистикалық деректерді өңдеуде компьютерлік технологияны пайдалану, тілдік мәселелердің шешімін тез арада табу, математикалық статистика әдістерін басқа ғылымдар тәсілдерімен (ақпараттық теория, математикалық логика және т.б.) ұштастыру қазіргі кездегі отандық және шетелдік тіл білімінің даму бағытына сай келеді.

Түркі мәтіндерін статистикалық тәсілдермен зерттеу, ғалымдардың пайымдауынша, бұл тілдердің ішкі табиғатына, яғни олардың жалғамалық (агглютинативтік) құрылымына негізделеді.

Тіл зерттеу тәжірибесінде сандық зерттеулердің қажеттігін математикалық лингвистика мамандарымен қатар дәстүрлі тіл білімінің аса көрнекті өкілдері де мойындап келеді. Олардың пікірінше, жазба әдебиет пен сөйлеу тіліндегі түрлі сөздердің әр ыңғайдағы қолданылу жиілігін анықтау тәсілі тілдің құрылымды-грамматикалық, кейде стильді-семантикалық айырымдарын ажыратумен қатар стильдерге тән қасиеттерді де анықтауға көмектеседі.

Қорыта айтқанда, барлық грамматикалық категорияларды сандық тәсілмен талдау олардың әдеби тілдегі функционалдық салмағын нақты көрсетуге мүмкіндік береді.

Осы оқу құралының қолжазбасымен танысып, құнды-құнды пікір айтқан әл-Фараби атындағы Қазақ ұлттық университеті жалпы тіл білімі кафедрасының меңгерушісі профессор Э.Д.Сүлейменоваға және осы кафедраның профессор-оқытушы құрамына автор өз ризашылығын білдіреді.

КІРІСПЕ

Тілдің қолданбалы аясы бұрыннан да өзінің кең және жан-жақтылығымен ерекшеленетін. Оның ескіден келе жатқан саласы – жазу (графика), ана тілі мен шет тілін оқыту әдістемесі және тілдің лексикографиялық жүйесі. Осылардан кейін барып – аударма, шифртану (дешифровка), орфография, транслитерация және терминологияны өңдеу. Қолданбалы лингвистиканың дәстүрлі бағыттарының бірі – мемлекеттің тіл саясатына қатынасу:

1) әліпбиді, орфографияны өңдеу, сауатсыздықты жою;
2) мемлекеттік тілді тағайындау;
3) мемлекеттік тілден басқа тілдердің орнын анықтау;
4) ұлттық терминологияны өңдеу, бір ізге келтіру және қалыптастыру;
5) қала, көше, алаңдарға ат қою мен оның бұрыннан келе жатқан атын өзгерту, яғни ономастика мәселелері.

Қолданбалы тіл білімнің осындай классикалық бағыт-бағдарын дамыту және мейлінше жетілдіру мәселелерімен бірге XX ғасырдың екінші жартысынан бастап оның бірқатар жаңа бағыттары орын ала бастады. Бұл бағыттар заманға сай қоғамдық, жаратылыстану және техникалық ғылымдардың өзара даму сипатынан туындайтын қолданбалы лингвистиканың тарихи дамуының логикалық жалғасының көрінісі еді.

Негізінен алғанда, адам баласының қызметінің әр түрлі аясындағы жұмыстарын жеңілдету бір ғана проблемалық мәселеге тіреледі. Ол – қоғам өміріндегі ақпарат жұмысын өңдеу. Бұндай ақпарат мәтін түріндегі жазба тілінде не үйреншікті сөйлеу тілінде болуы мүмкін.

Қоғамның практикалық қажеттігін өтеуде, әсіресе, ақпаратты сақтау мен оны тарату істерінің сырын ашуда зерттеуші оның ең алдымен тілдің ішкі табиғатындағы заңдылықтардан іздегені жөн.

Осымен байланысты өзекті мәселенің бірі – оператордың дауысымен басқарылатын станок пен құрал-аспаптарды өндіру. Ол үшін әр түрлі тілдердегі ақпараттарды автоматты түрде өңдеу мен оларды іздестіру, байланыс торабы жұмысын (телефон, радиобайланыс т.б.) жетілдіру, адам баласының сөйлеу мен есту қабілетінің бұзылуына қатысты ауруларды емдеуде дыбыстың фонетикалық деректеріне жүгіну және т.б.

Аталған проблемалық мәселелерді ЭЕМ-ды (компьютерді) кең түрде қолдана отырып, мәтінді (жазба, сөйлеу) автоматты түрде өңдеу қажеттілігі туындайды. Мәселен, әр типті ақпараттық жүйелерді тілдік деректермен қамтамасыз ету; машиналық аударма мәселесі; табиғи тілді түсінетін жүйе құру (жасанды интеллект жүйе-сіндегі тілдік мәселелер); сөйлеу кезіндегі дыбыстық сигналдарда бар ақпараттарды пайдалану негізінде арнайы жүйе құру т.б.

Соңғы жылдары әр түрлі ақпараттық жүйелерді тілдік деректермен қамтамасыз етумен байланысты терминтану мәселесін бірізділікке түсіріп, тұрақтандыру өзекті мәселеге айналып отыр. Себебі, әр білім саласының ғылыми және техникалық терминдерді қолдану қажеттігі барынша сұраныс тудыруда. Орине, мұндай мәселе лексикография саласынан тыс өз шешімін таппайтыны белгілі. Сондықтан қоғам өмірінде ЭЕМ-нің (компьютердің) кең қолдануымен тығыз байланысты автоматты лексикография саласы біртіндеп өз отауын құруда.

Практикалық мәселелерді шешуге байланысты табиғи тілде диалогты іске асыру үшін адам мен ЭЕМ арасында әсерлі қарым-қатынас орнату қажет. Мәселен, оларға жататындар: сұрақ-жауап жүйесін құру, роботтарды басқару жүйесін құру, аталған басқару жүйелерінде дұрыс шешім қабылдау үшін диалогтық процесті іске қосу. Мұндағы ең негізгі мәселе – автоматты құрылғының жазба не дыбыстық мәтінді түсіне білу (тану) проблемасы.

Әрбір тілдің қолданбалы мәселені шешуде өзіне тән ерекшелігі бар. Мұндай есептердің санын, түрін алдын ала білу

мүмкін емес. Өмірде олар жыл сайын, ай сайын ауысып, біреуі келіп, екіншісі кетіп жатады. Қолданбалы лингвистикаға қатысты іргелі зерттеулер ғылыми-техникалық, ұйымдастыру-басқарудың ауызша-жазбаша құжаттарының, сөздік түзу мен және т.б. түрлерінің фонетикалық, грамматикалық, семантикалық және статистикалық құрылымын сипаттау мен модельдеу жақтарын қамтиды. Жекелеп айтсақ, оған теориялық және қолданбалы лингвистиканың шекарасында жатқан мәтіндік бірліктердің формальды моделін (үлгісін) құру проблемасын жатқызуға болады.

Теориялық лингвистика, негізінен, тілді қалыптық, жүйелік тұрғыда қарастыратындықтан, қолданбалы тіл білімі тілді әрекет үстінде, яғни оның қарым-қатынас (коммуникация) кезіндегі табиғатын түсінуге тырысады. Көптен бері қолданбалы тіл білімі ғылыми-техникалық пен іскерлік прозадан өзін алшақ ұстағаны аян. Дегенмен, XX г. 70-жылдарында ғалымдардың мынадай шешім жасауына тура келді. Көптеген қолданбалы мәселелер таза лингвистикалық емес, тіпті дәлірек айтсақ, ол проблеманы шешу адамның іс-әрекеті мен ойлау процесін, тілдің семантикасы мен оның формальды және семантикалық әдіс-амалын синтездеу (жинақтау) жолдарын модельдеуге тікелей қатысты. Осыдан барып қолданбалы тұрғыдағы зерттеулердің іргелі проблемасы – білімді модельдеу (моделирование знаний) мәселесі анықталды.

Соңғы кезде бұл проблема (білімді модельдеу) бірнеше ғылымдар тоғысына, дәлірек айтқанда – логика, лингвистика, математика, психология, кибернетика салаларының аясына қатысты болып отыр. Көптеген ғылымдардың осы мәселеге қызығушылығының негізгі себебі біреу-ақ, ол – автоматты жасанды интеллект жүйесін құру. Солай бола тұра, «білім» бізге «тіл» арқылы берілетіні анық. «Білім» сөйлеу мәтінінде де (монолог, диалог, әр жағдайдағы реплика) және сонымен бірге жазба мәтіндерде де (ескі жазба ескерткіші, көркем әдебиет, ғылыми-техникалық әдебиет) көрініс табады. Тіл арқылы «білімді» біз ұрпақтан-ұрпаққа жеткіземіз. Сондықтан «тіл» - «білімді» сақтау формасы және оны тарату құралы десек те болады. Біздіңше, ғылым мен техникада мәтіннен тыс жатқан «білім» жоқ деуге болады, ал ғылыми-техникалық мәтіннің

семантикасын модельдеу дегеніміз – ол осы салалардың білім жүйесін модельдеу. Осылайша, біртіндеп келіп, білімді модельдеу проблемасы мәтін мазмұнын (мағынасын) модельдеумен ұштасып жатқанына көз жеткізуге болады. Бұл жерде ең негізгі шешімін табуды қажет ететін мәселе – мәтіннің семантикалық көрінісін (семантическое представление) құру.

Дегенмен, техникалық және іскерлік коммуникация аясындағы нақты қолданбалы мәселені шешу сол саладағы құжаттық мәтіндерінің грамматикалық, лексикалық, семантикалық құрылымының сипатталуына байланысты. Сонымен бірге, аталған салалардың терминологиялық сөздігінің болуына, мәтіннің статистикалық құрылымының зерттелуіне және осындай мәтіндер типтерінің семантикалық көрінісінің толық құрылуына да байланысты екенін айта кету қажет.

Қолданбалы лингвистиканың дамуына, яғни ғылыми жетістіктерге ие болуы тіл білімі теориясының дамуы да өз әсерін тигізбей қоймады. Мысалы, XX ғ. 20–30-жылдары практикалық қажеттіктен туындаған ғылыми-техникалық терминологияны бірізді және тұрақты ету жолдары жаңа лингвистикалық пән – «терминтану» ("терминоведение") саласын өмірге әкеледі. Сол сияқты, Кеңес Одағы халықтары тілінің әліпбиі мен жазуын құруға байланысты 30–40-жылдары жүргізілген кең көлемді әрі орасан күрделі, әрі зор тәжірибелік мәні бар жұмыстар тілдерді синхронды сипаттаудың әдістерін жетілдіруге себепші болды (анықталандырды) деуге болады.

Соңғы 20–30 жылда пайда болған қолданбалы лингвистиканың жаңа аспектілерінің бәрі бір ғана ортақ проблемаға қатысты екенін байқатты, ол – тілдің жазба не сөйлеу түрлерін автоматты өңдеу мәселесі еді. Мұндай проблемамен айналысу тілді талдау мен сипаттаудың жаңа әдіс-тәсілдерінің дамуына және тіл табиғаты мен тіл білімі құрылымына деген жаңаша көзқарастың пайда болуына мүмкіндік туғызды.

XX ғасырдың 50–70-жылдарында тілдік материалдарды автоматты түрде өңдеуге қатысты күрделі проблеманың шешім табуы қолданбалы лингвистиканың әрі қарай дамуына айтарлықтай әсер етті.

Қолданбалы лингвистиканың жаңа аспектілері теориялық тіл біліміне жаңадан ғана ене бастаған математикалық әдістерді,

әсіресе, теоретика-жиындық, формальды-логикалық, статистика-ықтималдық әдістердің қолдануын барынша жеделдетті.

Классикалық тіл білімінің, классикалық логиканың, психологияның, семантика мен математиканың қиылысуынан барып тілдегі модельдеу әдісінің өзінше бөлек ғылым ретінде көрініс табуының нәтижесінде қазіргі «құрылымдық лингвистика» деп аталып жүрген ғылыми сала өмірге келгені белгілі. Сол сияқты енді тіл біліміндегі дербес түрған теориялық бағыт ретінде «тілдің теориялық модельдері» атты жаңа тарау пайда болды.

Лингвистика мен математика салаларының аралығынан туындаған – «математикалық лингвистика» атты жаңа пән өмірге келді. Бұл пәннің XX ғ. 50–60-жылдары қалыптасуы бұрыннан да тіл білімінің барлық ішкі дамуының өзінен де сезіле бастаған болатын. Мәселен, бұл жайт тілдің құрылымына көңіл аударудан, тілдік жүйе ретінде қарастырудан және оның микрожүйелерден тұру күрделілігін жете түсінуден, лингвистикалық нысандар мен олардың атрибуттары арасындағы қатынастарды танып-білуден байқалып, математикалық лингвистика пәнінің дамуына объективті жағдай туғызды деуге болады.

Атап айтқанда, қазіргі теориялық тіл білімінің дамуына ең көп әсер еткен – құрылымдық лингвистика мен математикалық лингвистика пәндері. Бұндай әсер көбінде грамматиканың синтаксис және семантика салаларында айтарлықтай көрініс тапты.

Құрылымдық синтаксисте мынадай формальды екі синтаксистік модельдер жүйелі түрде жете зерттелді. Олар – тікелей құрастырушылар моделі мен өзара тәуелді модельдер деп аталады. Бұл аталған екі модель де машиналық аударма жасауда, автоматтанған синтаксистік талдау мен мәтінге жасалатын басқа да автоматты өңдеулерде кең түрде пайдаланылады.

Семантика тілдің барлық деңгейлеріне ортақ болғандықтан, тіл білімінде оны бөлек деңгей ретінде қарастырмайды. Семантикада, біріншіден, лингвистиканың өзіне ғана тән әдістің әсері байқалса, екіншіден, логикалық семантика мен теориялық классификациялау әдістерінің де әсері орын алады. Синтак-

ақпараттық семантикада сан жағынан көбірек зерттелген жайттар – қолданылмайтын семантикалық құрылымы жайлы түжырымдар (концепциялар).

Қоғамдық ғылымдарда ғылыми зерттеулерді автоматтандыру, ең алдымен, құжаттық және фактографикалық деректердің ауқымды қорын құру барысында қажет болды. Себебі, мұндай автоматтандыру типі ең алдымен қоғамдық ғылымдар мамандарының өз қалауынан туындап отыр. Мәселен, ғылым жолындағы қандай анықтамалар түрі қоғамдық саладағы қызметкерге қажеттірек? Біріншіден, мұндай сұраныс кітапханалардағы библиографиялық істе: белгілі бір мәселеге қатысты барлық отандық не шетелдік әдебиеттерді іздестіріп, тауып алу сияқты мәселелер. Мұндай материалдарға жатады: кітаптар, мақалалар, конференция тезистері, хроникалық шағын мақалалар және т.б.

Екінші кезекте, жиналған әдебиеттер ішінен қажетті деректерді жинастыру, оларды қажетті белгілеріне қарай сұрыптау, топтастыру және т.б. жұмыстар орындалуы керек.

Көптеген қоғамдық ғылымдарға қатысты мәселелер, әсіресе, тіл білімінде, этнографияда, антропологияда, тарихта карталық мәліметтерге тікелей қатысты болып келеді. Сондықтан, ЭЕМ-ге ондай деректерді енгізу мен қағаз бетіне шығару өте-мөте өзекті деп саналады.

Қоғамдық ғылымдар бойынша ғылыми-зерттеу жұмысын автоматтандыруды қажет ететін мәселелер:

а) әр тілдер бойынша белгілі тақырыпқа қатысты әдебиеттерді іздеп, табу;

ә) жазба ескерткіштер мен мәтіндік материалдар қоры бойынша таңдама жұмыстарын атқару;

б) табылған материалдар арқылы қажетті амал-әрекеттерді іске асыру;

в) материалды жан-жақты сұрыптау;

г) алдын ала белгілі өлшемдер бойынша реестрлер, каталогтар сипаттамаларын жинақтау;

д) статистикалық, картографикалық, сұрыптау теориясы мен жүйелік талдау әдістерін қолдану;

е) берілістерді (деректерді) сызба, сурет, карта түрінде көрсету.

Сонымен, ғылыми жұмысты автоматтандыру процесінде құрастырылған жүйе негізінде тұтынушы маманға қажетті тарихи деңгейде зерттеу нысанына қатысты барынша толық энциклопедиялық білім және библиографиялық мәліметтер берілуі қажет деп санаймыз.

Жоғарыда сөз болған мәселелерді қорыта келе айтатын болсақ, қолданбалы лингвистика, ең алдымен, кешенді пән. Себебі, ол философия, психология, физиология, математика, логика, әлеуметтану, информатика салаларымен үнемі тығыз қатынаста болады.

Жаңаша мағынадағы қолданбалы лингвистиканы жеке қарастыратын болсақ, ол көптармақты ғылыми сала болып табылады. Ғылыми әдебиеттерде оны «компьютерлік лингвистика», «инженерлік лингвистика», «автоматты лингвистика», «есептеу лингвистикасы», «квантитативті лингвистика», «статистикалық лингвистика» деген жарыспалы терминдермен атап жүр. Әрине, бұл атауларға сәйкес салалардың өздеріне тән бағыт-бағдары ажыратылады, алайда бәріне тән ортақ мақсат – адам баласының қарым-қатынасы үшін қажетті табиғи тіл қызметінің ең ыңғайлы, ең тиімді жолдарын іздестіру.

Қолданбалы лингвистика салаларының барлық тілдерде, соның ішінде қазақ тілінің жазба түріне қатысты зерттеулерде ең көп қолданыс тапқан тармағы – статистикалық лингвистика. Дәстүрлі әдіспен тілді зерттейтін тілші-ғалымдардың бәрі бірдей статистикалық әдіс-тәсілдерді қолдай бермейді, бірақ өз ізденістерінде осы саланың *көп, аз, мол, жиі, сирек, өнімді, өнімсіз* тәрізді терминдерін барынша пайдаланады және олардың статистикалық ғылым саласына қатыстылығын көп жағдайда аңғара бермейді.

Статистикалық лингвистика тілді зерттеудің тек әдісі не тәсілі болып қана қалмай, қазіргі кездегі тілтануда өзінше бір бөлек ғылыми пән дәрежесіне көтеріліп отыр. Бұл пән тілдік бірліктердің сапалық және сандық (мөлшерлік) мәліметтерін тілдің табиғатына сай қарастыруды жөн санайды.

Статистикалық лингвистика, негізінен, математикалық статистика әдістемесіне сүйенеді. Зерттеу барысында ондай әдістемені қолдану қатынас құралы ретіндегі «тілді» белгілі бір жүйе деп қарастырудан туындайды. Шынында да, «тіл»

дегеніміз ақиқат болмыста өмір сүретін көп өлшемді және белгілі бір ішкі заңдылықтардың негізінде топталған (реттелген) табиғи құбылыс деуге болады. Ал тілдік жүйе математикалық статистика заңдылықтарына бағыну үшін ондағы бірліктер тобы қайталанып отыратын және кездейсоқтық сипатта болатын тілдік элементтерден тұруы қажет. Бұл жерде айта кететін жайт, тіл қызметі кезінде (яғни тілдік қарым-қатынас кезінде) дыбыстар, сөздер, сөз тіркестері, сөйлемдер және одан да үлкен бөліктер қайталанбайтын сипатта болса, адамдар арасындағы тілдік қатынас бұзылып, ақпарат таратушы мен қабылдаушы арасында түсінбестік пайда болған болар еді. Сондықтан да тілдік бірліктердің қайталануы олардың тілімізге тән болуынан, олардың қолдану сипатына *жиі, сирек* деген ұғымдардың да тән екендігінен туындайды.

Табиғи ортадағы элементтер арасындағы қайталану сипатының негізінде математикалық статистика мен ықтималдық теория атты ғылым салаларында тілдік бірліктердің статистикалық заңдылықтары қалыптасады. Зерттеуге алынған бірліктің (единица, элемент) ондай заңдылыққа бағыну не бағынбауын айқындау үшін алынатын ортаның аумағы (мәселен мәтін көлемі) барынша мол болғаны жөн. Мысалы, кітаптың бір бетінде кездесетін сөздер мен оның жүз бетінде кездесетін сөздердің қайталану сипаты әр түрлі болатындығына ешкімнің күмәны жоқ. Ал заңдылықты айқындау үшін, кітаптың бір бетіндегі мәліметтен гөрі оның көптеген беттерінен алынған мәліметтердің шамасы шындыққа жақын болатындығы белгілі.

Тілге тән заңдылықтарды статистика тәсілімен анықтау негіздемесі мынада:

1) сандық (мөлшерлік) құбылыс тіл табиғатына әуелден-ақ тән болуы;

2) тіл құрылымындағы сандық және сапалық сипаттардың өзара байланыста болуы;

3) тілдің әр түрлі бірліктері сөйлеу ағымында статистикалық заңдылықтардың ең болмағанда біреуіне бағынуы тиіс деп үйғару керектігі.

Әдістердің индуктивті, дедуктивті болып ажыратылатынын ескере отырып, негізінен алғанда, статистикалық әдіс индуктивті әдіс арқылы, ал тілдің әр түрлі модельдерін жасау үшін

қолданылатын логика-математикалық әдістер – дедуктивті әдіс арқылы жүзеге асады.

Тілдің ықтималды-статистикалық моделін жасауға болатыны белгілі десек, ондай модель, біріншіден, белгілі бір тілдік жүйенің мәтіндері арқылы жасалады, ал екіншіден, тілдің логика-математикалық моделі сол тілдің толық жүйесін айқындай алады.

Сөйлесу арқылы қатынас жасау кезіндегі тілдік процесті ***байланыс өзегі*** (канал связи) не ***байланыс жолы*** деп есептеуге болады. Ал мұндай байланыс жолы арқылы таратылатын ақпарат әріп, дыбыс, морфема және т.б. лингвистикалық бірліктер негізінде іске асады. Бұл жердегі лингвистикалық бірліктер белгілі бір «кодтың» символдары ретінде есептеледі. Байланыс өзегіндегі хабарлаушы мен қабылдаушы, яғни айтушы мен тыңдаушы өзара бірдей «кодты» пайдаланулары шарт болуы қажет.

Тіл зерттеу тәжірибесінде статистиканың араласуы бұрыннан да бар десек, ондай әдіс тек жекелеген зерттеушілердің ізденістерінде ғана кездессе, қазір статистикалық лингвистика жеке ғылыми пәнге, ғылыми бағытқа айналып отыр. Осымен байланысты статистикалық лингвистика саласының қазақ тіліне қатысты өзіне тән алға қойған теориялық және практикалық мақсаты мен міндеті әлі де айқындалуда.

Қай ғылым саласын алсаңыз да онда қарастырылатын мәліметтер дәл және объективті болуы шарт, ал қоғамдық ғылымдар ішіндегі осындай дәлдікті көбірек қажет ететін сала – лингвистика.

Тіл зерттеу тәжірибесіне статистикалық әдіс-тәсілді ең алғаш ұсынған ғалым орыстың көрнекті математигі В.Я.Буняковский 1847 жылдың өзінде-ақ ықтималдық жіктеудің мүмкіншіліктерін аса бір білгірлікпен (көрегендікпен) атап, ол істің сәтті болуы үшін филологтар мен математика мамандарының бірлескен одағының қажеттігін айтқан болатын [86].

Тіл саласындағы кез келген статистикалық зерттеуде мынадай үш мәселенің басын ашып алған жөн:

1) Тілдің қай бірлігін, нені санау (есептеу) керек?

2) Таңдалып алынған бірліктерді неге санау (есептеу) керек?

3) Тілдік бірліктерді қалай санау (есептеу) керек?

Яғни зерттеушінің: ***«нені?», «неге?», «қалай?»*** деген үш сұраққа жауабы әр уақытта дайын болғаны жөн.

Бірінші сұрақтың жауабы зерттеушінің алға қойған мақсатына тікелей байланысты. Ең алдымен есептелуге тиісті тілдік бірлік анықталуы қажет және оның тілдегі сапалық сипаты, яғни ол жайлы тілдік мәліметтер жеткілікті түрде жиналуы керек. Саналуға тиісті әрбір бірліктің толық анықтамасы алдын ала айқын болуы талап етіледі. Мәселен, мәтін бойынан: сөйлем, сөзтіркес, сөзқолданыс, сөзформа, сөз, буын, әріп, тыныс белгілер және «бос аралық» (пробел) сияқты бірліктер ішінен қайсысы таңдалып алынса, солардың айырым белгілері, анықтамасы алдын ала белгілі болуы шарт.

Екінші сұрақтың маңыздылығы бірінші сұрақтан кем емес. Статистикалық деректері белгілі тілдік элементтердің бәріне бірдей лингвистикалық түсініктеме беру мүмкін бе, яғни олар лингвистикалық мағынаға ие ме, жоқ па деген сұрақтар туындайды. Сондықтан ***неге санаймыз*** деген сауалға дұрыс жауап іздеу үшін біз болашақ алынатын сандық мәліметтердің тілдік тұрғыдан түсінігі болу-болмауын алдын ала ескерілуі керек.

Әрине, ***неге санаймыз*** деген сұраққа әр кезде бірдей жауап беру көбінесе алға қойған мақсатқа байланысты, сондықтан әр мәселеге өзіне тән жауап іздестіру қажет болады.

Жоғарыда аталған бірінші және екінші сұрақтарға (***не себепті санаймыз)*** жауап беру қажеттігіне осы оқу құралының «БІРІНШІ ТАРАУЫНДА» арнайы тоқталмақпыз.

Енді үшінші – ***«қалай санау»*** сұрағына жауап беру үшін біз алдымен мектепте өткен арифметиканы еске түсіре отыра, ***ықтималдық теориясы және математикалық статистика*** аты жоғары математика пәндерінен хабардар болуымыз керек. Себебі, тілдегі бөліктердің жазба не сөйлеу тілінде аз немесе көп болып келуі кездейсоқтыққа жатады. Ал кездейсоқтықтың да өзіндік заңдылықтары болатынын ескерсек, ол заңдылықты зерттейтін пән – **«Ықтималдықтар теориясы».** Оқу құралының «ЕКІНШІ ТАРАУЫНДА» аталған теорияның ең қарапайым және ең қажетті-ау деген статистикалық құрал түрлері тіліші мамандар үшін арнайы қарастырылады.

Енді осы оқу құралының құрылымдық жағына қысқаша тоқталайық. Қазақстандық статолингвист-ғалымдардың негізгі статистикалық құралы – әр стильден түзілген жиілік сөздіктер материалдары болғандығын ескеріп, оқу құралының бірінші тарауы «Жиілік сөздіктер» деп аталып, жиілік сөздіктерге қатысты мәселелер сөз болды.

Ал екінші тарауда Ресейдің белгілі ғалымы Б.Н.Головиннің «Язык и статистика» атты оқу құралының ізімен филолог-студенттерге ең қажетті деген статистикалық минимум құралдар жайлы мәселе қарапайым түрде қарастырылды.

Кітаптың үшінші тарауы «қазақ мәтінін ықтималды-статистикалық модельдеу» деп аталып, онда қазақ тіліндегі жиілік сөздіктер материалдары негізінде «Ципф заңы» тексеріледі және жиі қолданыс табатын негізгі сөз таптарының мәтін бойында таралуы теориялық үлестіру зандарына бағынуы-бағынбауының сынау (бағалау) критерийлері арқылы қарастырылды. Соңғы – төртінші тарауда қазақ тілі мәтіндеріндегі сөзтұлғалардың лексика-морфологиялық құрылымын статистикалық әдіспен зерттеу үшін шартты белгі-кодты сәйкестендіру бағдарламасы да осы тарауда түңғыш рет көрініс тауып отыр. Мәтін бірліктеріне тән белгі-кодтарды сәйкестендіру «ұялы принципке» негізделді. Ал негізгі сөз таптары ретінде қазақ тілінің зат есім, етістік, сын есім сөздері тиісті белгі-кодқа сәйкестік кестелер арқылы берілді.

Оқу құралында қолданыс тапқан пәндік терминдердің көбіне автордың өз қалауымен алынғанына байланысты мемлекеттік деңгейде бекіп үлгермеген кейбір атаулардың кездесуі мүмкін екендігін ескерте кеткіміз келеді. Осы кітапта көтерілген мәселелер жайлы жайттарды, негізінен, орысша жазылған әдебиеттерден кездестіруге болатындықтан, кітаптың соңында «Қосымша» ретінде оқу құралының мәтінінде кездесетін терминдер мен олардың фразалық тіркестерінің қазақша-орысша сөздігі және басқа да қолданбалы лингвистика мен математикалық лингвистика пәндерінде кездесетін негізгі терминдерінің кейбір атауларының қысқаша орысша-қазақша сөздігі мен анықтамалары берілді.

Бірінші тарау

ЖИІЛІК СӨЗДІКТЕР

1.1. Қазақ тілтанымындағы статистикалық әдістің орны

Ғылыми-техникалық прогресс, электронды-есептеу техникасының дамуы, нақты әдіс-тәсілдердің түрлі білімдер салаларында қолданыс табуы тіл біліміне, соның ішінде түркітануға да өз ықпалын тигізді.

Өткен ғасырдың өзінде үндіеуропа тіл білімінде кең тараған статистикалық әдіс түркітануда, соның ішінде қазақ тілінде тек XX ғасырдың 60-жылдарынан бастап жүйелі және ғылыми түрғыда қолданыла бастады. Бірақ бұдан бұрын да аталған әдіс ішінара қолданыс тауып, тілдің әр деңгейіндегі бірліктердің сандық және пайыздық арасалмағын анықтаумен шектелін келген болатын.

Түркі тіліндегі мәтіндерді статистикалық тәсілмен зерттеу бұл тілдің ішкі табиғатына, яғни агглютинативті құрылымына негізделеді. Мамандардың пайымдауынша, кітаби және сөйлеу тіліндегі түрлі типті сөздердің әр ыңғайдағы қолдану жиілігін анықтайтын мұндай тәсіл тілдің құрылымдық-грамматикалық, кейде стильдік-семантикалық айырым белгілерін ажырата отырып, осы стильдерге тән қасиеттерді анықтауға мүмкіндік жасайды. Сонымен бірге грамматикалық категорияларды сандық тәсілмен талдау әдеби тілдің барлық аясындағы функционалдық салмағын корсетуде де маңызды.

Белгілі филолог-статист Р.Г.Пиотровский түркі тілдерін сандық деңгейде зерттеудің негізі олардың ішкі құрылымдық

табиғатына тән екендігін айта келе, бұл тілдерге үндіеуропа тіліне тән екпінді ассимиляция мен редукциялар және осы сипатты қасиеттер жат екенін баса көрсетеді. Бұл жағдай тілдік деректердің жай кезде қиындық туғызатын кейбір түстарын математикалық әдіспен тіркеу мен тануда аса маңызды деп саналады [75].

Тілдік мәселелерге қолданылатын математикалық әдісті интуитивті түрде тұжырымдалған, сол сияқты толық шешімі жоқ мәселелердің логикалық тұжырымдалуына бағынатын және алгоритмдік шешімі бар бір немесе бірнеше қарапайым математикалық есептермен ауыстыру деп ұғыну қажет. Тілдік мәселені осылайша математикалық жолмен шешу - лингвистикалық нысанның математикалық моделіне (үлгісіне) көшуді қажет етеді [31, 6-б.]. Тілдің аса күрделі түрдегі сан алуан параметрлі көп қабатты жүйе екенін ескерсек, онда ықтималдық теория, математикалық статистика, ақпарат теориясы және т.б. ғылыми салалар әдістерін қолданудың мүмкіндігі кең екені байқалады. Әсіресе, бұл әдістер сөйлеу тәжірибесінде (жазба не ауызша) жиі кездесетін белгілі бір кешенді шарттарға байланысты тілдік кездейсоқтық құбылыс заңдылықтарын қарастырғанда аса ыңғайлы. Тіл табиғатының нақты ғылымдар саласының ережелеріне бағына бермейтін түстары баршылық. Сондықтан осындай күрделі нысан ретіндегі тіл саласына ықтималдық теориясы мен математикалық статистиканы қолдану үшін, ең алдымен, «тілге» бірнеше шектеулер қою қажет. Мәселен, толығымен өзгермелі бірліктерден тұратын табиғатына жүйелілік сипат тән тілді көптеген зерттеуші «ашық жүйе» ретінде қарастыруды дұрыс деп санайды. Осы тұжырымның негізінде, тілге қойылатын бірінші шектеу бойынша, белгілі бір кезеңде тілдік жүйеде өзгермелілік сипат болмайды деп, оны «жабық жүйе» ретінде санау (модельдеу) ұйғарылды.

Екінші шектеу – тілдің ауызша және жазбаша түрлерінің аралық айырым белгілерін ескермей, оны арнайы түрде құрастырылған белгілер жүйесі немесе жазба тіл деп қарастыру. Мұндай ұйғарым мәтінді көру арқылы қабылдауға негізделген сипаттау грамматикасын құруға жеткілікті негіз бола алады [74].

Статистикалық зерттеу кезінде нақты мәтінді сөйлеу ақиқатығына белгілі дәрежеде сәйкес келеді деп үйғару негізінде оның үлгісі зерттеу нысаны ретінде алынады. Әрине, бұл жағдайда мәтін қасиеттері арнайы кеңейтіліп қарастырылады және мәтін үлгісі бір-біріне тәуелсіз дискретті бірліктердің тізбегі ретінде танылады.

Тілге қойылатын келесі шектеудің туындау себебі: тілдің қай саласында болмасын зерттеу арқылы сөйлеу қызметінің әр алуандығын қамту мүмкін еместігінде. Өйткені зерттеу нысаны ретінде алынған мәтінге қатысты мәліметтер тек сол мәтіндік стильге ғана тән болады. Зерттеу барысында мәтін типі мен оның сандық мөлшерін дұрыс таңдай білу зерттеу арқылы байқалған заңдылықтың дұрыстығына күмән туғызбаудың айғағы деп саналады.

Статистикалық әдіс басқа математикалық әдістерге қарағанда тіл зерттеу ісіне бұрынырақ және тұрақты сипатта енді деуге болады. Бұл әдістің қажеттілігі айқын, ол қазіргі жағдайда тілшілер арасында айтарлықтай күмән туғызбайды.

Белгілі орыс ғалымы Б.Н.Головин әуелден-ақ сандық сипаттың тіл категориялары мен оның элементтеріне объективті түрде тән болуынан, сонымен бірге сапалық және сандық сипаттарының өзара қатынастылығы тілдің қатынас құралы ретіндегі қызметі мен даму сатысындағы объективтілігінен тілді сандық тұрғыда зерттеу қажеттілігі туындады деп түсіндіреді [30]. Мысалы, морфемдік құрам создің санасына әсер ететіндігі сияқты морфологиялық деңгейдегі сандық құбылыстар оның синтаксистік деңгейдегі сапасына әсер етпей қоймайды. Осыдан барып әр қалам иесінің стилі де тек өзіне ғана тән тұрақты бірліктерінің қатынастары арқылы ерекшеленеді.

Заңдылықтарды тәжірибелік байқау мен оның нәтижелерін математикалық жолмен өңдеу арқылы анықтау тілдік жүйені ғаным-білудегі лингвистикалық статистикадағы әрі табиғи, әрі ғылыми бетбұрыс болып саналады. Нәтиженің дәлдігіне, жасалатын тұжырымдардың объективті болуына талпыну – барлық ғылымдар саласына тән жағдай. Ал қоғамдық ғылымдар саласындағы ерекше дәлдікті қажет ететін ғылым саласы – лингвистика деуге болады.

Статистикалық әдістің тіл саласындағы қолданылу тарихын сөз еткенде, осы әдісті тұңғыш ұсынған орыстың белгілі математик-ғалымы В.Я.Буняковскийдің (1804–1889) атын атамай кетуге болмайды [86]. Сонымен бірге басқа да ірі ғалымдардың тілді зерттеуде сандық деректерге жүгінудің қажеттілігін дәлелдеулеріне деректерді көптеп келтіруге болады (И.А.Бодуэн де Куртенэ, А.М.Пешковский, М.Н.Петерсон, Е.Д.Поливанов, В.В.Виноградов және т.б.).

Классикалық тіл білімінің өкілі В.В. Виноградов тілдің кітаби, сөйлеу түрлерінде және көркем әдебиет стилінің әр жанрларында сөздердің қолдану жиілігінің біркелкі еместігін айта келе, оларды анықтау үшін арнайы зерттеудің қажеттігін көрсетеді. Ғалымның тұжырымдауынша, мұндай зерттеулер түрлі стильдердің құрылымдық-грамматикалық, тіпті семантикалық айырмашылықтарын белгілеуге, сондай-ақ барлық грамматикалық категорияларына статистикалық талдау жүргізілген әр әдеби стильдің қатынастық-функционалдық салмағын айқындауға мүмкіндік туғызады [26].

Мәселен, Ф.П.Филин өзінің бір еңбегінде: «Ғылымның бірақ мақсаты болады, ол – әлі ашылмаған шындықты ашу немесе соның ашылуына ықпал жасау, сол арқылы қоғамға қызмет ету. Алға қойған мұндай мақсатқа әр түрлі жолмен жетуге болады. Соның ішінде зерттеудің дәлдігі мен объективтілігін қамтамасыз ететін тәсіл ғана ең дұрысы болмақ. Дәлдік пен объективтілікке ұмтылу – қазіргі ғылымның туы. Бұл тіл біліміне де қатысты», – деп жазса [89], В.Н.Головин тіл ғылымына сандық сипаттың тән екендігін, сондықтан тілші-ғалымдардың көпшілігінің тілдік деректерді сипаттағанда сандық ұғымдарға жүгінетінін айтады [30].

Тілдік фактілерді сандық ұғымдарға жүгініп баяндау – қазақ тіл білімін зерттеуші ғалымдарға да тән. Өйткені қай кездегі болмасын қазақ тілі жайлы зерттеулердің бәрінде дерлік ***аз, көп, көптеген, жиі, сирек, өнімді, өнімсіз, мол, жұық, тең, шамалас, көп рет*** т.б. шама, мөлшерді көрсететін сөздер кездесіп отырады және бұлардың мәнін нақты санмен көрсетуге талаптанушы ғалымдар да бар. Сөйтіп, қолданбалы тіл білімінің бүгінде «Статистикалық лингвистика» деп аталып жүрген жаңа саласының алғашқы нышандары мен белгілері міне осылардан

басталады. Ал қазақ тіл біліміндегі бұл саланың бастауында профессор Құдайберген Қуанұлы Жұбанов тұрды. Мұны белгілі тілші-статист ғалымдар – Қ.Б.Бектаев, С.Мырзабеков т.б. қазақ тілінің статистикасы туралы зерттеулері мен мақалаларында атап көрсеткен. Мәселен, С.Мырзабеков «Қазақ тілін зерттеуде санды пайдалану» атты мақаласында: «Қазақ совет тіл білімінің негізін қалаушылардың бірі, совет дәуірінде қазақ тілін тұңғыш рет ғылыми түрде зерттеген маман-лингвист Қ.Жұбанов ретіне қарай сандық деректерді де пайдаланған», – десе [69, 167-б.], ғалым Қ.Б.Бектаев «Лингвистикалық статистикаға қатысты элементтерді – сандық деректер мен мәліметтерді, оны пайдалану нәтижесіндегі кейбір пікірлер мен түжырымдарды біз профессор Қ.Жұбановтың еңбектерінен жиі кездестіреміз», – дейді [10, 110-б.]. Ғалымның сол кездің өзінде-ақ тіл білімінде сандық деректерді пайдаланудың қажеттілігіне ерекше деп қойғаны байқалады. Мәселен, ғалымның 1936 жылы жарық көрген «Қазақ тілінің грамматикасында» сөз мүшелерінің орны, дыбыстардың саны мен сапасы, түрлері туралы жазған мына түжырымы көңілге қонымды: «Дыбыстың саны бар да, сапасы бар. Дыбыстың сапасы бірдей бола тұрып, саны әлденешеу бола береді. ...Бір тілде жұмсалатын дыбыстардың жалпы санының үшы-қиыры жоқ, бірақ оның бәрінің бірдей сапасы әр түрлі бола бермейді. Сапасы әр түрлі болатын дыбыстар да көп болмайды». Сол сияқты «... Ілиястың небәрі 12 жол «Қойшы ойы» деген кішкене өлеңінде 488 дыбыс бар. Бірақ 488 дыбыс түрлі емес, мұнда небәры 28-ақ түрлі дыбыс бар. Осы 28 түрлі дыбыстың бірі 10; бірі 50 рет қайталанып барып 488 болған. ... Қазақ тілінде де сапасы әр түрлі болатын дыбыстардың арнаулы саны бар» [40. 183-б.].

Профессор Қ.Қ.Жұбановтың «Қазақ тілі жөніндегі зерттеулерінде» бұл жайлардан басқа да мәселелер, айталық: сөз тіркесінің формалары, сөйлем мүшелерінің орын тәртібі, буындар және ондағы дыбыс саны, буындарды оқыту әдістемесі т.б мәселелер сөз болған түста көптеген сандық (статистикалық) деректер мен мәліметтер келтірілген [40].

Шын мәніндегі статистикалық ізденістерді Қ.Жұбанов қазақ орфографиясын ғылыми негізде құру мақсатында қолданған. Ол әліпби құрамына енуге тиісті әріптердің санын

қазақ тіліндегі белгілі бір фонеманы іс жүзіндегі қажеттігіне, оған деген ділгерлікке, қолданыс жиілігіне бағындыра белгілеуді мақсат тұтқан [10, 114-б.].

Міне, бұл айтылғандардан проф. Қ.Қ.Жубановтың тілді зерттеуде, оның заңдылықтарын қарауда есеп-санаққа, реттi жерінде математикалық ойлауға, оның тәсілдеріне сүйенгендігінің тағы да бір дәлелі.

Қ.Қ.Жубанов еңбектеріндегі сандық деректерді пайдалану жайымен таныса отырып, мынадай қорытынды жасауға болады: проф. Құдайберген Қуанұлы Жубановтың қазақ тіл білімі туралы зерттеулерінде тілді сандық деректер мен мәліметтер негізінде қараудың алғашқы нышандары бар. Бүгінде тіл білімінің бір саласына айналып отырған лингвистикалық статистиканы ғалым сол кездің өзінде-ақ тани білген. Тілдің сандық сипатынан байқалатын сапалық қасиеттерін оны зерттеуде пайдалануға болатын тиімді әдіс көздерінің бірі екенін де байқаған. Сөйтіп, ол сандық деректердің негізінде тілге қатысты бірсыпыра жайларды (дыбыстар, буындар, сөздер, сөйлем мүшелерінің орны т.б.) нақты анықтап, кейбір тың тұжырымдар жасаған. Бұлардың маңыздылығы сонда – олар қазіргі кезде бұл саладағы ізденуші-зерттеушілерге үлгі-тірек, бағыт-бағдар болып отыр [10, 115–116-бб.].

Қазақ тілінің нормативті курсын жүйелеу жұмысына проф. Қ.Жубановтан кейін белгілі тілші-ғалымдар: Н.Т.Сауранбаев, І.К.Кеңесбаев, М.Б.Балақаев, А.Ы.Ысқақов және т.б. ғалымдар атсалысты. Бастауыш және жоғары сыныптарға және педучилицелерге арналған қазақ тілі оқулықтарының авторлары тілдің буындық құрылымын, дыбыс физиологиясын, тілдің акценттік және екпіндік және т.б. мәселелеріне қатысты зерттеулерінде сан мен сапаны бөлмей бірге қарастырады. Мәселен, І.К.Кеңесбаев қазақ тіліндегі буын типтерінің, М.Б.Балақаев қазақ тіліндегі синтаксистік сөз тіркестерінің табиғатын зерттеуде көптеген нақты мәтіндік материалдарды сандық қатынастар көмегімен қарастырған [8, 52]. Арнайы сандық талдауларды профессор Ғ.Мусабаевтың зерттеулерінен де кездестіруге болады [68]. Әрине, ғалымдардың мұндай ізденістері, яғни сандық қатынасқа көңіл бөлуі саналы түрдегі

ықыластан гөрі тілдің ішкі табиғатына (сырына) қатысты туындаған деп пайымдауға болады.

Осы кезеңде қазақ тілінің грамматикалық форманттарын статистикалық жолмен зерттеуге айрықша көңіл бөліне бастаған болатын. Қазақ лексикографтары сандық деректі ұлы ақын Абай тілінің сөздігін түзу кезінде жүйелі түрде толық пайдаланды. Түңғыш рет қазақ тіл білімінің тарихында бір ғана автордың (Абайдың) жазба түріндегі шығармалар тілі толық қамтылып, әр сөзі мен сөз тіркесінің қолдану жиілігіне көңіл бөлінді. Яғни Абайдың тіл байлығы, саны мен сапасы жайлы жайттар тілшілердің арнайы зерттеу нысанына айналды. Сөйтіп, орыстың ұлы ақыны Пушкин тілі сөздігі тәріздес – «Абай тілі сөздігі» жасалды [1]. Бұл сөздікті әрі түсіндірме сөздік, әрі оқыта-жиілік сөздік деуге болады. Сөздердің қолдану жиілігін көрсету сөздік түзудің негізгі мақсаттарының бірі болып саналды. Шығармадағы сөздердің жиілігін ескеру дәстүрлі картотека жасау жолымен іске асты.

1969 жылы Қазақстан Ғылым академиясы Тіл білімі институтында «Статистическое и информационное изучение тюркских языков» атты Бүкілодақтық ғылыми жиынның өтуі қазақ тіл білімі үшін елеулі құбылыс болды. Осы ғылыми форумның нәтижесі ретінде 1970 жылы Тіл білімі институтында статолингвистика саласының белгілі ғалымы Қалдыбай Бектаевтың жетекшілік етуімен «Тіл статистикасы және автоматтандыру» атты ғылыми топ құрылды. Осы кезеңнен бастап қазақ тілін статистикалық зерттеу ісі жүйелі сипатқа ауысты. Бірнеше жас ғалым статистикалық лингвистика саласынан кандидаттық диссертация қорғап, олардың ғылыми мақалалар жинағы, монографиялары, оқу құралдары жарық көрді. Кеңес Одағы кеңістігіндегі ғылыми конференцияларға кеңінен жол ашылды. Қазақстан статолингвистер зерттеулерінің нәтижесі ретінде 1973 жылы «Қазақ тексінің статистикасы» атты ғылыми жинақ жарық көрді [58]. Бұл жинақ сол жылға дейінгі қорғалған (және дайындалған) кандидаттық диссертациялардың негізгі мазмұнын ашады деуге болады.

1973-1974 жж. Р.Г.Пиотровский мен Қ.Б.Бектаевтың «Математические методы в языкознании» атты екі бөлімнен тұратын, жоғары оқу орындарына арналған оқу құралының

Қазақ мемлекеттік университетінен жарық көруі де осы жаңа саланың жетістігі еді [11].

Тілді зерттеу тәжірибесінде әр тілді жалпы типологиялық тұрғыдан бағалауда қолданылатын статистикалық критерийлердің үлес салмағын «өлшеуге» (анықтауға) қажетті жалпылама сандық дәреже әлі де ғалымдарды толық қанағаттандырмай келеді. Міне, осы бос кеңістікті толтыруға арналған іргелі ғылыми зерттеу ретінде 1978 жылы басылып шыққан Қ.Б.Бектаевтың «Статистико-информационная типология поркского текста» атты монографиясын атауға болады [12] Бұл зерттеу түркі мәтіндері үшін ықтималдық және ақпараттық үлгілеу әдістерін жүйелі түрде қолдануға арналды. Ғалым түркі тілдерінің типологиялық ерекшеліктерін үндіеуропа тілдерімен салыстыра зерттеуді мақсат етіп қойып, оны формальды статистикалық-ықтималдық әдіс арқылы іске асырды.

Осы кезеңдік зерттеулерге тән мәселелер – ол тілдің фонетика мен фонология саласы, сөз бен сөзтұлғаның буындық құрылымы, қазақ мәтінінің гүрлі функционалды стилінің лексика-морфологиялық құрылымы, қазақ тілінің ақпараттық сипаты және т.б. кешенді проблемалар болатын. Негізгі зерттеуге қажетті статистикалық құрал ретінде электронды-есептеу машиналары (ЭЕМ) көмегімен жасалатын әліпби-жиілік сөздік, жиілік сөздік және кері әліпби жиілік сөздіктер материалдары болды. Жиілік сөздіктерді ЭЕМ көмегімен автоматты түрде түзу мәселесі мен олардың құнды деректерін тілді зерттеу практикасында іске асыру мүмкіндігі жайлы Кеңес Одағы кеңістігі ғылыми орталықтарында ұйымдастырылған көптеген Бүкілодақтық ғылыми жиындар болып өтті. Қазақ тілін статистикалық әдіспен зерттеу және оны автоматтандыру мәселелері де Одақтық дәрежеде қарастырылды. Отандық түркітану ғылымы үшін ортағасырлық жазба ескерткіші «Кодекс Куманикус» мәтіні тұңғыш рет ЭЕМ жадына енгізілін, ол бойынша автоматты түрде «Куманша-қазақша жиілік сөздігінің» алынуы қазақ тілі тарихы үшін елеулі оқиға деуге болады [61].

Жоғарыда сөз болған қазақ тілі мәтіндерін статистика тәсілімен зерттеуге арналған кандидаттық диссертациялар тақырыбы 1979 жылы «Ғылым» баспасынан жарық көрген

М.О.Әуезовтің «Абай жолы» романының жиілік сөздіктеріне тікелей не салыстырмалы тұрғыда қатысты болды [13]. Ал ЭЕМ көмегімен алынған және 1995 жылы басылып шыққан «М.Әуезовтің 20 томдық шығармалар текстерінің жиілік сөздіктері» ұлы жазушының 100 жылдық мерейтойына арналған Қазақстандық статолингвистерінің арнайы сыйы деуге болады [14]. Жазушы тілінің лексикалық байлығын, автордың сөз қолданудағы стильдік тәсілінің өзіндік ерекшелігін, қазақ тілінің көркемдік және бейнелеуші мүмкіндіктерін меңгеру дәрежесін осы аталған лексикографиялық еңбек бойынша бағалауға болады. Осымен қатар бұл сөздік жалпы қазақ лексикасын статистикалық құрылым тұрғысынан зерттеуге бай тілдік материал бола алатыны сөзсіз. Кітап түрінде баспадан жарық көрмегенімен М.Әуезовтің әр томы бойынша компьютер жадында сақталған «Сөзнұсқағыш әліпби жиілік сөздік» атты электронды сөздік базасы да арнайы қаралатын лексикографиялық еңбек деуге тұрады.

Л.Ахабаевтың қазақ тілін нормативті тұрғыда тануға, қазіргі қазақ публицистикасы тілінің лексика-морфологиялық құрылымын статистикалық талдауға арналған ғылыми жұмысы да маңызды зерттеу болып табылмақ [3]. Бұл жұмыстың шағын бөлігі дублет сөздер вариантының нормасын статистика жолымен айқындау мәселесіне арналған. Статистикалық талдау негізінде А.Ахабаев сөздердің жиілік сипаты негізгі объективті критерий болатындығын сөз етеді. Жалпы алғанда, бұл ғылыми жұмыс 1965–1966 жж. газет мәтіндері мен «Абай жолы» романы (2-ші кітап) мәтініндегі зат есім мен есімдік жалғауларының форма-варианттарын статистика-морфологиялық тұрғыда салыстыра талдауға арналып, соның негізінде қазақ тілінің грамматикалық тұлғаларының нұсқалары стильдік айырым белгілер болатыны анықталған.

Статистика-морфологиялық зерттеудің келесі бір нысаны – қазіргі қазақ тілі етістік сөздерінің туынды тұлғаларын құрайтын негіз етістік пен оларға жалғанатын аффикстер. Осы зерттеуді М.Әуезовтің «Абай жолы» романындағы туынды түбір етістіктерінің құрылымына статистикалық талдау жасау арқылы ғалым С.Мырзабеков жүргізді. Ғалымның сандық деректері бойынша романда кездесетін барлық түбір етістіктің 80 пайызы –

туынды етістіктер екені және етістіктің туынды түбірлері негізгілерден төрт есе көп қолданылатыны анықталды [70]. Етістік құрылымына жасалған лингва-статистикалық талдаудың нәтижелері қазіргі қазақ тіліндегі етістік сөздерді танып-білуге мүмкіндік беретіні сөзсіз.

Қазақ тіліндегі жалғаулар мен олардың құрылымдық түлғалары, терминдер мен кірме сөздер, сөз тіркестері статистикасы сияқты мәселелер А.Белботаевтың ғылыми ізденістерінде қарастырылды [18].

Тіл ерекшелігі мәтіндердің жанрлық (стильдік) түрлеріне қарай да айырым табатыны белгілі. Осындай айырымдарды қазақ тілінің әр түрлі жанрларына қатысты статистикалық жолмен тағайындауды қазақстандық статистика тобының ғалымдары 70–80-жылдары өз мақсаты етіп қойды. Мысалы, Қ.Молдабеков пен Б.Е.Қалыбековтер балалар әдебиеті мен бастауыш сыныптарының оқулықтарын зерттеуді өз міндеттеріне алса, А.Р.Зекенова – М.Әуезовтің драмалық шығармалар тілін, А.Ахабаев – газет тілін, А.Белботаев – ғылыми-техникалық стиль, математикалық стиль тілдерін өз зерттеу нысандары етіп алып, алғашқы ғылыми нәтижелерге қол жеткізді [3, 18, 19, 43, 60, 67].

Қазақ тілінің фонетика саласы да статистика әдісінің назарынан тыс қалған жоқ. Фонетиканың дәстүрлі (есту) және құралды-тәжірибелік зерттеу әдісі бойынша фонологиялық және графикалық жүйелерге қатысты мәселелерін ықтималды-статистикалық әдіс-тәсілдермен шешу тиімді деп танылды [9, 32, 33, 53, 87].

Қазақ тіл білімінің қолданбалы саласының тың проблемалық мәселесі – қазақ мәтінінің әр әріпке (дыбысқа) шаққандағы ақпараттық өлшемін, яғни ***энтропияны*** анықтау. Осындай зерттеулер белгілі ғалым Қ.Б.Бектаев пен оның шәкірті Д.А.Байтанаеваның қазақ мәтініне қатысты ақпараттық-статистикалық ғылыми ізденістері арқылы іске асырыла бастады. Табиғи тілдің белгілер жүйесі үш түрлі деңгейде ***синтактикалық, семантикалық*** және ***прагматикалық*** аспектілерде қарастырылатыны мәлім. Мұндағы синтактика деңгейі әр түрлі белгілер жүйесінің синтаксисін, яғни сөз мағынасынан тыс, белгілердің тіркесімдік құрылымын және олардың туындау

ережелерімен қатар белгілер құрылымының ішкі қасиеттерін қарастырады. Ал белгілер жүйесінің сыртқы құрылымдық қасиеттерімен семантика мен прагматика айналысады. Дәлірек айтсақ, егер ***семантика*** белгілер жүйесін мағыналық көзқарас, яғни мәтін мазмұны тұрғысынан танып-білумен шұғылданса, ***прагматика*** – белгілер жүйесінің қабылдаушыға қатынасын анықтау мәселелерімен айналысады.

Қазіргі кезде ақпарат аясын теориялық және технологиялық тұрғыда зерттеудің синтактикалық деңгейі жақсы дамып келеді. Бұған К.Шеннонның символдар жүйелілігі мен комбинаторикасын қарастыратын статистикалық байланыс теориясы негіз болады. Бұл теория бойынша табиғи тілде жазылған мәтін белгілі бір код құрайтын дискретті белгілердің тізбегі ретінде қарастырылады. Ал мұндай мәтінді қабылдау кезінде дискретті белгілердің қай-қайсысының да пайда болуы белгісіздіктің шамасымен, яғни энтропиямен өлшенеді. Қ.Б.Бектаев пен Д.А.Байтанаеваның ғылыми ізденістерінде қазақ мәтінінің ең кіші бірлігі – әріпке түсетін энтропияның орта шамасын және қазақ тілінің толық және стильдік (жанрлық) айырымдарын анықтайтын жазба мәтіндеріндегі энтропия мен артық ақпарат шамаларын анықтау негізгі мақсат ретінде қойылды. Осымен қатар қазақ тілі бойынша анықталған осындай сипаттамаларды синтактикалық деңгейде үндіеуропа тілдері бойынша алынған белгілі деректермен салыстыра зерттеулер де іске асты. Қазақ мәтініндегі әрбір әріпке түсетін энтропияның ең үлкен (максимум) шамасын анықтау, әліпби әріптерін қабылдаудың ықтималдықтары тең және өзара тәуелсіз дәрежеде деген үйғарымға негізделетіні айқын [5]. Осы аталғандармен бірге Д.А.Байтанаева қазақ жазба мәтінінің буын бірлігінің статистика-ақпараттық қасиеттерін тәжірибелік мәліметтер негізінде айқындайтын статистикалық зерттеулер де жүргізді [6].

Кезінде Бүкілодақтық статистика тобы мүшелігінде Қазақстаннан басқа да түркі тілдес мемлекеттердің статист-тілшілері болды. Осының нәтижесінде 1988 жылы К.Дыйқановтың қырғыздың «Манас» эпосының 2 бөлімнен тұратын жиілік сөздігі жарық көрді. Ж.Жетешиков қырғыз публицистика тіліндегі зат есімдердің сөз өзгертуші аффикстерінің статистикалық деректерін

қазақ тілінің осындай мәліметтерімен салыстыра зерттеді [38]. С.И.Ибрагимов ғылыми-техникалық стильдері мәтіндері бойынша қырғыз тілінің зат есімдерінің синтаксисті-функционалды кластарын бөлек қарастырып, олардың басқа сөз таптарымен синтаксистік байланыс түрлерін статистикалық жолмен қарастырды [48].

Қырғыз тілі бойынша жүргізілген іргелі ғылыми-зерттеу жұмыстары қатарына Т.Садықовтың «Проблемы моделирования тюркской морфологии» (аспект прождения киргизской именной словоформы) атты монографиясын [84] қосуға болады. Бұл жұмыста қырғыз тілінің алдын ала берілген лексика-грамматикалық ақпараты бойынша есім сөзтұлғалардың туындау үлгісін (моделін) құру көзделді. Ол үшін қырғыз тілінің морфологиялық құрылымына жүйелі түрде талдау жасалды [84].

Қазіргі өзбек тілінің статистикалық құрылымын зерттеушілердің бірі С.А.Ризаев болды. Ол өзбек әдеби тіліндегі екі фонемдік тіркесім жағдайын қарастырып, олардың қатынастық жиілікке әсер ететін факторларын анықтады. Публицистика, ғылыми және көркем әдебиет (балалар ертегісі) стильдерінің түрлі жанрларындағы фонеманың сандық қатынастарын қазақ тіліндегі осындай деректермен статистика жолымен салыстыра зерттеп, олардың өзіне тән ерекшеліктері негізінде фонемалық топтарын бөліп алуға мүмкіндік туды.

Қазақ тілі мен өзбек тілдерінің фонемалық спектрлерін статистикалық талдауға А.Жүнісбеков пен Д.Маматов өз зерттеулерін арнады [34, 64]. Бұл жұмыстарда қазақ тілінің дауысты, ал өзбек тіліндегі күрделі дауыссыз дыбыстардың спектрлері қарастырылды. Осының негізінде дербес фонемалар мен олардың позициялық варианттары жайлы статистикалық деректер алынды. Осы тәріздес зерттеу жұмыстары қазақ, қарақалпақ пен әзірбайжан тілдері бойынша да жүргізілді [7, 23, 72].

С.А.Ризаев өзбек тілінің буын типтерін және олардың сөздердегі позициялық орналасуы мен ашық, жабық буындардың қатынастық сипаттары негізінде сөздердің тұрақты (канондық) түрпатын анықтады. Мұндай зерттеулер балалар әдебиеті бойынша да жүргізілді [79].

Қазіргі татар әдеби тіліндегі буын саны мен олардың буындық құрылым типтері, комбинаторлық қасиеті, сөздердегі

буындардың қолдану жиілігі жайлы мәселелерді Т.И.Ибрагимовтың ғылыми ізденістерінен кездестіруге болады [49]. Мәселен, татар тілі сөздеріндегі алдыңғы буын өзінен кейін келетін буын жайлы ақпараттың 47 пайызын өз бойында ұстайтындығы анықталды. Мұндай деректер тілдің құрылымдық және фонетикалық құрылымы жайлы көптеген мәлімет беретіні белгілі. Буынның ішкі құрылымына және сөздің буындық құрылымына ықтималды-статистикалық әдіс қолдану арқылы зерттеу өзбек және татар тілдерінен басқа қазақ тілі бойынша да молынан жүргізілді деуге болады. Мәселен, Қ.Б.Бектаев көркем әдебиет, публицистика, ғылыми-техникалық мәтіндер мен 2 томдық «Қазақ тілінің түсіндірме сөздігі» (1959, 1961 жж.) материалдары бойынша ЭЕМ арқылы қазақ тілі буындарының жиілік сөздігін жасады [9]. Ал Д.А.Байтанаева қазақ мәтіндеріндегі фонемалардың ғана емес сөздердегі буындардың да статистика-ақпараттық қасиеттерін тәжірибе жүзінде қарастырды [6].

Статистикалық әдіс түрлі тілдердегі грамматикалық түлғаларды салыстыра зерттеуде де елеулі нәтижелерге ие болды. Мәселен, Т.М.Гарипов башқұрт, қазақ, татар, чуваш түркі тілдерін лексика-семантикалық тұрғыда талдап, олардың сандық және сапалық ерекшеліктерін анықтауды мақсат етті [27]. Сол сияқты туыстас емес тілдер қазақ және орыс, қазақ және ағылшын тілдерін салыстырғандағы грамматикалық формалар да статистикалық зерттеу аясынан тыс қалмады [4, 36, 51].

1987 жылы ҚазКСР «Наука» баспасынан шыққан (осы жолдар авторының) «Квантитативная структура казахского текста» және 2002 ж. жарық көрген «Основные принципы формализации содержания казахского текста» атты монографияларында қазақ тілінің заманға сай өзекті зерттеу мәселелері сөз етілді. Бірінші аталған ғылыми еңбекте қазақ тіліндегі мәтіндерді автоматты түрде электронды-есептеу машиналары (ЭЕМ) арқылы статистикалық әдіспен зерттеу мүмкіндігі мен сөз тәртарының ықтималды-үлестірімді заңдылықтары зерттелсе, екінші монографияда – қазақ мәтінінің мәтін лингвистикасы тұрғысынан мазмұнын формальдау жолдары және мәтіннің туындау (порождение) проблемалық мәселелері қарастырылады [31, 39].

Сонымен, жоғарыда сөз болған ізденістер жеке автор шығармалары, функционалдық стильдер не толық әдеби тіл қамтылған мәтіндер бойынша статистикалық әдіс-тәсілдермен зерттеу тәжірибесінде әр түрлі жиілік сөздіктер жасаудың өзекті мәселе (актуалды) екендігін көрсетті. Себебі олардың негізінде аса маңызды және күрделі мәселелерді: тілді лексика-грамматикалық нормалау, түркі тілдерін өзара салыстырмалы-типологиялық зерттеу және олардың компьютерлік қорын жасауға зор мүмкіндік жасалады.

Статистикалық лингвистика саласына қатысты зерттеулер жайлы жоғарыда келтірілген қысқаша шолудан ең алдымен байқайтынымыз, мұндағы зерттеу аясының (нысанының) кеңдігі және бұл әдістің әмбебаптық сипаты. Солай бола тұра, біз қарастырған зерттеу жұмыстарының нәтижелері мен сапасы, деңгейі бірдей дәрежеде емес. Егер қазақ, қырғыз, әзірбайжан және өзбек тілдеріндегі ізденістер диссертациялық деңгейдегі іргелі зерттеулер қатарына жатса, ал кейбір басқа тілдерге қатысты статистикалық зерттеулер жаңадан ғана бастама алуда және олар жайлы тезис түріндегі қысқаша хабарламалар бойынша аз ғана пікір айтуға болады. Сол сияқты мұнда тіл білімінің барлық салалары да біркелкі қамтылмаған. Мәселен, қай түркі тілін алсаңыз да, олардың құрылымын статистикалық лингвистика тұрғысынан зерттеуде синтаксис пен семантика салалары тиісті дәрежеде қамтылмаған деуге болады. Әрине, аталған әдіс түркі тілдері үшін «жаңа» болғанымен, кейбір түркі тілдері үшін бұл әдіс тұрақтанып, түркітанушы ғалымдардың зерттеулерінен елеулі орын алған. Басқа түркі тілдерін айтпағанда, қазақ тілінің көптеген кезек күттірмейтін мәселелері, мәселен лексикография саласы бойынша компьютерлік қор (база) жасау, қазақ тілінен басқа тілге, не керісінше компьютерлік аударма жасау, ақпарат мазмұнынан қысқаша реферат, аннотация алу және қолданбалы лингвистика саласының басқа да мәселелері өз шешімін күтуде.

1.2. Жиілік сөздіктердің түрлері және олардың қолданбалы лингвистикадағы маңызы

Мәтін бірліктері және олардың сипаттамалары. Жазба мәтін – қағаз бетіне түскен жазба сөзі немесе соның бір үзіндісі, бөлігі. Мәтіндегі ой-пікірлер тек сөйлем түрінде ғана айтылғанда түсінікті бола алады. Ал сөйлем ойды айтып немесе жазып жеткізудің негізгі амалы. Сондықтан да мәтінді синтаксистік бірлік деп үйғаруға болады.

Мәтіннің негізгі қызметі – жазбаша (не сөйлеу) түрінде жай немесе күрделі мазмұнды ой-пікірлерді білдіру. Бұл жағдайлардың бәрі де сөйлемдер тобы арқылы жүзеге асады. Олай болса, сөйлем – белгілі қоғам мүшелерінің өзара пікір алмасуын, қарым-қатынас жасауын қамтамасыз ететін мәтін бірлігінің біртұтас түрі.

Лингвистикалық атау (термин) сөздердің мағына-мәндерін бірыңғай ашып алу – тілді зерттеу жұмысының ғылыми және сапалы болуының куәсі. Осындай принципке негізделінің шығарылған ғылыми нәтижелер мен топшылаулар, тілдік заңдылықтар тиянақты да тұжырымды болады деп саналады. Алайда мәтін бойында кездесетін бірліктердің бәрін бірдей қарастыру қиындық туғызады. Сондықтан көптеген зерттеу жұмыстарында қолданылатын лингвистикалық атаулардың – сөз, сөзтұлға (сөзформа), сөзқолданыс, морфема, сөз және оның морфемалық, тұрпаттық құрылысы сияқты түрлерінің айырым белгі-қасиетін алдын ала білген жөн. Бұл атаулар мәтіннің (жазба не сөйлеу) не сөздіктердің бірлігі тұрғысынан қарастырылады.

Енді жиілік сөздіктер және олардың түрлері мен сипаттамаларына қысқаша тоқталайық.

«Сөз» деген атауды ақиқат шындықтағы зат пен құбылыстың атауы, сол атаудың дыбысталуы мен мағына-ұғымының бірлігі және семантикалық жағынан ақиқат шындықтың элементі деп түсінген жөн. «Сөз» атауы жөнінде айтыс туғызатын жаңа мәселелер бола түрса да, соңғы кезде қазақ тіл білімі жаңа арнаға түсіп, дұрыс бағыт алып келеді. Біздің топшылауымызша, сөзтұлға дегеніміз – дербес немесе көмекші мағыналы сөздің түбір тұлғасы мен оған жалғанған тұлға

тудырушы морфемалардың сан алуан көріністері. Яғни түбір сөздің (негізгі түлғаның) өзі де, оған сөз түрлендіруші грамматикалық форманттардың үстелуінен пайда болған түлғалар да сол сөздің әр түрлі түрпаттары болып саналады.

Сөздің негізгі және туынды түрпатының, цифрлар мен әр түрлі символдардың және т.б. белгілердің мәтінде қайталанбай да, қайталанып та қолданылуы ***созқолданыс*** деп аталады.

«Сөз», «сөзтүлға», «негізгі түлға» және ***«созқолданыс»*** деген атаулардың (терминдердің) лексикалық мағына жағынан өзара ерекшеліктері болғанымен, олар бір-бірімен тығыз байланыста болатын әр түрлі тілдік бірліктер.

Сонымен, сөз бен оның түлғаларының және созқолданыстың мәтін бойындағы шекарасы – санауға жататын бірлік – екі ашық жер (пробел) арасындағы мәтіннің бөлігі (тыныс белгілерін есепке алмағанда). Бірақ мәтін бойында тек қана сөз және оның түлғалары ғана емес, сонымен бірге цифрлар (сандық белгі) мен неше түрлі басқа да белгілер кездесуі мүмкін. Мәтіннің бұл элементтері де шартты түрде созқолданыс деп аталады да, олардың жиіліктерінің сөздікте көрініс табуы зерттеу мақсатына қатысты болады.

Тілдің стильдік салаларының лексикалық құрылымын статистикалық тәсілмен жүйелі түрде зерттеу мақсатымен жиілік сөздіктердің бірнеше түрін түзу (жасау) қажет болады.

Жиілік сөздіктердің көп тараған түрлері мынадай:

1) әліпби-жиілік сөздік;

2) жиілік сөздік;

3) кері әліпби-жиілік сөздік;

4) сөзнұсқағыш әліпби-жиілік сөздік (сирек құрастырылатын жиілік сөздіктің түрі).

Осы аталған жиілік сөздіктер үзінділері: *1.1, 1.2, 1.3, 1.4, 1.5-кестелерде* көрініс тапты.

Ескерту. Сөздікке М.О.Әуезов шығармаларының 14-томындағы 30–31 беттердегі мәтін үзіндісі негіз болды және ондағы жол саны кітап жолдарымен сәйкес келмейді. Аталған мәтін үзіндісі тақырыпша соңында берілді. Енді жиілік сөздіктерге қысқаша сипаттама берейік.

1.3. Әліпби-жиілік сөздік

Жиілік сөздіктің бұл түрінде зерттеу мақсатына сай таңдалып алынған мәтіндердегі сөздер (сөзтұлғалар) түгелімен қамтылуы қажет және олар бір-бірімен салыстырылып, әрқайсысының қолдану саны анықталады. Осындай сандық көрсеткіш сөздің (сөзтұлғаның) мәтін ішіндегі ***қолдану жиілігі*** деп немесе сөздің ***«абсолютті жиілігі»*** деп аталады.

Әліпби-жиілік сөздікте көрініс табатын ең бірінші (қатарда) – сөздік бірлігінің реттік сандары, екінші – мәтіннен бөлініп алынған әр түрлі сөздің (сөзтұлғаның) қатаң әліпби тәртібімен орналасқан тұлғасы және үшінші – сөздің абсолютті жиілігі.

Жиілік сөздіктердің толық түрінде осы аталғандардан басқа: «жиынтық абсолютті жиілік», «қатынастық жиілік», «жиынтық қатынастық жиілік» және сонымен бірге бір топ сөздердің мәтінді қамтуының пайыздық шамалары беріледі. Жиілік сөздіктер жасау (түзу) арнайы компьютерлік бағдарламалар көмегімен іске асады. Сөздіктерді жасау барысының бастамасы сөздік бірлігінің әліпби тәртібіне келтірілмеген жиілікті сөзтізбе түрінде, яғни *1.1-кестеде* көрсетілген пішінде алынатынын ескеру қажет.

Осыдан кейін арнайы компьютерлік сұрыптау бағдарламасының негізінде басқа да сөздіктер түрлері түзіледі. Мәселен, әліпбилік сұрыптау арқылы *1.2-кестеде* көрсетілгендей «Әліпби-жиілік сөздік» алынады.

Әліпби-жиілік сөздік зерттеушіге қажет деген сөздің (оның туынды тұлғаларының да) қолдану дәрежесін, яғни оның жиілігін табуға, зерттеуге алынған мәтіндер лексикасының әліпбилік құрамының статистикасын анықтау үшін аса маңызды. Сондықтан, оны тілдің әліпбилік тұрғыдағы статистикалық үлгісі (моделі) деп те атауға болады.

Әліпби-жиілік сөздік, сөздіктің нормативті түріне жатпағандықтан, зерттеуші мәтіндегі немесе оның сөзтізбесіндегі сөздердің (сөзтұлғалардың) әдеби нормаға не орфографиялық жазылу нормасына жату-жатпауына жауап таба алмауы мүмкін. Оның себебі, сөздікте көрініс тапқан сөздер (не сөзтұлғалар), әдетте, қалам иесінің (автордың) не сол шығарманы жарыққа шығарушы жауапты адамның (редактордың) тіл игеру дәрежесіне қатысты да болады.

1.1-кесте (үзінді)

Әліпби тәртібіне келтірілмеген «жиілікті сөзтізбе»

СӨЗ (СӨЗТҰЛҒА)	Абсолютті жиілік	СӨЗ (СӨЗТҰЛҒА)	Абсолютті жиілік
1	2	1	2
Тіпті	1	ғана	4
тұқ	1	толас	1
білмейтін	1	тауып	1
бірақ	8	кішкене	2
оның	5	аласа	1
бұған	2	бөлмесінде	1
жұбанышты ғы	1	болымсыз	3
аз	2	әңгімесін	1
болды	6	тыңдап	1
қашарын	1	варенье	1
білмеді	1	жеп	1
қыстақтан	1	қана	1
шықпай	1	отыратын	1
бір	3	Анна	3
жыл	1	Павловнаның	1
тұрғанда	1	күтуші	1
сол	6	қыздарының	1
мезгіл	1	ішінде	1
оған	2	көркем	1
он	1	қыз	5
жылға	1	өңі	1
бергісіз	1	нәзік	1
тек	2	көзінде	1
шешесінің	1	момындық	1
қасында	1	бар	4

1.2-кесте (үзінді)

Әліпби-жиілік сөздік

Рет саны (Р/с)	СӨЗ (СӨЗТҰЛҒА)	Абсолютті жиілік	Рет саны (Р/с)	СӨЗ (СӨЗТҰЛҒА)	Абсолютті жиілік
1	2	3	1	2	3
1	Андреевич	3	25	ақырын	1
2	Андреевичке	1	26	ақырып	1
3	Андреевичтің	1	27	ал	2
4	Анна	3	28	аласа	1
5	ағылшын	1	29	алғашқы	1
6	адам	1	30	алдампаз	1
7	адамдар	2	31	алмады	1
8	Дидерот	1	32	алмай	1
9	Дмитрий	2	33	алмайды	1
10	аз	2	34	алуандас	1
11	Иван	10	35	алуға	1
12	айғай	2	36	алып	3
13	айдап	1	37	алыс	1
14	айтқандай	1	38	алысы	1
15	айтты	2	39	Маланья	3
16	айтып	2	40	Маланьяға	1
17	айыбы	1	41	Маланьяны	2
18	айыбын	1	42	Марфа	1
19	айыптады	1	43	ант	1
20	айыпты	1	44	ап	1
21	ақ	1	45	Павловна	1
22	ақша	1	46	Павловнаның	1
23	ақылды	1	47	апара	1
24	ақыр аяғы	1	48	апарып	1

Тіл мәдениетін зерттеуші ғалымдар мұндай әліпби-жиілік сөздіктен бағалы мәліметтер таба алады. Сол сияқты, әр стильдік мәтіндердің өзіндік тіл ерекшеліктерін анықтауға және жалпы тіл біліміне қатысты көптеген теориялық мәселелердің басын ашуға мүмкіндік туады деуге болады.

Сөздік бірлігі ***«сөзтұлға»*** болып есептелетін әліпби-жиілік сөздік бойынша, бірлігі – ***«сөз»*** болатын екінші әліпби-жиілік сөздікке көшу қиынға соқтырмайды, себебі «сөзтұлға» сөздігінде бірдей лексемалар құрайтын сөзтұлғалар бірімен-бірі жалғаса орналасқандықтан олардың жалғаулық қосымшаларын

оңай «қиып» тастау мүмкіндігі туындайды. Әліпби-жиілік сөздік үзіндісі *1.2-кестеде* көрсетілді.

Сонымен, әліпби-жиілік сөздіктердің бірлігі – сөз, сөзтұлға, түбірсөз, жай не фразеологиялық тіркес, морфема және т.б. бола алады және оларды жалпы атпен – «әліпби-жиілік тізім» немесе «жиілік тізім» деп атауға болады.

1.4. Жиілік сөздік

Сөздіктердің «жиілік сөздік» түрі әліпби-жиілік сөздіктен басқаша, дәлірек айтқанда, әр сөздің (сөзтұлғаның) қолдану жиілігінің дәрежесіне қарай орналасады: ең бірінші ретте орналасатын мәтіндегі ең жиі қолданылған сөз (не сөзтұлға), екінші, үшінші (тағы тағылар) ретте орналасатын сөздер кездесу жиіліктері бірте-бірте кемін отыратын сөздер (сөзтұлғалар). Егер бірнеше сөздің (не сөзтұлғаның) жиіліктері тұрақтылық сипатта болса, яғни біріне-бірі тең болып келсе, онда ол сөздер (сөзтұлғалар) жиілік сөздікте әліпби тәртібімен реттеледі. Осындай жиілік сөздіктің толық түрдегі үзінді түрпаты *1.3-кестеде* көрініс тапты.

Зерттеуші (мұғалім, оқушы) әліпби-жиілік сөздік бойынша өзін «қызықтыратын» сөздің қандай жиілікте қолданылатынын немесе керісінше, алдын ала «қызықтыратын» жиілік шамасының қандай сөздерге тән екендігін анықтай алады.

Жиілік сөздік (әліпби-жиілік сөздіктер тәрізді) әр түрлі тілдік бірліктер (сөз, сөзтұлға, сөзтіркес т.б.) бойынша жасалуы мүмкін. Біздің баяндауымызда қарастыратын бірліктеріміз негізінен сөзге не сөзтұлғаға сәйкес келетіндіктен, «сөздің жиілік сөздігі» не «сөзтұлғаның жиілік сөздігі» деген сөз тіркестер жиі-жиі кездесуі мүмкін. Ал бұл сөздіктердің де жалпы аты – «жиілікті сөзтізбе».

Ана тіліміздің немесе басқа да туыстас (не туыстас емес) тілдердің әр стильдік мәтіндеріне қатысты жасалған жиілік сөздіктер сол тілдің лексикалық құрылымын терең зерттеуде, лексиканың жалпы функционалдық стильге ортақ бөлігін ажыратуда септігі мол.

1.3-кесте

Жиілік сөздік (үзінді)

Рет саны	СӨЗ (СӨЗТҰЛҒА)	Абсолютті жиілік	Жинтық абсолютті жиілік	Қатынасты жиілік	Жинтық қатынасты жиілік	Бір топ сөздің мәнінің қамту пайызы
1	Иван	10	10	0,016077	0,016077	1,61
2	еді	9	19	0,014469	0,030547	3,05
3	бірақ	8	27	0,012862	0,043408	4,34
4	да	8	35	0,012862	0,056270	5,63
5	болды	6	41	0,009646	0,065916	6,59
6	сол	6	47	0,009646	0,075563	7,56
7	Петр	5	52	0,008038	0,083601	8,36
8	Петрович	5	57	0,008038	0,091640	9,16
9	әкесі	5	62	0,008038	0,099678	10,00
10	енді	5	67	0,008038	0,107717	10,77
11	қыз	5	72	0,008038	0,115756	11,57
12	ол	5	77	0,008038	0,123794	12,38
13	оның	5	82	0,008038	0,131833	13,18
14	Петровичке	4	86	0,006431	0,138264	13,83
15	баласына	4	90	0,006431	0,144695	14,47
16	бар	4	94	0,006431	0,151125	15,11
17	ғана	4	98	0,006431	0,157556	15,75
18	деп	4	102	0,006431	0,163987	16,40
19	өзінің	4	106	0,006431	0,170418	17,04
20	Андреевич	3	109	0,004823	0,175241	17,52
21	Анна	3	112	0,004823	0,180064	18,01
22	алып	3	115	0,004823	0,184887	18,49
23	Маланья	3	118	0,004823	0,189711	18,97
24	болатын	3	121	0,004823	0,194534	19,45
25	болымсыз	3	124	0,004823	0,199357	19,93
26	бұл	3	127	0,004823	0,204180	20,42
27	бір	3	130	0,004823	0,209003	20,90
28	деген	3	133	0,004823	0,213826	21,38
29	етті	3	136	0,004823	0,218650	21,86
30	жіберген	3	139	0,004823	0,223473	22,35
31	жігіт	3	142	0,004823	0,228296	22,83
32	мұның	3	145	0,004823	0,233119	23,31
33	адамдар	2	147	0,003215	0,236334	23,63
34	Дмитрий	2	149	0,003215	0,239550	23,95
35	аз	2	151	0,003215	0,242765	24,28
36	айғай	2	153	0,003215	0,245981	24,60
37	айтты	2	155	0,003215	0,249196	24,92
...	...	...	...	...	...	...
78	шығып	2	237	0,003215	0,381029	38,10
79	шықты	2	239	0,003215	0,384244	38,42
80	ішінен	2	241	0,003215	0,387460	38,75
81	Андреевичке	1	242	0,001608	0,389068	38,91
...	...	...	...	...	...	...
460	ішік	1	621	0,001608	0,998392	99,84
461	ішінде	1	622	0,001608	1,000000	100,00

Сондай-ақ, жалпы халықтық лексиканың, кірме сөздер мен терминдердің, неологизмдердің және басқа лексикалық топтардың тілдің жалпы лексика жүйесінен алатын орны жөнінде де жиілік сөздік бойынша көптеген сандық және сапалық мәліметтер алуға болады. Сөздердің саны мен түлғалары арқылы тілдің лексикалық байлығы мен стильдік ерекшеліктері анықталады. Мұндай сөздіктерден зерттеуші жаңа сөзжасам жүйесіне қатысты да бағалы деректер ала алады. Сонымен бірге, жиілік сөздік мәліметтері қазақ тілін туыстас түркі тілдері және туыс емес тілдермен салыстыру – типологиялық зерттеу үшін де аса қажетті екенін атап өткен жөн [12].

Жиілік сөздіктер күнделікті қолданбалы сипаттағы міндеттерді шешуде де маңызды. Мәселен, әр түрлі типтегі минимум сөздіктерін жасауда және шет тілдерін (екінші тілді) үйрену мен оқытуға қажетті статистикалық әдістерді қолдануда да қажетті. Сонымен бірге жиілік сөздіктердің тілдік және сандық мәліметтері оқулықтар мен оқу құралдарын құрастыру ісінде тілдік материалдарды тиімді түрде орналастыруда және әр қилы лексикалық топтар мен грамматикалық категорияларды орынды түрде оқыту процесінде де таптырмас тілдік құрал рөлін атқара алады.

Жиілік сөздіктер материалдарының тіл білімін зерттеудегі қажеттілігі мен бағалылығы жөнінде Л.Н.Засорина мынадай пікір айтқан болатын: «Материалы частотных словарей исключительно ценны и для собственно лингвистических исследований. Они оказывают влияние на судьбы традиционной лексикографии. Ими пользуются в решении основной проблемы общей лексикологии – выделение словарного фонда активного и периферического словаря; они полезны и при изучении вопросов стилистики, семантики и литературной нормы языка» [41, С. 3-4].

Жиілік сөздік өзге сөздіктердің көлеміне қатысты және басқа да лексикографиялық мәселелерді шешуде объективті негіз болатындықтан, екі немесе одан да көп тілдік аударма сөздік пен түсіндірме сөздік жасаушылар бұл сөздіктен қажетті деректер таба алады.

Жиілік сөздіктер телефон мен телеграф байланыстарын жетілдіру үшін де, стенография мен полиграфия мұқтаждықтары үшін де пайдаланылады.

Сайып келгенде, жалпы түркі тілдерінің, соның ішінде қазақ тілінің жиілік сөздіктерін жасау қажет деп саналады, өйткені олар көптеген лингвистикалық мәселелерге математикалық әдістерді қолданып зерттеуге мүмкіндік туғызады.

Қазақ тілінің сөз байлығын, яғни қазіргі қазақ әдеби тілінің жалпы лексикасын түгел қамтитын жиілік сөздік әлі жасала қойған жоқ. Мұндай сөздіктің құрастырылуы мен жарық көруі – жуық болашақтың ісі.

1.5. Кері әліпби-жиілік сөздік

Сөздіктердің басым көпшілігінде сөздер алдыңғы дыбыстар бойынша әліпби тәртібімен орналасса, кері әліпби сөздікте, керісінше, сөз бен сөзтұлғалар соңғы әріп-дыбысынан басталып әліпби тәртібіне келтіріледі және сөздердің соңғы жағынан тегістеліп беріледі. Сөздіктің бұл түрі «Кері әліпби-сөздік» немесе тек ***«Кері сөздік»*** деп аталып жүр. Егер белгілі бір мәтін бойынша түзілген мұндай сөздікте әр сөздің мәтіндегі қолдану жиіліктері де ескерілетін болса, ондай сөздікті «Кері әліпби-жиілік сөздік» деп атайды. Қазақ тілін зерттеу жұмыстарында мұндай сөздіктер де бірте-бірте қолданыс тауып келеді.

«Кері сөздіктің» алғашқы үлгісі осыдан бір ғасырдай бұрын (1873 ж.) баспадан шыға бастағаны белгілі. Бірақ біздің заманымызда бұл сөздіктерге қызығушылық тек XX ғасырдың 80-жылдарынан ғана басталды деуге болады. Мәселен, 1904 жылы латын тілі мен ескі иран тілінің кері сөздіктері жарияланды. 1944 жылы ескі грек тілінің кері сөздігі пайда болды. 1955 жылы ескі славян тілінің, ал 1957 жылы қазіргі румын тілінің кері сөздігі шықты. Чехославакияда 1935–1957 жылдары чех тілінің 9 томды сөздігі бойынша жасалған кері сөздік кішігірім кітапшалар түрінде басылып шықты.

Кері әліпби-жиілік сөздіктер жасауға айырықша көңіл бөлінудің нәтижесінде, Кеңес Одағы кеңістігінде де әр түрлі тілдер бойынша осындай сөздіктер жарық көре бастады. Мысалы, орыс тіліне қатысты «Обратный словарь русского

языка» (М., 1974), А.А.Зализняктың «Грамматический словарь русского языка» (М., 1977) және т.б.

Кері сөздіктердің ролі, әсіресе, агглютинативті тілдер үшін ерекше екенін еске сала отыра, олардың саны әлі де болса санаулы. Осындай сөздіктердің алғашқылары ретінде: Қ.Бектаевтың «Қазақ тілінің кері алфавитті сөздігі» (Алматы, 1971), Р.Кунгуров, А.Тихонов «Обратный словарь узбекского языка» (Самарканд,1969, өзбек тілінде) және т.б. кері сөздіктерді атауға болар еді. Аталған сөздіктер мәтінге тікелей қатыссыз, яғни бұрын жарық көрген түсіндірме, орфографиялық және т.б. сөздіктер негізінде жасалғандықтан, оларда сөздердің қолдану жиілігі көрсетілмеген. Ал белгілі бір мәтінге байланысты жасалған кері сөздіктер грамматикалық формантгардың жүмсалуын зерттеуге бірден-бір қажет екені айқын.

Қазақ тілінің кері әліпби-жиілік сөздігінің ең алғаш жарық көруі М.О.Әуезов шығармаларының ЭЕМ арқылы зерттеле бастауымен тікелей байланысты. 1979 жылы «Ғылым» баспасынан шыққан «М. Әуезовтің "Абай жолы" романының жиілік сөздігінің» және 1995 жылы «М.О.Әуезовтің 20 томдық шығармалар мәтіндерінің жиілік сөздіктерінің» жарық көргенін айрықша атап айтуға болады [13, 14]. Осымен қатар кейбір ғылыми, публицистикалық әдебиеттердің мәтіндері бойынша да ірілі-кішілі кері әліпби-жиілік сөздіктері жасалып, оларға біршама лингвостатистикалық талдау жасалды [58].

Жиілік сөздіктің кері әліпби түрінде реестрлік сөзге (сөзтұлғаға) қосылған біркелкі жалғаулар мен жаңа сөз жасаушы немесе сөз түрлендіруші жұрнақтар бір жерге жинақталып берілуі тілді зерттеушілер үшін аса құнды тілдік мәліметтер деп саналады. Кері әліпби-жиілік сөздік морфологиялық категорияларды зерттеу кезінде де және сөз тудырушы, сөз түрлендіруші морфемалардың құнарлы не құнарсыз екендігін анықтау үшін де, терминдер мен басқа тілдерден енген сөздерге жалғанған қосымшалардың құрылымын талдауда да аса бағалы. Осы айтылғандармен қатар, тіл зерттеушіге қажетті деген тілдік мәліметтерді әрі толық, әрі тез түрде тауып алу үшін кері әліпби-жиілік сөздік тілдік қазына деуге болады.

1.4-кесте (үзінді)

Кері әліпби-жиілік сөздік

Рет саны	Сөз (Сөзтүлға)	Абсолютті жиілік	Рет саны	Сөз (Сөзтүлға)	Абсолютті жиілік
1	2	3	1	2	3
1	қалаға	1	71	келуге	1
2	сахнаға	1	72	жіберуге	1
3	қораға	1	73	де	2
4	бақшаға	1	74	лезде	1
5	жолға	1	75	мүлде	1
6	жылға	1	76	білгенде	1
7	буларға	1	------	----------	------
8	тарихтарға	1	445	бірі	1
9	зорға	1	446	есі	1
10	дырдуға	1	447	әкесі	5
11	жазуға	1	448	келесі	1
12	алуға	1	449	шешесі	1
13	ұялтуға	1	450	кісі	1
14	Маланьяға	1	451	көршісі	1
15	да	8	452	тіпті	1
16	бойда	2	453	етті	3
17	түрғанда	1	454	жетті	2
18	мұнда	1	455	кетті	1
19	қалпында	1	456	Петровичті	1
20	басында	2	457	үйленбекші	1
----	----------	-----	458	түземекші	1
68	-	2	459	елші	1
69	үйге	1	460	күтуші	1
70	еденге бірге	1	461	Маланья	3

Кері әліпби-жиілік сөздіктердің лингвистикалық құндылығын И.Штиндлова өзінің мынадай ой-пікірімен білдіреді: «...Обратные словари могут быть использованы не только в качестве метариала для освещения вопросов словообразования, не только как словари рифм или как пособия для восстановления испорченного текста, но они ценны, в первую очередь, при сравнительном изучении родственных языков.

И были бы интересны и для другого типа, например для аналитического языка, или для агглютинативных языков» [97].

Кері әліпби-жиілік сөздіктің әрбір жұрнақ пен жалғаудың (олардың варианттарымен) және әрбір сөз табының белгілі бір мәтіндердегі қолданылу жиіліктерін анықтауда мүмкіндігі мол екендігін тағы да атай отыра, осы аталған сөздіктің үзінді көрінісін *1.4-кесте* арқылы беруді жөн көрдік.

1.6. Сөзнұсқағыш әліпби-жиілік сөздік

Тіл білімінің лексикография саласындағы сөздіктер түзу мәселесі, әсіресе, түсіндірме сөздікті құрастыру барысына тән ең алдымен неше түрлі сөздердің мағынасын ашатын картотекалық қор құру (жинақтау). Картотекалық қордағы әрбір карточкіде тілші-лексикографқа ең бірінші қажетті – «тірек сөз» (лексема) және, екінші ретте, сол сөздің мағынасына қатынасты мәнмәтін (контекст) келтіріледі. Осылармен бірге ол мәнмәтіннің қандай шығармадан алынғаны және оның шығу деректері: қалам иесі (автор), шығарма аты, басма аты, шыққан жылы, көлемі, мәнмәтіннің кездескен бет саны да бірге берілуі керек.

Мәселен, А.Байтұрсынұлы атындағы Тіл білімі институтында көп жылдардан бері жинақталған осындай картотекалық қордың жалпы көлемі 5 млн карточкі-цитата (мәнмәтін) құрайды. Осы қордың негізінде 2 томдық, 10 томдық «Қазақ тілінің түсіндірме сөздігі» [59], «Абай тілі сөздігі» [1] және басқа да сөздіктер жарық көрді. Қазіргі кезде Институттың лексикограф-мамандары осы және басқа да тың картотекалық қордың негізінде – 15 томдық «Қазақ тілінің түсіндірме сөздігін» құрастыру үстінде.

Мұндай көлемді картотекалық қорды толықтыру, түзету немесе ұзақ мезгілге сақтау мен пайдалану жұмыстары тілші-лексикограф үшін оңайға түспейді. Сондықтан да, заманға сай, қазақ лексикография саласын автоматтандыру (немесе компьютерлендіру) кешенді проблемасын шешудің бастамасы ретінде А.Байтұрсынұлы атындағы Тіл білімі институтында «ТІЛ ҚАЗЫНА» атты қазақ тілінің компьютерлік картотекалық базасы (қоры) іске қосылды. Оның негізгі мақсаты – қағаз бетіндегі 5 млн картотекалық қорды қайта сұрыптап, компьютер

жадына, яғни «ТІЛ – ҚАЗЫНА» қорына енгізу және әрдайым ол қорды жаңа деректермен жаңартып отыру.

1.5-кесте (үзінді)

Сөзнұсқағыш әліпби-жиілік сөздік

Рет саны	Сөз (Сөзтұлға)	Абсолютті жиілік	Сөзнұсқағыш (n_i–m_j) (i – кітап беті, j – жол саны)
1	Андреевич	3	030-28; 031-11; 031-39
2	Андреевичке	1	030-20
3	Андреевичтің	1	030-41
4	Анна	3	030; 9; 030-26; 031; 4
5	ағылшын	1	031; 2
6	адам	1	031-22
7	адамдар	2	031-16; 031-36
8	Дидерот	1	030-33
9	Дмитрий	2	031-25; 031-26
10	аз	2	030; 3; 030-18
11	Иван	10	030-11; 030-15; 030-26; 030-30; 030-40; 031-1; 031-10;
12	айғай	2	031-15; 031-23; 031-28
13	айдап	1	030-24; 031; 4
14	айғандай	1	031-14
15	айтты	2	031-22
16	айтып	2	030-35; 031-13
17	айыбы	1	030-40; 031-28
18	айыбын	1	030-37
19	айыпгады	1	030-38
20	айыпты	1	030-30
21	ак	1	030-36
22	акша	1	031; 3
23	акылды	1	031-21
24	акыр аяғы	1	030-11
25	акырын	1	030-19
26	акырып	1	030-13
27	ал	2	031; 9
28	аласа	1	031-14; 031-25
29	алғашқы	1	030; 7
30	алдампаз	1	030-12
31	алмады	1	030-29
32	алмай	1	030-37
33	алмайды	1	030-42

Осы аталған мәселені шешудің тағы бір жолы – «Сознүсқағыш әліпби-жиілік сөздігін» жасау. Бұл сөздіктің жоғарыда сөз болған «Әліпби-жиілік сөздіктен», «Жиілік сөздіктен» және «Кері әліпби-жиілік сөздіктен» негізгі айырмашылығы – баспадан шыққан шығарма мәтініндегі сөздердің кітап беттеріндегі орындарына нұсқау (мегзеу), яғни ол сөзтұлғаның (сөздің) қай бетте және сол беттің қай жолында кездесетініне сілтеме жасау. Сонда зерттеушіге қажет деген сөзге қатысты мәнмәтінді сол сілтеме арқылы кітап беттерінен оңай тауып алуына мүмкіндік туады. Қазақ тіл білімінде осындай әдіс тұңғыш рет А.Байтұрсынұлы атындағы Тіл білімі институтында М.О.Әуезовтің 20 томдық шығармалар жинағының жиілік сөздіктерін алу жағдайында қолданылды. Әрине, мұндай сөздіктің көлемінің үлкен болуы себепті, арнайы сөздік ретінде баспадан шығару оңайға түспесе керек. Ал ондай «Сознүсқағыш әліпби-жиілік сөздікті» тілші қауымының компьютерлік пішінде пайдалану мүмкіндігі мол. Аталған «Сознүсқағыш әліпби-жиілік сөздік» үзіндісімен *1.5-кесте* бойынша танысуға болады.

Жоғарыда аты аталған тәжірибелік жиілік сөздіктеріне мәтіндік нысан болған М.О.Әуезовтің 20 томдық шығармалар жинағының 14-томынан екі беттік үзінді *(Аударма лар, 30–31-бб.)*:

Француша тіпті тұк білмейтін, бірақ оның бұған жұбанышты ғы аз болды. Қашарын білмеді: қыстақтан шықпай бір жыл тұрғанда, сол мезгіл оған он жылға бергісіз болды. Тек шешесінің қасында ғана толас тауып, оның кішкене аласа бөлмесінде, болымсыз әңгімесін тыңдап, варенье жеп қана отыратын. Анна Павловнаның күтуші қыздарының ішінде бір көркем қыз болды, оның өңі нәзік, көзінде момындық бар, ақылды, Маланья дейтін қыз еді. Иван Петровичке сол қыз алғашқы көргеннен ұнады да, соған ғашық болды: қыздың сыпайы жүрісін, ұяң жауабын, ақырын үнін, үнсіз жымиюын бұл түгел ұнатты: күн санап қыз ыстық көрінді. Иван Петровичке қыз да, барлық орыс қызындай, өзінің бар жанымен берілді. Шынымен қосылып еді. Қыстақтағы бай үйінде ешбір сыр, ұзақ уақытқа жасырын боп қала алмайды. Аз уақытта жас мырзамен Маланья жақындығын жұрттың бәрі білді; сол хабар ақыр-аяғы Петр Андреевичке де жетті. Өзге уақыт болса, әкесі мұны болымсыз іс деп елемес еді, бірақ баласына

көптен зығыры қайнап жүрген әкесі енді Петербургтың білгішсіген сәнқой жігітін жерлеп, ұялтуға бекінді. Үлкен айғай, шаң-шуң шығып. Маланьяны апарып қарақшы үйге қамаптты, Иван Петровичті әкесі шақыртыпты. Бұл дырдуға Анна Павловна да келген болатын. Ол күйеуін тоқтаттақ еді, бірақ Петр Андреевич енді сөз тыңдамайтын болды. Баласына қырғидай соқтығып, оны ұятсыз, алдампаз, құдайдан безген деп айыптады, Иван Петрович басында үндемей шыдап көріп еді, бірақ әкесі мұның намысына тиіп, жазалайтынын көрген соң шыдамады. Ішінен "құдайдан безген Дидерот тағы сахнаға шықты ма? Ендеше тоқтай тұрыңыз, мен оны іске жұмсап, сізді таңқалдырайын" деп ойлады. Сол әкесіне дау айтты; ұятсыз деген айыпты мойнына алмады; бірақ ол мүлде айыбы жоқ демеді; енді сол айыбын болымсыз сөздерді елемей, түземекші; қысқасы ол Маланьяға үйленбекші екенін білдіріп, осы сөздерді айтып, Иван Петрович өзінің мұратына дәл жетті. Петр Андреевичтің жаңағы сөзге таңқалғандығы сонша, басында екі көзі бадырайып, үн шығара алмай қалды; бірақ артынан лезде есін жиып, үстіне тиіп ішік киіп тұрған, жылаңаяқ аяғына башмақ киген қалпында жұдырығын түйіп, Иван Петровичке тап берді; жігіт бүгін әдейі де емес, шашын сыпайы тарап, үстіне ағылшын фрак киіп, шашақты етік, ақ кербез ақ шалбар киген екен, Анна Павловна шошынғаннан айғай салып, бетін басты: жігіт зыта жөнелді, қашқан бойда бар үйден өтіп, қорага шығып, бақшаға қарай безіп, одан да өтіп жолға шықты: артына қарамай, зытып барып, әкесінің мұны қуған тысыры мен ауыр дем алысы басылғанша қашты. Әкесі ақырып: "Тоқта, бұзық, тоқта, теріс батамды берем", – деп еді, Иван Петрович тоқтамай кеткен бойда бір көшенің үйіне тығылды. Петр Андреевич қиналып, терлеп, демін зорға алып, үйге келе бере, баласына бата бермейтінін айтты; оның барлық құраған кітаптарын өртеттек болды; ал Маланья қызды қолма-қол алыс қыстақка айдап жіберуге бұйрық етті. Иван Петровичке дос-жар адамдар табылып, оған жайдың бәрін мәлім етті. Ол ұялса да қатты ызаланып, әкесінен кек алуға ант етті. Сол түнде, крестьян арбасына мінгізіп апара жаткан Маланьяны жолшыбай тартып алып, қасына отырғызып, жақындағы қалаға қашып барды да, өздерінің некесін қиғызды. Бұған ақша берген көршісі үнемі ішкілікте

босамайтын, бірақ мейірімді адам болатын. Ол өзі айтқандай осы алуандас ғажайып тарихтарға жаны құмар кісі. Иван Петрович келесі күні әкесіне шағымша тілмен, сыпайылау хат жазды, ал өзі шөберелес жақыны Дмитрий Пестовтың қыстағына кетті: Дмитрий Пестов бізге мәлім Марфа Тимофеевна атты қарындасымен бірге тұратын. Бұларға Иван Петрович өзінің бар жайын айтып, енді Петербургқа қызметке баратынын білдірген және мыналардан уақытша мұның әйелін қолдарына ұстауды сұраған. Әйелім деген жерде бұл жылап та жіберген, өзінің үлкен шаһардағы оқу төрбиесіне, философиясына қарамай, бишара ғана, мүләйім ғана орыстың бірі болып, туғандарының аяғына да жығылып бас иген, еденге маңдайын да тақ еткізген. Пестовтар қайырымды, мейірбан адамдар еді, мұның өтінішіне оңай ризалық білдірді; жігіт мұнда үш жұма тұрып, ішінен әкесінің жауабын күтіп еді, жауап келмеді, келуге мүмкін емес еді.

1.7. Мәтін мен оның жиілік сөздігі бірліктерінің арақатынасы

Тілдік бірліктердің деңгейі жоғарылаған сайын олардың формальды түрде жігін айыру қиындай түсетіні белгілі. Мұндай жағдай тіліміздегі лексика құрамының мағыналық аспектісінің көп жақтылық сипатына байланысты болады. Әсіресе, бұл сөздік қор бірліктерінің ішкі және сыртқы тілдік факторларының әсерінен бір-бірімен өзара тығыз және күрделі түрде әрекеттесуінің нәтижесінен болуы мүмкін. Сөздердің тіліміздегі қолданыс табуы кездейсоқтық оқиға деп есептесек, оның заңды не заңсыз құбылыс екендігі тек ықтималдықтар теориясы мен математикалық статистика әдістері арқылы ғана анықталады. Тек осы ғылымдар саласы ғана қайталанатын және біркелкілік сипаттағы оқиғалар аясымен шұғылдана алады. 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.

Қазақ тілінің кейбір сандық заңдылықтар сапасының сырын ашуға көмектесетін деректерін, негізінен, зерттеушіғалымдар М.Әуезовтің «Абай жолы» романының мәтіні мен жиілік сөздігі негізінде қарастырды. Алынған деректер қазақ мәтінінің құрылымы жайлы дәстүрлі көзқарасқа да және жаңаша түрғыдағы -- статистикалық жолмен зерттеуге де көптеген мәліметтер беруде.

Зерттеу нәтижелері негізінде М.Әуезовтің «Абай жолы» романының мәтіні *465966 сөзқолданыстан* тұратыны анықталды. Ал ол романның әрбір томы (кітабы) бойынша есептелген сөзқолданыстар саны: 1-ші кітапта – *105788 сөзқолданыс*, 2-ші кітапта – *124398*, 3-ші кітапта – *112727*, 4-ші кітапта – *123053*. Бұл сандық шамалар романның әрбір кітаптарындағы сөзтұлғалардың қайталануларын қоса есептеу негізінде алынған мәтін көлемдері (ұзындығы).

Жоғарыда сөз болған жиілік сөздіктер түрін еске алсақ, олар – әліпби-жиілік сөздік, жиілік сөздік және кері әліпби-жиілік сөздік, сөзнұсқағыш әліпби-жиілік сөздіктер деп аталған болатын. «Абай жолы» романының осы аталған сөздіктер түрлері компьютер (ЭЕМ) көмегімен алынып, әрқайсысы бөлек-бөлек лингва-статистикалық зерттеу нысанына айналды.

Осы аталған сөздіктер ішінен арнайы қарастыратынымыз – «Жиілік сөздік» түрі. Оның анықтамасы бойынша, сөздердің орналасу тәртібі жиіліктерінің кему тәртібімен сәйкес келеді және бірдей жиілікті сөздер қатаң әліпби тәртібімен орналасады.

Егер мәтін бірлігі – «сөзқолданыс» термині арқылы аталса, сөздік бірлігі – ***«сөзтұлға»*** не ***«сөз»*** деп аталатынын еске түсіре кетейік.

Сонымен, компьютер көмегімен анықталған деректер бойынша «Абай жолы» романының әрбір сөздігіндегі *сөзтұлға* саны мынадай: 1-ші кітап сөздігінде – *22642 сөзтұлға*, 2-шіде – *26418*, 3-шіде – *26530*, 4-шіде – *27447* ***сөзтұлға***, ал толық роман бойынша (5-ші сөздікте) – *61824 сөзтұлға.*

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 *«1»* санына жақын не алшақ болуына қарай пайымдауға болады [90, 42-43-бб.].

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 айырмашылыққа 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) және ағылшын, орыс, латын тілдерінің публицистика стильдерінің сөздіктері негізінде есептелген K_a=*L:N* қатынас шамалары *1.6-кестедегідей* көрініс тапты.

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.

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 да, ал егер ол «нольге» жуық шама болса, онда ол тілдің синтетикалық сипаты арта түседі деп пайымдалады. Түрлі тілдерге тән мұндай қасиетті *1.6-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.

1.6-кесте

Тілдердің аналитикалық не синтетикалық сипатын K_a коэффициенті арқылы анықтау

Тіл	Стиль	Создер саны	Сөзтүлға саны	K_a
Ағылшын тілі	публиц.	5197	8300	0, 62
Орыс тілі	публиц.	5225	14117	0, 37
Латын тілі	публиц.	13319	45211	0, 30
Қазақ тілі: «Абай жолы»	көркем әдебиет	16983	61424	0, 28

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 біз, өз баяндауымызда, «Абай жолы» романының жиілік сөздіктері мен мәтініне қатысты кейбір мәліметтеріне жүгінеміз.

М.Әуезовтің «Абай жолы» романында қанша сөз немесе сөзтүлға бар және олар қандай жиілікте қолданылған, мәтін мен жиілік сөздіктер бірліктерінің арақатынасы қандай деген мәселелердің басын ашу тек жиілік сөздіктер негізінде ғана іске асады. 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 созқолданыстардан компьютер арқылы жасалатын әліпби-жиілік сөздіктің бірлігі – сөзтұлға. Маман-лексикограф мұндай сөзтұлға сөздікті өз қалауынша бірлігі «сөз» болатын сөздікке (қосымшаларын «қию» арқылы), яғни реестрлік сөз түріне келтіруіне мүмкіндігі бар.

Жиілік сөздіктің құрылымы бойынша жиі қолданыстағы сөздер (сөзтұлғалар) сөздік бойының бас жағында орналасатындығын ескеріп, олардың мәтін мен сөздікті қамту пайызы жайлы мәліметтер *1.7-кестеде* берілді.

1.7-кесте

Бір топ сөзтұлғаның «Абай жолы» романы мәтіні мен сөздік бойын қамту пайызы

Бөліктер рет саны	Рангтық шекаралық	Мәтін бойын қамту пайызы	Сөздік бойын қамту пайызы
1	1	1	0,002
2	1–5	5	0,008
3	1 – 16	10	0,026
4	1 – 33	15	0,053
5	1 – 62	20	0,100
6	1 – 104	25	0,168
7	1 – 168	30	0,272
8	1 – 259	35	0,419
9	1 – 391	40	0,632
10	1 – 842	50	1,362
11	1 – 1749	60	2,829
12	1 – 3731	70	6,035
13	1 – 8430	80	13,635
14	1 – 61824	100	100

Көңіл бөлетін жайт «Абай жолы» романындағы ең жоғарғы жиілікті бірінші реттегі сөзтұлға мәтін көлемінің *1%*, ал сөздік бойының тек қана *0,002* пайызын қамтитыны. Сол сияқты, жиі қолданыстағы *5* сөзтұлға мәтіннің де *5* пайызын, ал олар сөздік бойының *0,008* пайызын ғана қамтитынына сөздік мәліметтері

бойынша көз жеткізуге болады. Сөздік бойының *1,4* пайызын құрайтын *842* сөзтұлға, мәтін ішінде қайталанып қолданудың негізінде мәтіннің *50* пайызын қамтиды екен, яғни бұл романның тең жартысы – *232983* сөзқолданыс. Осы төріздес басқа мәліметтер де *1.7-кестеде* көрініс тапты.

Енді сөздік бірлігі сөзтұлғадан «сөзге» ауысқан жағдайдағы сөздердің түрлі шығармаларда қолданылу жайындағы статистикалық деректеріне тоқталайық. Мәселен, Абай тілінің жиілік сөздігі бойынша Абай шығармалары тіліндегі ең жиі қолданылған *75* сөз барлық сөзқолданыстың *40,8* пайызын, *150* сөз *50,8* пайызын, *1000* сөз *80* пайызын құрайды екен. «Абай тілі жиілік сөздігі» арқылы аталған қолданыстағы сөздердің лингвистикалық табиғатын да, яғни тілдік сыр-сипатын да ашуға болады.

Қолданылу жиілігі *500-ден* жоғары болып кездесетін «сөз» саны «Абай жолы» романда – *144*. Бұл жалпы сөз санының *0,84* пайызын құрайды, ал олардың қолданылу жиілігі жалпы сөз қолданыстың *48* пайызын құрайды. Бұл сандық сипаттама үндіеуропа тілдерінен айтарлықтай алшақ емес. Шынында да, көлемі шамалас кез келген түркі және үндіеуропа тілдеріндегі мәтіндерді алып, ең жиі қолданылатын бір мың сөздің мәтін қамтуының сандық мәнін салыстырсақ, бұл түркі тілдерінде *63* пайыз бен *86* пайыз аралығында, ал үндіеуропа тілдерінде *64* пайыз бен *89* пайыз аралығында екендігі анықталды [12]. «Абай жолы» романындағы ең жиі қолданылған *1000* сөз (сөзтұлға емес) мәтіннің *77* пайызын қамтиды. Сонымен, бұл сандық сипаттама тілдік универсалия (әмбебаптық) міндетін атқаратындығының айғағы.

Бірдей жиілікті сирек қолданатын сөздерді (сөзтұлғаларды) жеке қарастырсақ, ең бірінші байқалатын жайт – жиілігі кеміген сайын ондай сөздердің (сөзтұлғалардың) санының да өсуі. Мәселен, ең сирек қолданыстағы сөздердің (сөзтұлғалардың) кездесу жиілігінің кемуі (азаюы), олардың сөздіктің бірлігі ретіндегі сандық мөлшерінің үлғаюына да әкеліп соғады екен. Сондықтан, *1* рет қолданыста болатын сөздердің саны *2* рет қолданыстан әрі көп, әрі мәтін мен сөздік бойын қамту жағынан да алдыда тұрады. Сол сияқты, жиіліктері *2* реттен болатын

сөздер саны мен мәтінді қамту мүмкіндігі *3* реттегіге қарағанда көптік сипатта. Осы айтылғандар *1.8-кестеде* көрініс тапты.

Сирек қолданыстағы сөздердің статистикалық деректері жайлы «Абай жолы» романы мәтінімен бірге басқа да шығармаларды қарастырайық. Мәселен, газет тіліндегі тек бір реттен қолданылған *5720* сөз (омонимдер ажыратылды) сөздіктегі барлық реестрлік сөздердің *(12424) 46,04* пайызын, ал мәтіндегі барлық сөзқолданыс санының *3,81* пайызын құрайтыны анықталды. Ал Абай шығармалары тілінің бір реттен жұмсалған *2975* сөз барлық сөздіктегі реестрлік сөздердің *(6017) 49,4* пайызын, ал жалпы сөзқолданыс көлемінің *6,35* пайызын құрайтыны анықталды. Бұл фактінің өзі тек публицистика тілі мен көркем шығармалар тілінің ерекшеліктерін ғана емес, Абайдың көркем сөздің шебері және тілінің соншалықты бай екендігін тағы да дәлелдей түскендей.

«Абай жолы» романның әліпби-жиілік сөздігі реестрінде де және жиілік сөздік реестрінде де бірдей *16983* сөз орналасқан. Бұл романда осынша әр түрлі сөз бар (қолданылған) дегенді де білдіреді.

«Абай жолы» романында (*1.9-кесте*) тек *1* және *2* рет қана қолданылған *8698* сөз барлық *16983* сөздің *50* пайызын құраса [13], Пушкин шығармаларында осындай жиіліктегі *9301* сөз барлық *21197* сөздің *44* пайызын [100], «Абай тілі сөздігі» бойынша *3877* сөз барлық *6017* сөздің *64* пайызын [1], Мамин-Сибиряктің «Приваловские миллионы» шығармасындағы *7312* сөз барлық *11283* сөздің *65* пайызын [101] құрайды. Мұндай сирек қолданылатын сөздердің сандық сипаттамасы жазушының тіл байлығы туралы мәлімет береді.

Жоғарыда аталған шығармалардың сөзқолданыс көлемі әр түрлі. Бірдей көлемдегі мәтіндерде сирек қолданылған сөздердің статистикалық сипаттамаларын салыстырып зерттеу арқылы құнды нәтижелер алуға болатыны сөзсіз. Ал ұлы жазушы М.О.Әуезовтің «Абай жолы» роман-эпопеясы бойынша жасалған жиілік сөздіктер деректері тек қазақ тілінің ғана емес, басқа түркі тілдерінің де ақын-жазушыларының сөздік қорын статистикалық лингвистика тұрғысынан зерттеуге бастама және түрткі болуы мүмкін.

«Абай жолы» романы мәтіні және жиілік сөздігі бойынша бірдей жиілікті сөзтұлғалардың қамту пайызы

Сөз-тұлға жиілігі	Реестрдегі сөзтұлға саны	Сөздік бойын қамту пайызы	Жиынтық абсолют-ті жиілік	Мәтін бойын қамту пайызы
1	34860	56, 3	34860	7, 48
2	8739	11, 4	17478	3, 75
3	4261	6, 89	12783	2, 74
4	2540	4, 11	10160	2, 18
5	1810	2, 93	9050	1, 94
6	1184	1, 92	7404	1, 52
7	937	1, 52	6559	1, 41
8	785	1, 27	6280	1, 35
9	638	1, 03	5742	1, 23
10	499	0, 81	4990	1, 07

Енді М.О.Әуезовтің 20 томдық шығармалар жинағындағы 2,1 млн сөзқолданыстан тұратын мәтіні мен 29483 реестрлік сөзден тұратын (жалқы есімді қоспағанда) жиілік сөздігі арасындағы сандық қатынастарға қысқаша тоқталайық. Мұндай сандық қатынастар *1.10-кестеде* көрсетілді.

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 (*1.10)* байқауға болады. Мысалы, жиілігі *10000*-нан астам *17* сөз тізбедегі (реестрдегі) барлық сөздердің *0,0006%* қамтығанымен, олар қайталана 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) не бары *0,122%*-ы екен. Жиілігіне сәйкес басқа топтардың қамту пайыздарын осылайша *1.10-кестеден* анықтауға болады.

Осы кестеден (не болмаса жиілік сөздіктің өзінен) барлық тілдерге (қазақ тілінде де) тән заңдылық байқалады. Ол – сөздердің жиілігі кеміген сайын реестердегі сөз санының

үлғаюы, 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.

1.9-кесте

«Абай жолы» романы жиілік сөздігі бойынша жиілікке сәйкес реестрлік сөз және сөзтұлға сандарының арақатынастары

Реестрлік бірліктің жиілігі	Жиілікке сәйкес сөз саны	Жиілікке сәйкес сөзтұлға саны	Реестрлік бірліктің жиілігі	Жиілікке сәйкес сөз саны	Жиілікке сәйкес сөзтұлға саны
1	2	3	1	2	3
2000-нан жоғары	27	11	11-20	1357	2525
1001-2000	33	23	10	219	499
501-1000	84	60	9	281	638
401-500	28	30	8	331	785
301-400	65	46	7	384	937
201-300	115	112	6	537	1184
101-200	332	318	5	680	1810
51-100	548	661	4	903	2540
41-50	260	350	3	1362	4261
31-40	353	480	2	2329	8739
21-30	591	955	1	6956	34460

Көбіне жазушының сөз байлығын осындай сирек кездесетін сөздердің санымен мөлшерлейді. Шынында да, М.Әуезовтің тың 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. Әрине, басқа да белгілі сөз шеберлерінің жиілік сөздіктері жасалса, олардағы осы құбылыстарды айқындап, салыстыра зерттесе, әр автордың өзіне ғана тән тіл ерекшеліктерін нақты айтуға болар еді. Мұндай зерттеулер болашақтың ісі.

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

1.11-кестедегі сандық көрсеткіштен мәтінде ең көп қайталанып қолданылатын сөз таптары алдымен етістік, екінші орында – зат есім, ал ең аз кездесетіндер одағай мен еліктеу 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, *фарсы, фигыл, фарасат, фақыр, фәле* т.б. Орыс сөздері де әр түрлі фонетикалық нұсқада ұшырайды. Мысалы. *прокурор - пыркорол, протокол – пуртөкөл, генерал – жанарал, приговор – пригоуор – пірғауар т.б.* Кейіпкерлердің тіліндегі айтылу ыңғайына орай жазылған әр түрлі фонетикалық нұсқадағы сөздерге де тізбеде (реестрде) орын берілді. Мысалы. *ажуа* (*55 рет* қайталанып қолданылған), *әжуа – 22, әжуала – 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 кездессе, 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.

1.10-кесте

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

1.11-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	Еліктеуіш (ел)	344	1, 45	2029	0, 12
	Кірме сөздер (бн)	5743	19, 40	45265	2, 65
	Қосындысы:	29483	***Жалқы есім мен кірме сөздерсіз***	1,70619	***Жалқы есімсіз***

М.Әуезовтің шығармалар жинағында жалқы есімдер *386729* 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 таңбаланады. 20 томдық шығармалар жинағы бірнеше жанрда, әр түрлі 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.

386729 рет қайталанған 20 томдық шығармадағы жалқы есімдер жайлы мәліметтер төмендегідей: адам есімдері – *173300* рет, жер-су аттары – *115396* рет, ру-тайпа 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 он бойында адам аттары мен есімдік сөздердің қайталану санының *(173300; 173495)* жуық болуы және атақ-данқ, лауазым атауларының сирек қолданылып, тек мақалалар мен зерттеулерде ғана аздап кездесетіндігі.

Сонымен, М.Әуезовтің 20 томдық шығармалар жинағында қайталанып, әр түрлі сөз таптары мен атаулардың сан мөлшерінің өзіндік қыры-сыры, тілдік таным-түсінігі бар.

Қорыта келе айтатынымыз. 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 болашақта жасалатын толық «Қазақ тілінің жиілік сөздігіне» қосылатын сүбелі үлес болатыны анық.

1.8. Жоғары жиілікті сөздердің лингвистикалық табиғаты

М.О.Әуезовтің «Абай жолы» романының барлық кітабында, шартты түрде алғанда, «аса жиі қолданылатын аймақ» деп аталатын сөзтұлғалардың жиынтығы лингвистикалық зандылық тұрғысынан қарастырылды. Біз осы аталған романның жиілік сөздігіндегі жиі қолданыстағы 100 сөзтұлғаны бөліп алып, оларға статистика-лингвистикалық тұрғыда сипаттама беруді мақсат еттік.

Морфологиялық құрамына қарай жиі қолданыстағы 100 сөзтұлға әр түрлі топтарға бөлінеді. Мәселен, жалғаулық

шылаулар *да//де, мен//пен, және;* сұраулық *ма, ме*, сондай-ақ *соң, сияқты, қарай, үшін, ал, ғана, бірақ, тағы, гой* сияқты сөз байланыстырушылар көмекші сөздерге жатады. Олар тілде жеке бірлік ретінде қолданылмайды, тек синтаксистік байланыс құралдары ретінде қызмет етеді де, тілдегі басқа элементтермен бірге жиі қолданылатындардың қатарына кіреді. Жазушының бұларды жиі қолдануының себебін қазақ тіліндегі сөздердің қатынасқа осы көмекші сөздердің көмегімен түсетіндігімен түсіндіруге болады.

Бұл шығармада жиі қолданылған етістіктер бөліп қарастыруды қажет етеді. Етістік негізді сөзтұлғаларға *е* және *де* түбірлерінен өрбіген *еді, екен, етін*, сондай-ақ *деп, деді, деген* сияқты көмекші етістіктер жатады. Аталған екі етістік те, яғни олардан жасалған көсемше, есімше тұлғалары да статистикалық қызметіне қарай жоғарыда аталған көмекші сөздер сияқты әр түрлі синтаксистік құрылымдарды ұйымдастыруда, мынадай модельдегі етістік+етістік, есім+етістік және басыңқы мен бағыныңқы сөйлемдерді байланыстыруда көмекші етістік қызметін атқарады. Және де «*де*» етістігінен жасалған тұлғалар құрмалас сөйлем ішіндегі төл сөздің ажырамас бөлігі ретінде де қолданылады. «*Е*» және «*де*» етістіктерінің статистикалық жұмсалымдығы оларды жиі қолданылатын тұлғалар қатарына қосып отыр. Қазақ грамматикасының заңдылығына сүйенсек, етістіктің «*е*» түбірі бұйрық рай мағынасында берілуі керек еді, бірақ қазіргі қазақ тілінде ол өзінің бастапқы лексикалық мағынасын жоғалтып алған да, енді қосымша аналитикалық формантты құрамында ғана қолданыла алады. Бірақ етістіктің «*е*» түбірінің бастапқы мағынасын жоғалтуы оның белсенді қолданылуына шек қойып отырған жоқ. Оған оның романдағы аса жиі қолданылысы дәлел бола алады.

Аса жиі қолданылатын аймаққа кезінде, жарты ғасыр бұрын профессор Қ.Қ.Жұбанов алғашқы болып теориялық сипатын көрсетіп берген төрт көмекші етістік те кіреді. Әлі түрлері толық анықталып болмаған барлық күрделі етістіктердің ішінен өте жиі қолданылғыштығы мен сан қырлылығы жағынан *отыр, жүр, тұр* және *жатыр* деген төрт етістікті бөліп айтуға болады. Бұл төрт етістік бір топқа кіріп, басқалардан үш морфологиялық қызметімен ерекшеленіп тұрады [40]. Автордың

айтуынша, бұл етістіктер жақ жалғауларын (I және II жақты) есімдер сияқты бірден қабылдап, нақ осы шақ мағынасын білдіреді және түбір күйінде бұйрық райдың II жағының және нақ осы шақтың III жағының түлғасында тұра береді. Олар көмекші сөз ретінде қолданыла отырып, көсемше түлғасында тұрған кез келген етістік түлғасына көмекші ретінде тіркесіп, нақ осы шақтың мағынасында қызмет атқарады. Бұндай кезде көсемше түлғалы негізгі етістік өзгеріссіз қалып, жақ жалғаулары өздерінің негізгі мағыналарынан ажыраған *отыр, тұр, жатыр, жүр* көмекші етістіктеріне жалғанады. Жазушы тілінде аталған етістіктерден өрбіген сөзтүлғалардың жазылуына осы етістіктердің осы жұмсалымдық қасиеттері себеп болған деуге болады.

Жиі қолданылатын бірінші жүздіктің ішіндегі етістіктердің қатарын грамматикалық көрсеткіштеріне қарай төмендегідей топтауға болады. Олар — *тын* есімшесі *(болатын); -ған//-ген* есімшесі *(келген, болған, алған, қалған); -ып, -іп, -п* көсемшесі *(алып, айтып, болып, келіп, беріп, қарап, боп, кеп); -е* көсемшесі *(кеше); -са* шартты рай түлғасы *(болса, қалса, алса, берсе, келсе, айтса, кетсе).*

Романдағы етістік түлғалардың жиі қолданылатынына қарап, жазушы аталған сөзтүлғалардың көмегімен және көсемше оралымдар, шартты бағыныңқылы құрмалас сойлемдер сондай-ақ ашық райлы баяндауышпен аяқталатын басқа да ерекше сипатталатын тұтастықтар құрастырған деген түжырымдар жасауға болады. Сондықтан ашық райдың септелу қатары көбіне баяндауыштың көрсеткіші ретінде қабылданады.

Жиі қолданылатын сөзтүлғалардың қатарына *бар, жоқ, аз, көп, керек* секілді модаль сөздерді де жатқызуға болады. Заттың нақты мөлшерін, санасын білдірмесе де, бұл сөздер тілде есім сөздердің де, етістіктердің де қызметін атқарады. Синтаксистік тұрғыдан алғанда олар сөйлемнің толық мүшесі ретінде қолданылады: *кітап жоқ, журнал бар, білім көп, ақысы аз, сия керек.* Бұнда олар оздері жеке тұрып та, тіркесіп келіп те күрделі баяндауыштың қызметін атқара береді: *мен айтқан жоқтың, естуім бар, көп болса, білсең керек* және т.б. Соңғы мысалдарда модаль сөздер басыңқы сыңарды мағыналық жағынан толықтырып, нақтылап тұрады. Осындай жұмсалымдығына орай осы

лексикалық топ басқа да лексика-семантикалық құрылымдармен қатар тілде өте жиі қолданылады.

Аса жиі қолданылатын сөзтұлғалардың қатарына есімдіктің түрлерін жатқызуға болады: жіктеу есімдіктері: *мен, сен, ол, оның;* сілтеу есімдіктері: *мынау, осы, бұл, сол;* өздік есімдіктері: *өз, өзі, өзінің,* сұрау есімдігі - *не?* белгісіздік есімдігі - *не.* Олардың жиі қолданылуын түсіндірудің өзі артық. Өйткені кез келген есімдік тілде қолданушының мақсатына қарай есімнің барлық түрінің, сондай-ақ етістіктен болатын есімдердің орнына қолданылады. Есімдіктер мағыналық жан-жақтылығына қарай сөйлемнің тұрлаулы және тұрлаусыз мүшесі де бола алады. Бұның өзі романдағы олардың жиі қолданысын анықтаған біздің статистикалық мәліметтеріміздің дұрыстығына дәлел бола алады.

Үстеулердің ішінен аса жиі қолданылатындардың қатарына негізінен «қашан?» деген сұраққа жауап беретін *енді, қазір, бүгін, әлі* сияқты мезгіл пысықтауыш қызметіндегі үстеулер мен «қандай?» деген сұраққа жауап беретін *ең, дәл* сияқты анықтауыш қызметіндегі үстеулер кіреді. Романның ішкі мазмұнында уақыт факторы маңызды рөл атқарады, сондықтан оның реңктік мәндерінің кейде айқын, кейде бәсең көрінуі заттың сипатын, қимылдың немесе белгілі бір құбылыстың мөлшері мен көлемін білдіруге жоғарыда келтірілген үстеулердің жазушының құралы ретінде жиі қолданылуы көркем шығарманың эстетикалық талаптарынан туындаған.

Аса жиі қолданылатын сөзтұлғалардың қатарына сын есімдер де кіреді. Сапалық – *үлкен, қатты, қалың, жақсы, қара, жас; жаңа, соң* түбірлерінен жасалған туынды сын есімдер *жаңағы, соңғы.* Тілде белгілі бір статистикалық бояуы жоқ болса да бұлардың барлығы да әсіресе сапалы сын есімдер заттар мен құбылыстардың сипатын білдіретіндігіне байланысты шығармада жиі қолданылады. Яғни бұл сөзтұлғаларды жазушы өмірдің бояуын, қоғамдағы құбылыстарды, табиғаттың тамаша сырларын бейнелеу үшін қажетінше жиі пайдаланған. Ал, жалпы, сын есімдердің өзі қолданылуына қарай басқа жиілік аймақтарда да кездесіп отырады.

Біз сипаттап отырған аймаққа зат есімдердің ішінен үш түрлі сөзтұлғалар енеді. Біріншісіне романның басты кейіпкер-

лерінің есімдері жатады – *Абай, Құнанбай, Базаралы, Әбіш, Ербол.* Тағы бір айта кететін жайт кейіпкерлердің *Абайдан* өзге барлығының аты тізімде септелмей берілген. *Абай* есімі атау түлғада келумен қатар ілік *(Абайдың)* және барыс *(Абайға)* түлғаларында келеді.

Зат есімдердің екінші түріне нақты заттардың атауын білдіретін сөздер жатады: *ел, кісі, үй* және абстрактілі зат есімдер: *сөз, күш.* Бұл сөздердің барлығы қазақ тілінің сөздік қорының актив қорына жатады. Тек *күш* сөзі ғана қазір көбіне *өзіндік күш* тіркесінде қолданылады. Осы сөздер бейтарап стильдің лексикасы ретінде шығарманың сөздік құрамын жасауда басты рөл атқарған деуге болады.

Зат есімдердің үшінші түріне *кезде, ішінде* сөздері кіреді. Олар мағынасына қарағанда үстеулерге жақын келгенімен, біз оларды *кез* және *іш* түбірлеріне тәуелдік жалғаулары *і – ің+і* мен табыс септігі *-де/-нде (кезде, ішінде)* жалғанып түрғандықтан зат есімге жатқызамыз. Бұл сөздердің осы жолмен грамматикализациялануы тілде көмекші сөздер сияқты бұлардың да жиі қолданылуына мүмкіндік жасайды.

Сан есімдердің ішінен жиі қолданылатын тізімге *бір, екі, жалғыз* сөзтүлғалары енді. «*Бір*» сөзі қазақ тілінде болсын, жалпы түркі тілінде болсын өзінің бастапқы мағынасын толық та, жартылай да сақтай отырып өте жиі жұмсалынады. Мәселен, *бір адам* деген тіркес адамның *біреу* екендігін және белгісіз *біреу* екендігін білдіреді. Оның қай мағынада екендігін мәнмәтіннен ажыратуға болады. Бұл сөздің осындай лексикалық әмбебап табиғаты оны әр түрлі стильдік мақсаттарда пайдалана беруге мүмкіндік береді. Бірақ *«бір»* сөзінің осы көп мағынасының ішінен статистикалық санауда негізінен сан есім мағынасы есепке алынды.

«Екі» сөзінің *«бір»* сөзі сияқты бірнеше мағынасы жоқ, сондықтан ол тілде тек сан есім мағынасында жұмсалады. Оның романдағы жиі қолданылуын автордың салыстыру, теңдестіру, қарсы қою және т.б. қолдану барысындағы шығармашылық тәсілі деп түсіндіруге болады.

Жалғызбастылықты білдіретін *жалғыз* және *жалқы* сөздерінің эмоционалдық бояуы қанық. Осындай қолданыстарға лайық сөз болғандықтан оның жиі ұшырауы заңды құбылыс деуге болады.

Алынған нәтижелерге байланысты мынадай сұрау туындауы мүмкін: романның әр түрлі кітаптарында сөздердің жиілігі бірқелкі болмаса, оның қосынды түріндегі жиілік сөздігіндегі сөздердің жиіліктік сипатының нақтылы мен дәлдігі қаншалықты? Бірақ жеке кітаптардағы мәліметтерді салыстырғанда, автор өз шығармасындағы стильдік колоритті аяғына шейін сақтағандығын байқаймыз. Жиі қолданылған лексикалық топтар әр кітапта бірдей екендігіне көз жеткізуге болады. Қолдағы мәліметтерге қарағанда 110 сөзтұлғаның романның әр бөлігіндегі қолданысы бірдей. Оларға төмендегідей етістіктер: *емес, болған, отырып, айтып, түсіп, алған, айтқан, алмай, барып, салып, кеткен, көріп, тартып, екенін, берген, барады, болсын, болар, айтады, сөйлеп, көрген, басып, бастап, бере, қалып, қойып, туған, біліп, салды, күле, тауып, айт, кетеді, ертіп, көтеріп, жетіп, десе, атып, жөнелді, кіріп, жеткен, сұрап, білген;* зат есімдер: *ел, күн, Абай, алды, жол, жерде, қасы, үсті, сөзі, көз, көзі, басын, үсті, сыр, ақын, Абайдан, сәт, әңгіме, елге, үйлер, ара, жақ, жел, қолын, бетін, қолына, қан;* есімдіктер: *бұл, өзі, мынау, бірі, соны, өзін, өзіне, кім, міне, оған, одан, қай, қайда, сені, осымен:* сын есімдер: *үлкен, рас, қара, ұзақ, басқа, сондай, бөлек, қазіргі, аппақ, жаман, ашық, артық:* үстеу: *бұрын, тез, әдейі, ақырын, қатар;* көмекші есімдер: *тағы, ме, қана, сайын;* модаль сөздер *керек*. Әр түрлі қолданылатын сөздердің тізіміне қолдану жиілігі орташа аймақтан да кіргізілді. Романның бөлігіндегі әр түрлі қолданыстар статистикалық мәліметтерге қарағанда онша көп емес. Мәселен аса жиі қолданылатын 100 сөзтұлғаның ішінен екінші кітаптан мына зат есімдер кірмей қалды: ***ән,*** кейіпкерлердің аты *Әйгерім, тек* (шектеу мағынасындағы) модаль сөзі, *жігіт* зат есімі, *бүгін* үстеуі және т.б.

Үшінші кітапта жиі қолданылатын 100 сөздің қатарына *Тәкежан, Дәрмен, Оразбай, Мағаш, Әзімбай, Оспан* сияқты кейіпкерлердің, халық атауы *қазақ* сөзі және көмекші сөздердің фонетикалық варианттары *(-та//-те)* және т.б. Алдыңғы кітаптардағы берілген тізімде бұрын кездеспеген *Дәрмен, Оразбай, Мағаш* деген есімдер аса жиі қолданылатын тізімге төртінші кітап бойынша енді. Романның соңғы кітабы бойынша басқа сөздерден алдыңғы еркін тізімге кірмеген *әсіресе, өзгеше* үстеулері мен *әр* есімдігі енді.

Сонымен, бұл келтірілген мысалдар жиілік сөздіктерді жасаудың қажеттігін, олардан алынған статистикалық сипаттамалардың теориялық және қолданбалы тіл білімінде қаншалықты құндылығын көрсетумен қатар, көптеген болжамдар береді. Бұл болжамдардың құнды болуы әр түрлі тілдерден, жеке автор шығармаларынан, белгілі стильдегі мәтіндерден т.б. жиілік сөздіктерді жасаумен және оларды ықтималды-статистикалық тәсілдермен зерттеуге де байланысты.

1.9. Жиілік сөздіктерді компьютер арқылы алудың біріккен және іріленген алгоритмі

Статистикалық лингвистика саласында қарастырылатын мәселелер, көбіне, мәтіннің әр түрлі бірліктерінің қолдану жиілігін білуді қажет етеді. Ол бірліктер – әріп, әріп тіркесі, буын, буын тіркесі, сөз, сөз тіркесі, сөзтұлға, сөзтұлғалар тіркесі және т.б. болуы мүмкін. Бірліктердің мәтіндегі қолдану жиілігін анықтау үшін ең алдымен сол мәтін бойынша жиілік сөздіктер түзіліп алынуы керек. Қарапайым түрде айтқанда, жиілік сөздік дегеніміз, ол қолдану жиілігі көрсетілген тілдік бірліктер тізімі. Жиілік сөздіктегі бірліктердің, яғни сөз не сөзтұлғалардың қалай сұрыпталуына байланысты олардың типтері ажыратылады. Алдыңғы тақырыпшаларда сөздіктер типтеріне (түрлеріне) сипаттамалар берілгенін еске түсірсек, олардың айырым белгілерін білу қиынға түспейді.

Бұл жұмысты «қолмен», яғни бөлек-бөлек қағаз парақтарына жазып, бірдей тілдік бірліктерді бөлектеп сұрыптап, олардың қайталану санын – жиілігін анықтауға және белгілі бір сипатта сұрыптауға да болар еді. Әрине, жиілік сөздікті «қолмен» түзу көп уақытты, күшті және аса ұқыптылықты қажет етеді. Көп жағдайда, ондай сөздіктердің нәтижесі зерттеушіні қанағаттандыра да бермейді. Сондықтан әрі жиі қайталанатын, әрі көп көлемді, әрі есептеуді және әр түрлі сұрыптауларды қажет ететін бірыңғай операцияға жататын міндеттерді (есептерді) адам көмекшісіне – компьютерге жүктеген жөн. Ол үшін жиілік сөздіктерді «қолмен» жасаудағы ретін сақтай отыра, тізбектелген амал-әрекеттің әр қадамы санадан тыс қалмайтындай етіп, «алгоритм» құрастырылады.

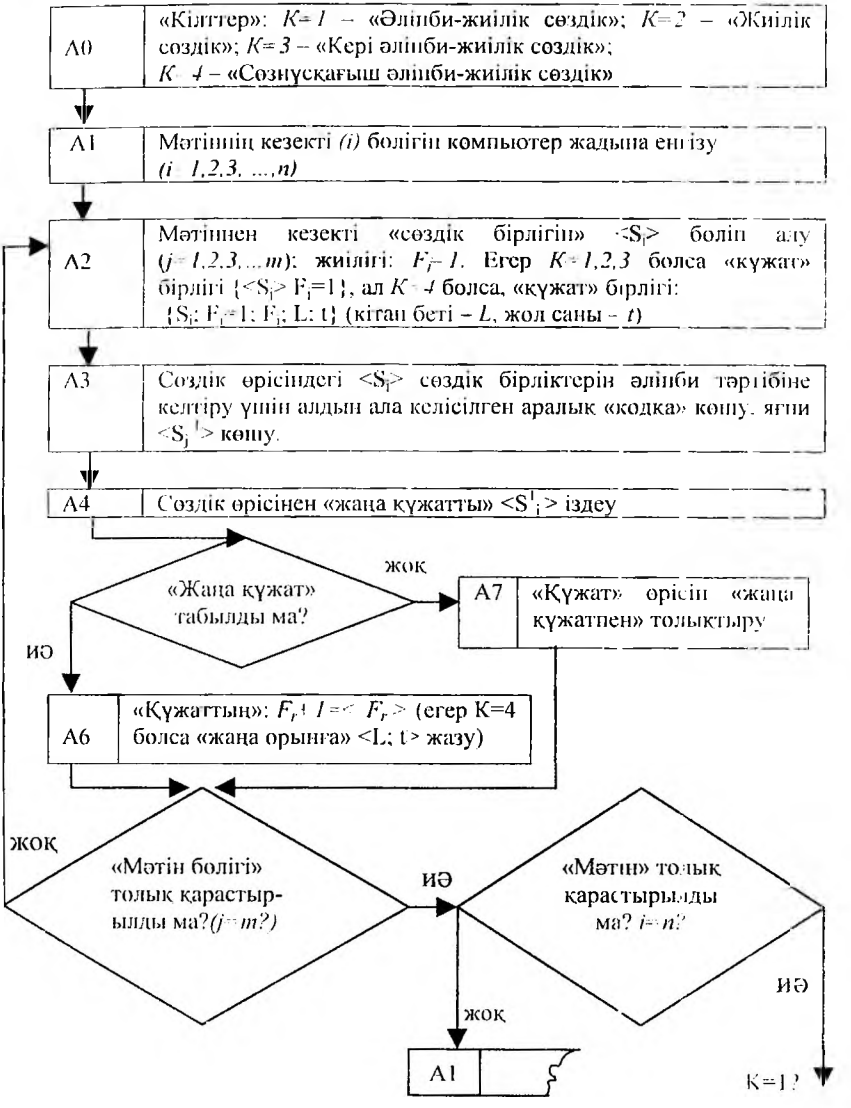
Осы алгоритм негізінде компьютерге арналған арнайы «жасанды тіл» көмегімен маман-бағдарламашы (программист) компьютерлік бағдарлама жазады. Бағдарламаның дұрыс-бұрыстығы тиянақты түрде тексеруден өткеннен кейін ғана (отладка) ол іске қосылады. Қазіргі компьютерлер аталған жиілік сөздіктерді әрі тез, әрі қатесіз, әрі зерттеушінің қалауына сай көрнекі түрде орындай алады.

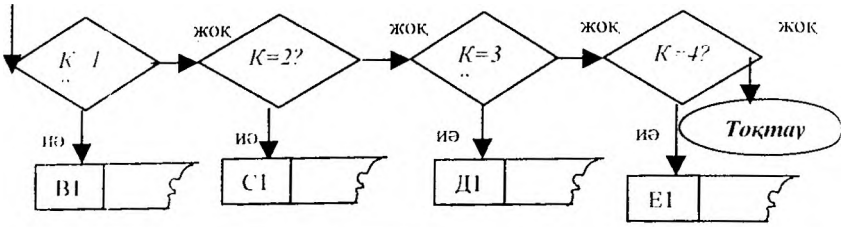
Оқырманды «алгоритм» ұғымымен және оны құрастыру мәселелерімен қарапайым түрде хабардар ету мақсатымен әліппи-жиілік сөздік, жиілік сөздік, кері әліппи-жиілік сөздік және сөзнұсқағыш әліппи-жиілік сөздіктердің біріктірілген және іріленген алгоритмінің көрінісін төмендегі «Сызба-топтама» (блок-схема) арқылы беруді жөн санадық. Әрине, бұл төрт түрлі сөздіктің әрбіреуіне бөлек-бөлек «алгоритмдер» сызба-топтама да жасауға болады. Біздің біріктірілген түрде ұсынуымыз әмбебаптық сипатты қалауды және қысқаша баяндауды көздеуден туындап отыр. Әр сөздіктің арнайы «кілт» (арнайы белгі) арқылы айырым табатынын ескерте отыра, олардың әрқайсысының және ортақ тұстарының сипаттамаларына тоқталайық.

Сызбада көрсетілген «А0» топтамада алдын ала келісілген сөздік типтерін ажыратуға қажетті «кілт» шамалары белгіленеді. Егер бағдарлама жұмысы әліппи-жиілік сөздік алуға бағыттал ған болса, «кілт» мәні $K=1$, ал жиілік сөздік үшін $K=2$, кері-әліппилі жиілік сөздік бойынша $K=3$, сол сияқты сөзнұсқағыш әліппи-жиілік сөздік түзу кезінде «кілт» мәні $K=4$ деп шартты түрде қабылданды. Ал «кілт» мәнінің осы төрт түрлі санның біреуіне тең болу шарты бұзылса – қателік, яғни күтілмеген жағдай (авария) деп саналады да бағдарлама жұмысы тоқтатылады.

Сөздік алынатын мәтін бір немесе бірнеше бөліктерден тұруы да мүмкін, сондықтан компьютер жадына енгізілуге тиісті бөліктер санын қадағалау сызбаның «А1» топтамасында іске асады. Мәтін бөліктері *«i»* мәні бойынша реттеледі.

Жиілік сөздіктерді компьютер арқылы алудың біріккен және іріленген алгоритмінің сызба-топтамасы





C1	Сөздік орісіндегі «құжат» жиілігінің *(F_r-ойң)* кему тәртібімен сұрыптау, ал $F_r = F_{r+1}$ болса, ондай сөздік бірліктерін әліпби тәртібімен сұрыптау және бастапқы «кодқа» ауыстыру
Д1	Сөздік орісіндегі «құжат» бірлігін < S_r > «теріс ретте» жазу (реверстеу), әліпби тәртібімен сұрыптау, бастапқы қалыпқа (ретке) және «кодқа» ауыстыру
E1	Сөздік орісіндегі «құжат» бірлігін S_r әліпби тәртібімен сұрыптау және бастапқы «кодқа» ауыстыру. Сөзнұсқағын сөздікті қалыптастыру: *r* – рет саны; S_r – сөздік бірлігі; *(L; l)* – орны; F_r – жиілігі
B1	Сөздік орісіндегі «құжат» S_r бірлігін әліпби тәртібімен сұрыптау және бастапқы «кодқа» ауыстыру
B2	Сөздікті қалыптастыру: *r* – рет саны; S_r – сөздік бірлігі, F_r; $\sum_{r=1}^{n} F_r$; $f_r = \frac{F_r}{N}$; $\sum_{r=1}^{n} f_r$.
B3	Сөздікті қағазға шығару және компьютер жадындағы «сөздік қорына» қосу

СОҢЫ

Мысалы, мәтін 5 бөліктен тұрса *(n=5), i=1, i=2, i=3, i=4, i=n=5* деп компьютер жадына өз кезегімен енгізіледі.

Сызбаның «А2» топтамасында мәтін бойынан кезекті сөздік бірлігін бөліп алу процесі орындалады. Мысалы, сөздік бірлігі сөз не сөзтұлға болса, оның мәтіндегі шекаралық формальды белгілері: «бос орын» (пробел), тыныс белгілері, тырнақша және т.б. деп саналады. Ал қажетті деген мәтін бірлігі осындай белгілердің арасында орналасқан әріптер тізбегі ретінде формальды түрде бөлініп алынады. Егер мәтін бөлігіндегі сөзтұлға санын *«m»* деп белгілесек, *j=1* болғанда бірінші сөзтұлға S_1, ал *j=2* болса, S_2 және осылайша *j* мәні есіп барып, *j=m* болғанда S_m болады, яғни бұл жағдайда мәтін бөлігіндегі сөзтұлғалар толығымен қарастырылды деген сөз. «Кілт» мәніне қарай, осы топтамада «құжат» бірлігі (жиілік сөздік бірлігі) қалыптасады. Егер, «кілт» мәні *K=1, 2, 3* болса, «құжат» бірлігі $\{S_j; F_j=1\}$, ал *K=4* болса – $\{S_j, F_j=1; L; t\}$ түрінде қалыптасады (*«L»* – кітап беті, *«t»* – беттегі жол саны).

Бүгінгі таңда компьютерлер орыс және латын әліпбилеріне ғана бейімделген, ал қазақ әріптерінің әліпби тәртібімен орналасуына әлі де толық мүмкіншілік жоқ. Осы себептен, қазақ әріптерін әліпбилеуге мүмкіндік жасайтын «аралық кодқа» уақытша көшуге мәжбүр боламыз. Сондықтан сызбаның «А3» топтамасында әрбір «құжат» бірлігіндегі сөзтұлғаны (не басқа бірлікті) аралық «кодқа» ауыстыру жүргізіледі, яғни $<S^1_j>$-ге ауыстыру іске асады. Енді «А4» топтамасының атқаратын қызметі, ол жиілігі бірге тең және аралық «кодта» жазылған «құжатты» осыған дейінгі түзілген (жазылған) сөздік өрісі ішінен іздеу. Нәтижесінде – не табылды, не табылмады деген жауап күтіледі. Осы жауаптар бойынша әрі қарайғы бағдарлама жұмысының бағыты «А5» топтамасында қарастырылады. Егер түзіліп жатқан сөздікте ондай бірлік табылса, онда «А6» топтамасында ол бірліктің жиілігіне «1» саны қосылады. Ал сөздік өрісінен іздеген құжатымыз табылмаған жағдайда «А7» топтамасында сөз жиілігі «1-ге» тең жаңа құжатпен толықтырылады. Бұл екі жағдайдан кейін, яғни «А6» және «А7» топтамалар жұмыстарынан кейін, әрі қарайғы бағыт «А8» топтамасында түйіседі. Бұл жерде мәтін бөліктері толығымен қарастырылып болу-болмауы тексеріледі (яғни $j \leq m$

болуы). Егер мәтін бөлігі толығымен қарастырылып болмаса, бағыт «А2» топтамасына қайтып оралып, жоғарыда аталған жұмыс түрлері (ішкі цикл ретінде) қайталанып, мәтін бөлігі толық аяқталғанға дейін жалғасады. Бұдан әрі қарайғы жұмыс бағыты екіге тармақталады. Біріншісі – мәтін бөлігі толық қарастырылып болған жағдай. Бұл бағыт бойынша бағдарлама жұмысы «А1» топтамасына барып, кезекті мәтін бөлігінен сөздік бірліктерін бөліп алу процесіне сыртқы цикл ретінде қайта оралады. Екінші бағыт, ол мәтін толық қарастырылып болған жағдайға қатысты *(i=n)*. Әрі қарайға жұмыс «кілттің» мәніне қатысты бағытталады.

1. Егер «кілт» мәні *K=1* болса, ол әліпби-жиілік сөздік түзуге байланысты жолмен жүруі қажет. Сондықтан түзілген сөздікті «В1» топтамасында әліпби тәртібімен сұрыптау іске асады да, сосын сөздік бірлігі (сөз не сөзтұлға) алғашқы «код» қалпына ауыстырылады. Ең соңында, яғни «В2» топтамасында қажетті деген сөздік пішіні қалыптастырылып, «В3»-те, зерттеушінің қалауы бойынша, нәтиже қағаз бетіне шығарылады және компьютерлік сөздік қорына қосылады. Соңғы нәтиже компьютер жадында орналасып, әліпби-жиілік сөздікпен (не сөздіктің басқа түрімен) зерттеуші-тілшінің пайдалануына көп жеңілдік келтіретіні айқын.

2. Егер «кілт» *K=2* болса, «С1» топтамасында бағдарлама «жиілік сөздік» түзу үшін, сөздік бірлігінің абсолюттік жиілігінің кему тәртібі бойынша, ал бірдей жиілікті бірліктерді (соз. сөзтұлғаларды) әліпби ретімен сұрыптауды іске асырғаннан кейін сөздік бірлігін алғашқы «код» мәніне ауыстырады. Осыдан кейін, бағдарлама жұмысы және оның қызметі жоғарыда айтылғандардан мәлім, «В2» топтамасына бағышталады.

3. Егер «кілт» *K=3* болса, жұмыс бағыты «D1» топтамасына бағытталады. Ал мұнда «кері әліпби-жиілік сөздік» алу мақсатымен сөздік өрісіндегі «сөз» не «сөзтұлғалар» теріс қарай қайта жазылады – реверстенеді. Бұл әрекет бірліктерді соңғы әріптері жағынан әліпби тәртібімен реттеуге мүмкіндік береді. Әліпби тәртібіне келтіру процесі аяқталғаннан соң, сөздік бірліктері алғашқы қалпына келтіріліп, бастапқы «кодына» қайта ауыстырылады. Ендігі жұмыс түрі жоғарыда айтыл-

ғандардан мәлім, ол «B2» және «B3» топтамаларында атқарылатын соңғы реттегі бағдарлама жұмысы.

4. «Кілт» *K=4* жағдайында әрі қарайғы бағдарлама жұмысы «E1» топтамасынан жалғас табады. Мұндағы жұмыс түрі «B1» топтамасындағыны қайталайды, тек одан айырмашылығы «E1» топтамасында «сөзнұсқағыш әліпби-жиілік сөздік» пішіні қалыптасады, сөздік бірліктерінің кітаптағы кездесу жиілігіне сай, олардың орны: кітап беті *(L)* мен жол саны *(l)* да көрініс табады. Аталған «B1» мен «E1» жұмыс топтамаларындағы қайталаулар сияқты сызбаның «B1», «C1», «D1», «E1» топтамаларында: жиілік сөздіктерді сұрыптау ісі аяқталғаннан соң, сөз не сөзтұлғаны аралық «кодтан» бастапқы «кодқа» келтіру жұмыстары қайталанып, бірдей әрекеттер жасалады.

Сызба-топтама жұмысын сипаттауды қорыта келе айтпақ ойымыз, осы сияқты «сызба» құрудың басқа да ұтымды жолдары болуы мүмкін [31, 30-40-бб.]. Сондықтан сөздік жасауға қажетті компьютерлік бағдарламаның сызба-топтамасы тек осылай ғана сызылу керек екен деген пікір тумауы керек. Ұтымды сызба-топтама жасау шығармашылықты, тапқырлықты, шеберлікті және т.б. қасиеттерді қажет етеді. Себебі әрі қарайғы «сызбаның» «жасанды тілде» көрініс табуы және оның компьютерде ұтымды іске асуы сызба-топтаманың сапалы құрылуына көп байланысты.

Екінші тарау

СТАТИСТИКАЛЫҚ ЛИНГВИСТИКА

2.1. Статистикалық заңдылық және ықтималдық

Біз өмір сүретін ортадағы заңдылықтардың екі түрін арнайы бөліп айтуға болады. Оның біріншісі динамикалық деп аталса, екіншісі статистикалық (ықтималдық) заңдылыққа жатады. Бірінші заңдылық бойынша оқиғаның болу-болмауын алдын ала дәлме-дәл айтуға болады. Мысалы, темірдің суда бататынын, ток қосылғанда электр шамының жарық беретінін, 100° С-да судың қайнайтынын, түннің күнге ауысатынын және т.б.

Ал енді заңдылықтың екінші түрінің нәтижесін алдын ала кесіп айту қиын, оны тек екі аралық арасында, бір *орта шамадан ауытқуын* ескере отырып қана айтуға болады. Бұл заңдылықтың түріне мысал ретінде адамның ми жұмысын, мектептің, үгіт-насихаттың, мәдениеттің адамға әсерін, баланың психикасының дамуын, сөйлеу қызметін, тілдің дамуы мен қызметін және т.т. айтуға болар еді.

Статистикалық заңдылықтың әсерін мынадай қарапайым екі мысалдан анық аңғаруға болады.

Бірінші: *ойынға бейімделген 6 жақты кубті бірнеше рет жоғары лақтыратын болсақ, оның 1-ден 6-ға дейінгі жақтары неше рет түсетінін алдын ала шамамен есептеп шығаруға бола ма?*

Екінші: *екі жақты тиынды бірнеше рет лақтырудан кейін, оның қай жағы неше рет жоғары қарап түсетінін алдын ала шамамен есептеп шығаруға бола ма?*

Сонымен, кубті 600 рет жоғары лақтырдық делік, сонда оның әрбір жағының неше рет жоғары қарап түсуін білу үшін 600 санын 6 санына бөлгенге тең шама болады екен, яғни 600:6=100 рет не болмаса оның шамасы осы 100 санынан не көп, не аз болып ауытқуы мүмкін. Сол сияқты, тиынды 500 рет жоғары лақтырсақ, оның әр жағы 500:2=250 рет шығуы мүмкін немесе осы санға шамалас болады.

Екі мысалда да күтілетін оқиғаның шығуына әр түрлі *кездейсоқ жағдайлар* себепші болады. Мысалы, кубик пен тиынды жасалу пішінінің дұрыс-бұрыстығына, ауа кедергісіне, тәжірибешінің лақтыру күшіне, түсетін жерінің тегістігіне және басқа да көптеген сырт жағдайлардың әсерінен кубтің таңдалған жағы 100 рет, ал тиынның бір жағы 250 рет шығу мүмкіндігінен ауытқиды. *Ауытқу шамасы* кездейсоқтыққа тәуелді болғандықтан, ол ықтималдық заңдылығына бағынады. Мұндай заңдылық математика жолымен, яғни ықтималдық теориясы мен математикалық статистика әдіс-тәсілдері арқылы тағайындалады [17, 29, 30].

Көптеген тәжірибелер жүргізіп, сосын барып ауытқудың орта шамасын есептеуге де болады. Бірақ ықтималдықтың теориясы негізінде тәжірибе жүргізбей-ақ, оқиғаның мүмкіндігін, яғни куб пен тиынды жоғары лақтырған кезде неше рет шығу ықтималдығын (мүмкіндігін) есептеп шығаруға болады.

Айталық, таңдалып алынған кубтің не тиынның кез келген жағының шығуын «оқиға» деп ат қойып, оның атын «*A*» деп белгілесек және «*n*» саны арқылы «оқиғаның» *тәжірибе үстінде* орындалу (қалаған жақтың шығу) санын белгілесек, «*n*» арқылы тәжірибенің (жоғары лақтыру) жалпы санын, яғни элементар оқиғалар санын белгілесек, онда *оқиғаның ықтималдығы* классикалық анықтама негізінде мына өрнекпен анықталады [17, 16-22-бб.]:

$$P(A) = \frac{m}{n}. \qquad (1)$$

Бұл теңдіктегі $m \le n$ болғандықтан $P(A) \le 1$.

Егер *күткен оқиға* (тиынның не кубтің таңдалған жағы) тәжірибе үстінде бірде-бір рет шықпаса (пайда болмаса), яғни

жағдайдың барлығы да күтетін оқиғамызға «қолайсыз» болса, онда $m=0$ және $P(A)=\frac{0}{n}=0$ болады.

Егер тәжірибенің өн бойы күткен оқиғамыз үнемі шығып отырса, яғни жағдайдың бәрі де күтілетін оқиға үшін қолайлы болса, онда $m=n$ және $P(A)=\frac{m}{n}=\frac{n}{n}=1$ болады.

Айтылғандарды қорыта келе (1) өрнекке қатысты мынандай анықтама беруге болады: егер $P(A)=0$ болса, ондай оқиға -- *«мүмкін емес оқиға»* деп, ал $P(A)=1$ болса, *«ақиқат оқиға»* деп аталады.

Сонымен, *ықтимал оқиғаның* барлық жағдайын қамтитын математикалық өрнек түрі, ол $0 \le P(A) \le 1$ деген теңсіздікпен көрініс табады.

Мысалы: *М. Әуезовтің «Абай жолы» романының 2-ші кітабында 124398 сөзқолданыс бар. Оның 11467 сөзқолданысы сын есім сөздер екен. Сонда 4 кітаптан тұратын «Абай жолы» романында сын есімнің кездесуінің ықтималдығы қанша?* – деген сұрақ туындауы мүмкін. Бұл сұраққа $P(A)=\frac{m}{n}$ өрнегі жауап береді: $n=124398$, $m=11467$, A – сын есімнің кездесуін білдіретін оқиға.

Сонда, сын есім сөздердің роман бойында *кездесу ықтималдығы* – $P(x)=\frac{m}{n}=\frac{11467}{124398}=0,092\approx 0,1$ шамасына тең болады. Бұл шаманың аз не көп екендігін білу үшін «0» мен «1» аралығымен салыстыру керек. Егер $P(x)$ мәні «1» санына неғұрлым жақындаған сайын, ол оқиғаның шығу мүмкіндігі артады.

Оқиғаның ықтималдығын анықтаудың ең қарапайым *статистикалық құралы* ретінде ол оқиғаның кездесу *«жиілігін»*, *«орта жиілігін»* және *«орта жіліктен ауытқу»* деп аталатын шамаларды табу жолдарын айтуға болады.

Бұл жердегі *«жиілік»* терминін *«болмыстың бөлігі»* ішіндегі байқауға алынған оқиғаның (қайсыбір тілдік бірліктің) кездесу саны деп түсіну қажет. Ал *«болмыстың бөлігі»* ретінде, мысалы, үлкенді-кішілі көлемдегі мәтін бөлігі алынуы мүмкін.

Айталық, ойын кубін *1000* рет жоғары лақтырғанда оның *«бір»* деген белгісі бар жағы *170* рет шықты десек, онда осы сан *(170) «бірдің кездесуі»* жайлы оқиғаның *«жиілігі»* деп есептеледі. Сол сияқты, М. Әуезовтің «Абай жолы» романында түбір есімдік *20811* рет, ал оған тәуелдік жалғауы жалғанған түлғасы (формасы) *2686* рет кездеседі делік. Осы келтірілген *20811* бен *2686* сандары есімдік сөздердің роман мәтіні бойынша есептелген *«жиілігі»* болып саналады.

Әдетте, статистиктер тілдік бірліктерге қатысты *оқиғаның заңдылығын* ашу кезінде бір саладағы барлық мәтіндерді түгелімен (*генералды жиын* немесе *бас жиын*) қамтуды мақсат етпейді, себебі ол мүмкін де емес. Сондықтан зерттеуші ол мәтіннен тек сынаққа түсетін белгілі көлемдегі үлгілерімен ғана қанағаттанады. Міне, осындай мәтін бөліктерін *«таңдама бөліктер»* немесе *«таңдама жиындар»* (выборки) деп атайды. Әрине, мұндай бөліктер көлемі жағынан бас жиыннан айтарлықтай аз мөлшерде болады және бөліктер саны бірнешеу болған жағдайда, олардың көлемдері тең не жуық шамада болуы керек. Осындай *«сынама үлгілер»* бойынша тағайындалған «оқиғаның» статистикалық заңдылығы бас жиынға да таралады. Мәтіннің осындай *сынама үлгілері* (немесе *таңдамалары*) бойынша есептелген «оқиғаның» кездесу жиілігін *«таңдама жиілік»* деп атайды. Мысалы, көркем әдебиет стилі *бас жиын* деп саналса, М.Әуезовтің 20 томдық шығармалар жинағы *таңдама бөлік* болып саналады. Ал егер бас жиын ретінде М.Әуезовтің 20 томдық шығармалар жинағының мәтіндері алынса, онымен салыстырғанда «Абай жолы» романының бір ғана томының мәтіні *«таңдама бөлік»* болып есептеледі. Осы *таңдама бөліктен* анықталған тілдік бірліктің жиілігі, мысалы түбір есімдіктің жиілігі, *таңдама жиілік* деп аталады. *Таңдама жиілік* жеке тұрып «оқиғаның» ықтималдығы мен статистикалық заңдылықтары жайлы зерттеушіге қанағаттандырғысыз ақпарат беруі мүмкін. Ал егер біз *«орта таңдама жиілік»* немесе *«орта жиілік»* деген терминдермен аталатын шаманы анықтайтын болсақ, бұл жағдайда *«оқиға»* туралы ақпарат ұлғая түседі.

Орта таңдама жиілікті есептеудің көптеген жолдары ішінен біз тек қана ең қарапайым түрін қарастырмақпыз [17; 29; 30, 22-б.].

Айталық, зерттеуге жататын мәтін ішінен құрылымы жағынан біркелкі келетін және бірдей көлемдегі (үзындықтағы) бірнеше *таңдама мәтін үзінділерін* зерттеу нысаны ретінде алдық делік. Мысалы, таңдалып алынған зерттеу нысаны – *500 негіз сөз* тұратын *10 мәтін бөліктері (үзінділері)* болсын. Енді зерттеу *«оқиғасы»* ретінде «зат есімнің кездесуін» «*x*» деп белгілейтін болсақ, оның *«орта жиілігін»* есептеу үшін *10* мәтін үзінділердің әрбіреуіндегі зат есім сөздердің кездесу жиілігін санап шығуымыз керек болады. Егер ондай жиіліктерді $x_1, x_2, \ldots, x_{10}$ деп белгілесек және тәжірибе жүзіндегі олардың сандық мәндері: $x_1 = 182$, $x_2 = 187$, $x_3 = 218$, $x_4 = 173$, $x_5 = 158$, $x_6 = 201$, $x_7 = 222$, $x_8 = 233$, $x_9 = 213$, $x_{10} = 194$ шамаларына тең деп есептейік.

Енді *«орта жиілікті»* $\overline{X}$ және мәтін үзінділерінің санын n деп белгілесек, *орта жиілікті* есептеп шығару үшін мынандай амалдарды орындауымыз керек:

а) оқиғаның кездесу жиіліктерінің қосындысын табу:

$$x_1 + x_2 + \ldots + x_{10} = 182 + 187 + \ldots + 194 = 1981;$$

ә) *«орта жиілік»* шамасын немесе *арифметикалық ортаны* табу:

$$\overline{X} = \frac{x_1 + x_2 + \ldots + x_{10}}{n}. \qquad (2)$$

Жоғарыда келтірілген есеп берілісі бойынша зат есімдерге қатысты *«орта жиілік»* (2) өрнек арқылы есептеп шығару керек болса, ол мына түрде табылады: $\overline{X} = \frac{1981}{10} = 198,1 \approx 198.$

Сонымен, алынған сандық нәтиженің мәні мынада: мәтін ішіндегі әрбір 500 негіз сөзге, орта есеппен алғанда, 198 зат есім сәйкес келеді деп саналу қажет.

Егер математика пәнінде қосындыларды мынадай «$\sum$» таңбамен белгіленетінін ескерсек, онда: $x_1 + x_2 + \ldots + x_{10} = \sum_{i=1}^{10} x_i$ түрінде жазылады. Онда *«орта жиілікті»* анықтаудың (2) өрнегін қысқа түрде былай жазуға болады:

$$\overline{X} = \frac{1}{n} \cdot \sum_{i=1}^{n} X_i. \qquad (3)$$

Бұл (3) өрнек *«орта таңдама жиілікті»* (немесе *«орта жиілікті»*) анықтау үшін қолданылады.

Есеп: *Жоғарыда қарастырған таңдама мәтіндер ішінен сын есімдердің кездесу жиіліктері:* $x_1=69$, $x_2=71$, $x_3=83$, $x_4=50$, $x_5=43$, $x_6=73$, $x_7=72$, $x_8=59$, $x_9=69$, $x_{10}=71$. *Сын есім сөздердің «орта жиілігін» табу керек. Ягни егер* $n=10$ *және сын есімнің кездесу жиіліктері белгілі болса,* $\overline{X}=?$ (жауабы: $\overline{X}\approx 67$).

Тілді статистикалық тәсілмен зерттеуде *«орта жиілік»* шамасын білу аса маңызды деп саналады. Зерттеушінің таңдап алған тілдік *«оқиғасының»* статистикалық заңдылығын айқындау немесе оның ықтималдығын білу осы (3) *«орта жиілік»* өрнегі арқылы іске асады. Сандық мәні белгілі болған *орта жиілікті* әрі қарайғы зерттеуімізге пайдалану үшін *таңдама жиіліктің* (x_i) *орта жиіліктен* ($\overline{X}$) *ауытқу* шамаларын анықтау қажет болады.

Егер *«таңдама жиілік»* мәні *«орта жиіліктен»* кіші болса, ауытқудың таңбасы – теріс (минус), ал керісінше жағдайда – оң (плюс) деп саналады. Мұндай ауытқу шамалары жеке тұрып, зерттеушіге беретін мағлұматы аз болуы мүмкін. Сондықтан ауытқуларды да жалпылайтын *«ортаны»* табудың қажеттігі туындайды. *Математикалық статистика* пәнінде мұндай *«ортаны»* табудың екі түрлі жолы кездеседі:

а) *ауытқулардың абсолют шамаларының ортасын* табу. Бұл жағдайда теріс таңбалы айырым шамалар оң таңбаға ауыстырылып, содан кейін барып ауытқулардың ортасы табылады. Ал ондай ортаны табу үшін абсолютті ауытқулардың қосындысын таңдама бөліктердің санына бөлу қажет:

$$\overline{a}=\frac{\sum / x_i-\overline{x} /}{n}; \qquad (4)$$

ә) немесе екіге дәрежеленген ауытқудың орта шамасын мынандай өрнекпен анықтау қажет:

$$\sigma=\sqrt{\frac{\sum (x_i-\overline{x})^2}{n}}. \qquad (5)$$

Мұндағы σ *(сигма) – квадраттық ауытқудың ортасы (среднее квадратичное отклонение)* және $x_i-\overline{x}$ *– таңдама жиіліктің орта шамадан ауытқуы* деп аталады.

Егер $x_i - \bar{x} = a_i$ арқылы белгілейтін болсақ, онда *квадраттық ауытқудың ортасы*:

$$\sigma = \sqrt{\frac{\sum a_i^2}{n}}. \qquad (6)$$

Бұл (6) өрнекті (формуланы) сынақ таңдамалардың (бөліктердің) ұзындығы (көлемі) өзара тең болған жағдайда пайдалану қажет (мысалы, *500* сынақ таңдамалардың әрбіреуі *100* негіз сөзден тұрады).

2. Жоғарыдағы (5) теңдіктің екі жағын бірдей квадраттайтын болсақ:

$$\sigma^2 = \frac{\sum (x_i - \bar{x})^2}{n}. \qquad (7)$$

Математикалық статистика пәнінде (7) өрнек «*дисперсия*» деген атпен аталады. Кейбір тәжірибе кезіндегі есеп-қисаптарда «квадраттық ауытқу ортасы» (6) өрнегінен гөрі *дисперсия* өрнегін (7) пайдаланған ыңғайлы деп саналады.

Есеп: *Бір мәтіннен 5 сынақ таңдама бөлік алынған. Әрбір бөлік 500 атаушы сөздерден тұрады. Оларда етістіктердің кездесу жиіліктері мынадай:* $x_1 = 95; x_2 = 87; x_3 = 94; x_4 = 104; x_5 = 100.$

Орта квадраттық ауытқуды табу керек.

Шешімі: *Орта квадраттық ауытқудың* өрнегі:

$$\sigma = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n}}; \qquad 1 \le i \le 5;\ n = 5;$$

$$\bar{X} = \frac{95 + 87 + 94 + 104 + 100}{5} = \frac{480}{5} = 96;$$

$x_1 - \bar{X} = 95 - 96 = -1;\ x_2 - \bar{X} = 87 - 96 = -9;\ x_3 - \bar{X} = 94 - 96 = -2;$
$x_4 - \bar{X} = 104 - 96 = 8;\ x_5 - \bar{X} = 100 - 94 = 4;\ (x_1 - \bar{X})^2 = 1;$
$(x_2 - \bar{X})^2 = 81;\ (x_3 - \bar{X})^2 = 4;\ (x_4 - \bar{X})^2 = 64;\ (x_5 - \bar{X})^2 = 16;$

$$\sigma = \sqrt{\frac{1 + 81 + 4 + 64 + 16}{5}} = \sqrt{\frac{166}{5}} = \sqrt{33,2} = 5,76.$$

$\sigma = 5,76$ – *орта квадраттық ауытқу.* $\sigma^2 = 33,2$ – *дисперсия.*

$$\overline{X} - \sigma \le \overline{X} \le \overline{X} + \sigma;\ 90{,}24 \le \overline{X} \le 101{,}76.$$

Осы есептің шешімін төмендегідей кесте арқылы көрнекі түрде де беруге болады:

Сынақ таңдамалар	Етістік создер		
	x_i	$a_i = x_i - \overline{X}$	$a^2_i = (x_i - \overline{X})^2$
1-ші	95	-1	1
2-ші	87	-9	81
3-ші	94	-2	4
4-ші	104	8	64
5-ші	100	4	16
Қосындысы	480	0	166
	$X = 96$		
	$\sigma = 5{,}76;$		$\sigma^2 = 33{,}2$

Есеп: *Алдыңғы есептің берілісі бойынша зат есім создердің жиілігі белгілі делік.*

Олар: $x_1 = 199,\ x_2 = 205,\ x_3 = 195,\ x_4 = 201,\ x_5 = 210,\ n = 5.$

$$\sigma = ? \qquad \sum_{i=1}^{5} x_i = 1010,\ \ \overline{X} = \frac{1010}{5} = 202,\ \ \sum_{i=1}^{5} a_i^2 = 132,$$

$$\sigma = \sqrt{\frac{132}{5}} = \sqrt{26{,}4} = 5{,}14.$$

Осы есептің шығару жолын кесте түрінде де көрсетуге болады.

Сонымен, жоғарыдағы баяндауымызда математикалық статистиканың ең маңызды әрі қарапайым түрдегі алты ұғымымен және олардың аталу терминдерімен таныс болдық.

Олар:

1) ықтималдық; 2) статистикалық заң; 3) таңдама жиілік; 4) орта таңдама жиілік; 5) ауытқулардың абсолютті ортасы; 6) орта квадраттық ауытқу (дисперсия).

Статистикалық әдісті қолданғысы келген тілші-зерттеушіге, алғашқы қадам ретінде, ықтималдықтар теориясы мен математикалық статистиканың осы аталған ұғымдары мен сипатталған қарапайым санау құралдары жеткілікті деуге болады.

2.2. Таңдама жиіліктер айырымдарын статистикалық бағалау

Жазба не сөйлеу тілін статистикалық жолмен зерттеу, әдетте, бірдей көлемдегі таңдама бөліктер мәтіндері негізінде жүргізіледі. Осындай жағдайда зерттеу нысаны ретінде алынған бір ғана «оқиға» (мысалы, қайсыбір тілдік бірлік) әр таңдама бөліктерінде кездесу жиіліктерінің (таңдама жиіліктерінің) әр түрлі шамада болуы сол бірліктің статистикалық заңдылығының ақпараттық деректері бола алады.

Айталық, қазақ тіліндегі ғылыми-техникалық шығармалар мәтінінен *5000* сөзқолданыстан тұратын таңдама сынаққа алынды делік. Осы мәтін көлеміндегі сөзқолданысты теңдей етіп *10-ға* бөлсек, *500* сөзқолданыстан тұратын *10* таңдама мәтін бөліктері шығады, яғни $N=5000$, $n=500$, $k=10$. Әрбір таңдама бойынша есептелген етістіктің таңдама жиіліктері: $x_1 = 98$, $x_2 = 87$, $x_3 = 102$, $x_4 = 105$, $x_5 = 123$, $x_6 = 108$, $x_7 = 85$, $x_8 = 78$, $x_9 = 110$, $x_{10} = 104$;

Етістік жиілігінің бұл ауытқулары тілші-тәжірибешіге қандай мәлімет беруі мүмкін немесе жағдай етістік сөздердің ғылыми-техникалық шығармалар стилі үшін заңды не заңсыз құбылысқа жата ма?

Егер жиіліктер арасындағы мұндай ауытқулар статистика тұрғысынан заңды болса, оларды кездейсоқтық жағдайға қатысты деуге болады. Ал егерде таңдама жиіліктің орта жиіліктен ауытқуы елеулі сипатқа енсе, онда оларды статистикалық заңдылыққа бағынбау салдарынан немесе етістік сөздердің қолдану ықтималдығының тұрақсыздығынан деп түсіну керек. Себебі, жиілік шамасының құбылуы кездейсоқтық сипатқа ие болмауынан және ауытқу дәрежесінің елеулі болып бір ғана ықтималдыққа бағынбауы салдарынан болып отыр. Сондықтан бұл оқиға (етістік сөздердің кездесуі) ешбір статистикалық заңдылыққа бағынбайды деп саналады. Тәжірибе кезінде зерттеуші осындай жағдайдың себебін анықтай білуі керек. Яғни оқиғаның (біздің мысалда етістіктің жиіліктері) қолданылуының орта жиіліктен ауытқуы елеулі ме, әлде ол кездейсоқтық жағдайға қатысты ма?

Міне, осындай сауалдың жауабын табу жолдарына тоқталайық.

Аталған мәселенің басын ашу математикалық статистика пәнінің міндеті. Дәлірек айтсақ, ол үшін аталған саланың *«χ^2 критерийі»* (оқылуы – *хи квадрат*) деп аталатын статистикалық құралын пайдалануымыз керек. Бұл критерий *«Пирсонның χ^2 келісім критерийі»* деп те аталады (үшінші тарауда бұл критерийге тағы да тоқталамыз).

Тәжірибелік мәтіннен алынған таңдама бөліктердің көлемі (жоғарыда баяндалған мысалдағыдай) тең болып келетін жағдайда χ^2 критерийінің өрнегі төмендегідей жазылады [17, 29, 30, 28-36-бб.]:

$$\chi^2 = \frac{\sum_{i=1}^{n} (x_i - \overline{x})^2}{\overline{x}} . \qquad (1)$$

Осы (1) өрнектегі x_i – таңдама жиіліктер, $\overline{X}$ – орта таңдама жиілік, Σ – қосу белгісі. Егер таңдама жиіліктің орта жиіліктен ауытқуын, яғни $x_i - \overline{X} = a_i$ деп белгілесек, онда χ^2 критерийі мынадай шағын түрде көрініс табады:

$$\chi^2 = \frac{\sum_{i=1}^{n} a_i^2}{\overline{X}} . \qquad (2)$$

Өрнектің оқылуы: *хи квадрат* – таңдама жиіліктің орта жиіліктен ауытқу квадраттарының қосындысын орта жиілікке бөлгенге тең.

Ықтималдықтың бір ғана тұрақты мәніне байланысты χ^2 шамасы *«жиіліктің үлестірілу заңдылығына»* бағынады. Яғни ықтималдықтың бір ғана тұрақты мәніне қатысты χ^2 *(хи квадрат)* шамасының әр түрлі жиі, сирек не тіпті сирек болып кездесу мәндері сәйкес келуі мүмкін екен. Осыны ескере отырып, математика мамандары *«хи квадрат»* мәндерінің теория бойынша мүмкін шамаларын анықтап, осы саладағы тиісті әдебиеттерде олар кесте деректері түрінде беріледі. Осы деректер арқылы зерттеуші таңдама жиіліктің орта жиіліктен ауытқуын бағалап, оның статистикалық заңға бағыну не бағынбау дәрежесін анықтай алады. Сондықтан да χ^2 критерий, кейде «χ^2 келісім критерийі» деп те аталады. Әрине, бұл жерде «немен» келісу керек екенін түсініп алу қажет. Енді осы мәселеге толығырақ тоқталайық.

Бұл сауалды тәжірибе жүзіндегі деректерге байланысты есептеп шығарылған χ^2 шамасын, оған сәйкестікте тұратын арнайы кестеде келтірілген χ^2-тың теориялық мәнімен салыстыра отырып, келісім жасалады деп түсіну қажет.

Таңдама жиіліктің орта жиіліктен ауытқуының әр түрлі мәндеріне байланысты χ^2-тың тиісті шамалары *2.1-кестеде* көрсетіледі.

Аталған *2.1-кестенің* χ^2-тың деректерін, яғни теориялық мәнін түсіндіру үшін жоғарыда келтірілген ғылыми-техникалық шығармалар мәтініндегі етістік сөздердің кездесу жиілігі бойынша қарастырайық: *(N=5000, n=500, k=10): x_1=98, x_2=87, x_3= 102, x_4= 105, x_5=123, x_6=108, x_7=85, x_8= 78, x_9=110, x_{10}=104;*

2.1-кесте

χ^2 критерийінің теориялық мәні

Еркіндік дәреже саны	Ықтималдықтың үлкен шамасына сай келетін χ^2 (хи квадраттың) теориялық мәні					
	0,95 (95%)	0,75 (75%)	0,50 (50%)	0,25 (25%)	0,10 (10%)	0,05 (5%)
1	-	0,10	0,45	1,32	2,71	3,84
2	0,10	0,58	1,39	2,77	4,61	5,99
3	0,35	1,21	2,37	4,11	6,25	7,81
4	0,71	1,92	3,36	5,39	7,78	9,49
5	1,15	2,67	4,35	6,63	9,24	11,07
6	1,64	3,45	5,35	7,84	10,64	12,59
7	2,17	4,25	6,35	9,04	12,02	14,07
8	2,73	5,07	7,34	10,22	13,36	15,51
9	3,33	5,90	8,34	11,39	14,68	16,92
10	3,94	6,74	9,34	12,55	15,99	18,31
14	6,57	10,17	13,34	17,12	21,06	23,68
15	7,26	11,04	14,34	18,25	22,31	25,00
19	10,12	14,56	18,34	22,72	27,20	30,14
20	10,85	15,45	19,34	23,83	28,41	31,41
24	13,85	19,04	23,34	28,24	33,20	36,42
25	14,61	19,94	24,34	29,34	34,38	37,65
29	17,71	23,57	28,34	33,71	39,09	42,56
30	18,49	24,48	29,34	34,80	40,26	43,77

Орта таңдама жиілікті анықтау өрнегі:

$$\overline{X} = \frac{x_1 + x_2 + \ldots + x_{10}}{n}; \quad \overline{X} = \frac{98 + 87 + \ldots + 104}{10} = \frac{1000}{10} = 100;$$

Енді (1) орнек бойынша: $\chi^2 = \frac{(-2)^2 + (-13)^2 + \ldots + 4^2}{100} =$

$$= \frac{4 + 169 + 4 + 25 + 529 + 64 + 225 + 484 + 100 + 16}{100} = \frac{1620}{100} = 16.2$$

Сонымен, есептеп шығарылған χ^2-тың тәжірибелік мәні 16,2 тең болды. Кестедегі бірінші «тік бағанадағы» сандар – *«число степеней свабоды»* немесе қазақша *«еркіндік дәреже саны»*, – деп аталады *(А.Ж.)*. Оның мәні таңдама мәтін бөлігінің санынан *(k)* бірді кеміткенге тең болады. Егер *еркіндік дәреже саны* – «ν» деп белгілесек, онда $\nu = k - 1 = 10 - 1 = 9$;

Ал тік бағанадан таңдалып алынған $\nu = 9$ санына сәйкес жатық жолдағы сандар ішіндегі χ^2 теориялық мәнімен тәжірибелік χ^2 мәні салыстырылады.

Нәтижесінде, тәжірибелік χ^2-қа тең не одан үлкендеу шамасы белгіленіп, сол бағана бойындағы (жоғарыдағы), ықтималдық таңдалады.

Еркіндік дәреже саны – $\nu = 9$ мәніне және тәжірибелік $\chi^2 = 16.2$ мәніне шамалас кестедегі теориялық χ^2-тың *16.92* мәні және ықтималдықтың *0,05 (5%)* сай келеді екен. Бұл жерде «хи квадрат» мәнінің тәжірибелік шамасы теориялықтан кіші екендігін ескере отырып, былай пайымдауға болады: оқиғаның шығуының барлық 100 мүмкіндігінің тек 5 жағдайында ғана етістік жиіліктерінің жоғарыда көрсетілгендей ауытқу шамаларының кездесуі мүмкін. Сондықтан ондай ауытқуларды *кездейсоқ оқиғаға* жатқызып, етістіктің таңдама жиіліктері статистикалық заңға қайшы келмейді және етістіктің 100-ге тең орта жиілігі 500 сөзқолданыстан тұратын мәтінге тән жиілік деуге болады. Ал етістік сөздердің мәтінде кездесу ықтималдығы $P(x) = \frac{100}{500} = 0,2$ шамасына тең. Сонымен, бұл есептің нәтижелері бойынша тәжірибелік ғылыми-техникалық шығармалар мәтінінің барлық сөзқолданысының 20 пайызы етістік сөздерден тұрады деп қорытындылауға болады.

Статистик мамандар таңдама жиіліктің орта жиіліктен ауытқуының *елеулі* не *елеулі еместік* сипатын анықтау үшін χ^2 мәнін кесте бойынша екі ықтималдық шамаларының аралығында қарастырады, яғни *0,95-тен 0,05* аралығы. Осы

аралықтан алынатын χ^2 мәндері оқиға ықтималдығының тұрақтылығын және ауытқудың елеулі еместігін білдіреді.

Зерттеуші мамандардың тұжырымдауынша, χ^2 критерийін таңдама жиіліктерінің шамасы 20-дан жоғары және тіпті жүздеп саналған жағдайларда қолданғанды дұрыс деп табады.

Қорыта айтқанда, χ^2 критерийін мынандай мазмұндағы есептерге пайдаланған жөн:

1) жоғарыда қарастырғандай, бір типті мәтіннен бірдей көлемдегі таңдама мәтіндер алынған жағдайда;

Мысалы, есеп берілісі мынадай: $n=500$, $k=5$, сын есім жиіліктері: $x_1=70$, $x_2=82$, $x_3=68$, $x_4=80$, $x_5=75$. $\overline{X}=?$ χ^2-?

Егер $V=k-1=5-1=4$ болса, *1-кестені* пайдаланып, ауытқудың елеулі сипатта еместігін дәлелдеу керек (жауабы – $\chi^2=1,97$).

2) екі түрлі (типті) жиыннан (қатар түрде) бірдей көлемдегі таңдама мәтіндер бойынша есептелген бір ғана «оқиғаның» (мысалы, зат есімдер) жиіліктерінің ауытқуы кездейсоқ екендігін немесе ауытқу дәрежесінің елеулілігін анықтауда.

Мәселен, есеп берілісі мынадай делік: бірдей көлемдегі ($L=1000$), стильдік ерекшеліктері бар екі түрлі мәтін таңдамасы беріледі. Бұл мәтін бөліктерінде зат есім сөздердің кездесу жиілігі: $x_1=270$ және $x_2=220$.

Жиіліктердің орта жиіліктен ауытқуының елеулі не елеулі еместігін анықтау керек.

Шешімі: Арифметикалық орта: $\overline{X}=\frac{270+220}{2}=245.$

Ал χ^2-тың тәжірибелік мәні:

$$\chi^2=\frac{(x_1-\overline{X})^2+(x_2-\overline{X})^2}{\overline{X}}=\frac{(270-245)^2+(220-245)^2}{245}=5,1.$$

Енді осы χ^2-тың тәжірибелік мәнін оның кестедегі теориялық мәнімен салыстыра бағалау қажет.

Ол үшін алдымен *еркіндік дәреже санын* анықтаймыз ($V=2-1=1$), ал сосын барып, χ^2-тың кестедегі теориялық мәнінің *3,84*-ке тең екендігіне көз жеткіземіз.

Енді χ^2 пен V мәндерінің негізінде кесте деректері бойынша $\chi^2=3,84$ екендігін және оның ықтималдық пайызы өте

аз дәрежеде – 5%, ал $5,1>3,84$ болғандықтан ауытқу елеулі сипатта және оқиға кездейсоқ емес деп қорытындылаймыз.

3) Салыстыра зерттейтін мәтіндердің көлемі тең болмаған жағдайға байланысты.

Мысалы: *Екі түрлі мәтіндер көлемдері:* $L_1=530$, $L_2=970$ *және әрбір мәтіндегі сын есім жиіліктері:* $x_1=75$ *және* $x_2=100$.

Таңдама жиіліктердің орта жиіліктен ауытқу шамасы елеулі ме әлде ол кездейсоқтық сипатта ма?

Шешімі:

а) $x_1+x_2=75+100=175$; *ә)* $L=L_1+L_2=530+970=1500$;

б) $p(x)=\frac{175}{1500}=0,116$; *в)* $\overline{x_1}=\frac{x_1+x_2}{L_1+L_2}\cdot L_1=0,116 \text{ x } 530=61,5$;

г) $\overline{x}_2=\frac{x_1+x_2}{L_1+L_2}\cdot L_2=0,116 \text{ x } 970=113,5$.

Енді $$\chi^2=\frac{(x_1-\overline{x}_1)^2}{\overline{x}_1}+\frac{(x_2-x_2)^2}{\overline{x}_2}=$$

$$=\frac{(75-61,5)^2}{61,5}+\frac{(100-113,5)^2}{113,5}=4,57.$$

Еркіндік дәреже саны: $l=2-1=1$; *Хи квадраттың* кестелік (теориялық) мәні: $\chi_k^2=3,84$ және оның тәжірибелік мәні: $\chi_t^2=4,57$; Тәжірибе жүзіндегі $\chi_t^2=4,57$ χ^2-тың кестедегі теориялық мәнінен үлкен, яғни $4,57>3,84$, сондықтан таңдама жиіліктің орта жиіліктен ауытқуы елеулі деп есептеледі және ол кездейсоқтық оқиғаға жатпайды.

2.3. Вариация коэффициенті (құбылу коэффициенті)

Вариация коэффициенті де математикалық статистикада «орта квадраттық ауытқу» тәрізді жиіліктің құбылу жағдаятының өлшемі ретінде қолданылады.

Анықтама. Вариация (құбылу) коэффициенті деп «орта квадраттық ауытқу» шамасын «орта жиілікке» бөліп, ол бөліндіні *100* санына көбейткенге тең болатын шаманы атаймыз.

Яғни бұл анықталған шама «орта квадраттық ауытқудың» «орта жиілікке» қатынасының пайыздық көрінісі деуге болады.

Егер «вариация коэффициентін» «*V*» деп белгілесек, онда:

$$V = \frac{\sigma}{\bar{X}} \cdot 100\%.$$

Бұл өрнектен ең алдымен байқайтынымыз, ол – орта квадраттық ауытқу шамасы неғұрлым көп және орта жиілік шамасы неғұрлым аз болса, вариация коэффициенті де соғұрлым үлкен шамаға ие болады.

Тәжірибеде, егер құбылу коэффициентінің шамасы 40 пайыздан көп болатын болса, таңдама жиіліктердің ауытқуы елеулі деп саналып, жиіліктің ауытқуының кездейсоқтығы жайлы жорамал теріске шығарылады. Әрине вариация коэффициенті «*V*» дәлдік жағынан «хи квадрат» критерийіне қарағанда төмендеу дәрежеде деп саналады.

Мысал ретінде жоғарыда сөз болған ғылыми-техникалық мәтініндегі етістіктердің жиіліктерін келтірейік: 98, 87, 102, 105, 123, 108, 85, 78, 110, 104; n=10; орта жиілік – $\bar{x} = 100$.

Ал квадраттық ауытқу ортасы: $\sigma = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n}} = 12{,}7.$

Енді вариация немесе құбылу коэффициенті:

$$V = \frac{\sigma}{\bar{x}} \cdot 100\% = \frac{12{,}7}{100} \cdot 100\% = 12{,}7\%.$$

Мұндай вариация коэффициенті жиіліктердің құбылу сипатының кездейсоқтығына айтарлықтай дәлел бола алады.

«Хи квадрат» критериясы мен «құбылу коэффициенті» мәндерінің ұтымды не ұтымсыз жақтарын тілдік және математикалық тұрғыда айқындай түсу үшін көптеген тәжірибелер жүргізу қажеттігін айта кеткенді жөн санаймыз.

2.4. Үлес ұғымы және оларды салыстыру

Анықтама. Үлес дегеніміз бақылауға қажетті «оқиғаның» барлық (осы тәріздес) қатарлар ауқымынан алатын орны (үлесі).

Үлес мәнін де ықтималдықты табу өрнегі арқылы анықтаймыз. Егер ықтималдықты латын әліпбиінің үлкен «*P*»

әрпімен белгілесек, үлесті соның кіші әрпі *«p»* арқылы белгілеуге болады және ол ықтималдықтың классикалық анықтамасы бойынша (өрнегімен) есептеледі.

Есеп. *1000 сөзқолданыстан тұратын мәтін ішінде 300 етістік сөздердің барына көзіміз жетсе, онда етістіктің мәтін бойындағы үлесі қанша?*

$P=\frac{m}{n}=\frac{300}{1000}=0,3$ не болмаса пайыздық салмағын білу үшін оны *100-ге* көбейту керек, яғни $0,3\times 100\%=30\%$.

Жауабы: етістік сөздердің мәтін бойынан алатын үлесі *0,3* немесе ол барлық мәтіннің 30 пайызы.

Жиіліктердің ауытқуы тәріздес үлестер де құбылып тұрады. Үлестің ауытқуын табу үшін тағы да статистика пәніндегі «квадраттық үлес ауытқуы» деп аталатын өрнекті пайдалану қажет. Ол үшін үлес ауытқуы тек бір ғана статистикалық заңдылыққа бағынуы керек.

Егер *«M»* әрпімен квадраттық үлестің ауытқуын белгілесек, онда: $M=\sqrt{\frac{pq}{n}}$, бұл жердегі p – зерттелетін «оқиғаның» үлесі, q – барлық басқа оқиғалардың жиынтығының үлесі, сондықтан $q=1-p$, ал $p=1-q$ болады, n – таңдама жиынның көлемі.

Енді жоғарыда келтірілген есептің берілісі бойынша, мәтін бойындағы барлық етістік емес сөзқолданыстардың үлесін табуға болады. Яғни оны q деп белгілесек, онда $q=1-0,3=0,7$.

Квадраттық үлес ауытқуының өрнегі екі түрлі таңдама жиындардағы бірдей «оқиғаның» үлестерін салыстыру мақсатымен қолданылады.

Мысалы, Ахмет Байтұрсынұлының және Мағжан Жұмабаевтың шығармаларындағы «баяндауыш» сөздердің үлесін анықтағымыз керек болды делік не болмаса көркем әдебиет пен публицистика стильдеріндегі сын есім үлестерін білгіміз керек болды десек, осы «квадраттық үлес ауытқуы» өрнегіне аздаған өзгеріс енгізіп, пайдалануымызға болады.

Аталған өзгеріс мынада:

$$\varepsilon_{1,2}=\sqrt{pq(\frac{1}{n_1}+\frac{1}{n_2})},$$

мұндағы $\varepsilon_{1,2}$ – квадраттық орта үлес ауытқу шамасының екі түрлі жиындар бойынша алынуы; $\bar{p}$ және $\bar{q}$ – анықтауға қажет және одан басқа да оқиғалардың жиынтығы бойынша есептелген орта үлестер; n_1 және n_2 – таңдама жиындардың көлемі.

Егер $3\varepsilon_{1,2} \leq p_1 - p_2$ болса, онда үлестердің ауытқуы елеулі дәрежеде болады, ал керісінші жағдайда – ауытқу кездейсоқ деп қорытынды жасау қажет.

Есеп: *1000 сөзқолданыстан тұратын екі түрлі мәтіндер жиындары бойынша етістіктердің кездесу жиілігі 200 және 150-ге тең.*

Етістіктің екі жиын бойынша анықталған үлестерінің статистикалық теңдігі жайлы ғылыми болжам жасауға болар ма еді?

Шешімі:

$$p_1 = 200:1000 = 0,20; \quad p_2 = 150:1000 = 0,15;$$

$$\bar{p} = \frac{p_1 + p_2}{2} = (0,20 + 0,15):2 = 0,35:2 = 0,175;$$

$$q_1 = 1 - 0,20 = 0,80; \quad q_2 = 1 - 0,15 = 0,85;$$

$$\bar{q} = \frac{q_1 + q_2}{2} = (0,80 + 0,85):2 = 1,65:2 = 0,825;$$

$$n_1 = n_2 = 1000;$$

$$\varepsilon_{1,2} = \sqrt{0.175 \cdot 0.825 \cdot 2/1000} = 0,017.$$

Егер $3 \times \varepsilon_{1,2} \leq p_1 - p_2$ болса, онда үлестердің ауытқуы елеулі дәрежеде болады, ал керісінші жағдайда – ауытқу кездейсоқ сипатта деп қорытынды жасау қажет.

Сонымен, қарастырып отырған жағдайда $3 \times \varepsilon_{1,2} = 0,051$ және $0,051 \approx 0,050$. Бұл айырым $(0,501 - 0,050 = 0,001)$ аз болғанымен етістіктің екі жиын бойынша үлестерінің статистикалық теңдігі жайлы ғылыми болжам қабылданбайды. Яғни мұндай жағдайда үлестердің ауытқуы кездейсоқтық сипаттағы заңдылыққа бағынбайды, ауытқу елеулі деген қорытынды жасалады.

2.5. Орта таңдама жиіліктерді салыстыру

Таңдама жиіліктерді және олардың үлестерін салыстыруға болатын сияқты, «орта таңдама жиіліктер» де өзара салыстырылады.

Есеп: *500 сөзқолданыстан тұратын А және В мәтіндерінен он-оннан таңдама бөліктер алынған. Бұл мәтіндер жайлы іштей біркелкі деген ұйғарым жасалды делік. Егер А және В мәтін таңдамаларында сын есім сөздердің жиіліктері төмендегідей болып кездессе:*

«А» бойынша: 72, 65, 78, 71, 70, 74, 80, 90, 68, 82.

«В» бойынша: 80, 93, 84, 83, 78, 67, 85, 86, 75, 89.

Онда осы «А» және «В» мәтіндері бойынша есептелген орта жиіліктер $\bar{X}_1$ *пен* $\bar{X}_2$ *аралырындағы ауытқу шамасы кездейсоқтық сипатта ма, әлде елеулі дәрежеде ме?*

Бұл сұраққа жауап беру үшін математикалық статистикадағы «Стьюдент критерийін» пайдалану қажет [17, 29, 30. 40-43-бб.]. Оның өрнегі мынандай:

$$t = \frac{\bar{x}_1 - \bar{x}_2}{S_{1,2}} \cdot \sqrt{\frac{k_1 . k_2}{k_1 + k_2}} \quad . \qquad (1)$$

Мұндағы $\bar{x}_1$, $\bar{x}_2$ – орта жиіліктер, k_1, k_2 – таңдама бөліктердің саны *(біздің мысалда* $k_1 = k_2 = 10$*)*; $S_{1,2}$ – квадраттық орта ауытқуларды бағалау параметрі. Аталған параметрді төмендегі (2) өрнекпен анықтайды:

$$S_{1,2} = \sqrt{\frac{\sum (x_{i1} - \bar{x}_1)^2 + \sum (x_{i2} - \bar{x}_2)^2}{k_1 + k_2 - 2}} \quad . \qquad (2)$$

Енді $\bar{x}_1$, $\bar{x}_2$, k_1, k_2, $S_{1,2}$ шамалары бойынша (1) өрнектегі t-ны тауып, сол мән арқылы ауытқудың кездейсоқтығы мен елеулілігінің ықтималдығын арнайы кесте (*2.2-кесте*) көмегімен анықтауға болады (осыған байланысты Б.Н.Головин «Язык и статистика» 40–50-бб. немесе басқа да оқулықтарды қараңыз).

Тақырыптың басында келтірілген есеп берілісіне қайта оралсақ, оның шешімі төмендегідей көрініс табады:

$$\bar{x}_1 = 75, \ \bar{x}_2 = 82, \ \Sigma (x_{i1} - \bar{x}_1)^2 = 508,$$

$\sum(x_{i2}-\bar{x}_2)^2 = 494$, $k = k_1 + k_2 - 2 = 10+10-2 = 18$; $K = 18$ *еркіндік дәрежесі саны;* K_1, K_2 – *тәжірибе саны.* Ал квадраттық орта ауытқуларды бағалау параметрі:

$$S_{1,2} = \sqrt{\frac{508 + 494}{18}} = 7,5.$$

Енді (1) Стьюдент критерийінің өрнегі бойынша:

$$t = \frac{\bar{x}_1 - \bar{x}_2}{S_{1,2}} \cdot \sqrt{\frac{k_1 \cdot k_2}{k_1 + k_2}} = \frac{82 - 75}{7,5} \cdot \sqrt{\frac{10 \cdot 10}{10 + 10}} = 0,93 \cdot 2,22 = 2,1.$$

2.2-кестедегі 18-жол бойынша $t=2,101$, ал бұл *5%* ықтималдыққа сай келеді. Бұл ықтималдық шамасы орта жиіліктерді ($\bar{x}_1$, $\bar{x}_2$), статистикалық тұрғыда тең екендігін терістеуге аз да емес, ал оны қолдауға айтарлықтай көп те емес. Яғни ауытқудың елеулі не кездейсоқ екендігін айту қиын, сондықтан қосымша байқау тәжірибесін жүргізген жөн болады.

Осы статистикалық есепті, яғни екі орта жиілікті салыстыруды басқа жолмен де шығаруға болады. Ол үшін квадраттық ауытқу мен олардың айырымын білу керек.

Аталған шаманы есептеп шығару үшін мына өрнек ыңғайлы деп саналады:

$$\varepsilon_{1,2} = \sqrt{\frac{\sigma_1^2}{\kappa_1} + \frac{\sigma_2^2}{\kappa_2}},$$

мұндағы σ_1^2 және σ_2^2 - екі таңдаманың дисперсиялары, K_1, K_2 – тәжірибе саны немесе таңдама бөліктер саны.

Жоғарыда дисперсия мына өрнек арқылы берілгенін еске түсірейік:

$$\sigma^2 = \frac{\sum(x_i - \bar{x})^2}{\kappa};$$

Есептеп шығарған квадраттық ауытқу мәні, яғни $\varepsilon_{1,2}$ екі орта жиіліктердің ($\bar{x}_1$, $\bar{x}_2$) айырымымен салыстырылады. Егер бұл айырым ($\bar{x}_1$ - $\bar{x}_2$) $\geq 3 \cdot \varepsilon_{1,2}$ болса, онда ауытқудың елеусіздігі жайлы болжам терістеледі.

Жоғарыда анықталғандар: $\bar{x}_1 = 75$, $\bar{x}_2 = 82$ және

$$\sigma_1^2 = \frac{\sum (x_{i1} - \bar{x}_1)^2}{\kappa_1} = \frac{508}{10} = 50.8;$$

$$\sigma_2^2 = \frac{\sum (x_{i2} - \bar{x}_2)}{\kappa_2} = \frac{494}{10} = 49{,}4.$$

Енді жаңадан ұсынылып отырған өрнек бойынша:

$$\varepsilon_{1.2} = \sqrt{\frac{50{,}8}{10} + \frac{49{,}4}{10}} = 3{,}17. \quad 3 \cdot \varepsilon_{1.2} = 3 \cdot 3{,}17 = 9{,}51.$$

2.2-кесте

Стьюдент критерийіндегі *«t»* мәнін кесте арқылы анықтау

Еркіндік дәреже саны ($k=k_1+k_2-2$)	Үлкен шаманың ықтималдығы				
	0.50 (50%)	0,20 (20%)	0.10 (10%)	0,05 (5%)	0.025 (2,5%)
1	1,000	3,078	6,314	12,706	25,452
2	0,816	1,886	2,920	4,303	6,205
3	0,765	1,638	2,353	3,182	4,176
4	0,741	1,533	2,132	2,776	3,495
5	0,727	1,476	2,015	2,571	3,163
6	0,718	1,440	1,943	2,447	2,969
7	0,711	1,415	1,895	2,365	2,841
8	0,706	1,397	1,860	2,306	2,752
9	0,703	1,383	1,833	2,262	2,685
10	0,700	1,372	1,812	2,228	2,634
11	0,697	1,363	1,796	2,201	2,593
12	0. 695	1,355	1,782	2,179	2,560
13	0. 694	1,350	1,771	2,160	2,533
14	0. 692	1,345	1,761	2. 145	2,510
15	0. 691	1,341	1,753	2,131	2,460
16	0. 690	1,337	1,746	2,120	2,473
17	0. 689	1,333	1,740	2,110	2,458
18	0. 688	1,330	1,734	2,101	2,445
19	0. 688	1,328	1,729	2,093	2,433
20	0. 688	1,325	1,725	2,086	2,423
21	0. 687	1,323	1,721	2,080	2,414
22	0. 686	1,321	1,717	2,074	2,406
23	0. 685	1,319	1,714	2,069	2,398
24	0. 685	1,318	1,711	2,064	2,391
25	0. 684	1,316	1,708	2,060	2,385

Ал, $X_2 - X_1 = 82 - 75 = 7$, яғни $7 < 9{,}51$. Яғни ол айырым $(X_2 - X_1)$ үш еселенген квадраттық ауытқудан $(3 \cdot \varepsilon_{1,2})$ кіші. Бұл жағдай екі орта жиіліктер $(\overline{x}_1, \overline{x}_2)$ арасындағы ауытқудың елеулі еместігінің куәсі.

Сонымен, бірінші рет критерий Стьюдент бойынша құман тудырған жайт, екінші әдіс бойынша анықталып, екі қатардағы жиіліктердің құбылуы бір ықтималдыққа бағынатын статистикалық заңдылық деп қорытындылауға болады.

2.6. Бақылау кезіндегі абсолюттік және қатынастық қателер мен таңдама мәтінінің көлемін анықтау

Статистикалық әдісті тілдік таңдама бөліктерге қолдану кезінде ықтималдық заңдылықтарға тән құбылмалықтың әсерінен зерттеушінің алған нәтижелері кейбір қателерге әкеледі. Сондықтан, мысалы, тілші анықтаған *орта таңдама жиіліктің* мәнін шындыққа сай келетін *«нағыз орта жиілік»* мәнімен және, сол сияқты, зерттеуші есептеп шығарған *«үлес»* мәнін *«нағыз үлес»* мәнімен салыстыруға бола ма? – деген сұрақ туындауы мүмкін.

Математикалық статистика пәні зерттеуші-тілшіге ондай жағдайды туғыза алады, яғни аталған шамалардың *«нағыз орта жиілік»* пен *«нағыз үлес»* мәндерінен құбылу шамасын дәл бір санмен көрсете алмағанымен, оның өзгеру шекарасын (ауқымын) анықтай алады. Ондай мүмкіндік мынандай *«бақылау қателігі»* анықтайтын өрнекке негізделеді [17, 29, 30, 40-50-бб.]:

$$L = \frac{t \cdot \sigma}{\sqrt{k}} . \qquad (1)$$

Бұл өрнекті басқа атпен *абсолютті қате* деп те атайды. Мұндағы t – арнайы кесте арқылы анықталатын *теориялық коэффициент*. Оның мәні *«еркіндік дәреже санына»*, яғни, басқаша айтқанда, таңдама бөліктердің санына байланысты анықталатын өзгермелі шама. Ал, σ – *квадраттық ауытқу ортасы*, k – *таңдама бөліктер саны (бақылау саны)*. Тәжірибе кезінде *квадраттық ауытқу ортасы* өрнегінің (σ-ның) орнына мынадай (2) өрнекті пайдаланған жөн деп саналады:

$$S = \sqrt{\frac{\sum (x_i - \bar{x})^2}{k-1}} \quad (2)$$

Енді *«бақылау қатесін»* анықтау үшін қажетті шамалардың барлығы да бар. Олар: σ, k және t. Арнайы кесте арқылы анықталатын *теориялық коэффициент* t-ның мәні ықтималдықтың дәлдігіне байланысты алынады. Әдетте, үлестің орта жиілігін және байқау қатесін есептеуде 95% сенімділік жеткілікті деп есептеледі, яғни ол *0,95* ықтималдыққа сәйкес келеді деген сөз. Енді *«t»* шамасын *2.3-кестедегі* таңдама саны мен *95%-ға* сәйкес келетін жатық және тік жолдардың қиылысындағы санның шамасын теориялық коэффициент t-ның мәні деп есептейміз.

Мысалы, таңдама бөліктердің саны $k=5$ болса, $t=2,78$, ал $k=10$ болса, $t=2,26$ тең деген сөз.

Кестеде көрсетілген коэффициент шамасы көп болған сайын, бақылау кезіндегі жіберген қате де дәлірек анықталады, яғни ықтималдық дәрежесі үлкен шама болады. Аса үлкен дәлдіктің қажеттігі болмаған жағдайда, тәжірибеде ондай коэффициентті кестесіз-ақ, $t=2$ деп алуға да болады деп есептейді. Ал, *95%* сенімділік дегеніміз жоғарыда көрсетілген өрнек бойынша есептелген бақылау кезіндегі қатенің шамасы *100 рет* жүргізілген статистикалық тәжірибеміздің бесеуінде ғана қайталануы мүмкін деген сөз. Осындай дәлдік мөлшері тілдерді зерттеуге байланысты жайттарда жеткілікті деп саналады.

Жоғарыда қарастырған есептің берілісін тағы да қайталайық.

Есеп: *А және В мәтіндерінен он-оннан таңдама бөліктер алынған. Бұл мәтіндер жайлы іштей біркелкі деген үйғарым жасалды делік. Ал осы мәтін таңдамаларында* ***сын есім создердің*** *жиіліктері төмендегідей:*

«А» бойынша: 72, 65, 78, 71, 70, 74, 80, 90, 68, 82;

«В» бойынша: 80, 93, 84, 83, 78, 67, 85, 86, 75, 89.

Осы екі жиіліктер қатары бойынша анықталған орта таңдама жиіліктердің: $\bar{x}_1 = 75$, $\bar{x}_2 = 82$ *мәндерінің тәжірибе кезіндегі* «бақылау қатесін» *анықтаңыздар.*

Есептің шешімпін табу үшін абсолютті қате өрнегіндегі ($L = \frac{t \cdot \sigma}{\sqrt{k}}$ - *t*), σ және k мәндері белгілі болуы керек.

Есептің берілісі бойынша бізге белгілі: $N = 500$, $k = k_1 = k_2 = 10$. *Квадраттық ауытқу ортасы* $\sigma = \sqrt{\frac{\sum (x_i - \bar{x})^2}{k}}$ өрнегі арқылы анықталады және олардың шамалары:

$$\sigma_1 = \sqrt{\frac{508}{10}} = 7,1; \quad \sigma_2 = \sqrt{\frac{494}{10}} = 7.$$

2.3-кесте бойынша $k = 10$ болса $t = 2,26$ екені жоғарыда дәлелденген болатын. Олай болса:

$$L_1 = \frac{t \cdot \sigma_1}{\sqrt{k}} = \frac{2,26 \cdot 7,1}{\sqrt{10}} = \frac{16,046}{3,16} = 5,1;$$

$$L_2 = \frac{t \cdot \sigma_2}{\sqrt{k}} = \frac{2,26 \cdot 7}{\sqrt{10}} = \frac{15,82}{3,16} = 5,0.$$

2.3-кесте

«Бақылау қатесі» өрнегіндегі «*t*» мәнін кесте арқылы анықтау

Таңдама бөлік саны	Бақылау қатесін анықтаудың сенімділігі (ықтималдық)					
	99% (0,99)	97,5 (0,975%)	95% (0,95)	90% (0,90)	80% (0,80)	60% (0,60)
3	9,93	6,21	4,30	2,92	1,89	1,06
5	4,60	3,50	2,78	2,13	1,53	0,94
6	4,03	3,16	2,57	2,02	1,48	0,92
7	3,71	2,97	2,45	1,94	1,44	0,91
8	3,50	2,84	2,37	1,90	1,42	0,90
9	3,36	2,75	2,31	1,86	1,40	0,89
10	3,25	2,69	2,26	1,83	1,38	0,88
15	2,98	2,51	2,15	1,71	1,35	0,87
20	2,86	2,43	2,09	1,73	1,32	0,86
25	2,80	2,39	2,06	1,71	1,32	0,86
30	2,76	2,36	2,05	1,70	1,31	0,85

$L_1 = 5,1$ және $L_2 = 5$ – *бақылау қателері*. Сондықтан, бақылау қателерінің шамасын ескере отыра, орта таңдама жиіліктің мәнінің өзгеру шекарасын төмендегідей жазуға болады:

«А» мәтіні үшін: $75-5,1 \le \bar{x}_1 \le 75+5,1$ немесе $\bar{x}_1 = 75$ емес, оның мәні $69,9 \le \bar{x}_1 \le 80,1$ аралығында;

«В» мәтіні үшін: $82-5 \le \bar{x}_2 \le 82+5$ немесе $\bar{x}_2 = 82$ емес, оның мәні $77 \le \bar{x}_2 \le 87$ аралығында.

Ықтималдықтың 95% сенімділікпен алынуы, орта жиіліктің есептелген аралықтан тыс жату мүмкіндігі 100 тәжірибенің бесеуінде ғана кездесуі мүмкін деген сөз.

Есеп: *А және Б мәтіндерінің көлемі N=500 сөзқолданысқа тең келетін және олар 5 таңдама бөлікке бөлінген екі түрлі мәтіндер бойынша сын есім сөздердің кездесу жиілігі төмендегідей:*

«А» – 55, 70, 76, 49, 45

«Б» – 52, 78, 88, 22, 25.

Тәжірибе кезіндегі бақылау қатесін табыңыз және «нағыз орта жиілік» қандай аралықта жататынын анықтаңыз.

Шешімі:

$$1)\ \bar{x}_1 = \frac{55+70+76+49+45}{5} = \frac{295}{5} = 59;$$

$$2)\ \bar{x}_2 = \frac{52+78+88+22+25}{5} = \frac{265}{5} = 53;$$

$$3)\ \sum_{i=1}^{5}(x_{i1} - \bar{x}_1)^2 = (55-59)^2 + (70-59)^2 + (76-59)^2 + (49-59)^2 + (45-59)^2 =$$

$$= 4^2 + 11^2 + 17^2 + (10)^2 + (-14)^2 = 16 + 121 + 289 + 100 + 196 = 722;$$

$$4)\ \sum_{i=1}^{5}(x_{i2} - \bar{x}_2)^2 = (52-53)^2 + (78-53)^2 + (88-53)^2 + (22-53)^2 + (25-53)^2 =$$

$$= 1 + 625 + 1225 + 961 + 784 = 3596.$$

5) Енді *ауытқудың квадраттық ортасын* анықтайық:

$$\delta_1 = \sqrt{\frac{\sum (x_{i1} - \bar{x}_1)^2}{K_1}} = \sqrt{\frac{722}{5}} = 12,00,$$

$$\delta_2 = \sqrt{\frac{\sum (x_{i2} - \bar{x}_2)^2}{K_2}} = \sqrt{\frac{3596}{5}} = 27.$$

6) *95% сенімділік* үшін, $k = 5$ болса, *2.3-кесте* бойынша $t = 2,78$;

7) *бақылау қателері*:

$$L_1 = \frac{t \cdot \delta_1}{\sqrt{k}} = \frac{2,78 \cdot 12}{\sqrt{5}} = \frac{33,36}{2.24} = 14,9;$$

$$L_2 = \frac{t \cdot \delta_2}{\sqrt{k}} = \frac{2,78 \cdot 12}{\sqrt{5}} = 33,7;$$

8) *«А»* мәтіні үшін $\bar{x}_1$ ауытқуларының аралығы:

$59 - 14,9 \le \bar{x}_1 \le 59 + 14,9; \quad 44,1 \le \bar{x}_1 \le 73,9;$

9) «Б» мәтіні үшін: $19,3 \le \bar{x}_2 \le 86,7;$

«Б» мәтініне байланысты ауытқу аралығы өте үлкен, сондықтан мұндай «жағымсыз» жағдай болмас үшін не мәтін көлемін, не ішкі таңдама бөліктер көлемін ұлғайту, не *«t»* мәнін кішірейту қажет болады.

Жоғарыда келтірген (1) өрнек, яғни $L = \frac{t \cdot \sigma}{\sqrt{k}}$ өрнегі бақылаудың *абсолютті қатесін* анықтау үшін қолданылады, себебі оның өлшемі таңдама орта жиіліктің өлшемімен бірдей болып келеді.

Бақылаудың *абсолютті қатесін* анықтау барлық жағдайларда зерттеушіні қанағаттандыра бермейді. Сол себепті статистиктер өз тәжірибесінде *абсолютті қателен* басқаша – *«қатынастық қате»* деп аталатын шаманы да қолданады. Егер *абсолютті қате* таңдама орта жиіліктің *«нағыз орта жиіліктен»* не көп, не аз болып ауытқуын білдіретін санға тең болатын болса, *қатынастық қате* – осы *абсолютті қатенің таңдама орта жиілікке* қатынасына тең болатын ондық бөлшек (немесе пайыздық шама).

Мысалы, абсолютті қате $L = 25$, ал орта жиілік $\bar{x} = 50$ десек, онда *қатынастық қатені* «δ» *(дельта)* деп белгілесек:

$$\delta = \frac{L}{\bar{x}} = \frac{t \cdot \sigma}{\bar{x} \cdot \sqrt{k}} = \frac{25}{50} = \frac{1}{2} = 0,5.$$

Яғни *«абсолютті қате» «орта жиіліктің»* $\frac{1}{2}$ бөлігін немесе *50* пайызын құрайды екен.

Ал $L=25$ және $\overline{x}=500$, $\delta=\frac{25}{100}=\frac{1}{20}=0,05$, яғни абсолютті қате «*L*» орта жиіліктің $\frac{1}{20}$ бөлігін немесе *5* пайызын құрайды екен.

Қатынастық қатені есептеу арқылы *абсолютті қате* мен *орта жиілікті* бір-бірімен салыстыруға мүмкіндік туады, яғни қате шамасының қай жағдайда – көп, ал қай жағдайда – аз екендігін ажырата аламыз.

Статистикалық жолмен тіл заңдылығын зерттеу тәжірибесінде қатынастық қатенің *5-10%* шамаларын қанағаттандырарлық деп санайды. Кейбір жағдайларда *30%-ға* дейінгі шамамен де келісуге болады, ал одан көбі контік етеді. Сонымен, *қатынастық қатені* есептеу өрнегі:

$$\delta=\frac{t\cdot\sigma}{\overline{x}\cdot\sqrt{k}}\ . \qquad (3)$$

Бұл өрнектегі $\frac{t\cdot\sigma}{\sqrt{k}}=L$ – *абсолютті қате*, $\overline{x}$ – *орта жиілік*.

Егер зерттеуші «оқиғаның» *орта жиілігін* емес, оның барлық зерттеу мәтінінен алатын *үлесін* білгісі келсе, ол төменде көрсетілетін өрнектер арқылы іске асады. Яғни тәжірибе жүзінде есептелген «*үлестің*» *«нағыз үлестен»* ауытқуын білу үшін *абсолютті* және *қатынастық қателерді* анықтау қажет болады.

Ал ондай қателер (4) және (5) өрнектер арқылы есептеледі:

$$L_p=\frac{2\sqrt{pq}}{\sqrt{N}}\ . \qquad (4)$$

Бұл (4) өрнек бойынша үлестің *абсолютті қатесін* табуға болады. Мұндағы «*2*» – тұрақты коэффициент, p және q – таңдама үлестері, N – соз санымен не басқа тілдік бірлік арқылы өлшенетін мәтін көлемі.

Келесі (5) математикалық өрнек арқылы *қатыпастық қателі* анықтауға болады:

$$\delta_p = \frac{2\sqrt{q}}{\sqrt{pN}}. \qquad (5)$$

Қатыпастық қателі анықтайтын (5) өрнек статистикалық тәжірибе үшін таңдама бөліктердің санын және олардың көлемін алдын ала анықтауға мүмкіндік туғызады.

Мәтін көлемін анықтау үшін мынадай амалдарды іске асырайық:

1) (5) өрнектегі теңдіктің екі жағын квадратқа алсақ:

$$\delta_p^2 = \frac{4q}{pN}, \qquad (6)$$

2) енді осы (6) өрнек бойынша мәтін көлемін анықтайтын (7) формуланы жазуға болады:

$$N = \frac{4q}{p\delta_p^2}. \qquad (7)$$

Бұл өрнектердегі p зерттеуге тиісті оқиғаның барлық мәтін бойынан алатын үлесі, ал q – басқа оқиғалардың үлесі, δ_p – осы үлеске қатысты *абсолютті қате* шамасы. Егер осы параметрлердің шамаларын алдын ала өз қалауымызша белгілесек, оларға сай мәтін көлемін (7) өрнек бойынша есептеп шығаруға болады.

Осындай мәтін көлемін *абсолютті қате* өрнегі бойынша да табуға болады: $L = \frac{t \cdot \sigma}{\sqrt{k}}$; $L^2 = \frac{t^2\sigma^2}{k}$; $k = \frac{t^2\sigma^2}{L^2}$.

Егер таңдама бөлік саны $k = N$ *болса,* яғни ол барлық мәтін көлеміне тең десек, онда бұл өрнек мәтін көлемін есептеп шығаруға қолданылады.

Төмендегі өрнектерді де мәтін көлемін анықтау үшін қолдануға болады:

а) *абсолютті қате* (L) бойынша мәтін көлемін анықтау:

$$N = \frac{4p \cdot q}{L^2};$$

б) *қатыпастық қате* (δ) бойынша мәтін көлемін анықтау:

$$N=\frac{4\cdot q}{\delta^2\cdot p}.$$

Есеп: *Алдыңғы есептің берілісі бойынша мәтін ішіндегі үстеу сөздердің үлесі 0,07. Егер абсолютті қате 0,005-тен аспау үшін қандай көлемді мәтін алғанымыз жөн.*

Шешімі:

$$p=0,07;\ q=1-0,07=0,93;\ L=0,005;$$

$$N=\frac{4\cdot 0,93\cdot 0,07}{(0,005)^2}=10,416;$$

Немесе мұны $N=10500$ сөзқолданыс деп шамамен (дөңгелетіп) жазуға болады.

Есеп: *1-ші есеп берілісі бойынша үстеу үлесі 0,07 болудың және қатынастық қателің $\delta=0,05$ аспау үшін мәтін көлемі «N» қандай болу керек*

Шешімі:

$$p=0,07;\ q=0,93;\ \delta=0,05;$$

$$N=\frac{4\cdot q}{\delta^2\cdot p}=\frac{4\cdot 0,93\cdot 0,07}{(0,005)^2}=21,257$$ сөзқолданыс,

яғни шамамен алғанда $N=21000$ сөзқолданыс болуы керек деп саналады.

Үшінші тарау

ҚАЗАҚ МӘТІНІН ЫҚТИМАЛДЫ-СТАТИСТИКАЛЫҚ МОДЕЛЬДЕУ

3.1. Ципф заңы және оны қазақ мәтіні бойынша түзілген жиілік сөздіктерге қолдану

Ципф заңының қысқаша тарихы. 1916 жылы француз ғалымы Эсту стенографиялық жазу ісін жетілдіру мақсатындағы тәжірибе үстінде мынадай заңдылықты байқаған болатын [37].

Мәселен, көлемі *N*-ге тең сөзқолданыс құрайтын мәтін бойынша түзілген жиілік сөздіктің реестрлік сөздер (сөзтүлғалар) саны *n*-ге тең болып, ол *3.1-кестедегідей* көрініс тапты делік.

3.1-кесте

Рет саны *(i)*	Реестрлік сөз (сөзтүлға) *(S_i)*	Абсолютті жиілік *(F_i)*	Қатынастық жиілік $P_i = \frac{F_i}{N}$
1	S_1	F_1	P_1
2	S_2	F_2	P_2
----	---	---	----
i	S_1	F_i	P_i

Егер осы сөздіктегі кез келген реестрлік сөздің рет санын (нөмірін) «*i*», ал оған сәйкес келетін сөздің абсолютті жиілігін F_i деп белгілесек:

$1 \le i \le n$; $F_i = F_1, F_2, F_3, \ldots, F_n$ және $F_i \ge F_{i-1}$, $P_i = \frac{F_i}{N}$ – қатынастық жиілік болатыны анық.

Сонда француз ғалымы Эсту мынадай жағдайдың күрсі болады: $F_1 \cdot 1 \approx F_2 \cdot 2 \approx F_3 \cdot 3 \approx \ldots \approx F_n \cdot n$, яғни $F_i \cdot i \approx F_{(i+1)} \cdot (i+1)$.

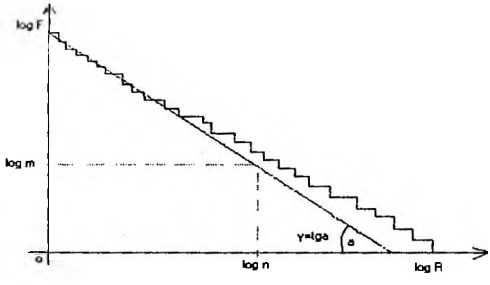
Бұл өрнектің мағынасы мынада: жиілік сөздіктегі кез келген сөздің (сөзтұлғаның) абсолютті жиілігі мен оның рет санының көбейтіндісі, шамамен алғанда, бір-біріне тең келеді. Егер ондай тең шамаларды тұрақты *«C»* санымен белгілесек:

$$F_1 \cdot 1 \approx F_2 \cdot 2 \approx F_3 \cdot 3 \approx \cdot \cdot \cdot \approx F_n \cdot n = C, \qquad (1)$$

яғни $F_i \cdot i \approx F_{(i+1)} \cdot (i+1) = C$, $F_i = \frac{C}{i}$, $(i = 1, 2, 3, 4, 5, 6, \ldots, n)$.

Сонымен, (1) өрнек мына заңдылықты аңғартады: жиілігі кему тәртібімен орналасқан жиілік сөздікте сөздің рет саны үлғайған сайын, оның кездесу жиілігі керісінше кеміп отырады, яғни сөздің жиілігі мен оның рет саны бір-бірімен өзара байланыста және кері пропорционалдық заңдылықта болады.

Ал 1928 жылы телефон компаниясының қызметкері Э.Кондон француз ғалымы Эстуге қатыссыз, мына заңдылықты ашады: егер билогарифмдік координат жүйесіндегі абсцисс өсіне жиілік сөздіктегі сөздердің рет санының *(i)* ондық логарифмдік шамасы *(ℓ gi)*, ал ординат өсіне сөздердің абсолют жиілігінің ондық логарифмдік шамасы *(ℓ gF$_i$)* салынатын болса, онда өстер жазықтығындағы нүктелер жүйесі – сынық сызықтардан тұратын сызба бейнесі төмендегі *1-суретте* көрсетілгендей түзу сызықтық пішінге ұқсас болып шыққан.



1-сурет

$P_i=\frac{c}{i}$ өрнегі билогарифмдік координатта мына түрде көрініс табады: $F_i=\frac{c}{i^{\gamma}}$. (2)

Бұл (2) өрнектегі C және γ – тұрақты сандар (константтар), ал $\gamma=tg\alpha$ – түзу сызық пен абсцисс өсінің арасындағы α бұрышының тангенсіне тең, C – түзу сызықтың ордината өсімен қиылысатын нүктесінің ордината шамасына тең.

Енді (2) теңдіктің екі жағын да бірдей N санына бөлетін болсақ: $\frac{F_i}{N}=\frac{c}{N\cdot i^{\gamma}}$, мұндағы $\frac{F_i}{N}=P_i$ – қатынастық жиілік, ал $\frac{C}{N}=K$ – тұрақты сан, яғни $P_i=\frac{k}{i^{\gamma}}=k\cdot i^{-\gamma}$. (3)

Осы (3) өрнек – *Ципф заңы* деп аталады. Бұл заңдылық бойынша, мәтіннің кез келген жерінен таңдалып алынған сөздің сөздіктегі орны, яғни рет саны зерттеушінің қалаған нөміріне тең болу ықтималдығы анықталады.

Ципф заңының мағыналық жағы мынада. Белгілі мәтін мен оның жиілік сөздігі бойынша K мен γ мәндерін бір-ақ рет анықтап алып, одан кейін кез келген сөздің сөздіктегі рет саны арқылы ол сөздің қолдану жиілігін анықтау мүмкіндігі және керісінше, сөздің абсолютті жиілігі бойынша ол сөздің сөздіктегі рет санын анықтау. Абсолютті жиілік мәні жоғарыда көрсетілген (3) өрнек бойынша былайша анықталады:

$$F_i=N\cdot k\cdot i^{-\gamma}. \quad (4)$$

Соңғы (4) көріністегі өрнек те Ципф заңы деп аталып жүр.

Ақпарат теориясы мен математикалық әдістердің XX ғасырдың 50-ші жылдары тілтаным ғылымына қарқындап араласуы көптілік ғалымдардың Ципф заңына қызығушылығын арттырды. Әсіресе, Ципф заңының бірнеше тілдердегі көрінісі мен тұрақты коэффициенттер шамасының әр тілге қатысты құбылуының мағыналық түсінігіне және т.б. осыған байланысты жағдайларға ғалымдар назар аудара бастады.

Ципф заңындағы (4) γ константасының мәтіннен мәтінге ауысқан сайын тұрақтылық сипатынан айырылатынын бел ілі ғалым Б.О.Мандельброт өзінше дәлелдеп, ол заңды мына түрде жазуды ұсынады [65]: $F_i = N \cdot K\,(i + \rho)^{-\gamma}$. (5)

Мұндағы ρ – тұрақты шама.

Енді осы соңғы жазылған (5) өрнек – Эсту-Ципф-Мандельброт заңдылығы деп атала бастады.

Р.М.Фрумкина орыс тілі мәтіндері негізінде γ шамасының тәжірибе үшін алынған бір мәтін ішінде де тұрақты бола бермей, өзгеріске ұшырайтынын дәлелдеді [90]. Ғалымның осы заңдылыққа қатысты тұжырымдаулары төмендегідей:

1. Эсту-Ципф-Мандельброт (5) өрнегіндегі γ – константасы тек белгі шекаралықта (интервалда) ғана тұрақты болады, ал одан сырт жағдайда γ рет саны i-ге тәуелді және өспелі функция. Мысалы, рет санының *50* мен *1500* аралығында, яғни $50 \le i \le 1500$ аралығында γ мәні тұрақты, ал рет санының одан сырт мәндерінде γ өзгеріске ұшырайды және оның мәні «*i*» шамасына қатысты өсіп отырады.

2. Әр мәтінге байланысты γ-ның тұрақтылық сақтайтын рет саны бойынша алынған шекарасы өзгеріске ұшырайды және ол мәтіннің статистикалық құрылымына тікелей қатысты.

3. γ шамасы мәтін көлемінің өзгеруіне қарай да құбылып отырады, яғни $\gamma = f(N)$. Сондықтан екі (не одан да көп) мәтіндерді салыстырып зерттеу кезінде ондай мәтіндердің көлемдері тең (не шамалас) болуы шарт.

Тәжірибе жүзінде Эсту-Ципф-Мандельброт заңдылығындағы γ, K және ρ мәндерін анықтау үшін мына теңдеулер жүйесін пайдалануға болады:

$$\begin{cases} F_1 = N \cdot K\,(i_1 + \rho)^{-\gamma}, \\ F_2 = N \cdot K\,(i_2 + \rho)^{-\gamma}, \\ F_{соңы} = N \cdot K\,(i_{соңы} + \rho)^{-\gamma}. \end{cases} \quad (6)$$

Мұндағы F_s сөздік үзіндісінің екі шеткі абсолютті жиіліктері – F_1 мен $F_{соңы}$ шамаларының геометриялық ортасы, ал i_s – рет сандары i_1 мен $i_{соңы}$ шамаларының геометриялық ортасы. K мен γ мәндерін алдын ала белгілі ρ бойынша «ең кіші квадраттар әдісі» деп аталатын әдіспен анықтауға болады.

Теңдеулер жүйесіндегі параметрлердің мәнін оңай табу үшін қажетті деген компьютерлік бағдарлама жасалуы қажет. Қазақстандық статолингвистер кезінде өз есеп-санақтары үшін Л.В.Зубов пен Э.Н.Хотяшовтың ЭЕМ-ға арнап құрастырған алгоритмдерін пайдаланған болатын [47].

М.Әуезовтің «Абай жолы» романы көлемдері шамалас 4 томнан (кітаптан) тұратыны және роман мәтінінің көлемі 465966 сөзқолданысты құрайтыны жайлы екінші тарауда айтылған болатын. Әрбір кітаптың мәтіндерін бөлек-бөлек қарастырып, олардан түзілген жиілік сөздіктер бойынша есептеліп шығарылған Эсту-Ципф-Мандельброт заңдылығы параметрлерінің салыстырма деректері *3.2-кестеде* берілді.

3.2-кесте

Эсту-Ципф-Мандельброт заңдылығы параметрлерінің салыстырма мәндері

Ципф заңының параметрлері	1-кітап	2-кітап	3-кітап	4-кітап	Роман «А.Ж.»
Мәтін көлемі (N_i)	105788	124398	112727	123053	465966
$i_{соң}$ $(F_i \geq 4)$	3270	3872	3570	4027	9614
ρ	4,770	5,490	4,760	6,380	7,835
γ	0,942	0,937	0,924	0,930	1,017
K	0,092	0,088	0,079	0,086	0,146

Кесте деректері жоғарыда Р.М.Фрумкинаның γ мәнінің мәтін көлеміне байланысты өзгеріске үшырайтынын дәлелдей түседі. Коптеген зерттеушілердің үйғаруынша, γ параметрінің шамасы, сол мәтіннің не «шағын тіл» (подъязык) мәтінінің лексикалық әр түрлілік көрсеткіші ретінде түсіндіріледі [90].

Мысалы, γ шамасы үлкен мәнге ие болған сайын, сол тілдің лексикалық қорының кедейлігі не керісінше, γ мәнінің «кіші» болуы лексикалық қордың байлығы деп саналады. «Абай жолы» романының әр томына қатысты есептелген ($\gamma < 1$) γ бірден кіші, ал балалар әдебиеті тілі бойынша жасалған жиілік сөздік деректерінде ол мән керісінше, яғни онда γ бір санынан үлкен шама қабылдайды екен (γ=1,6).

Жалпы алғанда, әр тіл немесе әр стиль бойынша есептелген γ мәндері не тіл типінің, не стильдер типінің айырымдық белгілері болады деп тұжырымдауға негіз бола бермейді. Яғни бұл параметр әмбебаптық қасиетке ие емес, ол тек лексикалық бірліктің жиілік сөздіктегі рет санына қарай өзгеретін шама деуге болады. Осыған көз жеткізу үшін «Абай жолы» романы жиілік сөздіктерінің он бойында γ мәні қалайша өзгеріске ұшырайтынына қысқаша тоқталайық.

Төменде *3.3-кестеде* көрсетілгендей, романның әрбір кітабының жиілік сөздіктерінің рет саны 10 түрлі пішінде екі-екі бөліктерге бөлінген. Яғни бірінші ондықта бастапқы бөлу нүктесінде $i=1$, ал екінші ондықта соңғы нүктеде $i=i_{соңы}$.

Әрбір бөлу нәтижелері бойынша Ципф заңының γ параметрі есептеліп, олардың шамалары айтылған кестеде көрініс тапты. Тіке сызықтық регрессия коэффициенттері арқылы әрбір кестеде көрсетілген зоналар бойынша және арнайы жазылған компьютерлік бағдарлама көмегімен γ мен $lg(N \cdot K)$ мәндері есептелді (келесі тақырыпшада берілген *3.3* пен *3.4-кестелерді* қараңыз).

Кестеде көрініс тапқан γ мәндері басқа тілдер бойынша жасалған тұжырымдарды дәлелдей түседі. Яғни γ мәні сөздіктің әр бөлігінде әр түрлі және олардың сандық мәні *1* санына шамалас келеді. γ мәтіннің кейбір *1* санынан барынша алшақтау мәндері сөздіктің бастапқы және соңғы бөліктеріне сәйкес келеді, ал ортаңғы бөлігінде γ параметрінің тұрақтылық қалпы байқалады.

«Абай жолы» романының толық мәтіні бойынша γ мәтінінің ортаңғы бөліктерінің арифметикалық орта шамасы *0,913* тең. Сол сияқты Ципф заңы өрнегіндегі K мәнін де тұрақты деуге бола бермейді, себебі сөздіктің ортаңғы бөлігімен салыстырғанда, шеткі бөліктерінде біршама ауытқулар байқалады.

Сонымен, қорыта айтқанда, Ципф заңдылығы үндіеуропа тілдерінен түзілген жиілік сөздіктері бойынша қалай орындалса, негізінен алғанда, түркі тілі, соның ішінде қазақ тілі мәтіндерінде де сол дәрежеде орындалады деуге әбден болады.

Егер $F_i=N\cdot K(i+\rho)$ заңдылығындағы K, ρ, γ параметрлері белгілі болса, онда әрбір реттік санға *(i)* қатысты абсолютті жиілік (F_i^{TM}) үшін, оған сәйкес келетін теориялық абсолютті жиілікті (F_i^T) анықтауға болады. Ал бұл Ципф заңының «тік сызықтық» пішіндегі теориялық жиіліктер үлестірілуін тәжірибе жүзінде алынған (эмпиризмдік) жиіліктер үлестірілуі арасындағы алшақтықты бағалауға мүмкіндік тудырады.

Әрбір рет саны *i*-ге сәйкес келетін F_i^{TM} мен F_i^T мәндерін салыстыру орта квадраттық ауытқу өрнегі арқылы жүзеге асады:

$$\sigma=\sqrt{\frac{1}{q}\cdot\sum_i\left(F_i^{TM}-F_i^T\right)^2}\ . \qquad (7)$$

Мұндағы q - байқау саны. Мысалы, «Абай жолы» романы мәтіндері бойынша түзілген жиілік сөздіктер бойынша орта квадраттық ауытқу мәндері мынадай:

$\sigma_1=18,280;\ \sigma_2=15,737;\ \sigma_3=15,550;\ \sigma_4=17,192$

және толық роман мәтіні бойынша $\sigma_5=66,261$.

3.2. Жиілік сөздіктегі ранг пен жиілік арасындағы корреляция

Тақырып атындағы «ранг» терминін жиілік сөздіктегі сөздердің рет саны деп, ал «корреляция» терминін екі өзгермелі шама арасындағы озара байланыстылық не арақатыстылық деп түсінген жөн.

Екі өзгермелі шама, яғни жиілік сөздіктегі сөздің (сөзтұлғаның) ранғі мен жиілігі арасындағы байланыстың тығыздығын немесе «күшін» және түрпатын (түрін) корреляциялық талдау негізінде анықтауға болады.

Егерде аталған байланыс түрін «түзу сызықтық» деп үйғарсақ, онда оның өрнегі $y=ax+b$ түрінде көрініс табу керек. Теңдеудегі $y=lgF_i$, $x=lgi$ билогарифмдік өстегі сызықтық функцияның ординат және абсцисс координаттары, i - ранг, F_i – рангке сәйкес сөздің абсолютті жиілігі. Яғни теңдеудің билогарифмдік өсіне қатысты түрі: $lgF_i=a\cdot lgi+b$ және оның жазықтықтағы сызбасы алдыңғы тақырыпшадағы *(1-суреттегі)* көрінісімен бірдей келеді.

Ендігі мақсат – тік сызықтың $y=ax+b$ теңдеуіндегі a және b параметрлерін анықтау. Қойылатын шарт – эмпиризмдік және теориялық сызықтар арасындағы кез келген екі жүп нүктелердің квадраттанған арақашықтықтары ең кіші дәрежеде болуы қажет.

Алға қойған мақсатқа жету үшін «ең кіші квадраттар әдісін» қолданып, ранг және жиілік арасындағы тәжірибе жүзіндегі - сынық сызық түрпаттағы кескінін «түзу сызықтық регрессия теориясы» бойынша сызықтық функция түріне келтіру керек. Ал регрессия теориясы ол үшін a мен b мәндерінің ең қолайлы деген мәндерін табуға мүмкіндік туғызады [2].

Мұндай жағдайда регрессия түзуінің теңдеуі былай жазылады:

$$y_x - \bar{y} = R\cdot\frac{\sigma_y}{\sigma_x}\cdot\left(x-\bar{x}\right), \qquad (1)$$

мұндағы $y_x=lgF_i$; $x=lgi$, x пен y билогарифмдік өске қатысты lgi мен lgF_i мәндерінің арифметикалық ортасы, ал σ_x пен σ_y – орта квадраттық ауытқулар, R – x пен y арасындағы байланыстың күшін анықтайтын корреляция коэффициенті.

Енді (1) теңдеуді $y=ax+b$ түріне келтіру үшін y_x анықтайтын болсақ, ол төмендегі (2) пішінді қабылдайды:

$$y_x = R \cdot \frac{\sigma_y}{\sigma_x} \cdot x + \left(\bar{y} - \bar{x} \cdot R \cdot \frac{\sigma_y}{\sigma_x} \right), \qquad (2)$$

ал a мен b мәндері: $a = R \cdot \frac{\sigma_y}{\sigma_x}$ және $b = \bar{y} - \bar{x} \cdot R \cdot \frac{\sigma_y}{\sigma_x}$.

3.3-кесте

«Абай жолы» романы жиілік сөздіктері бойынша Ципф заңындағы γ параметрі мен R корреляция коэффициентінің мәндері

Бөліктер	1-кітап		2-кітап		3-кітап		4-кітап	
	γ	R	γ	R	γ	R	γ	R
1-15	0, 97	0,97	0,47	0,94	0,48	0,97	0,38	0,99
1-50	0,98	0,98	0,59	0,98	0,59	0,99	0,56	0,98
1-100	0,99	0,99	0,62	0,99	0,63	0,99	0,59	0,99
1-300	0,99	0,99	0,71	0,99	0,70	0,99	0,69	0,99
1-1000	0,82	0,99	0,99	0,99	0,81	0,99	0,79	0,99
1-1500	0,85	0,99	0,84	0,99	0,84	0,99	0,82	0,99
1-2000	0,88	0,99	0,86	0,99	0,86	0,99	0,84	0,99
1-3000	0,91	0,99	0,89	0,99	0,88	0,99	0,87	0,99
1-4000	0,93	0,99	0,91	0,99	0,91	0,99	0,89	0,99
1-соңы	0,92	0,97	0,93	0,99	0,89	0,97	0,93	0,98
15-соңы	0,93	0,97	0,93	0,98	0,89	0,97	0,93	0,97
50-соңы	0,93	0,97	0,93	0,97	0,89	0,97	0,93	0,97
100-соңы	0,92	0,97	0,93	0,97	0,89	0,97	0,93	0,97
300-соңы	0,92	0,97	0,93	0,97	0,89	0,96	0,93	0,97
1000-соңы	0,89	0,95	0,91	0,96	0,86	0,95	0,91	0,96
1500-соңы	0,86	0,94	0,89	0,95	0,84	0,94	0,87	0,95
2000-соңы	0,83	0,93	0,87	0,94	0,81	0,93	0,89	0,94
3000-соңы	0,77	0,90	0,82	0,92	0,76	0,91	0,82	0,92
4000-соңы	0,67	0,86	0,76	0,89	0,70	0,87	0,77	0,89

Арнайы жазылған компьютерлік бағдарлама бойынша x, y, σ_x, σ_y, R шамаларын есептеп шығаруға болады. Енді a және b мәндері арқылы Ципф заңындағы γ және K параметрлерін де анықтауға болады.

«Абай жолы» романы толық мәтіні бойынша Ципф заңының γ параметрі, $lg(N \cdot K)$ және R корреляция коэффициенті

Зонаның р/с	Шекаралық бөліктер	γ	$lg(N \cdot K)$	R
1	1 –15	0,457	3,824	0,961
2	1 – 50	0,599	5,801	0,983
3	1 – 100	0,633	9,183	0,992
4	1 – 1000	0,815	4,271	0,995
5	1 – 4000	0,914	4,514	0,997
6	1 – 6000	0,950	5,616	0,996
7	1 – 10000	1,001	4,770	0,996
8	1 – 15000	1,045	4,912	0,995
9	1 – 25000	1,090	5,066	0,994
10	1 - соңы	1,065	4,951	0,982
11	15 - соңы	1,068	4,964	0,982
12	50 - соңы	1,071	4,978	0,981
13	100 - соңы	1,073	4,988	0,981
14	1000 - соңы	1,079	5,014	0,975
15	4000 - соңы	1,045	4,860	0,958
16	6000 - соңы	1,011	4,704	0,945
17	100000 - соңы	0,925	4,311	0,911
18	15000 - соңы	0,782	3,650	0,842
19	25000 - соңы	0,271	1,272	0,447

Ол үшін, $F_i=N \cdot K \cdot i^{\gamma}$ теңдеуінің екі жағы бірдей логарифмделсе жеткілікті:

$$lgF_i = -\gamma \cdot lgi + lg(N \cdot K).$$

Егер $y=lgF_i$ және $x=lgi$ екенін ескерсек, $a=-\gamma$ және $b=lg(N \cdot K)$ болатыны анықталады да, ал соңғы екі теңдіктен γ мен K мәндерін оңай табуға болады.

Жиілік сөздіктердегі сөздерді өз қалағанымызша зоналарға (бөліктерге) бөлуге болады және әр зонаға сай γ мен K мәндерін анықтауға мүмкіндік бар.

Алдыңғы тақырыпта қарастырған «Абай жолы» романы бойынша түзілген әр кітаптың және толық роман мәтінінің жиілік сөздіктерінің бөліктеріне қатысты есептелген корреляция коэффициенттері екі түрлі кестелер *(3.4* және *3.5)* арқылы көрініс тапты.

3.3-кесте мен *3.4-кестеде* келтірілген корреляция коэффициенттері жайлы деректерді қарастыра келе, мынадай қорытынды жасауға болады: түзу регрессия кезінде корреляция коэффициентінің *(R)* мәні *y* пен *x* шамаларының түзу сызықтық байланысының тығыздығын білдіретін және оның «күшін» бағалайтын параметр ретінде қарастырылады.

Сонымен, егер:

а) $0 \le R \le 1$ болса, *x* пен *y* аралығындағы байланыс – «тікелей байланыс» (прямая зависимость) болып саналады;

ə) $-1 \le R \le 0$ болса – «кері байланыс» (обратная зависимость) болып саналады;

в) корреляция коэффициентінің *(R)* мәні *1* санына жуықтаған сайын аталған байланыстың тығыздығы күшейе түседі.

Кестеде көрсетілген деректер бойынша (*R*-дің мәндері), «Абай жолы» романының әр кітабының жиілік сөздіктеріндегі сөзтұлға жиілігі мен оның рангісінің (реттік санының) арасындағы байланыс – кері байланыс екендігін $(-1 \le R \le 0)$ және олардың айтарлықтай «күшті» еместігін атауға болады. Корреляция коэффициентінің ең үлкен *(-0,994÷-0,996)* мәндері жиілік сөздіктің орта түстарына сәйкес келетіндігі анықталды. Сонымен бірге, сөздіктердің бас жағына қатысты байланыс күші, оның соңғы жақтарымен салыстырғанда күштілеу (тығыздау) болатындығы да байқалады. «Абай жолы» романы мәтінінен түзілген сөздіктердегі сөзтұлғалардың рет саны (ранг) мен жиіліктері арасындағы корреляциялық қатынастылық күшінің тығыздығын корреляция коэффициентінің арифметикалық ортасын анықтап та білуге болады:

$$\overline{R} = \frac{\sum_{j=1}^{19} R_j}{19}.$$

«Абай жолы» романының әрбір томы мен олардың қосындысы (роман) бойынша есептелген корреляция коэффициентінің орта шамасы мен жиіліктің тәжірибелік және теориялық мәндерінің орта квадраттық ауытқулары төмендегідей *(3.5-кестеде)* көрініс тапты:

3.5-кесте

Корреляция коэффициентінің орта шамасы, тәжірибелік және теориялық жиілік мәндерінің орта квадраттық ауытқулары

«Абай жолы» романы мәтіндері	1-том	2-том	3-том	4-том	Роман
R	- 0,966	-0,969	-0,967	-0,971	- 0,943
σ_R	0,00033	0,00028	0,00033	0,00028	0,00014

Корреляциялық коэффициент мәнінің сенімділігіне баға беру (оценка надежности) үшін «квадраттанған қатенің ортасын» анықтайтын мына өрнекті (формуланы) қолдануға болады [20, 35]:

$$\sigma_R = \frac{1 - R^2}{\sqrt{N}} \qquad (4)$$

Мұндағы R – жиілік сөздік бойын түгел қамтыған жағдайдағы корреляция коэффициенті, N – жиілік сөздік бойындағы әр түрлі сөзтұлғалар (сөздер) саны. Егер N айтарлықтай көп болса, онда корреляция коэффициентінің ақиқаттық шекаралық мәні мынадай интервал аралығында жататындығы дәлелденді:

$$R - 3\sigma_R \le R_{ақиқ} \le R + 3\sigma_R .$$

Біз қарастырып отырған «Абай жолы» романының сөздіктерінде мұндай интервалдар төмендегідей:

$$-0,975 \le R_1 \le -0,974;$$
$$-0,977 \le R_2 \le -0,975;$$
$$-0,973 \le R_3 \le -0,972;$$

$$-0,976 \leq R_1 \leq -0,975;$$
$$-0,982 \leq R_{роу} \leq -0,981.$$

Сонымен, корреляция коэффициентін сенімділік дәрежеге бағалау нәтижелері R мәнінің жоғары дәрежедегі дәлдік болатындығына дәлел. Ал роман материалдары бойынша алынған сөздіктердің өн бойына тән қасиет – корреляциялық байланыстың түзу сызықты сипатта болуы және ондай байланыстың сөздік бойының орта жағында «күшті», ал шеткі жақтарында «әлсіз» болатындығы.

3.3. Мәтін және оның жиілік сөздігі ішіндегі сөздің (сөзтұлғаның) ақпараттық сипаттамасы

Тіл – қатынас құралы. Сондықтан оның элементтеріне ақпараттық өлшем жүргізілуі қажет және оған тиісті баға берілуі керек деп ұйғарылады.

Осындай пайымдаулардан мынадай тұжырымдар туындайтыны анық:

1. Тіл арқылы қатынас жасау (речевое общение) – тілді байланыс торабына (каналына) жатқызуды қажет етеді. Ақпараттық байланыс тілдің әріп, дыбыс, морфема және басқа тілдік бірліктері арқылы іске асады.

2. Тіл арқылы қатынас жасау үшін тілдік бірліктерді белгілі бір «кодтың» символдары ретінде қарастыру керек. «Код» дегеніміз – «тіл», бірақ шектеулі тіл. Себебі, бұл жағдайда тілдегі тіркесімдер мен басқа да бірліктердің қолдану ықтималдығына қажетті шектеулер қойылады.

3. Тілдік қатынас кезінде хабарламаны тарату мен оны қабылдау аралығын қосу – сөйлеу каналы (торабы) арқылы іске асады. Бұл жағдайда айтушы мен қабылдаушы бірдей «кодты» пайдаланулары қажет.

Синхрондық лингвистика саласындағы әрбір зерттеушінің көздейтін мақсаты хабарлама ішіндегі лингвистикалық бірліктердің көрінісін реттейтін тілдік «кодтағы» шектеулерді анықтау және бағалау. Осындай бағалаудың ең бір ұтымды деп саналатын әсерлі әдісі, ол – лингвистикалық бірліктер мен қатынастарды осы тәріздес ақпараттық өлшем – энтропия, яғни

артықтық (избыточность) сияқты сандық дәрежедегі теориялық-ақпараттық шамалармен салыстыру.

Тілді ақпараттық тұрғыда бағалау мүмкіндігі тек нақты тілде жазылған мәтінді статистикалық зерттеу арқылы ғана іске асады. Мәтін арқылы таратылатын ақпарат мөлшерін анықтауда ондай мәтін белгілі бір «кодқа» енетін дискретті бірліктер тізбегі ретінде қарастырылады. Дискретті тілдік бірлік – ол сойлемдер, сөздер, буындар, фонемалар, әріптер. Осы аталған бірліктердің әрбіреуінің мәтін бойынан орын алуының көрінісі тәжірибе тізбегінің қайталану ретімен іске асады. Бұл жердегі «тәжірибе жүргізу» дегеніміз – бірліктер жиыны ішінен осы мәтін бөлігінде шығуы мүмкін болатын қайсыбір мәтін бірлігін ғандап алу деп түсінген жөн. Осылайша таңдалып алынған әрбір бірлік – айнымалы *X* шамасының әр түрлі ықтималдықпен қабылданатын мәндерінің бірі деп саналады.

Белгісіздік сипаттағы осындай айнымалы *X*-тің сандық мәні «энтропия» терминімен аталады. Айнымалы *X*-тің нақты мәніне сілтеме жасау дегеніміз, ол энтропияны (белгісіздікті) жою және сондай мөлшердегі ақпарат алу. Бұл түсінікте ақпараттың маңыздылық (семантикалық) жағы сөз болуы мүмкін емес. Сондықтан мұндай жағдайда тек селективті ақпарат жайлы сөз болады. Яғни мәтін бөлігіндегі ақпараттың лингвистикалық бірліктерін таңдау ықтималдық жүйе арқылы іске асатындығы ескерілуі қажет. Іс жүзінде тілдің синтаксис пен лексика салаларында элементарлы бірліктер саны (сөйлемдер, сөзтіркестер, сөздер) шексіздікке дейін өсуі мүмкін. Сол сияқты, тілдің фонетика саласындағы фонемаларды (дыбысты) дискретті бірліктерге ажырату проблемасының да шешілмеген жақтары баршылық. Сондықтан да теориялық-ақпараттық зерттеулерде, көбінесе, әріптік код көрінісіндегі және шекті дискретті бірліктен тұратын жазба тіл материалдары пайдаланылады. Осыған байланысты тілдегі ақпараттық өлшеуді іске асыру – мәтін бойындағы әр әріпке қатысты энтропияны анықтаудың әрекеті. Әрбір мағыналы мәтін өзара қатынастық сипаттағы және әр дәрежелі ықтималдықпен анықталатын әріптер тізбегі екені мәлім. Бірақ энтропияның әрбір әріпке шаққанда келетін «шындық мәніне» (истинное значение - *H*) жуық келетін мәнін

табу үшін мәтін әріптері бірдей ықтималдықта және өзара тәуелсіз сипатта болу қажет деген ұйғарым жасалуы керек.

Әрине, энтропиялық өлшемді іске асыру мәселесін толық түрде қарастыру бұл жұмыстың мақсатына жатпайды. Сондықтан бұл мәселе жайлы жайттарды толығырақ білгісі келген оқырманды «Энтропия языка и статистика речи» (Минск, 1966) [99] атты жинақпен және «Әдебиет» тізімінде көрсетілген басқа да ғылыми еңбектермен танысуына болады [76].

Келесі қарастырмақшы мәселеміз М.Әуезовтің «Абай жолы» романы мәтінінен түзілген «сөзтұлғалар жиілік сөздіктері» бойынша қазақ көркем әдебиет мәтініндегі бір сөзтұлғаға түсетін орташа ақпарат мөлшерін анықтау. Ол үшін бірдей дәрежедегі бірлікке түсетін «орташа ақпарат мөлшерінің» орташа энтропияға *(H)* тең болатындығын ескере отырып, К.Шеннонның белгілі формауласын пайдалануға болады:

$$I = H = -\sum_{i=L}^{L_{кол}} f_i \cdot \ell og_2 f_i , \qquad (1)$$

мұндағы f_i – қатынастық жиілік i – сөзтұлғаның сөздіктегі рет саны, $f_i \cdot \ell og_2 f_i$ – көбейтіндісі арқылы сөздіктегі кезекті рет саны үшін хабар бірлігіндегі «өлшенген ақпарат» (взвешенная информация) мөлшері анықталады. Ал ақпарат мөлшері *(I=H)* екілік санау жүйесіндегі «бит» өлшемімен анықталады.

Аталған (1) өрнекті қолдану үшін, жиілік сөздіктегі сөзтұлғаларды мәтін бойынан рет-ретімен жорамалмен таңдалып алынатын тәжірибе нәтижелерінің жиынтығы деп қарастыру қажет. Әрбір тәжірибе нәтижесі: мәтін ішіндегі кездесетін сөзтұлғаның қатынастық жиілігіне тең болатын нақты ықтималдық шамасын анықтау.

«Өлшенген жиынтық ақпарат» шамасын есептеу үшін (1) өрнекті басқаша түрде (жиілік сөздіктің *1* санынан *n* саны аралығын қоса есептейтіндей) қайта жазсақ:

$$I^* = H^* = \sum_{i=1}^{n} f_i \cdot \ell og_2 f_i . \qquad (2)$$

Бұл өрнек бойынша, сөзтұлға жиілігі мен «өлшенген ақпарат» санының тікелей қатыстығын анықтай аламыз. Ауызша не жазбаша мәтіндегі сөзтұлғаның қолдану жиілігі ұлғайған сайын, оның «өлшенген ақпарат» саны да артады. Басқаша айтқанда, сөзтұлғаның мәтінге «енгізетін» ақпараты арта түскен сайын, оның қолданысы да көбейе түседі және мәтін ішіндегі басқа сөздермен тіркесімдігі (немесе мәнмәтіндік қоршауы) де жиіліктік сипатта болады (және керісінше).

Сол сияқты, (2) өрнек арқылы жиілік сөздіктің әр бөлігіне қатысты «жиынтық өлшенген ақпаратты» және оның барлық сөздік бойынан алатын пайыздық үлесін де анықтауға мүмкіндік туады. Бұл жердегі пайыздық үлес – «ақпараттық салмақ» (информационный вес) деген терминмен аталып, мына түрде жазылады:

$$\eta = \frac{I^*}{I} \cdot 100\ \% . \qquad (3)$$

Мұндағы I – бір сөзтұлғаға қатысты «орташа ақпарат саны», ал I^* - бір топ сөзтұлғаларға қатысты жиынтық ақпарат салмағы.

М. Әуезовтің «Абай жолы» романы мәтінінің жиілік сөздігі бойынша бір топ сөзтұлғаның мәтінді қамту пайызына сәйкес жиынтық ақпарат саны мен ақпараттық салмақтың өсу қарқыны *3.6-кестеде* көрініс тапты. Бұл кестеде романның толық мәтіні мен 4-ші кітабы мәтінінің жиілік сөздіктеріндегі сөзтұлғаларға қатысты I^* мен η мәндері сөзтұлғалардың мәтін бойын әр түрлі *(1*-ден *100*-ге дейінгі)* аралықтағы пайыздық қамту шамаларына сәйкестендіріліп берілген. Пайыздық қамту қарқыны артқан сайын, жиынтық ақпарат салмағы *(I*)* шамасы да сондай шапшаңдықпен өседі. Мысалы, сөзтұлғаның 100% қамтуына *4*-ші кітап бойынша *12,5* дв.ед. ал роман бойынша – *12,7* дв.ед. сәйкес келеді. Бұл мәндер, орта шамамен алғандағы, бір сөзтұлғаға сәйкес келетін ақпарат мөлшерін анықтайды.

Жиілік сөздіктің алдыңғы жағынан соңына қарай есептеу нәтижесінде сөзтұлғалардың мәтінді қамту қарқыны ақпараттық салмақтың (η) өсу қарқынымен салыстырғанда шапшаңдау екенін байқауға болады. Сирек қолданыстағы сөзтұлға топтарына сәйкес келетін мәтінді қамту пайызы *(Q)* мен

ақпараттық салмақ (η) мәндері және олардың қосынды шамалары *3.7-кестеде* берілді. Осы кестедегі мәліметтер бойынша, егер жиі қолданыстағы сөздер үшін «ақпараттық салмақ» пайызы (η) «мәтінді қамту» пайызынан (Q) кіші болатын болса, сирек қолданатын сөздер үшін, керісінше, үлкендік сипатта.

Екі кестеде де әр түрлі көлемдегі мәтіндерден (V_1=123053; $V_{роман}$=465966) түзілген жиілік сөздіктер мәліметтері (I^*, η) қарастырылды. Оларды салыстыра қарастырудан шығатын қорытынды – сөздік бірлігінің ақпараттық өлшем дәрежесіне мәтін көлемінің айтарлықтай әсері болмауы.

3.6-кесте

М.Әуезовтің «Абай жолы» романы мәтінінің жиілік сөздігі бойынша бір топ созтұлғаның мәтінді қамту пайызына сәйкес жиынтық ақпарат саны мен ақпараттық салмақтың өсу қарқыны

Мәтінді қамту пайызы	4-кітап мәтіні бойынша		Роман мәтіні бойынша	
	I^*	η	I^*	η
1	0,0715	0,5727	0,0769	0,6036
5	2,2871	2,2996	0,3168	2,4867
10	0,7475	5,9874	0,7316	5,7427
15	1,1704	9,3748	1,1404	8,9515
20	1,6311	13,0650	1,5968	12,5183
25	2,1323	17,0796	2,0797	16,3246
30	2,6515	21,2383	2,6017	20,4220
35	3,2318	25,8865	3,1375	24,6277
40	3,7798	30,2759	3,7144	29,1561
50	4,9863	39,9399	4,9153	38,5825
60	6,2581	50,1270	6,2170	48,8041
70	7,7264	61,8879	7,6722	60,2227
80	9,2713	74,2625	9,2539	72,6383
100	12,4845	100,0	12,7397	100,0

Созтұлғаларды жеке және топтап қарастырып, олардың функционалдық мәнін анықтауда ақпараттық салмақ – түрпайы сипаттағы баға деуге болады. Себебі, мұнда сөйлеу қызметі

кезіндегі сөз бен сөз тіркестері арқылы байланысқа қатысты селективті ақпараттың 20–30 пайызын құрайтын лексика-грамматикалық ақпарат шамасы ескерілмей отыр. Солай бола тұра, көптеген зерттеушілердің пайымдауынша, сөздердің мәтінді қамту пайызымен салыстырғанда, «ақпараттық салмақ» мәні лексикалық бірліктердің функционалдық қарым-қатынасын мағыналы түрде сипаттайды.

3.7-кесте

Жиілік F_i	4-кітап мәтіні бойынша				Роман мәтіні бойынша			
	Q	$\sum Q$	η	$\sum \eta$	Q	$\sum Q$	η	$\sum \eta$
1	2	3	4	5	6	7	8	9
1	13,56	13,56	17,59	17,59	7,48	7,48	11,06	11,06
2	6,33	19,89	7,72	25,31	3,75	11,23	5,25	16,31
3	4,37	24,26	5,13	30,44	2,74	13,97	3,71	20,02
4	3,41	27,67	3,90	34,34	2,18	16,97	2,88	22,90
5	2,80	30,47	3,14	37,48	1,94	18,09	2,52	25,42
6	2,43	32,90	2,67	40,15	1,52	19,61	1,95	27,37
7	1,96	34,86	2,12	42,27	1,41	21,02	1,77	29,14
8	1,93	36,79	2,06	44,33	1,35	22,37	1,67	30,81
9	1,61	38,40	1,70	46,03	1,23	23,60	1,52	32,33
10	1,39	39,79	1,45	47,48	1,00	24,60	1,30	33,63

3.4. Лингвистикалық болжамды тексеру (сынау) критерийі

Қолданбалы лингвистиканың көптеген маңызды мәселелерінің бірі – мәтін ішіндегі тілдік бірліктердің орналасу үлгісінің (моделінің) ықтималды-статистикалық әмбебап сызбасын құру. Мұндай күрделі істі автоматты түрге келтіру: ақпаратты іздеу, тілдерді тиімді жолмен үйрету, базалық (негізгі) тілді құру, тілдің «нормалық» табиғатын түсіну, тілді стилистика тұрғысынан зерттеу және т.б. кешенді мәселелерді шешуде аса қажетті деп саналады [92].

Тілді математикалық модельдеу кезінде оның негізгі мәселесі болып саналатыны: тәжірибе жүзінде алынған сандық деректерге статистикалық талдау жүмысын жүргізу. Аталған

талдау арқылы кейбір дәстүрлі зерттеу кезінде тікелей байқауға жатпайтын бірқатар тілдік жайттардың ара-жігі анықталады. Ал тілдің ықтималдық үлгісі көмегімен мәтін ішіндегі қайсыбір тілдік бірліктердің сандық қасиеттерін сипаттауға мүмкіндік туады. Егер осы анықталған бірліктер сипаттамасын ақиқат шындықтың көрсеткіші ретінде қабылданса, онда модельден нақты нысанға өту іске асады. Ол үшін зерттеуге алынған мәтін бірліктерінің сандық параметрлерін анықтай білу қажет. Бұл параметрлерге «арифметикалық орта» ($\overline{X}$) мен «стандартты» (S) жатқызуға болады. Әрі қарай $\overline{X}$ пен S мәндері «ықтималдық модельдің» осы тәріздес параметрлерімен салыстырылады.

Тілдік бірліктердің сипаттамалары зерттеушіні сол тәжірибеге түскен шекті көлемдегі нақты мәтінге ғана қатысты қызықтырмайды, ол бас мәтін жиынтығы сипаттамасы тұрғысынан қарастырғанды қалайды. Бұл жердегі «бас мәтін жиынтығы» (генеральная собокупность текстов) дегеніміз тілдің толық түрі немесе бөлек бір стиль, тіл қабаттары, жазушы тілі және т.б. Бұлайша қарастырудың себебі аталған «бас мәтін жиынтығын» қамтитын толық түрдегі мәтін бойынша статистикалық талдау жүргізу тәжірибе жүзінде мүмкін бола бермейді. Яғни зерттеуге жататын нысан «бас мәтін жиынтығының» таңдалып алынған жеке бөліктері (частные выборки) немесе «таңдама бөліктер». Міне, осындай таңдама мәтін бөліктер параметрлерін бас мәтін жиынына қатысты статистика-математикалық зерттеу негізінде ықтималды модельден тілдік нысанның өзіне көшу (өту) жүзеге асады. Мұндай «өту» процесі нақты (белгілі) критерийлерге сүйеніп, статистика-лингвистикалық болжамды сынау негізінде бірліктердің (параметрлердің) мәтін ішінде кездейсоқ оқиға ретінде «үлестірілу» (таралу) түрпатын (түрін) анықтау арқылы іске асады.

Мәселен, мынадай статистика-лингвистикалық болжамды сынау (тексеру) қажет болды делік: мәтін ішіндегі зат есімге қатысты сөздер математикалық статистиканың «нормальды үлестірілу» деп аталатын теориялық заңдылығына бағынады. Мұндай болжамды сынау (тексеру) үшін зат есім сөздердің эмпирикалық (тәжірибелік) үлестірілу деректерінің теориялық

үлестірілуі «нормальды» заңдылығымен сәйкестік дәрежесін тексеру қажет. Ал ондай тексеру (сынау) тек қана зат есім сөздердің таңдама бөліктер ішінде қайталану жиілігінің орта шамасының бірнеше «сериялары» (мүмкіндіктері) бойынша қарастырылады.

Кездейсоқтық жағдайдағы лингвистикалық шаманың мәтін бойында таралу (үлестірілу) түрпатының не параметрлерінің сынауға ұсынылып отырған болжамға сәйкестігі нөлдік (негізгі) болжам (H_0) немесе оған қарама-қарсы қойылатын альтернативті болжам (H_1) ретінде тұжырымдалуы мүмкін. Негізгі – нөлдік (H_0) болжам белгілі критерий арқылы сыналады.

Критерийлер – статистикалық және реттік деп екіге ажыратылады. Алғашқысы (статистикалық) кездейсоқ лингвистикалық шаманың үлестірілуіне (қолдану жиілігіне) бағыттала отырып, нөлдік болжамды сынауға қолданылады, ал екіншісі аталған шамалардың реттік үлестірілуін қарастырады.

Әдетте, лингвистикалық болжамдар кездейсоқ таңдама бөліктер негізінде тексеріледі. Бірақ бұл тілдік бірліктің бас жиындағы үлестірілу көрінісінің дұрыстығының айғағы бола алмайды, себебі, мұндай жағдай «болжамды сынау» процесін жалған жолға бұрып жіберуі де мүмкін. Сондықтан дұрыс шешім орнына, мысалы, дұрыс нөлдік (H_0) болжам орнына, бұрыс (H_1) болжам (не керісінше) қабылданады. Осындай жайттарға байланысты зерттеушінің қай болжамды $(H_0$ не $H_1)$ негізгі деп қабылдауының да көп мәні бар.

Осы айтылғаннан жіберілетін қателердің мәнділік дәрежесі екі түрлі типпен ажыратылады. Олар – 1-ші және 2-ші дәрежелік мәнді қателер деп аталады. Тіл білімінде қатенің мәнділік дәрежелерін таңдау, көбінде, субъективті (жеке бастық) ой қорыту негізінде іске асады. Мысалы, түркі, монғол, тунгус-маньчжур тілдерінің «алтайлық гипотеза» деп аталатын генетикалық туыстығы жайлы болжам белгілі статистикалық критерий арқылы тексеруге болатындай тұжырымдалған [54].

Мұндағы нөлдік болжам бойынша барлық алтай тілдері бір ғана генетикалық туыстас жеріне, яғни негіз-тілге (язык-основа) келіп тіреледі екен (восходят). Осы тұжырымды теріске шығаратын өзіне нағыз кәміл сенетін алтайшы да аса қауіпті

1-ші дәрежелік мәндегі қате жіберген болар еді. Бұл жерде нөлдікке қарсы қабылданған болжамның өзі дұрыс болмаса да, мәнділік жағынан 2-ші дәрежелі қате қауіпсіздеу деуге болады.

Сондықтан, болжамды сынау критерийін таңдау, сонымен бірге сынау аясын таңдаумен үштасады деуге болады.

Сөйтіп қандай да болсын дұрыс шешім қабылдар түсында 2-ші типті қате жіберуден гөрі 1-ші типті қате жібермеу аса маңызды деп саналады. Қате дәрежесінің мәнділігін азайту мақсатымен болжамдар өзара сыналуы қажет.

Тәжірибелік құндылығына қарай, яғни 1-ші типті қатеден күтілетін салдарына қарай, зерттеушіні қанағаттандыратын маңыздылық деңгейі таңдалады. Басқаша айтсақ, жіберген қатеден туындайтын салдардың сипатына қарай маңыздылық деңгейін де азайтқан (төмендеткен) жөн болады.

Әдетте, лингвистикалық зерттеулерде маңыздылық деңгейінің $\alpha = 0,05$ немесе $\alpha = 0,01$ шамасы да жеткілікті деп саналады. Мәселен, алтай тілдеріне қатысты болжамды статистикалық тәсілмен тексеру (сынау) үшін көрсетілген маңыздылық деңгей шамасы қанағаттандырарлық деп саналады.

3.5. Қазақ мәтінінің ықтималды-статистикалық үлгісін (моделін) құру

Лингвистикалық статистика қазір өзінің алғашқы кезеңін, яғни лингвистикалық құбылыстарды сипаттауды аяқтап, келесі күрделі кезеңіне өтуде. Бұл – тілдік заңдылықтарды сандық түрғыда алдын ала болжай алатын теория құру мәселесі.

Сонымен, лингвистикалық бірліктердің мәтін ішінде үлестірілуінің математикалық түрпатын анықтау мәселесі тілдік құбылыстың моделін құрудағы ең бір маңызды іске айналып отыр [93].

Аталған мәселе бойынша қазақ тілі мәтіні ішінде сөз таптарының статистикалық үлестірілуінің кейбір теориялық заңдылықтарға сәйкес келу болжамын сынауға қатысты зерттеулер жүргізілді. Теориялық үлестірілу заңдылықтары ретінде математикалық статистикадағы «нормальды үлестірілуі», «Пуассон үлестірілуі» және «Шарлье үлестірілуі» заңдылықтары қарастырылды.

Қазақ тіліндегі мәтін нысаны ретінде М.Әуезовтің «Абай жолы» романынан 10 мың сөзқолданыс, газет мәтінінен 50 мың сөзқолданыс таңдама бөліктер тәжірибе ретінде алынды. Ал сөз таптары ішінен ең негізгілері деп: *зат есім, етістік, сын есім* және *үстеу* сөздердің мәтін ішіндегі статистикалық үлестірілуі сынаққа түсті.

Белгілі бір лингвистикалық бірліктің (мысалы, белгілі бір сөз табының) тілдегі қолданылу сипатын танып білу кезінде құбылу заңдылығы бұзылмайтын ең кіші (минимум) мәтін көлемін анықтау зерттеу ісінің маңызды шартына айналып отыр.

Осындай мақсатпен жүргізілген тәжірибе реті төмендегідей:

1) мәтін ішіндегі сөздерді сөз табына ажырату және оларды шартты белгілермен сәйкестендіру;

2) мәтіннің жалпы көлемі *25, 50, 100, 200* және *500 сөзқолданыстан* тұратын микробөліктерге, яғни сериялық таңдама бөліктерге бөлу;

3) өзгермелі шамаларды шартты белгілеуді ұстану: әр стиль бойынша алынған мәтіннің жалпы көлемі – *N;* микробөлік (серия) көлемі – *k;* әр серияға сай микробөлік саны – *n;*

4) әр жағдайдың жалпылама жазылу пішінін қабылдау – $B^{n}_{N}=k$ (теңдік белгісі шартты). Мұндағы $N=n\cdot k$. Мысалы, «Абай жолы» романы мәтіні бойынша алынған 10000 сөзқолданыстан тұратын мәтін көлемін үшке бөліп: $N_1=2500$, $N_2=5000$, $N_3=10000$ қарастырсақ, әрбір N_i көлемді $k=25, 50, 100, 200, 500$ сөзқолданыстан тұратын микробөліктерге (n) болсек, олардың саны өзгеріп отырады. Мәселен, егер $k=25$ болса:

$N=2500$ болса, $B^{25}_{2500}=100$ бөлікке тең;

$N=5000$ болса, $B^{25}_{5000}=200$ бөлікке тең;

$N=10000$ болса, $B^{25}_{10000}=400$ бөлікке тең.

Сол сияқты, егер $k=50$ болса: $B^{50}_{2500}=50$; $B^{50}_{5000}=100$; $B^{50}_{10000}=200$ немесе егер $k=100$ болса: $B^{100}_{2500}=25$; $B^{100}_{5000}=50$; $B^{100}_{10000}=100$.

Газет мәтіні бойынша тәжірибеге алынған *50000* сөзқолданыстан тұратын мәтін көлемі де, жоғарыда көрсеткендей, серия бөліктеріне бөлініп зерттелді;

5) әрбір вариантқа қатысты микробөліктер ішіндегі қажетті сөз табының жиілігі саналып, бірдей жиілікпен кездесетін сөз табының серия саны анықталды. Осы деректердің негізінде жиілік пен микробөлік санынан тұратын төмендегідей дискретті вариациялық үлестірілімдік қатар түзілді:

x_i	x_1	x_2	x_3	...	x_s
n_i	n_1	n_2	n_3	...	n_s

Мұндағы x_i *(i=1,2,3,...,s)* сөз табының жиілігі, n_i *(i= 1,2,3,...,s)* – әрбір x_i-ге сай микробөлік саны, $\sum_{i=1}^{s} n_i = n$.

Жүргізілген тәжірибе бойынша, көркем әдебиет пен газет мәтіндеріндегі негізгі сөз таптарының мәтін бойында таралуының вариациялық қатарларын (статистикалық үлестірілудің) кейбір теориялық заңдылықтармен үйлесімділік деңгейі зерттелді. Теориялық заңдылықтар ретінде Пуассонның, Шарльенің *(А, В түрлері)* және нормальдық, логарифмды-нормальдық үлестірілу заңдылықтары алынды.

Дискретті және үздіксіз вариациялық қатар құрудың мысалы ретінде «Абай жолы» романы мәтінінің *N=10000 сөзқолданыстан* тұратын таңдама бөлігін қарастырайық. Егер ішкі бөліктер (микровыборки) немесе серия таңдамаларының ұзындығы *k=100* сөзқолданыс десек, серия саны $B_N^n = k = B_{10000}^{100} = 100$ болады екен. Енді әр сериялық мәтін ішіндегі «зат есім» сөздердің жиілігін анықтап, олардың ең аз қолдану саны – *19*, ал ең көп қолдану саны – *43* екенін білдік делік. Әрі қарай, *19* бен *43* сандары аралығын немесе *(19; 43)* интервалы аралығын *«1»* санына (жиілікке) өсіру негізінде мынадай дискретті вариациялық қатар құруға болады:

x_i	*19*	*20*	*21*	*22*	*23*	*24*	*25*	...	*40*	*41*	*42*	*43*	n_i қосындысы:
n_i	*1*	*0*	*2*	*3*	*3*	*3*	*5*	...	*1*	*0*	*0*	*1*	*100*

Жоғарыда аталған теориялық үлестірілу заңдылықтардың *Пуассон* және *Шарлье-В түрі* дискретті вариациялық қатар құру арқылы, ал *нормальды, логарифмды-нормальды* және *Шарлье-А түрі* үздіксіз вариациялық қатар құру негізінде олардың теориялық жиіліктері мен статистикалық параметрлері анықталды.

Дискретті вариациялық қатардан үздіксіз вариация қатарына көшу үшін Δx-ке тең шама (қадам) таңдалып алынып, $(x_i; x_{i+\Delta x})$ интервалдық қатар құрылады. Интервалдың шекаралық мәндерінің арифметикалық ортасы – x_j' және олардың әрқайсысына сай келетін n_j' табуға болады:

$$x_j' = \frac{x_i + x_{i+\Delta x}}{2}, \qquad (1)$$

$$n_j' = \sum_{k=i}^{i+x_{i+\Delta x-1}} n_k . \qquad (2)$$

Мұндағы $i=1,2,3, \ldots, n$ – дискретті қатардың рет саны, $j=1,2,3,\ldots, m$ – үздіксіз қатардың рет саны.

Егер $\Delta x=3$, болса үздіксіз вариациялық қатар төмендегі *3.8-кестеде* көрсетілгендей болады:

3.8-кесте

$(x_i; x_{i+\Delta x})$	*(19; 22)*	*(22; 25)*	*(25; 28)*	...	*(40; 43)*	*(43; 46)*
x_j'	*20,5*	*23,5*	*26,5*	...	*41,5*	*44,5*
n_j'	*3*	*9*	*12*	...	*2*	*1*

Сонымен, осылайша құрған дискретті және үздіксіз вариациялық қатарлар мәтінге қатысты сөз таптары жиіліктерінің эмпирикалық үлестірілу заңдылығын анықтайды. Бұл деректер арнайы компьютерлік бағдарламалар негізінде теориялық жиілік пен статистикалық параметрлерді есептеп шығару үшін қажет.

Енді жоғарыда аталған теориялық үлестірілу заңдылықтардың екі ғана түрінің өрнектеріне (формуласына) тоқталайық.

3.5.1. Пуассонның үлестірілу заңдылығы

Лингвистикалық зерттеулерге қатысты ең бір маңызды теориялық заңдылық – Пуассонның үлестірілу заңдылығы. Бұл заңдылықты, көбінде, тәжірибе кезінде күтілетін әр оқиғаның шығу (пайда болу) ықтималдығы *(p)* өте аз, ал тәжірибенің саны барынша көп болған жағдайда қолданады.

Осы қойылатын шарттар қазақ тілінің мәтіндерінде кездесетін сөз таптарына қатысты да орындалады деп жорамал жасауға болады.

Пуассон заңдылығын қолдану үшін, жоғарыда айтылғандай, сөз таптарының статистикалық деректерінің негізінде дискретті вариациялық қатар құруымыз керек болады. Тәжірибе кезінде қалаған оқиғаның шығуын, мысалы, *зат есім сөздердің* кездесу мүмкіндігін (шамасын) x, «математикалық күтуді» (математическое ожидание) – a немесе оның орнына жүретін арифметикалық ортаны – $\bar{X}$, тәжірибе санын – n деп белгілесек, теориялық жиілік санын анықтайтын Пуассон заңының өрнегі төмендегідей жазылады [85]:

$$n^T = \frac{a^x \cdot e^{-a}}{x!} \cdot n. \qquad (3)$$

Бұл (3) өрнектегі «математикалық күтудің» орнына жүретін арифметикалық ортаның $(\bar{X})$ өрнегі мына түрде жазылады:

$$a \approx \bar{x} = \frac{1}{n} \cdot \sum n_i \cdot x_i' \,. \qquad (4)$$

3.5.2. Нормальды үлестірілу заңдылығы

Нормальды үлестірілу заңдылығы статистикалық зерттеулерде кеңінен қолданады деуге болады. Оның негізгі себебі, бұл заңдылықтың сызба пішінінің аса қарапайымдылығы және ол заңдылықтың көптеген *«жалғыз төбелік»* сипаттағы үлестірілулерге пішіндік жағынан жақындығы. *Нормальды үлестірілу заңдылығының* тағы бір ұтымды жері – тәжірибелік байқау кезіндегі деректерге қатысты әр түрлі математика-статистикалық өңдеу тәжірибесінің молдығы. Заңдылықтың ең

бір маңызды деп саналатын тағы бір жағы – ол мәтін бөлігінің көлемін анықтау мен статистикалық баға беруде (сынауда) қолдану мүмкіндігі.

Айнымалы және үздіксіз x шама үлестірілудің $\varphi(x)$ функцияның аргументі ретінде $-\infty$ *(минус шексіздік)* пен $+\infty$ *(плюс шексіздік)* аралығындағы шамаларды қабылдайды. Ал «x» шаманың үлестірілуінің нормальды заңдылығы төмендегі пішінде көрініс табады:

$$n^{III} = \varphi(x) \cdot n = \frac{n \cdot \Delta x}{\sigma} \cdot \frac{1}{\sqrt{2\Pi}} \cdot \ell^{-\frac{(x-a)^2}{2\sigma^2}}. \qquad (5)$$

Мұндағы Δx – дискретті вариациялық қатардан үздіксіз қатарға ауысуға қажетті интервал ұзындығы, a – «математикалық күту» (жоғарыдағы 4-ші өрнекпен анықталады). Ал σ^2 – дисперсия немесе ол *«екінші реттегі орталық момент»* деп те аталады және оның математикалық өрнегі томендегідей:

$$\sigma^2 = \mu_2 = \frac{1}{n-1} \cdot \sum_i x_i \cdot (x_i - a)^2. \qquad (6)$$

Математикалық күту *(a)* (4) мен дисперсия (σ^2) (6) статистикалық параметрлер арқылы үлестірілу заңдылығының маңызды жақтары сипатталады. Мысалы, олар арқылы өзгермелі x_i мәндерінің шашырамдық сипаты мен сызбадағы үлестірілудің қисық сызықтық пішінінің орналасу қалпын анықтауға болады. Ал тәжірибе жүзінде алынған үлестірілу сипатының теориялық «нормаль заңдылығына» үйлесу дәрежесін анықтау (сәйкестігін бағалау) үшін «ассиметрия» және «эксцесс» параметрлерін білу қажет деп саналады:

$$A = \frac{\mu_3}{\sigma^3}. \qquad (7)$$

$$E = \frac{\mu_4}{\sigma^4} - 3. \qquad (8)$$

Мұндағы σ – орта квадраттық ауытқу, μ_3 – мен μ_4 – үшінші және төртінші реттегі орталық моменттер:

$$\mu_3 = \frac{1}{n} \cdot \sum_i n_i \cdot (x_i - a)^3 \quad , \qquad (9)$$

$$\mu_4 = \frac{1}{n} \cdot \sum_i n_i \cdot (x_i - a)^4 \quad . \qquad (10)$$

Нормаль үлестірілу заңдылығы үшін μ_3 мен μ_4 мәндері нөлге тең болады. Сондықтан сынаққа түсіп отырған үлестірілу заңдылығы үшін «ассиметрия» *(A)* мен «эксцесс» *(E)* мәндері де нөлге жуық келетін шама болса, онда тәжірибелік үлестірілу *нормаль* заңына жақын деп жорамалдауға әбден болады. Керісінше, *A* мен *E* шамаларының нөлден алшақтығы өскен сайын, сынаққа түскен тәжірибелік үлестірілудің де *нормаль* заңдылығына үйлесімділігі алшақтай түседі.

Теориялық үлестірілу заңдылығының *Шарлье А, В* мен *логарифмдік нормаль* түрлері жайлы толығырақ мына әдебиеттерден танысуға болады: [80, 94].

3.5.3. Пирсонның χ^2 (хи квадрат) үйлесімдік критерийі

Жоғарыда сөз болған *«Пуассон»* мен *«нормаль»* және *«Шарлье А, В»* үлестірілу заңдылықтары *теориялық заңдылық* немесе *ықтималдық модель (үлгі)* деп аталады. Зерттеуге алынған тілдік бірліктің тәжірибе жүзінде анықталған тәжірибелік (эмпирикалық) үлестірілулердің жоғарыда сөз болған кейбір теориялық заңдылықтарға үйлесімділігі тілдік бірліктің мәтін бойында заңды таралуының айғағы деуге болады. Сондықтан осындай үйлесімдіктің (жақындықтың) дәрежесін «бағалайтын» мүмкіндіктің қажеттігі туады. Мұндай мүмкіндікке статистикалық лингвистика саласында бұрыннан да қолданылып жүрген Пирсонның χ^2 (хи квадрат) үйлесімдік критериясы (критерий согласия) жатады [15, 16, 93]. Ол үшін таңдама бөліктің эмпирикалық үлестірілуі бас жиынның теориялық үлестірілуіне сәйкес келеді деген «нөлдік» гипотеза (ғылыми болжам) – H_0 ұсынылуы қажет. Ол болжамды (гипотезаны) «қабылдау» немесе «терістеу» үшін тәжірибелік

(эмпирикалық) пен теориялық үлестірулер арасындағы алшақтықты анықтайтын мынадай статистикалық сипаттама құрастырылады:

$$W=\sum_{i=i}^{k}\frac{(n_i^{ЭМ}-n_i^{T})^2}{n_i^{T}} . \qquad (11)$$

Мұндағы $n_i^{ЭМ}$ және n_i^{T} – мәтіндегі лингвистикалық бірліктің тәжірибелік және теориялық жиіліктері. Егер мәтін көлемі *(N)* шексіздікке ұмтылса, яғни $N\to\infty$, онда статистикалық сипаттама *W*-ның χ^2 (хи квадрат) үлестірілу заңдылығын қанағаттандыратынына көз жеткізуге болады. Ол үлестірілудің еркіндік дәреже саны $\nu=t-\ell-1$. Мұндағы ℓ – сызықтық қатынаста тұратын параметрлер саны, t – тәжірибелік үлестірілу тобының саны.

Сонымен, тәжірибелік (эмпирикалық) және теориялық үлестірілулердің сәйкестігін эмпирикалық жиіліктердің ($n_i^{ЭМ}$) математикалық күтуінен $(n_i^{T}=NP_i)$ алшақтық дәрежесі бойынша бағалауға болады (p_i – ықтималдық, N – таңдама бөліктің көлемі).

Келесі математикалық өрнек χ^2 (хи квадрат) критерийі деп аталады:

$$\chi^2=\sum_{i=i}^{k}\frac{(n_i^{ЭМ}-n_i^{T})^2}{NP_i} . \qquad (12)$$

Егер есептеліп шығарылатын $\chi^2=0$ болса, онда эпирикалық (тәжірибелік) және теориялық жиіліктер біріне-бірі сәйкес келеді деп үйғарылады. Ал басқа жағдайларда, яғни χ^2 мәні нөлден айырықша болған жағдайда, аталған жиіліктердің сойкестік сипатының әлсіреуі χ^2 мәнінің нөлден алшақтығына тура пропорционал деп есептеледі. Дәлірек айтқанда, χ^2 мәнінің нөлден алшақтығы өскен сайын, тәжірибелік және теориялық жиіліктер сәйкестігі соншалықты әлсірей түседі.

Осымен бірге, алшақтық сипатының мәнді (существенное) және мәнсіз (несущественное) түрлері де ажыратылады. Мұндай жайт χ^2 мәні мен теориялық үлестірілудің басқа да параметрлері толық анықталғаннан кейін ғана белгілі болады. Басқаша айтқанда, тәжірибелік және теориялық жиіліктер арасындағы сәйкестік болжамын (H_0) χ^2 критерийі бағалай алады. Ол үшін «еркіндік дәреже санын» (ν) анықтау керек және осы критерий үшін пайыздық «мағыналық деңгей» (уровень значимости) – q таңдалып алыну қажет. Мысалы, ықтималдықтың 0,95 мәніне мағыналық деңгейдің $q=0,05$ немесе 5 пайыздық деңгейі сәйкес келеді. Бұл шама алдын ала қабылданған шекаралық аядан сырт шығуы кездейсоқтық жағдайға ғана тән екендігін және оның ықтималдығы $q=0,05$ тең немесе мұндай жағдайдың 5 пайызы ғана шындықтан тыс жатқанын білдіреді. Сонымен бірге зерттеуші арнайы кесте арқылы χ^2_q мәнін анықтай білуі керек. Мұндағы χ^2, еркіндік саны ν мен мағыналық деңгейге (q) тәуелді, яғни аргументтері ν және q болатын функция: $\chi^2_q=\chi^2(\nu, q)$; χ^2-тің есептеліп шыққан мәні мен оның кестелік мәнін (χ^2_q) салыстыру қажеттігі туындайды. Бұндай салыстыру кезінде екі түрлі теңсіздіктердің кездесуі мүмкін: $\chi^2 \geq \chi^2_q$ немесе $\chi^2 < \chi^2_q$. Бірінші жағдайда, яғни $\chi^2 \geq \chi^2_q$ болса, критерий мәні алшақтық дәрежесінің мәнділік аясына қатысты болады да тәжірибелік пен теориялық үлестірілулердің сәйкестігі жайлы H_0 болжам теріске шығарылып, қабылданбауы керек. Ал, екіншіде, яғни $\chi^2 < \chi^2_q$ болса, аталған үлестірілулердің арасындағы алшақтық – кездейсоқ (мәнсіз) жағдайға жатқызылып, H_0 болжам қабылдануы қажет.

H_0 болжамды «бағалау» осындай салыстырулардан басқа да жолмен іске асатынын айта кетейік. Ол әдіс $p(\chi^2 \geq \chi^2_q)$ ықтималдық шамасын білуді қажет етеді. Егер осындай

ықтималдық алдын ала берілген мағыналық деңгейден, мысалы, q 0,05 шамадан төмен жатса, онда ол кездейсоқ оқиға ретіндегі алшақтықтардың ықтималдығы өте аз екендігін білдіреді. Бұл жағдайда тәжірибелік пен теориялық үлестірілулердің арасындағы алшақтық кездейсоқ емес (мәнді), сондықтан сәйкестікті қолдайтын H_0 болжам қабылданбайды, ол – «жалған» деген сөз.

Керісінше, егер $p(\chi^2 \geq \chi_q^2)$ ықтималдық шамасы $q = 0,05$ санынан айтарлықтай үлкен болса, онда аталған екі үлестірілулер арасы жақын жатыр, сондықтан H_0 гипотезасы шындыққа сай келеді және ол қабылданады.

Сонымен, $p(\chi^2 \geq \chi_q^2)$ ықтималдық мәні H_0 болжамының (гипотезаның) қабылдану-қабылданбауына бағалық критерийі бола алады.

Математикалық статистика саласының белгілі маманы Т.Крамер бірнеше тәжірибелер нәтижелерін талдай келе, мынадай қорытындыға келеді: «Тәжірибелік және теориялық үлестірілулердің сәйкестігі «жақсы» деп бағаланады, егер $p(\chi^2 \geq \chi_q^2) = 0,17$ болса, ал осы ықтималдықтың $0,91$-ге тең мәнінде сәйкестік «өте жақсы» деп бағалануы керек. Екі үлестірілулер арасындағы алшақтық «мәнді дәрежеде» қалуы үшін $p(\chi^2 \geq \chi_q^2) = 0,03$ болуы жеткілікті» [81], – дейді.

Енді, осындай зерттеуді, яғни эмпирикалық үлестірілудің қалап алынған теориялық заңдылықпен сәйкестігінің мәнді не мәнсіз екендігін анықтау үшін жүргізілетін тәжірибе реті төмендегідей [31]:

1. Зерттеуге тиісті мәтінді және тілдік бірлікті таңдау. Мысалы, қазақ көркем әдебиет мәтіндеріндегі заг есім сөз табына қатысты сөздер.

2. $B_N^n = k$ көрінісіндегі таңдалып алынған параметрлеріне (N, n, k) сай етіп, зерттелетін мәтінді таңдама бөліктерге бөлу.

3. N, n, k параметріне және тілдік бірлікке сәйкес жиілік сөздіктер түзу.

4. Жиілік сөздіктер негізінде дискретті не үздіксіз вариациялық қатар құру.

5. Тиісті математикалық орнекті (формуланы) пайдаланып, жоғарыда сөз болған мына параметрлерді есептеп шығару:

а) математикалық күтілуді *(a)*;

ә) ауытқудың квадраттық ортасын (σ);

б) ассиметрия *(A)* мен эксцесті *(E)*.

в) тілдік бірліктің теориялық жиілігін $(n_i^T = NP_i)$;

г) әрбір «*i*» тобына қатысты «хи квадрат» (χ^2) мәнін.

6. Жиіліктің *5-тен* кіші мәндерін қажеттілікке сай біріктіру және $\sum_{i=1}^{k} \chi_i^2$ қосындысына қажетті өзгертулер енгізу (қолмен).

7. Еркіндік дәреже санын анықтау. Мысалы, үлестірілімнің Пуассон заңы үшін $\ell = 1$, нормаль мен логнормаль, Шарлье В заңдары үшін $\ell = 2$, Шарлье А үшін $\ell = 4$.

8. Есептеліп шығарылған χ^2 пен еркіндік дәреже саны *(V)* бойынша $p(\chi^2 \geq \chi_q^2)$ – ықтималдық интегралдың кестелік мәнін табу [81].

Тілдік бірліктің мәтін ішіндегі үлестірілу сипатын анықтау сериялар саны *(k)* мен таңдама бөліктер көлемін *(n)* дұрыс нормалауға байланысты болады. Себебі, көлемі *N*-ге тең бір мәтінді саны *k*-дан тұратын сериялarға әр түрлі жолмен бөлуге болады. Ол ішкі таңдама мәтін бөліктерінің көлеміне *(n)* байланысты. Мысалы, мәтін көлемі *N=5000* сөзқолданыстан тұрса, ал ішкі бөліктер көлемі *n=100* болса, серия саны *k=50; әл*ı сол сияқты *n=500* болса, *k=5* болады. Әр уақытта $N = k \cdot n$ және $k = N:n$, яғни $N = 5000 = 50 \cdot 100$ және $k = 5000:100 = 50$; $N = 5000 = 10 \cdot 500$; $N = 5000 = 5 \cdot 1000$ және т.б.

Бұл жерде ескерте кететін жайт, лингвистикалық зерттеулерде сериялар саны мәтіндегі сөзқолданыстар санынан басқа, кітаптың жол не бет санымен де өлшенуі мүмкін. Сол сияқты, параметрлерді нормалау кезінде мәтін бірліктерінің қайсысы және қандай көлемде алынатыны алдын ала нақтылануы қажет. Әрбір нормалау сипатына сай және өзіне ғана тән теориялық үлестірілу заңдылығы сәйкес келуі мүмкін.

Мәселен, нақты бір нормалауда сынаққа алынған тілдік бірлік «Пуассон үлестірілу» заңына, ал басқаша түрдегі нормалауда сол бірлік мәтін ішінде «Нормаль үлестірілу» заңына «бағынуы» немесе белгілі бір ішкі бөлік көлемінен бастап тек бір ғана теориялық заңдылыққа сәйкес келуі мүмкін екені анықталды [15, 63, 88]. Сондықтан үлестірілу заңдылықтарының параметрлерін салыстыруда және оған лингвистикалық тұрғыда түсініктеме беруде зерттеу нысанын (мәтінді) нормалау сипатына аса көңіл бөлінуі қажет.

Сонымен, мәтінді ішкі бөліктеудің ең тиімді жолын табу лингвистикалық модельдеу ісінде маңызды деп саналады.

Енді қазақ мәтіндеріне қатысты жүргізілген кейбір тәжірибе нәтижелеріне тоқталайық [31].

Сөз таптарының қолдану жиілігін сөз еткенде, барлық тілдерге де ортақ заңдылық – зат есім мен етістік сөздердің өнімді қолданылуы. Мысалы, М.Әуезовтің «Абай жолы» романы мәтінінде зат есім 35,36%, ал етістік сөздер 31,44%, яғни 100 пайыздың 70-ке жуығы тек осы екі сөз табына қатысты сөздер екен. Газет мәтінінің 45% – зат есім, 22 пайызы – етістік (Ахабаев). Сондықтан қазақ мәтіндеріндегі сөз үлестірілу заңдылығын зерттеу кезінде әр сөз табына байланысты алынатын таңдама мәтіндер осындай қарапайым және төуелсіз талаптарға жауап берулері керек.

Егер аталған шарттар орындалды деп үйғарылса, зат есім, етістік сөздердің (не басқа сөз таптарының) тәжірибелік (эмпирикалық) үлестірілудің теориялық (нормальды не басқа) заңдылыққа үйлесімді келеді (согласуется) деп *«H_0»* жорамалды жасауға болады. Ал жорамалды «бағалау» әр түрлі *N, n, k* шамаларына қатысты $p(\chi^2 \geq \chi_q^2)$ ықтималдықтарды салыстыру негізінде іске асады.

Біздің тәжірибемізде «Абай жолы» романынан алынған *N=50000* сөзқолданыс мынадай нормалауға сәйкес келді:

а) $N_1=2500$, $N_2=5000$, $N_3=10000$, $N_4=20000$, $N_5=25000$, $N_6=40000$ және $N_7=50000$ көлемдері ретінде бөлек-бөлек қарастырылды;

а) ішкі бөлік көлемі (n): $n_1=100$, $n_2=200$, $n_3=250$, $n_4=500$, $n_5=1000$ сөзқолданыстан тұратын көлемдері ретінде бөлек-бөлек қарастырылды;

б) барлық жағдайда $N=k\times n$ болатынын ескеріп, сериялар шамасы $k=N:n$ анықталды: $k_1=25$, $k_2=50$, $k_3=100$, $k_4=200$, $k_5=500$ сериялық сан мөлшері екені анықталды.

Осылайша нормалаудың нәтижесінде көркем әдебиет мәтіні бойынша зат есім сөздердің тәжірибелік жиіліктері оның теориялық жиіліктерімен сәйкес келудің «сенімділік дәрежесін» қанағаттандыру жағын алғанда 1-ші орында «Нормаль үлестірілу» заңы, ал 2-ші орында «Шарлье А үлестірілу» заң түрі екені анықталды. Газет мәтіні бойынша, сериялар саны 100-ден артқан жағдайда зат есім сөздердің теориялық заңдылықтарға сәйкес келу мүмкіндігі «логнормаль үлестірілу» заңдылығы екені дәлелденді (*3.9-3.12-кестелерді* қараңыз).

Көркем әдебиет және публицистика мәтіндеріндегі етістік сөздердің тәжірибелік үлестірілу деректерінің теориялық үлестірілу заңдарымен үйлесімділігін зерттей келе, олардың математикалық моделі (үлгісі) нормаль заңдылығы екені айқындалды. Ал егер, серия саны $k\leq 100$ болған жағдайда, етістік сөздердің үлестірілулері теориялық заңдылықтар ішінен «Шарлье А үлестірілу» типіне сәйкестігі айрықша байқалды.

Публицистикалық мәтіндердегі етістікке қатысты сөздердің тәжірибелік үлестірілуі параметрлерді нормалаудың мынандай түрінде: $\{N\geq 1000,\ n\geq 100,\ k\geq 100\}$, теориялық заңдылықтар ішінен «нормаль үлестірілуін» көбірек қанағаттандыратыны *3.13* және *3.14-кестелердегі* ықтималдықтың дәрежелерінен байқауға болады.

Сонымен, қазақ тіліндегі етістікке қатысты сөздердің тәжірибелік жиіліктерінің мәтін бойындағы үлестірілуі (аталған нормалауға сай) «Нормаль заңымен» үйлесімдік табады деп қорытындылауға болады.

Теориялық үлестірілу заңдарының Пуассон және Шарлье В типі қазақ тіліндегі *зат есім* мен *етістік* сөздердің тәжірибелік жиілік үлестірілу деректерімен үйлеспейтіндігі де байқалды.

Зат есім мен *етістік* сөздердің қатарында сын есім сөздердің де мәтіндегі үлестірілу заңдылығы қарастырылды. Қысқаша айтқанда, олардың теориялық заңдылықтарымен

сәйкестігі көркем әдебиет пен газет мәтіндерінде бірдей емес және нормалау түріне қарай, яғни N, n, k параметрлерін болжектеуге қатысты айырым табады. Дәлірек айтқанда, көркем әдебиет мәтіні үшін нормалау сипаты – $\{N \geq 2500,\ n \geq 1000,\ k \geq 25\}$ болғанда, «Пуассон үлестірілу» мен «Шарлье үлестірілу» заңдылығының А мен В типтерімен бірдей дәрежеде үйлесетіні байқалды, ал газет тіліне қатысты $\{N \geq 5000,\ n \geq 100,\ k \geq 50\}$ нормалау әдісі ғана аталған заңдылықтарды қанағаттандырады екен.

Үйлесімдік дәрежесін оның «артықшылық» түрғысынан қарастырсақ, әрине, *сын есім* сөздерге қатысты Пуассонның теориялық үлестірілу заңы бірінші орынға ие. Солай бола тұра, нормалаудың $\{N \geq 2000,\ n \geq 100,\ k \geq 200\}$ мәндерінен бастап, қазақ тілінің *сын есім* сөздері «нормаль заңдылығымен» де жақсы үйлесімдік табатынын байқаймыз *(3.15, 3.16-кестелер)*. Ал *3.17-кестеде* «*үстеу*» сөздердің 5 түрлі теориялық заңдылықтармен үйлесімдігін анғартатын $p(\chi^2 \geq \chi_q^2)$ шамалары көрініс тапты.

Сөз таптарының тәжірибелік жиіліктерінің теориялық үлестірілу заңдылықтарымен үйлесімдігі, көбіне, сериялар санына қатысты екені байқалады. *Үстеу* сөздерге қатысты ең «қолайлы» деп Пуассон үлестірілу заңын атауға болады. Ал сериялар саны $k > 500$ болғанда үстеу сөздер «нормаль заңымен» үйлеседі.

Сонымен, жоғарыда баяндаған жайттар мәтінді мөлшерлеу (нормалау) әдістерінің қазақ тіліндегі сөз таптарының тәжірибелік (эмпирикалық) үлестірілулерінің белгілі бір теориялық үлестірілу заңдылығына сәйкес келу не келмеу мүмкіндіктерін айқындауды көздейді. Бұл зерттеудің нәтижелері қазақ тіліндегі құбылыстарды модельдеу мәселелерінде аса маңызды рөл атқаратыны сөзсіз.

3.9-кесте

Көркем әдебиет мәтініндегі зат есім сөздер үлестірілуінің $p(\chi^2 \geq \chi_q^2)$ мәндері

Таңдама нормасы $B(N/n)=k$	Үлестірілу заңдары		
	Нормаль	Шарлье А	Логнормаль
$B(2500/100)=25$	0,660	0,746	0,175
$B(5000/200)=25$	0,734	0,548	0,006
$B(10000/400)=25$	0,862	0,713	0,0003
$B(5000/100)=50$	0,991	0,972	0,368
$B(10000/200)=50$	0,625	0,380	0,246
$B(10000/100)=100$	0,101	0,193	0,053

3.10-кесте

Публицистика мәтініндегі *зат есім* сөздер үлестірілуінің $p(\chi^2 \geq \chi_q^2)$ мәндері (k – тұрақты)

Таңдама нормасы $B(N/n)=k$	Үлестірілу заңдары		
	Нормаль	Шарлье А	Логнормаль
$B(2500/100)=25$	0,797	0,994	0,999
$B(5000/200)=25$	0,986	0,935	0,231
$B(12500/500)=25$	0,740	0,486	0,005
$B(25000/1000)=25$	0,979	0,880	0,000
$B(5000/100)=50$	0,687	0,781	0,619
$B(10000/200)=50$	0,302	0,118	0,064
$B(25000/500)=50$	0,775	0,398	0,692
$B(50000/1000)=50$	0,548	0,141	0,0003
$B(10000/100)=100$	0,626	0,563	0,839
$B(20000/200)=100$	0,983	0,976	0,750
$B(50000/500)=100$	0,991	0,990	0,413
$B(20000/100)=200$	0,310	0,456	0,477
$B(40000/200)=200$	0,575	0,658	0,677
$B(50000/250)=200$	0,695	0,699	0,756
$B(50000/100)=500$	0,802	0,649	0,809

3.11-кесте

Публицистика мәтініндегі *зат есім* сөздер үлестірілуінің $p(\chi^2 \geq \chi_q^2)$ мәндері (*n* – түрақты)

Таңдама нормасы	Үлестірілу заңдары		
B(N/n)=k	Нормаль	Шарлье А	Логнормаль
B(2500/100)=25	*0,797*	*0,994*	*0,999*
B(5000/100)=50	*0,687*	*0,781*	*0,619*
B(10000/100)=100	*0,626*	*0,563*	*0,839*
B(20000/100)=200	*0,310.*	*0,456*	*0,477*
B(50000/100)=500	*0,802*	*0,649*	*0,809*
B(5000/200)=25	*0,986*	*0,936*	*0,231*
B(10000/200)=50	*0,302*	*0,118*	*0,064*
B(20000/200)=100	*0,983*	*0,976*	*0,750*
B(40000/200)=200	*0,575*	*0,658*	*0,677*
B(12500/500)=25	*0,740*	*0,486*	*0,005*
B(25000/500)=50	*0,775*	*0,398*	*0,692*
B(50000/500)=100	*0,991*	*0,990*	*0,413*
B(25000/1000)=25	*0,978*	*0,880*	*0,000*
B(50000/1000)=50	*0,548*	*0,141*	*0,0003*
B(50000/250)=200	*0,695*	*0,699*	*0,756*

3.12-кесте

Публицистика мәтініндегі *зат есім* сөздер үлестірілуінің $p(\chi^2 \geq \chi_q^2)$ мәндері (*N* – түрақты)

Таңдама нормасы	Үлестірілу заңдары		
B(N/n)=k	Нормаль	Шарлье А	Логнормаль
B(5000/200)=25	*0,986*	*0,936*	*0,231*
B(5000/100)=50	*0,687*	*0,781*	*0,619*
B(25000/1000)=25	*0,979*	*0,880*	*0,000*
B(25000/500)=50	*0,775*	*0,398*	*0,692*
B(20000/200)=100	*0,983*	*0,976*	*0,750*
B(20000/100)=200	*0,310*	*0,456*	*0,477*
B(50000/1000)=50	*0,548*	*0,141*	*0,0003*
B(50000/500)=100	*0,991*	*0,990*	*0,413*
B(50000/250)=200	*0,695*	*0,699*	*0,756*
B(50000/100)=500	*0,802*	*0,649*	*0,809*
B(40000/200)=200	*0,575*	*0,658*	*0,677*
B(2500/100)=25	*0,797*	*0,994*	*0,999*
B(12500/500)=25	*0,740*	*0,486*	*0,005*
B(10000/200)=50	*0,302*	*0,118*	*0,064*
B(10000/100)=100	*0,626*	*0,563*	*0,839*

3.13-кесте

Көркем әдебиет мәтініндегі *етістік* сөздер үлестірілуінің $p(\chi^2 \geq \chi_q^2)$ мәндері

Таңдама нормасы	Үлестірілу заңдары		
B(N/n)= k	Нормаль	Шарлье А	Логнормаль
B(2500/100)= 25	*0,603*	*0,544*	*0,395*
B(5000/200)= 25	*0,530*	*0,788*	*0,539*
B(10000/400)= 25	*0,156*	*0,807*	*0,393*
B(10000/200)= 50	*0,319*	*0,547*	*0,679*
B(10000/100)= 100	*0,320*	*0,350*	*0,495*

3.14-кесте

Публицистика мәтініндегі *етістік* сөздер үлестірілуінің $p(\chi^2 \geq \chi_q^2)$ мәндері

Таңдама нормасы	Үлестірілу заңдары		
B(N/n)= k	Нормаль	Шарлье А	Логнормаль
B(2500/100)= 25	*0,151*	*0,080*	*0,224*
B(5000/200)= 25	*0,978*	*0,978*	*0,009*
B(25000/1000)= 25	*0,045*	*0,074*	*0,000*
B(5000/100)= 50	*0,469*	*0,213*	*0,005*
B(10000/200)= 50	*0,187*	*0,609*	*0,614*
B(25000/500)= 50	*0,174*	*0,369*	*0,0003*
B(10000/100)= 100	*0,926*	*0,729*	*0,140*
B(20000/200)= 100	*0,187*	*0,152*	*0,072*
B(50000/500)= 100	*0,238*	*0,048*	*0,100*
B(40000/200)= 200	*0,871*	*0,587*	*0,285*
B(50000/250)= 200	*0,855*	*0,636*	*0,067*
B(50000/100)= 500	*0,302*	*0,109*	*0,253*

3.15-кесте

Көркем әдебиет мәтініндегі *сын есім* сөздер үлестірілуінің $p(\chi^2 \geq \chi_q^2)$ мәндері

Таңдама	Үлестірілу заңдары				
нормасы	Норм.	Шар.-А	Логнор.	Пуас.	Шар.-В
B(2500/100)= 25	*0,215*	*0,019*	*0,078*	*0,024*	*0,044*

B(5000/200)= 25	0,000	0,013	0,000	0,014	0,005
B(10000/400)= 25	0,000	0,040	0,000	0,133	0,056
B(5000/100)= 50	0,384	0,623	0,141	0,848	0,749
B(10000/200)= 50	0,075	0,401	0,024	0,529	0,477
B(10000/100)= 100	0,93	0,925	0,661	0,674	0,548

3.16-кесте

Публицистика мәтініндегі *сын есім* сөздер үлестірілуі $[p(\chi^2 \geq \chi_q^2)]$

Таңдама нормасы	Үлестірілу заңдары				
	Норм.	Шар.-А	Логнор.	Пуас.	Шар.-В
B(2500/100)= 25	0,287	0,216	0,253	0,331	0,121
B(5000/200)= 25	0,966	0,999	0,000	0,851	0,989
B(12500/500)= 25	0,055	0,339	0,000	0,410	0,317
B(25000/1000)= 25	0,013	0,469	0,000	0,114	0,433
B(50000/2000)= 25	0,000	0,269	0,000	0,401	0,819
B(5000/100)= 50	0,227	0,277	0,012	0,188	0,079
B(10000/200)= 50	0,333	0,644	0,148	0,690	0,734
B(12500/250)= 50	0,602	0,779	0,130	0,481	0,420
B(25000/500)= 50	0,002	0,652	0,038	0,328	0,211
B(50000/1000)= 50	0,005	0,572	0,000	0,282	0,175
B(10000/100)= 100	0,017	0,054	0,172	0,104	0,159
B(20000/200)= 100	0,047	0,0002	0,162	0,065	0,123
B(25000/250)= 100	0,181	0,172	0,137	0,050	0,071
B(20000/100)= 200	0,965	0,874	0,760	0,450	0,407
B(40000/200)= 200	0,211	0,040	0,099	0,201	0,280
B(50000/250)= 200	0,253	0,083	0,021	0,130	0,070
B(50000/100)= 500	0,101	0,132	0,221	0,308	0,340

3.17-кесте

Публицистика мәтініндегі *үстеу* сөздер үлестірілуі $[p(\chi^2 \geq \chi_q^2)]$

Таңдама нормасы	Үлестірілу заңдары				
	Норм.	Шар.-А	Логнор.	Пуас.	Шар.-В
B(2500/100)= 25	0,000	0,000	0,00001	0,105	0,337
B(5000/200)= 25	0,000	0,00001	0,00003	0,535	0,699
B(12500/500)= 25	0,000	0,000	0,000	0,475	0,798
B(25000/1000)= 25	0,000	0,000	0,000	0,999	0,996
B(50000/2000)= 25	0,000	0,000	0,000	0,995	0,984
B(5000/100)= 50	0,010	0,0004	0,002	0,897	0,867

B(10000/200) 50	0, 050	0.199	0,0004	0,907	0,855
B(12500/250) 50	0,070	0,067	0,004	0,9992	0,997
B(25000/500) 50	0,001	0,045	0,005	0,863	0,902
B(50000/1000)=50	0,042	0,046	0,042	0,600	0,963
B(10000/100)=100	0,863	0,866	0,042	0,401	0,645
B(20000/200)=100	0,388	0,128	0,043	0,364	0,532
B(50000/500)=100	0,174	0,780	0,147	0,576	0,910
B(20000/100)=200	0,186	0,069	0,015	0,450	0,448
B(40000/200)=200	0,386	0,742	0,277	0,433	0,537
B(50000/200)=250	0,475	0,667	0,223	0,486	0,542
B(25000/50)=500	0,504	0,490	0,208	0,396	0,567

3.6. Колмогоровтың үйлесімдік критерийі арқылы лингвистикалық болжамды сынау (тексеру)

Мәтінді сипаттаудың аналитикалық түрпаты мен соған жуықтайтын моделін (үлгісін) табудың қолданбалы және теориялық маңыздылығы зор. Тіл білімінде алғаш рет орыс тілі мәтініне қатысты сөзтұлғаның жиілік сөздіктегі орнын анықтаудың ықтималдық үлестірілуін сипаттауға арналған Г.Г.Белоноговтың зерттеу жұмысын атауға болады [20]. Өз зерттеуінде ғалым Вейбуллдың формуласын пайдаланады [96]:

$$F(x)=1-e^{-cx^{k}}. \qquad (1)$$

Мұндағы x – сөзтұлғаның сөздіктегі рет саны, «*c* мен *k*» Вейбулл үлестірілуінің параметрлері, «*e*» – натурал логарифмнің негізі. Осы (1) өрнекпен анықталатын $F(x)$ функцияның мәні – мәтін ішінде кездесетін сөзтұлғаның сол мәтіннен түзілген жиілік сөздіктегі кез келген жиі кездесетін саны «*x*»-ке тең сөзтұлғалармен сәйкес келу ықтималдығын білдіреді. Ғалымдар, көп жағдайда, зерттеу нысаны ретінде орыс, ағылшын, неміс тілдерінің мәтіндерін қарастырып, аталған өрнекті сөзтұлғаның немесе сөз тіркестің «теориялық қатынастық жиіліктер жиынтығын» табу үшін қолданады. Ал біздің ізденісімізде ол өрнек түңғыш рет қазақ тіліндегі көркем әдебиет мәтіндеріндегі сөзтұлғаларды сипаттауда пайдаланылды.

Енді аталған (1) өрнекті басқаша түрпатта жазайық. Ол үшін $F(x)$ үлестірілу функциясының орнына теориялық

қатынастық жиіліктердің жиынтығын (қосындысын) f_i^{*Teop} аламыз.

Сонда ол $f_i^{*Teop}=1-e^{-cx^k}$ өрнегі арқылы жазылады.

Мұндағы $x=i$ – жиілік сөздіктегі сөзтұлғаның рет саны, $f_i^{*Teop} < 1$, ал $e^{-ci^k} \geq 0$ болғандықтан соңғы теңдіктің екі жағын логарифмдеуге мүмкіндік туады:

$$-ci^k \ell ne = \ell n(1-f_i^{*Teop})$$

немесе

$$-ci^k = -\ell n(1-f_i^{*Teop}). \qquad (2)$$

Енді (2) теңдіктегі $[-\ell n(1-f_i^{*Teop})]=y$ және $i=x$ деп белгілесек, аталған өрнек мынадай пішінді қабылдайды: $Cx^K=Y$, мұндағы $Y \geq 0$, сондықтан теңдіктің екі жағынан бірдей логарифмдей аламыз: $lnC+klnx=lny$.

Егер $lnC=A$, $lnx=X$, $lny=Y$ деп белгілесек:

$$A+kX=Y. \qquad (3)$$

Әрі қарайғы әрекетімізде «ең кіші квадраттар әдісі» (метод наименьших квадратов) бойынша A мен k мәндерін анықтап, C-ны табуға болады: [20, 47]:

$$C=l^A. \qquad (4)$$

Әрбір рет саны «i»-ге сай келетін k мен C-ны анықтап, оған сәйкес келетін f_i^{*Teop} табылады.

Сонымен, әрбір i-ге қатысты тәжірибелік $f_i^{*эмп}$ және теориялық $f_i^{*Teop.}$ қатынастық жиіліктердің жиынтықтары арқылы «орташа квадраттық ауытқуды» мынандай өрнекпен есептеп шығаруға болады:

$$\sigma=\sqrt{\frac{1}{q}\sum_i (f_i^{*эмп.}-f_i^{*Teop.})^2}, \qquad (5)$$

мұндағы q – тәжірибе саны.

Жоғарыда баяндалған жайттардың негізінде әрбір рет саны i-ге қатысты C, k, σ мәндерін арнайы компьютерлік бағдарлама арқылы есептеп шығаруға болады [24]. Осындай бағдарламаның негізінде М.Әуезовтің «Абай жолы» романының 5 түрлі жиілік сөздіктерінің соңғы рет санына $i_{соң}$ қатысты аталған параметрлердің мәндері *3.18-кестеде* көрсетілді.

Бұл кестедегі «орта квадраттық ауытқу» (σ) мәндері кездейсоқ шаманың «математикалық күтілу» *(а)* (математическое ожидание) аясында таралуын (рассеивание) сипаттайды. Кестедегі деректерге көңіл болсек, олардың таралу сипатының аздығы байқалады.

3.18-кесте

Параметрлер	Сөздіктер				
	1-ші кітап	2-ші кітап	3-ші кітап	4-ші кітап	5-ші роман
$i_{соң}$	*3276*	*3872*	*3570*	*4027*	*9614*
C	*0,0452*	*0,0429*	*0,0425*	*0,0374*	*0,0496*
K	*0,4196*	*0,4174*	*0,4150*	*0,4318*	*0,3876*
σ	*0,0064*	*0,0057*	*0,0061*	*0,0060*	*0,0050*

Бағдарлама көмегімен әр сөздікке қатысты есептелген $f_i^{*Теор.}$ мен тәжірибелік $f_i^{*эмп}$ мәндерін жай ғана (көзбен) салыстырғанның өзінде көп ұқсастықты байқауға болады. Бірақ бұлай салыстыру нәтижелері әр уақытта дұрыс бола бермеуі мүмкін. Сондықтан оған математикалық баға берілуі тиіс. Осындай қарапайым құрал ретінде біз «Колмогоровтың үйлесімдік критерийін» пайдалануды жөн санадық [24, 80, 81]. Аталған критерий статолингвистика саласындағы теориялық үлестірілуді таңдау жағдаятында жиі қолданылып жүр.

Біздің зерттеу барысындағы алға қойылған мақсат – «Абай жолы» романы мәтінінің сөзтұлға жиіліктерінің үлестірілуі H_0 жорамалы бойынша, кездейсоқ оқиғаның (x) алдын ала берілген $F(x)$ функциясының үлестірілуімен үйлесетіндігін тексеру (бағалау).

А.Н.Колмогоров тәсілі бойынша, теориялық пен тәжірибелік үлестірілулер арасындағы үйлесімдік (не үйлеспеушілік) өлшемі ретінде үлестірілудің эмпирикалық $f^{*эмп}(x)$ пен оған сәйкес келетін теориялық функция $F(x)$ айырым модулінің максимумдық шамасы алынады, яғни:

$$D = max\ /F(x)^{*эмп} - F(x)/\ . \qquad (6)$$

А.Н.Колмогоровтың тұжырымдауынша, егер тәуелсіз тәжірибелер саны «n» шексіздікке қарай өсетін болса, $P(D\sqrt{n} \geq \lambda)$ ықтималдығы мына шекке ұмтылады:

$$P(\lambda) = 1 - \sum_{k=-\infty}^{\infty} (-1)^k \cdot \ell^{-2k^2\lambda^2}\ . \qquad (7)$$

Ықтималдық теориясы мен математикалық статистика оқулықтарының «қосымшаларында» (7) өрнекпен есептелген әрбір 0,1 қадам сайын $0 \leq \lambda \leq 2$ мәніне сәйкес, $P(\lambda)$ ықтималдығының шамасы көрініс тапқан кестелер берілген [24, 80, 81].

Егер тәжірибелік үлестірілу функцияның мәндері (жиілік сөздіктегі қатынастық жиіліктер жиынтығы) және теориялық үлестірілу функция мәндері (Вейбулл өрнегімен анықталған қатынастық жиілік жиынтығы) белгілі болса, онда Колмогоров критерийін мынадай реттегі әдіспен қолданғанды жөн санаймыз:

1. Қатынастық жиынтық жиіліктерге сәйкес келетін эмпирикалық және теориялық жиіліктер мәндерінің айырым модульдерінің максимумдық шамасын анықтау:

$$D = max[\ f^{*эмп.}_{(x)} - f^{*Теор.}_{(x)}\ ].$$

2. Жиілік сөздіктің ең соңғы рет санына (n) қатысты теориялық жиілігі есептелген болса, $\lambda = D(n)$ шамасын анықтау.

3. Белгілі «λ» мәні бойынша, арнайы кесте арқылы $P(\lambda)$ ықтималдық шамасын табу.

4. Егер анықталған $P(\lambda)$ ықтималдық шамасы барынша аз болса, «H_0» болжам қабылданбайды, ал, керісінше, $P(\lambda)$ шамасы айтарлықтай үлкен болса, тәжірибелік пен теориялық үлестірілулер заңдарының үйлесімділігін, яғни H_0 болжамды қабылдауға болады.

Мысал ретінде М. Әуезовтің «Абай жолы» романы кітаптарының (4 кітап, 5 сөздік) мәтінінің сөзтұлға сөздіктерін қарастырайық. Болжам H_0 - роман мәтіндерінен түзілген жиілік сөздіктері негізінде анықталған сөзтұлға жиіліктерінің үлестірілуі Вейбулл функциясы мен берілген тәжірибелік үлестірілу аралықтары үйлеседі. Болжамды тексеру мақсатымен жоғарыда келтірілген әдіс ретімен есептеулер жүргізілу қажет. Арнайы компьютерлік бағдарлама негізінде 5 жиілік сөздіктер деректері бойынша есептеу нәтижелері *3.19-кестеде* орын алды.

3.19-кесте

Параметрлер	Сөздіктер				
	1	2	2	4	5
D	*0,0351*	*0,0336*	*0,0324*	*0,0309*	*0,042*
λ	*2,001*	*2,083*	*1,944*	*1,944*	*4,116*
$P(\lambda)$	*0,001*	*0,001*	*0,002*	*0,002*	*0,001*

Кесте деректерінен байқайтынымыз, роман бойынша түзілген 5 түрлі сөздіктерде де $P(\lambda)$ ықтималдығы нолге жуық сандар. Сондықтан, тәжірибелік жиіліктер үлестірілуі Вейбулл функциясының үлестірілуімен үйлеспейді, яғни H_0 болжам қабылданбайды. Бұлай тұжырымдаудың негізгі себебі – жиілік сөздік бойының алғашқы зонасы есепке алынды. Сынаққа алынған барлық сөздіктер бойынша айырым модулінің максимум шамасы сөздіктің бастапқы зоналарына сәйкес келеді. Ал егер есептеу процесіне жиілік сөздіктердің алғашқы 50-55 реттік сандарын қоспаған жағдайда $P(\lambda)$ ықтималдығы бірден жоғары көтерілетінін байқаймыз *(3.20-кесте).*

3.20-кесте

Параметрлер	Сөздіктер				
	1	2	3	4	5
D	*0,0107*	*0,0103*	*0,0106*	*0,0106*	*0,020*
λ	*0,612*	*0,641*	*0,634*	*0,673*	*1,962*
$P(\lambda)$	*0,864*	*0,864*	*0,864*	*0,711*	*0,502*

$P(\lambda)$ ықтималдығының аз шама қабылдауына сөздіктің соңғы зоналарының әсері бар екендігін ескере отырып, мына дай қорытынды жасауға болады:

Тәжірибелік қатынастық жиынтық жиіліктер үлестірілуінің функциясы $(f_i^{*эмп})$ Вейбуллдың теориялық қатынастық жиынтық жиілігімен $f_i^{*Теор.}$ тек рет саны 50-ден көп және тәжірибелік абсолютті жиіліктің шамасы 5-тен жоғары болған жағдайларда ғана «үйлеседі» (согласуется) екен.

Төртінші тарау

ТІЛДІҢ ЛЕКСИКА-МОРФОЛОГИЯЛЫҚ ҚҰРЫЛЫМЫНА СТАТИСТИКАЛЫҚ ӘДІСТІ ҚОЛДАНУДЫҢ АЛҒЫШАРТТАРЫ

4.1. Тілдік бірліктерді кодтау принципі

Қазіргі қазақ тіліндегі барлық сөздер үлкен үш топқа бөлінеді: атаушы сөздер, көмекші сөздер, одағай сөздер [98, 126-б.]. Бұл топтар, әрі қарай лексика-грамматикалық жақтарына сараланып, атаушы сөздер әуелі есімдер және етістіктер деп бөлінеді де, сосын есімдердің өздері іштей атаушы есімдер және үстеуші есімдер болып тағы да жіктеледі. Атаушы есімдер іштей зат есім, сын есім, сан есім, есімдік деген сөз таптарына сараланады да, ал үстеуші есімдер – үстеу сөздер мен еліктеу сөздерге бөлінетіні белгілі. Сөйтіп, жинақтап айтқанда, қазіргі қазақ тіліндегі сөздер 9 топқа бөлініп, олар «сөз табы» деп аталады: зат есім, сын есім, сан есім, есімдік, етістік, үстеу сөз, еліктеу сөз, көмекші (шылау) сөз, одағай.

Сөз табының мазмұны «лексика-грамматикалық» деп аталатын екі компоненттің бірлігінен құралады. Бұл атаудан лексикалық семантикамен бірге грамматикалық семантика деген ұғымды да бірге қабылдау керек. Ал қос сөздің екінші бөлігіндегі «грамматикалық» сөзінің мазмұнына белгілі бір сөзге тән грамматикалық категориялардың және олардың жасалу, түрлену тұлғаларының мағыналары енеді. Яғни бұл жердегі «грамматикалық» ұғымына сөздің жаңадан сөз тудыру, сөз түрлендіру, сөз байланыстыру жүйелерінің мағыналары мен

тұлғалары, демек, бүкіл сөздің морфологиялық және синтаксистік белгілері (сыр-сипаттары) түгел енеді [98, 126-134-бб.].

Қазақ тілінің қолданбалы саласының, дәлірек айтқанда, статистикалық лингвистиканың алдына қойған мақсаттарының бірі – сөз табының лексикалық (семантикалық), грамматикалық (морфологиялық және синтаксистік) белгілері арқылы, яғни осы үш белгінің негізінде статистикалық зерттеулер жүргізіп, сөздің мағынасын, морфологиялық формасын (тұлғасын) және синтаксистік қызметін сандық деректер арқылы тани білу. Соның нәтижесінде мәтін мазмұнын ашудың формальды жақтарын айқындау.

Сөз таптарының ішкі жіктелген бөліктерінің түрлері тілде біреуі аз не көп қолдануы мүмкін. Статистикалық зерттеу тәжірибесі бойынша олардың қолдану жиілігін анықтау үшін кейбір шараларды алдын ала белгілейтін бағдарламаларды іске қосу қажет.

Мысалы, әр түрлі тілдік стильдердегі сөз таптарының қатынастық сипатының статистикасын білу керек болса, әр сөз табына қатысты шартты белгіні (кодты) қоюдың зерттеушіге ыңғайлы жолдарын ойластыруымыз керек. Себебі, сөз таптарын не басқа тілдік бірліктерді мәтін ішінде түр-түрпатына қарай формальды әдіспен ажырату мүмкіндігі әлі де болса зерттеу аясынан тыс қалып жүр. Сондықтан статистикалық санақтарды компьютер арқылы жүргізуде аталған тілдік бірліктерді және олардың ішкі бөліктерін шартты белгілерге сәйкестендіру (кодтау) танымдық рөл үшін аса қажетті. «Кодтау» принципі алдын ала ойластырылған әмбебаптық сипатта болғаны жөн.

Сондықтан қазақ тілінің жоғарыда аталған сөз таптарын ажырату мен олардың статистикасын анықтау мақсатындағы зерттеулерде біздің төменде қарастыратын «кодтау» үлгісін пайдалануға болады:

1) цифрлар арқылы: зат есім – 01, сын есім – 02, сан есім – 03, есімдік – 04, стістік – 05, есімше – 06, көсемше – 07, үстеу – 08, шылау – 09, еліктеуіш сөз – 10, модаль сөздер – 11;

2) әріптер арқылы: зат есім – зт, сын есім – сн, сан есім – са, есімдік – ес, етістік – ет, есімше – еш, көсемше – кш, үстеу – үс, еліктеуіш сөздер – ел, модаль сөздер – мд;

3) цифрлар мен әріптерді араластыра белгілеу арқылы (аралас әдіс): зат есім – з1, сын есім – с1, сан есім – с2, етістік – е1, есімдік – е2, есімше – е3, еліктеуіш сөздер – е4, көсемше – к1, үстеу – ү1, модаль сөздер – м1.

Шартты белгілерді (кодтарды) қоюдың мүмкіндігі мол, дегенмен оларды әркім өз қалауынша белгілегеннен гөрі бірізділікке (стандартқа) ұмтылған дұрыс деп саналады. Әр сөз табының іштей морфологиялық, синтаксистік және семантикалық тармақтану түрлері мен жасалу жолдары да әр зерттеушінің алға қойған мақсатына сай әр түрлі сипатта болуы мүмкін. Сондықтан әрбір сөзге, сөзтұлғаға, сөзқолданысқа, сөзтіркеске немесе мәтіннің одан да үлкен бірліктеріне қойылатын «белгі-код» әр түрлі грамматикалық ақпаратты өзінде сақтай алады. Шартты белгі-кодтарды қабылдау тек бір ғана тілдік бірліктің бірнеше ақпараттарын сәйкестендіретін «ұялар» тізбегі ретінде ұйымдастыруға болады.

Әрине, көбінде, мәтін ішіне мұндай кешенді белгі-код енгізу статистикалық ізденістің бағыт-бағдарына байланысты атқарылады. Қазақ тілін статистикалық әдіспен зерттеу тәжірибесінде тіліміздегі сөз таптарының қолдану сипатын ажырату үшін (жиілігін) мәтінге шартты белгілер енгізудің түрлі мүмкіндіктері жоғарыда көрсетілді.

Енді осындай белгілеу әдісін күрделендіріп, әр сөз табының морфологиялық типтері мен құрылымы жағын ажырату, жасалу жолдарын ескеру, яғни сол тілдік бірлікке қатысты басқа да грамматикалық ақпараттарды барынша қамту және олардың әрбіреуінің статистикасын анықтау мақсатымен. Әрі қарайғы баяндауымызда «кодтаудың» кейбір үлгілеріне толығырақ тоқталамыз.

Орыс тілінің морфологиялық сипатына тән сөз таптарын шартты түрде белгілеу (кодтау) ресейлік ғалым Б.Н.Головиннің «Язык и статистика» атты кітабында да сөз болады [30].

Ал морфологиялық ақпарат белгілерін сәйкестендіру бағдарламасын құрудағы біз ұсынып отырған үлгінің ерекшелігі – қазақ тілі сөз таптарының ішкі бөліктерінің құрылымдық сипаты мен жасалу жолдары ескеріліп, әмбебаптық сипатты көздейді.

Бағдарлама жасау ісінің нысаны ретінде – зат есім, етістік, сын есім, сан есім, есімдік сияқты қазақ тілінің негізгі сөз таптары алынды. Аталған сөз таптарына морфологиялық ақпаратты белгі-код үлгісін сәйкестендіру бағдарламасы тиісті кестелер арқылы көрініс тапты.

4.2. Зат есімнің лексика-морфологиялық құрылымына ақпараттық белгі-кодты сәйкестендіру бағдарламасы

Зат есім сөздерге ғана тән морфологиялық ерекшеліктер жоқ емес. Олар өздерінің лексика-семантикалық сипаттарына қарай арнайы жалғаулар арқылы түрленіп және жұрнақтар арқылы түлғалық өзгеріске ұшырап, сөйлемдегі өзге сөздермен еркін қарым-қатынасқа түсіп отыратыны белгілі. Зат есім сөздердің морфологиялық, синтаксистік және семантикалық белгілерін ұштастыра қарағанда ғана олардың сыр-сипатын толық ашуға болады. Сол белгілер қатарына олардың мәтін ішінде жиі не сирек қолдану сипаты да жатады.

Қазіргі қазақ тілінің грамматикасында зат есім сөздерге ғана тән семантикалық және грамматикалық ерекшеліктері бар мынадай топтар бөліп қарастырылады: адамзат (кімдік) және ғаламзат (нелік) есімдер, жалқы есімдер, көптік мәнді есімдер, эмоциялы-экспрессивтік зат есімдер, көмекші есімдер. Зат есім сөздерді олардың морфологиялық сипаты, яғни құрылымдық, жасалу тәсілі және түрлену жүйесі арқылы да топтап қарастыруға болатыны да белгілі. Міне, осылайша зат есім сөздерді топ-топқа бөліп қарастыру қазақ тілін компьютерлік және статистикалық лингвистика әдістерімен зерттеуде аса маңызды деуге болады.

Кітаптың бірінші тарауында айтылғандай қазақ тіліндегі мәтіндерде ең жиі қолданыс табатын сөз таптары *зат есім* мен *етістік*. Олай болатын болса, олардың мәтін ішінде морфологиялық, синтаксистік жолмен түрленіп қолдануының да статистикасын танып білу «Қолданбалы лингвистика» пәнінің мақсаттарының бірі. Біздің баяндауымыздағы негізгі ұстаным – мәтін ішінде кездесетін *сөз*, *сөзтұлға* не *тұрақты сөзтіркеске*, яғни мәтін бірлігі деп аталатын – *сөзқолданысқа* сәйкес келетін грамматикалық ақпараттар *орны («ұясы»)* мен сол ақпаратты

танытатын шартты *белгі-кодты* тұрақтандыру. Қандай да болмасын тілдік бірліктің, мысалы, ең алдымен қарастырайын деп отырған – *зат есім* сөздердің орбіреуіне сәйкестендірілетін шартты параметрлерді («ұя» орны мен белгі-код) алдын ала жан-жақты ойластырып алғанды жөн санаймыз.

Айталық, *зат есім* сөздерге сәйкестендіретін (тиісті) «ұя» саны немесе топтама саны – сегізге тең делік, яғни n=1, 2, 3, 4, 5, 6, 7, 8. Ал әрбір «ұяда» жазылатын шартты *белгі-код* саны (топтама саны) мен «код мәні» *зат есім* сөздердің сол «ұяда» қарастырылатын сипатына қарай өзгеріп отыруы мүмкін.

Егер осы айтылғандарды қамтитын мәліметтер кестесін *«бағдарлама»* деп атасақ, *зат есім* сөз табына қатысты қысқаша топтама *4.1-кестедегі «бағдарлама»* түрінде көрініс табады.

Топтамадағы мәліметтердің кейбіреулеріне түсініктеме берейік.

Аталған кесте үш «бағанадан» тұрады:

1) *шартты белгілер орны* (топтама саны);

2) *зат есім сөздердің лексика-морфологиялық сипаты*;

3) *шартты белгі-код.*

Әрбір сөз табына қатысты «ұялар» топтамасының ең алғашқысы сол сөз табының қысқарған атымен белгіленеді (кодталады) және *1-ші орынға* ие болады деп ұйғару керек. Сонымен, «ұялар» тізбегінің ең алғашқысы (бірінші бағана) «1» санына тең (бірінші топтама), осы жатық жолдағы екінші бағана сөз табының сипаты жайлы мәлімет, яғни – *зат есім* атауы және үшінші бағана – мәліметтің «белгі-код» мәні – «ЗТ». Сол сияқты шартты белгілер орнының «2» санында, яғни екінші топтамасында *зат есім* сөздердің екі түрлі сипаттамасы берілген. Олардың біріншісі – *адамзат есімдері (кім?)* және екіншісі – *ғаламзат есімдері (не?)*. Шартты белгі орнында, яғни 2-ші орында қарастырылатын кезекті *зат есім* сөздің қай сұраққа *(кім? не?)* жауап беруіне қарай үшінші бағанадағы оған сәйкес келетін белгі-кодтың біреуі ғана таңдалып алынады. *4.1-кестеде* зат есім сөздерге қатысты осы тәріздес 8 топтама қарастырылды.

Сол топтамалардың енді 7-ші орынға қатыстысын сөз етсек, жетінші топтамада *зат есім* сөздің алты түрлі сипаты аталған: *1. Түбір морфема зат есім; 2. Туынды зат есім*

(түбір+жұрнақ); 3. Күрделі зат есім: а) біріккен сөз; ә) қосарланған сөз; б) құрама сөз; в) қысқарған сөз.

Үшінші бағанада олардың белгі-код мәндері үш орынды сандар: 070, 071, 072, 073, 074, 075 сандары арқылы берілген (*4.1-кестені* қараңыз).

Жоғарыда 2-ші орынға байланысты айтылған сияқты, бұл жолы да зерттеуге алынған кезекті сөзді (сөзтұлғаны) шартты түрде «кодтау» кезінде 7-ші орынға сәйкес келетін (070 пен 075 аралығы) алты белгі-код мәндерінің ішінен сөзтұлға сипатына тән келетін біреуі ғана алынады. Кестеде келтірілген басқа топтама орындарына да осылайша түсініктеме беруге болады.

Енді мысал ретінде кейбір зат есім сөздерді *4.1-кесте* деректері бойынша шартты белгі-код түрінде жазып көрейік. Мысалы, мына сөйлем ішіндегі зат есім сөздерді бөліп алып, белгі-код арқылы жазу керек: *Асан мектепке барарда кітап, қалам, дәптер және басқа да керек-жарақтарын аса ұқыптылықпен сөмкесіне салды.* Бұл сөйлемдегі зат есім сөзтұлғалар мыналар: *Асан; мектепке; кітап; қалам; дәптер; керек-жарақтарын; сөмкесіне.*

Шартты белгі-кодтарды осы сөздерге сәйкестендірейік. Яғни әрбір аталған *зат есім* сөздердің сегіз орындық «ұяларына» тиісті *белгі-код* мәнін қойып шығу керек деген сөз. Олардың барлығының да бірінші орындарында «ЗТ/» белгі-код тұратыны айғақ.

Енді *Асан* сөзін бөлек алып қарастырсақ, 2-ші орында – «020» (адам аты болғандықтан), үшіншіде – «030», төртіншіде – «040», бесіншіде – «050», алтыншыда – «060», жетіншіде – «070», сегізіншіде – «080». Сонымен, сөйлемдегі *Асан* сөзі мен басқа да зат есім сөздердің «кодтық» жазылуы төмендегідей болады:

Асан:	ЗТ/	020	030	040	050	060	070	080
мектепке:	ЗТ/	021	031	040	050	060	070	083
кітап:	ЗТ/	021	031	040	050	060	070	080
қалам:	ЗТ/	021	031	040	050	060	070	080
дәптер:	ЗТ/	021	031	040	050	060	070	080
керек-жарақтарын:	ЗТ/	021	031	042	050	060	073	084
сөмкесіне:	ЗТ/	021	031	040	050	060	070	083

4.1-кесте

Зат есім сөздерге лексика-морфологиялық ақпаратты сәйкестендіру бағдарламасы

Белгі-код орны	Зат есім сөздердің лексика-морфологиялық сипаты	Белгі-код
1	Зат есім сөз (сөзтұлға)	ЗТ/
2	Адамзат есімдері *(кім?)*	020
	Ғаламзат есімдері *(не?)*	021
3	Жалқы есім: а) кісі аттары (ономастика); ә) географиялық атаулар (топонимика)	030
	Жалпы есім	031
	Аралас мағыналы *(ай, күн, жер)*	032
4	Көптік мәнді емес (жалғаусыз қарастырғанда)	040
	Табиғи қос мәнді *(аяқ, құлақ, етік)*	041
	Көптік мәнді (жалғаусыз)	042
5	Реңсіз зат есім	050
	Эмоциялы-экспрессивтік реңді зат есім *(Әкей, Сәулеш)*	051
6	Көмекші зат есімге жатпайтын зат есім	060
	Көмекші зат есім *(алды, арты, қасы)*	061
7	Түбір морфема зат есім	070
	Туынды зат есім (түбір+жұрнақ)	071
	Күрделі зат есім: а) біріккен сөз	072
	ә) қосарланған сөз	073
	б) құрама сөз	074
	в) қысқарған сөз	075
8	Зат есімнің түрленуі:	
	Зат есімнің жалғаусыз түрі *(кім? не?)*	080
	Көптік жалғаулармен: *-лар, -лер,-дар,-дер,-тар,-тер*	081
	Тәуелдік жалғаулар арқылы: *-нікі,-дікі,-тікі -м,-ым, -ім; -ң, -ың, -ің, -ңыз, -ңіз, -ыңыз, - іңіз; -сы, -сі, -ы,-і*	082
	Септік жалғаулар арқылы: ілік *(кімнің? нені?)*; барыс *(кімге? неге? қайда?)*; табыс *(кімді? нені?)*; жатыс *(кімде? неде?)*; шығыс *(кімнен? неден?)*; көмектес *(кіммен? немен?)*.	083
	Жіктік жалғау арқылы: жекеше: *-мын, -мін, -сың, -сің, -сыз, -сіз;* көпше: *-мыз, -міз.*	084

Көрсетілген жолмен мына зат есімдерге қатысты сөзтұлғаларды белгі-кодпен сәйкестендірейік: *ғалым, Асандар, құлағым, Алматыға, жазушы, сақырау құлақтан, жаным-ай, жүрісіме.*

ғалым	ЗТ/	020	032	040	050	060	070	080
Асандар	ЗТ/	020	030	040	050	060	070	081
құлағым	ЗТ/	021	032	041	050	060	070	082
Алматыға	ЗТ/	021	030	040	050	060	070	083
жазушы	ЗТ/	020	032	040	050	060	071	080
сақырау құлақтан	ЗТ/	021	032	041	050	060	074	083
жаным-ай	ЗТ/	021	032	040	051	060	070	082
жүрісіме	ЗТ/	021	032	042	050	060	071	083

Белгілі бір мәтін сөзтұлғаларын белгі-код түріне ауыстырып барып, жиілік сөздіктер түзуге болады немесе әріп күйіндегі мәтіннің өзінен түзілген әліпби-жиілік, жиілік не кері-әліпби сөздіктердегі сөзтұлғаларды (бірліктерді) белгі-код түріне ауыстырып жазуға да болады. Егер ондай сөздіктерді тиісті компьютерлік бағдарлама бойынша «қысылған» түрге келтіріп жазатын болсақ, зат есім (не басқа сөз табы) сөзтұлғалардың «сипаттық жиілік сөздігін» алуға болады. Бұндай жиілік сөздік қазақ тілінің морфологиялық құрылымын зерттеу пысаны ретінде және әрі қарайғы мәтін лингвистикасына қажетті маңызды материал болатыны сөзсіз.

4.3. Етістіктің лексика-морфологиялық құрылымына ақпараттық белгі-кодты сәйкестендіру бағдарламасы

Етістік сөздер құрамының басқа сөз таптарына қарағанда күрделілігі оның аса өрісті лексика-семантикалық сипатымен, өте бай лексика-грамматикалық формаларымен және кең синтаксистік қызметімен тығыз байланысты. Етістіктің лексика-грамматикалық тұлғаларының бай болу себебі – олар әрекеттің болу мезгілін, жүзеге асу кезеңін, өту сипатын, яғни әрекеттің бағыты, қарқыны, тынуы тәрізді жайларды түгел қамтиды. Осымен қатар басқа сөз табына қатысты сөздерден етістік тудыратын синтетикалық, аналитикалық тәсілдер жүйелерін

қоса есептесек, етістіктің тұлғалану байлығына жететін сөз табы жоқ деуге болады. Осы айтылғандар оның өзіне ғана тән лексика-семантикалық, лексика-грамматикалық және грамматикалық категорияларынан айқын көрініп жатады. Етістіктің осындай әр қилы категорияларының мағыналары да, синтаксистік қызметтері де өзге сөз таптарына тән сөздермен қарымқатынасқа түскенде айқындала түседі.

Етістіктің лексикалық жүйесіндегі әрбір сөздің өзіне ғана тән лексикалық мағыналары бөлек болғанымен, семантикалық ерекшеліктеріне қарай бір сөз тобына телінеді. Етістіктің грамматикалық тұлғалары мен қызметтерін дұрыс айқындау үшін, оларды түбір формалардың, туынды синтетикалық және аналитикалық формалардың семантикалық құрылымына орай жіктеп, топтап барып сыр-сипатын ашу, олардың тұлғалық және мазмұндық арақатынастарын айқындауға әкеліп соғады. Себебі грамматикалық семантика сөздің құрамына көп байланысты болып келеді.

Сол себептен етістік формаларының морфологиялық құрылымының сыр-сипатын ашу мәнмәтіндік тұрғыда немесе мәтіндік лингвистика саласының зерттеу аясына жақындайды. Ал бұл тұрғыдан зерттеуді ұйымдастыру дегеніміз статистикалық лингвистика әдістерін молынан қолдану деген сөз. Туынды етістіктердің семантикалық құрылымының аса күрделі және түрлену сипатының бай болу себептерінің бірі – олардың сирек не жиі қолдануында деп түсіну қажет [70]. Сондықтан туынды етістіктердің морфологиялық құрылымын жеке бір жүйе ретінде қарастырып, әрбір өзіне тән сипатына қарай топтап, олардың ықтималды-статистикалық заңдылықтарын ашу үшін арнайы белгі-код жүйесін енгізіп зерттеу қажет деп санаймыз. Себебі, жоғарыда, зат есім сөздерге қатысты айтылғандай, қазақ тілінің тілдік бірліктерін, сөз таптарын формальды белгілері арқылы танып білу және соның нәтижесінде автоматты түрде жазба не сөйлеу мәтінінен оларды бөліп алу мәселесі әз шешімін таппай отыр. «Қолдан» енгізген шартты белгі-код негізінде етістіктің морфологиялық құрылымына статистикалық зерттеулер жүргізу көптеген жайттардың басын ашатыны белгілі. Солардың ішінде мәтіндік бірліктерді автоматты түрде айыра білу мен бөліп алу мәселесі де бар.

Сонымен бірге етістіктің лексика-грамматикалық құрылымының дәстүрлі әдіспен зерттеуде байқала бермейтін тұстарының да басы ашылуы (анықталуы) мүмкін. Қазіргі қазақ тіліндегі етістіктерді (түбір және туынды) осы тұрғыда қарастыра отырып, біз төменде олардың іштей бір-біріне мағына жағынан жақындықтарына, өзара функция жағынан орайластықтарына және лексика-морфологиялық құрылымына қарай топ-топқа бөліп, етістік сөздердің сипаты жайлы ақпараттарды алдын ала ойластырылған шартты белгі-кодтарға сәйкестендіру бағдарламасын ұсынамыз.

Мысалы, мағыналық тұрғыдан алып, етістік сөздерді іштей мына топтарға бөлуге болады [98, 223-226-бб.]: **амал-әрекет, қимыл-қозғалыс, қалып-сапа, ойлау-сөйлеу, осу-ону етістіктері** және олар *4.2-кестеде* көрсетілгендей жалғасын табады.

Әдетте, туынды сөздердің бір жүйе бойынша қалыптасқан белгілі морфологиялық құрылымы болатыны белгілі. Сол құрылымның бірінші компоненті дербес мағыналы сөз болады да, ал екінші компоненті – бірінші компонентті белгілі бір сөз табына айналдыратын жұрнақ болады *(тіс+те кел+тір, сабын+да)*. Туынды етістіктер, жасалатын негіздеріне қарай, есімдерден және етістіктерден жасалған етістіктер деген екі салаға бөлінеді. Олар есім негізді етістіктер, етістік негізді етістіктер деп аталып жүр. Біз осылардың тек алғашқысын ғана, яғни есім негізді етістіктердің жасалу жолдарын жекелеп 16 топқа бөліп қарастырып, оларды шартты белгі-код арқылы таңбалау үлгісін *4.2-кесте* арқылы ұсындық.

Зат есім сөздерге лексика-морфологиялық ақпаратты сәйкестендіру бағдарламасы төріздес, *4.2-кестедегі* бағдарлама да үш бағанадан тұрады:

1) белгі-код орны; 2) есім негізді туынды етістік сөздердің сипаты; 3) белгі-код таңбасы.

Бірінші бағананың белгі-код орны «1» (немесе №1-ші топтама), ал екінші бағана бойынша оның сипаттамасы – «етістік сөз» деп аталады. Егер мәтін бойынан зерттелетін кезекті сөз – *етістікке* жататын болса, онда үшінші бағанадан сол етістік сөздің *белгі-коды* – «ЕТ» болады. Екінші орынға немесе екінші топтаманың екінші бағанасында етістіктердің өздерін іштей мағыналық және өзара функционалдық жағынан

жақындастыратын 11 топтың әрбіреуіне сипаттама берілген. Ал үшінші бағанада осы сипаттамаларға сәйкестендірілген белгі-код таңбасын айыра білуге болады. Мысалы, белгі-кодтың «2-ші» орнының (2-ші топтамасының) үшінші бөлігі *қалып-сипат етістіктері* деп аталады, ал оған сәйкес келетін үшінші бағанада сол сипаттаманың үш орынды белгі-коды – «102» тұрады. Сол сияқты, туынды етістіктердің жұрнақтар арқылы есім сөздерден жасалу жолынан да мысал келтірейік. Туынды етістіктердің ***-ла (-ле, -да, -де, -та, -те)*** жұрнағы арқылы жасалу жолы ішкі 5 бөлікпен топталған атаулар сипатымен берілген:

1) дене мүшелері (*аяқта, жұдырықта, өкшеле*);

2) еңбек құралдары (*арала, балтала, арқанда);*

3) іс-әрекетті жүзеге асыруға объект болатын зат атаулары (*жүнде, майла, тұзда, алтында);*

4) мекен я орын, өлшеу, дыбыс т.т. (*мекенде, өрле; аршында, метрле; мыңқылда, шыңқылда);*

5) сын есім, сан есім, үстеу, еліктеуіш т.т. сөздерге жалғанған жағдайлар (*ақта, қарала, екеуле, ала, үнле).*

Осы 5 түрлі топтарға 5 үш орынды сандар (белгі-код) сәйкестендірілген: 113, 114, 115, 116, 117. Осы тәріздес, басқа топтама орындарында да өзіне тән сипаттамалар және соған сәйкес шартты белгі-кодтар *4.2-кестеде* көрініс тапты.

Енді осындай белгі-кодтар жиынын іс жүзінде қалай пайдаланамыз деген сұрақ туындауы мүмкін. Бұл жайында жоғарыда зат есім сөздерге қатысты да қысқаша айтылған болатын. Етістік сөздерге қатысты *4.2-кестені* пайдалануда да айтарлықтай айырмашылық жоқ деуге болады. Осы тұста айта кетерлік жайт, топтама орындар санын (белгі-код орнын) тағайындау (белгілеу), біріншіден, зерттеу мақсатына сай, екіншіден, әр сөз табының лексика-морфологиялық, семантикалық ерекшеліктеріне байланысты деуге болады. Мысалы, зат есім сөздерге қатысты *4.1-кестеде* сегіз орын жеткілікті деп алсақ, етістік сөздерге қатысты *4.2-кестеде* ондай орындар саны көптеу-ақ. Ал басқа сөз табына қатысты немесе зерттеу мақсатына сай, топтау орын саны басқаша болуы да мүмкін. Сол сияқты, *4.1* және *4.2-кестелердегі* деректерден байқайтынымыз әр топтаманың ішкі бөлік саны да әр түрлі болады екен. Енді осы аталған кестелерді пайдалану жайына қайта оралайық.

Алғашында, зерттеуге алынған мәтін бөлігіндегі сөзқолданыстарды сөз табына ажырататын шартты белгілер қойылғаннан кейін, олардан компьютер арқылы қажетті жиілік сөздіктер алынуы керек. Аталған жиілік сөздіктегі етістік сөзтұлғаларды (не басқа сөз табына қатысты сөздерді) бөліп алып, оларға *4.2-кесте* бойынша белгі-кодтарды сәйкестендіру қажет болады. Ескерту ретінде айта кететін жайт, зерттеуді жиілік сөздіксіз-ақ бірден мәтінмен де жүргізуге болады, бірақ бұл тәсілде қайталанатын мәтін бірліктері көп уақыт алады. Қалай болғанда да зерттеу жұмысы тізбек құрайтын сөзқолданыс жүйесінен етістік сөзтұлғаларды (немесе басқа сөз табы) кезек-кезегімен бөліп алып қарастыруды қажет етеді.

4.2-кесте

Етістік сөздерге лексика-морфологиялық ақпаратты сәйкестендіру бағдарламасы

Белгі-код орны	Етістік сөздердің лексика-морфологиялық сипаты	Белгі-код
1	2	3
1	**Етістік сөз**	ET/
2	**Магыналық және функционалдық жағынан топтастыру:**	
	1) Амал-әрекет етістіктері Мысалы: *босат, көтер, күрес, қи, сыз, өлше* т.т.	100
	2) Қимыл-қозғалыс етістіктері Мысалы: *ауна, аудар, домала, секір, бүкірей* т.т.	101
	3) Қалып-сана етістіктері. Мысалы: *жат, жантай, тұр, отыр, үлгай* т.т.	102
	4) Ойлау-сойлеу етістіктері. Мысалы: *айт, сөйле, де, ескер, жатта, ұмытпа* т.т.	103
	5) Өсу-өну етістіктері. Мысалы: *балала, жапырақта, гүлде, ос, қозыла* т.т.	104
	6) Бағыт-бағдар етістіктері. Мысалы: *бар, кел, қайт, әпер, әкет, қаш, түс, көтер* т.т.	105
	7) Көңіл-күй етістіктері. Мысалы: *жыла, қайғыр, өкін, күл, қуан, алақайла* т.т.	106
	8) Бейнелеу-еліктеу етістіктері. Мысалы: *жарқыра, күркіре, дүркіре, тарсылда* т.т.	107
	9) Дыбыс-сес етістіктері	108
	10) Кору-есту етістіктері.	109

1	2	3
2	11) Мінез-құлық етістіктері	110
	12) Етістіктің тізімге енбеген түрлері	111
3	**Түбір етістік** **Анықтама:** Етістік түбірі нешіе қилы формалар (есімше, көсемше, рай, етіс, шақ т.б.) тудыратын қосымшалардың барлығын алып тастағанда сақталатын бөлігі. Мысалы: *аз, ал, айт, алда, ат, ас* т.б.	112
	Есім негізді сөздерден жасалатын туынды етістіктер. Жұрнақтар арқылы жасалу жолдары: 1. ***–ла (-ле, -да, -де, -та, -те)*** жұрнағы арқылы жасалған туынды етістіктер: а) дене мүшелері атауларына жалғанады. Мысалы: *аяқта, жұдырықта, өкшеле.*	113
	ә) Еңбек құралдары атауларына жалғанады. Мысалы: *арала, балтала, арқанда.*	114
	б) Іс-әрекетті жүзеге асыруға объект болатын заттарға жалғанады. Мысалы: *жүнде, майла, тұзда, алтында.*	115
	в) Мекен я орын, өлшеу, дыбыс, қозғалыс, көрініс, топ және әлеуметтік, саяси, мәдени, тұрмыс, салт, табиғат т.т. атауларына жалғанады. Мысалы: *мекенде, өрле, төменде; аршында, метрле, тоннала; мыңқылда, шыңқылда; жарқылда, бұрқылда, ирелеңде; ботала, қозда; төрбиеле, еркеле, жазала, қаумала, үймеле* т.т.	116
	г) Сын есім, сан есім, үстеу, еліктеуіш, одағай сөздерге жалғанады. Мысалы: *ақта, қарала, жаманда; оңда, жүзде, екеуле; төменде, кейінде, ілгеріле; жымыңда, күлімде; аһла, үһле, ойбайла* т.т.	117
	2. ***–лан (-лен, -дан, -ден, -тан, -тен)*** жұрнағы арқылы жасалады. Мысалы: *ашулан, арлан, борышлан, целен …*	118
	3. ***–лас (лес, -дас, -дес, -тас, -тес)*** жұрнағы арқылы жасалған. Мысалы: *бірлес, көлектес, бәстес, достас, кезектес, сабақтас* т.б.	119
	4. ***–лат (-лет, -дат, -дет)*** жұрнағы арқылы жасалған. Мысалы: *дауылдат, тездет, тұнделет* т.б.	120
	5. ***–а (-е)*** жұрнағы арқылы. Мысалы: *аша, деме, жаса, өрте, тіле, сына, міне* т.б.	121
	6. ***–ай (-ей, -й)*** жұрнағы арқылы. Мысалы: *күшей, мұңай, қартай, кеңей, көбей, молай* т.б.	122

4.2-кестенің соңы

1	2	3
3	7. ***-қар (-гар, -кер, -гер)*** өнімсіз жұрнағы арқылы. Мысалы: *басқар, ескер, аңғар, теңгер, ақыр, қақыр, жазғыр, ысқыр, кекір* т.б.	123
	8. ***-ар (-ер, -р)*** көне жұрнағы арқылы. Мысалы: *жаңар, тазар, ескір, өзгер* т.б.	124
	9. ***-ал (-ел, -ал, -іл, -л)*** қосымшасы арқылы. Мысалы: *жоғал, оңал, тарыл, тірл, теңел* т.б.	125
	10. ***-ық (-ік)*** көне жұрнағы арқылы. Мысалы: *ашық, бірік, зарық, дәнік, өшік* т.б.	126
	11. ***-сы (-сі)*** және ***-ымсы (-імсі)*** көне жұрнақтары арқылы. Мысалы: *батырсы, босаңсы, көлгірсі, үлкенсі, апамсы* т.б.	127
	12. ***-сын (-сің)*** жұрнағы арқылы. Мысалы: *адамсын, білгішсін, көпсін, жамансын* т.б.	128
	13. ***-сыра (-сіре)*** жұрнағы арқылы. Мысалы: *айсыра, әлсіре, қансыра, жетімсіре* т.б.	129
	14. ***-ыра (-іре)*** жұрнағы арқылы. Мысалы: *барқыра, бұрқыра, дүркіре, күркіре* т.б.	130
3	15. ***-ырай, (-ірей)*** жұрнағы арқылы. Мысалы: *бақырай, шақырай, кішірей, шікірей,* т.б.	131
	16. ***-ы, -і; -шы, -ші; -ан, -ен, -ын, -ін, -н; -ырқа, -ірке, -ырқан, -іркен; -ына, -іне; -қа, -ке, -га, -ге*** т.б. көне жұрнақтар арқылы. Мысалы: *байы, жасы, желпі, кеңі; аунақшы, доңбекші; тасырқа, мүсірке, ашырқан, шіміркен; есіне, қатына, пысына; бүрке, ішке, қозға; басын, жирен, оян, үйрен* т.б.	132
	Етістік негізді сөздерден жасалатын туынды етістіктер	133
4	Күрделі етістіктер	140
	Жалаң етістік	141

Белгіленген бір орынға сол топтамадағы белгі-кодтар ішінен біреуі ғана таңдалып алынатынын үнемі есте ұстау қажет.

Мысал ретінде мына сөйлемде кездесетін етістік сөзтұлғаларды бөліп алып қарастырайық: *«Кешке әрі тоңып, әрі шаршап аһлап, үңілеп отырған кемпір өзін өлтіріп кете жаздаған кім екенін есіне алды»* (С.Көбеев).

Етістік сөзтұлғаларға жататындар: ***тоңып, шаршап, аһлап, үңілеп, отырған, өлтіріп кете жаздаған, алды***.

Енді *4.2-кестеде* көрсетілген деректер бойынша етістіктерге белгі-код мәндерін сәйкестендірейік:

тоқып	ЕТ/	106	133	141
шаршап	ЕТ/	106	133	141
ағлап	ЕТ/	106	133	141
үйлеп	ЕТ/	106	133	141
отырған	ЕТ/	102	133	141
өлтіріп	ЕТ/	100	133	141
кете жаздаған	ЕТ/	105	140	140
алды	ЕТ/	111	133	141

Сонымен, әр қилы етістік сөздер өздерінің жасалу және мағыналық ерекшеліктеріне қарай неше түрлі белгі-код мәндерін қабылдай алады.

4.4. Сын есімнің лексика-морфологиялық құрылымына ақпараттық белгі-кодты сәйкестендіру бағдарламасы

Қазіргі қазақ тілінің морфология саласында сын есім деп аталатын сөз табына мынадай анықтама берілген: *«Заттың сапасын, сипатын, қасиетін, көлемін, салмағын, түсін (түрлерін) және басқа да сыр-сипаттарын білдіретін лексика-грамматикалық сөз табы сын есім деп аталады»* [98, 166-172-бб.]. Морфологиялық сипаттары жағынан сын есімнің өзге есімдерге қарағанда өзіне ғана тән ерекшеліктері болатыны мәлім. Мысалы, зат есіммен, үстеумен және т.б. сөз таптарымен әрі ортақ, әрі тұлғалас сөз тудыратын жұрнақтарымен қатар тек туынды сын есім ғана жасайтын арнаулы жұрнақтары және сын есімнен сын есім тудыратын шырай жұрнақтары да бар. Бұл аталғандар да сын есімнің морфологиялық формалары болып есептеледі.

Сын есімдердің тағы да бір өзіндік қасиеті – заттардың алуан түрлі сыр-сипаттары мен белгілерін тікелей білдірумен қатар басқа заттардың қатынастары арқылы да сондай сипаттарды білдіре алады. Осыған байланысты, сын есімдер семантикалық мағыналары мен грамматикалық ерекшеліктеріне

қарай, *сапалық (негізгі)* және *қатыстық (туынды)* сын есім деп аталатын екі салаға бөлінетіні де белгілі.

Негізгі немесе сапалық сын есімдер жайлы айтатынымыз, олар, әдетте, ешбір қосымшасыз түрып-ақ, заттың әр килы сыр-сипатын білдіретін түбір сөздерден түрады. Бірақ, егер кейбір түбір сөз деп жүрген сөздерді олардың тарихи даму түрғысынан қарастырсақ, олар туынды сөз болып та шығуы мүмкін. Сондықтан, құрылымы жағынан негізгі сын есім деп есептелетін сөздердің саны да, сапасы да үдайы өзгеріп отырады. Ал сын есімдерді морфемалық құрылымына қарай, негізгі және туынды деп жіктеу шартты екендігін де есте ұстаған жөн.

Сонымен кез келген сөйлем ішінде (мәтінде) кездесетін сын есім сөздер, біріншіден, сапалық сын есімге жатуы мүмкін *(ақ, қара, сары, көк, сұр, биік, үлкен, аласа, жылы, суық, ірі, кіші* т.т.), екіншіден, морфологиялық (синтетикалық) тәсіл бойынша, яғни тиісті жұрнақтар арқылы жасалған туынды сын есімге жатуы мүмкін екен.

Үшіншіден, қарастырылатын кезекті сын есімге қатысты сөз синтаксистік (аналитикалық) тәсіл, яғни жалаң сын есімдердің бір-бірімен тіркесуі арқылы *(ақ сары, қызыл сары, ақ шұбар, ақ көйлекті, көп балалы, үлкен-кішілі* т.т.) және морфологиялық-синтаксистік (семантикалық) тәсіл бойынша жасалатын туынды сын есімдер де болуы мүмкін.

Соңғы аталған тәсіл бойынша туатын сын есімдер қолданылуы жағынан алғанда тіпті де өнімсіз деуге болады. Сондықтан сын есімдердің жасалу жолдарының ең негізгісі деп, ең алдымен морфологиялықты немесе синтетикалықты, ал одан кейін барып, аналитикалықты, яғни синтаксистік тәсілді айтуға болады.

Осы себептен, біз сын есімнің лексика-морфологиялық құрылымына ақпараттық белгі-кодты сәйкестендіру бағдарламасын құрастыруда тек қана сын есімнің синтетикалық тәсілмен жасалу жолына ғана, дәлірек айтқанда, есімдерден сын есім тудыратын өнімді жұрнақтар арқылы жасалу жолын негіз етпекпіз.

4.3-кесте

Сын есімнің лексика-морфологиялық құрылымына ақпараттық белгі-кодты сәйкестендіру бағдарламасы

Белгі -код орны	Сын есім сөздердің лексика-морфологиялық сипаты	Белгі-код
1	2	3
1	**Егер сөз (сөзтұлға) морфемалық құрамына қарай сапалық сын есім (негізгі сын есім) болса:**	СН1
2	1) заттың не құбылыстың түрі мен түсін анықтаса. Мыс.: *ақ, қара, қызыл, жасыл* т.б.	201
	2) заттың не құбылыстың сыры мен сапасын анықтаса. Мыс.: *жақсы, жаман, тәуір, нашар* т.б.	202
	3) заттың көлемі мен аумағын, ұзындығы мен салмағын анықтаса. Мыс.: *үлкен, кіші, ұзын, ауыр, қысқа, жеңіл* т.б.	203
	4) заттың дәмі мен иісін білдірсе. Мыс.: *ащы, тәтті, күлімсі* т.б.	204
	5) заттың не құбылыстың басқа да қасиет-белгілерін білдірсе.	205
1	**Егер кезекті сөз (сөзтұлға) морфемалық құрамына қарай қатыстық сын есімге жатса, яғни заттың белгісін басқа бір заттың қатысы арқылы білдірсе және ондай туынды сын есімдер өнімді (өнімсіз) жұрнақтар арқылы жасалса.**	СН2
2	Есімдерден туынды сын есім тудыратын өнімді жұрнақтар арқылы жасалып, заттың сыртқы түрі мен түсіне, кескіні мен келбетіне, сыры мен сынына, мекен мен мезгілге және т.б. белгілеріне қатысты сындық ұғым білдірсе. 1. ***-қы, -кі; -ғы, -гі*** жұрнақтары арқылы кейбір зат есімдерден, есімдіктерден, үстеулерден, сондай-ақ, жатыс және шығыс септіктегі сөздерден жасалатын туынды сын есімдер мынадай мағынада жұмсалса: а) егер мекендік ұғым білдіретін кейбір зат есімдерге, үстеулерге, сондай-ақ, жатыс (кейде шығыс) септік формаларындағы есімдерге жалғанса. Мыс.: *ауызғы, төргі, түпкі, ішкі, төменгі, соңғы, сыртқы, бүктеулі* т.б.	210

4.3-кестенің жалғасы

1	2	3
2	ә) егер мезгілдік ұғым білдіретін кейбір зат есімдер мен есімдіктерге, мезгілдік үстеулерге жалғанса. Мыс.: *кешкі, түскі, күзгі, жазғы, көктемгі, түнгі, күндізгі* т.б.	211
	2. ***-лы, -лі, -ды, -ді, -ты, -ті*** қосымшасы арқылы мынадай туынды сөздер жасалады:	
	а) белгілі бір заттың (я құбылыстың) бар екендігін білдіру үшін зат есімдерге жалғанса. Мыс.: *арлы, сұлы, әсерлі, гүлді, шкемді, байыпты, инабатты, балалы* т.б.	212
	ә) қосарланған зат есімнен, сын есімнен, сан есімнен, үстеуден күрделі сын есімдер осы қосымшалар арқылы жасалса. Мыс.: *ағалы-інілі, ойлы-қырлы, таулы-тасты, өзенді-сулы, үлкенді-кішілі, бұрынды-соңды* т.б.	213
	б) екі-үш сөзден құралып, суреттеме атаулар қызметіндегі күрделі сын есімдер жасалса. Мыс.: *ақ басты, қаз мойынды, ай қабақты, теке сақалды* т.б.	214
	3. ***-сыз (-сіз)*** жұрнағы есім сөздерге жалғанып, болымсыздық мағынадағы туынды сын есімдер жасалса. Мыс.: *баласыз, көліксіз, білімсіз, сенсіз, бізсіз, мұңсыз* т. б.	215
	4. ***-шыл (-шіл)*** қосымшасы зат есім, есімдік, модаль сөздерге (әр тараң) жалғанып, бейімділікті, икемділікті, құмарлықты білдіретін қатыстық сын есімдер жасалса. Мыс.: *ұйқышыл, ұйлышыл, өзімшіл, тұрашыл, ойшыл* т.б.	216
	5. ***-дай, (-дей, -тай, -тей)*** жұрнағы жалғанып, сөздерді салыстыру, үқсату мәнді туынды сын есімдер жасалса. Мыс.: *аттай, әкедей, мендей, сендей, жүздей, өлердей* т.б.	217
	6. ***-лық, -лік, -дық, -дік, -тық, -тік*** жұрнақтары арқылы мына мағыналағы туынды сын есімдер жасалса: а) зат есімдерге жалғанып, сөздердің нақтылылық мағына қасиеттерін білдіретін туынды сын есімдер жасалса. Мыс.: *орталық, қоғамдық, қалалық, азаматтық, жолдастық* т.б.	218

1	2	3
2	ә) мезгіл атаулары мен ор қилы бұйым атауларына жалғанып, мезгіл, өлшеу мөлшерімен байланысты туынды сын есімдер жасалса. Мыс.: *айлық, жылдық, апталық, тәуліктік, көйлектік, пальтолық, қайнатымдық, екі-үш асымдық* т.б.	219
	б) есімдіктерге жалғанып, олардан белгілі бір жаққа қатыстықты білдіретін туынды есімдер жасалса. Мыс.: *өзік, мендік, сендік, қандайлық, қаншалық* т.б.	220
	7. ***-лас, -лес, -дас, -дес, -тас, -тес*** жұрнақтары арқылы мына мағынадағы туынды сын есімдер жасалса:	221
	а) адам мінезіне байланысты зат есімге жалғанып, ондай сипаттың басқа адамдарға да қатысты екендігін білдірсе. Мыс.: *ниеттес, сырлас, тілеулес, пікірлес* т.б.	
	ә) адамзат, жан-жануарлар тегімен байланысты ұғымдағы зат есімдермен жалғанып, тұқым-туыстас екенін білдіретін сын есім жасалса. Мыс.: *туыстас, аталас, бауырлас, қарындас, тумалас* т.б.	222
	б) мекен ұғымдарына жалғанып, олармен қоныстастығын білдіретін мағынадағы туынды сын есімдер жасалса. Мыс.: *ауылдас, көршілес, жерлес, елдес, тумалас, ұялас, жансарлас, іргелес* т.б.	223
	в) кейбір зат есімдерге жалғанып, олардың нақтылы лексикалық мағыналарына қарай, түрлі қарым-қатынас, мерзім-мөлшер жағынан сыбайлас келетін сындық ұғымдарды білдірсе. Мыс.: *жолдас, замандас, сабақтас, кәсіптес, қызметтес, дәмдес, өкшелес* т.б.	224
	8. ***-шаң, -шең*** жұрнақтары арқылы жасалатын туынды сын есімдер: а) киім-кешек атауларына жалғанып, адамның бойындағы киімімен байланысты сыртқы бейне-көрінісін білдіретін сын есімдер жасалса. Мыс.: *көйлекшең, етікшең, пальтошаң, шалбаршаң* т.б.	225
	ә) кейбір зат есімдерге жалғанып, адамға не затқа тән белгілі бір ерекше қасиет білдіретін туынды сын есімдер жасалса. Мыс.: *ашушаң, созшең, бойшаң, тершең, кіршең* т.б.	226

4.3-кестенің жалғасы

1	2	3
2	9. Есімдерден туынды сын есім тудыратын өнімсіз жұрнақтар арқылы:	
	а) ***-дар, -дер, -тар, -тер*** қосымшалары кейбір зат есімдерге жалғанып, белгілі бір іс-әрекетіне дұшар болғандықты білдіретін сындық ұғым жасалса. Мыс : *қарыздар, хабардар, борышкер* т.б.	227
	ә) иран тілдерінен (парсы, тәжік, ауған) ауысқан кірме ***«-и»*** қосымшасы кейбір есімдерге жалғанып, туынды сын есім жасалса. Мыс.: *әскери, мәдени, тарихи* т.б.	228
	б) ***-паз*** жұрнағы жалғанып, сын есімге бейім жаңа сөз жасалса. Мыс.: *өнерпаз, аспаз, білімпаз, ойынпаз* т.б.	229
	в) ***-мпаз (-м+паз), -ымпаз (-ым+паз), -імпаз (-ім+паз)*** жұрнағы (және компоненттер) кейбір есім сөздерге жалғанып, туынды сын есім жасалса. Мыс.: *сезімпаз, алымпаз, білімпаз, жағымпаз, жасампаз* т.б.	230
	г) ***-қой, (-гой, -гөй)*** жұрнағы кейбір зат есім есім сөздерге жалғанып, амал-әрекет пен мінез-құлықтардың атауы болатын туынды сын есім тудырса. Мыс.: *көсіпқой, әзілқой, сәнқой, жәдігөй, әзілқой, сәнқой, жанжалқой* т.б.	231
	д) ***-қор*** жұрнағы зат есім есім сөздерге жалғанып, белгілі бір мақтықтанғанды білдіретін мағынадағы туынды сын есім жасалса. Мыс.: *жемқор, жалақор, айлақор, бейнетқор, қамқор, мансапқор, ызақор, намысқор* т.б.	232
	Етістіктерден өнімді жұрнақтар арқылы жасалатын туынды сын есім сөздер: 1. ***-қ, -к, -ық, -ік, -ақ, -ек:*** а) салт және сабақты етістікке жалғанып, түрлі сындық ұғымдағы атаулар жасайды. Мыс.: *ашық, тұнық, шірік, ілік, дөңгелек* т.б.; б) еліктеу сөздерге жалғанып, қатыстық сын есім тудырса. Мыс.: *бұлтақ, жалтақ, еңкек, жалпақ* т.б.; в) кейбір зат есімдерге жалғанып, туынды сын есім жасаса. Мыс.: *жолақ, қасырақ, ортақ, ирек* т.б.; г) етістіктен ***-уық (-үік)*** формасы арқылы белгілі бір іс-әрекетке бейімділікті білдіретін туынды сын есім жасаса. Мыс.: *жылауық, сойлеуік, сұрауық, сыбырлауық* т.б.	317

4.3-кестенің жалғасы

1	2	3
2	2. ***-ыңқы, -ңқы, -іңкі.*** Мыс.: *жатыңқы, салбыраңқы, көтеріңкі, кебіңкі, батырыңқы* т.б.	318
	3. ***-ынды, -інді, -нды, -нді.*** Мыс.: *асыранды, сертінді, жаттанды, шұбырынды, түйінді* т.б.	319
	4. ***–малы (-мелі, -балы, -белі, -палы, -пелі).*** Мыс.: *ауыспалы, көшпелі, таңдамалы, аумалы-төкпелі* т.б.	320
	5. ***-қыш, -кіш, -гыш, -гіш.*** Мыс.: *білгіш, оңғыш, тапқыш, сенгіш, айтқыш* т.б.	321
	6. ***-шақ (-шек).*** Мыс.: *мақтаншақ, еріншек, қызғаншақ, ашуланшақ, жасқаншақ, тартыншақ* т.б.	322
	7. ***-ымды (-імді, -мды, -мді).*** Мыс.: *жағымды, ұғымды, үйлесімді, жарасымды, сенімді* т.б.	323
	8. ***-улы, -үлі.*** Мыс.: *жанаулы, ертпеулі, үлгі* т.б.	324
	9. ***–қақ, -кек, -ғақ, -гек.*** Мыс.: *асқақ, піңғақ, жабысқақ, майысқақ, тайғақ, оңғақ, қатқақ, ұрысқақ* т.б.	325
	10. ***ма, -ме, -ба, -бе, -па, -пе.*** Мыс.: *жалдама (ақы), қызба (адам), көшпе (құм), сырма (бешпет), аспа (шам), сырма (бешпет)* т.б.	326
	Етістіктерден өнімсіз жұрнақтар арқылы жасалатын туынды сын есім сөздер: 1. ***–ыс, -іс, -с.*** Мыс.: *келіс, кетіс, ұқсас, таныс, тіркес, жалғас, талас* т.б.	327
	2. ***–қыр, -ғыр, -кір, -гір.*** Мыс.: *білгір, ұшқыр, үшкір, алғыр, өткір, білгір* т.б.	328
	3. ***–мыс (-мыш).*** Мыс.: *жасамыс, айтылмыш, тарамыс, жазылмыш, жаралмыш* т.б.	329
	4. ***–ымтал (-імтал).*** Мыс.: *ұғымтал, өсімтал, сезімтал* т.б.	330
	5. ***–қы (-кі, -ғы, -гі).*** Мыс.: *бұралқы, оралғы, жинақы, күлдіргі* т.б.	340
	6. ***–ыр (-ір, -ар, -ер, -р).*** Мыс.: *жұмыр, шір, обыр, қыңыр, тықыр, шымыр, құзар, былжыр, жалтыр, жылтыр, сылбыр, былбыр* т.б.	341
	7. ***–у.*** Мыс.: *жарау, қызу, тақау, жадау, жабырқау, тұзу, таяу, бітеу, баяу, көлбеу, қатаң* т.б.	342
	8. ***–ған (-ген).*** Мыс.: *тебеген, қабаған, сүзеген, қашаған, береген, алаған, безеген, көреген* т.б.	343
	9. ***–ын (-ін, -н).*** Мыс.: *ортан, бүтін, еркін, ұзын, жайын* т.б.	344
	10. ***–қалақ (-келек, -ғалақ, -гелек).*** Мыс.: *ұшқалақ, сасқалақ, қозғалақ* т.б.	345

4.3-кестенің соңы

1	2	3
2	11. ***–алақ (-елек).*** Мыс.: *еңкелек, қаңғалақ, бұлталақ, шыжалақ, жалталақ* т.б.	346
	12. ***–анақ.*** Мыс.: *сұғанақ, шұқанақ, шығанақ* т.б.	347
	13. ***–ғылықты (-гілікті, -қылықты, -кілікті).*** Мыс.: *жеткілікті, тұрғылықты, жергілікті* т.б.	348
	14. ***–мсақ (-мсек, -ымсақ, -імсек).*** Мыс.: *сұрамсақ, тілемсек, жарамсақ, берімсек, өлімсек* т.б.	349

Енді, осы *4.3-кестеде* қарастырылған сын есімнің лексика-морфологиялық құрылымына ақпараттық белгі-кодты сәйкестендіру бағдарламасы негізінде, кейбір негізгі және туынды сын есімдерге қатысты сөздерге де шартты белгі-код сәйкестендіру мысалдары *4.3.1* және *4.3.2-кестелерде* беріледі.

4.3.1-кесте

Сапалық (негізгі) сын есім сөздер	**Белгі-код орны**	
	1	**2**
қызыл, жасыл т.б.	CH1	201
жақсы, жаман т.б.	CH1	202
ұзын, қысқа, ауыр т.б.	CH1	203
тәтті, ащы т.б.	CH1	204

4.3.2-кесте

Қатыстық (туынды) сын есім сөздер	**Белгі-код орны**	
	1	**2**
сыртқы, төргі т.б.	CH2	210
кешкі, күндізгі т.б.	CH2	211
әсерлі, ынабатты т.б.	CH2	212
үлкенді-кішілі, өзенді-сулы т.б.	CH2	213
қаз мойынды, теке сақалды т.б.	CH2	214
мұңсыз, білімсіз т.б.	CH2	215
ойшыл, өзімшіл т.б.	CH2	216
балғадай, әкедей, шолмектей т.б.	CH2	217
қалалық, қоғамдық т.б.	CH2	218
айлық, жылдық т.б.	CH2	219
көйлектік, пальтолық т.б.	CH2	219

Қатыстық (туынды) сын есім сөздер	Белгі-код орны	
	1	2
өндік, қаншалық т.б.	CH2	220
сырлас, ниеттес т.б.	CH2	221
туыстас, қарындас т.б.	CH2	222
ауылдас, көршілес т.б.	CH2	223
жолдас, қызметтес т.б.	CH2	224
көйлекшең, пальтошаң т.б.	CH2	225
ашық, шірік, дөңгелек т.б.	CH2	317
көтеріңкі, жатыңқы т.б.	CH2	318
жаттанды, сөрпінді т.б.	CH2	319
таңдамалы, көшпелі т.б.	CH2	320
білгіш, оңғыш, сенгіш т.б.	CH2	321
мақтаншақ, еріншек т.б.	CH2	322
жағымды, үйлесімді т.б.	CH2	323
жиналулы, ерттеулі т.б.	CH2	324
ұрысқақ, тоңғақ т.б.	CH2	325
қызба (адам), сырма т.б.	CH2	326
таныс, келіс т.б.	CH2	327
сұрамсақ, өлімсек т.б.	CH2	349

4.5. Сын есімнің синтаксистік тәсіл арқылы жасалу типтерін шартты белгі-кодпен сәйкестендіру бағдарламасы

Сын есім сөздер құрамына қарай, жалаң және күрделі болып екіге бөлінетіні айқын. Жалаң сын есімдер бір ғана компоненттен тұрады, бірақ онда қосымша морфеманың болуы я болмауы шарт емес.

Ал күрделі сын есімдерге екі, үш және одан да көп компоненттерден құралып барып, синтаксистік (аналитикалық) жолмен жасалған сын есім сөздер жатады. Күрделі сын есімдер негізгі сын есім сөздерден және олардың тіркесуі, қосарлануы. бірігуі арқылы жасалып, тілімізде бір бүтін күрделі тұлға ретінде қызмет ететіні белгілі.

4.4-кесте

Сын есімнің синтаксистік тәсіл бойынша жасалу құрылымына ақпараттық белгі-кодты сәйкестендіру бағдарламасы

Белгі-код орны	Күрделі сын есімдер сипатының негізгі үлгілері	Белгі-код
1	**Күрделі сын есім**	СҚЗ
2	1. Сапалық (негізгі) сын есімдердің бірімен-бірі тіркеседі. Мысалы: *қара ала, сары ала, қызыл ала* т.б.: *қара көк, қара торы, қара күрең, қара шұбар* т.б.	400
	2. Бірыңғай я негізгі (сапалық), я туынды (қатыстық) сын есім сөздер не қайталанады, не қосарланады. Мысалы: *кішкене-кішкене, аппақ-аппақ, үлкен-үлкен* т.б.; *малды-малсыз, үлкенді-кішілі, қаралы-қопсылы* т.б.	401
	3. Негізгі сын есім сөзбен ***-лы, -лі*** жұрнақты туынды сын есім тіркеседі. Мысалы: *кең маңдайлы, қызыл шырайлы* т.б.; екі компоненті де не ***-лы, -лі*** не ***-сыз, -сіз*** жұрнақтары арқылы жасалған қос сөзден тұратын туынды (қатыстық) сын есімдер. Мысалы: *таулы-тасты, әкелі-балалы, ессіз-түссіз, жөнді-жөнсіз* т.б.	402
	4. Зат есім немесе сан есім сөздермен ***-лық (-лік, -дық, -тік, -тық, -тік)*** жұрнақтар арқылы жасалатын туынды сын есімдер тіркесін, күрделі сын есім жасалады. Мысалы: *халық аралық, бес жылдық, екі кісілік, екі-үш күндік* т.б.; *көңіл көтерерлік, қабырға қайысарлық* т.б.	403
	5. Екі компоненттің екеуіне де бірдей не ***-лық, -лік, -дық, -дік, -тық, -тік, -и*** не бір компонентті ***-и*** жұрнақтары арқылы жасалған күрделі сын есімдер. Мысалы: *әскери-саяси, ғылыми-әдеби, қоғамдық-тарихи, әлеуметтік-экономикалық, статистикалық-лингвистикалық* т.б.	404
	6. Тәуелділіктің 3-ші жағының қосымшасы қосылған зат есімге сын есім тіркесіп, күрделі сын есім есебіне қолданылады. Мысалы: *көзі ашық, тілі майда, басы бос, қолы қысқа, жүзі жылы* т.б.	405

Мәтін ішінде кездесетін сын есімнің синтаксистік тәсілі бойынша жасалатын сөздерді шартты белгілермен белгілеу немесе тиісті белгі-кодпен сәйкестендіру жоғарыда көрсетілген *4.4-кесте* арқылы іске асуы қажет деп үйғардық.

Аталған кестедегі мәліметтерді анықтай түсу үшін төменде сәйкестіктердің мысалы ретінде *4.4.1-кестені* ұсынып отырмыз. Мәтіндегі не жиілік сөздіктегі сын есімнің синтаксистік тәсіл арқылы жасалған түрлерін осылайша шартты түрде белгілеу негізінде оларға зерттеу кезінде ықтималды-статистикалық әдісті қолданудың мүмкіндігі туады.

4.4.1-кесте

Күрделі сын есім сөздер	Белгі-код	орны
	1	2
қара ала, сары ала, қара көк, қара торы, қара күрең т.б.	СН3	400
кішкене-кішкене, аппақ-аппақ, үлкен-үлкен т.б.; *малды-мансыз, үлкенді-кішілі, қоралы-қопсылы* т.б.	СН3	401
кең маңдайлы, қызыл шырайлы т.б.; *тауды-тасты, әкелі-балалы, ессіз-түссіз* т.б.	СН3	402
халық аралық, бес жылдық, екі кісілік, екі-үш күндік т.б.; *көңіл көтерерлік, қабырға қайысарлық* т.б.	СН3	403
әскери-саяси, ғылыми-әдеби, қоғамдық-тарихи, әлеуметтік-экономикалық, статистикалық-лингвистикалық т.б.	СН3	404
көзі ашық, тілі майда, басы бос, қолы қысқа, жүзі жылы т.б.	СН3	405

Сонымен, осы тарауда қарастырылған қазақ тілінің негізгі сөзтаптары ішінен зат есім, етістік және сын есімге қатысты сөз бен сөзгүлғаларды шартты түрде белгілеу, яғни олардың түр-түріне белгі-код сәйкестендіру бағдарламасы қазақ тілінің морфологиялық құрылымына статистикалық әдісті молынан қолдануға мүмкіндік тудырады. Мұндай бағдарламаны қазақ тілінің басқа да сөз таптары бойынша, оларға тән ерекшеліктерін ескере отырып, баяндалған жолмен іске асыруға болатындығына сеніміміз мол. Біз ұсынып отырған бағдарлама үлгісі әзірше теориялық ізденістердің бірі болғандықтан тәжірибе жүзінде толығынан қолданыс таба қойған жоқ. Бірақ мұндай статистикалық зерттеулер қазақ тілі үшін көп күттірмейтін жуық болашақтың ісі болмақ.

ҚОРЫТЫНДЫ

Компьютер көмегімен алынған қазақ тілінің көркем әдебиет және басқа да стильдер мәтіндерінен түзілген жиілік сөздіктерден (әліпбилік, жиілік, кері-әліпбилік, сөзнұсқағыш) лингвостатистикалық зерттеу арқылы алынған деректер қазақ мәтіндерін автоматты түрде өңдеуде, ұзақ мерзімге сақтауда және қажетті мәліметтерді іздеп, табуда аса қажет.

М.О.Әуезовтің қаламынан шыққан «Абай жолы» романының мәтіні тұңғыш рет статистикалық зерттеу нысанына айналып, көркем әдебиеттің бас жиыны (генералды жиын) немесе ықтималдық нысанының бас нұсқасы ретінде қазақ мәтінінің «электронды-ықтималдық сөздігінің» статистикалық үлгісіне айналды.

Өзге тілдерді зерттеуші ғалымдардың тұжырымдауы бойынша, бірқатар статистикалық сипаттамалар көптеген тілдерге ортақ болса, енді бірқатары тілдің өзіндік ерекшелігіне қарай, жанрлық айырымына, автордың тіліне ғана тән ерекшеліктеріне байланысты әр түрлі болады. Қазақ тілі мәтіндерінің де өзіне тән ерекшеліктері баршылық. Мәселен, әр түрлі стильдер мәтіндерінен түзілген жиілік сөздіктердегі сөзтұлғалар жиілігінің мәтінді не сөздікті қамту сипаты бірыңғай емес.

Түркі тілдеріндегі мәтіндерді статистикалық әдіспен зерттеудің теориялық та маңызы зор. Агглютинативтік құрылымды түркі тілдеріндегі зерттеулер нәтижелері флективті-аналитикалық құрылымдағы үндіеуропа тілдерінің бай статистикалық деректерімен салыстыра қарастыруға мүмкіндік береді. Мысалы мынандай жайтқа көңіл бөлейік. Орыс.

ағылшын, неміс, француз, румын, испан және т.б. тілдерде кез келген мәтіннің 50 пайызын қамту үшін ең жоғары жиілікті 100–150 сөзтұлға жеткілікті болса, ал түркі тілдерінде – жоғары жиілікті 700–800 сөзтұлға қажет екен.

Түркі тілдері бойынша ақпараттық сипаттама мөлшері де үндіеуропа тілдеріне қарағанда басқаша болуы мүмкін. Сондықтан лингвостатистикалық зерттеулер әрі типологиялық эксперимент аясына жатады деп түсінген жөн.

Еңбекте қазақ мәтіндері негізінде орындалған ықтималды-статистикалық және ақпараттық үлгі жасалып, олардың мүмкіндіктері белгіленді. Осындай зерттеу жұмысының нәтижелері бойынша, жиілік сөздіктің бастапқы жағынан соңына қарай қарастырғанда, сөзтұлғалардың мәтін бойын статистикалық және ақпараттық тұрғыда қамту сипатының өзіне тән ерекшеліктері тиісті теориялық критерийлер арқылы анықталды. Қазақ тілі мәтіндері бойынша түзілген жиілік сөздіктердегі сөзтұлғалардың көп жағдайда мәтінге тәуелсіз-ақ мағыналық дәрежеде болатындығы айқындалып отыр.

Қорыта айтқанда, кейбір тілдік құбылыстар өте әлсіз көрініс тауып, жасырын сипатта түрып, көп жағдайда зерттеушінің тікелей бақылауынан тыс қалады. Міне, осындай тілдік құбылыстарды анықтайтын математикалық құралды статистикалық зерттеулер арқылы табуға болады. Ондай құбылыстарға кейбір сөз таптарына қатысты *сөз, сөзтұлға, сөз тіркестердің* және т.б. бірліктердің тілдегі қолдану ерекшеліктері жатады.

Зерттеуімізде аталған математикалық құрал ретінде алынғандар:

а) айнымалы шаманың (тілдік бірліктер жиілігі) мәтін ішінде үлестірілуінің статистикалық заңдылықтары;

ә) тілдік бірліктердің статистикалық және ықтималдық сипаттарының арақатынастарын бағалайтын теориялық критерийлер.

Сонымен аталмыш оқу құралы түркі тілдеріндегі, соның ішінде қазақ тіліндегі мәтіндердің құрылымын объективті түрде квантитативтік әдіс-тәсілдермен зерттеуде көмегі болатындығына сеніміміз мол.

ҚОСЫМША

Оқу құралында пайдаланылған қазақша-орысша терминдер сөздігі

абсолютті ауытқу – абсолютное отклонение
абсолютті жиілік – абсолютная частота
абсолютті қате – абсолютная ошибка
абсцисс өсі – ось абсцисс
автоматты құрылғы – автоматическое устроиство
автоматты лингвистика – автоматическая лингвистика
автоматтандырылған түр – автоматизированный вид
айырым модулі – модуль разности
айнымалы шама – варирующая величина
ақиқат оқиға – достоверное событие
ақиқаттық аралық мән – действительное интервальное значение
ақпарат – информация
ақпаратты автоматты өңдеу – автоматическая обработка информации
ақпаратты автоматты іздеу (тауып алу) – автоматический поиск информации
ақпараттық өлшем – информационное измерение
ақпаратты сақтау – хранение информации
ақпаратты жолдау (тарату) – передача информации
ақпаратты ұзақ мерзімге сақтау – длительное хранение информации
ақпараттық артықтық (энтропия) – информационная избыточность
ақпараттық жүйе – информационная система
ақпараттық салмақ – информационный вес
ақпараттық салмақтың өсу қарқыны – темп (скорость) возрастания информационного веса
ақпарат теориясы – теория информации
ақпараттық үлгі (модель) – информационная модель
алгоритм – алгоритм
алшақтық дәрежесі – степень расхождения
алшақтық сипаты – характер расхождения
алшақтықты бағалау (сынау) – оценка расхождения

альтернативті (қарсы қойылатын) болжам – альтернативная гипотеза
аналитикалық (сараптамалық) түрпат – аналитическая форма
аралық код – промежуточный код
арифметикалық орта – средняя арифметическая
ауытқу шамасы – величина отклонения
ауытқудың орта шамасы – среднее значение от отклонений
ауытқулардың абсолют шамаларының ортасы – среднее от абсолютных отклонений
ашық жүйе – открытая система
әдіс – метод
әліпби-жиілік сөздік – алфавитно-частотный словарь
әліпби ретімен (бойынша) – по алфавиту
әліпбилік құрам – алфавитный состав
әмбебаптық сипат – универсальный характер
бағдарлама жұмысы – работа программы
базалық (негізгі) тіл – базовый язык
байқау саны – число наблюдений
байланыс күші – сила связи
байланыс торабы – узел связи
байланыс тығыздығы – теснота связи
бақылаулар тізбегі – последовательность испытаний
бақылау қатесі – ошибка наблюдения
бақылау (тәжірибе) саны – число испытаний
бас (генералды) мәтін жиынтығы – генеральная совокупность текстов
белгі (таңба) – знак
белгі-код – знак-код
белгілер жүйесі – система знаков
белгілеу (кодтау) әдісі – метод (прием) кодирования
билогарифмдік координат – билогарифмический координат
болжам – гипотеза
болжамды бағалау (тексеру) – оценка (проверка) гипотезы
болмыс бөлігі – часть действительности
бос аралық – пробел
буындық құрылым – слоговая структура
білім жүйесі – система знаний
білімді модельдеу (үлгілеу) – моделирование знаний

біртекті – однородный
біртектілік – однородность
бірлік – единица
бірліктер сипаттамасы – характеристика единиц
бірліктер арақатынасы – соотношение единиц
бірізділік (стандарт) – стандарт
генералды (бас) жиын (жиынтық) – генеральная совокупность
геометриялық орта – средняя геометрическая
грамматикалық ақпарат – грамматическая информация
грамматикалық белгі – грамматический признак
грамматикалық категория – грамматическая категория
грамматикалық көрсеткіш – грамматический показатель
ғылыми ақпарат – научная информация
дедуктивті – дедуктивный
дискретті вариациялық үлестірілу қатары – дискретный вариационный ряд распределения
дискретті бірліктер тізбегі – последовательность дискретных единиц
дисперсия (орта квадраттық ауытқу) – дисперсия
екінші (үшінші) реттегі орталық момент – центральный момент второго (третьего) порядка
елеусіз (сипаттағы) ауытқу – несущественное отклонение
елулі (сипаттағы) ауытқу – существенное отклонение
ең кіші квадраттар әдісі – метод наименьших квадратов
еркіндік дәреже саны – число степеней свободы
есептеу лингвистикасы – вычислительная лингвистика
жабық жүйе – закрытая система
жасанды интеллект – искусственный интеллект
жасанды тіл – искусственный язык
жиынтық абсолюттік жиілік – накопленная абсолютная частота
жиынтық ақпараттық салмақ – накопленный информационный вес
жиынтық қатынастық жиілік – накопленная относительная частота
жиілік – частота
жиілік сөздік – частотный словарь
жиілікті сөзтізбе – частотный словник

жиіліктердің үлестірілу заңдылығы – закон распределения частот
жоғары математика – высшая математика
жүйе – система
зерттеу нысаны – объект исследования
инженерлік лингвистика – инженерная лингвистика
индуктивті – индуктивный
интервал – интервал
интервалдық бөлік – интервальная часть
картотекалық қор – картотечный фонд
квадраттанған үлес ауытқуы – квадратичное отклонение доли
квантитативті құрылым – квантитативная структура
квантитативті лингвистика – квантитативная лингвистика
кері тәуелділік – обратная зависимость
кездейсоқ алшақтық – случайное расхождение
кездейсоқ жағдай – случайное обстоятельство
кездейсоқ лингвистикалық шама – случайная лингвистическая величина
кездейсоқ құбылыс – случайное явление
кездейсоқ оқиға (жағдай) – случайное событие
кездейсоқ таңдама – случайная выборка
кездесу жиілігі – частота встречаемости
кездесу ықтималдығы – вероятность встречаемости
келісім критерийі – критерий согласия
кему тәртібі – порядок убывания
кері әліпби-жиілік сөздік – обратно-алфавитный частотный словарь
кері пропорционал шама – обратно-пропорциональная величина.
кері сөздік – обратный словарь
кесте деректері – табличные данные
күткен оқиға – ожидаемое событие
кодтау принципі – принцип кодирования
комбинаторлық қасиет – комбинаторное свойство
компьютер – компьютер
компьютерлік аударма – компьютерный перевод
компьютерлік бағдарлама – компьютерная программа
компьютерлік қор – компьютерный фонд
компьютерлік лингвистика – компьютерная лингвистика

координат өстер жазықтығы – плоскость координатных осей
корреляция (өзгермелі шамалар арасындағы тәуелділік) – корреляция
корреляция коэффициенті – коэффициент корреляции
корреляциялық талдау – корреляционный анализ
кілт мәні (шамасы) – ключевое значение
қамту қарқыны – скорость (темп) покрываемости
қамту пайызы – процент покрываемости
қалыптастыру – формирование
қатынастық жиілік – относительная частота
қатынастық қате – относительная ошибка
қатынастық сипат – относительный характер
қатынастық-функционалдық салмақ – относительно-функциональный вес
қиылысу нүктесі – точка пересечения
қолайлы жағдай – благоприятствующий случай
қолайсыз жағдай – неблагоприятствующий случай
қолданбалы лингвистика – прикладная лингвистика
қолдану жиілігі – частота употребления
құбылу – варьирование (изменчивость)
құбылу (вариация) коэффициенті – коэффициент вариации
құжат – документ
құжат өрісі – поле документа
құрылымды-грамматикалық – структурно-грамматическая
лексика-морфологиялық құрылым – лексико-морфологическая структура
лексикалық белгі – лексический признак
лексикалық қор – лексический фонд
лингвистика (тіл білімі) – лингвистика
лингвистикалық бірлік – лингвистическая единица
лингвистикалық болжам – лингвистическая гипотеза
лингвистикалық мағына – лингвистическое значение
логика-математикалық – логико-математическая
маңыздылық ақпарат (семантикалық) – значимая информация (семантическая)
маңыздылық деңгей – уровень значимости
мағыналық жан-жақтылық – разнообразность значений
мазмұнды формальдау – формализация содержания

максимумдық шама – максимальное значение
математикалық баға – математическая оценка
математикалық лингвистика – математическая лингвистика
математикалық өрнек – математическая формула
математикалық статистика – математическая статистика
маңыздылық деңгей – уровень важности
машиналық аударма – машинный перевод
мәнмәтін – контекст
мәнмәтіндік қоршау – контекстное окружение
мәтін – текст
мәтін көлемі– объем текста
мәтін лингвистикасы – лингвистика текста
мәтін туындауы – порождение текста
мәтін ұзындығы – длина текста
мәтінді қамту – покрываемость текста
мәтінді нормалау – нормирование текста
мәтіндік бірлік – текстовая единица
мәтіннің семантикасы – семантика текста
мәтіндерді автоматты өңдеу – автоматическая обработка текстов
микротаңдама – микровыборка
микрожүйе (ең кіші жүйе) – микросистема
минимум сөздік – словарь минимум
модельдеуші мәтін – моделирующий текст
морфемдік құрам – морфемный состав
морфологиялық белгі – морфологический признак
морфологиялық деңгей – морфологический уровень
морфологиялық (синтаксистік, семантикалық) тармақтану – морфологическая (синтаксическое, семантическая) разветвление
морфологиялық түлға – морфологическая форма
мүмкін емес оқига – невозможное событие
нағыз орт жиілік – действительная средняя частота
нағыз үлес – действительная доля
нақты ғылымдар – точные науки
негізгі нысан – основной объект
нормалау сипаты – характер нормирования
нормалық табигат (тілдің) – природа нормы (языка)

нормальды үлестірілу – нормальное распределение
нөлдік болжам – нулевая гипотеза
нүктелер жүйесі – система точек
объективті – объективный
оқиға – событие
оқиғаның заңдылығы – закономерность события
оқиғаның ықтималдығы – вероятность события
оқиғаның шығуы (пайда болуы) – появление события
ординат өсі – ось ординат
орта ақпарат мәні – среднее информационное значение
орта ақпарат саны – среднее информационное число
орта жиілік – средняя частота
орта жиіліктен ауытқу – отклонение от средней частоты
орта квадраттық ауытқу – среднее квадратичное отклонение
орта таңдама жиілік – средняя выборочная частота
орта шамадан ауытқу – отклонение от средней величины
орта энтропия – средняя энтропия
өзгермелі сипат – изменчивый характер
өлшенген ақпарат – взвешенная информация
өрнек – выражение (формула)
өспелі функция – возрастающая функция
пайыздық деңгей – процентный уровень
параметрлерді нормалау – нормирование параметров
ранг (реттік нөмір) – ранг (порядковый номер)
регрессия түзуі – прямая регрессии
рет саны – порядковый номер
салыстырма деректер – сравнительные данные
санау құралы – инструмент подсчета
сандық деректер – числовые данные
сандық көрсеткіш – количественный показатель
сандық қатынас – количественное соотношение
сандық сипаттама – количественная характеристика
салыстырмалы-типологиялық – сравнительно-типологическая
селективті ақпарат – селективная информация
семантикалық белгі – семантический признак
семантикалық көрініс – семантическое представление
сенімділік – надежность

сенімділікті бағалау (тексеру) - оценка надежности
сериялық бөліктеу – деление на серий
символ – символ
символдар жүйелілігі – системность символов
синтаксистік белгі – синтаксический признак
синтаксистік бірлік – синтаксическая единица
синтаксистік деңгей – синтаксический уровень
синтаксистік қызмет – синтаксическая функция
синхрондық лингвистика – синхронная лингвистика
сипаттау грамматикасы – описательная грамматика
сөз байлық – словесное богатство
сөз статистикасы – статистика речи
сөздік бірлігі – единица словаря
сөздік қор – словарный фонд
сөздік құрылымы – структура словаря
сөздік өрісі – словарное поле
сөздік үзіндісі – отрывок словаря
сөздікті қағазға шығару – распечатка словаря
сөздікті қамту – покрываемость словаря
сөзқолданыс – словоупотребление
сөзнұсқағыш – словоуказатель
сөзнұсқағыш әліпби-жиілік сөздік – алфавитно-частотный словарь словоуказатель
сөзтіркес – словосочетание
сөзтұлға – словоформа
сөзформа (сөзтұлға) – словоформа
сөйлеу ақиқаттығы – речевая действительность
сөйлеу арнасы (байланыс арнасы) – канал речи (канал связи)
статистика-ақпараттық – статистико-информационная
статистикалық әдіс – статистический метод
статистикалық бағалау – статистическая оценка
статистикалық байланыс теориясы – статистическая теория связи
статистикалық бояу – статистическая окраска
статистикалық деректер – статистические данные
статистикалық заңдылық – статистическая закономерность
статистикалық құрал – статистический инструмент
статистикалық құрылым – статистическая структура

статистикалық лингвистика – статистическая лингвистика
статистикалық мәліметтер – статистические факты
статистикалық санақ – статистический подсчет
статистикалық сипаттама – статистическая характеристика
статистикалық үлестірілу – статистическое распределение
стильдік ерекшелік – стилевое отличие
стильдік-семантикалық – семантико-стилистическая
субъективті – субъективный
сызба – график
сызба-топтама - блок-схема
сызықтық пішін – прямолинейная форма
сызықтық функция – линейная функция
сынақ таңдамалары – оценочные выборки
сынама үлгі – испытательная модель
сынық сызықтық түрпат – ломанная прямая
сынық қисық – ломанная кривая
таңба – знак
таңдама бөлік – выборка
таңдама жиілік - выборочная частота
таңдама мәтіндер – выборочные тексты
таңдама жиіліктің орта жиіліктен ауытқуы – отклонение выборочных частот от среднего (значения)
тәуелді модельдер – зависимые модели
тәжірибе – эксперимент (испытание)
тәжірибе үстінде – во время испытания
тәжірибелік жиіліктер үлестірілуі – распределение эмпирических частот
теориялық-жиынтық – теоретико-множественная
теориялық-ақпараттық зерттеу – теоретико-информационное исследование
теориялық жиіліктер үлестірілуі – распределение теоретических частот
теориялық коэффициент – теоретический коэффициент
теориялық қатынастық жиынтық жиілік – теоретическая относительная накопленная частота
теориялық сызық – теоретическая линия
теориялық үлестірілу заңдылықтары – теоретический закон распределения

теориялық үлестірілу – теоретическое распределение
термин (атау) – термин
терминтану – терминоведение
техникалық және іскерлік коммуникация – техническая и деловая коммуникация
топтама – блок
түзу сызықтық байланыс тығыздығы – теснота прямолинейной связи
түлға – форма
тұрақты коэффициент – постоянный коэффициент
тұрақты сан (константа) – постоянное число (константа)
тұрақты шама – постоянная величина
тұрақтылық сипат – постоянный характер
тік сызықтық пішін – прямолинейный вид
тікелей байқау – непосредственное наблюдение
тікелей құрастырушылар моделі – модель непосредственно составляющих
тікелей сызықтық регрессия – прямолинейная регрессия
тікелей тәуелділік – прямая зависимость
тіл арқылы қатынас жасау (тілдік қатынас) – речевое общение
тілдесу – диалог
тілді ақпараттық бағалау – информационная оценка языка
тілдік ақпарат – языковая информация
тілдік бірлік – языковая единица
тілдік дерек – языковое данное
тілдік элемент – языковой элемент
тілдің ықтималдық үлгісі (моделі) – вероятностная модель языка
тірек сөз – опорное слово
үздіксіз вариациялық қатар – непрерывный вариационный ряд
үздіксіз шама – непрерывная величина
үйлесімдік дәреже – степень согласия
үйлесімдік дәрежені бағалау – оценка степени согласия
үйлесімдік критерийі – критерий согласия
үлгі (модель) – модель
үлес – доля
үлес ауытқуы – отклонение доли

үлестің тербелуі – колебание доли
үлестірілу (таралу) – распределение
үлестірілу түрпаты – форма распределения
үлестірілудің қисық сызықтық пішіні – положение кривой распределения
ұя – ячейка
ұялар топтамасы – блок ячеек
ұялар тізбегі – последовательность ячеек
функционалдық мән (мағына) – функциональное значение
функционалдық салмақ – функциональный вес
фонемалық спектр – фонемный спектр
фонемалық топ – фонемная группа
фонетикалық құрылым – фонетическая структура
формальды әдіс – формальный метод
формальды модель (формальды үлгі) – формальная модель
формула (өрнек) – формула
хабарламаны қабылдау – прием сообщения
хабарламаны тарату – передача (распространение) сообщения
шағын тіл – подъязык
шартты код – условный код
шартты кодтарға сәйкестендіру – ставить в соответствие условный код
шартты кодтау ұстанымы – придерживаться кодификации
шашырамдық дәреже – степень разбросанности
шекті дискретті бірлік – ограниченная дискретная единица
шектеулерді анықтау және бағалау – определение и оценка ограничений
шындық мән (мағына) – истинное значение
ықтималды-статистикалық әмбебап сызба – универсальная вероятностно-статистическая схема
ықтималды-статистикалық заңдылық – вероятностно-статистическая закономерность
ықтималды-статистикалық үлгі (модель) – вероятностно-статистическая модель
ықтималды-үлестірімді заңдылық – вероятностно-распределительная закономерность
ықтимал оқиға – вероятное событие
ықтималдық интеграл – интеграл вероятностей

ықтималдық қатынастар – вероятностные отношения
ықтималдық шама – вероятностная величина
ықтималдықтар теориясы – теория вероятностей
ЭЕМ – ЭВМ
электронды-есептеу машинасы – электронно-вычислительная машина
электронды-ықтималдық сөздік – электронно-вероятностный словарь
элемент – элемент
эмпирикалық сызық – эмпирическая линия
эмпирикалық (тәжірибелік) үлестірілу – эмпирическое распределение
энтропия – энтропия

Қолданбалы лингвистика мен математикалық лингвистика нәндерінде қолданылатын негізгі терминдердің орысша-қазақша сөздігі және анықтамалары

1. Алгоритм – *алгоритм.* Нақты мақсатқа жету үшін орындалатын амалдар реттілігі (тізбегі).

2. База данных – *деректер қоры.*

Ақпаратты автоматты түрде тарату, түсініктеме беру мен өңдеудің формальды көрінісінің қолайлы түрі.

3. База знаний – *білімдер қоры.*

Тілден сырт жатқан ақиқат шындық бөлшектері жайлы ілімнің формальды түсініктерінің элементтері.

4. Байт – *байт.*

Есептеу техникасындағы бірлік. Әдетте, біртұтас бірлік түрінде қарастырылатын екілік разрядтың тізбегі (көбінде тізбек саны 8-ге тең). Есептеу техникасы жадының өлшемі: 1 килобайт=1024 байт, 1мегабайт=1024 килобайт. Мәтінді өңдеу мен оны жадында сақтау жағдайында бір байт – бір символға сәйкес келеді.

5. Грамматика зависимостей – *тәуелділік грамматикасы.*

Құрамдастар араларында иерархиялық тәуелділігі бар сөйлем құрылымының формальды берілісі. Иерархия – төменгі дәрежедегі элементтердің жоғарғы дәрежедегілерге бағыну тәртібі.

6. Грамматика непосредственно составляющих – *тікелей құрастырушылар грамматикасы.*

Барынша тәуелсіз, сызықтық ішіндегі бір-бірімен қиылыспайтын элементтердің иерархиялық түрде біріне-бірі енетін сөйлем құрылымының формальды көрінісі.

7. Грамматика представлений – *түсініктер (көріністер) грамматикасы.* Әрбір сөздің екі жақтық (сол және оң жақ) қосылу мүмкіндіктері, яғни тіркесу мүмкіндігіне негізделген ережелер жүйесі.

8. Граф – *граф.*

Көп төбелі (нүктелі) және көп қабырғалы (байланысты) жиынның қос төбелерін қосатын (қосақтап) математикалық нысана.

9. Дерево предложения, дерево зависимостей – *сөйлем тармағы, тәуелділік тармағы.*

Тәуелді құрылымдардың құрамдас түйіндеріне сәйкестікке бейімделген граф түріндегі сөйлем құрылымының түсінігі.

10. Интерфейс – *интерфейс.*

Ақпараттар алмасуының амалдары мен тәсілдері. Адам мен машина аралығындағы интерфейс – адам мен ЭЕМ арасында диалогты ұйымдастырудың амалдары мен тәсілдері. Есептеу машиналары, бағдарламалар мен олардың жеке блоктары арасында да интерфейстер болады. Жасанды интеллектіге қатысты тілдік мәселенің бірі – табиғи-тілдік интерфейс құрастыру, яғни адамның ЭЕМ-мен (компьютермен) табиғи тілде сөйлесу мүмкіндігін іске асыру.

11. Квазиреферат – *квазиреферат.* Компьютер арқылы құрастырылған реферат (қысқаша мазмұн).

12. Конфигурационный анализ – *пішін үйлесімін талдау.*

Компьютерге енетін мәтінді екінші тілге аудару қажеттігіне қатысты тілдің алдын ала белгіленген синтаксистік «пішін үйлесімдер» жиыны негізінде талдау жүргізу. Егер компьютерге енетін мәтіннен жасалған «пішін үйлесімдігі» алдын ала белгілі «пішін үйлесімдігімен» сәйкес келсе, мәтіндік «пішін үйлесімдік» танылды деп саналады да, ол «шағымдалған пішінге» ие болып, әрі қарай да осындай ұстаныммен талданады. Жинақтау (синтездеу) кезінде талдау арқылы алынған осындай «пішін үйлесімділік» аударылуға тиісті тілдің «пішін үйлесімділігімен» салыстырылады. «Пішін үйлесімділік» сәйкестіктер алдын ала белгіленеді.

13. Лингвистический процессор – *лингвистикалық процессор.* Автоматтандырылған жүйелер үшін ЭЕМ-ге енетін табиғи тілдегі мәтіндерді өңдеу амал-әрекеттерінің жиынтығы.

14. Накопитель – *жинақтаушы.* Сыртқы құрылғы арқылы ақпаратты оқитын және жазатын ЭЕМ-нің (компьютердің) қондырғысы. Мысалы, магнит дискідегі жинақтаушы (дисковод), магнит таспадағы жинақтаушы.

15. Оконный интерфейс – *терезелі интерфейс.*

Соңғы кездегі интерфейс құру тәсілдерінің жетістігі болып саналады. Бұл тәсіл бойынша компьютер экранында қажетті ақпарат ерекше көрініс табады. Пішіні төрт бұрышқа тең ая, немесе «терезе» арқылы зерттеуші қажеттілікке сай компьютер жадына «енетін» және «шығатын» ақпараттарды орналастырады. Тілдесу (диалог) кезінде ондай «терезелердің» саны қажеттікке сай өсіп отырады.

16. Предсказуемостный анализ – *болжамдық талдау.*

Мәтін ішіндегі сөзқолданыстарды солдан оңға қарай талдай отыра зерттеу әрекеті (әдіс-тәсіл). Мәтіннен қарастырылатын әрбір келесі сөзқолданыстың алдын ала болжауға алынған сөзбен сәйкестігі тексеріледі.

17. Семантическое представление данных – *деректердің семантикалық берілісі.*

Семантикалық модельді сипаттауда қолданыс табатын ақпарат түсінігінің мазмұнын формальды түрде жазу тәсілі.

18. Словарь-конкорданс – *конкорданс-сөздік.*

Әрбір сөзтұлғаның көптеген мәтіндерде қолдану мүмкіндігіне сілтеме жасау үшін арнайы түзілген сөздік.

19. Тезаурус – *тезаурус.*

Тілдік бірліктер арасындағы семантикалық қатынасты көрсететін идеографикалық сөздік (идеография – әрбір таңбамен кез келген сөзді белгілеу). Тезаурустың құрылымдық негізі – нақты пәндік саланың иерархиялық ұғымдар жүйесі.

20. Файл – *файл (дерекжиын).* Біртұтас ретінде қарастырылатын компьютер жадындағы өзара байланыстағы жазулар жиыны.

21. Фрейм – *фрейм.* Мүмкін болатын құрамдар мен олардың арасындағы қатынас түрлерін атап көрсету үшін қажетті мәтін құрылымын не ситуацияны сипаттайтын амал.

22. Язык представления данных – *деректерді (берілістерді) сипаттау тілі.*

Ақпаратты оның құрылымы арқылы жазудың формальды әдісі. Деректерді сипаттаудың қайсыбір формальдық (математикалық) моделіне сүйенеді.

М.Әуезовтің «Абай жолы» роман-эпопеясының жиілік сөздігінен үзінді (ең жиі қолданылған 500 сөз)

Рет саны (i)	СӨЗ	Сөз табы	Абсолютті жиілік	Жиынтық абсолютті жиілік	Қатынастық жиілік	Жиынтық қатынастық жиілік
1	2	3	4	5	6	7
1	де	ет	9828	9828	0.02110	0.02110
2	бол	ет	9341	19169	0,02006	0.04116
3	е	ет	7844	27013	0,01684	0.05800
4	Абай	зт	6747	33760	0,01449	0.07249
5	өз	ес	5747	39507	0,01234	0.08483
6	да	шл	5653	45160	0,01214	0.09697
7	кел	ет	5175	50335	0,01111	0.10808
8	бұл	ес	4696	55031	0,01008	0.11816
9	ал	ет	4546	59577	0,00976	0.12792
10	де	шл	4506	64083	0,00968	0.13760
11	осы	ес	4379	68462	0,00940	0.14700
12	айт	ет	4301	72763	0,00924	0.15624
13	сол	ес	4175	76938	0,00897	0.16521
14	ол	ес	4117	81055	0,00884	0.17405
15	бір	са	3896	84951	0,00837	0.18242
16	отыр	ет	3568	88519	0,00766	0.19008
17	сөз	зт	2913	91432	0,00626	0.19634
18	қал	ет	2685	94117	0,00577	0.20211
19	бер	ет	2588	96705	0,00556	0.20767
20	жоқ	мд	2554	99259	0,00548	0.21315
21	үй	зт	2535	101794	0,00544	0.21859
22	түр	ет	2479	104273	0,00532	0.22391
23	кет	ет	2451	106724	0,00526	0.22917
24	бар	мд	2304	109028	0,00495	0.23412
25	мен	шл	2261	111289	0,00486	0.23898
26	екі	са	2254	113543	0,00484	0.24382
27	көп	мд	2226	115769	0,00478	0.24860
28	ет	ет	1954	117123	0,00420	0.25280
29	шық	ет	1946	119669	0,00418	0.25698
30	күн	зт	1890	121559	0,00406	0.26104
31	енді	үс	1858	123417	0,00399	0.26503

1	2	3	4	5	6	7
32	жүр	ет	1838	125255	0,00395	0,26898
33	көр	ет	1830	127085	0,00393	0,27291
34	ел	зт	1793	128878	0,00385	0,27676
35	ауыл	зт	1726	130604	0,00371	0,28047
36	бір	ес	1708	132312	0,00367	0,28414
37	қара	ет	1586	133898	0,00341	0,28755
38	бала	зт	1545	135443	0,00332	0,29087
39	бар	ет	1512	136955	0,00325	0,29412
40	бас	зт	1491	138446	0,00320	0,29732
41	ғана	шл	1484	139930	0,00319	0,30051
42	көз	зт	1472	141402	0,00316	0,30367
43	ат	зт	1464	142866	0,00314	0,30681
44	үлкен	сн	1452	144318	0,00312	0,30993
45	не	ес	1425	145743	0,00306	0,31299
46	іші	зт	1425	147168	0,00306	0,31605
47	біл	ет	1387	148555	0,00298	0,31903
48	бірақ	шл	1364	146619	0,00293	0,32196
49	сал	ет	1346	151265	0,00290	0,32486
50	Құнанбай	зт	1293	152558	0,00278	0,32764
51	жай	зт	1263	153821	0,00271	0,33035
52	тағы	шл	1232	155053	0,00265	0,33300
53	жер	зт	1220	156273	0,00262	0,33562
54	жақ	зт	1195	157468	0,00257	0,33819
55	ғой	шл	1177	158645	0,00253	0,34072
56	сен	ес	1123	159768	0,00241	0,34313
57	қатты	сн	1088	160856	0,00234	0,34547
58	қой	ет	1085	161941	0,00233	0,34780
59	баста	ет	1047	162988	0,00225	0,35005
60	қол	зт	1047	164035	0,00225	0,35230
61	адам	зт	997	165032	0,00214	0,35444
62	жат	ет	993	166025	0,00213	0,35657
63	жол	зт	984	167009	0,00211	0,35868
64	кісі	зт	976	167985	0,00209	0,36077
65	және	шл	963	168948	0,00207	0,36284
66	жет	ет	912	169860	0,00196	0,36480
67	мен	ес	909	170769	0,00195	0,36675
68	Базаралы	зт	895	171664	0,00192	0,36867

1	2	3	4	5	6	7
69	бәрі	ес	890	172554	0,00191	0,37058
70	топ	зт	890	173444	0,00191	0,37249
71	соң	шл	885	174329	0,00190	0,37439
72	Әбіш	зт	882	175211	0,00189	0,37628
73	жан	зт	875	176086	0,00188	0,37816
74	жігіт	зт	862	176948	0,00185	0,38001
75	сөйле	ет	856	177804	0,00184	0,38185
76	кез	зт	852	178656	0,00183	0,38368
77	алд	зт	822	179478	0,00177	0,38545
78	кас	зт	813	180291	0,00175	0,38720
79	ой	зт	789	181080	0,00169	0,38889
80	жақсы	сн	766	181846	0,00164	0,39053
81	сияқты	шл	756	182602	0,00162	0,39215
82	бар	ес	754	183356	0,00162	0,39377
83	күл	ет	753	184109	0,00161	0,39538
84	қалың	сн	751	184860	0,00161	0,39699
85	Тәкежан	зт	748	185608	0,00161	0,39860
86	тарт	ет	742	186350	0,00159	0,40019
87	дәл	үс	740	187090	0,00159	0,40178
88	Дәрмен	зт	730	187820	0,00158	0,40336
89	ат	ет	729	188549	0,00157	0,40493
90	үст	зт	729	189278	0,00157	0,40650
91	қазір	үс	726	190004	0,00156	0,40806
92	жас	зт	714	190718	0,00153	0,40959
93	жаңағы	сн	694	191412	0,00149	0,41108
94	қазақ	зт	689	192101	0,00148	0,41256
95	қала	зт	689	192790	0,00148	0,41404
96	жібер	ет	684	193474	0,00147	0,41551
97	қара	сн	678	194152	0,00146	0,41697
98	өт	ет	673	194825	0,00145	0,41842
99	жүз	зт	670	195495	0,00144	0,41986
100	ара	зт	669	196164	0,00144	0,42130
101	қыл	ет	668	196832	0,00143	0,42273
102	керек	мд	665	197497	0,00143	0,42416
103	ойла	ет	662	198159	0,00142	0,42558
104	Ербол	зт	660	198879	0,00142	0,42700
105	көңіл	зт	655	199474	0,00141	0,42841

1	2	3	4	5	6	7
106	аз	мд	652	200126	0,00141	0,42982
107	бет	зт	650	200776	0,00140	0,43122
108	бас	ет	647	201423	0,00139	0,43261
109	үн	зт	642	202065	0,00138	0,43399
110	кір	ет	639	202704	0,00137	0,43536
111	есті	ет	637	203341	0,00137	0,43673
112	әке	зт	633	203974	0,00136	0,43809
113	тап	ет	629	204603	0,00135	0,43944
114	Оразбай	зт	625	205228	0,00134	0,44078
115	қарай	шл	622	205850	0,00134	0,44212
116	өзге	сн	622	206472	0,00134	0,44346
117	мынау	ес	621	207093	0,00133	0,44479
118	барлық	ес	620	207713	0,00133	0,44612
119	халық	зт	614	208327	0,00132	0,44744
120	үшін	шл	608	208935	0,00131	0,44875
121	бүгін	үс	599	209534	0,00129	0,45004
122	Мағаш	зт	596	210130	0,00128	0,45132
123	жұрт	зт	595	210725	0,00128	0,45260
124	тек	шл	589	211314	0,00126	0,45386
125	үш	са	586	211900	0,00126	0,45512
126	түс	ет	572	212472	0,00123	0,45635
127	жалғыз	са	566	213038	0,00122	0,45757
128	ең	үс	564	213602	0,00121	0,45878
129	жатыр	ет	561	214163	0,00120	0,45998
130	пен	шл	560	214723	0,00120	0,46118
131	мал	зт	559	215282	0,00120	0,46238
132	қонақ	зт	558	215840	0,00120	0,46358
133	ұзақ	сн	554	216394	0,00119	0,46477
134	іс	зт	540	216934	0,00116	0,46593
135	аға	зт	539	217473	0,00116	0,46709
136	әлі	үс	538	218011	0,00116	0,46825
137	ән	зт	537	218548	0,00115	0,46940
138	мына	ес	537	218085	0,00115	0,47055
139	қыз	зт	533	219618	0,00114	0,47169
140	біз	ес	513	220149	0,00110	0,47279
141	ауыз	зт	513	220662	0,00110	0,47389
142	ақыл	зт	512	221174	0,00110	0,47499

1	2	3	4	5	6	7
143	көрін	ет	504	221678	0,00108	0,47607
144	соңғы	сн	504	222182	0,00108	0,47715
145	ал	шл	500	222682	0,00107	0,47822
146	қайт	ет	490	223172	0,00105	0,47927
147	болыс	зт	486	223658	0,00104	0,48031
148	Оспан	зт	482	224140	0,00103	0,48134
149	шақ	зт	479	224619	0,00103	0,48237
150	сұра	ет	475	225094	0,00102	0,48441
151	ес	зт	473	225567	0,00102	0,48543
152	бұрын	үс	468	226035	0,00100	0,48643
153	жауап	зт	466	225501	0,00100	0,48743
154	Әйгерім	зт	465	226966	0,00100	0,48843
155	сіз	ес	460	227426	0,00099	0,48942
156	күй	зт	453	227879	0,00097	0,49039
157	оқы	ет	453	228332	0,00097	0,49136
158	осындай	сн	451	228783	0,00097	0,49233
159	кім	ес	440	229223	0,00094	0,49327
160	та	шл	439	229662	0,00094	0,49421
161	тобықты	зт	434	230096	0,00093	0,49514
162	ақ	сн	429	230525	0,00092	0,49606
163	әйел	зт	428	230953	0,00092	0,49698
164	орта	зт	427	231380	0,00092	0,49790
165	таста	ет	426	231806	0,00091	0,49881
166	қайта	үс	421	232227	0,00090	0,49971
167	арт	зт	416	232643	0,00089	0,50060
168	сыр	зт	416	233059	0,00089	0,50149
169	бой	зт	408	233467	0,00088	0,50237
170	жиын	зт	407	233874	0,00087	0,50324
171	әңгіме	зт	406	234280	0,00087	0,50411
172	жыл	зт	401	234681	0,00086	0,50497
173	тіпті	үс	400	235081	0,00086	0,50583
174	жылқы	зт	399	235480	0,00086	0,50669
175	өлең	зт	396	235876	0,00085	0,50754
176	уақыт	зт	392	236268	0,00084	0,50838
177	жақын	сн	388	236656	0,00083	0,51004
178	тіл	зт	387	237043	0,00083	0,51087
179	аса	үс	385	237428	0,00083	0,51170

1	2	3	4	5	6	7
180	Дәркембай	зт	381	237809	0,00082	0,51252
181	әкел	ет	373	238182	0,00080	0,51332
182	әсіресе	үс	372	238554	0,00080	0,51412
183	сырт	зт	369	238923	0,00079	0,51491
184	үлық	зт	369	239292	0,00079	0,51570
185	қос	ет	363	239655	0,00078	0,51648
186	дос	зт	361	240016	0,00077	0,51725
187	тоқта	ет	359	240375	0,00076	0,51801
188	Бөжей	зт	358	240733	0,00076	0,51877
189	аш	ет	357	241090	0,00076	0,51953
190	талай	үс	357	241447	0,00076	0,52029
191	шеше	зт	357	241804	0,00076	0,52105
192	орын	зт	356	242160	0,00076	0,52181
193	ме	шл	355	242515	0,00076	0,52257
194	мінез	зт	354	242869	0,00076	0,52333
195	ту	ет	354	243223	0,00076	0,52409
196	үста	ет	354	243577	0,00076	0,52485
197	өл	ет	353	243930	0,00076	0,52561
198	түс	зт	353	244283	0,00076	0,52637
199	түн	зт	352	244635	0,00075	0,52712
200	ендігі	сн	351	244986	0,00075	0,52787
201	шығар	ет	350	245336	0,00075	0,52862
202	қат	ет	348	245684	0,00075	0,52937
203	ас	зт	344	246028	0,00074	0,53011
204	мін	ет	344	246372	0,00074	0,53085
205	кең	сн	339	246711	0,00073	0,53158
206	Баймағамбет	зт	338	247049	0,00072	0,53230
207	анық	сн	336	247385	0,00072	0,53302
208	бай	зт	336	247721	0,00072	0,53374
209	көрсет	ет	334	248055	0,00072	0,53446
210	дес	ет	333	248388	0,00071	0,53517
211	аңғар	ет	330	248718	0,00071	0,53588
212	те	шл	329	249047	0,00071	0,53659
213	мол	сн	328	249375	0,00070	0,53729
214	сонда	үс	328	249703	0,00070	0,53799
215	таны	ет	328	250031	0,00070	0,53869
216	жігітек	зт	327	240358	0,00070	0,53939

1	2	3	4	5	6	7
217	ба	шл	325	250683	0,00070	0,54009
218	хабар	зт	325	251008	0,00070	0,54079
219	күт	ет	323	251331	0,00069	0,54148
220	он	са	323	251654	0,00069	0,54217
221	соқ	ет	321	251975	0,00069	0,54286
222	Әзімбай	зт	317	252292	0,00068	0,54354
223	шап	ет	316	252608	0,00068	0,54422
224	ти	ет	315	252923	0,00067	0,54489
225	су	зт	313	253236	0,00067	0,54556
226	сәлем	зт	312	253548	0,00067	0,54623
227	істе	ет	312	253860	0,00067	0,54690
228	басқа	сн	309	254169	0,00066	0,54756
229	Шүбар	зт	308	254477	0,00066	0,54822
230	Жиренше	зт	307	254784	0,00066	0,54888
231	өмір	зт	307	255091	0,00066	0,54954
232	жүрек	зт	305	255396	0,00065	0,55019
233	орыс	зт	304	255700	0,00065	0,55084
234	жыла	ет	303	256003	0,00065	0,55149
235	шейін	шл	301	256306	0,00065	0,55214
236	жөнел	ет	301	256607	0,00065	0,55279
237	сөт	зт	301	256908	0,00065	0,55344
238	Майбасар	зт	300	257208	0,00064	0,55408
239	жаса	ет	298	257506	0,00064	0,55472
240	үнсіз	сн	298	257804	0,00064	0,55536
241	білдір	ет	297	258101	0,00064	0,55600
242	дала	зт	296	258397	0,00063	0,55663
243	бірге	үс	293	258690	0,00063	0,55726
244	бірі	ес	292	258982	0,00063	0,55789
245	жау	зт	292	259274	0,00063	0,55852
246	қабақ	зт	292	259566	0,00063	0,55915
247	қандай	ес	291	259857	0,00062	0,55978
248	рет	зт	291	260148	0,00062	0,56040
249	тез	үс	290	260438	0,00062	0,56102
250	Тоғжан	зт	290	260728	0,00062	0,56164
251	Үлжан	зт	290	261018	0,00062	0,56226
252	хат	зт	289	261307	0,00062	0,56288
253	жаңа	үс	286	261593	0,00061	0,56349

1	2	3	4	5	6	7
254	пе	шл	285	261878	0,00061	0,56471
255	сондай	сн	285	262163	0,00061	0,56532
256	тәрізді	шл	283	262446	0,00061	0,56593
257	қалай	ес	279	262725	0,00060	0,56653
258	есік	зт	278	263003	0,00060	0,56713
259	шай	зт	278	263281	0,00060	0,56773
260	Көкітай	зт	277	263558	0,00059	0,56832
261	біраз	үс	276	263834	0,00059	0,56891
262	жаз	ет	276	264110	0,00059	0,56950
263	қыс	зт	276	264386	0,00059	0,57009
264	ауыр	сн	273	264659	0,00059	0,57068
265	байлау	зт	268	264927	0,00058	0,57126
266	бері	шл	268	265195	0,00058	0,57184
267	шақыр	ет	266	265461	0,00057	0,57241
268	айнал	ет	265	265726	0,00057	0,57298
269	қон	ет	265	265991	0,00057	0,57355
270	үнде	ет	265	266256	0,00057	0,57412
271	аяқ	зт	264	266520	0,00057	0,57469
272	әуелі	үс	264	266784	0,00057	0,57526
273	же	ет	263	267047	0,00056	0,57582
274	бе	шл	261	267308	0,00056	0,57638
275	Семей	зт	259	267567	0,00056	0,57694
276	қой	зт	257	267824	0,00055	0,57749
277	на	шл	257	268081	0,00055	0,57804
278	әрі	шл	256	268337	0,00055	0,57859
279	ақын	зт	254	268591	0,00055	0,57914
280	сойлес	ет	254	268845	0,00055	0,57969
281	біт	ет	252	269097	0,00054	0,58023
282	кітап	зт	252	269349	0,00054	0,58077
283	от	зт	252	269601	0,00054	0,58131
284	ағайын	зт	251	269852	0,00054	0,58185
285	кок	сн	250	270102	0,00054	0,58239
286	көтерс	ет	248	270350	0,00053	0,58292
287	әдейі	үс	245	270595	0,00053	0,58345
288	арнаулы	сн	244	270839	0,00052	0,58397
289	жаңа	сн	244	271083	0,00052	0,58449
290	қош	ет	244	271327	0,00052	0,58501

1	2	3	4	5	6	7
291	алыс	сн	243	271570	0,00052	0.58553
292	сәл	үс	243	271813	0,00052	0.58605
293	бүгінгі	сн	241	272054	0,00052	0.58657
294	бес	са	240	272294	0,00052	0.58709
295	жатақ	зт	240	272534	0,00052	0.58761
296	ояз	зт	240	272774	0,00052	0.58813
297	жи	ет	239	273013	0,00051	0.58864
298	суық	сн	238	273251	0,00051	0.58915
299	жел	зт	237	273488	0,00051	0.58966
300	ата	ет	236	273724	0,00051	0.59017
301	бойынша	шл	236	273960	0,00051	0.59068
302	дүние	зт	236	274196	0,00051	0.59119
303	қан	зт	236	274432	0,00051	0.59170
304	кеш	зт	233	274665	0,00050	0.59220
305	іш	ет	233	274898	0,00050	0.59270
306	Абылғазы	зт	232	275130	0,00050	0.59320
307	Мәкен	зт	232	275362	0,00050	0.59370
308	жөн	зт	230	275592	0,00049	0.59419
309	сүйсін	ет	230	275822	0,00049	0.59468
310	үзын	сн	230	276052	0,00049	0.59517
311	кедей	зт	229	276281	0,00049	0.59566
312	туралы	шл	228	276509	0,00049	0.59615
313	байқа	ет	227	276736	0,00049	0.59664
314	қана	шл	227	276963	0,00049	0.59713
315	Павлов	зт	227	277190	0,00049	0.59762
316	алыс	ет	226	277416	0,00049	0.59811
317	қақ	ет	223	277642	0,00048	0.59859
318	Шыңғыс	зт	222	277865	0,00048	0.59907
319	ата	зт	222	278087	0,00048	0.59955
320	қызыл	сн	222	278309	0,00048	0.60003
321	ыстық	сн	222	278531	0,00048	0.60051
322	ие	зт	221	278752	0,00047	0.60098
323	тос	ет	221	278973	0,00047	0.60145
324	ас	ет	220	279193	0,00047	0.60192
325	құдай	зт	220	279413	0,00047	0.60239
326	теңқ	зт	220	279633	0,00047	0.60286
327	шын	сн	220	279853	0,00047	0.60333

1	2	3	4	5	6	7
328	артық	сн	219	280072	0,00047	0,60380
329	түй	ет	218	280290	0,00047	0,60427
330	шет	зт	218	280508	0,00047	0,60474
331	анау	ес	217	280725	0,00047	0,60521
332	тыңда	ет	217	280942	0,00047	0,60568
333	біреу	ес	216	281158	0,00046	0,60614
334	ерт	ет	216	281374	0,00046	0,60660
335	бүйрық	зт	215	281589	0,00046	0,60706
336	қалып	зт	213	281802	0,00046	0,60752
337	қу	ет	212	282014	0,00046	0,60798
338	Сүйіндік	зт	212	282226	0,00046	0,60844
339	жаман	сн	211	282437	0,00045	0,60889
340	сүй	ет	211	282648	0,00045	0,60934
341	құлақ	зт	209	282857	0,00045	0,60979
342	өмір	зт	208	283065	0,00045	0,61024
343	қадал	ет	208	283273	0,00045	0,61069
344	қыстау	зт	208	283481	0,00045	0,61114
345	асық	ет	207	283688	0,00044	0,61158
346	қамшы	зт	206	283894	0,00044	0,61202
347	Сармолла	зт	206	284100	0,00044	0,61246
348	ер	зт	204	284304	0,00044	0,61290
349	жуан	сн	204	284508	0,00044	0,61334
350	соншалық	үс	204	284710	0,00044	0,61378
351	үр	ет	202	284912	0,00043	0,61421
352	заман	зт	201	285113	0,00043	0,61464
353	би	зт	200	285313	0,00043	0,61507
354	үк	ет	200	285513	0,00043	0,61550
355	бөлек	сн	199	285712	0,00043	0,61593
356	дау	зт	199	285911	0,00043	0,61636
357	алған	үс	198	286109	0,00043	0,61679
358	ер	ет	198	286307	0,00043	0,61722
359	жеткіз	ет	198	286505	0,00043	0,61765
360	өзгеше	үс	198	286703	0,00043	0,61808
361	бұрыл	ет	197	286900	0,00042	0,61850
362	көрі	сн	197	287097	0,00042	0,61892
363	кейде	шл	197	287294	0,00042	0,61934
364	сез	ет	197	287491	0,00042	0,61976

1	2	3	4	5	6	7
365	айнала	зт	195	287686	0,00042	0,62018
366	айтыл	ет	195	287881	0,00042	0,62060
367	қайда	ес	195	288076	0,00042	0,62102
368	аралас	ет	194	288270	0,00042	0,62144
369	ақырын	үс	193	288463	0,00041	0,62185
370	кейін	үс	193	288656	0,00041	0,62226
371	қар	зт	193	288849	0,00041	0,62267
372	қағаз	зт	192	289041	0,00041	0,62308
373	қай	ес	192	289233	0,00041	0,62349
374	салқын	сн	192	289425	0,00041	0,62390
375	тақа	ет	192	289617	0,00041	0,62431
376	тілек	зт	192	289809	0,00041	0,62472
377	ай	зт	191	290000	0,00041	0,62513
378	бақ	ет	191	290191	0,00041	0,62554
379	косыл	ет	191	290382	0,00041	0,62595
380	төрт	са	190	290572	0,00041	0,62636
381	арыз	зт	189	290761	0,00041	0,62677
382	биік	сн	189	290950	0,00041	0,62718
383	бөкенші	зт	189	291139	0,00041	0,62759
384	жолдас	зт	189	291328	0,00041	0,62800
385	қора	зт	189	291517	0,00041	0,62841
386	құр	ет	189	291706	0,00041	0,62882
387	қоңыр	сн	188	291894	0,00040	0,62922
388	бойы	шл	187	292081	0,00040	0,62962
389	бірнеше	ес	186	292267	0,00040	0,63002
390	Ырғызбай	зт	185	292452	0,00040	0,63042
391	ажар	зт	184	292636	0,00040	0,63082
392	қарсы	сн	183	292819	0,00040	0,63122
393	өнер	зт	183	293002	0,00040	0,63162
394	ізде	ет	183	293185	0,00040	0,63202
395	кішкене	сн	182	293367	0,00040	0,63242
396	түс	зт	182	293549	0,00040	0,63282
397	тыс	зт	182	293731	0,00040	0,63322
398	міне	ес	181	293912	0,00039	0,63361
399	ақ	ет	179	294091	0,00038	0,63399
400	бауыр	зт	179	294270	0,00038	0,63437
401	алғашқы	сн	178	294448	0,00038	0,63475

1	2	3	4	5	6	7
402	қауым	зт	178	294626	0,00038	0,63513
403	мәлім	сн	178	294804	0,00038	0,63551
404	түгел	үс	177	294981	0,00038	0,63589
405	зор	сн	176	295157	0,00038	0,63627
406	Михайлов	зт	176	295333	0,00038	0,63665
407	еңбек	зт	174	295507	0,00037	0,63702
408	жай	ет	172	295679	0,00037	0,63739
409	қи	ет	172	295851	0,00037	0,63776
410	ертең	үс	171	296022	0,00037	0,63813
411	ет	зт	171	296193	0,00037	0,63850
412	мойын	зт	171	296364	0,00037	0,63887
413	азамат	зт	170	296534	0,00037	0,63924
414	ашық	сн	170	296704	0,00037	0,63961
415	қарсы	үс	170	296874	0,00037	0,63998
416	есеп	зт	169	297403	0,00036	0,64034
417	қысыл	ет	169	297212	0,00036	0,64070
418	ойлан	ет	169	297381	0,00036	0,64106
419	тас	зт	169	297381	0,00036	0,64142
420	қымыз	зт	168	297550	0,00036	0,64178
421	орал	ет	168	297718	0,00036	0,64214
422	ендеше	шл	167	297886	0,00036	0,64250
423	көрші	зт	167	298053	0,00036	0,64286
424	қыр	зт	167	298220	0,00036	0,64322
425	дауыс	зт	166	298387	0,00036	0,64358
426	ит	зт	166	298553	0,00036	0,64394
427	төре	зт	166	298719	0,00036	0,64430
428	жайлау	зт	165	298885	0,00035	0,64465
429	кездес	ет	165	299050	0,00035	0,64500
430	немесе	шл	165	299215	0,00035	0,64535
431	салмақ	зт	165	299380	0,00035	0,64570
432	тіле	ет	165	299710	0,00035	0,64605
433	Ысқақ	зт	165	299875	0,00035	0,64640
434	Байдалы	зт	164	300039	0,00035	0,64675
435	Бат	ет	164	300203	0,00035	0,64710
436	Ділдә	зт	164	300367	0,00035	0,64745
437	аңда	ет	163	300530	0,00035	0,64780

1	2	3	4	5	6	7
438	қаш	ет	163	300693	0,00035	0,64815
439	өлім	зт	163	300856	0,00035	0,64850
440	сайын	шл	163	301019	0,00035	0,64885
441	сонымен	шл	163	301182	0,00035	0,64920
442	тау	зт	163	301345	0,00035	0,64955
443	жиыл	ет	162	301507	0,00035	0,64990
444	осылай	үс	162	301669	0,00035	0,65025
445	сары	сн	162	301831	0,00035	0,65060
446	төсек	зт	162	301993	0,00035	0,65095
447	е	од	161	302154	0,00035	0,65130
448	ар	сн	160	302314	0,00035	0,65165
449	ана	зт	159	302473	0,00034	0,65199
450	белгі	зт	159	302632	0,00034	0,65233
451	байла	ет	158	302790	0,00034	0,65267
452	жұмыс	зт	158	302948	0,00034	0,65301
453	қой	шл	158	303106	0,00034	0,65335
454	қоныс	зт	158	303264	0,00034	0,65369
455	қоста	ет	158	303422	0,00034	0,65403
456	киім	зт	157	303579	0,00034	0,65437
457	патша	зт	157	303736	0,00034	0,65471
458	шапшаң	үс	157	303893	0,00034	0,65505
459	ағай	зт	156	304049	0,00033	0,65538
460	бел	зт	156	304205	0,00033	0,65571
461	кейде	үс	156	304361	0,00033	0,65604
462	қазіргі	сн	156	304517	0,00033	0,65637
463	аппақ	сн	155	304672	0,00033	0,65673
464	таң	зт	155	304827	0,00033	0,65706
465	тара	ет	155	304982	0,00033	0,65739
466	денелі	сн	154	305136	0,00033	0,65772
467	жар	зт	154	305290	0,00033	0,65805
468	кеуде	зт	153	305443	0,00033	0,65838
469	күлкі	зт	153	305596	0,00033	0,65871
470	үш	ет	153	305749	0,00033	0,65904
471	бөлме	зт	152	305901	0,00033	0,65937
472	ашыл	ет	151	306052	0,00032	0,65969
473	мұң	зт	151	306203	0,00032	0,66001

	2	3	4	5	6	7
474	атаулы	үс	150	306353	0,00032	0,66033
475	жандарал	зт	150	306503	0,00032	0,66065
476	жүз	са	150	306653	0,00032	0,66097
477	қайыр	ет	150	306803	0,00032	0,66129
478	айтыс	ет	149	306952	0,00032	0,66161
479	ұры	зт	149	307101	0,00032	0,66193
480	арман	зт	148	307249	0,00032	0,66225
481	таныт	ет	148	307397	0,00032	0,66257
482	жаз	зт	147	307544	0,00032	0,66289
483	солай	үс	147	307691	0,00032	0,66321
484	уақ	зт	147	307838	0,00032	0,66353
485	бүр	ет	146	307984	0,00031	0,66384
486	Мағыш	зт	146	308130	0,00031	0,66415
487	қараңғы	сн	144	308274	0,00031	0,66446
488	амандас	ет	143	308417	0,00031	0,66477
489	ауыс	ет	143	308560	0,00031	0,66508
490	жаға	зт	143	308703	0,00031	0,66539
491	сақал	зт	143	308846	0,00031	0,66570
492	ауру	зт	142	308988	0,00030	0,66600
493	әже	зт	142	309130	0,00030	0,66630
494	кезек	зт	142	309272	0,00030	0,66660
495	еркек	зт	140	309412	0,00030	0,66690
496	тоқтат	ет	140	309552	0,00030	0,66720
497	шеш	ет	139	309691	0,00030	0,66750
498	көпшілік	зт	138	309829	0,00030	0,66780
499	үнемі	үс	138	309967	0,00030	0,66810
500	іні	зт	138	310105	0,00030	0,66840

ӘДЕБИЕТ

1. Абай тілі сөздігі. Алматы, 1968. 734-б.

2. *Алексеев П. М.* Распределение лексических единиц по длине в тексте и словаре // Квантитативная лингвистика и автоматический анализ текстов. Гарту, 1986. Вып. 745. С. 3-28.

3. *Ахабаев А.* Статистический анализ лексико-морфологической структуры языка казахской публицистики. Автореф. дисс. канд. филол. наук. Алма-Ата, 1971. С. 24.

4. *Ахметов Е.Г.* Статистическая характеристика родительного падежа имени существительного в английском и казахском языках // Тезисы докладов и сообщений Всесоюзного семинара «Статистическое и информационное изучение тюркских языков». Алма-Ата, 1969. С. 51-54.

5. *Байтанаева Д.А., Бектаев К.Б.* Энтропия казахского текста // Статистика казахского текста. Алма-Ата, 1973. С. 664-696.

6. *Байтанаева Д.А.* Энтропия распределения слогов казахского письменного текста // Тезисы докладов и сообщений Всесоюзного семинара «Статистическое и информационное изучение тюркских языков». Алма-Ата, 1969. С. 80-82.

7. *Байтанаева Д.А.* Информационные характеристики казахского текста. Алма-Ата, 1985. С. 18.

8. *Балакаев М.* Современный казахский язык Алма-Ата, 1985.

9. *Бектаев К.Б.* Алфавитно-частотный словарь слогов казахского языка // Статистика казахского текста. Алма-Ата, 1973. С. 566-612.

10. *Бектаев Қ., Белботаев А., Смайылов Т.* Қазақ тіл біліміндегі лингвистиканың алғашқы көріністері // Құдайберген Жұбанов және қазақ совет тіл білімі. Алматы: «Ғылым». 1990. 108-115-бб.

11. *Бектаев К.Б., Пиотровский Р.Г.* Математические методы в языкознании. Ч.1. Теория вероятностей и моделирование нормы языка. Алма-Ата, 1973. 281 с.; Математическая статистика и моделирование текста. Алма-Ата, 1974. 334 с. Ч. II. Алма-Ата, 1974.

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

13. *Бектаев Қ.Б., Жубанов А.Қ., Мырзабеков С., Белботаев А.Б.* М.Әуезовтің «Абай жолы» романының жиілік сөздігі. Алматы, 1979. 334-б.

14. *Бектаев Қ.Б., Жубанов А.Қ., Мырзабеков С., Белботаев А.Б.* М.О.Әуезовтің 20 томдық шығармалар текстерінің жиілік сөздіктері. Алматы-Түркістан, 1995. 346-б.

15. *Бектаев К.Б., Зубов А.В., Ковалевич Е.Ф.* и др. К исследованию законов распределения лингвистических единиц // Статистика текста. Минск, 1969. С. 131-162.

16. *Бектаев К.Б., Лукьяненков К.Ф.* О законах распределения единиц письменной речи // Статистика речи и автоматический анализ текста. Л., 1971. С. 47-112.

17. *Бектаев Қ.* Ықтималдықтар теориясы және математикалық статистика. Алматы: «Рауан», 1991. 432 с.

18. *Белботаев А.Б.* Лингвостатистические характеристики частей речи казахского текста. Автореф. дисс. ... канд. филол. наук. Алма-Ата, 1992. С. 32.

19. *Белботаев А.Б.* Ғылыми-техникалық стильдегі сын есім аффикстерінің қолданылуы туралы. 225-230-бб.; Математика тексінің алфавитті-жиілік сөздігі. 543-552-бб. // Қазақ тексінің статистикасы. Алматы: «Ғылым», 1973.

20. *Белоногов Г.Г.* О некоторых статистических закономерностях в русской письменной речи // ВЯ. 1962. №1. С. 100-101.

21. *Беляева Л.Н.* Применение ЭВМ в лингвистических исследованиях и лингводидактике. Л., 1986.

22. *Бодуэн де Куртенэ И.А.* Языкознание, или лингвистика XIX века // Избранные труды по общему языкознанию. М., 1963. Т.2.

23. *Вейсалов Ф. Е., Керимов Я.К., Сафаров М.Г.* Комбинаторика фонем современного азербайджанского языка // Вопросы азербайджанского языкознания. Баку, 1981.

24. *Вентцель Е.С.* Теория вероятностей. М., 1969. 576 с.

25. *Винер Н.* Кибернетика. М., 1983.

26. *Виноградов В.В.* Современный русский язык. Введение в грамматическое учение о слове. 1938. Вып. 1. 160 с.

27. *Гарипов Т. М.* Лексикостатистические отношения поркских языков Урало-Поволжья // Языковые контакты в Башкирии. Уфа, 1972. С. 248-267.

28. *Геро А.С.* Предмет и основные направления прикладной лингвистики. // Прикладное языкознание. Ст.-Пб., 1996. С. 5-14.

29. *Гмурман В.Е.* Теория вероятностей и математическая статистика. М.: «Высшая школа», 1972. 368 с.

30. *Головин Б.Н.* Язык и статистика. М., 1971. 191 с.

31. *Джубанов А.Х.* Квантитативная структура казахского текста (опыт лингвистического анализа на ЭВМ). Алма-Ата. 1987. С. 147.

32. *Джубанов А.Х.* К вопросу о графемной статистике казахского текста // Вопросы казахской фонетики и фонологии. Алма-Ата, 1979. С. 79-86.

33. *Джубанов А.Х., Джунисбеков А.* Преобразование с помощью ЭВМ казахских текстов в фонетическую запись // Материалы семинара «Статистическая оптимизация преподавания языков и инженерная лингвистика». Чимкент. 1980. С. 317-319.

34. *Джунусбеков А.* Гласные казахского языка. Алма-Ата. 1972. С. 94.

35. *Длин А. М.* Математическая статистика в технике. М., 1958. 466 с.

36. *Ергалиев А.* Статистическая характеристика форм страдательного залога в английском и казахском языках // Тезисы докладов и сообщений Всесоюзного семинара «Статистическое и информационное изучение тюркских языков». Алма-Ата, 1969. С. 62-63.

37. *Estoup J.* Gammes stenographiques. Paris, 1916.

38. *Жетешиков Ж.* Статистический анализ словоизменительных аффиксов имен существительных киргизского языка (на материале газетных текстов)//Материалы семинара «Статистическая оптимизация преподавания языков и инженерная лингвистика». Чимкент, 1980. С. 25.

39. *Жубанов А.К.* Основные принципы формализации содержания казахского текста. Алматы, 2002. 250 с.

40. *Жубанов Қ.* Қазақ тілі жөніндегі зерттеулер. Алматы: «Ғылым», 1999. 581 б.

41. *Засорина Л.Н.* Частотные словари и вопросы лексико-статистики // Межвузовская конференция по вопросам частотных словарей и автоматизации. Л., 1966. С. 3-4.

42. *Звегинцев В.А.* 1) История языкознания XIX – XX веков о очерках и извлечениях. Ч. II. М., 1965; 2) Теорическая и прикладная лингвистика. М., 1968.

43. *Зекенова А.* М.О. Әуезов пьесаларындағы есім сөздердің морфологиялық құрылысы жайында // Известия АН КазССР. Сер. филолог. 1976. С. 56-62.

44. Знание языка и языкознание. М., 1991.

45. *Зубов А.В.* Основы лингвистической информатики: Учеб. пособие. Ч.I. Минск, 1991; Ч. II. 1992; Ч. III. 1993.

46. *Зубов А.В., Лихтарович А.А.* ЭВМ анализирует текст. Минск, 1989.

47. *Зубов А.В., Хотяшов Э.Н.* Статистический анализ текста с помощью ЭВМ // Энтропия языка и статистика речи. Минск, 1966. С. 118-166.

48. *Ибрагимов С.И.* Некоторые статистические характеристики имен существительных киргизского языка // Тезисы докладов и сообщений всесоюзного семинара «Статистическое и информационное изучение тюркских языков». Алма-Ата, 1969. С. 34-38.

49. *Ибрагимов Т.И.* Некоторые статистические данные о слогах татарского языка: вероятностные методы и кибернетика // Учен. зап. Казанского ун-та. Вып. 4. Т. 25. Кн. 6. 1966. С. 74-78; Его же. Изучения слогов и структуры их сочетаний в татарском литературном языке. Казань, 1970.

50. Использование ЭВМ в лингвистических исследованиях / Под ред. В.И. Перебейнос. Киев, 1990.

51. *Иманов К.И., Садиков А.В., Мырзабеков С.* Частотно-сопоставительная характеристика падежных форм русского и казахского языков // Тезисы докладов и сообщений Всесоюзного семинара «Статистическое и информационное изучение тюркских языков». Алма-Ата, 1969. С. 56-58.

52. *Кецесбаев І., Мусабаев Ғ.* Қазіргі қазақ тілі. Алматы, 1962. 315-б.

53. *Кенесбаева С.С.* К вопросу исследования протетического звука в словах арабского происхождения статистическим методом // Тезисы докладов и сообщений Всесоюзного семинара «Статистическое и информационное изучение тюркских языков». Алма-Ата, 1969. С.43-44.

54. *Клоусон ДЖ.* Лексико-статистическая оценка алтайской теории // ВЯ. 1969. №5. С.22-41.

55. Компьютеризация общества и человеческий фактор / Под ред. *А.И. Ракитова. М., 1988.*

56. *Кондратов А.М.* Электронный разум. М., 1987.

57. *Котов Р.Г., Якушин Б.В.* Языки информационных систем. М., 1979.

58. Қазақ тексінің статистикасы. Алматы, 1973. 731-б.

59. Қазақ тілінің түсіндірме сөздігі. 1-10-томдар. Алматы: «Ғылым», 1974-1986 жж.

60. *Қалыбеков Б.Е.* 50-жылдардағы бастауыш сынып оқулықтарының лексикалық жүйесі мен морфологиялық құрылымының статистикасы. Афтореф. дисс. ...канд. наук. Алматы, 2003. 25 с.

61. *Құрышжанов А.Қ., Жубанов А.Қ., Белботаев А.Б.* Куманша-қазақша жиілік сөздік. Алматы: «Ғылым», 1978. 277 б.

62. Лингвистический энциклопедический словарь. М., 1990.

63. *Лукьяненков К.Ф.* Использование схем Пуассона и Гаусса в исследовании распределения лингвистических единиц текста // Вопросы лингвостатистики и автоматизация лингвистических работ. М., 1970. Вып. 2. С. 3-14.

64. *Маматов Д.* Количественный анализ сложных согласных // Опыт экспериментального и структурного изучения языка. Ташкент, 1982. С. 108-111.

65. *Мандельброт Б.* О рекуррентном кодировании, ограничивающем влияние помех// Теория передачи сообщений. М., 1957. С. 139-157.

66. *Марчук Ю.Н.* Математические методы в языкознании. М., 1990.

67. *Молдабеков К.* Лингвостатистические исследования казахских текстов для младших школьников (на материале текстов учебников начальных классов и литературы для детей). Дисс. канд. филол. наук. Алма-Ата, 1985, 259 с.

68. *Мусабаев Ғ.* Қазіргі қазақ тіліндегі сын есімнің шырайлары. Алматы, 1951. 90 б.

69. *Мырзабеков С.* Қазақ тілін зерттеуде санды пайдалану // Қазақ тілі грамматикасы бойынша зерттеулер. Алматы, 1975.

70. *Мырзабеков С.* Статистико-лингвистический анализ структуры глагола современного казахского языка. Автореф. дисс. ... канд. филол. наук. Алма-Ата, 1973. С. 32.

71. *Налимов В.В.* Вероятностная модель языка. М., 1974.

72. *Насыров Д.Х., Аймбетов А.* О статистике фонем каракалпакского языка // Материалы семинара «Статистическая оптимизация преподавания языков и инженерная лингвистика». Чимкент, 1980. С. 273-274.

73. Новое в зарубежной лингвистике. Вып. XXIV. Компьютерная лингвистика. М., 1989.

74. *Пиотровская А.А.* Автоматическое приведение именных словоупотреблений к канонической форме // НТИ. Сер. 2. 1977. №1. С.32-36.

75. *Пиотровский Р.Г.* Экстралингвистические и внутриязыковые вопросы при переработке текста в системе «человек-машина-человек»// Вопросы социальной лингвистики. Л., 1969. С.40-64.

76. *Пиотровский Р.Г.* Информационные измерения языка. Л., 1969. 116 с.

77. *Пиотровский Р.Г.* 1) Текст, машина, человек. Л., 1978;

2) Инженерная лингвистика и теория языка. Л., 1979.

78. Прикладные аспекты лингвистики. М., 1989.

79. *Ризаев С.А.* Из опыта исследования статистической структуры слога в узбекском языке // Общественные науки в Узбекистане. Ташкент, 1972. №1; Его же. Статолингвистический анализ структуры слова в современном узбекском языке // Материалы семинара «Статистическая оптимизация преподавания языков и инженерная лингвистика». Алма-Ата, 1980. С. 180-182.

80. *Романовский В.И.* Математическая статистика. Ташкент, 1961. Кн. 1. 637 с.; 1963. Кн. 2. 794 с.

81. *Романовский В.И.* Применение математической статистики в опытном деле. М.–Л., 1947. 247с.

82. *Рогожникова Р.П., Чернышева Л.В., Кузнецова Е.* Основные направления автоматизация лингвистических исследований. Л., 1988.

83. *Рождественский Ю.В.* Техника, культура, язык. М., 1993.

84. *Садыков Т.* Проблемы моделирования тюркской морфологии (аспект порождения киргизской именной словоформы). Фрунзе, 1987. С. 121. Стуктурная и прикладная лингвистика. Вып. / Под ред. А.С.Герда. Л., 1978.

85. *Смирнов Н.В., Дунин-Барковский И.В.* Курс теории вероятностной и математической статистики. М., 1969. 511с.

86. *Струве П.Б.* Кто первый указал на применение статистики к филологическим исследованиям? // Изв. Российской АН. Сер. 6. 1918. №13.

87. *Татубаев С.С.* Статистические подходы к исследованию певческой фонетики казахского языка // Тезисы докладов и сообщений Всесоюзного семинара «Статистическое и информационное изучение тюркских языков». Алма-Ата, 1969. С. 38-40.

88. *Феллер В.* Введение в теорию вероятностей и ее приложения. М., 1967. Т.1. 498 с.

89. *Филин Ф.П.* О некоторых философских вопросах языкознания // Ленинизм и теоретические проблемы языкознания. М., 1970.

90. *Фрумкина Р.М.* Статистические методы изучения лексики. М., 1964. 115 с.

91. *Фрумкина Р.М.* Роль статистических методов в современных лингвистических исследованиях // Математическая лингвистика. М., 1973. С. 156 и след.

92. *Фрумкина Р.М.* Вероятность элементов текста и речевое поведение. М., 1971. 168 с.

93. *Фрумкина Р.М.* О законах распределения слов и классов слов // Структурно-типологические исследования. М., 1962. С. 124-133.

94. *Хальд Л.* Математическая статистика с техническим приложением. М., 1956. 644 с.

95. *Шингарева Е.А.* Семиотические основы лингвистической информатики. Л., 1987.

96. *Шор Я.Б.* Статистические методы анализа и контроля качества и надежности. М., 1962. 552 с.

97. *Штиндлова И.* Обратные словари // Автоматизация в лингвистике. М.; –Л., 1966. 87-88-бб.

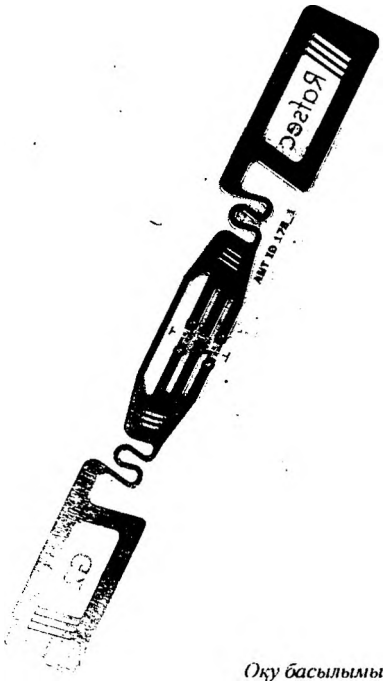
98. *Ысқақов А.* Қазіргі қазақ тілі. Алматы: «Ана тілі», 1991. 384 б.

99. Энтропия языка и статистика речи. Минск, 1966.

100. Словарь языка Пушкина. М., 1956-1961.Т. 1-4.

101. *Генкель М.А.* Частотный словарь романа Д.Н. Мамина-Сибиряка «Приваловские миллионы». Пермь, 1974.

МАЗМҰНЫ



Оқу басылымы

Жұбанов Асқар Құдайбергенұлы

ҚОЛДАНБАЛЫ ЛИНГВИСТИКА: ҚАЗАҚ ТІЛІНІҢ СТАТИСТИКАСЫ

Оқу құралы

Шығарушы редакторы *Агила Шуриева*
Компьютерде беттеген *Улбосын Әбдіқайымова*
Мұқабасын өңдеген *Қарлығаш Өмірбекова*

ИБ №2138

Басылуға 20.12.2004 жылы қол қойылды. Пішімі 60x84 1/16.
Көлемі 13,06 б.т. Офсетті қағаз. RISO басылыс. Тапсырыс №3177.
Таралымы 500 дана. Бағасы келісімді. Әл-Фараби атындағы
Қазақ ұлттық университетінің «Қазақ университеті» баспасы.
480078, Алматы қаласы, әл-Фараби даңғылы, 71
«Қазақ университеті» баспаханасында басылды.