Ғылыми мақала
ӘОЖ 81-114.2
https://doi.org/10.31489/2026PHI2/106-117
Мақаланың редакцияға түскен күні: 02.02.2026 ж.
Қабылданған күні: 27.03.2026 ж.

А.М. Фазылжан[1], Л.Р. Қойшығұлова[2*], Ж.Қ. Өмірбекова[3]

[1]А. Байтұрсынұлы атындағы Тіл білімі институты, Алматы, Қазақстан;
[2,3] Л.Н. Гумилев атындағы Еуразия ұлттық университеті, Астана, Қазақстан
(e-mail: afazyljan20@gmail.com, rahimjan2305@mail.ru, ozhamila@inbox.ru)

Корпустың репрезентативтілігіне қойылатын талаптар және олардың Қазақ тілінің ұлттық корпусындағы көрінісі

Мақалада корпустық лингвистиканың өзекті мәселелерінің бірі — тілдік корпустың репрезентативтілігі теориялық және практикалық тұрғыдан қарастырылады. Зерттеуде репрезентативтілік ұғымының ғылыми мазмұнын айқындалып, оған қойылатын негізгі талаптар жүйеленді және осы талаптардың Қазақ тілінің ұлттық корпусы (ҚТҰК) материалында жүзеге асу деңгейі талданады. Тілдік корпус белгілі бір мақсатта ұйымдастырылған, метабелгіленіммен және тілтанымдық аннотациялармен қамтамасыз етілген электрондық мәтіндер жиынтығы ретінде сипатталып, оның тілдің нақты қолданыстағы динамикалық бейнесін танытатын дереккөз екені көрсетіледі. Осы тұрғыдан репрезентативтілік корпустың ғылыми құндылығын, тілдік жүйені шынайы бейнелеу мүмкіндігін және зерттеу нәтижелерінің сенімділігін қамтамасыз ететін басты сапалық параметр ретінде бағаланады. Зерттеу барысында талдау, сипаттау, салыстыру әдістері қолданылды. Корпустық лингвистика зерттеушілерінің ғылыми еңбектері сараланып, соның негізінде корпус репрезентативтілігіне қойылатын сегіз негізгі талап айқындалды: стиль мен жанрдың әралуандығы, стильдік және регистрлік теңдік, аймақтық қамту, диахрония, әлеуметтік стратификация, мәтін көлемінің жеткіліктілігі, метадерек пен құжаттама, мәтіннің нақтылығы. Зерттеу нәтижесінде Қазақ тілінің ұлттық корпусының құрылымы қазақ тілінің функционалдық қабаттарын қамтуға бағытталғаны анықталды. Негізгі корпус пен арнайы ішкорпустар арқылы көркем, публицистикалық, ғылыми, ресми-іскери, ауызша, тарихи, поэтикалық және терминологиялық мәтіндер жүйеленгені байқалды. Сондай-ақ регистрлік теңгерімнің толық сақталмауы, әлеуметтік метадеректердің бірізді берілмеуі сияқты бірқатар мәселенің бары анықталып, диахронды және аймақтық қабаттардың әлі де толықтыруды қажет ететіні көрсетілді. Репрезентативтілік тілдік корпустың формальды белгісі ғана емес, оның ғылыми жарамдылығын, практикалық маңызын және тілдің шынайы қолданысын дәл бейнелеу мүмкіндігін айқындайтын басты қағида екені тұжырымдалды. Репрезентативті корпус тіл білімінің түрлі салаларындағы зерттеулерге, сөздік жасауға, білім беру ісіне және тіл саясатын ғылыми негізде жоспарлауға сенімді эмпирикалық база бола алады. Мақалада ҚТҰК-нің жанрлық және стильдік тұрғыдан біршама жүйеленген ұлттық цифрлық ресурс екені, алайда заманауи коммуникация мәтіндерін, әлеуметтік желі, чат-коммуникация, жастар тілі сияқты инновациялық ғылым салалары материалдарын көбірек енгізу қажеттігі де атап көрсетіледі. Сондай-ақ корпус көлемінің ұлғаюы сапалық теңгеріммен ұштасқанда ғана оның ғылыми әлеуеті артатыны, метадеректік жүйені тереңдету мен стратификациялық параметрлерді кеңейту болашақта қазақ тілінің толыққанды эмпирикалық бейнесін ұсынатын жоғары деңгейдегі ғылыми корпус қалыптастыруға мүмкіндік беретіні негізделеді. Бұл тұжырымдар корпустық лингвистиканың теориялық базасын нығайтып, ұлттық корпус құру ісінің әдіснамалық бағдарларын одан әрі нақтырақ айқындайды.

Кілт сөздер: репрезентативтілік, баланс, стратификация, метадерек, регистр.

Кіріспе

Тілдік корпус — жүйелі түрде жинақталған, белгілі бір мақсатта ұйымдастырылған, арнайы іздеу жүйесі мен қолданушыға және зерттеу бағытына қарай ізденушіге ыңғайлы интерфейсі бар метабелгіленіммен, түрлі тілтанымдық аннотациямен қамтамасыз етілген электрондық мәтіндердің біртұтас жиынтығы. Оның репрезентативтілігі мен оған қойылатын талаптар — зерттеу нысаны ретінде тілтанушы үшін күрделі әрі қызықты тақырыптың бірі. Бұл бағыттағы ізденістер отандық тіл біліміндегі корпустық лингвистика саласындағы зерттеулер қатарын кеңейте түседі.

Зерттеу нысанына алынған лингвистикалық корпус (әрі қарай — Корпус) — тілдің қолданыстағы нақты көрінісін жинақтап көрсететін, құрылымы алдын ала жоспарланған, электронды түрде сақталатын тегтелген мәтіндер базасы. Корпус мәтіндерді сақтап қана қоймай, оларды қандай да бір белгіленім бойынша (жанр, стиль, автор т.б.) жіктеп, арнайы бағдарламалар арқылы іздеуге,

* Хат-хабар үшін автор.e-mail: *rahimjan2305@mail.ru*

талдауға мүмкіндік береді. Осы тұрғыдан алғанда, корпус — тілдің динамикалық бейнесін айқындайтын, тілтанымдық зерттеулерге негіз болатын деректер қоры. Тіл біліміндегі жаңа бағыттың бірі — корпус лингвистикасы XX ғасырдың екінші жартысында дербес сала ретінде қалыптасты. Корпус лингвистикасы дегеніміз — тілдік бірліктерді үлкен көлемдегі мәтіндер жиынтығы негізінде зерттейтін, сандық әрі сапалық талдауды ұштастыратын ғылым саласы.

Мақалада корпус репрезентативтілігіне қойылатын талаптар әлемдік корпус құру тәжірибесіне сәйкес анықталды, сонымен бірге Қазақ тілінің ұлттық корпусы материалында талданды. Корпус репрезентативтілігі ұғымы теориялық және тәжірибелік негізде қарастырылды. Осыған байланысты корпустық лингвистикадағы репрезентативтілік ұғымына талдау; оның зерттелу тарихына шолу жасау; корпус репрезентативтілігіне қойылатын талаптарды анықтау міндеті алға тартылды.

Әдістер мен материалдар

Мақаланы жазу барысында талдау, сипаттау, синтездеу әдістері қолданылды. Таңдалған әдістер корпус лингвистикасының репрезентативтілігі ұғымын сипаттауға, оған қойылатын талаптарды анықтауға, соның негізінде корпус репрезентативтілігін талдауға мүмкіндік берді. Корпустық лингвистиканың кешегісі мен бүгінгісін шолып қарағанда, репрезентативтілікке қатысты ізденістердің барлық кезеңде өзекті болғаны анық байқалады. Корпустық лингвистикадағы зерттеулер Ж. Синклэйр мен Т. МакЭнери еңбектерінен бастау алғаны белгілі. Одан кейінгі Д. Бибер, В.П. Захаров еңбектерінде корпус лингвистикасының негізгі ерекшеліктері сөз болды. Отандық тіл білімінде А. Жұбанов, Э. Сүлейменова, А. Жаңабекова, А. Фазылжан, С. Құлманов, А. Барменқұлова, А. Қожахметова, К.Қ. Пірманова, С. Иманқұлова және т.б. ғалымдар корпустық лингвистиканың өзекті мәселелерін әр қырынан қарастырды. Мақалада көзделген мақсат-міндетке жету үшін тілдік корпус, репрезентативтілік талаптарына қатысты ғылыми тұжырымдар сарапталды; осы бағыттағы еңбектерге шолу жасалды; репрезентативтілік қағидасы пысықталды.

Нәтижелер мен оларды талдау

«Лингвистикалық корпус» түсінігі 1950 жылдары қалыптасқанымен, корпус лингвистикасы бірден дами қойған жоқ. Ф. Чарлес оған кедергі болған бірнеше себепті көрсетеді:

1) қуаттағыш мәшинелердің болмауы;

2) сол кездегі жетекші лингвист ғалымдар мен өзге де ғылым өкілдерінің корпус лингвистикасын қабылдамауы;

3) ғылыми зерттеулердің біраз бағыты дамымағандықтан, лингвистикалық корпусты жасаушылардың бұл саланың болашағына сенбеуі [1; 3]. Соған қарамастан компьютерлік ғылымның дамуына байланысты миллиондаған мәтіннен құралған лингвистикалық корпус қарқынды дами бастады. Корпус лингвистикасының теориялық негізін қалыптастырған ағылшын ғалымдары — Ж. Синклэйр мен Т. МакЭнери корпусты тілдің «нақты қолданыстағы бейнесін» көрсететін дереккөз ретінде сипаттады [2; 105]. Ал қазақ тіл білімі үшін бұл — салыстырмалы түрде жаңа салалардың бірі, өйткені қазақ тіліндегі корпус құру тәжірибесі 2010 жылдардан бастау алады.

2012 жылдан бастап қазақ тілінің ұлттық корпусын әзірлеу түбегейлі қолға алынды. Дегенмен қазіргі кезде бұл сала жақсы дамып келеді. Оған дәлел — корпус лингвистикасына қатысты негізгі терминдер жинақталған сөздіктің, осы тақырып аясында жүргізілген зерттеулердің болуы және корпус лингвистикасының жоғары оқу орындарында пән ретінде оқытылуы.

Алғашқы электрондық корпустар 1960-1970 жылдары пайда болды. Солардың ішінде ағылшын тіліндегі «Броун корпусы» тұңғыш толық көлемді корпус ретінде ғылымда айрықша орын алды [3; 22]. Бұл корпус 515 бет базасында жасалған 1 миллион сөзқолданыстан құралып, регистрлік әртараптылықты сақтады. Кейіннен Британ ұлттық корпусы мен қазіргі ағылшын корпусы әзірленіп, корпустың репрезентативтілік сипатын жаңа деңгейге көтерді.

Репрезентативтілік — тілдік корпус туралы зерттеулердің объектісі, себебі ол тілдік корпустың басты алғышарты, міндетті параметрі. Ол — кез келген корпустың табиғи тілдің бейнесін және сол тілді жасап отырған тілдік қауымдастықтың шынайы бейнесін дәл әрі ақиқат дүниедегі көрінісіне барабар қалыпта көрсете алу қасиетін қамтамасыз етеді. Ол жанр, регистр, аймақ, уақыт және әлеуметтік қабат секілді өлшемдерді қамтиды, яғни корпустағы материал «тіл осылай қолданылады» деген ғылыми тұжырым жасауға сай болуы керек. Американдық зерттеуші Д. Бибер репрезентативтілікті регистрлік және жанрлық стратификация арқылы қамтамасыз ету қажет екенін атап көрсетті. Оның пікірінше, корпустың көлемі емес, ішкі құрылымындағы жүйенің бірізділігі

корпустың нақтылығын айқындайды [4; 13]. Шетелдік корпустардың көбі осы қағидаға сүйеніп жасалады. Орыс тіл білімінде де бұл параметр корпус жасауда басшылыққа алынды. Орыс ұлттық корпусын қалыптастыруға үлес қосқан В.П. Захаровтың ойынша, корпус лингвистикасының маңызды ұғымы — репрезентативтілік. Репрезентативтілік дегеніміз — әртүрлі кезеңдегі, жанрдағы, стильдегі және авторлық т.б. сипаттағы мәтіндердің жеткілікті және пропорционалды ұсынылуы. Репрезентативтілікті анықтаушы әртүрлі тәсіл бар, жалпы тілдік (ұлттық) корпусқа қатысты бұл ұғымды қатаң есептеу және сипаттау қиын деп айтуға болады, математикалық тұрғыдан алғанда, бұған корпусты жобалау кезеңінде де, оны пайдалану кезеңінде де ұмтылуға болады әрі ол аса қажет [5; 5]. Өкілдікті анықтаудың әртүрлі тәсілі бар. Солардың бірі — жалпы тілдік (ұлттық) корпусқа қатысты бұл ұғымды қатаң математикалық жолмен сипаттау. Корпусты жобалау кезеңінде де, оны пайдалану кезеңінде де бұған ұмтылу маңызды.

Корпус тек белгілі бір жанр не стильді ғана қамтыса, ол бүкіл тілдік жүйені толыққанды сипаттай алмайды. Мәселен, корпус тек газет мәтіндерінен тұрса, онда қазақ тілі тек публицистикалық жағынан ғана қабылданатын еді, ал ол шындыққа жанаспайды, сондықтан репрезентативтілік — корпус сапасын айқындайтын басты көрсеткіш. Оның корпус құрудағы маңыздылығы төмендегі белгілері арқылы анықталады:

1) нақты ғылыми қорытынды жасауға мүмкіндік береді, себебі репрезентативтілік болмаса, жиілік, коллокация, грамматикалық ерекшеліктер біржақты ғана сипатталатын еді;

2) тіл саясаты мен білім беру жүйесін жетілдіруге септігін тигізеді, өйткені корпус негізінде сөздіктер, оқулық, түрлі зерттеулер жасалатындықтан, оның нақты болуы маңызды;

3) салыстырмалы зерттеулерге жол ашады, өйткені корпус репрезентативті болғанда ғана басқа тіл корпустарымен салыстыру объективті болады. Мысалы, ағылшын тіліндегі екі уақыт кеңістігіндегі корпустың репрезентативтілігі негізінде ағылшын тілінің әр дәуірі мен жанры бойынша салыстыру жасалады.

Репрезентативтілік ұғымы кең көлемде алғаш рет Д. Бибер еңбектерінде қарастырылды. Ол корпустағы мәтіндерді стильдік және жанрлық жағынан топтастыру қажет екендігін түсіндірді [4; 14]. Кейін Ж. Егберт пен Б. Грей репрезентативтілікті абсолютті өлшем емес, зерттеу мақсатына қарай салыстырмалы бағаланатын қағидат ретінде көрсетті [6; 50]. А. Килгарриф бастаған зерттеушілер репрезентативтілікті қамтамасыз етудің технологиялық жолдарын — мәтіндерді іріктеу, қайталанатын тақырыптарды анықтау тәсілдерін көрсетті [7; 810]. М.В. Пименова осы тақырыптағы зерттеу еңбегінде репрезентативтілікті тілдің функционалдық стильдерін теңгерімді қамтумен байланыстырды. Сондай-ақ ол тілдің тек құрылымдық емес, этномәдени қырын да қамту керектігін айтты [8; 120]. Қазақ тіл біліміндегі корпус идеясының негізін қалаған А. Жұбанов тілдік деректерді жанрға бөліп талдау арқылы корпусқа тән ұстанымдардың негізін қалады [9; 48]. К. Пірманова лингвистикалық зерттеулерде корпустық материалдарды пайдалану технологиясын қарастырады [10; 83-93]. Ғалымдардың пікірлері мен тұжырымдарын салыстыра отырып, корпус репрезентативтілігіне қойылатын талаптарды төмендегідей жүйелеуге болады.

1. Стиль мен жанрдың әралуандығы. Корпус репрезентативтілігі сақталуы үшін корпус мәтіндерінің әртүрлі стиль мен жанрды қамтуы маңызды, яғни көркем әдебиетпен ғана шектелмей, публицистикалық, ресми-іскери, ғылыми және ауызша сөйлеу стильдері, сондай-ақ поэзиямен қатар проза мәтіндері болуы тиіс. Г. Лич пен Л. Бурнард жанрлық жобалау корпус репрезентативтілігінің қаңқасы екенін, жанр үлестері алдын ала есептеліп, іріктеу сол квотаға бағынуы керектігін атап көрсетеді [11; 31]. Корпус — мәтіндерді электронды түрде жинауға негізделген ақпараттық-анықтамалық жүйе. Анықтамада айқындалып тұрғандай, біріншіден, корпус — бұл мәтіндерді электронды (цифрландырылған) түрде жинау; екіншіден, корпус анықтамалық-ақпараттық жүйе ретінде қолданыла алады [11; 42]. Э. Сүлейменова корпус көрнектілігімен/шамаластығымен, яғни теңдестірілгендігімен, тіл дамуының әртүрлі кезеңдеріндегі жазба және ауызша мәтіндердің барлық типтерінің тең көлемділігімен және репрезентативтілігімен, яғни тілдің өмір сүруінің қазіргі және/немесе оның өмір сүруінің барлық кезеңіндегі мәтіндердің сан алуан жанры, стилі, аумақтық және әлеуметтік варианты т.б. жазба және/немесе ауызша түрде беріледі деп атап өтті [12; 77], О. Жұбаева ұлттық корпус талабы ретінде жанрлық әртараптылықты атап көрсетсе [13; 487], А.Ә. Жаңабекова жанр белгіленімін метабелгіленіммен ұштастырған [14; 59].

2. Стильдік және регистрлік теңдік. Әр стиль мен регистр корпуста белгілі бір үлеспен теңдей қамтылуы тиіс. Жазба/ауызша, ресми/бейресми, академиялық/көпшілік регистрлерінің үлесі тең болуы керек. О. Ляшевская регистрлік теңдік бұзылса, жиілік өлшемдерінің сыртқы тұтастығы

құлдырайтынын көрсетеді [15; 44]. А. Плунгян ұлттық корпустарды «көпфункционалды стильдің ортақ алаңы» деп сипаттап, корпустағы мәтіндердің үлесі стильдік жағынан әртүрлі болуы корпустың құрылымын бұзатындығын айтқан [16; 21]. А. Барменқұлова регистрге тәуелді морфологиялық белгіленім статистикасын көрсетіп, теңсіздіктің белгіленім сапасына әсерін талдайды. Әр регистр үшін сөз саны мен мәтін саны белгіленуі қажеттігін, теңдік бұзылса, оны ескерту түрінде беру керектігін атап өтеді [17; 234].

3. Аймақтық қамту. Корпустың репрезентативтілігі сақталуы үшін белгілі бір тілдің барлық өңірлік ерекшелігі қамтылуға тиіс. Өңірлік алуантүрлілік корпуста дербес қабат болып көрінуі қажет. В. Лабов пен П. Трудгилл: «Жалпыға ортақ норманы беру тілдің әлеуметтік географиясын жасырады, сәйкесінше, тіл жайлы нақты деректерді бере алмайды», – дейді [18; 75]. Сондықтан аймақтық метадеректің болуы өте маңызды. Бұл тілді салыстырмалы түрде және жан-жақты зерттеуге әрі толыққанды тануға мүмкіндік береді.

4. Диахрония. Репрезентативтілік тілдің тарихи кезеңдерін қамтуды да талап етеді. Мәселен, XX ғ. басындағы, кеңестік кезеңдегі, тәуелсіздік тұсындағы тілдік ерекшеліктер жеке қамтылуы керек. Шетелдік корпус құру тәжірибесіне сәйкес тілдің тарихи кезеңдерін енгізбей, семантикалық және грамматикалық ерекшеліктерді нақты бағалау мүмкін емес. А. Сейітбекова қазақ тілінің ұлттық корпусы қазақ тілінің байырғы графикасы мен жазу ерекшелігін, автор қолтаңбасын, лексикасы мен сөз саптауын, өлең құрылысын, стильдік сипатын және тұрақты сөз орамдарын танытатын тілтанымдық дерек ұсынатын тарихи-поэтикалық ішкорпуспен де толықтырылуы қажет екендігін атап өтеді [19; 142].

5. Әлеуметтік стратификация. Репрезентативті корпус тілдің әлеуметтік қабаттарын қамтуы тиіс. Орыс тіл білімінде корпус лингвистикасын зерттеген М.В. Пименова әлеуметтік факторлардың (жас, жыныс, кәсіп) тілдік қолданысқа әсерін ерекше атап өтеді [8; 193]. Жас, жыныс, білім, кәсіп, әлеуметтік орта сынды қабаттар корпусқа енгізілуі тиіс. Әсіресе, қазіргідей қостілділік күшейген, жастар тілінде басқа тілден енген түрлі сөздер көбейген заманда олардың тілін әлеуметтік индексация тұрғысынан түсіндіруге болады. Тілдегі мұндай ерекшеліктерді байқау үшін мәтіндерге әлеуметтік тег енгізілуі керек.

6. Мәтін көлемінің жеткіліктілігі. Корпус көлемі тілдік жүйенің әртүрлілігін көрсетуге жеткілікті болуы керек. Мәтін көлемі артқан сайын сөздікқор да арта түседі, яғни шағын корпус сөз байлығын толық көрсете алмайды. Репрезентативтілік көлемнің жеткіліктілігімен қатар оның мазмұндық байлығымен де бағаланады. Сондықтан мәтін көлемін арттыруда олардың лексикалық әралуандығына да назар аударған жөн. Бұл репрезентативтілік талаптарының бір-бірімен өзара байланысты екендігін анық көрсетеді, яғни мәліметтің мөлшері сандық сапаны арттырғанымен, оның сындық сапасы да аса маңызды. Сирек кездесетін тілдік бірліктер мен жаңа сөз қолданыстарын тұрақты қадағалап отыру үшін мәтін көлемін арттыру қажет. Бірақ оның сапасына да назар аударып, белгілі бір сүзгіден өткізген жөн.

7. Метадерек пен құжаттама. Корпус әр мәтіннің толық сипаттамасын беретін метадерекпен қамтамасыз етілуі тиіс. Деректердің толық құжаттамасының болуы маңызды. Метадерексіз корпус деректер қоймасына айналады, бірақ ғылыми база бола алмайды. Сондықтан әр мәтіннің авторы, шыққан жылы, жанры секілді мәліметтер міндетті түрде белгіленуі керек. Әрбір тілдік корпустың метабелгіленімге арналған арнайы технологиясы қалыптасқан деуге болады. Осы деректер корпустың репрезентативтілігіне жан-жақты ақпарат береді.

8. Нақтылық. Корпус тек шынайы қолданыстағы, яғни аутентикалық мәтіндерден тұруы керек. Бұл — корпус сапасының негізі. Ойдан шығарылған немесе редакцияланған мәтіндер тілдің табиғилығын бұзады. Репрезентативтілікті сақтау үшін корпусқа тек түпнұсқа мәтіндер, халықтың шынайы тілдік қолданысын көрсететін материалдар енгізілуі тиіс.

Осы сипаттама бойынша Ахмет Байтұрсынұлы атындағы Тіл білімі институты әзірлеген Қазақ тілінің ұлттық корпусына талдау жасалды. Талдау нәтижесі 1-кестеде берілді.

1 - к е с т е

Қазақ тілінің ұлттық корпусының репрезентативтілігін талдау

№№№ і	Репрезентатив-тілікке қойылатын талаптар	Қазақ тілінің ұлттық корпусында берілуі	Түйін
1	2	3	4
11	Стиль мен жанрдың әралуандығы	ҚТҰК ішкорпустары арнайы мамандандырылған мәтіндік база және ақпараттық дереккөз ретінде қазақ тілінің алуан түрлі стильдік және жанрлық қабаттарын қамтыған инновациялық-цифрлық ресурс. Корпус — ашық жүйе, сондықтан ол үнемі толығып, кеңейіп отырады. Негізгі корпус — қазақ тілінің 5 стиль мәтіндерін түгел қамтыған корпус. Оның құрамында 30 миллион сөзқолданыстан — орфографиялық сөздікте бірігіп берілген, 515 бет тұратын мәтін бар. Әр стиль бойынша мәтіндердің көлемі: публицистикалық стиль — 11 950 995; көркем проза — 7 167 348; поэзия — 1 млн; Ғылыми стиль — 7 779 000; іскери стиль — 2 208 448; сөйлеу стилі — 1 млн; Оқулық мәтіндер –1 млн; Жалпы көлемі — 30 105 791. Негізгі корпустан басқа 19 ішкорпус қазақ тілінің стильдік және жанрлық қабаттарын толық дерлік қамтитын мәтіндік контентті құрайды. Негізгі корпуста көркем әдебиет, публицистикалық, ғылыми стиль мәтіні біршама, ал іскери стиль мәтіндерінің көлемінің аз болуы (жалпы іскағаз мәтіні — 669 бет) стандартты шаблондарға негізделуіне байланысты болса керек. Ауызша ішкорпус мәтіндері негізінде сөйлеу стилі жанрлық т.б. талапқа сай толық қамтылған, яғни репрезентативтілік талабының үдесінен шығатындай ұйымдастырылған. Ауызша ішкорпустың базасы дәстүрлі сөзсаптам үлгілерінен бастап күнделікті ауызекі сөйлеуге дейінгі 10 түрлі сөз стилін қамтитын сұхбаттардан тұрады. Жазушы ішкорпусында қазақ жазушыларының түрлі жанрда жазған шығармалары қамтылған. Поэтикалық корпус поэзия жанрының өлең, поэма, жыр т.б. типтерін қамтиды. Терминологиялық корпусқа мәтіндік база ретінде салынған ғылыми мәтіндер. Корпус репрезентативтілігін ғылыми стиль бойынша толықтырады. Сонымен қатар қазіргі кезде жаратылыстану саласы бойынша ғылыми мәтіндік контент жинақталып жатыр. Параллель ішкорпуста қазақ және орыс, екітілді көркем стиль мен іскери стиль мәтіндері теңестірілген. Бұндағы іскери мәтіндер ҚТҰК-нің стил бойынша репрезентативтілігін қамтамасыз етеді. Жарнама ішкорпусы ерекше жанрлық өріс ретінде қазақ тіліндегі түрлі типтегі жарнамалық мәтіндерді қамтиды. Мәдени-репрезентативті мәтіндер ішкорпусы — ғылыми-көпшілік, фольклор, көркем әдебиет сияқты ұлттық мәдени мәтіндік контентті құрайды [20].	Корпус жанрлық тұрғыдан әралуан, дегенмен заманауи жанрлық түрленім әлі де ескерілмеген: жасанды интеллект, ақпараттық технология, роботтандыру сияқты т.б. инновациялық ғылым саласының мәтіндері (чат-коммуникация, әлеуметтік желі мәтіндері, спичрайтинг мәтіндері, жастар тілінің мәтіндері т.б.)

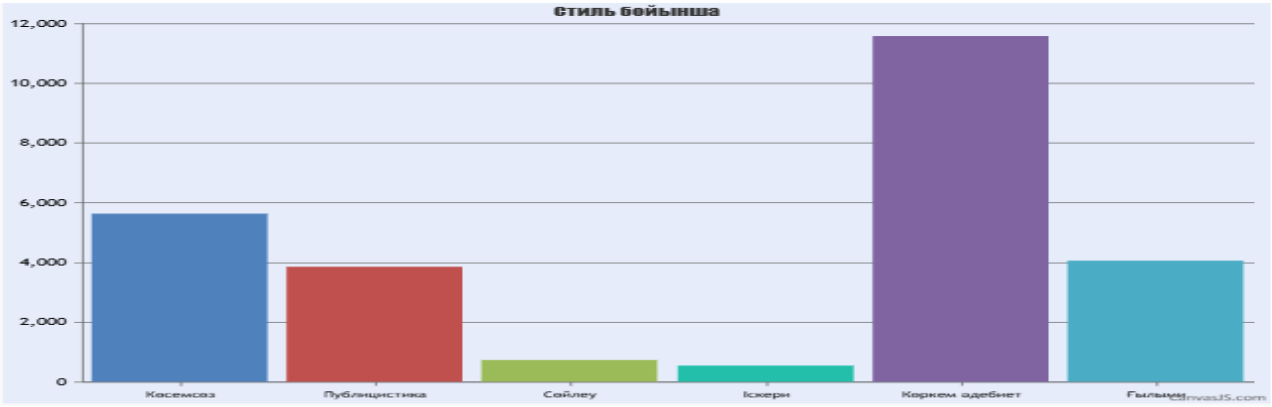
1-кестенің жалғасы

1	2	3	4
22	Стильдік және регистрлік теңдік	Корпустық лингвистикада Ұлттық корпусқа қойылатын талаптардың бірі — стильдік және регистрлік теңдік, басқаша айтқанда, теңгерімділік (сбалансированность). Алайда ешбір тілде мұндай баланс сақталмаған. Бұл талап көлемі шектеулі корпустар құрастыруда, яғни алғашқы кезеңдері сақталған. Мәселен, ҚТҰК тарихында алғашқыда 6 миллион сөзқолданыстан тұратын мәтіндік контент болған. Сол кезде бұл талап толық орындалды. Теңгерімділікті сақтау Корпус көлемін шектеуге әкелуі мүмкін. Сол себепті барлық тілдің ұлттық корпустарында бұл талап қатаң сақтала бермейді. ҚТҰК-де де осы үрдіспен мәтіндік контенттің тілдегі жазылу шегіне қарай кейбір стильдердің аз-көп болуы — қалыпты жағдай.	Регистрлік теңдік сақталмаған, сондықтан жиілік және қолданыс талдауларында кейбір деректер тілдің объективті көрінісіне сәйкес келмей қалуы мүмкін.
3 33	Аймақтық қамту	ҚТҰК-нің Диалектілік ішкорпусы әзірге кезінде диалектолог-мамандардың өңірлік экспедициялары арқылы жинақталған Аймақтық сөздікке енгізілген жергілікті мәтіндік контентті қамтиды. Аймақтық сөздік материалдары жергілікті ерекшелігі бар сөздер қолданылған ауызекі сөйлеу стилі мен жергілікті газеттерден жиналған және өңірлік ерекшеліктер көрініс беретін көркем шығармалардан алынған. Бұл экспедициялар Қазақстанның барлық өңірі бойынша жасалғандықтан, Диалектілік ішкорпусқа енгізілген мәтіндік контент тілдің аумақтық таралуын қамтамасыз етеді деуге болады [23; 105]. Сонымен қатар Қытай, Моңғолия қазақтарының т.б. диаспоралық ерекшеліктер де қамтылған. Жалпы Диалектілік корпусты өңірлік ауызша сөйлеу тілінен аудио/видео таспаға жазып алу арқылы толықтыру көзделіп отыр.	Қазақ тілі (әдеби стандартты тіл) ауызша тараған жалпыхалықтық тіл негізінде қалыптасқан [21; 13-14], жергілікті ерекшелігі терең көрінбейтін, яғни диалектілік түрленімі жоқ монолитті тіл [22; 97-98] болғандықтан қазақ тілі лексикасында ғана кездесетін кейбір ерекшіліктерді ғалымдар говор деп танып жүр. Сондықтан ҚТҰК диалектілік ішкорпус атына сай тілтанымда анықталған шын мәніндегі диалектіні білдірмейді. Дегенмен өңірлік ерекшеліктер жоқ емес, сол себепті бұл ішкорпусты байытатын өңірлік ауызша сөз үлгілерін жинауды тоқтатпау керек. Сол сияқты автохтонды тіл ретінде қазақ тілінің Қытай, Ресей, Түркия, Өзбекстан, Иран т.б. өзге мемлекеттегі этномәдени қазақ қауымдастықтарының тілдік мәліметі жеткіліксіз, солармен де ішкорпусты толықтыру керек.
44	Диахрония	ҚТҰК-нің «Тарихи ішкорпусының» мәтіндік базасына: 1. V-IX ғғ. көне түркі дәуіріндегі жазба ескерткіштері; 2. XI-XV ғғ. орта түркі дәуіріндегі жазба мұралар; 3. XVI-XIX ғғ. ескі қазақ жазба үлгілері; 4. XX ғ. басындағы жазба мұралардың мәтінін салу жоспарланған. Қазіргі кезеңде араб графикасында жазылған XII-XX ғғ. аралығындағы жазба мұралардың мәтіндері іріктеліп енгізілген. Орта түркі жазба ескерткіштерінен XII ғғ. Ахмет Йүгнекидің «Ақиқат сыйы», Ахмет Ясауидің «Диуани хикмет», XVII ғғ. Әбілғазы Баһадүр ханның «Түркі шежіресі», XVIII-XIX ғ.ғ. ресми-іс қағаздары үлгілерінің факсимелесі, транскрипциясы, аудармасы параллель түрде берілген	ҚТҰК-де XII ғасырдан бастап XX ғасырға дейінгі кезеңдегі қолжазбалар мәтіні қамтылады. Дегенмен де тілдің тарихи дамуын, эволюциясын көрсетуге мүмкіндік әлі де шектеулі. Диахронды стратификацияны руникалық жазба ескерткіштер мәліметтерімен толықтырып, араб графикалы мәтін базасын көбейту керек.

1-кестенің жалғасы

1	2	3	4
		[19; 85-86]. Тарихи-поэтикалық ішкорпусқа енгізілетін мәтіндер базасын әзірлеудің алғашқы кезеңінде XV-XX ғғ. аралығындағы халық ауыз әдебиеті үлгілерінің поэтикалық мәтіндері іріктеліп, енгізілді. Олардың *қатарында хикаялық дастандар, діни дастандар, батырлар жыры, ғашықтық жырлар, қаһармандық дастандар* қамтылды. Аталмыш ішкорпуста халық ауыз әдебиеті үлгілерінің араб графикасында жазылған факсимелесі, аудармасы параллель түрде ұсынылды.	
55	Әлеуметтік стратификация	ҚТҰК-ге енгізілген мәтіндердің метабелгіленімдерінің құрамында бірнеше әлеуметтік дерек бар: Автордың жынысы; Аудитория жасы; Ауызша корпуста сөзсаптам иелерінің әлеуметтік деңгейі т.б. Осы деректер бойынша қолданушы іздеу жасай алады [10; 90].	Корпустың әлеуметтік стратификацияны көрсету құралдарын жетілдіруді ұсынамыз. Себебі ол барлық ішкорпус метабелгіленімінде бірдей көрініс таба бермейді, бұл қоғамдағы стратификациялық ерекшеліктерін талдауға кедергі келтіреді.
66	Мәтін көлемінің жеткіліктілігі	Қазақ тілінің ұлттық корпусы қазіргі кезде белгілі бір мақсатта арнайы әзірленген 20 ішкорпустан тұрады. Жалпы сөзқолданыс саны — 100 000 000, 2026 жылдың басында оның көлемінің бұдан 2 есеге өсетіні айтылған.	Кемінде 200–300 миллион сөзге жеткенде ғана тілдің лексикалық әртараптылығы шынайы көрінеді.
77	Метадерек пен құжаттама	Метабелгіленім — корпусқа енгізілген әрбір мәтін туралы толыққанды ақпарат. ҚТҰК-дегі метабелгіленімдер параметрі 20-дан асады және ішкорпустардың ерекшелігіне қарай түрлендірілген [3; 435]. Әрине, кейбір мәтіндерде автордың жынысын көрсету (автордың аты-жөнінен тану мүмкін емес жағдайда), мәтін тақырыбын көрсету (көркем шығармалар шытырман оқиғаға құрылады) және кейбір мәліметтердің жоқ болуына, белгісіздігіне байланысты қиындықтар туындайды. Алайда корпуста мұндай кемшіліктер үнемі түзетіліп, толықтырылып отырады.	Негізгі корпустың метабелгіленімі ауқымды. Басқа ішкорпустарды түгел сол үлгімен ауқымды түрде беру керек. Толық метадеректсіз корпус ғылыми зерттеулерге толықтай жарамсыз болуы мүмкін.
88	Нақтылық	ҚТҰК базасына қазақ тілінде жазылған, жарық көрген түрлі стильдегі шынайы мәтіндер салынған. Метадеректері мәтін дереккөзінің шынайылығын нақтылай түседі. Мәтіндер өзгеріске түспеген, түпнұсқалары сақталған.	ҚТҰК нақтылық параметрі толық сақталған, себебі тек объективті өмірде бар тілдік дерек жинақталған. Жасанды мәтін жоқ.

1-суретте берілген диаграмма ҚТҰК стильдік және жанрлық алуантүрлілікті қамтамасыз ете отырып, тілдік зерттеулер жүргізіп, сөздік түзуге сенімді эмпирикалық база бола алтынын көрсетеді.



1-сурет. Қазақ тілінің ұлттық корпусындағы сөздердің стиль бойынша статистикасы

2-суретте көрсетілген ҚТҰК-дегі қазіргі мәтіндер базасындағы сөздердің жылдар бойынша статистикасына назар аударар болсақ, 2000 жылдан кейін базадағы сөздердің саны айтарлықтай өскені байқалады. Бұл қазақ тіліндегі тілдік материалдың жыл өткен сайын кеңейіп, толыға түскенін білдіреді. 2013 жылдан кейінгі көрсеткіштің төмендеуі корпусқа енгізілетін мәтіндердің көлемі мен сапасындағы өзгерістерді, жоғарыда берілген талаптардың толықтай сақталмауын көрсетеді.

2-сурет. Қазақ тілінің ұлттық корпусындағы қазіргі мәтіндер базасындағы сөздердің жылдар бойынша статистикасы

Қорытынды

Репрезентативтілік — корпустың басты сапа өлшемі. Репрезентативтіліксіз корпус дерекқор ғана болып қалады, ғылыми база бола алмайды; тілдің жанрлық, әлеуметтік, аймақтық бейнесі көрінбейді; білім беру жүйесіне, оқулық, сөздік жасау ісіне сенімді негіз болмайды.

Корпустың репрезентативтілігіне қойылатын мақалада көрсетілген сегіз негізгі талап-бағыт мынадай: стиль мен жанрдың әралуандығы; стильдік және регистрлік теңдік; аймақтық қамту; диахрония; әлеуметтік стратификация; мәтін көлемінің жеткіліктілігі; метадерек пен құжаттама; нақтылық. Бұл талаптар бір-бірімен өзара тығыз байланысты. Олар толық орындалғанда ғана корпус тілдің толық әрі объективті сипаттамасын бере алады.

Репрезентативтілік — тек көлемдік көрсеткіш емес, ең алдымен, жанрлық, стильдік, әлеуметтік, аймақтық және тарихи теңгерімділікті қамтамасыз ететін сапалық та категория болғандықтан, талапқа сай жасақталған корпус зерттеушіге тілдің шынайы қолданысын, жиілік заңдылықтарын, стильаралық айырмашылықтарды, семантикалық өрістер мен прагматикалық құрылымдарды дәл талдауға мүмкіндік береді. Мұндай корпус тілдің когнитивтік бейнесін, мәдени кодтарын және коммуникативтік нормаларын айқындауда сенімді база бола алады. Керісінше, репрезентативтілігі төмен корпус тек кездейсоқ деректер жиынтығы болып, тілдің толық бейнесін бере алмайды. Ондай жағдайда тілдің әлеуметтік стратификациясы, диахронды дамуы, функционалды регистрлері бұрмаланып көрінеді; зерттеу нәтижелері жалпылама сипат алады және әдіснамалық тұрғыдан

сенімсіз болады. Сондықтан репрезентативтілік — корпус құрастырудың формалды талабы ғана емес [23, 629], оның ғылыми жарамдылығы мен нәтижелілігін қамтамасыз ететін басты шарт. Репрезентативті корпус тілдің нақты қолданысын бейнелеп, тіл білімінің түрлі салалары — морфология, синтаксис, лексикология, стилистика, когнитивтік және әлеуметтік лингвистика бағыттарындағы зерттеулер үшін эмпирикалық негіз болады. Сонымен қатар ол білім беру, оқулық пен сөздік жасау, тіл саясатын жоспарлау ісінде де стратегиялық құралға айналады. Демек репрезентативтілік — тілдің жан-жақты сипатын дәл көрсететін, ғылыми талдау мен тәжірибелік қолданысты байланыстыратын негізгі категория [24; 15-16]. Тек осы қағидат толық орындалған жағдайда ғана корпус нағыз ғылыми ресурс ретінде қызмет ете алады.

Өзінің қалыптасу кезеңінде ҚТҰК репрезентативтілік тұрғысынан мынадай бастапқы тұжырым жасауға мүмкіндік береді: қазіргі даму сатысында корпус жанрлық және стильдік қамту жағынан едәуір жүйеленген, ішкорпустар арқылы тілдің функционалдық қабаттарын көрсетуге бағытталған құрылымға ие, сонымен қатар негізгі корпус пен арнайы ішкорпустардың болуы қазақ тілінің әртүрлі стильдік тармақтарын, тарихи кезеңдерін және ішінара әлеуметтік-аймақтық ерекшеліктерін қамтуға жағдай жасайды.

Репрезентативтілік талаптарының барлығы толық деңгейде қамтамасыз етілді деуге әлі де ерте. Әсіресе Корпусты регистрлік теңгерім, әлеуметтік стратификацияның тереңдігі, диахронды қабаттардың пропорционалдығы мен аймақтық метадеректердің ауқымдылығы бағытында жетілдіру қажеттілігі байқалады. Дегенмен Корпус көлемінің ұлғаюы мен мәтіндік базаның кеңеюі сапалық теңгеріммен ұштасқанда ғана оның ғылыми әлеуеті арта түсетінін ескеріп, көлемі ұлғайған сайын ҚТҰК репрезентативтілігі де уақыт өте артатынына сенім бар. Ал Корпус өзі құрылған 2022 жылдан бергі аз уақытта, яғни алғашқы кезеңде репрезентативтілік қағидаттарын институционалдық деңгейде негіздеген, құрылымдық моделі айқындалған ұлттық цифрлық ресурс ретінде бағаланады. Алдағы уақытта стратификациялық параметрлерді тереңдету мен метадеректік жүйені жетілдіру оның тілдің толыққанды эмпирикалық бейнесін ұсынатын жоғары деңгейдегі ғылыми корпусқа айналуына мүмкіндік береді.

Зерттеу жұмысы BR24993244 «Қазақ тілі ұлттық корпусын Smart-мәтіндер мегажобасы және қазақтілді жасанды интеллект негізі ретінде жетілдіру, ішкорпустарын әзірлеу» ғылыми бағдарламасын орындау аясында жазылған.

Әдебиеттер тізімі

1 Meyer C.F. English Corpus Linguistics: An Introduction / C.F. Meyer. — Cambridge: Cambridge University Press, 2004. — 168 p.

2 Sinclair J. Corpus, Concordance, Collocation / J. Sinclair, T. McEnery. — Oxford: Oxford University Press, 2021. — 320 p.

3 Қазақ тілінің ұлттық корпусын әзірлеу тәжірибесі (1-кезең) / ред. А. Фазылжан. — Алматы: ЖК Асыл, 2023. — 446 б.

4 Biber D. Representativeness in corpus design / D. Biber // Literary and Linguistic Computing. — 2023. — Vol. 8, No. 4. — P. 243–257.

5 Захаров В.П. Корпусная лингвистика: учебно-методическое пособие / В.П. Захаров. — Санкт-Петербург: Санкт-Петербургский государственный университет, 2005. — 168 с.

6 Egbert J. Corpus-Based Applied Linguistics / J. Egbert, B. Gray. — Cambridge: Cambridge University Press, 2022. — 256 p.

7 Kilgarriff A. The Sketch Engine: Ten years on / A. Kilgarriff, V. Baisa, J. Bušta, M. Jakubíček, V. Kovář, J. Michelfeit, P. Rychlý, V. Suchomel // Lexicography. — 2014. — Vol. 1, No. 1. — P. 7–36.

8 Пименова М.В. Репрезентативность корпуса: функционально-стилистический аспект / М.В. Пименова // Вестник Кемеровского государственного университета. — 2010. — № 2. — С. 118–123.

9 Жұбанов А.Қ. Тәуелсіздік құндылықтарының бірі ретіндегі қазақ тілінің мәтіндер корпусын жасаудың теориялық негіздері / А.Қ. Жұбанов // Мемлекеттік тіл — тәуелсіздік кепілі. — Алматы: Дайк-Пресс, 2011. — 45–81-б.

10 Пірманова К.Қ. Ұлттық корпустарға негізделген лингвистикалық зерттеулер жүргізу / К.Қ. Пірманова, А.Ә. Жаңабекова, А. Барменқұлова // Қазақ ұлттық университетінің хабаршысы. Филология сериясы. — 2022. — № 3. — 83–93-б.

11 Leech G. The British National Corpus Users' Reference Guide / G. Leech, L. Burnard. — London: British Library, 2020. — 320 p.

12 Сүлейменова Э.Д. Қазақ тілі үшін ұлттық корпус керек пе? / Э.Д. Сүлейменова // Тілтаным. — 2015. — № 2. — 42–45-б.

13 Жұбаева О.С. Қазақ тілінің ұлттық корпусын түзуде қойылатын талаптар / О.С. Жұбаева // «Ақпараттық инновациялық парадигмасының қалыптасуы жағдайындағы Қазақстанның қоғамдық ғылымдары» атты Халықаралық ғылыми конференцияның материалдары. — Алматы, 2012. — 485–491-б.

14 Жаңабекова А.Ә. Ұлттық корпус дегеніміз не? / А.Ә. Жаңабекова // «Мәдениеттер тоғысындағы тіл, әдебиет, аударма және журналистика мәселелері» атты Халықаралық ғылыми-практикалық конференция материалдары. — Алматы: ҚазҰУ, 2012. — Т. 1. — 57–61-б.

15 Ляшевская О.Н. Репрезентативность и стратификация в корпусной лингвистике / О.Н. Ляшевская // Вопросы языкознания. — 2015. — № 4. — С. 40–52.

16 Плунгян В.А. Национальные корпуса как инструмент межстилевого анализа / В.А. Плунгян // Филологические науки. — 2018. — № 6. — С. 19–24.

17 Барменқұлова А. Мәтіндер корпусына морфологиялық аннотация жасаудағы статистикалық мәліметтер алу мүмкіндіктері / А. Барменқұлова // «Қазақ тілін оқытудың заманауи әдіснамасы: үдеріс, сапа, жетістік» атты ғылыми-практикалық конференция материалдары. — Алматы: Елтаным, 2012. — 232–236-б.

18 Labov W. Sociolinguistic Patterns / W. Labov, P. Trudgill. — Philadelphia: University of Pennsylvania Press, 2022. — 344 p.

19 Сейітбекова А.А. Тарихи поэтикалық ішкорпус: ескі қазақ поэтикалық мәтіндер базасы / А.А. Сейітбекова, Н. Елесбай // Тілтаным. — 2024. — № 3 (95). — 140–150-б.

20 Фазылжанова А. Национальный корпус казахского языка как инновационно-информационная база государственного языка (об особенностях и задачах первого этапа разработки и создания) / А. Фазылжанова // Alatoo Academic Studies. — 2016. — № 3. — С. 103–109.

21 Сыздықова Р. Қазақ әдеби тілінің тарихы / Р. Сыздықова. — Алматы: Ана тілі, 1993. — 320 б.

22 Малов С.Е. К истории казахского языка / С.Е. Малов // Известия Академии наук СССР. Отделение литературы и языка. — 1941. — № 3. — С. 97-98.

23 Жұбанов А.Қ. Қолданбалы тіл білімі мәселелері / А.Қ. Жұбанов. — Алматы: Арыс, 2008. — 647 б.

24 Плунгян В.А. Корпус как идеология: о некоторых уроках современной корпусной лингвистики / В.А. Плунгян // Русский язык в научном освещении. — 2008. — № 2. — С. 7–20.

А.М. Фазылжан, Л.Р. Койшыгулова, Ж.К. Умирбекова

Требования к репрезентативности корпуса и их реализация в Национальном корпусе казахского языка

В статье рассматривается одна из актуальных проблем корпусной лингвистики — репрезентативность языкового корпуса в теоретическом и практическом аспектах. Уточняется научное содержание понятия репрезентативности, систематизируются основные требования к ней, а также оценивается степень их реализации на материале Национального корпуса казахского языка. Языковой корпус определяется как структурированная совокупность электронных текстов, сформированная для исследовательских целей и снабжённая метаданными и лингвистической аннотацией; он интерпретируется как ресурс, отражающий динамику функционирования языка в реальном употреблении. Репрезентативность рассматривается как ключевой качественный параметр, обеспечивающий научную валидность корпуса, адекватность моделирования языковой системы и достоверность результатов исследований. В работе использованы методы анализа, описания и сопоставления. На основе обобщения теоретических подходов в корпусной лингвистике выделены восемь критериев репрезентативности: жанрово-стилевое разнообразие, баланс стилей и регистров, региональная представленность, диахроническое покрытие, социальная стратификация, достаточный объём текстов, наличие метаданных и документации, а также аутентичность материала. Результаты анализа показывают, что структура Национального корпуса казахского языка ориентирована на репрезентацию функциональных слоёв языка. Основной корпус и специализированные подкорпуса включают художественные, публицистические, научные, официально-деловые, устные, исторические, поэтические и терминологические тексты. Вместе с тем выявлены ограничения, в частности, недостаточная сбалансированность регистров и неполная представленность социальных метаданных, что указывает на необходимость расширения диахронных и региональных компонентов. Обосновывается, что репрезентативность следует рассматривать не только как формальную характеристику языкового корпуса, но и как фундаментальный критерий, определяющий его научную валидность, практическую релевантность и способность адекватно моделировать реальное функционирование языковой системы. Репрезентативный корпус выступает в качестве надёжной эмпирической базы для проведения лингвистических исследований, разработки лексикографических ресурсов, реализации образовательных задач и осуществления научно обоснованной языковой политики. Устанавливается, что Национальный корпус казахского языка в значительной степени представляет собой структурированную цифровую систему, организованную с учётом жанрово-стилевой дифференциации текстов. Вместе с тем аргументируется необходимость его дальнейшего развития за счёт включения текстов современной коммуникации, в том числе материалов социальных сетей, чат-

дискурса, молодежной речи, а также текстов, отражающих новые научные направления. Подчеркивается, что количественное расширение корпуса способствует росту его научного потенциала лишь при условии сохранения внутреннего качественного баланса. Развитие и детализация системы метаданных, а также углубление стратификационных параметров рассматриваются как ключевые факторы формирования высокоуровневого корпуса, обеспечивающего репрезентативное и всестороннее эмпирическое описание казахского языка и уточнение методологических принципов его дальнейшего изучения.

Ключевые слова: репрезентативность, баланс, стратификация, метаданные, регистр.

A.M. Fazylzhan, L.R. Koishygulova, Zh.K. Omirbekova

Requirements for the representativeness of a corpus and their implementation in the National Corpus of the Kazakh Language

The article examines a key issue in corpus linguistics — the representativeness of a linguistic corpus — from theoretical and practical perspectives. The study clarifies the scientific meaning of representativeness, systematizes its main requirements, and analyzes their implementation in the National Corpus of the Kazakh Language. A linguistic corpus is defined as a collection of electronic texts organized for specific purposes, enriched with metadata and linguistic annotations, and serving as a source reflecting the dynamic, actual use of the language. Representativeness is considered a core qualitative parameter ensuring the scientific value of the corpus, the accurate depiction of the linguistic system, and the reliability of research outcomes. The research applied analytical, descriptive, and comparative methods. Based on a review of corpus linguistics scholarship, eight primary requirements for corpus representativeness were identified: diversity of styles and genres, stylistic and register balance, regional coverage, diachrony, social stratification, sufficient text volume, metadata and documentation, and textual accuracy. The study shows that the National Corpus of the Kazakh Language is structured to cover the language's functional layers. Through the main corpus and specialized subcorpora, literary, journalistic, scientific, official-business, oral, historical, poetic, and terminological texts are systematized. However, some issues were identified, including incomplete register balance, inconsistent social metadata, and underdeveloped diachronic and regional layers. Representativeness is thus not merely a formal feature but a principal criterion defining a corpus's scientific validity, practical relevance, and capacity to reflect real language use. A representative corpus provides a reliable empirical basis for linguistic research, lexicography, education, and evidence-based language policy. The article notes that while the National Corpus is a relatively well-organized digital resource in terms of genre and style, it requires more contemporary communication texts, including social media, chat language, youth language, and materials from emerging scientific fields. Expanding corpus size alongside qualitative balance, deepening metadata, and extending stratification parameters will allow the creation of a high-level scientific corpus offering a comprehensive empirical representation of Kazakh. These conclusions strengthen the theoretical foundations of corpus linguistics and refine the methodological principles for national corpus development.

Keywords: representativeness, balance, stratification, metadata, register.

References

1 Meyer, C.F. (2004). *English Corpus Linguistics: An Introduction*. Cambridge: Cambridge University Press.

2 Sinclair, J., & McEnery, T. (2021). *Corpus, Concordance, Collocation*. Oxford: Oxford University Press.

3 Fazylzhan, A. (Ed.). (2023). *Qazaq tilinin ulttyq korpusyn azirleu tazhiribesi (1-kezen)* [Experience of developing the national corpus of the Kazakh language (stage 1)]. Almaty: ZhK Asyl [in Kazakh].

4 Biber, D. (2023). Representativeness in corpus design. *Literary and Linguistic Computing*, *8*(4), 243–257.

5 Zakharov, V.P. (2005). *Korpusnaia lingvistika: Uchebno-metodicheskoe posobie* [Corpus linguistics: A teaching manual]. Sankt-Peterburg: Sankt-Peterburgskii gosudarstvennyi universitet [in Russian].

6 Egbert, J., & Gray, B. (2022). *Corpus-Based Applied Linguistics*. Cambridge: Cambridge University Press.

7 Kilgarriff, A., Baisa, V., Bušta, J., Jakubíček, M., Kovář, V., Michelfeit, J., Rychlý, P., & Suchomel, V. (2014). The Sketch Engine: Ten years on. *Lexicography*, *1*(1), 7–36.

8 Pimenova, M.V. (2010). Reprezentativnost korpusa: funktsionalno-stilisticheskii aspekt [Representativeness of the corpus: functional and stylistic aspect]. *Vestnik Kemerovskogo gosudarstvennogo universiteta — Bulletin of Kemerovo State University*, 2, 118–123 [in Russian].

9 Zhubanov, A.K. (2011). Tauelsizdik qundylyqtarynyn biri retindegi qazaq tilinin matinder korpusyn zhasaudyn teoriialyq negizderi [Theoretical foundations of creating a text corpus of the Kazakh language as one of the values of independence]. *Memlekettik til – tauelsizdik kepili — State language – the guarantee of independence* (pp. 45–81). Almaty: Daik-Press [in Kazakh].

10 Pirmanova, K.K., Zhanabekova, A.A., & Barmenqulova, A. (2022). Ulttyq korpustarga negizdelgen lingvistikalyq zertteuler zhurgizu [Conducting linguistic research based on national corpora]. *Qazaq Ulttyq Universitetinin Khabarshysy. Filologia seriasy — Bulletin of the Kazakh National University, Philogogy, 3,* 83–93 [in Kazakh].

11 Leech, G., & Burnard, L. (2020). *The British National Corpus Users' Reference Guide*. London: British Library.

12 Suleimenova, E.D. (2015). Qazaq tili ushin ulttyq korpus kerek pe? [Is a national corpus needed for the Kazakh language?]. *Tiltanym — Linguistics*, 2, 42–45 [in Kazakh].

13 Zhubaeva, O.S. (2012). Qazaq tilinin ulttyq korpusyn tuzude qoiylatyn talaptar [Requirements for the formation of the national corpus of the Kazakh language]. «*Aqparattyq innovatsiialyq paradigmasynyn qalyptasuy zhagdaiynda Qazaqstannyn qogamdyq gylymdary» atty Khalyqaralyq gylymi konferentsianyn materialdary — Proceedings of the International Scientific Conference "Social sciences of Kazakhstan in the conditions of formation of information innovation paradigm"* (pp. 485–491). Almaty [in Kazakh].

14 Zhanabekova, A.A. (2012). Ulttyq korpus degenimiz ne? [What is a national corpus?]. «*Madenietter togysyndagy til, adebiet, audarma zhane zhurnalistika maseleleri» atty Khalyqaralyq gylymi-praktikalyq konferentsia materialdary — Proceedings of the International Scientific and Practical Conference "Issues of language, literature, translation and journalism at the crossroads of cultures"* (Vol. 1, pp. 57–61). Almaty: Qazaq ulttyq universiteti [in Kazakh].

15 Liashevskaia, O.N. (2015). Reprezentativnost i stratifikatsiia v korpusnoi lingvistike [Representativeness and stratification in corpus linguistics]. *Voprosy yazykoznaniia — Topics in the Study of Language*, 4, 40–52 [in Russian].

16 Plungian, V.A. (2018). Natsionalnye korpusa kak instrument mezhstilevogo analiza [National corpora as a tool for inter-style analysis]. *Filologicheskie nauki — Philological Sciences*, 6, 19–24 [in Russian].

17 Barmenqulova, A. (2012). Matinder korpusyna morfologiialyq annotatsiia zhasaudagy statistikalyq malimetter alu mumkindikteri [Possibilities of obtaining statistical data in morphological annotation of text corpora]. «*Qazaq tilin oqytudyn zamanaui adisnamasy: uderis, sapa, zhetistik» atty gylymi-praktikalyq konferentsia materialdary — Materials of the scientific and practical conference "Modern methodology of teaching the Kazakh language: process, quality, achievement"* (pp. 232–236). Almaty: Eltanym [in Kazakh].

18 Labov, W., & Trudgill, P. (2022). *Sociolinguistic Patterns*. Philadelphia: University of Pennsylvania Press.

19 Seiitbekova, A.A., & Elesbai, N. (2024). Tarikhi poetikalyq ishkorpus: eski qazaq poetikalyq matinder bazasy [Historical poetic subcorpus: a database of old Kazakh poetic texts]. *Tiltanym — Linguistics*, 3(95), 140–150 [in Kazakh].

20 Fazylzhanova, A. (2016). Natsionalnyi korpus kazakhskogo yazyka kak innovatsionno-informatsionnaia baza gosudarstvennogo yazyka (ob osobennostiakh i zadachakh pervogo etapa razrabotki i sozdaniia) [National corpus of the Kazakh language as an innovative information base of the state language (on the features and tasks of the first stage of development and creation)]. *Alatoo Academic Studies*, 3, 103–109 [in Russian].

21 Syzdykova, R. (1993). *Qazaq adebi tilinin tarikhy* [History of the Kazakh literary language]. Almaty: Ana tili [in Kazakh].

22 Malov, S.E. (1941). K istorii kazakhskogo yazyka [On the history of the Kazakh language]. *Izvestiia Akademii nauk SSSR. Otdelenie literatury i yazyka — Proceedings of the Academy of Sciences of the USSR. Department of Literature and Language*, 3, 97-98 [in Russian].

23 Zhubanov, A.K. (2008). *Qoldanbaly til bilimi maseleleri* [Issues of applied linguistics]. Almaty: Arys [in Kazakh].

24 Plungian, V.A. (2008). Korpus kak ideologiia: o nekotorykh urokakh sovremennoi korpusnoi lingvistiki [Corpus as ideology: on some lessons of modern corpus linguistics]. *Russkii yazyk v nauchnom osveshchenii —Russian Language and Linguistic Theory*, 2, 7–20 [in Russian].

Information about the authors

Fazylzhan, Anar — Candidate of Philological Sciences, Associate Professor, A. Baitursynuly Institute of Linguistics, Almaty, Kazakhstan. E-mail: afazyljan20@gmail.com; ORCID: https://orcid.org/0000-0002-0562-4846

Koishygulova, Laura Rakhymzhanovna — PhD student, L.N. Gumilyov Eurasian National University, Astana, Kazakhstan. E-mail: rahimjan2305@mail.ru; ORCID: https://orcid.org/0009-0005-4340-0308

Omirbekova, Zhamilya Kaldibekovna — Candidate of Philological Sciences, Associate Professor, L.N. Gumilyov Eurasian National University, Astana, Kazakhstan. E-mail: ozhamila@inbox.ru; ORCID: https://orcid.org/0000-0002-3162-0515