



ХҒТАР 16.31.21
Ғылыми мақала

<https://doi.org/10.32523/2616-678X-2025-152-3-11-27>

Қазақ фольклорлық мәтіндерінің корпусы: ғылыми талдау және зерттеу мүмкіндіктері

Н.Н. Аитова^{1*}, М. Бақытқызы²

Л.Н. Гумилев атындағы Еуразия ұлттық университеті, Астана, Қазақстан

(E-mail: ¹nurlykhan.an@gmail.com; ²0002992@mail.ru)

Аңдатпа. Гуманитарлық саланы цифрландыруды көздейтін мемлекетіміздің негізгі бағдарына сәйкес соңғы жылдары мәтіндік базалар құруға деген талпыныс артуда. Бұл шын мәнінде әлемдік маңызы бар қазіргі тенденциялар қатарында танылады. Корпустық ізденістердің, атап айтқанда, табиғи тілді модельдеу, тілдік алгоритмдер түзу, мәтін өңдеудің қазіргі лингвистиканың басым бағытына айналуы шынайы қажеттіліктерден туындап отыр. Бұл мақалада «Қазақ фольклорлық мәтіндерінің корпусын әзірлеудің мәтінді ғылыми талдау және зерттеуге беретін мүмкіндіктері қарастырылады. Жұмыстың негізгі бағыттары мәтіндерді жинақтау, цифрлық ортаға енгізу, морфологиялық және семантикалық таңбалау, конкорданс құру және тілдік бірліктерді статистикалық талдау әдістерін қамтиды. Зерттеудің ғылыми маңызы – қазақ фольклорын жаңа деңгейде зерделеуге, оның мәтіндік құрылымын заманауи әдіснама арқылы талдауға жол ашатын лингвистикалық корпус базасын қалыптастырудың теориялық негіздемесін ұсынып, тілдің бай қабаттарын ғылыми айналымға енгізуді көздеуінде. Негізгі нәтижелер ретінде қазақ фольклорлық мәтіндерінің корпустық моделі екшеліп, тілдік деректерді кешенді талдау мүмкіндіктері көрсетіледі. Бұл жұмыс корпус лингвистикасының теориясы мен практикасын ұлттық фольклорға бейімдеу арқылы қазақ тіл білімінде жаңа зерттеу кеңістігін ашады. Жүргізілген зерттеу фольклор мәтіндерінің тілдік табиғатын жан-жақты сипаттап, қазақ фольклортану ғылымына әдіснамалық үлес қосады.

Түйін сөздер: қазақ фольклоры, мәтіндік корпус, мәтін, корпустық талдау, корпустық лингвистика.

Түсті: 07.08.2025; Жөнделді: 19.08.2025; Мақұлданды: 17.09.2025; Онлайн қолжетімді: 25.09.2025

*хат-хабар үшін автор

Кіріспе

Қазақ тілінің мәтіндік корпустары соңғы жылдары еліміздегі түрлі ғылыми ұйымдар мен жекелеген жобалардың аясында жасалып келеді. Дегенмен, фольклорлық мәтіндерге арналған арнайы корпус әзірге Қазақстанда қолға алынған жоқ. Ауыз әдебиетінің аса бай қоры бар көп томдық «Бабалар сөзі» сериясын (Бабалар сөзі, 2013) корпусқа енгізу ғылыми тұрғыдан талай зерттеудің тиімді жүргізілуіне сеп болары анық. Ал оны мүмкіндік шегінде параллель корпус ретінде түзсе, тіпті зор мәнге ие болатыны түсінікті. Қазақ фольклоры ұлттық тілдік-мәдени қазынаны сақтап, ұрпақтан ұрпаққа жеткізген мол мұра ретінде кең көлемде зерттелгенімен, мәтіндерді корпустық үлгіде жүйелеу мен ғылыми талдау ісі әлі кенже қалып келеді. Оған корпустың жасалмауы ғана емес, зерттеушілер арасында жалпы корпус материалдарын зерттеу үшін қолданылуының аздығы, тіпті корпус базаларының функционалдығының кемшіндігі де әсер етеді деп ойлаймыз.

Әлемдік лингвистикада табиғи тілді модельдеу, тілдік алгоритмдер құру және мәтінді автоматты өңдеу бағыттары ғылыми зерттеулердің жаңа деңгейін көрсетуде. Қазақстан ғылымында да бұл үрдіс қарқын алып, ұлттық мұра үлгілерін, соның ішінде фольклорлық мәтіндерді заманауи әдістермен ғылыми айналымға енгізу қажеттілігі артып отыр.

Бұған дейінгі Ш.Уәлиханов, Ә. Марғұлан, Ә. Диваев, М. Әуезов, С. Қасқабасов, Р.Бердібаев, Ә.Қоңыратбаев, Ж. Жақыпов және т.б. ғалымдардың еңбектерінде қазақ фольклоры әдебиеттанулық, этнографиялық немесе мәтінтану тұрғысынан қарастырылғаны, бұл тұрғыдан қазақ фольклорының зерттелуі ілгері түрғаны белгілі. Ал корпус лингвистикасының әдістерін қолдана отырып кешенді зерттеу – отандық ғылыми кеңістігіндегі тың бағыт.

Зерттеу мақсаты мен міндеттері

Зерттеу мақсаты – қазақ фольклорлық мәтіндерін корпус лингвистикасы әдістері негізінде жүйелеу және олардың тілдік, құрылымдық, семантикалық, стилистикалық қырларын айқындау. Мұндай ұстаным зерттеу тезисін нақтылайды: фольклорды корпустық талдау – ұлттық мұраның тілдік болмысын жаңа қырынан танудың пәрменді құралы.

Зерттеудің мақсатына сәйкес мынадай міндеттер алға қойылды: – әлемдік тәжірибедегі фольклорлық мәтіндер корпустарына шолу жасау; – қазақ фольклорлық мәтіндерін жинақтап, цифрлық ортаға енгізу амал-тәсілдерін сипаттау; – фольклорлық мәтіндерді корпустық тұрғыдан зерттеудің ғылыми маңызын негіздеу; – тілдік бірліктерді статистикалық және корпустық талдаудың үлгілерін ұсыну.

Зерттеу материалы мен әдістері

Фольклорлық мәтіндерді цифрландырып, корпус базасына енгізу, оларды ғылыми ортаға қолжетімді ету және білім беру саласына бейімдеу біршама жұмысты талап етеді. Себебі бұл үдеріс тек лингвистика саласымен шектелмей, фольклортану мен ақпараттық технологиялар ғылымдарымен өзара байланыста жүзеге асады. Сондықтан фольклорлық мәтіндер корпусын құру пәнаралық зерттеудің айқын көрінісі болып табылады.

Жұмыс барысында төмендегідей әдістер қолданылды: сапалық және сандық талдау әдістері (мәтіндік деректерді жиілік, коллокация, құрылымдық ерекшеліктері тұрғысынан қарастыру); ақпаратты өңдеу әдістері (мәтіндерді жинақтап, цифрлық ортаға енгізу, морфологиялық, семантикалық және құрылымдық тұрғыдан кодтау); жинақтау әдісі (фольклорлық мәтіндерді бірізділікке түсіріп, жүйелі корпус базасын қалыптастыру).

Зерттеу гипотезасы. Қазақ фольклорлық мәтіндерін корпустық негізде жүйелеу олардың тілдік табиғатын жаңа деңгейде зерделеуге мүмкіндік береді және ұлттық мұраны цифрлық-гуманитарлық ғылымдар кеңістігінде жаңа қырынан танытады. Демек, фольклорды корпус арқылы талдау дәстүрлі фольклортанулық зерттеулерді заманауи лингвистикалық және ІТ құралдарымен ұштастыру арқылы жаңа ғылыми нәтижелерге қол жеткізуге жағдай жасайды.

Зерттеу қазақ фольклорын цифрландыру арқылы ұлттық мұраны жаңа ғылыми айналымға енгізуді және корпус лингвистикасы мен лингвофольклортану тоғысында жаңа әдіснамалық негіз қалыптастыруды көздейді. Жұмыс нәтижелері қазақ тіл білімінің теориялық базасын толықтырып қана қоймай, фольклортану ғылымына да қосымша үлес қосады. Сонымен қатар, зерттеу нәтижелері қазақ тілінің ақпараттық технологиялардағы қолданысын (NLP, жасанды интеллект, машиналық аударма) кеңейтеді, білім беру саласында тіл үйрету мен фольклорды меңгертудің тиімді құралына айналады және ұлттық мұраны мәдени-танымдық әрі цифрлық мұрағат ретінде сақтауға мүмкіндік береді, фольклорды ұлттық тілдің байлығын айқындайтын электрондық ресурс ретінде ғылыми айналымға енгізуді қолдайды.

Әдебиеттерге шолу

Фольклорлық мәтіндердің корпустық базасын құруда, әсіресе, батыс елдерінде біршама тәжірибе бар. Мұндай мәтіндік корпустар өздерінің мақсаты мен құрылымы бойынша әр алуан. Солардың ішінде кейбір ірі жобалардың бірі Финляндияда жасалған “Suomen kansan vanhat runot (SKVR, The Ancient Songs of the Finnish People” корпусы (Suomen kansan vanhat runot, n.d.). Бұл – дәстүрлі фин халық поэзиясын қамтитын ауқымы кең архив. Корпус мақсаты – фин халқының поэтикалық мұраларын сақтау және зерттеу жұмыстарын оңайландыру. SKVR (Suomen kansan vanhat runot) дерекқоры 34 томдық басылым (1908-1948, 1997) негізінде шамамен 89 мың поэтикалық мәтінді цифрландыру арқылы жасалып, 1998 жылдан бастап Фин әдебиет қоғамының бастамасымен ашық қолжетімді инфрақұрылымға айналған, Жобаның елеулі артықшылығы жанрлық және географиялық метадеректерді жүйелі тіркеуі мен зерттеуге ыңғайлы интерфейстерінде.

Корпус базасы негізінде жасалған зерттеулерінде Lauri Harvilahti (2013) SKVR (Suomen Kansan Vanhat Runot (фин. «Фин халқының көне жырлары») дерекқорының құрылымын, қамтылым көлемін және Semantic Kalevala («Калевала» – фин халқының ұлттық эпосы) жобасымен байланысын сипаттайды, сондай-ақ, бұл модель мағыналық-онтологиялық индекстеу арқылы тақырыптық шолу ұсынады (Harvilahti, 2013). Зерттеу – корпус инфрақұрылымы мен оны қолдану әдістемесіне арналған іргелі жұмыс, корпус деректерін қолданудың бір қырын көрсетеді. Ал зерттеушілер Maciej Janicki, Kati Kallio & Mari Sarv

(2023) болса, SKVR (және онымен сабақтас фин-эстон жыр базалары) негізіндегі калевала өлшеміндегі өлең жолдарының автоматты эквивалентін айқындау әдісін ұсынды (жол – мағына – сөз сәйкесімін есептеу), нәтижесінде интертекстуалдық картографиялау жүзеге асырылған; IT шешімі SKVR көлеміндегі деректерге қолданылып, зерттеушіге мәтінаралық ұқсастықтарды жүйелі шолуға мүмкіндік береді (Janicki et al., 2023). Мақалада зерттеу интерфейсі ретінде Runoregi (SKVR-мен бірге қолданылатын фин жыр дереккөздері) қолданылғаны да көрсетіледі. Runoregi – Финляндиядағы халық поэзиясы мен жырларын цифрлық ортада іздеуге және зерттеуге арналған онлайн каталог пен деректер қоры. Осы базаны қолданған зерттеулердің тағы бірі – Tuukka Karlsson (2021) еңбектері. Калевала-өлшемді арбау мәтіндеріндегі регистрлік ерекшеліктерді Виена Карелиясы материалдары негізінде талдайды (Karlsson, 2021). Мәтін базасы ретінде фин-карел дереккөздері, соның ішінде SKVR аясындағы материалдар пайдаланылған. Жанр/регистр ұғымдары арқылы формулалық құрылымдар мен дискурстық қызметті сипаттайды. Осы жерде нақтылай кетсек, фольклорлық мәтіндерді корпусқа енгізу олардың құрылымдық табиғатын жаңа қырынан тануына мүмкіндік беретінін байқаймыз. Корпустық технологиялар мәтінді жанр және регистр бойынша таңбалауға, ал формулалық құрылымдарды автоматты түрде іздеуге жағдай жасайды. Конкорданс құру, жиілік санау, колокациялық талдау сияқты әдістер арқылы қайталанатын тұрақты тіркестерді сүзіп, олардың қолданылу өрісін анықтаған зерттеулер бұған дәлел.

Келесі бірі – Норвегиядағы «The Norwegian folklore archives» фольклорлық мәтіндерді жинау, сақтау және зерттеуге арналған ұлттық архив. Бұл корпустық мұрағат негізінен XIX мен XX ғасырларда жинақталған халық поэзиясы, аңыздар мен балладалардан тұрады. Басты ерекшелігі – халық мәдени мұрасын жинау мен сақтауға бағытталуы (The Norwegian folklore archives, 2025).

Әртүрлі тілдердегі халық ертегілерінің мәтіндерін қамтитын көптілді мәліметтер қоры ретінде әзірленген «The Multilingual Folk Tale Database» (MFTD) корпусы да жұртшылыққа белгілі. Бұл база да халық ертегілерінің толық мәтіндерін іздеуге қолайлы, негізінен халық мұрасын көпшілікке қолжетімді етіп, зерттеу мүмкіндіктерін кеңейту мақсатын көздейді (The Multilingual Folk Tale Database, 2025). Бұдан өзге Америка фольклор орталығы АҚШ (American Folklife Center, 2025), Тыва тілінің электронды корпусы (Ресей) (Тыва тілінің электронды корпусы, 2025), Эстония, Қытай т.б. елдердің фольклорлық мәтіндер базалары танымал.

Әрбір архив өзінің мақсаты мен бағыттарына қарай ерекшеленеді. Кейбіреуі тек фольклорлық мәтіндермен шектелсе, кейбірі фольклорды кеңірек зерттеумен айналысады және олардың базасы этнологиялық материалдарды да қамтиды. Бұл фольклорлық корпустар қазіргі таңда сол тілдердегі мәтіндік зерттеулердің қарқынды дамуына, тілдік мұраларды білім беруге пайдалану мен насихаттауда кеңінен пайдалануға зор ықпал етіп отыр.

Сонымен қатар, корпустық платформада әр мәтінге байланысты метадеректердің тіркелуі (жазылып алынған уақыты, орны, орындаушысы, жанры) зерттеушілерге бір ғана формуланың әртүрлі регистрлерде қандай дискурстық қызмет атқаратынын салыстыра зерттеуге мүмкіндік береді. Осылайша, корпус тек мәтіндерді сақтаудың құралы емес, олардың формулалық және дискурстық табиғатын жүйелі түрде сипаттаудың заманауи әдісі болып отыр.

Осындай тәжірибелерге сүйене отырып, қазақ тілінің фольклорлық мәтіндер корпусын әзірлеу, көпшіліктің пайдалануына оның интерфейстерін ыңғайлы ету, тілдік мұраларға корпустық әдіспен зерттеу жұмыстарын жүргізу – уақыт талабы әрі қажетті іс.

Нәтижелер мен талқылау

Түркі халықтары ішінде тыва тілінің ғана фольклорлық мәтіндер корпусы жасақталған. Бұл фольклорлық электронды база – тыва тіліндегі эпостар мен фольклорлық мәтіндерге бағытталған арнайы корпус (Тыва тілінің электронды корпусы, 2025). Мұнда морфологиялық талдау, клише-деректер және жанр бойынша іздеу құралдары бар. Ішкорпусқа эпостар, ертегілер, салттық поэзия және тыва қаһармандық эпосы мәтіндері енгізілген. Лексемалар жиілігі мен есім/етістік негіздерінің дерекқоры электронды жиілік сөздік құруға ыңғайлы. Фольклорға тән метадеректері – жанр, нұсқа, жарияланған дерек, формулалық клишелердің тізімі мен сипаттамасы жүйелі берілген. Бұл жоба тілдік материалдарды жүйелі жинақтау, таңбалау және талдау мүмкіндігін ұсына отырып, фольклордың сандық ресурстар арқылы зерттелуіне ықпал етті. Фольклорлық мәтіндер корпусын түзуде түркі тілдері үшін бағдар етуге болатын корпус ретінде өзіндік орны барын атап өтуді орынды деп санаймыз.

Қазақ тілінің ұлттық корпусына қатысты айтар болсақ, жоғарыда атап өткендей, фольклорлық мәтіндер корпусының отандық үлгісін әзірлеу қазір ғана бастау алып отыр. Әзірге бұл бастама ашық қолжетімді деңгейге көшпегенімен, дайындық жұмыстары жүргізілуде. Солардың алғашқы қадамы ретінде ішкі қолданысқа арналған «Ертегілер» электрондық сайты жасалды. Бұл ресурс – фольклорлық мәтіндердің (ең алдымен ертегілердің) базасы болып, бейімделген үлгі сипатында қызмет атқаруда. Аталған базада мәтіндік талдау тәжірибелері үшін қазіргі таңда кеңінен қолданылатын Voyant Tools құралы пайдаланылды. Әрине, әзірленіп жатқан негізгі база корпустың толық сипаттамасына сай аннотацияланады.

Зерттеу барысында біз осы үлгілік базаның алғашқы нәтижелерін дәйектемелік материал ретінде ұсына отырып, фольклорлық мәтіндер корпусын құрудың қажеттілігін дәлелдей түсеміз. Бұл қадам ұлттық мұраны цифрлық-гуманитарлық ғылымдар кеңістігінде жаңа әдіснамалық негізде талдаудың моделін ұсынады.

Аталған ішкі параллель ресурс жобасы базасына 40 ертегінің қазақ-орыс тілдеріндегі мәтіні енгізілді. Ертегі мәтіндерінің тілдік бірліктерінің сандық, сапалық талдамасын автоматты көрсету үлгісі төменде берілді:

1-кесте. Ертегі мәтіндері ұзақтығы мен лексикалық көлемі/ тығыздық көрсеткіші

Санат (категория)	Мәтін атауы	Ұзындығы (сөзқолданыс саны)/ тығыздығы
Ең ұзағы	Ер Төстік	3504
	Қара үйрек	1861
	Керқұла атты Кендебай	1823
	Көкжал	1754
	Тазша бала	1689

Ең қысқасы	Ақыл, ғылым, білім	141
	Сауысқан мен көкек	154
	Қарлығаш пен Дәуіт	159
	Тышқан мен жылан	166
	Уақыт патшалығы	181
Лексикалық тығыздық (ең жоғары)	Елеңке батыр	0.841
	Қоянның айланы	0.801
	Тышқан мен жылан	0.795
	Мысық неліктен тышқан...	0.773
		0.773

Ескерту: деректер негізінде құрастырылған (Ертегілер параллель корпусы, 2025) (ішкі жабық ресурс).

Бұдан өзге, фольклорлық мәтіндерді бір ортақ базада жинақтап, сандық форматта ұсыну зерттеушілерге деректерді жартылай автоматты түрде алу мен талдауды мүмкін етеді. Мұндай тәсіл корпус лингвистикасының стандарт әдістеріне сүйенеді (Baker, 1995). Параллель корпус ретінде ұйымдастыру мәдениеттер тоғысында мәтіндерді салыстырмалы-типологиялық тұрғыдан қарастыруға және аудармалық трансформацияларды тікелей салыстырып зерттеуге жол ашатынын ғалымдар жиі еске салады (McEnery & Hardie, 2012). Техникалық жағынан мәтіндерді бірізді таңбалау үшін TEI-негізді белгіленім қолданылады, зерттеушілердің пайымынша, ол мәтіндердің дерек сапасын, қайта пайдаланылуын және репродуктивтілігін күшейтеді (Ide & Véronis, 1995). Қазақ фольклорлық мәтіндерінің параллель үлгісін жасау тәжірибесі түркі халықтарының басқа да жобаларымен сабақтас келеді. Мәселен, Тыва фольклорлық корпусында эпос мәтіндерінің дерекқоры алғашқы үлгі ретінде ұсынылған болатын (Ондар, 2016).

Фольклорлық мәтіндердің корпусын параллель корпус ретінде жасаудың өзіндік артықшылықтарына тоқтала кетсек, біріншіден, мәдениеттер тоғысы кеңістігінде мәтіндерді зерттеуге мүмкіндік береді. Екіншіден, аудармашыларға көмегі зор, әр тілдік бірліктің аударылу тәсілі мен ерекшеліктерін электронды нұсқада тікелей талдауға, зерттеуге ыңғайлы. Үшіншіден, білім беруде, тіл үйретуде бейімдеп қолдануға жол ашылады. Оның бер жағында лексикалық минимумдарды автоматты түзу, лексикографиялық материалдар әзірлеуде дерек болады. “Ер Төстік” ертегісінің параллель теңестірілген нұсқасы ұсынылды:

Төмендегі суретке қараңыз.

"Ер Төстік"

Search in table

Page 1 of 16

	Ерте заманда Ерназар деген кісі болыпты. Ерназардың өзі бай болыпты. Төрт түлік малы сай болыпты. Қора толған қойлары болыпты. Келе-келе түйелері болыпты. Өріс толған жылқылары болыпты.	Давным-давно жил-был человек по имени Ерназар. Он был очень богат: у него был скот всех четырех видов. Его загоны были полны овцами, позже появились верблюды, а пастбища были забиты табунами лошадей.
1	Ерназардың сегіз ұлы болыпты. Бір жылы үлкен жұт болып, ел малын алысқа, отарға айдап кетіпті. Ерназардың сегіз ұлы да соның ішінде кетіпті. Бір қыстық азығын алып, кемпірі мен Ерназар үйінде қалыпты.	У Ерназара было восемь сыновей. Однажды случилась сильная засуха, и люди увели свой скот далеко на отары. В числе этих людей были и все восемь сыновей Ерназара. Сам Ерназар с женой остались дома, взяв припасы на одну зиму.
2	Ерназардың сегіз ұлы сол кеткеннен хабарсыз кетеді, айлар өтеді, келмейді. Ерназардың азығы таусылады, ашығады, жейтін тамақ таппайды. Кемпірі екеуінің тұруға елдері әрең-әрең келеді.	С тех пор о восьми сыновьях не было вестей: проходили месяцы, а они всё не возвращались. Припасы у Ерназара закончились, они начали голодать, еду добыть не могли. Оба — и он, и его жена — едва держались на ногах.
3	Бір күні кешке жақын кемпірі төсегінен тұрып, үйінің түңдігін ашады. Тұруға мұршасы келмей жатқан шал шаңыраққа қараса, шаңырақтың күлдіреуішінде керулі тұрған кер биенің төстігі көзіне түседі.	Однажды к вечеру жена Ерназара поднялась с постели и приоткрыла верхнее отверстие юрты. Старик, еле-еле приподняв голову, взглянул на купол и увидел в перекрестии верхних жердей кусок грудинки от чалой кобылы.
4	Ерназар қуанып, есі шығады. «Кемпір, сүйінші! Кер биенің керулі тұрған төстігі майлы көрінеді, боп, жылдам асып жіберші!», — дейді. Сөйтіп, бұлар төстікті қазанға салады. Төстікті асып жеп, шал мен кемпір өлденеді.	Ерназар обрадовался и закричал: «Кемпір, радость! Грудинка от нашей чалой кобылы выглядит жирной, скорее свари её!» Они положили грудинку в казан, сварили и, поев её, немного окрепли.
5	Ұзамай кемпір жүкті болады. Мезгілі толып босанады. Бір ұл туады, оның атын төстік жеген соң біткен бала деп Төстік қояды.	Вскоре жена забеременела. Когда пришло её время, она родила мальчика, которого назвали Төстік, потому что он был зачат после съеденной грудинки.
6	Төстік өзгеше бала боп өседі: бір айда бір жастағы баладай, екі айда екі жастағы баладай, үш айда үш жастағы баладай, бір жылда он бестегі жасөспірімдей болып, екі жасында алысқан кісілерін алып ұратын бала болады.	Төстік рос необычным ребёнком: за месяц он становился как годовалый, за два месяца как двухлетний, за три — как трёхлетний ребёнок, а к году достиг силы пятнадцатилетнего подростка. В два года он стал побеждать взрослых мужчин в борьбе.
7	Садақ тартып үйренеді, тартқан садағын тоғыз қабат кетпеннен өткізеді. Аң аулап, құс атып, әкесі мен кемпір шешесін асырайды.	Он научился стрелять из лука, стрела которого могла пробить девять слоёв мотыги. Төстік охотился на зверей и птиц, кормя тем самым отца и мать.

1-сурет. Ертегілердің қазақша-орысша параллель корпусындағы "Ер Төстік" ертегісі мәтінінің теңестірілген нұсқасы

Ескерту: деректер негізінде алынған (Ертегілер параллель корпусы, 2025) (ішкі жабық ресурс).

Кейбір зерттеушілер фольклор тілі арқылы ұлттың құндылық жүйесін зерттеу қазақ халқының дүниетанымдық өзіндік ерекшеліктерін анықтауға мүмкіндік беретінін, осы тұрғыда фольклор халық танымының базалық негізі бола алатынын, соның ішінде оны бүгінгі заман талабымен, бүгінгі үрдіспен байланыстыра талдау ұлттық ғылымымыз бен мәдениетіміздің заманауи бағытта дамуына үлкен әсері ететінін айтады (Балабекова & Отарбекова, 2019). Бұл пікірді корпустық талдауға да қатысты қолдансақ, дұрыс болады.

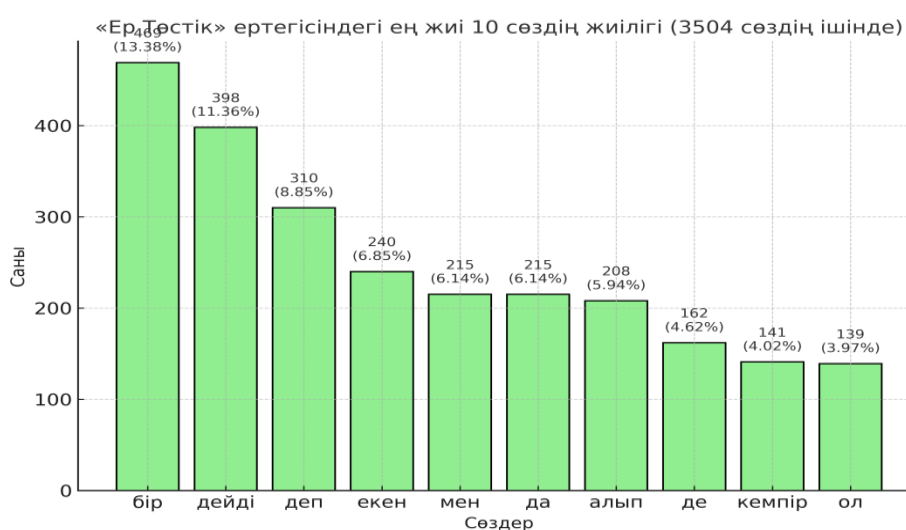
Мәтіндердегі сөздердің жиілігі бойынша мәліметтер де корпус базасындағы талдағыштар негізінде автоматты сүзіп алынады. Кестеде "Ер Төстік" ертегісі мәтініндегі жоғары жиіліктегі алғашқы он сөз берілген.

2-кесте. "Ер Төстік" ертегісіндегі сөздер жиілігі

lemma	freg
бір	469
дейді	398
деп	310
екен	240
мен	215
да	215
алып	208
де	162
кемпір	141
ол	139

Ескерту: деректер негізінде құрастырылған (Ертегілер параллель корпусы, 2025) (ішкі жабық ресурс).

Мұнда ізделген сөздерді кездесу жиілігіне қарай талдай аламыз. Ертегі мәтінінде барлығы 3504 сөз бар. Осыған сүйенсек, жоғарыдағы кестедегі ең жиі кездесетін 10 сөздің үлесі: $2497/3504 \times 100 \approx 70\%$. Демек, бар болғаны 10 сөз ертегі мәтінінің маңызды бөлігін қамтиды. Бұл көрсеткіш – фольклорлық мәтіндерге тән формулалық, қайталамалы сипаттың айқын дәлелі. Ертегідегі «бір», «дейді», «деп» сияқты сөздер мәтіннің құрылымдық өзегін құрап, баяндау стилін тұрақтандырады. Қалған жиі қолданылатын бірліктер («екен», «мен», «да») мәтіндегі грамматикалық және дискурстық байланыстарды қамтамасыз етуге жұмсалады. Бұл сөздердің жалпы мәтін ішіндегі үлес салмағын төмендегі диаграммадан көруге болады.



2-сурет. “Ер Төстік” ертегісіндегі сөздердің жиілік үлесі

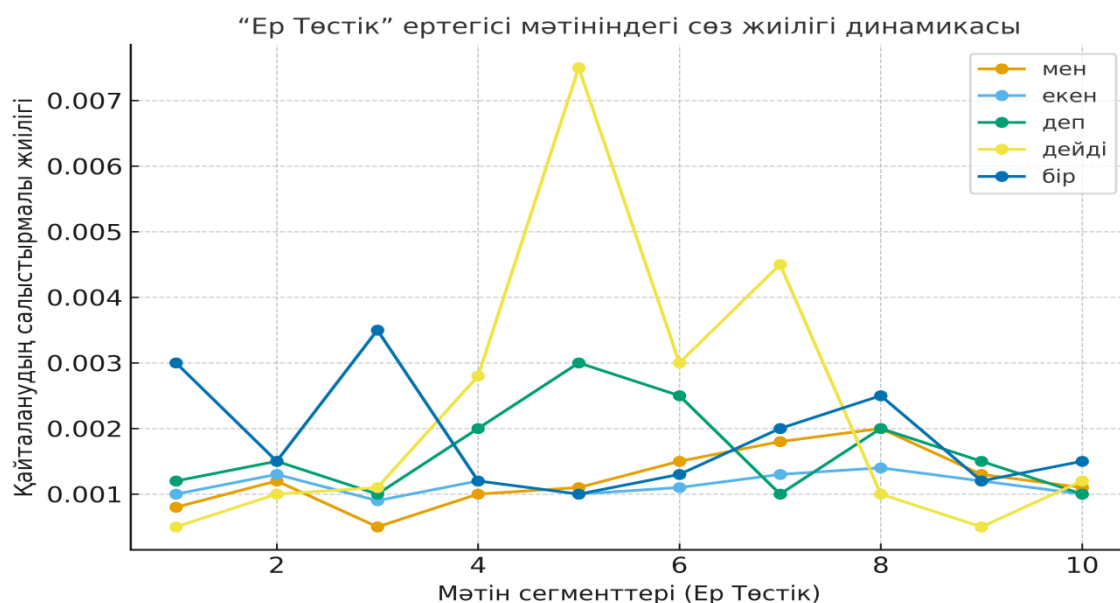
Ескерту: деректер негізінде алынған (Ертегілер параллель корпусы, 2025) (ішкі жабық ресурс).

Бағандық диаграмма нәтижелері «Ер Төстік» ертегісі мәтіндегі сөздік құрамның құрылымдық ерекшеліктерін айқындайды: «бір» (469 рет, 13,4%) – ең жиі қолданылған сөз. Бұл фольклорлық мәтінге тән тұрақты құрылымдық басқы шендегі («Баяғыда бір...», «бір шал...», «бір қыз болыпты...», т.с.с.) баяндау дәстүрін көрсеттеді. «Дейді» (398 рет, 11,36%) мен «деп» (310 рет, 8,85%) – негізгі дискурстық маркерлер. Олар ертегіде диалогтық құрылымды қалыптастырып, баяндаушының сөз беру тәсілін көрсетеді. «Екен» (240 рет, 6,85%) – баяндаудың сенімділігін, нақтылығын арттыратын грамматикалық форма. «Мен» және «да» (әрқайсысы 215 рет, 6,14 %) – мәтін ішіндегі логикалық және синтаксистік байланыстарды қамтамасыз етеді. Қалғандары («алып», «де», «кемпір», «ол») салыстырмалы түрде төмен болғанымен, сюжеттік дамудың, кейіпкер сипаттаудың және оқиғаның динамикасын көрсетудің маңызды құралдары болып табылады.

Диаграмма ертегі мәтінінде қайталанымдылық пен формулалық құрылымның жоғары екенін айқындайды. Бар болғаны 10 сөз мәтіндегі жалпы сөздердің үштен бірін құрап, қазақ фольклорының дискурстық ерекшелігін нақты көрсетеді.

Жалпы қазақ фольклорын тек әдеби мұра ретінде ғана емес, ұлттық кодтар мен мәдени символдарды сақтаушы тілдік кеңістік ретінде қарастыратын еңбектерде фольклорды тек мазмұндық деңгейде емес, тілдің ішкі жүйесі арқылы түсіндіруге ден қояды (Zhakirov et al., 2020). Мұндай зерттеу нәтижелерін корпустық талдаулармен сабақтастыруға болады.

Ал енді “Ер Төстік” ертегісіндегі кейбір жоғары жиіліктегі сөздердің корпустық талдауына назар аударсақ:



3-сурет. “Ер Төстік” ертегісі мәтінінің сегменттерінде сөздердің қайталану жиілігі

Ескерту: деректер негізінде алынған (Ертегілер параллель корпусы, 2025) (ішкі жабық ресурс)

Жоғарыдағы сурет “Ер Төстік” ертегісі мәтіні сегменттерінде жиі қолданылатын сөздердің салыстырмалы динамикасын көрсетеді. Мәтін 10 сегментке бөлініп, сөздердің қайталану жиілігі әр бөлім бойынша өлшенген. Талдауға алынған бірліктер: «мен», «екен», «деп», «дейді», «бір». Жиілік динамикасының ерекшелігі неде? Көріп отырғанымыздай, бастапқы сегменттерде «бір» сөзінің жиілігі айқын байқалады. Бұл ертегінің формулалық басталу дәстүріне («бір хан болыпты», «бір күні») сәйкес келетінін жоғарыда айттық. Қалған сөздер сиректеу кездеседі. 2-сегмент кезінде «бір» сөзінің қолданысы төмендеп, орнына «деп» және «екен» біршама белсенді көрінеді. Мұнда баяндаушы оқиға желісін нақтылау үшін шылауларды қосымша пайдаланғанын көреміз. 3-сегмент «бір» сөзі қайтадан шарықтап, жиілігі жоғары деңгейге жетуін білдіреді. Бұл тағы да ертегінің баяндаудың циклдік формуласына байланысты. 4-сегменте байқасақ, «деп» және «дейді» сөздері алдыңғы қатарға шығады. Диалогтық құрылымдар күшейіп, кейіпкерлердің сөзіне сілтеме жасау жиіленеді. «Дейді» сөзі 5-сегментте ең жоғары

жиілікке жетіп, мәтіннің басқа бөліктерінде де айтарлықтай қолданысқа ие болған. 6-сегментте «дейді» сөзінің жиілігі сақталғанымен, салыстырмалы түрде азаяды. «Деп» тұрақты қолданылуын жалғастырады. «Мен» және «екен» сөздерінің жиілігі салыстырмалы түрде төмен, алайда мәтін соңына қарай біршама арта түседі. «Мен» есімдік/шылауы және «екен» баяндаушы мәтінге көмекші реңк береді. Негізінен бұл жерде «мен» омонимиялық ажырату болмағандықтан автоматты түрде бірге саналып тұр. Сондықтан оның шылау ретінде немесе есімдік ретінде қолданылғанын қолмен санау қажет болады. Дегенмен, сөз соңында жиі келгендігіне сүйеніп, олардың көбіне көмекші мәнде жұмсалғанын аңғарамыз. 7-сегментке келгенде, «дейді» сөзі қайтадан күшейіп, екінші шарықтау шегіне жетеді. Бұл ертегінің орталық бөлігіндегі оқиғалардың шиеленісуімен байланысты. «Бір» сөзі де елеулі деңгейде қолданылған. Ертегінің басында жоғары қолданыста болып, кейін біртіндеп төмендейді. Бұл көрсеткіш ертегі дискурсында көрсетілген баяндау формуласының («Бір күні...», «Бір хан болыпты...») маңызды орын алатынын дәлелдейді. Ал «деп» сөзінің қолданылуы тұрақты сипатқа ие. Әсіресе, 4-6-сегмент аралығында жиіліктің артуы кейіпкерлер сөзін жеткізуде осы бірліктің жетекші қызмет атқаратынын айғақтайды.

8-сегмент – барлық сөздердің қолданылуы біршама теңескен кезең. «Бір», «деп», «дейді» шамалас деңгейде, ал «мен» мен «екен» біршама жоғарылайды. Бұл мәтіннің соңына қарай сөздік құрылымдардың теңгерімді бола бастайтынын көрсетеді. 9-сегментте барлық сөздердің жиілігі төмендейді. Ертегінің аяқталу бөлігіне тән баяндау ықшамдалады, сондықтан формулалық қайталанулар да азаяды. Мәтіннің соңғы сегментінде «деп» пен «дейді» тағы да алдыңғы қатарда, бірақ алдыңғы шыңдарынан төмен. «Мен», «екен», «бір» қалыпты деңгейде. Бұл ертегінің қорытындылау бөлігіне тән сипаттама боп саналады.

Жалпы алғанда, сөз жиілігінің динамикасы ертегі мәтінінің құрылымдық-композициялық ерекшеліктерін көрсетеді. «Дейді» мен «деп» баяндау мен диалогты жеткізудің негізгі көрсеткіштері болса, «бір» ертегінің формулалық тұрақтылығын айқындайды. Ал «мен» мен «екен» тәрізді бірліктердің сирек қолданылуы олардың стилистикалық-қызметтік ерекшеліктерімен байланысты.

Ертегі формулаларының қызметтері туралы ғалымдардың пікірі сан алуан. Соның ішінде мынадай ғылыми тұжырымдарға назар аудартқымыз келеді. Зерттеу жұмысында авторлар фольклорлық мәтіндердің бастапқы және аяққы шендерінің жалпы мәтіндегі қызметін қарастырады. Олар фольклорлық жанрлардың белгілі бір синтаксистік формулалармен басталып, солайша аяқталуы дәстүрлі типтік құбылыс ретінде түсінілетінін, фольклортану ғылымы мұны жаттанды нәрседей қабылдайтынын, алайда бұл позицияларды лингвофольклортану аспектісінен қараудың біраз мәселенің мәнін айқындайтынын көрсеткен. Зерттеуде бұл шендердің мәтіннің басталуы мен баяндауға өту, тақырып пен идеяны байланыстыру, сюжетті түйіндеу, ғибрат беру және мәтінді түсіндіруге ықпал ету сияқты функциялары айқындалады. Сонымен бірге олардың прагматикалық рөлі – оқырманды қызықтыру, оқиғаға тарту, мәтін бөліктерін межелу және эстетикалық әсер етуімен сипатталады. Бұл қызметтер әр жанрда түрлі тілдік бірліктер арқылы жүзеге асатынын дәлелдейді (Жақыпов & Доғдырбай, 2018). Ғалым Ж.Жақыповтың (2020) фольклортану бағытындағы еңбектерінде де бұл тұжырымдар нақты дәйектелген.

Тереңірек үңілсек, фольклорлық мәтіндердің формулалық құрылымдарының қызметі мәтін бөліктеріндегі семантикалық мазмұнның ұйымдасу, берілу, құрылу, таралу және сақталу динамикасын реттеуші ретінде де жұмыс жасайды. Оны корпустық, семантикалық, құрылымдық кешенді талдау арқылы жіті қарастырып, дәйектеу мүмкіндігі бар.

Мәтіндік корпустың тек дереккөздік құрал ретінде қабылданып, отандық гуманитарлық ғылымда кең қолданыс таппауы салдарынан корпустық талдаулардың мәтінге терең үңілуге беретін мүмкіндіктері кем бағаланып жүргені жасырын емес.

Ал сайып келгенде, тұтас ертегі мәтінінің сөздік бұлт суретінің де өзіндік тәпсірі бар.



4-сурет. “Ер Төстік” ертегісі бойынша сөздік бұлт (word cloud)

Ескерту: деректер негізінде алынған (Ертегілер параллель корпусы, 2025) (ішкі жабық ресурс)

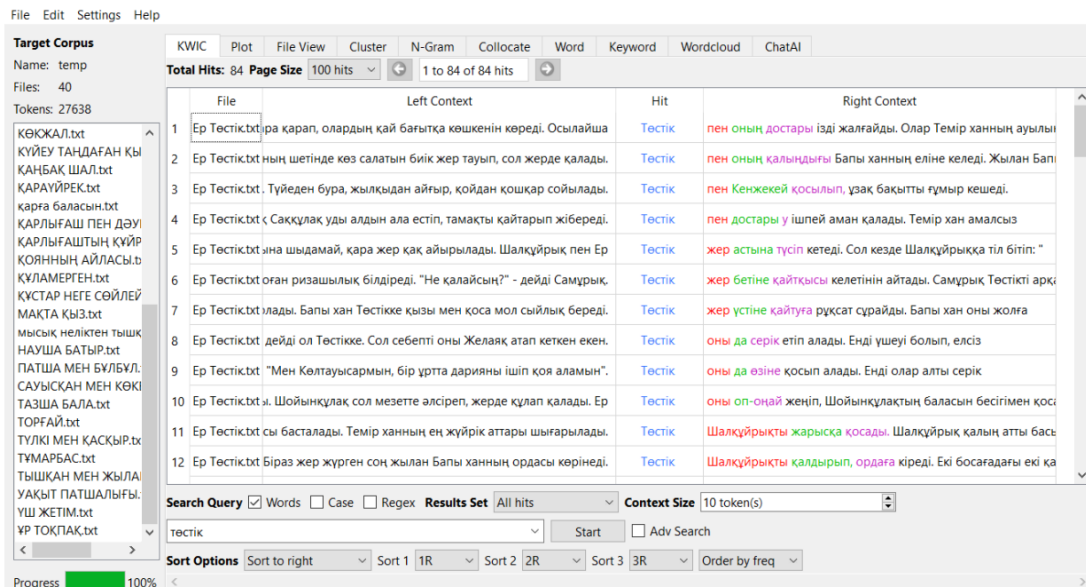
Суреттегі “Ер Төстік” ертегісінің бұл сөздік бұлты (word cloud) мәтінде жиі кездесетін сөздердің көлемдік бейнесін береді. Мұндағы әр сөздің өлшемі оның жиілігіне сәйкес: жиі кездескен сөздер ірі, ал сирек кездескендері ұсақ етіп көрсетіледі. Ірі жазылған сөздер: бір, дейді, деп, екен, мен, алып – мәтіннің дискурстық өзегін құрайды. Мысалы, шеткері төменде орналасқан «алып» – көбінесе әрекет динамикасын жеткізуші. Орташа көлемді сөздер: да, де, кемпір, ол, бала, береді сияқты сөздер сюжеттік дамуға қатысып, кейіпкерлерді, іс-әрекетті және байланыстарды сипаттаған. Ұсақ жазылған сөздер: қыз, келіп, сонда, бар, үш, енді сияқты бірліктер мәтіннің мазмұндық толықтығын көрсеткенімен, жиілігі төмен.

Ең қызығы жиілігі төмен сөздер мәтіндегі бірегей формаларды құрайды. Олардың ішінде таңдаулы құрылымдар, метафора, окказионал бірліктер, т.б. болуы мүмкін.

Бұлтты визуализация мәтіннің қайталанымға негізделген табиғатын дәл көрсетеді: шектеулі бірқатар сөздер мәтіннің негізгі бөлігін құрап, қалғандары қосалқы сипатта қолданылады. Мұндай дискурстық құрылым фольклорлық мәтіндердің формулалық сипатын, дәстүрлі баяндау тілі мен стильдік тұрақтылығын айғақтайды.

«KWIC (KeyWord In Context)» модулі – корпустық интерфейстің негізі, корпус лингвистикасында жиі қолданылатын мәтін талдау құралы. Бұл модуль ізделінген сөздің мәтін ішінде барлық қолданылған мәнмәтіндерін көрсетеді. Іздеу нәтижесі сөйлемнің ішінде орталықтандырылған кілт сөзді және оның сол және оң жағындағы сөздерді

қамти беріледі. Мысалы, «Төстік» сөзін іздегенде, оның қандай сөйлемдерде, қандай мағынада қолданылғанын, қандай контекстермен тіркескенін 5-суреттен көре аламыз. Мұнда «Ер Төстік» ертегісі бойынша ізделген сөз («Төстік») конкордансы көрсетілген.



5-сурет. “Ер Төстік” «ертегісіндегі “Төстік» конкордансы

Ескерту: деректер негізінде алынған (Ертегілер параллель корпусы, 2025) (ішкі жабық ресурс)

Конкордансты талдаудың мәнін түсіндіре кетсек: Target Corpus – зерттеу жүргізіліп отырған корпус (temp) құрамында 40 файл, 27638 токен бары көрсетілген. Іздеу нысаны – Төстік сөзі. Барлық табылған сөзқолданыстар – 84 hits (ізделген сөз). Контекст мөлшері: 10 токенге дейін (сол және оң жақта). Бұл параметрлерді зерттеу мақсатына қарай әр іздеуші өзі белгілей алады. Әр жолда үш бөлік берілген: Left Context – ізделген сөздің алдындағы мәтін. Hit – ізделген сөздің өзі, біздің жағдайымызда “Төстік”, яки конкорданс. Right Context – ізделген сөзден кейінгі мәтін. Мұны да қажеттіңізге байланысты мәнін өзіңіз кеңейтіп көре аласыз. Басқаша айтқанда бұлар – конкорданстың оң, сол жақтарынан келетін сөздер. Мұны фразаларды, валенттіліктерді тексеру, анықтау үшін қолданады. Немесе конкорданстың сөйлемдегі синтаксистік позициясын айқындауда қажет: Мысалы: «Осылайша/ Төстік пен оның достары ізді жалғайды...» – мұнда «Төстік» баяндауыш субъектісі ретінде қолданылып тұр.

Бұл суреттен «Ер Төстік» мәтінінде “Төстік” сөзінің дискурстың орталық тірегі ретінде көрінгенін байқаймыз. Оның әртүрлі септік формалары мәтіннің синтаксистік құрылымын, сюжеттік дамуын анық көрсетеді. Конкорданс нәтижелері бойынша, “Төстік” сөзі әрекетті жүзеге асырушы субъект және әрекетке қатысушы объект ретінде жиі қолданылған. Бұл дерек корпус лингвистикасында кейіпкердің дискурстық рөлін, сөз тіркесі байланысын және мәтіндік жиілікті зерттеуге негіз бола алады.

Талдаулардан көріп отырғанымыздай, фольклорлық мәтіндерді корпустық әдіспен талдау дәстүрлі фольклортанулық зерттеулерді заманауи лингвистикалық және IT құралдарымен ұштастыру арқылы жаңа ғылыми нәтижелерге қол жеткізуге жағдай жасайды.

Қорытынды

Зерттеу нәтижелері қазақ фольклорлық мәтіндерін корпусық лингвистика әдістерімен зерделеудің теориялық және қолданбалы маңызын дәлелдейді. Ұлттық фольклорды заманауи цифрлық форматта ғылыми айналымға енгізу қазақ тіл білімінің жаңа парадигмасын қалыптастыруға жол ашады. Фольклорлық мәтіндер корпусын құру тек деректерді жинақтау емес, олардың құрылымдық, семантикалық және прагматикалық деңгейлерін кешенді зерттеудің жаңа ғылыми негізін қалыптастырады.

Зерттеу барысында мәтіндерді цифрлық ортаға көшіру, морфологиялық және семантикалық таңбалау, конкорданс жасау және статистикалық талдау әдістері сынақтан өтіп, фольклор мәтіндерінің ішкі заңдылықтары мен жанрлық ерекшеліктері айқындалды. Бұл деректер әдебиеттану, лингвистика және IT салаларымен тоғысып, пәнаралық зерттеулерге негіз болады.

Алынған нәтижелер фольклорлық мәтіндердің құрылымдық ерекшеліктері мен сөздік қорын автоматтандырылған жүйелерде сақтауға, әрі оларды тілдік алгоритмдер, мәтін тану және жасанды интеллект бағытындағы зерттеулерге тиімді қолдануға мүмкіндік береді. Практикалық тұрғыдан бұл жұмыс фольклорды заманауи әдістермен сақтап қана қоймай, қазақ тілін халықаралық деңгейде таныстыруға, жаңа буын оқыту құралдарын, цифрлық сөздіктер мен лингвистикалық платформаларды жасауға мүмкіндік береді.

Қорыта келгенде, қазақ фольклорлық мәтіндерінің корпусын әзірлеу – ұлттық тіл ғылымында жаңа зерттеу кеңістігін қалыптастыратын кешенді ғылыми жоба. Алдағы уақытта корпус деректерін кеңейту және пәнаралық байланыстарды дамыту арқылы қазақ фольклортану ғылымы мен цифрлық гуманитарлық зерттеулердің сапасын арттыруға мүмкіндік туғызады.

Мүдделер қақтығысы, алғыс айту және қаржыландыру туралы ақпарат

Мақала бойынша мүдделер қақтығысы жоқ.

Бұл зерттеу жұмысы Қазақстан Республикасы Ғылым және жоғары білім министрлігі Ғылым комитеті тарапынан гранттық қаржыландырылған “АРО26103909-Фольклорлық мәтіндер корпусы: қазақ тілін цифрландырудағы лингвистикалық және IT шешімдер” ғылыми жобасы аясында орындалды (Жоба жетекшісі – Қ.Сарекенова, орындау мерзімі – 2025-2027 ж.).

Авторлардың қосқан үлесі. Н.Н. Аитова мақаланы жазып, идеялық, мазмұндық жағынан орындады, хат-хабар үшін жауапты. М. Бакытқызы мақаланың рәсімделуіне елеулі үлес қосты, зерттеу материалдарын жинақтады. Авторлар тең құқылы және жауапты.

Әдебиеттер тізімі

Бабалар сөзі: жүз томдық. (2013). Астана: Фолиант.

Балабекова, К.Ш., & Отарбекова, Ж.К. (2019). Қазақ фольклорындағы дәстүрлі ментальдік және ұлттық таным. Абай атындағы ҚазҰПУ хабаршысы. Филология ғылымдары сериясы, 1(67), 30-35.

Ертегілер параллель корпусы. (2025, мамыр 8). <https://ertegilerparallelcorpus.tilda>

Жақыпов, Ж., & Доғдырбай, А. (2018). Фольклорлық жанрлар мәтінінің бастапқы және аяққы шендері. Л.Н. Гумилев атындағы ЕҰУ хабаршысы. Филология сериясы, 124(3), 116–124. <https://doi.org/10.32523/2616-678X-2018-124-3-116-124>

Жақыпов, Ж. (2020). Қазақ лингвофольклортануы: тілдік-жанрлық зерттеу. Л.Н. Гумилев атындағы ЕҰУ.

Ондар, М. В. (2016). База данных текстов тувинского героического эпоса: первый этап. Новые исследования Тувы, 4. <https://nit.tuva.asia/nit/article/view/616>

Тыва тілінің электронды корпусы. (2025, May 2). <https://www.tuvancorpus.ru/?q=content/folklornyy-korpus>

American Folklife Center (2025, May 2). <https://www.loc.gov/research-centers/american-folklife-center/about-this-research-center/>

Baker, M. (1995). Corpora in Translation Studies: An overview and some suggestions for future research. Target, 7(2), 223–243.

Harvilahti, L. (2013). The SKVR Database of Ancient Poems of the Finnish People in Kalevala meter and the Semantic Kalevala. Oral Tradition, 28(2).

Janicki, M., Kallio, K., & Sarv, M. (2023). Exploring Finnic written oral folk poetry through string similarity. Digital Scholarship in the Humanities, 38(1), 180–194. <https://doi.org/10.1093/llc/fqac034>

Ide, N., & Véronis, J. (1995). Text Encoding Initiative: Background and context. Kluwer Academic Publishers. <https://doi.org/10.1007/978-94-011-0325-1>

Karlsson, T. (2021). Register features in Kalevala-metric incantations. Language & Communication, 78, 40–53. <https://doi.org/10.1016/j.langcom.2021.02.005>

McEnery, T., & Hardie, A. (2012). Corpus linguistics: Method, theory and practice. Cambridge University Press. https://assets.cambridge.org/97805218/38511/frontmatter/9780521838511_frontmatter.pdf

Suomen kansan vanhat runot (SKVR, The Ancient Songs of the Finnish People). (n.d.). <https://www.finlit.fi/en/about-us/kalevala-european-heritage-label>

The Norwegian folklore archives. (2025, May 2). <https://www.traditionarchives.org/archives/norway#h.whj763x3t279>

The Multilingual Folk Tale Database (MFTD). (2025, May 2). https://psychemedia.github.io/storynotes/MFTD-Multilingual_Folk_Tale_Database.html

Zhakupov, Zh., Meirambekova, L., Aimukhambet, Zh., Akhmetova, M., Yespekova, L., & Syzdykova, G. (2020). Linguistic representation of time and space in Kazakh folklore genres. Opción, 36(91), 523–535. <http://produccioncientificaluz.org/index.php/opcion/article/view/31864>

Н.Н. Аитова*, М. Бақытқызы

Евразийский национальный университет имени Л.Н. Гумилева, Астана, Казахстан

Корпус казахских фольклорных текстов: научные подходы и перспективы исследования

Аннотация. В соответствии с основной стратегией нашего государства, направленной на цифровизацию гуманитарной сферы, в последние годы наблюдается возрастающий интерес к созданию текстовых баз. Это, по сути, признается частью современных мировых тенденций. Развитие корпусных исследований, в частности, моделирование естественного языка, построение языковых алгоритмов и обработка текста, превращается в приоритетное направление современной лингвистики и продиктовано реальными потребностями.

В данной статье рассматриваются возможности научного анализа и исследования текста на основе разработки «Корпуса казахских фольклорных текстов». Основные направления работы включают сбор текстов, их цифровизацию, морфологическую и семантическую разметку, создание конкордансов и статистический анализ языковых единиц.

Научная значимость исследования заключается в формировании предпосылок для создания лингвистической корпусной базы, которая открывает новые возможности для изучения казахского фольклора на современном уровне, анализа его текстовой структуры с применением актуальной методологии и введения в научный оборот богатых пластов языка. В качестве основных результатов представлен корпусный модуль казахских фольклорных текстов и показаны его аналитические возможности. Настоящее исследование открывает новое научное пространство в казахском языкознании посредством адаптации теории и практики корпусной лингвистики к национальному фольклору и вносит методологический вклад в развитие казахской фольклористики.

Ключевые слова: казахский фольклор, лингвистический корпус, текст, корпусный анализ, корпусная лингвистика.

N.N. Aitova*, M. Bakytzyzy

L.N. Gumilyov Eurasian National University, Astana, Kazakhstan

The Kazakh folklore corpus: scientific approaches and research perspectives

Abstract. In line with the main strategy of our country aimed at the digitalization of the humanities, in recent years, there has been a growing interest in creating textual databases. This trend is recognized as part of global developments of contemporary importance.

This article explores the opportunities for scientific text analysis and research offered by the development of "The Kazakh Folklore Corpus." The study aims to analyze the significance of systematizing samples of Kazakh oral literature through corpus linguistics methods in order to identify their linguistic, structural, and semantic features. The main areas of the study include the collection of texts, their digitization, morphological and semantic annotation, concordance creation, and statistical analysis of linguistic units.

The scientific significance of the research lies in establishing the prerequisites for building a linguistic corpus database that enables the study of Kazakh folklore at a new level, analyzing its textual structures using modern methodologies, and introducing the rich layers of the language into academic circulation. The key results include the presentation of a corpus model of Kazakh folklore texts and the demonstration of its potential for comprehensive linguistic analysis. This study opens up a new research space in Kazakh linguistics by adapting the theory and practice of corpus linguistics to national folklore and contributes a methodological basis to the development of Kazakh folklore studies.

Keywords: Kazakh folklore, linguistic corpus, text, corpus analysis, corpus linguistics.

References

- Babalar sozi [The Words of the Ancestors]. (2013). In one hundred volumes. Astana: Foliant. (in Kazakh)
- Balabekova, K. Sh., & Otarbekova, Zh. K. (2019). Kazak folkloryndagy dasturli mental'dik zhane ulttyk tanym [Traditional mental and national perception in Kazakh folklore]. Bulletin of Abai KazNPU. Series

of Philological Sciences, 1(67), 30–35. <https://bulletin-philology.kaznpu.kz/index.php/ped/issue/view/24/36> (in Kazakh)

Ertegiler parallel korpusy [Parallel corpus of fairy tales]. (2025, May 8). <https://ertegilerparallelcorpus.tilda> (in Kazakh)

Zhakupov Zh., & Dogdyrbai A. (2018). Folklorlyk zhanrlar matininin bastapky zhane aiakky shenderi [Initial and Final Positions of the Text of Folklore Genres]. Bulletin of L.N. Gumilyov Eurasian National University. Philology Series, 124(3), 116–124. <https://doi.org/10.32523/2616-678X-2018-124-3-116-124> (in Kazakh)

Zhakypov, Zh. (2020). Kazak lingvofol'klortanuy: tildik-zhanrlik zertteu [Kazakh linguistic folklore studies: Linguistic-genre research]. Astana: L.N. Gumilyov ENU. (in Kazakh)

Ondar, M. V. (2016). The structure of an entry in the National corpus of Tuvan language. The New Research of Tuva, (4). Retrieved from <https://nit.tuva.asia/nit/article/view/616>

American Folklife Center (2025, May 2). <https://www.loc.gov/research-centers/american-folklife-center/about-this-research-center/>

Baker, M. (1995). Corpora in Translation Studies: An overview and some suggestions for future research. Target, 7(2), 223–243.

Harvilahti, L. (2013). The SKVR Database of Ancient Poems of the Finnish People in Kalevala meter and the Semantic Kalevala. Oral Tradition, 28(2).

Janicki, M., Kallio, K., & Sarv, M. (2023). Exploring Finnic written oral folk poetry through string similarity. Digital Scholarship in the Humanities, 38(1), 180–194. <https://doi.org/10.1093/llc/fqac034>

Ide, N., & Véronis, J. (1995). Text Encoding Initiative: Background and context. Kluwer Academic Publishers. <https://doi.org/10.1007/978-94-011-0325-1>

Karlsson, T. (2021). Register features in Kalevala-metric incantations. Language & Communication, 78, 40–53. <https://doi.org/10.1016/j.langcom.2021.02.005>

McEnery, T., & Hardie, A. (2012). Corpus linguistics: Method, theory and practice. Cambridge University Press. https://assets.cambridge.org/97805218/38511/frontmatter/9780521838511_frontmatter.pdf

Suomen kansan vanhat runot (SKVR, The Ancient Songs of the Finnish People). (n.d.). <https://www.finlit.fi/en/about-us/kalevala-european-heritage-label>

The Norwegian folklore archives. (2025, May 2). <https://www.traditionarchives.org/archives/norway#h.whj763x3t279>

The Multilingual Folk Tale Database (MFTD). (2025, May 2). https://psychedia.github.io/storynotes/MFTD-Multilingual_Folk_Tale_Database.html

Zhakupov, Zh., Meirambekova, L., Aimukhambet, Zh., Akhmetova, M., Yespekova, L., & Syzdykova, G. (2020). Linguistic representation of time and space in Kazakh folklore genres. Opción, 36(91), 523–535. <http://produccioncientificaluz.org/index.php/opcion/article/view/31864>

Авторлар туралы мәлімет:

Аутова Нурлыхан Нуруллаевна – автор для корреспонденции, кандидат филологических наук, ассоциированный профессор, Евразийский национальный университет им. Л.Н. Гумилева, Астана, Казахстан. E-mail: nurlykhan.an@gmail.com, ORCID: <https://orcid.org/0000-0002-9606-4564>

Бақытқызы Молдир – PhD докторант, Евразийский национальный университет им. Л.Н. Гумилева, Астана, Казахстан. E-mail: 0002992@mail.ru, ORCID: <https://orcid.org/0009-0008-7654-4959>

Аитова Нұрлыхан Нуруллаевна – хат хабар үшін автор, филология ғылымдарының кандидаты, қауымдастырылған профессор, Евразийский национальный университет им. Л.Н. Гумилева, Астана, Қазақстан. E-mail: nurlykhan.an@gmail.com, ORCID: <https://orcid.org/0000-0002-9606-4564>

Бақытқызы Мөлдiр – PhD докторант, Евразийский национальный университет им. Л.Н. Гумилева, Астана, Казахстан. E-mail: 0002992@mail.ru, ORCID: <https://orcid.org/0009-0008-7654-4959>

Aitova Nurlykhan – corresponding author, Candidate of Philology, Associate Professor, L.N. Gumilyov Eurasian National University, Astana, Kazakhstan. E-mail: nurlykhan.an@gmail.com, ORCID: <https://orcid.org/0000-0002-9606-4564>

Bakytkyzy Moldir – PhD student, L.N. Gumilyov Eurasian National University, Astana, Kazakhstan. E-mail: 0002992@mail.ru, ORCID: <https://orcid.org/0009-0008-7654-4959>



Copyright: © 2025 by the authors. Submitted for possible open access publication under the terms and conditions of the Creative Commons Attribution (CC BY NC) license (<https://creativecommons.org/licenses/by-nc/4.0/>).