

ФИЛОЛОГИЯ

ӘОЖ 81'271; 82.085; 808.5; МҒТАР 16.21.61
<https://doi.org/10.47526/2025-4/2664-0686.269>Ф.Ш. ОРАЗБАЕВА¹, А.Д. РЫСҚҰЛОВА²¹педагогика ғылымдарының докторы, профессор, ҚР ҰҒА академигі
(Қазақстан, Алматы қ.), e-mail: f_orazbaeva@mail.ru²PhD докторант, e-mail: ainagulr@mail.ru^{1,2}Абай атындағы Қазақ ұлттық педагогикалық университеті**ЖАСАНДЫ ИНТЕЛЛЕКТ ЖӘНЕ ҚАЗАҚ ТІЛІНІҢ ҒЫЛЫМИ ДАМУЫ:
КОРПУС, ТЕРМИНОЛОГИЯ ЖӘНЕ КОНТЕНТТІ АВТОМАТТАНДЫРУ**

Аңдатпа. Мақалада жасанды интеллект технологиялары мен тілдік корпустарды үйлесімді ұштастыру арқылы қазақ тілінің ғылыми-теориялық әлеуетін арттыру жолдары жан-жақты зерделенеді. Зерттеу мақсаты – қазақ тілін цифрлық кеңістікте тиімді қолдану үшін тілдік ресурстарды жасанды интеллект құралдарымен біріктіре отырып, терминологиялық жүйелеу мен автоматтандырылған өңдеу үдерістерін тілдің морфологиялық, семантикалық, синтаксистік және лексикалық деңгейінде талдау.

Зерттеу аясында алдымен қазақ тілін автоматты өңдеуге арналған ұлттық деректер қорын қалыптастыру қажеттілігі ғылыми-теориялық тұрғыда негізделеді. Бұл қорды құрастыруда морфологиялық белгілеу, сөз табын автоматты тану, түбір мен қосымшаның жіктелуі сияқты тілдік бірліктердің ішкі құрылымы ескеріледі. Сонымен қатар терминдер жүйесін біріздендіру мақсатында семантикалық өріс, ұғымаралық байланыс және сөздердің функционалдық жүктемесі сипатталады.

Тәжірибелік талдау барысында веб-скрепинг әдісі арқылы іріктелген мәтіндерден жиілік сөздер анықталып, олардың қолданыс жиілігі мен контекстуалдық қолданыс ерекшеліктері конкорданс әдісімен зерделенді. Салыстырмалы-диахрониялық тәсіл арқылы белгілі бір терминдердің әр кезеңдегі қолданысы мен мағыналық өзгерісі, сондай-ақ грамматикалық құрылымдағы орны сараланды. Жасанды интеллектінің семантикалық модельдеу, машиналық оқыту алгоритмдері аталған процестерді жүйелеуге мүмкіндік берді.

Нәтижесінде қазақ тіліндегі ақпараттық технологияларға қатысты жаңа терминдердің қалыптасу қарқыны, мағыналық аясын кеңейту тетіктері мен олардың бірізділігі анықталды. Бұл деректер қазақ тілінің цифрлық экожүйесін жетілдіру мақсатында нақты ұсыныстар ұсынуға негіз болды. Мақала қазақ тілінің ғылыми қолданысын арттыруға, терминологиялық

*** Бізге дұрыс сілтеме жасаңыз:**

Оразбаева Ф.Ш., Рысқұлова А.Д. Жасанды интеллект және қазақ тілінің ғылыми дамуы: корпус, терминология және контентті автоматтандыру // *Ясауи университетінің хабаршысы*. – 2025. – №4 (138). – Б. 7–19. <https://doi.org/10.47526/2025-4/2664-0686.269>

***Cite us correctly:**

Orazbaeva F.Sh., Ryskulova A.D. Zhasandy intellekt zhane qazaq tilinin gylymi damuy: korpus, terminologia zhane kontentti avtomattandyru [Artificial Intelligence and the Scientific Development of the Kazakh Language: Corpus, Terminology, and Content Automation] // *Iasau universitetinin habarshysy*. – 2025. – №4 (138). – Б. 7–19. <https://doi.org/10.47526/2025-4/2664-0686.269>

Мақаланың редакцияға түскен күні 18.06.2025 / жарияланған күні 30.12.2025

реттеуді жетілдіруге және тілдің құрылымдық деңгейлерінің (морфология, синтаксис, лексика, семантика) өзара байланысын жүйелі түрде сипаттауға бағытталған.

Зерттеу нәтижелері лингвистика мен ақпараттық технологиялар тоғысында жүзеге асырылатын тіл саясаты мен инновациялық бастамаларға негіз бола алады.

Кілт сөздер: қазақ тілінің корпусы, терминологияны автоматтандыру, жасанды интеллект, машинамен оқыту, семантикалық талдау, контентті автоматтандыру, цифрлық экожүйе.

F.Sh. Orazbayeva¹, A.D. Ryskulova²

¹*Doctor of Pedagogical Sciences; Professor,
Academician of National Academy of Sciences of the Republic of Kazakhstan, e-mail: f_orazbaeva@mail.ru*

²*PhD Doctoral Student, e-mail: ainagulr@mail.ru*

^{1,2}*Abai Kazakh National Pedagogical University (Kazakhstan, Almaty)*

Artificial Intelligence and the Scientific Development of the Kazakh Language: Corpus, Terminology, and Content Automation

Abstract. This article provides a comprehensive analysis of effective strategies for enhancing the scientific and theoretical development of the Kazakh language through the integration of artificial intelligence technologies and linguistic corpora. The primary aim of the study is to examine the processes of terminological standardization and automated language processing across the morphological, semantic, syntactic, and lexical levels of the Kazakh language through the integration of linguistic resources with AI tools in a digital environment.

The theoretical section substantiates the necessity of developing a national language data repository designed for automated Kazakh language processing. The construction of this repository involves morphological annotation, automatic part-of-speech recognition, and the segmentation of word stems and affixes. Additionally, semantic fields, interconceptual relations, and the functional load of lexical units are analyzed to support consistent terminological structuring.

In the practical part of the study, texts collected via web scraping were analyzed to identify high-frequency terms and their contextual usage using concordance methods. A comparative diachronic approach was employed to trace the semantic evolution and grammatical roles of specific terms across different time periods. Modern AI tools, including semantic modeling techniques and machine learning algorithms, facilitated the systematic analysis of these processes.

As a result, the study identified the rate of term formation, the mechanisms of semantic expansion, and the level of terminological standardization in the field of information technology in the Kazakh language. The findings are supplemented with specific proposals aimed at improving the digital ecosystem of the Kazakh language. This research makes a significant contribution to expanding the scientific use of the Kazakh language and strengthening terminological regulation by systematically describing the interrelationships between morphology, syntax, semantics, and lexicon.

Keywords: Kazakh language corpus, terminology automation, artificial intelligence, machine learning, semantic analysis, content automation, digital ecosystem.

Ф.Ш. Оразбаева¹, А.Д. Рыскулова²

¹*доктор педагогических наук, профессор, академик НАН РК, e-mail: f_orazbaeva@mail.ru*

²*PhD докторант, e-mail: ainagulr@mail.ru*

^{1,2}*Казахский национальный педагогический университет имени Абая (Казахстан, г. Алматы)*

Искусственный интеллект и научное развитие казахского языка: корпус, терминология и автоматизация контента

Аннотация. В статье всесторонне рассматриваются эффективные пути повышения научно-теоретического потенциала казахского языка посредством интеграции технологий искусственного интеллекта и языковых корпусов. Цель исследования – проанализировать процессы терминологической стандартизации и автоматизированной обработки казахского языка на морфологическом, семантическом, синтаксическом и лексическом уровнях через объединение языковых ресурсов с инструментами ИИ в условиях цифровой среды.

В теоретической части обоснована необходимость создания национальной базы данных для автоматизированной обработки казахского языка. При формировании корпуса учитываются морфологическая разметка, автоматическое определение части речи, структурный анализ основы и аффиксов. Также описаны семантические поля, межпонятийные связи и функциональная нагрузка лексем с целью систематизации терминов.

В эмпирической части исследования при помощи метода веб-скрейпинга были отобраны тексты, из которых с использованием конкорданс-анализа выявлены частотные слова и проанализированы особенности их употребления в контексте. Сравнительно-диахронический подход позволил проследить эволюцию значений терминов и их грамматическое функционирование на разных исторических этапах. Алгоритмы машинного обучения и семантического моделирования способствовали систематизации полученных данных.

В результате определена скорость формирования новых терминов, расширение их семантического содержания и уровень стандартизации в сфере информационных технологий на казахском языке. Полученные данные дополнены конкретными предложениями по совершенствованию цифровой экосистемы казахского языка.

Статья вносит значительный вклад в расширение научного употребления казахского языка и укрепление терминологической упорядоченности, описывая взаимосвязь между морфологией, синтаксисом, семантикой и лексикой.

Ключевые слова: корпус казахского языка, автоматизация терминологии, искусственный интеллект, машинное обучение, семантический анализ, автоматизация контента, цифровая экосистема.

Кіріспе

Қазіргі кезеңдегі цифрлық технологиялардың үдемелі дамуы тілдің қызмет ету кеңістігін бұрынғыдан едәуір кеңейтіп, оны жаңа ортада жан-жақты пайдалануға мүмкіндік беріп отыр. Әлемдік ғылыми қауымдастықта жасанды интеллекттің ықпалы күшейген сайын, кез келген тілдің сандық жүйелерге икемделуі, деректермен жұмыс істейтін алгоритмдерге бейімделуі және ақпараттық ортада еркін қолданылуы маңызды талапқа айналды. Мұндай жағдайда тіл тек коммуникация құралы ғана емес, қоғамдағы интеллектуалдық әлеуетті танытатын стратегиялық ресурс ретінде бағаланады.

Қазақ тілінің қазіргі даму жағдайы бұл бағыттағы жұмыстың жүйелі жүргізілуін қажет етеді. Соңғы жылдары лингвистикалық ресурстарды цифрландыруға, терминологиялық қорды жаңартуға және тілдің қолданыс аясын кеңейтуге арналған түрлі жобалар қолға алынғанымен, ұлттық көлемдегі ауқымды корпус толық қалыптаспай тұр. Сонымен бірге салалық терминдердің бірнеше нұсқада қатар қолданылуы, тілдік материалдарды автоматты түрде талдайтын арнайы платформалардың жеткіліксіздігі қазақ тілінің цифрлық кеңістікте орнықты қызмет етуіне белгілі бір дәрежеде кедергі келтіреді. Бұл жағдай қазақ тілінің

ғылыми, техникалық және әлеуметтік-коммуникативтік бағыттарда толыққанды дамуына ықпал ететін жаңа әдістерді қажет етеді.

Жасанды интеллектке негізделген технологиялар тілдік деректерді өңдеудің жаңа мүмкіндіктерін ұсынып отыр. Алайда қазақ тілінің морфологиялық күрделі құрылымы, сөзжасамдық ерекшеліктері, жалғау-жұрнақ арқылы мағынаның үнемі түрленіп отыруы алгоритмдерді тікелей пайдалануға мүмкіндік бере бермейді. Сондықтан қазақ тілін автоматты өңдеуге бейімделген әдістерді қалыптастыру, шағын көлемдегі мәтіндер негізінде машиналық талдаудың ықтимал жолдарын көрсету және тілдің ішкі заңдылықтарын ескеретін жаңа үлгілерді ұсыну – ғылыми тұрғыдан да, практикалық тұрғыдан да аса маңызды міндет.

Мақаланың мақсаты – қазақ тілін жасанды интеллект әдістерімен талдау және автоматтандырудың ғылыми-тәжірибелік негіздерін айқындау, сондай-ақ тілдің цифрлық кеңістікте қолданысын кеңейтуге бағытталған тиімді тәсілдерді ұсыну.

Мақсатқа сәйкес төмендегідей міндеттер белгіленді:

1. Шағын мәтіндік корпустар мен қолда бар лингвистикалық ресурстарды машиналық талдау әдістерімен ұштастырып, қазақ тілін автоматты өңдеудің ықтимал бағыттарын көрсету;
2. Ақпараттық технологиялар, медицина, экономика сияқты салалардағы терминдердің қолданыс ерекшеліктерін анықтап, көпнұсқалықтың себептерін ғылыми тұрғыда түсіндіру;
3. Контентті автоматтандыру мен семантикалық талдау әдістерін қазақ тілінің аглютинативті құрылымына бейімдеу мүмкіндіктерін сипаттау;
4. Тілдің цифрлық инфрақұрылымын жетілдіруге арналған практикалық ұсыныстар әзірлеу.

Зерттеу объектісін қазақ тіліндегі ғылыми-публицистикалық материалдар, БАҚ мәтіндері, әдеби үзінділер, блогтар, терминологиялық сөздіктер және қолданыстағы лингвистикалық платформалар құрайды. Зерттеу пәні – осы материалдарға сүйенген талдау нәтижесінде айқындалатын лексикалық өзгерістер, терминдердің қолданылу динамикасы және автоматты өңдеу барысында көрінетін тілдік заңдылықтар.

Мақаланың ғылыми жаңалығы бірнеше бағыттан көрініс табады. Біріншіден, толыққанды ұлттық корпус қалыптаспай тұрған жағдайда да шағын мәтіндік деректерді жасанды интеллект әдістерімен талдаудың жұмысқа жарамды үлгісі ұсынылады. Екіншіден, соңғы онжылдықтардағы терминологиялық қолданысқа салыстырмалы талдау жүргізіліп, көпнұсқалықтың нақты көріністері жүйеленеді. Үшіншіден, қазақ тілінің морфологиялық ерекшеліктерін ескеретін автоматтандырылған талдау тәсілдерінің мүмкіндіктері айқындалып, олардың тілдік зерттеулерге қосатын үлесі көрсетіледі.

Түйіндей айтқанда, қазақ тілінің цифрлық кеңістікте тиімді қызмет етуін қамтамасыз ету – тек ғылыми қажеттілік ғана емес, тілдің жалпы даму бағытын айқындайтын стратегиялық мақсат. Осы мақалада ұсынылған тәсілдер мен алынған нәтижелер ұлттық корпус қалыптастыруға, салалық терминологияны біріздендіруге және қазақ тілінің цифрлық инфрақұрылымын жетілдіруге бағытталған болашақ ғылыми зерттеулер үшін берік әдіснамалық негіз бола алады.

Зерттеу әдістері мен материалдар

Бұл зерттеуде қолданылған материалдар қазақ тілінің қазіргі қолданысын әр қырынан танытатын мәтіндік дереккөздерден жинақталды. Ұлттық деңгейдегі толық корпус әзірге қалыптасу үстінде болғандықтан, зерттеу мақсатына сәйкес шағын көлемдегі мәтіндер жиынтығы пайдаланылды. Материалдар бірнеше дереккөзден алынды: ашық қолжетімді

Qazcorpus, KazNLP және TILQURAL платформаларындағы мәтіндер; жаңалық сайттары, әдеби блогтар мен кәсіби порталдарда жарияланған материалдар; түсіндірме және терминологиялық сөздіктердің қағаз және электронды нұсқалары; сондай-ақ ғылыми-публицистикалық мақалалар мен әдеби шығармалардан іріктелген үзінділер. Бұл мәтіндер жанры, уақыты, көлемі және тақырыптық сипаты жағынан әртүрлі болғандықтан, қазақ тілінің әр стильдегі қолданысын кең ауқымда қарастыруға мүмкіндік берді.

Материалдарды сипаттау барысында олардың шыққан жылы, жанры, сөз саны және негізгі мазмұны ескеріліп, арнайы кестеге түсірілді. Ғылыми мақалалар қазақ тілінің ресми-ғылыми қолданысын көрсетсе, БАҚ материалдары қоғамдық өмірдегі лексикалық және стильдік өзгерістерді тануға жол ашты. Әдеби мәтіндер авторлық стиль мен көркемдік амалдарды саралауға мүмкіндік берді, ал терминологиялық сөздіктер жаңа атаулардың ресми нұсқасын салыстыруға негіз болды. Әртүрлі жанрдағы материалдарды қатар қамту тіл дамуының жалпы бағытын бағалауға және тілдің әлеуметтік, ғылыми қолданысындағы ерекшеліктерді байқауға жағдай жасады.

Зерттеу әдістері кешенді түрде қолданылды. Таңдамалы мәтіндер әдісі зерттеу мақсатына сай мәтіндерді іріктеп алуға мүмкіндік берді. Лексикографиялық және терминологиялық шолу арқылы салалық сөздіктер мен глоссарийлердегі деректер жүйеленіп, терминдердің қалыптасу ерекшеліктері мен қолданылу бірізділігі қарастырылды. Уақыт аралығындағы лексикалық өзгерістерді талдау үшін салыстырмалы-диахрониялық тәсіл қолданылып, 1990 және 2020 жылдар аралығындағы жазба материалдар салыстырылды. Мәтіндерді техникалық тұрғыдан өңдеу үшін AntConc, Voyant Tools және Python ортасындағы талдау құралдары пайдаланылды. Бұл бағдарламалар мәтіндегі сөздердің қайталануын, тіркесімділік өрнектерін, құрылымдық ерекшеліктерін және терминдердің қолданылу жағдайларын белгілеуге мүмкіндік берді.

Зерттеу барысы бірнеше дәйекті кезеңнен тұрды. Алдымен әртүрлі жанрдағы мәтіндер іріктеліп, сипаттамалық кестеге енгізілді. Одан кейін мәтіндер техникалық белгілерден тазартылып, қажет болған жағдайда орфографиялық түзетулер енгізілді. Метадеректер (жылы, авторы, дереккөзі) тіркелгеннен кейін талдау әдістері іске асырылып, мәтіндер жиілік, тіркесімділік және кезеңдік айырмашылық тұрғысынан сараланды. Жиналған деректер өзге лексикографиялық көздермен салыстырылып, қажет болған тұстарда салалық мамандардың пікірлері ескерілді.

Болашақ зерттеулерде қазақ тілінің табиғи қолданысын тереңірек тану үшін NVivo, Sketch Engine және ELAN сияқты заманауи платформаларды пайдалану ұсынылады. NVivo сапалық деректерді кодтауға және дискурстық құрылымды тақырыптық тұрғыдан талдауға мүмкіндік береді; Sketch Engine жанр бойынша салыстырмалы жиілік пен контекст көрсеткіштерін анықтауға қолайлы; ELAN сөйлеу тілін аннотациялау арқылы интонациялық, синтаксистік және уақыттық ерекшеліктерді талдауға мүмкіндік береді. Бұл бағдарламаларды қазақ тілінің табиғи заңдылықтарына бейімдеу болашақта корпустық, дискурстық және мультимодальды зерттеулердің әдістемелік мүмкіндігін кеңейтеді.

Тіл білімінің қазіргі даму кезеңінде цифрлық технологиялардың қарқынды енуі лингвистикалық зерттеулердің теориясы мен әдіснамасына жаңа сапалық серпін беріп отыр. Тілді тек құрылымдық жүйе ретінде қарастырудан бөлек, оны үлкен деректермен жұмыс істейтін динамикалық құбылыс ретінде сипаттау бағытындағы ізденістер күшейді. Осындай жаңа бағыттардың ішінде corpus-based және corpus-driven зерттеу тәсілдері ерекше орын алады. Corpus-based әдіс тілді алдын ала қалыптасқан теориялық жорамалдарды тексеру үшін қолданылса, corpus-driven тәсілі тілдік заңдылықтарды эмпирикалық дерек негізінде индуктивті түрде анықтауға мүмкіндік береді. Осы екі модельдің артықшылықтары қазіргі әлемдік лингвистикада кеңінен талқыланып, интуитивті пайымдауға емес, нақты дерекке

сүйенетін ғылыми дәлдікті қамтамасыз ететіні көрсетілген [1, 12–14-бб.]. Қазақ тілінің цифрлық кеңістікте қызмет ету мүмкіндігін кеңейтуге бағытталған қазіргі зерттеулер де осы әдіснамалық ұстанымдарға негізделі бастады.

Тіл корпусын синтаксистік, морфологиялық және лексикалық құрылымдарды табиғи қолданыстың өзінде жүйелі сипаттайтын әмбебап құрал ретінде McEnergy мен Hardie жан-жақты сипаттайды. Авторлар корпус деректері тілдің даму динамикасын, стильдік ерекшеліктерін, мәтіндер арасындағы функционалдық айырмашылықтарды анықтауға мүмкіндік беретінін атап өтеді [2, 21–24-бб.]. Алайда қазақ тіліндегі Qazcorpus және KazNLP платформалары, құнды бастама болғанымен, қазіргі ғылыми талаптарды толық қанағаттандырмайды: олардың жанрлық ауқымы тар, мәтін саны жеткіліксіз, техникалық және ауызша дискурс үлгілері қамтылмаған. Сондықтан қазақ тілінің жан-жақты сипатталған толық корпусын жасау – уақыт талабы және тіл саясатының стратегиялық міндеттерінің бірі [3].

Аглютинативті тілдердің ерекшеліктері корпус жасауда айрықша қиындық туғызатыны белгілі. Қазақ тілінің морфологиялық жүйесі қосымшалардың көпқабатты құрылымына, реңктік айырмашылықтардың молдығына, сөзжасамдық модельдердің кең өрісіне сүйенеді. Рәбиға Сыздық өз еңбектерінде қазақ тілінің тарихи морфологиясы мен сөздің ішкі мағыналық жіктелімін терең талдай отырып, автоматтандырылған талдауда жиі байқалатын қателердің басым бөлігі дәл осы морфологиялық күрделілікті ескермеуден туындайтынын атап көрсетеді [4, 55–57-бб.]. Демек, қазақ тілінің корпусын құруда морфологиялық аннотацияның сапасы – зерттеудің негізгі ғылыми шарттарының бірі.

Терминология саласындағы халықаралық зерттеулер де тілдің ғылыми дамуы үшін ерекше маңызды. Терминнің басты міндеті – ұғым мен атау арасындағы тұрақты, нақты және қайшылықсыз байланыс орнату. Алайда қазіргі кезеңде ғылым мен техниканың үдемелі дамуы жаңа терминдердің жаппай енуіне әкеліп, олардың тұрақтану үдерісін күрделендіре түсуде. Sabré терминдерді «арнайы лексика» ретінде қарастырып, олардың жүйелік, семантикалық және құрылымдық сипаттарын сақтау қажеттігін айқындайды [5, 33–35-бб.]. Ал Рәбиға Сыздық қазақ терминжасамында тілдің табиғи мүмкіндіктерін, сөзжасамдық үлгілерді, ұлттық дүниетаныммен сабақтас семантикалық өрістерді пайдаланудың маңызын ерекше атап өтеді [4, 201–205-бб.]. Терминдердің қазақ тіліне ену жолдары – калькалау, транслитерация және төл сөздер негізінде жаңа атау жасау – қазақ тіл біліміндегі салалық зерттеулерде жан-жақты талданады [6, 224–232-бб.].

Терминологияны автоматтандыру мәселесі заманауи лингвистиканың жаңа бағытына айналды. Кім түркі тілдерінде машиналық оқыту модельдерін қолданудың ерекшелігін талдай отырып, морфологиялық күрделі тілдерге бейімделген алгоритмдер жасау қажеттігін дәлелдейді [7, 254–257-бб.]. Қазақ тіліндегі салалық терминдердің қолданысы мен нормаға сәйкестігіне қатысты бірқатар мәселелер отандық зерттеулерде де көтеріледі. Мәселен, Мединаева кейбір тұрақты қолданылып жүрген терминдердің мағыналық шегі, аударма нұсқалары мен орфографиялық берілуіндегі алшақтықтарды талдай отырып, оларды жүйелеу қажеттігін атап өтеді [8, 222–228-бб.].

Контентті автоматты түрде өңдеу – мәтінді қысқарту, тақырыптық жіктеу, семантикалық өрістерді белгілеу, мәтіндер арасындағы байланысты анықтау сияқты күрделі лингвистикалық міндеттерді атқара алатын технологиялар жиынтығы. Maybury автоматты мәтін өңдеу жүйелері ғылыми ақпараттың құрылымдық айқындығын арттыруға және мәтінді тиімді ұйымдастыруға мүмкіндік беретінін көрсетеді [9, 492–497-бб.]. Ал нейрондық аударма жүйелерінің тиімділігі туралы Wu және әріптестері ірі көлемді әрі сапалы таңбаланған дерек болмаса, модель тілдің табиғи құрылымын меңгере алмайтынын нақты дәлелдейді [10, 6–10-

бб.]. Бұл тұжырым қазақ тілі үшін де маңызды: толыққанды корпуссыз автоматты аударма да, автоматтандырылған талдау да сапалы нәтиже бере алмайды.

Корпустық деректердің лексикографиядағы маңызын Bowker мен Pearson кеңінен талдап, сөзді контекстен бөліп қарастыру оның мағынасының толық ашылуына мүмкіндік бермейтінін дәлелдейді [11, 112–118-бб.]. Сондықтан қазақ тілінің заманауи сөздіктері де контекстік мысалдарға, нақты қолданым деректеріне сүйенуі тиіс. Тілдің танымдық табиғаты туралы Chomsky еңбектеріндегі тұжырымдар да цифрлық кеңістіктегі тілдік өзгерістерді түсіндіруге мүмкіндік береді. Автор тілдің адамның ойлау жүйесімен тығыз байланысты когнитивтік қабілет екенін, сондықтан жаңа технологиялық орта тілдің жаңа қызметтік рөлдерін қалыптастыратынын көрсетеді [12, 41–45-бб.]. Біздің зерттеу материалдары да қазақ тілінің жаңа лексикалық қабаттарды жылдам игеретінін, терминологиялық және орфографиялық өзгерістердің қарқыны жоғары екенін дәлелдейді. Осылайша, теориялық дереккөздер мен эмпириялық талдаулардың сабақтастығы қазақ тілінің цифрлық кеңістікте дамуын түсіндірудің берік ғылыми негізін қалайды. Цифрландыру үдерісі тілдің құрылымдық, стильдік, терминологиялық және орфографиялық жүйелеріне ықпал етіп, жаңа лингвистикалық мүмкіндіктер мен міндеттерді айқындап отыр.

1-кесте – Материалдары туралы жалпы мәлімет

Мәтін атауы / дереккөз	Жанр	Жылы	Сөз саны	Қысқаша мазмұны
«Қазақ тілінің келешегі»	Әдеби публицистика	2020	4500	Автордың қазақ тіліне қатысты субъективті ойлары
«ІТ және білім беру жүйесі»	Ғылыми-публицистика	2019	4000	Білім саласындағы компьютерлік технологиялардың рөлі
«Экономика және тіл мәселесі»	БАҚ (талдау мақаласы)	2021	2500	Экономикада қазақ тілінің қолданылуын сөз етеді
«Терминологиялық сөздік» (ІТ)	Салалық терминология	2018	~6000	Компьютерлік технологияларға қатысты ресми бекітілген сөздер
«Медицинадағы жаңа ұғымдар»	БАҚ (танымдық мақала)	2020	3000	Денсаулық сақтау саласының заманауи терминдері
«Көркем шығарма үзінділері»	Әдебиет (әңгіме, повесть)	2021	5000	Әдеби стильдегі тілдік ерекшеліктерін талдау

1-кесте зерттеу базасының жанрлық әркелкілігін айқын көрсетеді. Ғылыми-публицистикалық мәтіндерде ғылыми терминология мен талдамалық құрылымдар жиі кездессе, БАҚ материалдарында қоғамдық-саяси және әлеуметтік-экономикалық лексика басым екені байқалады. Әдеби үзінділер авторлық стильдің, метафоралық қолданыстардың, бейнелі сөз өрнектерінің молдығын танытады, ал терминологиялық сөздіктер ресми бекітілген атауларды, нормаланған лексиканы тіркейді. Осындай түрлі материалдарды бір корпус аясында қарастыру қазақ тілінің әр саладағы қызметін салыстыра бағалауға, жаңа лексикалық қабаттар мен стильдік өзгерістердің түйісетін тұстарын анықтауға мүмкіндік береді.

Жиілік талдауының нәтижесінде әр жанрға тән негізгі лексикалық топтар айқындалды. Алдымен көмекші сөздер сүзгіден алынып, мәтіннің мазмұнын айқындайтын атаулар мен етістіктер талданды. Ғылыми мәтіндерде ұғымдық сипаттағы терминдер мен ғылыми категорияларды білдіретін лексемалар көш бастаса, БАҚ-та қоғамдық өмірді, саяси-экономикалық процестерді сипаттайтын атаулар белсенді қолданылады. Әдеби мәтіндерде кейіпкердің күйін, эмоциясын, ішкі әлемін бейнелейтін психологиялық-экспрессивті сөздер

айрықша жиілікке ие. Осы деректер қазақ тілінің функционалдық стильдер жүйесінің корпус деңгейінде де айқын көрінетінін, және стилдік айырмашылықтарды жиілік арқылы дәлелдеуге болатынын көрсетеді.

Қазақ тілінің цифрлық кеңістіктегі даму бағытын айқын көрсететін маңызды көрсеткіштердің бірі – ақпараттық технологиялар саласына қатысты терминдердің нақты мәтіндік қолданысы. IT лексикасы қазақ тілінің қазіргі кезеңдегі ең жылдам жаңаратын қабаты ретінде танылып отыр. Сол себепті зерттеу аясында IT саласына тән бірқатар негізгі терминдердің мәтіндік жиілігі арнайы шағын корпус негізінде талданды.

Бірнеше жанрдан (БАҚ, ғылыми-публицистика, блог, техникалық нұсқаулықтар) алынған 50 мәтіндік жиынтық IT-терминдердің табиғи қолданысын бағалауға мүмкіндік берді. Алдымен мәтіндер техникалық қалдықтардан тазаланып, кейін AntConc және Python талдау құралдары арқылы жиілік тізімі алынды. Талдау барысында терминдердің тек сандық жиілігі емес, сондай-ақ олардың контекстік қолданысы, стильдік ерекшелігі, мәтін ішіндегі функциясы және қазақша баламалармен сәйкестігі де қарастырылды.

2-кесте – Ақпараттық технологиялар саласындағы терминдердің жиілігі (N = 50 мәтін)

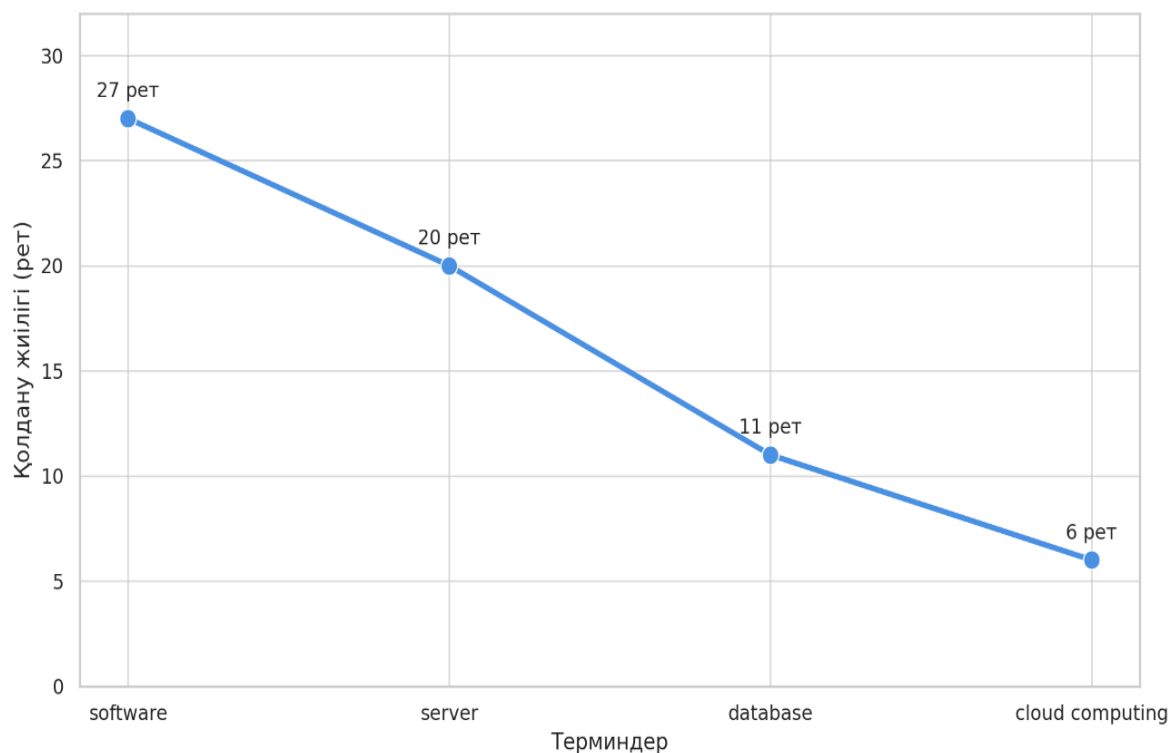
Термин (ағылш.)	Қазақша нұсқасы	Қолданыс жиілігі	Ескертпе
software	бағдарламалық жасақтама	27	Кей мәтіндерде «софт» ретінде қолданылған
server	сервер	20	«серверлік жүйе» тіркесімен де кездесті
database	дерекқор	11	Кей мәтіндерде «база данных» нұсқасы да қолданылған
cloud computing	бұлтты есептеу	6	Сирек қолданыста; «клауд» нұсқасы бейресми стильде көрінеді

2-кесте IT терминдерінің қазақ мәтіндеріндегі таралу деңгейін салыстырмалы түрде бағалауға мүмкіндік береді. Software терминінің жиілігі басқа терминдерден айтарлықтай жоғары. Бұл оның ақпараттық технологиялар саласындағы негізгі ұғым және қазақ тілді медиада кең таралған категория екенін көрсетеді. Екінші орында тұрған server термині инфрақұрылымдық жүйелерді сипаттайтын мәтіндерде жиі кездеседі. Оның «серверлік жүйе» тіркесімен қолданылуы қазақша терминжасам үрдісінің контекстке бейімделуін дәлелдейді. Database термині бірізді қолданыла бермейді: бірқатар мәтіндерде «дерекқор», енді бірінде «база данных» нұсқасы қолданылған. Бұл қазақ тіліндегі терминологиялық норма әлі тұрақтану үстінде екенін және орыс тілінің ықпалы кей салаларда сақталғанын көрсетеді. Ал cloud computing термині сирек қолданыста — бұның себебі салалық тақырыптардың қазақ тілді медиада аз берілуімен және бұл ұғымның терминологиялық жүйеде кеш пайда болуымен байланысты.

Диаграмма кестеде (1-сурет) келтірілген деректерді визуалды түрде көрсетіп, IT лексикасының таралу деңгейлерін салыстыруға мүмкіндік береді. Software терминінің басым қолданылуы графикте айқын көрінеді; server термині одан сәл төмен деңгейде орналасқан; database пен cloud computing терминдері төменгі жиілікті көрсетеді.

Бұл жиілік талдауы IT саласындағы қазақ тілінің даму бағыты туралы бірнеше маңызды тұжырым жасауға мүмкіндік береді. Ең алдымен, терминдердің таралуы біркелкі емес: кей ұғымдар кеңінен қолданылса, басқалары әлі тұрақты мәтіндік қолданысқа орныға қойған жоқ. Қазақша балама беру тәжірибесі де әркелкі. Әсіресе *database* → *дерекқор* / *база данных* сияқты параллель нұсқалардың қатар жүруі терминологиядағы нормаланбаған аймақтардың

бар екенін көрсетеді. Жаңа терминдердің медиада таралуы контенттің жанры мен типіне тәуелді: техникалық блогтарда және ғылыми-публицистикалық мәтіндерде олардың қолданылу жиілігі айтарлықтай жоғары. Сонымен бірге лексикалық бейімделу қарқыны да жоғары: *software* → *бағдарламалық жасақтама* сияқты баламалар кеңінен қолданылып, жаңа тілдік нормалардың қалыптасуына негіз болып отыр. IT саласына қатысты лексикалық өзгерістер қазақ тілінің жалпы даму бағытының бір көрінісі ғана. Бұл үрдістер тілдің технологиялық ортаға бейімделу қабілетінің жоғары екенін және жаңа ұғымдарды тез қабылдай алатын динамикалық жүйе екенін дәлелдейді.



1-сурет – IT-терминдердің жалпы жиілігі (50 мәтінде)

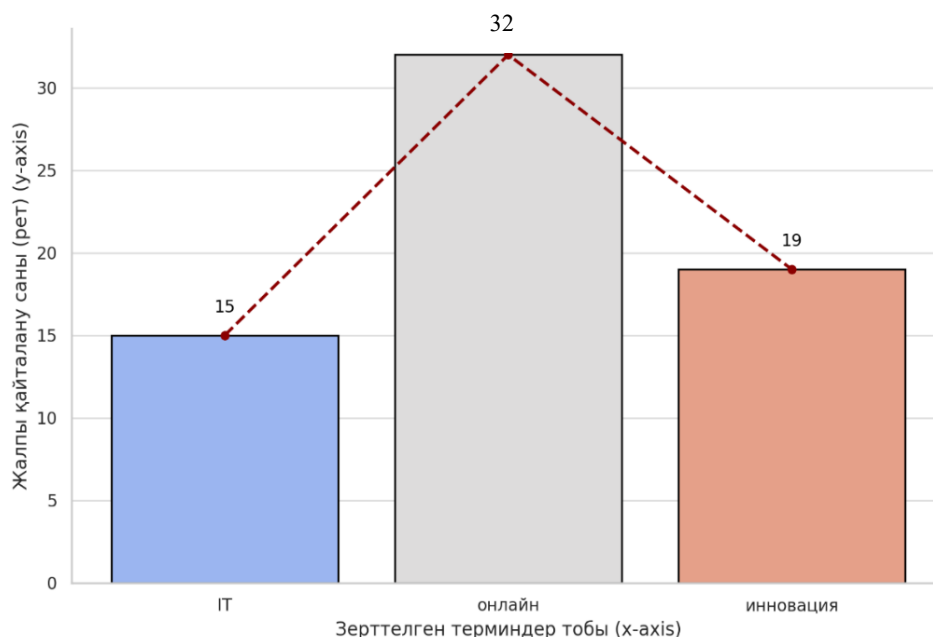
Терминологиялық көпнұсқалылық та зерттеудің маңызды нәтижелерінің бірі. Бір ұғымның бірнеше баламамен берілуі – қазақ тіліндегі терминжасам үдерісінің әлі тұрақтанбағанын көрсетеді. Мысалы, *software* терминінің «бағдарламалық жасақтама», «бағдарламалық қамтамасыз ету», «бағдарламалар жиынтығы» және «софт» түрінде қолданылуы салалық терминологияның әркелкілігін аңғартады. Мұндай көпнұсқалылық коммуникациядағы түсінбеушілікке де себеп болуы мүмкін.

Орфографиялық талдау жаңа сөздердің жазылымында бірізді норма толық қалыптаспай отырғанын көрсетті. Материалдар ішінде «онлайн оқыту», «онлайн-оқыту», «онлайноқыту» сияқты үш түрлі жазылым қатар кездеседі. Ең жиі қолданылатын нұсқа – «онлайн оқыту», дефисті нұсқа сиректеу, ал тұтас жазылған нұсқа жекелеген мысалдармен шектеледі. Бұл емле нормасының тұрақтану кезеңінде екенін дәлелдейді.

Диахрониялық талдау қазақ тілінің кейінгі отыз жылдағы өзгерісін нақты көрсетті. 1990 жылдары сирек кездескен «компьютер», «интернет», «онлайн» сияқты сөздер бүгінгі күні негізгі сөздік қорға енген. «Әлеуметтік желі», «вебинар», «цифрлық орта» тәрізді сөздер қазіргі кезеңде жиі қолданылатын лексикаға айналды. Бұл құбылыстар қазақ тілінің жаңа технологиялық, мәдени және әлеуметтік ортаға тез бейімделетін қабілетін көрсетеді.

Заманауи ақпараттық технологиялар мен жаһандану үдерістері қазақ тілінің лексикалық жүйесіне айтарлықтай ықпал етіп отыр. Соңғы онжылдықтарда «интернет», «онлайн», «инновация» сияқты сөздердің жылдам таралуы олардың тұрмыстық сөйлеуде де, ғылыми мәтіндерде де белсенді жұмсалыуына жағдай жасады. Бұл құбылыс жаңа лексиканың тек салалық ортада емес, жалпы тілдік қолданыста орнығып келе жатқанын көрсетеді.

Мысалы, 2-суретте 100 мәтінде «онлайн» сөзінің орташа жиілігі 32 рет болса, «инновация» сөзі орташа 19 рет кездескені байқалды. Бұл деректер жаңа терминдердің қолданылуы біршама тұрақты және өсіп келе жатқанын дәлелдейді.



2-сурет – Жаңа терминдердің жиілік көрсеткіштері (N = 100 мәтін)

Ескерту: Диаграммада (y-axis) әр терминнің жалпы қайталану саны көрсетілген; (x-axis) – зерттелген терминдер тобы.

Зерттелген мәтіндердің бірқатарында бір ұғымға бірнеше түрлі аударма нұсқасы қатар қолданылады. Мысалы, «software» – «бағдарламалық жасақтама», «бағдарламалық қамтамасыз ету», «бағдарламалар жиынтығы» немесе «софт» деп те кездеседі. Мұндай алақұлалық тілдік коммуникацияны күрделендіреді және бірыңғай стандарт қабылдаудың өзектілігін көрсетеді.

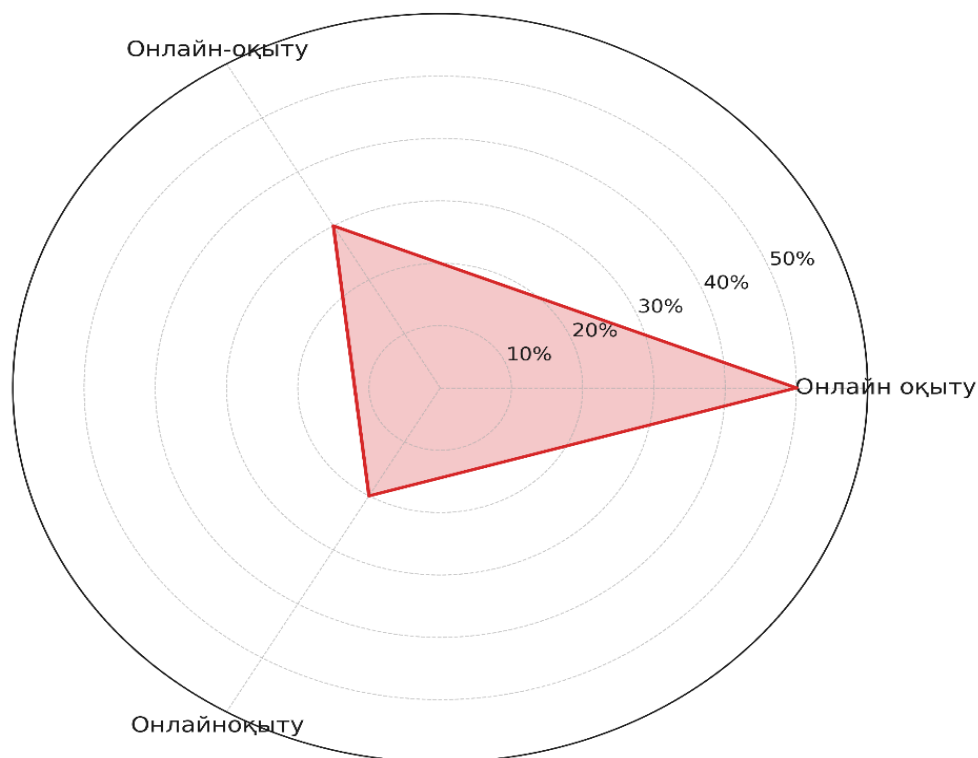
3-кесте – “Software” терминінің қазақша аудармалары және жиілігі (N = 50 мәтін)

Аударма нұсқасы	Қолданыс жиілігі (рет)
Бағдарламалық жасақтама	27
Бағдарламалық қамтамасыз ету	15
Бағдарламалар жиынтығы	6
Софт	9

3-кестеден көрінгендей, «бағдарламалық жасақтама» нұсқасы ең көп қолданылса, «бағдарламалық қамтамасыз ету», «софт» және «бағдарламалар жиынтығы» да мәтіндерде қатар кездеседі. Бұл – бір ұғымға байланысты бірнеше атаудың қатар өмір сүріп жатқанын

және терминологиялық жүйеде көпнұсқалықтың әлі де сақталып отырғанын көрсететін дәлел.

Жаңа сөздердің түрліше жазылуы – емле нормадағы бірізділіктің әлі толық қалыптаспауынан туындайды. «Онлайн-оқыту» мен «онлайн оқыту», «онлайноқыту» сияқты нұсқалар бір уақытта қатар жүруде (3-сурет). Бұл құбылыстарға тілдік реформалар, латын графикасына көшу талпынысы, жаңа сөздердің халық арасында тез таралуы сияқты факторлар да ықпал етеді.



3-сурет – Орфографиялық нұсқалардың жиілік үлесі

Ескерту: “n=100” – түрлі БАҚ және блог мәтіндерінен жиналған мысалдар саны.

Бұл деректер, орташа есеппен алғанда, әрбір 5 мәтіннің 2-інде «онлайн оқыту» (сыртқы жалғаусыз), 1-інде «онлайн-оқыту» (дефис арқылы), ал сирек жағдайда (~20%) «онлайноқыту» деп жазылатынын айқындады. Осылайша, «орфографиялық көпнұсқалық» жөніндегі гипотезамыз да расталды.

Ескі мақалалар мен қазіргі заман мәтіндерін салыстыру арқылы қазақ тілінің соңғы 30 жыл ішіндегі трансформациясын байқауға болады. Бұрынғы бірен-саран жаңа сөздер енді көпшілік қолданысқа шығып, орын тепті. Бұл тілдің жаһандық үрдістерге тез икемделіп отыратындығын дәлелдейді.

Қорытынды

Жүргізілген талдаулар мен кестелік-диаграммалық иллюстрациялар қазақ тілінің соңғы кезеңдердегі қарқынды өзгерістерге бейім екендігін дәлелдейді. Ғылыми-публицистикалық, бұқаралық ақпарат құралдары, әдеби мәтіндер және салалық терминологиялық сөздіктер негізінде жасалған салыстырулар:

- Жаңа сөздердің өте жылдам қалыптасатынын және кең таралатынын (мысалы, «онлайн», «блокчейн»);

- Терминологиядағы бірнеше нұсқаның қатар өмір сүретінін (мысалы, «дерекқор» ↔ «база данных», «бағдарламалық жасақтама» ↔ «софт»);

- Орфографиялық нормаларды қатаң реттеудің қажеттігін («онлайн оқыту» ↔ «онлайн-оқыту» ↔ «онлайноқыту»);

- Уақыт ағымындағы (1990–2020 жылдар) тілдік қолданыстың өзгерісін (жаңа технология мен жаһандану әсері) нақты мысалдармен растайтынын көрсетеді.

Зерттеу нәтижелері қазақ тіліндегі терминжасам, орфография, неологизмдерді игеру және диахрониялық трансформация салаларында бірқатар өзекті проблемалардың бар екенін айқындаса да, бұл мәселелерді кешенді түрде шешу мүмкін екенін де дәлелдейді. Мемлекеттік тіл саясаты, ғылыми орталар, терминком, БАҚ және білім беру жүйесі бірлесе әрекет еткен жағдайда тілдің даму барысы бірізділеніп, қалыпты арнаға түседі. Ашық деректер саясатын күшейту, шағын корпустарды біріктіріп, болашақта ұлттық корпус жасау – қазақ тілін жаңа деңгейге көтерудің негізгі тетіктерінің бірі болмақ.

Сөйтіп, осы жұмыста ұсынылған мысалдар мен кестелер, диаграммалар мен блок-схемалық визуализациялар – зерттеудің тек шағын көрінісі. Толыққанды нәтижеге қол жеткізу үшін, бір жағынан, материалды кеңейту, екінші жағынан, мамандандырылған тілдік технологияларды (машиналық аударма, автоматты морфологиялық талдау, семантикалық іздеу жүйелері) жетілдіру қажет. Дегенмен, бүгінгі қолдағы мүмкіндіктерді де орынды пайдаланса, қазақ тілінің заманауи даму үрдістерін бағамдауға жеткілікті деңгейде ақпарат жинауға болатыны анық. Міне, осындай жан-жақты талдау арқылы қазақ тілінің болашағын айқындауға үлес қосу – негізгі мақсаттардың бірі.

Осылайша, *тіл корпусын арнайы жасақтамай-ақ* қазақ тілін талдауға бағытталған зерттеу жұмысы көптеген қызықты нәтижелер мен қорытындылар береді. Ұсынылып отырған әдістер – *іріктемелі талдау, веб-скрепинг, диахрониялық салыстыру, лексикографиялық және терминологиялық шолу* – шағын ресурстармен де үлкен мәселелерді шешуге көмектеседі. Тілдің цифрлық ортадағы үлесін арттыру үшін тек мемлекеттік органдар ғана емес, тілтанушылар, IT-мамандар және қоғам белсенділері бірлесіп еңбек етуі қажет. Осы үдерісте алынатын ғылыми-тәжірибелік нәтижелер болашақта ауқымды корпус құруға да, терминологияны біріздендіруге де, контентті автоматтандыруға да мықты негіз болатындығы даусыз.

Сондықтан қазіргі қолда бар мүмкіндіктерді тиімді қолданып, қазақ тілінің өрісін кеңейту және оның бәсекеге қабілеттілігін нығайту – ең басты міндеттердің бірі. Бұл міндетті абыроймен атқару үшін ұсынылған зерттеу әдістері мен нәтижелерінің көмегі айтарлықтай болмақ.

ҚР ҒжЖБМ Ғылым комитетінің бағдарламалық-нысаналы қаржыландыру жобасы «Үздіксіз білім берудегі халықаралық және ұлттық стандарттарға сәйкес лингвокоммуникация мен нейролингвистика негізінде мемлекеттік тілді деңгейлік меңгерудің әлеуметтік-педагогикалық парадигмасы» (BR28713427) аясында жарияланды.

ПАЙДАЛАНЫЛҒАН ӘДЕБИЕТТЕР ТІЗІМІ

1. Hunston S. Corpora in Applied Linguistics. – Cambridge: Cambridge University Press, 2002. – 214 p.
2. McEnery T., Hardie A. Corpus Linguistics: Method, Theory and Practice. – Cambridge: Cambridge University Press, 2012. – 278 p.

3. Makhambetov O., Kokenbayev Y., Yessenbayev Z., et al. Assembling the Kazakh Language Corpus // Proceedings of ACL. – 2021. – P. 1–12.
4. Сыздық Р. Тіл табиғаты және оның зерттелуі. – Алматы: Ғылым, 2000. – 320 б.
5. Cabré M.T. Terminology: Theory, Methods and Applications. – Amsterdam: John Benjamins, 1999. – 356 p.
6. Ақылбекова Г.К., Жанғабылова З.К. Салалық терминология және терминжасам қағидаттары // ҚР Заңнама және құқықтық ақпарат институтының хабаршысы. – 2023. – №1 (72). – Б. 224–232.
7. Kim H. Machine Learning Methods for Turkic Languages // Language Resources and Evaluation. – 2022. – Vol. 56, No. 3. – P. 251–268.
8. Мединаева А.А. Қазақ тіліндегі кейбір терминдердің мәселелеріне орай // ҚР Заңнама және құқықтық ақпарат институтының хабаршысы. – 2023. – №2 (73). – Б. 222–228.
9. Maybury M. New Directions in Automated Text Summarization // Information Processing and Management. – 1999. – Vol. 35, No. 4. – P. 491–514.
10. Bowker L., Pearson J. Working with Specialized Language: A Practical Guide to Using Corpora. – London: Routledge, 2002. – 256 p.
11. Wu Y., Schuster M., Chen Z., Le Q.V., Norouzi M., & Macherey W. Google's Neural Machine Translation System: Bridging the Gap Between Human and Machine Translation // arXiv preprint. – 2016. – arXiv:1609.08144.
12. Chomsky N. Aspects of the Theory of Syntax. – Cambridge, MA: MIT Press, 2015. – 270 p.

REFERENCES

1. Hunston S. Corpora in Applied Linguistics. – Cambridge: Cambridge University Press, 2002. – 214 p.
2. McEnery T., Hardie A. Corpus Linguistics: Method, Theory and Practice. – Cambridge: Cambridge University Press, 2012. – 278 p.
3. Makhambetov O., Kokenbayev Y., Yessenbayev Z., et al. Assembling the Kazakh Language Corpus // Proceedings of ACL. – 2021. – P. 1–12.
4. Syzdyq R. Til tabigaty zhane onyn zerttelui [The nature of language and its study]. – Almaty: Gylym, 2000. – 320 b. [in Kazakh]
5. Cabré M.T. Terminology: Theory, Methods and Applications. – Amsterdam: John Benjamins, 1999. – 356 p.
6. Aqylbekova G.K., Zhangabylova Z.K. Salalyq terminologia zhane terminzhasam qagidattary [Industry terminology and terminology principles] // QR Zannama zhane qyqyqyq aqparat institutynyn habarshysy. – 2023. – №1 (72). – Б. 224–232. [in Kazakh]
7. Kim H. Machine Learning Methods for Turkic Languages // Language Resources and Evaluation. – 2022. – Vol. 56, No. 3. – P. 251–268.
8. Medinaeva A.A. Qazaq tilindegi keibir terminderdin maselelerine orai [In connection with the problems of some terms in the Kazakh language] // QR Zannama zhane quqyqyq aqparat institutynyn habarshysy. – 2023. – №2 (73). – Б. 222–228. [in Kazakh]
9. Maybury M. New Directions in Automated Text Summarization // Information Processing and Management. – 1999. – Vol. 35, No. 4. – P. 491–514.
10. Bowker L., Pearson J. Working with Specialized Language: A Practical Guide to Using Corpora. – London: Routledge, 2002. – 256 p.
11. Wu Y., Schuster M., Chen Z., Le Q.V., Norouzi M., & Macherey W. Google's Neural Machine Translation System: Bridging the Gap Between Human and Machine Translation // arXiv preprint. – 2016. – arXiv:1609.08144.
12. Chomsky N. Aspects of the Theory of Syntax. – Cambridge, MA: MIT Press, 2015. – 270 p.