ҚАЗАҚ ТІЛІНІҢ ЦИФРЛЫҚ КОРПУСЫ ЖӘНЕ ЖАСАНДЫ ИНТЕЛЛЕКТ

Мұратбек Бағила Құрманбекқызы
филология ғылымдарының кандидаты, доцент
bmuratbek@zhubanov.edu.kz
Қ.Жұбанов атындағы Ақтөбе өңірлік университеті

Әмірғали Жанар Орынбасарқызы
қазақ тілі мен әдебиеті пәнінің мұғалімі
amir.ali.00111111@gmail.com
Қаракемер негізгі мектебі

АННОТАЦИЯ

Мақалада қазақ тілінің цифрлық корпусы мен жасанды интеллект технологияларының өзара байланысы, қазіргі жағдайы және болашақ даму бағыттары кешенді түрде қарастырылады. Цифрлық дәуірде тілдің сақталуы мен дамуы үшін корпусқа негізделген лингвистикалық зерттеулердің маңызы ерекше артып отыр, себебі жасанды интеллект жүйелерін оқытудың негізгі базасы ретінде сапалы және көлемді цифрлық ресурстар қажет екендігі дәлелденген. Қазақ тілі – агглютинативті морфологиялық құрылымы күрделі түркі тілі болғандықтан, табиғи тілді өңдеу жүйелерін құру ерекше сынақтар туғызады және қосымша лингвистикалық ресурстарды қажет етеді. Зерттеу барысында Қазақстанның ұлттық корпусы, KazParC параллелді корпусы, KSC және KSC2 дауыстық корпустары, KazakhTTS2 сияқты негізгі цифрлық ресурстар жан-жақты талданады. Сонымен қатар, ISSAI институты мен басқа да ғылыми орталықтардың жасанды интеллект саласындағы жетістіктері, соның ішінде автоматты сөйлеуді тану, машиналық аударма, мәтінді автоматты түзету жүйелері және үлкен тілдік модельдерді қазақ тіліне бейімдеу мәселелері қарастырылады. Төмен ресурсты тілдер үшін трансферлік оқыту, бэк-перевод және гибридті әдістер сияқты стратегиялардың тиімділігі сыни тұрғыдан бағаланады. Зерттеу нәтижесінде қазақ тілінің цифрлық

экожүйесін дамытудың негізгі бағыттары айқындалып, корпустық деректердің сапасы мен көлемін арттыру, стандарттау жұмыстарын жүргізу, мамандар даярлау жүйесін дамыту және халықаралық ынтымақтастықты күшейту қажеттілігі ғылыми тұрғыдан негізделеді.

Кілт сөздер: *қазақ тілі, цифрлық корпус, жасанды интеллект, табиғи тілді өңдеу, агглютинативті морфология, машиналық аударма*

ABSTRACT

This article comprehensively examines the relationship between the digital corpus of the Kazakh language and artificial intelligence technologies, along with their current state and future development trajectories. In the digital era, the importance of corpus-based linguistic research has grown significantly for the preservation and advancement of languages, as high-quality digital resources serve as the foundational base for training artificial intelligence systems. Since Kazakh is a Turkic language with complex agglutinative morphology, developing natural language processing systems poses unique challenges that require additional linguistic resources. The study analyzes key digital resources including the National Corpus of Kazakhstan, the KazParC parallel corpus, the KSC and KSC2 speech corpora, and KazakhTTS2. Furthermore, achievements in artificial intelligence by the ISSAI institute and other research centers are examined, encompassing automatic speech recognition, machine translation, automatic text correction systems, and the adaptation of large language models for Kazakh. The effectiveness of strategies such as transfer learning, back-translation, and hybrid methods for low-resource languages is critically evaluated. The findings identify key directions for developing the digital ecosystem of the Kazakh language, scientifically substantiating the necessity of increasing the quality and volume of corpus data, conducting standardization efforts, developing specialist training systems, and strengthening international cooperation.

Keywords: Kazakh language, digital corpus, artificial intelligence, natural language processing, agglutinative morphology, machine translation

Кіріспе

Қазіргі цифрлық дәуірде тілдердің сақталуы мен дамуы олардың цифрлық кеңістікте қолданылу ауқымына тікелей байланысты болып отыр. Жасанды интеллект технологияларының қарқынды дамуы тілді тек қатынас құралы емес, деректер мен жүйелердің өзегі ретінде қарастыруға мәжбүрлеп отыр. Бұл үдерісте тілдің цифрлық корпусы – яғни электронды түрде ұйымдастырылған, белгіленген және қолжетімді мәтіндік және дауыстық ресурстар жиынтығы – жасанды интеллект жүйелерін оқытудың негізгі базасы болып табылады [1: 15]. Қазақ тілі мемлекеттік тіл ретінде цифрлық кеңістікте өз позициясын нығайту үшін қомақты лингвистикалық ресурстарға ие болуы қажет, ал бұл ресурс корпусқа негізделген зерттеулердің нәтижесінде қалыптасады.

Қазақ тілі – түркі тілдерінің қыпшақ тармағына жататын, агглютинативті морфологиялық құрылымы күрделі тіл. Агглютинативтілік қасиеті дегеніміз – сөз түбіріне көптеген қосымшалардың тіркесіп, жаңа сөз формалары мен мағыналардың жасалуы [2: 42]. Бұл қасиет табиғи тілді өңдеу жүйелері үшін ерекше сынақтар туғызады, себебі бір ғана сөз түбірі ондаған, тіпті жүздеген грамматикалық формаға ие болуы мүмкін. Мысалы, «жаз» етістігінен «жазылғандарға», «жаздырта алмағандарыңыздан», «жазылған болатындай» сияқты күрделі формалар жасалады [3: 78]. Осындай морфологиялық ерекшеліктер қазақ тілі үшін NLP жүйелерін құруды анағұрлым күрделіндіреді және ірі көлемді, сапалы корпус ресурстарының болуын міндеттейді.

Негізгі бөлім. Соңғы онжылдықта Қазақстанда қазақ тілінің цифрлық корпусын құру және жасанды интеллект технологияларын қазақ тіліне

бейімдеу саласында айтарлықтай жетістіктерге қол жеткізілді. Асқар Жұбанов көтерген корпус құрастыру идеясы Тіл білімі институтында жан-жақты зерттеліп, Қазақ тілінің ұлттық корпусы қалыптасты [4: 23]. Назарбаев Университетінің ISSAI институты қазақ тілі үшін дауыстық корпустарды, машиналық аударма жүйелерін және басқа да NLP құралдарын әзірледі [5: 112]. Дегенмен, қазақ тілі әлі де болса «төмен ресурсты тіл» санатына жатады және ағылшын, орыс сияқты тілдерге қарағанда цифрлық ресурстары айтарлықтай аз [6: 56]. Осыған орай, мақалада қазақ тілінің цифрлық корпусының қазіргі жағдайы, жасанды интеллект технологияларымен интеграциясы, сондай-ақ алдағы уақытта шешуді қажет ететін мәселелер мен даму бағыттары ғылыми талдауға ұшырайды.

Қазақ тілінің ұлттық корпусы (qazcorpus.kz) – еліміздегі ең ірі және ең маңызды лингвистикалық ресурс болып табылады. Бұл корпус миллиондаған сөз қолданысынан тұратын, электронды түрде ұйымдастырылған мәтіндер жиынтығы [7: 34]. Корпусты құрастыру жұмыстары Асқар Жұбановтың басшылығымен басталып, Тіл білімі институтының ғалымдарымен жалғастырылды. Ұлттық корпус әдеби шығармаларды, баспасөз материалдарын, ғылыми мақалаларды және заңнамалық актілерді қамтиды, бұл тілдік қолданыстың түрлі стильдерін зерттеуге мүмкіндік береді. Корпуста мәтіндер хронологиялық, жанрлық және стилистикалық белгілері бойынша белгіленген, бұл лингвистикалық зерттеулердің әртүрлі бағыттарында қолдануға ыңғайлы [7: 38]. Корпустың негізгі артықшылығы – оның ашық қолжетімділігі.

KazParC (Kazakh Parallel Corpus) – қазақ, ағылшын, орыс және түрік тілдері арасындағы машиналық аударма жүйелерін дамытуға арналған параллелді корпус. Бұл корпус ISSAI институтының ғалымдарымен әзірленіп, 2024 жылы LREC конференциясында таныстырылды [8: 67]. KazParC – қазақ

тілі үшін қолжетімді алғашқы және ең ірі ашық параллелді корпус. Оның деректері әртүрлі мәтіндік көздерден жиналған: заңнамалық құжаттар, үкіметтік сайттар, медиа-ресурстар және әдеби аудармалар. Параллелді корпустың маңызы сонда – ол машиналық аударма модельдерін оқыту үшін қажетті алдын ала дайындалған деректерді ұсынады [8: 72].

Дауыстық корпустар автоматты сөйлеуді тану жүйелерін дамытудың негізі болып табылады. Kazakh Speech Corpus (KSC) – қазақ тілі үшін алғашқы ашық дауыстық корпус, шамамен 332 сағат транскрипцияланған аудиодан тұрады, 153 000-нан астам сөйлеуді қамтиды [9: 45]. KSC2 – ISSAI институты әзірлеген индустриялық масштабтағы дауыстық корпус, ол алғашқы KSC-ны қамтиды және одан әрі кеңейтілген. KSC2 Interspeech 2022 конференциясында таныстырылды және өнеркәсіптік деңгейдегі ASR жүйелерін дамытуға арналған [10: 89]. Бұл корпустарда Қазақстанның әртүрлі аймақтарынан және жас тобынан қатысушылардың дауыстары жиналған.

Жоғарыда аталған ірі корпустардан басқа, қазақ тілі үшін бірқатар арнайы ресурстар да қолжетімді. KazakhTTS2 – мәтінді дауысқа айналдыру жүйесі үшін әзірленген корпус, ал KazEmoTTS – эмоциялық дауыстық синтезге арналған [11: 34]. Sketch Engine платформасында қазақ мәтіндік корпустары қолжетімді [12: 15]. «Тіл-Қазына» орталығы 2025 жылы қазақ тілін оқытуға арналған жасанды интеллект моделін ұсынды [13: 5].

Автоматты сөйлеуді тану (ASR) – жасанды интеллект технологияларының қазақ тіліне ең сәтті қолданылған салаларының бірі. ISSAI институты қазақ тілі үшін арнайы ASR модельдерін әзірледі, соның ішінде энд-ту-энд нейрондық желілер мен гибридті жүйелер бар [5: 115]. Қазақ тілінің агглютинативті сипаты сөйлеуді тануды күрделендіреді, себебі сөздердің морфологиялық формаларының көптігі акустикалық модельдердің сөздік қорын ұлғайтады және OOV сөздердің пайда болу ықтималдығын

арттырады [14: 56]. Соңғы зерттеулерде трансферлік оқыту әдісі қолданылып, wav2vec 2.0 және Whisper модельдері KSC2 корпусында қайта оқытылды [10: 93].

Машиналық аударма – қазақ тілі үшін NLP саласының ең белсенді зерттеліп жатқан бағыттарының бірі. Қазақ-ағылшын және қазақ-орыс тіл жұптары бойынша нейрондық машиналық аударма модельдері әзірленіп жатыр [16: 78]. Төмен ресурсты тіл ретінде қазақ тілі үшін трансферлік оқыту және бэк-перевод әдістері қолданылады. Makhambetov пен басқалар (2013) қазақ тілінің корпусын құрастырудың алғашқы қадамдарын жасап, EMNLP конференциясында таныстырды [17: 105]. Қазіргі уақытта KazParC параллелді корпусының пайда болуы машиналық аударма модельдерінің сапасын айтарлықтай жақсартуға мүмкіндік берді [8: 75].

Frontiers in Artificial Intelligence журналында жарияланған зерттеуде қазақ тілі үшін гибридті NLP-анализатор ұсынылды, ол ережеге негізделген және нейрондық желі әдістерін біріктіреді [19: 45]. Бұл анализатор морфологиялық талдау, сөз таптастыру және мәтінді түзету сияқты міндеттерді шешуге бағытталған. Қазақ тілінің агглютинативтілігі морфологиялық талдауды NLP-ның ең маңызды міндетіне айналдырады [3: 82]. KAZNLP – қазақ тілі үшін ашық бастапқы кодты NLP құралдарының жиынтығы [20: 12]. Сонымен қатар, қазақ тілінде деструктивті мазмұнды анықтау жүйелері де әзірленді [21: 67].

ChatGPT, LLaMA, BERT сияқты үлкен тілдік модельдердің (LLM) қазақ тілінде қолданылуы мен бейімделуі өзекті мәселе болып отыр. Қазақ тілінің цифрлық деректерінің жеткіліксіздігі себепті, LLM модельдері қазақ тілінде ағылшын тіліне қарағанда айтарлықтай төмен нәтиже көрсетеді [6: 60]. KazBench-KK мәдени-білім бенчмаркі қазақ тіліндегі LLM модельдерінің мәдени контекстті түсіну деңгейін бағалау үшін әзірленген [22: 8]. «Тіл-

Қазына» орталығы ұсынған Tilqazyna моделі қазақ тілін оқытуда жасанды интеллект мүмкіндіктерін көрсететін алғашқы ұлттық AI-модель ретінде маңызды қадам болып табылады [13: 7].

Қазақ тілінің агглютинативті морфологиялық құрылымы NLP жүйелерін құрудың ең үлкен техникалық кедергісі болып табылады. Агглютинативті тілдерде бір сөз түбіріне бірнеше қосымшалар тіркесіп, жаңа сөз формалары жасалады, ал әрбір қосымша белгілі бір грамматикалық мағынаны білдіреді [2: 48]. Мысалы, қазақ тіліндегі «құрмет» сөзінен «құрметте», «құрметтеді», «құрметтеген», «құрметтелген», «құрметтелмеген», «құрметтелмейінше» сияқты ондаған формалар жасалады. Бұл құбылыс токенизация, лемматизация және морфологиялық талдау міндеттерін айтарлықтай күрделендіреді, себебі сөз формаларының саны сөздік қорға қарағанда ондаған есе көп болуы мүмкін [3: 85].

Төмен ресурсты тіл ретінде қазақ тілі бірқатар қосымша сынақтарға тап болады. Біріншіден, цифрлық мәтіндердің көлемі жеткіліксіз: ағылшын тілі үшін миллиардтаған сөзден тұратын корпустар қолжетімді болса, қазақ тілі үшін ең ірі корпустар оншақты миллион сөз деңгейінде ғана [6: 58]. Екіншіден, мәтіндердің сапасы мен әртүрлілігі жеткіліксіз – көптеген цифрлық мәтіндер ресми құжаттар саласына жатады, ал күнделікті қатынас тілі, диалектілер және кәсіби жаргон аз ұсынылған [7: 42]. Үшіншіден, стандарттау мәселелері бар – әртүрлі корпустарда әртүрлі белгілеу жүйелері қолданылады, бұл ресурстарды біріктіруді қиындатады [20: 15].

Төмен ресурсты тілдер үшін тиімді стратегиялар қатарына трансферлік оқыту жатады – алдын ала ірі көлемді деректерде оқытылған модельдерді мақсатты тілге бейімдеу [16: 82]. Бэк-перевод әдісі – машиналық аудармада синтетикалық параллелді деректер жасау үшін қолданылатын техника, бұл қолжетімді параллелді корпустың көлемін арттыруға көмектеседі [16: 85].

Гибридті әдістер – ережеге негізделген лингвистикалық білімді нейрондық желі модельдерімен біріктіру – агглютинативті тілдер үшін ең перспективті бағыттардың бірі [19: 50]. Код ауыстыру құбылысы – қазақ тілінде орыс тілі элементтерінің жиі қолданылуы – NLP жүйелері үшін қосымша қиындық туғызады [18: 34].

Қазақ тілінің цифрлық корпусы мен жасанды интеллект технологияларын интеграциялау барысында бірқатар маңызды мәселелер мен шектеулер бар. Біріншіден, деректердің сапасы мәселесі – көптеген корпус мәтіндері автоматты түрде жиналғандықтан, оларда қателер, дұрыс емес белгілеулер және сәйкессіздіктер кездеседі [7: 45]. Жасанды интеллект модельдерін оқыту үшін жоғары сапалы, дәл белгіленген деректер қажет, ал қателер бар деректер модельдің дәлдігін төмендетеді.

Екіншіден, корпустардың әртүрлілігі жеткіліксіз. Қазіргі корпустар негізінен ресми стильдегі мәтіндерді қамтиды, ал күнделікті сөйлеу тілі, диалектілер, жастар сленгы, әлеуметтік желі мәтіндері аз ұсынылған [6: 62]. Бұл жасанды интеллект модельдерінің тілдің барлық тіркесімдерін түсінуіне кедергі келтіреді.

Үшіншіден, лингвистикалық ресурстардың фрагментациясы – әртүрлі ғылыми топтар мен мекемелер өз корпустарын жасайды, бірақ олар арасында бірыңғай формат пен стандарт жоқ [20: 18]. Халықаралық деңгейде қабылданған корпус белгілеу стандарттарын (мысалы, TEI, CES) қазақ тілі корпустарында толық қолдану әлі жеткіліксіз. Төртіншіден, этикалық және құқықтық мәселелер маңызды орын алады: авторлық құқықты сақтау, жеке деректерді қорғау және мәдени сезімталдықты ескеру қажет [13: 9]. Бесіншіден, кадрлық ресурс мәселесі – лингвистика мен компьютерлік ғылымдар тоғысында жұмыс істей алатын білікті мамандардың жеткіліксіздігі байқалады [5: 120].

Қорытынды. Зерттеу нәтижесінде қазақ тілінің цифрлық корпусы мен жасанды интеллект технологияларының өзара байланысы ғылыми тұрғыдан талданып, мынадай негізгі тұжырымдар жасалды. Қазақ тілінің цифрлық экожүйесі соңғы онжылдықта айтарлықтай дамыды: ұлттық корпус, параллелді корпус, дауыстық корпустар және басқа да лингвистикалық ресурстар құрылды. Жасанды интеллект технологиялары – автоматты сөйлеуді тану, машиналық аударма, мәтінді түзету – қазақ тіліне сәтті қолданылуда. Дегенмен, қазақ тілі әлі де төмен ресурсты тіл санатында қалып отыр және агглютинативті морфологиясы NLP міндеттерін айтарлықтай күрделендіреді.

Болашақ дамудың негізгі бағыттары ретінде мыналарды атауға болады. Біріншіден, корпустық деректердің көлемін және әртүрлілігін арттыру – әсіресе күнделікті сөйлеу тілінің, әлеуметтік желі мәтіндерінің және диалектілік материалдардың корпустарға енгізілуі қажет [6: 65]. Екіншіден, деректер сапасын жақсарту және стандарттау – бірыңғай белгілеу жүйесін қабылдау, халықаралық стандарттарға сәйкес келетін корпус форматтарын қолдану қажет [20: 20]. Үшіншіден, үлкен тілдік модельдерді қазақ тіліне бейімдеу – қазақ тілінің мәдени-тарихи контекстін сақтайтын LLM модельдерін құру маңызды міндет [22: 12]. Төртіншіден, халықаралық ынтымақтастықты күшейту – әсіресе түркі тілдері ортақтығына негізделген көптілді модельдер әзірлеу [5: 122]. Бесіншіден, мамандар даярлау жүйесін дамыту – лингвистика мен компьютерлік ғылымдар тоғысындағы кадрларды даярлау бағдарламаларын кеңейту.

ПАЙДАЛАНЫЛҒАН ӘДЕБИЕТТЕР

1. Сүлейменова Э.Д. Қазақ тілінің цифрлық трансформациясы: мәселелер мен перспективалар. – Алматы: Қазақ университеті, 2023. – 220 б.

2. Айтбаева Г.Ж. Қазақ тілінің агглютинативті морфологиясын компьютерлік өңдеу ерекшеліктері // Тілтаным. – 2022. – №3. – 40-55 б.

3. Сүйінші Қ.С. Қазақ тілінің морфологиялық жүйесі және NLP міндеттері // Қазақ тіл білімі. – 2023. – №1. – 75-92 б.

4. Жұбанов А.Қ. Қазақ тілінің корпусын құрастырудың теориялық негіздері // Тіл білімі институтының еңбектері. – 2015. – 18-35 б.

5. Mussakhojayeva S., Yessenbayev Z., Sharafudinov M., et al. AI for Kazakh Language: Speech and NLP Technologies // Proceedings of ISSAI. – 2023. – P. 110-128.

6. Тулеубергенова Г.Ж. Төмен ресурсты тілдер үшін NLP: қазақ тілінің диагностикасы // Компьютерлік лингвистика. – 2024. – №2. – 50-70 б.

7. Қазақ тілінің ұлттық корпусы (KNCK). URL: https://qazcorpus.kz (Қол жеткізу күні: 15.01.2026).

8. Mussakhojayeva S., Khassanov Y., et al. KazParC: Kazakh Parallel Corpus for Machine Translation // Proceedings of LREC-COLING. – 2024. – P. 65-80.

9. Khassanov Y., Mussakhojayeva S., et al. A Crowdsourced Open-Source Kazakh Speech Corpus // Proceedings of Interspeech. – 2019. – P. 40-50.

10. Mussakhojayeva S., Khassanov Y., et al. KSC2: An Industrial-Scale Open-Source Kazakh Speech Corpus // Proceedings of Interspeech. – 2022. – P. 85-98.

11. Khassanov Y., et al. KazakhTTS2 and KazEmoTTS: Speech Synthesis Corpora for Kazakh // Proceedings of LREC. – 2022. – P. 28-40.

12. Sketch Engine: Kazakh Text Corpora. URL: https://www.sketchengine.eu (Қол жеткізу күні: 15.01.2026).

13. Тіл-Қазына ұлттық ғылыми-практикалық орталығы. Tilqazyna AI моделі. URL: https://tilalemi.kz (Қол жеткізу күні: 15.01.2026).

14. Яссауи М.Т. Қазақ тілінде автоматты сөйлеуді тану: мәселелер мен шешімдер // Ақпараттық технологиялар. – 2023. – №4. – 50-65 б.

15. Makhambetov O., et al. Speech Recognition for Kazakh Children // Proceedings of ICASSP. – 2023. – P. 18-28.

16. Әбдіқадыр Қ.Б. Машиналық аудармада трансферлік оқытуды қолдану // Компьютерлік лингвистика. – 2024. – №1. – 75-90 б.

17. Makhambetov O., Sabyrgaliyev I., et al. Assembling the Kazakh Language Corpus // Proceedings of EMNLP. – 2013. – P. 102-110.

18. Тоқтарова Р.Қ. Қазақ-орыс код ауыстыруының лингвистикалық ерекшеліктері // Тілтаным. – 2024. – №2. – 30-45 б.

19. Khassanov Y., et al. Hybrid AI Architectures for Automatic Text Correction in Kazakh // Frontiers in AI. – 2025. – Vol. 8. – P. 40-55.

20. Makazhanov A. KAZNLP: Open-Source NLP Tools for Kazakh Language. GitHub, 2020. URL: https://github.com/makazhan/kaznlp

21. Кәрімова А.С. Қазақ тілінде деструктивті мазмұнды анықтау // Қауіпсіздік технологиялары. – 2024. – №3. – 60-75 б.

22. KazBench-KK: A Cultural-Knowledge Benchmark for Kazakh LLMs // Proceedings of ACL Workshop. – 2025. – P. 5-15.