ISSN 2788-7979 (online)

МҒТАР 16.01.11 DOI: 10.59102/kufil/2026/iss3pp228-242

Г.Ш. Шахманова[1], О.А. Жумагулова[1]
[1]Ш. Уәлиханов атындағы Көкшетау университеті, Көкшетау, 020000, Қазақстан

BALAUSA: ҚАЗАҚТІЛДІ БАЛАЛАРДЫҢ СӨЙЛЕУ ТІЛІНІҢ КОРПУСЫН ҚҰРУДЫҢ ТЕОРИЯЛЫҚ, ӘДІСТЕМЕЛІК ЖӘНЕ ТИПОЛОГИЯЛЫҚ НЕГІЗДЕРІ

Мақала заманауи онтолингвистиканың өзекті бағыттарының бірі – ерте жастағы қазақтілді балалардың сөйлеу тілінің корпусын құрудың ғылыми-теориялық, әдіснамалық және практикалық негіздері айқындалады. Зерттеу шеңберінде «Balausa» аудиожазбалар корпусы ұсынылып, оның ерте тілдік даму заңдылықтарын зерделеудегі эмпирикалық маңызы негізделеді. Зерттеу нысаны ретінде табиғи қарым-қатынас жағдайында жазылып алынған 1-4 жас аралығындағы 12 баланың сөйлеу тілінің аудиожазбалары алынды. Деректерді өңдеу барысында халықаралық CHILDES базасының стандарттарына сәйкес транскрипцияланып, көпқабатты морфосинтаксистік аннотациялау әдістері қолданылды.

Қазақ тілінің агглютинативті табиғаты ескеріліп, түркі тілдес балалар корпустарының тәжірибесіне сүйене отырып, қазақ тіліне тән параметрлерді таңбалаудың ерекшеліктері қарастырылды. Сонымен қатар, мақалада балалар сөйлеу тілінің корпусын құрудың негізгі кезеңдері сипатталып, жобаның қазіргі таңда бастапқы сатысында жүзеге асырылып жатқаны көрсетіледі. Зерттеу қазақ тілінің цифрлық онтолингвистикалық моделін қалыптастыруға және балалардың тілдік құзыреттілігінің даму траекториясын айқындауға мүмкіндік береді.

Кілт сөздер: онтолингвистика, онтогенез, корпустық зерттеу, балалар тілі, аудиокорпус, CHILDES, транскрипциялау.

НЕГІЗГІ ЕРЕЖЕЛЕР

Балалар тілі – адамзат тілінің табиғатын, құрылымдық заңдылықтарын және онтогенездегі қалыптасу кезеңдерін зерттеуде ерекше мәнге ие күрделі құбылыс. Ерте жастағы балалардың тілдік дамуы адамның когнитивтік, әлеуметтік және коммуникациялық қабілеттерінің қалыптасуымен тығыз байланысты. Яғни, бұл қарапайым дыбыстардан басталып, күрделі ойлау жүйесіне дейін жалғасатын көп сатылы процесс.

«Тіл – күрделі мәселе. Дегенмен, 5 жасқа дейінгі кішкентай балалар өз тілінің грамматикасында кездесетін күрделі жүйенің көпшілігін біледі. Олар сандарды бір-біріне қосуды немесе аяқ киімнің бауын байлауды үйренбестен бұрын, етістіктер мен зат есімдердің орнын ауыстырып, сұрақтарды жүйелеп, сөйлемдерді болымсыз түрге айналдырып, есімдіктерді дұрыс қолданып, бұрын естімеген сөйлемдерді қиындықсыз ұға алады» [1]. Бұл құбылыс баланың тілдік қабілеті өмірдің алғашқы жылдарында қарқынды дамитынын көрсетеді. Сондықтан ерте жастағы балалардың сөйлеу тілін зерттеу олардың тілдік жүйені қалай меңгеретінін, сөздік қоры мен грамматикалық құрылымдарының қалай қалыптасатынын анықтауға мүмкіндік береді.

Қазақ тіл білімінде бала тілін зерттеу онтолингвистикалық бағыт ретінде жүйелі қалыптаспағанмен, балалардың тілдік дамуы мен сөйлеу әрекеттері жайлы зерттеулер бар. Мәселен, қазақ тіл білімінде сөйлеу әрекетінің онтогенезі мен бала тілінің қалыптасу, даму үдерісі ф.ғ.к., профессор Т.Т. Аяпованың ғылыми еңбектерінде дербес зерттеу нысаны ретінде қарастырылды. Ғалым Таңат Аяпованың «Развитие синтаксиса речи детей (на материале казахского языка)» [2] атты еңбегінде бала тілін синтаксистік семантика тұрғысынан талдап, баланың сөйлеуіндегі бір сөзді құрылымдардан бастап, екі және үш сөзден тұратын сөйлемдерге дейінгі даму кезеңдерін жүйелі түрде сипаттайды.

ISSN 2788-7979 (online)

Сөйлеу әрекетінің онтогенезін зерделеу аясында Б.Б. Баймұратованың «Мектеп жасына дейінгі балаларды сөйлеуге үйрету» [3] атты еңбегін атап өтуімізге болады. Зерттеуші ерте мектеп жасына дейінгі балалардың лексикалық дамуының өзіндік ерекшеліктерін талдап, қазақ тілінде тәрбие берілетін балабақшалардағы 3–5 жас аралығындағы және мектепке дейінгі жастағы балалардың сөйлеу дағдыларын қалыптастырудың психологиялық-педагогикалық негіздерін айқындайды.

Сонымен қатар қалалық балабақшалар жағдайында орыс және қазақ тілінде сөйлейтін балаларды қазақ тіліне оқыту мәселесіне арналған зерттеулер жүргізіп, мемлекеттік тілдің болашағын қамтамасыз етудің маңызды алғышарты ретінде қазақ тілін оқытуды ерте жастан бастаудың қажеттігін атап көрсеткен зерттеуші А.Б. Бакраденованың [4] және ерте және мектеп жасына дейінгі қалыпты дамудағы балалардың қазақ тіліндегі сөйлеу әрекетінің онтогенезін зерттеп, сөйлеу тілінің даму кезеңдерін айқындап зерделеген зерттеушілер Г.Ж. Есенжолова мен А.Д. Сейсенованың еңбектерін де атап өтуімізге болады.

Балалар тілін зерделеуге қатысты соңғы жылдары шыққан ғылыми негізделген еңбектердің бірі – Қ.Қ. Өмірбекова, Ж.Т.Тасжурекова, А.Д.Сейсенова, Г.Б. Ибатова, А.С. Оспанова авторлығымен жарық көрген «Онтогенез барысында ана тілі жүйесінің қалыптасуы» атты ұжымдық монография [5]. Арнайы және инклюзивті білім беру саласындағы зертхананың ғылыми қызметкерлері зерттеген бұл еңбекте қазақ тілін ана тілі ретінде меңгеретін балалардың сөйлеу тілінің қалыптасу заңдылықтары онтогенез тұрғысынан кешенді түрде қарастырылып, сөздік қордың дамуы, грамматикалық құрылымдардың игерілуі және сөйлеудің психолингвистикалық негіздері ғылыми деректермен сипатталады.

Сондай-ақ, қазақ тіл білімінде қалыпты дамыған балалардың сөйлеу тілі мен әртүрлі бұзылыстары бар мектеп жасына дейінгі балалардың сөйлеу тілін салыстырмалы түрде зерттеген ғалымдар қатарына Қ.Қ. Өмірбекова, Г.Н. Тулебиева, Г.Б. Ибатова, Т.Ж.Байдильдинов, Г.А. Рахимбаева және А.Г. Жұбатхан жатады. Бұл зерттеулерде бала тілінің қалыпты және қалыптан тыс дамуын салыстыра зерттеу арқылы сөйлеу әрекетінің құрылымы, дыбыстық құрам, грамматикалық және лексикалық бірліктердің қалыптасу ерекшеліктері талданған.

Зерттеуші А.К.Шаяхмет көрсеткендей, «қазақстандық лингвистикада балалар тілі қазақ тілі материалдары негізінде бірнеше бағытта зерттелген. Атап айтқанда, балалар сөйлеуінің синтаксистік ерекшеліктері қарастырылып, билингвизмнің бала тұлғасының дамуына ықпалы зерттелген, мектеп жасына дейінгі балалардың етістік формаларын меңгеруіне психолингвистикалық талдау жасалған, сондай-ақ білім беру мен педагогикалық бағыттағы мониторингтік зерттеулер жүргізілген. Алайда аталған еңбектер өзара сабақтастықта қарастырылмаған және зерттеу мақсаттары мен міндеттері, сондай-ақ пайдаланылған тілдік материалдары жағынан әртүрлі бағыттағы дербес ғылыми ізденістер болып табылады» [6]. Яғни, қазақ тіл білімінде балалар тілінің ерекшеліктері арнайы онтолингвистикалық пән ретінде емес, жалпы тіл дамуы, сөйлеу мәдениеті, психолингвистика аясында ғана қарастырылып келеді. Бұл зерттеулер көбіне сипаттамалық деңгеймен шектеліп, табиғи сөйлеу деректеріне негізделген эмпирикалық базаға сүйенбейді.

Онтолингвистика ғылымының негізін қалаған ғалым С.Н.Цейтлин атап көрсеткендей: «Балалар тілі, бір жағынан, нормативті тілдің айнасы ретінде қарастырылуы тиіс. Ересектердің сөйлеуінен балалар тілінде ненің қалай енетінін және ол қалай өзгеретінін, соны анықтау нормативті тілдің құрылымы туралы жаңа мәлімет алуға мүмкіндік береді немесе, ең болмағанда, басқа зерттеу әдістерімен алынған ақпаратты тексеріп, нақтылауға мүмкіндік береді. Табиғат өзі жүзеге асыратын осы керемет лингвистикалық эксперименттің деректерін елемеу ақылға қонбайды» [7].

Осы тұрғыдан алғанда, С.Н. Цейтлин ұсынған тұжырым балалардың ерте сөйлеу деректерін ғылыми корпус ретінде жүйелі түрде жинау мен сипаттаудың теориялық маңызын айқындайды. Балалар тілі – ересек тілінің жай ғана қысқартылған көшірмесі емес, ол тілдік жүйенің қалыптасу механизмдерін, грамматикалық категориялардың ассимиляциялану реттілігін және семантикалық құрылымдардың табиғи эволюциясын көрсететін динамикалық

ISSN 2788-7979 (online)

модель. Сондықтан, ерте жастағы балалардың сөйлеу тілін корпустық әдістерге негізделе зерттеу тілдің нормативтік құрылымының когнитивтік және функционалдық негіздерін жаңа тұрғыдан түсіндіруге мүмкіндік береді.

КІРІСПЕ

Қазақтілді балалардың сөйлеу тілі – түркі тілдеріне тән агглютинативті морфологияны, бай синтаксистік құрылымдарды және фонетикалық ерекшеліктерді меңгеру үдерісін көрсету арқылы тіл онтогенезін тереңірек түсінуге мүмкіндік беретін құнды лингвистикалық дерек. Қазақ тілі – жалғамалы жүйедегі сөзжасамдық ерекшелігі айқын байқалатын тілдердің бірі, сондықтан қазақтілді балалардың тіл дамыту жолы халықаралық зерттеулерде кең тараған үндіеуропалық тілдермен салыстырғанда өзгеше сипатқа ие екені белгілі. Осы аталған ерекшеліктерді жүйелі түрде сипаттау мен тіл онтогенезінің заңдылықтарын дәл әрі объективті анықтау үшін ерте жастағы балалар сөйлеуін құрылымдалған, аннотацияланған деректер базасы негізінде зерттеу қажеттілігі туындайды.

Итальян тіліндегі мектепке дейінгі балалар корпусын қарастырған зерттеушілер Киара Болоньези және әріптестері көрсеткендей: «...тілдік корпустар үлкен көлемдік материалды қамтитындықтан, тілдің қолданыстық ерекшеліктерін зерттеудің тиімді құралы болып табылады. Әсіресе, тіл меңгеру үдерісін зерттеуде корпусқа негізделген әдістер кеңінен қолданылады. Сонымен қатар, тілдің табиғатын негіздейтін әмбебап грамматика теориясымен қатар, тілдік тәжірибе мен қолданысқа сүйенетін көзқарастар да корпус деректері арқылы дәлелденіп келеді» [8] тілдік корпустың практикалық маңыздылығы жоғары, себебі ол тіл туралы деректерді жүйелі түрде жинақтап, ғылыми және қолданбалы мақсатта пайдалануға мүмкіндік береді.

Балалар сөйлеуінің мультимедиялық корпусы балалар мен ата-аналар арасындағы өзара әрекетескен аудио немесе бейнежазбаларынан тұрады. Мұндай деректер балалардың сөйлеуін ең ерте кезеңдерінен бастап зерттеуге, талдау барысында бастапқы тілдік деректерді, яғни, балаға бағытталған сөйлеуді ескеруге, сондай-ақ тілдік емес факторларды қарастыруға мүмкіндік береді. Балалар сөйлеу корпустарының ең танымал әрі ең толық дерекқоры – CHILDES халықаралық базасы. 1983 жылы құрылған бұл база қазіргі таңда 40-тан астам тілді қамтитын 230-дан аса корпусты (әрбір бала үшін бір немесе бірнеше ішкі корпустарды қоса алғанда) біріктіреді және біртілді де, көптілді де қауымдастықтардағы ана тілін меңгеру үдерісі туралы жан-жақты ақпарат ұсынады. Питтсбург университетінде құрылған бұл жоба ауызша сөйлеуді, балалар мен олардың ата-аналарының спонтанды (табиғи) сөйлеуін талдау, сондай-ақ екінші тілді игеру үдерісін зерттеу үшін құрылған. Мәтіндік жазбалардың көлемі мол болумен қатар, онда психолингвистика, лингвистика, бірінші және екінші тілдің игерілу теориясы бойынша кең ауқымды библиография, материалды енгізу ережелері және оны талдауға арналған бағдарламалар да қамтылған [9].

Корпустық лингвистика әдістері балалардың сөйлеу деректерін сақтау мен басқаруда жоғары тиімділікке ие. Бұл әдістер бір ғана деректер жиынтығы негізінде көптеген зерттеулер жүргізуге және жаңа гипотезаларды тексеруге мүмкіндік береді. Ғалымдар балалар сөйлеу корпустарын фонология, морфология және синтаксис сияқты әртүрлі тілдік деңгейлердегі ерекшеліктердің меңгерілуін зерттеу үшін кіріс деректерінің балалар сөйлеуінің әртүрлі аспектілеріне ықпалын талдау үшін, сондай-ақ тілдің дискурстық қырларын зерделеу үшін кеңінен пайдаланып келеді [10].

Қазіргі уақытта әлемнің көптеген тілдері бойынша балалар сөйлеу тілін зерттеуге арналған арнайы тілдік корпустар қалыптастырылған. Мысалы, жоғарыда атап өткен CHILDES деректер қорында ағылшын тілі бойынша 95 корпус (шамамен 3000 файл), француз тілі бойынша 16 корпус (549 файл), қытай тілі бойынша 23 корпус (1453 файл) тіркелген [10].

Алайда мұндай жүйелі тілдік ресурстар барлық тілдер үшін бірдей деңгейде қалыптасқан деп айтуға болмайды. Әсіресе түркі тілдері, оның ішінде қазақ тіліндегі балалар сөйлеу тілін кешенді түрде қамтитын корпустары әлі де жеткіліксіз. Мәселен, CHILDES

ISSN 2788-7979 (online)

базасында түрік тілі бойынша Аксу (54 файл) және Altinkamis (15 файл) сияқты шектеулі ғана корпустар ұсынылған [https://talkbank.org/].

CHILDES деректер базасында орналасқан Аксу Коч (Aksu Koç) [11] – түрік тілді балалар корпусы. Бұл деректер базасы 1972-1973 жылдары Стамбул қаласында Д.Слобиннің жетекшілігімен, The Grant Foundation қорының қолдауымен жиналған. Зерттеуге кемінде бір ата-анасының жоғары білімі бар, қалалық отбасылардан шыққан балалар қатысты. Қатысушылар 4 айлық жас интервалдары бойынша іріктеліп, 2;0-4;4 жас аралығындағы 33 баланы қамтыған. Кейбір балалар 4 айдан кейін қайта бақыланып, жалпы жас ауқымы 2;0-4;8 жасқа дейін кеңейтілген. Балалар өз үйлерінде немесе мектепке дейінгі мекемелерде бір апта бойы бақыланып, Слобин (1982) сипаттаған когнитивтік және тілдік тапсырмалар кешенін орындаған. Зерттеу түрік, серб-хорват, итальян және ағылшын тілдерін қамтыды, ал түрік тілі бойынша кезең ғалым Айхан Ақсу-Кочпен бірлесе жүзеге асырылды. Транскрипттер бақылау барысында тіркелген ересек пен бала арасындағы спонтанды және бағытталған диалогтарды қамтиды. Балалар айтылымдары морфемалық деңгейде таңбаланған, транскрипцияда стандартты лингвистикалық белгілер қолданылған.

Түрік тіліндегі балалар сөйлеуіне қатысты тағы бір маңызды тілдік ресурс – Алтынкамис (Altinkamis) корпусы[12]. Зерттеу материалы бір қыз баланың 1 жас 4 айдан бастап 2 жас 4 айға дейінгі аралықтағы тілдік дамуын қамтитын лонгитюдтік бақылаулар негізінде жинақталған. Деректер баланың табиғи тілдік ортада ересектермен өзара әрекеттесу барысында пайда болған сөйлеу үлгілерін тіркеу арқылы алынған.

Корпустың авторы әрі құрастырушысы – Н.Фейза Алтынкамыш (Гент университеті), балалардың түрік тілін бірінші тіл (L1) ретінде меңгеруі, ерте сөздік қоры, грамматикалық даму үдерістері, сондай-ақ балалар тіліндегі «лексика–грамматика» континуумы мәселелерін зерттеумен айналысатын маман. Деректер құрамына ересек пен бала арасындағы табиғи диалогтардың транскрипттері енген. Зерттеудің негізгі бағыты – түрік тілінің ана тілі ретінде игерілуі, ерте лексикалық қордың қалыптасуы, грамматикалық құрылымдардың біртіндеп дамуы және лексика мен грамматиканың өзара байланысын анықтау. Материал CHILDES жобасы аясында қабылданған CHAT форматында транскрипцияланған.

Соңғы жылдары балалардың табиғи сөйлеуін ұзақ мерзімді бақылау негізінде жинақталған корпустық жобалар белсенді дамуда. Әсіресе, славян тілдеріндегі онтолингвистикалық зерттеулерде балалар тілінің деректерін жүйелеуге бағытталған ірі корпустар құрылды. Соның бірі – Мәскеу қаласындағы Жоғары экономика мектебінің Нейролингвистика зертханасының (қазіргі «Тіл және ми» орталығы) бір топ ғалымдары әзірлеген RusLan-M (doi:10.21415/1XMG-D508) корпусы. 2016-2019 жылдары екі отбасындағы бала мен ата-ананың өзара әрекеттесуі жүйелі түрде бейнежазбаға түсіріліп CHILDES/TalkBank платформасында қолданылатын CHAT форматы бойынша, Брайан Макуинни ұсынған стандарттарға сүйене отырып транскрипцияланған. Монолингвалды орыс балаларының табиғи сөйлеуі жүйелі түрде жинақталған бұл корпус ерте тіл дамуының динамикасын ұзақ мерзімді бақылау негізінде сипаттауға, тілдік бірліктердің қалыптасу ретін анықтауға және балалар сөйлеуін көпдеңгейлі аннотациялау тәжірибесін жүйелеуге мүмкіндік береді.

Эксперименталдық әдістеме негізінде қазақтілді балалардың сөйлеу тілін алғаш жүйелі түрде қарастырған зерттеулердің бірі – Ж.Смағұлованың [13] еңбегі. Автор қалалық қазақ-орыс қостілді отбасыларында балаларға бағытталған сөйлеу (baby talk) тілдік әлеуметтік тәжірибелердің маңызды компоненті екенін көрсетеді. Зерттеу ата-аналардың ерте тілдік өзара әрекеттесуін талдай отырып, қазақ тілінің балаға бағытталған сөйлеуде кеңінен қолданылатынын, алайда әлеуметтік бедел мен кең таралған тіл мәртебесіне байланысты балалардың негізгі өмірлік контексте орыс тілін меңгеруге бейім екенін көрсетті. Мақалада қазақ тілінің қаладағы бала тілі ретінде тіркелуі мен тіл ауысуы (language shift) арасындағы өзара байланыс қарастырылып, балалардың тілдік таңдауы мен тілдік даму траекториясына ықпал ететін әлеуметтік және идеологиялық факторлар талданады. Бұл зерттеу қазақ тілінің

ISSN 2788-7979 (online)

қалалық орталарда сақталу және дамуы мәселелерін түсінуге, сонымен қатар ерте тілдік интервенциялар мен тілдік саясатты ғылыми негізде жобалауға мүмкіндік береді.

МАТЕРИАЛДАР МЕН ӘДІСТЕР

Зерттеу барысында 1 жастан 4 жасқа дейінгі балалардың табиғи коммуникативтік жағдайдағы сөйлеу әрекеттерін қамтитын аудиожазбалары пайдаланылды. Балалардың ерте сөйлеу тілінің дамуын айқындау мақсатында жүйелі бақылау жүргізіліп, эмпирикалық деректер жинау әдістері қолданылды. Тілдік деректерді жүйелеу мен талдау үдерісіндегі корпустық әдістер қолданылып, жиналған материалдар арнайы корпус форматына келтіріліп ұйымдастырылды.

Сонымен қатар, зерттеу барысында тілдік материалдарды өңдеу мен интерпретациялауда лингвистикалық сипаттау әдісі, балалар тілінің даму динамикасын айқындауға бағытталған лонгитюдтік бақылау әдісі және салыстырмалы талдау әдісі қолданылды.

НӘТИЖЕЛЕР

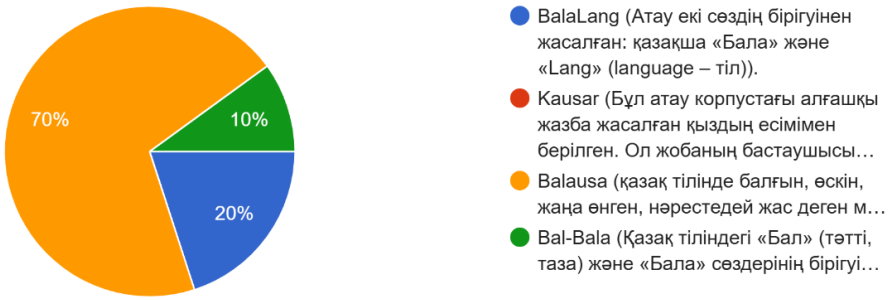
Аудиокорпус – сөйлеудің табиғи жағдайында жазылып алынған және арнайы белгіленген (аннотацияланған) дыбыстық деректер жиынтығы. Балалар тілін зерттеуде оның маңызы ерекше, себебі ол баланың сөйлеуіндегі фонетикалық, лексикалық, просодикалық ерекшеліктерді толық сақтап, толық емес сөздер, дыбыстық алмастырулар мен контекстке тәуелді мағыналарды дәл тіркеуге мүмкіндік береді. Аудиокорпус балалар сөйлеуінің фонологиялық, морфологиялық және синтаксистік деңгейлердегі бірліктерін зерттеуде, сондай-ақ тілдік кіріс деректерінің балалардың сөйлеуіне ықпалын анықтауда кеңінен қолданылады.

Корпустық тәсілдердің басты артықшылығы – балалардың сөйлеуін табиғи ортада тіркеуі. Бұл олардың экологиялық валидтілігін арттырып, зерттеу нәтижелерін нақты өмірлік жағдайларға жақындатуға мүмкіндік береді [10].

Осы бағыттағы алғашқы бастама ретінде біз «Balausa» атты қазақтілді балалардың сөйлеу тілін зерттеуге арналған аудиокорпусын құру жұмыстарын жүзеге асырудамыз. Корпус атауы да кездейсоқ алынған жоқ. Алғашында ерте жастағы қазақтілді балалардың сөйлеу тілін зерттеу үшін төрт түрлі атау ұсынылды: Balalang, Kausar, Balausa, Bal-bala (Диаграмма-1). Қазақ тілі мен әдебиеті кафедрасының оқытушыларымен кеңесу және дауыс беру нәтижесінде жобаның ғылыми мазмұны мен концептуалдық мақсаттарына ең сай келетін атау ретінде «Balausa» атауы таңдалды. Атау қазақ тіліндегі bala (бала) түбірімен байланысты болғандықтан, корпус нысаны – 1–4 жас аралығындағы балалардың табиғи (спонтанды) сөйлеуі екенін бірден білдіреді. «Balausa» қазақ тілінде *жас өркен, бүршік, жаңа өсінді* мағынасында қолданылады. Бұл метафора тілдің ерте кезеңдегі қалыптасуын, алғашқы сөздердің пайда болуын және сөйлеудің табиғи түрде өсуін бейнелейді.

Сонымен қатар «Balausa» атауы ұлттық мәдени контекске тән *өсіп-өну, қалыптасу, бастау* сияқты концептілермен үйлеседі және халықаралық ғылыми ортада оңай қабылданады деген ойдамыз.

1-диаграмма. Корпус атауын таңдау



Зерттеу аясында 1–4 жас аралығындағы 12 қазақтілді бала іріктеліп алынды. Балалардың аудиожазбаларын заңды түрде пайдаланып зерттеу үшін және оларды CHILDES халықаралық базасында ашық жариялау үшін Ш.Уәлиханов атындағы Көкшетау университетінің этикалық комитетінің талабына сай ата-аналардан ақпараттық келісім мен сауалнама алынды. Деректерді жинауды бастамас бұрын, ата-аналарға егжей-тегжейлі нұсқаулық берілді. Нұсқаулықта жобаның қысқаша сипаттамасы, қаншалықты жиі жазу керек, не жазу керек (баламен әдеттегі өзара әрекеттесу), қалай жазу керек екені жайлы толыққанды ақпарат берілді.

1-кесте. «Balausa» корпусының бастапқы сипаттамасы.

Бала	Арнайы код	Жынысы	Бастапқы жасы[a]	Аяқталу жасы	Жазбалар саны	Жазбалар ұзақтығы (минут)
Airis	AA-01	Қыз	1;10.24	Аяқталмаған	2	33
Marlen	ME-01	Ұл	1;9.2	Аяқталмаған	1	25
Alfiya	AE-01	Қыз	1;8.16	Аяқталмаған	6	66
Ayat	AN-01	Ұл	2;0.19	Аяқталмаған	1	15
Kamshat	KH-02	Қыз	1;11.28	Аяқталмаған	4	62
Tauekel	TZ-01	Ұл	2;5.29	Аяқталмаған	2	30
Aisafi	AM-01	Қыз	2;7.27	Аяқталмаған	2	15
Aisana	AS-01	Қыз	2;6.1	Аяқталмаған	2	56
Alinur	AB-01	Ұл	2;6.7	Аяқталмаған	1	60
Zhazira	ZQ-01	Қыз	3;0.19	Аяқталмаған	7	118
Sardar	SQ-01	Ұл	3;6.1	Аяқталмаған	2	30
Kausar	KH-01	Қыз	3;10.26	Аяқталмаған	3	104

[a] *Жасы стандартты жыл; ай; күн форматында көрсетілген.*

Balausa аудиокорпусы – лонгитюдті зерттеу ретінде жоспарланған балалардың сөйлеу дамуының динамикасын бақылауға, лексикалық құрамдағы өзгерістерді, жаңа грамматикалық формалардың пайда болуын және коммуникативтік стратегиялардың трансформациясын тіркеуге мүмкіндік беретін, ұзақ мерзімді қамтитын жоба.

Концептуалдық деңгейде Balausa тек эмпирикалық ресурс қана емес, сонымен қатар цифрлық қазақ онтолингвистикасының базасын қалыптастыруға бағытталған платформа болып табылады.

Жобаның ғылыми жаңалығы – ерте жастағы қазақ тіліндегі балалардың сөйлеу тілінің аудиокорпусын құру. Ол стандартизацияланған транскрипцияны, аннотацияны және кейінгі халықаралық корпус ресурстарына интеграциялауды көздейді. Бұл тілдер арасындағы салыстырмалы зерттеулер жүргізуге мүмкіндік беріп, сөйлеу онтогенезді зерттеулерінің типологиялық базасын кеңейтеді.

Аудиокорпус жасау бірнеше кезеңнен тұрады, олардың әрқайсысы нақты мәселелерді шешуге бағытталған. Негізгі кезеңдер келесідей сипатталады:

ISSN 2788-7979 (online)

1. Аудиожазбаларды жинау – зерттеу материалдарын жинау және балалардың сөйлеу әрекетін аудио форматта жазу. Алдын ала берілген нұсқаулыққа сәйкес ата-аналар балаларымен әңгімелесу барысын жазып отырады. Деректерді сақтаудың біркелкілігін қамтамасыз ету үшін барлық аудио файлдар MP4 немесе WAV форматына түрлендіріледі. Әрі қарай аудиожазбалар үшін стандартталған атау форматын жасадық. Мысалы, AA-01_001.wav файлында А әрпі баланың атының бірінші әрпін (Айрис), ал сандар ішкі корпустағы реттік нөмірін білдіреді. Аудиожазбалар әзірге бір мәрте ғана жиналды, алайда зерттеу жоспарына сәйкес деректерді жинау үдерісін үш ай сайын жүйелі түрде қайталап отырамыз.

2. Автоматты транскрипция – жазбаларды мәтіндік форматқа бастапқы автоматты транскрипциялау. Аудиожазбаларды транскрипциялау үшін одан әрі өңдеу үшін біз тек форматты өзгертіп, сол атауларды қолдандық. Мысалы, бұл аудиожазбаның транскрипциясы AA-01_001.cha деп аталады. Аудиожазбалар Whisper автоматты сөйлеуді тану моделі арқылы өңделіп, BatchAlign2 бағдарламасының көмегімен транскрипцияланды.

3. Тексерілген транскрипция – автоматты транскрипция нәтижелерін қолмен тексеру және қателерді түзету. «Қазіргі бағдарламалық жасақтама корпустармен жұмыс істеу кезінде операциялардың бір бөлігін, соның ішінде транскрипциялап белгілеу процедурасын автоматтандыруға мүмкіндік береді. Дегенмен, автоматты белгілеу бағдарламасының нәтижелері зерттеуші тарапынан тексеруді және түзетуді қажет етеді. Корпустың көлемі неғұрлым үлкен болса, автоматты өңдеуден кейінгі кезең соғұрлым көп еңбекті қажет ететін болады» [11].

Бұл процесс әсіресе қазақтілді аудиоматериалдарға өте маңызды, себебі автоматты жүйелер көбіне фонетикалық ерекшеліктерді, сөздердің морфологиялық формаларын немесе дыбыстық вариацияларды дұрыс танымайды. Қолмен тексеру транскрипциядағы қателерді азайтып, корпус деректерінің сапасын арттыруға мүмкіндік береді.

4. Псевдонимизация – жеке деректерді қорғау үшін қатысушылардың есімдері мен идентификаторларын ауыстыру. Балалар мен олардың ата-аналары арасындағы күнделікті қарым-қатынас жазбаларында көбінесе жеке немесе құпия ақпарат болғандықтан, анонимдеу маңызды қадам болып табылады. Біз жазбаларда пайда болған барлық жеке ақпарат (мысалы, аты-жөні, мекенжайы) міндетті түрде анонимделеді.

5. Морфосинтаксистік талдау – транскрипцияланған мәтіндерді морфологиялық және синтаксистік деңгейде автоматты түрде өңдеу. CHAT форматындағы транскриптерді CSV файлға түрлендіріп, морфологиялық және синтаксистік талдау жүргізіледі. Әр сөз үшін лемма жасалып, сөз табы анықталып, толыққанды талдау жасалады.

6. Тексерілген морфосинтаксистік талдау – автоматты талдау деректерін қолмен тексеру және түзету арқылы корпус материалдарының сапасын қамтамасыз ету. Mystem көпмәнді аннотациялар жасайтындықтан, нәтижелерді қолмен тексере отырып, морфологиялық және синтаксистік таңбалану (разметка) жасалады.

Грамматикалық белгіленімдері (разметки) бар балалар сөйлеу тілінің корпусы өзекті грамматиконды модельдеудің тиімді құралы болып табылады, өйткені ол келесі мүмкіндіктерге жол ашады:

- ерте онтогенез кезеңіндегі (екі жастан бес жасқа дейін) балалардың монологтық және диалогтық сөйлеуіндегі айтылымдар жиынтығын (массивін) ұсыну;
- нақты бір жастағы баланың типтік коммуникативтік жағдайлардағы айтылымдарында грамматикалық формалардың жиілік үлесін анықтау;
- белгілі бір жастағы тілдік тұлғаның грамматикалық профилін (сөз түрлендіру формаларының бөлінісін) қарастыру [11].

Барлық кезеңдерден өту нәтижесінде балалар сөйлеу тілінің деректерін сандық көрсеткіштер негізінде кешенді түрде талдауға мүмкіндік аламыз. Мәселен, MLU (Mean Length of Utterance (айтылымның орташа ұзақтығы) – көрсеткіші арқылы қазақтілді баланың сөйлеуіндегі орташа құрылымдық бірліктерінің көлемі айқындалса, IpSyn (Index of Productive

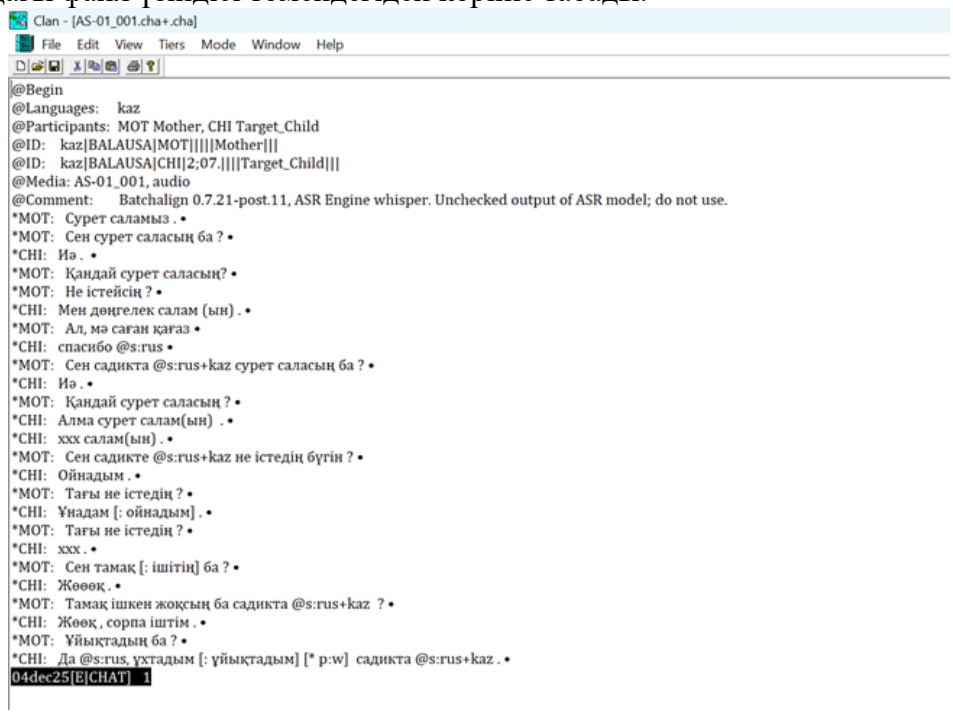
ISSN 2788-7979 (online)

Syntax (синтаксистік өнімдік индексі) – көрсеткіші сөйлеу тілінде синтаксистік құрылымдардың күрделілік деңгейі мен өнімділігін бағалауға мүмкіндік береді.

ТАЛҚЫЛАУ

Аудиожазбаларды транскрипциялау *Тілдік компьютерлік талдау* (CLAN) бағдарламасында, оның ішінде жоғарыда атап өткеніміздей CHAT форматында орындалады. Әр файл метадеректерден, сөйлеу транскрипциясынан және түсіндірмелерден тұрады. Баланың сөйлеуі мен ата-ананың балаға бағытталған сөйлеуі кирилл қарпімен транскриптелінеді.

«CHAT жүйесі кемінде екі деңгейде қызмет етеді. Оның ең қарапайым нұсқасы – minCHAT. Бұл нұсқаны пайдалану күрделі кодтауды қажет етпейді. Мұндай транскрипциялау тәсілі балалар тілін зерттеу және дискурсты талдау салаларында қолданылатын дәстүрлі интуитивті транскрипцияларға жақын болып келеді» [15]. CHILDES базасының негізін қалаушы психолингвист ғалым Брайан Макуинни атап өткендей, CHAT жүйесі кемінде екі деңгейде жұмыс істеуге арналған. Оның ең қарапайым нұсқасы minCHAT деп аталады. CHAT форматындағы файл үзіндісі төмендегідей көрініс табады.



1-сурет. CLAN бағдарламасы арқылы транскрипцияланған .CHA файлы

Аудиоматериалдарды транскрипциялау жүзеге асырылғанымен, транскрипцияда көптеген қателер бар екені байқалды. Бұл қателердің негізгі себептері ретінде бағдарламаның қазақ тіліндегі фонетикалық ерекшеліктерді толық меңгермеуі, балалардың сөйлеу тіліндегі артикуляциялық вариациялар, дауыссыз және дауысты дыбыстардың алмастырылуы, сондай-ақ буындық құрылымның ауытқуы мен былдырлау кезеңіндегі дыбыстық созылымдардың әртүрлілігін атауға болады. Сондай-ақ, фонетикалық ұқсастықтары жоғары дыбыстардың қатар келуі де қателіктердің жиілеуіне әсер етеді. Сонымен қатар, қазақ тілінің ғылыми тіл ретінде табиғи тіл өңдеу (NLP) саласында толықтай қалыптаспағандығы да транскрипцияның дәлдігіне әсер етеді. Осы жағдайларды ескере отырып, транскрипция нәтижелерін қолмен тексеру және түзету қажеттілігі туындайды. Бұл әдіс балалардың нақты дыбыстық, буындық жүйесін дұрыс бағалауға мүмкіндік береді. Брайан Макуинни авторлығымен құрастырылған CHAT транскрипция стандартының нұсқаулығында (Сөйлеу тілін талдау құралдары. 1-бөлім: CHAT транскрипциялау форматы) [15] берілген қағидалар мен ескертулерге сүйене отырып, транскрипциялау барысында жиі кездесетін қателіктерді төмендегідей белгіледік.

ISSN 2788-7979 (online)

«Сөз деңгейіндегі қателер сөзден кейін [*] символы арқылы беріледі. Егер алмастыру элемент, мысалы [: because] берілсе, ол қателік кодынан бұрын жазылады. Қайта түзетілген жазбаның бастапқы бөлігінде қате орын алған жағдайда [*] символы, қате элементтен кейін, ал [/] белгісінен бұрын қойылады» [15].

2-кесте. Аудиожазбаларды транскрипциялау барысында жиі кездесетін қателіктерді кодтау

[*p]	Фонологиялық қателіктер. *[* p:w] – қолданыста бар сөзге ұқсата айтылатын фонологиялық қателік.* *[* p:n] – қолданыста жоқ сөзге ұқсата айтылатын фонологиялық қателік.* *[* p:m] – метатеза: дыбыс немесе буындардың орындарын алмастырып айту.* Мысалы: *Міне, нандай , нандай [:мынадай] [* p:w]* *омарал [: орамал] [* p:m]*
[* s]	семантикалық қателіктер.
[* m]	морфологиялық қателіктер.
[* n]	жаңа сөздер, яғни неологизмдер.
@o	Ономатопея (дыбыстық еліктеу, сөз жасау) Мысалы: *Әтеш кукарекууу @o дейді ! (Айрис, 1 жас 10 ай)*
&=	Сөйлеуден тыс жасалатын іс-әрекеттер (күлу, жылау, айқайлау, әндету, түшкіру, жөтелу және т.б.) *Мысалы: &=laughs күлу*
xxx	Түсініксіз, ойдан шығарылған сөздерді айтқан кезде
(.)	Сөйлеу барысында ұзақ кідірісті көрсету үшін қолданылады.
[/]	Бір сөзді немесе сөз тіркесін бірнеше мәрте қайталаған кезде. Мысалы: Сосын [/] сосын [/] сосын (бала бір сөзді контексте қатарынан үш рет қолданды)
0:	Сөздің немесе сөз тіркесінің түсіп қалуы (омиссия)
%com	Комментарий. Жазбаға түсініктеме беру үшін қолданылады (бір нәрсені нақтылау, түсініктеме беру және т.б.).
&=imit:	Шуды, қозғалысты және т.б. бейнелейтін дауысты дыбыстар үшін қолданылады.

Жинақталған материалдар қазақтілді балалардың ерте лексикалық дамуындағы бірнеше ерекшеліктерді айқындауға мүмкіндік берді.

Біріншіден, баланың қоршаған ортаға қатысты номинативтік лексиканың басымдығы байқалды. Баланың тілдік қолданысында жиі кездесетін сөздерге отбасы мүшелерінің (*ата, апа, мама, аға, бөпе*), тұрмыстық заттардың (*қасық, үй, сурет, қарындаш*), ойыншықтар (*машина, доп*) мен тағамдардың (*су, нан, май*) атаулары жатады. Зерттеушілер А.Н.Гвоздев, В.Штерннің жұмыстарында 1 жас 6 айдан 2 жасқа дейінгі уақытта балалар заттың атына байланысты «Бұл не?» сұрақ қою кезеңі деп белгіленді. Бұл жаста бала қоршаған ортаға бейімделе бастайды, соған орай біртіндеп сөздік қоры да молаяды [5].

Екіншіден, фонетикалық дыбыстардың қысқаруы, буындардың түсіп қалуы және артикуляциялық өзгерістерге ұшырауы арқылы көрініс табады. Мысалы, *«қарындаш»* сөзіне қатысты *«қандараш»* түріндегі формалар немесе етістік формаларындағы метатездер байқалады, бұл фонологиялық даму кезеңдерінің өтпелі сатыларын көрсетеді. Мұндай формалар әртүрлі балаларда жүйелі түрде қайталанатыны олардың фонологиялық дамудың қалыпты кезеңдері ретінде қарастырылуына мүмкіндік береді.

Үшінші маңызды ерекшелік – ономатопеялық лексиканың белсенді қолданылуы болып табылады. Балалар сөзінде дыбысқа еліктеу арқылы жасалған атаулар кеңінен таралған: ав-ав (ит), мяу (мысық), бух (құлау, соғылу) пи-пи (көлік дабылы). Аталған бірліктер көптеген тілдердегі балалар сөйлеуіне ортақ универсал/әмбебап сипатқа ие және көбіне заттың немесе әрекеттің атауын алмастырады.

Ал қазақтілді балалардың сөйлеуінде ономатопеялық лексиканың қолданылуы қазақ тілінің төл дыбыстардың фонетикалық жүйесіне негізделуімен де ерекшеленеді. Атап айтқанда, жануарлар дыбыстарын беру барысында қазақ тіліндегі созылыңқы дауыстылар

ISSN 2788-7979 (online)

белсенді қолданылады: сиыр – *мө-ө-ө*, қой – *мә-ә-ә*, қарға – *қарқ-қарқ*, тауық – *құт-құт* және т.б. Мұндай бірліктер баланың тек дыбысқа еліктеуін ғана емес, сонымен қатар ана тілінің фонетикалық үлгілерін ерте меңгере бастайтынын көрсетеді.

Зерттеу барысында балалардың тілдік қарым-қатынасында орыс және қазақ тілдеріне тән сөздердің аралас қолданысы байқалды. Балалардың тілдік қарым-қатынас барысында орысша лексикалық бірліктерді тұрақты қолдануы ерте жастан билингвалды құзыреттіліктің қалыптаса бастағанын айғақтайды. Бұл үдерісті шартты түрде біз үш бағытқа бөліп қарастырдық:

1. **Лексикалық интерференция (тікелей ену).** Орыс тілді лексикалық бірліктердің ешқандай өзгеріссіз, сол қалпында қолданылуы. Яғни, мұндағы сөздер семантикалық тұтастығын сақтай отырып, прагматикалық қызмет атқарады. Әсіресе, модаль сөздер мен шылаулардың өзгеріссіз тікелей енуін көрсетеді. Мысалы: *да, давай, потом, быстро, и еще, не хочу.*

*MOT: Бұл күнің көзі ма ?
*CHI: Күннің көзі .
*CHI: Давай @s:rus мен істейм (ін) .
*MOT: Кәне, сен сал енді .
*MOT: Ол не ?
*CHI: Күні көзі [:күннің көзі] .
*MOT: Күннің көзінің түсі қандай болады ?
CHI: Міне, нандай , нандай [:мынадай] [p:w] .
MOT: Міне нандай .[:мынадай] [p:w]
*MOT: Түсі қандай ?
*CHI: Күннің көзі .
*MOT: Қандай түсі ?
*MOT: Күннің көзінің түсі қандай ?
CHI: Сады [: сары] [p:w]. *(*AS-01*, 2 жас 6 ай)*

2. **Фоно-морфологиялық бейімделу.** Балалардың тілдік қолданысында кірме сөздердің қазақ тілінің фонетикалық және морфологиялық нормалары мен заңдылықтарына сәйкес бейімделу жағдайлары байқалды. Мысалы: *кататься → кәттену, качели → кәшеля, носки → нәски, тарелка → тәрелке*

*CHI: Мам ?
*CHI: Қазір далаға шығамыз ба ? .
*MOT: Далаға шыққың келіп тұр ма? .
*CHI: иә
*MOT: Қазір көреміз .
*CHI: иә, көреміз .
*MOT: Не істейсің далаға шығып сен ?
*CHI: Ақша табу керек , магазинге @rus+kaz барамыз сосын .
*CHI: Сосын кәшеля @rus [: качели] кәттиміз. *(*КН-01*, 3 жас 10 ай)*

3. **Морфологиялық интеграция (Гибридті құрылымдар).** Орыс тіліндегі сөздерге қазақ тілінде қосымшаларды жалғау немесе керісінше қазақ тіліндегі сөздерді орысша қосымшаларды қолдану да тіркелді. Бұндай мысалдары кодтау үшін төмендегілер белгілер қолданылды:

@s:rus+kaz – орыс тіліндегі сөзге қазақша септік жалғауының жалғануы
@s: kaz+rus – қазақ тілінде сөзге орысша қосымшаларды қосу

Бұл жүйе балалардың тілдік араласуын дәл тіркеп, код ауысудың морфологиялық ерекшеліктерін талдауға мүмкіндік береді.

3-кесте. Балалар сөйлеу тіліндегі аралас қолданыстар

Тілдік қарым-қатынас орыс +қазақ (@s:rus+kaz) Орыс тілді сөзге қазақша септік жалғауларын қоса қолдану	**Қазақ тілді сөзге орысша қосымша қосу (kaz+@s:rus)**
Садикқа, садикта	*Апашка, апашкаға*
Магазинге	*Аташка*
Кукланы	*Әжека, әжеканы, Әжеша*
Виноградты	*Мадиношка*
Пижаманы	*Аташканы*
Машинкамен	
Цветокты	

ҚОРЫТЫНДЫ

Қорыта айтқанда, қазақтілді балалардың ерте сөйлеу тілінің корпусын құру – жай ғана техникалық база жинақтау емес, бұл ұлттық онтолингвистиканың іргетасын қалау процесі, қазақ тіл білімінде аса өзекті әрі стратегиялық маңызды ғылыми жоба. Мұндай ауқымды аннотацияланған ресурс бізге бала тілінің даму динамикасын нақты цифрлармен және ғылыми фактілермен сөйлетуге мүмкіндік береді. Бұл корпус бүгінгі күнге дейін кешенді зерттелмеген қазақ балаларының когнитивтік және тілдік дамуының табиғи үлгілерін жүйелі түрде жинап, халықаралық стандартқа сай өңдеуге мүмкіндік беретіні сөзсіз.

Корпустың жасалуы қазақ тілінің онтолингвистикалық моделі, психолингвистикалық нормасы мен тілдік технологияларын дамытуға зор үлес қосады. Сонымен қатар логопедия, сөйлеу диагностикасы, нейротілтану және ерте жастағы тілдік білім беру салалары үшін құнды мәліметтер береді.

Balausa жобасы онтолингвистика, психолингвистика, социолингвистика, логопедия және тіл оқыту әдістемесін біріктіретін эмпирикалық платформа ретінде қарастырылуы мүмкін. Оның ерекшелігі – нормативтік педагогикалық модельдерден шығып, баланың тілдік дамуының нақты когнитивтік механизмдерін көрсететін модельдерге көшу мүмкіндігінде.

Зерттеудің ғылыми жаңалығы тек қазақ тіліндегі алғашқы балалар сөйлеуінің аудиокорпусын жасауда ғана емес, сонымен бірге қазақ тілін ерте тілдік онтогенезді жүйелі корпус зерттеу нысанына айналдыратын жаңа зерттеу парадигмасын қалыптастыруда.

Алайда, бұл жоба бірнеше күн немесе бірнеше айдың ішінде толыққанды аяқталуы мүмкін емес екенін ашық түсінуіміз қажет. Біз қазір жобаның ең жауапты – алғашқы қалыптасу кезеңіндеміз. Балалардың табиғи сөйлеуін жазып алып, оны транскрипциялау және халықаралық стандарттарға сай көпқабатты аннотациялау – ұзақ уақытты, төзімділікті және жоғары кәсіптілікті талап ететін жүйелі процесс.

Қазіргі жұмыс нәтижелері толық аяқталған корпус емес, оның ғылыми негізін қалыптастыруға бағытталған бастапқы қадамдар ретінде қарастырылады. Бұл бағыттағы жүйелі зерттеулердің жалғасуы қазақ тіл білімінің, оның ішінде бала тілін зерттеу парадигмасын жаңа деңгейге көтеріп, отандық онтолингвистиканың деректік базасын қалыптастыруға мүмкіндік береді.

Мақаланы жазу барысында мәтінді техникалық реттеу мақсатында және генерациялауда, әдебиеттерді реттеу кезінде ChatGPT жасанды интеллект құралы қолданылды.

ӘДЕБИЕТТЕР ТІЗІМІ

1 Фромкин В., Родман Р., Хайамс Н. Тіл біліміне кіріспе. 10-басылым, Халықаралық басылым. – Алматы: Ұлттық аударма бюросы, 2018 жыл – 608 бет.

2 Аяпова Т.Т. Развитие синтаксиса речи детей (на материале казахского языка): автореф.дис. ... канд. филол.наук. – Алма-Ата, 1988.
3 Баймұратова Б. Мектеп жасына дейінгі балаларды сөйлеуге үйрету. – Алматы.: Мектеп, 1981.
4 Бакраденова А.Б. Мектеп жасына дейінгі балалардың қазақша сөйлеу тілін дамыту. – Алматы, 1993.
5 Өмірбекова Қ.Қ., Тасжурекова Ж.Т., Сейсенова А.Д., Ибатова Г.Б., Оспанова А.С / Онтогенез барысында ана тілі жүйесінің қалыптасуы: монография. – Алматы: АИБД ҰҒПО, 2021. – 166 б.
6 Шаяхмет А.К. Формирование билингвальной личности в Казахстане // Многоязычие в образовательном пространстве. Серия «Языковое и межкультурное образование».– Ижевск, 2014. – Том. 6.
7 Цейтлин С.Н. Возвратные глаголы и детская речь. Опубликовано в сб.: Проблемы теории грамматического залога АН СССР., Ин-т языкознания – Л., 1978. С.128- 130.
8 Bolognesi, C., Cinini, A., Cutugno, P., Ferretti, M., & Chiarella, D. Towards a preschooler corpus of Italian: an experimental journey. Research Methods in Applied Linguistics, 2025. Article 100252. https://doi.org/10.1016/j.rmal.2025.100252
9 Зырянова Е. В. Система CHILDES как метод сбора материалов и изучения детской речи / Е. В. Зырянова. – Текст : электронный// Известия РГПУ им. А. И. Герцена. – 2008. – № 76-2. – URL: https://cyberleninka.ru/article/n/sistema-childes-kak-metod-sbora-materialov-i-izucheniya-detskoy-rechi
10 Lelik, V., Diachkova, M., Dorofeeva, S. V., Lopukhina, A., Dragoy, O., & Sekerina, I. A. (2025). 2025 CHILDES RusLan-M Corpus (TalkBank/CHILDES). TalkBank. https://doi.org/10.21415/1XMG-D508
11 Aksu-Koç, A., & Slobin, D. I. (1978). Turkish child language data. In CHILDES: The Child Language Data Exchange System. TalkBank. https://talkbank.org/childes/access/Other/Turkish/Aksu.html (қолданылған күні: 21.01.2026).
12 Altınkamış N. F. Turkish Altınkamış corpus [Электрондық ресурс]. CHILDES: Child Language Data Exchange System, TalkBank. URL: https://talkbank.org/childes/access/Other/Turkish/Altinkamis.html (қолданылған күні: 21.01.2026).
13 Smagulova, J. Early language socialization and language shift: Kazakh as baby talk. Journal of Sociolinguistics. – 2013. 17(3), 309–336. https://doi.org/10.1111/josl.12030
14 Краснопёрова, Е. С. Грамматический корпус детской речи как инструмент исследования актуального грамматикона ребенка / Е. С. Краснопёрова. – Текст : непосредственный // Филологический класс. – 2024. – Т. 29, № 2. – С. 108–118. – DOI: https://doi.org/10.26170/2071-2405-2024-29-2-108-118
15 MacWhinney, B. The CHILDES Project: Tools for Analyzing Talk. 3rd Edition. Mahwah, NJ: Lawrence Erlbaum Associates, 2000

Материал 28.03.2026 баспаға түсті

BALAUSA: теоретические, методологические и типологические основы корпуса детской речи казахскоязычных детей

Г.Ш. Шахманова[1], О.А. Жумагулова[1]
[1]Кокшетауский университет им. Ш. Уалиханова, Кокшетау, 020000, Казахстан

В статье определены научно-теоретические, методологические и практические основы создания корпуса разговорной речи казахскоязычных детей раннего возраста, одной из актуальных областей современной этнолингвистики. В рамках исследования представлен корпус аудиозаписей «Balausa» и обоснована его эмпирическая значимость в изучении закономерностей раннего языкового развития. Объектом исследования стали аудиозаписи разговорной речи 12 детей в возрасте 1-4 лет, записанные в естественных условиях общения.

ISSN 2788-7979 (online)

При обработке данных они были транскрибированы в соответствии со стандартами международной базы данных CHILDES, и использовались многослойные морфосинтаксические методы аннотации.

С учетом агглютинативной природы казахского языка и на основе опыта работы с корпусами тюркоязычных детей были рассмотрены особенности маркировки параметров, характерных для казахского языка. Кроме того, в статье описаны основные этапы создания корпуса разговорной речи детей и указано, что проект в настоящее время реализуется на начальной стадии. Исследование позволяет сформировать цифровую онтолингвистическую модель казахского языка и определить траекторию развития языковой компетенции детей.

Ключевые слова: онтолингвистика, онтогенез, корпусные исследования, детский язык, аудиокорпус, CHILDES, транскрипция

СПИСОК ЛИТЕРАТУРЫ

1 Фромкин В., Родман Р., Хайамс Н. Тіл біліміне кіріспе, 10-басылым, Халықаралық басылым. Алматы: Ұлттық аударма бюросы, 2018 жыл – 608 бет.

2 Аяпова Т. Т. Развитие синтаксиса речи детей (на материале казахского языка): автореф. дис. … канд. филол. наук. Алма-Ата, 1988.

3 Баймұратова Б. Мектеп жасына дейінгі балаларды сөйлеуге үйрету. – А.: Мектеп, 1981.

4 Бакраденова А.Б. Мектеп жасына дейінгі балалардың қазақша сөйлеу тілін дамыту. – Алматы, 1993.

5 Өмірбекова Қ.Қ., Тасжурекова Ж.Т., Сейсенова А.Д., Ибатова Г.Б., Оспанова А.С / Онтогенез барысында ана тілі жүйесінің қалыптасуы: монография. – Алматы: АИБД ҰҒПО, 2021. – 166 б.

6 Шаяхмет А.К. Формирование билингвальной личности в Казахстане // Многоязычие в образовательном пространстве. Серия «Языковое и межкультурное образование». – Том. 6. – Ижевск, 2014.

7 Цейтлин С.Н. ВОЗВРАТНЫЕ ГЛАГОЛЫ И ДЕТСКАЯ РЕЧЬ Опубликовано в сб.: Проблемы теории грамматического залога АН СССР., Ин-т языкознания- Л., 1978. С.128- 130.

8 Bolognesi, C., Cinini, A., Cutugno, P., Ferretti, M., & Chiarella, D. (2025). Towards a preschooler corpus of Italian: an experimental journey. Research Methods in Applied Linguistics, 2025 08 25, Article 100252. https://doi.org/10.1016/j.rmal.2025.100252

9 Зырянова Е. В. Система CHILDES как метод сбора материалов и изучения детской речи / Е. В. Зырянова. – Текст : электронный// Известия РГПУ им. А. И. Герцена. – 2008. – № 76-2. – URL: https://cyberleninka.ru/article/n/sistema-childes-kak-metod-sbora-materialov-i-izucheniya-detskoy-rechi

10 Lelik, V., Diachkova, M., Dorofeeva, S. V., Lopukhina, A., Dragoy, O., & Sekerina, I. A. (2025). 2025 CHILDES RusLan-M Corpus (TalkBank/CHILDES). TalkBank. https://doi.org/10.21415/1XMG-D508

11 Aksu-Koç, A., & Slobin, D. I. (1978). Turkish child language data. In CHILDES: The Child Language Data Exchange System. TalkBank. https://talkbank.org/childes/access/Other/Turkish/Aksu.html (дата рассмотрения: 21.01.2026).

12 Altınkamış N. F. Turkish Altınkamış corpus [Электрондық ресурс]. CHILDES: Child Language Data Exchange System, TalkBank. URL: https://talkbank.org/childes/access/Other/Turkish/Altinkamis.html (дата рассмотрения: 21.01.2026).

13 Smagulova, J. (2013). Early language socialization and language shift: Kazakh as baby talk. Journal of Sociolinguistics, 17(3), 309–336. https://doi.org/10.1111/josl.12030

14 Краснопёрова, Е. С. Грамматический корпус детской речи как инструмент исследования актуального грамматикона ребенка / Е. С. Краснопёрова. – Текст : непосредственный // Филологический класс. – 2024. – Т. 29, № 2. – С. 108–118. – DOI: 10.26170/2071-2405-2024-29-2-108-118.

ISSN 2788-7979 (online)

15 MacWhinney, B. The CHILDES Project: Tools for Analyzing Talk. 3rd Edition. Mahwah, NJ: Lawrence Erlbaum Associates, 2000.

Материал поступил в редакцию журнала 28.03.2026

BALAUSA: Theoretical, methodological and typological foundations for the creation of a corpus of Kazakh-speaking children's speech

G. Shakhmanova[1], O. Zhumagulova[1]
[1]Sh. Ualikhanov Kokshetau University, Kokshetau, 020000, Kazakhstan

This article defines the scientific, theoretical, methodological, and practical foundations for creating a corpus of spoken language from children of early Kazakh descent, a topical area of modern ethnolinguistics. The study presents the "Balausa" audio recordings corpus and substantiates its empirical significance for studying patterns of early language development. The study focused on audio recordings of spoken language from 12 children aged 1-4 years, recorded in natural settings. Data processing involved transcribing the audio in accordance with the standards of the international CHILDES database, using multilayered morphosyntactic annotation methods.

Taking into account the agglutinative nature of the Kazakh language and drawing on experience working with corpora of Turkic-speaking children, the article examines the marking of parameters specific to the Kazakh language. Furthermore, the article describes the main stages of creating a corpus of children's spoken language and notes that the project is currently in its initial stages. This study allows us to create a digital ontolinguistic model of the Kazakh language and determine the developmental trajectory of children's linguistic competence.

Key words: ontolinguistics, ontogenesis, corpus studies, children's language, audio corpus, CHILDES, transcription.

REFERENCES

1 Fromkin, V., Rodman, R., & Hyams, N. (2018). *An introduction to language* (10th ed., international ed.). Almaty: National Translation Bureau.
2 Ayapova, T. T. (1988). *The development of children's speech syntax (based on the Kazakh language)* (Abstract of Candidate of Philological Sciences dissertation). Alma-Ata.
3 Baimuratova, B. (1981). *Teaching speech to preschool children*. Almaty: Mektep.
4 Bakradenova, A. B. (1993). *Development of Kazakh speech in preschool children*. Almaty.
5 Omirbekova, K. K., Taszhurekova, Zh. T., Seisenova, A. D., Ibatova, G. B., & Ospanova, A. S. (2021). *Formation of the native language system during ontogenesis*. Almaty: National Scientific and Practical Center for Special and Inclusive Education.
6 Shayakhmet, A. K. (2014). Formation of a bilingual personality in Kazakhstan. In *Multilingualism in educational space. Series "Language and intercultural education"* (Vol. 6). Izhevsk.
7 Tseitlin, S. N. (1978). Reflexive verbs and child speech. In *Problems of grammatical voice theory*. USSR Academy of Sciences, Institute of Linguistics (pp. 128–130). Leningrad.
8 Bolognesi, C., Cinini, A., Cutugno, P., Ferretti, M., & Chiarella, D. (2025). Towards a preschooler corpus of Italian: An experimental journey. *Research Methods in Applied Linguistics*, Article 100252. https://doi.org/10.1016/j.rmal.2025.100252
9 Zyryanova, E. V. (2008). The CHILDES system as a method for collecting data and studying child speech. *Izvestia Herzen State Pedagogical University of Russia*, 76-2. https://cyberleninka.ru/article/n/sistema-childes-kak-metod-sbora-materialov-i-izucheniya-detskoy-rechi
10 Lelik, V., Diachkova, M., Dorofeeva, S. V., Lopukhina, A., Dragoy, O., & Sekerina, I. A. (2025). *CHILDES RusLan-M corpus*. TalkBank. https://doi.org/10.21415/1XMG-D508

ISSN 2788-7979 (online)
11 Aksu-Koç, A., & Slobin, D. I. (1978). Turkish child language data. In *CHILDES: The Child Language Data Exchange System*. TalkBank. https://talkbank.org/childes/access/Other/Turkish/Aksu.html
12 Altınkamış, N. F. (n.d.). *Turkish Altınkamış corpus*. CHILDES: Child Language Data Exchange System. TalkBank. https://talkbank.org/childes/access/Other/Turkish/Altinkamis.html
13 Smagulova, J. (2013). Early language socialization and language shift: Kazakh as baby talk. *Journal of Sociolinguistics*, 17(3), 309–336. https://doi.org/10.1111/josl.12030
14 Krasnoperova, E. S. (2024). Grammatical corpus of child speech as a tool for studying the child's actual grammaticon. *Philological Class*, 29(2), 108–118. https://doi.org/10.26170/2071-2405-2024-29-2-108-118
15 MacWhinney, B. (2000). *The CHILDES project: Tools for analyzing talk* (3rd ed.). Mahwah, NJ: Lawrence Erlbaum Associates.

Received: 28.03.2026