

Қ.Б. Кудеринова¹, Г. Борибаева², Ә.М. Кожакмет^{3*}

^{1,2}А. Байтұрсынұлы атындағы Тіл білімі институты, Алматы, Қазақстан;

³Әл-Фараби атындағы Қазақ ұлттық университеті, Алматы, Қазақстан

(e-mail: kuralai.kuderinova@gmail.com, boribayeva.gulnara@gmail.com, sdu.aseem@gmail.com)

Ауызша ішкорпустарды әзірлеудегі әлемдік және отандық тәжірибе

Тіл білімінде тілдердің ұлттық корпусын құрастыру мәселесі жан-жақты қарастырылып жүр. Ал жасанды интеллектінің дамыған заманында ауызша сөйлеу үлгілерінің лингвистикалық деректер базасы болып саналатын ауызша ішкорпустарды зерттеу — өзектілігі жоғары, назар аударуды қажет ететін тақырыптардың бірі. Мақалада алғашқы шетелдік және отандық ауызша корпустардың құрылу тарихы сипатталып, корпустық зерттеулердің қазіргі жағдайы, Ахмет Байтұрсынұлы атындағы Тіл білімі институты әзірлеген «Қазақ тілінің ауызша ішкорпусының» сипаты, сөз жинау тәсілдері көрсетілген. Орыс тілі ұлттық корпусының ауызша ішкорпусы, Санкт-Петербург мемлекеттік университетінің ғалымдары құрастырған орыс ауызша сөзінің корпусы, британ ұлттық корпусының ауызша ішкорпусы, араб сөзінің корпусы сияқты ауызша ішкорпустардың құрастырылу қағидалары, олардағы сөзқолданыстар мен дереккөздер саны салыстырылып баяндалған. Ауызша корпустардағы метабелгілем, тілдік деректерді орфоэпиялық және орфографиялық аннотациялау үлгілеріне аналитикалық шолу жасалған. Мақалада, сонымен бірге А. Байтұрсынұлы атындағы Тіл білімі институтының ғалымдары әзірлеген «Қазақ тілінің ұлттық корпусының» бірі «Ауызша ішкорпустың» мазмұны, сөз жинау әдістері туралы толық сипаттама берілген. Ауызша ішкорпус алдымен шаршытоп алдындағы сөздер, сұхбаттар, ресми, бейресми сөздермен толықты, қарапайым еңбек адамдары мен жергілікті тұрғындарының сөйлеу тілінен бөлек, ақын, жазушы, қоғам қайраткерлерінің аудио және бейнеформаттағы сөзі жинақтала бастады. М. Әуезов, Ғ. Мүсірепов, С. Мұқанов сынды жазушылардың 60-жылдары жиналыстарда сөйлеген сөзінен бастап, сөзі үлгі болып келген басқа да танымал тұлғалардың сөздерінің орфографиясы жазылып, орфоэпиясы түсіріліп, просодикасы қойыла бастады. Бұл, бір жағынан, қазақ ауызша сөзінің әр жылдардағы қоры түрінде болса, екінші жағынан, белді тұлғалардың таспаларда ғана ауызша нұсқада сақталып келген сөздерінің жазба нұсқасын алуға мүмкіндік берді. 2024 жылдан бастап ауызша ішкорпусқа қазіргі ауызекі сөйлеу тілінің үлгілері, жергілікті сөздер алына бастады. Жоба идеясы бойынша 2025 жылы Ахмет Байтұрсынұлы атындағы Тіл білімі институты ауызша корпус әзірлемешілері қазақ тілінің дәстүрлі сөзсаптамын білетін, ауызша сөз машығына төселген жергілікті тұрғындардың айтқан әңгімелерін жинау, таспаға басу, қазақ тілінің ұлттық корпусына енгізу және зерттеу мақсатында Қазақстанның әр түрлі өңірлеріне сөзтанымдық экспедицияға шықты. Сапар барысында қазақ тілінің дәстүрлі сөзсаптамын, машықсыз, дайындықсыз, шынайы, табиғи сөйленген сөзі, елдімекендегі әңгіме айтуға шебер, ойы жинақы, айтары бар, сөзге шешен тұлғалардың кез келген ортадағы сөйлеген сөздері бейне-аудиотаспаға түсіріліп, «Қазақ тілінің ұлттық корпусына» енгізілді. Қазақ тілінің ұлттық корпусы – қазақ тілінің тілдік жүйесі толық қамтылған, аннотацияланған мәтіндер жинағы. Ұлттық корпусты әзірлеу тіл білімі ғана емес, жалпы ғылым саласындағы басым бағыттардың бірі және мемлекеттік деңгейдегі маңыздылығы жоғары жоба деуге болады.

Кілт сөздер: корпустық лингвистика, ұлттық корпустар, ауызша ішкорпустар, аннотациялау, метабелгіленім.

Kіpіcne

Корпустық лингвистика саласындағы зерттеулер тіл білімі саласында үлкен серпіліс жасады. Зерттеу нысаны ретінде ауқымды тілдік деректер қоры қызмет ететін корпустық лингвистика зерттеулері тілді тану мен түсінуде новаторлық құралға айналды. Тіл корпустарын пайдалану арқылы тіл мамандары тілдің заңдылықтарын, тілдік жүйелердің құрылымын, тілдің қолданыс аясын, тілдің даму жолдарын тереңірек зерттеуге мүмкіндік алып отыр.

Корпустық зерттеулер қолданбалы сипатымен ерекшеленеді, себебі корпустық лингвистика тілдің қыр-сырын зерделеуде нақты мысалдарға сүйенеді. Үлкен көлемдегі мәтінді немесе ауызша сөйлеуді талдау үшін корпустық лингвистика арнайы алгоритмдер мен есептеу құралдарын пайдаланады және теориялық мәліметтерді растайтын не жоққа шығаратын эмпирикалық дәлелдер ұсына алады.

Қазіргі корпус лингвистикасында зерттеулер бірнеше заманауи парадигмаларға негізделеді. Ең алдымен, **қолданысқа негізделген (usage-based)** тәсіл тілдік бірліктердің жиілігі мен контекстегі қолданысын талдау арқылы тіл жүйесін түсіндіруге бағытталған. Сонымен қатар **корпусқа негізделген (corpus-based)** және **корпусқа бағытталған (corpus-**

* Хат-хабар үшін автор. E-mail: sdu.aseem@gmail.com

driven) тәсілдер кеңінен қолданылады: біріншісі алдын ала қойылған гипотезаларды тексеруге бағытталса, екіншісі деректердің өзінен заңдылықтарды анықтауға негізделеді.

Сондай-ақ қазіргі зерттеулерде көпдеңгейлі талдау (лексикалық, морфологиялық, синтаксистік, прагматикалық) мен үлкен деректерді (big data) өңдеу қағидалары маңызды орын алады. Корпус лингвистикасы бүгінде когнитивтік және психолингвистикалық бағыттармен тығыз байланыста дамып, тілдік сананы, дискурс құрылымын және тілдік өзгерістерді зерттеуге мүмкіндік береді.

Цифрлық гуманитаристика (Digital Humanities) әдістері бұл үдерістерді жаңа деңгейге көтереді. Олардың қатарына мәтіндерді автоматты түрде өңдеу (NLP), аннотациялау, мәтіндік деректерді белгілеу (tagging), жиілік талдауы, коллокациялық талдау, кластерлеу және желілік (network) талдау жатады. Сонымен қатар визуализация құралдары тілдік деректерді график, карта немесе семантикалық желі түрінде көрсетуге мүмкіндік береді.

Соңғы онжылдықтарда корпустық зерттеулердің нәтижесі халықаралық деңгейдегі конференцияларда талқыланып, оларға арналып оқулықтар мен монографиялар жарияланып жатса да ауызша және жазбаша мәтіндер корпустарының құрылуы, қызмет етуі мен дамуы туралы сұрақтар әлі де шешімен табуды қажет етеді.

Десек те, қазіргі корпустық зерттеулердің ауқымы едәуір кеңейіп, мамандардың назары ауызша ішкорпустарды құрастыру мәселесіне арналып отыр. Ауызша ішкорпустардың дамуы тіл үйрену, жасанды интеллект және сөйлеу технологиясын қоса есептегенде тіл білімі үшін де, жалпы ғылым үшін де аса маңызды. Себебі ауызша корпустарды әзірлеу тілдің заманауи қолданыстық жағдайын тіркеу үшін қажет. Ауызша сөйлеу үлгілері мынадай тілдік өзгерістерді анық көрсететіні сөзсіз:

- Тілдің фонетикалық, морфологиялық, синтаксистік, сөзжасамдық, стилистикалық нормаларының өзгеру немесе даму, жетілу онтогенезі;
- тілге жаңа сөздердің кіруі;
- фонеманың жаңа варианттары мен вариацияларының пайда болуы;
- грамматикалық, лексикалық жаңа тіркесімдердің пайда болуы;
- грамматикалық нормалардың өзгеруі;
- сөз тіркестерінің жаңа типтерінің пайда болуы;
- жиі қолданыстағы сөздердің сипаты;
- геосаяси факторларға байланысты тілдің трансформациялануы;
- әлеуметтік факторлардың сөйлеуге ықпалы;
- аймақтық, жергілікті сөйлеу ерекшеліктері;
- сөз интонациясындағы өзгерістер және т.б.

Әдебиеттерге шолу

Корпустық лингвистика термині ғылыми айналымға XX ғасырдың 80-ші жылдары енгенімен [1], тілді корпустық тұрғыда зерттеу мәселесінің тарихы одан әлдеқайда тереңде. Тілді корпустық тұрғыда зерттеу әдіснамасы 1940 жылы Боас сияқты ғалымдардан бастау алып, Сепир, Блумфилд, Пайк сияқты ғалымдардың еңбегінде жалғасын тапты [2]. Ол кездегі корпустық зерттеулер үшін ғалымдар қағаз бетіндегі мәтіндерді жинап өндесе, қазіргі технология мен жасанды интеллектің дамуы арқасында бұл тақырып аясындағы зерттеулер аса өзекті болып отыр.

Есептеу техникасы мен ауқымды жадысы бар қуатты компьютерлердің пайда болуы бүгінгі таңда көлемді корпустармен жұмыс істеуге мүмкіндік береді. Корпустар мен компьютерлік технологиялардың тоғысуы корпустық әдіснамаға деген қызығушылықты арттыра түсті. Алғашқы корпус болып ағылшын тілінің америкалық нұсқасы бойынша 1960-жылы Браун университетінің ғалымдары құрастырған Браун корпусы есептеледі. Аталған корпус шыққаннан кейін ұлттық корпустарды құрастыру мәселесінде серпіліс орын алды деуге болады [3]. Браун корпусына әрқайсы екі мың сөзден тұратын 500 мәтін фрагменті енгізілген. Нәтижесінде корпуста енген 1 млн сөзқолданыс корпус құрастырудағы стандарттарды белгіледі. Қазіргі әлемдік корпустардағы сөзқолданыстардың саны миллиардтап есептеледі [4].

1970 жылдары Л.Н. Засорина Браун корпусының үлгісіне ұқсас орыс тілінің жиілік сөздігін құрастырды. Бұл корпустың да көлемі 1 млн сөзқолданысты құрады. Корпуста тең көлемде көркем әдебиет мәтіндері, әлеуметтік-саяси және ғылыми бағыттағы мәтіндер енді. Осы үлгі бойынша жасалған тағы бір корпус 1980 жылдары Швецияның Уппсала университетінде құрастырылған орыс тілінің корпусы болды [5].

Корпус – арнайы лингвистикалық критерийлер бойынша іріктеліп, тілдің үлгісі ретінде қолданылуы үшін жинақталған тілдік фрагменттер жиынтығы [6]. Корпустарды мәтіндер жинағы мен мұрағаттардан ерекшелендіретін қыры осы критерийлер бойынша іріктелуі:

1. Репрезентативтілігі. Корпуста енетін мәтіндер таңдалған тілдің шынайы көрінісін бере алатындай болып іріктелуі қажет. Сондықтан корпуста жүктелетін мәтіндер тілдің барлық ерекшелігін көрсететін сипаттарға ие болып, әртүрлі жанрлардан жинақталады.
2. Теңгерімділігі. Әртүрлі типтегі мәтіндердің үлесі барынша тең болуы тиіс.
3. Корпус көлемі. Корпустың көлемі әртүрлі болуы мүмкін. Оған байланысты қандай да бір шектеу жоқ. Алайда корпуста енген сөзқолданыстардың саны үнемі есептеліп отыруы керек. Бұл корпустағы сапалық мәліметтерді сандық форматқа көшіру үшін маңызды. Ал деректерді сандық форматқа көшіру лингвистикалық статистика бойынша күрделі формулаларды жүзеге асыру үшін керек.
4. Электронды формат. Қазіргі корпустардың басты сипаты олардың сандық форматта болуы. Корпустарды сақтау мен өңдеудің электронды форматы корпустардан ақпарат іздеуде ыңғайлы және уақыт үнемдеуде тиімді. Мәтіндерді өңдеудегі жылдамдықтан басқа, электронды корпустармен жұмыс істеуде қате жіберу қаупі төмен. Себебі мәтін фрагменттерін өңдеуде мамандар бір немесе бірнеше мәрте қате жіберу мүмкін. Ал бұл сияқты монотонды әрекеттерді орындауда компьютерлердің қате жіберуі мүмкін емес.
5. Белгіленім мен аннотацияның болуы. Корпустарды басқа мәтіндер жинағынан ерекшелейтін тұсы дәл осы белгіленімге байланысты. Метабелгіленім мен лингвистикалық аннотацияның жасалуы корпустағы сөзқолданыстарды

тануға көмектеседі. Корпустағы мәтіндерді аннотациялау кезінде оларды орфографиялық белгіленім бойынша бөлу, яғни токен жасау, одан кейін лемматизациялау және морфологиялық талдау жасау мәселесі ең маңызды [7].

Компьютерлік корпустар кездейсоқ мәтіндер жиынтығы емес. Корпусқа жинақталатын мәтіндер арнайы мақсатпен іріктеледі және олардың репрезентативтілігіне аса мән беріледі [8].

Соңғы жылдары корпус лингвистикасы мен сөйлеу технологиялары жасанды интеллект пен нейрондық модельдердің дамуына байланысты жаңа деңгейге көтерілді. Қазіргі зерттеулер корпус деректерін *deep learning* әдістерімен өңдеуге, мультимодальды талдауға және автоматты аннотациялауға бағытталған [9-10]. Сонымен қатар сөйлеу технологиялары саласында үлкен көлемді көптілді корпустар [CoVoST 2; Speak & Improve Corpus 2025]. автоматты сөйлеуді тану және синтездеу жүйелерін дамытуда маңызды рөл атқарады [11-12]. Бұл үрдістер корпус лингвистикасының цифрлық гуманитаристика және AI бағыттарымен интеграциясын күшейтіп отыр [13].

Зерттеу материалдары мен әдістері

Зерттеу материалы ретінде қазақ және орыс тілі ұлттық корпусының ауызша ішкорпусы, Санкт-Петербург мемлекеттік университеті ғалымдары құрастырған орыс ауызша сөзінің корпусы, Ланкастер университетіндегі ESRC CASS орталығы мен Cambridge University Press әзірлеген британ ұлттық корпусының ауызша ішкорпусы [Spoken BNC2014] сондай-ақ араб сөзінің корпусы алынды [14].

Зерттеуде қолданылған әдістер нақты деректерге бағытталып қолданылды. Талдау әдісі әрбір корпусқа қатысты корпус көлемін, мәтін типтерін, аннотация деңгейлерін және метадеректер құрылымын сипаттауға қолданылды. Мысалы, орыс және британ корпустарындағы морфологиялық, семантикалық және социоллингвистикалық белгіленімдер жеке-жеке қарастырылды, ал қазақ ауызша ішкорпусында орфоэпиялық және просодикалық аннотация ерекшеліктері талданды.

Жүйелеу және сараптау әдістері әртүрлі корпустардың функционалдық мүмкіндіктерін (іздеу жүйесі, интерфейс, сүзгілеу параметрлері) біріздендіріп, оларды салыстырмалы сипатта топтастыру үшін пайдаланылды. Осы әдіс арқылы корпустардың құрылымдық және қолданбалы ерекшеліктері ортақ параметрлер бойынша жіктелді.

Синтездеу әдісі корпус лингвистикасына қатысты негізгі ұғымдарды (ауызша корпус, аннотация, метадеректер) жалпылап, зерттеу аясында теориялық негіз қалыптастыру үшін қолданылды.

Салыстыру әдісі нақты көрсеткіштер бойынша жүзеге асырылды: корпус көлемі (сөзқолданыс саны), аннотация деңгейлері, метадеректердің толықтығы, сондай-ақ қолданушыға қолжетімділік деңгейі. Осы салыстыру нәтижесінде әрбір корпустың артықшылықтары мен шектеулері айқындалды.

Нәтижелер мен талқылау

Тіл білімінде корпус деп машинаның оқуына ыңғайлы, лингвистикалық тұрғыда аннотацияланған жазбаша немесе ауызша формадағы мәтіндердің жиынтығы танылады [15]. Жасанды интеллекті технологияларының дамуы сан түрлі корпус пен ішкі корпусты жасауға мүмкіндік береді.

Бүгінгі лингвистикалық зерттеулердің басты нысаны дискурс, сөйлеу және жалпы коммуникативтік бірліктер екенін ескерер болсақ, ауызша ішкорпустарды құрастыру мен дамыту мәселелері аса өзекті.

Ауызша ішкорпус немесе сөйлеу корпустары — сөйлеу аудиофайлдарының мәліметтер базасы. Жинақталған аудиофайлдарға мәтіндік транскрипция жасалады және ол транскрипциялардың көмегімен акустикалық модельдер жасалады. Ал акустикалық модельдерді сөйлеуді тану механизмдерімен байланыста қолдануға болады [16].

Ауызша ішкорпустардың екі негізгі түрі ажыратылады [17]:

1. Оқылып, дыбысталған корпус. Мысалы, кітаптардан үзінді; жаңалықтар; кез келген сөздер тізбегі; сандық деректер және т.б.

2. Дайындықсыз жүзеге асатын спонданды (машықсыз) сөйлеу. Мысалы, диалогтар (екі не одан да көп тұлғалар арасындағы тілдесулер); жеке сөз, яғни монолог (қандай да бір адамның оқиғаларды баяндауы, әңгімесі); сұхбаттар мен басқа да телевизиялық жобалар және т.б.

Ауызша ішкорпустардағы мәтіндердің ғылыми әлеуетін арттырып, зерттеу материалы ретіндегі тартымдылығын арттыру үшін оларды аннотациялау қажет. Аудиофайлдарды аннотациялау келесі белгіленімдерді қамтиды:

- біріншіден дыбысталған сөздер стандартты орфография ережелеріне сай транскрипцияланады;
- екіншіден буын саны, дыбыс сипаты, интонация туралы қосымша ақпарат беріледі;
- үшіншіден, бос және гезитациялық кідірістер жайлы ақпарат беріледі.

орфографиялық және фонетикалық транскрипциялар, әдетте, қарапайым мәтіндік файлдар күйінде болады [18].

Ауызша корпустарды пайдалану арқылы жүргізілетін фонетикалық зерттеулер келесі сипатта болуы мүмкін:

- әртүрлі диалектілер, жергілікті ерекшеліктер, сөйленістер арасындағы сөйлеу үлгілерін салыстыру;
- ерлер мен әйелдердің сөйлеу үлгілерін салыстыру;
- машықсыз сөйлеу кезіндегі нормасыздықтың сипаты;
- кідірістердің жалпы сөз мағынасына, сөз нормалылығына әсері;
- тілдердің фонетикалық, просодикалық өзгерістер сипатын зерттеу;
- жаңа дыбыстар мен дыбыстық реңктердің пайда болуын зерттеу;
- дыбыстардың түсіп қалуы, орфографияда таңбалану-таңбаланбау негіздерін айқындау;
- тілдің палатизациялану құбылысын қарастыру;
- интонациялық үлгілерді құрастыру [19]

Ауызша корпусқа арналған мәтіндерді жинақтау арнайы реті бар рәсімнен тұрады. Бұл рәсімнің орындалуы ауызша корпустардың сапасына тікелей әсер етеді. Сондықтан корпустарды толтыру үдерісі өндірістегі ISO жүйесі секілді автоматтандырылған болуы тиіс.

Төменде ауызша корпустарды жинақтаудың жалпы стандартталған, әлемдік корпустық лингвистикада мойындалған үлгі берілген. Негізінен, реттілікті сақтау маңызды, алайда бірнеше әрекет қатар орындалуы мүмкін немесе әр ұлттық корпус өз тілінің мүддесінен өз мақсатына қажеттісіне алмастыруы мүмкін.

1. Жоба дизайнын әзірлеп, мақсатын айқындау. Бұл кезеңде корпус спецификациясы анықталып, корпус көлемі, сөйлеушілер саны, сөйлеу стилі және таспалауға арналған құрылғылар айқындалады.
 2. Дайындық кезеңі. Жоспар бойынша дайындық жұмыстарын бастау. Мәтіндерді енгізуге арналған нұсқаулықтарды әзірлеу, бағдарламалық қамтамасыз етумен құрылғыларды іске қосып тексеру, сөйлеушілерді табу.
 3. Мәтін жинаудың алдындағы кезең. Егер ауызша корпус көлемі өте үлкен болса, әртүрлі үлгілерді жинау аса маңызды. Бұл негізгі мәтін жинау кезеңінде қателерді болдырмау үшін қажет.
 4. Алдын ала тексеру кезеңі. Жинақталған мәтіндерді алдын ала тексеріп, бағалау кезеңі.
 5. Мәтін жинау кезеңі.
 6. Аннотациялау.
 7. Лексикографиялық сөздік құру.
 8. Екінші мәрте тексеру кезеңі. Корпустың критерийлерге сай келуін тексеру. Корпусты мақұлдау.
 9. Таныстыру және тарату [20].
- Корпусты құрастыру процедурасы ғана емес, корпустың өзі де стандартталған болуы керек.

1 - к е с т е

Ауызша корпусты құрастырудың спецификациясы

№	Спецификация	Түсіндірме
1	Сөйлеушілер спецификациясы	Сөйлеушінің жасы, жынысы, білімі, жұмыс тәжірибесі, дауыс сапасы, тілі және акценті ескеріледі
2	Корпус дизайнының спецификасы	Корпустың ұйымдастырылуы мен мазмұнының сипаттамасы. Мысалы, мәтінді оқу, монолог және диалогтардың сценарийі бойынша нұсқаулықтарды әзірлеу. Корпус сценарийін таңдау үшін қолданылатын лингвистикалық алгоритмдерді әзірлеу
3	Таспалау спецификациясы	Таспалайтын құрылғыларға, қоршаған орта жағдайларына, таспалайтын платформаларға және дискретизация жиілігі мен сөйлеу толқынының форматы сияқты мәліметтерді сақтау стратегияларына арналған техникалық сипаттамаларды беру
4	Аннотациялау спецификациясы	Символдар транскрипциясында дыбыстарды аннотациялау ережесін жазу. Синтаксистік аннотациялау қағидаларын анықтау
5	Тексеру критерийлері	Корпус сай келуі тиіс талаптарды айқындау. Корпусты мақұлдау кезінде тексерілуі тиіс сипаттамаларды бекіту
6	Тарату спецификациясы	Корпустардың мақсатты аудиториясын, таралу аймағын айқындау

Орыс тілі ұлттық корпусының ауызша ішкорпусы

Орыс тілі ұлттық корпусының ауызша ішкорпусы 2007 жылдан бастап дербес түрде жұмыс істеп келеді. Ішкорпустың авторлары — Ресей ғылым академиясының ғалымдары. Аталған корпусқа күнделікті ауызекі сөйлеу үлгілері мен алдын ала дайындалған көпшілік алдында сөйлеу үлгілері, кинофильмдер, театр қойылымдары транскриптілері және авторлар мен кәсіби дикторлардың оқыған көркем проза мәтіндері енген.

Ауызша ішкорпуста орыс тілінің стандартты орфографиясы қолданылады. Материалдар транскрипт күйінде беріледі, дыбыстық мәтін қолжетімсіз. Бұл корпустың көлемі 14 млн сөзқолданысты құрайды және уақыт диапазоны 1900-2020-жылдар аралығы. Мәтіндер функционалдық тұрғыда әртүрлі және Мәскеу, Санкт-Петербург, Воронеж, Саратов, Таганрог, Екатеринбург, Ульяновск, Новосибирск сияқты кең географияны ескерген [21].

Ауызша ішкорпуста орыс тілінің ұлттық корпусында қолданылған стандартты морфологиялық, семантикалық, метамәтіндік және ауызша ішкорпусқа арналған әлеуметтанулық белгіленім пайдаланылған. Әрбір фраза немесе репликаға сөйлеуші туралы ақпарат берілген: жасы, жынысы, туған жылы және еңбек ету саласы. Корпустан лексикалық, морфологиялық және семантикалық ақпаратты іздеу мүмкіндігі бар. Сонымен қатар қолданушылар өз қажеттіліктері бойынша біртекті мәтіндерді іріктеп, әлеуметтанулық параметрлер бойынша өздерінің ішкорпустарын қалыптастыра алады. Мысалы, ауызша ғылыми мәтіндер ішкорпусын, әңгімелесулер ішкорпусын, тұрмыстық микротілдісіндерді, әртүрлі жас өкілдерінің немесе Ресей Федерациясының әртүрлі аймағында орын алған әңгімелерді іріктеп ала алады (ruscorgo.ru) [22].

Орыс ауызша сөзінің корпусы

Санкт-Петербург мемлекеттік университетінің ғалымдары құрастырған тағы бір ауызша корпус — «Орыс ауызша сөзінің корпусы». Бұл орфографиялық және акустика-фонетикалық транскрипциясы бар дыбысталған мәтіндер жиынтығы. Аталмыш корпусты құрастырудағы қажеттілік тіл иесі болып саналатын сөйлеушілердің сөйлеуді қабылдау үдерісін зерттеумен байланысты. Жинақталған материал сөйленісті қабылдаудың функционалды үлгісін жасауда және ең бастысы, табиғи сөйлесу барысында тіл иесі қандай сигналды өңдейтінін анықтауда қолданылады.

Қазіргі таңда корпуста 22 мың сөзқолданыс бар: дыбысталған кәсіби дикторлар оқуы, қатардағы тіл иелмендерінің оқуы, балалар сөзі, ересектер сөзі және т.б. Орфографиялық аннотациялауда орыс тіліндегі «ё» әрпі қалпына келтірілген. Корпуста XX ғасырдың 40 жылдарынан бүгінгі күнге дейінгі мәтіндер жинақталған. Корпус аясында орфографиялық сөз формаларының (12 мың) жиілік сөздігі құрастырылған [<https://russpeech.spbu.ru/>] [23].

Британдық ауызша ұлттық корпусы [Spoken BNC2014]

Ланкастер университетіндегі Экономикалық және әлеуметтік зерттеу кеңесінің (ESRC) Әлеуметтік ғылымдар бойынша корпустық зерттеулер орталығы (CASS) мен Кэмбридж университетінің баспасы (Cambridge University Press) британдық ағылшын тілінің ауызша нұсқасының корпусын әзірледі. Аталған корпус бейресми мәнмәтіндерден тұрады. 2014 жылы құрастырылған бұл корпус қоғамда Ауызша британ ұлттық корпусы атауымен белгілі (Spoken BNC).

Бұл корпус 2012 мен 2016 жылдар аралығындағы мәліметтерді қамтиды. 2017 жылдан бастап Ланкастер университетінің COP web серверінде жұмыс істейді. Ал XML негізгі файлдары 2018 жылдың соңынан бастап қолжетімді. Ауызша корпус 11,5 млн сөзқолданыстан және 1251 таспадан тұрады. Барлығы 668 сөйлеуші қатысқан.

Spoken BNC2014 Британ ұлттық корпусының (1994) ауызша корпусына негізделген [24]. Британ ұлттық корпусының ауызша компоненті британдық ағылшын тілінің ауызша ең ауқымды корпустарының бірі болып есептеледі [25].

BNC1994 құрастырудың мақсаты – жалпы тіл саласындағы зерттеулердің мүмкіндік аясын кеңейту. Бұл корпус 1991 және 1994 жылдар аралығында құрастырылып, екі бөлімнен тұрды [26].

BNC1994DS бейресми, күнделікті қарым-қатынаста қолданылатын сөйлеу үлгілерінен тұрады. Корпус қатысушылары – өз сөздерін таспаға басып жүруге келісімін берген еріктілер. Қатысушылар жас ерекшелігі, жынысы, әлеуметтік рөлі мен географиялық аумақтары бойынша іріктеліп алынған. 124 ересек адам жалпы саны 1000 адамның сөзін таспаға жазған. BNC1994DS 4,2 миллион транскрипцияланған әңгімелер жинағынан тұрады.

Корпустың мәмләтіндік басқарылатын бөлімі ресми кездесулерден институционалды келісімдерге сай таспаға жазылып алынған сөздерден тұрады. Жинақталған сөздер тақырыптар бойынша категорияларға бөлінген.

Лондон жасөспірімдер тілінің Берген корпусы

BNC1994DS құрамына Лондон жасөспірімдер тілінің Берген корпусы кіреді. Аталған корпус 1993 жылы Лондон қаласы айналасындағы елді мекендерден 13 және 14 жас аралығындағы жасөспірімдердің әңгімелерінен жинақталған жарты миллионға жуық сөзден тұрады [27].

BNC1994 ауызша корпусының дизайны мен метамәліметтері тұрғысынан кемшіліктері болғанымен, соңғы онжылдықтарда лингвистикалық зерттеулер үшін өнімділігі жоғары ресурс ретінде қызмет етіп келеді. Бұл корпус грамматика, әлеуметтік лингвистика, прагматика және тілдерді оқыту салаларындағы зерттеулерге ықпал етті.

BNC1994 ауызша корпусының кеңінен қолданылуы корпустың тегін қолжетімділігімен байланысты.

Британдық ағылшын тілінің басқа да бейресми немесе мамандандырылмаған ауызша корпустары бар. Ондай корпустарға деген қолжетімділік төмен. Кейбір корпустар тек коммерциялық мақсатта құрастырылып, көпшілікке ұсынылмаған. Мысалы, *Кэмбридж және Ноттингем дискурс корпусы* тек Кэмбридж университетінің баспасына тиесілі және ғылыми қауымдастық үшін қолжетімсіз. Корпустардың кейбіреуі лицензиялық жарнаны төлеген соң ғана қолжетімді болады. Ал Collins Publishers Word Banks Online 57 млн сөзқолданысы бар Bank of English ішкорпусына ақылы түрде қолжеткізуді ұсынады. Bank of English ішкорпусы америкалық және британдық ағылшын тілінің 61 млн сөзінен тұрады.

Ағылшын тілінің халықаралық корпусының британдық компоненті (ICE-GB) де ақылы корпустардың бірі. Корпуста ауызша да, жазбаша да деректер бар [28].

Ағылшын тілі корпустарының арасында тегін, бірақ британдық ағылшынның аймақтық нұсқалары бойынша ғана мәлімет жинақталған түрлері бар. Мысалы, мәтіндер мен сөйлеудің шотландтық корпусы. Бұл корпусты тегін қолдануға болады, бірақ тек шотландтық ағылшын бойынша ғана ақпарат берілген.

Деректер қоры, қолданушылар саны жағынан BNC1994 ауызша корпусымен бәсекелесе алатын корпустардың болмауы оның төмендегідей күшті жақтарының болуымен байланысты:

- орфографиялық транскрипциясы қамтылған мәліметтер;
- ауқымды мәліметтер қоры;
- британдық ағылшын тілін толық қамтуы;
- тегін қолжетімділік.

Ағылшын тілінің басқа ауызша корпустары жоғарыдағы төрт параметрге қатар жауап бере алмайды.

Араб сөзінің корпусы

Деректер қоры аз болғанымен, бірақ қолданылу жағынан танымалдығы жоғары корпустардың бірі – араб сөзінің корпусы. Араб сөзінің корпусы сөйлеуді синтездеуге арналған қазіргі стандартты араб тілінің (MSA) сөйлеу корпусы. Корпус қазіргі стандартты араб тілінің 3,7 сағаттан асатын сөйлеу үлгілерінен тұрады. Корпус аннотациясы жекелеген фонемалардағы екпіндерді белгілеу арқылы жасалған.

Араб сөзінің корпусы Саутгемптон университетіндегі докторлық диссертациялардың бірінің аясында жасалған. Бұл зерттеуді MicroLinkPC қаржыландырған және коммерцияландыру лицензиясы да аталмыш компанияға тиесілі. Коммерциялық емес мақсатта корпусты Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International лицензиясы арқылы жұмыс істейтін *араб сөзінің корпусы* сайтында қолдануға болады [29].

Корпус негізінен сөйлеуді синтездеу үшін құрастырылған. Алайда ол НММ негізінде дыбыстар шығару үшін де қолданылып келеді. Сонымен қатар бұл корпус басқа да корпустардың фонетикалық транскрипциясын түзету үшін пайдаланылған.

Араб сөзінің корпусында:

- ауызша сөйленімдерден тұратын 1813 wav файл;
- мәтіндік сөйленімдерден тұратын 1813 lab файл;
- wav файлдарда кездесетін шекаралары белгіленген фонема белгілері бар 1813 TextGrid файлдар;
- әрбір мәтін жолында phonetic-transcript.txt бар;
- әрбір мәтін жолында orthographic-transcript.txt. [30].

Орфография Buckwalter форматында берілген. Орфографияның бұндай форматта жасалуы араб алфавитін оқымайтын бағдарламалық қамтамасыз етумен жұмыс істеуде ыңғайлы. Корпус орфография негізінде жасалған екпін белгілерін автоматты түрде анықтау – қазіргі стандартты араб тілін (MSA) синтездеудің сапасын арттыратынын дәлелдеу үшін қолданылған.

Қазақ тілі ұлттық корпусының ауызша ішкорпусы

Қазақ тілінің ұлттық корпусы – қазақ тілінің тілдік жүйесі толық қамтылған, аннотацияланған мәтіндер жинағы [31]. Ұлттық корпусты әзірлеу тіл білімі ғана емес, жалпы ғылым саласындағы басым бағыттардың бірі және мемлекеттік деңгейдегі маңыздылығы жоғары жоба деуге болады.

Қазақ тілінің ұлттық корпусын А. Байтұрсынұлы атындағы Тіл білімі институтының ғалымдары әзірлеген. Корпуста 65 млн сөзқолданыс бар. Корпусты жетілдіру нәтижесінде қазіргі таңда 19 ішкорпус қоса жұмыс істеп тұр [32].

Сондай ішкорпустардың бірі – ауызша ішкорпус. Ауызша ішкорпус – қазақ тіліндегі ауызша сөйлеу үлгілерінің аннотацияланған электронды форматтағы жинағы.

Қазақ тілінің ауызша ішкорпусын жасау Ахмет Байтұрсынұлы атындағы Тіл білімі институты директоры Анар Фазылжанованың идеясымен қолға алынып, әлемдік ауызша ішкорпустардан өзіндік ерекшеліктерімен сипатталады.

Ауызша ішкорпусты жасауды бастаған алғашқы жылдар, яғни 2021 жылдары ішкорпуста материал ретінде шаршытоп алдындағы сөздер, сұхбаттар, ресми, бейресми сөздер алынды.

Кез келген тілдің ауызша сөйлеу нормаларының қалыптасып кодификациялануында қоғамдағы беделді, танымал тұлғалардың рөлі зор. Сол себепті қазақ тілінің ауызша ішкорпусына қарапайым еңбек адамдары мен жергілікті тұрғындарының сөйлеу тілінен бөлек, ақын, жазушы, қоғам қайраткерлерінің аудио және бейнеформаттағы сөзі жинақтала бастады. М. Әуезов, Ғ. Мүсірепов, С. Мұқанов сынды жазушылардың 1960-жылдары жиналыстарда сөйлеген сөзінен бастап, сөзі үлгі болып келген басқа да танымал тұлғалардың сөздерінің орфографиясы жазылып, орфоэпиясы түсіріліп, просодикасы қойыла бастады. Және сонымен бірге сол сөздің әдеби тілдегі нормаланған орфоэпиясы қоса берілді. Бұл, бір жағынан, қазақ ауызша сөзінің әр жылдардағы қоры түрінде болса, екінші жағынан, белді тұлғалардың таспаларда ғана ауызша нұсқада сақталып келген сөздерінің жазба нұсқасын алуға мүмкіндік берді.

1940-80-жылдардағы М. Әуезов, Ғ. Мүсірепов, С. Мұқановтың таспада жазылған сөздері М. Әуезов атындағы Әдебиет және өнер институтының сирек қолжазбалар қорынан алынды. Бұл ретте институт директоры профессор **Кенжехан Матыежанов** мырзаға осындай мүмкіндік бергені үшін зор алғысымызды білдіреміз.

2024-жылдан бастап ауызша ішкорпуста қазіргі ауызекі сөйлеу тілінің үлгілері, жергілікті сөздер алына бастады. Жоба идеясы бойынша, мұндай сөз үлгілері ел ішінде, жергілікті сөйлеу тілінде сақталған болуы керек деген оймен 2025-жылы Ахмет Байтұрсынұлы атындағы Тіл білімі институты ауызша корпус әзірлемешілері қазақ тілінің дәстүрлі сөзсаптамын білетін, ауызша сөз машығына төселген жергілікті тұрғындардың айтқан әңгімелерін жинау, таспаға басу, қазақ тілінің ұлттық корпусына енгізу және зерттеу мақсатында Маңғыстау облысының Шетпе ауданына, Атырау облысының Индер ауданына, Алматы облысының Нарынқол, Кеген аудандарына, Павлодар облысының Баянауыл ауданына, Ақмола облысының Зеренді, Бурабай аудандарына сөзтанымдық экспедицияға барды. Сапар барысында қазақ тілінің дәстүрлі сөзсаптамын, машықсыз, дайындықсыз, шынайы, табиғи сөйленген сөзі, әр ауылдағы, елдімекендегі әңгіме айтуға шебер, ойы жинақы, айтары бар, сөзге шешен тұлғалардың (20-90 жас арасындағы) кез келген ортадағы сөйлеген сөздері бейне-аудиотаспаға түсіріліп, «Қазақ тілінің ұлттық корпусына» мәтіндік жазбасы жазылып, сөздің орфоэпиясы, просодикасы, метабелгіленімі қоса қойылып, салына бастады.

Қазақстанның басқа да өңірлерін қалыс қалдырмау мақсатында ауызша әңгімелер, шаршысөз, көсемсөздерге республика бойынша конкурс жарияланды. Конкурста қатысу үшін сөз жинаудың мынадай шарттары белгіленді:

а) ауызша сөз кез келген коммуникативтік қатыста сөйленген сөз болуы керек (үйде, көшеде, тойда, кештерде, сахнада, студияда);

ә) дайындықсыз, машықсыз сөйленген, жаттанды болмауы, шынайы, табиғи сөз болуы маңызды;

б) сөз немесе әңгіменің бас-аяғы болуы, өзі туралы, отбасы, ауыл-елі туралы, ондағы белгілі тұлғалар туралы, ауылының, ондағы жер-су аттарының тарихы туралы, елден естіген әңгімелер, жас күніндегі өмір тіршіліктері, жастық, балалық шағы туралы, қазіргі көрген-білгенін және осы секілді айтары бар барлық бейресми тақырыптар болғаны жөн делінді.

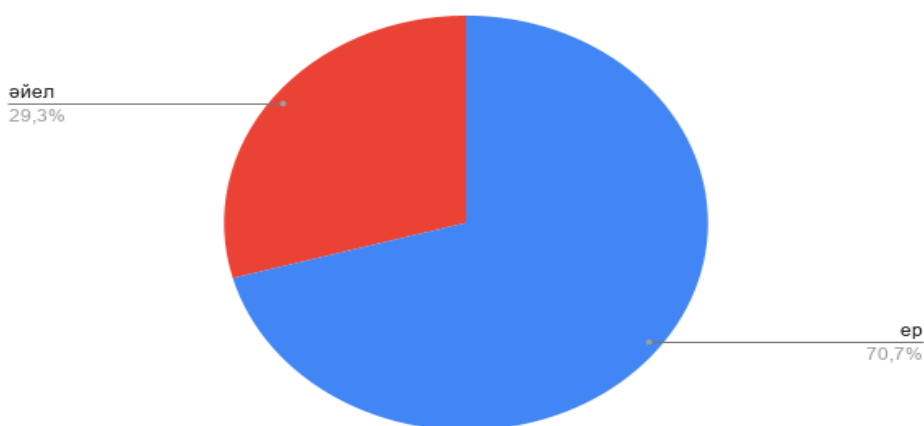
Сөз жинаудың бұл тәсілі арқылы да ауызша ішкорпустың сөз қоры біршама көбейді.

Сонымен, қазақ тілінің ауызша ішкорпусында бүгінде 1 500 000 сөзқолданыс, 300 сұхбат, 200-ден астам сағат сөз салынған. Оның 9,4 %-ы ауызекі сөйлеу стилінде, 64,4 %-ы шаршысөз, 15,4 %-ы сөйлеу тілі, 2,7 %-ы көркемсөз, 2,7 %-ы іскери, 1,3 %-ы ғылыми стильде, 1 %-ы ресми, 1 %-ы діни стиль материалдары болып отыр. Ауызша ғылыми сөз, ауызша көркемсөз үлгілері әзірге аздық етеді.

Қазақ тілінің ауызша корпусына енгізілген 300 сұхбатта ер адамдардың үлесі басым екені анықталды (70,7%). Бұл жағдай зерттеудің бастапқы кезеңінде гендерлік теңгерімді сақтау мақсат болмауымен байланысты. Негізгі назар тілдік материалдың табиғилығы мен репрезентативтілігін қамтамасыз етуге бағытталды.

Дегенмен қазіргі зерттеу талаптарын ескере отырып, мұндай теңгерімсіздік гендерлік тілдік ерекшеліктерді толық талдауға белгілі бір шектеулер тудырады. Осыған байланысты ішкорпусты толықтыруда әйел респонденттердің үлесін арттыруды, гендерлік теңгерімді сақтауды ойластырып отырмыз.

Гендірлік үлес көрсеткіші



1-диаграмма. Қазақ тілінің ауызша ішкорпусындағы гендерлік үлес көрсеткіші

Ішкорпус осы айтылған сипаттағы сұхбаттардан басқа қазақ тіліндегі фильм, мультфильм үзінділері мен бейнеконтенттен тұрады.

«Сұхбаттар» бөлімінің іздеу жүйесі қажет сөзді жазу және сұрыптап іздеу арқылы жүзеге асады. Сұрыптап іздеу кезінде сұхбат берушіні, сөз стильдерін, лингвоәлеуметтік деңгейін, тақырыпты және аймақтың бірін тандап, іздеу батырмасын басқанда сұрау бойынша табылған нәтижелер экранға шығады. Іздеу жүйесінің қай түрін тандалса да, сұхбат беруші туралы мәлімет және сол сөздің орфографиясы, сөйлеушінің орфоэпиясы мен сөздің нормаға түскен орфоэпиясы қолжетімді. Сонымен қатар сол сұхбаттың бейнежазбасын көріп, аудиосын тыңдауға мүмкіндік бар. Сөздің орфографиясы бөліміндегі әр сөзді басқанда фонетика-фонологиялық, просодикалық талдаулар көрсетілген.

«Фильм үзінділері» бөлімінде А1, А2 тілдік деңгейлерге сай таңдалып алынған 40 фильм, 60 мультфильм үзінділері қамтылған. Бұл бөлімнің де іздеу жүйесі қажет сөзді жазу және сұрыптап іздеу арқылы жүзеге асады. Сұрыптап іздеуде атауы, тип, тақырып және деңгей бойынша сұрыпталады. Іздеу нәтижесінде әр үзіндінің орфографиясы, сөйлеуші және норма орфоэпиясы беріледі. Сонымен қатар фильм, мультфильм туралы мәліметтерді оқып, бейнежазбасын көруге болады.

«Бейнеконтент» бөлімінде корпус әзірлемешілері өздері сценарийін жазып, өздері рөлдерде ойнаған А1, А2 тілдік деңгейлерге арналған субтитры бар бейнероликтер орналастырылған.

«Құралдар» бөлімінде ауызша сөз машығын меңгеруге көмекші құрал ретінде орфоэпиялық конвертор бар.

Анар Фазылжанова идеясы бойынша, қазақ тілі ауызша ішкорпусын құрастыру дәл қазіргі кезекте қазақ сөзінің әдеби нормаға сай айтылу заңдылықтарын насихаттауға қызмет ету болуы керек. Яғни қазақ тілінде дұрыс, таза сөйлеймін деген тіл иелмендері мен тіл тұтынушылардың ауызша тілін жетілдіруде қолданатын құралы, сөз үлгісін танытатын аннотацияланған ресурстардың қызметін атқаруы тиіс.

Қазіргі корпус лингвистикасында аннотациялау бірнеше деңгейден тұратыны белгілі: орфографиялық транскрипция, морфологиялық белгілеу (POS-tagging), синтаксистік талдау, семантикалық және прагматикалық белгілеу. Сонымен қатар халықаралық тәжірибеде аннотацияны стандарттау мақсатында **TEI (Text Encoding Initiative), Linguistic Annotation Framework (ISO TC37), EAGLES** сияқты стандарттар кеңінен қолданылады. Бұл стандарттар мәтіндерді XML/SGML форматында белгілеуді, аннотацияның көпдеңгейлі құрылымын және деректердің өзара үйлесімділігін қамтамасыз етеді.

Қазақ тілінің ауызша ішкорпусында қолданылатын аннотациялау жүйесі де осы халықаралық үрдістермен үйлеседі. Корпуста:

- орфографиялық және орфоэпиялық транскрипция;
- фонетика-фонологиялық және просодикалық белгілеу;
- морфологиялық және метамәліметтік аннотация;

• сөйлеушінің әлеуметтік сипаттамалары (жасы, жынысы, аймағы, стилі) сияқты көпдеңгейлі белгілеу түрлері қолданылады. Бұл құрылым халықаралық корпустардағы (мысалы, ICE, BNC) көпқабатты аннотация модельдеріне сәйкес келеді, мұнда мәтіндік, грамматикалық және метадеректер деңгейлері біріктіріледі.

Осылайша қазақ тілінің ауызша ішкорпусындағы аннотациялау жүйесі халықаралық корпус лингвистикасындағы стандарттармен үйлесімді дамып, тілдік деректерді кешенді сипаттауға мүмкіндік береді.

Қазақ тілінің ауызша ішкорпусын құрастыру барысында халықаралық корпус лингвистикасында кеңінен қолданылатын әдістер мен құралдарға сүйену маңызды болды. Ауызша деректерді транскрипциялау және аннотациялау үдерісінде көпдеңгейлі белгілеу принциптері пайдаланылды. Жалпы, қазіргі тәжірибеде мұндай жұмыстарды жүзеге асыру үшін **EXMARaLDA, Praat, ELAN** сияқты сөйлеу корпусын өңдеуге арналған бағдарламалар қолданылады, олар транскрипциялау, аннотациялау және метадеректерді басқаруға мүмкіндік береді. Сонымен қатар мәтіндік және лингвистикалық деректерді өңдеуде корпус құралдары (мысалы, WordSmith, Korpusomat) жиілік, коллокация және статистикалық талдау жүргізуге мүмкіндік береді.

Қазақ тілінің ауызша ішкорпусы мен мақалада сипатталған басқа да ішкорпустардың көлемі, аннотация деңгейлері, метадеректер құрылымы бойынша салыстырмалы талдау 2-кестеде берілген.

Ауызша ішкорпустар құрылымының салыстырмалы талдауы

Корпус атауы	Корпус көлемі	Аннотация деңгейлері	Метадеректер құрылымы
Орыс тілі ұлттық корпусының ауызша ішкорпусы	~14 млн сөзқолданыс (1900–2020)	Морфологиялық, семантикалық, метамәтіндік, социолінгвистикалық	Жасы, жынысы, туған жылы, кәсібі, аймақ
Орыс ауызша сөзінің корпусы (СП6МУ)	~22 мың сөзқолданыс	Орфографиялық, акустика-фонетикалық	Сөйлеуші типі (диктор, бала, ересек), мәтін түрі
Spoken BNC2014 (британдық ауызша корпус)	~11,5 млн сөз, 1251 жазба, 668 сөйлеуші	Орфографиялық, дискурстық, социолінгвистикалық	Жасы, жынысы, әлеуметтік рөлі, география, қатысым түрі
BNC1994 (ауызша компонент)	~4,2 млн сөз (бейресми бөлігі)	Орфографиялық, метамәтіндік, дискурстық	Жасы, жынысы, әлеуметтік мәртебе, аймақ
Араб сөзінің корпусы	~3,7 сағат сөйлеу, 1813 файл	Фонетикалық, орфографиялық, епіндік	Файл типтері, фонема деңгейі, транскрипция түрлері
Қазақ тілі ұлттық корпусының ауызша ішкорпусы	~1,5 млн сөз, 300 сұхбат, 200+ сағат	Орфографиялық, орфоэпиялық, фонетика-фонологиялық, просодикалық, метабелгіленім	Жасы, жынысы, аймақ, стиль, тақырып, лингвоәлеуметтік сипат

Мақалада қарастырылған корпустардың функционалдық мүмкіндіктерін нақтылау мақсатында олардың іздеу жүйесі, интерфейс және қолданушы сценарийлері қосымша сипатталды.

Мысалы, **Орыс тілі ұлттық корпусының ауызша ішкорпусында** қолданушы лексикалық немесе грамматикалық бірлікті іздеу кезінде морфологиялық, семантикалық және социолінгвистикалық параметрлер бойынша сүзгілеу мүмкіндігіне ие. Іздеу нәтижелері сөйлем контекстінде ұсынылып, әрбір репликаға сөйлеуші туралы толық метадеректер (жасы, жынысы, кәсібі) қоса беріледі.

Spoken BNC2014 корпусында интерфейс пайдаланушыға сөздер мен тіркестердің жиілігін, коллокацияларын және дискурстық қолданысын талдауға мүмкіндік береді. Қолданушылар белгілі бір әлеуметтік топтың (мысалы, жастар немесе белгілі бір аймақ өкілдері) сөйлеу ерекшеліктерін сүзгілер арқылы бөліп қарастыра алады.

Орыс ауызша сөзінің корпусында дыбыстық файлдар мен фонетикалық транскрипция қатар ұсынылады. Бұл қолданушыға бірліктің орфографиялық түрін ғана емес, оның нақты айтылуын, интонациясын және акустикалық параметрлерін зерттеуге мүмкіндік береді.

Қазақ тілінің ұлттық корпусының ауызша ішкорпусында іздеу жүйесі сөзді тікелей енгізу немесе параметрлер бойынша сұрыптау арқылы жүзеге асады. Қолданушы сұхбат берушінің жасын, аймағын, сөйлеу стилін және тақырыбын таңдай отырып, қажетті материалды іріктей алады. Нәтижелерде сөздің орфографиясы, орфоэпиясы, нормаланған нұсқасы, сондай-ақ аудио/бейнежазба қолжетімді. Сонымен қатар әрбір сөзге фонетика-фонологиялық және просодикалық талдау ұсынылады.

Осылайша, қарастырылған корпустардың функционалдық мүмкіндіктері оларды тек деректер ретінде ғана емес, тілдік талдауға арналған көпдеңгейлі зерттеу құралы ретінде пайдалануға мүмкіндік береді.

Қорытынды

Корпустық лингвистика тіл білімінің басқа салаларына қарағанда тілдік деректерді сақтау, өңдеу және іздеуде цифрлық технологиялардың мүмкіндіктері мен жетістіктері негізінде жинақталып, өңделген эмпирикалық материалдың орасан зор массивтерін пайдалануға мүмкіндік беретін тіл білімінің ерекше саласы ретінде қалыптасуда. Мұның бәрі лингвистикалық зерттеулердегі шынайылықты қамтамасыз етудің жаңа әдіснамасын зерделеуге және сөйлеу үлгілерін ғылыми материал ретінде қолдануға мүмкіндік береді. Осы орайда ауызша корпустарды құрастырудың өзектілігі айқындала түседі деп ойлаймыз.

Мақалада бірқатар ауызша ішкорпустарға шолу жасалды. Корпустардың деректер қоры, акпарат іздеу алгоритмі, аннотациялануы корпус танымалдығын арттырып, қолданушылар санын арттырудың негізгі параметрлерін құрайтындығы анықталды. Ауызша ішкорпус – сөйлеуді тану, сөйлеуді синтездеу, ауызша және жазбаша сөзді сәйкестендіру сияқты үдерістерді автоматтандыру және жасанды интеллектің қазақша сөйлетуде қолданылатын цифрлық құрал. Кез келген тіл ауызша қарым-қатынас кезінде танылатын ескерер болсақ, ауызша ішкорпустарды құрастыру, толтыру және тарату мәселесін зерттеу аса қажет.

Жүргізілген зерттеу нәтижелері ауызша ішкорпустардың құрылымы, аннотациялау жүйесі және функционалдық мүмкіндіктері әртүрлі болғанымен, олардың барлығы тілдік деректерді көпдеңгейлі талдауға бағытталғанын көрсетті. Әсіресе, қазақ тілінің ауызша ішкорпусы орфоэпиялық, просодикалық және метадеректерді біріктіретін кешенді аннотациясымен ерекшеленеді.

Болашақ зерттеулер екі бағытта дамуы мүмкін. Біріншіден, ауызша корпус деректерін жасанды интеллект және мәшинелік оқыту әдістерімен біріктіру арқылы автоматты аннотациялау, сөйлеу үлгілерін тану және тілдік модельдерді жетілдіру мүмкіндігі бар. Екіншіден, корпус материалдарын автоматты сөйлеуді тану (ASR) және сөйлеуді синтездеу (TTS) жүйелерімен интеграциялау қазақ тіліндегі цифрлық ресурстарды дамытуға ықпал етеді.

Практикалық тұрғыдан алғанда, зерттеу нәтижелері білім беру саласында (тіл үйрету платформалары), орфоэпиялық нормаларды қалыптастыруда, сондай-ақ цифрлық лексикография мен интеллектуалды тілдік жүйелерді әзірлеуде қолданылуы мүмкін.

Ауызша корпустарды дамыту және оларды заманауи технологиялармен интеграциялау тіл білімінің қолданбалы мүмкіндіктерін кеңейтіп, ұлттық тіл ресурстарын жаңа деңгейге көтеруге мүмкіндік береді.

Мақала ҚР БҒМ қаржыландырып отырған BR24993244 «Қазақ тілі ұлттық корпусын Smart-мәтіндер мегажобасы және қазақтілді жасанды интеллект негізі ретінде жетілдіру, ішкорпустарын әзірлеу» іргелі зерттеу бағдарламасы аясында орындалды.

Авторлардың үлесі

Кудеринова Қ.Б. — тұжырымдаманы әзірлеу, әдіснама, қаржыландыруды тарту, жобаны басқару, ғылыми жетекшілік, ресурстармен қамтамасыз ету, валидация, мәтінді редакциялау және толықтыру.

Борибаева Г.А. — зерттеу жүргізу, деректерді жинақтау және басқару, формальды талдау, әдіснама, ресурстармен қамтамасыз ету, валидация, визуализация, мәтінді редакциялау және толықтыру.

Кожахмет А.М. — зерттеу жүргізу, деректерді жинақтау және басқару, формальды талдау, бағдарламалық қамтамасыз ету, визуализация, бастапқы мәтінді жазу, мәтінді редакциялау және толықтыру.

Author contributions

All authors have read and approved the final manuscript and agreed to the published version of the paper.

Әдебиеттер тізімі

- 1 Spoken corpora and linguistic studies / ed. T. Raso, H. Mello. — Amsterdam: John Benjamins Publishing Company, 2014. — Vol. 61.
- 2 Biber D. Styles of stance in English: Lexical and grammatical marking of evidentiality and affect / D. Biber, E. Finegan // Text-interdisciplinary journal for the study of discourse. — 1989. — Vol. 9, No. 1. — P. 93–124.
- 3 Ide N. Preparation and analysis of linguistic corpora / N. Ide // A Companion to digital Humanities. — 2004. — P. 289–305.
- 4 Davies M. Corpora: an introduction / M. Davies // The Cambridge handbook of English corpus linguistics. — 2015. — P. 11–31.
- 5 Грудева Е.В. Корпусная лингвистика / Е.В. Грудева. — М.: Флинта, 2012. — 165 с.
- 6 Sinclair J. Trust the text: Language, corpus and discourse / J. Sinclair. — London: Routledge, 2004.
- 7 Aston G. Learning with corpora: An overview / G. Aston // Learning with corpora. — 2001. — P. 7–45.
- 8 Leech G. Corpora and theories of linguistic performance / G. Leech // Directions in corpus linguistics. — 1992. — P. 105–122.
- 9 Incelli E. Exploring the future of corpus linguistics: Innovations in AI and social impact / E. Incelli // International Journal of Mass Communication. — 2025. — Vol. 3. — P. 1–10.
- 10 Al-onazi B.B. Integration of corpus linguistics and deep learning for language analysis [Electronic resource] / B.B. Al-onazi et al. // Scientific Reports. — 2025. — Access mode: <https://www.nature.com/articles/s41598-025-28929-z>.
- 11 Wang C. CoVoST 2: A massively multilingual speech-to-text translation corpus [Electronic resource] / C. Wang, Y. Tang, X. Ma, A. Wu, D. Okhonko, J. Pino. — Access mode: <https://arxiv.org/abs/2007.10310>.
- 12 Knill K. Speak & Improve corpus 2025: A large-scale speech corpus for language learning and assessment [Electronic resource] / K. Knill. — Access mode: <https://arxiv.org/abs/2412.11986>.
- 13 Pérez-Paredes P. Corpus linguistics and AI in the reconfiguration of language learning ecologies [Electronic resource] / P. Pérez-Paredes // Language Teaching. — 2026. — Access mode: <https://www.cambridge.org/core/journals/language-teaching/article/corpus-linguistics-and-ai-in-the-reconfiguration-of-language-learning-ecologies/A0AD76A187BD0316A1D615D51741AA32>.
- 14 Corpus studies: Current trends and future directions. — [Electronic resource]. — Access mode: https://www.researchgate.net/publication/400502891_CORPUS_STUDIES_CURRENT_TRENDS_AND_FUTURE DIRECTIONS.
- 15 Meijs W. Linguistic corpora and lexicography / W. Meijs // Annual Review of Applied Linguistics. — 1996. — Vol. 6. — P. 99–114.
- 16 Adolphs S. Building a spoken corpus / S. Adolphs, D. Knight // The Routledge handbook of corpus linguistics. — 2010. — P. 38–52.
- 17 Tseng S.C. Spoken corpora and analysis of natural speech / S.C. Tseng // Taiwan Journal of Linguistics. — 2008. — Vol. 6, № 2.
- 18 Wichmann A. Corpora and spoken discourse / A. Wichmann // Corpus Linguistics 25 Years on. — Brill, 2007. — P. 73–86.

- 19 Dobrushina N. Spoken corpora of Slavic languages / N. Dobrushina, E. Sokur // *Russian Linguistics*. — 2022. — Vol. 46, № 2. — P. 77–93.
- 20 Li A. RASC863 - A Chinese Speech Corpus with Four Regional Accents / A. Li, Z. Yin, T. Wang, Q. Fang, F. Hu // *ICSLT-o-COCOSDA*. — New Delhi, 2004.
- 21 Гришина Е.А. Устная речь в Национальном корпусе русского языка / Е.А. Гришина // *Национальный корпус русского языка: 2003–2005*. — М.: Индрик, 2005. — С. 94–110.
- 22 Национальный корпус русского языка. — [Электронный ресурс]. — Режим доступа: <http://ruscorpora.ru>.
- 23 Корпус русской устной речи. — [Электронный ресурс]. — Режим доступа: <https://russpeech.spbu.ru/>.
- 24 Crowdy S. Spoken corpus design / S. Crowdy // *Literary and linguistic computing*. — 1993. — Vol. 8, No. 4. — P. 259–265.
- 25 Lam P. The making of a BNC customised spoken corpus for comparative purposes / P. Lam // *Corpora*. — 2009. — Vol. 4, No. 2. — P. 167–188.
- 26 Aston G. The BNC handbook: exploring the British National Corpus with SARA / G. Aston, L. Burnard. — Edinburgh: Edinburgh University Press, 1998.
- 27 Stenström A.B. Trends in teenage talk: Corpus compilation, analysis and findings / A.B. Stenström, G. Andersen, I.K. Hasund. — Amsterdam: John Benjamins Publishing Company, 2002.
- 28 Ludańska Z. Technology propels speech-language research and theory [Electronic resource] / Z. Ludańska // *Neuroscience & Biobehavioral Reviews*. — 2025. — Access mode: <https://www.sciencedirect.com/science/article/pii/S014976342500199X>.
- 29 Halabi N. Modern standard Arabic phonetics for speech synthesis: Doctoral dissertation / N. Halabi. — Southampton: University of Southampton, 2016.
- 30 Halpern J. Word stress and vowel neutralization in modern standard Arabic / J. Halpern // *2nd International Conference on Arabic Language Resources and Tools*. — Cairo, 2009. — P. 1–7.
- 31 Жұбанов А. Қазақ тілінің ұлттық корпусы және метабелгіленім мәселесі / А. Жұбанов // *Тілтаным*. — 2015. — № 1. — Б. 21–29.
- 32 Qazcorpus.kz. — [Электрондық ресурс]. — Қолжетімділігі: <https://qazcorpus.kz/>.

К.Б. Кудеринова, Г. Борибаева, А.М. Кожамбет

Мировой и отечественный опыт разработки устных подкорпусов

В современной лингвистике активно развивается направление, связанное с формированием национальных языковых корпусов. В условиях стремительного прогресса технологий, включая искусственный интеллект, особую актуальность приобретает изучение устных подкорпусов — специализированных лингвистических баз данных, содержащих образцы спонтанной и нормативной устной речи. Эти подкорпусы служат важным инструментом для анализа фонетических, просодических, синтаксических и прагматических характеристик языка. В данной статье рассматривается история создания первых устных корпусов как за рубежом, так и в отечественной практике. Освещается текущее состояние исследований в области корпусной лингвистики, уделяется внимание современным подходам к разработке и аннотированию устных корпусов. На основе сравнительного анализа описываются принципы построения устных корпусов, включая устный подкорпус Национального корпуса русского языка, Корпус русской разговорной речи (СПбГУ), устный корпус Британского национального корпуса, а также корпус арабской речи. Особое внимание уделяется аналитическому обзору практик метаразметки, а также орфоэпических и орфографических моделей аннотирования языковых данных. Кроме того, в статье представлено полное описание методов сбора лексического материала и содержания устного подкорпуса — одного из корпусов Национального корпуса казахского языка, подготовленного научными сотрудниками Института языкознания им. А. Байтұрсынова. Устный подкорпус изначально был пополнен речами перед аудиторией, интервью, официальными и неофициальными высказываниями. Наряду с устной речью жителей, начался сбор аудио- и видеозаписей выступлений поэтов, писателей и общественных деятелей. С 1960-х годов стали собираться материалы выступлений таких писателей, как М. Ауэзов, Г. Мусрепов, С. Муканов, а также других известных личностей. Их тексты подвергались орфографической фиксации, орфоэпической разметке и просодическому анализу. Это, с одной стороны, позволило создать корпус устной казахской речи разных лет, а с другой — получить письменные версии речей видных деятелей, ранее существовавших только в устной форме на магнитных носителях. С 2024 года в устный подкорпус начали включаться образцы современной разговорной речи, а также региональные варианты. Согласно проектной концепции, в 2025 году разработчики устного корпуса при Институте языкознания им. Ахмета Байтұрсынова предприняли экспедиции по различным регионам Казахстана с целью сбора рассказов местных жителей, владеющих традиционными нормами казахской речи и устной культуры. Эти рассказы фиксировались, записывались и включались в Национальный корпус казахского языка для последующего исследования. В ходе экспедиций особое внимание уделялось фиксации традиционного словоупотребления, неподготовленной, спонтанной, естественной речи; собирались аудио- и видеозаписи выступлений мастеров слова —

людей с выразительной речью и умением импровизировать. Все собранные материалы включены в «Национальный корпус казахского языка».

Ключевые слова: корпусная лингвистика, национальные корпуса, устные подкорпуса, аннотация, метатаразметка.

K.B. Kuderinova, G. Boribayeva, A.M. Kozakhmet

World and domestic experience in the development of spoken corpora

Modern linguistics places significant emphasis on the development of national language corpora. In the era of rapid technological advancement, particularly the rise of artificial intelligence, the study of spoken subcorpora – linguistic databases containing samples of spoken language—has become increasingly relevant and requires focused scholarly attention. These subcorpora are essential tools for analyzing phonetic, prosodic, syntactic, and pragmatic aspects of language use. This article explores the history of the creation of the earliest spoken corpora, both internationally and within domestic linguistic traditions. It provides an overview of the current state of corpus linguistic research, with particular attention to contemporary methodologies for compiling and annotating spoken corpora. The article compares and describes the construction principles of several major spoken corpora, including the Spoken Subcorpus of the Russian National Corpus, the Corpus of Russian Everyday Speech (developed by scholars at St. Petersburg State University), the Spoken Component of the British National Corpus, and the Arabic Speech Corpus. Special emphasis is placed on the analytical review of metadata annotation practices, as well as the use of orthoepic and orthographic models in the annotation of linguistic data. In addition, the article provides a comprehensive description of the methods of collecting lexical material and the content of the Oral Subcorpus—one of the components of the National Corpus of the Kazakh Language, prepared by the researchers of the A. Baitursynuly Institute of Linguistics. The Oral Subcorpus was initially enriched with public speeches, interviews, and both formal and informal utterances. Alongside the spoken language of ordinary workers and local residents, audio and video recordings of poets, writers, and public figures began to be collected. Since the 1960s, materials from speeches by writers such as M. Auezov, G. Musrepov, S. Mukanov, and other prominent figures have been gathered. Their texts underwent orthographic transcription, orthoepic annotation, and prosodic analysis. This, on the one hand, made it possible to build a corpus of Kazakh oral speech from different periods, and on the other, to obtain written versions of speeches by prominent figures that had previously existed only in oral form on magnetic tapes. Since 2024, samples of contemporary colloquial speech and regional varieties have been added to the Oral Subcorpus. According to the project's concept, in 2025 the developers of the Oral Corpus at the A. Baitursynuly Institute of Linguistics launched expeditions across various regions of Kazakhstan to collect stories from local residents familiar with traditional norms of the Kazakh language and oral culture. These stories were documented, recorded, and included in the National Corpus of the Kazakh Language for further research. During these expeditions, special attention was given to capturing traditional word usage, unprepared, spontaneous, and natural speech. Audio and video recordings were made of master speakers—individuals with clear thinking, expressive language, and skill in improvisation. All collected materials have been incorporated into the «National Corpus of the Kazakh Language».

Keywords: corpus linguistics, national corpora, spoken subcorpora, annotation, metadata.

References

- 1 Raso, T., & Mello, H. (Eds.). (2014). *Spoken corpora and linguistic studies*. Vol. 61. Amsterdam: John Benjamins Publishing Company.
- 2 Biber, D., & Finegan, E. (1989). Styles of stance in English: Lexical and grammatical marking of evidentiality and affect. *Text-interdisciplinary journal for the study of discourse*, 9(1), 93–124.
- 3 Ide, N. (2004). Preparation and analysis of linguistic corpora. *A Companion to digital Humanities* (pp. 289–305).
- 4 Davies, M. (2015). Corpora: an introduction. *The Cambridge handbook of English corpus linguistics* (pp. 11–31).
- 5 Grudeva, E.V. (2012). *Korpusnaia lingvistika* [Corpus linguistics]. Moscow: Flinta [in Russian].
- 6 Sinclair, J. (2004). *Trust the text: Language, corpus and discourse*. London: Routledge.
- 7 Aston, G. (2001). Learning with corpora: An overview. *Learning with corpora* (pp. 7–45).
- 8 Leech, G. (1992). Corpora and theories of linguistic performance. *Directions in corpus linguistics* (pp. 105–122).
- 9 Incelli, E. (2025). Exploring the future of corpus linguistics: Innovations in AI and social impact. *International Journal of Mass Communication*, 3, 1–10.
- 10 Al-onazi, B.B. et al. (2025). Integration of corpus linguistics and deep learning for language analysis. *Scientific Reports*. Retrieved from <https://www.nature.com/articles/s41598-025-28929-z>
- 11 Wang, C., Tang, Y., Ma, X., Wu, A., Okhonko, D., & Pino, J. (2020). CoVoST 2: A massively multilingual speech-to-text translation corpus. *arxiv.org*. Retrieved from <https://arxiv.org/abs/2007.10310>
- 12 Knill, K. (2025). Speak & Improve corpus 2025: A large-scale speech corpus for language learning and assessment. *arxiv.org*. Retrieved from <https://arxiv.org/abs/2412.11986>

- 13 Pérez-Paredes, P. (2026). Corpus linguistics and AI in the reconfiguration of language learning ecologies. *Language Teaching*. Retrieved from <https://www.cambridge.org/core/journals/language-teaching/article/corpus-linguistics-and-ai-in-the-reconfiguration-of-language-learning-ecologies/A0AD76A187BD0316A1D615D51741AA32>
- 14 Corpus studies: Current trends and future directions (n.d.). *researchgate.net*. Retrieved from https://www.researchgate.net/publication/400502891_CORPUS_STUDIES_CURRENT_TRENDS_AND_FUTURE_DIRECTIONS
- 15 Meijs, W. (1996). Linguistic corpora and lexicography. *Annual Review of Applied Linguistics*, 6, 99–114.
- 16 Adolphs, S., & Knight, D. (2010). Building a spoken corpus. *The Routledge handbook of corpus linguistics* (pp. 38–52).
- 17 Tseng, S.C. (2008). Spoken corpora and analysis of natural speech. *Taiwan Journal of Linguistics*, 6(2).
- 18 Wichmann, A. (2007). Corpora and spoken discourse. *Corpus Linguistics 25 Years on* (pp. 73–86). Brill.
- 19 Dobrushina, N., & Sokur, E. (2022). Spoken corpora of Slavic languages. *Russian Linguistics*, 46(2), 77–93.
- 20 Li, A., Yin, Z., Wang, T., Fang, Q., & Hu, F. (2004). RASC863 - A Chinese Speech Corpus with Four Regional Accents. *ICSLT-o-COCOSDA*. New Delhi.
- 21 Grishina, E.A. (2005). Ustnaia rech v Natsionalnom korpuse russkogo yazyka [Spoken speech in the Russian National Corpus]. *Natsionalnyi korpus russkogo yazyka: 2003–2005 — Russian National Corpus: 2003–2005* (pp. 94–110). Moscow: Indrik [in Russian].
- 22 Natsionalnyi korpus russkogo yazyka (n.d.). [Russian National Corpus]. *ruscorpora.ru*. Retrieved from <http://ruscorpora.ru> [in Russian].
- 23 Korpus russkoi ustnoi rechi (n.d.). [Corpus of Russian Spoken Language]. *russpeech.spbu.ru*. Retrieved from <https://russpeech.spbu.ru/> [in Russian].
- 24 Crowdy, S. (1993). Spoken corpus design. *Literary and linguistic computing*, 8(4), 259–265.
- 25 Lam, P. (2009). The making of a BNC customised spoken corpus for comparative purposes. *Corpora*, 4(2), 167–188.
- 26 Aston, G., & Burnard, L. (1998). *The BNC handbook: exploring the British National Corpus with SARA*. Edinburgh: Edinburgh University Press.
- 27 Stenström, A.B., Andersen, G., & Hasund, I.K. (2002). *Trends in teenage talk: Corpus compilation, analysis and findings*. Amsterdam: John Benjamins Publishing Company.
- 28 Ludańska, Z. (2025). Technology propels speech-language research and theory. *Neuroscience & Biobehavioral Reviews*. Retrieved from <https://www.sciencedirect.com/science/article/pii/S014976342500199X>
- 29 Halabi, N. (2016). Modern standard Arabic phonetics for speech synthesis [Modern standard Arabic phonetics for speech synthesis]. *Doctor's thesis*. Southampton: University of Southampton.
- 30 Halpern, J. (2009). Word stress and vowel neutralization in modern standard Arabic. *2nd International Conference on Arabic Language Resources and Tools* (pp. 1–7). Cairo.
- 31 Zhubanov, A. (2015). Qazaq tilinin ulttyq korpussy zhane metabelgilenim maselesi [National corpus of the Kazakh language and the issue of metatagging]. *Tiltanyim – Linguistic Studies*, 1, 21–29 [in Kazakh].
- 32 Qazcorpus.kz (n.d.). [Qazcorpus.kz]. *qazcorpus.kz*. Retrieved from <https://qazcorpus.kz/> [in Kazakh].

Information about the authors

Kuderinova Kuralai Bimoldinovna — Doctor of Philological Sciences, professor, Deputy Director General for Research Affairs, Akhmet Baitursynuly Institute of Linguistics, Almaty, Republic of Kazakhstan. E-mail: kuralai.kuderinova@gmail.com; ORCID <https://orcid.org/0009-0002-4913-2840>

Boribayeva Gulnar Abdikasymovna — PhD, Institute of Linguistics named after Akhmet Baitursynuly, Almaty, Kazakhstan. E-mail: boribayeva.gulnara@gmail.com; ORCID <https://orcid.org/0000-0003-3225-776X>

Kozhakhmet Assem Mukhitovna — PhD student, Al-Farabi Kazakh National University, Almaty, Kazakhstan. E-mail: sdu.aseem@gmail.com; ORCID <https://orcid.org/0009-0003-1871-0166>