Ғылыми мақала
ӘОЖ 81.11, 81'33
https://doi.org/10.31489/2025Ph3/88-98

Мақаланың редакцияға түскен күні: 28. 03.2025 ж.
Қабылданған күні: 05.06.2025 ж.

Ж.М. Қоңыратбаева*

Л.Н. Гумилев атындағы Еуразия ұлттық университеті, Астана, Қазақстан
E-mail: zhanarkon@mail.ru

Қазақ тілінің академиялық корпусын әзірлеу: мақсаты, міндеті, маңызы (гуманитарлық ғылымдар бағытында)

Академиялық корпус — білім беру және ғылыми кеңістіктің негізгі бөлігінің бірі. Корпус әзірлеу ісі жоғары білім беру жүйесінің сапасын арттырып, ғылыми әлеуетін дамытуға бағытталған кешенді үдеріс ретінде танылады. Мақалада гуманитарлық ғылымдар бағытындағы қазақ тілі академиялық корпусын әзірлеу мәселесі қарастырылған: корпусты жасақтаудың мақсат-міндеттері мен маңызы талданған. Академиялық корпус әзірлеу ісінің маңыздылығы ұлттық тілдегі ғылыми мәтіндердің кешенді эмпирикалық базасын құру, сол арқылы ұлттық тілді ғылым тілі ретінде зерделеу саналады. Зерттеу жұмысының мақсаты — гуманитарлық ғылымдар бағытында жарық көрген қазақ тілі деректері корпусының мақсат-міндетін, оның қазақ ғылыми тілін дамытудағы атқарар рөлін, мазмұндық сипатын талдау. Корпус әзірлеу ісіне гуманитарлық ғылымдардың философия, тарих, діntану, археология және этнология, шығыстану, теология, түркітану, шығыстану, мұражай ісі, шетел филологиясы, аударма ісі, қазақ филологиясы сынды бағыттарының деректері алынады. Академиялық корпус әзірлеу ісі оған қойылатын талап-меже тұрғысынан сипатталады. Корпус әзірлеудің табиғи тілді өңдеу жүйелерін дамытудың маңызды тетігі екендігіне назар аударылады. Қазақ тілінің академиялық лексика, академиялық морфология, академиялық синтаксис ерекшеліктерін жинақтауға мүмкіндік беретіндігі сипатталады. Корпус жасағына енгізілетін қысқаша аңдатпа, тезис, баяндама, мақаладан бастап кең көлемді монография, диссертация, оқулық, оқу құралы кешендерінің ғылыми деректерді талдау үшін аса қажет болатындығы талдауға түседі. Тілдік деректерді жинау, редакциялау, цифрландыру кезеңіндегі метадеректер (метабелгіленім) базасының маңызы әлемдік тәжірибе мысалында қарастырылады. Метадеректер аясында қазақ тілін ғылым тілі ретінде зерттеуді дамытудың оңтайлылығы көрсетіледі. Академиялық корпус әзірлеудің ғылыми-практикалық негізін зерттеу барысында сипаттау, салыстыру, филологиялық сараптау, талдау секілді әдістер қолданылған. Әлемдік тәжірибелерге шолу жасауда сипаттау, салыстыру әдістері, тілдік деректерді мазмұндық, құрылымдық жағынан қарастыруда сипаттау, талдау әдістері пайдаланылған. Алынған нәтижелерді терминология, лексикография, когнитивтік тіл білімі, гендерлік лингвистика, аударматану және т.б. ғылым салаларында жаңаша зерттеу жүргізуге мүмкіндік беретіндігі талданады. Гуманитарлық ғылымдар бағытын зерттеу нысаны бойынша іштей топтастырып, тілдік деректерді жинауда атқарылар ғылыми-практикалық жұмыс барысында ескеруге, пысықтауға тиіс мәселелер қатары да сөз етіледі. Атап айтқанда, корпус әзірлеудің мақсат-міндеттерін айқындау кезеңінде, тілдік деректерді жинау, редакциялау, цифрландыру кезеңінде, сондай-ақ белгіленім (разметка) кезеңінде ғылыми-әдіснамалық тұрғыдан аса мән берілуі тиіс жайттардың орын алатындығы көрсетіледі. Академиялық корпустың мақсатына сай оған енгізілетін мәтіндер теңгерімділігін (сбалансированность) сақтаудың пәнаралық ерекшелігі айқындалған. Кейбір пәндердің ұлттық тілдегі үлес салмағынан төрі өзге (ағылшын, орыс) тілдердегі жарияланым жиілігі басымдығының теңгерімділік шартына әсер ету жайы талданған.

Кілт сөздер: академиялық корпус әзірлеу, корпус лингвистикасы, академиялық мәтін, гуманитарлық ғылымдар, қазақ тілі, ғылыми стиль, метадеректер.

Кіріспе

Заманауи тіл білімі ғылымының ең жас саласы ретінде корпус лингвистикасы соңғы уақытта қарқынды дамып отырғаны белгілі. Бұл мәселе ғылым мен технологияның цифрландырылуымен тікелей байланысты. Әлемдік ғылымда корпус түрлерінің сандық, типтік жиілігі де артып келеді.

Корпус лингвистикасының тарихында материалдарды қомақты шоғырландыруда ұлттық корпустар маңызды рөл атқарып отыр. Ағылшын тілінің Британ ұлттық корпусы (BNC, 100 млн.) (http://www.natcorp.ox.ac.uk/), американдық ұлттық корпус (ANC, 15 млн.) (https://anc.org), түрік тілінің ұлттық корпусы (TNC, 50 млн.) (https://www.tnc.org.tr), орыс тілінің ұлттық корпусы (RNC, 2 млрд.) (http://www.ruscorpora.ru) және т.б. корпустар — тіл ресурстарын бір платформа аясына жинай алған ірі құрылымдар.

*Хат-хабарларға арналған автор. E-mail: *zhanarkon@mail.ru*

Отандық ғылымда қазақ тілінің ұлттық корпустары (qazcorpus.kz және qazcorpora.kz), қазақ тілі корпусы (ISSAI) және Алматы қазақ тілі корпусы (АҚТК) қалыптасу үстінде. Қазақ тілінің ұлттық корпусында (А. Байтұрсынұлы атындағы Тіл білімі институтында әзірленіп жатқан) 70 миллион бірлік қарастырылған. Корпус өзге ұлттық корпустар секілді (ағылшын, орыс, түрік, т.б.) іштей бірнеше ішкорпустан тұрады: негізгі корпус, Ахмет Байтұрсынұлы ішкорпусы, ауызша ішкорпус, тарихи ішкорпус, параллель ішкорпус, мәдени-репрезентативті ішкорпус, жарнама ішкорпусы, диалектілік ішкорпус, мақал-мәтелдер ішкорпусы, фразеологизмдер ішкорпусы, ономастикалық ішкорпус, жазушы мәтіндері ішкорпусы, терминологиялық ішкорпус, оқу ішкорпусы және ағылшын тілділерге қазақ тілін үйренуге арналған Learner Corpus оқу ішкорпусы [1]. Қазақ тілінің ұлттық корпусының екінші бір нұсқасы Ш. Шаяхметов атындағы «Тіл-Қазына» ұлттық ғылыми-практикалық орталығында әзірленген. Корпус, негізінен, қазақ тілі публицистикалық мәтіндерінің кіші корпусынан тұрады. Онда «Егемен Қазақстан», «Ана тілі», «Түркістан», «Қазақ әдебиеті», «Заң» және басқа да мерзімді басылымдар материалдары қамтылған. Корпус базасына 24 миллиондай сөзқолданыс тіркелген [2]. Негізгі мақсаты — табиғи тіл ресурстарын жинақтап, тілді нормаландыру, жүйелендіру, тұтынушыларға ұтымды цифрландыру платформасын ұсыну. Келесі корпус түрі — Назарбаев университеті Ақылды жүйелер мен жасанды интеллект институты (ISSAI) әзірлеген қазақ тілі корпусы. Корпус базасында 2000-нан астам адамнан алынған 300 сағаттық ауызша дерек жиналған. Корпус жұмысы қазақша сөз тануға және оны синтездеу технологиясын әзірлеуге арналған [3]. Тағы бір корпус түрі — Алматы қазақ тілі корпусы. Онда 40 миллион бірлік қамтылған. Алматы қазақ тілі корпусында көркем әдебиет, ғылыми және көркемсөз стильдері мәтіндері жинақталған [4]. Атап көрсетілгендей, ұлттық корпустар өз ішінен түрлі ішкорпустарға жіктеліп, тұтас ұлттың бай қазынасын жинақтауға ықпалдасады.

Корпус лингвистикасының зерттеу әдіс-тәсілдері дамыған сайын корпус түрлерін жасақтау мәселесі де алғы шепке шыға бастады. Соның ішінде білім беру жүйесінде академиялық мәтіндер корпусын әзірлеу ісі өзекті мәселеге айналып отыр. Бұрын ғылыми-зерттеу жұмыстарын жүргізу мысал жинау, сол бойынша картотека жасау, оларды іріктеу, топтастыру, жіктеу, салыстыру, т.с.с. ұзақ уақыт бойғы қол жұмыстарын талап етсе, қазіргі таңда компьютердің көмегімен корпус түзу ісі уақыт үнемділігі мен жұмыс өнімділігін арттыруға мүмкіндік береді, яғни *картотекадан → корпусқа* ұстанымы жолға қойыла бастады. Корпустың бұл түрі академиялық корпус деп аталады. Корпус лингвистикасында бұндай корпус түрін білім беру немесе оқу корпусының ішінде қарастырады. Білім беру/оқу корпусының бір түрі тіл үйренушілерге оқу материалдарын ұсыну үшін әзірленетін болса, екіншісі — оқу үдерісінде білім алушылардың пайдалануына арналған ғылыми мәтіндер ресурсы [5; 134]. Академиялық корпус осылардың екіншісіне жатады.

Академиялық корпус әзірлеу, жалпы білім беру/оқу корпусын құру ісі XX ғасырдың соңы — XXI ғасырдың басында қарқын ала бастады. Алғашқы белгілі корпус қатарына 1990 жылдары Бельгияда әзірленген халықаралық ағылшын тілі корпусын (ICLE) жатқызуға болады. Онда ағылшын тілін бастапқы деңгейден жетік деңгейге дейін меңгерген білім алушылардың эссе жұмыстары қамтылған. Корпус жұмысын Лувен-ла-Неве университетінің профессоры С. Гранжер 1990 жылдары бастаған. 2009 жылы екінші нұсқасы (ICLEv2), ал 2020 жылы корпустың кеңейтілген үшінші нұсқасы (ICLEv3) әзірленген. Алғашқы нұсқасында 14 ұлттық тіл тұтынушыларының эсселері енгізілген. Соңғы нұсқасында 26 түрлі ішкорпус қамтылған. Корпус деректері арқылы ағылшын тілінің ерекшеліктерін тереңірек зерттеумен қатар, әртүрлі тілдердің өзара ықпалдастығын да (тілдік аударма) танып меңгеруге болады. Өзіндік веб-интерфейсі жасақталған аталмыш корпус ағылшын тіліндегі 5 миллион сөзден тұрады [6].

Танымал оқу (академиялық) корпусының ең ірісі ретінде ағылшын академиялық корпусын атауға болады. Британ академиялық жазбаша ағылшын тілі корпусы (BAWE) — Ұлыбританиядағы университеттерде жазылған ғылыми жұмыстардың кешенді жинағы. Онда өнер ғылымы, гуманитарлық ғылымдар, әлеуметтік ғылымдар, өмір жайлы ғылым және физика ғылымы салаларының мәтіндері мен оқу деңгейлері (бакалавриат және магистратура) ресурстары біркелкі тәртіппен қамтылған. Корпус 500-ден 5000 сөзге дейінгі тілді меңгеру деңгейі бойынша бағаланатын білім алушылардың (студенттер мен магистранттардың) жазбаша жұмыстарынан тұрады. Білім алушыларға арналған 3000-дай тапсырма ұсынылған. BAWE корпусында барлығы 6,9 миллион сөз тіркелген. 30 негізгі пән қамтылған [7].

Орыс тілінің білім беру/оқу корпусы (RLC) соңғы он-он бесжылдықта қалыптасты. Онда академиялық та, академиялық емес те деректер тіркелген. Корпус құрамында академиялық мәтін

тіліне арналған субкорпус (RULEC) жеке жұмыс істейді [8]. Ол екінші тілді үйренуге қызығушылық танытатындарға арналған. Корпустың көлемі шағын, қысқа абзацтардан бастап 8 беттен тұратын ғылыми жұмыстарға дейінгі 3800-дей жазбаша жұмыстарды қамтиды.

Отандық ғылымда академиялық корпус арнайы қалыптаспағанымен, Ұлттық корпустың ішкорпусы ретінде қарастырылады. Оқу ішкорпусын әзірлеуде қазақ тілін деңгейлік оқыту принципі ұстанылған. Тілдік деңгейлерге қарай оқу ішкорпусының базасы 1) Практикалық қазақ тілі; 2) Кәсіби қазақ тілі; 3) Теориялық қазақ тілі; 4) Іскери қазақ тілі; 5) Шешендік қазақ тілі секілді оқу пәндерімен жасақталған [1].

Қазақ тілінің оқу корпусы мәселесі корпус лингвистикасының ғылыми-әдіснамалық негізін салған ғалым А. Жұбановтың еңбектерінен бастау алады [9], [10]. Кейінгі жылдары оқу корпусын әзірлеу ісі жөнінде Н. Аитованың ғылыми мақаласында [11], сондай-ақ А. Жаңабекованың ұжымдық мақаласында арнайы сөз етіледі [5]. Бұл еңбектерде оқу корпусын әзірлеудің маңыздылығы мен қазақ ғылым тілін зерттеуді оңтайландырудағы жаңашылдығы сипатталады.

Қазақ тілінің академиялық корпусын әзірлеудің маңыздылығы әлемдік тәжірибеде көрсетілгендей, ұлттық тілдегі ғылыми мәтіндер кешенін жасап, сол арқылы қазақ тілін ғылым тілі ретінде әр қырынан зерттеуде жатыр.

Академиялық корпус әзірлеудің негізгі мақсаты — академиялық мәтін тілінің ерекшеліктерін зерттеудің әдіс-тәсілдерін оңтайландыру. Ол мақсатты орындаудың басты межесіне ғылыми стильде жазылған мәтін тілінің негізгі тетіктерін айқындау жатады. Демек, ғылыми ортада жұмсалатын академиялық мәтін тілінің лексикалық, синтаксистік, терминологиялық жиілігін, стилистикалық ерекшелігін анықтау міндеттері жүзеге асырылады. Мақалада гуманитарлық ғылымдар бағытында жарық көрген қазақ тілді деректердің кешені бойынша әзірленіп жатқан академиялық корпустың мақсат-міндеті, қазақ ғылыми тілін дамытудағы атқарар рөлі, мазмұндық сипаты қарастырылады.

Зерттеу әдістері мен материалдар

Мақаланың зерттеу материалы ретінде корпус лингвистикасы, оның ішінде академиялық корпус лингвистикасы саласы бойынша жарық көрген отандық (А. Жұбанов, А. Жаңабекова, Г. Мәдиева, С. Бектемірова, Ж. Күзембекова, Н. Аитова, т.б.) және шетелдік зерттеушілердің (С. Гранжер, В.П. Захаров, М. Заки, М. Матте, Э. Штумпф, т.б.) ғылыми-әдістемелік еңбектері қарастырылды. Талдаудың эмпирикалық материалына әзірленіп жатқан қазақ тілінің гуманитарлық бағыттағы академиялық корпусының кейбір деректері алынды. Зерттеу барысында Халықаралық ағылшын тілі корпусы (ICLE), Британ академиялық жазбаша ағылшын тілі корпусы (BAWE), орыс тілінің білім беру/оқу корпусы (RLC), қазақ тілінің ұлттық корпусы (ҚТҰК) базалары дереккөз ретінде қызмет атқарды.

Мақалада зерттеу мақсат-міндетіне байланысты ғылыми және эмпирикалық әдіс-тәсілдердің бірқатары пайдаланылды. Қазақ тілінің академиялық корпусын әзірлеудің ғылыми-практикалық негізіне шолу жасауда сипаттау, салыстыру, филологиялық сараптау, талдау әдістері пайдаланылды. Атап айтқанда, әлемдік тәжірибедегі академиялық корпус кешендеріне шолу жасауда сипаттау, салыстыру әдістері, жинақталған тілдік деректерді мазмұндық, құрылымдық жағынан қарастыруда сипаттау, талдау әдістерімен, корпус деректеріне қысқаша шолу жасауда филологиялық сараптау әдісімен жұмыс жасалды.

Нәтижелер мен талқылау

Академиялық корпус әзірлеу мәселесін қарастыруда «Академиялық мәтін дегеніміз не?», «Оның құрамына қай мәтін түрі жатады?» деген сұрақтарға тоқтала кеткен жөн. Академиялық мәтін — белгілі бір ғылым немесе білім беру саласындағы түрлі мәселелерді саралап, талдап, ғылыми болжамдар келтіру, оны дәлелдермен дәйектеу мақсатында әзірленетін жазбаша құрылым. «Академиялық мәтіннің мазмұнында артық сөзге, эмоцияға, идеологиялық немесе діни нанымдарға жол берілмейді. Онда бәрі оқырманға (ізденушіге, ғылыми қызметкерге) қажетті ақпаратты тез тауып, олардың шынайылығына көз жеткізуге бағынады» [12; 20]. Демек, ол өзінің логикалық нақтылығымен, қатаң стилімен, объективті баяндауымен ерекшеленеді.

Академиялық мәтін бастапқы және екінші реттік жанрларға бөлінеді. Бастапқы жанрға мақала, монография, монография тарауы, рецензия, оқулық, оқу құралы, баяндама немесе хабарлама, презентация, дәріс, диссертация жатады, ал екінші реттік түрі бастапқы жанрдың мазмұнын қысқарту арқылы жасалған мәтіндерді қамтиды. Мысалы, аннотация (мақала, баяндама немесе диссертация

туралы қысқаша ақпарат), автореферат (диссертацияның қысқартылған мазмұны), тезистер (мақала немесе сөз сөйлеудің қысқаша сипаттамасы), конспект (бір немесе бірнеше бастапқы мәтіннің негізгі тұжырымы) [13; 18].

Қазақ тілінің академиялық корпусы жасағында бастапқы және екінші реттік жанрлардың негізгі түрлері, атап айтқанда, *монография, диссертация, оқулық, оқу құралы, мақала, баяндама, тезис, аңдатпа* қамтылатын болады.

Қамтылатын мәтін түрлері мен оларды зерттеудің ғылыми және қолданбалы міндеттеріне байланысты корпустың біршама түрі қалыптасып үлгерді. Олар алға қойған мақсат-міндеттері мен жіктелім белгілеріне қарай бірнеше типке топтастырылып беріледі [14; 12] (1-кесте).

1 - к е с т е

Тілдік корпустардың жіктелімі (В.П. Захаров бойынша):

Белгісі:	Корпус түрлері:
Деректер түрі	Жазбаша, ауызша, аралас
Мәтіндер тілі	Орысша, ағылшынша және т.б.
Параллель	Біртілді, екітілді, көптілді
Әдеби Мамандандырылған	Әдеби, диалектілік, ауызекі, Терминологиялық, аралас
Жанры	Әдеби, фольклорлық, драматургиялық, аралас
Қолжетімділігі	Қолжетімді, коммерциялық, жабық
Мақсаты	Зерттеу, иллюстративті
Динамикасы	Динамикалық, статикалық
Белгіленімі	Белгіленген, белгіленбеген
Белгіленім сипаты	Морфологиялық, синтаксистік, семантикалық, просодикалық және т.б.
Мәтіндердің көлемі	Толық мәтін, мәтін үзіндісі
Хронологиялық белгісі	Синхронды, диахронды
Жалпылығы	Жалпы, бір жазушынікі
Құрылымы	Орталық және мұрағаттық, орталық және перифериялық

Тілдік корпустардың көрсетілген осы жіктелімі негізінде қазақ тілінің академиялық корпусының құрылымын былайша жіктеуге болады: корпусқа тек қазақ тіліндегі жазбаша деректер енгізіледі; яғни біртілді, ғылыми мәтіндер кешені құрылады; корпус ашық түрде жұмыс істейді, кез келген тұтынушыға қолжетімді; қамтылған мәтіндер сан-салалы бағытта зерттеуге арналады; тілдік деректер белгіленген (разметка) сипатта жүзеге асады; белгіленімі жағынан морфологиялық, синтаксистік, семантикалық және т.б. қыры талданады; корпус базасында толық мазмұнды мәтіндер жинақталады, бұл өз кезегінде академиялық мәтін тілін әртүрлі ғылыми-әдістемелік мақсатта пайдалануға мүмкіндік береді (2-кесте):

2 - к е с т е

Қазақ тілі академиялық корпусының сипаты

Белгісі:	Корпус түрі:
Деректер түрі	Жазбаша
Мәтіндер тілі	Қазақша
Параллель	Біртілді
Мәтін стилі	Ғылыми
Қолжетімділігі	Қолжетімді
Мақсаты	Зерттеу
Белгіленімі	Белгіленген
Белгіленім сипаты	Морфологиялық, синтаксистік, семантикалық
Мәтіндердің көлемі	Толық мәтін

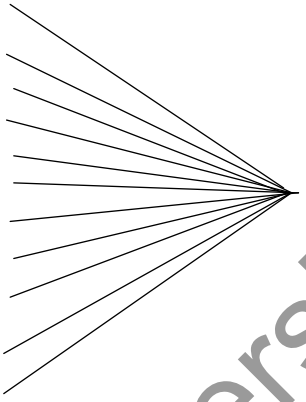
Гуманитарлық ғылымдар бағытында қазақ тілінің академиялық корпусын әзірлеу жұмысы да өзге корпус түрлеріне қойылатын талап-меже тұрғысынан қарастырылады [15]. Оның басты сипаты мынадай кезеңдерден тұрады:

- Корпус әзірлеудің мақсат-міндеттерін айқындау кезеңі;

- Тілдік деректерді жинау, редакциялау, цифрландыру кезеңі;
- Белгіленім (разметка) кезеңі;
- Корпус менеджерін таңдау кезеңі.

Академиялық корпуста ЖОО-ында оқытылатын гуманитарлық ғылымдар бағыты бойынша деректер қамтылу көзделеді. Корпус базасы кемінде 5 000 000 сөзден тұратын болады, оның ішінде 50 000 сөз аннотацияланады. Қазіргі кезде гуманитарлық бағыттың әр пәні бойынша 2000 тілдік деректен жинақталды [15]. Алдағы уақытта төмендегі он екі ғылыми бағыттың мәтіндері (қысқаша аңдатпадан бастап кең көлемді монографияларға дейін) жинақталып, аннотациясы берілетін болады:

5В020100 –	Философия
5В020300 –	Тарих
5В020600 –	Дінтану
5В020800 –	Археология және этнология
5В020900 –	Шығыстану
5В021100 –	Теология
5В021200 –	Түркітану
5В021500 –	Исламтану
5В041900 –	Мұражай ісі және ескерткіштерді қорғау
5В021000 –	Шетел филологиясы
5В020700 –	Аударма ісі
5В020500 –	Филология



аңдатпа, тезис, мақала, баяндама, оқулық, оқу құралы, диссертация, монография

Академиялық корпус әзірлеудің маңызды кезеңінің бірі — гуманитарлық ғылымдар бағыттарын зерттеу нысаны бойынша іштей топтастырып, тілдік деректерді жинау, оларды word және pdf құжаттарда сақтау. «Бұл кезеңде мәтіндердің қай жанр, қандай стильде екендігі, сонымен қатар хронологиялық кезеңі қарастырылуы қажет» [16; 99]. Сонымен қатар аталған бастапқы кезеңде тілдік деректерді іздестіру, жүйелеу жұмысымен қатар бірқатар техникалық мәселелерді пысықтап алу маңызды. Академиялық мәтіндерде кездесетін графикалық кескіндемелер, түрлі кестелер, яғни тілдік деректерге жатпайтын өзге мәліметтерді корпуска енгізу/енгізбеуге байланысты сұрақ туындауы мүмкін. Мысалы, мәтін бойында тілдік деректің негізгі мазмұнын ашу мақсатында ұсынылған графикалық кескіндемені, суретті (скриншот) не кестені сол күйінде қалдыру корпус құрудың IT технологиясын пайдалану тетігімен сәйкес келе бермейді. Сол себепті көп жағдайда корпусқа тек мәтіндік деректер ғана алынады.

Корпус әзірлеу ісінің негізгі бөлігінің бірі — метадеректер (метабелгіленімдер) базасы. Метадеректер құрамына, негізінен, *мәтін типі, жанры, авторы, тақырыбы, мәтіннің жариялану мерзімі, баспасы* және т.б. мәліметтер жатады. Бұл жөнінде А. Жұбанов корпус құрастыруда метабелгіленімдер (оқулық атауы, авторы, жылы, шыққан орны, стилі, жанры, көлемі және т.б.) мен грамматикалық белгіленімдерді мәтінге енгізудің әдіс-тәсілдерін, түрлерін, компьютерлік бағдарламаларын, т.б. зерттеудің маңызы зор екендігін атап көрсетеді [10; 81]. Демек, корпус метадеректері (метабелгіленімдері) мазмұнында ғылыми-әдіснамалық тұрғыдан жан-жақты зерттеу жүргізуге мүмкіндік туады.

Академиялық корпуста қамтылған метадеректерді түрлі бағытта пайдаланудың тиімділігі зерттеулерде қарастырылады. Мәселен, араб тілі академиялық тезистеріндегі метадискурс пен риторикалық тәсілдердің қолданылуын зерттеген арнайы еңбекті атауға болады. Зерттеу авторы М. Заки араб тілінде жазылған ғылыми мақалалар мен диссертациялық жұмыстар бойынша 400 тезистен тұратын корпус құрады [17]. Метадискурс қызметін қолданудың өзгерістерін екі бағытта алып қарайды: 1) тезистің түрлеріне (диссертациялар мен зерттеу мақалалары) қарай және 2) авторлардың жынысына (ер, әйел немесе аралас) қарай. Тезистердің бойында түрлі маркерлер (өтпелі, фрейм), дәлел, түсіндірмелердің кездесетіндігі, оның ішінде, әсіресе, маркерлер мен фреймдердің басым қолданысқа ие екендігі сарапталып беріледі. Зерттеу нәтижесі ғылыми ортаны, оқырманды қызықтыру үшін араб ғалымдарының қолданған тілдік ерекшеліктерін тануға мүмкіндік береді.

Корпус метадеректерінде мәтін типі, жанры, авторы(лары), тақырыбы, баспасы, т.б. мәліметтермен қатар, алынған мәтіннің не туралы екендігін баяндайтын қысқаша аннотация берілуі тиіс.

Қазақ тілінің академиялық корпусының метадеректерінде он алты түрлі мәлімет қамтылады, олар: мәтін деңгейі (кімдерге арналғандығы, яғни бакалавриат, магистратура, докторантура, профессор-оқытушылар құрамы); пән саласы; жанры; пәні (мамандығы); шифры; академиялық мәтін тақырыбы; баспасы; авторлардың аты-жөні; авторлар коды (метадеректе тіркелген); жарияланым мерзімі; деңгейі; мәтіннің сөз саны; мәтіннің бет саны; авторлардың саны; авторлардың жынысы (ер, әйел, аралас тип) және корпусқа тіркелген күні (1-сурет).

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
1	деңгей (Б/М/Д/ПО)	пән саласы	жанр/регистр	пән	шифр, мамандық	тақырып	баспа/ұйым атауы	автор(лар)дың аты-жөні	ID автор(лар)дың код ы	жарияланған мерзімі	деңгей	сөз саны	бет саны	авторлар саны	жынысы	жиналған мерзімі
	ПО		баяндама	Аударма ісі	Аударма ісі	ІС-ТӘЖІРИБЕСІ	Хрестоматия. Құрастырушы Г. Қ. Қазыбек	А. Курелла	10343	Алматы «Қазақ	ПО	9055	32	1	ерлек	04.01.2025
647	ПО	ТӨ (тілдер және әдебиет)	баяндама	Аударма ісі	Аударма ісі	АБАЙ ШЫҒАРМАЛАРЫНЫҢ ОРЫС ТІЛІНЕ АУДАРЫЛУЫ	ӘЛ-ФАРАБИ атындағы ҚАЗАҚ ҰЛТТЫҚ УНИВЕРСИТЕТІ. АУДАРМА ТАҒЫЛЫМЫ. Хрестоматия. Құрастырушы Г. Қ. Қазыбек	А. Көпбасарова	10344	Алматы «Қазақ	ПО	8126	32	1	әйел	04.01.2025
648	ПО	ТӨ (тілдер және әдебиет)	баяндама	Аударма ісі	Аударма ісі	ТӘРЖІМЕШІ ТАРАЗЫСЫ	ӘЛ-ФАРАБИ атындағы ҚАЗАҚ ҰЛТТЫҚ УНИВЕРСИТЕТІ. АУДАРМА ТАҒЫЛЫМЫ. Хрестоматия. Құрастырушы Г. Қ. Қазыбек	В. Ларбо	10345	Алматы «Қазақ	ПО	794	3	1	ерлек	04.01.2025
649	ПО	ТӨ (тілдер және әдебиет)	баяндама	Аударма ісі	Аударма ісі	КЛЕР КЛЕРМОНТ – ҚАЗАҚ ТІЛІНДЕ НЕМЕСЕ «ВЛАДИМИР МЕН ЗАРА» ПОЭМАСЫ	ӘЛ-ФАРАБИ атындағы ҚАЗАҚ ҰЛТТЫҚ УНИВЕРСИТЕТІ. АУДАРМА ТАҒЫЛЫМЫ. Хрестоматия. Құрастырушы Г. Қ. Қазыбек	М. Неталиев	10346	Алматы «Қазақ	ПО	2262	9	1	ерлек	04.01.2025
650	ПО	ТӨ (тілдер және әдебиет)	баяндама	Аударма ісі	Аударма ісі	МЕН ТҮРГЕНЕВТІ АҒЫЛШЫН ТІЛІНЕ ҚАЛАЙ ТӘРЖІМЕЛЕДІМ	ӘЛ-ФАРАБИ атындағы ҚАЗАҚ ҰЛТТЫҚ УНИВЕРСИТЕТІ. АУДАРМА ТАҒЫЛЫМЫ. Хрестоматия. Құрастырушы Г. Қ. Қазыбек	А. Паймен	10347	Алматы «Қазақ	ПО	4160	15	1	ерлек	04.01.2025
651	ПО	ТӨ (тілдер және әдебиет)	баяндама	Аударма ісі	Аударма ісі	СИНХРОНДЫ АУДАРМАНЫҢ ҚОЛДАНЫЛУЫ	ӘЛ-ФАРАБИ атындағы ҚАЗАҚ ҰЛТТЫҚ УНИВЕРСИТЕТІ. АУДАРМА ТАҒЫЛЫМЫ. Хрестоматия. Құрастырушы Г. Қ. Қазыбек	Ә. Тарақов	10348	Алматы «Қазақ	ПО	1069	4	1	ерлек	04.01.2025
652	ПО	ТӨ (тілдер және әдебиет)	баяндама	Аударма ісі	Аударма ісі	АУЫЗША АУДАРМА ТАРИХЫ	ӘЛ-ФАРАБИ атындағы ҚАЗАҚ ҰЛТТЫҚ УНИВЕРСИТЕТІ. АУДАРМА ТАҒЫЛЫМЫ. Хрестоматия. Құрастырушы Г. Қ. Қазыбек	Ә. Тарақов	10349	Алматы «Қазақ	ПО	1079	5	1	ерлек	04.01.2025
653	ПО	ТӨ (тілдер және әдебиет)	баяндама	Аударма ісі	Аударма ісі	И.А. КРЫЛОВ МЫСАЛЫНДАҒЫ ДИАЛОГТАРДЫҢ ҚАЗАҚ КӨРКЕМ АУДАРМАСЫНДА БЕРІЛУІ	ӘЛ-ФАРАБИ атындағы ҚАЗАҚ ҰЛТТЫҚ УНИВЕРСИТЕТІ. АУДАРМА ТАҒЫЛЫМЫ. Хрестоматия. Құрастырушы Г. Қ. Қазыбек	Қ.С. Телгожаева	10350	Алматы «Қазақ	ПО	1714	8	1	әйел	04.01.2025
654	ПО	ТӨ (тілдер және әдебиет)	баяндама	Аударма ісі	Аударма ісі	АУДАРМА ЛИНГВИСТИКАЛЫҚ ЗЕРТТЕУ НЫСАНЫ РЕТІНДЕ ЖӘНЕ ОНЫҢ ЛИНГВОМӘДЕНИ МӘСЕЛЕЛЕРІ	ӘЛ-ФАРАБИ атындағы ҚАЗАҚ ҰЛТТЫҚ УНИВЕРСИТЕТІ. АУДАРМА ТАҒЫЛЫМЫ. Хрестоматия. Құрастырушы Г. Қ. Қазыбек	Г. Тіхенбай	10351	Алматы «Қазақ	ПО	1259	6	1	ерлек	04.01.2025
655	М/Б	ТӨ (тілдер және әдебиет)	баяндама	Аударма ісі	Аударма ісі	ИРОНИЯ ЖӘНЕ ОНЫ АУДАРУ ТӘСІЛДЕРІ	Абылай хан атындағы Қазақ халықаралық қатынастар және әлем тілдері университеті	Алимбекова Л.У., Бөлекбай А.М.	10352	2022	М/Б	1227	3	2	аралас	04.01.2025
656	М/ПО	ТӨ (тілдер және әдебиет)	баяндама	Аударма ісі	Аударма ісі	ФЗНТЕЗИ ЖАНРДАҒЫ КӨРКЕМ ШЫҒАРМАНЫҢ БЕЙНЕЛІ ТУЫНДЫЛАРЫН АУДАРУДАҒЫ ЛИНГВОМӘДЕНИ ЕРЕКШЕЛІКТЕРІ	Абылай хан атындағы Қазақ халықаралық қатынастар және әлем тілдері университеті	Усали А.К, Курмамбаева А.М.	10353	2022	М/ПО	1369	3	2	аралас	04.01.2025

1-сурет. Қазақ тілінің академиялық корпусының метадеректері базасының үлгісі
(гуманитарлық ғылымдар бағыты бойынша)

Академиялық корпус метадеректерінің базасы белгілі бір автордың (ғалымның/зерттеушінің) ғылыми дискурсын корпустық зерттеуде де нақты мәліметтер қоры бола алады. Корпусқа тіркелген автордың ғылыми еңбектерін жинақтау, сұрыптау, талдау арқылы ғылыми-әдістемелік тұрғыдан зерттеу мүмкіндігі туады. Корпус лингвистикасында бұндай тәжірибе бар. Қазақ тілінің ұлттық корпусының Ахмет Байтұрсынұлы мәтіндерінің ішкорпусын [18], сондай-ақ Ахмет Байтұрсынұлының қазақша-орысша параллель корпусын [19] осындай ізденістердің оң нәтижесі деуге болады. А. Байтұрсынұлы шығармаларын («Қырық мысал», «Маса» жинақтарын) корпустық зерттеу нәтижесі ғылыми зерттеулерде баяндалады [20]. Авторлардың пайымдауынша, мәтіндердің белгілі бір тарихи кезеңдер мен қазіргі жай-күйіндегі сөздердің жиілігін автоматты сараптау біріншіден, белгілі кезеңдердегі сөзқолданыс жиілігінің сандық ақпаратын білуге мүмкіндік береді; екіншіден, тілдік қолданыстардың белгілі бір уақыт аралығындағы теңгерімділігін, қоғамның түрлі саласында өмір сүру қабілетін, т.с.с. сапалық, сандық өзгерістерді анықтауға, тәпсірлеуге және болжауға септігі тиеді [20; 14].

Гуманитарлық ғылымдар бағытында әзірленіп жатқан академиялық корпус метадеректері негізінде белгілі бір автордың (ғалымның) ғылыми дискурсының ішкорпусын, терминдік-глоссарийлік корпусын, академиялық жазылымының жиілік цифрлық базасын және т.б. бағыттарда автоматтандырылған өнімдер әзірлеуге мүмкіндік алуға болады.

Сонымен, академиялық корпустың метадеректері аясында:

- академиялық мәтіндердің тілі мен терминдер базасын зерттей отырып, қазақ ғылым тіліне еніп жатқан жаңа терминдерді, оларды біріздендіру жайын анықтауға;

- гуманитарлық ғылымдар бойынша жарық көрген академиялық мәтін тақырыптарының өзекті бағытын айқындауға;

- әртүрлі ғылым бағыттары мәтіндерін салыстыра қарастырып, пәнаралық зерттеу байланыстарын, олардың мәнін, себебін, нәтижесін, маңыздылығын ашуға;

- академиялық мәтін авторларының үлес салмағын бақылай отырып, қазақ гуманитарлық ғылымдарындағы «ер» және «әйел» көпцептісінің үлес салмағын гендерлік лингвистика тұрғысынан зерттеуге және т.с.с. өзекті мәселелерді ғылымның жаңа әдіс-тәсілдері арқылы зерттеуге мүмкіндік

туады. Демек, метадеректер (метабелгіленімдер) академиялық мәтіннің ішкі құрылымдануы бойынша да, сондай-ақ ғылыми зерттеулерге әсер ететін сыртқы ықпалды да бірдей қарастыруға көмектеседі.

Корпус деректерінің теориялық еңбектерді эмпирикалық материалдармен толықтыруда мүмкіндігі зор. Жалпы алғанда, корпустар тілдік зерттеулер барысында пайдалануға болатын үш типті дерек бере алады: *эмпирикалық қолдау, жиілігі бойынша ақпарат, экстралингвистикалық ақпарат (метаақпарат)* [21; 85]. Корпус базасына енгізілген мысалдар легі зерттеушілердің болжамдарына, белгілі бір түйіндеріне эмпирикалық қолдау бола алады. Сонымен қатар корпус деректері сандық зерттеу жүргізу үшін сөздердің, фразалар мен сөз тіркестерінің қолдану жиілігі туралы да ақпарат бере алады. Ал метаақпарат әртүрлі мәтін типтері мен әртүрлі сөйлеушілер топтарын салыстыруға мүмкіндік береді [21; 86].

Академиялық корпус әзірлеу нәтижесінің тағы бір маңыздылығы — корпус құрамына енгізілген тілдік деректер бойынша түрлі тақырыпта зерттеу жүргізуге мүмкіндік алу. Ол зерттеу жұмысы тілдің фонетикалық қабатынан грамматикалық құрылымына дейінгі кез келген нысан болуы мүмкін. Бұл мәселе соңғы уақытта әртүрлі Ұлттық корпус деректері аясында жүргізіле бастады. Сондай ізденістің бірі ретінде М. Матте мен Э. Штумпфтың португал тіліндегі ғылыми мақалаларда кездесетін етістіктердің қолданылуын қарастырған зерттеу мақалаларын атап өтуге болады. Авторлар Португал тілі корпусынан «autor» сөзін алып, одан соң етістіктерді жинақтайды. Жинақтау кезеңінен соң корпустағы жаратылыстану және гуманитарлық ғылымдарда жиі кездесетін етістіктерді құрылымдық және семантикалық қолданылуы жағынан саралайды. Нәтижесінде екі ғылым саласында жұмсалатын етістіктердің арасында айтарлықтай ұқсастықтың бар екені анықталған. Сонымен қатар авторлар ғылыми мақалалардағы дереккөздерде осы шақ пен өткен шақ етістіктерінің бірлікте қолданылатынын дәйектеуге қол жеткізгенін көрсетеді [22; 68-69].

Етістіктердің академиялық мәтін тіліндегі өзіндік сипатын, қызметін ұлттық корпустың деректері бойынша ашуға бағытталған зерттеу жұмыстары отандық ғылымда да бар. А.Е. Бижкенова мен Р. Кенжебекованың ғылыми мақалаларында әр типтес тілдердегі (қазақ, орыс, ағылшын тілдері) қимыл етістіктерінің қолданылу жиілігі салыстырыла қарастырылады [23]. Нәтижесінде авторлар үш тілдегі қимыл етістіктерінің тілдегі қозғалыс динамикасының әртүрлі екендігін айқындауға мүмкіндік алған. Орыс тіліндегі қимыл етістіктерінің қазақ және ағылшын тілдерімен салыстырғанда айырмашылығы бар екені: қазақ және ағылшын тілдерінде префикстердің жоқ екендігі деректер негізінде талдауға түскен. Авторлар корпустармен жұмыс істеудің тиімділігін: олар алынған нәтижелердің шынайылығын, олардың қаншалықты шынайы екендігін тексеруге мүмкіндік бере алатындығын атап көрсетеді [23; 92-94].

Академиялық корпус мәтіндері өзге корпустардағы секілді белгілі бір жүйе бойынша іріктеліп алынады. Демек, корпустың мақсатына сай оған енгізілетін мәтіндер теңгерімділігі (сбалансированность) сақталуы қажет. Бұл — корпус жасаудағы негізгі шарттың бірі. Бұл жерде туындайтын мәселе бар. Академиялық корпусқа мәтін жинақтауда «Қазақ тілі мен әдебиеті», «Аударма ісі», «Шетел тілдері: екі шет тілі» мамандықтары бойынша мәтін теңгерімділігі біраз қиындық туындатуы мүмкін. Себебі аталған мамандықтар бойынша жарияланған қазақ тілді мәтіндер саны біркелкі емес. Егер «Қазақ тілі мен әдебиеті» мамандығы мәтіндері, негізінен алғанда, қазақ тілінде жарияланса, ал «Аударма ісі», «Шетел тілдері: екі шет тілі» мамандықтары жарияланымдарының ұлттық тілдегі үлес салмағы анағұрлым аз, көбінесе ағылшын, орыс тілдерінде жарық көрген. Сондықтан теңгерімділік шартына жауап беру үшін «Аударма ісі», «Шетел тілдері: екі шет тілі» мамандықтары мәтіндерінің шегін ескеру керек. Бұл жайт өз кезегінде корпустың көлемділігіне әсер етуі мүмкін. «Әртүрлі жанрдағы мәтіндердің бірінің көлемі аз, бірінің көлемі үлкен болуы белгілі бір дәрежеде сандық, сапалық мәліметтер алуға кедергі келтіреді» [10; 52]. Сондықтан академиялық корпус әзірлеу ісінде ғылыми-әдістемелік тұрғыдан пысықталуы тиіс мәселелер қатары орын алады.

Қорытынды

Сонымен, қазақ тілі академиялық корпусын әзірлеу ісінің отандық ғылым үшін берер нәтижесін мыналар деп тануға болады:

- біріншіден, қысқаша аңдатпадан кең көлемді монографиялық еңбектерге дейін жинақталатын академиялық корпус базасы мәтіндер кешені ғана емес, осы бағытта ғылыми жұмыспен айналысатын зерттеушілер мен ғалымдар үшін бай репрезентативті жүйе ретінде қызмет ете алады;

- екіншіден, қазақ тілінің академиялық жазылым тілі лексикасын іріктеп, қазақ тілін ғылым тілі ретінде бекітуге атсалысады;

- үшіншіден, гуманитарлық ғылымдар бағытындағы академиялық мәтіндердің тілін өзара салыстыра зерттей отырып, ғылымның әртүрлі бағытындағы ортақ терминдер мен әдіс-тәсілдерді, олардың қазақ ғылым тілінің дамуына қосар үлесін айқындауға мүмкіндік береді. Әсіресе заманауи ғылым кеңістігіндегі антропоөзектік парадигма шеңберіндегі зерттеулердің ортақтығы мен олардың маңызын ашуға, келешектегі бағыт-бағдарын анықтауға оң ықпалдасады;

- төртіншіден, корпус метадеректері (метабелгіленімі) арқылы гуманитарлық ғылымдар бағытының автоматтандырылған жүйесін жетілдіруге мүмкіндік береді және т.с.с. өзекті мәселелерді қарастырудың ғылыми-әдіснамалық амал-жолын жеңілдетуге үлес қосады.

Қорыта келгенде, академиялық корпус әзірлеу ісі ғылымның соңғы технологиялық жетістіктерін барынша пайдалануды талап ететін күрделі де кешенді жұмыс саналады. Мақалада гуманитарлық ғылымдар бағытындағы қазақ тілінде жарық көрген академиялық мәтіндер корпусын құрудың мақсат-міндеттері мен маңыздылығы талдауға түсті. Алдағы уақытта корпусқа енгізілетін эмпирикалық материалдарды арттырып, оларды редакциялап, цифрландырып, аннотацияларын әзірлеп, қолжетімді түрде ғылыми-көпшілік ортаға ұсыну міндеті тұр.

Бұл зерттеу жұмысы Қазақстан Республикасы Ғылым және жоғары білім министрлігі тарапынан гранттық қаржыландырылған ЖТН AP23488585 «Цифрлық гуманитарлық ғылымдар: Академиялық қазақ тілінің корпусын әзірлеу» ғылыми жобасы аясында әзірленді.

Әдебиеттер тізімі

1 Қазақ тілінің ұлттық корпусы. — [Электрондық ресурс]. — Қолжетімділігі: https://qazcorpus.kz

2 Қазақ тілі ұлттық корпусы. — [Электрондық ресурс]. — Қолжетімділігі: https://qazcorpora.kz/

3 Қазақ тілі корпусы. — [Электрондық ресурс]. — Қолжетімділігі: http://test-test-issai.nu.edu.kz/kz-speech-corpus/

4 Алматы қазақ тілі корпусы. — [Электрондық ресурс]. — Қолжетімділігі: http://webcorpora.net/KazakhCorpus/search/?interface_language=ru

5 Жаңабекова А.Ә. Қазақ тілінің оқу корпусын әзірлеу мәселелері / А.Ә. Жаңабекова, Г.Б. Тлегенова, А.Ж. Мұқатаева, М.С. Жолшаева // Тілтаным. — 2024. — № 1 (93). — Б. 133–143. doi.org/10.55491/2411-6076-2024-1-133-143

6 Granger S. International Corpus of Learner English. — [Electronic resource]. — Access mode: https://www.researchgate.net/publication/353637356

7 British Academic Written English Corpus. — [Electronic resource]. — Access mode: https://www.sketchengine.eu/british-academic-written-english-corpus

8 Учебный корпус русского языка. — [Электронный ресурс]. — Режим доступа: http://www.web-corpora.net/RLC/rulec

9 Жұбанов А.Қ. Компьютерлік лингвистикаға кіріспе / А.Қ. Жұбанов. — Алматы, 2013. — 204 б.

10 Жұбанов А.Қ. Корпустық лингвистика / А.Қ. Жұбанов, А.Ә. Жаңабекова. — Алматы: Қазақ тілі, 2017. — 336 б.

11 Аитова Н.Н. Оқу корпусын жасау және мектеп терминологиясын зерттеуді оңтайландыру / Н.Н. Аитова // «Мектеп оқулықтары: пән терминдерін жетілдіру және біріздендіру» атты Республикалық ғылыми-әдістемелік конференция материалдары**.** — Астана, 2023. — Б. 8–11.

12 Короткина И.Б. Академическое письмо: процесс, продукт и практика / И.Б. Короткина. — Москва: Юрайт, 2024. — 349 с.

13 Ярская-Смирнова Е. Создание академического текста / Е. Ярская-Смирнова. — Москва, 2013. — 156 с.

14 Захаров В.П. Корпусная лингвистика / В.П. Захаров. — Санкт-Петербург: СПб., 2005. — 48 с.

15 Қазақ тілінің академиялық корпусы. — [Электрондық ресурс]. — Қолжетімділігі: https://academickazakhcorpus.kz/

16 Мәдиева Г.Б. Лингвистикалық зерттеулердегі компьютерлік технологиялар / Г.Б. Мәдиева, С.Б. Бектемірова, Ж.Ж. Күзембекова. — Алматы: Қазақ университеті, 2014. — 119 б.

17 Zaki M. The metadiscourse of Arabic academic abstracts: A corpus-based study / M. Zaki // Research in Corpus Linguistics. — Spain, 2022. — 10/2. — P. 113–146. //https://DOI 10.32714/ricl.10.02.06

18 А. Байтұрсынұлы мәтіндерінің ішкорпусы. — [Электрондық ресурс]. — Қолжетімділігі: https://qazcorpus.kz/findahmeti/

19 А. Байтұрсынұлының қазақша-орысша параллель корпусы. — [Электрондық ресурс]. — Қолжетімділігі: http://baitursynuly-corp.kz/

20 Аитова Н.Н. А. Байтұрсынұлы шығармаларын корпустық зерттеу («Қырық мысал», «Маса» жинақтары негізінде) / Н.Н. Аитова, Г.Ж. Байшукурова, А.Б. Иргебаева // Қарағанды университетінің хабаршысы. Филология сериясы. — 2024. — 29-т., 3(115)-шығ. — Б. 14–23. https://doi.org/10.31489/2024Ph3/14-23

21 Пірманова К.Қ. Ұлттық корпустарға негізделген лингвистикалық зерттеулер жүргізу (қазақ, орыс, ағылшын тілі материалдары негізінде) / К.Қ. Пірманова, А.Ә. Жаңабекова, А. Барменқұлова // Әл-Фараби атындағы ҚазҰУ хабаршысы. Филология сериясы. — 2022. — № 3(187). — Б. 83–93. https://doi.org/10.26577/EJPh.2022.v186.i2.01

22 Matte M. A corpus-based study of reporting verbs in academic Portuguese / M. Matte, E. Stumpf // Research in Corpus Linguistics. — Spain, 2022. — 10/2. — P. 46–69. //https://DOI 10.32714/ricl.10.02.04

23 Бижкенова А.Е. Семантика глаголов движения в русском, казахском и английском вариантах (с опорой на корпусные и словарные данные) / А.Е. Бижкенова, Р. Кенжебекова // Вестник Карагандинского университета. Серия филология. — 2024. — Т. 24, № 1(113). — С. 80–95. https://doi.org/10.31489/2024Ph1/80-95

Ж.М. Коныратбаева

Разработка академического корпуса казахского языка: цель, задачи, значимость (в направлении гуманитарных наук)

Академический корпус — одна из ключевых составляющих образовательного и научного пространства. Его разработка рассматривается как комплексный процесс, направленный на повышение качества системы высшего образования и развитие научного потенциала. В статье рассматривается проблема создания академического корпуса казахского языка в области гуманитарных наук. Анализируются цели, задачи и значение формирования такого корпуса. Актуальность проекта заключается в создании комплексной эмпирической базы научных текстов на государственном языке, что способствует изучению казахского языка как языка науки. Цель исследования — анализ функций и задач корпуса гуманитарных текстов на казахском языке, его роли в развитии казахского научного языка, а также характеристик корпуса. Для его создания были использованы тексты по направлениям: философия, история, религиоведение, археология и этнология, востоковедение, теология, тюркология, музеология, иностранная филология, переводоведение и казахская филология. В статье описывается процесс формирования корпуса с учётом предъявляемых к нему требований и критериев. Отдельное внимание уделено значению корпуса как инструмента для развития технологий обработки естественного языка. Корпус позволяет выявить особенности академической лексики, морфологии и синтаксиса казахского языка. Обоснована необходимость включения в корпус различных видов научных источников: аннотаций, тезисов, докладов, статей, а также монографий, диссертаций, учебников и учебных пособий — для всестороннего анализа научных данных. Особое значение придаётся метаданных (метаразметке), важным на всех этапах — от сбора до цифровизации материалов. Метаданные рассматриваются в контексте международного опыта, подчёркивается их роль в дальнейшем изучении казахского языка как языка научной коммуникации. В методологической части применялись методы описания, сравнения, филологической экспертизы и анализа. При изучении международного опыта использовались методы описания и сопоставления, а при анализе языковых данных с содержательной и структурной стороны — методы описания и анализа. Результаты исследования открывают новые перспективы для научных изысканий в таких областях, как терминология, лексикография, когнитивная и гендерная лингвистика, переводоведение и другие смежные дисциплины. Отмечается, что на всех этапах — от постановки задач до разметки — необходимо учитывать научно-методологические аспекты. Также поднимаются вопросы, требующие доработки в рамках научно-практической деятельности, с учётом внутренней классификации гуманитарных наук по объекту исследования. Выявлена междисциплинарная специфика соблюдения сбалансированности текстов, включаемых в корпус, в зависимости от его назначения. Анализируется влияние преобладания публикаций на английском и русском языках по ряду дисциплин на соблюдение принципа сбалансированности корпуса по отношению к казахскому языку.

Ключевые слова: разработка академического корпуса, корпусная лингвистика, академический текст, гуманитарные науки, казахский язык, научный стиль, метаданные.

Zh.M. Konyratbayeva

Development of the academic corpus of the Kazakh language: goal, task, significance (in the direction of humanities)

The academic corpus is one of the key components of the educational and scientific sphere. Its development is recognized as a comprehensive process aimed at improving the quality of higher education and enhancing scientific potential. The article addresses the issue of creating an academic corpus of the Kazakh language in

the field of the humanities. It analyzes the goals, objectives, and significance of developing such a corpus. The relevance of the project lies in the creation of a comprehensive empirical database of scholarly texts in the state language, which contributes to the study of Kazakh as a language of science. The aim of the study is to analyze the functions and objectives of the corpus of humanities texts in the Kazakh language, its role in the development of the Kazakh scientific language, as well as the characteristics of the corpus. For its creation, texts were used from the following fields: philosophy, history, religious studies, archaeology and ethnology, oriental studies, theology, Turkology, museology, foreign philology, translation studies, and Kazakh philology. The article describes the process of corpus development in accordance with the relevant requirements and criteria. Particular attention is given to the significance of the corpus as a tool for the advancement of natural language processing technologies. The corpus makes it possible to identify the features of academic vocabulary, morphology, and syntax of the Kazakh language. The inclusion of various types of scientific sources in the corpus — such as abstracts, theses, conference papers, journal articles, as well as monographs, dissertations, textbooks, and teaching materials — is justified for a comprehensive analysis of scholarly data. Particular emphasis is placed on metadata (meta-annotation), which plays a crucial role at all stages — from the collection to the digitization of materials. Metadata is examined in the context of international practices, with its importance highlighted for the continued study of the Kazakh language as a medium of scientific communication. The methodological section employs methods of description, comparison, philological examination, and analysis. In the study of international practices, descriptive and comparative methods were used, while the analysis of linguistic data from both content-related and structural perspectives involved descriptive and analytical methods. The research results open up new prospects for scholarly inquiry in fields such as terminology, lexicography, cognitive and gender linguistics, translation studies, and other related disciplines. It is emphasized that at all stages — from defining objectives to annotation — it is essential to take scientific and methodological aspects into account. The article also raises issues that require further refinement within the framework of scientific and practical activities, considering the internal classification of the humanities based on their objects of study. The interdisciplinary specificity of maintaining a balanced selection of texts included in the corpus — depending on its intended purpose — has been identified. The study analyzes the impact of the predominance of publications in English and Russian in a number of disciplines on the implementation of the principle of balance in relation to the Kazakh language within the corpus.

Keywords: development of academic corpus, corpus linguistics, academic text, humanities, Kazakh language, scientific style, metadata.

References

1 Qazaq tilinin ulttyq korpusy [National Corpus of the Kazakh Language]. Retrieved from https://qazcorpus.kz/find [in Kazakh].

2 Qazaq tili ulttyq korpusy [National Corpus of the Kazakh Language]. Retrieved from https://qazcorpora.kz/ [in Kazakh].

3 Qazaq tili korpusy [Corpus of the Kazakh Language]. Retrieved from http://test-test-issai.nu.edu.kz/kz-speech-corpus/ [in Kazakh].

4 Almaty qazaq tili korpusy [Almaty Kazakh Language Corpus]. Retrieved from http://webcorpora.net/KazakhCorpus/search/?interface_language=ru [in Kazakh].

5 Janabekova, A.A., Tlegenova, G.B., Mukatayeva, A.Zh., & Jolshayeva, M.S. (2024). Qazaq tilinin oqu korpusyn azirleu maseleleri [Issues of Development of the Kazakh Language Training Corpus]. *Tiltanym — Linguistics*, 1(93). 133–143. DOI: doi.org/10.55491/2411-6076-2024-1-133-143 [in Kazakh].

6 Granger, S. International Corpus of Learner English. Retrieved from https://www.researchgate.net/publication/353637356.

7 British Academic Written English Corpus. Retrieved from https://www.sketchengine.eu/british-academic-written-english-corpus.

8 Ushebnyi korpus russkogo yazyka [Russian learner corpus]. Retrieved from http://www.web-corpora.net/RLC/rulec [in Russian].

9 Jubanov, A. (2013). *Kömpüterlik lıngvıstıkağa kirispe* [Introduction to computational linguistics]. Almaty [in Kazakh].

10 Jubanov, A., & Janabekova, A.A. (2017). *Korpystyq lıngvıstıka* [Corpus linguistics]. Almaty: Qazaq tili [in Kazakh].

11 Aitova, N.N. (2023). Oqu korpusyn zhasau zhane mektep terminologiasyn zertteudi ontailandyru [Development of an academic building and optimization of the study of school terminology]. *«Mektep oqulyqtary: pan terminderin zhetildiry zhane birizdendiry» atty Respublikalyq gylymi-adistemelik konferensia materialdary — Materials of the Republican scientific and methodological conference «School textbooks: improvement and unification of subject terms»* (pp. 8–11). Astana [in Kazakh].

12 Korotkina, I.B. (2024). *Akademisheskoe pismo: protsess, product i praktika* [Academic Writing: process, product and practice]. Moscow: Yurait [in Russian].

13 Yarskaya-Smirnova, E. (2013). *Sozdanie akademisheskogo teksta* [Creation of Academic text]. Moscow [in Russian].

14 Zakharov, V.P. (2005). *Korpusnaia lingvistika* [Corpus linguistics]. Saint Petersburg [in Russian].

15 *Qazaq tilinin akademialyq korpusy* [Academic Corpus of the Kazakh language]. Retrieved from https://academickazakhcorpus.kz/ [in Kazakh].

16 Madieva, G.B., Bektemirova, S.B., & Kuzembekova, J.J. (2014). *Lingvistikalyq zertteulerdegi kömpüterlik tekhnologialar* [Computer technologies in linguistic research]. Almaty: Qazaq University [in Kazakh].

17 Zaki, M. (2022). The metadiscourse of Arabic academic abstracts: A corpus-based study. *Research in Corpus Linguistics*. 10/2, P. 113–146. Spain. https://DOI 10.32714/ricl.10.02.06.

18 A. Baitursynuly matinderinin ishkorpusy [The corpus of A. Baitursynuly's texts]. Retrieved from https://qazcorpus.kz/findahmeti/ [in Kazakh].

19 A. Baitursynulynyn qazaqsha-oryssha parallel korpusy [Kazakh-Russian parallel corpus of A. Baitursynuly]. Retrieved from http://baitursynuly-corp.kz/ [in Kazakh].

20 Aitova, N.N., Baishukurova, G.Zh., & Irgebayeva, A.B. (2024). A. Baitursynuly shygarmalaryn corpustyq zertteu («Qyryq mysal», «Masa» zhinaqtary negizinde) [Corpus-based study of the works of A. Baitursynov (based on the collections "Kyryk mysal" ("Forty examples") and "Masa" ("Mosquito")]. *Qaragandy universitetinin Qabarshysy. Filologiia seriiasy — Bulletin of the Karaganda University. Series Philology, 29, 3*(115), 14–23. DOI: https://doi.org/10.31489/2024Ph3/14-23 [in Kazakh].

21 Pirmanova, K.K., Janabekova, A.A., & Barmenkulova, A. (2022). Ulttyq korpustarğa negizdelgen lingvistikalyq zertteuler zhurgizu (qazaq, opus, ağylshyn tili materialdary negizinde) [Conducting linguistic research based on national corpus (based on materials of Kazakh, Russian, English)]. *Al-Farabi atyndağy Qazaq Ulttyq Universitetinin KHabarshysy. Filologiia seriiasy — Bulletin of Al-Farabi Kazakh National University*, 3(187), 83–93. https://doi.org/10.26577/EJPh.2022.v186.i2.01 [in Kazakh].

22 Matte, M. & Stumpf, E. (2022). A corpus-based study of reporting verbs in academic Portuguese. *Research in Corpus Linguistics*. 10/2, P. 46–69. Spain. https://DOI 10.32714/ricl.10.02.04.

23 Bizhkenova, A.E. & Kenzhebekova, R. (2024). Semantika glagolov dvizhenia v russkom, kazakhskom i angliiskom variantakh (s oporoi na korpusnye i slovarnye dannye) [Semantics of verbs of motion in Russian, Kazakh and English versions (based on corpus and dictionary data)]. *Vestnik Karagandinskogo universiteta. Seriia Filologiia — Bulletin of the Karaganda University. Series Philology, 24*, 1(113), 80–95. https://doi.org/10.31489/2024ph1/80-95 [in Russian].

Information about of author

Konyratbayeva Zhanar Moldalykyzy — Candidate of Philological Sciences, Professor, L.N. Gumilyov Eurasian National University, Astana, Kazakhstan; e-mail: zhanarkon@mail.ru