ӘОЖ 81’33
ХҒТАР 16.21.09:
DOI 10.48371/PHILS.2022.66.3.012

ТАРИХИ ІШКОРПУС ӘЗІРЛЕУДІҢ ӘЛЕМДІК ТӘЖІРИБЕСІНЕН

*Сейітбекова А.А.,[1] Сейдамат Ә.Қ.[2]
[1]ф.ғ.к., ғылыми жетекші, Ахмет Байтұрсынұлы атындағы Тіл білімі институты, Алматы, Қазақстан,
[2]кіші ғылыми қызметкер, Ахмет Байтұрсынұлы атындағы Тіл білімі институты, Алматы, Қазақстан,
*[1]e-mail: Ainurseit@mail.ru,
[2]e-mail: asel.seidamat@tilbilimi.kz

Аңдатпа. Ақпараттық технология дамыған заманда тілдің жазбаша формасы қағаз форматында ғана өмір сүрмейді, оның электрондық нұсқасын даярлау – қазіргі заман талабы. Әлемнің көптеген тілдері жеке ұлттық корпустарын әзірлеуде. Бұндай ауқымды зерттеу жұмыстары қазақ тіл білімінде де «корпустық лингвистика» саласында орын алып келе жатыр. Бүгінгі таңда қазақ тілінің ұлттық корпусына мәтіндер базасы жинақталып, инновациялық-ақпараттық база ретінде үздіксіз жетілдіріп отыр. Ұлттық корпус ұғымы тек синхрониялық емес, сонымен қатар диахроникалық зерттеулердің құралы болып табылады. Осыған орай аталмыш мақалада қазақ тілінің ұлттық корпусының аясында «тарихи ішкорпустың» құрастыру қажеттілігі айтылады. Себебі «тарихи ішкорпус» – V-XX ғғ. қамтитын жазба мұралардың тілін, тарихын, мәдениетін, әдебиетін зерттеуде, қажетті материалды іздеуде кез келген пайдаланушыға онлайн-режімде бірден-бір пайдалы, аса қажетті тілтанымдық құрал болып есептеледі. Сондықтан атамыш мақаланың мақсаты – көне және ортағасыр дәуіріндегі жазба мұралардың мәтіндерін жинақтау, электронды форматқа көшіру және корпусқа ақпараттық және тілтанымдық белгілерімен енгізу. Қолданбалы лингвистика саласында «тарихи ішкорпус» тұңғыш рет жасалғалы жатыр. Алайда түрлі графикамен жеткен қолжазбаларды цифландырудың өзіндік қиындықтары да жоқ емес. Мақалада тарихи ішкорпусты әзірлеудің әлемдік тәжірибедегі құрылымы мен практикалық қолданысы зерттеледі. Яғни, әлем тілдерінің тарихи ішкорпусы (орыс тілінің диахрондық және неміс тілінің тарихи ішкорпусы) әзірлену барысы талданады. Өзге елдердің тәжірибелерін ескере отырып, қазақ тілінің тарихи корпусын әзірлеудің негізгі атқарылатын жұмыстары нақтыланады. Әр дәуірдегі жазба ескерткіштер мәтіндер базасын анықтау, жинау; жинақталған материалдардың сапасы мен құрамын сұрыптау, топтастыру, өңдеу; оларды корпусқа енгізу және демонстрациялау мәселесі, әрбір жазба мұраға ақпараттық белгіленімдердің құрылымы айқындалады. Сонымен қатар тарихи ішкорпус әзірлеудегі қиындықтардың бірі – лингвистикалық белгіленімдердің қойылуы мәселесі назарға алынады. Ұсынылған зерттеу жұмысын кез келген тілдің тарихи ішкорпусын әзірлеуде пайдалануға болады.

Тірек сөздер: корпустық лингвистика**,** тарихи ішкорпус, мәтіндер базасы, транскрипция, факсимиле, лингвистикалық белгіленім, араб графикасы, жазба мұралар.

Негізгі ережелер

Қазақ тілінің тарихи ішкорпусы Қазақ тілінің ұлттық корпусы аясында алғаш рет әзірленіп жатыр. Корпусқа енетін ерте дәуірде, ортағасырда жазылған барлық жазба мұралар мәтіндерінің базасы және

ондағы лигвистикалық мәліметтер шығыстанушыларға, тарихшыларға, түркітанушыларға, жалпы көпшіліке таптырмас пайдалы құрал бола алады.

Кіріспе

Бүгінгі жаһандану кезеңінде өзге елдермен саяси-экономикалық қарым-қатынастарға байланысты ақпарат ағымы күннен-күнге көбеюде. Осыған байланысты әлемдік лингвистикада тілдік корпус мәселесі жан-жақты талқыланып, өзекті мәселеге айналып отыр. Қазақ тіл білімінде қолданбалы лингвистика саласы 1970 жылдардан бастау алады. Тіл зерттеуді автоматтандыру мәселесі профессор Қ. Бектаевтың есімімен байланысты [1]. Одан кейін оның шәкірті профессор А. Жұбановтың «Основные принципы формализации содержания казахского текста» монографиясы қолданбалы лингвистиканың дамуына өз үлесін қосты [2]. Содан бері Ахмет Байтұрсынұлы Тіл білімі институтында «Қолданбалы лингвистика» бөлімі ашылғалы бері тілдік корпусты ғылыми-теориялық және практикалық тұрғыдан зерттеу және қазақ мәтіндерінің базасын өңдеп, корпусқа енгізу мәселесімен айналысып келеді. Электронды тілдік корпус халық игілігі үшін, кез келген тілдік ақпаратты пайдаланушы оңай әрі жылдам табу үшін арналады. Қазіргі таңда әлемдік қолданбалы лингвистикасы саласы бойынша тілдік корпустың барлық функционалды стильдерін қамтитын ішкорпустары әзірленуде. Атап айтқанда, «екі тілді немесе үштілді параллель корпус», «тарихи ішкорпус», «ауызша тіл ішкорпусы», «лексика-семантикалық ішкорпусы», т.б. Жалпы «тарихи ішкорпусты» әзірлеу аталған басқа ішкорпустарға қарағанда аса үлкен жауапкершілікті қажет етеді. Себебі ерте кезеңдегі жазба ескерткіштер мәтіндері әртүрлі графикада және әртүрлі көлемде сақталған. Осыған орай қазіргі таңда «тарихи ішкорпустың» әлемдік тәжірибедегі қолданысын зерттей отырып, қазіргі қоғамға өзінің тарихи мәдени құндылықтарынан мол ақпарат ала алатындай қазақ тілінің тарихи ішкорпусын әзірлеудің маңыздылығы зор. Тарихи ішкорпуста қазіргі қазақ тілі тарихының әртүрлі кезеңдері жазба мұраларынан, тіл тарихынан, сөздің этимологиясынан, т.б. тілтанымдық тұрғыдан ақпарат беретін мәліметтер жиналады. Тарихи мәтіндер базасын жинаудың алғашқы жұмыстары 1928 жылы басталған. Осы жылы ортағасырлық жазба ескерткіш «Кодекс куманикус» мәтіні алғаш реті ЭЕМ жадына енгізілген. Отандық түркітану ғылымы үшін «Куманша-қазақша жиілік сөздігінің» алынуы қазақ тіл білімі үшін аса бағалы зерттеу жұмысы болған [3]. Қазіргі жаңа компьютерлік технологиялар тарихи материалдарды өңдеудің жаңа корпусын қалыптастыруға мүмкіндік береді. А. Фазылжанова: «Тарихи субкорпустың болуы шарт, тарихи кезеңдердегі сол тілдің көрінісін білдіретін мәтіндердің белгіленім қойылған базасы болу керек. Мәселен, ол база болса, XVI-XVIII ғасырлардағы қазақ тілінен хабар ала аласыз немесе бір сөздің тарихи кезеңдердегі көрінісін таба аласыз», – деп атап көрсетті [4].

Материалдар және әдістер

Зерттеу материалдары ретінде қазақ тілінің ұлттық корпус мәселесімен айналысқан отандық ғалымдардың К.Бектаев, А.Жұбанов, А.Фазылжанова, А.Жаңабекова еңбектері басшылыққа алынды. Орыс, түркі, шетел ғалымдарының еңбектері пайдаланылды. Атап айтқанда, В.В. Рыков, С.О. Савчук, Д.В. Сичинава, неміс тілінің тарихи корпусын құрастырған жоба жетекшілері Marcel Erdal, Jost Gippert, Klaus Röhrborn т.б. Зерттеу барысында тарихи ішкорпус әзірлеудің өзге елдің тәжірибелерін зерттеуге, саралауға байланысты *салыстырмалы-салғастырмалы әдіс*, тарихи кезеңдерді белгілеу, жүйелеу, мәтін жанрларын ажырату, топтастыру, транскрипциялау, аудару мәселелеріне қатысты *филологиялық әдіс*, тарихи мәтіндердің компьютерлік көрінісі және олардың өңдеуге байланысты *ақпараттық-технологиялық әдіс* қолданылды.

Нәтижелер мен талқылау

Қазақ тілінің жазу тарихында бірнеше (көне, араб, латын) графика болғаны белгілі. Сол жазулардың ішінде бүгінге дейін араб графикасымен жазылған жазба мұралар қоры мол сақталған. Оның өзін қадим, жадид, төте жазуымен сақталған қолжазбалар деп бөлуге болады. Бүгінгі таңда мәтіндер базасын жаппай цифрландыру кезеңінде араб графикасымен жазылған жазба мұраларды электронды форматқа айнылдыру аса оңай жұмыс емес екені баршаға мәлім. Бұл жайында түрік ғалымы Халит Ерен: « ... коммуникациялық кеңістікте түркі халықтарына ортақ жазба мұраламыздың мәтіндерін теріп, енгізудің лингвистикалық программалаудың жолдарын қарастыру қажет», – дейді [4]. Жазба мұралар мәтіндерінің электронды нұсқалары тіл тарихымен айналысатын зертттеушіге немесе қызығушылық танытып жүрген кез келген пайдаланушыға қолжетімсіз болатыны жасырын емес. Олардың көбі шетелдік немесе отандық мұражайлар мен кітапханаларының сирек қолжазбалар қорында сақтаулы.

Өзге елдердің тарихи ішкорпусын әзірлеудің тәжірибелерін қарастырғанымызда, орыс тілінің ұлттық корпусы өзге тілдердің корпусына қарағанда ертерек басталғанын, әрі жан-жақты қамтылғанын сайттарынан көруге болады [6]. Орыс тілінің *тарихи корпусы* іштей *көне орыс тілі, қабық хаттар* (аққайыңның қабықтарына жазылған көне жазулар), *ескі орыс тілі, шіркеулік славян* болып бөлінеді. Мәселен, 2007 жылы көне орыс тілінің қабыққа жазылған барлық грамоталары корпус базасына қамтылған. Базада әр грамотаның нөмірі, табылған жері, құжаттың түрі, грамотаның мәтіні, транскрипциясы, аудармасы енгізілген [7].

Ескі орыс тілі ішкорпусы XV-XVIII ғасырлар аралығын қамтитын үш миллион сөзқолданыстан тұрады. Корпусқа әдеби шығармалар, шежірелер, іскери хаттар, мақалалар тізімі, қабық хаттар, тұрмыстық

деңгейде жазылған хаттардың мәтіндері енген. Орыс тілінің тарихи корпусының ерекшелігі – пайдаланушы өзіне қажетті ақпаратты тауып қана қоймайды, конкордастағы кез келген сөзге меңзермен нұсқағанда ұяшықтан сол сөз туралы лингвистикалық ақпараттар ала алады. Айталық, «яблоко» сөзін меңзермен басқанда, «яблоко» сөзіне грамматикалық (сущ, вин, с, ед) анықтама беріледі және бұл сөз қанша рет және қанша құжатта кездесетіні туралы толық ақпарат ұсынылады [8].

Орыс тілінің тарихи ішкорпусы жыл сайын жаңа жазба мұралармен толығып, жаңарып тұрады. Мәселен, 2015 жылы 7 мың сөзқолданыстан тұратын XIV-XVII ғғ. жаңа мәтіндер (шежіре, іскери құжаттар) енген болса, 2020 жылы 90 мың сөзқолданыстан тұратын XIII-XVI ғғ. шығыс славяндардың іскери мәтіндері қамтылған. Мәтіндер морфологиялық белгіленімсіз, тек [*] символы арқылы ашылады.

Жалпы әлем тілдерінің корпусын барласақ, көптеген тілдерде тарихи ішкорпусының мәтіндер базасы жасалмаған. Бұны орыс тілінің ұлттық корпусының сайтында берілген тізімінен байқауға болады. Мәселен, сайтта әлем тілдеріне жасалған классификация бойынша орыс тілінің тарихи ішкорпусын есептемегенде, *герман тілдерінен*: американ ағылшын тілінің тарихи корпусы, ескі нидерланд тілінің тарихи корпусы, исланд тілінің тарихи корпусы; *роман тілдерінен*: орта ғасыр француз тілінің корпусы, *үндіеуропа тілдерінен:* валлы тілінің тарихи корпусы, *үнді-еуропа емес тілдерден*: оксфорд көне жапон корпусы, грузин тілінің корпусы, *түркі тілдерінен:* татар тілінің тарихи корпусы жасалғанын көрсетеді. Сондай-ақ, аталмыш тілдердегі онлайн іздеу жүйесі бойынша олардың бірі қолжетімді болса, кейбірі қолжетімсіз болып тұр.

Неміс тілінің ғалымдары, атап айтқанда, проф. Марсель Эрдал, проф. Клаус Рёрборн және проф. Петер исламға дейінгі көне түркі тілдерінің электронды корпусын әзірлегенін «https://vatec2.fkidg1.uni-frankfurt.de» [8] сайтынан анықтадық. 1999-2003 жж. жазба мәтіндер HTML форматында енгізіп және қайта редакцияланғаны туралы ақпарат бар. Корпусқа Алтын Ярлұқ, соғды, көне түркі, ұйғыр жазуымен жазылған мәтіндерді, т.б. көне түркі тілдерінде жазылған жазба ескерткіштердің мәтіндері қамтылған. Жазба ескерткіш туралы метасипаттамасы мен лингвистикалық белгіленімдері жүйелі берілген. Мәселен, Алтын Ярлұқ жазба ескерткішінің мәтінін бөліктерге бөліп, сілтеме арқылы көрсеткен. Мәтінде кездесетін сөздердің транслитерациясын yylq̈y, транскрипциясын yılkı, морфемасын yılkı, аудармасын Vieh кестеге салып, аудармасын жанына көрсеткен. Бұдан түйетініміз, жазба ескерткіштегі әрбір сөзге кесте арқылы берілген талдаулары кез келген пайдаланушыға өте бағалы құрал болатыны аңғарылады.

Қазақ тілінің тарихи мәтіндер базасын төмендегі кезеңдер бойынша, яғни тарихи ішкорпусқа топтастыруға болады:

1. V-IX ғғ. көне түркі дәуіріндегі жазба ескерткіштерінің мәтіндік базасы.

2. X-XV ғғ. орта ғасыр дәуіріндегі жазба мұралардың мәтіндік базасы.

3. XVI-XIX ғғ. ескі қазақ жазба үлгілерінің мәтіндік базасы.

4. XX ғасырдың басындағы жазба мұралардың мәтіндік базасы.

5. Латын графикасымен жазылған жазба мұралардың мәтіндік базасы.

Тарихи ішкорпусқа енгізілетін аталмыш материалдардың көбі араб графикасымен оның ішінде қадим, жадид, төте жазуларымен жазылған. Бұндай түрлі графикада сақталған жазба ескерткіштер мәтіндеріне қызығушылық танытатындар – зерттеушілер (түркітанушылар, тарихшылар, әдебиетшілер).

Араб графикасымен жазылған жазба ескерткіштер мәтіндерін корпусқа енгізу мәселесін қарастырайық:

Корпусқа әр дәуірдегі көлемі әртүрлі жазба ескерткіштер алынады;

- әр бір жазба мұраға ақпараттық белгіленімдер көрсетіледі;
- мәтіндегі әрбір сөзге лингвистикалық белгіленімдер қойылады.

Қазақ тілінің тарихи корпусы алғаш рет жасалып жатқандықтан, XII-XX ғасырлардағы әр дәуірден екі-үш жазба мұрадан ғана алынады. Әзірше араб графикасымен жазылған қолымызда бар жазба мұралармен жұмыс жасалады. Материалдарды тезірек жинақтауда, алдымен, мәтіндердің дайын электрондық нұсқаларын ашық электронды кітапханалардан алыну керек. Әрине, араб графикасымен сақталған материалдардың, әсіресе жарық көрмеген жазба мұралардың көбі сирек қолжазбалар қоры сақталатын кітапханалардан табуға болады.

Қазақ тілінің тарихи мәтіндер базасы әртүрлі жанрдан тұратыны белгілі. Сондықтан базаға шежіре, дастан, жыр-өлең, проза, оқу құралы, мақала, хат, жарлықтар, т.б. жанр түрлері енеді. Осындай түрлі жанрдан тұратын жазба мұралардың қолжазбаларын демонстрациялауда http://gramoty.ru/ «Древнерусские берестяные грамоты» сайтындағы алгоритімін ескеруге болатынына көз жеткіздік [9]:

біріншіден, осы уақытқа дейін әлемдік және отандық түркітанушы ғалымдар араб графикасымен жазылған қолжазба мәтіндерін барынша өз тілдеріне жақын және әртүрлі графикамен транскрипциялап келді. Кейбір зерттеушілер жазба ескерткіштер мәтіндерімен жұмыс жасағанда, дайын транкрипцияланған мәтіндерді пайдаланады. Нәтижесінде шынайы зерттеу жұмыстары жасалмайды. Сондықтан корпусқа қолжазбаның факсимилесімен берілетін мәтіндер – ізденушінің мәтіндердің шынайы қалпын көруге, талдауына мүмкіндік береді.

екіншіден, араб графикасымен жазылған мәтіндерде бұлыңғыр жазылған әріптер, сөздер, тіпті сөйлемдер жиі кездеседі, бұл сөздердің графикасын, мағынасын анықтауда қиындық келтіреді. Салыстырмалы жұмыс жасау үшін араб графикасымен жазылған қолжазбаның факсимилесі, кирил графикасындағы транскрипциясы, аудармасы жеке-жеке бетте қатар беріліп, демонстрацияланады.

Әр бір жазба мұраға берілетін *ақпараттық белгіленімдері* толық, дәл болуы керек. Әлемдік тарихи корпустың тәжірибелерін ескере отырып, әрбір жазба мұраға берілетін ақпараттық белгіленім төмендегідей болуын жоспарлап отырмыз:

- мәтін туралы қысқаша аңдатпа;
- мәтін авторы (аты-жөні);
- мәтін тақырыбы;
- мәтін графикасы: көне жазу, араб-қадим, араб-жадид, араб-төте;
- мәтіннің алғашқы жарияланымы;
- мәтін стилі: ресми, публицистикалық, тарихи шежіре, көркем, ғылыми, педагогикалық дискурс;
- мәтін жанры: проза, поэзия, (аралас: поэзия, проза), сөздік, хат, жарлық, газет, өтініш, арыз, оқулық, мақала, ғылыми ақпараттық, ғылыми көпшілік.
- мәтін дереккөзі;
- мәтіннің нұсқалары: Қазан нұсқасы, Демезон нұсқасы,
- мәтін транскрипциясының авторы;
- мәтін аудармасының авторы;
- мәтіннің бет саны;
- аудитория жасы;
- мәтінді корпусқа енгізген автор;
- мәтінді корпусқа енгізген мерзімі;
- Ескерту: жоғарыдан төмен берілген сандар сөйлемнің қатарын көрсетеді.

Тарихи корпус әзірлеудегі қиындықтардың бірі – мәтіндердегі сөздерге *лингвистикалық белгіленімдердің қойылуы*. Жазба ескерткіштер мәтіндеріндегі кез келген сөз туралы лингвистикалық ақпарат көрсететтетін конкордансты жасау қиындық тудыруы мүмкін. Себебі араб графикасымен жазылған мәтіндердің қолжазбасы әртүрлі каллиграфиямен берілетіндіктен, таныла бермеуі мүмкін. Өйткені мәтіннің танылуында мынадай қиындықтар кездеседі: кейбір диакритикалық белгілердің (кесра, сукун, фатха), графикалық символдардың анық көрінбеуі немесе түсіп қалуы; кейбір сөздердің бірігіп жазылуы. Оған арнайы автоматтындырылған белгіленім түрлерін қарастыруды қажет етеді. Бұл мәселе орыс ғалымдарының күн тәртібінен түспеген. Тарихи корпуспен айналысқан Д.В. Сичанова «хлеб» сөзінің морфологиялық белгіленімін қойғанда, көне *хлѣбъ* нұсқасы таныла ма, әлде қазіргі *хлеб* нұсқасы таныла ма? деген мәселе туындағанын айтады [10, 209 б]. Мәтіндегі кез келген сөзге автоматты түрде көрсетілетін лингвистикалық (фонетикалық, грамматикалық, семантикалық) белгіленімдерді түркітанушылар, шығыстанушылар мен корпустық бағдарламалармен қамтамасыз ететін мамандармен бірлесіп анықтау, талдау – алдымызда тұрған міндет.

Қорытынды

Сонымен ғылым мен техниканың үлкен қарқынмен даму кезеңінде жаңа техникалық мүмкіндіктерді пайдалана отырып, тарихи мәтіндердің ішкорпусын құрастыру – кезек күттірмейтін мәселе. Тарихи ішкорпус әзірлеудің басты артықшылықтары мынада:

– лингвистикалық белгіленімдер зерттеушілер үшін тек ақпараттық құрал қызметін ғана емес, өздеріне қажетті анықтамалық дереккөз де бола алады. Зерттеуші кез келген жерде және кез келген уақытта отырып, бастаған зерттеу жұмысын жалғастыра алады;

– факсимиленің әр бетіне сәйкес берілетін мәтіннің транскрипциясы мен аудармасы зерттеушінің тарихи-салыстырмалы, диахронды-сихронды зерттеулер жасауына мүмкіндік береді;

– нұсқалар арасында лингвотекстологиялық зерттеулер, лексикографиялық жұмыстар жасауға болады. Бұл әрі қарай архаикалық тілдік (фонетикалық, морфологиялық, лексикалық, фразеологиялық), т.б. белгіленімдерінің модельдерін нақтылауға көмектеседі.

Қазақ тілінің тарихи ішкорпусын әзірлеудің мәселелерін қарастырғанымызда, мынадай нәтижелерге қол жеткізілді:

біріншіден, қазақ тілінің тарихи ішкорпусын әзірлеудің әлемдік тәжірибесі ескерілді;

екіншіден, қазақ тілінің тарихи мәтіндер базасын *көне түркі, орта түркі, ескі қазақ жазба үлгілері, XX ғасырдың басындағы жазба мұралар* деп 4 кезеңге топтастыруға болады;

үшіншіден, пайдаланушыға ыңғайлы болу үшін сканерленетін жазба мұраның факсимилесінің әр беті жеке-жеке беріліп, сол мәтіннің *транскрипциясы* мен *аудармасы* демонстрацияланады;

төртіншіден, V-XX ғғ. жазба мұралардың стилі, жанры, т.б. ерекшеліктеріне қарай ақпараттық белгіленімдердің тізімі ұсынылады.

Зерттеу жұмысы BR11765619 «Мемлекеттік тілдің ақпараттық-инновациялық базасы ретіндегі қазақ тілінің ұлттық корпусын әзірлеу: ғылыми-зерттеу және оқыту интернет-ресурсы» атты бағдарламалық-нысаналы зерттеу аясында әзірленді.

ӘДЕБИЕТ

[1] Бектаев К.Б. Статистико-информационная типология тюркского текста. – Алма-Ата: Изд-во. "Наука" КазССР, 1978. – 184 с.

[2] Жубанов А. Основные принципы формализации содержания казахского текста. – Алматы, 2002. – 250 с.

[3] Жұбанов А.Қ., Жаңабекова А.Ә. Корпустық лингвистика. – Алматы: Қазақ тілі, 2017. – 336 б.

[4] Фазылжанова А. Үштілділік ұлттық мүддеге қызмет етуі тиіс // http://www.ikitap.kz. –Тамыз, 2016.

[5] Редькин И.О. Формирование корпуса текстов и определение частотности слов в арабском языке: проблемы и решения // Вестник Санкт-Петербургского университета. Востоковедение и африканистика. – 2014. – № 1. – С. 14-22.

[6] Сичинова Д.В., Дышкант А.Н. Базы данных некнижной письменности, исторический корпус, словоуказатель: интеграция // Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной международной

конференции «Диалог» (Москва, 17-20 июня 2020 г.). Вып. 19 (26), дополнительный том. –М.: Изд-во РГГУ, 2020. – С. 1138-1148.

[7] Национальный корпус русского языка // URL: https://ruscorpora.ru/new.

[8] Vorislamische Alttürkische Texte: Elektronisches Corpus // URL: https://vatec2.fkidg1.uni-frankfurt.de

[9] Древнерусские берестяные грамоты. http://gramoty.ru.

[10] Сичинова Д.В. Старорусские/среднерусские тексты в НКРЯ: от экстенсивной коллекции к корпусу // Written Heritage and Digital Technologies: VI International Scientific Conference, Vilnius, 2016. August 22-28, pp. 208-210.

REFERENCES

[1] Bektaev K.B. Statistiko-informacionnaja tipologija tjurkskogo teksta (Statistical and informational typology of the Turkic text). Alma-Ata: Izd-vo. "Nauka" KazSSR, 1978. 184 p. [In Rus.]

[2] Zhubanov A. Osnovnye principy formalizacii soderzhanija kazahskogo teksta(Basic principles of formalization of the content of the Kazakh text). Almaty, 2002. 250 p. [In Rus.]

[3] Jūbanov A.Q., Jañabekova A.Ä. Korpustyq lingvistika(Corpus linguistics). Almaty: Qazaq tılı, 2017. 336 p. [in Kaz.]

[4] Fazyljanova A. Üştıldılık ūlttyq müddege qyzmet etuı tiıs(Trilingualism should serve the national interest) // http://www.ikitap.kz. Tamyz, 2016. [in Kaz.]

[5] Red'kin I.O. Formirovanie korpusa tekstov i opredelenie chastotnosti slov v arabskom jazyke: problemy i reshenija(Formation of the corpus of texts and determination of the frequency of words in the Arabic language: problems and solutions)// Vestnik Sankt-Peterburgskogo universiteta. Vostokovedenie i afrikanistika. 2014. № 1. P. 14-22. [In Rus.]

[6] Sichinova D.V., Dyshkant A.N. Bazy dannyh neknizhnoj pis'mennosti, istoricheskij korpus, slovoukazatel': integracija (Non-book writing databases, historical corpus, word index: integration) // Komp'juternaja lingvistika i intellektual'nye tehnologii: Po materialam ezhegodnoj mezhdunarodnoj konferencii «Dialog» (Moskva, 17-20 ijunja 2020 g.). Vyp. 19 (26), dopolnitel'nyj tom. M.: Izd-vo RGGU, 2020. S. 1138-1148. [In Rus.]

[7] Nacional'nyj korpus russkogo jazyka (Russian National Corpus) // URL: https://ruscorpora.ru/new. [In Rus.]

[8] Vorislamische Alttürkische Texte: Elektronisches Corpus // URL: https://vatec2.fkidg1.uni-frankfurt.de

[9] Drevnerusskie berestjanye gramoty. http://gramoty.ru. [In Rus.]

[10] Sichinova D.V. Starorusskie/srednerusskie teksty v NKRJa: ot jekstensivnoj kollekcii k korpusu (Old Russian/Middle Russian texts in the National Corpus of the Russian Language: from an extensive collection to a corpus) // Written Heritage and Digital Technologies: VI International Scientific Conference, Vilnius, 2016. August 22-28, pp. 208-210. [In Rus.]

ИЗ МИРОВОГО ОПЫТА РАЗРАБОТОК ИСТОРИЧЕСКОГО ПОДКОРПУСА

*Сейтбекова А.А.,[1] Сейдамат А.К.[2]

[1]к.ф.н., ведущий научный сотрудник, Институт языкознания имени Ахмета Байтурсынулы, г. Алматы, Казахстан,

[2]младший научный сотрудник, Институт языкознания имени Ахмета Байтурсынулы, г. Алматы, Казахстан,

*[1]e-mail: Ainurseit@mail.ru,

[2]e-mail: asel.seidamat@tilbilimi.kz

Аннотация. В эпоху развития информационных технологий подготовка письменных форм в электронном формате стало требованием времени. Во многих странах мира разрабатываются свои собственные национальные корпусы. Такая масштабная исследовательская работа ведется и в казахском языкознании («корпусная лингвистика»). На сегодняшний день Национальный корпус казахского языка собрал внушительную базу текстов. Он непрерывно совершенствуется как инновационно-информационный источник. Понятие национального корпуса является инструментом не только синхронических, но и диахронических исследований. В данной статье рассмотрена необходимость создания «исторического подкорпуса» в рамках Национального корпуса казахского языка. «Исторический подкорпус» представляет собой один из самых востребованных лингвистических инструментов для любого пользователя в онлайн-режиме, в целях поиска им необходимых материалов для изучении языка, истории, культуры, литературы письменного наследия V-XX вв. Цель данной статьи – сбор, оцифровка и внесение в корпус с информационными и языковыми метаобозначениями текстов древнего и средневекового письменного наследия. В области прикладной лингвистики такой исторический подкорпус создается впервые. Есть особые сложности в оцифровке дошедших до нас рукописей, имеющих различную графику. В статье исследуются структура и практическое применение в мировой практике разработки исторического подкорпуса: анализируется процесс разработки исторического подкорпуса мировых языков, в частности исторический подкорпус русского и немецкого языков. С учетом опыта других стран будут определены основные направления по разработке исторического корпуса казахского языка: вопросы выявления и сбора текстов письменных памятников разных эпох; сортировка, классификация, обработка качества и состава собранных материалов, введение и демонстрация текста, определение структур информационных метаобозначений для каждого текста. Также принимается во внимание проблема постановки лингвистических обозначений – это одна из трудностей в разработке исторического подкорпуса. Представленная исследовательская работа может быть использована при разработке исторического подкорпуса любого языка.

Ключевые слова: корпусная лингвистика, исторический подкорпус, база текстов, факсимиле, транскрипция, лингвистическая разметка, арабская графика, письменные памятники.

FROM THE WORLD EXPERIENCE OF THE DEVELOPMENT OF THE HISTORICAL SUBCORPUS

*Seitbekova A.A.,[1] A.K. Seidamat[2]

[1]Candidate of Philological Sciences, leading researcher, Institute of Linguistics named after Akhmet Baitursynuly, Almaty, Kazakhstan,

[2]junior researcher, Institute of Linguistics named after Akhmet Baitursynuly, Almaty, Kazakhstan,

*[1]e-mail: Ainurseit@mail.ru,

[2]e-mail: asel.seidamat@tilbilimi.kz

Abstract. In the era of information technology development, the preparation of written forms in electronic format has become a requirement of the time. Many countries of the world are developing their own national buildings. Such large-scale research work is also being carried out in Kazakh linguistics ("corpus linguistics"). To date, the National Corpus of the Kazakh language has collected an impressive database of texts. It is continuously being improved as an innovative information source. The concept of a national corpus is a

tool not only for synchronic, but also for diachronic research. This article discusses the need to create a "historical subcorpus" within the National Corpus of the Kazakh language. "Historical Subcorpus" is one of the most popular linguistic tools for any user in online mode, in order to find the necessary materials for them to study the language, history, culture, literature of the written heritage of the V-XX centuries. The purpose of this article is to collect, digitize and add texts of ancient and medieval written heritage to the corpus with informational and linguistic meta–meanings. In the field of applied linguistics, such a historical subcorpus is being created for the first time. There are special difficulties in digitizing extant manuscripts with different graphics. The article examines the structure and practical application in the world practice of the development of the historical subcorpus: the process of developing the historical subcorpus of world languages, in particular the historical subcorpus of the Russian and German languages, is analyzed. Taking into account the experience of other countries, the main directions for the development of the historical corpus of the Kazakh language will be determined: issues of identification and collection of texts of written monuments of different eras; sorting, classification, processing of the quality and composition of collected materials, introduction and demonstration of the text, definition of structures of informational meta-meanings for each text. The problem of the formulation of linguistic designations is also taken into account – this is one of the difficulties in developing a historical subcorpus. The presented research work can be used in the development of the historical subcorpus of any language.

Keywords: corpus linguistics, historical subcorpus, database of texts, facsimile, transcription, linguistic markup, arabic graphics, written monuments.

Статья поступила 11.05.2022