ҒТАМР 16.01.09 doi.org/10.55491/2411-6076-2023-3-163-172

А. Сейітбекова[1]*, А. Фазылжанова[1], Ғ. Аязбаев[1]

[1]Ахмет Байтұрсынұлы атындағы Тіл білімі институты, Алматы, Қазақстан
*e-mail: ainurseit@mail.ru

ТАРИХИ ІШКОРПУС МӘТІНДЕРІНІҢ ЛЕКСИКА-ГРАММАТИКАЛЫҚ БЕЛГІЛЕНІМІ

Аннотация. Белгілі бір ұлттық корпустың мәтіндер басазын әзірлеуде тек метасипатта мен метадеректік ақпараттарды әзірлеумен шектелмейді. Сондай-ақ, лингвистикалық белгіленімдерді де әзірлеуді қажет етеді. Лингвистикалық белгіленім – мәтіндегі әрбір лексикалық бірлікке орфографиялық, фонетикалық, лексикалық, грамматикалық тұрғыдан берілетін лингвистикалық ақпарат. Алайда тарихи ішкорпус мәтіндеріне лингвистикалық белгіленім әзірлеу терең ізденісті қажет ететін аса күрделі жұмыстардың бірі. Себебі тарихи корпусқа енген мәтіндер базасының басым көпшілігі араб графикасымен жазылған. Ал араб графикасымен жазылған мәтіндер кирилл графикасына әртүрлі транскрипцияланған. Тарихи ішкорпусты әзірлеуде басқа ішкорпустарға қарағанда теориялық және техникалық тұрғыдан өзіндік қиындығы бар. Осығай орай мақаланың *мақсаты* – Ахмет Байтұрсынұлы атындағы Тіл білімі институтында алғаш рет әзірленіп жатқан тарихи ішкорпус мәтіндеріне лингвистикалық белгіленімдерді қою мәселесін қарастыру. *Міндеттері*: транскрипцияланған мәтіндер үшін лингвистикалық, оның ішінде лексика-грамматикалық белгіленімдерді таңдау, анықтау; лексика-грамматикалық белгіленімдерді әзірлеуде өзге елдердің тәжірибесін ескеру; арабграфикалы мәтіндерінің кирилл графикасындағы транскрипциясын талдау; транскрипцияланған сөздердің вариантарын анықтау; транскрипцияланған мәтіндерге арналған лексика-грамматикалық белгіленімдер бағдарламасының жұмыс істеу механизмдері туралы сипаттау.

Зерттеу барысында сипаттау, тарихи-салыстырмалы, лингвотекстологиялық *әдістер* қолданылады.

Зерттеу *нәтижесінде* лексика-грамматикалық белгіленім әзірлеуде орыс тілінің тарихи ішкорпусы әзірлеу тәжірибесі қарастырылады; әр кезеңде жазылған арабграфикалы мәтіндерінің транскрипциясы талданады; транскрипцияланған сөздердің варианттары анықталады; транскрипцияланған мәтіндер үшін лексика-грамматикалық белгіленімдер айқындалады; транскрипцияланған мәтіндерге арналған лексика-грамматикалық іздеу жүйесінің механизмдері сипатталады.

Практикалық маңыздылығы. Тарихи корпусқа енетін арабграфикалы мәтіндердің транскрипциясына жасалған лексика-грамматикалық белгіленімдердің әзірленуі – белгілі бір лексикалық бірліктің эволюцисын білу үшін пайдалы құрал бола алады.

Тірек сөздер: тарихи ішкорпус, араб графикасы, транскрипция, лексика-грамматикалық белгіленім, жазба ескерткіштер.

А. Сейтбекова[1*], А. Фазылжанова[1], Г. Аязбаев[1]

[1]Институт языкознания имени Ахмета Байтурсынова, Алматы, Казахстан
*e-mail: ainurseit@mail.ru

ЕКСИКО-ГРАММАТИЧЕСКАЯ РАЗМЕТКА ТЕКСТОВ ИСТОРИЧЕСКОГО ПОДКОРПУСА

Аннотация. База определенного национального корпуса не ограничивается информацией в виде метаописаний и метаданых введеных текстов. Также необходимо разработать лингвистические разметки. Лингвистическая разметка – это лингвистическая информация, охарактеризованная каждой лексической единице в тексте по орфографическим, фонетическим, лексическим, грамматическим признакам. Однако разработка лингвистической разметки исторического подкорпуса является одной из сложных задач требующих глубокого исследования. Это связано с тем, что большинство текстов внесенные в исторический подкорпус написаны арабской графикой. А тексты средневековья написанные арабской графикой транскрибировались по-разному. Разработка исторического подкорпуса имеет сложности как теоретически, так и технически, по сравнению с другими подкорпусами. В связи с этим *целью* статьи является рассмотрение вопроса лингвистических разметок текстов исторического подкорпуса, которые впервые разрабатываются в Институте языкознания имени Ахмета Байтурсынова. *Задачи*: определить лингвистических, из них лексико-грамматических разметок для транскрибированных текстов; учитывать опыты других стран при разработки лексико-грамматических разметок; анализировать транскрибированнных текстов с арабской графики на кириллическую графику; выявить вариативности транскрибированных слов; описать механизм функционирования программы лексико-грамматических разметок.

В ходе исследования используется описательный, историко-сравнительный, лингвотекстологический, лингвостатистические методы.

В результате исследования при разработке разметок рассматривались опыты разработок исторического подкорпуса руского языка; аназировались транскрибирования текстов написанные арабской графикой разного периода

средневековья; определены лексико-грамматические разметки для траскрибированных текстов; описаны механизмы лексико-грамматической поисковой системы для транскрибированных текстов.

Практическая значимость. Разработка лексико-грамматических разметок для транскрибированных текстов внесенные в исторический подкорпус будет полезным лингвистическим инструменом для изучения эволюции определенной лексической единицы.

Ключевые слова: исторический подкорпус; арабская графика, транскрипция, лексико-грамматическая разметка, письменные памятники.

A. Seitbekova[1]*, A. Fazyljanova[1], Ғ. Aiazbayev[1]

[1]Akhmet Baitursynuly Institute of Linguistics, Almaty, Kazakhstan
*e-mail: ainurseit@mail.ru

LEXICAL AND GRAMMATICAL MARKUP OF TEXTS OF THE HISTORICAL SUBCORPUS

Annotation. The base of a certain national corpus is not limited to information in the form of meta descriptions and metadata of entered texts. It is also necessary to develop linguistic markup. Linguistic markup is a linguistic information that is cataractarized to each lexical unit in the text according to spelling, phonetic, lexical, grammatical features. However, the development of the linguistic markup of the historical subcorpus is one of the complex tasks that require in-depth research. This is due to the fact that most of the texts included in the historical subcorpus are written in Arabic graphics. And the texts of the Middle Ages written in Arabic graphics were transcribed in different ways. The development of a historical subcorpus has difficulties both theoretically and technically, compared with other subcorpuses. In this regard, the purpose of the article is to consider the issue of linguistic markings for the texts of the historical subcorpus, which are being developed for the first time at the Institute of Linguistics named after Akhmet Baitursynula. Tasks: to identify linguistic, lexical and grammatical markup for transcribed texts; to take into account the experiences of other countries in the development of lexical and grammatical markup; to analyze transcribed texts from Arabic graphics to Cyrillic graphics; to identify the variability of transcribed words; to describe the mechanism of functioning of the lexical and grammatical markup program.

The study uses descriptive, historical-comparative, linguotextological, linguostatistical methods. As a result of the study, when developing the markup, the experiments of the development of the historical subcorpus of the Russian language were considered; transcribing texts written in Arabic graphics of different periods of the Middle Ages were anasized; lexical and grammatical markup for transcribed texts were determined; the mechanisms of a lexical and grammatical search system for transcribed texts were described.

Practical significance. The development of lexical and grammatical markup for transcribed texts included in the historical subcorpus will be a useful linguistic tool for studying the evolution of a certain lexical unit.

Keywords: historical subcorpus; Arabic graphics, transcription, lexico-grammatical markup, written monuments.

Кіріспе

Корпустық лингвистика қажетті әрі сұранысқа ие салалардың біріне айналды. Осыған байланысты қазіргі жаһандану кезеңінде ғылым мен техниканың қарыштап дамуына байланысты қазақ тіліндегі материалдар базасын жасақтап, оның тілдік корпусын әзірлеу – өзекті мәселелердің бірі. Бүгінгі таңда тілдің ішкі мүмкіншілігіне орай ұлттық корпусты кеңейту, тереңдету мақсатында түрлі ішкорпустар әзірленіп жатыр. Сондай ішкорпустардың бірі – тарихи ішкорпус. Тарихи ішкорпус әзірлеуде белгілі бір графикамен жазылған қолжазбаның тек метасипаттамасы ғана берілмейді, сонымен қатар мәтіндегі әр лексикалық бірлікке лингвистикалық ақпараты бар белгіленімдер қойылады. Белгіленім – корпустың басты сипаттамасы. Сондықтан белгіленімнің түрлері неғұрлым көп болса, соғұрлым корпустың құндылығы артады.

Жалпы белгіленімдер атқаратын функциясына қарай сыртқы және ішкі деп бөлуге болады. *Сыртқы белгіленімге* мәтін туралы ақпарат, яғни авторы, тақырыбы, жылы, жарны, стилі, көлемі, т.б. Оларды библиографиялық, типологиялық, тақырыптық, әлеуметтік, формальдық (мәтін, тарау, бөлім, абзац, сөйлем т.б.) және техникалық (орындаушылар, электрондық нұсқа алынған дереккөз, өңделген күні, кодталған кезі т.б.) деп те бөледі. *Ішкі белгіленімге* мәтіндегі әр сөзге берілетін лингвистикалық (фонетикалық, лексикалық, сөзжасамдық, синтаксистік, лингвомәдени, т.б.) ақпарат (Жаңабекова, 2021). Бұндай лингвистикалық белгіленімдері бар тарихи ішкорпустар, әсіресе, тілтанушы-зерттеушілер, мектеп мұғалімдері мен оқушылар үшін өте пайдалы ақпарат беретен дереккөз болатыны сөзсіз. Қазақ тілінің тарихи ішкорпусына енгізілетін мәтіндер базасы, олардың графикасы мен транскрипциясы, оларға қойылатын лингвистикалық белгіленім мәселелері алғаш рет қарастырылып жатыр. Бұл зерттеу жұмысының өзектілігін көрсетеді.

Материал және әдістер

Жұмыстың зерттеу материалдары ретінде орта ғасырдың әр кезеңіндегі арабграфикалы жазба ескерткіштер мәтіндері пайдаланылды. Атап айтқанда, XI ғасырдағы Ахмет Иасауиді «Диуани хикмет», XII ғасырдағы Ахмет Йүгнекидің «Хабатул хақайық», XIII ғасырдағы Хорезмидің «Мұхаббатнамесі», XVII ғасырдағы Әбілғазы баһадүр ханның «Түркі шежіресі» жазба ескерткіштері. Зерттеу барысында әлемдік тәжірибесіндегі лингвистикалық белгіленім қоюдағы *сипаттау, салыстырмалы әдістер,* жазба ескерткіштер мәтіндерінің транскрипциялау ерекшелектерін талдауда *тарихи-салыстырмалы*, өзара мәтіндерді салыстыруда *лингвотекстологиялық*, лингвистикалық белгіленімдерді таңдауда, анықтауда *топтастыру, жүйелеу*, графикалық, орфографиялық варианттарын анықтауда *лингвостатистикалық әдістер* қолданылады. Сондай-ақ, көне грек, испан, португал, венгер тілдерінің тарихи корпустарының сайттары зерттеу материалдары ретінде пайдаланылды.

Әдебиетке шолу

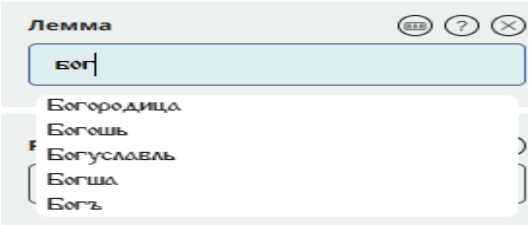
Тарихи ішкорпус жадына енгізілетін мәтіндердегі сөздерге лингвистикалық белгіленімдерді жүйелі құрылыммен қоюдың теориялық және практикалық әдіс-тәсілдерін меңгергенде ғана мүмкін болады. Осыған байланысты жалпы тарихи ішкорпус әзірлеу бойынша әлемдік қолданбалы лингвистика саласында құрылымы мен мазмұны тұрғысынан озық келе жатқан орыс тілінің тарихи ішкорпусын атауға болады. Аталмыш ішкорпус іштей көне орыс тілі, қабық хаттар, ескі орыс тілі және шіркеулік славян ішкорпустары болып бөлінеді. Барлық осы аталған ішкорпусының мәтіндеріндегі әрбір сөзге автоматтандырылған сөздің бастапқы тұлғасы (леммасы) мен грамматикалық белгіленімдері қойылған. Сондықтан зерттеуде орыс тілінің тарихи ішкорпус әзірлеушілердің, атап айтқанда, Т.С. Гаврилова, Т.А. Шалганова, О.Н. Ляшевская, С.О. Савчук, Д.В. Сичинова, сондай-ақ, қалмақ ғалымдарының Е.В. Бембеев, отандық ғалымдарының А. Жұбанова, А. Жаңабекованың, араб графикасынан кирилл графикасына транскрипциялаған Р. Сыздық, Э. Фазылов, Э. Наджип, Б. Сағындықов, Б.Әбілқасымов, С. Иванов, Г. Сердюченконың еңбектері басшылыққа алынды.

Нәтижелер және талқылау

Кез келген тілдің тарихи ішкорпусының метадеректері мен метасипаттамаларынан бөлек лингвистикалық белгіленімдерінің таңдалуы мен қойылуы аса үлкен қажырлы еңбек пен жауапкершілікті қажет етеді. Бұл жайында орыс тілінің тарихи ішкорпусын әзірлеушілері мынадай пікір айтады: «Ескі орыс кезеңіндегі мәтіндерді автоматтандырылған морфологиялық талдауды қиындататын бірқатар мәселелер белгілі. Біріншіден, бұл әртүрлі грамматикалық және лексикалық қабаттарды біріктіретін тілдің өтпелі кезеңі. Көптеген мәтіндер алдыңғы кезеңдегі (XI–XIV ғасырлар) тілдің ерекшеліктерін көрсетеді. Мәтіндерде шіркеу славян тілінің лексикалық немесе грамматикалық элементтері де бар. Әртүрлі аймақтарда жазылған мәтіндердің диалектілік ерекшеліктерін де назардан тыс қалдыруға болмайды. Мұның бәрі ескі орыс мәтіндерін сәтті талдау үшін компьютерлік бағдарлама «гибридті» морфологияға негізделуі керек, атап айтқанда, ол ескі орыс және қазіргі орыс тілдерін ғана емес, сонымен қатар шіркеу славян лексемалары мен парадигмаларын да ескеру керек. Екіншіден, ескі орыс тіліндегі мәтіндердің емлесі тұрақсыз. Сондықтан корпус мәтіндерінде сөздің бірнеше орфографиялық варианттары болады. Мұндай варианттар әртүрлі факторларға байланысты болуы мүмкін, бастапқы мәтіннің құрылуына немесе мәтін жасалған аймақтың диалектілік ерекшеліктеріне, сөйлеушіні сауаттылыққа үйреткен ұстазына байланысты. Осыған орай лингвистикалық белгіленімдерде орфографиялық варианттар неғұрлым көп ескерілсе, мәтіндер соғұрлым толық және дәл талданады. Үшіншіден, техникалық қиындық – ескі орыс мәтіндерін автоматты түрде белгілеу үшін ресурстардың болмауы. Қазіргі орыс тіліндегі мәтіндерді өңдеу үшін жасалған анализаторлар ескі орыс мәтіндеріне жоғарыда аталған мәселелерге байланысты морфологияның гетерогенділігі және орфографиялық варианттар үшін қолайсыз болып отыр» (Гаврилова Т. С., Шалганова Т. А., Ляшевская О. Н., 2016).

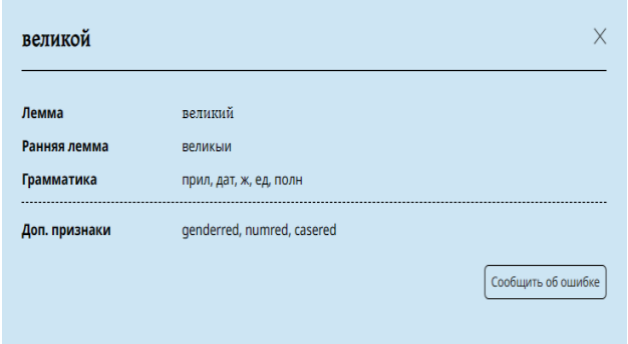
Сонымен көне орыс тілі, қабық хаттар, ескі орыс тілі, шіркеулік славян ішкорпустарының лексика-грамматикалық іздеу жүйесінің құрылымы өзара ұқсас болғанмен, айырмашылықтары да бар. Мәселен, көне орыс тілі ішкорпусы лексема, түсіндіру, грамматикалық, қосымша белгіленімдерден тұрады. Қабық хаттар ішкорпусы көне мәтіннің түпнұсқа және аудармасы, лемма, бастапқы лемма, сөзқолданыс, грамматикалық және қосымша белгіленім, түсіндіру, негізгі (орыс

тіліндегі аудармасымен) семантикадан тұрады. Ескі орыс тілі ішкорпусы лемма, көне лемма, сөзқолданыс, грамматикалық және қосымша белгіленімдерден тұрады. Шіркеулік славян ішкорпусы лемма, сөзқолданыс, грамматикалық және қосымша белгіленімдерден тұрады. Осы аталған ішкорпустардың барлығына ортақ іздеу жүйесі грамматикалық және қосымша белгіленімдердің болуы. Грамматикалық белгіленімде лексеманың немесе сөзформасының қажетті морфологиялық сипаттамалары көрсетіледі. «Таңдау» сілтемесі арқылы ашылатын қызмет терезесінде грамматикалық категорияларға бөлінген морфологиялық сипаттамалардың тізімі көрсетілген: сөз табы, септіктер, тегі, етіс, сан есім және т.б. «Қосымша белгіленімде» көрсетілген белгілердің көмегімен сөздерді белгілі бір позицияда іздеуге болады. Мәселен, тыныс белгісінің алдында және одан кейін, сөйлемнің басында немесе соңында. «Түсіндіру» іздеу жүйесінде омонимдерді ажыратуға мүмкіндік беретін көмескі мағынадағы сөздердің қысқаша түсіндірмелері көрсетілген. Сондай-ақ, сөздің бастапқы формасы (лемма) іздеу жүйесінде таңдалған сөздің лексемасы немесе сөз формасы көрсетіледі.



1-Сурет. – «Бог» сөзнің леммасы мен сөзформасы.

Лексикалық-грамматикалық іздеу жүйесі сөздің бастапқы формасынан (лемма) бөлек грамматикалық (сөз табы, септік, жеке, көпше түрі) сипаттамаларын анықтаудан тұрады. Пайдаланушы мәтіндегі кез келген сөзге меңзермен нұсқағанда ұяшықтан сол сөз туралы лингвистикалық ақпараттар ала алады. Айталық, «великий» сөзін меңзермен басқанда, мынадай грамматикалық (прил, дат, ж, ед. полн) анықтама беріледі.



2-Сурет. – «Великий» сөзінің грамматикалық анықтамасы.

Қалмақ тілтанушылары ескі қалмақ жазба ескерткіштері мәтіндеріндегі «көне» графемаларды өңдеу барысында UNICODE стандарттарына сәйкес кейбір символдарды кодтау мүмкін еместігін айтады. Тағы бір мәселе – сөздерді жазудың фонетикалық принципі, яғни жазба ескерткіштерде сөздер қалай айтылса, солай жазылған. Бір жағынан, бұл сол кезеңдегі қоғамның сөйлеу ерекшелігін көрсетеді. Бұл, әсіресе, лингвистер үшін маңызды. Ал бағдарламашылар үшін белгілі бір сөздің бірнеше варианттылығы лексикалық бірліктің танымдылығын көрсетуде қиындық туғызады (Бембеев, 2012).

Кейбір еуропа тілдерінің жазуы латын графикасында жазылғандықтан, тарихи корпустарының лингвистикалық белгіленімдері қиындықсыз әзірленген болса керек. Мәселен, португал тілінің тарихи корпусы 88 мәтіні лингвистикалық аңдатапасының жүйесі екі кезеңнен тұрады: 58 мәтіндегі сөздерге сөз табының белгісөзі қойылған, 27 мәтінге синтаксистік белгісөзі қойылған (Corpus of Historical Portuguese). Көне грек корпусы 820 цифрландырылған көне грек мәтіндерінің базасынан

тұрады. Корпус мәтіндеріндегі әрбір сөзі морфологиялық ақпараттармен қамтылған (Vatri A., McGillivray B., 2018). Ескі венгер корпусы XII-XVII ғасыр аралығындағы мәтіндерді қамтиды. Мамандар мәтіндерді өңдеуде екі принципті басшылыққа алған: 1) қазіргі венгер тілінде қолданылмайтын барлық сөздерді морфологиялық ерекшеліктерімен сақтаған; 2) фонологиялық және орфографиялық ерекшеліктері бар барлық сөздерді алып тастаған. Яғни, әртүрлі нұсқада жазылатын сөздер қазіргі венгер тіліне жақындатылып, біріздендірілген. Бір мағынаны білдіретін екі түрлі жазылатын сөздер кездескен жағдайда «Венгер тілінің тарихи-этимологиялық сөздігіне» сүйенген. Мәселен, «бірақ» деген мағынаны білдіретін *кедиг* пен *педиг* шылаулары екі түрлі сөз ретінде жазылған (Симон, 2014).

Ал қазақ тілінің тарихи ішкорпусы мәтіндеріндегі жекелеген лексемаларға лингвистикалық белгіленімдер мен іздеу жүйесі қалай әзірленеді?

Орта ғасыр жазба ескерткіштері араб графикасымен жазылғанын ескерсек, арабграфикалы мәтіндерге лингвистикалық белгіленімдердің қойылуы өте қиын. Себебі араб графикасымен жазылған мәтіндердің қолжазбасы әртүрлі каллиграфиямен берілетіндіктен, іздеу жүйесінде танылa 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. Сондықтан жазба ескерткіштер мәтіндеріндегі кез келген сөз туралы лингвистикалық ақпарат көрсететін конкорданстыжасау қиындық тудыруы мүмкін. Корпусқа енген арабграфикалы жазба ескерткіштер мәтіндерін компьютерде термейінше, скан (көшірме) нұсқасына лексикалық немесе грамматикалық белгіленім қою мүмкін емес. Сондықтан алғашқы уақытта араб графикасынан кирилл графикасына траскрипцияланған нұсқаларына лингвистикалық белгіленім қоюға болады. Алайда араб әліпбиіндегі дауысты дыбыстардың үш-ақ түрі болғандықтан, түркі тілдеріндегі барлық дауыстыларды үш таңбаға сыйдыру мүмкін бола бермейді. Арабграфикалы жазба ескерткіштер мәтіндерін транскрипциялауда белгілі бір лингвистикалық нормаға бағыну қажеттігін В.В.Радлов, В.Л.Васильев және К.Г. Залеман, И.Ю. Крачковскийдің ұсынған кирилл түркологиялық әліпбиінен бастау алады (Сердюченко, 1967). Десек те, татар, өзбек т.б. түркітанушы ғалымдар ескерткіштер мәтіндерін кирилл таңбасымен бергенде өз тілдеріне жақындатып транскрипциялайды. Мәселен, өзбек ғалымы Э.Фазылов Хорезм мәтіндерін отандық түркологтардың белгілер жүйесі пайдаланыланады, алайда өз тарапынан аздап толықтырулар жасап транскрипциялайды (Фазылов, 1971). С.Н. Иванов Әбілғазының «Түркі шежіресін» транскрипциялағанда академиялық (радловская) транскрипцияны пайдаланады (Иванов, 1969). Аталмыш жазба ескерткішті Б.Әбілқасымов араб тіліндегі дауысты дыбыстарды қазіргі қазақ тілінің дыбыстарына салып транскрипциялайды (Әбілқасымов, 2001). Р. Сыздық та Йасауи қолжазбасына транскрипциялауда қазақ тілінің фонетикалық жүйесі мен дыбыстық заңдарына (мысалы, ы-і, а-ә, ұ-ү қатарларын айырып, дауыстылардың жуан-жіңішкелігін, яғни дауыстылар үндестігін сақтап көрсету) сәйкестендіре транскрипциялайды (Сыздықова, 2004).

Ғалымдардың транскрипциялаған әр ғасырдың мәтіндерінен үзінді келтірейік. XI ғасырдағы Ахмет Иассауидің «Диуани хикметі» Самарқанд (Залеман) қолжазбасының транскрипцияланған нұсқасынан үзінді:

Күлбари хали / йігірмә бешкә йетмәй
Mahy сали/ насихатлар қылар пир-у жууанны/
Өзі фәһм етмәгән йахшы йаманны/
Қуруб дамин кәззаблар кәни би пир/ тілі
Мәкр-у хилә қылғаны тасхир/ аны тәзуир
Дүр шайтандын а‘лә/ қопр йүзі қара

XII ғасырдағы Ахмет Йүгнекидің «Хабатул хақайық» шығармасынан үзінді:

Илаһи өкүш хамд айурмен сеңа
Сениң рахматиндин умармен өң-а
Сана му айуғай бу тилим
Унарча айайын йары бер меңа
Сениң барлығыңға тануғлуқ берүр
Жумат жануар учған йүгүрген нең-а

Сениң бирлығыңға далил арқаған
Булур биър нең ичра далиллар мың-а.

XIV ғасырдағы Сейф Сараидің «Гүлістан бит-түрки» шығармасынан үзінді:

Алтун қанатын ачты есä субх шунқары
Көк көлгä батты жумлä кävакиб кäбутäри.
Äсфändäри Рум чықып тиксä 'äлäм
Болуп көруп хäзимät аны Шам läшкäри.
Төкти қырмызы қан, оқытып чирх тасына
Туннуң йурäк тамырын ачып, субх шиштäри
Қушлар көруп хавада аны башлады räван

XVII ғасырдағы Әбілғазы bahадүр ханның «Түркі шежіресі» шығармасынан үзінді: *Артұқ бөке бин Толы хан нәсіліндеи ерді хан көтерділер Ғали ойратке Бағдад хәкімі ерді Мосы бин Ғали бин Байду бин Тарағай бин Һалаку ханны хан қылдылар Арба хан бірлән ұрұшыб ғалыб келді тақы Арба ханны өлтүрді мемлекетіні тахт тәсарафға кіркүзді ол уақытда Шейх Хасан Жалайрке аңа Шейх Хасан Бузруг дерлер Румның хәкімі ерді ол Мосы болғанын ешітіб Махмұд бин йол құтлұқ бин Тимур бин Анбаржу бин Менку бин һалаку ханны хан қылыб Румдын уа Кәржиден ләшкері ғазім жамиғ қылыб Иран тарафыға мәтаужаһ болды.*

Берілген үзінділерден көріп отырғанымыздай, авторлар бірқатар сөздерді әртүрлі транскрипциялаған. Мәселен, XI ғасырдағы Ахмет Иассауидің «Диуани хикметінде» қыпшақ тіліне тән *ы-і, а-ə, ұ-ү* дыбыстарының үндестігі барлық уақытта сақталмаған. XII ғасырдағы Ахмет Йүгнекидің «Хабатул хақайық» мәтіні *ы-і, а-ə, ұ-ү* дыбыстар үндестігімен транскрипцияланбаған. XIV ғасырдағы Сейф Сараидің «Гүлістан бит-түрки» мәтінінде сөздердің жіңішкелік үндестігін сақтау үшін *ä ö* сияқты умлаут белгісі қолданылған. «Жамиғ-ат тауарихта» мәтінінде و графемасы брефиспен ў, ا графемасы макронмен ā таңбасымен транскрипцияланған. Мысалы, اولایت – ўилāйат, т.б. Араб сөздеріндегі ع (айын) дыбысы кейбір мәтіндерде апострофпен ‘ , кебйр мәтіндерде *ғ* таңбасымен транскрипцияланған. Мысалы, *та‘ ала – тағала, ‘ақл – ғақыл,* т.б.

Қарахандықтар және Алтын Орда дәуірі жазба ескерткіштері мәтіндері өзбек әдеби тілінің дыбыстық ерекшелігіне сәйкес транскрипциялау басымдық алған. XVII ғасырдағы Әбілғазы bahадүр ханның «Түркі шежіресі» транскрипциялауда өзге шығармаларға қарағанда дыбыстардың жуан-жіңішкелігі ескерілген. Осыған байланысты жазба ескерткіштер мәтіндерінде бір мағынадағы сөздің әртүрлі транскрипцияланған варианттары пайда болған. Оларды графикалық, фонетикалық, грамматикалық варианттар деп бөлуге болады. Бұндай ерекшелік орыс тілінің корпусында да кездеседі: «Орфографилық нормасы мен оның варианттары: *естли* {если*}, *зделать*{сделать*}, *доволно*{довольно*}» (Савчук, 2008).

1-Кесте. – Жазба ескерткіштер мәтіндеріндегі вариант сөздер

Орфографиялық вариант		**Фонетикалық вариант**		**Грамматикалық вариант**	
норма	**вариант**	**норма**	**вариант**	**норма**	**вариант**
бірлән	бірлə/білән	Чыңғыз	Шыңғыз	екінің	Екі нің
йол	иол	алтұн/алтун	алтын	бірісі	бірісі
барұб	баруп/барұп	деб	теб	һәр түрлү	һәр түрлү
негоһ	неках/некох	дағы	тағы	сусулықдан	сусулық дан
бесиор	бесиар	мұндағ	мұндақ	хабарлы	хабарлы

маңа/баңа	меңа	тірік	тірі	ачлықдан	ач лық дан
саңа	сеңа	құтлығ	құтты	қамашлар	қамаш лар
аңа	анга	уа	ва	hауаға	hауа ға

Орфографиялық варианттар араб тіліндегі үш созылыңқы және үш қысқа (фатха, кесра, дамма) дауысты дыбыстарын кирилл графикасымен транскрипциялауда дауыстыларды дәлме-дәл бере алмайды. Сол сияқты ج (ж), ك (к), ب (б) дауыссыздары жазба ескерткіштер мәтіндерінде екі-үш дыбыстың функциясын атқарған. Мәселен, ج (ж) әрпі бірде *ж*, бірде *ч,* бірде *ш* дыбыстарының функциясын, ك (к) әрпі бірде *к*, бірде *г* және *ғ* дыбыстарының функциясын, ب (б) әрпі сөздің соңғы позициясында бірде *б*, бірде *п* графемаларымен таңбаланған. Сондай-ақ, орфографиялық варианттар түрін көбейтетін кірме араб, парсы сөздерін әртүрлі нұсқада транскрипциялауда көрініс табады: «көп» сөзінің синонимі ретінде парсы сөзі бірде *бесиор*, бірде *бесиар,* бірде *бисйар* деп транскрипциялана берген. Орта ғасыр жазба ескерткіштерінің тілі аралас тілде жазылғандықтан, қыпшақтық, қарлұқтық, оғыздық белгілері мен архаикалық элементтеріне қарай фонетикалық варианттар да пайда болған. Кейбір мәтіндерінде қосымшалар түбір сөзден бөлек жазылған. Оның себебі араб графикасында өзінен жалғастырып жазбайтын жеті таңба бар. Сондықтан болса керек, мәтінде қалай жазылса, солай транскрипцияланған. Осындай тұрақты орфографиялық норма болмағандықтан, траскрипциялауда да ала-құлалық тудырған. Сондықтан ортағасыр транскрипцияланған мәтіндердің танымдылығын арттыруда лингвистикалық белгіленімдерінің графикалық, орфографилық, фонетикалық, орфографиялық варианттарды ескеруіміз қажет. Яғни, олардың түрлі варианттарын корпустың лингвистикалық белгіленім жүйесіне енгізуіміз қажет. Осылайша корпуста лексика-грамматикалық іздеу жүйесінде варианттардың барлық түрі беріледі. Сондай-ақ, әртүрлі нұсқада транскрипцияланған сөздердің қай варианты басым екендігін анықтауды қажет етеді. Лингвостатистикалық талдау нәтижесінде ең көп кездесетін орфографиялық варианттар – *а* мен *ә, ы* мен *і, у* мен *ұ* дыбыстарды таңбалайтын графемалар аралығында. Осындай транскрипциялаудағы ала-құлалық, әсіресе, қарахандықтар мен Алтын Орда дәуіріндегі жазба ескерткіштер мәтіндерінде көбірек кездеседі.

Сонымен араб жазулы жазба ескерткіштеріне лингвистикалық белгіленім жүйесін арабграфикалық мәтіндеріне емес, транскрипцияланған мәтіндеріне қойылады. Лингвистикалық белгіленім лексикалық және грамматикалық белгіленімдерін қамтиды. Корпусқа енетін жазба ескерткіштер мәтіндеріне лексикалық және грамматикалық белгіленім мынадай ретте көрініс табады:

Сөзтізбе. Транскрипцияланған мәтіндегі әрбір лексикалық бірлік сөзтізбеге алынады. Онда сөздің негізгі нұсқасы беріледі.

Вариант. Мәтіндегі бір мағынадағы әрбір сөздің түрлі нұсқадағы көрінісі. Сөзтізбедегі орфографиялық, фонетикалық варианттары беріледі. Сөздердің варианттары неғұрлым көп берілсе, соғұрлым іздеу жүйесінде танылуына мүмкін болады.

Лемма. Сөздің бастапқы тұлғасы, түбірі беріледі.

Лексикалық қабаты. Лексикалық бірліктің этимоны беріледі.

Мағынасы. Лексикалық бірліктің мәнмәтіндегі мағынасы беріледі.

Сөз табы. Лексикалық бірліктің сөз табына қатыстылығы анықталады.

Тұлғасы. Негізгі және туынды тұлғасы анықталады.

Қосымша. Көптік, тәуелдік, септік, жіктік жалғаулары мен сөзтудырушы жұрнақтары анықталады.

Арабграфикалы мәтіндерге арналған лингвистикалық белгіленімдер бағдарламасын жұмыс істеу механизмдерін қалай әзірлейміз? Тегтер дегеніміз – белгілі бір мәтіннің лексикалық-грамматикалық аннотациясын автоматты түрде жүргізуге қабілетті бағдарлама. Бағдарламаның көмегімен мәтіндегі әрбір грамматикалық дұрыс сөзге оның бастапқы формасы (лемма), сөйлеу бөлігі және грамматикалық көрсеткіштер жатады. (мысалы, септіктер, сөздің жеке немесе көпше түрі, етіс, шақ, етістік категориялары). лексикалық-грамматикалық аннотациясын автоматты түрде енгізудің мынадай бағдарламалар қолданылады:

- Google кестесі;
- Python бағдарламалау тілі;
- Django Web сервер бағдарламасы;
- MySQL мәліметтер қорын басқару жүйесі.

Google кесте онлайн бағдарламасы – мәтіндерді жиңактауға және оларды теңестіруге ыңғайлы құрал. Бұндай жиналған мәтіндер Google кесте файлынан тікелей MySQL мәліметтер қорын басқару жүйесіне жүктеуге де жеңіл. Сонымен қатар Notepad++ мәтіндік процессор бағдарламасы қолданылады. Бұл бағдарлама тарихи мәтіндерді сұрыптауға және статистикасын алуға көмектеседі. Осылайша тарихи ішкорпусты интернетке жариялау үшін Django вебсервері қоланылады. Бұл сервер қолданушының сұрауларын қамтамасыз етеді. Серверді MySQL мәліметтер қорын басқару жүйесімен байланыстыратын ортақ Processing.py бағдарламасы жасақталды. Бұл бағдарлама Python тілінде жазылған және тарихи ішкорпусының іздеу функцияларын атқарады. Осылайша ішкорпустың әр сөзіне лингвистикалық паспорт әзірленді. Оған арнайы паспорт форматы мен бағдарламасы жасақталды. Ішкорпустың лингвистикалық паспортының сұлбасы 1-суретте көрсетілген.

2-Кесте. – Тарихи ішкорпустың лингвистикалық паспорты

Сөзтізбе	**Сөз**	hәр	тұрұрлар
	Лемма	hәр	тұр
	Вариант	хәр	турурлар
Лексикалық белгіленім	**Лексикалық қабаты**	парсы	төл
	Мағынасы	әр	тұр
Грамматикалық белгіленім	**Сөз табы**	есімдік	етістік
	Тұлғасы	түбір	негізгі
	Қосымша	-	-лар: көптік жалғау

Паспортта әр тарихи сөздің *сөз, лемма, вариант, лексикалық қабаты, мағынасы, сөз табы, тұлғасы, қосымша* сипаттары (разметка) көрсетіледі. Бұл сипаттардың бәрі алдың ала анализатордан өткізіліп, мәліметтер қорын басқару жүйесіне де жылдам іздеуге ыңғайлы етіп сақталады. Жалпы ішкорпустың бағдарламалық құрылымы 1-суретте көрсетілген.

1-Сурет. – Тарихи ішкорпустың бағдарламалық құрылымы

Тарихи ішкорпусының бағдарламалық архитектурасы client-server және MVC (Model-View-Controller) технологияларын пайдаланы отырып, әзірленеді. Мұндағы Model – тарихи ішкорпустың мәліметтер қорын жинақтайтын жүйе. View – пайдаланушыға жұмыс істеуге арналған интерфейс сұлбасы. Controller іздеу функцияларын орындауға және View мен Model-ді өзара біріктіруге қызмет етеді.

Қорытынды

Қорыта келгенде, тарихи ішкорпусқа араб графикасынан кирилл графикасына транскрипцияланған мәтіндерге лингвистикалық метабелгіленімдерді енгізу үшін төмендегідей зерттеулер жасалды:

- корпусқа еніп отырған XI–XVII ғғ. жазба ескерткіштері мәтіндерінің араб графикасынан кирилл графикасына транскрипцияланған мәтіндерді саралағанымызда, бір кезеңге жататын мәтіндердің өздері әртүрлі транскрипцияланған. Олардың ішінде Қарахандықтар және Алтын Орда дәуірі жазба ескерткіштері мәтіндері өзбек әдеби тілінің дыбыстық ерекшеліктеріне сәйкес транскрипцияланған;
- транскрипциялауда әртүрлі орфограмманы тудырған дауысты, дауыссыз дыбыстардың негізінде орфографиялық, фонетикалық, грамматикалық варианттар анықталды;
- лингвостатистикалық талдау нәтижесінде ең көп кездесетін орфографиялық варианттар – *а* мен *ә, ы* мен *і, у* мен *ұ* дыбыстарды таңбалайтын графемалар;
- транскрипцияланған мәтіндер бойынша лексика-грамматикалық белгіленімдер жүйесі әзірленді;
- транскрипцияланған мәтіндерге арналған лингвистикалық белгіленімдер бағдарламасының жұмыс істеу механизмдері сипатталды.

Сонымен тарихи ішкорпусқа енетін мәтіндерге лексикалық, семантикалық белгіленімдердің қоюлуы, талданған мәтіндердің көпшілікке қолжетімді болуы – диахрондық лингвистиканы сапалы зерттеуге мүмкіндік береді.

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.

Әдебиеттер

Гаврилова Т. С., Шалганова Т. А., Ляшевская О. Н. (2016) К задаче автоматической лексико-грамматической разметки старорусского корпуса XV-XVII вв // Вестник ПСТГУ. Серия III: Филология. 2016. Вып. 2 (47). С. 7 – 25.

Бембеев Е.В. (2012) Коллекции рукописей на старокалмыцком (ойратском) языке XVII–XIX вв. в свете компьютерной обработки: постановка проблемы // Информационные технологии и письменное наследие. El'Manuscript-2012: Материалы IV международной научной конференции (Петрозаводск, 3–8 сентября 2012 г.). Петрозаводск, Ижевск, 2012. С. 31–34.

Corpus of Historical Portuguese. https://www.clarin.eu/resource-families/historical-corpora. https://www.clarin.eu/resource-families/historical-corpora.

Vatri A, McGillivray B. (2018) The Diorisis Ancient Greek Corpus Linguistics and Literature. Research Data Journal for the Humanities and Social Sciences. Издатель: Brill E-ISSN:2452-3666. 2018. page 55–65 /https://brill.com/view/journals/rdj/3/1/article-p55_55.xml

Симон Э. (2014) Здание корпуса из древневенгерских кодексов. In: Каталин Э. Поцелуй (ред.): Эволюция функциональной левой периферии в венгерском синтаксисе. Оксфорд: Издательство Оксфордского университета, 2014 г.

Сердюченко Г. П. !967) Русская транскрипция для языков зарубежного востока. – Москва: Наука, 1967. С 359.

Фазылов Э.И. (1971) Староузбекский язык. Хорезмские памятники XIV века. – Т.2. Ташкент: Фан, 1971. – 778 с.

Иванов С. Н. (1969) Родословное древо тюрок Абу-л-Гази-хана. – Ташкент: изд.«Фан» Узбекской ССР, 1969 г. С 202.

Әбілқасымов Ә. (2001) Әбілғазы ханның «Түркі шежіресі» және оның тілі. – Алматы: Арыс, 2001. – 246 б.

Сыздықова Р. (2004) Ясауи «Хикметтерінің» тілі. – Алматы, Сөздік-Словарь, 2004. – 552 б.

Савчук С.О. (2008) Корпус текстов XVIII века в составе национального корпуса русского языка: проблемы и перспективы. 25 июнь 2008 / http://textualheritage.org/lt/el-manusctipt-08-/xviii.html

References

Gavrilova T. S., Shalganova T. A., Lyashevskaya O. N. (2016) K zadache avtomaticheskoy leksiko-grammaticheskoy razmetki starorusskogo korpusa XV-XVII vv [On the problem of automatic lexico-grammatical marking of the Old Russian corpus of the XV-XVII centuries] // Vestnik PSTGU. Seriya III: Filologiya. 2016. Vyp. 2 (47). S. 7 – 25. [in Russian]

Bembeyev Ye.V. (2012) Kollektsii rukopisey na starokalmytskom (oyratskom) yazyke XVII–XIX vv. v svete komp'yuternoy obrabotki: postanovka problemy [Collections of manuscripts in the Old Kalmyk (Oirat) language of the 17th–19th centuries. in the light of computer processing: problem statement] // Informatsionnyye tekhnologii i pis'mennoye naslediye. El'Manuscript-2012: Materialy IV mezhdunarodnoy nauchnoy konferentsii (Petrozavodsk, 3–8 sentyabrya 2012 g.). Petrozavodsk, Izhevsk, 2012. S. 31–34. [in Russian]

Corpus of Historical Portuguese. https://www.clarin.eu/resource-families/historical-corpora. [in English]

Vatri A, McGillivray B. (2018)The Diorisis Ancient Greek Corpus Linguistics and Literature. Research Data Journal for the Humanities and Social Sciences. Издатель:Brill E-ISSN:2452-3666. 2018. page 55–65 /https://brill.com/view/journals/rdj/3/1/article-p55_55. Xml [in English]

Ester Simon. (2014) Zdaniye korpusa iz drevnevengerskikh kodeksov [The building of the building from the ancient Hungarian codes] In: Katalin E. Potseluy (red.): Evolyutsiya funktsional'noy levoy periferii v vengerskom sintaksise. Oksford: Izdatel'stvo Oksfordskogo universiteta, 2014. [in English]

Serdyuchenko G.P. (1967) Russkaya transkriptsiya dlya yazykov zarubezhnogo vostoka [Russian transcription for languages of the foreign east] – Moskva: Nauka, 1967. S 359. [in English]

Fazylov E.I. (1971) Staryy uzbekskiy yazyk. Pamyatniki Khorezma v 14 veke. [Old Uzbek language. Khorezm monuments of the 14th century] - T.2. Tashkent: Fan, 1971. - 778 s. [in Russian]

Ivanov S. N. (1969) Genealogicheskoye drevo tyurka Abu-l-Gazi-khana [Family tree of the Turks Abu-l-Ghazi Khan] - Tashkent: izd-vo "Fan" Uzbekskoy SSR, 1969. S 202. [in Russian]

Äbilqasımov Ä. (2001) Äbilğazı xannıñ «Türki şejiresi» jäne onıñ tili. – Almatı: Arıs, 2001. – 246 b. [in Kazakh]

Sızdıqova R. (2004) Yasawï «Xïkmetteriniñ» tili [Yasawi "Hikmetterinin" tili] – Almatı, Sözdik-Slovar, 2004. – 552 b. [in Kazakh]

Savchuk S.O. (2008) Korpus tekstov XVIII veka kak chast' natsional'nogo korpusa russkogo yazyka: problemy i perspektivy. 25 iyunya 2008 g. [Corpus of texts of the 18th century as part of the national corpus of the Russian language: problems and prospects] // http://textualheritage.org/lt/el-manusctipt-08-/xviii. Html. [in Russian]

