МРНТИ 16.31.02

А.Қ.Жұбанов

А. Байтұрсынұлы атындағы Тіл білімі институтының бас ғылыми қызметкері,
филология ғылымдарының докторы, профессор
Алматы қаласы, Қазақстан

ҚАЗАҚ ТІЛІНІҢ ҰЛТТЫҚ КОРПУСЫ ЖӘНЕ МЕТАБЕЛГІЛЕНІМ МӘСЕЛЕСІ

Аннотация. Мақала оқырманды корпустық лингвистиканың негізгі ұғымдарымен, терминдерімен және корпусты жобалау, тілдік материалдарды іріктеу мен өңдеу, мәтінге метабелгіленім енгізу тәсілдерінің технологиялық процестерімен жалпылама түрде таныстырады. Сонымен бірге мәтіндер корпусынан зерттеушіге қажетті тілдік бірліктерді іздептабу үдерісін қамтамасыз ететін корпустық менеджер жұмысы да қысқаша сипатталады. Мақала иесі оқырманды корпустық лингвистика пәнінің концепциясымен, корпустық технологиялардың негізімен және пәннің басқа да ақпараттық технологиялар ішінен алатын орнымен таныстыруды мақсат етеді.

Мақалада тілдік бірліктер жайында әртүрлі анықтамалық және статистикалық деректер алу мүмкіндіктері қысқаша сипатталады. Мәселен, корпустар негізінде сөзформалардың, лексемалардың, грамматикалық категориялардың қолдану жиілігі жайында және олардың кезеңдік өзгеріске ұшырау сипаты немесе лексикалық және т.б. бірліктердің тіркесе қолдану жиіліктері жайындағы деректерді алу мүмкіндіктері қарастырылады. Сонымен бірге тілдік бірліктерді белгілі бір кезеңдерге қатысты сипаттайтын деректер тілдік қордың динамикасын зерттеуге және әр жанр мен әр авторға қатысты мәтіндер бойынша тілдің лексика-грамматикалық сипатын талдауға болатындығы сөз болады.

Мақалада тіл білімінің синтаксис, семантика және әлеуметтік лингвистика сияқты салалары тілдік сипаттауды немесе тілдік құрылымды бағалауды, немесе тілдік қолданысты зерттеуді мақсат етсе, корпустық лингвистика тілдік зерттеудің көптеген аспектілеріне қолдануға болатындығы қарастырылады.

Тірек сөздер: корпустық лингвистика, мәтіндер корпусы, ұлттық корпус, морфологиялық белгіленім, синтаксистік белгіленім, семантикалық белгіленім, метабелгіленім, сөзформа жиілігі, грамматикалық категориялар жиілігі, лексика- грамматикалық сипаттама, контекст, кезең, жанр.

А.К.Жубанов

главный научный сотрудник Института языкознания имени
А. Байтурсынулы, доктор филологических наук, профессор
Алматы, Казахстан

НАЦИОНАЛЬНЫЙ КОРПУС КАЗАХСКОГО ЯЗЫКА И ПРОБЛЕМЫ МЕТАРАЗМЕТКИ

Аннотация. Статья знакомит читателей с основными понятиями и терминами корпусной лингвистики, а также описывает в общих чертах технологические процессы, связанные с их проектированием, отбором и обработкой языкового ма- териала, способами

метаразметки. Также кратко описываются корпусные менеджеры, обеспечивающие поиск в корпусе необходимых для пользователей лингвистических единиц. Автор статьи ставит цель познакомить читателей с концепциями корпусной лингвистики с тем, чтобы помочь им освоить основы корпусных технологий, приобрести навыки работы с корпусами, определить место дисциплины и собственно корпусов в ряду информационных технологий.

В статье описываются возможности использования корпусов для получения разнообразных справок и статистических данных о языковых и речевых единицах. В частности, на основе корпусов можно получить данные о частоте словоформ, лексем, грамматических категорий, проследить изменение частот и контекстов в различные периоды времени, получить данные о совместной встречаемости лексических единиц и т.д. Также очевидно, что представительный массив языковых данных за определенный период позволяет изучать динамику процессов изменения лексического состава языка, проводить анализ лексико-грамматических характеристик в разных жанрах и у разных авторов.

В статье рассматриваются такие области лингвистики, как синтаксис, семантика и социолингвистика, которые могут быть применены ко многим аспектам лингвистического исследования, если целью является изучение языкового описания или оценки языковой структуры, или языкового употребления.

Ключевые слова: корпусная лингвистика, корпус текстов, национальный корпус, морфологическая разметка, синтаксическая разметка, семантическая разметка, метаразметка, частота словоформ, частота грамматических категорий, лексико-грамматическая характеристика, контекст, период, жанр.

A.K.Zhubanov

Chief researcher of the Institute of Linguistics named after A. Baitursynuly,
Doctor of Philological sciences, professor
Almaty, Kazakhstan

NATIONAL CORPUS OF THE KAZAKH LANGUAGE AND METAMARKING PROBLEMS

Annotation. This article discusses the basic concepts and terminology of Corpus Linguistics, as well as outlines the technological processes related to their design, the selection and processing of linguistic material, the metamarkup ways. The brief descrip- tion is given of corpus managers providing the search in the corpus of linguistic units necessary for users. The author of the articles tries to acquaint readers with the concepts of Corpus Linguistics, so that they can learn the basics of corpus technologies, acquire skills to work with the corpus, determine the place of the branch of science and corpuses in the series of information technologies.

The article describes the possibilities of using corpuses for receiving a variety of information and statistic data on language and speech units. In particular, on the basis of corpuses, data on the frequency of word forms, tokens, grammatical categories can be obtained, changes in the frequency and contexts at different times can be traced, data on co-occurrence of lexical units, etc. can be received. It is also evident that a representative array of language data for a specific period allows us to study the dynamics of the proces- ses of changing of the lexical structure of the language, to conduct the analysis of lexical and grammatical characteristics of different genres and different authors.

Keywords: Corpus Llinguistics, corpus, national corpus, morphological markup, syntax markup, semantic markup, metamarkup, frequency of word forms, frequency of tokens, frequency of grammatical categories, lexical-grammatic characteristics, context, time period, genres.

Тіл білімінің жаңа саласы «Корпустық лингвистиканың» мән-мағынасына тоқталатын болсақ, оның «Компьютерлік лингвистика» деп аталатын саланың бір бөлімі екендігін айта кеткен жөн. Корпустық лингвистика – компьютерлік технологияларды қолдана отырып, лингвистикалық корпустарды (мәтіндер корпустарын) құру мен оны пайдалану-дың жалпы приципгерін зерттемелейді. Ал **лингвистикалық** немесе **тілдік, мәтіндер корпусы** дегеніміз нақты тілдік мәселелердің шешімін табуға арналған аса үлкен көлемдегі мәшине (компьютер) оқи алатындай түрде көрініс табатын, бірыңғайланған, құрылымдалған, белгіленген (шартты белгілер қойылған), филологиялық тұрғыда компетентті саналатын тілдік деректер ауқымы.

Кез келген корпусты жобалау оны құру кезеңдері мен әрі қарайғы даму жолдарын алдын ала ескеруді қажет етеді. Корпус ұғымы тілшілердің үнемі қолданысында болған дәстүрлі картотека қорын жасау жұмысының жалғасы деуге болады. XX ғасырда бұл картотекалар компьютер жадына енгізіліп, көпшіліктің пайдалануына мүмкіндік жасалды. Корпустық тіл білімінің қалыптасуына Интернет желісі айтарлықтай қызмет атқарды десек те болады. Интернеттің дамуы арқасында үлкен көлемді мәтіндік материалдарға қолжетімді болып, әртүрлі лингвистикалық зерттеулер жүргізуге мүмкіндік туды. Осындай жағдайда, сөздіктер мен грамматикалар құрастыруда негіз болатын дәстүрлік мәселе – тілдік материалдардың репрезентативтілігі мен теңгерімдігі. Корпустық лингвистиканың жетістіктерін өзіне сақтаған аса дамыған корпус түрі – Ұлттық корпус. Аталған корпус Ұлттық тілді белгілі дәрежеде толық түрде бейнелейді. Ұлттық корпустың репрезентативтілігі (тұлғалылығы) – сол тілдің жазба және сөйлеу түріндегі мәтіндерінің барлық типтерінің қамтылуы. Ұлттық корпустың айтарлықтай дәрежеде көлемді (ондаған, жүздеген миллион сөзқолданыс) болуы репрезентативтілікке жету-дің қажетті шарты болып саналады. Ұлттық корпустың ажыратылмас бөлігі оның белгіленген (аннотацияланған, мазмұндалған) бейнесі.

Қолданбалы тіл білімінің өкілі В.В. Рыков мәтіндер корпусын мәтіндер жиынтығы ретінде есептей келе, оны түпкі негізінде логикалық ой мен логикалық идея жатқан және ондай мәтіндерді корпусқа біріктіріп тұрған ережелер мен корпус мәтіндерін талдауға арналған алгоритмдер мен программалар идеологиясы және әдістемелері деп анықтама береді [1].

Заманауи корпустық лингвистикаға әсер еткен негізгі бағыт салыстырмалы-тарихи тіл білімінен келді. Бұған таңғалуға да болмайды, себебі тарихи зерттеулермен айналысатын тілшілер әр уақытта мәтіндерді немесе мәтіндер жинағын негізгі айғақ ретінде санайды. Аса ертедегі тілдерді реконструкциялау үшін немесе тіларалық байланысты орнату үшін XIX ғасырда дамыған көптеген технологиялар осы кезге дейін қолданыс тауып жүр. Үндіеуропалық дәстүр бойынша тілдік өзгерістер мен реконструкциялау талаптары ерте кездегі мәтіндерге немесе кор-пустарға (тарихи ескерткіштерге) байланысты болған. Младограмматиктер диалектілерде орын алатын қазіргі тілге зерттеу жүргізгендері жайлы (ертедегі мәтіндерді зерттегендерінен басқа) өздерінің манифестерінде жария еткен болатын, міне бұл да аса зор маңыздылыққа ие болды.

XIX ғасырдан бастап дамып келе жатқан көптеген идеялар мен технологиялар корпустық лингвистикада қолданыс тауып, сосын дамыған болатын. Тарихи корпустарды құрастыру ісі бұрынғыша зор қызығушылық тудырумен бірге, электрондық пішіндегі қолжетімді алғашқы корпустар ішінде тарихи корпустар екені сала мамандарына мәлім.

Электрондық форматта қолжетімді аса көп мәтіндердің пайда болуы көлемді тілдік деректерді барынша тез арада жинақтап алуға мүмкіндік берді. Ал бұл жағдай тілшілерге лингвистикалық талдауда статистикалық әдіс пен модельдеу әдістерін қолдану мен дамыту арқылы айтарлықтай ұтымдылыққа ие болуға мүмкіндік тудырды.

XIX ғасырдағы грамматистер өздерінің пайымдауларын белгілі авторлардың шығармаларындағы мысалдармен дәлелдеп отырған. Қазіргі кезде сөйлеу тілінің грамматикасына көп көңіл бөлінуде. Тілді грамматикалық деңгейде сипаттауда корпустарды әртүрлі грамматикалық бірліктер варианттарының, регистрлердің және т.б. қолдану жиіліктері жайлы да ақпарат алуға болатындығы мәлім.

Өлі тілдердің көптеген сөздіктерінен қажетті сөздерді кездестіретін контекстерден цитаталар алуға болатын сияқты қазіргі кездегі корпустық лингвистикада компьютер көмегімен қажетті мысалдарды іздеу мен классификациялау және көпсөзді бірліктерді бөліп алу әрекеттері барынша жеңілдеп отыр.

Тілді оқыту немесе үйрету кезінде корпустар студенттердің қызығушылығы мен өз бетімен тілді зерттеп-үйренуге және тілдесу кезінде қолдануға ұмтылысы артады. Корпустық деректерді маңызды қолдану – Computer-Assisted Language Learning (CALL), мұнда корпустың программалық қамтамасыз етуіне негізделуінен студенттердің компьютер көмегімен орындайтын интерактивті оқыту әдісі қолданылады.

ҚР БҒМ ҒК А. Байтұрсынұлы атындағы Тіл білімі институтында қазақ тілінің корпусын құрастыру мәселесі «Мәдени құндылықтар ретіндегі қазақ тіліндегі мәтіндер корпусы және сөздіктердің «Тіл – қазына» атты Ұлттық компьютерлік қоры» атты тақырыпқа қатысты зерттеу жұмыстарынан бастама алған болатын (2009-2011 жж.). Аталған зерттеу жұмысының негізгі мақсаты – қазақ тілінің мәдени құндылығы болып саналатын толық мәтіндеріне, қажеттілікке сай, грамматикалық белгі-кодтар енгізіп, оның дербес түрдегі «Тіл – қазына» атты мәтіндер корпустарының компьютерлік базасын құру. Алғашында (2009-2011 жж.) толық мәтіндердің компьютерлік қорының нысандары ретінде М. Әуезовтің, Ә. Кекілба-евтың, М. Мақатаевтың, М. Мағауиннің және т.б. толық шығармаларынан тек таңдама мәтіндер ғана алынды. Ал басқа қазақ классиктерінің, ғылыми мәтіндердің, публицистикалық шығармалардың мәтіндер корпусын жасау ісі «Қолданбалы лингвистика» бөлімінде өз жалғасын табуда.

Корпустарды құру мен оны пайдаланудың мақсаттылығы келесі алғышарттармен айқындалатындығы мәлім:

1) барынша үлкен көлемді корпус (репрезентативтілік) деректердің типтілігіне кепілдік береді және барлық тілдік құбылыстар шоғырына толықтық сипат қамтамасыз етеді;

2) корпустарда әртүрлі деректер типтері өзінің табиғи контекстік формада сақталуы оларды жан-жақты және объективті зерттеуге мүмкіндік тудырады;

3) бір-ақ рет құрылған және дайындалған деректер қорын талай мәрте және әртүрлі зерттеушілер әртүрлі мақсатпен пайдалануға болады.

«Мәтіндер корпусы» ұғымына мәтіндер мен тілдік деректерді басқару жүйесі де енеді. Соңғы кезде мамандар аталған жүйені ***корпустық менеджер*** (немесе корпус-менеджер) (ағыл. corpus manager) деп жиі атап жүр. Бұл жүйені корпустағы тілдік деректерді іздестіру мен статистикалық ақпараттарды алу үшін және табылған нәтижелерді ыңғайлы пішінде пайдалануға ұсыну мақсатындағы программалық құралдарды қамтитын мамандандырылған іздестіру жүйесі деуге болады.

Мәтіндер корпусындағы кез келген сөзді іздеу, сол сөздің конкордансын құруға мүмкіндік туғызады, яғни сол сөздің барлық қолданысы бар мәтін үзінділері қамтылған шығармалар тізімін және оларға сілтеме берілген мәліметтермен танысуға мүмкіндік туады. Корпустарды жазба және сөйлеу тілдері бірліктері жайлы неше түрлі анықтамалар мен статистикалық мәліметтер алу үшін пайдалануға болады. Мәселен, корпустар негізінде сөзформалар, лексемалар, грамматикалық категориялардың қолдану жиілігі жайында, әртүрлі кезеңдерге қатысты олардың қолдану жиілігінің және контекстердің өзгеріске ұшырауын бақылап отыруға, тілдік бірліктердің бірлесіп қолдануы жайындағы

деректерді және т.б. мәліметтерді алуға болады. Белгілі кезеңдегі тілдік деректердің өкілді жиыны бойынша тілдің лексикалық құрамының өзгеріске ұшырау динамикасын зерттеуге, әртүрлі жанрдың және әр автордың шығармаларындағы лексика-грамматикалық сипатына талдау жүргізуге мәтіндер корпустары зор мүмкіндік береді. Сонымен бірге корпустар дереккөз ретінде де және көпаспектілі лексикографиялық жұмыстар жүргізуде, мысалы, әртүрлі тарихи және заманауи сөздіктерді дайындауда да зор қызмет ете алады. Корпус деректері грамматикаларды құрастыру мен нақтылау және тілді оқыту мақсатында пайдалануда мүмкіндігі мол. Табиғи тілдегі мәтіндерді өңдеуге арналған қазіргі интеллектуалды программалық жүйелердің дамуы да көлемді эксперименталды тілдік қорды қажет етеді.

Егер тіл білімінің синтаксис, семантика және әлеуметтік лингвистика сияқты салалары тілдік сипаттауды немесе тілдік құрылымды бағалауды немесе тілдік қолданысты зерттеуді мақсат етсе, корпустық лингвистика тілдік зерттеудің көптеген аспектілеріне қолдануға болатын әдістемелігімен ерекшелінеді. Кейбір жағдайда корпустық лингвистиканы «әртүлі тілдік зерттеу аясының әдістер шоғыры» деп те атайды [2]. Тілдік талдау әдісі ретінде корпустық лингвистика тілдер ара-сындағы жалпылық және даралық фактілерді анықтауға бағытталған, тілдерді салыстырмалы зерттеу барысындағы диалектілердің немесе тіл варианттарының ерекшелігін анықтайтын контрастивтік зерттеулерге де қатысы бар [3]. Тілдік талдаулардың көптеген түрлері тұрақты және аумақты эмпирикалық деректер қоры негізінде ең ұтымды түрде дамып келеді.

Ғалым Э. Финеган корпустық лингвистикаға «табиғи тілдің қолданысын зерттеуге бағытталған корпусты құрастыру мен қолдануды қажет ететін қызмет түрі» деп анықтама береді [4]. Бұл анықтамада корпустық лингвистиканың жасампаздық бағытына көбірек көңіл бөлінеді. Корпустық лингвистиканың екіжақтылық сипатын (корпусты құру және оны пайдалану) оның ***нысанының*** екіжақтылығымен түсіндіруге болады. Себебі, бірінші-ден, нысан деп отырғанымыз, корпустық лингвистика үшін де және басқа да тілдік пәндер үшін де – сөйлеу материалы, ал, екіншіден, корпустық лингвистика қызметінің де нәтижесі.

Орыс тілі үшін де және түркі тілдері үшін де бүгінгі күні корпустық лингвис-тика пәні терминологиясының қалыптасуына қатысты проблемалық мәселелер туындауда. Оның себебі, негізінен, ол пәннің «дүниеге келуінің» қысқа мерзімдігі болса, екіншіден, оның АҚШ пен Ұлыбританиядан бастау алып, терминологияның ағылшын тілінде қалыптасуы себеп болып отыр. Орыс тілі корпустарының терминдері, негізінен, ағылшын корпусы терминдерінен алынған кірме терминдер, ал олардың кейбіреулері басқа мағынада орыс тілінде бұрыннан да қолданып келе жатқандығы белгілі.

Корпус авторларының міндеті – зерттеушілердің талабына қарай, сол тілдік саладан барынша көп мәтіндер жинау. Корпус – сол тілдің (немесе шағын тілдің) кішірейтілген моделі. ***Репрезентативтік*** ұғымы – барынша жеткілікті және әртүрлі кезеңдердегі мәтіндердің жанрларға, стильдерге, авторларға және т.б. қатысты пропорционалды көрінісі, яғни тілдік зерттеу аясының барлық қасиетін өзінде сақтайтын мәтін көлемі [1]. ***Репрезентативтікке*** анықтама беру әртүрлі амалдарға сүйене беріледі. Мысалы, жалпытілдік (ұлттық) корпустарға қатысты ***репрезентативтік*** ұғымын қатаң математикалық тәсілмен есептеп шығу және оны сипаттау мүмкін емес деген пікір бар, бірақ корпусты жобалау кезінде де және оны пайдалану кезінде де оның шешімін табуға ұмтылу қажет.

Тәжірибеден байқағанымыздай, корпустық лингвистика ең кем дегенде екі түрлі типтегі нысандармен (мәтіндер корпустарымен) әрекет етеді:

1. Бірінші типтегі корпустар әмбебаптық сипатта, олар өзінде тілдік әрекеттің барлық түрлерінің молдық қасиетінің көрінісі.

2. Екінші типтегі корпустар қайсыбір тілдік немесе мәдени феноменнің қоғамдық тілдесу тәжірибесінде «өмір кешетінінің» (бытование) көрінісі, олар ***ad hoc*** (арнайы мақсатпен) құрылған, мысалы, мақалдар корпусы немесе газет тіліндегі саяси метафоралар корпусы [1].

Екі жағдайда да ***репрезентативтік*** тек мәтіндер корпусында зерттеу аясының барлық қасиеттерінің қамтылуына байланысты статистикалық баға беруге қатысты қарастырылған. Солай бола тұра, бағалаудың статистикалық критерийлері бірден-бір айқындаушы белгі бола алмайды, себебі корпус қайсыбір нысан ретінде өзінен сырт ақиқаттықтың моделін бейнелейді. Корпустың ***репре-зентативтігі*** ғана одан алынған нәтижелердің нақтылығының айғағы. Бұл мәселені аса ірі мәтіндер ауқымының немесе сөйлеу әрекетінің басқа да ірі көлемді фрагменттерінің, көлем жағынан барынша кіші, мәтіндер корпусының адекватты көрінісі ретіндегі жеке бір проблемалық мәселе деп қарастыруға болады. Сөйлеу ақиқаттығы айтарлықтай дәрежеде алуан түрлі, яғни түрлі-түрлі фактурамен көрініс табады (ауызша, жазбаша, баспасөз және т.б.) және оларда орын алған тілдік құбылыстардың саны ұшан-теңіз деуге болады. *Бірінші түрге* жататын мәтіндер корпустары 60-жылдары әмбебаптық қасиетке ұмтылыста болғаны мәлім [5]. Мысалы, баспа сөзін бейнелеу үшін 1960 жылы АҚШ-та сол кезеңге сай қанағаттанарлық репрезентативтік дәрежеде Браундық мәтіндер корпусы құрастырылған болатын. 15 жанрды (регистрлерді) бейнелуді қажет ететін 6-дан 80-ге дейінгі қарапайым таңдама мәтіндер іріктеліп алынды:

1) баспасөз;
2) пресса: бас мақала;
3) пресса: шолу;
4) діни мәтіндер;
5) дағды, сабақ, әуестік;
6) ғылыми-көпшілікке арналған әдебиет;
7) беллетристика, өмірбаяндар, эссе;
8) әртүрлі (үкіметтік құжаттар, кәсіпорын есептері, өнеркәсіп есептері, колледждердің каталогтары);
9) ғылыми шығармалар;
10) көркем әдебиет;
11) мистика мен детективтер;
12) ғылыми проза;
13) қызық оқиғалы әдебиет және вестерндер;
14) сүйіспеншілік жайлы романдар;
15) әзіл-оспақ шығармалар.

Екінші типтегі корпустарда репрезентативтілік критерийі ретінде қызықтыратын құбылыстың қолданыста барын барынша объективті елестету болып табылады. Мәселен, ағылшын тілін иеленушінің сөйлеу тәжірибесінде белгілі бір кезеңдегі және географиялық аймақтағы максимал репрезентативтікке ие ағылшын тіліндегі мақалдардың корпусы [1].

Корпус кең түрдегі қолданушы үшін құрылады және әртүрлі мәселелердің шешімін табуға, соның ішінде барынша «экзотикалық», мысалы, шеттілдік графиканы қолданатындардың орыстілдік мәтіндерді зерттеуге арналған корпус және т.б. Бастапқы мәтін ішінен не «қалады», не «алып тасталады» деген сұрақ та дұрыс шешімін табуды қажет етеді. Мәселен, мәтін ішіндегі суреттер тілдік материалдарға жатпайды, сондықтан

оларды алып тастауға болады. Ал мәтін ішіндегі кесте, дәйексөз, төл сөз, шет тілдерден енген сөз, өлшем бірліктеріне келгенде жағдай күрделірек.

Бұл аталған мәселелер корпусты жобалау кезінде қарастырылуы қажет. Кейбір мәселелер біртіндеп, құрастыру мен тәжірибелік пайдалану кезінде де шешілуі мүмкін. Осындай жағдайда корпусты пайдаланудың басынан-ақ пайдаланушы әрекеті алдын ала ескерілуі қажет.

Корпусты құрудың технологиялық процесі мынадай қадамдардан (немесе кезеңдерден) тұрады:

1. Дерекнама тізіміне сәйкес мәтіндердің түсімін қамтамасыз ету.

2. Мәтіндерді машине оқи алатын пішінге келтіру. Корпусты құрастыру үшін мәтіндерді электрондық пішінге келтіру әртүрлі тәсілдермен жүзеге асырылады – қолмен енгізу, сканерлеу, авторлық көшірме, сыйлау және айырбас, Интернет, баспадан түпнұсқа-макет түрінде және т.б. тәсілдер арқылы.

3. Мәтіндерді талдау және алдын ала өңдеу. Бұл кезеңде әртүрлі дерекнамалардан алынған мәтіндер филологиялық тұрғыда тексеру мен түзетуден өтеді. Мәтіндерді технологиялық сипаттауға дайындау оларды библиографиялық және экстралингвистикалық сипаттауды өзіне қосады [4].

4. Мәтін белгіленімі (Разметка текста). Мәтінге белгіленім жүргізу кезінде мәтінге және оның құрамдас бөліктеріне қосымша ақпарат (метадеректер) тіркеліп жазылады. Метадеректерді 3 типке бөлуге болады: барлық мәтінге қатынасы бар экстралингвистикалық дерек; мәтіннің құрылымына қатысты деректер; мәтіннің элементтерін сипаттайтын тілдік метадеректер. Корпус мәтіндерін метасипаттау деректердің мағыналы элементтерімен (библиографиялық деректер, мәтіннің жанрлық және стильдік ерекшеліктерін сипаттау, автор жайлы мәліметтер) бірге деректердің формалды (файлдың аты, кодтау параметрлері, белгіленім тілінің нұсқасы, жұмыс кезеңдерін орындаушылар) элементтері. Бұл мәліметтер әдетте қол жұмысы арқылы орындалады да, ал құжаттың құрылымына қатысты белгіленім (абзацты, сөйлемді, сөздерді бөліп алу) және тілдің өзіне тікелей қатысты белгіленім әдетте автоматты түрде жүзеге асады.

5. Автоматты белгіленім нәтижелерін түзету: қателерді түзету және бірізділікке келтіру (қолмен немесе жартылай автоматты түрде жүзеге асады).

6. Белгіленген мәтіндерді мамандандырылған ақпараттық-іздеу жүйесімен (corpus manager) айырбастау арқылы көпаспектілі іздеу мен статистикалық өңдеу тез арада қамтамасыз етіледі (қорытынды кезең).

7. Корпусқа қолжетімділікті қамтамасыз ету. Корпус дисплейлік класс аясында, компакт-диск бойынша және ғаламтор желісі арқылы таралуы мүмкін. Корпусты пайдаланушылардың әртүрлі категорияларына қарай, олар әртүрлі құқық пен мүмкіндікке ие болады.

8. Корпусты құру мен оны пайдалану жайлы әртүрлі аспектілер сипатталатын құжаттық қамтамасыз етуді жүзеге асыру. Мұндай құжатпен қамтамасыз ету дегеніміз метадеректер негізінде сауалдар-тілі бойынша корпус-менеджерді және т.б. іздеп табуға мүмкіндік туады.

Көлемді корпустар үшін автоматтанған синтаксистік талдағыштарды (анализаторларды, парсерлерді) құру компьютерлік лингвистиканың ең бір маңызды саласы болып табылады. Көптеген тәсілдер өлшеудің сапалық және сандық жақтарын қоса қарастырады. Синтаксистік шежірелерге (tree-banks) қолмен енгізілген белгілер арқылы жаттығатын әртүрлі статистикалық әдістермен бірге көптеген синтаксистік талдағыштар ережелерге немесе шектеулі тәсілдерге негізделген тәсілдерді қолдана отырып, ерекше тілдік теорияны бірден модельдейді. Мұндай синтаксистік талдағыштарды зерттеу осы

теориялардың дамуымен ілісіп-шатасып жатыр. Кез келген теорияда көптеген ұсыныстар біркелкі (бірізді) болмауынан, ережелер негізінде (немесе шектеулер тізімі) әркелкілік жағдайды жою стратегиясы зерттелуі қажет. Көптеген бұл сияқты «жоюға» қатысты стратегиялар сандық деректерге сүйенеді – осы корпустағы белгілі құрылымның жиілігі, корпустық деректерден алынған немесе бөлектелген лексикалық бірлікті таңдап алуға шектеме және т.б.

Корпустарды алдын ала өңдеуді талқылау кезінде екі түрлі шартты қарастыру қажет болады:

1. Мәтінді өңдеуге байланысты дайындық жасаудағы әрбір қадам корпус құрушының келесі қадамдарына әсер етеді және корпусты бағалауға қатысты тиісті лингвистикалық шешім қабылдауға мәжбүрлейді. Әрине корпусты пайдаланушы өз іздегеніне қол жеткізу үшін осы шешімдермен хабардар болуы қажет. Мысалы, мәтінді құрамдық бөліктерге бөлетін маман *New York* және *Baden Baden* сияқты типтердің бір немесе екі сөзге қатыстылығы жайлы мәселені шешіп алуы қажет. Лексемаларды осы тәріздес айқындау үшін құрастырушы адам мұндай құбылыстармен не істеу керектігін, сөз алды қосымшасын бөліп алуға болатын неміс тілінің етістігі сияқты қарастыруы қажет.

2. Түпкілікті тұтынушыны алдын ала өңдеу кезіндегі қандай жұмыстар орындалғанын және мүмкін болатын қателіктермен хабардар етуіміз қажет, себебі алдын ала өңдеу кезінде және мүмкін болатын кез келген ала-құлалықтар, мысалы, кодтау кезіндегі қателер, әсіресе, жүйелі сипаттағы қателер пайдаланушының қол жеткізетін нәтижелеріне теріс әсер ететіні сөзсіз.

Сөз соңында, мәтіндер корпусын кімдер және не үшін пайдаланатыны жайлы қысқаша тоқталайық.

Корпустарды пайдаланушыларды, ең алдымен, тілшілерді, әдетте, нақты мәтіндердің мазмұнынан гөрі, оларды метамәтіндік ақпарат пен қайсыбір тілдік элементтер мен олардың құрылымдық қолданыстарының мысалдары көбірек қызықтырады. Корпустар арқылы жүргізілген ең алғашқы тілдік зерттеулер әртүрлі тілдік элементтердің мәтіндегі қолдану жиіліктерін анықтаумен ғана саятын. Статистикалық әдістер машиналық аударма, сөйлеу тілін танып білу мен оны синтездеу, орфографияны тексеретін құралдар мен грамматикалар және т.б. осы сияқты күрделі лингвистикалық мәселелердің шешімін табуда қолданылады. Мәселен, корпус материалында статистикалық әдістермен қай сөздер әрдайым бірге қолданатынын білу, оларды тұрақты сөз тіркестеріне жатқызуға болатындығының айғағы деуге болады. Семантикалық тұрғыда қарастыратын болсақ, тұрақты сөз тіркестері тұтас сипаттағы (бөлінбейтін) мағыналық бірлік, ал мұндай жағдай лексикография саласы мен мәтінді автоматты өңдеу жүйелерінде ескерілуі аса маңызды деп саналады. Лексикография мен грамматиканы зерттеу ісінде корпустар аса бай дереккөз болып табылады. Лексикографиялық зерттеулер мен семантика саласындағы зерттеулер өте тығыз байланыста болып келеді. Корпустағы қайсыбір лингвистикалық бірліктің қоршауын бақылау негізінде ондай тілдік бірлікті сипаттайтын семантикалық белгілерін анықтауға болады.

Тілші-теоретиктер корпустарды өз болжамдарын тексеруге және теорияларын дәлелдеуге қолданады. Қолданбалы лингвистер (мұғалімдер, аудармашылар және т.б.) компьютерлік корпустарды тілдерді үйретуге және өздерінің кәсіби мәселелерін (есептерін) шешуге пайдалануда. Корпустарды пайдаланушылардың айрықша санатына компьютерлік лингвист мамандарын жатқызуға болады: олар компьютерлік тілдік модельдерді құру үшін мәтінде орын алатын статистикалық және лингвистикалық заңдылықтарды айқындау мен пайдалануды мақсат етеді. Тілге қатысты басқа мамандар (әдебиетшілер, редакторлар) да корпус арқылы өздерін қызықтыратын сұрақтар бойынша қанағаттандырарлық жауап ала

алады. Қоғамдық ғылымдар саласының мамандары (тарихшылар, социологтар), өздерінің зерттеу нысанын кезең, автор немесе жанр деп аталатын мәтін параметрлерін тіл арқылы зерттеуге мүмкіндік алады. Әдебиетшілер корпусты стилеметрлік зерттеулер үшін пайдаланады. Ең соңында, корпустар әртүрлі автоматтанған жүйелерді (машиналық аударма, сөйлеу тілін тану, ақпараттық ізденіс) зерттеу үшін пайдаланылады.

А.Байтұрсынұлы атындағы Тіл білімі институтында 2012-2014 жылдары «Қазақ тілінің аннотацияланған Ұлттық корпусы» атты ғылыми тақырып бойынша зерттеу жұмыстары жүргізілсе, ал енді 2015-2017 жылдар аралығында «Қазақ тілінің Ұлттық корпусындағы метамәтіндік белгіленімдер ұстанымдары мен әзірлемесі» атты ғылыми тақырыпты қолға алды. Егерде корпус құрушыларға қажетті деген жағдайлар (орындаушылар саны мен олардың айлықтары және т.б.) орын алып жатса, болашақта қазақ тілінің мынадай корпустары жүзеге асырылады деген ойдамыз:

1) Қазақ тілінің қазіргі кездегі (немесе кезеңдік) бұқаралық ақпарат құралдары (газет, журнал бетіндегі) мәтіндерінің жеке корпусы;

2) Қазақша сөйлеу тілі жазбасының (мәтінінің) жеке корпусы (орыс тілінің «Корпус живой русской речи» тәріздес);

3) Қазақ тілінің мультимедиалық корпусы (корпустың негізін мәтіндердің видео- және аудиожазбалары құрайды);

4) Қазақ тілімен қатар (параллель) тілдердің жеке корпусы (түркітілдес және үндіеуропа тілдері), мысалы, қазақ-қырғыз, қырғыз-қазақ, қазақ-өзбек, өзбек-қазақ және т.б., сол сияқты, қазақ-орыс, орыс-қазақ, қазақ-украин, украин-қазақ және т.б. қатар тілдер корпусы;

5) Қазақ тілінің поэтикалық мәтіндерінің жеке корпусы (орыс тілінің «Корпус русских поэтических текстов» тәріздес);

6) Қазақ тілінің білім беру корпусы (орыс тілінің «Обучающий корпус русского языка» тәріздес).

ӘДЕБИЕТТЕР ТІЗІМІ:

[1] Рыков В.В. Корпус текстов как реализация объектно-ориентированной парадигмы // Труды Международного семинара Диалог-2002. – М.: Наука, 2002. С. 12-13

[2] Lüdeling A., Kytö M., eds. Corpus Linguistics. An International Handbook. Volumes 1, 2. – Berlin & New York: Walter de Gruyter, 2008. – http://alknyelvport. nytud.hu/muhelyek/elte.../HSK-Corpus-Linguistics.../file. 25 p.

[3] Гвишиани Н.Б. Практикум по корпусной лингвистике: Учеб. пособие по английскому языку. – М.: Высшая школа, 2008. 65 p.

[4] Finegan E. LANGUAGE: its structure and use. N.Y.: Harcourt Brace College Publishers, 2004. 16 p.

[5] McEnery T., Wilson, A. Corpus Linguistics. Edinburgh: Edinburgh University Press, 2001. 47 p.

[6] Баранов А.Н. Введение в прикладную лингвистику. М., 2007. – 78 с.

[7] Клименко С.В., Рыков В.В. Логические индукция и дедукция как принципы отражения предметной области в корпусе текстов // Труды Международного семинара Диалог – 2001 по компьютерной лингвистике и ее приложениям. Аксаково, 2001. – 43 с.