МРНТИ 16.21.

А.Ә. Жаңабекова

А.Байтұрсынұлы атындағы Тіл білім институты,
Қолданбалы лингвистика бөлімінің меңгерушісі, филология ғылымдарының докторы.
Алматы қаласы, Қазақстан

ТІЛДІК КОРПУСТАРДЫҢ ҒЫЛЫМИ-ЗЕРТТЕУ ӘЛЕУЕТІ

Аннотация. Мақалада тілдік корпустардың ғылыми-зерттеу әлеуеті берілген. Сондай-ақ тілдік корпустарды тіл білімінің барлық салаларында қолдануға болатындығы жайында жан-жақты қарастырылған.

Тірек сөздер: тілдік корпус, морфология, лексика, грамматика.

А.А. Жанабекова

Институт языкознания имени А. Байтурсынулы,
Заведующий отделом прикладной лингвистики, доктор филологических наук
Алматы, Казахстан

НАУЧНО-ИССЛЕДОВАТЕЛЬСКИЙ ПОТЕНЦИАЛ ЯЗЫКОВЫХ КОРПУСОВ

Аннотация. В статье рассматривается научно-исследовательский потенциал языковых корпусов. Также подробно рассматривается возможность использования языковых корпусов во всех областях языкознания.

Ключевые слова: языковой корпус, морфология, лексика, грамматика.

A.A. Zhanabekova

Institute of Linguistics named after A. Baitursynuly,
Head of the Department of Applied Linguistics, Doctor of Philology
Almaty, Kazakhstan

THE RESEARCH POTENTIAL OF LANGUAGE CORPUSES

Annotation. The article discusses the research potential of language corpuses and also the theoretical and practical importance of the compilation of linguistic corpuses in the Kazakh language. The tasks dealing with the problem of the compilation of the National Corpus of the Kazakh language are defined.

Keywords: language corpus, morphology, lexics, grammar.

Тілдік корпустарды зерттеу мен жасау сала мамандары үшін ғана емес, қоғамдық-әлеуметтік мәселе ретінде де аса маңызды. Шетел, орыс ғалымдары еңбектерінде корпус жасалғаннан кейін академиялық сөздіктер мен грамматикаларының қайта құрастырылып, қайтадан жазылып шыққандығы туралы мәліметтер берілген. Корпустық лингвистика саласының ғылыми-зерттеу әлеуеті орасан зор. Сондықтан қазақ әдеби тілі мәтіндері бойынша тілдік корпустарды ғылыми-зерттеу жұмыстарына пайдаланудың мүм-

кіндіктерін қарастырып, сол мүмкіндіктерді жүзеге асырудың ғылыми-теориялық негіздемесін жасау қажет. Өйткені тілдік корпустарды құрастыру барысында алынған нәтижелерді қолданысқа енгізу, нақты айтқанда, әртүрлі сөздіктер құрастыру ісінде, ғылыми грамматикаларды қайта жазуда немесе қандай да бір тілдік құбылыстарды айқындауда пайдалану корпустық лингвистиканың болашақ дамуы үшін ғана емес, ғылыми-зерттеу жүргізудің жаңа технологиясын қалыптастыру үшін де өзекті.

Корпус материалдары төмендегідей нәтижелерге қол жеткізуге мүмкіндік береді:

- тілдің лексика, грамматика салалары бойынша ғылыми-зерттеу жұмыстарын жүргізу үшін мол тілдік материал болу қызметі;

- тілдің бірнеше кезеңдік тарихына қатысты мәліметтер бере алуы: тілдік бірліктердің белгілі бір кезеңге қатысты тілдік қолданысы, жиілігі, белсенділігі, семантикалық өрісі туралы мәліметтер, өзгерісі, даму сатысы т.б.;

- тіл білімінің барлық салаларына қатысты анықтамалық құрал қызметі, яғни дереккөз қызметін атқару мүмкіндігі;

- автоматты түрде жиілік сөздіктер алуға;

- екі тілді (аударма) корпустар негізінде автоматты аударма жасау мүмкіндігі т.б.

- түсіндірме, екі тілді сөздіктер жасау мүмкіндігі;

- грамматикаларды қайта жазу мүмкіндігі;

- тілді оқытуда немесе тілүйренім мүмкіндігі, яғни оқу құралдары мен оқу бағдарламаларына негіз болу қызметі т.б.

Келешекте ізденушілер, ғалымдар, зерттеушілер кез келген ғылыми-зерттеушілік жұмыстарын тілдік корпустар материалдары негізінде жасайтын болады. Сондықтан жаңа заман өкілдеріне тілдік корпустарды жасап қана қоймай, оны пайдаланудың лингвистикалық және инженерлік технологиясын меңгеру қажеттігі дау тудырмайды.

Қазақ тіл білімінде А.Байтұрсынұлы атындағы Тіл білімі институты корпустық лингвистика мәселелерімен соңғы жылдары айналыса бастады. Бұл ретте қазіргі уақытта қазақ тілінің әртүрлі жанрларынан алынған таңдама мәтіндер бойынша морфологиялық деңгейде белгіленім қою жұмысы жүргізіліп жатыр. Морфологиялық деңгейде Резюмеленген тілдік корпустардың өзі келешек автоматтандырылып, бірорталықты басқару жүйесі арқылы сайтқа енгізілсе және оны ғылыми-зерттеу жұмыстарында пайдаланудың лингвистикалық және инженерлік технологиясы жасалатын болса, тіліміздің грамматикалық жүйесіне қатысты қыруар ғылыми жаңалық ашылған болар еді. Ал мұндай тәжірибе тіпті түркі тілдері бойынша қолданбалы бағыттағы үлкен ғылыми жаңалық болатындығы сөзсіз.

Ол үшін мынадай міндеттер шешілуі тиіс:

- шетел және орыс тіл біліміндегі лингвистикалық белгіленім қойылған тілдік корпустарды ғылыми-зерттеу жұмыстарында пайдаланудың ғылыми-теориялық және инженерлік технологиясына қатысты еңбектермен танысу; тәжірибе жинақтау;

- қазақ әдеби тілінің әртүрлі жанрларынан (көркем проза, драматургия, ғылыми-публицистикалық және газет пен журналдар мәтіндері) алынған таңдама мәтіндерге әртүрлі ғылыми-зерттеу жұмыстарын жүргізуге қажетті лингвистикалық белгіленімдердің түрлі модельдерін жасау;

- айқындалған лингвистикалық (стильдік, семантикалық т.б.) модельдер бойынша таңдама мәтіндерге жартылай автоматты тәсілмен белгіленімдер қою жұмысын жүзеге асыру;

- шетел және орыс тіліндегі ғылыми-зерттеу жұмыстарын жүргізуге қажетті лингвистикалық модельдердің компьютерлік алгоритміне (бағдарламасына) қатысты WEB сайттармен жұмыс істеу мүмкіндіктерін игеру;

- ғылыми-зерттеу жұмыстарын жүргізуге қажетті лингвистикалық модельдердің компьютерлік алгоритмін (бағдарламасын) жазу; электронды пішінге келтіріліп, компьютерлік бағдарламасы жазылған тілдік корпустарды ғылыми-зерттеу жұмыстарына пайдалану мүмкіндіктерін сыннан өткізу;

- морфологиялық белгіленімдер қойылған тілдік корпустардан жиілік сөздіктер алу мүмкіндіктерін қарастыру және сыннан өткізу;

- морфологиялық белгіленім қойылған тілдік корпустардан түсіндірме сөздіктерге қажетті сөздік бірліктерді автоматты жолмен айырып алу мүмкіндіктерін қарастыру; сонымен қатар әрбір сөздің мағыналық мүмкіндіктерін айқындау;

- морфологиялық белгіленімдер қойылған тілдік корпустардан автоматты жолмен фразеологиялық сөздіктер құрастыруға қажетті тұрақты тіркестерді шығарып алу жолдарын қарастыру;

- әртүрлі стильдік белгіленімдер (жерг., этн., көне., ескі кіт. т. б.) қойылған тілдік корпустардан аймақтық, этнографиялық, көне сөздер сөздіктерін құрастыруға қажетті сөздік бірліктерді автоматты жолмен шығарып алудың әдіс-тәсілдерін айқындау;

- әртүрлі лингвистикалық белгіленім қойылған тілдік корпустар материалдары негізінде тілдің әр саласы бойынша ғылыми-зерттеу жұмыстарын жүргізудің әдіс-тәсілдерін айқындау, тілдік корпус материалдарын грамматикалық зерттеулермен сабақтастыра қарастыру және нақты ғылыми нәтижелер алу т.б.

Тілдік корпустарды ғылыми-зерттеу жұмыстарына пайдаланудың теориялық және практикалық мәселелерін шешу арқылы мынадай нәтижелер алынады:

- электронды тілдік корпустарды лексикографиялық зерттеулерде пайдалану мүмкіндіктері, яғни оны әртүрлі сөздіктер (терминологиялық, этнографиялық, аймақтық, түсіндірме т.б.) құрастыру тәжірибесінде пайдаланудың лингвистикалық және инженерлік технологиясы, әдіс-тәсілдері айқындалады; сөздік құрастыру ісін автоматтандырудың ғылыми методологиясы жасалады;

- тілдік корпустардан қазақ тілінің әртүрлі жиілік сөздіктерін алудың ғылыми-инженерлік технологиясы жасалады;

- тілдік корпус материалдары негізінде қазақ тілінің құрылымдық жүйесіндегі тілдік заңдылықтар қайта қарастырылып, бұрын-соңды жасалған ғылыми тұжырымдар нақты тілдік фактілер бойынша тиянақталады немесе ғылыми жаңартпалар енгізіледі; лингвистикалық зерттеулер жүргізудің әдіс-тәсілдері айқындалады;

- таңдама мәтіндерге жасалған морфологиялық, синтаксистік, семантикалық, сөзжасамдық т.б. белгіленімдер қою тілдік жүйедегі осындай талдаулар жүйесімен тығыз сабақтастықта қарастырылып, тілдік талдаудың автоматты түрі енгізілетін болады т.б.

Әдетте қазақ тіліндегі сөздердің сөз таптарына қатысын көрсету, яғни қай сөз табына жататындығын көрсету оңай сияқты болып көрінеді. Корпус құрастыру барысында морфологиялық белгіленімдер енгізуде компьютерлік бағдарламаға реестр сөздер тізімі салынады. Ең бірінші грамматикалық белгісі олардың сөз табына қатысын көрсету болып табылады. Реестр сөздер түсіндірме сөздіктерден алынғанмен, соның ішінде де кейбір сөздердің сөз табын қоюда даулы мәселелер көп. Мысалы, *бар, жоқ* сөздерін қай сөз табына жатқызамыз? Бір кездері олар жиілік сөздіктерде әртарап сөздер деп те берілген. Сондай-ақ реестрде сөз табын қою қиындық тудырған сөздер көбінесе үстеу ретінде көрсетіле берген. Тілдік корпустарға автоматты түрде морфологиялық талдау жасайтын бағдарламалар енгізуде ол түсіндірме сөздіктер реестрін пайдаланып қана қоймай, керісінше, түсіндірме сөздіктерді жаңа сөздермен толықтырады. Миллиондаған сөзқолданыстан тұратын үлкен массивті материалдар болғандықтан, оның ішінен сөздікте берілмеген, жоқ сөздерді тауып алуға болады және де ол сөздер иллюстративтік материалымен (мысалмен) қоса

алынады. Сөйтіп, тілдік корпустар әртүрлі түсіндірме сөздіктер жасаудың бірден-бір материалдық құралы қызметін атқара алады. Бұған қоса түсіндірме сөздіктерде сөздердің мағыналық мүмкіндіктерінің барлығы қамтыла бермейді. Қолданыста болғанмен, жадта сақталмайды, яғни сөздікшінің ойына келмеуі мүмкін. Корпус материалдары бір сөздің қолданыстағы қаншама контекстерін көрсететіндіктен, бұрын сөздікте көрсетілмеген мағыналық реңкі беретін мысалдарды да шығара алады. Осы мысалдар арқылы сөздіктің мағыналық парадигмасын байытады.

Сонымен қатар автоматты морфологиялық талдауда грамматикалық сөздік жасалады, мұндай сөздік жасаудың бір түрі – түбір сөздерге жалғанатын түрленім сөздігін жасау, яғни сөзформалар сөздігін жасау. Агглютинативтік тілдерде түрленім қосымшалар бірінің үстіне бірі синтагматикалық ретпен жалғанады. Мысалы, *лар-ы+мен; ғыз+дыр+ды* т.б. Осындай етістіктердің, зат есімдердің өзіндік түрлену жүйесі, парадигмасы бар. Түрленімге қатысты да даулы сұрақтар туып отырады. Мысалы, сөзжасам деп анықталып жүрген кейбір қосымшалар (-дай/дей, -лық/лік, -сыз/сіз) т.б. түрленім формаларынан кейін жалғанады. Мысалы: *бар+ған+дай, айт+у+ы+нсыз, жаса+ған+дық* т.б. Осы жекелеген қосымшалардың грамматикалық сипаттамасын жасағанда оларды қандай атаумен белгілейміз? Бұл проблема да көп уақыт бойы даулы болып келген «қос функциялы» қосымшаларды нақты бір тілдік қабатқа жатқызудың, оны біржақты шешудің уақыты келгендігін көрсетеді. Мысалы, жоғарыдағы есімше формасына кейін жалғанған *-дай* жұрнағын – салыстыру формасы деп атауға болады. Өйткені ол бұл жерде сын есім тудырып тұрған жоқ.

Міне, тілдік корпустарды жасау, бір жағынан, массивті тілдік материалдар арқылы ғылыми-зерттеу жұмыстары үшін эмперикалық материал болса, екінші жағынан, тіл білімінде шешімін таппаған проблемаларды айқындауға мүмкіндік береді, түрткі болады.

Сонымен қатар тілдік корпустар арқылы сөздердің тіркесім мүмкіндігін айқындауға болады. Мұны тіл білімінде сөздердің дистрибуциясы деп атайды. Тіркесімділіктің үлкен массивті материалдар арқылы жүйесін табу тіл біліміндегі көптеген синтаксистік құбылыстарды айқындауға мүмкіндік береді. Сөздердің тіркесімділік жүйесін анықтау арқылы басқа да тілдік құбылыстарды тануға болады. Мысалы, тілдік корпустарда мәтінге автоматты морфологиялық талдаулар жасағанда, компьютер берілген мәтіндегі омонимдер мен омографтарды ажырата алмайды. Мұны корпустық лингвистикада омонимдерді ажырату (снятие омонимии) деп атайды. Омоним сөздердің екі жағындағы сөздер тіркесінің заңдылығын модельдеу арқылы омонимдерді автоматты түрде ажырататын бағдарлама жасау өте өзекті мәселе. Бұл басқа тілдерде де жасалмаған.

Тілдік корпустарды тіл білімінің барлық салаларында қолдануға болады. Нақты айтқанда, әртүрлі сөздіктер (жиілік, түсіндірме, аймақтық, фразеологиялық, этнографиялық) құрастыруда, ғылыми грамматикалар жазуда, оқыту жүйесінде және оқулықтар құрастыруда, әдістемелік құралдар жазуда, аударма жұмыстарында т.б. Тілдік корпустар компьютерлік базаға салынып, бір орталықты басқару жүйесі бойынша жұмыс істейтіндіктен, тілдік зерттеулердің барлығын дерлік нақты фак-тологиялық материалдармен қамтамасыз етеді. Тілдік корпустармен жұмыс істеу ғылыми-зерттеу жұмыстарын автоматтандырып, ғылымды дамытуға орасан зор үлес қоспақ.

Корпустар бір тілдің барлық стилін қамтыса, ұлттық корпус немесе кейде үлкен корпус деп аталады. Өз ішінде осындай ұлттық корпустарға салынған мәтіндерді жанрлық, стильдік немесе кезеңдік тұрғыдан жіктеу арқылы жеке шағын корпустарды бөліп алуға да болады. Мәселен, қазақ газет мәтіндерінің корпусын жасаудың мынадай әдіс-тәсілдері бар.

Корпус құрастыру негізінен қолданбалы бағытта жүзеге асырылатындықтан, жұмыс бір мезгілде екі бағытта қатар жүргізіледі. *Бірінші бағытта* қазақ тілінде жарық көрген

алғашқы қазақ газеттерінен бастап, соңғы кездері жарық көріп жатқан газеттер жинақталып, электронды пішінге келтіріліп, компьютер жадына салынады. Жинақталған газет мәтіндерінің шыққан жылы, атауы, мақала авторы, шыққан орны, стилі т.б. сипаттары бойынша метабелгіленімдер әзірленіп, автоматты іздеу бағдарламасына сәйкестендіріліп, корпусқа енгізіледі. Осындай метабелгіленімдер арқылы зерттеуші өзіне қажетті газет мәтіндерін қол жұмысы арқылы газет нөмірлерін ақтарып жатпай-ақ, бір мезетте тауып алатын болады. *Екінші бағытта* қатар атқарылатын жұмыс – газет мәтіндер корпусына салынатын лингвистикалық белгіленімдерді әзірлеу. Авоматты түрде лингвистикалық сипаттама жасалатын бағдарламалар салынған мұндай корпустар әсіресе, газет мәтіндерін, БАҚ тілін зерттейтін зерттеушілер үшін өте құнды материал болады. Мұндай газет мәтіндерінің Резюмеленген корпустары жасалатын болса, алғашқы газеттерден бастап қазірге дейінгі газет тіліндегі әртүрлі тілдік өзгерістер, тілдің тарихи даму үрдістері, терминдердің қолданысы, әртүрлі сөзжасамдық, терминжасамдық үдерістер – бәрі-бәрі компьютерде көз алдымызда көрініп тұратын болады. Сонымен қатар сан алуан лингвистикалық ақпараттарды іздеу бағдарламасы арқылы тез әрі оңай тауып алуға болады. Нақты нәтиже – метабелгіленімдер мен лингвистикалық белгіленімдер енгізілген автоматты компьютерлік бағдарламамен жұмыс істейтін қазақ газет мәтіндерінің электронды жинағы (корпусы) болып табылады.

Осы әдіспен қазақ әдеби тілінің басқа да стильдері мен жанрлары бойынша жеке-жеке шағын корпустар түрлерін жасауға болады. Мысалы, қазақ прозалық мәтіндер корпусы, поэтикалық корпус, ауызша корпус, публицистика тілінің корпусы, драматургия мәтіндері корпусы, ғылыми-техникалық мәтіндер корпусы, жарнама тілінің корпусы т.б.

Сонымен, тілдік корпустар жасау мәселесі – қазақ тіл білімінде күні бүгінге дейін толық шешімін таппай келе жатқан өзекті мәселелердің бірі. Әлем тілдеріндегі компьютерлік (қолданбалы) лингвистика жетістіктерін ұлттық тіліміздің қажетіне пайдалану осы сала мамандарының алдында тұрған жауапты іс.

ӘДЕБИЕТТЕР ТІЗІМІ:

[1] Беляева Л.Н., Герд А.С., Убин И.И. Автоматизация в лексикографии // При- кладное языкознание (учебник) // Отв. ред. А.С. Герд. – СПб.: Изд-во С.Петерб. университета, 1996. С. 318-333.

[2] Фрэнсис У. Н. Проблемы формирования и машинного представления большого корпуса текстов // Новое в зарубежной лингвистике. Вып. XIV. Проблемы и методы лексикографии. М., 1983. С.334-353.