



ХҒТАР 16.01.011

Ғылыми мақала

<https://doi.org/10.32523/2616-678X-2025-150-1-172-185>

Корпустың құрылымдық сипаты және қызметі

С.С. Шоқабаета^{*1}, Ф.Ж. Серікжан^{*2}

¹Л.Н. Гумилев атындағы Еуразия ұлттық университеті, Астана, Қазақстан

²Қарағанды облыстық білім басқармасының “Дарын” мамандандырылған мектеп-лицей-интернаты, Қарағанды, Қазақстан

(E-mail: ¹shokabaeva@yandex.kz*, ²farida.serikzhan@mail.ru)

Аңдатпа. Тілді модельдеудің бір бағыты – компьютерлік аударма ісі бойынша зерттеулер жүргізу қазір аса маңызды және уақыт талабы болып отыр. Зерттеудің мақсаты – корпус, параллель корпус, олардың мақсаты мен қызметі, құрылым ерекшеліктері жайында мәлімет беру, корпустардың машиналық аударма ісіндегі қызметін талдау. Негізгі бағыттары – корпус түзудің негізгі талаптарына, жалпы құрылымына ақпараттық шолу жасау. Негізгі идеясы – машиналық аударма ісінде компьютерлік лингвистика әлеуетін ұтымды пайдалану жолдарын ұсыну. Корпус құрылымы, жұмыс істеу алгоритмі, қызметі бойынша берілген мәліметтер машиналық аударма саласында жүргізілетін әртүрлі көлемдегі зерттеулерге сай теориялық әрі тәжірибелік негіз бола алады. Зерттеу әдістемелері: теориялық талдау, жинақтау, сауалнама. Зерттеудің негізгі нәтижелері, талдаулары, қорытындылары: сауалнама нәтижесінен бақылағанымыз: Қазақ тілінің ұлттық корпусының сан алуан мүмкіндіктеріне сұраныс көп. Корпустың міндеттері мен мүмкіндіктері әлдеқайда күшті екені белгілі. Дәлірек айтқанда, машиналық аударма ісінде негізгі құрал болуы керек екенін айрықша атаған орынды деп есептейміз. Өйткені бүгінде Қазақ тілінің ұлттық корпусында тек мәтінді өңдеудің бірнеше түрі ғана іс жүзінде қолданылады. Осының өзінде техникалық әрі базалық ақау аз емес. Ал машиналық аударманы жетілдіру үшін корпустың ішінде параллель корпустар құрылуы қажет.

Жүзеге асырылған зерттеудің мәні Қазақ тілінің ұлттық корпусының дамуына үлес қосады. Жұмыс нәтижесінің практикалық мәні: Корпус қызметі, жұмыс істеу алгоритмі, құрылымы бойынша алынған нәтижелер кейінгі зерттеушілерге тәжірибелік жұмыстар жүргізу кезінде жетекші нұсқаулық қызметін атқара алады.

Түйін сөздер: компьютерлік лингвистика, корпус, корпустың құрамы, ұлттық корпус, корпус құрастыру, лексикографиялық зерттеулер.

Түсті: 26.11.2024; Жөнделді: 05.12.2024; Мақұлданды: 19.12.2024; Онлайн қолжетімді: 30.03.2025

*Хат-хабар үшін автор

Кіріспе

Корпус – компьютерлік лингвистика саласының аса маңызды бір құралы. Бұл мәселені дәстүрлі картотека қорын жасау мен оны қолдану әрекетінің жалғасы деуге де болады. Өткен ғасырда компьютер жадына картотекалар енгізіліп, соған сәйкес электрондық қорға айналып, қолданушылардың пайдалануына мүмкіндік берілді. Десек те, кейін лингвистикалық зерттеулерге қажет мәліметтерді жинақтауға, сақтауға байланысты компьютерлік технологияны пайдаланудың жаңа тәсілдеріне қажеттілік туды.

Компьютер жадына енгізілген мәтіндерді алғашқы корпусқа жинауды тілшілер 1960 жылы бастаған. Корпустық лингвистика АҚШ-та 1963 жылы У.Фрэнсистің жетекшілігімен құрылған Браун корпусынан басталады.

Браун университетінде қазіргі ағылшын тілінің американдық нұсқасының Браун корпусының компьютерлік нұсқасы бар. Бас кезінде корпуста бір миллион сөзқолданыс болған, оған әрбіреуі екі мың сөзқолданысқа тең, морфологиялық белгіленімдермен жабдықталған 500 мәтін енгізілген. Осы корпусқа енгізілген мәтіндер алғаш 1961 жылы шыққан ағылшын тілінің әдеби нұсқасына жатады. Браун корпусының индекстелген, яғни аннотацияланған нұсқасы 1980 жылы шықты. Онда сөзформалар лемматизацияланып, синтаксистік қызметтері белгіленеді, т.б.

Корпустың авторлары У. Фрэнсис пен Г. Кучера кейін алғаш рет корпусты статистикалық өңдеуден өткен жиілікті, әліпбилі-жиілікті сөздіктер және әртүрлі статистикалық үлестірімдер, т.б. материалдармен толықтырады. Авторлар корпусты құрастыруда бес арнайы принциптер мен критерийлерді ұстанған [1, 16].

Алғашқы Браун корпусын жасау жазба ағылшын тілінің әр жанрын жеке, жүйелі әрі салыстыра зерттеуді, саралауды іске асыруды мақсат етті. Браундық корпустың құрастырылуы көпшіліктің қызығушылығымен бірге түрлі пікірталас тудырды. Мұндай пікірталастар, ең алдымен, корпуста шешілетін мәселелерге, мәтінді іріктеу принциптеріне байланысты болды. Оның статистикалық әдістердің негізінде құрылып, қолданылуы корпус авторының кәсіби тұрғыдағы интуициясына негізделді. Осындай қиын процестің мүмкіндігінше объективті болуы үшін неғұрлым формалданған, тексеру мен бақылауға барынша айқындылық, нақтылықты қамтамасыз ету керек болды.

Кейін Браундық корпус осындай корпустар құрастыруға қатысты, соған арналған зерттеулердің кең тараған үлгісіне, нысанына, стандартына айналды деуге негіз бар. Зерттеуші ғалымдар арасында көптеген лингвистикалық зерттеулерді сапалы жүргізу үшін тек көлемді тілдік материалдар негізінде жұмыс істеу керек екені қабылдана бастады. Бұның барлығы мәтіндерді корпус түрінде ұйымдастыру ережелерін, оларды талдау әдіснамасын зерттейтін бағыттың пайда болуына алғышарт болып, осы салаға ғалымдардың қызығушылығын арттырып, тың идеялар ұсынуға, жаңалық ашуға одан әрі жол ашты.

Әдебиеттерге шолу

Корпустың өзіндік даму жолы бар. “...1961 жылы алғаш рет мәтіндер корпусын еуропалық зерттеушілер Ұлыбританияда құрастырды. Браундық корпус сияқты бұл

корпус та 1 млн. сөзқолданыстан тұратын ағылшын тілінің британдық жобасын ұсынды. Оның ішінде 2000 сөзқолданыстан тұратын 500 мәтінді 15 жанрдан іріктеп алған. Бұл корпус Ланкастер-Осло-Берген (The Lancaster-Oslo-Bergen Corpus, британдық және екі норвеждік университеттер немесе қысқаша LOB) корпусы деген атауға ие. Браундық корпус сияқты теңгерімделген корпустар зерттеушілер үшін тілдік талдау мен зерттеулер жасау үшін аса маңызды құрал саналады. Lancaster/Oslo-Bergen Corpus-тың, яғни Ұлыбританияда құрастырылған корпус құрамында синтаксистік белгіленіммен жабдықталған екі кішігірім корпус бар: Leeds-Lancaster Treebank және Lancaster Parsed Corpus” [1].

Демек, алғашқы екі ірі корпустар – ол жазба тілінің америкалық және британдық нұсқаларының ағылшын тіліне арналған корпустар. Екі корпус та қазіргі уақытта да маңызды, өйткені бұл екі корпус негізінде ағылшын тіліндегі көптеген сапалы зерттеулер жасалған.

Неміс-ағылшын машиналық аударма жүйесі жобасы бойынша жасалған LIMAS-корпусының жобасын әзірлеуге Браун корпусының пайда болуы себеп болды.

Лингвистика тарихында аса көлемді француздық жобалардың бірі «Француз тілінің қазынасы» еңбегі болды. Ол 90 млн. сөзқолданыстан тұратын мәтіндер корпусын қамтыды. 1963 жылы жобаны әзірлеу басталды. 1968 жылы негізгі материалды енгізу аяқталды. Оның қорына XIX-XX ғасырлардағы мәтіндер енгізілді. XIX-XX ғасырлардағы француз тілінің 80 мың лексикалық бірліктен тұратын 15 томдық сөздігін жасауда осы корпус пайдаланылған [1, 24].

Бұл корпустарды жасағаннан кейінгі өткен бірнеше онжылдықтан кейін компьютерлер арзандап, әлеуеті де бұрынғыдан артып, бағасы да арзан әрі сенімді сканерлер мәтіндерді компьютерде теру қажет болмайтындай етті. Осылайша, жаңа техниканың арқасында корпустарды құрастыру үдерісі де жеңілдеді. Қазіргі корпустар соның нәтижесінде миллиардтаған сөзқолданысты қамтиды.

Тілде барынша жиі кездесетін вариациялық және өзгермелі құбылыстарды зерттеуде корпус көлемінің пайдасы зор. Ол мына бағыттар бойынша сенімді нәтижелер алуға мүмкіндік береді:

1. есім, етістіктердің, т.б. морфологиялық нұсқаларын, олардың эволюциясын зерттеу;
2. сөзжасамдық нұсқаларды, оларға қатысты паронимдердің, сөзжасамдық үлгілердің өнімділігін, құралдарын зерттеу;
3. қабысу, меңгеру варианттарының өзгеріске ұшырауын зерттеу;
4. орыс тілінің акцентологиялық нұсқалары мен акценттік жүйедегі өзгерістерін зерттеу;
5. лексикалық жарыспалықты, мәселен, синонимдік қатарлардың және тақырыптық топтардың құрамдарын, сондай-ақ олардағы семантикалық арақатынастарды зерттеу.

Корпус ұғымының анықтамасына тоқталсақ, «корпус» дегеніміз – компьютер жадына енгізілген, лексикалық, грамматикалық бірліктері қамтылған мәтіндер жиынтығы. Ғалым А.Қ. Жұбанов пен А.Ә. Жаңабекованың «Компьютерлік лингвистика» еңбегінде «әр тілдегі электронды пішінге келтірілген, яғни бір басқару орталығынан автоматты түрде жұмыс істейтін, лингвистикалық ақпараттарды қамтитын мәтіндер жинағы» деген

анықтама берілген [1, 14]. Осы еңбектегі енді бір анықтама мынадай: «мәтіндер корпусы дегеніміз – нақты тілдік мәселелердің шешімін табуға арналған аса үлкен көлемдегі машина, яғни компьютер оқи алатындай түрде көрініс табатын, бірыңғайланған, құрылымдалған, белгіленген (арнайы шартты белгілер қойылған), филологиялық тұрғыда компетентті саналатын тілдік деректер ауқымы» [1, 15].

Шетелдік ғалым Э.Финеган «корпус» ұғымына мынадай анықтама берген: «Корпус – машина (компьютер) оқи алатын форматта және мәтін туралы ақпарат берілген, яғни айтушы, автор, адресат (тыңдаушы) немесе дәрісхана (тыңдаушылар) жайлы ақпараттар қамтылатын мәтіндердің репрезентатты (көлемді) жиыны» [2]. Ғалым Т. Мак Энери мен Э. Вилсон ұсынған анықтама мынадай: «Корпус – ол тілдік модель ретінде қолданылатын нақты тілдік критерийлерге сәйкес келетіндей таңдалып алынған тілдік фрагменттердің жиынтығы» [3]. Зерттеуші ғалым В.В. Рыков «олар мәтіндерді құрастыру ережелері, мәтіндерді талдаудың алгоритмі мен бағдарламасы және осымен түйіндес идеология мен әдіснамалар арқылы жүзеге асады», – деп сипаттайды [4]. Ғалым В.П. Захаров «Корпусная лингвистика» оқу құралында корпус ұғымын тар мағынада түсіндіреді: «Под названием лингвистический, или языковой, корпус текстов понимается большой, представленный в электронном виде, унифицированный, структурированный, размечанный, филологический компетентный массив языковых данных, предназначенный для решения конкретных лингвистических задач» [5].

Жоғарыда келтірілген тұжырымдардан корпустың қандай да бір мәтіндердің емес, белгілі мақсатта жинақталған, арнайы талаптарға сай сұрыпталған мәтіндер жиынтығы екенін түсінеміз. М. Стабстың еңбегінде корпусқа тән бірнеше талап аталған:

1. корпус – кез келген мәтіндердің жиыны емес, компьютерге енгізілген, компьютерде өңделетін мәтіндер;
2. корпусқа ауызша мәтіндер де енгізіледі, яғни диалогтар;
3. корпусқа енгізілетін мәтін жайлы толық ақпарат болуы тиіс: авторы, шыққан жылы, жанры және т.б. [6, 138].

Белгілі бір принциптерге сәйкес құрастырылған мәтіндер жиынтығы – корпус. Ол принциптер лингвистикалық сипатта да, экстралингвистикалық сипатта да болуы ықтимал. Корпус қандай да бір тілдің бір немесе бірнеше түрлерімен, белгілі бір уақыт аралығымен немесе осы параметрлердің комбинациясымен, мәтіндердің белгілі бір түрлерімен шектелуі мүмкін.

Корпустың құрамы жайындағы мәлімет, әдетте, зерттеушіге оның интерфейсі арқылы қолжетімді болады. Демек, корпус – әр тілдегі электронды қалыпқа келтірілген, яғни бір басқару орталығынан автоматты түрде жұмыс істейтін, лингвистикалық ақпараттардан тұратын мәтіндер жинағы. Мәтіндер корпусының негізін мәліметтер корпусы құрайды. Феномендері тілдік сипаттауды қажет ететін, қандай да бір ережелерге сай құрастырылған, тілдік жүйеде жүзеге асатын деректер жинағын мәліметтер корпусы дейміз. Ол бірөлшемді (тілдік) ғана болады.

Мәтіндер корпусы – бірліктері есебінде мәтіндер немесе олардың көлемді үзінділерін қамтитын мәліметтер корпусының көрінісі [1].

Корпус (лат. corpus – «body») саны жағынан бір жиыннан асатын кез келген мәтіндер жиынтығынан құралады.

Келтірілген анықтамаларды негізге ала отырып, қазіргі мәтіндер корпусының негізгі белгілері деп мыналарды атауға болады:

1) машина, яғни компьютер оқи алатын форматта болуы;

2) ерекше іріктеу әрекетінің нәтижесі ретіндегі репрезентаттық, яғни көлемділік сипаты;

3) металингвистикалық ақпараттың ұсынылуы.

Корпустарды жасап, оны пайдалану үшін бірнеше алғышарттар қажет:

1) аса ірі корпус (репрезентативтілік) мәліметтердің типтілігіне кепілдік береді әрі тілдік құбылыстардың кез келген саласы жайлы жан-жақты, терең, толық ақпаратпен қамтамасыз етеді;

2) корпустарда түрлі деректер типтері табиғи мәнмәтіндік түрінде сақталатындықтан, олар объективті, жан-жақты зерттеуге мүмкіндік береді;

3) бір-ақрет жасалған деректер жиынын әртүрлі зерттеушілер түрлі мақсатта, қолайлы уақытта бірнеше рет пайдалануларына болады. Корпустарды түрлі анықтамалар мен статистикалық мәліметтер алуға, тілдің лексикалық құрамының уақыт өте келе өзгерісу динамикасын зерттеуге мүмкіндік береді, көпәспектілі лексикографиялық жұмыстар жасау кезінде дереккөз ретінде де, сөздіктерді әзірлеуде де зор қызмет атқарады. Оның мәліметтерін грамматикалық жүйелер құрастыруда, тілді оқыту үшін, машиналық аударма жасауда қолдануға болады.

Негізінен қолданбалы бағытта жүзеге асырылатындықтан, корпус құрастыру жұмысы бір уақытта 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 атауы, мақала авторы, шыққан жылы, 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 бағыттағы жұмыс – газет мәтіндер корпусына салынатын лингвистикалық белгіленімдерді әзірлеу. Автоматты 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 [7, 127].

Өз қызметін толық атқару үшін корпус мынадай бірнеше талаптарға сай болуы тиіс:

– релевантты құбылыстардың толық ескерілуі;

– корпустың үнемділік қасиеті, яғни пәндік аяны зерттейтін маманның іс-әрекетін үнемдеуі;

– тілдік материалдарды құрылымдау жолының пайдаланушы үшін де аса маңыздылығы;

– корпус деректерін өңдеуге қажетті компьютерлік бағдарламамен қамтамасыз етілуі: статистикалық талдау, конкорданс құрастыру, сөздікті автоматты өңдеу – толық және ішінара сөзтізбелер алу: әліпбилі, жиілікті сөздіктер алу, лемматизациялау әрекеттері.

Мәтіндер корпусын қолдану аясы алуан түрлі болуына байланысты ғалым А.Жұбанов пен А.Жаңабекованың еңбегінде олар лингвистикалық, филологиялық және педагогикалық деген үш салаға жіктелген. Осындай топтарға жататын пәндер: лексикография, грамматика, лингводидактика, мәтін лингвистикасы, типология, контрастивтік лингвистика, аударма жасаудың теориясы мен практикасы, әдебиеттану ғылымы.

Әдіснама

Зерттеу барысында алғашқы кезеңде тақырыпқа қатысты теориялық мәліметтер жинақталып, талданды, сараланды, сондай-ақ қолданылатын әдіс-тәсілдер анықталды. Екінші кезеңде қазақ тілінің ұлттық корпусының бүгінгі мүмкіндіктерін бағалау мақсатында сауалнамаға қатысушылар таңдалып, жоо білім алушыларына (28 адамға) сауалнама жүргізілді. Бұл қазақ тілінің ұлттық корпусының бүгінгі мүмкіндіктерін бағалауға мүмкіндік берді. Зерттеудің соңғы кезеңінде сауалнама нәтижелері талданып, қорытынды жасалды.

Нәтижелер мен талқылау

Қазіргі уақытта корпустық лингвистиканың қолданылу өрісі бұрынғыдан да кеңейген. Бейнелейтін аясына қарай, корпустағы сақталатын ақпарат бірлігі тұрғысынан, оны әртүрлі негізде топтастыру мен құрастыру, т.б. принциптерді ұстаным етуіне қарай корпустық лингвистиканың негізгі түсінігін түрліше ұғуға болады. Пәндік саланың белгілі бір жиынтығы – корпуста сақталатын бірлік. Мәселен, оларға сөз, сөзформа, сөз тіркесі, сөйлем немесе толық мәтіндер жиынтығы жатады. Корпустағы деректер құрылымданған және құрылымданбаған пішімдерде сақталуы мүмкін. Біріншіден, яғни құрылымданған пішім жағдайында арнайы шартты белгілермен белгіленген мәтін туралы сөз болады. Ол белгілер семантикалық, морфологиялық, синтаксистік болуы мүмкін. Мысалы, қазақ тілінің ұлттық корпусында да сөздердің сөзжасамдық құрамы, морфологиялық құрылымы, лексика-семантикалық сипаты мен синтаксистік құрылымына қатысты біршама ақпарат енгізілген.

Тіліміздің ұлттық корпусында белгіленімнің бес түрі бар: фонетика-фонологиялық, лексикалық, семантикалық, сөзжасамдық, морфологиялық.

Фонетика-фонологиялық белгіленім сөздегі әрбір фонеманың (тіл дыбысының) жасалым (артикуляциялық), айтылым (акустикалық) белгілерін көрсетеді.

Лексикалық белгіленім корпустан ізделген сөздің сөздікте тіркелген барлық мағынасын көрсетеді. Егер ізделген сөз омоним болса, басқа да мағыналары бірге беріледі.

Семантикалық белгіленім ізделген сөздің қандай мағыналық топқа жататынын көрсетеді.

Сөзжасамдық белгіленім ізделген сөздің жасалу жолын көрсетіп, сөздің түбірі мен жұрнағын ажыратады.

Морфологиялық белгіленім ізделген сөздің сөз табын анықтап, жұрнақ, жалғаудың түрлерін ажыратып береді. Жұрнақ, жалғау, сөз табының атаулары корпуста шартты белгілермен қысқартылып берілген. Іздеген сөз корпустан табылмаған жағдайда табылмады, белгісіз деген мағынаны білдіретін «Unknown» жазуы шығады.

Қазақ тіліндегі сөздердің сөз таптарына қатысын көрсету оңай емес. Корпус құрастыру барысында, морфологиялық белгіленімдер енгізуде компьютерлік бағдарламаға реестр сөздер тізімі салынады. Ең бірінші грамматикалық белгісі – олардың сөз табына қатысын көрсету. Реестр сөздер түсіндірме сөздіктерден алынғанымен, олардың ішінде де кейбір сөздердің сөз табын қоюда даулы мәселелер көп. Мысалы, бар, жоқ сөздерін қай сөз табына жатқызамыз? Олар бір кездері жиілік сөздіктерде әртарап сөздер деп те берілген. Бұған қоса реестрде сөз табын қою қиындық тудырған сөздер көбінесе үстеу ретінде көрсетіле берген. Тілдік корпустарға автоматты түрде морфологиялық талдау жасайтын бағдарламалар енгізу түсіндірме сөздіктер реестрін пайдаланып қана қоймай, түсіндірме сөздіктерді жаңа сөздермен толықтырады. Миллиондаған сөзқолданыстан тұратын ірі массивті материалдар болғандықтан, оның ішінен сөздікте берілмеген сөздерді тауып алуға болады және ол сөздер иллюстративтік материалымен (мысалмен) қоса алынады. Демек, тілдік корпустар әртүрлі түсіндірме сөздіктер жасаудың бірден-бір материалдық құралы қызметін атқара алады. Сондай-ақ түсіндірме сөздіктерде сөздердің барлық мағыналары қамтыла бермейді. Тілде қолданыста болғанымен, 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. Бұл мысалдармен сөздіктің мағыналық парадигмасын байытады.

Метабелгіленім ізделген сөздің қандай мәтіннен алынғанын, мәтіннің авторын, тақырыбын, дереккөзін, жарияланған мерзімін, таралу типін, стилін, т.б. жан-жақты ақпарат береді.

Ғалым А.Қ. Жұбанов пен А.Ә. Жаңабекованың зерттеуінше, “тіліміздің ұлттық корпус материалдары мынадай нәтижелерді алуға мүмкіндік береді:

- тілдің лексика, грамматика салалары бойынша ғылыми-зерттеу жұмыстарын жүргізу үшін мол тілдік материалдық қызмет;

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

- тіл білімінің барлық салаларына қатысты анықтамалық құрал, 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 тіл үйрену мүмкіндігі, яғни оқу құралдары мен оқу бағдарламаларына негіз болу қызметі, т.б.” [1, 302].

Болашақта ғалымдар, зерттеушілер, ізденушілер кез келген ғылыми-зерттеу жұмыстарын тілдік корпус материалдары негізінде жасайтындықтан, жаңа дәуір адамдарына тілдік корпустарды жасаумен бірге оны пайдаланудың лингвистикалық, инженерлік технологиясын меңгеру қажет болады.

“Мәтіндер корпусы” ұғымы мәтіндер мен тілдік деректерді басқару жүйесін қамтиды. Корпусты пайдаланушы мен корпус арасындағы әрекет «корпустық менеджерлер» атты компьютерлік бағдарламалардың көмегімен орындалады.

“Корпустық менеджер – бұл жүйе корпустағы деректерді іздеп-табуға арналған бағдарламалық құралдарды өз қатарына қосумен бірге, статистикалық ақпаратты алуды және пайдаланушыға алынған нәтижелерді ыңғайлы түрде ұсынуды көздейтін мамандандырылған іздестіру жүйесі болып табылады [1, 22].

Мәтіндер корпусындағы қандай да бір сөзді іздеу оның конкордансын жасауға мүмкіндік береді. Ізделген сөздің барлық қолданысын кездесетін мәтін үзінділері қамтылған шығармалар тізімі мен оларға сілтеме берілген мәліметтер конкорданс деп аталады.

Корпустық лингвистиканың пайда болуы тілшілердің тілді зерттеу әдісіне елеулі өзгеріс әкелді. Қазір корпусты пайдаланушылар үшін өте ауқымды мәтіндерден сөз бен сөз тіркестері бойынша мысалдар келтіру, іздеп-табу бірер секундта іске асатын болды. Демек, жаңа ақпаратты қамтыған тілдік сөздіктердің жаңартылған нұсқасын бұрынғыдан гөрі тез арада баспадан шығаруға мүмкіндік пайда болды. Аса көлемді мысалдарды зерделеп жасалған сипаттамалардың да толық, дәл болып, сапасы жоғарылайтыны сөзсіз. Іріктелген мысалдарды мағыналық тобы жағынан да жинақтауға мүмкіндік бар.

Сондай-ақ корпустық деректердің аса көп метамәліметтерден тұратынына қарамастан, жеке сөздер мен сөз тіркестерінің қайдай стильге, жергілікті тілдік ерекшеліктерне, т.б. тән қолданыстарын топ-топқа бөліп қарастыру да әлдеқайда жеңілдеді. Үнемі дамып, көлемі де өсіп отыратын мониторлық корпустардың сөздік құрылысының сапасын жетілдіруде маңызды зор. Өйткені ол лексикографтарға тілге еніп жатқан жаңа сөздерді немесе жаңа мағынаға ие болып жатқан сөздерді, тіпті олардың стильдік қолданыстарының ерекшеліктерін қадағалауға, лексика саласына қатысты статистикалық сипаттама жасау, сөздердің жиілігін анықтауға да зор мүмкіндік береді. Корпус белгілі бір тірек сөздің бірнеше туыс тілдерде қолданылу мүмкіндігінің, статистикалық үлесін есептеп шығаруда да тиімді. Жеке сөздің ғана емес, олардың тіркесімдік позицияларын жүйеге түсіру мүмкіндігі де бұрынғы дәстүрлі зерттеулерден өзгеше болатыны сөзсіз.

Грамматикалық және лексикографиялық модельдер арасында өзара жүйелі байланыс бар. Мәселен, лексикографиялық зерттеулердің корпустық тәсілдерді қолдану арқылы өзара қатынастағы сөздердің әртүрлі жағдаятта, әртүрлі контекстерде қалай қолдануға болатынын көрсету мүмкіндігі бар болса, дәстүрлі әдістер синонимдік сөздер тобын ғана анықтай алады.

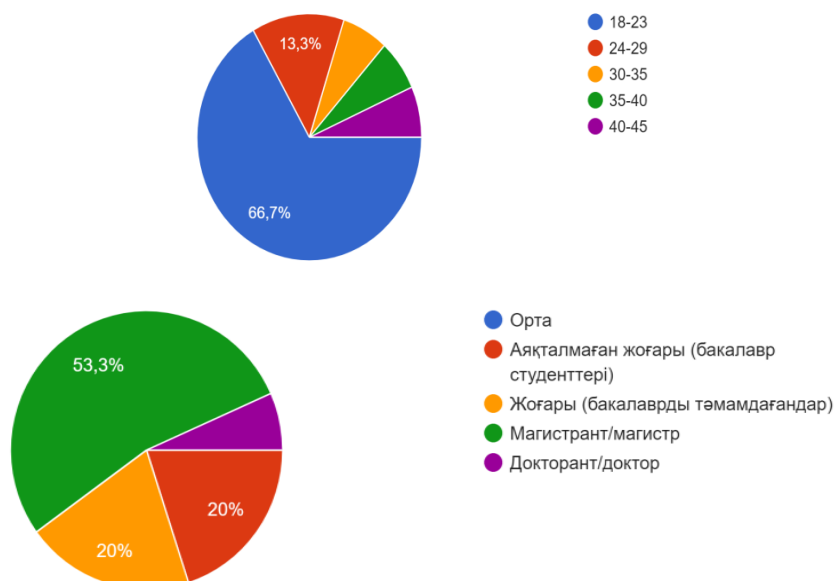
Корпустық тәсілмен зерттеудің лексикографиядағы тиімді тұсы – корпустан берілген сөз қолданылатын контекстер жиынын кең түрде алу мүмкіндігі. Осыдан кейін конкорданстық тізімдердің көмегімен мәнмәтіндерден ізделген сөзді ассоциациялайтын түрлі мағыналарды іріктеп алуға болады.

Ғалым А.Жаңабекованың еңбегінде корпус түрлерін жасап шығу үшін шешілуі тиіс міндеттер мен тілдік корпустарды ғылыми-зерттеу жұмыстарына пайдаланудың теориялық және практикалық мәселелерін шешу арқылы алынатын нәтижелер атап көрсетілген [7, 128].

Тіліміздің ұлттық корпусындағы мәтіндер әдеби тілдің 5 стилінен (ғылыми стиль, публицистикалық стиль, көркем стиль, ісқағаз стилі, сөйлеу стилі) алынған.

Корпустан керек сөзді іздеу кезінде конкорданс, лемматизация, лингвистикалық белгіленімдер енгізу бағдарламасы сияқты компьютерлік бағдарламалар жұмыс істейді.

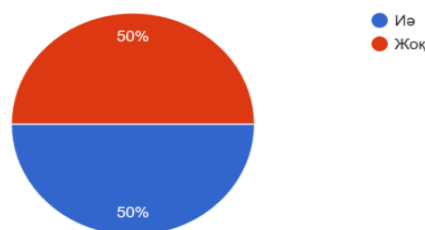
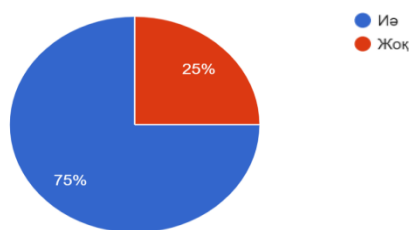
Ұлттық корпусымыздың бүгінгі мүмкіндіктерін бағалау үшін сауалнама жүргізілді. Нәтижесі мынадай:



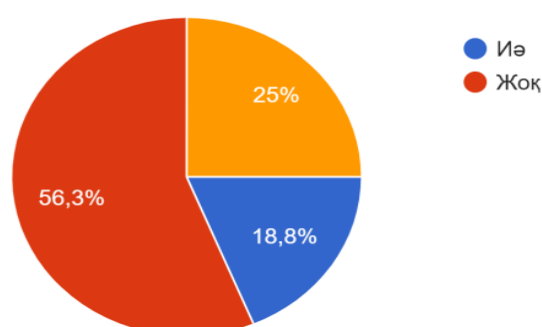
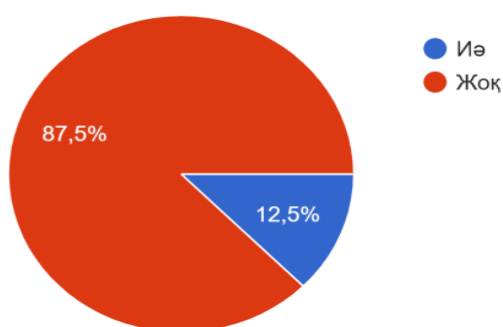
Сауалнамаға жалпы 28 адам қатысты.

Жас ерекшелігіне қарай 18-23 жас аралығында қатысқандар саны басым, яғни 68, 7 % болды.

Білім деңгейі бойынша 53, 3 % – магистр дәрежесін алғандар мен магистранттар, қалған топ бакалаврда білім алып жатқандар және аяқтағандар.



«Қазақ тілінің ұлттық корпусы жайлы естуіңіз бар ма?» деген сауалға сауалнамаға қатысушылардың 75 %-ы «иә», 25 %-ы «жоқ» деп жауап берсе, «Қазақ тілі ұлттық корпусын пайдаланасыздар ма?» деген сауалға жауап бергендердің үлес салмағы тең – 50 %-ды құрады.



Қатысушылардың 12, 5 %-ы ғана «Қазақ тілі ұлттық корпусын аударма жасау үшін пайдаландыңыздар ма?» деген сауалға «иә» деп жауап берген.

«Қазақ тілінің ұлттық корпусының мүмкіндіктері мен бүгінгі сапасын қалай бағалайсыз?» сауалына жауап бергендердің ішінде 18, 7 %-ы жоғары бағаласа, 25 %-ы – төмен, 56, 3 %-ы орташа деп бағалаған.

Сауалнама нәтижесінен мынаны байқаймыз: Ұлттық корпусымыздың түрлі мүмкіндіктері жоғары сұранысқа ие. Алайда, корпусты тек деректер базасы деп түсіну басым. Оның міндеттері мен мүмкіндіктері әлдеқайда мол. Оның ішінде машиналық аударма ісінде басты құрал болуы тиіс екенін ерекше атап өтуге болады. Өйткені Қазақ тілінің ұлттық корпусында қазіргі уақытта мәтінді өңдеудің санаулы түрі ғана жүзеге асырылады. Бұған қоса базалық та, техникалық та ақау көп. Машиналық аударманы жетілдіру үшін корпустың ішінде параллель корпустар құрылуы керек. Өкінішке орай, Қазақ тілінің ұлттық корпусында әзірге ондай мүмкіндік жоқ. Сондықтан бұл мәселе өзекті және оны шешу өте маңызды.

Зерттеу жұмысында жинақтау, талдау, жалпылау, топтастыру сияқты ғылыми, статистикалық әдіс-тәсілдер қолданылды.

Қорытынды

Қорытындылап айтқанда, корпусты біртұтас кешен деуге болады. Тілдік корпустарды түрлі сөздіктер (жиілік, түсіндірме, аймақтық, фразеологиялық, этнографиялық) құрас-

тыруда, ғылыми грамматикалар жазуда, оқыту жүйесінде және оқулықтар құрастыруда, әдістемелік құралдар жазуда, аударма жұмыстарында, тілдің барлық салаларында, т.б. қолдануға болады. Лингвистикалық корпустар тілдік зерттеулердің барлығын дерлік нақты фактологиялық материалдармен қамтамасыз етеді. Лингвистикалық корпустармен жұмыс жасау ғылыми-зерттеу жұмыстарын автоматтандырып, ғылымды дамытуға аса елеулі үлес қоса алады. Корпус – компьютерлік лингвистика саласының ірі жетістігі, аса сәтті құралы.

Мүдделер қақтығысы

Мүдделер қақтығысы жоқ.

Авторлардың қосқан үлесі

С.С. Шоқабаева мақаланың барлық бөліктерінің тұтастығына байланысты мәселелерді дұрыс зерттеуге және шешуге, мақаланың тұжырымдамасына, рәсімделуіне, мәтін мазмұнын сыни тұрғыдан қайта қарауға, жариялау үшін мақаланың соңғы нұсқасын бекітуге, жұмыстың барлық аспектілері үшін жауапты болуға келісім беруге елеулі үлес қосты. **Ф.Ж. Серікжан** зерттеу нәтижелерін жинау, талдауға, мәтін жазуға, сауалнама жүргізуге елеулі үлес қосты.

Әдебиеттер тізімі

1. Жұбанов А.Қ., Жаңабекова А.Ә. Корпустық лингвистика. – Алматы: «Қазақ тілі» баспасы, 2016. – 336 б.
2. Finegan E. Language: its structure and use. – New York: Harcourt Brace College Publishers, 2004. – P. 573.
3. McEnery T., Wilson A. Corpus Linguistics. – Edinburgh: Edinburgh University Press, 1996. – 209 p.
4. Рыков В.В. Корпус текстов как реализация объектно-ориентированной парадигмы //Труды Международного семинара Диалог-2002. – Москва: Наука, 2002. – С. 12-13.
5. Захаров В.П. Корпусная лингвистика: Учебно-метод. пособие. – Санкт-Петербург, 2005. – 48 с.
6. Stubbs M. Corpus analysis: The state of the art and three types of unanswered questions / G. Thompson & S. Hunston (eds.). System and Corpus. – London: Equinox, 2006. – S. 15-36
7. Жаңабекова А. Тілдік корпустардың ғылыми зерттеу әлеуеті // Тілтаным ғылыми журналы. – 2015. – №2. – Б.110-114
8. Biber D., Reppen R. Benchmarks in language and linguistics, 1, Sage Publications, pp. 23-45, 2012, Lexical Studies. <https://u-paris.hal.science/hal-01220026/document>

С.С. Шокабаева*¹, Ф.Ж. Серикжан²

¹Евразийский национальный университет им. Л.Н. Гумилева, Астана, Казахстан

²Специализированная школа-лицей-интернат "Дарын" управления образования Карагандинской области, Караганда, Казахстан

Конструктивный характер и функции корпуса

Аннотация. Одно из направлений моделирования языка – проведение исследований по машинному переводу – сегодня приобретает все большее значение и востребованность. Цель исследования – дать представление о корпусе, параллельном корпусе, их назначении и функциях, особенностях строения, проанализировать роль корпусов в деле машинного перевода. Основные направления – информативный обзор основных условий, общей структуры создания корпуса. Идея состоит в том, чтобы предложить способы эффективного использования возможностей компьютерной лингвистики в деле машинного перевода. Данные по структуре корпуса, алгоритму работы, деятельности могут служить качественной теоретической и практической основой для крупных и малых исследовательских работ на тему машинного перевода. Методы исследования: теоретический анализ, обобщение, анкетирование. Основные результаты, анализ, выводы исследования: из итогов анкетирования следует, что востребованы различные возможности Национального корпуса казахского языка. Известно, что у корпуса гораздо больше задач и возможностей. Особо хотелось бы отметить, что машинный перевод должен стать ведущим инструментом в деле. Потому что в Национальном корпусе казахского языка сейчас осуществляется лишь несколько видов обработки текста. В нем много дефектов, как базовых, так и технических. Чтобы внести свой вклад в машинный перевод, внутри корпуса должны быть созданы параллельные корпуса. Значение проведенного исследования вносит вклад в развитие Национального корпуса казахского языка. Практическая значимость результатов работы: данные по структуре корпуса, алгоритму работы, функции могут служить основным руководством для будущих исследователей при выполнении практических работ.

Ключевые слова: компьютерная лингвистика, корпус, состав корпуса, Национальный корпус, сборка корпуса, лексикографические исследования.

S.S. Shokabayeva*¹, F.Zh. Serikzhan²

¹L.N. Gumilyov Eurasian National University, Astana, Kazakhstan

²Specialized boarding school-lyceum "Daryn" of the education department of the Karaganda region, Karaganda, Kazakhstan

The constructive nature and functions of the corpus

Abstract. The study aims to explore the concept of corpora and parallel corpora, their purpose, structural features, and role in machine translation. It provides an overview of corpus creation and proposes ways to leverage computational linguistics effectively for machine translation. Information on corpus structure and operational algorithms serves as a solid theoretical and practical foundation

for research in this field. The research employs methods such as theoretical analysis, generalization, and questionnaires. The main findings indicate a high demand for the National Corpus of the Kazakh Language's capabilities. However, the corpus currently supports only a limited range of text processing functions and has several basic and technical shortcomings. Machine translation should become a key tool in enhancing the corpus's functionality. To advance this, it is essential to develop parallel corpora within the existing framework. The study significantly contributes to the improvement of the National Corpus of the Kazakh Language. Its practical value lies in offering guidance on corpus structure, algorithms, and functionality, which can aid future researchers in conducting practical work on machine translation.

Key words: computational linguistics, corpus, corpus composition, National Corpus, corpus assembly, lexicographic studies.

References

1. Zhubanov A.K., Zhanabekova A.A. Korpustyk lingvistika [Corpus linguistics] (Kazakh Tili, Almaty, 2016, 336 p.) [in Kazakh]
2. Finegan E. Language: its structure and use (Harcourt Brace College Publishers, New York, 2004, 573)
3. McEnery T., Wilson A. Corpus Linguistics (Edinburgh University Press, Edinburgh, 1996, 209 p.)
4. Rykov V.V. Korpus tekstov kak realizacija obiektno-orientirovannoy paradigmy [Text corpus as a realization of object-oriented paradigm], Trudy Mezhdunarodnogo seminar Dialog-2002 [Proceedings of the International Seminar Dialogue-2002], 12-13 (2002) [in Russian]
5. Zaharov V.P. Korpusnaya lingvistika: Uchebno-metod. posobie. [Corpus Linguistics: Teaching aid] (Saint Petersburg, 2005, 48 p.) [in Russian]
6. Stubbs M. Corpus analysis: The state of the art and three types of unanswered questions. In: G. Thompson & S. Hunston (eds.). System and Corpus. – (Equinox, London, 2006, 15-36)
7. Zhanabekova A. Tildik korpustardyk gylimi zertteu aleueti [Scientific research potential of language corpora], Tiltanym, 2, 110-114 (2015) [in Kazakh]
8. Biber D., Reppen R. Benchmarks in language and linguistics. Lexical Studies, 1, Sage Publications, 23-45 (2012). – Available at: <https://u-paris.hal.science/hal-01220026/document>

Авторлар туралы мәлімет:

Шоқабаета С.С. – хат-хабар үшін автор, филология ғылымдарының кандидаты, доцент м.а., Л.Н.Гумилев атындағы Еуразия ұлттық университеті, Астана, Қазақстан. E-mail: shokabaeva@yandex.kz, ORCID: 0000-0003-2705-276X.

Серікжан Ф.Ж. – педагогика ғылымдарының магистрі, Қарағанды облыстық білім басқармасының “Дарын” мамандандырылған мектеп-лицей-интернаты, Қарағанды, Қазақстан. E-mail: farida.serikzhan@mail.ru, ORCID: 0009-0009-1033-9079.

Шокабаева С.С. – автор для корреспонденции, кандидат филологических наук, и.о. доцента, Евразийский национальный университет имени Л.Н. Гумилёва, Астана, Казахстан. E-mail: shokabaeva@yandex.kz, ORCID: 0000-0003-2705-276X.

Серікжан Ф.Ж. – магистр педагогических наук, специализированная школа-лицей-интернат «Дарын» Карагандинского областного управления образования, Караганда, Казахстан. E-mail: farida.serikzhan@mail.ru, ORCID: 0009-0009-1033-9079.

Shokabayeva S.S. – corresponding author, Candidate of Philology, Acting Associate Professor, L.N. Gumilyov Eurasian National University, Astana, Kazakhstan. E-mail: shokabaeva@yandex.kz, ORCID: 0000-0003-2705-276X.

Serikzhan F.Zh. – Master of Education, Specialized Boarding Lyceum School "Daryn" of the Karaganda Regional Education Department, Karaganda, Kazakhstan. E-mail: farida.serikzhan@mail.ru, ORCID: 0009-0009-1033-9079.



Copyright: © 2025 by the authors. Submitted for possible open access publication under the terms and conditions of the Creative Commons Attribution (CC BY NC) license (<https://creativecommons.org/licenses/by-nc/4.0/>).