Торайғыров университетінің
ҒЫЛЫМИ ЖУРНАЛЫ

НАУЧНЫЙ ЖУРНАЛ
Торайгыров университета

ТОРАЙҒЫРОВ УНИВЕРСИТЕТІНІҢ ХАБАРШЫСЫ

Филологиялық серия
1997 жылдан бастап шығады

TORAIGHYROV UNIVERSITY

ВЕСТНИК ТОРАЙГЫРОВ УНИВЕРСИТЕТА

Филологическая серия
Издается с 1997 года

ISSN 2710-3528

№3 (2024)

Павлодар

НАУЧНЫЙ ЖУРНАЛ
ТОРАЙГЫРОВ УНИВЕРСИТЕТА

Филологическая серия
выходит 4 раза в год

СВИДЕТЕЛЬСТВО
О постановке на переучет периодического печатного издания,
информационного агентства и сетевого издания
№ KZ30VPY00029268
выдано
Министерством информации и общественного развития
Республики Казахстан

Тематическая направленность
публикация материалов в области филологии

Подписной индекс – 76132

https://doi.org/10.48081/NCYE9704

ҒТАМР 16.01.11

https://doi.org/10.48081/ZRHT9356

Г. Ә. Сәрсеке
Л. Н. Гумилев атындағы Еуразия
ұлттық университеті,
Қазақстан Республикасы, Астана қ.
*e-mail: sarseke_ga@enu.kz
ORCID: https://orcid.org/0000-0002-1670-5131

КОРПУС ЛИНГВИСТИКАСЫ ТЕРМИНДЕРІН ҚАЗАҚ ТІЛІНДЕ ҚАЛЫПТАСТЫРУ ЖӘНЕ БІРІЗДЕНДІРУ

Мақалада корпус лингвистикасы терминдерін қазақ тілінде қалыптастыру мәселесі қарастырылады. Зерттеудің мақсаты – қазақ тіліндегі корпус лингвистикасымен байланысты терминологиялық лексиканы ғылыми және оқу әдебиеттері материалдары негізінде салыстырмалы талдау. Бұл зерттеу жұмысының міндеттері – корпус лингвистикасы терминдерінің жасалу жолдарын талдау, сондай-ақ корпустарды және олармен жұмыс жасау кезеңдерін сипаттауда пайдаланылатын қолданыстағы терминдерді салыстыру, терминдерді біріздендіру және қазақ тілінің терминологиялық қорын кеңейту бойынша ұсыныстар жасау. Қазақ тілінің корпус лингвистикасы терминдеріне талдау бұл саладағы сөздердің негізінен ағылшын тілінен енген кірме терминдер екенін көрсетті. Сонымен қатар корпус лингвистикасының кейбір ұғымдарын атауда қазақ тілінде бұрыннан қалыптасқан төл терминдердің де қолданылатынын көрсетті. Дегенмен төл терминдерді жұмсауда әртүрлілікті жойып, біріздендіру қағидатын ұстану қажет. Қазақ тіл білімінде корпус лингвистикасының зерттелуі соңғы онжылдықта ғана басталып, дамып келе жатқанын ескерсек, сала терминдерінің қалыптасуы белгілі бір уақытты қажет етеді және зерттеулер саны артқан сайын сапасы да артады. Мақала корпус лингвистикасы терминдерін дұрыс қалыптастырудың, саланың ұғымдарын аударуда мүмкіндігінше тілде бар ұлттық терминдерді орнықтырудың және қолданыстағы терминдерді бірізді жұмсаудың маңыздылығына баса назар аударады.

Кілтті сөздер: корпус, корпус лингвистикасы, қазақ тілі, термин, аударма.

Кіріспе

Корпус лингвистикасы – қазақ тіл білімі үшін салыстырмалы түрде алғанда жаңа сала. Қазақ тілінің корпустары 2010 жылдардан бері жасала бастады. Алғаш 2012 жылы Алматы Қазақ тілі корпусын құру ісі басталды [1]. Кейіннен Қазақ тілінің ұлттық корпусын құру ісі қолға алынды [2]. Соңғы жылдарда Қазақ сөйлеу тілінің корпустарын құру ісі жанданды [3]. Қазірде қазақ тіл білімінде корпустар құру тәжірибесі бар, сондай-ақ корпус лингвистикасының негізгі терминдері мен ұғымдарының сөздігі бар [4]. Сонымен қатар корпусқа негізделген талдаулар, корпустарды түрлі зерттеулер жүргізуде және тілдерді оқытуда пайдалану аз да болса бар. Корпус лингвистикасын пән ретінде оқыту да жоғары оқу орындарында педагогикалық және гуманитарлық ғылымдардың білім беру бағдарламаларында іске асырылып келеді. Зерттеулер жүргізу және оқыту ісі тәжірибеге енгеннен соң, оның ұғым-түсініктерін дұрыс қалыптастыру және мүмкіндігінше ұлттық терминдер негізінде орнықтыру да маңызды. Сала қазақ тіл білімі үшін біршама жаңа, соған орай жарық көретін зерттеулер саны әлі де көбейеді, сол қалыптасу және даму процесінде корпус лингвистикасының ұғымдары мен терминдерінің мейлінше дәлді берілуі және аударылуы маңызды. Корпус лингвистикасының терминдерін дұрыс қалыптастырып жүрміз бе деген сұрақ менің зерттеуімнің басты сұрағы болмақ. Мақаланың мақсаты – ғылыми және оқу-әдістемелік материалдар негізінде қазақ тіліндегі корпус лингвистикасымен байланысты терминологиялық лексиканы салыстырмалы талдау. Бұл зерттеудің міндеттері корпус лингвистикасы терминдерінің жасалу тәсілдерін талдау, сондай-ақ корпустарды және олармен жұмыс жасау кезеңдерін сипаттау үшін қолданылатын бұрыннан қалыптасқан терминдерді салыстыру, терминдерді біріздендіру және қазақ тілінің терминологиялық қорын кеңейту жөнінде ұсыныстар енгізу. Мақалада корпус лингвистикасының кейбір терминдерінің мағыналары түсіндіріле отырып, олардың қазақ тіліне аударылуы және қазақ корпус лингвистикасының терминологиялық қорының қалыптасуы талданады.

Материалдар мен әдістері

Зерттеудің материалдарын таңдау критерийлері бойынша корпус лингвистикасы саласында жарық көрген негізінен қазақ және орыс тіл біліміндегі басты ғылыми және оқу-әдістемелік зерттеулер іріктеліп алынды. Ақпарат көздері корпус лингвистикасының терминологиясына қатысты ағылшын және қазақ тілдеріндегі сөздіктер, ағылшын тіліндегі корпус лингвистикасының глоссарийі, қазақ тіліндегі корпустар орналастырылған веб-сайттар, қазақ тіліндегі терминдерді біріздендіретін веб-сайт болып табылды. Барлық дереккөздерді іздеу стратегиялары корпус лингвистикасы

саласында жарияланған қазақ, орыс және ағылшын тілдеріндегі ғылыми зерттеулерді табуға және корпус лингвистикасының терминдерін түрлі дереккөздерден, жоғарыда аталған ақпарат көздерін қоса алғанда, жекежеке іздеуге құрылды. Зерттеу әдісі аталған дереккөздерді және жекелеген терминдерді сыни талдау болып табылады.

Нәтижелер және талқылау

Ең алдымен, пәннің қазақ тіліндегі атауына қатысты ой өрбітсек. Мақаланың атауынан да, кіріспеден де байқауға болады, мен бұл саланың, сәйкесінше пәннің атауын «корпус лингвистикасы» деп атауды жөн санаймын. Бұл саланың атауы қазақ тілді дереккөздерде «корпустық лингвистика» деп аталып жүр. Бұл орыс тіліндегі «корпусная лингвистика» атауынан тікелей аударылған. Бұл терминнің ағылшын тіліндегі атауы – «corpus linguistics». «Корпус» сөзі тек зат есім. Бұл сөзде сындық мағына жоқ. Ағылшын тіліндегі сөздіктерде де реестр сөз ретінде тек зат есім тұлғасында тіркелген: британдық сөздікте *«corpus*, зат есім, жазбаша және ауызша мәтіндердің жинағы» [5] немесе американдық сөздікте «corpus, зат есім, тіл білімі. Репрезентативті болып есептелетін және лексикалық, грамматикалық немесе басқа лингвистикалық талдау үшін қолданылатын сөздер немесе сөйлемдер түріндегі айтылыстар жиынтығы» [6]. Түпнұсқа терминді қазақ тіліне «корпустық» деп аударатындай бұл сөзде сындық мағына жоқ. Термин «корпус лингвистикасы» деп аударылуы қажет. Лингвистиканың қай түрі десек, корпусты зерттейтін лингвистика түсінігі шығады. Корпус лингвистикасының басты нысаны – корпус, сала корпусқа қатысты мәселелерді қарастырады. Сонда пәннің атауын орыс тіліндегі нұсқасынан емес, ағылшын тіліндегі тұпнұсқа терминнен тікелей аударсақ, әлдеқайда дұрыс болады. «Корпусная» сөзі орыс тілінің құрылымдық ерекшелігінен туындаған. Орыс тілінде зат есіммен тіркесе келген сын есім сөз сол зат есімнің тұлғасына бейімделіп, өзгеретіні белгілі: *корпусная лингвистика, корпусной лингвистике* т.т. Ағылшын тілінде олай емес, екі түбір зат есім қатар келеді, қазақ тілінде де дәл солай. Қазақ тілінде екі түбір зат есім қатар келіп, анықтаушы-анықталушы қатынаста қатыстық сындық мағына білдіре алады. «Корпус лингвистикасы» атауының дұрыстығының басты дәлелі кілт терминнің – «корпус»-тың - пәннің негізгі зерттеу нысанын дәл ашуында: бұл пәннің зерттеу нысаны – корпус. Корпус дегеніміз мәтіндердің электронды жинағы. Корпус құрылған соң, онда әртүрлі лингвистикалық амалдар жасалады. Кэмбридж сөздігімен айтсақ, корпус дегеніміз тілдік деректер қоры, «жазбаша немесе ауызша материалдардың жинағы, ол компьютерде сақталады және тілдің қалай қолданылатынын білу мақсатында пайдаланылады» [7]. Ендеше, корпусты зерттейтін сала *корпустық лингвистика* емес, *корпус лингвистикасы* деп аталады.

Салыстырмалы түрде басқа тілдерде бұл термин қалай аударылды деп зерттесек, бурят және халха-моңғол тілдерінде ешбір қосымшасыз түбір күйінде жұмсалатынын байқадық. Моңғол тіліндегі «корпус лингвистикасы» терминінің жұмсалуына қатысты Ю. Абаеваның келтірген мысалдарында бурят тілінде де, халха-моңғол тілінде де «корпус» болып қолданылатынын көреміз. Қараңыз: *«Буряад хэлэнэй корпус* 'Корпус бурятского языка'; бур. *корпус хэлэ шэнжэлэлгэ,* х.-монг. *корпус хэл шинжлэл* 'корпусная лингвистика': х.-монг. 1980-аад оноос корпус хэл шинжлэл нь ихээхэн хурдацтай хөгжиж … / «С 1980-х гг. корпусная лингвистика очень быстро развивалась…» [8, 163]. Бурят және халха-моңғол тіл білімдерінде де жұмсалған термин – «корпус» («корпустық» емес), яғни ағылшын тілінен (орыс тілі арқылы емес немесе орыс тіліндегі аудармасы негізінде емес) тікелей аударылған, сөйтіп тіл білімі саласының атауы «корпус лингвистикасы» деп қолданылады. Айта кетер жайт, халха-моңғол тілінде «корпус» кірме сөзінен гөрі оның ұлттық баламалары жиі жұмсалады екен. Бұл тілдерде «корпус» сөзін монғолдың «хөмрөг» және «сан» сөздері алмастырады. Қараңыз: «БНХАУ-ын монголчуудын хувьд хөмрөг байгуулалтын ажил нь нэлээд эрт эхэлж… / «Что касается монголов КНР, то работа по созданию корпуса у них началась намного раньше…»; «*Компьютер хэл шинжлэл, тэр дундаа хөмрөгийн хэл шинжлэлийн судалгаа харьцангуй өндөр түвшинд хийгдэж буй* / «Компьютерная лингвистика, в частности исследования корпусной лингвистики, находится на относительно высоком уровне» [8, 163]. Салыстырмалы түрде бурят тілінде кірме сөз нұсқасы ғана қолданылады екен: «корпус хэлэ шэнжэлэлгэ» [8, 163]. «Корпус» мағынасында *сан* сөзі терминденіп, *материалын сан, хэлний материалын сан* ('досл. хранилище языкового материала') анықтауыштарымен қолданылады: «Например: Хятад улсын хувьд хэлний материалын сан байгуулах ажил 1980-аад оноос эхэлсэн байна / «Что касается Китая, то работа по созданию языкового корпуса началась в 1980-х гг.» [8, 163]. Бұл екі сөздің де білдіретін мағыналары ұқсас: «хөмрөг» сөзі «қойма (астық үшін), қамба, астық қоймасы; қазынашылық; қор» мағыналарын берсе, «сан» сөзі «қазына, қор, қамба» мағыналарын білдіреді [8, 163]. Зерттеуші мысалдарына қарасақ, осы екі сөздің ішінде «корпус лингвистикасы» атауының баламасы болып негізінен «хөмрөг», ал «корпус» сөзінің баламалары ретінде «хөмрөг» пен «сан» ұлттық терминдері жұмсалады. Ендеше, «корпус» терминіне қазақша балама ұсынып, қолданыста орнықтыру мүмкіндігі бар.

Алдыңғы терминмен ұқсас аударылып қолданысқа енген келесі атау – «корпустық менеджерлер» термині. Бұл да «корпустық лингвистика» термині сияқты орыс тілінен тікелей аудармамен алынған атау. В. Захаров пен С. Богданова оқулығында [9] «корпусные менеджеры» атауы жұмсалған да, сол оқулықтың құрылымы мен мазмұнымен бірдей шыққан А. Жұбанов

пен А. Жаңабековалардың еңбегінде де [10] «корпустық менеджерлер», сөздікте де [4] дәл осы нұсқа берілген. Бұл терминнің дұрыс аудармасы «корпус менеджерлері» болуы керек. Бұл терминді де ағылшын тіліндегі түпнұсқа атаудан – «corpus manager» - тікелей алып, «корпус менеджер(лер) і» деп орнықтырған дұрыс. Корпус менеджері – мәтіндер корпустарын басқару үшін, яғни корпустарды құру, өңдеу, аннотациялау және іздеу үшін пайдаланылатын бағдарлама. *Корпусты басқару* сияқты балама сөз тіркесімімен алмастыра қарасақ та, «корпус менеджері» атауының дұрыстығы көрінеді. Осылардың ізімен «корпустық» болып кеткен басқа да атаулар мен қолданыстарды түзеткен дұрыс. Ондай атаулар қатары біраз: *корпустық деректер, корпустық тексеру, корпустық материал* т.т. Бұларды *корпус деректері, корпус тексеруі, корпус материалы* деп аударып орнықтырған дұрыс және қазақ тілінің құрылымына сай, сонымен бірге терминнің түп мағынасын жеткізуде дәлдірек.

Корпус лингвистикасы терминдерін ұлттық тілде беруде және оларды орнықтыруда талқылануы тиіс атаулардың бірі – «mark-up/markup». «Mark-up» – корпус файлына ақпарат қосуды білдіретін термин. Mark-up екі түрге бөлінеді: құжатты белгілеу және аннотациялар. Mark-up – корпус жасауда жүзеге асырылатын аннотация процесінің маңызды бөліктерінің бірі. Ағылшын тіліндегі «mark-up» термині орысша «разметка» деп аударылса, қазақ тілінде «белгі», «белгіленім», «белгілену», «белгілеу» сияқты мағыналары бір-бірінен алшақ түспейтін сөздермен түрліше беріліп жүр. Орыс тілінде *понятие разметки, типы разметки, морфологическая разметка, синтаксическая разметка, семантическая разметка, анафорическая разметка, просодическая разметка, экстралингвистическая разметка* [9] т.т. терминдік тіркестерде жұмсалса, қазақ тілінде белгіленім ұғымы, белгіленім түрлері, морфологиялық белгіленім, синтаксистік *белгіленім, семантикалық белгіленім, анафоралық белгіленім, просодикалық белгіленім, экстралингвистикалық белгіленім* [10] т.т. тіркестерде қолданылған. Монография авторлары [10] корпус лингвистикасының бұл терминіне бір ғана «белгіленім» атауын ұсынса, сөздіктен [4] бұл сөздің төрт түрлі аудармасын ұшыраттық. Бірі – «белгі»: Корпустың британдық компоненті (ICE-GB) толығымен дайындалған, оның мәтіндері морфологиялық және синтаксистік белгілермен жабдықталған, Анафорикалық белгі – референттік,.. [4, 8], лингвистикалық белгі [4, 29]; екіншісі – «белгілену»: белгіленуі бойынша зерттеуге... [4, 13], морфологиялық белгілену, мәтіннің морфологиялық белгіленуі [4, 29]; үшіншісі – «белгілеу»: егер анафорикалық, просодикалық белгілеулер автоматты жүйелерді жасау өте қиын және негізгі бөлігі жұмыс қолмен жүргізілсе ... [4, 13], Лингвистикалық белгілеу үдерісіндегі сөздерге арналған код [4, 25], автоматты түрде морфологиялық белгілеу, автоматты

белгілеу [4, 29]; төртіншісі – «белгіленім»: Лингвистикалық белгіленім типіне қарай корпустарды бөлу... [4, 12], Синтаксистік белгіленімді қамтитын корпус лексикалық... [4, 12], Лингвистикалық белгіленім (лингвистическая разметка) [4, 18] т.т. Сөздікте «белгіленім» сөзі «белгі» сөзімен синонимдес те жұмсалған: «Лингвистикалық белгіленім (лингвистическая разметка) – лингвистикалық бірлік процесі негізіндегі лингвистикалық сипаттамалар атрибуциясы. Лингвистикалық белгілердің келесі түрлері бөлінеді: морфологиялық, синтаксистік, семантикалық, анафориялық, просодикалық, дискурстық және т.б.» [4, 18]; «Морфологиялық белгіленім (part-of-speech tagging (POS-tagging) – бұл сөйлеу бөлігінің белгісі ғана емес, сонымен қатар сөйлеу бөлігіне тән грамматикалық санаттардың белгілерін қамтитын морфологиялық белгі, лингвистикалық таңбаның негізгі түрі» [4, 20]. Байқап отырғанымыздай, бір ғана терминге қазақ тілінде бірнеше баламалар жарыса қолданылып жүр. Қазақ тілімен салыстырғанда бұл терминнің орыс тіліндегі көрінісі бірізді: «разметка» термині орныққан. Ендеше, бірнеше жарыспалы аудармалардың ішінен біреуін ғана орнықтыру қажет. Осы жарыспалы терминдердің ішінен «белгілеу» терминін орнықтырған дұрыс. «Белгіленім» сөзі терминнің мағынасын жеткізіп тұр, дегенмен -ім жұрнағымен тағы бір жасанды сөз (айтылым, жазылым, тыңдалым т.т. сияқты) жасағанша, тілде бұрыннан бар «белгілеу» сөзін пайдаланған дұрыс. Тағы бір дәлел мынада. «Mark-up» та, «белгілеу» де негізінен информатика, ақпараттық технология саласының терминдері. «Белгілеу» термині – информатика және есептеуіш техника саласында бұрыннан бекіген термин [11]. Корпус лингвистикасының аталған салаларымен тығыз байланысты екенін ескерсек, қазақ тіліндегі термин атауы «белгілеу» болып қалғаны дұрыс.

Корпус лингвистикасының талқылауға тиіс тағы бір термині – «лемма» және одан туындаған сөздер. Лемма – сөздің сөздікке енгізілетін бастапқы тұлғасы. Осы терминнен туындаған зат есім сөз ағылшын тілінде – «lemmatisation», орыс тілінде «лемматизация» деп аталады. Бұл ұғым қазақ тілінде корпус лингвистикасының терминдері мен түсініктері сөздігінде «леммалдандыру» (– «бұл басқа сөздік формаларынан сөздің өзіндік формасын қалыптастыру үдерісі») [4, 18] деп берілген болса, біріккен авторлықтағы еңбекте «лемматизация» және «леммаға ажырату» («яғни сөзформаларды алғашқы негіз сөз пішініне (формасына) келтіру әрекеті») [10, 92] болып аударылған және түсіндірілген. Орыс тіліндегі «лемматизация» ағылшын тіліндегі «lemmatisation» сөзінің дәлме дәл калькалануы. Сөздіктегі «леммалдандыру» баламасын қабылдау қиынырақ, оның орнына «леммалау» десек те, терминнің мағынасын беруге болады. Леммалау – берілген сөздің леммасын анықтаудың алгоритмдік процесі, мысалы, *барған, барса, бармақ* сөз тұлғаларын *бар* леммасына дейін

қысқарту. «Леммаға ажырату» аудармасы сырттай қарағанда терминнің мағынасын ашатын сияқты, дегенмен оның ішкі қайшылығы бар. «Леммаға ажырату» аудармасының қайшылығы терминнің мәнін дәл бере алмауында. Леммалау леммаға ажырату процесі емес, лемманы табу немесе анықтау процесі. Және бұл процесс – машинада автоматты орындалатын процесс. Егер біз корпустың іздеу жолына «барған» сөзін теріп іздесек, «бар» леммасын табуымыз екіталай. Егер іздеуге «бар» леммасын қойсақ, онда «бар» леммасынан тараған барлық сөздер шығады. Леммалау сөзді түбірі мен қосымшаларына ажырату емес, сөзді сөздік формасы түбірге дейін, яғни леммаға дейін қысқарту. Леммалаудың мақсаты - әрбір сөздің флексия түрлерін ортақ негізге немесе түбірге келтіру. Бұл сөздің етістік тұлғасы да ағылшын тілінде «lemmatise/lemmatize» сөзімен белгіленеді. Егер «леммалау» терминін зат есім ретінде орнықтырсақ, онда корпус лингвистикасында оны қимыл (етістік) мағынасында да жұмсай аламыз.

Қазақ корпус лингвистикасына енген кірме терминдердің бірі – «стем» сөзі. «Стем» – ағылшын тіліндегі «stem» сөзінің дәлме-дәл аудармасы. Орыс тілінде де «стем». В. Захаров пен С. Богданова орыс тілінің корпус лингвистикасында «стем» терминін қолданып, ұғымды түсіндіруде оны «основа» деп айқындайды: «Процесс, несколько отличный от лемматизации, называется ***стеммингом,*** он состоит в нахождении стема (основы) слова» [9, 39]. Бұл термин «Корпустық лингвистика: негізгі терминдер мен түсініктердің оқу сөздігінде» «стем сөз» деп аударылған: ***«Стем Сөз*** – сөздің негізі, сөздің өзгермейтін бөлігі» [4, 24]. Стем деген не? Алдымен бұл терминнің мағынасын ашып алайық. Стем - ағылшын тілінде сөздің флективті аффикстерін алып тастағанда қалатын негізгі бөлігі. Мысалы, «stays», «stayed», «staying» сөз тұлғаларының стемі «stay» сөзі, «writes», «writing» және «written» сөз тұлғаларының стемі «writ» болып табылады. Ағылшын тілінде деп айтуымыздың себебі бар, себебі ағылшын тілінде «stem», «root», «base» терминдерінің өз ерекшеліктері бар [12]. Стем түбірмен сәйкес түсуі мүмкін: стем=түбір, сәйкес түспеуі де мүмкін: стем=түбір+(аффикс). Мысалы, «әнші» сөзінің стемі «ән» және түбірі де «ән». Бұл жағдайда стем мен түбір сәйкес түседі. *Әншінің, әншіге, әншімен* сөздерінің стемі «әнші», ал түбірі – «ән». Бұл мысалдарда стем түбір және аффикстен тұр. *Стем* сөздің *негізіне* ұқсайды. Бірақ ағылшын тілінде *стем* мен *негіз* ажыратылады. Ол ажырату аффикстерге қатысты болады: егер сөзге аффикстің флективті түрі жалғанса, онда ол стем, егер деривациялық түрі жалғанса, онда ол негіз болады. *Құтқарушыларды, құртқарушымен, құтқарушының* сөздерінің стемі - *құтқарушы*, негізі - *құтқару*, түбірі - *құтқар.* Сондықтан «стем»-нің мағынасын қазақ тіл білімінде *түбір* де, *негіз* де бере алмайды. Корпус лингвистикасында «стем» сөзінен туындайтын «стемминг», «стеммер»

сөздері жұмсалады. Стемминг - сөзді стеміне дейін қысқарту процесі, яғни стемдеу. Ағылшын тілінің сөздіктерінде және «Корпус лингвистикасы глоссарийінде» бұл сөз жеке термин ретінде тіркелмейді [13]. Ол «стем» сөзінің өзгерген тұлғаларының бірі ғана: «present participle of stem» [14]. Ағылшын тілінің «stemmer» сөзінен калькаланған «стеммер» сөзі стемминг процесін іске асыратын құралды білдіреді. Стеммерлердің бірнеше түрлері бар.

Егер жоғарыда аталған терминдер қазақ тілі үшін кірме терминдер және олардың ұлттық тілде берілуі белгілі бір мөлшерде қиындық келтіретін болса, кейбір терминдер тілде бұрыннан бар сөздер мен сөз тіркестеріне негізделеді. Ондай терминдерді орнықтыруда жаңадан сөздер ойлаудың қажеті жоқ. Тілде бұрыннан бар аудармаларды пайдаланған жөн. Сондай терминдердің бірі – «дереккөз» және «көз». «Источник» сөзінің және «источник данных» тіркесті сөздің аудармасы ретінде екеуі де корпус лингвистикасында контекске байланысты қолданыла алады. Дереккөз – корпус құру ісінде қажетті маңызды ұғымдардың бірі. Корпус құру үшін түрлі дереккөздерден мәтіндер жиналады. «Дереккөз» қазақ тіл білімінде орнын тапқан термин, ендеше корпус лингвистикасында да осы сөзді орнықтырған дұрыс. Сөздікті [4] тексеріп шыққанда «источник» сөзінің аудармасы ретінде «дереккөз» термині ғана жұмсалғанын анықтадық. В. Захаров пен С. Богданова оқулығындағы «Отбор источников. Критерии отбора» тараушасының атауы А. Жұбанов пен А. Жаңабековалардың бірлескен монографиясында «Дерекнаманы іріктеу» деп аударылған [10]. Тілде «дереккөз» сөзі тұрғанда, оған жарыса нама-мен балама берудің қажеттілігі жоқ деп ойлаймын. Бұл екі сөздің – «дереккөз» бен «(деректер) көзі» – қолданысын контекске байланысты шешу керек. Айталық, «На этом этапе все тексты, полученные из разных источников, проходят филологическую выверку и корректировку» [9, 34] сөйлемін аударғанда «дереккөз», «... түрлі дереккөздерден алынған барлық мәтіндер...», сөзін жұмсау қажет. Ал мынадай сөйлемдерді – «Текстовые массивы Интернета широко используются как источник данных для формирования корпусов» [9, 71], «Корпусы являются богатым источником данных для исследований по лексикографии и грамматике» [9, 94] – қазақ тіліне «... корпустарды қалыптастыру үшін деректер көзі ретінде кеңінен қолданылады», «... зерттеулер үшін деректердің бай көзі ...» «деректер(дің) көзі» деп аударған орынды.

«Дерек» сөзімен келетін корпус лингвистикасындағы келесі бір атау - ағылшын тіліндегі «database» сөзі. Орыс тіліндегі аудармасы - «база данных». Бұл күрделі сөздің қазақ тіліндегі «деректер қоры» немесе «дерекқор» аудармаларын қолданып, орнықтыра беру жөн. Data – «деректер», base – «қор». «Дерекқор» – қазақ тілінде бұрыннан бар

термин [11]. Дегенмен корпус лингвистикасының терминдері мен түсініктерін анықтайтын сөздікте [4] «деректер базасы» деп берілген: *индекстік деректер базасын* [4, 23], *Іздеу деректер базасы (ағыл. index database)* [4, 27]. «База» сөзін «қор» сөзімен беру халха-моңғол тілінің корпус лингвистикасында да бар. Моңғол тілінде орыс тіліндегідей *«дата бааз»* деп жиі жұмсалады екен. Дегенмен халха-моңғол тілінде бұл терминді беруде ұлттық тілдің «хөмрөг» пен «сан» сөздері де қолданылады. «База» сөзін «қор» мағынасын беретін «хөмрөг» сөзі алмастырады: «дата хөмрөг». *«Одоо үеийн монгол хэлний дата хөмрөгийн программ. / «Программа для создания базы данных современного монгольского языка* [8, 163]. Сонымен қатар «сан сөзі де алмастырады: *«мэдээллийн сан* или *хэлний мэдээллийн сан,* где *мэдээлэл* ‘информация’, *мэдээллийн* ‘информационный’: … *database-ийг мэдээллийн сан хэмээн орчуулсан байна.* / «…database перевели как *«мэдээллийн сан»»* [8, 163]. Көріп отырғанымыздай, халха-моңғол тілінің төл сөздері «хөмрөг» пен «сан» «корпус» мағынасын берумен қатар, «база данных» терминінің мағынасын беруде де жұмсалады.

Қаржыландыру туралы ақпарат

Бұл зерттеуді Қазақстан Республикасы Ғылым және жоғары білім министрлігінің Ғылым комитеті қаржыландырды (Грант № AP23488585 «Цифрлық гуманитарлық ғылымдар: Академиялық қазақ тілінің корпусын әзірлеу»).

Қорытынды

Корпус лингвистикасы қазақ тіл білімінде соңғы онжылдықта қалыптасып, дамып келе жатқан сала болғандықтан, оның терминдерінің ұлттық тілде қалыптасуы да уақытты қажет етеді. Бұл салада қазақ тілінде жарық көрген монографиялар, оқулықтар, сөздіктер, зерттеулер ағылшын тілді зерттеулермен салыстырғанда аз. Қазақ тілінде шыққан зерттеулердің басымын қазақ тілінің ұлттық және сөйлеу корпустарын құру тәжірибесін ортаға салатын еңбектер қатары құрайды. Бұл салада әлі де жаңа зерттеулер саны артатыны сөзсіз. Саланың қалыптасу және даму кезеңдерінде корпус лингвистикасының ұғым-түсініктерін қазақ тіліне дәлді әрі бірізді аудару, сол арқылы кірме және ұлттық терминдерді дұрыс қалыптастыру маңызды. Екіұшты және жарыспалы терминдерді мүмкіндігінше болдырмау қажет. Әр термин ұғымның мағынасын дәл бере алатындай болу керек.

Корпус лингвистикасы терминдерінің қазақ тілінде берілуі қандай деген сұрақта негізінен терминдердің кірме сөздер екенін байқауға болады. Кірме терминдер белгілі бір ұғымдардың мағынасын дәл ашатын сөздер ұлттық сөздік қорда болмаған жағдайда пайдаланылады. Корпус лингвистикасының кейбір ұғымдарын беруде қазақ тілінің өз сөз әлеуеті де жетеді. Мақалада талданған терминдер кірме де, ұлттық сөздер қатарынан да болып отыр.

Терминдерді аудару ісіне орыс тілі арқылы келмеу керек. Терминдердің түпнұсқа тіліне қарау керек. Корпус лингвистикасындағы терминдердің басым көпшілігі ағылшын тілінде (арғы түбі – латын, грек сөздері) қалыптасқандықтан, оларды қазақ тіліне аударғанда сол түпнұсқа атауларды басшылыққа алу қажет. Әрине, аударғанда қазақ тілінің өз заңдылықтары мен ерекшеліктері ескерілетіні сөзсіз. Сонымен қатар корпус лингвистикасы пәнаралық сала болғандықтан, онда жұмсалатын кейбір атаулар ақпараттық технология, информатика және есептеуіш техника, тілді машинада өңдеу сияқты салаларға ортақ терминдер болып келеді. Мақалада талданған «белгілеу», «стеммер» сөздері соған дәлел. Сол себепті ортақ ұғымдардың аталған салаларда қалыптасқан атауларын корпус лингвистикасында да орнықтырған жөн. Ұлттық терминдерді бірнеше баламамен жарыспалы жұмсамай, бірізді қолданып, орнықтыру маңызды.

Пайдаланылған деректер тізімі

1 Алматы Қазақ тілі корпусы [Электронный ресурс]. – Сетевой режим доступа: http://web-corpora.net/KazakhCorpus/search/?interface_language=kz

2 Қазақ тілінің ұлттық корпусы [Электронный ресурс]. – Сетевой режим доступа: https://qazcorpus.kz/

3 Kazakh Speech Corpus 2 [Электронный ресурс]. – Сетевой режим доступа: https://issai.nu.edu.kz/kz-speech-corpus/

4 Корпустық лингвистика : негізгі терминдер мен түсініктердің оқу сөздігі [Текст] / құраст. : Г. Б. Мәдиева, С. Б. Бектемірова, Н. А. Исмайлова. – Алматы : Қазақ университеті. – 2018. – 40 б.

5 Corpus. Oxford Learner's Dictionaries [Электронный ресурс]. – Сетевой режим доступа: https://www.oxfordlearnersdictionaries.com/definition/english/corpus?q=corpus

6 Corpus. Dictionary.com [Электронный ресурс]. – Сетевой режим доступа: https://www.dictionary.com/browse/corpus.

7 Corpus. Cambridge Dictionary [Электронный ресурс]. – Сетевой режим доступа: https://dictionary.cambridge.org/dictionary/english/corpus

8 **Абаева, Ю.** Термины корпусной лингвистики в халха-монгольском и бурятском языках [Текст] // Филологические науки. Вопросы теории и практики. – 2019. – № 10. – Т. 12. – С. 162–166.

9 **Захаров, В. П., Богданова, С. Ю.** Корпусная лингвистика : учебник для студентов гуманитарных вузов [Текст]. – Иркутск : ИГЛУ. – 2011. – 161 с.

10 **Жұбанов, А. Қ., Жаңабекова, А. Ә.** Корпустық лингвистика : монография [Текст]. – Алматы, «Қазақ тілі» баспасы. – 2017. – 336 б.

11 Termincom.kz [Электронный ресурс]. – Сетевой режим доступа : https://termincom.kz/search/?termin=%D0%A0%D0%B0%D0%B7%D0%BC%D0%B5%D1%82%D0%BA%D0%B0&cid=0

12 **Payne, Th.** *Exploring Language Structure: A Student's Guide* [Текст]. – Cambridge : Cambridge University Press,. – 2006. – 365 p.

13 **Baker, P., Hardie, A., McEnery, T. A.** Glossary of Corpus Linguistics [Текст]. – Edinburgh : Edinburgh University Press Ltd. – 2006. – 192 p.

14 Stemming. Cambridge Dictionary [Электронный ресурс]. – Сетевой режим доступа: https://dictionary.cambridge.org/dictionary/english/stemming

References

1 Almaty` Kazak tili korpusy` [E`lektronny`j resurs] [Almaty Corpus of the Kazakh language]. – Setevoj rezhim dostupa: http://web-corpora.net/KazakhCorpus/search/?interface_language=kz

2 Kazak tilinin ultty`k korpusy` [E`lektronny`j resurs] [National Coprus of the Kazakh language]. – Setevoj rezhim dostupa: https://qazcorpus.kz/

3 Kazakh Speech Corpus 2 [E`lektronny`j resurs]. – Setevoj rezhim dostupa: https://issai.nu.edu.kz/kz-speech-corpus/

4 Korpusty`k lingvistika: negizgi terminder men tusinikterdin oku sozdigi [Corpus Linguistics : A Dictionary of Key Terms and Concepts] [Text]. – Almaty` : Kazak universiteti, 2018. – 40 p.

5 Corpus. Oxford Learner's Dictionaries [E`lektronny`j resurs]. – Setevoj rezhim dostupa : https://www.oxfordlearnersdictionaries.com/definition/english/corpus?q=corpus

6 Corpus. Dictionary.com [E`lektronny`j resurs]. – Setevoj rezhim dostupa: https://www.dictionary.com/browse/corpus

7 Corpus. Cambridge Dictionary [E`lektronny`j resurs]. – Setevoj rezhim dostupa: https://dictionary.cambridge.org/dictionary/english/corpus

8 **Abaeva, Yu.** Terminy` korpusnoj lingvistiki v xalxa-mongol`skom I buryatskom yazy`kax [The terms of corpus linguistics in the Khalkha-Mongolian and Buryat languages] [Text]. – Almaty` : Kazak universiteti, 2018. – 40 p.

9 **Zaxarov, V. P., Bogdanova, S. Yu.** Korpusnaya lingvistika : uchebnik dlya studentov gumanitarny`x vuzov [Corpus linguistics : a textbook for students of humanities at universities] [Text]. – Irkutsk : IGLU, 2011. – 161 p.

10 **Zhubanov, A. K., Zhanabekova, A. A`.** Korpusty`k lingvistika : monografiya [Corpus Linguistics : a monograph] [Text]. – Almaty`, «Kazak tili» baspasy`, 2017. – 336 p.

11 Termincom.kz [E`lektronny`j resurs]. – Setevoj rezhim dostupa : https://termincom.kz/search/?termin=%D0%A0%D0%B0%D0%B7%D0%BC%D0%B5%D1%82%D0%BA%D0%B0&cid=0

12 **Payne, Th.** *Exploring Language Structure: A Student's Guide*. – Cambridge : Cambridge University Press, 2006. – 365 p.

13 **Baker, P., Hardie, A., McEnery, T. A.** Glossary of Corpus Linguistics. – Edinburgh : Edinburgh University Press Ltd, 2006. – 192 p.

14 Stemming. Cambridge Dictionary [E`lektronny`j resurs]. – Setevoj rezhim dostupa : https://dictionary.cambridge.org/dictionary/english/stemming

28.01.23 ж. баспаға түсті.
09.04.24 ж. түзетулерімен түсті.
26.08.24 ж. басып шығаруға қабылданды.

Г. А. Сарсеке
Евразийский национальный
университет имени Л. Н. Гумилева,
Республика Казахстан, г. Астана.
Поступило в редакцию 28.01.23.
Поступило с исправлениями 09.04.24.
Принято в печать 26.08.24.

ФОРМИРОВАНИЕ И УНИФИКАЦИЯ ТЕРМИНОВ КОРПУСНОЙ ЛИНГВИСТИКИ В КАЗАХСКОМ ЯЗЫКЕ

В статье рассматривается вопрос формирования терминов корпусной лингвистики на казахском языке. Цель исследования – сравнительный анализ терминологической лексики, связанной с корпусной лингвистикой, в казахском языке на материале научной и учебной литературы. Задачи данного исследования заключаются в анализе способов образования терминов корпусной лингвистики, а также сравнении уже существующих терминов, используемых при описании корпусов и этапов работы над ними, внесении предложений по унификации терминов и расширению терминологической базы казахского языка. Анализ терминов корпусной лингвистики казахского языка выявил, что слова данной области в основном представляют собой заимствованные термины из английского языка. Кроме того, было показано, что при наименовании некоторых понятий корпусной лингвистики используются и национальные термины, уже сформировавшиеся в казахском языке. Однако при использовании национальных терминов необходимо исключать

многообразие и придерживаться принципа унификации. Учитывая, что исследования корпусной лингвистики в казахском языкознании начались и развились лишь в последнее десятилетие, формирование терминов в данной области требует определенного времени, а по мере увеличения количества исследований растет и качество. В статье подчеркивается важность правильного формирования терминов корпусной лингвистики, максимального закрепления национальных терминов в языке при переводе понятий данной области, а также унификации имеющихся терминов.

Ключевые слова: корпус, корпусная лингвистика, казахский язык, термин, перевод.

**G. A. Sarseke*
L. N. Gumilyov Eurasian National University,
Republic of Kazakhstan, Astana.
Received 28.01.23.
Received in revised form 09.04.24.
Accepted for publication 26.08.24.

FORMATION AND UNIFICATION OF THE TERMS OF CORPUS LINGUISTICS IN KAZAKH

The article deals with the formation of the terms of corpus linguistics in Kazakh. The aim of the study is a comparative analysis of terminological vocabulary associated with corpus linguistics in Kazakh based on scientific and educational literature. The objectives of this study are to analyse the ways of forming of the terms of corpus linguistics, compare existing terms used in describing corpora and stages of work on them, and make proposals on unifying terms and expantion the terminological base of Kazakh. An analysis of the terms of corpus linguistics in Kazakh shows that the words in this area are mainly borrowed terms from English. In addition, it is shown that national terms that have already been formed in the language are also used to name some concepts of corpus linguistics. However, when using national terms, it is necessary to eliminate diversity and adhere to the principle of unification. Considering that research into corpus linguistics in Kazakh linguistics began and developed only in the last decade, the formation of terms in this area takes some time, and the more the number of studies increases, the more the quality rise up. The article emphasises the importance of correct formation of the terms of corpus linguistics, maximum consolidation of national terms in the language when translating concepts in this field, as well as unification of existing terms.

Keywords: corpus, corpus linguistics, Kazakh, term, translation.

Теруге 26.08.2024 ж. жіберілді. Басуға 26.09.2024 ж. қол қойылды.
Электронды баспа
4,12 МБ RAM
Шартты баспа табағы 30,39. Таралымы 300 дана. Бағасы келісім бойынша.
Компьютерде беттеген: А. К. Темиргалинова
Корректор: А. Р. Омарова, М. М. Нугманова
Тапсырыс № 4273

Сдано в набор 26.08.2024 г. Подписано в печать 26.09.2024 г.
Электронное издание
4,12 МБ RAM
Усл. печ. л. 30,39. Тираж 300 экз. Цена договорная.
Компьютерная верстка: А. К. Темиргалинова
Корректор: А. Р. Омарова, М. М. Нугманова
Заказ № 4273

«Toraighyrov University» баспасынан басылып шығарылған
Торайғыров университеті
140008, Павлодар қ., Ломов к., 64, 137 каб.

«Toraighyrov University» баспасы
Торайғыров университеті
140008, Павлодар қ., Ломов к., 64, 137 каб.
67-36-69
e-mail: kereku@tou.edu.kz
www.vestnik.tou.edu.kz