ҚАЗАҚСТАН РЕСПУБЛИКАСЫ
БІЛІМ ЖӘНЕ ҒЫЛЫМ МИНИСТРЛІГІ
ҒЫЛЫМ КОМИТЕТІ
А.БАЙТҰРСЫНҰЛЫ АТЫНДАҒЫ ТІЛ БІЛІМІ ИНСТИТУТЫ

Жұбанов А.Қ., Жаңабекова А.Ә.

КОРПУСТЫҚ ЛИНГВИСТИКА

Оқу құралы

Алматы
2016

ӘОЖ 811.512.122
КБЖ 81.2 Қаз
Ж 80

А.Байтұрсынұлы атындағы Тіл білімі институтының
ғылыми кеңесінде мақұлданған

Жауапты редактор: филология ғылымдарының докторы, профессор **М. Малбақов**

Пікір жазғандар: филология ғылымдарының докторы, профессор **Н. Уәли**

филология ғылымдарының кандидаты **А. Фазылжанова**

Жұбанов А.Қ.
Ж 80 **Корпустық лингвистика/** А.Қ. Жұбанов, А.Ә. Жаңабекова. – Алматы: «Қазақ тілі» баспасы, 2016. – 336 б.

ISBN 978-601-7293-43-7

Оқырман назарына ұсынылып отырған оқу құралы корпустық лингвистиканың негізгі ұғымдарымен таныстыра отырып, корпустық технологиялардың негіздерін меңгеруге, корпустармен жұмыс істеу дағдысын қалыптастыруға және корпустық лингвистика пәнінің ақпараттық технологиялар қатарынан алатын орнын айқындауға мүмкіндік тудырады.

Мұндай зерттеу жұмысы қазақ тіл білімінде бұрын-соңды қолға алынбағандықтан, біріншіден, зерттеу нәтижесі корпустық лингвистика бойынша алғашқы теориялық тәжірибелердің негізін қаласа, екіншіден, қазақ тілінің ұлттық корпусын құрастыруға бағыт-бағдар береді.

Оқу құралы филология мамандықтарындағы студенттерге, магистранттар мен PhD докторанттарға, компьютерлік лингвистика, корпустық лингвистика пәнінен дәріс оқитын оқытушыларға және ақпараттық технология саласындағы мамандарға арналған, сонымен қатар корпусты құрастырушылар мен оны пайдаланушыларға әдістемелік құрал бола алады.

ӘОЖ 811.512.122
КБЖ 81.2 Қаз

ISBN 978-601-7293-43-7

© Жұбанов А.Қ., Жаңабекова А.Ә., 2016
© «Қазақ тілі» баспасы, 2016

АЛҒЫ СӨЗ

Оқу құралының мақсаты – оқырманды корпустық лингвистиканың негізгі ұғымдарымен, корпустармен жұмыс істеу жолдарымен, әдіс-тәсілдерімен таныстыру, корпустық технологиялар негіздерін меңгеруге мүмкіндік жасау және бұл пәннің басқа да ақпараттық технологиялар қатарынан алатын орнын айқындап беру.

Мақсатқа сай, оқу құралында мынадай міндеттер көзделеді: оқырмандарды корпустық зерттеулер тарихымен таныстыру; оқырмандарды корпустық лингвистиканың тілдік және программалық құралдарын зерттеуге бағыттау; оқырмандарды корпустық лингвистиканың программалық құралдарымен және ақпараттық ресурстарымен жұмыс істеу дағдысын қалыптастыру; оқырмандарды корпустық деректерге негізделген зерттеулермен таныстыру және нақты тілдік материалдармен корпустар арқылы жұмыс істеудің тез және тиімді жүзеге асатындығын көрсету.

Оқу құралы төрт бөлімнен тұрады. «Корпустық лингвистикаға кіріспе» атты *бірінші бөлімде* корпустық лингвистиканың негізгі ұғымдары, терминдері, тіл білімінің жеке саласы ретінде қалыптасу тарихы, тіл білімінің басқа салаларымен байланысы және қолданыстағы корпустар түрлері туралы мәлімет берілген.

«Корпустар құрастыру мәселелері» деп аталатын *екінші бөлімде* лингвистикалық белгіленім (разметка) түрлері, сонымен қатар корпустарды жобалауға, іріктеуге және тілдік материалдарды өңдеуге, белгіленім тәсілдеріне, стандарттауға қатысты технологиялық үдерістер сипатталады.

«Корпустарды пайдалану» атты *үшінші бөлімде* корпустардан қажетті мәліметтерді іздестіруді қамтамасыз ететін корпустық менеджерлер мен корпус материалдары негізінде жүргізілген әртүрлі лингвистикалық зерттеулердің (лексикография, грамматика, дискурс) нәтижелері сипатталады.

«Қазақ тілінің Ұлттық корпусын құрастыру тәжірибесінен» деп аталатын *төртінші бөлімде* Қазақ тілінің ұлттық корпусын құрастырудағы алғашқы тәжірибе нәтижесі, морфологиялық белгіленімнің лингвистикалық әзірлемесі, мәтіндерге морфология-

лық белгіленім қоюды автоматтандыру, аннотацияланған корпустардың ғылыми-практикалық маңызы туралы баяндалады.

Бұл оқу құралын жазуға авторлардың «Компьютерлік лингвистика» мен «Корпустық лингвистика» пәндері бойынша оқыған дәрістері және «Қазақ тілінің Ұлттық корпусын құрастыруға» қатысты теориялық зерттеулерінің (мақалаларының) материалдары мен корпус құрастыру тәжірибесі негіз болды. Сонымен бірге, аталған пән бойынша орыс тілінде жарық көрген филолог-ғалымдар, доцент Виктор Павлович Захаров (Корпусная лингвистика: Учебно-метод. пособие. – СПб., 2005) пен профессор Светлана Юрьевна Богдановамен бірігіп жазған (Корпусная лингвистика: учебник для студентов гуманитарных вузов. – Иркутск: ИГЛУ, 2011. – 161 с.) еңбектері және т.б. осы сала бойынша жазылған ғылыми еңбектер материалдары да негіз болды. Әсіресе, Қазақ тілінің ұлттық корпусы жасалу үстінде болғандықтан, үшінші тарауда ағылшын тілі негізінде жасаған В.Захаровтың корпусты пайдалануға қатысты материалдары түпнұсқа күйінде алынды.

Сонымен бірге осы оқу құралының авторының бірі А.Қ.Жұбановтың қазақ тілін зерттеуге арналған жаңа сала «Қолданбалы лингвистика» бойынша жүргізген ғылыми зерттеу жұмыстары да ұсынып отырған кітапқа теориялық және практикалық жағынан негіз болды деуге болады.

Атап айтқанда, автордың «Қолданбалы тіл білімінің мәселелері» атты монографиялық жұмысы ҚР БжҒМ А.Байтұрсынұлы атындағы Тіл білімі институты, ҚР Мәдениет және Ақпарат министрлігі Тіл комитетінің қолдауымен және «Арыс» қорының «Қазақ ғылымының озық үлгілері» атты мегажобасының аясында жарық көрді (Алматы: «Арыс» баспасы, 2008. – 640 бет). Аталған монографияның «Қолданбалы лингвистика: қазақ тілінің статистикасы» деп аталын бірінші бөлімінде қазақ тілі мәтіндеріндегі сөзтұлғалардың лексика-морфологиялық құрылымын статистикалық әдіспен зерттеу үшін шартты белгі-кодты сәйкестендіру (белгіленім) бағдарламасы көрініс тапқан. Мәтін бірліктеріне тән белгі-кодтарды сәйкестендіру «ұялы принципке» негізделіп, негізгі сөз таптары ретінде қазақ тілінің зат есім, етістік, сын есім сөздері тиісті белгі-кодқа сәйкестік кестелер арқылы берілген.

Сол сияқты, аталған монографияның ең өзекті саналатын екінші бөлімі «Основные принципы формализации содержания казахского текста» деп аталып, мұнда «Қазақ тілінің атаушы сөз таптарын семантикалық топтастыру» мәселесі және қазақ мәтінінің семантикалық-мәнді бірліктері мен қисындасу ережелерін таңдап алу мәселесі де сөз болады.

А.Қ.Жұбановтың ғылыми зерттеулері қазақ тілінің мәтіндік корпусын жасау мәселесіне теориялық тұрғыдан да, практикалық тұрғыдан да құнды негіз болмақ. Оқу құралын жазу барысында пайдаланылған автордың мұндай ғылыми жұмыстары қазақ тіл білімінде бұрын-соңды қаралмағандықтан, біріншіден, зерттеу нәтижесі корпустық лингвистика бойынша алғашқы теориялық тәжірибелердің негізін қаласа, екіншіден, тілдік материалды модельдеу нәтижесінде көптеген тілдік ақпараттарды алып, оларды тіл білімінің кез келген саласында нақты тілдік дерек ретінде пайдалануға, ұстануға мүмкіндік жасайды.

Қазақ тілінің мәтіндік корпусын зерттеуге арналған бұл жұмыстың нәтижелерін жоғарғы оқу орындарында оқытылатын «Қазақ тілінің функционалды грамматикасы», «Компьютерлік лингвистика», «Корпустық лингвистика», «Морфология», «Мәтін лингвистикасы» т.б. салалар бойынша дәрістер мен семинар сабақтарында пайдалануға болады.

Филолгия ғылымдарының докторы,
профессор
Мырзаберген МАЛБАҚОВ

КІРІСПЕ

XX ғасырда басталған ғылыми-техникалық «революция» әлемнің кез келген мемлекетінің ішкі-сыртқы саясатына, әсіресе экономикалық әлеуетіне ерекше серпін беріп қана қоймай, Тәуелсіз Қазақстан Республикасы сияқты дамушы елдердің жас мемлекет ретінде қалыптасуында айрықша рөл атқарды. Қоғамдық қызметтің қай саласында да қолданбалы бағыт басымдық алды. Осы орайда лингвистиканың қолданбалы саласы да қалыптасып, дәстүрлі тіл білімінің бағыттарын өз әдіс-тәсілдерімен зерттеуге кірісті.

Ғылым-білімнің бүгінгі даму сатысында жаңа инновациялық бағытта зерттеліп, жазылып, жасалып жатқан жұмыстарға үлкен мән беріліп отыр. Әсіресе, жаһандану үдерісінің бір көрінісі болып табылатын ақпараттық технологияларды барлық ғылым салаларында кеңінен қолдану ісі кезек күттірмейтін өзекті мәселелердің біріне жатады. Осыған байланысты Қазақстан Республикасының Президенті Н.Назарбаев өзінің «Қазақстан жолы» еңбегінде: «Алдағы онжылдықта жаңа технологиялық сипатты дамудың шешуші бағыттары био- және нанотехнологиялар, жасанды интеллект жүйелері, жаһандық ақпарат тораптары және жоғары жылдамдықтағы көлік жүйелері, қуат үнемдеуші технологиялар болып табылады. ....Өндірісті одан әрі интеллектендіру, үзіліссіз инновациялық үрдіске көшу жүзеге асырылады. Білімге негізделген жаңа үлгідегі қоғамға біртіндеп көшу аяқталады», – деп айтқан болатын. Міне, осы аталғандарға қатысты инновациялық зерттеулер қазақ тіл білімінде Кеңес одағы кезінде-ақ Қазақ ССР Ғылым академиясының Тіл білімі институтында 1969 жылы өткен «Түркі тілдерін статистикалық және ақпараттық зерттеу» атты Бүкілодақтық ғылыми жиыннан бастама алды деуге болады. Көп ұзамай, 1970 жылы математикалық статистика саласының маманы, белгілі математик-ғалым Қалдыбай Бектаевтың жетекшілік етуімен институтта «Тіл статистикасы және автоматтандыру» атты ғылыми топ құрылды. Осы кезеңнен бастап қазақ тілін статистикалық тұрғыдан зерттеу ісі жүйелі түрде қолға алынып, өз нәтижелерін бере бастады. Олардың алғашқысы ретінде 1973 жылы

жарық көрген «Қазақ тексінің статистикасы» атты ғылыми жинақты атауға болады.

Қазақ тіл білімінің қолданбалы лингвистика саласы бойынша 70-жылдары тілдік материалдарды автоматты өңдеу мәселесімен байланысты математикалық, формалды-логикалық, статистикалық-ықтималдық әдістерді игеру қажеттігі туды. Соның нәтижесінде қазақ тіл білімінде математикалық лингвистика саласы дүниеге келіп, тілдік бірліктерді модельдеуге қатысты докторлық, кандидаттық диссертациялар қорғалды және жас ғалымдардың монографиялары жарияланды.

1973–1974 жж. Одақтық «Тіл статистикасы» тобының ғылыми жетекшісі Раймонд Генрихович Пиотровский мен Қазақстандық профессор ғалым Қалдыбай Бектайұлы Бектаевтың «Математические методы в языкознании» атты екі бөлімнен тұратын, жоғары оқу орындарына арналған оқу құралының қазақ елінде жарық көруі де осы жаңа саланың жетістігі деуге болатындай.

Сол сияқты, 1978 жылы отандық түркітану ғылымы үшін ортағасырлық жазба ескерткіші «Кодекс Куманикус» мәтіні тұңғыш рет ЭЕМ жадына енгізіліп, ол бойынша ЭЕМ көмегімен «Куманша-қазақша жиілік сөздігінің» алынуы қазақ тіл білімі үшін аса бағалы зерттеу жұмысы болды. Ал 1979 жылы «Ғылым» баспасынан жарық көрген М.О.Әуезовтің «Абай жолы» романының жиілік сөздігі» атты ұжымдық лексикографиялық еңбек те қазақ ғалымдарының статистикалық лингвистика саласына қосқан айтулы еңбегінің нәтижесі екендігі даусыз.

Қазақ тіл білімінде алғаш тіл зерттеуді автоматтандыру мәселесі белгілі ғалым Қ.Бектаевтың бастамасымен статистикалық әдіспен зерттеліп, нәтижелер алған болса, одан бері де статистикалық зерттеулердің өрісі кеңейіп, әр қырынан, тіл білімінің түрлі салалары (морфология, синтаксис т.б.) мен түрлі жанрлары (стильдер) бойынша қарастырылып келеді.

Қазақ елінің тәуелсіздігі жарияланған 1991 жылы А.Байтұрсынұлы атындағы Тіл білімі институтында «Түсіндірме сөздік» бөлімі мен «Тіл статистикасы және автоматтандыру» ғылыми тобы біріктіріліп, «Қолданбалы лингвистика» бөлімі ретінде жаңаша аталып, қайтадан құрылды. Міне, сол кезден бері жаңа бөлімнің

зерттеу мақсаттары мен міндеттері ұлғая түсті. Қолданбалы лингвистика бөлімі қызметкерлері және ғылыми меңгерушісі Асқар Құдайбергенұлы Жұбанов өз ұстазы проф. Қ.Б.Бектаевтың ғылыми жолын жалғастыру мақсатымен, қазіргі қазақ тілінің жаңа саласы қолданбалы тіл білімінің теориялық және практикалық негізін қалау ісіне айтарлықтай үлес қосты деуге негіз бар. Мәселен, өткен ғасырдың 90-шы жылдары 20 томнан тұратын М.Әуезовтің шығармалар жинағы ЭЕМ жадына енгізіліп, соның нәтижесінде 1995 жылы «М.Әуезовтің 20 томдық шығармалар текстерінің жиілік сөздіктері» атты жиілік сөздіктер жүйесі жарық көргенін айтуға болады. Міне, осындай жиілік сөздіктердің негізінде ұлы жазушының тіл байлығы мен стильдік ерекшелігі жайында бірнеше ғылыми зерттеулер жүргізіліп, диссертациялар қорғалды.

2002 жылы А.Жұбановтың «Основные принципы формализации содержания казахского текста» атты докторлық монографиясы қазақ қолданбалы лингвистикасының болашақ дамуына қосылған елеулі үлес ретінде саналды.

Сондай-ақ А.Байтұрсынұлы атындағы Тіл білімі институтындағы картотекалық қорды компьютер жадына енгізу және терминологиялық түсіндірме сөздіктер мен екі тілді терминологиялық сөздіктерді, қазақ тілінің біртомдық, онтомдық, екітомдық сөздіктерін, онбестомдық «Қазақ әдеби тілінің сөздігін» т.б. еңбектерді автоматты пішінге келтіру істері атқарылды.

Бұрын негізінен статистика мен лексикографиялық зерттеулер бойынша қолданылған тілді автоматтандыру мәселесі тіл білімінің жекелеген салалары үшін де қажеттілік тудырды. Бүкіл тіл жүйесін біртіндеп компьютер жадына енгізіп, оларды белгілі бір автоматты басқару орталығы арқылы іске қосу, әсіресе, тіл тұтынушыларының практикалық қажеттілігі үшін өте маңызды. Басқаша айтқанда, ғылым-білім саласындағы қоғамдық-әлеуметтік сұраныстарды ***жеңіл*** әрі ***тез*** өтейтін бірден-бір ақпараттық жүйе ретіндегі орны ерекше.

Қазіргі кезде компьютерлік лингвистика саласы бойынша осындай ақпараттық сұраныс жүйесін қалыптастыру мақсатында «тілдік корпустарды» зерттеу мен оларды қазақ тілі материалдары негізінде қолданысқа енгізу мәселесі қызығушылық тудыруда.

Бүгінгі таңда тілдік сипаттамаларды теориялық тұрғыдан ғана негіздеп қоймай, сол тілдік жүйедегі заңдылықтар мен қағидаттарды, тұрақты категориялық сипаттамаларды автоматтандыру ісімен байланысты қарастыру қажеттігі дау тудырмайды. Бұл әсіресе, бүгінгі дамығын мемлекетіміздің ғылым мен техниканы үйлестіру мақсатын көздейтін негізгі міндеттерінің бірі.

Компьютерде мәтіндерге лингвистикалық ақпараттар беру, корпустық зерттеулер біраз уақыттан бері жалпы тіл білімінде, соның ішінде шетел тіл білімінде лингвистикалық зерттеулердің негізгі тәсілдерінің бірі ретінде қолға алынып келеді. Мұндай зерттеу жұмыстары қазақ тіл білімінде бұрын-соңды зерттеу нысанына ілінген жоқ, тіпті орыс тіл білімінің өзінде бұл мәселе лингвистиканың басқа салалары мен бағыттарына қарағанда кейіндеп қалған.

Қазіргі жаһандану кезеңінде әртүрлі саяси-әлеуметтік, экономикалық қарым-қатынастарға байланысты ақпарат ағыны бұрын-соңды болмаған қарқынмен өршуде. Ал қоғам өміріндегі мұндай ақпарат ағымының таралуы табиғи тілде жүзеге асатындықтан, тіл білімінің қызметі күннен-күнге кеңеюде. Осыған байланысты ұшы-қиырсыз ақпарат ағынын игеру мақсатында шетел және орыс тіл білімінде орасан зор нәтиже беріп отырған тілдік корпустарды қазақ тіл білімінің материалдары негізінде жасау бүгінде үлкен сұранысқа ие болып отыр.

Осы ретте қазақ әдеби тіліндегі әртүрлі функционалдық стильдерді (көркем проза, драматургия, ғылыми-техникалық мәтіндер, публицистикалық мәтіндер) қамтитын лингвистикалық ақпаратпен толық аннотацияланған (мазмұндалған) ұлттық корпусты құрастыру – бүгінгі таңда аса өзекті мәселе.

2009–2011 жылдар «Компьютерлік лингвистика» ғылыми тобы қызметкерлері «Мәдени құндылықтар ретіндегі қазақ тіліндегі мәтіндер корпусы және сөздіктердің «Тіл – қазына» атты ұлттық компьютерлік қоры» атты ғылыми-зерттеу тақырыбын өз зерттеу нысаны ретінде алып, 2012–2014 жылдары «Қазақ әдеби тілінің аннотацияланған ұлттық корпусы», 2015–2017 жылдары «Қазақ тілінің ұлттық корпусындағы метамәтіндік белгіленімдер ұстанымдары мен әзірлемесі» деген тақырыптарда зерттеу жүргізуде. Зерттеу нәтижесінде Қазақ тілінің ұлттық корпусын құрас-

тырудың алғышарттарын қалыптастырып, метабелгіленіммен және морфологиялық, лексикалық, морфо-семантикалық белгіленімдермен жабдықталған 7 миллионға жуық сөзқолданыстан тұратын бастапқы мәтіндер корпусын құрастырды.

Компьютерлік лингвистика мамандарының тұжырымдауынша, компьютерлік тілдік қор дегеніміз – ғылым адамының өз зерттеу нысанына жаңаша тұрғыда көз салу мүмкіндігі. Мұндай тілдік қор неғұрлым қомақты болса, солғұрлым тіл құрылысының сыры тереңірек ашылады, зерттелетін нысан жөніндегі түсініктердің аумағы кеңейеді. Сол сияқты, зерттеуші адамның мүмкіндік-қабілеті әлденеше есе артады, шығармашылық қуат көздері ашыла түседі, сөйтіп, бұл жаңа мүмкіндіктер қазақ тілінің жүйесін жетілдіруге және тіл жүйесін мұқият тануға жұмсалатыны сөзсіз. Сонымен бірге тілдік процестерді автоматтандыру мәселесі тоқырап қалмай, үнемі жетілдіру жағынан өзінің жалғасын тауып тұруы қажет.

Корпустық лингвистика қазақ тіл білімінің де ерекше саласы ретінде қалыптасатын болса, қазақ тілшілеріне көлемді тәжірибелік материалдарды пайдалануға, қажетті деген тілдік деректерді тез тауып алуға және оларға тиісті деген өңдеулер жүргізуге мүмкіндік тудырады. Осының бәрі қазақ тіліне қатысты зерттеулердің шынайылыққа (ақиқаттыққа) жетудің эмпирикалық тәсілдеріне жаңаша көзқараспен қарауға және ғылыми айналым аясына аса маңызды тілдік материалдарды енгізуге жағдай жасайды.

Қазақ тілін зерттеуде корпустық лингвистика саласына көп мән берілуі және оның әлемдік дәрежедегі теориялық және практикалық жақтарын зерттеу қажеттігі туындауда. Әлем бойынша мәтіндер корпустарын құрастыру мен олардың қызметіне қатысты жалпы және нақты мәселелелерге арналған мақалалар жарық көрген ғылыми журналдардың арнайы басылымдары да шыға бастады. Бірақ әлде де қазақ тіл білімі үшін корпустық лингвистикаға қатысты көптеген мәселелер арнайы зерттеуді қажет ететіні белгілі. Оған жататындар: корпустық лингвистика мен оның негізгі ұғымдарының анықтамалары, корпустық лингвистиканың тіл білімі құрылымында алатын орны, әдіс-тәсілдері және т.б. Сонымен бірге жаңа бағыттың теориялық негізін ұғыну мәселесі кор-

пустарды нақты зерттеулерде пайдалануға қарағанда белгілі дәрежеде қалыс қалып келе жатқаны да байқалады.

Корпустық лингвистиканы дамытудың өзектілігін сапалы грамматикалық өңдеуден өткен және көлемді тілдік деректерге жеңіл қол жеткізуді қамтамасыз ететін ресурстарды құрастыру қажеттілігімен түсіндіруге болады. Осындай қажеттіліктің ең негізгісі деп қолданбалы лингвистика мәселелерінің ішінен бір тілден екінші тілге компьютер арқылы аудару жүйесін тілдік материалдармен қамтамасыз ету әрекетін атауға болады. Корпустық лингвистикада қолданыс табатын компьютерлік бағдарламалар әртүрлі табиғи тілдерді өңдеуге бағытталған. Қазіргі кезде түрлі тілдердің мәтіндер корпусын құрастыру әрекетінде қолданылатын мұндай бағдарламалар (NLP – natural language processina) барынша жетілдірілуде.

Сондықтан тіл білімінің осындай аса қызығушылық туғызып отырған әлемдік деңгейдегі жаңа саласы – корпустық лингвистиканың зерттеу нысанына нелер жатады, тілдік корпус дегеніміз не, мәтіндер корпусын құрастыру не үшін қажет және ол қандай ғылыми-теориялық мәселелерді шешуге септігін тигізеді деген мәселелерден көпшілікті хабардар етпекпіз.

1. КОРПУСТЫҚ ЛИНГВИСТИКАҒА КІРІСПЕ

1.1. Корпустық лингвистиканың негізгі ұғымдары, зерттеу нысаны

Бүгінгі жаһандану заманында ғылым-білімнің барлық саласын автоматтандыру кезек күттірмейтін өзекті мәселеге айналып отыр. Тілдік материалдарды автоматтандыру қажеттігі қазіргі кезде тіл білімінің «Компьютерлік лингвистика» атты саласының дамып, қалыптасуына мүмкіндік берді. Тіл білімінің жаңа саласы болып табылатын «Корпустық лингвистика» – «Компьютерлік лингвистиканың» бір саласы.

«Компьютерлік лингвистика» термині (computational linguistics), негізінен, тіл қызметін модельдеуге қатысты пайдаланылатын компьютерлік бағдарламалар мен тілдік деректерді өңдеу, сол үшін қажетті компьютерлік технологияны жинақтау және компьютердің араласуына қатысты басқа да тілді зерттеу жұмыстарымен байланысты қолданылады. Кең мағынада алғанда, тіл білімі аясына қатысты теориялық және практикалық мәселелердің шешімін табуға байланысты компьютердің араласу жағдайларының бәрін бірдей компьютерлік лингвистика саласына жатқызуға болады.

Корпустық лингвистика – қолданбалы лингвистиканың заман сұранысынан туындаған салаларының бірі. Корпустық лингвистика тілді ақпараттандыру, техника мен технологияны пайдалана отырып зерттеуді қажет ететін маңызды ғылым саласы болып табылады.

Корпустық лингвистика компьютерлік технологияларды қолдана отырып, лингвистикалық корпустар (мәтіндер корпустарын) құрастыру мен оны пайдаланудың жалпы ұстанымдарын зерттейді. 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. Басқаша айтқанда, қазақ тілінің мәтіндер корпусы (тілдік корпусы) – аса ірі көлемдегі жү-

йелі түрде құрылымдалған, тиісті белгіленімдер енгізілген нақты тілдік мәселелерді компьютер арқылы шешуге арналған тілдік деректер ауқымы (массиві).

Қазіргі кезде «корпус» ұғымының бірнеше анықтамалары бар. Мысалы, Э.Финеганның оқулығында мынадай анықтама берілген: корпус – мәшине (компьютер) оқи алатын форматта және мәтін туралы ақпарат берілген, яғни айтушы, автор, адресат (тыңдаушы) немесе дәрісхана (тыңдаушылар) жайлы ақпараттар қамтылатын мәтіндердің репрезентатты (көлемді) жиыны [1].

Википедия желісі корпустарға статистикалық талдау мен болжамдардың шынайылығын, белгілі бір лингвистикалық салалар бойынша тілдік ережелердің негізділігін тексеру үшін қолданылатын үлкен көлемді және құрылымдалған электронды пішіндегі мәтіндер жиынтығы деген анықтама береді [2]. Т. Мак Энери және Э. Вилсон мынадай анықтама ұсынады: корпус – ол тілдік модель ретінде қолданылатын нақты тілдік критерийлерге сәйкес келетіндей таңдалып алынған тілдік фрагменттердің жиынтығы [3]. Ғалым В.В. Рыков мәтіндер корпусын, негізінде логикалық түпкі ой, логикалық идея жатқан мәтіндер жиынтығы деп түсіндіреді де, ал олар мәтіндерді құрастыру ережелері, мәтіндерді талдаудың алгоритмі мен бағдарламасы және осымен түйіндес идеология мен әдіснамалар арқылы жүзеге асады, – дейді [4].

В.П.Захаровтың корпус ұғымын тар мағынада түсіндіруі қазіргі ғылыми түсінікте эволюциялық тұрғыда көрініс тапқан. Мысалы, ғалым «Корпусная лингвистика» атты оқу құралында былай дейді: «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].

Сонымен, **корпус дегеніміз** – әр тілдегі электронды пішінге келтірілген, яғни бір басқару орталығынан автоматты түрде жұмыс істейтін, лингвистикалық ақпараттарды қамтитын мәтіндер жинағы.

Мәтіндер корпусын сөз етер алдында, ең алдымен, мәліметтер корпусы ұғымын түсініп алу қажет болады. Мәліметтер кор-

пусы дегеніміз – феномендері лингвистикалық сипаттауды қажет ететін, белгілі ережелерге сай құрастырылған және тілдік жүйеде іске асатын деректер таңдамасы. Ол тек бірөлшемді – тілдік (речевое).

Мәтіндер корпусы – ол бірліктері ретінде мәтіндер немесе олардың құнды-құнды салаларының ауқымды үзінділері қамтылатын мәліметтер корпусының көрінісі. Негізінде, саны жағынан бір жиыннан асатын кез келген мәтіндер жиынтығын корпус деп атауға болады (лат. corpus – «body»). Әдеби және тілдік талдаулардың түрлеріне қарай жеке мәтіндердің жиі қолданылатыны белгілі. Алайда электрондық лингвистиканың негізін құрайтын корпус ұғымы жеке-жеке мәтіндерді қарастырудан қалайда айырым табады.

Жоғарыда келтірілген анықтамалар негізінде қазіргі мәтіндер корпустарының негізгі белгілері ретінде мыналарды атап айтуға болады:

1) мәшине оқи алатын форматта болуы;

2) айрықша іріктеу әрекетінің нәтижесі ретіндегі репрезентаттық (көлемділік) сипаты;

3) металингвистикалық ақпараттың берілуі.

Корпустар жасау мен оны пайдаланудың қажеттігі келесі алғышарттармен айқындалады:

1) барынша үлкен көлемді корпус (репрезентативтілік) деректердің типтілігіне кепілдік береді және тілдік құбылыстардың барлық саласы туралы толық ақпаратпен қамтамасыз етеді;

2) корпустарда әртүрлі деректер типтері өзінің табиғи контекстік формасында сақталатындықтан, олар жан-жақты және объективті зерттеулер жүргізуге мүмкіндік тудырады;

3) бір-ақ рет жасалған және дайындалған деректер массивін талай мәрте және әртүрлі зерттеушілер, әртүрлі мақсатта пайдалануларына болады.

Корпустарды тілдік және сөйлеу бірліктері жайлы неше түрлі анықтамалар мен статистикалық мәліметтер алуға пайдалануға болады. Мәселен, корпустар негізінде сөзформалар, лексемалар, грамматикалық категориялардың қолдану жиілігі жайында, әртүрлі кезеңдерге қатысты олардың қолдану жиілігін және контекстер-

дің өзгеріске ұшырауын бақылап отыруға, тағы да басқа әртүрлі деректерді алуға болады. Белгілі кезеңдегі тілдік деректердің ауқымды жиыны тілдің лексикалық құрамының өзгеріске ұшырау динамикасын зерттеуге, әртүрлі жанрдың және әр автор шығармалары тілінің лексика-грамматикалық сипатына талдау жүргізуге мүмкіндік береді. Сонымен бірге корпустар көпаспектілі лексикографиялық жұмыстар жүргізуде дереккөз ретінде де, әртүрлі тарихи және замани сөздіктерді дайындауда да зор қызмет атқарады. Корпус деректерін грамматикаларды құрастыру мен нақтылау және тілді оқыту мақсатында пайдалануға болады.

Қазіргі кезде корпустық лингвистиканы компьютер көмегімен және электрондық корпустар арқылы «шынайы өмірдегі» тілдің қолданысын зерттеумен айналысатын тіл біліміндегі жаңа тәсіл деп тану жиі кездеседі.

Егер тіл білімінің синтаксис, семантика және әлеуметтік лингвистика сияқты бөлімдері тілдік сипаттауды, тілдік құрылымды, тілдік қолданысты зерттеуді мақсат етсе, ал корпустық лингвистика тілдік зерттеудің көптеген аспектілеріне қолдануға болатын әдістемесімен ерекшелінеді. Кейбір жағдайда корпустық лингвистиканы «әртүлі тілдік зерттеулердің әдістер шоғыры» деп те атайды. Тілдік талдау әдісі ретінде корпустық лингвистиканың тілдер арасындағы жалпылық және даралық фактілерді анықтауға бағытталған, тілдерді салыстырмалы зерттеу барысындағы диалектілердің немесе тіл варианттарының ерекшелігін анықтайтын контрастивтік зерттеулерге де қатысы бар. Өйткені тілдік талдаулардың ұтымды жасалуы тұрақты және аумақты эмпирикалық деректер қоры негізінде ғана мүмкін болады.

Ғалым Э.Финеган корпустық лингвистиканы табиғи тілдің қолданысын зерттеуге бағытталған корпусты құрастыру мен қолдануды қажет ететін қызмет түрі деген анықтама береді [1]. Корпустық лингвистиканың екіжақтылық сипатын (корпусты құрастыру және оны пайдалану) оның ***нысанының*** екіжақтылығымен түсіндіруге болады. Себебі, біріншіден, нысан деп отырғанымыз, корпустық лингвистика үшін де және басқа да тілдік пәндер үшін де – тілдік материал, екіншіден, корпустық лингвистика саласы бойынша жасалатын жұмыстардың нәтижесі.

Корпустық лингвистика саласының **зерттеу пәні** – әртүрлі салалардағы пайдаланушылардың тілдік зерттеулер жүргізуіне арналған тілдік деректердің көлемді ауқымын жасау мен оны пайдаланудың теориялық негіздері және практикалық механизмдері.

Орыс тілі үшін де және түркі тілдері үшін де бүгінгі күні корпустық лингвистика пәнінің терминологиясының қалыптасуына қатысты проблемалық мәселелер бар. Оның себебі, негізінен, ол пәннің дүниеге келуінің қысқа мерзімдігі болса, екіншіден, оның АҚШ пен Ұлыбританиядан бастау алып, терминологиясының ағылшын тілінде қалыптасуы болып табылады. Орыс тілі корпустарының терминдері негізінен ағылшын корпусы терминдерінен алынған кірме терминдер, ал олардың кейбіреулері басқа мағынада орыс тілінде бұрыннан қолданыста бар. Мысалы, орыс тіліндегі «корпус» сөзі корпустық лингвистика саласының термині болудан көп бұрын көпмағыналық сипатта орыс тілінде де қолданылған. Орыс тілінде бұл сөздің зат есім формасы көптік мағынада «корпусы» және «корпуса» формаларында қолданылып, проблемалық мәселеге айналған. Сол сияқты, осы сөздің екпінге қатысты да шешілмеген мәселері болған («корпуса́», «ко́рпусы», «ко́рпусный», «корпусно́й», «корпусна́я). Қазақ тілінде де әлемдік халықаралық термин ретінде «корпус» терминін қолдану қажет деп санаймыз.

Корпустық лингвистиканы осы саланың мамандары тілдік корпустарды құрастыру мен оны пайдалану жағдайын зерттейтін тіл білімінің бір бөлімі ретінде қарастырып келеді. Сонымен бірге, ол таңдама мәтіндер (мәтін үзінділері) түрінде көрініс табатын тілді зерттеу мен оны танып білуге қажетті маңызды материал деуге де болады. Кейбір ғалымдар ол пәннің түсінігін тар шеңберде қарастырып, оны тек компьютерлік лингвистика саласының аясында ғана түсіндіреді: «Корпусная лингвистика – раздел компьютерной лингвистики, занимающийся разработкой общих принципов построения и использования лингвистических корпусов (корпусов текстов) с использованием компьютерных технологий» – дейді [5]. Ал компьютерлік лингвистика ұғымын, әдетте, компьютерлік құралдарды пайдаланудың кең мүмкіндігі ретінде түсіндіруге болады. Бұл жердегі «компьютерлік құралдар» деп

отырғанымыз – компьютерлік бағдарламалар, тілдік деректерді өңдеу мен компьютерлік технологияны орынды ұйымдастыру жұмыстары және т.б.

Корпустық лингвистика компьютерлік технологияларды қолдана отырып, тілдік мәліметтердің лингвистикалық корпусын құрастыру мен оны қолданудың жалпы принциптерін зерттеумен айналысады. Ол мынадай үш тәсілді қолданады:

– корпус ішінен тілге қатысты мәліметтерді автоматты түрде таңдап алу;

– ақпаратты өңдеу;

– өңделген мәліметтерді тексеру және оларға түсініктеме беру.

Алғашқы екі тәсіл толығымен автоматтанған, ал үшінші тәсілдің толық шешімі әлі де табыла қойған жоқ, ол көбінесе жартылай автоматты түрде адамның қатысуымен орындалады. Автоматты лингвистикалық түсініктеме беруге ыңғайлы белгіленім – ол тілдердің барлығында да негізінен морфологиялық белгіленім.

Ғалым В.В. Рыков зерттеудің мынадай ретін ұсынады:

– шынайы сөйлеу тілінің құрылымын көрсету (ұсыну) қажет;

– корпусты құрастыруға қатысты қандай материалдық шектеулер (қиындықтар) барын айқындау;

– мәтіндерді таңдау;

– корпусты құрастыру (компиляциялау).

Корпустық лингвистиканың негізін құрайтын тіл – ол толығымен әлеуметтік құбылыс және оны тәжірибе жүзінде негізделген (яғни сөйлеу актісі бойынша) мәліметтермен сипаттауға болады. Мұның өзі, сөйлеушінің немесе тыңдаушының сөзді, сөйлемді немесе мәтінді дыбыстағанда (айтқанда) немесе қабылдағанда (естігенде) оны қалайша түсінетінін біз біле бермейміз, ал кейде тіпті білуді қажет деп те санамаймыз. Табиғи тіл әлеуметтік феномен ретінде өзін жазып алуға, сипаттауға және талдауға болатындай түрде көрініс табады. Адамның ойында пайда болатын (ішкі) дыбыссыз тілді де мәтін деп санауға болады, бірақ оларды біз бақылай алмаймыз, міне, сол себепті олар әлеуметтік құбылысқа жатпайды. Көптеген мәтіндер сөйлеу актілері түрінде, яғни тілдік қоғам мүшелерінің арасындағы өзара әрекеттестік ретінде

кездеседі. Көпке дейін адамдар сөйлеу актілерін жазып алуға мүмкіндіктері болмағаны белгілі, ал қазіргі кезде неше түрлі таспаға жазып алу құралдарының арқасында жазылып алынған сөйлеу актілерінің саны күннен-күнге артуда. Енді жинақталған барлық ақпараттарға қажетті өңдеу жүргізу проблемасы туындайды. Мұндай ақпарат ағымы көлем жағынан қысылған түрде көрініс табуы керек.

Табиғи тілдің бірден-бір ерекшелігі – оның динамикалылығы. Сөйлеу актісінің талқыға салу, теріске шығару және қабыл алмау сияқты прагматикалық сипаттары бар. Сонымен бірге пікірталас тудыратын мәселе, ол – мағына. Корпустық лингвистика жеке табиғи тілдерді, жеке тілдік мәселелерді сипаттайды. Жазба және сөйлеу тілдерінің элементтері қайсыбір жиілікпен және ықтималдықпен сипатталатын болғандықтан, тілдік мәтіндерді математикалық статистика пәні әдістерімен талдау жүргізуге және зерттеуге болатындығы қазір еш дау туғызбайды. Ғалым-тілшілер тілдік элементтердің статистикалық сипаттамаларын назардан тыс қалдыруы мүмкін де емес, осы ретте тілді зерттеуде «лингвостатистика» бағыты пайда болғаны белгілі. Демек, корпустық лингвистика тілдік қоғамның табиғи болмысын дәл сипаттау үшін барынша толық мәліметті қамтуға тырысады. Корпустық лингвистика үшін тіл – ол мәтіндердегі адам тілінің моделі ретінде танылатын виртуалды әрекеттегі құбылыс.

Компьютерлік (корпустық) лингвистиканың негізгі мақсатын ғалым В.В.Рыков қысқаша түрде былай тұжырымдайды:

«Бұл сала – сол тілдегі толық және жүйелі түрдегі мағыналық қарым-қатынастың көрінісі. Зерттеу бағытының негізгі ерекшелігі лингвистикалық зерттеулердің ақпараттық-семиотикалық бағытына жататындығы және оны лингвистиканың қолданбалы мәселелерін қатаң коммуникативтік үдеріс ретінде қарастыратын тәсіл деуге болады. Бұл ретте, негізгі назар жүйе құрайтын табиғи тілге аударылмайды және формалдаудың да проблемасы емес, ол сол тілдегі мағыналық қатынас жасау үдерісі және, мүмкіндігінше, информатиканың ғылыми-техникалық мәселелерін шешуге қолдануға болатындай оның дәл сипаттамасы. Корпустық лингвистика тілге қатысты бұрынырақ алынған зерттеу нәтижелері мен

жасалған қорытындыларды нақтылау ісіне және лингвистикалық зерттеулердің эмпирикалық тілдік материалдарын аса кең және жүйелі тұрғыда қамти отырып, жаңа зерттеулер жүргізуге үлкен мүмкіндік жасады. Корпустық лингвистика негізгі назарын тілдік тұлға мәселесіне аударады, яғни тұлғаның тілдік әрекеті, жаппай коммуникация және сипаттау мәселелері деуге болады» [4].

Корпустарды құрастырудың және қолданудың барлық тілдерге ортақ бірыңғай әдістемесі жоқ, себебі ол тілдер бір-бірінен айырым табады. В.В.Рыков мынадай негізгі талаптарды бөліп қарастырғанды жөн санайды:

1. Корпусты пайдаланатын кім? (жеке адам, топ, лингвистикалық қоғам).

2. Корпустың негізінде қандай логикалық идея жатыр?

3. Корпусты құрастырғанда біз қандай көлемдегі мәліметтермен жұмыс істейміз? Қаншалықты бұл қажет және қаншалықты ақиқаттыққа сай келеді?

4. Мәтіндерден үзінділер алына ма, әлде толық мәтін алына ма немесе екеуі де қолданыс таба ма?

5. Корпусқа мәтіндерді таңдау тәртібі. Түрлі мақсатқа түрліше:

– тілдік материалды зерттеу;

– мәтіндерді сканерлеу;

– корпусты құрастыру.

6. Салалық стандарттар деңгейінде корпусты стандарттау, яғни барлық корпустың өнім ретіндегі көрінісі:

– барлық мәтінді толығымен аннотациялау;

– мәтіндегі сөздер мәліметінің бірізді көрінісі.

7. Мәтіндегі сөздер материалын аннотациялау, индекстеу.

Корпустық лингвистиканың негізгі міндеттеріне мыналарды жатқызуға болады:

1. Аталған бағыттың теориялық негіздерін жете зерттеу;

2. Корпустар түрлерін құрастыру мен қолдану тәжірибесін талдау;

3. Корпусқа қойылатын жалпы талаптарды тұжырымдау;

4. Әртүрлі зерттеулер мен оқыту міндеттерін атқаратын корпустар құрастыру;

5. Тіл білімінің әртүрлі салаларында мәтіндер корпустарын

қолданудың тиімді тәсілдерін қалыптастыру. Бұл міндет зерттеуші қауымына практикалық тұрғыда аса құнды болып саналады.

Корпустық лингвистика әдістері ішінен келесі бөліктерді жеке қарастыруға болады:

1) филологиялық әдістер;

2) теориялық-лингвистикалық әдістер;

3) математикалық (статистикалық) әдістер;

4) ақпараттық технологиялық әдістер.

Бірінші топ әдістеріне жанрға, мәтін авторларына, жанрлар мәтіндерінің үлесіне және т.б. мәселелерге қатысты әдістер жатады. Екінші топта – мәтін таңдаудың және корпустағы лингвистикалық ақпараттың берілуінің жалпы және жеке лингвистикалық мәселелерімен байланысты әдіс-тәсілдер. Үшінші топты таңдама мәтіндердің санына, оның репрезентативтілігін (көлемділігін) қамтамасыз ету аспектілеріне қатысты математикалық статистиканың әдістері құрайды. Төртінші топқа корпус деректерінің компьютерлік көрінісі мен оны өңдеу әрекеттерін қамтамасыз ететін ақпараттандыру әдістері жатады.

Корпустық лингвистиканы жаңа білім аясы деп түсінуге байланысты «әдіс» ұғымы түрліше ұғынылуы мүмкін. Мәселен, корпустық лингвистиканы компьютерлік лингвистиканың бір бөлігі деп қарастыратын болсақ, онда төртінші топқа қатысты ақпараттану әдістері басымдық етер еді. Алайда В.П.Захаровтың жоғары оқу орындарында «Корпустық лингвистика» пәнін оқыту курсында мәтінге лингвистикалық шартты белгілер енгізуге, корпустарды құру технологиясына, корпустан қажетті ақпаратты іздеу әдістеріне (конкордансерлерге, корпустық менеджерлерге, сұрату (язык запросов) тілдеріне) көбірек көңіл бөлінген [5].

А.П.Барановтың тұжырымдауынша, корпустық лингвистикада негізгі назар тілдік жүйенің сөйлеу арқылы жүзеге асуына көп көңіл аударылады: «Как правило, в корпусной лингвистике языковой аспект фактически игнорируется, поскольку изначально фиксируется область привлекаемых данных – реализаций языковой системы» [6]. Солай бола тұра, корпусты құру мен оны түрлендіру жағын зерделеуде тілдік аспектіні ескеру қажеттігі күмән тудырмайды.

Корпусқа қойылатын басқа да талаптар бар, олар мыналар:

– релевантты құбылыстардың толық ескерілуі;

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

– тілдік матералдарды 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** неше түрлі болғандықтан, олар лингвистикалық, филологиялық және педагогикалық салаларға топтастырылуы мүмкін. Ондай топтарға жататын пәндер: лексикография, грамматика, мәтін лингвистикасы, типология, контрастивтік лингвистика, аударма жасаудың теориясы мен практикасы, әдебиеттану ғылымы, лингводидактика. Қазіргі кезде корпустық лингвистиканың қолдану аясы бұрынғыдан да кеңейе түсті.

Корпустық лингвистиканың негізгі түсінігін оның бейнелейтін аясына қарай, корпустағы сақталатын ақпарат бірлігі тұрғысынан, оны әртүрлі негізде топтастыру мен құрастыру және т.б. принциптерді ұстаным етуіне қарай әртүрлі ұғынуға болады.

Корпуста сақталатын бірлік – ол пәндік саланың қайсыбір жиынтығы. Мысалы, оларға жататындар: сөз, сөзформа, сөзтіркес, сөйлем немесе толық мәтіндер жиынтығы. Корпустағы деректер құрылымданған және құрылымданбаған пішімдерде сақталуы мүмкін. Біріншіден, яғни құрылымданған пішім жағдайында шартты белгілермен белгіленген мәтін туралы сөз болады. Ондай белгілер семантикалық (лексика-семантикалық), морфологиялық, синтаксистік т.б. Мысалы, орыс тілінің ұлттық корпусында орын алған морфологиялық белгілер барлық сөздерді сөз табына, септелу категориясына, тегіне (род – муж., жен.), жіктелуіне, жанды-жансыздығына, етіс және етістік түріне, шырай және т.б. морфологиялық сипаттамаларына қарай арнайы белгі қою жүргізілген. Сол сияқты семантикалық талдау арқылы да мәтін ішіне қажетті белгілердің қойылатынын айтуға болар еді. Мысалы, орыс тілінде

«Таксономияға», «бағалауға», «сөзжасамдық сипаттамаға» қатысты белгілеулер орын алған. Мұндағы «Таксономия» белгісі: орынды, бағытты, қашықтықты, уақытты, жылдамдықты, санына және т.б. жүйелеулерге қатысты семантикалық сипаттамаларды білдіреді. Қазақ тілінде де осы тәрізді сөздердің морфологиялық құрылымы, сөзжасамдық құрамы, лексика-семантикалық сипаты мен синтаксистік құрылымына қатысты да ақпарттар енгізуге болады.

Корпустық менеджер және конкорданс. «Мәтіндер корпусы» ұғымына мәтіндер мен тілдік деректерді басқару жүйесі де енеді. Пайдаланушы мен корпус арасындағы әрекеттестік: корпусты пайдаланушының сұранысы бойынша орындалатын қажетті тілдік ақпаратты іздеп-табу әрекеті арнайы жазылған «корпустық менеджерлер» (немесе корпус-менеджер) (ағыл. corpus manager) атты компьютерлік бағдарламалар арқылы жүзеге асады. Олар бойынша корпус негізінде іздеп-табылған нәтижелерді сұрыптаудан, статистикалық санақтан, конкорданстар мен сөзтібе құрастыру жұмыстары қамтамасыз етіледі.

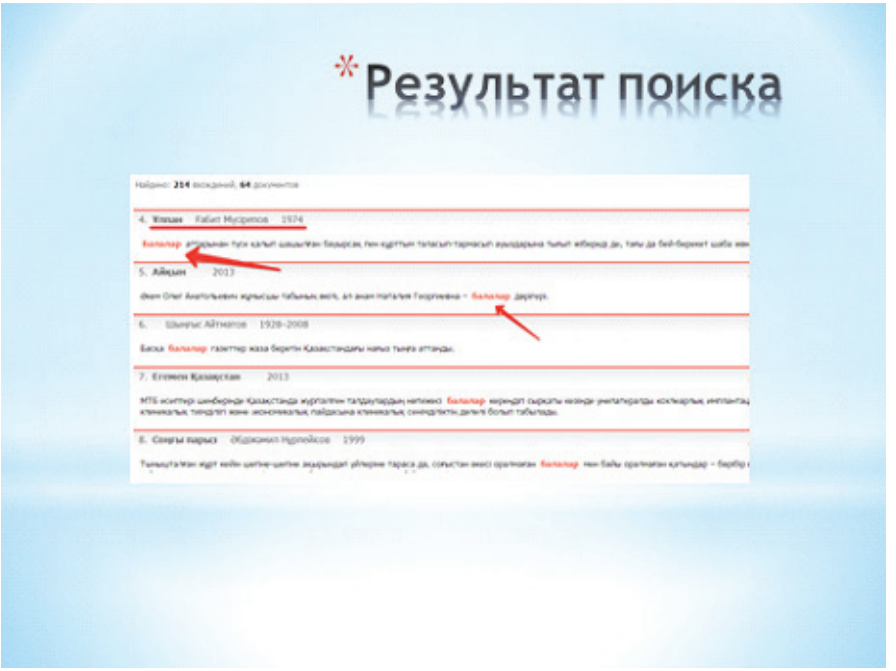
Корпустық менеджер – бұл жүйе корпустағы деректерді іздеп-табуға арналған бағдарламалық құралдарды өз қатарына қосумен бірге, статистикалық ақпаратты алуды және пайдаланушыға алынған нәтижелерді ыңғайлы түрде ұсынуды көздейтін мамандандырылған іздестіру жүйесі болып табылады. Корпус бойынша зерттеуші өзіне қажетті сөздерді, сөз тіркестерін, грамматикалық формаларын арнайы құрастырылған іздестіру жүйесі арқылы іздеп-тауып, олардың мағыналарын аша алады.

Қазіргі кездегі корпустық менеджерлердің көпшілігі ақпараттың бірнеше түрлерін іздеуді жүзеге асырады: нақты сөзформаларды іздеу; леммаға сәйкес келетін сөзформаны іздеу (бір ғана лемманың мәтінде кездескен барлық формаларын іздеу); тұрақты және тұрақсыз сөзтіркестерді іздеу. Егер корпуста қажетті деген грамматикалық деңгейдегі белгіленім және басқа да ақпараттық белгіленім орын алатын болса, солар бойынша іздеуді жүзеге асыруға болады. Метабелгіленімнің түрлеріне қарай пайдаланушы өзіне қажет деген мәтіндер подкорпусын (субкорпус) жасауға мүмкіндігі бар. Мысалы, ондағы мәтіндер шығарманың жанрына, тақырыбына, шығарманы жазған уақытына және т.б. қарай бір-

бірінен айырма табуы мүмкін. Мұндай субкорпустардан алынған мысалдар шығарма көзі жайлы ақпаратпен қамтамасыз етіледі. Кейбір корпустардан сұраныс болған тілдік бірліктің статистикалық ақпараты жайлы да мәлімет: корпустағы барлық мәтіндерге шаққандағы қатынастық жиілігі, тілдік бірліктің жанрға немесе кезеңдік деңгейге қатысты қолдану үлестірімі және бірліктің тіркесіп келуі жайлы ақпарат алуға болады.

Мәтіндер корпусындағы кез келген сөзді іздеу, сол сөздің конкордансын жасауға мүмкіндік тудырады.

Конкорданс дегеніміз – сол сөздің барлық қолданысы бар мәтін үзінділері қамтылған шығармалар тізімі және оларға сілтеме берілген мәліметтер. Мысалы, «корпустан іздеу» жүйесі арқылы қандай да бір сөзге мысалдар іздесек, корпус осы сауалға сай мысалдар тізімін шығарады, осындай тізімді конкорданс деп атайды.



1-сурет. «Бала» сөзінің конкордансынан үзінді

1.2. Корпустық лингвистиканы өмірге әкелуге себепші болған бағыт: «картотекадан корпусқа»

Көптен бері тілді зерттеу аздаған көлемдегі және сан жағынан да көп емес мәтіндер ішінен қолмен көшіріліп алынған материалдарды негіз етіп жүзеге асқаны көпшілікке мәлім. Ондай тілдік материалдардың жиынтығын картотекалық қор деп атап, зерттеушілер кезінде өз қажеттігіне пайдаланған болатын. Зерттеу тақырыбына байланысты зерттеуші ғалымдар уақыттарының көбін қажетті деген мысалдар жинауға, оларды кәртішкелерге көшіріп, кітаптың атын, авторын, баспа атын, шыққан жылын, кітаптың бетін және т.б. мәліметтерді көрсетіп, қажетті деген сұрыптаулардан өткізіп барып, өз қажеттігіне пайдалануға жұмсайтын. Мәселен, қазақ лекикографтары екітомдық, онтомдық «Қазақ тілінің түсіндірме сөздігін» құрастырудан бұрын көлемі 5 млн.-дай болатын картотекалық қор жинап алғаны белгілі. Картотекалық қордағы кәртішкілер әліпби тәртібімен орналасса да, қордан қажетті сөзге не сөз тіркесіне қатысты кәртішкілерді іздеп табу да аз уақыт алмайтыны белгілі.

Корпус мәселесі тілшілерге таныс дәстүрлі картотека қорын жасау мен қолдану әрекетінің заңды жалғасы деуге болады. XX ғасырда картотекалар компьютер жадына енгізіліп, электрондық қорға айналып, көпшілік қолдануына мүмкіндік жасалды.

Технологиялық құралдар ретінде компьютерлік технологияны қолданатын тіл білімінің басқа салаларымен қатар, корпустық лингвистика тілдік материалды жинау мен сақтаудың дәстүрлі әдіс-тәсілдеріне – жазба мәтінді қол күшімен өңдеу, лұғаттық картотекаларды жинау мен сақтау және т.б. түрлеріне тікелей әсер етті. Дәстүрлі тәсілдердің кемшілік жақтарына жататындар: аса көп уақыт шығыны, зерттеу нысанының көлем жағынан шектелуі, жаңартудың күрделігі, статистикалық өңдеудің қолайсыздығы, қашықтық жағдайындағы (дистанционный) ақпарат алмасу мүмкіндігінің болмауы. Тілдік зерттеулерге қажетті материалдарды жинау мен сақтауға қатысты компьютерлік технологияны пайдалану жаңа тәсілдің артықшылығын толығымен пайдалануға мүмкіндік тудырды. Атап айтқанда, тілдік материалға «қол жеткізу»

жедел түрде іске асады, интернет желісін пайдалану ақпарат алмасу қашықтығын жояды, ақпарат көлемі талай мәрте өседі, деректерді сақтайтын және өңдейтін компьютерлік бағдарламалар арқылы қажет деген ақпаратпен танысу мен оны өңдеу әрекеттері барынша тез жүзеге асады.

Сол сияқты, корпустық әдістің қалыптасуына Интернет желісі айтарлықтай рөл атқарды. Мысалы, корпус жадына енгізілетін мәтіндердің көлемін және әртүрлі лингвистикалық зерттеулерге қажетті мәтін түрлерін арттыруға мүмкіндік туды. Аса кең электронды ресурстарды пайдалану мүмкіндігі лингвистикалық зерттеулердегі қажетті материалдарды жинау әрекетін айтарлықтай жеңілдетті.

Мәтін таңдау, корпусқа мәтін енгізу үдерісінде жинақталған тілдік материалдарға қатысты репрезентаттық (репрезентативность) пен теңгерімділік (сбалансированность) мәселелері туындайды. Әсіресе бұл мәселе ұлттық корпустарды қалыптастыру кезінде аса қажет. Корпустың репрезентаттығы – мәтіндік материалдардың көлемі мен әртүрлілігі жағынан ықтималдықтар теориясының қанағаттандыру шартына бағынады.

Корпустық лингвистика – тіл біліміндегі эмпирикалық деректермен жұмыс істейтін ресурстар мен әдістердің жиынтығы. Қазіргі кездегі корпустық лингвистиканың әдістемелік ғылым тұрғысынан дамуы (өркендеуі) оның эмпирикалық ғылым ретінде тіл білімінің тарихымен тығыз байланысты деуге болады.

Корпустық лингвистикада қолданылатын технологиялар электрондық компьютерлерден де көп бұрын пайда болған: олардың көпшілігі тіл білімі ең алғаш «нақты ғылым саласы» немесе эмпирикалық ғылым деп жарияланған XVIII ғасырдың аяғы мен XIX ғасырлардан басталады. Корпустық лингвистиканың негізіне жататын төмендегі үш салада қолданылған технологиялар қазіргі кездегі корпустық лингвистиканың дамуына әсер етті (және керісінше) [7].

Тарихи лингвистика: салыстырмалы-тарихи әдіс. Заманауи корпустық лингвистикаға әсер еткен басты бағыт салыстырмалы-тарихи тіл білімі болды, себебі тарихи зерттеулермен айналысатын тілшілер әруақытта мәтіндерді немесе мәтіндер жи-

нағын негізгі айғақ ретінде қолданып отырған. Ертедегі тілдерді реконструкциялау үшін немесе тіларалық байланысты анықтау үшін XIX ғасырда пайда болған көптеген технологиялар осы кезге дейін қолданыс тауып жүр. Үндіеуропалық дәстүр бойынша тілдік өзгерістерді анықтау мен реконструкция жасау мәселесі ерте кездері жазылған тарихи ескерткіштерге байланысты болған. Ғалым Я. Гримм және одан кейінгі кездегі младограмматиктер тілдердің тарихы мен грамматикалары жайындағы өз пайымдауларын мәтіндерден алынған дәйексөздермен дәлелдеп отырған.

XIX ғасырдан бастап дамып келе жатқан көптеген идеялар мен технологиялар корпустық лингвистикада қолданыс тауып, біртіндеп дамыды. Тарихи корпустар жасау ісі бұрынғыша әлі де қызығушылық тудырып отыр. Шындығында да, электрондық пішінде қолжетімді алғашқы корпустар ішінде тарихи корпустар да болды.

Электрондық форматтағы қолжетімді көптеген мәтіндердің пайда болуы көлемді тілдік деректерді барынша тез арада жинақтап алуға мүмкіндік тудырды. Ал бұл жағдай тілшілерге статистикалық әдістер есебінен лингвистикалық талдаулар жасауға, сондай-ақ зерттеулер үшін жаңа әдістер мен модельдерді қалыптастыру мен дамытуға мүмкіндік берді. Қазіргі кезде тарихи электрондық корпустары деректері негізінде тілдік өзгерістердің күрделі математикалық модельдерін шығарып алуға мүмкіндік бар.

Грамматикаларды жазу, лексикография және тілге үйрету. XIX ғасырдың грамматистері өз тұжырымдарын дәлелдеу үшін танымал авторлардың шығармаларынан алынған мысалдарды келтірген. Мысалы, Г.Пауль өзінің неміс тілінің грамматикасында фонология, морфология және синтаксис салалары бойынша айтылған пікірлерін дәлелдеу үшін неміс классиктерінің шығармаларынан алынған мысалдар беріп отырған. Бүгінгі күні де грамматиканы құрастырушылар корпустық тәсілді қолдануларына болады, қазіргі корпустарға тек классиктерді ғана емес, басқа да мәтіндер типтерін енгізуге болады. Сондай-ақ, қазіргі кезде сөйлеу тілінің грамматикасына көп көңіл бөлінуде. Тілді грамматикалық деңгейде сипаттауда корпустарды әртүрлі грамматикалық бірліктер варианттарының, түрлерінің және т.б. қолдану жиіліктері жайлы да ақпарат алу үшін пайдалануға болады.

Лексикография саласы бойынша жасалған еңбектерде де корпустық тәсілді пайдаланылғанын байқауға болады. Мәселен, XVIII ғасырдың ортасында С. Джонсон ағылшын тілінің түсіндірме сөздігін жазу үстінде (Dictionary of the English language, 1755), кітаптар ішінен «цитаталар» деп атаған иллюстрациялық сөйлемдер таңдап алып, сөздерді ағылшын авторлары қалайша қолданатынын мысалдар арқылы көрсеткен. С.Джонсон шығарма мәтінін оқу кезінде сөз мағынасы ашыла түсетін сөйлемдерді таңбалап, белгілеп қоятын болған. Содан соң оның ассистенттері белгіленген сөйлемдерді қағаз бетіне жазып алып отырған. Ал Джонсон болса, ол мәліметтерді сөздік мақалаларындағы сілтемелер ретінде пайдаланып, сөздік құрастырған.

Өлі тілдердің көптеген сөздіктерінен қажетті сөздер кездесетін контекстерден цитаталар алуға болатыны сияқты қазіргі кездегі корпустық лингвистикада бұл әдіс пішін жағынан конкорданспен сәйкес келеді. Қазіргі кезде сөздіктер құрастыруда корпустық конкорданстар арқылы мысалдарды жүздеп табуға болады. Компьютер көмегімен мысалдарды іздеу мен классификациялау және көпсөзді бірліктерді бөліп алу әрекеттері жеңілдегеніне қарамай, корпустан мәтіндерді пайдалану әдісі әлі де болса, компьютерлік технология өмірге келмей тұрғандағы лексикографтар мен филологогтардың зерттеу әдістеріне ұқсас.

Дәстүрлі мектеп грамматикалары мен оқулықтарында көбінесе жасанды құрастырылған немесе редакцияланған мысалдар беріледі. Болашақта олардың студенттерге көмегі аз болуы да мүмкін, себебі тілдік бірліктердің тілдесімдегі шынайы қолданысы мүлде басқаша. Осыған қатысты корпустар эмпирикалық дереккөз ретінде лингводидактикада маңызды рөл атқарады. Тілді оқыту немесе үйрету кезінде корпустар студенттердің қызығушылығын арттырады, тілді өз бетінше зерттеп-үйренуге ықпал етеді. Корпустық деректердің маңызды қолданысы – Computer-Assisted Language Learning (CALL), бұл компьютерлік бағдарлама жүйесі корпустар негізінде құрастырылып, студенттерді компьютер көмегімен интерактивті әдіспен оқытуда қолданылады.

Әлеуметтік лингвистика: тілдік әртүрлілік. Вариациялық лингвистика диалектологиялық карта және диалектілік мәтіндер жи-

нағын құрастырумен айналысқан XIX ғасырдың соңғы үштігінен бастау алады. Оның әдістері сол кезде қолданыста болған тарихи лингвистиканың әдістеріне ұқсайды, тек бір ғана маңызды белгісімен ерекшеленеді: диалектілік корпустар жүйелі түрде белгілі критерийлер бойынша құрастырылады. Күні бүгінге дейін корпусқа диалектілерді, диалектілік ерекшеліктерге негізделген мәтіндерді енгізу-енгізбеу туралы даулы мәселе бар.

Қазіргі кезде электрондық корпустар тілдік әртүрлілікті (диалектілер, жаргондар, сленгтер, жанрлар т.б.) зерттеу үшін жиі қолданылып жүр. Математикалық әдістер (мысалы, мультифакторлық талдау) толығымен осындай деректердің қолжетімділігіне сүйенеді.

Қазіргі кездері корпустық лингвистика осындай әдістерді қолданады және оларды әрі қарай дамытуда. Үлкен көлемді электрондық пішіндегі мәтіндер мен заманауи компьютерлік техника қолжетімді болған жағдайда, тілдік зерттеулер жүргізу де әлдеқайда жеңіл әрі нәтижелі болар еді. Табиғи тілдегі мәтіндерді өңдеуге арналған қазіргі интеллектуалды программалық жүйелердің дамуы да көлемді эксперименталдық тілдік қорды қажет етеді. Ал мұндай тілдік қор корпустар құрастыру арқылы ғана мүмкін болмақ. Қазіргі ғылым мен техниканың дамуы кезеңінде корпустық деректерге сұраныс артып отыр. Оған сәйкес келетін техникалық мүмкіндіктер де жеткілікті. Сондықтан, техникалық мүмкіндіктерді пайдаланып, лингвистикалық зерттеулер жүргізудің теңдесі жоқ құралы болып табылатын корпустар құрастыру ісі бүгінгі күннің аса өзекті мәселесі болып табылады.

1.3 Лингвистикалық корпустардың тарихы

«Корпустық лингвистика» Н.Хомскийдің көзқарасына қарама-қарсы бағытта пайда болды деуге әбден болады. Себебі, оның оппозициялық «құзырет/түсінігі» («компетенция/представление») бойынша тіл жайындағы деректер тіл иесінің құзыретін құрайды да, сондықтан да әртүрлі аса көлемді тілдік нысанды зерттеудің орнына сөйлеушінің тілдік интуициясына (ішкі түйсігіне) сүйенгенін жеткілікті деп санайды. Корпустық лингвистика мен оған

қатысты мәселелер теориялық және эмпирикалық білім аяларының ең ұтымды арақатынастарын іздестіру жолын бейнелейді деуге болады.

Осы кезге дейін тіл-тілдерде жинақталған мәтіндер корпусын әзірлеу тәжірибесінің ішінен төмендегідей іргелі мәтіндер корпустарын атауға болады.

Лингвистер компьютерленген мәтіндерді алғашқы корпусқа 1960 жылы жинаған болатын. Корпустық лингвистика У.Фрэнсистің жетекшілігімен 1963 ж. АҚШ-та өмірге келген Браун корпусынан (The Brown Standard Corpus of American English) бастама алады. Браун университетінде қазіргі ағылшын тілінің американдық нұсқасының Браун корпусының компьютерлік нұсқасы да, микроафишалары да бар. Бастапқыда бұл корпустың көлемі 1 млн. сөзқолданыстан тұрып, оның құрамында әрбіреуі 2 мың сөзқолданысқа тең 500 мәтін қамтылған және ол морфологиялық белгіленіммен жабдықталған. Корпусқа алғаш 1961 жылы шыққан және ағылшын тілінің әдеби нұсқасына жататын мәтіндер енгізілген. 1980 жылы Браун корпусының индекстелген (аннотацияланған) нұсқасы шықты, онда сөзформалар лемматизацияланып, олардың үстіртін синтаксистік қызметтері белгіленеді және т.б. Бұл корпустың авторлары У. Френсис (W. Francis) және Г. Кучера (H. Kucera) корпусты көптеген алғаш рет статистикалық өңдеуден өткен материалдармен (жиілікті, әліпбилі-жиілікті сөздіктермен және әртүрлі статистикалық үлестірімдермен) толықтырады.

Ғалымдар У.Н.Фрэнсис пен Г.Кучера бұл корпусты құрастыру үшін мынадай принциптер мен критерийлерді ұстаным етті:

1. Мәтін иесі ретінде американдық авторлар мен ағылшын тілінің американдық нұсқасын қолданатын авторлар шығармаларын алу;

2. Мәтіндерді синхрондау (мезгілдік үйлестіру) үшін АҚШ-та 1961 жылы алғаш рет баспадан шыққан еңбектерді алу;

3. Жанрлардың арақатынасын таңдау жұмыстары үшін ерекше ықтималдықтар рәсімін қолдану: барлығы 15 жанр, 9 – «ақпараттық» проза, 6 – көркем проза; әрбір жанрдан 6-дан 80-ге дейін таңдама мәтін алу;

4. Мәтіндерді таңдауға қойылатын талаптарды айқындау мен

тұжырымдау: әр мәтіннің көлемі статистикалық тұрғыда, оның стильдік ерекшеліктерін ақиқат дәрежеде бейнелеуін ескеру.

5. Жанрлардың құрамы мен арақатынасы, олардың стильдік ерекшеліктері мен қатынастық салмағы барабарлық деңгейде көрініс табуы қажет.

Браундық корпусты құрастырудың мақсаты – жазба ағылшын тілінің жеке жанрларын жүйелі түрде зерттеу және оларды салыстыра зерттеуді қамтамасыз ету болды. Браундық корпустың өмірге келуі жалпы көпшіліктің қызығушылығымен бірге қызу айтыс, пікірталас та тудырды. Ең бірінші ретте, олар мәтінді іріктеу принциптеріне және мұндай корпуста шешілетін мәселелерге қатысты болды. Бір жағынан ол статистикалық әдістердің негізінде құрылса, екіншіден, статистикалық әдістің қолданылуы корпус авторының кәсіби интуициясына да негізделді. Мұндай күрделі процестің барынша объективтілігіне жету үшін максималды формалданған, тексеру мен бақылауға барынша мүмкін болатын айқындылық, нақтылықты қамтамасыз ету қажет болды [4].

Сонымен, Браундық корпус осыған ұқсас корпустар құрастыруға қатысты зерттеулердің кеңінен тараған нысаны мен стандартына айналды. Ғалымдар арасында көптеген лингвистикалық зерттеулерді сапалы жүргізу тек ауқымды тілдік материалдар негізінде ғана жүзеге асатындығын ұғыну байқала бастады. Осы айтылғандардың барлығы мәтіндерді корпус түрінде ұйымдастыру ережелері мен оларға талдау жүргізу әдіснамасын зерттейтін бағыттың пайда болуына себепші болып, ғалымдарды одан әрі ынталандыра түсті.

Кейінірек еуропалық зерттеушілер 1961 жылы Ұлыбританияда алғаш рет мәтіндер корпусын жасады. Бұл корпус та Браундық корпус тәріздес 15 жанрды қамтитын 500 мәтіннен 2000 сөзқолданыстан алып, 1 млн. сөзқолданыстан тұратын ағылшын тілінің британдық жобасын ұсынды. Оны Ланкастер-Осло-Берген (The Lancaster-Oslo-Bergen Corpus, британдық және екі норвеждік университеттер, немесе қысқаша LOB) корпусы деп атайды. Браундық корпус сияқты теңгерімделген (балансталған) корпустар зерттеушілер үшін тілдік талдау мен зерттеулер жасау үшін аса маңызды саналады. Lancaster/Oslo-Bergen Corpus-тың құрамында син-

таксистік белгіленіммен жабдықталған екі кішігірім корпус бар: Leeds-Lancaster Treebank және Lancaster Parsed Corpus.

Сонымен, екі алғашқы үлкен көлемді корпус – ол жазба тілінің американдық және британдық нұсқаларының ағылшын тіліне арналған корпусы. Бұл екі корпус та осы күнге дейін маңызды саналады, себебі ағылшын тіліндегі көптеген сапалы зерттеулер осы екі корпусқа негізделген.

Германиядағы корпустық лингвистика. Браун корпусының пайда болуы неміс-ағылшын машиналық аударма жүйесі жобасының аясында жасалған LIMAS-корпусының жобасын әзірлеуге түрткі болды (LIMAS – 1965 жыл мен 1976 жыл аралығында Бонн университетінің Коммуникация және фонетика проблемаларын зерттеу институында жұмыс істеген Linguistische und MAschinelle Sprachübersetzung зерттеу тобының қысқарған атауы).

Неміс сөйлеу тілі мәтіндерінің корпустары. Сөйлеу тілі мәтіндерінің корпусын жасау Германияда ертерек басталды. Сөйлеу тілі корпусын сақтаудың негізгі формасы магниттік таспаға жазып алу болды. Алайда кейінірек акустикалық форма фонетикалық транскрипцияға ауыстырылды. Фонетикалық транскрипцияның толық нұсқасы транскрипциялаудың және пайдаланушының қабылдауының қиындығынан сәтті болмай шықты. Сөйлеу тілін графикалық түрге енгізу бойынша жасалған бірнеше тәжірибенің нәтижесінде фонетикалық транскрипцияны айтудың кейбір ерекшеліктерін сақтаған және сөйлеу актілерінің ізбе-іздігін, яғни бір уақытта жүруін бейнелейтін жартылай нұсқасы жасап шығарылды. Фонетикалық транскрипцияның қысқартылған нұсқасында интонацияны, айтылымның аймақтық ерекшеліктерін, сөйлеудің айқындылық дәрежесін және т.б. бейнелейтін құралдар қолданылды.

Неміс тілі жазбаша мәтіндерінің корпустары. Қазіргі уақытта лингвистердің қолында машина оқитын формадағы мәтіндердің мол қоры бар. Алайда неміс тілінде корпустағы проблемалық аумақты бейнелеу принциптері ашық талқыланбайды.

Неміс мәтіндерінің корпусы туралы айтқанда *DeReKo* (das Deutsche Referenz Korpus) корпусы жайлы айту қажет, ал оған интернет желісі бойынша кіруді http://www.ids-mannheim.de/kl/projekte/korpora/ адресі бойынша жүзеге асыруға болады. Ман-

геймдегі (Германия) неміс тілі Институтының ғылыми жобасы бойынша құрастырылған электронды жиын беллетристика, ғылыми және публицистикалық мәтіндерді қамтитын көлемі 4 млрд-тан астам сөзқолданыстан тұрады (16. 08. 2010 ж. бойынша). Бұл әлем бойынша ең үлкен көлемді корпус болып саналады. Бірақ ол неміс тілінің жеке шағын корпустары (подкорпустары) түрінде жасалған. Корпус TEI ұсыныстарына сәйкес зерттеліп, SGML-ге негізделген морфосинтаксистік белгіленімді пайдаланған. COSMAS II корпустық менеджермен қамтамасыз етілген неміс корпусы лексикалық бірліктер мен созформалардың морфологиялық белгілері арқылы қажетті деген іздестірулерді жүзеге асыра алады.

Франциядағы корпустық лингвистика. Тіл тарихында аса қомақты француздық жобалардың бірі 90 млн. сөзқолданыстан тұратын мәтіндер корпусын қамтыған «Француз тілінің қазынасы» атты еңбек болды. Жобаны әзірлеу 1963 жылы басталды. Негізгі материалды енгізу 1968 жылы аяқталды. XIX-XX ғасырлардағы мәтіндер енгізілген. Корпус XIX-XX ғасырлардағы француз тілінің 80 мың лексикалық бірліктен тұратын 15 томдық «Tresor de la langue francaise» сөздігін дайындауда пайдаланылған.

Осы корпустарды құрастырғаннан кейінгі өткен бірнеше онжылдықтардан соң компьютерлер арзандап, әлеуеттілігі де бұрынғыдан артып, сол сияқты бағасы арзан, әрі сенімді сканерлер мәтіндерді компьютерде теруді қажет етпейтіндей етті. Сөйтіп, жаңа техника көмегімен корпустарды құрастыру үдерісі де жеңілдеді. Соның арқасында қазіргі корпустар миллиардтаған сөзқолданысты қамтиды.

1990 жылдары әлемде 600-ге жуық компьютерлік корпус жасалды. Олардың әр жылдардағы өсімін, шамамен, былай көрсетуге болады.

1-кесте

Әр жылдары құрастырылған корпустардың сандық үлестірілімі

Жылдар	1965	1966–1970	1971–1975	1976–1980	1981–1985	1986–1990
Корпус саны	10	20	30	80	160	320

Кестеден әсіресе соңғы жылдары корпустардың саны көбейіп, корпустардың түрлері де тек артып отырғанын байқаймыз.

Қазіргі американдық ағылшындық корпус (The Corpus of Contemporary American English – COCA) ағылшын тілінің ең көлемді корпусы болып саналады. Аталған корпус баланстандырылған және өзіне теңдесі жоқ ағылшын тілінің американдық нұсқасы. Оған мына http://corpus.byu.edu/coca/ адреспен еркін кіріп, пайдалануыңызға болады. Корпус 2008 жылы ғалым М.Дэвистің жетекшілігімен құрастырылған (Brigham Young University, США) және 1990 жылдан осы кезге дейінгі мәтіндерді бірге есептегенде 410 млн. сөзқолданыстан тұрады. Бұл корпуста ауызша тіл, көркем әдебиет, көпшілікке арналған журналдар, газеттер және ғылыми әдебиет мәтіндері біркелкілік сипатта орын алған.

Корпус базасы жылына екі рет жаңартылып тұрады және тілде болып жатқан өзгерістерді бақылап тұруға аса ыңғайлы.

Қазіргі кездегі ағылшын тіліндегі корпустар ішінен ең танымалы: **Британдық ұлттық корпус** (British National Corpus – BNC), Халықаралық ағылшын тілінің корпусы (International Corpus of English – ICE), лингвистикалық ағылшын тілінің Банкі (Bank of English), Қазіргі американдық ағылшын корпусы (Corpus of Contemporary American English – COCA) және т.б. Қазіргі кезде әлемнің көптеген тілдерінің корпустары жасалған.

Корпустық лингвистиканың ағылшын тіл білімінде кең етек алуын ғалымдар АҚШ-та компьютерлік техника мен XX ғасырдың 60-80 жылдары британ лингвистикасындағы интеллектілік ахуалдың белсенді дамуымен түсіндіреді. Осы кездерде тілдік зерттеулердің ең көп бөлігі компьютерленген мәтіндік корпустарға лингвистикалық талдау жүргізуге бағытталғандығы мәлім. Мұндай зерттеулердің нәтижесі 2001 жылы Бирменгем университетінде корпустық лингвистика кафедрасын ашуға және International Journal of Corpus Linguistics журналын баспадан шығарып тұруға мүмкіндік туғызды. Бірнеше тілдің материалдары бойынша құрастырылған, түрі мен қызметі жағынан ерекшеленетін корпустар және солардың негізінде неше түрлі сөздіктер түзіліп, грамматикалар жазылды.

Британдық ұлттық корпус (BNC) ең бір үлкен эталондық корпустар қатарына жатады және онда 100 млн.-нан арта ағылшын тілінің сөйлеу тілі мен жазба тілінің мәтіндеріндегі сөзқолданыстар қамтылған. Британдық ұлттық корпус Оксфорд университетінде және Ланкастер университеті мен Британдық кітапхана қызметкерлерінің қатысуымен құрастырылған болатын. Бұл корпусты құрастыруға байланысты іс-шаралар 1991 жылдан бастап 1994 жылға дейін жалғасты. Оның 90% жазба материалдардан, ал 10% сөйлеу тілі материалдарынан тұрады. Ағылшынның жазба тілін бейнелейтін шағын корпус (немесе орыс. подкорпус) мынадай жанрлардан тұрады: газеттер, мерзімдік ғылыми шығармалар және әр жастағы адамдарға арналған журналдар, көпшілікке арналған ғылыми фантастика, баспадан шыққан және шықпаған хаттар, мектеп және жоғары оқу орындары оқулықтары және т.б. Ауызша тілдің шағын корпусына (подкорпусы) өз еркімен жобаға қатысқысы келген әр жастағы, Ұлыбританияның әр бөлігінде тұратын және әртүрлі әлеуметтік топқа қатысты адамдардың сөйлеу тілі мәтіндері енгізілген болатын. Сөйлеу тілі неше түрлі контекстік қоршауда кездеседі: іскер адамдардың сөзі немесе үкімет адамдарының сөздерінен бастап, радиошоулар мен телефон арқылы сөйлеген сөздер т.б.

Корпустағы мәтіндер морфологиялық деңгейдегі белгіленімнен өткен. Британдық ұлттық корпустағы барлық мәтіндер сөйлемдерге сегменттелген. Сөйлем ішіндегі сөздерге, олардың грамматикалық класқа немесе сөз табына қатыстылығына қарай, алдын ала келісілген шартты белгілер (маркелер) сәйкестендіріледі. Мәтіндегі тыныс белгілеріне де сәйкес келетін маркерлер ұсынылады. Ланкастер университетінде зерттеліп, құрастырылған CLAWS компьютерлік программа бойынша сегменттеу мен сөздерге автоматты түрде тэгтерді (маркерлерді) қою жүзеге асырылған. Тэгтерді автоматты түрде белгіленім қою әрекетінде қате жіберу шамамен 1,7% құрайды. Егер белгіленімді автоматтандыру кезінде біріздiлік болмай, бір сөзге екі маркер (немесе одан көп) сәйкес келіп жатса, оларды сөз соңына дефис арқылы қою келісілген (мысалы, VVD-VVN – бірінші – өткен шақ етістік, ал екін-

шісі – өткен шақ есімше). Мұндай «синонимдік» маркерлер барлық корпустың 4,7% алады.

Корпус тек Ұлыбританияда қолданылатын қазіргі ағылшын тіліндегі сөздерден ғана тұрады. Бірақ Ұлыбританияда қолданылғанымен, корпуста шыққан тегі британдық емес және шетелдерден енген сөздер де кездеседі. Сол сияқты корпустағы мәтіндер әртүрлі стильдерге қатысты және тақырыптық жағынан да шектелмеген, мәселен, корпуста жазба және ауызша тілдерден де мысалдар кездеседі.

Британдық ұлттық корпусқа мәтіндер негізгі үш критерий бойынша таңдалып алынған: уақыт, берілген мәтін сипатталатын облыс және баспаның типі. Уақыты бойынша барлық мәтіндер, шамамен алғанда, 1975 жылдан басталатын бір кезең. Ерекшелік ретінде, тек ойын-сауыққа қатысты әдебиеттер қазірге кезеңге де қатысты. Британдық ұлттық корпуста (БҰК) 1964 жылдан бастау алатын әдеби шығармалар да орын алған, ал бұл корпустың синхрондығының бір дәлелі. Ойын-сауыққа қатысты әдебиеттер барлық мәтіннің 25%, ал қалған жазба мәтіндердің 75 % ақпараттық басылымдардан алынған (ғылым, өнер, коммерция мен финанстар, бос уақыт, әлеуметтану, әлемдік шолу). Сонымен бірге мына мәселелер де ескерілген: көлем (сөздер саны), мәтінде талқыланатын тақырып, автордың аты-жөні, жасы, жынысы, туылған жері, мекен-жайы, мәтін арналған жас тобы және мәтіннің күрделілік деңгейі.

Барлық 10 миллиондық ауызша тілдің субкорпусы (шағын корпусы) шамамен бірдей екі бөлікке бөлінген: 1) «спонтанды» (дайындықсыз), табиғи диалогтар транскрипциясын сақтайтын *демографиялық* бөлік; 2) маңызды рөл атқаратын *контекстік-басқарылатын* бөлік. Бұл бөлікте қайсыбір көпшілікке арналған шаралардан жазылып алынған ауызша мәтіндер бар.

1) Демографиялық бөлік. Диалогтарға өз еркімен қатысқан Ұлыбритания жерінің барлық аумағынан 124 адам қатысқан. Олар бірнеше күн бойы өздерімен бірге магнитофон аппараттарын алып жүріп, сөйлеген сөздерін таспаға жазып алған және ол әрекеттерін қалта кітапшасына жазып отырған, яғни әңгімелерінің қандай жағдайда болғанын, диалогқа қатысқан коммуниканттар жайлы

мәліметтерді және т.б. тіркеп отырған. Өз еркімен қатысушылар шамамен бірдей деңгейде таңдалып алынған, яғни белгілі бір жас жағынан шамалас ер адам мен әйел адам саны жағынан және әртүрлі әлеуметтік топтан таңдалған. Олардың айтқан сөздері толығымен жазылып алынған. Содан кейін бұл магниттік таспадағы жазбалар өңделіп, мәтіндері әдеттегі ағылшын орфографиясына сай көшіріліп жазылған. Бұл құжаттар қазіргі кезде ауызша тілдің негізі ретінде қолданыс тауып, нәтижелері әрі пайдалы, әрі қызығушылық тудырады [1].

2) *Контекстік-басқарылатын* бөлік. Корпусты құрастырушылар келесі төрт әлеуметтік контекстің айтарлықтай кең сипаттағы категорияларын таңдап алып, олардан тең көлемді жазбаларды жинақтауды мақсат еткен:

– білім беру мен ақпараттық жиналыстар: дәрістер, жаңалықтар бағдарламасы, кейбір мәселелерді талқылау, семинарлар;

– көрме, консультация, интервью, сауда мекемелеріне қатысты іскери уақиғалар;

– діни уағыздар, саяси отырыстағы сөздер, парламент мәжілісі сияқты көпшілік уақиғалар;

– бос уақытқа қатысты тақырыптар, мысалы, спорттық комментарийлер, клубтық кездесулер.

Корпусты құрастырушылар SGML атты белгіленім негізінде өздерінің SARA (SGML Aware Retrieval Application) атты программаларын жазып шыққан. Ең басында SARA программасы клиент/сервер қызметін атқаратын жүйе ретінде жасалған болатын, яғни бір немесе бірнеше компьютер желі бойынша орталық сервер жүйесімен қатынас жасау қызметін атқарады. Ал қазіргі кезде XAIRA (XML Aware Indexing and Retrieval Architecture) атты корпустық менеджердің жаңа түрі жасалды.

Келесі ең танымал жалпы типті корпустардың бірі – **Чехтың ұлттық корпусы** (Český národní korpus) (әрі арай ЧҰК). Чех тілінің 100 миллиондай сөзқолданыстан тұратын мәтіндер корпусын айтарлықтай көлемді деуге болады. Бұл қазіргі кездегі чех тілін бейнелейтін синхрондық типтегі морфологиялық деңгейде белгіленген корпус. Аталған корпус проф. М.Чермактың ғылыми жетекшілігімен Чехтың Ұлттық корпусы (ЧҰК) институтын өмір-

ге әкелген болатын. Бұл Институт 1994 жылы Прагадағы Карлова университетіндегі философия факультеті негізінде ашылып, Білім беру министрлігінің қолдауымен гранттар мен спонсорлардың қаржыландыруының арқасында қызмет атқаруда. Институттың жұмысымен және корпустың өзімен мына http://uchk.ff.cuni.cz сайт бойынша танысуға болады.

Корпуста жинақталған чех тіліндегі мәтіндер массиві синхрондық және диахрондық бөліктер болып ажыратылады. Синхрондық бөлік, өз ретінде, мынадай бөліктерден тұрады: жазба мәтіндер (100 млн. аса сөзқолданыс), сөйлеу тілі (750 мың сөзқолданыс) және диалекті тілі. Диахрондық бөліктің көлемі – 1750 мың сөзқолданыс. Одан басқа жеке қор ретінде сөйлеу тілінің Прагалық корпусы жасалуда.

ЧҰК-ын қалыптастырғанда корпустың репрезентативтілігі мәселесіне көп назар аударылды. Корпустың негізгі бөлігін 1990-2000 жылдар аралығындағы мәтіндер қамтылуы керек болған және 1950 жылға дейінгі чех әдебиетіне қатысты шығармаларды бейнелейтін қосымша ретроспективтік бөлікті де қоса есептеу қажет деп шешім шығарылған.

Кітаптану саласындағы зерттеулердің нәтижесінде ЧҰК-дағы мәтіндердің жанрлық және тақырыптық құрылымы 2-кестедегідей көрініс табады.

2-кесте

ЧҰК-ның фрагменттік және тақырыптық құрылымы

Көркем әдебиет мәтіндері	15 %
Ақпараттық мәтіндер	85 %
Оның ішінде:	
Публицистикалық	60 %
Ғылыми	25 %
Оның ішінде:	
Әлеуметтік ғылымдар	3,6 %
Жаратылыс ғылымдар	3,4 %
Өнертану	3,4 %
Техникалық ғылымдар	4,6 %
және т.б.	

Барлық мәтіндер үш түрде сақталады: мәтіндер мұрағаты (бастапқы қалпында), деректер банкі (SGML тілінде өңделген мәтіндер) және корпустың өзі (арнайы форматтағы мәтіндер және оларды пайдалану құралдары). Бастапқы күйдегі мәтіндер конвертирленудің (айырбастаудың) бірнеше кезеңінен өтеді, яғни мұндай әрекеттің нәтижесінде олар қайта кодтау (қажеттікке қарай), мәтінді құрылымдау, мәтіндік емес және шет тілдік элементтерді, қайталауларды жою және белгіленімнің өзі жүзеге асырылады. Нәтижесінде мынадай құрылым қалыптасады: файл, құжат, абзац, сөйлем, сөзформа. Әрбір құрылымдық бірлік SGML тілінің элементі ретінде қалыптасады. Әрбір файлдың бас тақырыбы конвертирлеу үдерісінің сипаттамасын бейнелейді, құжаттың бас тақырыптық жағы библиографиялық және типологиялық белгілерден тұрады (автор, шығарма аты, дереккөз, басылым жылы, мәтін типі, жанр және т.б.). Лингвистикалық белгіленім ұғымы – сөзформаларды лемма түріне келтіру (лемматизациялау) және оларға, SGML тілінің жеке элементі ретінде позициялық форматта морфологиялық сипаттамаларды «тіркеп» жазу (приписывание) үдерісінен тұрады.

Чех тілі мәтіндерінің синхрондық корпусы пайдаланушыларға екі түрде бейнеленеді: SYN2000 толық корпус (100 млн. сөзқолданыс) және SYN2000-тің ішкі жиыны ретінде ондағы барлық жанрдың пайыздық қатынасын сол қалпында сақтайтын PUBLIC корпус түрінде (20 млн. сөзқолданыс). PUBLIC корпус Интернет желісінде еркін қолжетімді және ол функционалдық қызметі жағынан мынадай шектеулерге ие:

– іздеу тек бір сөзформа бойынша;

– берілген сөздің қоршауындағы контекст шектеулі (≤ 60 белгіден);

– бастапқы сөздің морфологиялық сипаттамасын қарау мүмкін емес.

ЧҰК институты сайтында контекстен сырт тұрған берілген сөздің сипаттамасын алу мақсатында морфологиялық талдаушыға (анализатор) жеке кіру (қатынас жасау) мүмкіндігі қарастырылған. Интернеттегі SYN2000 корпусына да тегін қол жеткізуге болады, тек ол үшін коммерциялық емес пайдалану мүмкіндігі жайлы жаз-

баша келісім хатпен заңдастыру қажет болады. Корпуспен жұмыс істеу BONITO (қара п. 3.1) корпустық менеджер арқылы жүзеге асады.

Корпуспен бірге әрекет ететін лингвистикалық бағдарламалық қамтамсыз ету құралы мәтін конкордансын құрастыратын бағдарлама арқылы сөзформалардың қолданысын бейнелейтін контекстерімен бірге сол сөзформаның корпуста кездесу жиілігін де қоса тауып беру мүмкіндігі бар. Сонымен бірге, морфологиялық талдау жасайтын бағдарлама мәтіндегі грамматикалық омонимия мәселесінің шешімін табу үшін, өзінің тікелей міндетімен бірге, контекстік талдауды да іске асыра алады.

Орыс тілінің алғашқы корпустары. Алғашқыда орыс тілінде репрезентативті (тұлғалы) корпус болмады. Себебі, орыс тілі үшін жеке мәтіндер жинағын әртүрлі тілдік ішкі жиындардың корпусы деп қарастыруға болады. Мұндай жиындар әртүрлі электрондық кітапханаларға жинақталған. Бұлардың кейбіреулері қор құрастыруда ғылыми тәсілдерге сүйенеді: ғаламдық қамту және ұсынылған материалдардың толықтығы, жариялынымдардың жоғары филологиялық және технологиялық стандарттарға сай келуі. Мысалы, мұндай кітапханалар қатарына Орыстың виртуалды кітапханасы (ОВК) (русская виртуальная библиотека) және Фундаментальды электрондық кітапханасы (ФЭК) (фундаментальная электронная библиотека) жатады. ОВК-да классикалық және орыс әдебиетінің қазіргі шығармаларының авторлық дерекнамалары мен қосымша ретінде ақпараттық-коментаторлық аппараты берілген. ФЭК-та XVI-XX ғасырлардағы орыстың ауызша және жазбаша тілі (русская словесность) және орыс фольклоры, әдебиеттану ғылымы мен тарихи-өмірбаяндық жұмыстары бойынша толықмәтінді ақпараттық жүйе көрініс тапқан. ФЭК кітапханасы әрбіреуі жеке авторға, жанрға немесе шығармаға арналған электрондық ғылыми басылымдардан тұрады. ОВК пен ФЭК кітапханалары белгілі бағдарламалық-тілдік құралдармен жабдықталған. Оған ЯНДЕКС атты іздестіру мәшинесі негізіндегі іздестірудің лемматизациялау жүйесін жатқызуға болады. Сонымен бірге ОВК орыс тілінің сөзформаларын туындататын кіріктірілген модульмен жабдықталған. Мұндай модуль берілген құжаттар жиынтығы үшін сөзформалар

мен синтагмалардың индекстерін құруға мүмкіндік жасайды. ОВК мәтіндері оның құрылымын (бөлімдер, тараулар, абзацтар және т.б.) сипаттайтын белгіленімді өз құрамында сақтайды. TEI жобасының шеңберінде зерттелген электрондық құжаттарға белгіленім жүргізу сызбасына сәйкес SGML құралдары бойынша мәтін бірліктеріне белгіленім үрдісі жүзеге асырылады. Бірақ басқа да кітапханалар сияқты, ОВК пен ФЭК ресурстары тілдік тұрғыда «шикі» болып табылады, яғни мәтіндер белгіленімнен өтпеген күйде көрініс тапқан. Сондықтан сол кездері корпустық лингвистиканың алдында тұрған негізгі мәселе, бірішіден, орыс тілінің барлық функционалды-стильдік қабаттары қамтылған мәтіндердің (және сөйлеу тілі жазбаларының) репрезентативті корпусын құрастыру, екіншіден, тілдік зерттеулерге қажет қолдану жағын ескере отырып, корпустағы лингвистикалық белгіленімнің, ең алдымен, морфологиялық түрін жасау міндеті қойылды.

Орыс тілінің бірінші корпусы *Упсальский машинный фонд русского языка* (Upsal'skij korpus russkix tekstov) деп аталып, Швецияның Упсаль университетінде құрастырылған болатын. Ол корпуспен Интернеттен мына адрес бойынша http://www.slaviska.uu.se/korpus.htm танысуға болады. Корпус 600 мәтіннен тұрады және оның көлемі 1 млн. сөзқолданыс. Ондағы мәтіндер арнаулы және көркем әдебиеттер нұсқалары арасында тең түрде бөлінген. Құрастырушылардың мақсаты бойынша қазіргі (современный) орыс тілінің корпусын жасау көзделген. Корпусты қалыптастыру мақсаты – ең алдымен, әдеби тілді корпуста бейнелеу болатын, сол себепті онда сөйлеу тілінің мәтіндері орын алмаған. Бұл корпусқа 1985–1989 жылдар аралығындағы арнайы мәтіндер (мәтіндердің фрагменттері емес, толық мәтіндер алынған), 1960–1988 жылдар аралығындағы көркем әдебиет мәтіндері таңдалып алынады. Көркем әдебиет мәтіндеріндегі өзгерістер қысқа мерзімде байқала қоймайтындықтан, олар корпуста ұзақ кезеңдік аралықты қамтып жатыр. Корпустың аннотациясында, корпусты құрастырушылардың көзқарасы бойынша, арнайы мәтіндердің ішінде тақырыптарға айырықша мән берілген (which *felt* to be more important), ал көркем әдебиет мәтіндері ішінен көпшілікке белгілі авторлардың шығармаларына артықшылық берілген. Соған орай, осы тарауға қатысты

корпустағы мәтіндердің көлемі айтарлықтай дәрежеде басқа тараулардан артық [8, 129 б.].

Арнайы әдебиеттерді саралау (классификациялау) тақырып бойынша іске асқан (барлығы 25 тақырыптық ая: биология, химия, физика және т.б.), ал көркем әдебиеттер – авторлар бойынша (барлығы 40 автор). Егер шығармалар авторы тақырып бойынша әртүрлі болса, онда таңдамаға барлық маңызды деген тақырыптық аялар алынған. Корпустағы әрбір мәтінге формалды сипаттаулар сәйкестікке келтірілген, яғни онда мәтіннің типі жайында ақпарат (арнайы vs. көрем әдебиет), дерекнама туралы, тақырып, автор жайында ақпараттар берілген [8]. Бастапқы мәтіндердің кирилица әліпбиі латын қарпі арқылы беріледі, ал басқа таңбалар пунктуацияны белгілеуге пайдаланылады (3-кесте).

Упсаль корпусы «Орыс мәтіндерінің Тюбинген корпусы» деп аталатын корпусқа жатады. Ол корпус 1990-жылдары Тюбинген университетінің арнайы SFB 441 ғылыми-зерттеу секторы жұмыстарының аясында онлайн-іздеу (http://www.sfb441.uni-tuebingen.de/b1/rus/korpora.html#uppsalakorpus) мүмкіндігін пайдалану үшін құрастырылды.

3-кесте

Упсаль корпусындағы пунктуацияны белгілейтін таңбалар

@	beginning and end of a main headline
@@	beginning and end of a subsection
&	new paragraph
*	capital letter marked as permanent, i.e., which should never become lower case
\	comments, e.g. \tablica\ instead of the table
{	marks the beginning of a string of Latin alphabet characters

Аталған корпустар морфологиялық аннотация тэгтерімен белгіленген. Белгіленім статистикалық тэгтер (TnT) көмегімен жүзеге асқан. Іздеу процесі сөзформалар бойынша да және морфологиялық тэгтер бойынша да жүргізіледі. Мәтіндерді олардың белгіленімімен бірге шығаруға (қағаз және экран беттеріне) бо-

лады. Іздеу сөйлемшесін енгізу үшін және іздеп табылған мәтінді шығару үшін келесі кодтаулардың біреуін таңдап алуға болады: кирилицаны (KOI8 немесе Windows-1251) немесе латын әріптерімен транслитерациялауды.

Корпуста екі түрлі іздеуді қолдануға мүмкіндік бар: қарапайым мәтіндік іздеу және күрделі іздеу. Күрделі іздеу мәтіндердің қосымша корпустарын қолдануға және аса күрделі іздеу сөйлемшелерін пайдалануға мүмкіндік жасайды. Қарапайым іздеуде Упсаль корпусы және сұхбат мәтіндерінің корпусы қолжетімді, ал күрделі іздеуде – Упсаль корпусы, сұхбат мәтіндерінің корпусы және басқа да көптеген мәтіндер қолжетімді, олар: қазіргі кездегі мәтіндер (көбінде публицистикалық), XX және XIX ғасырлардағы әдебиеттер.

Күрделі іздеу CQP программасы көмегімен жүзеге асады. CQP программасы Штутгарт университетінің Тілді мәшинелік өңдеу институында жасалып, көлемді корпустарды басқаруға арналған жүйе ретінде қолданылады.

XX-ғасырдың соңындағы орыс газет мәтіндерінің компьютерлік корпусы 2000–2002 жылдары МГУ-дің Филология факультетінің Жалпы және компьютерлік лексикология мен лексикография зертханасында құрастырылған болатын. Корпус үшін аса көп газет материалдарын таңдау (жалпы көлемі 11 млн. сөзқолданыстан тұратын 23110 мәтін) ісі орыс тіліндегі 13 Ресей газетінің (Правда, Завтра, Новая газета, Литературная газета) 1994–1997 жылдардағы толық нөмірлерін қамту негізінде жүзеге асты.

Мәтіндердің әртүрлі типтерін (мысалы, әртүрлі жанрлар және жанрлар типтері), олардың бірліктерін және өзара қатынастарын таңдаудың бұл принциптері біршама объективті және сенімді нәтижелерге қол жеткізуге мүмкіндік жасайды.

Құрастырылған корпусқа талдау жүргізу мен оны басқару Диктум-1 жүйесінің негізінде жүзеге асады. Бұл жүйе МГУ-дің Жалпы және компьютерлік лексикология мен лексикография зертханасында зерттеліп іске қосылған болатын. Жүйенің көмегімен корпустың мәтіндері мен бірліктері автоматты және жартылай автоматты түрде әртүрлі маркерлермен маркіленеді: мәтіндер (оның әрбір сөзқолданыстары) – газет-дереккөз, мәтін көлемі,

оның жанры, жарияланым датасы және т.б. маркерлерімен; сөзқолданыстар – грамматикалық, лексикалық, морфемдік және басқа категориялардың маркерлерімен белгіленеді.

Корпустың интернетте көрсетілім (демонстрация) вариантын дайындауда жалпы көлемі 200 мыңнан асатын сөзқолданыстан тұратын мәтін фрагменті бөлініп алынып, корпустың сөзқолданыстарын автоматты лемматизациялау мен морфологиялық сипаттама жасау және сонымен бірге сөзформалар мен лексемаларды морфемдік сегменттеу әрекеті жүргізіледі.

Жанрлық сипаттамаларды жүйелегеннен кейін кейбір нақты жанрлар біріктіріліп, 9 жанрлық тип айқындалған:

1. Таза ақпараттық жанрлар – авторлық көзқарасы берілмейтін, максималды объективті формада көрсетілген мазмұндағы ақпарат.

2. Ақпараттық-публицистикалық жанрлар – ақпараттың объективті түрде баяндалуы оның субъективті көзқараспен, түсіндіруімен бірге жүреді, сондықтан эмоционалды немесе интеллектуалды бағалау көрініс табады. Сонымен бірге, газеттерге тән емес жанрлар да осы топ ішіне енген: ғұмырнама, өтініш, ырымдар.

3. Таза публицистикалық жанрлар – оның мазмұны, автордың өзі өңдеген ақпарат: қайсыбір жағдайды дәлелдеу, пікір, сезімін білдіру және т.б. Оқырманға объективті жаңа ақпарат бұл жерде екінші дәрежелі рөл атқарады.

4. Көркемдік-публицистикалық жанрлар – көркем мәтіндерді құрастыруда әртүрлі бейнелі тәсілдер қолданылатын жанрлар.

5. Жарнамалық жанрлар

6. Көркемдік жанрлар.

7. Сөйлеу түріндегі жанрлар.

8. Ресми-іскерлік жанрлар.

9. Өзге де жанрлар – ойын, кроссворд, гороскоп және т.б., діни насихаттық жанр, қай жанрға жатқызу қиындық тудыратын жанрлар.

Сонымен, 1990 жылдардың бірінші жартысында корпустық лингвистика тіл ғылымының дербес бағыты ретінде толығымен қалыптасты деуге болады. Мұны 1992 жылы Я.Свартвиктің корпустық лингвистика жайында өткен бірінші Нобельдік симпозиум

материалдарының алғы сөзін «Корпусная лингвистика достигла зрелости» деген тақырыппен атауынан көруге болады. Корпустық лингвистика компьютерлік лингвистиканың бір саласы ретінде оның жетістіктерін қолдана отырып, өз кезегінде, компьютерлік лингвистиканың да жетістіктерін арттыра түсті деуге болады.

Орыс тілінің қазіргі корпустары. Көп уақыт бойы Интернет желісінде тілшілер пайдаланатын әрі көпшілікке қолжетімді, көлемді және белгіленім жүргізілген орыс тілінің корпусы қолданысқа енгізілмеді. Мұндай корпусты құрастыру жұмысына тікелей кірісу тек 2000 жылдан басталды деуге болады. Алайда корпус құрастыру ісіне байланысты кейбір іс-шаралар 1980-жылдардан-ақ бастама алған болатын [4]. Орыс тілінің Ұлттық корпусы – электронды түрдегі орыс тілінің мәтіндер жиынтығына негізделген ақпараттық-анықтамалық жүйе. Алғаш рет бұл корпус 2004 жылдың сәуір айында http://ruscorpora.ru/ атпен Интернет сайтына салынған болатын. Корпус орыс тіліне қатысты көптеген сауалдарға қызығушылық танытатын көпшілік қауымға бағытталып жасалған, атап айтқанда: кәсіби тілшілерге, орыс тілінен сабақ беретін мұғалімдерге, оқушылар мен студенттерге, орыс тілін үйреніп жүрген шетелдік азаматтарға т.б. Орыс тілінің Ұлттық корпусы теңгерімділік, көлемділік (репрезентативтік) критерийіне және қазіргі корпустарға қойылатын тағы да басқа талаптарға толық жауап береді. Оған мына сипаттамалар дәлел бола алады:

1) ОТҰК-ның көлемі 2011 жылдың ақпан айында 176 млн. сөзқолданыстан тұрған болатын (http://ruscorpora.ru/ сайттың деректері бойынша);

2) мәтіндердің жанрлық әртүрлілігі орыс тілінің қолдану аясының негізгі өрістерін қамтиды: ғылыми, ресми-іскерлік, публицистикалық, шіркеулік-діни, көркемдік, ауызша және электрондық коммуникацияны қоса санағанда тұрмыстық-сөйлеу жанрлары;

3) негізгі әлеуметтік параметрлер бойынша айтарлықтай әртүрлі (жасы бойынша, білім деңгейі мен тілді меңгеруі жағынан, мамандығы, тіл мәдениеті типіне қарай) мәтіндердің қамтылуы, корпусқа енген шығармалардың авторлық құрамы (20 мыңнан көп);

4) ОТҰК-ға енетін мәтіндер әр кезеңдік сипатта болуы, бұл тілдік құбылыстардың қолданыстағы өзгерістерін байқауға және ондай өзгерістердің динамикасын анықтауға мүмкіндік туғызады [9].

Корпустарды шартты түрде екі бөлікке бөліп қарастыруға болады – қазіргі және диахрондық. Орыс тіліндегі қазіргі мәтіндер корпусын құрайтын мәтіндер сол кездері 1951–2007 жылдар аралығында жазылған мәтіндер болған. Бұл бөлікті құрайтын мәтіндер көлемі – 97,5 млн. сөзқолданыстан тұрады, ал оның жартысы 2000 жылдан кейінгілер. Корпустың диахрондық бөлігінің мәтіндер көлемі шамамен 53 млн. сөзқолданыстан тұрады және ол XVIII ғасырдың (1,1 млн. сөзқоладнысын), XIX ғасырдың (прозалық мәтіндерінің 23,3 млн. сөзқолданысын және поэтикалық корпустың 2,5 млн. сөзқолданысын) және XX ғасырдың 1-жартысының (25,4 млн. сөзқолданысын) мәтіндерін қамтиды.

ОТҰК-да жинақталған мәтіндердің негізгі массиві 200 жылдық кезеңді қамтып жатыр, сондықтан ол тілдегі қысқа (бірнеше онжылдықтар) және орта (1-2 жүзжылдық) мерзімдегі өзгерістерді зерттеуге барынша лайықталған.

Корпус көлемі тілдегі барынша жиі кездесетін вариациялық және өзгермелі құбылыстарды зерттеуге пайдасы мол және сонымен бірге мына бағыттар бойынша сенімді нәтижелер алуға мүмкіндік береді:

1. есімдердің, етістіктердің және т.б. морфологиялық варианттарын және олардың эволюциясын зерттеуге;

2. сөзжасамдық варианттарды және осыларға байланысты паронимдердің, сөзжасамдық модельдердің өнімділігі мен сөзжасамдық құралдарын зерттеуге;

3. меңгеру, қиысу, қабысу варианттарының өзгеріске ұшырауын зерттеуге;

4. орыс тілінің акцентологиялық варианттары мен олардың акценттік жүйедегі өзгерістерін зерттеуге;

5. лексикалық вариациялықты, мысалы, синонимдік қатарлардың және тақырыптық топтардың құрамдарын, сонымен бірге олардағы семантикалық арақатынастарды зерттеуге [9].

ОТҰК мәтіндері мезгілдік кезеңдер бойынша келесі кестедегідей үлестірілген (4-кесте).

4-кесте

ОТҰК-дағы мәтіндердің кезеңдік үлестірімі

Кезең	Млн. сөзқолданыстар	Кезең	Млн. сөзқолданыстар	Кезең	Млн. сөзқолданыстар
18 ғасыр	1,1	1911–1920	3,9	1961–1970	5,8
1800–1830	2,1	1921–1930	8,6	1971–1980	7,0
1831–1870	11,3	1931–1940	4,9	1981–1990	7,2
1871–1900	10,3	1941–1950	3,3	1991–2000	23,2
1901–1910	3,4	1951–1960	3,5	2001–2007	46,8

Орыс тілінің Ұлттық корпусы қазіргі кезде келесі шағын корпустардан (субкорпустардан) тұрады:

– әрбір сөйлемге толық сипаттағы морфологиялық және синтаксистік құрылымдары (тәуелділік сызбасы) (деревозависимость) салынған терең аннотацияланған корпус;

– орыс немесе ағылшын тіліндегі белгілі бір сөздің не сөз тіркесінің барлық аудармаларын іздеп табуға мүмкіндік беретін мәтіндердің орысша-ағылшынша параллель корпусы;

– Ресейдің әртүрлі аймақтарының диалектілік сөйлеу тілінің жазбасын жадында сақтайтын және олардың грамматикалық ерекшеліктерін ескере отырып, диалектілік морфология бойынша арнайы іздеу үдерісі қарастырылған диалектілік мәтіндер корпусы;

– тек лексикалық және грамматикалық белгілер ғана емес, сонымен бірге өлеңдерге тән белгілер бойынша іздеу (мысалы, сонета, эпиграмма, өлеңдер және т.б. белгілі тіркестерді іздеу) мүмкіндіктері қарастырылған поэтикалық мәтіндер корпусы;

– орыс тілінің мектеп бағдарламасына бағытталған белгіленімі бар, омонимдер қатары ажыратылған корпус – орыс тілінің оқыту корпусы;

– көпшілік және жекелік сөйлеу тілінің, сонымен бірге 1930–2000 жылдардағы кинофильмдер транскриптерінің магнитофондық жазбаларының мағынасы ашылған орыс тілінің сөйлеу тілі корпусы [10].

Орыс тілінің Ұлттық корпусы басқа да тілдердің Ұлттық корпустары сияқты мынадай маңызды екі ерекшелікке ие:

– біріншіден, корпусқа аса көлемді және әр кезеңдер бойынша шамалас көлемдегі мәтіндердің алынуы (орыс тілінің Ұлттық корпусына барлық жазбаша және ауызша мәтіндер (көркем әдебиеттің әртүрлі жанрлары, көсемсөз жанры, оқу, ғылыми, іскағаздары, сөйлеу тілі, аймақтық тіл және т.б.) қамтылып, олар әр кезең бойынша сәйкес көлемде алынады);

– екіншіден, корпус құрамындағы мәтіндер айрықша сипаттағы қосымша ақпаратқа ие болады. Мұндай ақпарат шартты түрдегі белгіленімдер арқылы беріледі.

Орыс тілі Ұлттық корпусының даму барысы XIX ғ. басынан XXI ғ. бастапқы кезеңін қамтиды деуге болады. Корпустық қорға көркем әдебиеттің (проза мен драматургия, поэзия) мәдени маңызы зор және тілдік тұрғыда тілші-ғалымдардың қызығушылығын тудыратын түпнұсқа шығармалары енгізілген. Бірақ Ұлттық корпус тек көркем әдебиетке қатысты мәтіндерден ғана тұрмайды, ол сонымен бірге мәтін үлгілерінің басқа да жазба нұсқаларын (мемуарлар, эсселер, көсемсөз стильдері, ғылыми-көпшілікке арналған және ғылыми әдебиеттер, жұрт алдында сөйлеген сөздер, жеке адамдар арасындағы хат алысу, күнделіктер, құжаттар және т.б.) қамтиды.

«Орыс тілінің Ұлттық корпусын» құрастыруға Ресей Ғылым академиясының В.В. Виноградов атындағы Орыс тілі институты ғалымдары ғана емес, оған Ресейдегі аса ірі ғылыми топтардың, атап айтсақ, Мәскеу, Санкт-Петербург, Казань, Воронеж, Саратов және басқа да ғылыми орталықтардың көптеген ғалымдары қатысқан. 2003–2010 жылдары «Орыс тілінің Ұлттық корпусын» құрастыруға: 1) Ресей ғылым академиясының тарих-филология («Филология және ақпараттану») бөлімі; 2) Ресей қоғамдық ғылым қоры; 3) Білім берудің федералды агенттігі сияқты ғылыми мекемелер қолғабыс еткен.

Сондай-ақ, В.В.Виноградов атындағы Орыс тілі институтының мамандарымен бірге жобаға:

1) РҒА-ның Тіл білімі институты [ИЯз РАН], РҒА-ның Ақпарат тарату мәселелері институты;

2) РҒА-ның Бүкілресейлік ғылыми және техникалық ақпарат институты [ВИНИТИ РАН];

3) Санкт-Петербургтегі РҒА-ның лингвистикалық зерттеулер институты;

4) Казань (Приволжский) федералды университеті;

5) Воронеж мемлекеттік университеті;

6) Саратов мемлекеттік университеті сияқты ғылыми мекемелер де атсалысқан.

Аталған жобаға осындай бірнеше ғылыми мекеме мен топтардың қатысуына орыс тілінің жазба және сөйлеу тілі мәтіндерінің негізгі корпусын құрастыруда төмендегідей мәселелерді қарастыру керек болған:

1) XVIII ғасырдағы жазба мәтіндерінің тұлғалы (көлемді) корпусын құрастыру;

2) XIX–XX ғасырдың бірінші жартысы аралығындағы жазба мәтіндерінің көлемді корпусын құрастыру;

3) қазіргі кезеңнің (XX ғ. ортасы – XXI ғ. басы) жазба мәтіндерінің көлемді корпусын құрастыру;

4) жазба мәтіндерінің корпустарын теңгеру (баланстау) үшін морфологиялық және сөзтудырушы-семантикалық белгіленім қағидаттарына (принциптеріне) зерттеу жүргізу және компьютерлік бағдарламалар мен корпустарға белгіленім жасауды қамтамасыз ету.

Осы аталған мәселелер жоғарыдағыдай ғылыми мекемелердің бірнеше мамандарының бірлескен жұмысының нәтижесінде шешімін тапқан. Мысалы, мәтіндерге морфологиялық белгіленім жүргізудің ортақ принциптерін айқындау мәселесіне ғана 5 ірі ғалым атсалысқан (В. А. Плунгян, Г. И. Кустова, А. Е. Полякова және Д. В. Сичинава).

Сол сияқты орыс тілі корпусына қажетті морфологиялық белгіленімді автоматтандыруды компьютерлік бағдарламамен қамтамасыз ету мәселесімен Mystem (Яндекс бірлестігі) және Dialing атты программалық қорын құрастырушы 20-ға жуық программист-ғалым айналысқан. Ал осы компьютерлік бағдарламаларға қатысты морфологиялық талдаудың теориялық қағидаттарын зерттеуді Л.Л.Иомдин, В.З.Санников (Mystem), Н.Н.Леонтье-

ва (Dialing) сияқты белгілі филолог-ғалымдар өз міндеттеріне алған. Сонымен, орыс тілінің Ұлттық корпусын құрастыру мен оны жетілдіру ісін қажетті компьютерлік бағдарламалармен қамтамасыз ету мәселесіне, яғни іздестіру жүйесін, метамәтіндік белгіленімді, морфологиялық, синтаксистік, семантикалық белгіленімді және т.б. жетілдіру ісінің әр кезеңінде және алынған нәтижелерді сараптамадан өткізу мәселелеріне көптеген ірі ғалымдармен бірге Мәскеудің жоғары оқу орындарындағы филолог-студенттер мен аспиранттар, магистранттар қатысып, өз үлестерін қосып отырған.

Орыс тілінің Ұлттық корпусын құрастыруда белгіленім түрлерімен айналысатын орындаушылар тобының саны әртүрлі. Мысалы, корпусқа сөзжасамдық-семантикалық белгіленім енгізу мәселесін зерттейтін ғылыми топ 9 орындаушыдан тұрса, семантикалық белгіленімнің компьютерлік бағдарламалық құрамдау тобы 2 ғалымнан (А.Е. Поляков, А.И. Зобнин) тұрады екен. Келесі ғылыми топ метамәтіндік белгеленім мен мәтіндерді таңдаудың жалпы қағидаттарын зерделеумен шұғылданған ғалымдар саны 7-еу. Ал әр ғылыми топтың орыс тілінің Ұлттық корпусы мәтіндеріне жүргізіп жатқан метамәтіндік белгіленімдерді үйлестіретін ғалымдар тобы 20 шақты орындаушыдан тұрған.

Сонымен, аталған міндеттерді орындайтын ғалымдар тобының саны мен әр топтағы ғалымдар саны да өне бойы өсіп отыратынын байқауға болады. Сондай-ақ орыс тілінің Ұлттық корпустарын құрастыру ісіне көптеген ғылыми-лингвистикалық, техникалық орталықтар, баспасөз, баспа, жоғары оқу орындары т.б. атсалысып, бірігіп атқарып отырған.

Орыс тілінің ауызша корпустары. Ауызша корпус – басқа ұлттық корпустарда кездесе бермейтін корпустың жаңа типі болып табылады. Қазіргі кезде орыс тілінің мультимедиялық корпусын құрастыруға байланысты белсенді жұмыстар атқарылып жатыр. Мультимедиялық корпус, бұл – сөйлеу жағдаятына байланысты болып, дыбысталу арқылы жүзеге асатын ауызша тілді зерттеуге арналған корпус. Мұндай корпус коммуникация процесіне қажетті мәтіндік құрамнан басқа, видео- және аудио жазбаларды да қамтиды. Сөйлеу тілі мәтіндерінің мағыналары-

ның ашылуы арқылы тек тілдік бірліктер ғана зерттеліп қоймай, сонымен бірге тілдік қатынастың әртүрлі жағдаятындағы тілдік және тілдік емес әрекеттері де (мимикалары, ымдары, кейіп-қалыптары) зерттеледі [9].

Табиғи диалогтің вербальді және вербальді емес құралдарының өзара арақатынасын зерттеу тұрғысынан алғанда мультимедиялық корпустардың келешегі зор деуге болады. Себебі ауызша тіл, дәлірек айтсақ, көптеген ғалымдардың пайымдауынша, көпшілік алдында емес, дайындықсыз айтылған ауызша тіл сөйлеудің ең маңызды түрі болып табылады. Мұндай тіл түрі тілдің өзегіне жақын орналасады және тіл үлгілерінің ішінен оның табиғи болмысын барынша шынайы бейнелейді [11]. Енді орыстың ауызша тіл корпусын пайдалану мүмкіндіктеріне тоқталайық.

Мәселен, мультимедиялық корпусты қолдану арқылы жүргізілген бір зерттеудің мақсаты – ымдарды иллюстрациялаушылардың қандай белгілері дискурс сегменттерінің шекарасы барын көрсететіндігін анықтау [12]. Зерттеудің мақсатына сәйкес орыс тілінің *Ауызша әңгімелер корпусы* құрастырылады. Оны құрастыруға негіз болған «Фильм о грушах» («Pear film») деп аталатын 6 минуттық видеосюжет болады. Осы фильм бойынша МГУ студенттерінің 8 әңгіме бойынша жасаған фильм үзіндісі орта есеппен 20 минутқа созылады. Корпуста барлығы 595 қарапайым дискурстық бірлік сақталып, олар әдеттегі жай сөйлемдермен сәйкес келеді және Г.Е. Крейдлиннің тәсіліне сәйкес, ол бірліктер ақпарат иесін бейнелейтін 327 иллюстративтік ымдар арқылы көрініс тапқан [13, 10б.]. Зерттеушілер ауызша тілдегі әңгімелер корпусының материалдары арқылы ымдардың дискурсты құрастыруға қатысын, сөйлеушінің күйі мен коммуникация үдерісінің жеке белгілері мен сөйлеуші қолдарының әртүрлі қимыл қозғалыстарының қосымша ақпарат беру мүмкіндіктері бар екендігін көрсетеді. Мысалы, ымдар аралығындағы қолдың тыныштық қалпынан өзгеріске ұшырауы жүйелі түрде нарратив (сипаттама) сегменттер аралығының шекарасын көрсететіні айқындалады. Мультимедиялық корпус арқылы ауызша нарратив құрылымы мен иллюстративтік ымдар арасындағы байланыстарды зерттеу мүмкіндіктерін осы мысал бойынша көруге болады [12].

Санкт-Петербургтың Мемлекеттік университетінің филология және өнер факультетінің филологиялық зерттеу Институты құрастырған «дыбыстық» корпус «*Один речевой день* (ОРД)» деп аталады. Бұл орыс тілінің күнделікті қатынастағы «дыбыстық» корпусының мақсаты – табиғи жағдайдағы тіл иесінің нақты сөйлеу тілін зерттеу болған. Аталған корпустың көптеген басқа сөйлеу тілі корпустарынан айырмашылығы да осы мақсатында деуге болады [14]. Оның ең алғашқы дыбысжазу сериясы 2007 жылдың күзінде жүзеге асқан. Ол үшін әлеуметтік тобы және жас жағына қарай Санкт-Петербург тұрғындарынан 30 адамнан құралған ақпарат беруші коммуниканттар таңдалып алынған. Информанттар өздерінің сөйлеу қатынастарын таспаға жазып алудың әдістемесі жайында бір тәулік бойы толық нұсқамалық алып, әлеуметтік анкеталарды толтырған және психологиялық тестілеуден өткен. Корпуста ақпарат берушілердің сөздерінен басқа олардың әртүрлі жастағы және әр мамандық иелеріндегі (туыстары, жолдастары, әріптестері, таныстары, жат адамдар) 520 коммуниканттың сөздерінің де жазбалары орын алған. ОРД жазбалары жалпы 233 сағаттан тұратын 266 файл түрінде ұсынылады. Жазбадан 2202 эпизод бөлектеліп алынып, оларға тиісті талдаулар жүргізіліп, 134-нің мағынасы ашылған (бұл 2009 жылдың қазан айындағы деректер бойынша 4500-ден асатын қақпа сөз (реплика), 39 мың сөзқолданыс).

Бұл дыбыстық корпус жазылып алынған материалдағы лингвистикалық динамиканы зерттеуге мүмкіндік жасайды.

1.4 Корпуста қамтылатын мәтіндер және корпустардың репрезентативтігі

«Корпус» термині әдетте белгіленім жүргізілген көлемді мәтіндер жиыны дегенді білдіреді. Корпуста қамтылатын мәтіндер көбінесе жазба тілін негізге алады, яғни проза, поэзия т.б. стильдер қарастырылады. Алайда ұлттық корпустар белгілі бір тілдің болмысын түгел танытуы қажет. Соған сәйкес корпустар стильдік, кезеңдік т.б. ерекшеліктеріне қарай үнемі толықтырылып отырады және корпус құрастырудың бағдарламасы да осы жаңғыртуға, кеңейтуге мүмкіндік беретіндей етіп жасалады.

Корпусқа мәтіндер белгілі бір жүйе бойынша енгізіледі. Корпус мәтіндері жасалып отырған корпустың мақсатына сай белгілі бір стильдер мен жанрлар немесе кезеңдер бойынша теңгерімді мөлшерге сай болуы қажет. Яғни мәтінде жасалып отырған корпус түріне қарай сәйкестендіріледі. Нақтырақ айтсақ, мәселен, корпустар өз ішінде субкорпустарға бөлінеді: *поэзия тілінің корпусы, проза тілінің корпустары, газет мәтіндерінің корпусы, ғылыми-техникалық мәтіндер корпусы, ұлттық корпус.* Сонымен қатар лингвистикалық белгіленімдер енгізілу сипатына қарай да *морфологиялық белгіленімдер енгізілген корпустар, синтаксистік белгіленімдер енгізілген корпустар* т.б. Ал ұлттық корпус дегеніміз – бір тілдің қолданылу мүмкіндігін бейнелейтіндей дәрежеде жасалған, электронды сипаттағы, түрлі компьютерлік бағдарламалармен қамтамасыз етілген үлкен корпус болып табылады, яғни ұлттық корпустарда бір тілдік қолданыс аясындағы барлық жанрлар мен стильдер түгел қамтылуы қажет. Мұны корпус жасаудың бірден бір шарты – ауқымдылық немесе тұлғалылық (көлемділік) (представительность) деп атайды. Бұл әсіресе ұлттық корпустарға қойылатын басты шарт болып табылады.

Корпустардың қай түрінде болмасын мәтіндер теңгерімділігі сақталады, яғни енгізілетін мәтіндер бейберекет алынбайды. Белгілі бір стиль, жанрдағы т.б. сипаттары бойынша мәтіндер көлем жағынан мүмкіндігінше біркелкі алынады. Әртүрлі жанрдағы мәтіндердің бірінің көлемі аз, бірінің көлемі үлкен болуы белгілі сандық, сапалық мәліметтер алуға кедергі келтіреді. Сондықтан әртүрлі сипаттағы мәтіндер теңгерімділігі (арасалмағы) (сбалансированность) сақталу қажет. Демек, корпус жасауда енгізілетін мәтіндердің көлемі айқын болуы шарт.

Кейбір тілдерде жазба мәтіндерді ғана емес, ауызша формадағы мәтіндерді де корпусқа енгізген. Корпусқа енгізілген жазба мәтіндер мен ауызша мәтіндер сөйлеудің осы екі формасының арасындағы айырым белгілерді, өздеріне тән ерекшеліктерін айқындауға мүмкіндік береді. Сондай-ақ жазба мәтіндер дәстүрлі ұстанымдарға, тілдік нормаға сәйкес жазылса, ауызша мәтін тілдің динамикалық сипатын танытады деуге болады. Яғни ауызша

мәтіндер – табиғи тірі тілдің бірден бір көрінісі. Осы орайда корпустарға ауызша формадағы мәтіндер де енгізілуде.

Ауызша формадағы мәтіндерге радиодан берілетін мәтіндер, интервьюлер, дөңгелек үстел материалдары, қоғам қайраткерлерімен әңгімелесу сияқты ресми мәтіндерді, сонымен қатар тұрмыстық деңгейдегі әңгімелесу кезіндегі (үйде, көшеде қолданылатын) мәтіндерді жатқызуға болады. Сондай-ақ жергілікті ерекшеліктерді танытатын аймақтық сипаттағы ауызша мәтіндерді, тіпті әлеуметтік топтар тіліндегі ерекшеліктерді айқындау үшін жастар тілі, кәсіби маман тілі, студент тілі, оқушы тілі, ер адамдар тілі, әйел адамдар тілі сияқты ерекшеліктерді танытатын ауызша формадағы таспаға жазылып алынған мәтіндерді де енгізуге болады.

Корпусқа енгізуге болатын келесі мәтіндер тобы – электронды пішіндегі жазбалар. Олар: хаттар, блогтар т.б. Мұндай электронды сөйлесулер жазбаша формада жазылғанмен, ауызша сөйлеу элементтерін молынан қамтиды. Ауызша және жазбаша форманың аралық ерекшеліктерін танытады. Алайда электронды мәтіндерді аннотациялау өте қиын. Себебі электронды пошталарды көбінесе жастар қолданады. Жастардың тілінде біріншіден, көбінесе, орфографиялық нормалар сақтала бермейді, олар саналы түрде сөзді ойнату мақсатында да қолданылады. Екіншіден, жастар тілінде жаргондар, кірме сөздер – орыс, шетелдік сөздер көп кездеседі. Мысалы: қазіргі жастар тілінде *«шашу, жырту, косякқа кіргізу* т.б. сияқты сөздер жиі қолданылады және олар электронды жазбаларда да көрініс тапқан. Бұған қарамастан, мұндай тілдік ерекшеліктерді айқындау лингвистер үшін қажет. Сондықтан, электронды мәтіндердің қиындығына қарамай, оларға мүмкіндігінше лингвистикалық белгіленімдер қою керек. Өйткені олар қазіргі тілдік сленгтерді зерттеу үшін қажет саналады.

Корпустар құрастыруда оның түріне қарай (ұлттық, стильдік, кезеңдік) мәтіндер таңдалып алынады. Оларды **таңдама мәтіндер** деп атайды. Корпусқа енгізілетін мәтіндер көбінесе ақын-жазушылар шығармаларынан алынады. Поэзия мәтіндерін алуға корпус құрастырушылар шек қойып отырады, себебі мұндай мәтіндерде

сөздердің орын ауысуы, мағынасы тұрақсыз болып келеді. Ал драматургия жанры көбінесе диалогқа құрылады. Тыныс белгілері, адам аттары жиі кездеседі т.б. Корпустар құрастыру тәжірибесінде әсіресе, проза жанры басым. Сондықтан стильдік жағынан алғанда корпустарды «проза жанрына орталықтандырылған» (литературацентричный) деуге болады.

Уақыт өте келе, корпустың көлемі мен құрылымы өзгеруі мүмкін. Корпустың тұлғалылығы, яғни оның жеке бөліктері көлемінің арақатынастары (әртүрлі сипаттамалары бойынша) ***репрезентативтілік*** немесе ***теңгерімділік*** (***сбалансировнность***) деп аталады. Алғашқы корпустардың көлемі 1 млн. сөзқолданыстан тұрғандығы белгілі (Браун корпусы, Ланкастер-Осло-Берген корпусы, Орыс тілінің Упсаль корпусы). Корпустың мұндай көлемі тілді жан-жақты сипаттай алмады. Қазіргі кезде жалпы тілдік (ұлттық) корпустардағы мәтін көлемі 100 млн. сөзқолданыстан кем болмауы керек деп саналады. Ұлттық корпусты сол тілдің белгілі кезеңдегі әр алуан жанрларының, стильдерінің, аймақтық, әлеуметтік және т.б. қолданыстарының (немесе кезеңдерін) көрінісі деуге болады (мысалы, ОТҰК, бұған толық ену мына адреспен: http://ruscorpora.ru, BNC, шектеулі ену мына адреспен: http://www.natcorp.ox.ac.uk/ немесе http://sara.natcorp.ox.ac.uk). Қазіргі кездегі барлық тілдік зерттеулер және сөздіктер мен грамматикаларды құрастыру жұмыстары сол тілдің көлемді (репрезентативті) мәтіндер корпустары негізінде жасалады.

Корпус авторларының міндеті – корпус мақсатына сәйкес салалардан барынша көп мәтіндер жинау. Корпусты сол тілдің кішірейтілген моделі деуге болады. ***Репрезентативтік*** ұғымы – барынша жеткілікті және әртүрлі кезеңдердегі мәтіндердің жанрларға, стильдерге, авторларға және т.б. қатысты пропорционал шамамен алынған көлемі, яғни тілдік зерттеу аясын жан-жақты ашатын мәтін көлемі [4]. Репрезентативтікке әдетте әртүрлі анықтама 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 теңгерімділік (репрезентативтілік) принципін

сақтауға ұмтылу қажет. Репрезентативтіктің бұл сипаты сандық аспект ретінде қарастырылады да, ал оны зерттеп анықтау математикалық әдіспен жүзеге асады. Сонымен бірге репрезентативтіктің филологиялық әдіспен зерттелетін сапалық жағы да орын алады. Ал ондай сапалықты корпустағы мәтіндер жиынтығының жанрлар мен функционалдық стильдермен теңдестік сипатта болуы деп ұғыну қажет.

А.Н.Барановтың тұжырымдауынша, репрезентативтікті жүзеге асырудың ең қарапайым жолы – пәндік салалар бойынша алынатын таңдама мәтіндерді сәйкес деңгейде шағындау әдісі. Міне, тек осы жағдайда ғана корпусты ұйымдастыруда сәйкестік стратегиясы сақталатын болады [6].

Сонымен, репрезентативтілік (жан-жақтылық) – корпустың пәндік аяның барлық қасиетін бейнелей алу мүмкіндігі деуге болады яғни лингвистикалық зерттеу типіне қатысты пәндік аядағы құбылыстың кездесу жиілігі тілдік бірліктерді бір-бірінен ажырата алатындай мәнде болуы қажет [6, 118-б.].

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

1. Бірінші типтегі корпустар әмбебаптық сипатта болады, олар тілдік әрекеттің барлық саласын қамтиды.

2. Екінші типтегі корпустар арнайы мақсатпен жасалған, мысалы, мақалдар корпусы немесе газет тіліндегі саяси метафоралар корпусы [3].

Екі жағдайда да репрезентативтік критерийі сақталады. *Бірінші түрге* жататын мәтіндер корпустары 60-жылдары әмбебап сипатта жасалды. Мысалы, АҚШ-та сол кезеңге сай қанағаттанарлық репрезентативтік дәрежеде Браундық мәтіндер корпусы құрастырылған болатын. Оған 15 жанрды бейнелейтін 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 (хобби);

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 нысанға алынған құбылыстың қолданыста барын барынша объективті түрде бейнелеу мақсаты қойылады. Мәселен, ағылшын тілін иеленушінің сөйлеу тәжірибесіндегі белгілі бір кезеңдегі және географиялық аймақтағы максимал репрезентативтікке ие ағылшын тіліндегі мақалдардың корпусы, мысалы, ағылшын тіліндегі саяси метафораларды зерттеу үшін репрезентативтікке ие емес [8].

XXI ғасырдың басында газеттердегі тақырыптар корпусы, машиналық аударма жүйесін дұрыстауға арналған ағылшын тіліндегі мәтіндер корпусы, саяси метафоралар корпустары арнайы мақсатқа құрылған корпустар қатарына жатады [8]. Әрине, мұнда корпусқа мәтіндерді іріктеу критерийін корпусты құрастырушы өзінің практикалық қажеттілігі немесе ғылыми зерттеу мақсатына қарай өзі белгілейді. Демек, корпус құрастыруда мәтінді іріктеу корпустың түріне байланысты болуы қажет.

Бірінші типтегі корпусты құрастырудың әдістемесі дедукция принципіне (жалпыдан – тіл иелерінің объективті өмірде бар сөйлеу тәжірибесінен оны бейнелейтін дербес мәтіндер корпусына) негізделеді

Екінші типтегі корпустарды құрастыру әдістемесі бойынша мәтіндер корпусындағы дербес, бірлі-жарым лингвистикалық феномендерді аса дәлдікпен бейнелеуді көздейді [15].

Аталған екі типтегі корпустар көпшілік жағдайда біріктірілген (құрамдастырылған) түрде қолданыс табады.

1.5. Корпустық лингвистика мен дәстүрлі тіл білімінің зерттеу нысандарының айырмашылықтары

Корпустық лингвистика мен дәстүрлі лингвистика арасындағы айырмашылық тұстарын мына кестеден байқауға болады:

5-кесте

Корпустық лингвистика мен дәстүрлі лингвистиканың арақатынасы

Корпустық лингвистика	**Дәстүрлі лингвистика**
1	2
Негізгі назар – сөйлеу тілі (речь)	Негізгі назар – тіл (язык)
Мақсат – арнайы таңдалып алынған мәтіндер корпусы бойынша сөйлеуде көрініс тапқан тілді сипаттау, яғни тілді қолданымдық тұрғыдан сипаттау	Мақсат – тілді сипаттау және тілге түсініктеме беру
Өз зерттеулерінде мәтіндер корпусының мәліметтеріне сүйенеді	Өз зерттеулерінде теориядан оған түсініктеме беру мен тіл деректері бойынша оны дәлелдеу бағытында жүреді
Квантитатиті (сандық) әдіске артықшылық беру	Квалитативті (сапалы) әдіске артықшылық беру
Өзін эмпирикалық әдістерге негізделген дәстүрдің бөлігі ретінде санау	Өзін рационалистік әдістерге негізделген дәстүрдің бөлігі ретінде санау
Мәтінді қайсыбір физикалық мән ретінде қарастыру	Мәтінді қайсыбір абстракция ретінде қарастыру
Нақты тілдердің грамматикасын құрастыру	Тіл универсалияларын зерттейді
Негізгі назарды тұрпатқа бөледі	Негізгі назарды – тұрпатқа ғана емес, мазмұнға да бөледі
Мәтіндерді ғаламдық болашақта қарастырады	Мәтіндерді локальды (жергілікті) болашақта қарастырады

1	2
Мәтінге деген өзінің назарын, ешбір догмамен шектелмеген, барынша кең көзқарас арқылы шоғырландырады	Қайсыбір нақтылы, жасанды түрде шектелген проблемалық саланы анализдейді (зерттейді, талдайды)
Өзінің қорытындысында мәтін түрінде көрініс тапқан сөйлеу әрекетіне қатысты зерттеулеріне сүйенеді	Өз зерттеулеріндегі тілдік материалдарды, эмпирикалық материалдарды таңдауда интуицияға сүйенеді
Тілдік материалдарды алғашқы кезеңде өңдеуде ықтималдықтар теориясының әдістері мен статистикалық әдістерді жиі қолданады	Логикалық пайымдау әдісіне артықшылық береді
Мәтіндік контексте қалай кездессе, сол күйінде лингвистикалық бірліктермен (сөз қолданыстармен) зерттеу жұмысы жүргізіледі	Мәтіндерден дараланған сөз қолданыстарды, жасанды мысалдарды қарастырады
Эмпирикалық ауызша материалдарды өңдеуде индуктивті әдістерді қолдануды таңдайды және оларды ғылыми әдістің негізі деп санайды	Эмпирикалық ауызша материалдарды өңдеуде дедуктивті әдістерді қолдануды таңдайды
Эмпирикалық материалдарды өңдеу арқылы ашылған ғылыми жаңалықтарға сенім артады	Деректерді өңдеуде рәсімдерге, бағалауға, салыстыруларға және т.б., яғни көпғасырлық зерттеулер нәтижелеріне негізделген ғылыми жаңалықтарға сенім артады

1.6. Корпустық лингвистиканың тіл білімі салаларымен байланысы

Корпустық лингвистика тіл білімінің жеке саласы ретінде өзімен іргелес жатқан тіл ғылымы пәндерімен жанасып жатады, яғни математикалық лингвистика, дискурстік анализ және лексикография салаларымен тығыз байланысты. Корпустық лингвистиканың басқа тіл ғылымы пәндерімен қарым-қатынаста болу ерекшелігі, бір жағынан алғанда, мәтіндер корпусының корпустық лингвисти-

ка қызметінің нәтижесі ретінде болса, ал екіншіден, лингвистикалық пәндердің басқа түрлеріне де бастапқы эмпирикалық материал болу мүмкіндігінде. Міне, дәл осы жағдай корпустық лингвистиканың фонетика, лексикология, грамматика және стилистика салаларымен де байланыста болатындығына негіз бола алады.

Корпустық лингвистика және сөздік қор (лексикология). Қазіргі таңда тіл білімінде сөздік құрамды анықтау, жазушы тіліндегі қолданыстардың ерекшеліктерін ашу, лингвоелтаным, лингвомәдениеттаным, этнолингвистика, мәтін лингвистикасы сияқты мәселелерге ерекше назар аударылып келеді. Соған байланысты тілдің ұлттық қабатын (корпус) анықтау мәселесіне де зерттеушілер ерекше ден қоя бастады. Ұлттық корпустарды құрастыруда да тілдің лексикалық қабатындағы бірліктерді анықтау, яғни белгілі бір тілдің өзіндік сөздік қорын, қолданыс ерекшеліктерін, қолдану жиілігін анықтау мақсат етіледі.

Эмпирикалық деректер корпустық лингвистиканың өмірге келуінен көп бұрын қолданыста болған. Мысалы, Сэмьюэл Джонсон өзінің сөздігін әдебиеттердегі әртүрлі мысалдармен дәйектеген болатын, ал XIX ғасырдағы Оксфорд сөздігінде сөздердің қолданылу мүмкіндігін зерттеу үшін оларды мысалдармен дәлелдеу мақсатында көптеген шығармалардан қажетті цитаталар келтірілген. Корпустық тіл білімі саласының өмірге келуі лингвистердің тілді зерттеу әдісіне айтарлықтай өзгеріс әкелді деуге болады. Енді корпусты пайдаланушы тілші-ғалымға миллиондаған сөзден тұратын мәтіндерден сөз бен сөз тіркестері бойынша мысалдар келтіру бірнеше секунд арасында жүзеге асады. Сонымен, корпустың арқасында, тілге қатысты жаңа ақпаратпен қамтамасыз етілген сөздіктердің жаңартылған нұсқасын бұрынғы кездерге қарағанда өте тез арада баспадан шығаруға мүмкіндік туады. Мол көлемді мысаларды зерттеп, жасалған сипаттамалар да толықтық және дәлдік тұрғысынан алғанда жоғары сапалы болатыны сөзсіз.

Корпустан таңдалып алынған мысалдарды мағыналы топтарға жинақтауға да мүмкіндік бар. Мысалы, нақты сөздің қолдану мүмкіндігінің барлық дербес жағдайларын қарастыру мақсатымен сол сөзге қатысты контекстерді әліпби тәртібімен реттеу әрекеті де жеңілдейді. Сонымен бірге, корпустық деректердің аса көп

метамәліметтерге ие болуынан – автор, күнтізбелік күні (жылы, айы, күні), стиль, аймақтық айырмашылықтар, сөз табы және т.б. – жеке сөздер мен сөз тіркестерінің қайсыбір стильге, аймақтық айырмашылықтарға және т.б. тән қолданыстарын топ-топқа бөліп қарастыру ісі де айтарлықтай қиындық тудырмайды. Әрдайым дамып, көлем жағынан өсіп отыратын мониторлық корпустар сөздік құрылысының сапасын арттыруда үлкен рөль атқарады. Себебі, тілге еніп жатқан жаңа сөздерді, немесе жаңа мағынаға ие болып жатқан сөздерді немесе олардың қолданысының стильдер бойынша арақатынастарын қадағалауына лексикографтардың зор мүмкіндігі бар. Лексика саласына қатысты сандық (статистикалық) сипаттама жасау, сөздердің жиілігін анықтау мәселесінде де айтарлықтай рөль атқарады. Корпус арқылы белгілі бір тірек сөздің бірнеше туыстас тілдерде қолдану мүмкіндігінің, статистикалық үлесін есептеп шығарудың тиімді екендігін де айтуға болады. Сол сияқты тек жеке сөз ғана емес, олардың тіркесімдік позицияларын жүйеге түсіру мүмкіндігі де бұрынғы дәстүрлі зерттеулерден айырым табады.

Корпустық лингвистика және грамматика. Корпус мәліметтерін барынша көп қолданатын лексикалық зерттеулерден кейінгі ең бір жиі қолданыстағы сала – грамматикалық зерттеулер. Корпустың мәліметтері грамматикалық зерттеулер үшін де ең бір тиімді құрал болып отыр, себебі корпус грамматикалық теориялардан енген болжамдарды тексеру үшін баға жетпес эмпирикалық мәліметтер береді.

Корпусты қолданатын көптеген грамматикалық ілімдер өз зерттеу аясына сандық мәліметтеріне талдау (анализдеу) жүргізуді енгізіп отыр. Қазіргі кезде грамматикалық бірліктердің статистикасын математикалық тәсілдермен зерттеу қызығушылық тудыруда.

Зерттеушілер мәтіндер корпусын тек грамматикалық теорияны сипаттау немесе индукциялық қорытындылау үшін ғана пайдаланып қоймай, оны рационалистік тұрғыдағы грамматикалық теорияны тексеруге де пайдаланған. Мысалы, Nijmegen университетінде алғашқы рационалистік тұрғыдағы грамматиканы компьютерлік корпуста нақты қолданыста бар тіл бойынша тексерді, мә-

селен, формалды грамматиканы компьютерік талдағышқа (анализаторға) енгізілгеннен кейін олардың корпустағы мәліметтермен қаншалықты сәйкес келуін тексеру мақсатында корпустағы барлық мәтіндер бойымен «жүгіртіп» сәйкестігін тексеріп шығады. Осындай әрекеттің нәтижесінде формалды грамматиканың дұрыс-бұрыс тұстары айқындалып, олар қайтадан өңделіп, өзгеріске ұшырайды.

1.7. Корпус мәтіндерінің электрондық жинақтағы мәтіндерден айырмашылығы

Корпус дегеніміз – әр тілдегі электронды пішінге келтірілген, яғни бір басқару орталығынан автоматты түрде жұмыс істейтін мәтіндер жинағы. Қазіргі кезде әртүрлі стильдегі мәтіндердің электронды нұсқалары бар. Кітаптар, газет-журналдар т.б. көп-ақ. Бірақ мұндай мәтіндердің электронды нұсқаларынан корпусқа жинақталған мәтіндердің айырмашылығы бар. Мәселен, жай бір кез келген мәтіннің (шығарманың) электронды нұсқасы арқылы іздеген сөзімізді «іздеп табу» нұсқаулығы арқылы тауып алуымызға әбден болады. Сонымен қатар оны қол жұмысы арқылы біртіндеп немесе арнайы бағдарлама енгізу арқылы автоматты түрде бір демде-ақ шығара аламыз. Алайда корпусқа арнайы енгізілген мәтіндердің автоматты мүмкіндігі мен жай мәтіндер жинағының айырмашылығы бар.

Біріншіден, корпусқа енгізілген мәтіндер арнайы бағдарламалар арқылы жұмыс істейді. Арнайы енгізілетін бағдарламалар арқылы жеке сөздерді ғана емес, кез келген сөз формасын, сондай-ақ ол сөзформаның әртүрлі позициядағы қолданысын тауып алуға және жиілігін айқындауға болады. Ал мұндай жұмысты аннотацияланбаған (белгіленбеген) жай ғана мәтіндер жинағы атқара алмайды.

Екіншіден, корпусқа енгізілген мәтіндердегі әрбір тілдік бірлікке, ол не жеке сөз не қосымша болсын, лингвистикалық ақпарат беріледі. Тілде мұндай лингвистикалық ақпарат беруді аннотациялау деп атайды. Аннотация дегеніміздің өзі қысқаша мазмұн дегенді білдіреді. Соған сәйкес корпусқа енгізілген тілдік бірлік-

тердің аннотациясы да шартты белгілер арқылы қысқаша беріледі. Қазақ тілінде мұндай шартты белгілер «белгі-код», «белгіленім» деген терминдермен қолданылып жүр. Орыс тілінде «разметка» деп аталады. Мәтіндер бірліктеріне берілген лингвистикалық ақпараттардың мол болуы осы белгіленімдердің әртүрлілігіне байланысты. Белгіленімдер тілдегі жекелеген деңгейлерге тән тілдік мәліметтерді қамтиды. Айталық, мәтіндер корпусы туралы жазылған ғылыми еңбектерде фонетикалық, морфологиялық, семантикалық, синтаксистік белгіленімдердің енгізілетіндігі туралы айтылады. Бірақ корпус жасау барысында аталған белгіленімдердің барлығын бір уақытта енгізу қиындық тудырады. Осы орайда шетел, орыс тіл біліміндегі мәтіндер корпусында лингвистикалық белгіленімдер енгізу ісі біртіндеп жүзеге асырылған. Сонымен қатар арнайы компьютерлік бағдарламада корпустағы лингвистикалық мәліметтерді жаңғырту немесе жаңа мәліметтермен толықтыру мүмкіндігі ескеріліп отыратын болған. Демек, мәтіндер корпусы лингвистикалық ақпараттармен қамтамасыз етілуі жағынан бір ғана белгіленім енгізілген болуы да, немесе бірнеше белгіленім енгізіліп, кеңірек аннотацияланған болуы да мүмкін. Толық лингвистикалық ақпарат берілген корпустарды «терең аннотацияланған» (глубоко аннотированный) деп атайды.

Екіншіден, корпус мазмұнының күрделілігі терең аннотацияланумен қатар әртүрлі стильді қамтуына да байланысты. Әдетте, корпус құрастыруда, сондай-ақ жиілік сөздіктер жасауда да көбінесе төрт түрлі стиль қамтылады. Олар: көркем стиль (проза, поэзия), драматургия, газет-журнал (публицистикалық), ғылыми-техникалық стильдер. Бұлардан басқа ауызекі стильден де корпус мәтіндері жинақталады.

Үшіншіден, корпустардың сапасы ондағы қамтылған сөзқолданыс мөлшерімен де өлшенеді. Жалпы тіл білімінде алғашқы корпустардың кемінде 1 млн. сөзқолданыстан бастап жасалғандығы айтылады. Қазіргі кездері сол алғаш 1 млн. сөзқолданыстан жасалған мәтіндер корпусының көлемі 20 миллионнан 100 млн.-ға дейін жетеді екен. Қазіргі орыс тілінің корпусында 350-500 миллион сөзқолданыстан тұратындығы айтылады. Демек, корпустың құрастырылуы туралы мәселе сөз болғанда, ең алдымен

оның көлемі туралы нақты деректер беріледі. Осы орайда қайсібір тілде болмасын корпустар жасау ісінде әртүрлі лингвистикалық ақпараттармен қамтамасыз етілген мәтіндерді компьютер жадына мүмкіндігінше молынан енгізу қажет. Мұндай әрбір тілдегі барлық стильді қамтыған көлемді корпустарды ұлттық корпус деп атайды. Ұлттық корпустар әр тілдің тілдік заңдылықтарын, оның ішінде тұрақты лексика-грамматикалық жүйесін толық қамтыған (терең аннотацияланған) миллиондаған сөзқолданыстан тұратын электронды пішіндегі көлемді мәтіндер жинағы болып табылады. Мұндай ұлттық корпустар барлық стильді қамтуы арқылы тұтас бір тілдің табиғатын толық бейнелесе, әрбір стильге тән тілдік ақпараттар арқылы жекелеген стильдердің өзіндік ерекшеліктерін айқындауға болады. Мысалы, бейнелі сөздер көркем стильде көп қолданылса, ғылыми-техникалық стильде сирек қолданылады. Әрбір стильге тән мұндай ерекшеліктер бұрыннан белгілі болғанымен, корпус материалдары тілдің лексикалық қабаты туралы ғана емес, грамматикалық жүйесі туралы да мол ақпарат береді. Корпус тілдегі заңдылықтарды нақтылай түседі немесе қайта жаңғыртады, тіпті мүлде жаңа заңдылықтарды айқындауға мүмкіндік береді.

1.8. Корпустарды әртүрлі негізде классификациялау

Корпустардың бірнеше түрін кездестіруге болады, мысалы, бір ғана автордың корпусы, бір ғана кітаптың корпусы, бір ғана стильдің корпусы және т.б. Корпустың көптүрлігіне қарамай, оларды негізгі екі топқа бөліп қарастыруға болады:

1) барлық тілге қатысты (тілдердің белгілі кезеңдеріне қатысты), тілдердің ішкі түрлеріне қатысты (жанр, стиль, тіл иелерінің жасына немесе әлеуметтік тобына қарай (жазушы тілі немесе ғалымның еңбектері және т.б.) корпустарды жіктеу;

2) корпустарды тілдік белгіленім (разметка) типіне қарай бөлу. Белгіленім типтерінің көптігіне қарамай, корпустардың жүзеге асқан нақты түрлері морфологиялық немесе синтаксистік белгіленім жүргізілген типтеріне қатысты (синтаксистікті ағылшынша әдебиеттерде treebanks деп атайды, оның қазақша аудармасы

«синтаксистік құрылымды банк»). Солай бола тұра, синтаксистік белгіленім жасалған жерде морфологиялық белгіленімі де міндетті түрде қамтылады.

Корпустың алға қойған мақсатын және классификациялау белгілеріне қарай әртүрлі типтерін бөліп көрсетуге болады (6-кесте).

Тілдік деректердің типіне қарай корпустар **жазба тіл**, **сөйлеу тілі** және **аралас** тілдік корпустар деп бөлінеді. Жазба корпустарда сөйлеу тілі мәтіндері алынбаған (Браундық корпус, LOB), ауызша корпустарда сөйлеу тілінен жазылып алынған мәтіндер беріледі, ал аралас корпустар, әдетте, белгілі кезеңдегі тілдің қолданысын бейнелейтін ұлттық корпустарда көрініс табады (НКРЯ, BNC және т.б.).

Параллельдік критериясы бойынша корпустар біртілдік, екітілдік және көптілдік деп бөлінеді. Біртілдік корпустарда диалектілер, варианттар қарама-қарсы қойылады. Мысалы, ағылшын тілінде көптеген тілдік мәселелер жаңа технологиялар өмірге келгенше ғылыми тұрғыдан қызығушылық тудырмаған. Жаңа технологияны қолданудың арқасында көптеген салыстыруды қажет ететін шығармалар тілі контрастивтік талдау әдістерімен зерттелді.

Екітілді және көптілді корпустарда екі немесе бірнеше тілде жазылғандарына қарамай (мысалы, белгілі ғылыми проблемаға арналған, әртүрлі елдерде және бірнеше тілде өткен конференция материалдары корпусы), бір тақырыпқа қатысты мәтіндер біріктіріліп беріледі. Мұндай корпустар терминдермен жұмыс істеу кезінде ұтымды және оны аудармашылар жиі қолданады. Екітілді немесе көптілді корпустардың тағы бір түрі – қайсыбір бастапқы тілде жазылған мәтін-түпнұсқалардың жинағы және осы бастапқы мәтіндердің бір немесе бірнеше тілдерге аударма-мәтіндер жинағы корпусы. Мұндай корпус салыстырмалы-салғастырмалы және аударма теориясы бойынша зерттеулер жүргізуге септігін тигізеді. Адамдарды аударма жасау ісіне үйретуде және машиналық аудармаға құнды материал береді.

Әдебилік критериясы бойынша мынадай: әдеби тіл, диалект тілі, сөйлеу тілі, терминологиялық, аралас корпус түрлеріне бөлінеді. Сөйлеу тілі корпусына мысал ретінде Санкт-Петербургте құрас-

тырылған «Один Речевой День (ОРД)» атты корпусты келтіруге болады [14].

Мақсатына қарай корпустарды құрастыру көпмақсатты және мамандандырылған корпустарға бөлінеді. Көпмақсатты корпустар әдетте әртүрлі жанр мәтіндерін жинақтайды (бұған ұлттық корпустар да жатады), ал мамандандырылған корпустар бір ғана жанрмен немесе жанрлар тобымен ғана шектелуі мүмкін.

Мәтіндер корпустары жанрлар бойынша *әдеби, фольклорлық, драматургиялық, публицистикалық* және т.б. жанрлаға жіктеледі. Публицистикалық корпустарға мысал ретінде XX ғасырдың соңындағы орыс газеттерінің компьютерлік корпусын (http://www.philol.msu.ru/~lex/corpus/) және саяси метафоралар корпусын келтіруге болады [8].

Корпусты пайдаланушылар үшін корпустың *қолжетімділік* сипаты маңызды критерий болып саналады. Еркін қолжетімді корпустар кез келген уақытта корпустың толық көлемдегі барлық мәтіндерінен on-line режимінде қалаған ақпаратты алуға мүмкіндік тудырады. Кейбір жағдайларда еркін қол жеткізу сипаты корпус деректерінің тек жеке бөліктеріне ғана мүмкін болады. Коммерциялық корпустармен жұмыс істегенде оны on-line режимінде немесе компакт-дискідегі көшірмесін пайдалану үшін арнайы рұқсат алуға тура келеді. Көлем жағынан кішігірім корпустар ғана ашық жүйеде тұрады. Жабық корпустар, яғни кез келген пайдаланушыға қолжетімсіз корпустар арнайы мақсаттармен құрастырылады.

Қызметіне қарай зерттеуге арналған және иллюстрациялық мақсатта құрастырылған корпустарды ажыратып айтуға болады. *Зерттеу жұмысына арналған корпустар* тіл қызметінің әртүрлі аспектілерін зерттеу мақсатында құрастырылады. Корпустардың бұл типі тілдегі әлі де шешімін таппаған өзекті мәселелерді шешуге бағытталады. Мұндай мәтіндер корпустары бірнеше ондаған миллионнан жүздеген миллионға дейінгі сөзқолданыстарды қамтиды. *Иллюстрациялық корпустар* ғылыми зерттеуді жүргізгеннен кейін құрастырылады: олардың мақсаты жаңа деректерді айқындау емес, алынған нәтижелерді растау мен оларды негіздеу. Олар, бұрын дәстүрлі әдістермен алынған, тілдік фактілерді (ауызша, мәтіндік), нәтижелердің дұрыстығын дәйектеуге қызмет етеді. Иллюстра-

циялық корпус туралы «Путеводитель по дискурсивным словам русского языка» [3] атты жұмыстан қарап, танысуға болады, мұнда демеуліктерге семантикалық талдау мен оқырманға ұсынылған семантикалық түсіндірмені тексеруге мүмкіндік беретін көлемді мәтіндік материалдар тіркеліп беріледі [17, 2].

«Динамикалық» (мониторлы) критерий бойынша корпустар динамикалық және статикалық деп бөліп қарастырылады. Алғашқы мәтіндер корпустары белгілі уақыт аралығындағы тілдік жүйені сипаттайтын статикалық мәтіндердің көрінісі ретінде құрастырылатын. Статикалық корпустар қайсыбір аздаған уақыт аралығындағы мәтіндерді жинақтайды [17]. Мұндай корпустардың типтік өкілі ретінде авторлық корпустарды –жазушылар шығармалары мәтіндерінің топтамасын атауға болады.

Тіл дамуымен бірге тілдік құбылыстар да өзгермелі болып отыратындықтан, тілдік құбылыстардың өзгерісін сипаттайтын динамикалық корпустар құрастыру да қажет болды. Мысалы, сөз мағынасының өзгеруі, қайсыбір синтаксистік конструкциялардың қолдану жиіліктерінің және т.б. өзгеруі. Проблемалық саланың процессуалдық аспектілерін бейнелеу үшін мәтіндердің динамикалық корпусын құрастыру мен оны пайдалану мақсатында жаңа технология пайдаланылды [8]. Динамикалық корпустарды мониторлық немесе мониторингтік корпустар деп те атайды. Мониторлық корпустардың мақсаты – компьютер жадындағы сан жағынан үнемі өсіп отыратын мәтіндерді «қаттау» («складировать»). Мұндай корпустарда алдын ала белгіленген уақыт аралығында корпустағы мәтіндер жиыны жаңартылып және/немесе толықтырылып отырады. Шексіз (әрдайым өсіп отыратын) сипаттағы мониторлық корпустар лексикографиялық сөздіктер құрылысының сапасын арттыруда маңызды рөл атқарады. Себебі лексикографтарға тілімізге жаңадан енетін сөздерді немесе мағынасы өзгеріске ұшырайтын бұрыннан бар сөздерді және стильдік өзгеріске қарай сөздердің қолданысының теңгерімін үнемі қадағалап тұруға мүмкіндік тудырады. Динамикалық мәтіндер корпустарында, әдетте, ұзақ мерзім аралығындағы жазба дереккөздер сақталады, сондықтан, олар неше түрлі диахрондық зерттеулер жүргізуге арналған [17].

«Белгіленім» критерийі бойынша корпустар белгіленген және белгіленбеген корпустарға ажыратылады. Мұны басқа да терминдермен атайды: индекстелген және индекстелмеген, аннотацияланған және аннотацияланбаған, таггирленген және таггирленбеген. Белгіленген корпустарда сөздерге немесе сөйлемдерге *белгіленімнің сипатына* қарай тиісті белгілер (тэгтер) қойылады: морфологиялық, синтаксистік, семантикалық, просодикалық және т.б.

«Мәтіндер көлемі» критериясы бойынша толықмәтінді және фрагменттімәтіндік корпустар деп екіге ажыратылады. Мәтін ұзындығы 2000 сөзге (сөзқолданысқа) сай келу критериясына қатаң түрде сай келетін мәтіндер тәжірибе жүзінде жоқ деуге де болады. Ондай корпустарды фрагменттімәтіндік деп атайды. Толықмәтіндік корпустарға белгілі қалам иелерінің кейбір шығармалар мәтіндерінің корпустарын, қысқа мәтіндер корпустарын, мысалы, газет тақырыптары корпусын жатқызуға болады.

Жоғарыда айтылған баяндауларды қысқаша мына кесте арқылы беруге болады.

6-кесте

Корпустарды классификациялау

Белгі	**Корпус түрлері**
1	2
Тілдік деректердің түрлері	Жазба Ауызша Аралас
Параллельдік	Біртілді Екітілді Көптілді
Әдебилік	Әдеби Диалект тілі Сөйлеу тілі Терминологиялық Аралас
Мақсат	Көпмақсатты Мамандандырылған

1	2
Жанр	Әдеби Фольклорлық Драматургиялық Публицистикалық
Қол жетімділік	Қол жетімді (Ашық) Коммерциялық Жабық
Тағайындалуы	Зерттеу үшін Иллюстрациялық
Динамикалық	Динамикалық (мониторлы) Статикалық
Белгіленім (Разметка)	Белгіленген Белгіленбеген
Белгіленім сипаты	Морфологиялық Синтаксистік Семантикалық Просодикалық және т.б.
Мәтін көлемі	Толық мәтінді «Фрагменттік мәтіндер»

Алға қойған мақсатқа сай мәтіндер корпусының тағы бір жіктемесі төмендегіше көрініс табады:

1. *Сақтаудың формасына (тұрпатына) қарай:*

– дыбыстық формада (сөйлеу түріндегі (фонетикалық) корпустар;

– жазба түрінде (жазба түрдегі корпустар);

– аралас (дыбыстық және жазба) түрінде

2. *Мәтіндердің тілдік айырмашылығына қарай:*

– біртілді;

– көптілді.

3. *Жанрлық айырым-белгісіне қарай:*

– әдеби;

– диалектілік;

– сөйлеу тілі;

– публицистикалық;

– аралас.

4. *Корпусқа кіру мүмкіндігіне қарай:*

– еркін кіру;

– коммерциялық кіру;

– кіруге рұқсат жоқ.

5. *Мақсатына қарай (по назначению):*

– зерттеуге арналған;

– көрнекілік үшін.

6. *Динамикалығына қарай:*

– динамикалық (мониторлық);

– статикалық.

Статикалық корпустар тілдік жүйенің белгілі қалпын бейнелейді.

Динамикалық (мониторлық) корпустар пәндік аяның процессуалды аспектісіндегі тілдік феномендердің белгілі уақыт ішіндегі қызметін айқындау мәселелерін қарастырады.

7. *Қосымша ақпараттың болуына қарай:*

– аннотацияланған (белгіленген);

– аннотацияланбаған (белгіленбеген).

Ғалым В.В. Рыков мәтіндер корпусының басқаша классификациясын келтіреді:

1. *Ұйымдастыру мен құрылымдану дәрежесіне қарай:*

– электронды мұрағат – бұл электрондық таспалардағы мәтіндер, бірақ олардың машиналық таспалардағы формасы (тұрпаты) стандарттанбаған және бейімделмеген (не унифицирована);

– Электрондық кітапхана – мұндағы мәтіндер біркелкілік және стандарттанған түрде көрініс тапқан;

– мәтіндер корпусы – формасы (тұрпаты) стандарттанған және бейімделген (унифицирован), мәтіндер лингвистикалық ақиқаттықтың бір бөлігін бейнелеп көрсетуді мақсат етеді;

– субкорпус – бұл корпустың қайсыбір автономды бөлігі.

2. *Хронологиялық белгісі бойынша:*

– синхрондық;

– мониторлық (тілдің ағымдағы күйін қадағалау);

– диахрондық.

3. *Индекстелуіне қарай:*

– жай (индекстелмеген);
– аннотацияланған (индекстелген).

Қазіргі кездегі корпустар тек аннотацияланған болып келеді, себебі тек белгіленген корпустардан ғана қажетті ақпаратты іздестіруге, пайдалануға мүмкіндік бар.

4. *Тілдік айырмашылығына қарай:*
– біртілді;
– екітілді;
– көптілді.

5. *Корпусты қолдану мен пайдалану тәсілдеріне қарай:*
– зерттеу үшін;
– көрнекілік үшін;
– параллель тілдер үшін.

Зерттеу корпустары көбінесе тілдік жүйелер қызметінің әртүрлі аспектілерін зерттеуге арналған. Иллюстративтік (көрнекілік) корпустары зерттеу жүргізілгеннен кейін ғана құрылады; бұл корпустардың мақсаты – алынған нәтижелерді растау және негіздеу. Мұндай қарама-қарсы қоюшылық корпустардың ерте даму кезеңіне ғана тән болғаны белгілі, себебі қазіргі кездегі корпустар әрі зерттеу жүргізу үшін, әрі көрнекілік үшін де қолданылуы мүмкін.

6. *Корпустың динамикалығына қарай:*
– динамикалық;
– статикалық.

Мәтіндер корпустары алғашында табиғи тілді өңдеу үшін жазылатын бағдарламалармен және оларды тестілеу кезіндегі өңделмеген (шала) күйінде көрініс табады. Бұл жердегі мәтіндер ұғымына ауызша сөйлеу тілі де және табиғи коммуникациядағы тіл де қамтылады. Барлық бүгінгі күнге белгілі корпустар мынадай төрт нұсқада жүзеге асады:

1) коммуникацияның әртүрлі аясын қамтитын мәтіндер (монолингвальдық корпус) – ұлттық корпус;

2) жоғарыда аталған корпусқа ұқсас құрастырылатын (репрезентативтігі сәйкес және ортақ корпус-менеджерімен) бірнеше ұлттық корпустарды қамтитын – салыстырмалы (немесе контрастивтік) корпус;

3) бір тілдегі мәтіндерді және оның басқа тілдегі (тілдердегі) аударма мәтінін қамтитын – параллель корпус немесе параллель мәтіндердің корпусы;

4) ұлттық корпустың құрамына еніп қана қоймай, өзінше жеке тұрып та корпус құрайтын – сөйлеу тілінің корпусы.

Сөйлеу тілі корпустары (кең мағынада – ауызша) ауызша тілдің тек аудиожазбасын (мысалы, қайсыбір тілдің диалектілер корпусы) немесе, керісінше, оның тек транскрипциялық формасын қамтуы мүмкін (мысалы, Британдық ұлттық корпус массивінің бір бөлігі). Үйлесімді нұсқа ретінде корпуста аудиожазба мен оның стенограммасының (транскрипцияланған немесе орфографиялық түрде) қатар орналасқаны дұрыс деп саналады [4].

Аталған нұсқалардың әрбіреуі өзіне тән ерекшелігі бар мәселелерді ғана шеше алады. Мысалы, салыстырмалы корпус контрастивтік аспектідегі тілдерді үйренуге мүмкіндік береді, параллель корпустар тәріздес, қазіргі машиналық аударма жүйелерінде деректер базасы (аударма мысалдарының базасы) ретінде қолданылады. Ал контрастивтік зерттеу кезінде салыстырмалы корпусты қолдану мақсатқа сай келе бермейді. Себебі екітілді параллель мәтіндерінде аударма жасауда сөзсіз орын алатын тіларалық интерференция (екітілдік жүйенің бір-біріне тигізер әсері) көрініс тауып, салыстыру нәтижесіне айтарлықтай дәрежеде кері әсер етуі мүмкін.

Кез келген корпус **аннотацияланған корпус** (шартты түрдегі грамматикалық белгілерді енгізу) болғандықтан, ол мәтіннің лингвистикалық сипатының көрінісін айқындай алады. Әдетте ондай сипаттама корпус құрамындағы бірліктерге қатысты мүмкін болатын тілдік талдау үлгісін қамтиды. Аннотацияланған корпус, оны пайдаланудың қарапайымдылығы мен көпфункционалдық мүмкіндікке ие болуымен де аса құнды. Көпшілік жағдайда, корпустарды аннотациялау әрекетінің автоматты және «қолмен» жүргізу түрлері қатар жүргізіледі. Ал автоматтанған аннотациялау әрекетін қанағаттандырарлық деп табу үшін қателік шегі үш пайыздан аспау керектігін ескеру қажет саналады [2, 63].

Қазіргі кезде кейбір тілдерге қатысты жүзеге асырылған компьютерлік бағдарламалар қанағаттандырарлықтай дәрежеде

жұмыс істейді. Мәселен, сөзформаларды лемматизациялау (лексема түріне келтіру) және оның қай сөз табына қатыстылығын анықтау әрекеттері ағылшын, француз және испан тілдері бойынша толығымен автоматтандырылған [4,455]. Бірақ автоматтанған белгілеу нәтижелерін «қолмен» (автоматтандырылмаған) тексеру жұмыстары әлі де болса орын алып келеді. Мысалы, Британдық ұлттық корпустың негізгі массиві осылайша, яғни әрі автоматты, әрі қолмен тексеру арқылы аннотацияланған (белгіленген) болатын. Әзірше, көптеген тілдер бойынша, тілдік тұрғыда күрделі болып саналатын мәтіндерге шартты белгілерді енгізу жұмысы толығымен «қолмен» (автоматсыз) атқарылуда. Мысалы, анафоралық және катафоралық референцияны құрастыруда қолмен белгі қою орын алады. Сонымен, мәтіндер корпустарына автоматты түрде белгі қою жұмысы әлі де болса өзін ақтай алмай келеді, яғни қате кету мүмкіндігі мол [2, 65].

Ғалым Энтони МакЭнери өз зерттеуінде [3] шетелдік дереккөздерге сүйене отырып, әртүрлі белгі қою мүмкіндіктерін 7-кестеде көрсетілгендей сипаттайды. Осы сияқты корпустарға берілген аннотацияларды ғаламтордағы (интернеттегі) Ұлыбританияның Ланкастер университетінің мына сайтынан табуға болады [2, 65]:

http://www.Ling.lancs.ac.uk/monkey/ihe/linguistics/corpus2/2fral.htm.

7-кесте

Корпустарды белгілеу түрлері

Аннотация (белгілеудің) түрлері	**Анықтама**	**Сілтеме**
1	2	3
Лемматизация	Сөзформаны лексемаға келтіру, лексеманы анықтау	**Кең тараған процедура.** Оның сипатталуын орыс тілінің Ұлттық корпусы сайтынан (www.ruscorpora.ru) және орыс әдеби тілінің корпусы сайты бойынша (*www.narusco.ru*) танысуға болады.

1	2	3
Сөз табына қатыстылығы (тэггер)	**Морфологиялық белгілеу:** грамматикалық формаларды анықтау, грамматикалық мағыналарды кодтау (шартты белгілер қою)	**Кең тараған процедура.** Оның сипатталуын орыс тілінің Ұлттық корпусы сайтынан (www.ruscorpora.ru) және орыс әдеби тілінің корпусы сайты бойынша (www.narusco.ru) танысуға болады.
Парсинг	**Морфосинтаксистік және синтаксистік белгілеу:** тәуелділік таралымын (деревье зависимости) құру, синтаксистік конструкцияларын анықтау	**Кең тараған процедура.** Оның сипатталуын орыс тілінің Ұлттық корпусы сайтынан (www.ruscorpora.ru) және орыс әдеби тілінің корпусы сайты бойынша (www.narusco.ru) танысуға болады.
Семантикалық белгілеу	Семантикалық категорияларды және олардың лексикалық пен грамматикалық көрсеткіштерін бейнелеу тәсілдерін анықтау	**Әзірше кең таралмаған процедура.** Мұны Сингапурлық Корпус бойынша танысуға болады *Ng H.T., Lee H.B.* Integrating multiple knowledge sources to disambiguate world sense: an example-based approach. Proceedings of the 34th Annual Meeting of the Association for Computational Linguistics (Santa Cruz, Calif.) 1996. P. 40-47
Дискурстық белгілеу	Дискурстық параметрлерді анықтау (дискурс түрлері және т.б.), екпін, үзіліс, түсіндірме (огаворка) түрлері арқылы просодикалық белгілеулер қамтылады	**Жедел дамып келе жатқан сала** *Botley S., McEnery A*. (eds.) Discourse Anaphora and Resolution. Studies in Corpus Linguistics. Amsterdam: John Benjamins. 2000

1	2	3
Прагматикалық белгілеу	Прагматикалық параметрлерді анықтау (сөйлеу актілерінің түрлері), анафоралық белгілеулер түрлері қамтылады (есімдіктер референциясын анықтау)	**Жедел дамып келе жатқан сала** *Short M.,, Culpeper J., Semino E.* Using a corpus for srylistics research: speech presentation // *Short M and Tomas J.* (eds). Using Corpora for Language Research. London: Longman, 1996
Ақпаратты шығарып алу (IE – orlented annotation)	Корпус мәтінінің мазмұнына талдау жүргізу негізінде ақпаратарды іздеу және оны компьютер жадынан шығарып алу	*Aone C., Bennett S.W.* Discourse tagging and discourse tagged multilingual corpora. Proceeding of International Workshop on Sharable Natural Language Resources (Nara), 1994. P. 71-77

1.9. Корпустардың ерекше типтері

Параллель корпустар. Қазіргі кезде мәтіндер корпустарының бірнеше түрі кездеседі. Солардың ішінде «қатар тілдер» (параллель) корпустары бір тілден екінші тілге аударма жасауға қатысты талдау жұмыстарын жүргізуге аса қолайлы болып келеді. Мысалы, «Орыс тілінің Ұлттық корпусында» орыс тіліндегі мәтінге оның басқа тілге аударылған үлгісі және, керісінше, шет тілдеріндегі мәтіндерге орысша аудармасы сәйкестендірілген.

Параллель корпустарды мынадай негізгі екі топқа бөлуге болады:

1) параллель корпустар – қайсыбір негізгі тілде жазылған түпнұсқа-мәтіннің және осы негізгі мәтіннің бір немесе бірнеше басқа тілдерге аударма-мәтіндерінің жиынтығының электрондық қоры;

2) параллель корпустар – екі немесе бірнеше тілде жазылғанына қарамай, бір ғана тақырыптық салаға қатысты мәтіндердің электрондық қоры.

Аталған екі корпус та тілдерге салыстырмалы зерттеулер жүргізуге арналып қарастырылады (лексикология, грамматика, стилистика, аударматану және т.б.) және аударманың нәтижелі әдістерін, соның ішінде, мәшинелік аударма әдісін де зерттеп, дамыту үшін қажет.

Мәтіндердің параллель корпустарының *бірінші типін* дайындауда және оларды өңдеуге арналған бағдарламалар пакеттерін зерттеуде түпнұсқа мәтін мен оның аудармасы арасындағы сәйкестікті айқындауда проблема туындайды [8]. Бұл мәселенің шешімін табу үшін мәтіндерді автоматты түрде ***теңестіру*** (выравнивание, alignment) әдісі қолданылады. Әдістің негізінде түпнұсқа мәтін мен оның аудармасын сөйлемдерге, клаузаларға (грамматикалық құрылыстар бойынша), сөзтіркестер мен сөздерге параллельдік сипатта сегменттеу (бөліктеу) әрекеті жатыр. Сөйлем деңгейінде теңестіруде қолданбалы тіл білімінің мамандары А.В. Зубов пен И.И. Зубованың оқулығында [17] жазылған әдісті қолдануға болады, яғни төменде көрсетілген екі параллель мәтіндер сөйлемдері арасындағы мүмкін болатын алты түрлі сәйкестіктік түрлері кездеседі:

1) қарастырып отырған бір сөйлем бір сөйлем ретінде аударылады;

2) қарастырып отырған екі сөйлем бір сөйлем ретінде аударылады;

3) қарастырып отырған бір сөйлем екі сөйлем ретінде аударылады;

4) қарастырып отырған екі сөйлем екі сөйлем ретінде аударылады, бірақ ол сөйлемдердің ішкі шекарасы түпнұсқа мәтіні мен аударма мәтінінде сәйкес келмейді;

5) қарастырып отырған сөйлем аударылмайды;

6) аударма мәтініндегі сөйлемнің түпнұсқа мәтінінде баламасы жоқ.

Бұл проблемалық мәселені шешуде аралық-тіл немесе әмбебаптық тіл тәсілдерін қолданатын мәшинелік аударма жүйелер технологиясын пайдалану теориялық тұрғыда негізді болмақ [8].

Тәжірибе жүзінде, сөйлем ұзындықтарын салғастыру, мәтінді абзацтарға бөліктеу, тыныс белгілерін талдау, сыртқы сөздікті және т.б. факторлардың сәйкес келу негізінде мәтіндерді автоматты түрде салғастыру мен теңестірудің әртүрлі компьютерлік программалары орын алады. Бұл программаларда автоматты теңестіру нәтижелері адамның қатысуымен редакцияланады.

Түпкі және аударма мәтіндердің бірліктері арасында «теңестіру» нәтижесінде арнайы қарастырылған шаралар бойынша сәйкестік жүзеге асады. ***Теңестірілген паралельді корпус*** – ол ғылыми зерттеулердің, әсіресе, аударма жасаудың теориясы мен практикасының аса тиімді құралы.

Мәтіндердің параллель корпустары ауқымды ақпарат алуға мүмкіндік туғызады. Олардың көмегімен мынадай жұмыстарды жүзеге асыруға болады:

– екітілдік және көптілдік аударма сөздіктерін құрастыру;

– мәшинелік аударма жүйелері үшін сөздіктер құрастыру және оларды толықтыру;

– ұзындығы жағынан сөйлем ұзындығынан артық болатын көпмағыналы сөздердің контекстік қоршауын пайдалана отырып, компьютер көмегімен лексикалық бірліктердің полисемиясын жою.

– мәтіннің терминологиялық және фразеологиялық бірліктерін аудару;

– аудармалық жады бар жүйелер деп аталатын мәшинелік аударманың жаңа жүйелерінің аясында, әртүрлі деңгейде өзара теңестірілген мәтіндер корпустары мен олардың аудармаларын компьютер жадында жинақтау және соның негізінде автоматты аударманы толық жүзеге асыру.

Мұндай жаңа жүйе мәтіндерді аудару кезінде кезекті аударуға тиісті сөйлемді немесе оның фрагментін бастапқы параллель мәтіндердің массивінен тауып алуға ұмтылады. Егер іздеген сөйлем немесе оның фрагменті бастапқы түпнұсқа-мәтін массивінен табылса, онда жаңа жүйе аударылған мәтіндер массиві ішінен сол сөйлемнің немесе оның бөлігінің аудармасын бөліктеп алады [17].

Параллель корпустарды, сонымен бірге екінші типті корпустарды зерттеу кезінде лексиканы автоматты классификациялау құралдарын ұтымды қолдануға болады. Лексиканы автоматты түрде классификациялау мәтінді автоматты түрде түсінудің ең бір өзекті әдісі болып табылады [18]. Мұндай әдіс мәтін құрылымын формалдау және мәтін элементтері арасындағы семантикалық байланысты сандық деңгейде бағалау (сөздер, леммалар мен сөзформалар) арқылы жүзеге асады. Сөздердің қолданысы және олардың өзара семантикалық жақындығын сипаттайтын сандық деректердің салыстырмалы талдауы әртүрлі тілдердің лексика-семантикалық және тақырыптық топтар ішіндегі лексикалық бірліктерінің үлестірімін анықтауға көмектеседі.

Екінші типті параллель корпустардағы түпнұсқа және аударма мәтіндерін қатар өңдеу кезінде кластерлер элементтерінің қатынасы жайлы ақпарат аударма адекваттығын анықтау мен контрастивтік зерттеулер жүргізуде аса жоғары бағаланады. Лексиканы автоматты түрде классификациялау модульдерін қолдану параллель корпустардағы іздеу процесінің тиімділігін арттырады, көптілді сөздіктерді толықтыру мен өңдеу жұмыстары үшін қажетті деректерді алуға, мәшинелік аударма жүйесі жұмысының және олардың білім берудегі сапасын бағалауға мүмкіндік береді [19, 7].

Аудармалық жады бар жүйелер аударма процесін автоматтандыруды барынша дамыту үшін қажет. Мәшинелік аударма жүйелері туыстас табиғи тілдердің синтаксистік ұқсастығына негізделеді. Мысалы, бұл чех және словак тілдері үшін, ал түркі тілдерінде қазақ, қырғыз, қарақалпақ және т.б. тілдерге қатысты деуге болады.

Жоғарыда сөз болған параллель корпустардың екі типі де тілдерді салыстырмалы зерттеулерді (лексикология, грамматика, стилистика, аударматану салаларында) жүзеге асыру үшін құрастырылады және қолданыс табады. Сонымен бірге, бұл корпустың келесі мақсаты бір тілден екінші тілге аударуға қатысты ұтымды әдіс-тәсілдерді жете зерттеу және машиналық (компьютерлік) аударма қатысты мәселелердің шешімін табу, аударма сапасын арттыру.

Параллель корпустар көбінесе көптілді қоғамдастықтырдағы мәтіндер негізінде құрылады, мысалы, Біріккен Ұлттар Ұйымы (БҰҰ), ЕвроОдақ және Канада сияқты екітілді мемлекеттер.

Сөйлеу тілінің корпусы. Компьютерлік лингвистика мен корпустық зерттеулерде прагматика басқа тілдік салаларға қарағанда соншалықты терең зерттелмеген, сол себепті сөйлеу тілінің репрезентативті корпусын құрастыру аса күрделі мәселеге айналды. Солай бола тұра, әдеп модельдерін, рөлдердің ауысу сипатын және т.б. мәселелердің (коммуникативтік) модельдерін жасау қажет болды [1].

Корпусты құрастырушылар корпус арқылы шешімін табатын лингвистикалық мәселелердің барлығын бірдей тани бермейді. Солардың ішінде аса маңыздысы тілді танудың негізін құрайтын ауызша мәтіндерді зерттеу болып табылады. Лондон-Лунд (The London-Lund Corpus) корпусын құрастыру «Обзор употребления английского языка» (The Survey of English Usage) атты жоба аясында зерттелген. Жобаның мақсаты – ағылшын тілді білімді ересек адамның ағылшынша сөйлеу тілінің грамматикалық жүйе ерекшеліктерін мүмкіндігінше толық түрде жазып алу болған. Аталған жоба Лондон университетінің колледжінде Р. Квирктың ғылыми жетекшілігімен 1960 жылдан зерттеле бастайды. Корпус көлемі – 1 млн. сөзқолданыс. Оған ауызша тілдің мәтіні ретінде радиохабар жазбалары, мәжілістер және бейресми әңгіме жазбалары алынған. Корпустың мәшинелік (электрондық) нұсқасы Лунд университетінде (Швеция) жасалып, 1979 жылы пайдалануға беріледі. Ауызша тілдің тек Лондон-Лунд корпусы ғана ең алғашқы мәшине оқи алатын корпустар қатарына жатады. Ол жасырын түрде ауызша дискурстан жазылып алынған әңгімелер түріндегі 34 мәтіннен тұрады және Дж. Свартвик пен Р. Квирктің «Корпус английского разговора» атты кітабында 1980 жылы басылып шыққан болатын [20]. Бұл кітап компьютерлік корпустар әлі де болса көпшілік назарына кең түрде таралмай тұрған және ауызша тілдің транскрипциясы оқырманға қиындық туғызатын кезде жарық көрген аса маңызды әрі құнды еңбек еді [21]. Ауызша тіл корпусын құрастырудың күрделігінен бұл корпус ағылшын сөй-

леу тілін компьютер көмегімен зерттеу құралы ретінде ұзақ уақыт бойы аса маңызды дерекнама болып қала берді.

Лондон-Лунд корпусының жарық көруі лексика, грамматика, сөйлеу просодиясы, әсіресе дискурс құрылымы мен оның қызметіне қатысты көптеген зерттеулерді өмірге әкелді. Нәтижесінде, ағылшын тіліндегі *actually, really, you know, you see, I mean, well* сөздерінің қолданулары зерттелді, ағылшынша сөйлеу тіліндегі сұрақ қою мен жауап беру мәселесі, пассив құрылымдар және ағылшынша сөйлеудің просодикалық моделін қолдану және т.б. Лондон-Лунд и Ланкастер-Осло-Берген корпустары қоры негізінде ауызша және жазбаша ағылшын тілдері салыстырмалы зерттеулердің нысаны бола бастады; мысалы, модальдылық, күрделі сөйлемдердегі байланыстар (связи), болымсыздық, жоққа шығару, терістеу модалдылықтары. Қазіргі кезде корпустық лингвистерді ауызша тілде эмоцияны білдіру, таңғалуды сөзбен білдіру тәсілдері және т.б. қызықтырады. Мұндай мәселелерге арналған зерттеулерді жүргізуге мүмкіндік беретін корпустарға мысал ретінде орыс тілінің ұлттық корпусы (НКРЯ – ОТҰК) құрамындағы мультимедиялық шағын корпусты (субкорпусты) атауға болады.

Электрондық форматтағы ауызша және жазбаша материалдарға қолжетімдік теңгерімі әлі де ұзаққа созылады. Оның себебін әртүрлі себептермен ауызша тіл корпустарын құрастыру жазба тіл корпустарын құрастырудан барынша баяу жүріп жатқандығымен түсіндіруге болады. Мұны жолға қою үшін ең алдымен, ауызша тілді жазып алудың (есте сақтаудың) бір амалын табу керек – мысалы, магниттік лента көмегімен, цифрлік жазбалар арқылы немесе видеокассеталарға жазу. Содан кейін, оларды әріптермен жазып алу қажеттігі туындайды, әрине мұндай әрекет барынша жалықтыратын жұмыс болуы және қымбатқа түсуі мүмкін, себебі, ондай жазбаның сапасы табиғи жағдайдағы сырт ортаның шулы-шусыз жағдайына байланысты болады.

Фонетикалық лингвистикалық ресурстарды құрастырудағы басты күрделілік ауызша тілді транскрипциялау қажеттігіне байланысты. Мұндай жағдайда келесі мәселелер туындайды:

1. Транскрипциялау үшін қандай алгоритм қолдану қажет?

2. Айтылу кезінде сөйлеуші факторы есепке алына ма?

3. Ауызша мәтін толық ескерілу керек пе, әлде тек оның фрагменті ғана ма?

4. Сөздің айтылуында диалектілік варианттар ескеріле ме?

5. Айтылу кезінде сөздердегі екпін ескеріле ме?

6. Фразалар айтылғанда просодиялық белгілер ескеріле ме?

7. Тыңдау кезінде танылмаған сөздерді арнайы белгілеу қажет пе?

8. Фонетикалық корпус үшін паралингвистикалық құбылыстарды, ілеспелі дыбыстарды (кідіріс, күлкі, жөтел, міңгірлеу және т.б.) белгілеу қажет пе? [17].

Қазіргі кезде мәшине оқитын фонетикалық корпустарды құрастыру үшін сөйлеу тілі дыбыстарының орфографиялық көрінісіне негізделген транскрипция қолданылады. Қажеттікке қарай, просодикалық, паралингвистикалық және т.б. бейнелейтін қосымша белгілер де транскрипция үшін пайдаланылған.

Фонетикалық корпустарды құрастырудың қиындығына қарамай, әлемде мұндай тұлғалы корпустардың көптеген түрлері жасалғаны мәлім. Мәселен, А.В. Зубов пен И.И. Зубованың оқулығында сипатталғандай, XX ғасырдың 70-жылдары АҚШта Х.Далем және оның әріптестері «Корпус устной речи американского варианта английского языка» атты фонетикалық корпус құрастырады. Бұл корпустың көлемі психоаналитикалық сеанстар кезіндегі жазба мәтінінен алынған 1 млн. сөзқолданыстан тұрады. Корпусты құрастырушылар иелігінде болған 15 таспаның әрбіреуінде кездейсоқтықпен сеанс мәтіндерінен таңдалып алынған 225 жазба мәтін болған. Онда АҚШ-тың 9 қаласынан қатысқан 8 әйелдің және 21 ер адамның дауыстары жазылған. Іріктеліп алынған жазбалар ағылшын орфографиясының стандарты негізінде транскрипция процесінен өтеді. Айтылымдарда кездескен диалектілік нұсқалар есепке алынбаған, танылмаған сөздер жазбада Z әріпімен белгіленген, сол сияқты екпіндер мен сөздердің басқа да просодикалық сипаттары ескерілмеген. Солай бола тұра, ауызша сөйлеу тілінің орфографиялық жазбасында пауза, күлкі, дем, жөтел және

басқа да паралингвистикалық құбылыстар түсіндірме ретінде арнайы белгіленген [17].

Британ ұлттық корпусын құрастырушылардың бірі Л. Бернардтың пайымдауынша, корпусты құрастыру кезінде (1990-шы жылдары) ауызша дерекөздерден 10 млн. сөзді іріктеп алу жазба мәтіндерден іріктеп алынған 50 млн. сөзбен бірдей болған екен [22].

Орыс тілінің Ұлттық корпусы құрамында (Орыс тілінің корпусы – ОТК деп те аталады) 2008 жылдың қаңтарында бар болғаны 3,9% ауызша мәтін болды. Мәтіндер корпусының «Ауызша» құрамдас бөлігі мынадай типтерге бөлінеді: бұқаралық сөйлеу (публичная речь) (64,3%), бұқаралық емес сөйлеу (непубличная речь) (8,1%), кино сөзі (27,6%) [10].

Сонымен, мәтіндердің ұлттық корпусын құрастыруда барынша маңызды проблемалық мәселе ауызша сөйлеу мәтінімен толықтырудың жеткіліксіздігі, әсіресе ол ауызша сөйлеуге, кездейсоқ сұхбатқа және т.б. қатысты.

Білім беру корпустары. XXI ғасыр білімге қатысты түбегейлі жаңа көзқарасты қажет етеді. Білім беру әрекеті жеке оқушының ойлау қабілетін сыни және шығармашылық тұрғыда қарастырғанда әр уақытта даму процесінде болуы қажет. Бірақ бұл үшін білім беру жүйесінде тек қана мұғалім мен оқулықтың болуы жеткіліксіз. Оған оқушыны өз бетінше ойлауға ынталандыратын кең тараған ақпараттық әрекет алаңы, бірнеше ақпарат көзі және бір ғана мәселеге қатысты бірнеше көзқарастар жиынтығы да қажет болады.

Интернетте білім беру мақсатында бірнеше ауқымдағы ресурстар пайдаланылады. Осындай ауқымды жобалардың бірі ретінде әртүрлі тілдердердің Ұлттық корпустарын айтуға болады. Білім беру аясындағы корпустардың ең бір негізгі қызметі – кез келген тілді оқыту мүмкіндігі. Сол сияқты, корпус мұғалімдерге жеке жоспарды құрастыруда, лексика саласы бойынша бақылау мен тексеру тапсырмаларын құрастыруда, морфология, синтаксис салалары бойынша және т.б. қажетті материалдарды іздеп-табуда баға жетпес көмек көрсете алады.

Осы айтылғандарға қатысты В.А. Плунгян былай дейді: «Использование Корпуса в преподавании гуманитарных дисциплин в

школе поможет успешно решать и чисто научные, и собственно педагогические задачи» [16].

Арнайы корпустар. Мәтіндердің арнайы корпусы – ол, әдетте, белгілі бір зерттеу міндетіне орай жасалған және көпшілік жағдайда, корпусты құрастырушының мақсатына сай қолдануына арналған көлемі жағынан шағын (бірнеше мың сөзқолданыс) болып келетін теңгерімделген (баланстандырылған) корпус. Мысал ретінде А.И. Герцен атындағы РМПУ-дың Қолданбалы лингвистика кафедрасы құрастырған *Петербург мектеп оқушыларына арналған ағылшын мәтіндерінің корпусын* атауға болады. Бұл корпусты құрастырудың негізгі мақсаты – орыс мектеп оқушылары ойлап тапқан ағылшын тіліндегі мәтіндердің ерекшеліктерін зерттеу болған. Түпнұсқалық (аутентті) мәтіндік материалдар 2007 жылдың қараша айы мен желтоқсан айы арасындағы Санкт-Петербург мектептерінен жиналады. Мәтіндер авторлары 9-11 сыныптарында оқитын алдын ала арнайы тестілеуден өткен 78 оқушы еді. Оқушылардың ағылшын тілін меңгеру деңгейі орташа / intermediate (26%) және ортадан жоғары / upper-intermediate (74%) екендігі анықталған. Аталған корпустың көлемі 38 мың сөзқолданысты қамтиды. Корпус негізіндегі зерттеу нәтижелері бойынша, өте қарапайым құрылымға және стандартты ағылшын тілінің барынша табиғи үлгісіне артықшылық беру әсіресе тіл үйренушілерлердің сөздік қорына, ағылшынша сөйлеу құрылымдарын меңгеруге зиянын тигізеді. «Петербург мектеп оқушыларына арналған ағылшын мәтіндерінің корпусында» грамматикалық құрылымда «өтпелі грамматиканы» (интертіл) бейнелейтін жағдайлар кездесетіні айқындалған [23].

Құрастыру мен стандарттау тұрғысынан алғанда тарихи корпустар да арнайы корпустарға жатады, бұл корпус тарихи ескерткіштерді зерттеуге арналады. Оған орыс тілінің *XVI-XVII гг. Агиографиялық мәтіндердің Санкт-Петербург Корпусын* (АМСК) жатқызуға болады (http://project.phil.pu.ru/skat). АМСК – ол СпбМУ-дың филология факультетінің математикалық лингвистика кафедрасында құрастырылған орыстың ежелгі агиографиялық әдебиеті ескерткіштері мәтіндерінің электрондық корпусы

болып табылады. Агиографиялық шығармалар тілі көпшілік жағдайда XV–XVII ғасырлардағы орыс әдеби тілін сипаттайды. 2006 жылы аталған корпустың көлемі 500 мың сөзқолданысты құраған [24].

Мәтіндердің арнайы корпусының сұранысы, негізінен алғанда, филологияға, лексикографияға, лингвистикаға бағытталған ұлттық корпустардың қажеттігінен кем емес. Кез келген салалық мәтіндердің арнайы корпусы тілшілерге де және басқа сала мамандарына да салалық терминдер жүйесіне қатысты мәлімет бере алу мүмкіндігіне ие [25].

Ұлттық корпус. Корпустық лингвистиканың жетістіктерін өзіне сақтаған аса дамыған корпус түрі – ұлттық корпус. Ұлттық корпус сол тілдің өмір сүрген белгілі кезеңіндегі сан алуан жанрын, стилін, аймақтық, әлеуметтік нұсқасын және т.б. да түрлерін қамтиды. Ұлттық корпус деген термин бір қарағанда белгілі бір ұлттың, белгілі бір ұлтқа тән корпус дегенді де білдіреді. Алғаш ұлттық корпустар деп осындай белгілі бір тілдің корпустарын атаған. Мысалы, 1990 жылдары жасалған Британ ұлттық корпусын осы мәнде атаған. Кейін келе, ұлттық корпус деп белгілі бір тілдің табиғатын толық бейнелей алатын көлемді корпустарды атай бастаған. Мұндай корпус белгілі дәрежеде ұлттық тілді толық түрде бейнелейді. Ұлттық корпустың репрезентативтілігі (тұлғалылығы) – сол тілдің жазба және сөйлеу түріндегі мәтіндерінің барлық типтерінің бейнеленуі. Ұлттық корпустың айтарлықтай дәрежеде көлемді (ондаған, жүздеген миллион сөзқолданыс) болуы репрезентативтілікке жетудің қажетті шарты болып саналады. Ұлттық корпустың ажыратылмас бөлігі оның белгіленген (аннотацияланған, мазмұндалған) бейнесі.

Ұлттық корпустардың осы мәнде қолданылуы белгілі бір кезеңдерді қамтитын мәтіндерден құрылған Чех ұлттық корпусынан басталады. Сондай-ақ ұлттық корпус деп атаудың екінші бір критерийі оның миллиондаған сөзқолданыстан тұратын көлемді болуы, үшіншіден, ұлттық корпустарда белгілі біл тілдегі мәтіндердің барлық түрі, яғни барлық стильде жазылған мәтіндер қамтылу керек.

1.10. Қазақ тілінің ұлттық корпусын құрастыру мәселесі

Қазақ тілінің ұлттық сипаттағы «тұлғалы» тілдік корпустарын құрастыру мәселесі қазіргі кезде Қазақстанның бірнеше ғылыми-қолданбалы бағыттағы орталықтарында (А.Байтұрсынұлы атындағы Тіл білімі институтының Қолданбалы лингвистика бөлімінде, Еуразия университетінде «Жасанды интеллект», әл-Фараби атындағы ҚазҰУ-де т.б.) қолға алынып, дербес жұмыс істеп жатыр. Алайда олардың барлығы да орыс тілі тәжірибесіне сүйеніп, корпус құрастыру мәселесін өзінше шешемін деп талаптанғанымен, ауқымды мәтіндерді компьютер жадына енгізу, лингвистикалық белгіленімдер талдамасын жасау ісінде шашыраңқылық бар.

Өйткені, біріншіден, әр мекемеде жасалып жатқан корпустардағы лингвистикалық белгіленімдер мен олардың моделі, шартты белгілері бірізді емес.

Екіншіден, автоматты түрде лингвистикалық белгіленім қою мәселесі әлі де болса толық шешімін таппаған. Тілдік талдаулардың өзінде де даулы мәселелер баршылық. Сондықтан аннотацияланған тілдік корпустарды құрастыру ісіне білікті, тәжірибелі лингвистерді тарту қажеттігін туындатады.

Үшіншіден, жоғарыда сөз болғандай, миллиондаған сөзқолданыстан тұратын корпустар құрастыру үшін аса көлемді мәтіндердің электронды нұсқасы қажет болады. Ал оларды «қолдан» енгізу көп уақытты қажет ететіні белгілі. Осы орайда бұл мәселе Қазақстан аумағындағы кітап, газет-журнал шығаратын баспалармен келісімге келе отырып шешілетін мәселе деп санаймыз.

А.Байтұрсынұлы атындағы Тіл білімі институты қазақ тілінің корпусын құрастыру мәселесі бойынша 2009 жылдан бастап зерттеу жүргізіп келеді. Қазақ тілінің әртүрлі стильдері бойынша интернет желілерінен, электронды кітапханалардан т.б. дереккөздерден мәтіндер жинақталып, автоматты түрде мәтіндердегі сөздерді түбір сөздер мен сөз түрлендіруші қосымшаларға ажырататын және оларға морфологиялық белгіленім қоятын бағдарлама жасалды. Қазіргі уақытта басқа да лингвистикалық белгіленім түрлері бойынша зерттеулер жүргізіліп жатыр.

Корпус құрастыру ісін орталықтандырып, үлкен күшпен мемлекет тарапынан қолдау көрсетілетін болса, мынадай корпус түрлерін жасап шығуға болар еді.

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 бейне және дыбыс жазбалары құрайды);

4) Қазақ тілімен 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).

Бір айта кететін жайт, «Орыс тілінің Ұлттық корпусы» бойынша ақпарат іздеу жүйесін құру әрекетіне «Яндекс» компаниясы қолдау көрсеткен. Сол сияқты «Қазақ тілінің Ұлттық корпусынан» ақпарат іздестіруге және оның интернеттегі сайтының дизайнына да қолдау көрсететін компаниялар табылып, бірігіп жұмыс істеу қажет.

Қорыта келе айтарымыз, Ұлттық тіл мәтіндерінің компьютерлік корпусын құру жобасы бір ғана ғылыми ұйымның шешетін мәселесі емес және ол зерттеу жұмысы 3-5 жылда аяқтала қояды деуге де болмайды. Себебі бұл аса күрделі және оның нәтижелері әлемдік дәрежедегі аса маңызды ғылыми жұмыс болып саналады. Зерттеу жұмысының мақсатына сай орындалатын міндеттері де сала-салаға, кезең-кезеңге бөлініп, тек қана ғалымдар қауымдастығын құру арқылы ғана ауқымды нәтижеге ие болары сөзсіз.

Қазақ тілінің тілдік корпустарын жасау – көп болып жұмылып атқаратын ұлттық құндылығымыз. Сондықтан Қазақстанның

әр жерінде бір-бірінен дербес атқарылып жатқан корпус жасау ісін орталықтандыру керек.

– оған Қазақстанның әр жерінде ғылыми-педагогикалық қызмет атқарып жүрген ғалымдардан арнайы лингвистикалық топ құру керек. Өйткені тілдік бірліктерді модельдеу – өте күрделі мәселе. Сондай-ақ тілдік корпустар құрастырудың өзі ең алдымен лингвистикалық белгіленім талдамасын жасауға тіреледі;

– осы кезге дейін жасалып жатқан корпус жасау тәжірибесіндегі нәтижелерді бір орталыққа жинақтау керек;

– кітап, газет-журнал шығаратын баспалармен шартқа отырып, олардан мәтіндердің электронды нұсқаларын алу қажет.

Сонымен, мұндай ауқымды іске мемлекет тарапынан қолдау көрсетілсе, біртұтас қазақ тілінің Ұлттық корпусын құрастыру ісі алға басатыны сөзсіз. Мұндай Ұлттық құндылықты жасап шығару бүгінгі қазақ тіл білімінің ғана емес, қоғамның болашаққа қояр талаптарының бірі деп білеміз.

2. КОРПУСТАР ҚҰРАСТЫРУ МӘСЕЛЕЛЕРІ

2.1. Корпус құрастыруды жобалау және технологиялық үдеріс

Кез келген корпус жобасы оны құрастырудың кезеңдері мен әрі қарайғы даму жолдарын айқындап алуды қажет етеді. Корпус ұғымы тілшілердің үнемі пайдаланған дәстүрлі картотека қорын жасау жұмысының жалғасы деуге болады. XX ғасырда бұл картотекалар компьютер жадына енгізіліп, көпшіліктің пайдалануына берілді. Корпустық әдістің қалыптасуына Интернет желісі айтарлықтай рөл атқарды, интернеттің дамуы арқылы үлкен көлемді мәтіндік материалдарға қолжетімді болып, әртүрлі лингвистикалық зерттеулер жүргізуге мүмкіндік туды.

Осындай жағдайда сөздіктер мен грамматикалар құрастыруға негіз болатын дәстүрлі мәселе – тілдік материалдардың репрезентаттығы мен теңгерімдігі, әсіресе, ұлттық корпустарды құрастыру кезінде де туындайды. Корпустың репрезентаттығы мәтіндік материалдардың жеткілікті көлемімен және оның әртүрлі мәтіндерді қамтамасыз етуімен түсіндіріледі.

Корпус мәтіндерінің жанрлық-тақырыптық құрылымынан басқа да көптеген маңызды мәселелерді шешуге тура келеді, оларға жататындар:

1. Корпустағы мәтін дегеніміз не? Мысалы, газеттегі шағын хабарландырулар жеке мәтін ретінде енгізіле ме, әлде оларды біріктіріп бір мәтін ретінде енгізіле ме?

2. Газет бетіндегі мақала жеке мәтін бола ма? Немесе газеттің бір санындағы барлық мақалалар бір мәтін болып санала ма?

3. Өлеңдер жинағындағы әр өлең бір мәтін бе немесе жинақ тұтастай бір мәтін бе?

4. Хат жазысудағы әр хат бір мәтін бе немесе жауап хат екеуі бір мәтінге жата ма әлде ондай хаттардың жиыны бір мәтін құрай ма?

Бұл аталғандардан маңызы кем түспейтін нәрсе – корпустағы мәтіндердің хронологиялық мәселесі. Әр жанрға қатысты корпустағы хронологиялық шеңбер бірдей болуы керек пе, әлде әр жанрға бөлек пе?

Корпус көптеген пайдаланушы үшін және әрқилы мәселелердің шешімін табу мақсатында құрылатындықтан, мәтіндегі қандай дерек қажет, ал қайсысын алып тастаса да болатындығын білген жөн. Мысалы, мәтін ішіндегі суреттер тілдік материалдарға қажеті жоқ, сондықтан оларды алып тастауға болады. Ал мәтінде орын алатын кестелер, цитаталар, кірме сөздер және өлшем бірліктерін алып тастай беруге болмайды. Мұндай мәселелерді алдын ала корпусты жобалау кезінде шешіп алған жөн болады. Аталғандар немесе және т.б. кейбір аталмаған мәселелер біртіндеп, корпусты құру барысында, тәжірибелік пайдалану (корпус қызметін сыннан өткізу) кезінде олардың корпуста қалу-қалмау жайттары өз шешімін табуы керек.

Корпусты құрастырудың технологиялық үдерісі мынадай кезеңдерден тұрады:

1. *Мәтіндерді корпусқа енгізу алдын ала жасалған дерекнама тізіміне сәйкес келуі.*

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

3. *Мәтіндерді талдау және алдын ала өңдеу.* Бұл кезеңде әртүрлі дерекнамалардан алынған мәтіндер филологиялық тексеру мен түзетуден өтеді. Мәтіндерді технологиялық тұрғыдан дайындау кезінде библиографиялық және экстралингвистикалық тұрғыдан сипаттау жұмысы іске асырылады.

4. *Айырбастау (конвертациялау) және графемалық талдау.* Кейбір мәтіндер бір немесе бірнеше кезеңдік алдын ала жүргізілетін мәшинелік өңдеуден өтеді. Олардың нәтижесінде (қажеттікке қарай) қайта кодтау жүзеге асырылады, сол сияқты мәтіндік емес элементтерді (суреттер, кестелер) жою немесе оларды өзгерту, тасымал белгілерін жою, сызықшалардың жазылуын бірізденді́ру және т.б.

Графематикалық талдау келесі амалдарды орындауды көздейді: корпусқа енгізілетін мәтіндерді элементтерге бөліктеу

(сөздер, бөлгіш белгілері және т.б.), мәтіндік емес элементтерді жою, стандартқа сай емес (лексикалық емес) элементтерді бөліктеу және дайындау (ресімдеу), арнайы мәтіндік элементтерді өңдеу (адамдарың аты мен инициалдары), латын қарпімен жазылған шетелдік лексемалар, суреттердің атаулары, ескертпелер, форзац беттері, сызықтар, титулдық беттер, әдебиет тізімі және т.б. Әдетте, бұл амалдар автоматты режімде орындалады және осы кезеңде мәтіннің құрамдас бөліктерін сегменттеу процесі де жүзеге асырылады.

5. *Мәтін белгіленімі (Разметка текста).* Мәтінге белгіленім жүргізу кезінде мәтінге және оның құрамдас бөліктеріне қосымша ақпарат (метадеректер) тіркеліп жазылады. Метадеректерді 3 типке бөлуге болады:

а) барлық мәтінге қатысы бар экстралингвистикалық деректер;

ә) мәтіннің құрылымына қатысты деректер;

б) мәтіннің элементтерін сипаттайтын тілдік метадеректер.

Корпус мәтіндерін метасипаттау (метабелгіленім) деректердің мазмұндық 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) элементтерін де қамтиды. 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.

6. *Автоматты белгіленім нәтижелерін түзету:* қателерді түзету және бірізділікке келтіру (қолмен немесе жартылай автоматты).

7. *Белгіленімнен өткен мәтіндерді мамандандырылған лингвистикалық ақпаратты-іздестіру жүйесінің құрылымына ауыстыру (corpus manager).* Бұл жүйе тез және көпаспектілі іздестіру мен статистикалық өңдеу жұмысын қамтамасыз етеді (соңғы кезең).

8. *Корпусқа қолжетімділікті қамтамасыз ету.* Корпус дисплейлік класс аясында, компак-диск бойынша және ғаламтор желісі арқылы таралуы мүмкін. Корпусты пайдаланушылардың катего-

рияларына қарай, олар корпусты пайдалануда әртүрлі құқық пен мүмкіндікке ие болады.

9. *Корпусты құрастыру мен пайдаланудың әртүрлі аспектілері жайында құжаттамалық қамтамасыз ету.* Атап айтқанда, метадерек бойынша іздеу әрекетіне мүмкіндік жасайтын белгіленім жайында, корпус-менеджердің сауалдар тілі туралы және т.б. мәлімет келтірілетін құжаттамалық қамтамасыз ету.

Әрине, әрбір нақты жағдайда корпус құрастыру кезінде іске асатын үдерістердің құрамы мен саны жоғарыда аталғандардан басқаша болуы мүмкін немесе корпус құрастыру кезіндегі технология аса күрделі болуы да мүмкін.

2.2. Дерекнаманы іріктеу

Мәтіндер корпусы қайсыбір тілдің кездейсоқтық біріктірілген мәтіндер жиыны емес. Корпусты құрастыру кезінде көптеген проблемалар туындайды. Олардың негізгілері мыналар:

1. Мәтіндер корпусының негізгі бірлігі ретінде не алыну қажет?

2. Мәтіндер корпусының көлемі қандай болуы қажет (ол қанша бірліктен тұруы қажет)?

3. Мәтіндер корпусында қандай жазба мәтіндер дерекнамалары қамтылуы және оның саны қанша болуы керек?

4. Мәтіндер корпусы құрамына енгізілетін мәтіндерді қандай бастапқы тілдік саладан таңдау қажет?

Бұл сұрақтардың алғашқы жауаптарын профессор Р.Г.Пиотровский мен оның шәкірттерінің 1965-1980 жылдардағы жиілік сөздікке мәтін таңдауға қатысты және лингвостатистикалық зерттеулерінде кездестіруге болады. Осы мәселелер Л.Н.Засоринаның редакциялығымен шыққан (1977) жиілік сөздігінің алғы сөзінде де айтылады [26]. Дәл сол кездерде ең алғаш рет бас (негізгі) жиынтық таңдамасын, таңдама көлемін, таңдама порциясын (элементар таңдаманың) және т.б. бағалау үшін әртүрлі статистикалық әдістер қолданылған болатын [17].

Мәтіндер корпусының негізгі бірлігі ретінде *сөзқолданыстар* (әдетте оларды *сөз* деп атайды), *негіздер* (түбір, леммалар) және

сөйлемдер болуы мүмкін. Құрастырылатын корпус көлемі алға қойған мақсатқа байланысты. Мысалы, әріп, әріптер тіркесі, дыбыс, дыбыстар тіркестерінің қолдану жиіліктерін зерттеуде олардың көлемі онша көп болмайды. Ал лексиканы, морфологиялық құбылыстарды және мәтіндердің синтаксистік немесе стильдік ерекшеліктерін зерттеу үшін мәтін көлемі айтарлықтай көп болуы қажет [17].

Келесі мәселелер де проблемалық мәселелер болып табылады:

1. Қандай функционалдық жанрларды мәтіндер корпусы жадына енгізуге болады (көркем проза, драма, өлең, ғылыми мәтіндер, газеттер, журналдар, техникалық сипаттамалар және т.б.)?

2. Қандай уақыттық кезеңдер аралығындағы мәтіндерді корпус жадына енгізуге болады (қазіргі, 10 жыл бұрынғы, 50 жыл бұрынғы, ескі кезең және т.б.)?

3. Корпусқа әдеби тіл мәтіндерін ғана енгізу қажет пе немесе тағы да басқа дереккөз типтерін де енгізуге бола ма? Әдеби тіл деп нені санаймыз?

Осы сауалдарға жауап беру үшін мәтіндер корпусын құрастырушылар әдетте тілші мамандар мен лингвостатистерден кеңес алады немесе анкеталық әдісті пайдаланады. Өз зерттеу тәжірибесіне сүйене отырып, мамандар мәтіндер корпусының жалпы көлемін, мәтіндердің баспадан шыққан мезгілін, мәтіндер саны мен элементар таңдаманың өлшемін, іріктеліп алынған мәтіндер жанрлары мен олардың санын, әр жанр бойынша элементар таңдамалардың санын анықтайды. Анкеталық әдісті мамандардың тәжірибесімен бірге қолдана отырып, оны «Американский корпус наследия» (The American Heritage Intermediate Corpus) атты мәтіндер корпусын құрастыруда пайдаланған. Корпус көлемін 5 млн. сөз (сөзқолданыс) деп анықтап, және корпусқа ағылшын тіліндегі балалар мен жасөспірімдер әдебиетінің 22 бөлімнен (жанрынан) тұратын мәтіндерді және АҚШ-тың 221 мектебіне анкеталар таратып, қандай мәтіндерді енгізудің дұрыстығын көрсетуге өтініш жасайды. Анкеталар арнайы зерттеліп, 19 мың кітап аттарының тізімі жасалады және олардан 1045 мәтін таңдалып алынады. Осының негізінде әрбіреуі 500 сөзқолданыстан тұратын 10 мың элементар таңдама мәтін корпусқа енгізуге дайындалады.

2.3. Табиғи тілді өңдеудің негізгі рәсімдері: токендеу (токенизация), леммаға ажырату (лемматизация), стемминг, парсинг

Корпус дегеніміз – ол әр уақытта компьютерлік деректер базасы және оны құрастыру кезінде арнайы рәсімдер мен программаларды қолдану әрекеті. Мысалы, ***токенизация***, яғни табиғи тілдегі символдар ағымын жекелеген мағыналы бірліктерге (токендер, сөзформалар) бөліктеу табиғи тілді өңдеудің қажетті шарты болып табылады. Егерде тілдер аса жетілген пунктуацияға ие болған жағдайда, токенизация айтарлықтай қиындық туғызбас еді – тіпті ең қарапайым компьютерлік бағдарлама да сөздер аралығындағы бос орын (пробел), тыныс белгілерге сүйеніп, мәтінді сөздерге ажырата алар еді. Шындығына келгенде, табиғи тілдер ондай пунктуацияға ие емес, сол себепті токенизация мәселесі қиындық туғызады. Мысалы, ағылшын тілінде бірізді токендеу мүмкін болмайтын жағдайлар кездеседі. Салыстырыңыз: мәтін ішіндегі *chap*. жолы қысқартылып жазылған *chapter* сөзінің формасы немесе сөйлем соңында кездесетін *chap* сөзі болуы мүмкін. Сол сияқты кітап жолында кездесетін *Jan.* деген сөзді *January* сөзінің қысқарған формасы немесе сөйлем соңында кездесетін жалқы есім деп те қарастыруға болады. Бірінші жағдайда, нүкте қаншалықты сөзге қатысты болса, соншалықты сол токенге де қатысы бар, ал екінші жағдайда, нүкте өзі тәг ретінде жеке бөлініп алынуы керек. Сонымен бірге, мұндай қиындықтар шектеулі болғандықтан, мәтінді өңдеу кезінде олар ескерілмеуі мүмкін (мысалы, аббревиатуралар (қысқарған сөздер) мен күрделі сөздер ескерілмейді) немесе оларды арнайы алгоритм көмегімен өңдеуге болады. Қазақ мәтіндерінде де белгілі ұғымдарға қатысты қысқартулар кездесуі әбден ықтимал.

Морфологиялық талдаудың өзіне ғана тән мәселесі – ол ***лемматизация***, яғни сөзформаларды алғашқы негіз сөз пішініне (формасына) келтіру әрекеті. Көптеген тілдерде сөз әртүрлі флексиядағы бірнеше формамен кездеседі. Мысалы, ағылшын тіліндегі *'walk'* етістігі мынадай формада көрініс табуы мүмкін: *'walk', 'walked', 'walks', 'walking'*. Мұндағы негізгі форма *'walk'* сөздік реестрінде беріледі де ***лемма*** деп аталады. Лемматизациялау дегеніміз –

бір негіз сөздің бірнеше флективті формаларын топтап жинастыру және бірақ ол формаларды лемматизациялағанда бәрі де бір негіз сөзді – лемманы бейнелеуі қажет, яғни жоғарыда келтірген мысал бойынша *'walk'* сөзін түбір ретінде таниды.

Қазақ тілінде *балаларға, балалардан, балалармен, балаларына* т.б. сөзформалардың барлығы да лемматизациялағанда, яғни леммаға келтіргенде олардың барлығына ортақ *бала* сөзі лемма болып саналады.

Лемматизацилаудан аздап айырым табатын әрекет – ол ***стемминг*** деп аталады және бұл да сөздің (сөзформаның) ***стема***сын (негізін) табу әрекетін жүргізеді. Бұл ***стеммер*** әрекетінің ***леммати-зациялау*** әрекетінен айырмашылығы – жеке сөзді (сөзформаны) контекссіз жеке қарастырады, сондықтан бұл сөздің қай сөз табына қатыстығы және мағынасы да ажыратылмайды деген сөз. Әрине, ***стеммерлер*** әдетте өзі қарапайым болып келеді, жылдам табылады, алайда контекстік мағыналарды көрсетпейді. Ал төменгі дәлдік кейбір қосымша зерттеулерде шешуші рөл атқармайтыны мәлім.

Төменде ***стемминг*** пен ***лемматизацияға*** қатысты мысалдар келтірілді. Мысалы, ағылшын тілінде осы екі терминді ажыратуға мынадай сөйлемді мысал ретінде береді:

[The] [quick] [brown] [fox] [jumped] [over] [the] [lazy] [dogs].

Ең көпшілікке танымал стеммерлердің бірі – SnowballAnalyzer мынадай стемаларға бөледі: *[quick] [brown] [fox] [jump] [over] [lazy] [dog].*

Бұл сөйлемдегі сөздердің леммалары келесідей:

[the] [quick] [brown] [fox] [jump] [over] [the] [lazy] [dog].

Лемматизация әдісі сөздердің қай сөз табына қатыстылығына байланысты және корпустағы сөздерді оларға сәйкес келетін лексемаға дейін қысқартады. Тек лемматизация тәсілі ғана зерттеушіге жеке лексеманың барлық варианттарын бөліп алып, оларды зерттеуге мүмкіндік тудырады.

Парсинг дегеніміз – тілдегі лексемалардың (сөздердің, токендердің) сызықтық реттілігін оның формалды грамматикасымен салғастыру әрекеті. Әдетте, оның нәжижесі тәуелділіктер шежіресі (синтаксистік шежіре) болып табылады. Көлемді корпустар үшін

автоматтанған синтаксистік талдағыштарды (анализаторларды, парсерлерді) жасау компьютерлік лингвистиканың ең бір маңызды саласы болып табылады. 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 біте қайнасып жатыр.

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

1. Мәтінді өңдеуге байланысты дайындық жасаудағы әрбір қадам корпус құрастырушының келесі қадамдарына әсер етеді және корпусты бағалауға қатысты тиісті лингвистикалық шешім қабылдауға мүмкіндік береді. Әрине корпусты пайдаланушы өз іздегеніне қол жеткізу үшін осы шешімдерден 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. Қазақ тілінде де мұндай мәселелер жеткілікті. Бұл күрделі сөздер мен күрделі етістіктер, тұрақты тіркестерге қатысты. Лексемаларды осы тәріздес айқындау үшін құрастырушы адам мұндай құбылыстармен не істеу керектігін шешу қажет.

2. Түпкілікті тұтынушыны алдын ала өңдеу кезіндегі қандай жұмыстар орындалғанын және мүмкін болатын қателіктермен хабардар етілуі 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 етуі мүмкін [1].

2.4. Белгіленім ұғымы. Белгіленім түрлері

Корпуста сақталатын бірлікке сөз, сөзтіркес, сөйлем немесе толық мәтіндер жиынтығы жатады. Мәселен, корпусқа енгізілген мәтіндердегі әрбір тілдік бірлікке, ол не жеке сөз, не қосымша болсын, лингвистикалық ақпарат беріледі. Тілде мұндай лингвис-

тикалық ақпарат беру «аннотациялау» деп аталады. (Бұл туралы жоғарыда да сөз болған). Аннотация дегеніміздің өзі «қысқаша мазмұн» дегенді білдіретіндіктен, корпусқа енгізілген тілдік бірліктердің аннотациясы да шартты белгілер арқылы қысқаша беріледі. Орыс тілінде «разметка» деп аталатын мұндай шартты белгілер қазақ тілінде «белгі-код», «белгіленім» деген терминдермен аталып жүр. («Белгіленім» терминінің авторы А.Жұбанов). Мәтіндер бірліктеріне берілген лингвистикалық ақпараттардың толық сипатта болуы осы белгіленімдердің әртүрлілігіне байланысты. Белгіленім тілдегі жекелеген деңгейлерге тән тілдік мәліметтерді бейнелейді. Ондай белгілер семантикалық (лексика-семантикалық), морфологиялық, синтаксистік сипатта болуы мүмкін. Мәселен, орыс тілінің Ұлттық корпусында сөздерді сөз табына, септелуіне, тегіне (род – муж., жен.), жіктелуіне, жанды-жансыздығына, етіс және етістік түріне, шырай және т.б. морфологиялық сипаттамаларына қарай арнайы морфологиялық белгілер қойылған. Сол сияқты семантикалық талдау арқылы да мәтін ішіне қажетті белгілердің қойылатынын айтуға болады.

Мәтіндер корпусы туралы жазылған ғылыми еңбектерде фонетикалық, морфологиялық, семантикалық, синтаксистік белгіленімдер енгізілетіндігі туралы айтылады. Бірақ корпус жасау барысында аталған белгіленімдердің барлығын бір уақытта енгізу қиындық тудырады. Сондықтан шетел, орыс тіл біліміндегі мәтіндер корпустарында лингвистикалық белгіленімдер енгізу ісі кезең-кезеңмен жүзеге асырылған. Аталған белгіленімдердің бәрі қойылып, толық лингвистикалық ақпарат берілген корпустар «терең аннотацияланған» (глубоко аннотированный) деп аталады.

Табиғи тілді өңдеуге арналған арнайы компьютерлік бағдарламалардың ішінен мәтінге белгіленім енгізуді автоматты түрде жүзеге асыратын бағдарламалар айырықша орын алады. Корпустарға белгіленім жүргізу (tagging, annotation) ең бір күрделі жұмыс деуге болады, әсіресе қазіргі кездегі корпустар көлемінің үлкендігі бұл жұмысты одан сайын күрделендіре түседі. Егер корпус мәтіндеріне морфологиялық және синтаксистік деңгейдегі белгіленімдер жүргізу үшін қазіргі кезде неше түрлі тэггерлер (taggers) және парсерлер (parsers) деп аталатын программалық құралдар қолданы-

латын болса, ал анафоралық және просодикалық деңгейлердегі белгіленімдерді жүргізетін автоматтанған жүйелерді құру әзірше мүмкін болмай, негізгі жұмыс қол күшіне жүктеліп отыр. Морфологиялық талдауды (тэггерлер) автоматтандыратын компьютерлік бағдарламалар жұмысының нәтижесінде мәтін бойындағы әрбір лексикалық бірлікке олардың грамматикалық сипаттамалалары «тіркеліп» беріледі. Мысалы, сөз табының белгісі, лемма және граммемдер (аналық-аталық тек, орта тек (бұл орыс тіліне қатысты), септік, жанды/жансыз, салт, сабақты және т.б.). Автоматтанған синтаксистік талдауға арналған компьютерлік бағдарлама жұмысының нәтижесінде сөздер мен сөзтіркестер арасындағы синтаксистік байланыстарға және синтаксистік бірліктерге (сөйлем типі, сөзтіркестердің синтаксистік қызметі және т.б.) сәйкес келетін сипаттамалар беріледі.

Бірақ табиғи тілді талдау автоматты түрде еш қатесіз жүргізіледі деуге болмайды, себебі, ол, әдетте, бір лексикалық бірліктің (сөз, сөзтіркес, сөйлем) бірнеше талдау нұсқасын ұсынады. Мұны грамматикалық омонимия деп атайды. Осындай проблемалық мәселелерді (морфологиялық, синтаксистік) жою компьютерлік лингвистика саласының ең бір маңызды және күрделі мәселелеріне жатады. Корпустарды құрастыру кезінде мұндай ауытқуларды (көпжақтылық, көпмағыналық) жою үшін жартылай автоматтанған әдіс қолданады, яғни компьютердің «қатесі» адам араласуы арқылы түзетіледі. Корпустардың жаңа буыны көлемі жүз миллионнан асатын сөзқолданыспен жұмыс істейді, сондықтан компьютер жұмысы нәтижесіне адам араласу жағын корпус жасаушылар барынша азайту қағидатын ұстаным етеді. Морфологиялық немесе синтаксистік проблемаларды автоматты жолмен шешу, әдетте, статистикалық әдісті қоса қолданатын жоғары деңгейдегі ақпаратты (синтаксистік, семантикалық) пайдалануға негізделеді.

Әртүрлі тілдік мәселелердің шешімін табу үшін мәтіндердің ауқымды көлеміне ғана сүйену жеткіліксіз. Сонымен бірге мәтіндерде айқын түрде көрсетілген әр типті қосымша лингвистикалық және экстралингвистикалық ақпараттар берілуі қажет. Мәселен, Браундық корпус тәріздес корпустар материалында сөздердің қолдану жиілігін, яғни олардың белгілі бір мәнмәтіндерде жүйелі

түрде (әрдайым) қолданылуын анықтауға болады. Бірақ бұл жиілік токендер (сөзформалар) қолданысы болады. Ал лексемалардың қолдану жиілігін анықтау үшін мәтіндегі әрбір сөзге олардың леммалары тіркеліп берілуі қажет.

Грамматикалық категориялардың жиілігін есептеу үшін мәтінде олар да белгіленуі қажет. Осындай ақпаратпен қамдалған ағылшын тілі корпусында, мысалы, зат есімнің шартты белгісі (тэгі) – *noun,* ал етістіктікің шартты белгісі (тэгі) – *verb* және т.б. Алдын ала белгілі шартты таңбалар арқылы белгіленген мұндай тэгтер сөздердің грамматикалық категорияларының категориялық сипаттамасын анықтауға мүмкіндік береді. Егер үлкен көлемді мәтіндер корпусында сөздерді тэгтермен қамтамасыз етуді қол күшімен енгізсек, ол өте көп уақытты алар еді, сондықтан зерттеушілер корпустағы мәтіндерді тэгтермен қамтамасыз етуді автоматты түрде қоюдың әдісін ойлап тапты. Оның ең бір қарапайым тәсілі – белгіленімі жоқ мәтіндер корпусын грамматикалық және лексика-семантикалық категориялары алдын ала белгіленген аса көлемді сөздік қорымен қатар қолдану әдісі. Содан кейін белгіленбеген мәтіндер корпусының әр сөзіне электрондық түрдегі алдын ала белгіленген сөздік қоры деректері негізінде автоматты түрде тэг таңбалары қойылады. Сонымен, егер *information* және *distribution* сөзформалары корпуста да, сөздікте де кездесетін болса, онда сөздіктегі сөздердің тэгтері автоматты түрде корпустағы сөздерге де ауысады (қойылады). Осы тәріздес, корпустағы *lexical* және *frequent* сөзформалары сөздік деректері бойынша сын есім таңбасына ие болады, ал *the* және *a* артикль тэгіне, *identify* және *see* – етістік тэгіне ие болады және т.б. [1]. Мұны қазақ тілі материалдары бойынша да айтуға болады. Алдын ала сөз табы қойылған түбір (негіз) сөздер сөздігін жасап, белгіленбеген мәтінге сол сөздікті пайдалану арқылы автоматты түрде мәтінде кездескен сөздерді лемматизациялауға және сол леммалардың сөз табына қатысты белгіленімін қоюға болады.

Корпус пен сөздік арасындағы сәйкестікке негізделетін бұл әдіс мәтін бірліктерінің барлық категорияларына әмбебаптық сипатта қолданыла бермейді. Бұл «морфологиялық көпмағыналық (ambiguity)» (немесе бір мағыналы еместік) мәселесіне қатысты.

Мысалы, ағылшын тіліндегі *words, forms, can, use, present* және *process* сөздер әрі зат есім, әрі етістік болуы мүмкін. Ағылшын тілінде (басқа тілдерде де) көптеген сөзформа бір ғана категорияға жата бермейтіндіктен, сөздерді корпус пен сөздік сәйкестіктеріне қарай автоматты әдіспен белгіленім жүргізу мүмкін бола бермейді, мұндай жағдайда оларды тек аса күрделі рәсімдер арқылы ғана белгіленім үдерісінен өткізуге болады. Қазақ тілінде де бұл омонимия құбылысында кездеседі. Мысалы, *қара* сөзі әрі етістік, әрі сын есім болуы мүмкін.

Әрине, сөзформа қолданыстағы контексте (мәнмәтінде) тек бір ғана категорияға қатысты екені мәлім. Сондықтан, ағылшын корпусында белгіленімнің дәлдігіне жету үшін сөзформаның қолданған контексі бойынша немесе одан да жоғары деңгейде талдау жүргізу арқылы: морфологиялық белгіленім үшін синтаксистік талдау, ал синтаксистік белгіленім үшін семантикалық талдау жүргізу қажет.

Мысал ретінде ағылшын тіліндегі *deal* сөзін қарастырайық. Бұл сөз сөзформа ретінде әрі зат есім және етістік те болуы мүмкін. Айталық, корпуста *«a good deal of trouble»* фраза кездесті дейік және бұл фразадағы *good* сөзін электронды сөздік қорымен сәйкестендіру кезінде ол сөзді сын есімге жатқызу қажет. Енді *deal* сөзін зат есімге немесе етістікке жатқызу үшін *good deal* сөз тіркесін қарастырайық. Бұл жерде ағылшын тілінде, әдетте, сын есімге жататын сөз зат есім болатын сөз алдында кездесе ме, әлде етістік алдында кездесе ме деген сауалдар туындайды. Ағылшын тілінде сын есімге қатысты сөз көпшілік жағдайда зат есім болатын сөздің алдында қолданатынын ескере отырып, *deal* сын есім сөзі *«a good deal of trouble»* фразасында қолданғанда зат есім сөз деп белгіленуі керек. Басқа сөзбен айтқанда, *good* сөзі еш даусыз сын есім болғандықтан, корпус пен сөздікті сәйкестендіру кезінде-ақ *adjective* ретінде белгіленеді. Егер белгіленім тек бір категорияға ғана жататын сөздер бойынша жүргізіліп, сосын осы ақпаратты біркелкі емес жағдайларды айқындау үшін пайдаланатын болсақ, көптеген күрделі мәселелердің шешімін табуға болар еді. Күнделікті тәжірибеде, сөздерге ең алдымен барлық сөз таптары бойынша белгіленім жүргізіледі де, ал сосын барып тіркес сөздер-

дің категориялары бірнеше белгілерге ие көпмағыналы сөздердің категорияларын анықтау үшін пайдаланылады.

Қазақ тілінде де морфологиялық белгіленімдерді автоматты түрде қою бағдарламасына омонимияны ажырататын тіркесім шарттарын салу маңызды болып табылады. Сөздердің қоршауын (дистрибуциясын) көрсететін сөздік құрастыру арқылы сөздердің контекстегі сөз табына қатысын да табуға болады. Мысалы, қазақ тілінде де сын есім негізінен зат есімдердің алдында тұрады. Ал етістіктер көбінесе сөйлем соңында тұрады. Яғни *қара* сөзінен кейін келетін сөз зат есім болса, онда *қара* сөзін сын есім деп тануға болады. *(қара көйлек, қара шаш)*. Ал *қара* сөзінің алдында барыс, табыс, шығыс, көмектес септіктерінде тұрған зат есімдер тұрса, оны етістік деп тануға болады *(үйге қара, баланы қара, дүрбімен қара, терезеден қара)*. Сонымен қатар *қара* етістік мәнінде нөлдік тұлғада қолданылғанда бұйрық райдың екенші жағында тұрып, көбінесе сөйлемді аяқтайды, яғни *қара* етістігінен кейін нүкте қойылады. Демек, пунктуациялық белгілерді де омонимдерді ажыратудағы бір тәсіл ретінде қарауға болады.

Сонымен, ***белгіленім*** жүргізу дегенді мәтін бірліктеріне алдын ала шартты түрде арнайы қабылданған тэгтерді «тіркеп» жазу әрекеті деп ұғыну қажет. Бұл жердегі **тэг** дегеніміз: лексикалық, грамматикалық және басқа да мәтін элементтеріне таза *лингвистикалық* деңгейде сипаттау үшін және сыртқы *экстралингвистикалық* (автор жайлы мәлімет және мәтін жайлы мәлімет: қалам иесі, шығарма аты, басылым жылы мен баспа орны, жанр, тақырыбы) деңгейде сипаттау үшін арнайы алынған шартты белгілер болып табылады.

Белгіленімнің лингвистикалық типі ішінде мынадай түрлерін ажыратуға болады: морфологиялық, синтаксистік, семантикалық, анафоралық, просодиялық, дискурстік және т.б. Бұлардың бәрі бірдей келесі ұстанымдарға сәйкес жүзеге асады:

1. Белгіленім сызбасын (схемасын) сипаттау (негіздеу);
2. Тілдік ұғымдардың жалпыға ортақ жүйесі;
3. Пайдаланушыға (тұтынушыға) таныс талдау сызбасы;
4. Параметрлерді енгізудің уәжділігі (мотивированность);
5. Белгіленім сызбасының теориялық бейтараптылығы (дәстүрлігі)
6. Халықаралық стандартты сақтау.

Корпустың электрондық нұсқадағы мәтіндер жиынтығынан бірден-бір айырмашылығы – мәтін өз ішінде анық түрдегі әртүрлі лингвистикалық ақпаратқа ие болуы. Осы талапқа сай корпустық лингвистикада арнайы белгіленген (размеченный) корпус идеясы пайда болды. Белгіленім (разметка) корпустың негізгі сипаттамасы деуге болады. Осындай сипаттама арқылы корпус көптеген қарапайым түрдегі мәтіндер жиынтығынан (немесе мәтіндер кітапханасынан) айырым табады. Мәтіннің грамматикалық сипатына қатысты белгі түрлерінің көп болуы корпустың ғылыми және оқыту саласындағы құндылығын арттыра түседі. Мәселен, орыс тілінің ұлттық корпусында қазіргі кезде белгілеудің үш түрі кең қолданыс табады. Біріншісі – метамәтіндік, яғни мәтінді жалпы түрде қарастыруға қатысты сипаттаулар: мәтін авторы, мәтін жанры және т.б. ерекшеліктер; екіншісі мен үшіншісі – морфологиялық және семантикалық сипаттаулар: мәтінді жалпы түрде қарастырмай, оның әрбір сөзіне қатысты грамматикалық ерекшеліктерін шартты белгілер арқылы белгілеу.

Шындығында, статистикалық санақ деңгейінің өзінде, егер мәтіндегі сөздерде оның қай сөз табына жататындығы жайында ақпарат болса, онда тек сөздердің қолдану жиілігін ғана біліп қоймай, қайсыбір сөз табының да жиілігін анықтауға мүмкіндік туады. Лингвистикалық белгіленімді мәтін сөздеріне сәйкестендіретін арнайы түрдегі шартты белгі-код қою деп түсіну керек. Әрбір белгі-кодқа сол сөзді сипаттайтын грамматикалық белгілердің топтамасы сай келеді. Мұндай белгі-кодтар «тэги» сөзі (ағыл. tag – ярлық, таңба) арқылы оқырманға белгілі, ал сөздерге белгі-код қою әрекеті «тэггинг» (ағыл. tagging) сөзіне сәйкес келеді.

Қазіргі кезде көпшілік мақұлдаған мәтіндерге тілдік немесе басқа да ақпарат түрлерінің белгіленімін қою стандарттары жоқ.

Text Encoding Initiative (TEI) атты арнайы халықаралық жоба осындай стандартталған құралдарды зерттеуге арналған. Ол үшін көпшілік мақұлдаған белгіленім қоюдың халықаралық тілі SGML немесе оның бір түрі ретіндегі XML қолданылады. Корпусқа қойылатын белгіленім түрлерін шартты түрде лингвистикалық және одан сырт түрі деп екіге бөліп қарастыруға болады. Белгіленімнің лингвистикалықтан сырт түріне мыналар жатады:

1) мәтінді форматтау ерекшеліктерін бейнелейтін белгіленім (тақырыптар, абзацтар, бос жер және т.б.);

2) автор мен мәтінге қатысты мәліметтерді бейнелейтін белгіленім. Сонымен бірге автор жайлы мәлімет тек оның аты-жөні ғана емес, оның жасын, жынысын, өмір сүрген жылдарын және т.б. болуы мүмкін. Ал мәтін туралы мәлімет, әдетте, тақырыптан басқа оның қай тілде жазылғаны, жылы, баспа орны, аты және т.б. Мұндай ақпараттың корпуста орын алуы мәтіндер қорынан іздеу әрекетін тәптіштеп іздестіруге мүмкіндік жасайды және сонымен бірге олар тиісті құжатты теңестіру әрекеті үшін де қажетті құрал бола алады. Мұны метабелгіленім деп те атайды.

2.4.1. Морфологиялық белгіленім

Шетелдік терминологияда part-of-speech tagging (POS-tagging) қолданылады, сөзбе-сөз аударсақ, орысшасы – частеречная разметка, ал қазақшасы – сөз таптары белгіленімі.

Шындығына келгенде, морфологиялық белгілер тек сөз таптарын ғана анықтап қоймайды, ол сонымен бірге сол сөз табына ғана тән грамматикалық категорияларды да қамтиды. Бұл белгіленімдер типінің негізгісі болып саналады. Себебі, біріншіден, көлемді корпустардың көпшілігі морфологиялық белгіленім жүргізілген корпустар, екіншіден, морфологиялық талдау әрі қарайғы талдау формаларына – синтаксистік пен семантикалық белгіленімдер жүргізуге негіз болып саналады және үшіншіден, компьютерлік морфологияның жетістіктері үлкен дәлдікпен көлемді корпустарға автоматты түрде белгіленімді жүзеге асыруға мүмкіндік жасайды.

Белгіленім туралы деректер белгілі құрылымда көрініс табады және олардың түрлері мынадай: лемма, сөз табы белгісі, грамматикалық категориялар белгісі. 1980 жылы Браун корпусының белгіленімінен өткен нұсқасы өмірге келді және бұл корпуста сөзформалар лемма түріне келтіріліп, үстірт-синтаксистік функцияларға (поверхностно-синтаксическая функция) маркалау және т.б. жүзеге асқан. Браун корпусының морфологиялық белгіленімі төмендегідей көрініс тапқан:

the_AT jury_NN further_RB said_VBD in_IN term-end_NN presentments_NNS that_CS the_AT *city_NP *executive_NP *committee_NP ,_, which_WDT had_HVD over-all_JJ charge_NN of_IN the_AT election_NN ,_, deserves_VBZ the_AT praise_NN and_CC thanks_NNS of_IN the_AT *city_NP of_NP *atlanta_NP for_IN the_AT manner_NN in_IN which_WDT the_AT election_NN was_BEDZ conducted_VBN |

Мысал ретінде орыс тіліндегі «*Звонили к вечерне. Торжественный гул колоколов*» мәтін үзіндісін және оған АОТ белгілегіш негізінде XML-форматында қойылған морфологиялық белгіленімді қарастырайық (*2-сурет)*.

2-суретте мынадай тэгтер көрініс тапты: – текст (мәтін), – абзац (абзац), – предложение (сөйлем), – словоупотребление (сөзқолданыс), – знак пунктуации (тыныс белгі). Аталған тэгтер ішіне тағы да ішкі тэгтер салынған және олардың атрибуттары: – лемма (лемма), – часть речи (сөз табы), – набор граммем (граммем жиынтығы).

```
<?xml version=»1.0» encoding=»windows-1251» ?> <text> <p>
<s> <w>Звонили<ana lemma=»ЗВОНИТЬ» pos=»Г»
gram=»мн,нс,нп,дст,прш,» /></w> <w>к<ana lemma=»К»
pos=»ПРЕДЛ» gram=»» /></w> <w>вечерне <ana
lemma=»ВЕЧЕРНЯ» pos=»С» gram=»жр,ед,дт,пр,но,» /> <ana
lemma=»ВЕЧЕРНИЙ» pos=»П» gram=»ср,ед,кр,» /></w>
<pun>.</pun> </s>
<s><w>Торжественный<ana lemma=»ТОРЖЕСТВЕННЫЙ»
pos=»П» gram=»мр,ед,им,вн,» /></w> <w>гул<ana lemma=»ГУЛ»
pos=»С» gram=»мр,ед,им,вн,но,» /></w> <w>колоколов <ana
lemma=»КОЛОКОЛ» pos=»С» gram=»мр,мн,рд,но,» /> <ana
lemma=»КОЛОКОЛОВ» pos=»С» gram=»мр,фам,ед,им,од,» /></
w> .........................<pun>.</pun> </s></p></text>
```

2-сурет. Орыс тіліндегі мәтін үзіндісіне қойылған морфологиялық белгіленімге мысал

Морфологиялық белгіленім сызбасының шарты бойынша, біріншіден, тэгтердің (таңбалардың) топтамасының болуын, екіншіден, олардың әрқайсысының сипаттамасының болуын, үшіншіден, мәтін бірліктеріне тэгтерді (таңбаларды) сәйкестендіру ережесі болуын қажет етеді. Әртүрлі корпустарда қолданылатын таңбалар (тэгтер) топтамысының көлемі құбылып тұрады. Әрине, таңбалар топтамасы неғұрлым көп болса, соғұрлым оның көмегімен мәтінге жасалатын талдау да толық сипатта болады. Бірақ мәтіндер корпусының көлемі ұлғайған сайын морфологиялық таңбалар санының қысқару үрдісі байқалады. Белгі-код қоюдың ықшамдалған жүйесін қолдану артық қате жібермеуге, ретсіздікті, морфологиялық біріздiлікті бұзбауға және миллиондаған сөзден тұратын мәтінге белгіленім жүргізу үрдісін тездетуге мүмкіндік туғызады.

Орыс тіл білімінде морфологиялық белгілеу үш кезең бойынша іске асырылады: анализ-талдау (парсинг); сұрыптау; омонимдерді ажырату. Мұндағы бірінші кезеңде жасалатын лингвистикалық (морфологиялық) белгілеу әдісі MYSTEM программасы бойынша жүзеге асады. Бұл программа А.А.Зализняктың «Орыс тілінің грамматикалық сөздігіндегі» мәліметтерге негізделген және мұнда әрбір сөзформаның морфологиялық белгілерінің мүмкін нұсқалары ирек жақша ішіне жазылады.

Бірінші кезеңде сөзформаның мүмкін варианттары автоматты түрде түгелдей қамтылса, екінші кезеңде омоним сөзформалар анықталып, олардың тұсына сұрақ белгілері қойылып отырады. Бұл да автоматты түрде GRAMBAT программасына салынады. Үшінші кезеңде анықталған омонимдер, яғни бірдей сөзформалардың (ол сөз не сөздің формасы, яғни грамматикалық формалар болуы мүмкін) лексикалық не грамматикалық мағыналары мәтіндерден іздестіріліп, қай сөз табы екендігі автоматты түрде емес, қол жұмысы арқылы жүзеге асырылады. Ал омонимдерді автоматты түрде ажырататын алгоритм ойлап табу немесе осы мәселенің шешімін тапқан тілдер болса тәжірибе алмасу қажеттігі дау тудырмайды.

Қазақ тілі корпусы бойынша автоматты морфологиялық белгіленім қойылған мәтіннен үзінді берейік:

«XIX ғасырда мәдениет мәселелерімен тығыз айналысқан ғалымдардың бірі, ағылшын ғұламасы Эдуард Бернетт Тайлор тұжырым жасаған» («Айқын»).

Осы сөйлемді автоматты морфологиялық белгіленім қою бағдарламасы төмендегіше талдау жасайды:

XIX/ XIX// [Lat]
ғасырда/ғасыр/да/ [ЗТ/да/ЖС]
мәдениет/мәдениет [зт]
мәселелерімен / мәселе/лерімен/ [зт/лер/КЖ+і/ТЖ-3+мен/КС]
тығыз/тығыз// [сн]
айналысқан / айналыс/қан [ЕТ/қан/Е+АР+ЖЖ-3]
ғалымдардың / ғалым/дардың/ [ЗТ/дарКЖ+дың/ІС]
бірі / бір/і [ЕС/і/ТЖ-3+АС]
бірі/ бір/і [СА/іТЖ-3+АС]
ағылшын/ағылшын [зт]
ғұламасы / ғұлама/сы [СН/сы/ТЖ-3+АС]
Эдуард/Unknown
Бернетт/ Unknown
Тайлор/ Unknown
тұжырым /тұжырым/ [зт]
жасаған/жаса/ған/ [ЕТ/ған/Е+АР+ЖЖ-3]

Мұндағы Unknown – деген табылмаған сөз. Табылмаған сөз ол бұл мәтінде жалқы есім болып тұр. Негізінде жалқы есімдер тізімі корпус жадына салынса, жадта бар жалқы есімдер туралы ақпарат берілген болар еді. Ал қалған сөздердің бәрін корпус бағдарламасы автоматты түрде лемматицация жасап, сонымен қатар олардың сөз табы мен белгілі бір категориялық мағыналарына сәйкес шартты белгілерін қойып беріп отыр.

Ізделген сөзге қажетті ақпаратты бір жерде жинақтап беруге де болады, яғни сөзге қатысты тек морфологиялық белгіленім ғана емес, лексикалық сипаттама, семантикалық белгіленімін де көрсету мүмкіндігі бар. Мысалы, институттың А.Жұбанов, А.Жаңабекова, А.Фазылжанова және т.б. ғалымдар жасаған қазіргі 7 миллион сөзқолданыстан тұратын Қазақ тілінің ұлттық корпусының қанатқақты жобасы бойынша «бала» сөзіне жүздеген мысал конкардансы шығады. Ал лингвистикалық белгіленім/сипаттама мына түрде көрініс тапқан.

балаға

	Сипаттау
Лемма	бала
Морф. белгіленім	ЗТ|ға/БС
Лексик. семантика	«1. Ата-ананың перзенті, ұрпақ; тұқым. 2. Нәресте, сәби, бөбек. 3. Жан-жануарлардың күшігі, құстардың балапаны 4. а у ы с. Ер жетіп ысылмаған; жас, балғын. « «1. Ата-ананың перзенті, ұрпақ; тұқым. 2. Нәресте, сәби, бөбек. 3. Жан-жануарлардың күшігі, құстардың балапаны 4. а у ы с. Ер жетіп ысылмаған; жас, балғын. «
Морфо-семант белгіленім	жалпы, дара, түбір, деректі

Демек, жоғарыда «балаға» сөзі алдымен түбірі көрсетіліп, лемматизация жасалған. Екіншіден, морфологиялық белгіленімі берілген, яғни зат есім сөз табына жататындығы және жалғанған қосымшаның барыс септік жалғауы екендігі сипатталған. Үшіншіден, лексикалық семантикасы ретінде 4 түрлі мағынасы сипатталған. Төртіншіден, түбірдің морфосемантикалық сипаттамасы берілген, яғни жалпы есім екендігі, құрамына қарай дара зат есім, жасалымына қарай түбір екендігі, сонымен қатар деректі зат атауы екендігі сипатталған.

2.4.2. Корпустағы морфологиялық белгіленімдердің тілдік талдаулармен сабақтастығы

Тілдік корпустардың мәтінтүзуші бірліктеріне берілетін лингвистикалық ақпараттар жүйесі тіл ғылымының әрбір саласы бойынша айқындалған тілдік теорияларға негізделеді. Тілдік ақпараттар тіл дамуының бүгінгі қалып-күйіне сәйкес алынады. Яғни соңғы тілдік талдауларды басшылыққа алады. Сонымен қатар электронды пішіндегі тілдік корпустар тілдік ақпараттар-

ды мейлінше нақты, шартты белгіленімдер арқылы қысқа беруді мақсат етеді.

Тілдік корпустарға енген лингвистикалық белгіленімдер тіл оқырмандары үшін ақпараттық құрал қызметін ғана атқармайды, сонымен қатар анықтамалық дереккөз де болып табылады. Дәлірек айтқанда, тілші-оқырман тілдік корпустар арқылы қандай да бір тілдік бірлік туралы мәлімет алып қана қоймайды, оны зерттеу, ізденістеріне қажетті бірден бір дереккөз ретінде де пайдалана алады. Бұрын ғылыми зерттеулер мысал жинау, кәртішке жазу, жинақтау, топтау сияқты қол жұмыстары арқылы атқарылып, ұзақ уақытқа созылған ғылыми ізденісті қажет еткен болса, тілдік корпустар мұндай жұмыстарды барынша жеңілдетеді. Айталық, бұрынғы ғылыми тақырыптарға мысал жинақтау үшін қаншама жазушылар шығармаларынан, газет-журналдардан мысал жинау жұмысы айлап, жылдап жүргізілетін. Лингвистикалық ақпарат салынған тілдік корпустар жасалатын болса, мұндай материал жинақтау ісін өте тез арада атқаруға болады. Электоронды корпустар тілдегі (мәтіндегі) лексикалық бірліктер туралы ақпараттар беріп қана қоймайды, оның әрбір құрамдық бөлшегі туралы да мәлімет бере алады. Дәлірек айтсақ, тілдегі әрбір сөзге тән лексика-семантикалық, грамматикалық (морфологиялық, синтаксистік), тіпті фонетикалық ерекшеліктерді де бейнелей алады. Тілдік корпустардың ақпараттық мүмкіндігі оны басқарушы, жасаушы тілші-мамандар мен инженер-программистердің ғылыми-шығармашылық шеберлігіне байланысты болады. Өйткені тілдік корпустарды жасауда тілші-мамандар әрбір мәтін бірлігінің тілдік сипатын белгілеу жағымен айналысса, инженер-программист компьютерлік бағдарламалар арқылы оны басқарудың математикалық алгоритімін құрастырады. Элекронды пішінге келтірілген тілдік корпустарды осындай ғылымдар тоғысында сала мамандарының бірлескен шығармашылық жұмыстары негізінде ғана жасап шығуға болады.

Морфология – сөздердің құрылысын зерттейтін ғылым. Морфология үшін сөз бөлшектерінің лексикалық мағынаны білдіру-білдірмеуі маңызды емес, өйткені ол тіл бірліктерінің барлығын формалды құрылымы бойынша қарастырады. Мазмұн межесі тұрғысынан грамматикалық мағыналарды ғана нысанға алады. Мә-

селен, сөздердің заттық, сындық, сандық мағыналары да грамматикалық мағынаға жатады. Тілші-ғалым С.Исаев бұл мағыналарды жалпы грамматикалық мағына деп атаған болатын [28]. Морфологиялық белгіленім қоюда осы жалпы грамматикалық мағыналары арқылы топтастырылатын *сөз таптарын* дәл айқындау маңызды болып табылады. Өйткені сөздердің толық морфологиялық бейнесі сөздердің жалпы грамматикалық мағыналарына, яғни сөз таптарына тікелей қатысты. Сөз таптарына қатысты белгіленім дұрыс қойылмаса, одан кейінгі грамматикалық формалардың белгіленімі де қате құрылады.

Тілдік корпустарға енгізілетін мәтіндерге қойылатын морфологиялық белгіленімдер осы сала бойынша жасалатын тілдік талдаулар жүйесіне негізделеді. Өйткені тілдік корпустар тіл теориясының объективті көрінісі болып табылады. Демек, морфологиялық белгіленімдер тілдің морфологиялық жүйесін толық бейнелеуі тиіс. Осымен байланысты тілді оқытуда қолданылып жүрген тілдік талдаулар мен электронды пішінге келтірілген тілдік белгіленімдер бір-бірімен сабақтасып жатады.

Әдетте, қазақ тілі сабақтарында, оқулықтарда морфологиялық талдаулар сөздердің сөз таптарына қатысына, тұлғасына, құрамы мен құрылысына, мағынасына қарай жасалады. Корпустарда да морфологиялық белгіленімдер осы критерийлерді ұстанады. Бірақ электронды корпустарда бұл морфологиялық ақпараттар шартты белгілер арқылы кодталып, шартты белгімен белгіленіп ықшамдалып беріледі. Өйткені, тілдік бірліктің грамматикалық бейнесін оқулықтардағыдай түсіндіріп жату мүмкін емес.

Бүгінгі оқыту жүйесінде тілдік талдаулар Н.А.Оразахынованың жалпы білім беретін мектепке арналған қазақ тілін сатылай кешенді талдау технологиясы бойынша оқытылып жүр. [29]. Мұндай морфологиялық талданымда сөздердің құрамы, тұлғасына қарай негізгі-туынды, құрылысына қарай дара не күрделі, жалпы грамматикалық мағынасына қарай сөз табына қатысы, екінші категориялық мағынасына қарай грамматикалық түрленім мағыналары, мағынасына қарай жалпы-жалқы, жанды-жансыз, деректі-дерексіз т.б. белгілері бойынша сипатталып, талдау жасалады. Тілдік корпустар осындай морфологиялық білімдер жүйесін бейнелеуде де осы тілдік жүйені негізге алады.

Мысалы, М.Әуезов шығармаларынан алынған мәтін бірліктеріне қолдап енгізілген нұсқа бойынша, морфологиялық белгіленім моделі төмендегіше беріледі: *Абай Қарқаралыдан қайтқан соң атқа мініп ел аралаған жоқ.*

Абай {**Абай** – зат: жалқы, дара, түбір, деректі + Ø жекеше + Ø септік, атау}

Қарқаралыдан {**Қарқаралы** – зат: жалқы, дара, туынды, деректі + Ø жекеше + септік, шығыс}

қайтқан {**қайт** – етістік: дара, түбір, негізгі, салт + есімше + Ø рай, ашық + Ø жіктік, 3-жақ, Ø жекеше}

соң {**соң** – шылау: дара, түбір, септеулік}

атқа {**ат** – зат: жалпы, дара, түбір, деректі + Ø жекеше + септік, барыс}

мінген {**мін** – етістік: дара, түбір, негізгі, салт + есімше + Ø рай, ашық + Ø жіктік, 3-жақ, Ø жекеше}

ел {**ел** – зат: жалпы, дара, түбір, деректі + Ø жекеше + Ø септік, табыс}

аралаған {**арала** – етістік: дара, туынды, негізгі, сабақты + есімше + Ø рай, ашық + Ø жіктік, 3-жақ, Ø жекеше}

жоқ {**жоқ** – әртарап сөз + Ø жіктік, 3-жақ + Ø жекеше}

Мұндай талдау үлгісі талдауға кететін уақыт шығыны жағынан да өте тиімді. Оқушы жазылым бойынша жасайтын талдауға кететін уақыт ішінде бір сөйлемде берілген сөздердің әрқайсысына морфологиялық талдау жасай алады. Мұндайда морфологиялық белгіленімге қатысты қойылатын барлық белгіленім түрлері, модельдері компьютер жадына енгізіледі. Оқушы сол жадтағы белгіленімдердің өзіне берілген мәтінге қажетті моделін таңдап алып, автоматты түрде қойып отыруы қажет. Дайын модельдерді тауып қоюдың өзі де оқушылардың аялық біліміне байланысты көрініс тауып отырады. Демек, автоматты жүйенің өзі адамның басқаруы арқылы жүзеге асады.

Оқыту жүйесіне корпустық талдауларды енгізу талдау үлгісінің екінші деңгейіне жатады. Жазылым, айтылым дағдылары бойынша талдау жасап үйренгеннен кейін оқушыларды компьютер арқылы белгіленім қоюға үйретуге болады. Бұл оқушыларды жаттықтырып, білімдерін жетілдіріп, ұштай түсуіне мүмкіндік

береді. Көру қызметі (көрілім) арқылы оқушы берілген сөздердің морфологиялық бейнесін жадта (зрительный память) қабылдап алады әрі сөздерге қойылған морфологиялық белгіленімнің дұрыс-бұрыстығын бір көргеннен айқындай алады. Корпустық белгіленімдер бойынша оқушы дағдысын қалыптастыру келешекте оқыту тәжірибесіне енгізілуі қажет.

Автоматты талдаулар оқытушының жазба жұмыстарын тексеру ісін жеңілдету үшін де тиімді болып табылады. Бұл жағдайда морфологиялық белгіленімдер жасалған мәтіндер алдын ала дайындалады. Ондай мәтінге қойылған белгіленімдер негізінде оқушылардың талдау жұмыстарын тексеріп, дұрыс-бұрыстығын көрсете алатын алгоритмдер құрастырылады. Бұл бағдарламадан кейін компьютер қате варианттарды бірден айқындайды. Қысқасы, тілдік белгіленімдер енгізілген мәтіндер корпустары тілүйренуші үшін ғана емес, тілүйретуші үшін де тиімді әдіс.

2.4.3. Зат есімнің лексика-морфологиялық құрылымына ақпараттық белгі-кодты сәйкестендіру бағдарламасы

Зат есім сөздер өздерінің лексика-семантикалық сипаттарына қарай арнайы жалғаулар арқылы түрленіп және жұрнақтар арқылы тұлғалық өзгеріске ұшырап, сөйлемдегі өзге сөздермен еркін қарым-қатынасқа түсіп отыратыны белгілі. Зат есім сөздердің морфологиялық, синтаксистік және семантикалық белгілерін ұштастыра қарағанда ғана олардың сыр-сипатын толық ашуға болады. Сол белгілер қатарына олардың мәтін ішінде жиі не сирек қолдану сипаты да жатады.

Қазіргі қазақ тілінің грамматикасында зат есім сөздерге ғана тән семантикалық және грамматикалық ерекшеліктері бар мынадай топтар бөліп қарастырылады: адамзат (кімдік) және ғаламзат (нелік) есімдер, жалқы/жалпы есімдер, көптік мәнді есімдер, эмоциялы-экспрессивтік зат есімдер, көмекші есімдер. Зат есім сөздерді олардың морфологиялық сипаты, яғни құрылымдық, жасалу тәсілі және түрлену жүйесі арқылы да топтап қарастыруға болатыны да белгілі. Зат есім сөздерді топ-топқа бөліп қарастыру қазақ тілін компьютерлік және статистикалық лингвистика әдістерімен зерттеуде аса маңызды деуге болады [88, 114-117].

Қазақ тіліндегі мәтіндерде ең жиі қолданыс табатын сөз таптары *зат есім* мен *етістік* болса, олардың мәтін ішінде морфологиялық, синтаксистік жолмен түрленіп қолдануының да статистикасын танып білу «Қолданбалы лингвистика» пәнінің мақсаттарының бірі. Біздің негізгі ұстаным – мәтін ішінде кездесетін *сөз*, *сөзтұлға* не *тұрақты сөзтіркеске*, яғни мәтін бірлігі деп аталатын – *сөзқолданысқа* сәйкес келетін грамматикалық ақпараттар *орны («ұясы»)* мен сол ақпаратты танытатын шартты *белгі-кодты* тұрақтандыру. Қандай да болмасын тілдік бірліктің, мысалы, ең алдымен қарастырайын деп отырған – *зат есім* сөздердің әрбіреуіне сәйкестендірілетін шартты параметрлерді («ұя» орны мен белгі-код) алдын ала жан-жақты ойластырып алғанды жөн санаймыз.

Айталық, *зат есім* сөздерге сәйкестендіретін (тиісті) «ұя» саны немесе топтама саны – сегізге тең делік, яғни n=1, 2, 3, 4, 5, 6, 7, 8. Ал әрбір «ұяда» жазылатын шартты *белгі-код* саны (топтама саны) мен «код мәні» *зат есім* сөздердің сол «ұяда» қарастырылатын сипатына қарай өзгеріп отыруы мүмкін [88, 114-117].

Егер осы айтылғандарды қамтитын мәліметтер кестесін *«бағдарлама»* деп атасақ, *зат есім* сөз табына қатысты қысқаша топтама 8-кестедегі *«бағдарлама»* түрінде көрініс табады.

8-кесте

Зат есім сөздерге лексика-морфологиялық ақпаратты сәйкестендіру бағдарламасы

Белгі-код орны	Зат есім сөздердің лексика-морфологиялық сипаты	Белгі-код
1	2	3
1	Зат есім сөз (сөзтұлға)	ЗТ/
2	Адамзат есімдері *(кім?)*	020
	Ғаламзат есімдері *(не?)*	021
3	Жалқы есім: а) кісі аттары (ономастика); ә) географиялық атаулар (топонимика)	030
	Жалпы есім	031
	Аралас мағыналы *(ай, күн, жер)*	032

1	2	3
4	Көптік мәнді емес (жалғаусыз қарастырғанда)	040
	Табиғи қос мәнді *(аяқ, құлақ, етік)*	041
	Көптік мәнді (жалғаусыз)	042
5	Реңсіз зат есім	050
	Эмоциялы-экспрессивтік ренді зат есім *(Әкей, Сәулеш)*	051
6	Көмекші зат есімге жатпайтын зат есім	060
	Көмекші зат есім *(алды, арты, қасы)*	061
7	Түбір морфема зат есім	070
	Туынды зат есім (түбір+жұрнақ)	071
	Күрделі зат есім: а) біріккен сөз	072
	ә) қосарланған сөз	073
	б) құрама сөз	074
	в) қысқарған сөз	075
8	Зат есімнің түрленуі:	
	Зат есімнің жалғаусыз түрі *(кім? не?)*	080
	Көптік жалғаулармен*: -лар, -лер,-дар,-дер,-тар,-тер*	081
	Тәуелдік жалғаулар арқылы: *-нікі,-дікі,-тікі -м,-ым, -ім; -ң, -ың, -ің, -ңыз, -ңіз, -ыңыз, - іңіз; -сы, -сі, -ы,-і*	082
	Септік жалғаулар арқылы: ілік *(кімнің? ненің?);* барыс *(кімге? неге? қайда?);* табыс *(кімді? нені?);* жатыс *(кімде? неде?);* шығыс *(кімнен? неден?);* көмектес *(кіммен? немен?).*	083
	Жіктік жалғау арқылы: жекеше*: -мын, -мін, -сың, -сің, -сыз, -сіз;* көпше*: -мыз, -міз.*	084

Топтамадағы мәліметтердің кейбіреулеріне түсініктеме берейік.

Аталған кесте үш «бағанадан» тұрады:

1) *шартты белгілер орны* (топтама саны);

2) *зат есім сөздердің лексика-морфологиялық сипаты*;

3) *шартты белгі-код.*

Әрбір сөз табына қатысты «ұялар» топтамасының ең алғашқысы сол сөз табының қысқарған атымен белгіленеді (кодталады)

және *1-ші орынға* ие болады деп ұйғару керек. Сонымен, «ұялар» тізбегінің ең алғашқысы (бірінші бағана) «1» санына тең (бірінші топтама), осы жатық жолдағы екінші бағана сөз табының сипаты жайлы мәлімет, яғни – *зат есім* атауы және үшінші бағана – мәліметтің «белгі-код» мәні – «ЗТ». Сол сияқты шартты белгілер орнының «2» санында, яғни екінші топтамасында *зат есім* сөздердің екі түрлі сипаттамасы берілген. Олардың біріншісі – *адамзат есімдері (кім?)* және екіншісі – *галамзат есімдері (не?)*. Шартты белгі орнында, яғни 2-ші орында қарастырылатын кезекті *зат есім* сөздің қай сұраққа *(кім? не?)* жауап беруіне қарай үшінші бағанадағы оған сәйкес келетін белгі-кодтың біреуі ғана таңдалып алынады. 8-кестеде зат есім сөздерге қатысты осы тәріздес 8 топтама қарастырылды.

Сол топтамалардың енді 7-ші орынға қатыстысын сөз етсек, жетінші топтамада *зат есім* сөздің алты түрлі сипаты аталған: *1. Түбір морфема зат есім; 2. Туынды зат есім (түбір+жұрнақ); 3. Күрделі зат есім: а) біріккен сөз; ә) қосарланған сөз; б) құрама сөз; в) қысқарған сөз.*

Үшінші бағанада олардың белгі-код мәндері үш орынды сандар: 070, 071, 072, 073, 074, 075 сандары арқылы берілген (8-кестені қараңыз).

Жоғарыда 2-ші орынға байланысты айтылған сияқты, бұл жолы да зерттеуге алынған кезекті сөзді (сөзтұлғаны) шартты түрде «кодтау» кезінде 7-ші орынға сәйкес келетін (070 пен 075 аралығы) алты белгі-код мәндерінің ішінен сөзтұлға сипатына тән келетін біреуі ғана алынады. Кестеде келтірілген басқа топтама орындарына да осылайша түсініктеме беруге болады.

Енді мысал ретінде кейбір зат есім сөздерді 8-кесте деректері бойынша шартты белгі-код түрінде жазып көрейік. Мысалы, мына сөйлем ішіндегі зат есім сөздерді бөліп алып, белгі-код арқылы жазу керек: *Асан мектепке барарда кітап, қалам, дәптер және басқа да керек-жарақтарын аса ұқыптылықпен сөмкесіне салды.* Бұл сөйлемдегі зат есім сөзтұлғалар мыналар: *Асан; мектепке; кітап; қалам; дәптер; керек-жарақтарын; сөмкесіне.*

Шартты белгі-кодтарды осы сөздерге сәйкестендірейік. Яғни әрбір аталған *зат есім* сөздердің сегіз орындық «ұяларына» тиісті

белгі-код мәнін қойып шығу керек деген сөз. Олардың барлығының да бірінші орындарында «ЗТ/» белгі-код тұрады.

Енді *Асан* сөзін бөлек алып қарастырсақ, 2-ші орында – «020» (адам аты болғандықтан), үшіншіде – «030», төртіншіде – «040», бесіншіде – «050», алтыншыда – «060», жетіншіде – «070», сегізіншіде – «080». Сонымен, сөйлемдегі *Асан* сөзі мен басқа да зат есім сөздердің «кодтық» жазылуы төмендегідей болады:

Асан:	ЗТ/	020	030	040	050	060	070	080
мектепке:	ЗТ/	021	031	040	050	060	070	083
кітап:	ЗТ/	021	031	040	050	060	070	080
қалам:	ЗТ/	021	031	040	050	060	070	080
дәптер:	ЗТ/	021	031	040	050	060	070	080
керек-жарақтарын:	ЗТ/	021	031	042	050	060	073	084
сөмкесіне:	ЗТ/	021	031	040	050	060	070	083

Көрсетілген жолмен мына зат есімдерге қатысты сөзтұлғаларды белгі-кодпен сәйкестендірейік: *ғалым, Асандар, құлағым, Алматыға, жазушы, саңырау құлақтан, жаным-ай, жүрісіме.*

ғалым	ЗТ/	020	032	040	050	060	070	080
Асандар	ЗТ/	020	030	040	050	060	070	081
құлағым	ЗТ/	021	032	041	050	060	070	082
Алматыға	ЗТ/	021	030	040	050	060	070	083
жазушы	ЗТ/	020	032	040	050	060	071	080
саңырау құлақтан	ЗТ/	021	032	041	050	060	074	083
жаным-ай	ЗТ/	021	032	040	051	060	070	082
жүрісіме	ЗТ/	021	032	042	050	060	071	083

Белгілі бір мәтін сөзтұлғаларын белгі-код түріне ауыстырып барып, жиілік сөздіктер түзуге болады немесе әріп күйіндегі мәтіннің өзінен түзілген әліпби-жиілік, жиілік не кері-әліпби сөздіктердегі сөзтұлғаларды (бірліктерді) белгі-код түріне ауысты-

рып жазуға да болады. Егер ондай сөздіктерді тиісті компьютерлік бағдарлама бойынша «қысылған» түрге келтіріп жазатын болсақ, зат есім (не басқа сөз табы) сөзтұлғалардың «сипаттық жиілік сөздігін» алуға болады. Бұндай жиілік сөздік қазақ тілінің морфологиялық құрылымын зерттеу нысаны ретінде және әрі қарайғы мәтін лингвистикасына қажетті маңызды материал болар еді. Корпус мәтіндерін аннотациялауда да осындай цифрлық кодтар арқылы белгіленім қою мүмкіндігі бар.

Сонымен қатар осы аталған зат есімнің кестеде көрсетілген топтарға сандық белгі код қана емес, халықаралық шартты белгілермен кодтауға да болады. Ол белгі кодтар корпуста сөздердің семантикалық топқа қатысын сипаттайтын болады.

2.4.4. Етістіктің лексика-морфологиялық құрылымына ақпараттық белгі-кодты сәйкестендіру бағдарламасы

Етістік сөздер құрамының басқа сөз таптарына қарағанда күрделілігі оның аса өрісті лексика-семантикалық сипатымен, өте бай лексика-грамматикалық формаларымен және кең синтаксистік қызметімен тығыз байланысты. Етістіктің лексика-грамматикалық тұлғаларының бай болу себебі – олар әрекеттің болу мезгілін, жүзеге асу кезеңін, өту сипатын, яғни әрекеттің бағыты, қарқыны, тынуы тәрізді жайларды түгел қамтиды. Осымен қатар басқа сөз табына қатысты сөздерден етістік тудыратын синтетикалық, аналитикалық тәсілдер жүйелерін қоса есептесек, етістіктің тұлғалану байлығына жететін сөз табы жоқ деуге де болады. Осы айтылғандар оның өзіне ғана тән лексика-семантикалық, лексика-грамматикалық және грамматикалық категорияларынан айқын көрініп жатады. Етістіктің осындай әрқилы категорияларының мағыналары да, синтаксистік қызметтері де өзге сөз таптарына тән сөздермен қарым-қатынасқа түскенде айқындала түседі.

Етістіктің лексикалық жүйесіндегі әрбір сөздің өзіне ғана тән лексикалық мағыналары бөлек болғанымен, семантикалық ерекшеліктеріне қарай бір сөз тобына телінеді. Етістіктің грамматикалық тұлғалары мен қызметтерін дұрыс айқындау үшін, оларды

түбір формалардың, туынды синтетикалық және аналитикалық формалардың семантикалық құрылымына орай жіктеп, топтап барып сыр-сипатын ашу, олардың тұлғалық және мазмұндық арақатынастарын айқындауға әкеледі. Себебі грамматикалық семантика сөздің құрамына көп байланысты болып келеді.

Сол себептен етістік формаларының морфологиялық құрылымының сыр-сипатын ашу контекстік тұрғыда немесе мәтін лингвистикасы саласының зерттеу аясына жақынырақ келеді. Ал бұл тұрғыдан зерттеуді ұйымдастыру дегеніміз – статистикалық лингвистика әдістерін молынан қолдану деген сөз. Туынды етістіктердің семантикалық құрылымының аса күрделі және түрлену сипатының бай болу себептерінің бірі – олардың сирек не жиі қолдануында деп түсіну қажет. Сондықтан туынды етістіктердің морфологиялық құрылымын жеке бір жүйе ретінде қарастырып, әрбір өзіне тән сипатына қарай топтап, олардың ықтималды-статистикалық заңдылықтарын ашу үшін арнайы белгі-код жүйесін енгізіп зерттеу қажет деп санаймыз. Қазақ тілінің тілдік бірліктерін, сөз таптарын формальды белгілері арқылы танып білу және соның нәтижесінде автоматты түрде жазба не сөйлеу мәтінінен оларды бөліп алу әлі де іске аспай отыр. «Қолдан» енгізген шартты «белгі-код» негізінде етістіктің морфологиялық құрылымына статистикалық зерттеулер жүргізу көптеген мәселелердің басын ашатыны белгілі. Сол мәселелердің ішіне мәтіндік бірліктерді автоматты түрде ажырата білу мен олады бөліктеп алу мүмкіндігінің де қатынасы бар. Сонымен бірге етістіктің лексика-грамматикалық құрылымының дәстүрлі әдіспен зерттеуде байқала бермейтін тұстарының да басы ашылуы (анықталуы) мүмкін. Осы тұрғыда қарастыра отырып, біз төменде қазіргі қазақ тіліндегі етістік (түбір және туынды) сөздерді іштей бірбіріне мағына жағынан жақындықтарына, өзара функция жағынан орайластықтарына және лексика-морфологиялық құрылымына қарай топ-топқа бөліп, олардың сипаты жайлы ақпараттарды арнайы ойластырылған шартты белгі-кодтарға сәйкестендіру бағдарламасын беріп отырмыз [88, 117-122].

Мысалы, мағыналық тұрғыдан алып, етістік сөздерді іштей мына топтарға бөлуге болады [30, 384 б.]: *амал-әрекет, қимыл-қозга-*

лыс, қалып-сапа, ойлау-сөйлеу, өсу-өну етістіктері және т.б. 9-кестеде көрсетілгендей жалғасын табады.

Әдетте, туынды сөздердің бір жүйе бойынша қалыптасқан белгілі морфологиялық құрылымы болады. Сол құрылымның бірінші компоненті дербес мағыналы сөз болады да, ал екінші компоненті – бірінші компонентті белгілі бір сөз табына айналдыратын жұрнақ болады *(тіс+те, кел+тір, сабын+да)*. Туынды етістіктер, жасалатын негіздеріне қарай (бірінші компоненті), есімдерден және етістіктерден жасалған етістіктер деген екі салаға бөлінеді. Олар есім негізді етістіктер, етістік негізді етістіктер деп аталып жүр. Біз осылардың тек алғашқысын ғана, яғни есім негізді етістіктердің жасалу жолдарын жекелеп топ-топқа бөліп қарастырып, оларды шартты белгі-код арқылы таңбалау үлгісін 9-кесте арқылы ұсындық.

9-кестедегі бағдарлама да үш бағанадан тұрады:

1) белгі-код орны;

2) есім негізді туынды етістік сөздердің сипаты;

3) белгі-код таңбасы.

Бірінші бағананың белгі-код орны «1» (немесе №1-ші топтама), ал екінші бағана бойынша оның сипаттамасы – «етістік сөз» деп аталады, егер мәтін бойынан зерттелетін кезекті сөз – *етістікке* жататын болса, онда үшінші бағанадан сол етістік сөздің *белгі-коды* – «ЕТ» екенін білеміз. Екінші орынға немесе екінші топтаманың екінші бағанасында етістіктердің өздерін іштей мағыналық және өзара функционалдық жағынан жақындастыратын 11 топтың әрбіреуіне сипаттама берілген. Ал үшінші бағанада осы сипаттамаларға сәйкестендірілген белгі-код таңбасымен танысуға болады. Мысалы, белгі-кодтың «2-ші» орнының (2-ші топтамасының) үшінші бөлігі *қалып-сапа етістіктері* деп аталады да, ал оған сәйкес келетін үшінші бағанада сол сипаттаманың үш орынды белгі-коды – «102» берілген. Сол сияқты, туынды етістіктердің жұрнақтар арқылы есім сөздерден жасалу жолынан мысал келтірейік. Туынды етістіктердің ***-ла (-ле, -да, -де, -та, -те)*** жұрнағы арқылы жасалу ішкі 5 бөлікпен топталған атаулар сипатымен берілген:

1) дене мүшелері (*аяқта, жұдырықта, өкшеле*);

2) еңбек құралдары (*арала, балтала, арқанда);*

3) іс-әрекетті жүзеге асыруға объект болатын зат атаулары (*жүнде, майла, тұзда, алтында);*

4) мекен я орын, өлшеу, дыбыс т.т. (*мекенде, өрле; аршында, метрле; мыңқылда, шыңқылда);*

5) сын есім, сан есім, үстеу, еліктеуіш т.т. сөздерге жалғанған жағдайлар (*ақта, қарала, екеуле, аһла, үһле)*.

Осы 5 түрлі топқа 5 үш орынды сандар (белгі-код) сәйкестендірілген: 113, 114, 115, 116, 117. Осы тәріздес, басқа топтама орындарында да өзіне тән сипаттамалар және соған сәйкес шартты белгі-кодтар 9-кестеде көрініс тапқан.

Енді, осындай белгі-кодтар жиынын қалай пайдаланамыз деген сұрақ туындауы мүмкін. Етістік сөздерге қатысты 9-кестені пайдалануда да айтарлықтай айырмашылық жоқ деуге болады. Осы тұста айта кетерлік жайт, топтама орындар санын (белгі-код орнын) тағайындау (белгілеу), біріншіден, зерттеу мақсатына сай, екіншіден, әр сөз табының лексика-морфологиялық, семантикалық ерекшеліктеріне байланысты деуге болады. Мысалы, зат есім сөздерге қатысты сегіз орын жеткілікті деп алсақ, етістік сөздерге қатысты 9-кестеде ондай орындар саны төртеу-ақ. Ал басқа сөз табына қатысты немесе зерттеу мақсатына сай, топтау орын саны басқаша болуы да мүмкін.

Енді кестеде көрініс тапқан деректерді пайдалану жағын сөз етелік. Алғашында, зерттеуге алынған мәтін бөлігіндегі сөзқолданыстарды сөз табына ажырататын шартты белгілер қойылғаннан кейін, олардан компьютер арқылы қажетті жиілік сөздіктер алынуы керек.

Аталған жиілік сөздіктегі етістік сөзтұлғаларды (не басқа сөз табына қатысты сөздерді) бөліп алып, оларға 9-кесте бойынша белгі-кодтарды сәйкестендіру қажет болады. Ескерту ретінде айта кететін жайт, зерттеуді жиілік сөздіксіз-ақ бірден мәтінмен де жүргізуге болады, бірақ бұл тәсілде қайталанатын мәтін бірліктері зерттеушінің көп уақытын алуы мүмкін. Қалай болғанда да зерттеу жұмысы тізбек құрайтын сөзқолданыс жүйесінен етістік сөзтұлғаларды (немесе басқа сөз табы) кезек-кезегімен бөліп алып қарастыруды қажет етеді.

9-кесте

Етістік сөздерге лексика-морфологиялық ақпаратты сәйкестендіру бағдарламасы

Белгі-код орны	Етістік сөздердің лексика-морфологиялық сипаты	Белгі-код
1	2	3
1	Етістік сөз	ЕТ/
2	Мағыналық және функционалдық жағынан топтастыру:	
	1) Амал-әрекет етістіктері Мысалы: *босат, көтер, күрес, қи, сыз, өлше* т.т.	100
	2) Қимыл-қозғалыс етістіктері. Мысалы: *ауна, аудар, домала, секір, бүкірей* т.т.	101
	3) Қалып-сапа етістіктері. Мысалы: *жат, жантай, тұр* т.т.	102
	4) Ойлау-сөйлеу етістіктері. Мысалы: *айт, сөйле, де, ескер, жатта, ұмытпа* т.т.	103
	5) Өсу-өну етістіктері. Мысалы: *балала, жапырақта, гүлде, өс, қозыла* т.т.	104
	6) Бағыт-бағдар етістіктері. Мысалы: *бар, кел, қайт, әпер, әкет, қаш, түс, көтер* т.т.	105
	7) Көңіл-күй етістіктері. Мысалы: *жыла, қайғыр, өкін, күл, қуан, алақайла* т.т.	106
	8) Бейнелеу-еліктеу етістіктері. Мысалы: *жарқыра, күркіре, дүркіре, тарсылда* т.т.	107
	9) дыбыс-сес етістіктері	108
	10) көру-есту етістіктері.	109
	11) мінез-құлық етістіктері	110
	12) етістіктің тізімге енбеген түрлері	111
3	Түбір етістік. Анықтама: Етістік түбірі әртүрлі формалар (есімше, көсемше, рай, етіс, шақ т.б.) тудыратын қосымшалардың барлығын алып тастағанда сақталатын бөлігі. Мысалы: *аз, ал, айт, алда, ат,* т.б.	112

1	2	3
	Есім негізді сөздерден жасалатын туынды етістіктер. Жұрнақтар арқылы жасалу жолдары: 1. *-ла (-ле, -да, -де, -та, -те)* жұрнағы арқылы жасалған туынды етістіктер: а) дене мүшелері атауларына жалғанады. Мысалы: *аяқта, жұдырықта, өкшеле.*	113
	ә) Еңбек құралдары атауларына жалғанады. Мысалы: *арала, балтала, арқанда.*	114
	б) іс-әрекетті жүзеге асыруға объект болатын заттарға жалғанады. Мысалы: *жүнде, майла, тұзда, алтында.*	115
	в) Мекен я орын, өлшеу, дыбыс, қозғалыс, көрініс, төл және әлеуметтік, саяси, мәдени, тұрмыс, салт, табиғат т.т. атауларына жалғанады. Мысалы: *мекенде, өрле, төменде; аршында, метрле, тоннала; мыңқылда, шыңқылда; жарқылда, бұрқылда, ирелеңде; ботала, қозда; тәрбиеле, еркеле, жазала,қаумала, үймеле* т.т.	116
	г) Сын есім, сан есім, үстеу, еліктеуіш, одағай сөздерге жалғанады. Мысалы: *ақта, қарала, жаманда; онда, жүзде, екеуле; төменде, кейінде, ілгеріле; жымыңда, күлімде; аһла, үһле, ойбайла* т.т.	117
	2. *-лан (-лен, -дан, -ден, -тан, -тен)* жұрнағы арқылы жасалады. Мысалы: *ашулан, арлан, борышлан, иелен ...*	118
	3. *-лас (-лес, -дас, -дес, -тас, -тес)* жұрнағы арқылы жасалған. Мысалы: *бірлес, көмектес, бәстес, достас, кезектес, сабақтас* т.б.	119
	4. *-лат (-лет, -дат, -дет)* жұрнағы арқылы жасалған. Мысалы: *дауылдат, тездет, түнделет* т.б.	120
	5. *-а (-е)* жұрнағы арқылы. Мысалы: *аша, деме, жаса, өрте, тіле, сына, міне* т.б.	121
	6. *-ай (-ей, -й)* жұрнағы арқылы. Мысалы: *күшей, мұңай, қартай, кеңей, көбей, молай* т.б.	122
	7. *-қар (-ғар, -кер, -гер)* өнімсіз жұрнағы арқылы. Мысалы: *басқар, ескер, аңғар, теңгер, ақыр, қақыр, жазғыр, ысқыр, кекір* т.б.	123
	8. *-ар (-ер, -р)* көне жұрнағы арқылы. Мысалы: *жаңар, тазар, ескір, өзгер* т.б.	124

1	2	3
	9. *-ал (-ел, -ал, -іл, -л)* қосымшасы арқылы. Мысалы: *жоғал, оңал, тарыл, тіріл, теңел* т.б.	125
	10. *-ық (-ік)* көне жұрнағы арқылы. Мысалы: *ашық, бірік, зарық, дәнік, өшік* т.б.	126
3	11. *-сы (-сі)* және *-ымсы (-імсі)* көне жұрнақтары арқылы. Мысалы: *батырсы, босаңсы, көлгірсі, үлкенсі, апамсы* т.б.	127
	12. *-сын (-сін)* жұрнағы арқылы. Мысалы: *адамсын, білгішсін, көпсін, жамансын* т.б.	128
	13. *-сыра (-сіре)* жұрнағы арқылы. Мысалы: *айсыра, әлсіре, қансыра, жетімсіре* т.б.	129
	14. *-ыра (-іре)* жұрнағы арқылы. Мысалы: *барқыра, бұрқыра, дүркіре, күркіре* т.б.	130
	15. *-ырай, (-ірей)* жұрнағы арқылы. Мысалы: *бақырай, шақырай, кішірей, шікірей,* т.б.	131
	16. *-ы, -і*; *-шы, -ші*; *-ан, -ен, -ын, -ін, -н; -ырқа, -ірке, -ырқан, -іркен*; *-ына, -іне*; *-қа, -ке, -га, -ге* т.б. көне жұрнақтар арқылы. Мысалы: *байы, жасы, желпі, кеңі; аунақшы, дөңбекші; тасырқа, мүсірке, ашырқан, шіміркен; есіне, қатына, пысына; бүрке, иіске, қозға; басын, жирен, оян, үйрен* т.б.	132
	Етістік негізді сөздерден жасалатын туынды етістіктер	133
4	Күрделі етістіктер	140
	Жалаң етістік	141

Белгіленген бір орынға сол топтамадағы белгі-кодтар ішінен біреуі ғана таңдалып алынатынын үнемі есте сақтау қажет.

Мысал ретінде мына сөйлемде кездесетін етістік сөзтұлғаларды бөліп алып қарастырайық: *«Кешке әрі тоңып, әрі шаршап аһлап, үһлеп отырған, кемпір өзін өлтіріп кете жаздаған кім екенін есіне алды»* (С.Көбеев).

Етістік сөзтұлғаларға жататындар: *тоңып, шаршап, аһлап, үһлеп, отырған, өлтіріп кете жаздаған, алды.*

Енді 9-кестеде көрсетілген деректер бойынша етістіктерге белгі-код мәндерін сәйкестендірейік:

10-кесте

Етістіктерге белгі-код мәндерінің сәйкестігі

тоңып	ЕТ/	106	133	141
шаршап	ЕТ/	106	133	141
аһлап	ЕТ/	106	133	141
үһлеп	ЕТ/	106	133	141
отырған	ЕТ/	102	133	141
өлтіріп	ЕТ/	100	133	141
кете жаздаған	ЕТ/	105	140	140
алды	ЕТ/	111	133	141

Сонымен, әртүрлі етістік сөздер өздерінің жасалу және мағыналық ерекшеліктеріне қарай неше түрлі белгі-код мәндерін қабылдай алатынын аңғаруға болады. Етістік топтарына берілген бұл сандық белгі кодтарды келешекте корпусты құрастыру кезінде халықаралық стандартқа сәйкестіндіріп, семантикалық тэгтерді әзірлеу қажет.

2.4.5. Сын есімнің лексика-морфологиялық құрылымына ақпараттық белгі-кодты сәйкестендіру бағдарламасы

Грамматика саласындағы сөз табының мазмұны «лексика-грамматикалық» деп аталатын екі компоненттің бірлігінен құралатыны белгілі. Бұл атаудан лексикалық семантикамен бірге грамматикалық семантика деген ұғымды да бірге қабылдау керек болады. Ал қос сөздің екінші бөлігіндегі «грамматикалық» сөзінің мазмұнына белгілі бір сөзге тән грамматикалық категориялардың және олардың жасалу, түрлену тұлғаларының мағыналары енеді. Яғни бұл жердегі «грамматикалық» ұғымына сөздің жаңадан сөз тудыру, сөз түрлендіру, сөз байланыстыру жүйелерінің мағыналары мен тұлғалары, демек, бүкіл сөздің

морфологиялық және синтаксистік белгілері (сыр-сипаттары) түгел енеді [30, 384 б.].

Қазақ тілінің қолданбалы саласының, дәлірек айтқанда, статистикалық лингвистиканың алдына қойған мақсаттарының бірі – сөз табының лексикалық (семантикалық), грамматикалық (морфологиялық және синтаксистік) белгілері арқылы, яғни осы үш белгінің негізінде статистикалық зерттеулер жүргізіп, сөздің мағынасын, морфологиялық формасын (тұлғасын) және синтаксистік қызметін сандық деректер арқылы тани білу. Соның нәтижесінде мәтін мазмұнын ашудың формальды жақтарын айқындау.

Сөз таптарының ішкі жіктелген бөліктерінің түрлері тілде біреуі аз не көп қолдануы мүмкін. Статистикалық зерттеу тәжірибесі бойынша олардың қолдану жиілігін анықтау үшін кейбір шараларды алдын ала белгілейтін бағдарламаларды іске қосу қажет. Мысалы, әртүрлі тілдік стильдердегі сөз таптарының қатынастық сипатының статистикасын білу керек болса, әр сөз табына қатысты шартты белгіні (кодты) қоюдың зерттеушіге ыңғайлы жолдарын ойластыруымыз керек. Себебі, сөз таптарын не басқа тілдік бірліктерді мәтін ішінде түр-тұрпатына қарай формальды әдіспен ажырату мүмкіндігі әлі де болса зерттеу аясынан тыс қалып жүр.

Орыс тілінің морфологиялық сипатына тән сөз таптарын шартты түрде белгілеу (кодтау) ресейлік ғалым Б.Н.Головиннің «Язык и статистика» атты кітабында да сөз болады [31, 191 б.].

Морфологиялық ақпарат белгілерін сәйкестендіру бағдарламасын құрудағы біз ұсынып отырған үлгінің ерекшелігі – қазақ тілі сөз таптарының ішкі бөліктерінің құрылымдық сипаты мен жасалу жолдары ескеріліп, әмбебаптық сипатты көздейді. Мысалы, зат есіммен, үстеумен және т.б. сөз таптарымен әрі ортақ, әрі тұлғалас сөз тудыратын жұрнақтарымен қатар тек туынды сын есім ғана жасайтын арнаулы жұрнақтары және сын есімнен сын есім тудыратын шырай жұрнақтары да бар. Бұл аталғандар да сын есімнің морфологиялық формалары болып есептеледі [88, 122-130].

Сын есімдердің тағы да бір өзіндік қасиеті – заттардың алуан түрлі сыр-сипаттары мен белгілерін тікелей білдірумен қатар басқа

заттардың қатынастары арқылы да сондай сипаттарды білдіре алады. Осыған байланысты, сын есімдер семантикалық мағыналары мен грамматикалық ерекшеліктеріне қарай, *сапалық (негізгі)* және *қатыстық (туынды)* сын есім деп аталатын екі салаға бөлінетіні де белгілі.

Негізгі немесе сапалық сын есімдер жайлы айтатынымыз, олар, әдетте, ешбір қосымшасыз тұрып-ақ, заттың әртүрлі сыр-сипатын білдіретін түбір сөздерден тұрады. Бірақ, егер кейбір түбір сөз деп жүрген сөздерді олардың тарихи даму тұрғысынан қарастырсақ, олар туынды сөз болып та шығуы мүмкін. Сондықтан, құрылымы жағынан негізгі сын есім деп есептелетін сөздердің саны да, сапасы да ұдайы өзгеріп отырады. Ал сын есімдерді морфемалық құрылымына қарай, негізгі және туынды деп жіктеу шартты екендігін де есте ұстаған жөн. Яғни кез келген сөйлем ішінде (мәтінде) кездесетін сын есім сөздер, біріншіден, сапалық сын есімге жатуы *(ақ, қара, сары, көк, сұр, биік, үлкен, аласа, жылы, суық, ірі, кіші* т.т.), екіншіден, морфологиялық (синтетикалық) тәсіл бойынша, яғни тиісті жұрнақтар арқылы жасалған туынды сын есімге жатуы мүмкін.

Үшіншіден, қарастырылатын кезекті сын есімге қатысты сөз синтаксистік (аналитикалық) тәсіл, яғни жалаң сын есімдердің бір-бірімен тіркесуі арқылы *(ақ сары, қызыл сары, ақ шұбар, ақ көйлекті, көп балалы, үлкен-кішілі* т.т.) және морфологиялық-синтаксистік (семантикалық) тәсіл бойынша жасалатын туынды сын есімдер де болуы мүмкін.

Соңғы аталған тәсіл бойынша туатын сын есімдер қолданылуы жағынан алғанда тіпті де өнімсіз деуге болады. Сондықтан сын есімдердің жасалу жолдарының ең негізгісі деп, ең алдымен морфологиялық, яғни синтетикалық тәсілді, ал одан кейін барып, аналитикалық, яғни синтаксистік тәсілді айтуға болады.

Осы себептен, біз сын есімнің лексика-морфологиялық құрылымына ақпараттық белгі-кодты сәйкестендіру бағдарламасын құрастыруда тек қана сын есімнің синтетикалық тәсілмен жасалу жолына ғана, дәлірек айтқанда, есімдерден сын есім тудыратын өнімді жұрнақтар арқылы жасалу жолын негіз етпекпіз.

11-кесте

Сын есімнің лексика-морфологиялық құрылымына ақпараттық белгі-кодты сәйкестендіру бағдарламасы

Сын есім сөздердің лексика-морфологиялық сипаты	**Белгі-код**
1	2
Егер сөз (сөзтұлға) морфемалық құрамына қарай сапалық сын есім (негізгі сын есім) болса:	СН1
1) заттың не құбылыстың түрі мен түсін анықтаса. Мыс.: *ақ, қара, қызыл, жасыл* т.б.	201
2) заттың не құбылыстың сыры мен сапасын анықтаса. Мыс.: *жақсы, жаман, тәуір, нашар* т.б.	202
3) заттың көлемі мен аумағын, ұзындығы мен салмағын анықтаса. Мыс.: *үлкен, кіші, ұзын, ауыр, қысқа, жеңіл* т.б.	203
4) заттың дәмі мен иісін білдірсе. Мыс.: *ащы, тәтті, күлімсі*	204
5) заттың не құбылыстың басқа да қасиет-белгілерін білдірсе.	205
Егер кезекті сөз (сөзтұлға) морфемалық құрамына қарай қатыстық сын есімге жатса, яғни заттың белгісін басқа бір заттың қатысы арқылы білдірсе және ондай туынды сын есімдер өнімді (өнімсіз) жұрнақтар арқылы жасалса.	СН2
Есімдерден туынды сын есім тудыратын өнімді жұрнақтар арқылы жасалып, заттың сыртқы түрі мен түсіне, кескіні мен келбетіне, сыры мен сынына, мекен мен мезгілге және т.б. белгілеріне қатысты сындық ұғым білдірсе. 1. *-қы, -кі; -ғы, -гі* жұрнақтары арқылы кейбір зат есімдерден, есімдіктерден, үстеулерден, сондай-ақ, жатыс және шығыс септіктегі сөздерден жасалатын туынды сын есімдер мынадай мағынада жұмсалса: а) егер мекендік ұғым білдіретін кейбір зат есімдерге, үстеулерге, сондай-ақ, жатыс (кейде шығыс) септік формаларындағы есімдерге жалғанса. Мыс.: *ауызғы, төргі, түпкі, ішкі, төменгі, соңғы, санаулы, бүктеулі* т.б.	210
ә) егер мезгілдік ұғым білдіретін кейбір зат есімдер мен есімдіктерге, мезгілдік үстеулерге жалғанса. Мыс.: *кешкі, түскі, күзгі, жазғы, көктемгі, түнгі, күндізгі* т.б.	211

1	2
2. *-лы, -лі, -ды, -ді, -ты, -ті* қосымшасы арқылы мынадай туынды сөздер жасалады:	
а) белгілі бір заттың (я құбылыстың) бар екендігін білдіру үшін зат есімдерге жалғанса. Мыс.: *арлы, сулы, әсерлі, гүлді, икемді, байыпты, инабатты, балалы* т.б.	212
ә) қосарланған зат есімнен, сын есімнен, сан есімнен, үстеуден күрделі сын есімдер осы қосымшалар арқылы жасалса. Мыс.: *ағалы-інілі, ойлы-қырлы, таулы-тасты, өзенді-сулы, үлкенді-кішілі, бұрынды-соңды* т.б.	213
б) екі-үш сөзден құралып, суреттеме атаулар қызметіндегі күрделі сын есімдер жасалса. Мыс.: *ақ басты, қаз мойынды, ай қабақты, теке сақалды* т.б.	214
3. *-сыз (-сіз)* жұрнағы есім сөздерге жалғанып, болымсыздық мағынадағы туынды сын есімдер жасалса. Мыс.: *баласыз, көліксіз, білімсіз, сенсіз, бізсіз, мұңсыз* т. б.	215
4. *-шыл (-шіл)* қосымшасы зат есім, есімдік, модаль сөздерге (әр тарап) жалғанып, бейімділікті, икемділікті, құмарлықты білдіретін қатыстық сын есімдер жасалса. Мыс.: *ұйқышыл, ұйымшыл, өзімшіл, турашыл* т.б.	216
5. *-дай, (-дей, -тай, -тей)* жұрнағы жалғанып, сөздерді салыстыру, ұқсату мәнді туынды сын есімдер жасалса. Мыс.: *аттай, әкедей, мендей, сендей, жүздей, өлердей* т.б.	217
6. *-лық, -лік, -дық, -дік, -тық, -тік* жұрнақтары арқылы мына мағынадағы туынды сын есімдер жасалса: а) зат есімдерге жалғанып,	
сөздердің нақтылылық мағына қасиеттерін білдіретін туынды сын есімдер жасалса. Мыс.: *орталық, қоғамдық, қалалық, азаматтық, жолдастық* т.б.	218
ә) мезгіл атаулары мен әр қилы бұйым атауларына жалғанып, мезгіл, өлшеу мөлшерімен байланысты туынды сын есімдер жасалса. Мыс.: *айлық, жылдық, апталық, тәуліктік, көйлектік, пальтолық, қайнатым-дық, екі-үш асымдық* т.б.	219
б) есімдіктерге жалғанып, олардан белгілі бір жаққа қатыстықты білдіретін туынды есімдер жасалса. Мыс.: *өздік, мендік, сендік, қандайлық, қаншалық* т.б.	220

1	2
7. *-лас, -лес, -дас, -дес, -тас, -тес* жұрнақтары арқылы мына мағынадағы туынды сын есімдер жасалса: а) адам мінезіне байланысты зат есімге жалғанып, ондай сипаттың басқа адамдарға да қатысты екендігін білдірсе. Мыс.: *ниеттес, сырлас, тілеулес, пікірлес* т.б.	221
ә) адамзат, жан-жануарлар тегімен байланысты ұғымдағы зат есімдермен жалғанып, тұқым-туыстас екенін білдіретін сын есім жасалса. Мыс.: *туыстас, аталас, бауырлас, қарындас, тумалас* т.б.	222
б) мекен ұғымдарына жалғанып, олармен қоныстастығын білдіретін мағынадағы туынды сын есімдер жасалса. Мыс.: *ауылдас, көршілес, жерлес, елдес, тумалас, ұялас,* т.б.	223
в) кейбір зат есімдерге жалғанып, олардың нақтылы лексикалық мағыналарына қарай, түрлі қарым-қатынас, мерзім-мөлшер жағынан сыбайлас келетін сындық ұғымдарды білдірсе. Мыс.: *жолдас, замандас, сабақтас, кәсіптес, қызметтес, дәмдес, өкшелес* т.б.	224
8. *-шаң, -шең* жұрнақтары арқылы жасалатын туынды сын есімдер: а) киім-кешек атауларына жалғанып, адамның бойындағы киімімен байланысты сыртқы бейне-көрінісін білдіретін сын есімдер жасалса. Мыс.: *көйлекшең, етікшең, пальтошаң, шалбаршаң* т.б.	225
ә) кейбір зат есімдерге жалғанып, адамға не затқа тән белгілі бір ерекше қасиет білдіретін туынды сын есімдер жасалса. Мыс.: *ашушаң, сөзшең, бойшаң, тершең, кіршең* т.б.	226
9. Есімдерден туынды сын есім тудыратын өнімсіз жұрнақтар арқылы:	
а) *-дар, -дер, -тар, -тер* қосымшалары кейбір зат есімдерге жалғанып, белгілі бір іс-әрекетіне душар болғандықты білдіретін сындық ұғым жасалса. Мыс.: *қарыздар, хабардар, борыштар* т.б.	227
ә) иран тілдерінен (парсы, тәжік, ауған) ауысқан кірме *«-и»* қосымшасы кейбір есімдерге жалғанып, туынды сын есім жасалса. Мыс.: *әскери, мәдени, тарихи* т.б.	228
б) *-паз* жұрнағы жалғанып, сын есімге бейім жаңа сөз жасалса. Мыс.: *өнерпаз, аспаз, білімпаз, ойынпаз* т.б.	229
в) *-мпаз (-м+паз), -ымпаз (-ым+паз), -імпаз (-ім+паз)* жұрнағы (және компоненттер) кейбір есім сөздерге жалғанып, туынды сын есім жасалса. Мыс.: *сезімпаз, алымпаз, білімпаз, жағымпаз, жасампаз* т.б.	230

1	2
г) *-қой, (-гой, -гөй)* жұрнағы кейбір зат есім есім сөздерге жалғанып, амал-әрекет пен мінез-құлықтардың атауы болатын туынды сын есім тудырса. Мыс.: *кәсіпқой, әзілқой, сәнқой, жәдігөй, әзілқой, сәнқой, жанжалқой* т.б.	231
д*) -қор* жұрнағы зат есім есім сөздерге жалғанып, белгілі бір машықтанғанды білдіретін мағынадағы туынды сын есім жасалса. Мыс.: *жемқор, жалақор, айлақор, бейнетқор, қамқор, мансапқор, ызақор, намысқор* т.б.	232
Етістіктерден өнімді жұрнақтар арқылы жасалатын туынды сын есім сөздер: 1. *-қ, -к, -ық, -ік, -ақ, -ек:* а) салт және сабақты етістікке жалғанып, түрлі сындық ұғымдағы атаулар жасайды. Мыс.: *ашық, тұнық, шірік, ілік, дөңгелек* т.б.; б) еліктеу сөздерге жалғанып, қатыстық сын есім тудырса. Мыс.: *бұлтақ, жалтақ, еңкек, жалпақ* т.б.;	317
в) кейбір зат есімдерге жалғанып, туынды сын есім жасаса. Мыс.: *жолақ, қасырақ, ортақ, ірек* т.б.; г) етістіктен *-уық (-уік)* формасы арқылы белгілі бір іс-әрекетке бейімділікті білдіретін туынды сын есім жасаса. Мыс.: *жылауық, сөйлеуік, сұрауық, сыбырлауық* т.б.	
2. *-ыңқы, -ңқы, -іңкі.* Мыс.: *жатыңқы, салбыраңқы, көтеріңкі, кебіңкі, батырыңқы* т.б.	318
3. *-ынды, -інді, -нды, -нді.* Мыс.: *асыранды, серпінді, жаттанды, шұбырынды, түйінді* т.б.	319
4. *-малы (-мелі, -балы, -белі, -палы, -пелі).* Мыс.: *ауыспалы, көшпелі, таңдамалы, аумалы-төкпелі* т.б.	320
5. *-қыш, -кіш, -гыш, -гіш.* Мыс.: *білгіш, оңғыш, тапқыш, сенгіш,* т.б.	321
6. *-шақ (-шек).* Мыс.: *мақтаншақ, еріншек, қызғаншақ, ашуланшақ, жасқаншақ, тартыншақ* т.б.	322
7. *-ымды (-імді, -мды, -мді).* Мыс.: *жағымды, ұғымды, үйлесімді, жарасымды, сенімді* т.б.	323
8. *-улы, -улі.* Мыс.: *жинаулы, ерттеулі, үюлі* т.б.	324
9. *-қақ, -кек, -гақ, -гек.* Мыс.: *асқақ, тоңғақ, жабысқақ, майысқақ, тайғақ, оңғақ, қатқақ, ұрысқақ* т.б.	325
10. *ма, -ме, -ба, -бе, -па, -пе.* Мыс.: *жалдама (ақы), қызба (адам), көшпе (құм), сырма (бешпет), аспа (шам), сырма (бешпет)* т.б.	326
Етістіктерден өнімсіз жұрнақтар арқылы жасалатын туынды сын есім сөздер: 1. *-ыс, -іс, -с.* Мыс.: *келіс, кетіс, ұқсас, таныс, тіркес, жалғас* т.б.	327

1	2
2. *-қыр, -ғыр, -кір, -гір*. Мыс.: *білгір, ұшқыр, үшкір, алғыр, өткір* т.б.	328
3. *-мыс (-мыш)*. Мыс.: *жасамыс, айтылмыш, тарамыс, жазылмыш, жаралмыш т.б.*	329
4. *-ымтал (-імтал)*. Мыс.: *ұғымтал, өсімтал, сезімтал* т.б.	330
5. *-қы (-кі, -ғы, -гі)*. Мыс.: *бұралқы, оралғы, жинақы, күлдіргі* т.б.	340
6. *-ыр (-ір, -ар, -ер, -р)*. Мыс.: *жұмыр, иір, обыр, қыңыр, тықыр, шымыр, құзар, былжыр, жалтыр, жылтыр, сылбыр, былбыр* т.б.	341
7. *-у*. Мыс.: *жарау, қызу, тақау, жадау, жабырқау, түзу, таяу, бітеу, баяу, көлбеу, қату* т.б.	342
8. *-аған (-еген)*. Мыс.: *тебеген, қабаған, сүзеген, қашаған, береген, алаған, безеген, көреген* т.б.	343
9. *-ын (-ін, -н)*. Мыс.: *ортан, бүтін, еркін, ұзын, жайын* т.б.	344
10. *-қалақ (-келек, -ғалақ, -гелек)*. Мыс.: *ұшқалақ, сасқалақ, қозғалақ* т.б.	345
11. *-алақ (-елек)*. Мыс.: *еңкелек, қаңғалақ, бұлталақ, шыжалақ* т.б.	346
12. *-анақ*. Мыс.: *сұғанақ, шұқанақ, шығанақ* т.б.	347
13. *-ғылықты (-гілікті, -қылықты, -кілікті)*. Мыс.: *жеткілікті, тұрғылықты, жергілікті* т.б.	348
14. *-мсақ (-мсек, -ымсақ, -імсек)*. Мыс.: *сұрамсақ, тілемсек*	349

Енді, осы *11-кестеде* қарастырылған сын есімнің лексика-морфологиялық құрылымына ақпараттық белгі-кодты сәйкестендіру бағдарламасы негізінде, кейбір негізгі және туынды сын есімдерге қатысты сөздерге де шартты белгі-код сәйкестендіру мүмкіндігі бар.

Жоғарыда зат есім, сын есім, етістік сөз таптарының лексика-морфологиялық топтарына қатысты цифрлар арқылы белгі-кодтар қойылған кестелер берілді. Сөз таптарының мұндай лексика-морфологиялық топтары бойынша корпус мәтіндерін сипаттауда осы жоғарыда көрсетілгендей цифрлық кодтау арқылы да немесе басқа да әріп таңбалары, яғни қысқартулар арқылы да немесе латын қаріптері арқылы да белгі-кодтар жасауға болады. Тіпті корпус мүмкіндігі өте кең болғандықтан, мұндай лингвистикалық сипаттамаларды толық берудің де мүмкіндігі бар.

Семантикалық белгіленімдерді енгізуде алдымен тіл жүйесіндегі осындай мағыналық топтар анықталып, сөздер сол топтарға сараланады. Әр топқа халықаралық шартты белгілерді (тэгтерді) сәйкестендіреді. Мұндай лексика-семантикалық топ басқа сөз таптарына да қатысты анықталады. Алайда тілде мұндай топтар жоғарыда аталған үш сөз табында ең көп кездеседі.

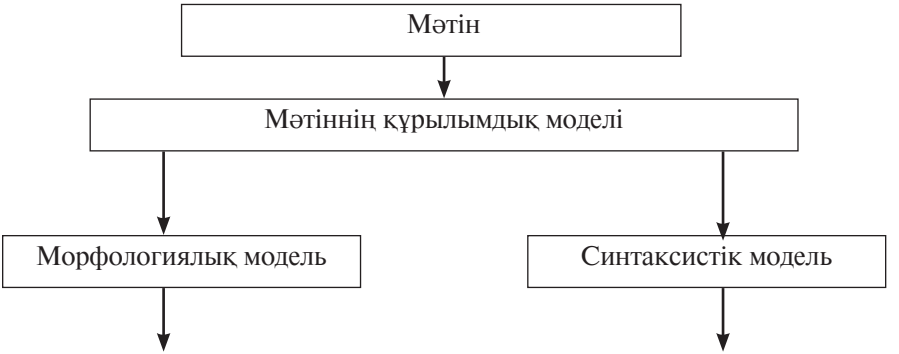
2.4.6. Синтаксистік белгіленім

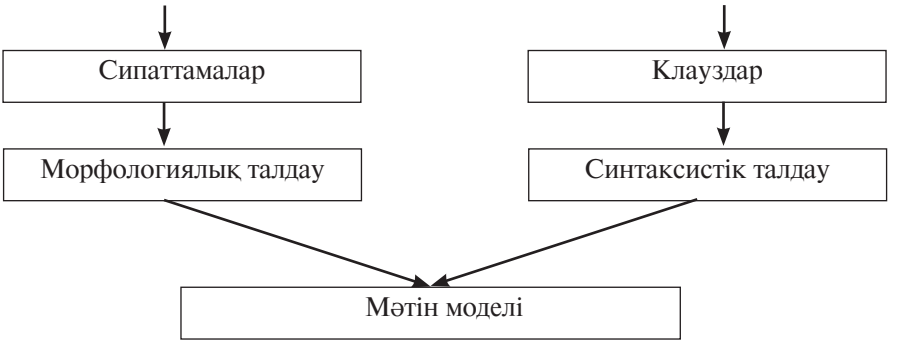
Қазіргі кезде қоғам талаптарына сәйкес қайсібір істі болмасын, заманауи технологияға бейімдеп, қол жұмысын жеңілдету, ғылым-білімді мүмкіндігінше автоматтандыру қажет екендігі белгілі. Бұл ретте тіл ғылымы да техника мүмкіндіктерін барынша тиімді пайдалануға бағытталып отыр. Тіл адамдардың қарым-қатынас құралы болғандықтан, адам миындағы ойлау мүмкіндіктері қаншалықты зор болса, соғұрлым ойды берудің құралдары да алуан түрлі болып құбыла береді. Сөз (сөйлеу) тілдік құралдардың ұйымдасып, белгілі бір ойды жеткізу ғана емес, оның ішкі формасында ой жеткізудің коммуникативтік мақсатын көрсететін экстралингвистикалық факторлармен де біте қайнасып жатады. Осындай ой жеткізудің, тілдік қатынасқа түсудің формаларын сипаттайтын терминдер ретінде қазіргі кезде «дискурс», «прагматика» сияқты ұғымдар пайда болып, тіл құрылысы әр қырынан зерттеле бастады. Тілдесімнің алдындағы қатынасқа дайындық кезеңі, тілдесім кезіндегі коммуникативтік жағдаяттар тікелей тілдік құралдардың интралинвистикалық семантикасынан көрініс таба бермейді. Ал экстралингвистикалық факторларды айқындау адамның ойлау, ми жұмысының қызметі арқылы, байқау, сезіну, ой түю сияқты үдерістер арқылы жүзеге асады. Тіл қолданыста ой жеткізу барысында сөйлеушінің мақсатына қарай қаншалықты құбылып отырса да, ой жеткізу құралдары белгілі бір амал-тәсілдерге бағынады. Яғни ойды жеткізудің қалыптасқан синтаксистік құрылымдары бар. Мысалы, сөйлеуші белгілі бір іс-әрекетке қарсылығын білдіру үшін, «Жоқ», «болмайды», «мүмкін емес», «қажеті жоқ», «дұрыс емес», «бұлай етпей-ақ қойсақ қайтеді», «қарсымын» т.б. осы сияқты қалыптасқан синтаксистік құрылымдарды қолда-

нады. Демек, қайсібір тілдік қолданыстың болмысын қалыптасқан формуласы, моделі бар. Тілдік бірліктерді моделдеудің ең қиын әрі күрделісі – сөйлемдер де белгілі бір синтаксистік қатынастарға құрылады.

Синтаксистік қатынас теориясының бүгінгі таңдағы аса өзектілігі жаңа ғылыми технологияға негізделген тілдік корпустар құрастыру ісінде де айқын байқалады. Ғылым-білімнің бүгінгі жетістігі оның барлық салаларына компьютерлік технологияның енгізілуін талап етеді. Ғылыми-зерттеу жұмыстарындағы дәстүрлі қалыптасқан материал жинауға қатысты қол жұмысы бүгінде тілдік корпустарды жасаумен байланысты автоматтандырылып, адам атқаратын жұмысты жеңілдетуге баса назар аударылуда. Бұл ретте компьютер жадына ғылыми-зерттеу жұмыстарына қажетті материалды салып, үлкен массивті мәтіндерді лингвистикалық талапқа сәйкес өңдеп шығару қажет. Ал компьютер жадына салынатын лингвистикалық ақпараттар тіл білімінде «разметка» (белгіленім) терминімен қолданыс тапқан. Олар тілдік деңгейлерге сәйкес түр-түрге бөлінеді. Мысалы, морфология деңгейіне (немесе талдау түріне) сәйкес – морфологиялық белгіленім, синтаксис деңгейіне сәйкес – синтаксистік белгіленім, бұдан басқа да белгіленім түрлері көп. Бірақ бұл екеуі негізгілері болып табылады.

Белгіленімдер корпус мәтіндерін лингвистикалық тұрғыдан өңдеу үшін салынады. Айналып келгенде, белгіленімнің қай түрі болмасын, мәтінді сипаттайды, мәтіннің лингвистикалық моделін құрайды. Өрнектеп көрсетсек:





3-сурет. Мәтіннің лингвистикалық моделі

Лингвистикалық белгіленімдердің ішінде компьютер жадына морфологиялық белгіленімдерді енгізу барлық тілдерде алдымен қолға алынған. Және морфология сөз бөлшектерін қарастыратындықтан, олар тілдік жүйеге түскен. Ал синтаксистік белгіленімдерді компьютер жадына енгізу мәселесі тілдердің барлығында бірдей шешімін таппаған.

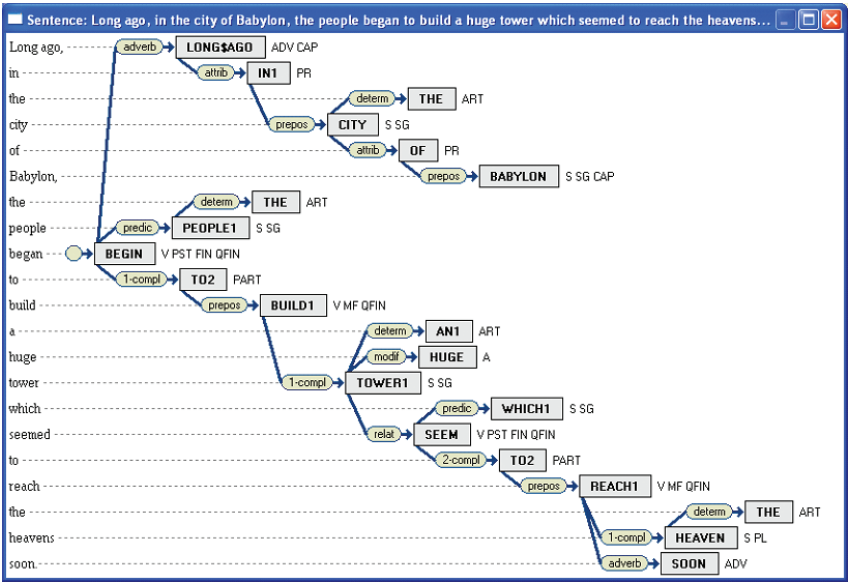
Синтаксистік белгіленім синтакситік талдаудың немесе парсингтің (ағыл. parsing) нәтижесі болып табылады. Белгіленімнің бұл түрі лексикалық бірліктер мен әртүрлі синтаксистік конструкциялар арасындағы синтаксистік байланысты сипаттайды (мысалы, бағыныңқы, етістікті сөз тіркесі және т.б.). Ал сөздердің сөйлем ішіндегі байланысы әрқашан тұрақты емес. Жайылма сөйлемдерде сөздердің байланысуы қатар тұрған сөздерде ғана емес, араға бірнеше сөз салып та алшақ байланыста тұра береді. Сөздер арасындағы мұндай синтаксистік байланыстарды корпустар құрастыруда орыс тіл білімінде «деревозависимость» деген терминмен атайды. Сөздердің байланысуына ұйтқы болатын басыңқы сөз – «вершина» болса, бағыныңқы сөздер «ветьва», яғни тармақтар түрінде сипатталады.

Көпшілік жағдайда, оның негізінде тікелей құрастырушылардың құрылым грамматикасы (грамматика структур непосредственно составляющих) жатыр. Сөйлем мүшелері арасындағы синтагматикалық қатынастардың сызбасы тәуелділіктер таралымы (дерево

зависимостей) түрінде көрініс табады. Мәтінде олар ашық және жабық тік жақшалар арқылы белгіленеді. Тік жақшалар ішінде әртүрлі синтаксистік конструкцияларды сипаттайтын таңбалар (кодтар) жазылады. Бір қос жақша ішіне екінші қос жақша және т.т. енуі мүмкін, ал жоғарғы деңгейдегі элемент ретінде «S» символомен белгіленетін сөйлем саналады. Синтаксистік белгіленім қойылған мәтіндер «treebanks» ретінде танылады.

Морфологияға қарағанда, синтаксистік құрылым мен синтаксистік қатынастарды бейнелейтін тәсілдер айтарлықтай бірыңғайланбаған деуге болады.

Орыс тілі бойынша жүргізілген синтаксистік талдау көбінде тәуелділік (дерево зависимость) құрылымымен бейнеленген. 4-суретте ағылшын тіліндегі «*Long ago, in the city of Babylon, the people began to build a huge tower which seemed to reach the heavens soon*» сөйлемінің тәуелділік таралымы (дерево зависимость) түріндегі бейнесі көрініс тапқан.



Сурет 4. Синтаксистік талдауға мысал
(тәуелділік грамматикасы, ЭТАП-3 жүйесі)

Морфологиялық деңгейдегідей, синтаксистік деңгейде де грамматикалық белгіленім сызбасын тәптіштеуді азайту үрдісі – жылдамдықты көбейту мен мәтінді талдаудағы реттілікті ұлғайту мақсатында жүргізіледі. Осындай үрдістің нәтижесінде пайда болған әдіс skeleton parsing деген атқа ие болды.

Қазақ тілінде де мәтіндерге синтаксистік белгіленімдер енгізуге болады. Алайда синтаксис морфология сияқты модельдеуге көне бермейді, яғни автоматты түрде синтаксистік белгіленім қою қиын. Мәтіндерге синтаксистік белгіленім қою процесі көбінесе қолдап жүргізіледі. Сол себепті синтаксистік белгіленген корпустарда сөзқолданыс саны да көп болмайды.

Тілдің синтаксистік жүйесі моделдеуге біршама қиын әрі күрделі болғандықтан, синтаксистік белгіленімдер морфологиялық белгіленімдер сияқты барлық тілдердің корпустарына бірдей енгізілмеген. Орыс тілінің корпустық лингвистика саласының белгілі маманы П.В.Захаров бұл туралы «В отличие от морфологии, способы представления синтаксической структуры и синтаксических отношений не столь унифицированы», - деп көрсетеді [5]. П.В.Захаров синтаксистік талдауларды бірнеше бағытта жүргізіп, модельдеуге болады дейді. Атап айтқанда:

- грамматикалық қатынастар немесе байланыстар (грамматика зависимостей);
- құрамдық грамматика (грамматика непосредственно составляющих);
- құрылымдық сызба грамматикасы (грамматика структурных схем);
- сөйлем мүшелері туралы дәстүрлі зерттеулер (традиционные синтаксические учения о членах предложения);
- функционалды грамматика (функциональная грамматика);
- семантикалық синтаксис (семантический синтаксис и др) және т.б.

Хельсинкский университетінен М.В.Копотев пен Петрозаводск мемлекеттік университетіненің корпустық лингвистика мамандары Г. Б. Гурин «Принципы синтаксической разметки Хельсинкского аннотированного корпуса русских текстов Ханко» деп

аталатын мақалаларында корпусқа енгізілетін синтаксистік белгіленімдер енгізудің әртүрлі формаларын атай отырып, дәстүрлі сөйлем мүшелеріне қатысты талдаулар негізінде белгіленімдер енгізуге болатындығын, алайда тіл жүйесіндегі сөйлем мүшелеріне қатысты кейбір мәселелердің әлі күнге дейін бір шешім таппай отырғандығын айта келе, сөйлем мүшелеріне қатысты белгіленімдерді толық 100 пайызға модельдеу мүмкін емес деген қорытынды жасайды. Өз сөзімен берсек: «Опыт последовательного применения разметки по членам предложения показывает, что эта теория имеет существенные недостатки, главными из которых можно назвать вложенность однопорядковых компонентов (напр., определение в обстоятельстве) и пересекаемость классификационных признаков (количество случаев теоретической неоднозначности доходит до 30%). Возможным решением этих проблем могло бы стать частичная разметка корпуса, то есть выделение только хорошо определяемых, «чистых» случаев, однако это привело бы к отсутствию разметки для значительной части синтаксических единиц. С другой стороны, часть описанных проблем неотъемлема от диффузной природы языка и будет, следовательно, возникать и при применении альтернативных синтаксических теорий» [32].

Синтаксистік белгіленімдерді толық автоматтандыру мүмкін еместігін И.С.Чардин «Лингвистические корпусы с разметкой на основе грамматики зависимостей и их применение при автоматическом синтаксическом анализе» деп аталатын кандидаттық диссертациясында былай деп тұжырымдайды: «Многие виды разметки требуют участия экспертов-лингвистов в подготовке корпуса. К ним относится и разметка синтаксическая, которая представляет собой особо трудоемкий вид разметки. Это связано с тем, что точность результатов предварительной автоматической синтаксической разметки далека от 100%, особенно при различении большого числа типов отношений» [33].

Синтаксистік белгіленімдер қоюда оның минимальді бірлігі сөз емес, синтаксистік бірліктер болады. Олар жай сөйлем немесе сөйлем бөліктері (предикативті бірліктер) немесе клаузалар болуы мүмкін.

Корпустық лингвистикада клауза дегеніміз - күрделі (құрмалас) сөйлем құрамындағы ең кіші сөйлем, яғни сөйлем бөліктері сөйлемдегі етістікке бағындырылады. Мұны орыс тілі корпустарында «вершина» деп атайды. Орыс тілі корпустары бойынша жазылған еңбектердегі пікірлерге сүйенсек, етістік болмаған жағдайда «вершинаны» етістік (баяндауыш) қызметінде тұрған сөйлемнің басқа мүшесі атқарады. Құрмалас сөйлемдердің сабақтаса және салаласа байланысуына қарай, клаузалар да сабақтас және салалас сөйлемдер құрайды. Ал орын тәртібі мен интонация арқылы берілетін клаузалар байланысы (шылаулар қатыспайтын) шылаусыз байланыс ретінде топтастырылады.

Синтаксистік модельдерде мәтінді клаузаларға бөлетін нұсқаны пайдалану шешімін тапқан, ал басқа позицияларды осы нұсқаларға жатқызуға болады. Құрылымдық-формалдық тұрғыдан клаузаларды былай көрсетуге болады. Мәтін ең алдымен абзацтарға бөлінеді. Ол формалдық тұрғыдан жаңа жол арқылы ажыратылады. Олар сөйлемдерге бөлінеді. Сөйлемдерді бөлшектеуге тыныс белгісі (нүкте) меже болады. Яғни нүкте арқылы және үлкен әріп дербес сөйлемдерді анықтайды. Ал жекелеген сөйлемдер ішкі бөліктерге бөлшектенеді. Сөйтіп, сол бөліктердің жиынтығы ретінде көрінеді. Алайда бұл бөліктердің кез келгені клауза болып танылмайды. Екі жағдайды ескеру қажет:

1) әрбір клауза бір сөйлемнің ішінде бірнеше бөліктен тұруы мүмкін.

2) Сөйлемнің кейбір бөліктері біруақытта бірнеше клауза-ларға енуі мүмкін.

Демек, корпустарды құрастыру және зерттеу барысында синтаксистік талдаудың моделі дайындалады. Бұл модель бойынша, мәтіндегі сөйлемдер сөйлем бөліктеріне бөлінеді. Ал бұл бөліктер клаузаларға топтастырылады. Клаузалардың реті маңызды болып табылады. Мысалы, 4 бөлікке бөлінген сөйлемді алып, *Part1, Part2, Part3, Part4* деп белгілейік. Егер біз *Part1* мен *Part3,* содан соң *Part3* пен *Part4* бөліктерін клаузаларға біріктірсек, онда тұтас *Part1-Part3-Part4* клаузасын аламыз. Егер *Part1* мен *Part3,* содан соң *Part1 мен Part4* бөліктерін клаузаға біріктірсек*, Part1-Part3* және

Part1-Part4 екі клаузасы құрылады. Сөйтіп, байланыс ұйымдасады. Осы алгоритмнің негізінде қандай да болмасын клаузаларды бөліп шығаруға болады делінеді.

Орыс тілінің корпустарына енгізілген синтаксистік белгіленімдер жоғарыда аталған «деревозависимость» деп аталатын сөздер арасындағы қатынасқа негізделген. Мұны кейде «грамматиказависимость» деп те атайды. Синтаксистік талдауларды автоматтандыруда синтаксистік қатынастарға негізделген түрінің тиімді екендігін И.С. Чардин да атап өтеді:

«При этом особое внимание уделяется корпусам с синтаксической разметкой на основе грамматики зависимостей. Подавляющее большинство исследований по автоматическому синтаксическому анализу за рубежом опирается на грамматики составляющих и корпусы, размеченные на их основе. Однако в последние годы все больше специалистов приходит к выводу, что использование грамматик зависимостей предоставляет более широкие возможности для разработки программ автоматического синтаксического анализа. С этим связано использование этого типа грамматик в качестве основы при подготовке многих вновь создаваемых корпусов с разметкой» [33].

Синтаксистік белгіленім қоюдың бұл ұстанымы алғаш Прага корпусынан бастау алады. Прага университетінде Карлов құрастырған чех мәтіндерінің аннотацияланған корпусы 1998 жылы жасалып, осы грамматикалық байланыс ұстанымына негізделген болатын. Алайда бұл корпуста синтаксистік қатынастардың майда топтары берілмей, негізінен ірі топтары ғана көрсетіледі. Ал орыс тілінің корпусында интернет мәліметтері бойынша, 78 синтаксистік қатынас типтері бар және чех корпусына қарағанда синтаксистік қатынастар терендетіліп берілген. Грамматикалық байланысқа негізделген мұндай корпустарға неміс (NEGRA) және голланд (Alpino) тілінің корпустарын атауға болады.

Орыс тілінің синтаксистік аннотация жасалған корпус мәтіндерін екі топқа бөлуге болады. Бірінші топ орыстың белгілі Упсаль корпусына енген көркем әдебиеттен алынған синтаксистік аннотация жасалған шамамен 10000 сөйлемнен тұрады. Екінші топ

қоғамдық-саяси, мәдени, экономика, ғылыми-техникалық стильдерден, интернеттен, интернеттегі жаңалықтардан әрқайсысынан 30 сөйлемнен алынған қысқа мәтіндерден тұрады. Мәтіндердің бұл екінші тобы орыс әдеби тілінің қазіргі жағдайын сипаттайды. Орыс тілі мәліметтері бойынша тағы да екінші топтағы стильдегі синтаксистік аннотация жасалған 2000 сөйлемнен тұратын мәтіндер енгізіліп жатқаны туралы мәлімет берілген.

Мәтін белгіленімдері ерекше маркерлер арқылы сипатталады. Оларды «теги» деп атайды. Тег-сөздер және тег-контейнерлерден тұрады. Бұлардың біріншісі мәтіндегі сөздерді сипаттаса, екіншісі мәтін бөліктерін сипаттайды.

Морфологиялық белгіленім қойылғанда мәтіндердегі жекелеген сөздер түбір мен қосымшаға ажыратылып, олардың грамматикалық сипаттамасы берілсе, синтаксистік белгіленім қою кезінде, жоғарыдағы орыс тілі корпустарында қолданылған «деревозависимость» принципине сүйенсек, мәтіндегі жекелеген сөйлемдердің құрамындағы сөздердің байланысы көрсетіліп, олардың арасындағы синтаксистік қатынастар сипатталады. Мысалы, *Оқушы мектептен шықты* деген жай сөйлемді алсақ, сөйлемдегі синтаксистік қатынасқа түскен сөздер:

Оқушы шықты.

Мектептен шықты.

Бұл тіркестердің біріншісі – субъект-предикатты қатынасқа негізделсе, екіншісі – жанама объектілік қатынасты көрсетеді. Ал сөйлемдегі *оқушы мектептен* тіркесінің арасында синтагмалық байланыс жоқ, сондықтан синтаксистік қатынас та жоқ. Себебі, «компоненттер арасындағы семантикалық байланыс оны құрайтын синтагмаларға, олардың тұлғалық ерекшелігіне тәуелді болады» [34].

Қазақ тіл білімінде әзірге синтаксистік белгіленім әзірлемесі жасалу үстінде болғандықтан, оны әзірлеудің шарттары ғана жоғарыда сөз болды.

2.4.7. Синтаксистік белгіленімнің синтаксистік қатынас теориясымен сабақтастығы

Қазақ тілінде мәтіндерге автоматты түрде синтаксистік белгіленім қоя алатын тілдік корпустарды жасау үшін тілдік жүйедегі синтаксистік талдауларды нақты формалдық жүйеге түсіру керек. Өйткені компьютер тілдік бірліктерді жадқа салынған формалдық белгіленімдерге қарай ажыратады. Сондықтан синтаксистік белгіленімдер қоюдағы принциптік ұстаным болып табылатын сөздер арасындағы синтаксистік қатынастардың тұрақты жүйесін модельге түсіру қажет. Ол үшін ең алдымен синтаксистік қатынас мәселесінің грамматикадағы тілдік табиғатын айқындау өзекті.

Синтаксистік қатынас дегеніміз – құрамындағы компоненттерінің арасындағы мағыналық байланыс болып табылады. Синтаксистік қатынастар жай сөйлем немесе құрмалас сөйлем құрамындағы жеке сөздердің арасында да, сол сияқты құрмалас сөйлем құрамындағы жай сөйлемдер арасында да көрініс табады. Синтаксистік қатынастар арқылы сөз бен сөз, сөйлем мен сөйлем байланысады. Сондықтан синтаксистегі сөйлемді модельдеуге қажетті негізгі табан тірейтін ұғым да осы – синтаксистік қатынастар болып табылады.

Синтакситік қатынастарадың синтаксис саласына қатысты ең негізгі ұғым ретінде қарау қажеттігі дау тудырмайды. Өйткені тіл білімінің әрбір саласының өзіне тиесілі зерттейтін объектілері бар. Фонетика – тіл дыбыстарын, дыбыстардың тіркесімін (морфонология), лексика – сөздің, сөз тіркестерінің мағынасын, морфология сөздің құрам-құрылысын, сөздің түрленімін зерттесе, синтаксистің зерттеу объектісі – сөздің тіркесі, сөздердің байланысы, сөйлем. Ал бұл тілдік бірліктердің болмысы синтаксистік қатынастар арқылы ғана айқындалады. Мысалы, матаса байланысу – анықтауыштық, яғни атрибутивтік қатынасты, меңгере байланысу – объектілік қатынасты, қиыса байланысу – предикативтік қатынасты сипаттайды.

Синтаксистік қатынастар мәселесі тіл білімінің синаксис саласында қазақ тілін зерттеудің алғашқы кезеңдерінен бастап-ақ сөз бола бастаған.

Т.Сауранбаев С.Аманжоловпен бірігіп жазған «Қазақ тілі грамматикасында» синтаксистік қатынас ұғымын сөз тіркесіне қатысты қарастырған. «Бұл еңбегінде автор сөз тіркестерін «синтаксистік қатынастар» деп атай отырып, оларды субъекті, предикаттық, объектілік, атрибуттық және сипаттауыштық қатынас деп бірнеше түрге бөледі. Автор бастауыш пен баяндауыш арасындағы қатынасты – субъекті-предикаттық қатынас деп атаса, ал пысықтауыш сөздердің етістікпен байланысын – объектілік, ал анықтауыш пен анықталатын мүшелерінің байланысын – атрибуттық қатынас деп бөледі [35].

Синтаксистік қатынастың бір түрі болып табылатын объектілік қатынас мәселесі қазақ ғалымдары еңбектерінде алдымен тура толықтауышқа қатысты сөз болса, осы мәселені арнайы зерттеген А.Қ. Нұрмұханбетованың кандидаттық диссертациясында объектілік қатынастың берілу жолдары дәстүрлі грамматикадағы жанама толықтауыш мүше ұғымына сәйкес жанама объектілік қатынас тұрғысынан қарастырылады. Зерттеуші бұл зерттеуінде объектілік қатынастың атау және ілік септігі арқылы берілмейтіндігін атап көрсетеді, сондай-ақ тура объектілік қатынастың табыс септігі арқылы, ал жанама объектілік қатынастың барыс, жатыс, шығыс, көмектес септігі арқылы берілуіне арнайы тоқталады. Функционалды өріс тұрғысынан табыс септігінің объектілік қатынасты білдірудің ядросы, ал қалған септіктер жанама тілдік құралдары екендігін тұжырымдайды [35].

Н.Т.Сауранбаев еңбектерінде объект категориясы объект пен предикат арасындағы синтаксистік қатынас деп аталып, оның септік жалғаулары мен демеуліктер арқылы берілетіндігі айтылады [36]. Мәселен, ғалымның айтуынша, *«Төсекке жатып, ұйқыға кетер алдында, Майбасарды қайта-қайта есіне алып, оның Ділдә аулына барып қайтқанын ойлап, Абай ең алғаш рет өз қалыңдығы туралы құпия, ыстық бір сезімді сезгендей болды»* (М.Әуезов) деген сөйлемдегі *төсекке жатып, ұйқыға кетіп, есіне алып, аулына барып, қайтқанын ойлап, Ақтөбеге дейін, кеме жасау туралы, үйрену үшін* деген тіркестер объектілік қатынасты білдіреді [36].

Синтаксистік қатынас теориясына қатысты мәселені Н.Т. Сауранбаев «деп» сөзформасына қатысты да қарастырып,

оны жалғаулық ретінде таниды. *Деп* жалғаулығын жалғаулық қатарына жатқызып, оның құрмалас сөйлемдер құруға қатысатынын айтады.

Н.Т.Сауранбаев *деп* жалғаулығы арқылы берілетін объектілік қатынасты мынадай мысалдармен көрсеткен: *Хабар келе ме деп, күн сайын күтеді. Осы алдап қолға түсіргелі жүр ме деп ойлаған Амантай уәде берген жоқ* [37].

Р.С.Әмір және Ж.Р.Әміровалардың еңбектерінде де *деп* жалғаулығының жалғаулықты салалас құрмалас сөйлемдер құрамындағы мағыналық қатынасты білдіретіндігі айтылады.

Себептік қатынас: *Асан бұзылады деп, біз де бұзыламыз ба?* (Б.Майлин);

Объектілік қатынас: *Басқаның реті келмес деп, өздері де ойлап отыр* (Б.Майлин) [38].

Меншіктік қатынас ұғымы қазақ тіл білімінде изафеттік тіркес тұрғысынан кеңірек зерттелген. Синтаксистік қатынастың бұл түрі де Н.Т.Сауранбаев еңбектерінде лексикалық және синтаксистік изафет деп аталып, атрибуттар категориясы, яғни анықтауыштық қатынаста тұрған сөздер тұрғысынан қарастырылған. Ғалым еңбектеріндегі лексикалық изафеттердің біріншісі – әуелде меншіктік қатынаста жұмсалып, бірігіп атауға айналып кеткен сөздер (*желбау, баскиім, қаламұш*), екіншісі – ілік жалғауы түсіріліп айтылатын күрделі атаулар (*ауыл шаруашылығы, Арал теңізі, мектеп үйі* т.б.). Ғалым синтаксистік изафетті синтаксистік мағынада ғана жұмсалатын, яғни ілік жалғауы түсірілмей қолданылатын топ деп көрсетеді. Ғалымның пайымдауынша, бұл топ құрмалас сөйлемдер мен жай сөйлемдердің құрамында жеке сөйлемдер мен жеке синтагмаларды бір-бірімен матастырушы, қабыстырушы тәсіл есебінде жұмсалады [36].

Меншіктік қатынас ұғымы қазақ тіл білімінде көптеген зерттеушілер еңбектерінде жалғасын тапты. Мәселен, зерттеушілердің бірі қазақ тілінде меншіктілік қатынасты Р.Әмірдің жетекшілігімен арнайы қарастырған С.К.Нұрпейісова меншіктілік қатынасты білдіретін негізгі тұлға – ілік септік жалғауы, тәуелдік жалғау көмекші қызмет атқарады деп көрсетеді. Сондай-ақ меншіктілік қатынастың негізгі және басты объектісі – тәуелдеу категориясы

деп, функционалды грамматика тұрғысынан осы мағыналық қатынасты білдіретін негізгі тілдік құрал ретінде таниды [39].

Сол сияқты құрамалас сөйлемдердегі жай сөйлемдердің арасындағы мағыналық қатынас, оларды мағыналық топқа бөлудің критерийі ретінде де ғылымда орныққан. Мысалы, мезгіл құрмаластар – мезгілдік қатынасқа, – қарсылықты құрмаластар – қарсылықты қатынасқа негізделеді.

Синтаксистік қатынас мәселесін сөз еткен ғалымдар оның сөз тіркесінің ғана емес, құрмалас сөйлемдер құрамындағы жай сөйлемдер арасындағы мағыналық байланысқа қатысты да қолданады. Бұл туралы Н.Т.Сауранбаев: «Жай сөйлемге енген жеке сөйлемдердің қызметі мен құрмалас сөйлемге енген жеке сөйлемдердің қызметі бірдей. Дәлірек айтқанда, құрмаласқа енген бағыныңқы сөйлемдердің қызметі жай сөйлемдегі тұрлаусыз мүшелердің қызметімен бірдей. Осыдан барып құрмаласқа енген жай сөйлемдердің өзара байланысу тәсілі жай сөйлем мүшелерінің байланысу тәсілімен бірдей, жай сөйлемдегі мүшелер қандай синтаксистік қатынаста болса, олар қандай грамматикалық тәсілмен байланысса, сабақтас құрмаласқа енген жеке сөйлемдер де сондай синтаксистік қатынаста тұрып, сол тәсілдер арқылы байланысады», – деп, ғалым бұл құбылысты «тәсілдес принцип» деген терминмен атайды.

Жоғарыдағы ғалым көрсеткен синтаксистік қатынастар сөйлем мүшелерінің атқаратын қызметімен байланысты қарастырылған. Мысалы, ілік септігінде тұрған сөздің анықтауыштық қызметіне сәйкес – атрибуттық қатынас, барыс, табыс, жатыс, шығыс, көмектес септігінде тұрған сөздердің толықтауыштық қызметіне сәйкес – объектілік қатынас. Құрмалас сөйлемдерді жіктеуге алғаш негіз болған сөйлем мүшелерінің қызметіне сәйкес топтастыру сөз болып отырған «синтаксистік қатынас» ұғымымен сабақтас. Қазақ тіл білімінде мұндай жіктеме А.Байтұрсынұлы, Қ.Жұбанов, сондай-ақ Н.Т.Сауранбаевтар классификациясына негіз болған.

Міне, көріп отырғанымыздай, синтаксистік қатынас ұғымы – синтаксистің ең негізгі ұғымдарының бірі және оны қазақ ғалымдары ертеден-ақ таныған. Оның сөздер арасындағы тіркесім-

ге, сонымен қатар сөйлемдер арасындағы мағыналық байланысқа негіз болатын ұғым екенін көре білген. Синтаксистік қатынастар синтаксистік талдаулар жасаудың ең негізгі бірлігі болып табылатындықтан, кейбір ғалымдар еңбектерінде құрмалас сөйлемдерді жіктеуге негіз болатын ұстаным ретінде алынған және ол қазіргі жаңа функционалды бағыттардың талаптарына сәйкес келеді. Осы ретте белгілі синтаксис мамандарының бірі профессор Т.Ермекова құрмаластарды жіктеуде синтаксистік қатынастарға негізделген ұстанымды басшылыққа алу керектігін үнемі айтып жүр. Демек, синтаксистік талдаудың ең бір маңыздысы осы синтаксистік қатынастар (*субьект предикаттық қатынас; анықтауыштық (атрибутивтік қатынас); меншіктік қатынас; тура обьектілік қатынас; жанама обьектілік қатынас; пысықтауыштық қатынас; мезгілдік қатынас; шарттылық қатынас; себептік қатынас,; қарсылықты қатынас* т.б.) бойынша жүргізу қажеттігі дау туғызбайды.

Орыс және басқа да шетелдердегі корпустар құрастыруда лингвистикалық аннотация жасаудағы ең күрделісі болып табылатын синтаксистік белгіленімдер моделін құрастыру қазақ тілінің ұлттық корпустарын құрастыруда аса қажет. Ол үшін қазақ тіл біліміндегі жоғарыда сөз болған синтаксистік қатынастар теориясын егжей -тегжейлі зерттеп, тілдік табиғатын аша түсуді қажет етеді.

2.5. Семантикалық белгіленім

Кез келген қазақ тіліндегі мәтіндер корпусы аннотациямен (шартты түрдегі грамматикалық белгілермен) жабдықталады, яғни кез келген аннотацияланған корпус, оның лингвистикалық сипатының көрінісін айқындай алады. Әдетте ондай сипаттама корпус құрамындағы бірліктерге қатысты мүмкін болатын тілдік талдау үлгісін қамтуы керек. Аннотацияланған корпус, оны пайдаланудың қарапайымдылығы мен көпфункционалдық қасиеттеріне ие болуымен де аса құнды. Көпшілік жағдайда, корпустарды аннотациялау әрекетін автоматты және «қолмен» жүргізу түрлері қатар жүргізіледі. Ал автоматтанған аннотациялау әрекетін қанағаттандырарлық деп санау үшін қателік шегі үш пайыздан аспау керектігін ескеру қажет болады.

Семантикалық тэгтер (шартты белгілер), көпшілік жағдайда, қарастырып отырған сөздің немесе сөзтіркесі жататын семантикалық категорияны және мағыналық реңкке ажырайтын одан да шағын категорияларды білдіреді. Корпустарды семантикалық деңгейде белгіленім жасау сөз мағынасын айрықшалау, омонимия мен синонимия мәселелерін шешу, сөздерді категориялау (топтау), тақырыптық топтарды, каузативтік белгілерді, сөздердің бағалауыштық және таралым (деривациялық) сипатын анықтау және т.б мәселелерді қарастырады.

Семантикалық белгілер дегеніміз – сөздердің әртүрлі сөз табына жататындығына қарай іштей лексика-семантикалық кластарға бөлінуі.

Орыс тілінің Ұлттық корпусын құрастырушылар сол тілге тән семантикалық белгіленім нұсқасын ұсынады. Бұл корпуста әр сөзформаға үш түрлі белгі тіркеліп жазылады:

1. Разряд (жалқы есім, өздік есімдік және т.б.);

2. Лексика-семантикалық сипаттамалары (лексеманың тақырыптық тобы, каузативтік белгілері, бағалаулар және т.б.);

3. Деривациялық (таралымдық, сөзжасамдық) сипаттамалары («диминутив», «сын есімнен жасалған (отадъективті) үстеу және т.б.).

Лексика-семантикалық тэгтер мынадай топтарға жіктелген:

– таксономия (лексеманың тақырыптық тобы) – зат есімге, сын есімге, етістіктер мен үстеулерге;

– мереология («бөлік – бүтін», «элемент – жиын») – заттық және заттық емес атауларға;

– топология (белгілейтін нысанның топологиялық статусы) – заттық атаулар үшін;

– каузация – етістіктер үшін;

– қызметтік статус – етістіктер үшін;

– бағалау – заттық және заттық емес атаулар үшін, сын есім мен үстеулер үшін.

Сөзжасамдық сипаттамалар бірнеше типке бөлінеді:

• морфо-семантикалық сөзжасамдық белгілер (мысалы, «каритив», «семельфактив»);

• туындайтын сөздің разряды (мысалы, етістіктен жасалған зат есім немесе адъективтен (сын есімнен) жасалған үстеу);

• туындайтын сөздің лексика-семантикалық (таксономиялық) типі (мысалы, сын есімнен жасалған үстеу);

• сөзжасамның морфологиялық типі (субстантивтену, күрделі сөз) (толығырақ қара: http://ruscorpora.ru, «Семантика» бөлімі).

Қазақ тілінде де сөздердің семантикалық топтарға қатысын корпустар мәтіндерінде аннотациялау үшін тіл жүйесіндегі әрбір сөз табының лексика-семантикалық топтарын жіктеп, тілдегі сөздердің барлығын осы топтарға қатысына қарай сипаттап базаға салу керек. Сол үшін әрбір сөз табының лексика-семантикалық топтарын анықтап, сол сөз тобына жататын сөздердің тізімін жасау қажет. Мәселен, зат есім, етісті, сын есім сөздердің лексика-семантикалық топтары төмендегідей болып келеді.

2.5.1. Қазақ тілі лексикасын семантикалық классификациялау ұстанымдары

Сөз – тілдік функциялардың барлығына қатысты болып келеді. Ең алдымен заттар мен құбылыстарды, олардың белгісі т.б. сипаттарын атап, номинативтік қызмет атқарады. Сонымен қатар сөздің сигнификатты, коммуникативті, прагматикалық қызметтері бар. Әсіресе зат есім, етістік, сын есімнің аталымдық және сипаттамалық функциялары күшті болып келеді. Сондықтан да олар сөздік құрамның негізгі бірліктері болып табылады.

Қазақ тілі лексикасын семантикалық классификациялауда, яғни сөздердің лексика-семантикалық топтары мен субтоптарын анықтау күрделі мәселе. Бұл әлемдік тіл білімінде де бір шешімін таппаған деуге болады. Сөздердің лексика-семантикалық топтарын жіктеу және олардың саны кейде зерттеу мақсатына, лингвистикалық мектептерге және басқа да факторларға байланысты болып келеді [79, 70-72 бб.].

Қазақ тілінде семантикаға қатысты жазылған еңбектер де көп емес. Бұл ретте М. Оразовтың «Қазақ тілінің семантикасы» [40; 41] және т.б. ғалымдардың қазақ тілі семантикасына қатысты жазған еңбектерінің маңызы зор. Қазақ және орыс тілдері материалдары

негізінде тілдік бірліктердің семантикалық қырларын салғастырмалы аспектіден қарастыратын Э.Д.Сүлейменованың еңбектерін атауға болады [42]. Сонымен қатар семантика мәселесі жас ғалымдар зерттеулерінде жалғасын тапты [43, 44, 45, 46, 47, 48, 49, 50].

Қазақ тілі лексикасын семантикалық классификациялауда үш түрлі ұстанымға сүйенуге болады [51; 52, 53]:

1) денотативті-десигнативті;

2) парадигматикалық;

3) синтагматикалық.

Денотативті-десигнативті ұстаным ақиқат дүниенің онтологиялық бөлінісін негізге алады. Бұл ұстаным бойынша сөздер ақиқат дүниені білдіруі мен белгілеуіне қарай топтарға біріктіріледі. Мысалы, жан-жануарлар атаулары, өсімдік атаулары, заттардың белгілі бір түрлерінің атаулары және т.б.

Парадигматикалық ұстаным бойынша сөз мағынасы құрамындағы белгілі бір мағыналық компонентіне қарай топтастырылады. Мысалы, «модальды етістіктер», «каузативті етістіктер» және т.б.

Синтагматикалық ұстаным бойынша сөздер тіркесімдік саны мен сапасына қарай топтастырылады [54; 53, 189-192 бб.].

Біздің зерттеуде сөздерді семантикалық классификациялау негізіне денотативті-десигнативті ұстаным басшылыққа алынды. Алайда кейбір жағдайларда парадигматикалық және синтагматикалық ұстанымдар да ескеріліп отырды.

2.5.2. Зат есімнің семантикалық классификациясы

Қазақ тіліндегі зат есімдерді жекелеген лексика-семантикалық топтарға бөлудің негізгі крийтериі басқа тілдердегідей мағыналық құрамында денотат пен сигнификаттың көрініс табатын номинативтік мағынасымен тікелей байланысты.

А. А. Уфимцева, «Типы словесных знаков» деп аталатын еңбегінде сөздің таңбалық мағынасының компоненті ретінде сигнификатты алты ұғымға біріктіреді [55, 102 б.].

1. Денотаты анық зат атауларын атайтын нақты ұғымдар. Мысалы, *қала, жел, кітап* және т.б.

2. Әрекетті, белгіні, қатынасты білдіретін денотаты нақты емес

абстракты ұғымдар. Мысалы, *данышпандық, ер жету, надандық, қараңғылық, суықтық, жан тыныштық* және т.б.

3. Ойлау категорияларын білдіретін абсолютті, жалпы ұғымдар. Мысалы, *белгі, нышан, қасиет, сан.*

4. Ақиқат дүниенің категорияларын білдіретін абстракты ұғымдар. Мысалы, *сезім, оқиға, қозғалыс, әрекет, сүйықтық* және т.б.

5. Ирреалды заттардың көрінісі болып табылатын жекелеген ұғымдар. Мысалы, *кентавр, леший, русалка.* Қазақ тілінде:*диуана, пері, жын-шайтан* және т.б.

6. Әртүрлі білім (ғылым, техника, өнер) салалары терминдерін білдіретін метатілдік ұғымдар.

Бұл классификациядан байқайтынымыз, денотативті және денотативті-сигнификативті мағыналы есімдер көбінесе нақты лексиканы білдірсе, ал сигнификатты және сигнификатты-денотативті мағынағы негізделген сөздер абстракты лексиканы жасайды. Екіншіден, есім лексиканың семантикасында денотат пен сигнификатты мағына міндетті түрде болады. Егер сөздік таңба арнайы атау, яғни термин болса, денотат сигнификаттан басым болады, керісінше, егер сөздік таңба абстракты ұғым атауы болса, сигнификат денотаттан басым болады немесе сөздік таңба таза сигнификатты мағынаға ие болады [55,103 б.].

Қазақ тілінде де басқа да тілдердегідей зат есімдер заттар мен ұғымдардың әртүрлі атаулары болып келеді. Олар: жалқы есімдер: кісі есімдері, қала, өзен, көл т. атаулары; жалпы есімдер: адам атаулары (кісі, ана, әйел); жануарлар атаулары; өсімдік атаулары; әртүрлі заттар мен нәрселер атаулары; дерексіз ұғымдар атаулары (ел, ру, тағдыр); ғылыми, техникалық терминдер атаулары; іс-әрекет, құбылыс, қалып атаулары (жүріс, хабар, ұйқы); іс-әрекет нәтижелері атаулары (жиналыс, кеспе); құрал атаулары (бұрғы, сыпырғыш, ара) және т.б.

Түбір зат есімдер қазақ тілінде негізінен түбірге сәйкес келеді. Бұл бір жағынан сөздерді автоматты бөлшектеуде де жеңілдік тудырады. Зат есімдер дербес лексика-грамматикалық категория болып табылады

М. Оразовтың тұжырымдауынша, қазақ тіліндегі зат есімдердің ішінде жалпы зат есімдер мағыналық жағынан күрделі болып

келеді. Мұндай зат есімдер мағыналық жағынан өзгермелі, кейде денотатитвтік мағына, кейде сигнификатты мағына басым болады, ал кейді екеуі жарыспалы түрде тең дәрежеде қолданыс табады. Сонымен қатар қазақ тілінде ауыспалы мағынада сөздер жиі қолданылады.

Сандық жағынан тақырыптық топтар басқа тілдердегі сияқты қазақ тілінде де өте көп. М. Оразов «Қазақ тілінің семантикасы» еңбегінде мұндай тақырыптық топтардың 8 түрін көрсетеді. Олар:

1) *туыстық атаулар;*
2) *дене мүше атаулары;*
3) *төрт түлік мал атаулары;*
4) *киім-кешек атаулары;*
5) *құс атаулары;*
6) жабайы аң атаулары;
7) *ыдыс-аяқ атаулары;*
8) *адам баласына тән ауру атаулары* [53, 191 б.].

А.Жұбанов докторлық диссертациясында («Основные принципы формализаии содержания казахского текста» монографияда) [79, 72-77 бб.] зат есімдердің семантикалық топтарын төмендегіше көрсеткен болатын (***12-кесте***). Ғалым классификациясында зат есімдердің ірі семантикалық топтарын жіктеуге денотативті-десигнативті ұстаным, ал ұсақ лексика-семантикалық топтарды (дене мүшелері, әскери атақтар, апта атаулары) бөлуге парадигматикалық ұстаным негіз болған.

12-кесте

Қазақ тілі мәтіндерін туындату моделінде қолданылатын
зат есімнің семантикалық топтары

Ақиқат дүниенің сегменттері	Семантикалық топтың атауы	Топ коды (шартты белгісі)
1	2	3
1. Ғылым	Жалпыфилософиялық ұғымдар	ЗТ101
	Жалпығылыми ұғымдар	ЗТ102
	Жалпымедициналық ұғымдар	ЗТ103
	Психологияның негізгі ұғымдары	ЗТ104

1	2	3
	Нейропсихологияның негізгі ұғымдары	ЗТ105
	Логиканың негізгі ұғымдары	ЗТ106
	Геометрияның негізгі ұғымдары	ЗТ107
	Негізгі географиялық ұғымдар	ЗТ108
	Тіл білімінің негізгі ұғымдары	ЗТ109
	Әдебиеттанудың негізгі ұғымдары	ЗТ110
	Бейнелеу өнерінің негізгі ұғымдары	ЗТ111
	Музыка өнерінің негізгі ұғымдары	ЗТ112
	Заң мен қылмыс салаларының негізі ұғымдары	ЗТ113
	Театрға қатысты негізгі ұғымдар	ЗТ114
	Авиацияғақатысты негізі ұғымдар	ЗТ115
	Құрылысқа қатысты ұғымдар	ЗТ116
	Механикалық нысанның жеке құрылысына байланысты ұғымдар	ЗТ117
	Ақпараттық ұғымдар	ЗТ118
	Жинақтық ұғымдар	ЗТ119
	Электроника менэлектротехниканың негізгі ұғымдары	ЗТ120
	Әкімшілік ұғымдар	ЗТ121
	Мекемелердің, газеттердің, журналдардың және т.б. аттары	ЗТ122
	Абстракт есімдер	ЗТ123
	Баға	ЗТ124
2. Табиғат	Табиғат заттары	ЗТ201
	Табиғи құбылыстар	ЗТ202
	Табиғат бұйымдары – артефактар	ЗТ203
	Табиғаттың қалыптары	ЗТ204
	Заттар құрылысы мен оның бөліктерін сипаттайтын ұғымдар	ЗТ205
	Заттар, металдар	ЗТ206
	Табиғат заттарының әрекеттері	ЗТ207

1	2	3
3. Адам және оның әлемі	Адам және оның дене мүшелері	ЗТ301
	Адам беті мен басының бөліктері	ЗТ302
	Адамға тән күй мен сезім	ЗТ303
	Адамға тән әрекет	ЗТ304
	Адамдарды сипаттау үшін қолданатын ұғымдар	ЗТ305
	Адамның киімі мен аяқ киімдеріне қатысты заттар және оларға байланысты ұғымдар	ЗТ306
	Гигиеналық, әшекейлік және т.б. мақсатта адам қолданатын артефакт-заттар	ЗТ307
	Адам орын ауыстыруында қолданатын нысандар және оларға қатысты ұғымдар	ЗТ308
	Адам пайдаланатын азық-түлік заттары	ЗТ309
	Адамның тұрғын үйге қатысты ұғымдары	ЗТ310
	Адам өлімі мен жерлеуге қатысты ұғымдар	ЗТ311
	Адамдардың аты, әкесінің аты мен тегі	ЗТ312
	Адамның ұлты	ЗТ313
	Адам қолданатын ыдыс-аяқтар және оларға байланысты ұғымдар	ЗТ314
	Адамдардың демалыс кезінде, аңшылық, балық аулау кезінде пайдаланатын заттары және оларға қатысты ұғымдар	ЗТ315
	Мамандық, лауазым, ғылыми атақ, дәреже атаулары	ЗТ316
	Оқу мен еңбек ету және т.б. кезінде пайдаланатын артефакт-заттар	ЗТ317
	Адамдардың арақатынасына қатысты ұғымдар	ЗТ318
	Адамдар арасындағы туысқандық қатынастарға байланысты ұғымдар	ЗТ319
	Адам марапаттары	ЗТ320

1	2	3
4. Жануарлар әлемі	Ірі қара мал және олардың дене мүшелері	ЗТ401
	Ұсақ үй жануарлары және олардың дене мүшелері	ЗТ402
	Ірі жабайы аңдар және олардың дене мүшелері	ЗТ403
	Жәндіктер және олардың дене мүшелері	ЗТ404
	Құстар және олардың дене мүшелері	ЗТ405
	Жануарларға тән әрекеттер	ЗТ406
	Жануарлар, құстар мен жәндіктердің қорегіне қатысты заттар	ЗТ407
5. Өсімдіктер	Өсімдіктер және оның бөліктері	ЗТ501
	Ағаштар және оның бөліктері	ЗТ502
6. Уақыт	Уақытқа қатысты ұғымдар	ЗТ601
	Апта күндері	ЗТ602
7. Дін	Дінге қатысты ұғымдар	ЗТ701
8. Соғыс және оған қатысты ұғымдар	Соғыс техникасының заттары, қару-жарақ және оның бөліктері	ЗТ801
	Әскери мамандықтардың атаулары	ЗТ802
	Әскери атақтар	ЗТ803
	Соғыс кезіндегі көмек көрсетуге байланысты ұғымдар	ЗТ804
	Әскери бөлімшелердің аттары	ЗТ805
	Соғысқа байланысты ұғымдар	ЗТ806
9. Оқу мен еңбек	Оқыту мен еңбекке қатысты мекемелер	ЗТ901
	Оқыту мен еңбекке қатысты ұғымдар	ЗТ902

1	2	3
10. Өлшемдер	Ұзындық, салмақ, жылдамдық және т.б. өлшемдер	ЗТ911
11. География	Елді мекендердің аттары	ЗТ921
	Елдер, материктер, жер шары бөліктері, аралдар, таулардың және т.б. аттары	ЗТ922
	Өзендер, теңіздер, көлдер, бұғаздардың аттары	ЗТ923

12-кестенің соңғы бағанасында зат есім сөздерінің семантикалық топтарының туындау моделінде қабылданған *зат есім* сөзі орнына шартты қысқартылған ЗТ белгісі алынды.

2.5.3. Қазақ тіліндегі етістіктердің семантикалық классификациясы

Зат есімдер заттар мен құбылыстарды білдірсе, етістіктер зат есім (сөздер) арасындағы қатынастарды білдіреді. Сондықтан зат есімдер мағыналары абсолютті және дербес болып келеді, ал етістіктер мағыналары қатыстық тұрғыдан алынады және олардың мағыналары іс-әрекеттің объекті мен субъектіге синтагматикалық қатынасын білдіреді.

Етістік лексемалардың мағынасын айқындау күрделі мәселе. Олар мынадай үш фактормен анықталады:

1) ақиқат шындыққа сәйкестігі;

2) етістіктермен тіркесетін заттық атаулардың семантикасы;

3) әрекет пен оның субъектісі және объектісі арасындағы мағыналық қатынастар типтері [55, 113].

Етістіктердің зат есімдерден айрмашылығы етістікке қатысты лексемалар топтарын классификациялауда олардың таңбалық мағынасы тұрғысынан жіктеу біршама қиындық тудырады.

Етістіктер семантикасы бойынша дәстүрлі тұрғыдан үш топқа бөледі:

1) амал-әрекет етістіктері;

2) қалып етістіктері;

3) қимыл-қозғалыс [55, 120].

Әрбір лексика-семантикалық топтың парадигматикалық параметрлері мен семантикалық ерекшеліктері бар. Мысалы, қозғалыс етістіктері көптеген тілдерде жақсы зерттелген және мынадай семантикалық сипаттары анықталған: бағытталған әрекет, әрекет жылдамдығы, қозғалу сипаты, құрал немесе қозғалыс ортасы және т.б. Ал эмотитві етістіктер семантикасында басқа параметрлер негізге алынады. Онда әсер ету, қабылдау, бағалау, субъективтік көзқарасты білдіру етістіктері бар [55, 148 б.].

М. Оразов етістіктің лексикалық мағынасының нақтылығы оның денотативті және сигникативті мағыналарына қатысты дегенді айтады [53, с.99]. Етістіктердің денотативті мағыналары туралы сөз болғанда іс-әрекеттің өту мерзімі, өту орны, әрекетті орындаушы субъектінің қатысы, сол сияқты амал-әрекеттің мақсаты сияқты басқа да элементтері болады. Мысалы: *Шаңғышылар селоға кіргенде көз байланып та қалған еді(М. Сүндетов)* – деген сөйлемдегі «кір» етістігінің мағыналық құрылымында мынадай семалар бар:

1) қозғалыстың болуы;

2) бірнеше заттың немесе адамның (субъект) қозғалысқа қатысы;

3) әрекеттің жерде өтуі;

4) белгілі бір кеңістік шегінен өтуі;

5) ішке ену семасы.

Осы *кір* етістігі басқа сөздермен тіркескенде мүлде басқа мағынаны білдіруі де мүмкін. Мысалы, *Мен 1962 жылы партия ұйымына кірдім*. Бұл жерде *кір* етістігі қозғалысты емес, партия ұйымына мүше болу деген мағынаны білдіріп тұр. Соған сәйкес *кір* етістігінің мағыналық құрылымы да өзгеріске түседі. Сол сияқты қозғалыс етістіктері кейбір тіркестерде қарым-қатынас етістігіне жатса, енді бірде ойлау етістіктерінің құрылымына енеді. Мысалы, *үйге келу, қарсы келу,ойға келу*, *тілге келу* және т.б.

Сөздерді тақырыптық, лексика-семантикалық топтарға жіктеуде олардың жалпылауыш семалары негіз болады. Бұл мәселеде үндіеуропа тілдерінде жеткілікті зерттеу тәжірибесі болған-

мен, қазақ тілінде мұндай зерттеулер аз. Бұл мәселе қазақ тілінде М.Оразов еңбектері мен кандидаттық диссертацияларда (көбінесе салғастырмалы аспектіден) қарастырылады.

А.Жұбанов «Основные принципы формализации содержания казахского текста» атты докторлық диссертациясында [79, 78-88 бб.] етістік сөздерді төмендегіше лексика-семантикалық топтарға жіктейді (***13-кесте)***.

13-кесте

Қазақ тілі мәтіндерін туындату моделінде қолданылатын етістіктің семантикалық топтары

Етістіктің семантикалық субтобы	**Мысалдар**	**Код (шартты белгі)**
1	2	3
Семантикалық топ: 1. Болмыс етістігі		
1.Болмыс етістіктері: *Бар*	*Менде жақсы оқуға мүмкіншілік* ***бар****, бірақ оған түк ынтам* ***жоқ.***	ЕТ101
Жоқ	*Оның жасы әлі кәмелетке толған* ***жоқ.***	ЕТ102
Бол (болу, болмау)	*Сөзіңді айтып* ***бол****.* *Кеше жиналыс* ***болды***	ЕТ103
Е (еді, екен, емес)	*Алматы Қазақстанның астанасы* ***еді.*** *Өзі бір қызық адам* ***екен.*** *Оның аты қара-торы* ***емес.***	ЕТ104
Семантический топ: 2. Физикалық әрекеттер етістігі		
1. Физикалық өзгерістер етістігі: жасау және өндіру, біріктіру және бөліктеу және т.б.	*Жасау, соғу, тігу, тұрғызу, салу, бөлшектеу, екшеу, ыдырату* және т.б.	ЕТ201
2. Белсенді физикалық әрекеттер етістігі: қирату және бүлдіру	*Қирату, бүлдіру, бұзу, сындыру, құлату* және т.б.	ЕТ202

1	2	3
3. Эмоционалды физикалық әрекеттер етістіктері	*Аузы ашылу, мәз болу, күлу, ыржиу, шапалақтау, секіру (қуанғаннан)*	ЕТ203
4. Адам мен жануарлар әлеміне тән тамақ ішу әрекетін бейнелейтін етістіктер	*Ішу, жеу, ұрттау, жұту, асау, дәм тату, сорапта*у және т.б.	ЕТ204
5. Адам мен жануарлар әлеміне тән басқа да әрекеттерді бейнелеуде қолданатын етістіктер	*Жүгіру, секіру, ырғу, құсу, таптау, жылжу, қыймылдау* және т.б.	ЕТ205
6. Өлі табиғатты бейнелейтін физикалық әрекеттер етістіктері	*Күркіреу, сарқырау, арсылдау,сілкіну* және т.б.	ЕТ206
7. Өлі табиғатқа тән басқа да физикалық әрекеттер етістіктері	*Ағу, бату (күн), құлау, қайнау, діріл*деу, ысу, суу, *гүрсілдеу* және т.б.	ЕТ207
8. Қоршаған ортаның дыбыстарына қатысты етістіктер, дыбысбейнелеу етістіктері	*Тарсылдау, сатырлау, сарқырау, ызыңдау, гүрсілдеу, қоңқылдау* және т.б.	ЕТ208
9 Жануарлар әлеміне ғана тән физикалық әрекеттер етістіктері	*Ұлу, ү*ру, *ырылдау, мөңку, мөңіреу, жорту, шабу* және т.б.	ЕТ209
Семантическалық топ: 3. Қозғалыс етістіктері		
1. Жер бетінде болатын қозғалыстарды білдіретін етістіктер	*Жүру, жүгіру, еңбектеу, домалау,сырғанау* және т.б.	ЕТ301
2. Су үстінде (су ішінде де) болатын қозғалыс етістіктері	*Жүзу, сүңгу, малту, қалқу* және т.б.	ЕТ302
3 Ауада болатын қозғалыстарды білдіретін етістіктер	*Ұшу, самғау, қалықтау, шүйлігу, зымырау* және т.б.	ЕТ303
4. Қозғалыс бағытын білдіретін қозғалыс етістіктер і	*Өрлеу, ылдилау, келу, кету, жақындау, ұзау*	ЕТ304
5. Ретсіз қозғалысты білдіретін қозғалыс етістіктері	*Аралау, шарлау, кезу, қыдыру, атқу*және т.б.	ЕТ305

1	2	3
Семантическалық топ: 4. Адамдар, жануарлар мен өлі табиғат арасындағы қатынастарды білдіретін қозғалыс етістіктері		
1. Теңестіру етістіктері	*Алу, қосу, теңестіру, салыстыру, бөлу* және т.б.	ЕТ401
2. Меншіктеу етістіктері	*Иемдену, меншіктеу, иелену, бауырау* және т.б.	ЕТ402
3. Амандасу және қоштасу етістіктері	*Амандасу, сәлемдесу, көрісу, қоштасу* және т.б.	ЕТ403
4. Құрметтеу етістіктері	*Құрметтеу, қадірлеу, ізеттеу, сыйлау* және т.б.	ЕТ404
5. Билік жүргізу етістіктері	*Басқару, әмір ету, боп-салау, өкім ету*және т.б.	ЕТ405
6. Бағынышты етістіктер	*Көну, бас иіу, ырқына көну, бағыну* және т.б.	ЕТ406
Семантическалық топ: 5. Психикалық әрекет етістері		
1. Сезім етістіктері а) жалпы мағынадағы сезім етістіктері	*Сезу, сезіну, байқау, абайлау, аңғару, пайымдау, түйсіну* және т.б.	ЕТ501
б) Көру етістіктері	*Көру, көз салу, көз тастау, үңілу, бажырайу* және т.б.	ЕТ502
в) дәмді білдіретін етістіктері	*Кермектену, ашқылтымдану, тату* және т.б.	ЕТ503
г) сезіну етістіктері (иіске байланысты)	*Бықсу, мүңку, аңқу, сасу, борсу, қаңсу* және т.б.	ЕТ504
д) адам мен жануарлар дене терісіне қатысты сезіну етістіктері	*Сыздау, ашу, қышу, тызылдау, сырқырау* және т.б.	ЕТ505
е) есту әрекетіне қатысты етістіктер	*Тыңдау, есіту, құлақ салу, дауыстау, естілу* және т.б.	ЕТ506
2. Ойлау етістіктері а) жалпы мағынадағы ойлау әрекетіне қатысты ойлау етістіктері	*Ойлау, пайымдау, ой жүгірту, ақылға салу, болжау, шамалау* және т.б.	ЕТ507

1	2	3
б) болжам етістіктері (шамалау)	*Шамалау, болжау, тұспалдау, межелеу* және т.б.	ЕТ508
в) <u>қорытынды</u>; <u>тұжырым</u>; <u>түйін</u>, жалпылауға қатысты етістіктер	*Қорытындылау, шешілу, кесіп айту* және т.б.	ЕТ509
г) сенім білдіру етістіктері	*Сену, достасу, дос көру, жақындасу, сенім арту.*	ЕТ510
д) адамның есте сақтау әрекетіне қатысты етістіктер	*Ескер, ұмытпа, есте сақта, зерделеде қалдыру ...*	ЕТ511
3. Психикалық әсер ету етістіктері	*Қуанту, күлдіру, қорқыту, ойландыру* және т.б.	ЕТ512
4. Зейін қою, тілек білдіру етістіктері		ЕТ513
5 Белсенді психикалық әсер ету етістіктері	*Айқайлау, ұрысу, зар жылау, қарқылдау* және т.б.	ЕТ514
6. Эмоционалды етістіктер:		
а) таңдануетістіктері	*Таңдану, қуану, таңырқау, аузы ашылу* және т.б.	ЕТ515
б) қуаныш етістіктері	*Шаттану, қуану, мәз болу ...*	ЕТ516
в) қанағаттану етістіктері	*Рахаттану, қанағаттану, шаттану* және т.б.	ЕТ517
г) қайғыру етістіктері	*Қапалану, қайғыру, қабырғасы қайысу және т.б.*	ЕТ518
7. Сөйлеу етістіктері а) ақпараттық	*Айту, хабарлау, сөйлеу, баяндау* және т.б.	ЕТ519
б) сұхбаттасу әрекетіне қатысты етістіктер	*Сөйлесу, әңгімелесу, сырласу, сұбхаттасу және т.б.*	ЕТ520
в) «сұрақ-жауап» әрекеті бойынша қолданылатын етістіктер	*Сұрау, жауап беру, кінәласу, беттесу* және т.б.	ЕТ521
г) адамның эмоционалды күйіне байланысты етістіктер	*Мыңқылдау, көку, оттау, шалқу, лепіру* және т.б.	ЕТ522

1	2	3
Семантикалық топ: 6. Физикалық күй мен нормаға қатысты етістіктер		
1. Өлі табиғаттың элементтеріне тән (түсінің өзгеруі)	*Жылтырау, жарқырау, ағару, қараю, сарғаю, көгеру, сәулелену* және т.б.	ET601
2. Физикалық күйдің сақталуын білдіретін етістіктер	*Сақталу, бұзылмау, ашымау, қартаймау...*	ET602
а) адамның ішкі психологиялық күйін білдіретін етістіктер	*Қасарысу, ашулану, қайғыру, толқу, қуану, ойлану, арманда*у және т.б.	ET603
б) адамның субъективтік ерекшелігін (сыртқы) білдіретін етістіктер	*Кісімсу, адамсу, баласыну, сасқалақтау* және т.б.	ET604
3. Заттың көлем өзгерісін білдіретін етістіктер	*Шөгу, ұзару, солу, толысу, зораю, кеңею* және т.б.	ET605
4. Заттың сандық өзгерісін білдіретін етістіктер	*Азаю, көбею, молаю, ортаю, қосылу* және т.б.	ET606
5. Адамның немесе заттың ішкі қасиеттік өзгерістерін білдіретін етістіктер	*Іру, шіру, үгілу, тозу, босау, өзгеру* және т.б.	ET607
6. Адамдардың әдеттеріне қарай қолданатын етістіктер	*Әдеттену, үйрену, дағдылану, жалқаулану ...*	ET608
7. Адамдар мен жануарлардың тууы мен өсу үрдісіне тән етістіктер	*Туу, өсу, өну, ержету, жігіт болу* және т.б.	ET609
Семантикалықтоп: 7. Психикалық күйге қатысты етістіктер		
1. Адамға ғана тән етістіктер	*Қайғыру, күңірену, ренжу, күлу, жылау, қуану*	ET701
2. Табиғи құбылыстарға қатысты етістіктер	*Азынау, гуілдеу, борандату, желдету* және т.б.	ET702
3. Адамға және жануарлар әлемінің ерекше түрлеріне тән етістіктер	*Қайғыру, қуану, ашулану, ызалану* және т.б.	ET703
4. Өлі табиғат элементтеріне тіркелетін етістіктер	*Түнеру, жадырау...*	ET704
5. Адамға, жануарлар әлемінің ерекше түрлеріне және кейбір өлі табиғат элементтеріне тіркелетін етістіктер	*Қайғыру, мұңаю, жадырау, күңірену, қуану, күлімсіреу* және т.б.	ET705

1	2	3
Семантикалық топ: 8. Фазалық және модальды етістіктер		
1. Әрекеттің басы	*Бастау, кірісу, жасаға-лы-, кіріскелі-* және т.б.	ЕТ801
2. Әрекеттің жалғасы	*Жалғастыру, созу, атқару, -ып, -ip (отыр)...*	ЕТ802
3. Әрекеттің соңы (аяғы)	*Аяқтау, қою, бітіру, тоқтату, болу* және т.б.	ЕТ803
Модальды қалып етістіктері :		
Отыр	*Балалар сабақ оқып отыр.*	ЕТ804
Жатыр	*Оқушылар мектептен келе жатыр.*	ЕТ805
Тұр	*Әнші өлең айтып тұр.*	ЕТ806
Жүр	*Оқымыстылар космосты зерттеп жүр.*	ЕТ807
Семантикалық топ: 9. Каузативті етістіктер		
1. Таза каузативті етістіктер: – мәжбүрлеу	*Мәжбүрле-, күште-, көндір-, душар ет-* және т.б.	ЕТ901
– шақыру, келтіру, тигізу	*Келтір-, туғыз-, оят-, тигіз-, қоздыр-* және т.б.	ЕТ902
– түрткі болу, қозғау салу	*Түрткі бол-, көндір-, көңілін аудар-, тарт-* және т.б.	ЕТ903
– қамтамасыз ету	*Жеткіз-, әкел-, апар-, түсір-* және т.б.	ЕТ904
– ықпал ету	*Ет-, ұста-, тұтын* және т.б.	ЕТ905
– мүмкіндік	*(мүмкіндік) бер-* және т.б.	ЕТ906
2. Дербес (частно)-каузативті етістіктер Нысандарды жасау (құру) етістіктері:		
– бөліктерін жалғау бекіту	*Тік-, астарла-, тепші-, жапсыр-, құра-* және т.б.	ЕТ907
– материалды нығыздау, біріктіру	*Өр-, иір-, тоқы-, ес-, иле-, пысықта-, қарпы-, шират-* және т.б.	ЕТ908

1	2	3
– бір нәрсенің бетін білгілі тәсілмен өңдеу	*Еге-, жаңқала-, жаны-, сүргіле-, қайра-* және т.б.	ET909
– әрекет құралы	*Бұрғыла-, қаз- и др.*	ET910
– материалдың қасиеті	*Құй-, жапсыр-, үрле- ...*	ET911
– салу, тұрғызу	*Сал-, тік-, қала ...*	ET912
– тамақ әзірлеу	*Пісір-, қуыр* және т.б.	ET913
Нысанды жою етістіктері		
Мазмұнында әрекет тәсілі:		
– отпен	*Өрте-, жандыр-* және т.б.	ET914
– ауадан шабуылдап	*Бомбыла-, атқыла- ...*	ET915
– жарылыспен	*Қопар, құлат ...*	ET916
– сындырып	*Бұз-, қират-, сындыр-...*	ET917
– екпіндеп, күшпен	*Шақ-, талқан қыл ...*	ET918
– қатты ашумен, түп-тұқылын қоймай күшпен қирату	*Астан-кестен қыл-, тас-талқан ет* және т.б.	ET919
3. Туу мағынасындағы етістіктер	*Босандыр-, тап-, шағы-лыстыр-, төлдет-* және т.б.	ET920
4. Біреуді өлтіру мағынасындағы етістіктер	*Өлтір-, мерт қыл-, бауызда-, ас-, ат-* және т.б.	ET921
5. Нысандарды өзгерту мағынасындағы етістіктер		
Мазмұнында әрекет тәсілі: а) алғашқы тұрпатын өзгерту	*Бүкте-, қайыр-, бүк-, жалпайт-, ез-, мыжғыла-, умажда-* және т.б.	ET922
б) белгілі тұрпатқа келтіру	*Пішінде-, дөңгелекте-, үй-, түй-, өр-* және т.б.	ET923
6. Нысанның болмыс белгілеріне қатынасын білдіретін етістіктер		
Сандық сипаттама: а) көбею–азаю	*Көбейт-, молайт-, ұлғайт-, азайт-* және т.б.	ET924
б) ұзарту–қысқарту	*Ұзарт-, қысқарт-* және т.б.	ET925

1	2	3
в) кеңейту—тарылту	*Соз-, тарылт-, кеңейт, жуандат-,жіңішкерт-...*	ET926
г) тереңдету—тайыздату	*Тереңдет-, тайыздат-....*	ET927
Уақыттық сипаттама: а) мерзімін ұзарту	*Соз-, жылжыт-, ұзарт, қысқарт-* және т.б.	ET928
б) мерзімін қысқарту	*Жылдамдат-, тездет-..*	ET929
Жергілікті (локальды) сипаттама: а) орақтастық – бөлектестік	*Жина-, жинақта-, шашырат-, бөліп_ отырғыз-*	ET930
б) нысандарды белгілі кеңістікте орналастыру	*Орнат-, орналастыр-, іл-, тұрғыз-* және т.б.	ET931
Етістік лексемаларының парадигматикалық қатары: тазарту, тазалау, жуу, өшіру және т.б.	*Тазала-, тазарт-, тазартқыз-, жу-, жуын-, жуыс-, жуындыр-* және т.б.	ET932
Дифференциалды белгілері бар етістіктер: өзгерту арқылы басқа күйге ауыстыру	*Балқыт-, ез-, еріт- , қорыт-, жасыт-, жылыт-, ысыт-, қатыр-....*	ET933
Заттар контактарына қатысты каузация мағынасын білдіретін етістіктер: біріктіру, жалғау: а) жаңа нысан құрылады	*Жалга-, жалғастыр-, тірке-, ұластыр-, біріктір-, қос-* және т.б.	ET934
б) жаңа нысан құрылмайды	*тік-, астарла-, тепші-*	ET935
Функтивті қатынастарды білдіретін каузация мағынасындағы етістіктер: а) жұмыс істеуге, пайдалануға дайын ету	*Дайында-, әзірле-, даярла-, сайла-* және т.б.	ET936
б) Бірдеңені пайдалануға қайта жарамды қылып істеу	*Жөнде-, түзе-, дұрыс-та-, бүтінде-, өңде-.....*	ET937
в) нысандардың жұмыс істеп тұруын қамтамасыз ету	*Күзет-, қарауылда-, бақыла-, тексер-* және т.б.	ET938

1	2	3
Қозғалысқа қатысты каузация мағынасындағы етістіктер: 1) бейтарап мағынасындағы қозғалыстың бағыттылығы	*Ағыз-, қозғалт-, көшір-, секірт-, қыдырт-, қадамдат-, саяхат-та-* және т.б.	ЕТ939
2) горизанталь және вертикаль бағытын білдіретін қозғалыстар	*Бар-, кел-, кет-, қайт-, шық-, кір-, мін-, түс-...*	ЕТ940
3) қозғалыстардағы тоқсаулдарды еңсеру	*Ас-, кеш-, өт-, қарғы-* және т.б.	ЕТ941
4) қозғалыс қарқынын белгілейді:	*Қу-, айда-, қалқы-, шап-* және т.б.	ЕТ942
– материалдық ресурстардың қамтамасыз етілуіне қатысты каузация (жабдық – құрал-сайман, қару-жарақ, техникалық құралдар)	*Жабдықта-, әбзелде-, безендір-, сайла-, газдандыр-, электрлен-дір*-және т.б.	ЕТ943
– материалды емес нысандардың каузациясы (білім, дағдылар және т.б.)	*Оқыт-, үйрет-, дайында-, даярла-,білгіз-* және т.б.	ЕТ944

13-кестенің соңғы бағанасында қазақ мәтінін туындататын модельде қолданылатын қазақ лексикасының семантикалық топтарының шартты белгілері, яғни кодтары берілген. Етістік сөзі қысқарған – ЕТ түрде алынды.

2.5.4. Қазақ тілі мәтіндерін туындату моделінде қолданылатын сын есімдердің семантикалық топтары

Қазақ тілінде сын есімдердің өзіне тән формалары мен семантикалық сипаты жағынан басқа сөз таптарынан ерекшеленеді. Сын есімдер заттанып зат есімдердің грамматикалық формаларымен түрленгенмен, өзіне тән грамматикалық категориясы – сын есімнің шырай категориясы болып табылады. Олар сөйлемде негізінен анықталушы зат есімдердің алдында тұрады. Сөйлеу кезінде немесе поэтикалық шығармаларда сын есімдер сөйлем соңында да тұра алады. Мысалы, *Күн жылы. Ай жарық.*

Сын есімдерді сапалық және қатыстық сын есімдерге бөлу

тұрақталған. Алайда олардың ішкі мағыналық топтарын жіктеуде бірізді классификация жоқ [56, 211-215 бб.; 57, 9-28; 58 бб.].

Қазақ грамматикаларындағы сын есімдерді сапалық және қатыстық деп бөлу сапа мәнді сөздердің таза семантикасына ғана негізделмеген. Өйткені сапалық сын есімдер мен қатыстық сын есімдердің бір-бірінен айырмашылығы олардың семантикалық сипаттарына ғана қатысты емес. Олар бір-бірінен, бірінші кезекте, құрылымы мен мағынасы жағынан ерекшеленеді. Семалық талдауды негіз етіп алсақ, онда сын есімдерді сапалық/қатыстық деп бөлу «таза» семантикалық жіктеу бола алмайды. Бұл жіктеме сын есімдерді топтастыру үлгілерінің бірі болып табылатын лексика-морфологиялық принципке негізделеді. Бұл принциптің мәні, біріншіден, сапа мәнді сөздердің семантикасы, екіншіден, құрылымы, яғни формасы. Сапалық сын есімдер негізінен түбір сөздер болып келеді. Ал қатыстық сын есімдер қосымша жалғану арқылы, яғни сөзжасамның синтетикалық тәсілі арқылы жасалып, туынды сын есімдер болып табылады.

Сапалық сын есімдер заттық мәннен алшақтап, абстрактанған, ал қатыстық сын есімдер заттар мен құбылыстарға қатысты әртүрлі қатынастарды білдіре алады. Ғалым А.Жұбанов қазақ мәтінін туындатуға қатысты докторлық диссертациясында [79, 97-103 бб.] сын есімдердің семантикалық классификациясын жасауда денотативті-десигнативті ұстанымға сүйеніп, сапалық және қатыстық сын есімдердің бір топ лексика-семантикалық топтарын жіктеп көрсетеді. Бұл классификациялар ***14-кесте*** және ***15-кестеде*** берілді.

14-кестенің 4-бағанасында *-рақ/-рек, -ырақ/-ірек, -лау/-леу, -дау/-деу, -тау/-теу, -қыл/-гыл, -қылт/-ғылт, -тым,-шыл/-шіл, -қай, -аң* және т.б. аффикстер арқылы жасалатын салыстырмалы шырайға жататын сапалық сын есімдерге мысалдар келтірілген. 6-бағанада сын есімдер мен үстеулердің тіркесінен жасалатын күшейтпелі шырайға жататын *өте, аса, тым, тіпті, тіптен, асқан, ең, нағыз* және т.б. мысалдар келтірілген; сын есімнің алғашқы буынының қайталауынан жасалған сын есімдерге, мысалы, *қып-қызыл, сап-сары, біп-биік* және т.б.

14-кестенің 3, 5 және 7-бағаналарында қазақ мәтіндерін туындататын модельде қолданыс табатын семантикалық субтоптарының шартты белгілері – кодтары берілген, ал 7-бағанада *сапалық сын есім* сөз тіркесінің қысқарған түрі – СС орын алды.

14-кесте

Қазақ мәтінін туындату моделінде қолданылатын сапалық сын есімдердің семантикалық топтары

	Сын есімдердің түрлері, мысалдар және шартты белгілері (кодтары)					
Сын есімнің семантикалық субтобы	Жай шырай	Суб-топтың коды	Салыстырмалы шырай	Суб-топтың коды	Күшейтпелі шырай	Суб-топтың коды
1	2	3	4	5	6	7
Семантикалық топ: 1. Адам және оған қатысты белгілерді білдіретін сапалық сын есімдер						
1. Адамның ішкі сапасын білдіреді	*Жақсы, жаман, тәуір, алғыр, жүйрік* және т.б.	СС001	*Жақсырақ, жамандау, жүйрігірек ...*	СС041	*Өте жақсы, тіпті жаман, тым салақ* және т.б.	СС091
2. Адамның физиологиялық қасиетін білдіреді	*Семіз, арық, шолақ, ақсақ, соқыр* және т.б.	СС002	*Семіздеу, арықтау, соқырлау* ж.т.б.	СС042	*Өте семіз, тым арық, тым соқыр* және т.б.	СС092
3. Адамның мінез-құлқын білдіреді	*Жуас, тентек, пысық, қыңыр, жұмсақ* және т.б.	СС003	*Жуастау, тентектеу, жайдарырақ*	СС043	*Өте жуас, тым тентек, өте пысық, өте қорқақ* және т.б.	СС093
4. Қоршаған ортаның температурасының өзгерістеріне байланысты қолданылады	*Қызу, ыстық, жылы, суық, салқын, дым-қыл, құрғақ...*	СС004	*Қызулау, ыстықтау, жылылау, суығырақ ...*	СС044	*Аса қызу, тым ыстық, өте жылы, тым суық, тым салқын* және т.б.	СС094

14-кестенің жалғасы

1	2	3	4	5	6	7
5. Адамның жасына қарай қолданылады	*Қарт, жас, кәрі, балғын, сәби* және т.б.	СС005	*Қарттау, жастау, кәрілеу* және т.б.	СС045	*Тым қарт, өте жас, тым кәрі, тым сәби* және т.б.	СС095
6. Адамның тұрмыстық ахуалына қатысты қолданады	*Бай, кедей, жарлы* и т.д	СС006	*Байырақ, кедейлеу, жарлылау* ж.т.б.	СС046	*Өте бай, өте кед-ей,тым кедей, тым жарлы* және т.б.	СС096
7. Адамның сыртқы сұлулығына байланысты қолданылады	*Сұлу, әсем, әдемі, келісті, көркем* және т.б.	СС007	*Сұлуырақ, сұлу-лау, әсемірек, әдемілеу*	СС047	*Тым сұлу, өте сұлу, өте әсем, тым әдемі* және т.б.	СС097
Семантикалық топ: 2. Жануарлар әлемінің ерекше түрлеріне тән белгілерді білдіретін сапалық сын есімдер						
1. Жануарлардың физиологиялық қасиетін білдіреді	*Семіз, арық, жүдеу, ақсақ, тапал, ортық*	СС011	*Семіздеу, арық-тау, биігірек, ақсақтау*	СС051	*Аса сұлу, тым арық, өте жүдеу, тым ақсақ* және т.б.	СС101
2. Жануарлардың мінез-құлқын білдіреді	*Асау, жуас, қорқақ, қыңыр тентек, үркек*	СС012	*Асаулау, жуастау, қорқақтау ...*	СС052	*Тым асау, өте жуас, тым қорқақ, өте қыңыр* және т.б.	СС102

14-кестенің жалғасы

1	2	3	4	5	6	7
3. Жануарлардың жасына қарай қолданылады	*Кәрі, жас* және т.б.	СС013	*Кәрілеу, жастау* ж.т.б.	СС053	*Тым кәрі, өте кәрі, тым жас, өте жас* ж.т.б.	СС103
4. Жануарлардың сыртқы әлпетіне қатысты қолданылады	*Торы, бурыл, ақ, боз, қара, шұбар, тарғыл*	СС014	*Торылау, бурылдау, ақтау, боздау*	СС054	*Тым торы, тым бурыл, аппақ, қап-қара* және т.б.	СС104
Семантикалық топ: 3. Заттарды сипаттауда қолданылатын белгілерді білдіретін сапалық сын есімдер						
1. Заттардың түстерін білдіреді	*Ақ, көк,қызыл, сары, сұр, боз, шұбар, жасыл*	СС021	*Ақшыл, көкшіл, қызылырақ, қызылдау*	СС061	*Аппақ, көкпеңбек, қып-қызыл, сап-сары, сұп-сұр* және т.б.	СС111
2. Азық-түлік өнімдерінің дәмдік сапасын білдіреді	*Тәтті, ащы, тұщы, ышқыл* және т.б.	СС022	*Тәттілеу, ащы-лау, ащырақ, қышқылдау. . .*	СС062	*Өте тәтті, тым ащы, аса дәмді* және т.б.	СС112
3. Бұйымдардың көлемін білдіреді	*Дәу, ірі, үлкен, кіші, қалың, жұқа, жуан...*	СС023	*Іріленеу, үлкен-ірек, үлкендеу, кішірек* ж.т.б.	СС063	*Тым ірі, өте дәу, тым кіші, өте қалың, өте жұқа* және т.б.	СС113

14-кестенің жалғасы

1	2	3	4	5	6	7
4. Заттардың салмағын білдіреді	*Ауыр, жеңіл, зіл* и т.д	СС024	*Ауырлау, жеңілдеу, жеңіл-ірек, зілдей. . .*	СС064	*Өте ауыр, тым ауыр, өте жеңіл, тым жеңіл*	СС114
5. Заттың табиғи жай-күйін білдіреді	*Қалың, сирек, жуан, жіңішке, селдір,қаба*	СС025	*Қалыңдау, сирегірек, жуан-дау, селдірлеу*	СС065	*Қап-қалың, тым сирек, өте жуан, сеп-селдір* және т.б.	СС115
6. Қоршаған ортаның температуралық өзгерісіне қарай, заттың белгісін білдіреді	*Ыстық, суық, салқын, жылы, дымқыл, құрғақ* және т.б.	СС026	*Ыстықтау, суығырақ, салқындау, құрғақтау. . .*	СС066	*Ып-ыстық, тым ыстық, сұп-суық, тым суық, сап-салқын* және т.б.	СС116
7. Заттың сыртқы тұрпатына қарай қолданылады	*Дөңгелек, сопақ, үшкіл, тегіс, жұмыр ...*	СС027	*Дөңгелектеу, сопақтау, үшкілдеу* және т.б.	СС067	*Дөп-дөңгелек, тым сопақ, өте үшкір, теп-тегіс* және т.б.	СС117
8. Табиғи құбылыстарға қарай қолданылады	*Жарық, қараңғы, ашық, күңгірт, жылы* және т.б.	СС028	*Жарықтау, қараңғылау, ашығырақ* және т.б.	СС068	*Жап-жарық,тым жарық, өте ашық, тым күңгірт* және т.б.	СС118

14-кестенің жалғасы

1	2	3		4	5		6
9. Заттардың сандық қатынастарына қарай қолданылады	*Аз, көп, мол шағын, қыруар, кішкене* және т.б.	СС029	*Аздау, көптеу, көбірек, молдау, молырақ* және т.б.		СС069	*Өте аз, тым аз, өте көп, шап-шағын, моп-мол* және т.б.	СС119
10. Заттың бағасына қарай қолданылады	*Арзан, қымбат, тегін, бос* және т.б.	СС030	*Арзанырақ, қымбаттау* және т.б.		СС070	*Ап-арзан, тым арзан, теп-тегін* және т.б.	СС120
Семантикалық топ:4. Адамға, жануарлар әлемінің айырықша түрлеріне тән және затты сипаттауға қолданылатын сапалық сын есімдер							
1. Заттардың түсі мен сапасын белгілейді	*Ақ, қара, жақсы,жаман тәуір* және т.б.	СС031	*Ақшыл, қаралау,жақсырақ, жамандау ...*		СС081	*Аппақ, қап-қара, тым қара, өте жақсы, тәп-тәуір ...*	СС131
2. Заттардың көлемдік мөлшеріне қарай қолданылады	*Үлкен, кіші, дәу* и т.д	СС032	*Үлкендеу,мол-ырақ, молдау, кішірек* және т.б.		СС082	*Үп-үлкен, өте үлкен, тым кіші, өте дәу* және т.б.	СС132
3. Заттардың салмағына қарай қолданылады	*Ауыр, жеңіл, зіл* және т.б.	СС033	*Ауырлау, жеңілірек, жеңіл-деу, зілдей ...*		СС083	*Ап-ауыр, өте ауыр, жеп-жеңіл, тым жеңіл* және т.б.	СС133
4. Жанды және жансыздардың жасына қарай қолданылады	*Қарт, кәрі, жас* және т.б.	СС034	*Қартаңдау, кәрілеу, жасы-рақ, жастау*		СС084	*Тым қарт, өте кәрі, тым кәрі, жап-жас*және т.б.	СС134

15-кесте

Қазақ мәтінін туындату моделінде қолданылатын қатыстық сын есімдердің семантикалық топтары

Қатыстық сын есімдерге мысалдар	Топ коды (шартты белгі)	Салыстырмалы шырай	Топ коды (шартты белгі)	Күшейтпелі шырай	Топ коды (шартты белгі)
1	2	3	4	5	6
Семантикалық топ: 1. Қайсыбір жансыз объектіге (нысанға) қатысты субъектіні сипаттайтын сын есімдер					
Тершең, кіршең, малды, безеулі, сақалды, азулы, қанатты,үйлі, бассыз	ҚС011	*Тершеңдеу, кіршеңірек, малдылау, безеулілеу, сақалдырақ, азулырақ ...*	ҚС012	*Өте тершең, тым кіршең, нағыз малды, тым безеулі, тым сақалды* және т.б.	ҚС013
Семантикалық топ: 2. Қайсыбір жанды субъектіге қатысты объектіні (нысанды) сипаттайтын сын есімдер					
Таулы, сулы, желді, күзгі, жазғы, түнгі, күндізгі, өзенді, сулы ...	ҚС021	*Таулылау, сулырақ, күзгілеу, жазғылау, сулырақ, таусыздау, сусыздау* және т.б.	ҚС022	*Ең таулы, тым сулы, тіпті желді, нағыз жазғы, ең таулы-тасты* және т.б.	ҚС023
Семантикалық топ: 3. Уақытқа қатысты объектіні (нысанды) сипаттайтын сын есімдер					
Кешкі, түскі, күзгі, түнгі, жазғы, айлық	ҚС031	*Кешкірек...*	ҚС032	*Тым кешкі, нағыз түскі, нағыз жазғы* және т.б.	ҚС033

15-кестенің жалғасы

1	2	3	4	5	6
Семантикалық топ: 4. Материалға қатысты объектіні (нысанды) сипаттайтын сын есімдер					
Көйлектік, пальтолық, тағамдық, шайлық, согымдық және т.б.	ҚС041		ҚС042	*Нағыз көйлектік, нағыз пальтолық, нағыз едендік* және т.б.	ҚС043
Семантикалық топ: 5. Қайсыбір әрекетке қатысты нысанды сипаттайтын сын есімдер					
Жылауық, сыбырлауық, асыранды, жаттанды, жасырынды, ауыспалы	ҚС051	*Жылауықтау, жаттандылау, ауыспалылау, жылжымалылау, тапқыштау*	ҚС052	*Нағыз жылауық, тіпті сыбырлауық, ең жаттанды, тым жасырынды* және т.б.	ҚС053
Семантикалық топ: 6. Объектіні (нысанды), оның орнына қатысты сипаттайтын сын есімдер					
Жерлес, ауылдас, көршілес, елдес, ұялас, іргелес, мекендес және т.б.	ҚС061	*Жерлесшіл, ауылдасшыл, көршілестеу, ұяластау, іргелестеу, мекендестеу*	ҚС062	*Нағыз жерлес, нағыз ауылдас, тіпті көршілес, нағыз ұялас, өте іргелес* және т.б.	ҚС063
Семантикалық топ: 7. Дерексіз ұғымға қатысты объектіні (нысанды) сипаттайтын сын есімдер					
Аршыл, ойшыл, сезімтал, түйсіксіз, ақылсыз, еріксіз, санасыз, пікірлес, мұңдас, ниеттес, сырлас және т.б.	ҚС071	*Аршылдау, ойшылдау, сезімталырақ, түйсіксіздеу, ақылсыздау, санасыздау, пікірлестеу, мұңдасырақ, ниеттестеу* және т.б.	ҚС072	*Аса аршыл, өте ойшыл, тым сезімтал, нағыз түйсіксіз, тіптен ақылсыз, өте санасыз, нағыз пікірлес,өте сырлас* және т.б.	ҚС073

15-кестенің 3-бағанасында *-рақ/-рек, -ырақ/-ірек, -лау/-леу, -дау/-деу, -тау/-теу, -қыл/-ғыл, -қылт/-ғылт, -тым, -шыл/-шіл, -қай, -аң* және т.б. аффикстер арқылы қатыстық сын есімдерден салыстырмалы шырай жасалатын мысалдар келтірілген.

15-кестенің 5-бағанасында *өте, аса, тым, тіпті, тіптен, асқан, ең, нағыз* және т.б.сын есімдер мен үстеулердің тіркесі арқылы жасалатын сын есімнің күшейтпелі шырайына мысалдар келтірілген. Ал ***2-ші, 4-ші, 6-шы*** бағаналарда қатыстық сын есімдердің семантикалық субтоптарына біздер ұсынған шартты белгілер (кодтар) орын алды. Мұндағы, ҚС – *қатыстық сын есім* сөз тіркесінің қысқарған түрі.

2.5.5. Қазақ тілі мәтіндерін туындату моделінде қолданылатын үстеулердің семантикалық топтары

Үстеу сөз табының семантикалық құрылымы жағынан әркелкі болып келеді. Контекске қарай да әртүрлі мағыналық реңке ие болады. Қазақ тілінде үстеуге жататын сөздерді мынадай топтарға бөледі:

1) Анықтауыш үстеулер, олар іс әрекеттің сапасын, санын, өлшемін, тәсілін т.б. білдіреді.

2) Пысықтауыш үстеулер, олар іс әрекеттің уақытын, орнын, себебін, мақсатын білдіреді.

Қазақ тілінде сапалық үстеулер жеке топ құрамайды. Алайда кейбір үстеулер орыс тіліндегі сапалық үстеулердің мағынасын білдіреді (медленно – шабан, бәсең; быстро – жылдам, тез, шапшаң.

Біздің статистикалық зерттеулеріміз бойынша, қазақ тілі мәтіндерінде сын-қимыл үстеулері жиі кездеседі. Содан кейін мезгіл үстеулері, мекен үстеулері, мөлшер үстеулері. Ал әлсірету, күшейту мәнін үстейтін үстеулер басқаларынан сиректеу. Жоғарыда аталғандардан басқа қазақ тілі грамматикаларында топтау үстеулері деген де топ аталады. Жоғарыда аталған топтар мен тіл біліміндегі үстеулерді классификациялау ұстанымдарын негізге алып, сонымен қатар эксперименттік зерттеулерімізге сүйеніп, қазақ мәтіндерін туындату моделін жасауда пайдалануға болатын

үстеулердің семантикалық классификациясын ғалым А.Қ. Жұбанов ұсынғандай [79, 104-106 бб.] төмендегіше жіктейміз.

Төменде үстеудің семантикалық классификациясы 16-кестеде көрініс тапты.

16-кесте

Қазақ мәтінін туындату моделінде қолданылатын үстеулердің семантикалық топтары

Үстеулердің семантикалық субтобы	Үстеулердің семантикалық субтоптарына мысалдар	Субтоп коды
1	2	3
Семантикалық топ: 1. Анықтауыш үстеулер		
1. Сын – бейне үстеулері	*Осылай, осылайша, сөйтіп, өйтіп, бүйтіп, олай-бұлай, бірден, бірдей, бірге, бірте-бірте, біртіндеп, ойша, өзінше, іле, қолма-қол, кезек-кезек, көрнеу, өздігінен, өзді-өзді, жатпай-тұрмай, үнемі* және т.б.	YC011
2. Мөлшер үстеулері	*Онша, сонша, осынша, оншалық, соншалық, осыншалық, оншама, мұншама, анағұрлым, бірен-саран, ондаған, көптеген, бірталай* және т.б.	YC012
3. Анықтауыш сапалық үстеулер: а) күшейту не әлсірету үстеулері	*Ең, әбден, ылғи, кілең, сәл, өңкей, тіпті, тым, нақ, нағыз, әнтек, мүлдем, дәл, керемет, қабағат, мейлінше, сонша, төтенше, жөнсіз, орасан, өрен, аса, өте* және т.б.	YC013
б) Топтау үстеулері	*Екеулеп, үшеулеп, он-ондап, бір-бірлеп, он-оннан, қырық-елуден, аз-аздан, көп-көптен, топ-тобымен, рет-ретімен, бас-басына, үйді-үйіне* және т.б.	YC014

1	2	3
Семантикалық топ: 2. Пысықтауыш үстеулер		
1. Пысықтауыш мезгіл үстеулері	*Бүгін, былтыр, ертең, таңертең, кешке, қазір, енді, әлі, ендігәрі, екіншәрі, бұрын, әуел-баста, ерте, кеш, күндіз, бүрсігүні, әсте, күнімен, ертеден, күні бүгін, күні кешке, бүгін таңда, күн ара, күні бойы, қыстыгүні, қыстай, жаздай, ала жаздай, күні-түні, анда-санда, енді-енді, осындайда, апақ-сапақта* және т.б.	YC021
2. Себеп-салдар үстеулері	*Жоққа, босқа, құр босқа, бекерге, амалсыздан, лажсыздан, шарасыздан, сөйле-сөйлей, көре-көре, оқый-оқый* және т.б.	YC022
3. Пысықтауыш мақсат үстеулері	*Әдейі, әдейілеп, жорта, қасақана, қітапқа бола, оқуға бола* және т.б.	YC023
4. Пысықтауыш мекен үстеулері	*Ілгері, ілгеріде, әрі, әріде, әрмен, жоғары, төмен, жолшыбай, жол-жөне-кей, алға, артта, алды-артынан, тысқары, одан-бұдан* және т.б.	YC024

Үстеудің семантикалық субтоптарына қатысты ***16-кестеде*** келтірілген мысалдар А. Ысқақовтың («Қазіргі қазақ тілі». Алматы: Ана тілі, 1991. –340-343 б.) кітабынан алынды.

2.5.6. Қазақ тілі мәтіндерін туындату моделінде қолданылатын есімдіктердің семантикалық топтары

Есімдіктер семантикалық сипаты жағынан дейктикалық сөздерге жатады. Қазақ тілінде жіктеу және сілтеу есімдіктері ең жиі қолданылатын сөздерге жатады. Есімдіктердің зат есім, сын есім, сан есімдерден айырмашылығы жалпы, ортақ мағынаны білдіреді,

есім сөздердің орнына қолданыла береді. Есімдіктер тілде сілтеп көрсету қызметін атқарады. Оның семантикалық құрылымы жалпы болып келеді. Затты, сапаны білдірмейді, тек солардың орнында жүреді. Зат есімнің не басқа сөз табының орнында қолданылғанмен, сол сөз табының барлық мағынасын білдірмейді. Есімдіктер сол сөз таптарының бір мағынасын ғана білдіре алады. Мәтіндегі баяндауға байланысты белгілі бір субъектіге тәуелді болып тұрады, сөйлеу жағдаятына бейімделеді. Мәтіндерде автор оларды сөздерді қайталамау мақсатында қолданады. Есім сөздердің синонимі сияқты болып келеді. Сондықтан есімдіктердің денотативті және сигникативті мағынасы көптеген факторларға байланысты болып келеді.

Қазақ тілінде есімдіктер парадигмасы бірізді жүйеге түскен. Ал саны жағынан ғалымдар түрліше пікір айтады. Мысалы, А.Исхаков олардың санын – 60-70 деп көрсетсе [30, с.210], И.Ұйықбаев – 111 дейді [59].

А.Жұбанов М.Әуезовтің «Абай жолы» романында кірме және диалектілік формаларымен қоса есептегенде – 150 деп анықтайды. Ал газет мәтіндерінде – 102, Абай тілі сөздігінде 68 [60]. Жалпы алғанда, басқа сөз таптарымен салыстырғанда есімдіктер саны көп емес деуге болады. Есімдіктер санының аз болуы оларды семантикалық топтарға толығымен жіктеп көрсетуге мүмкіндік береді.

Қазақ тілінде есімдіктер 7 топқа топтастырылады:

1) жіктеу;
2) сілтеу;
3) сұрау;
4) белгісіздік;
5) өздік;
6) жалпылау;
7) болымсыздық.

Есімдіктер сөйлемдер мен абзацтарды, яғни мәтін бөліктерін байланыстырушы қызмет атқарады. Мысалы жіктеу есімдіктері мәтіндегі сөздердің контекстік синонимі болып келеді. Төменде аталған есімдіктер жекелеген семантикалық кодтарға салынып, ***17-кестеде*** берілді.

17-кесте

Қазақ мәтінін туындату моделінде қолданылатын
есімдіктердің семантикалық топтары

Есімдіктің семантикалық тобы	Топтың есімдіктері	Есімдіктің коды (шартты белгі)
1	2	3
1. Жіктеу есімдіктері	*Мен*	ЕС101
	Біз	ЕС102
	Ол	ЕС103
	Сен	ЕС104
	Сіз	ЕС105
	Біздер	ЕС106
	Олар	ЕС107
	Сіздер	ЕС108
	Сендер	ЕС109
2. Өздік есімдіктер	*Өз*	ЕС201
	Өзі, өзім, өзің, өзіңіз, өзіміз, өздеріміз, өздерің, өздеріңіз, өздері және т.б.	ЕС202-ЕС300
3. Сілтеу есімдіктері**:**		
а) сөйлеушіге біршама жақын тұрған заттар мен құбылыстарға нұсқайтын есімдіктер	*Осы*	ЕС301
	Бұл	ЕС302
	Мына	ЕС303
	Мынау	ЕС304
	Осынау	ЕС305
	Міне	ЕС306
	Әне	ЕС307

1	2	3
б) сөйлеушіден біршама алыстау тұрған заттар мен құбылыстарға нұсқайтын есімдіктер	*Ол*	ЕС311
	Сол	ЕС312
	Ана	ЕС313
	Анау	ЕС314
	Сонау	ЕС315
	Сона	ЕС316
	Түу	ЕС317
4. Жалпылау есімдіктері (анықтайтын есімдіктер)	*Әр*	ЕС401
	Әрбір	ЕС402
	Әркім	ЕС403
	Ең	ЕС404
	Бар	ЕС405
	Барлық	ЕС406
	Барша	ЕС407
	Бәрі	ЕС408
	Бүкіл	ЕС409
	Бүтін	ЕС410
	Күллі	ЕС411
	Түгел	ЕС412
5. Сұраулық есімдіктер	*Кім*	ЕС501
	Кімдікі	ЕС502
	Қай	ЕС503

1	2	3
	Қайсы	ЕС504
	Қайда	ЕС505
	Қайдан	ЕС506
	Қайдағы	ЕС507
	Қалай	ЕС508
	Қандай	ЕС509
	Қанша	ЕС510
	Қашан	ЕС511
	Қашаннан	ЕС512
	Қашанғы	ЕС513
	Не	ЕС514
	Немене	ЕС515
	Неше	ЕС516
	Нешеу	ЕС517
	Нешінші	ЕС518
6. Болымсыздық есімдіктер	*Еш*	ЕС601
	Ешкім	ЕС602
	Ешбір	ЕС603
	Ештеме	ЕС604
	Ештеңе	ЕС605
	Дәнеңе	ЕС606
	Ешқайда	ЕС607
	Ешуақытта	ЕС608
	Ешнәрсе	ЕС609
	Ешқашан	ЕС610
	Ешқандай	ЕС611
	Ешқайсы	ЕС612

17-кестенің жалғасы

1	2	3
7. Белгісіздік есімдіктер	*Бір*	ЕС702
	Біреу	ЕС703
	Бірнеше	ЕС704
	Бірдеме	ЕС705
	Кей	ЕС706
	Кейбір	ЕС707
	Кейбіреу	ЕС708
	Қайсыбір	ЕС709
	Әр	ЕС710
	Әрбір	ЕС711
	Әркім	ЕС712
	Әрне	ЕС713
	Әрқайсы	ЕС714
	Әрқалай	ЕС715
	Әлде	ЕС716
	Әлдебір	ЕС717
	Әлдекім	ЕС718
	Әлдене	ЕС719
	Әлдеқайда	ЕС720
	Әлдеқайдан	ЕС721
	Әлденеше	ЕС722
	Әлдеқалай	ЕС723
	Әлдеқашан	ЕС724
	Пәлен	ЕС725

Нақты есімдіктер мен есімдіктердің семантикалық субтоптары ***17-кестенің 3-***бағанасында шартты түрде қабылданған белгілермен (кодтар) көрініс тапты. **ЕС** – «**есімдік**» сөзінің қысқарған түрі.

2.6. Анафоралық белгіленім және просодикалық белгіленім

Мәтінді автоматты түрде өңдеу үшін референциялардың барлық түрлерінің ішінде ең күрделі саналатыны – есімдіктер. Мәселен, машиналық (компьютерлік) аударма жүйелерінің көпшілігі мәтінді жеке сөйлемдер ретінде қарастырудан аударманың мағыналық байланысы ақсайды. Сондықтан егер есімдіктердің орнына қолданылатын референция (референция местоимений-заместителей) дұрыс айқындалған жағдайда мұндай жүйелердің сапасы арта түсер еді. Атап айтқанда, анафоралық белгіленім аталған тиімділікті арттыру үшін қолданылып жүр. Әдетте, антецедент орнына қолданылатын атау сөз тіркесі нөмірленген жақшалар ішіне алынады да, ал есімдік-орынбасар жанына (рядом с местоимением-заместителем) нөміріне сәйкес антецедентке сілтейтін арнайы белгі қойылады.

Сөйлеу тілінің транскипция жасалған корпустарында екпін мен интонацияны сипаттайтын белгілер қолданылады. Ауызша сөйлеу тілінің корпустарында просодиялық белгіленім кідіріс, қайталау, ескертулерді және т.б. белгілейтін ***дискурстік*** белгіленіммен қатар жүреді. Просодикалық белгіленімдер жасау көбінесе ауызша корпустарға негізделеді. Қажетті сөзді іздегенде ауызша корпус конкордансынан шығатын әрбір мысалдың аудиожазбасы/видеожазбасы қасында тұрады. Өйткені просодикалық белгілер негізінен ауызша сөйлеудегі фонеткалық үдерісті бейнелейді. Мысалы, синтагма, пауза, екпін, интонация. Жазбаша мысалда ауызша жазбаның просодикалық сипаты белгіленеді.

Анафоралық, просодикалық деңгейлердегі белгіленімдер үшін автоматтанған жүйе құру мәселесі әзірше өте күрделі болғандықтан, негізгі белгі-код қою жұмысы «қолмен» атқарылады. Ал морфологиялық және синтаксистік талдаулар үшін компьютерлік неше түрлі бағдарламалық құралдар жүзеге асырылды. Оларды тэггерлер (taggers) және парсерлер (parsers) деп атау дәстүрге айналды. Алайда, бұл жүйелердің көпшілігі тағы да зерттеушінің араласуын, яғни қосымша редакциялауды қажет етеді. Себебі морфологиялық омонимия мен синтаксистік бірмәнділіктің бұзылуы бірнеше

шешім ішінен қажеттісін таңдау ісі зерттеушінің өзіне қалдырылады. Дегенмен, қазіргі заманның корпустары ондаған миллионнан тұратын сөздерді қамтитын болғандықтан, қосымша редакцияны жасамауды, мәтін көлемін шектемеуді және белгіленім үрдісін толығымен автоматтандыратын жүйе құруды талап етеді.

2.7. Экстралингвистикалық белгіленім

Корпустар атқаратын функциясына қатысты түрлі шағын топтарға (подкорпус) бөлінеді. Әсіресе лингвистикалық зерттеулер үшін лингвистикалық белгіленімдер қойылған аннотацияланған корпустар құрастыру өте маңызды әрі пайдалы. Осындай белгіленім қойылу-қойылмауына қарай корпустар белгіленім қойылған (аннотацияланған) және белгіленім қойылмаған (аннотацияланбаған) болып екіге бөлінеді. Белгіленім қойылған корпустар құрастыру лингвистикалық зерттеулер үшін маңызды болғанмен, ең алдымен корпусқа енгізілетін мәтіндерді сұрыптаудан өткізіп, олар туралы нақты ақпараттар беру қажет. Мұны кейде экстралингвистикалық белгіленімдер деп те атайды, метамәтіндік белгіленім (метатекстовая разметка – метаразметка) деп те қолданылады. Лингвистикалық белгіленімдер ішкі белгіленім деп аталса, метамәтіндік белгіленімдерді сыртқы белгіленім деп бөлетіндер де бар. Олар: *мәтін және* автор туралы ақпарат: *автор, атауы, жылы, шыққан жері, мәтіннің жанры, тақырыбы, стилі, көлемі т.б.* Оларды библиографиялық, типологиялық, тақырыптық, әлеуметтік, формальдық (мәтін, тарау, бөлім, абзац, сөйлем т.б.) және техникалық (орындаушылар, электрондық нұсқа алынған дереккөз, өңделген күні, кодталған кезі т.б.) деп те бөледі.

Экстралингвистикалық белгіленім немесе метадеректер қатарына «сыртқы», «интеллектуалдық» белгіленімдер (библиографиялық сипаттама, типологиялық сипаттама, тақырыптық сипаттамалар, әлеуметтік сипаттамалар), «формалды» құрылымдық белгіленімдер (мәтін, бөлім, тарау, тарауша, абзац, сөйлем), сонымен бірге техника-технологиялық белгіленімдер (кодтау, өңдеу даталары, орындаушылар, электрондық нұсқаның дереккөзі) жатады. Метадеректер жиыны мәтіндер корпустарындағы зерттеушілерге

ұсынылатын мәліметтер мүмкіндіктерін айқындайды. Сондықтан деректерді таңдау кезінде зерттеудің мақсатын және тілшілердің сұранысын және мәтінге қосымша белгілерді енгізу мүмкіндіктерін басшылыққа алу қажет болады.

Метамәтіндік белгіленімдер корпустың қай түріне болмасын енгізілуге тиісті маңызды ақпараттар болып табылады, сонымен қатар лингвистикалық белгіленімдер қойылатын корпустар құрастыру жұмысынан бұрын, корпусқа әртүрлі стильдерден алынған мәтіндерді енгізу барысында атқарылады.

Зерттеу жұмыстарының белгілі бір жүйе бойынша жүргізілетіні белгілі. Өйткені тіл – қарым-қатынас құралы, қолданыс аясы өте кең күрделі жүйе болғандықтан, тілдік зерттеулер белгілі автор шығармалары бойынша немесе белгілі бір кезеңде жарық көрген көркем не баспасөз немесе тарихи ескерткіштер бойынша, сол сияқты белгілі бір жанр, стиль немесе белгілі бір тақырып, мәселе бойынша тарам-тарам болып жіктеліп кете береді. Осындай әртарапта шашылып жатқан материалдарды корпус жадына жай ғана енгізбей, белгілі бір жүйеге бағындырып енгізсе, зерттеу жұмыстарының да тиімді орындалуына ықпал етеді. Ал бұлайша корпус жадына енгізілген мәтіндерді жүйелі құрылыммен беру метабелгіленімдер қоюдың теориялық және практикалық әдіс-тәсілдерін меңгергенде ғана мүмкін болады. Сондықтан корпустың қайсібір түрінде болмасын, метабелгіленімдер қою мәселесі дұрыс жүргізілуі керек. Бұл да теориялық бағытта зерттеліп, әрі қолданбалы бағытта жүзеге асырылатын көлемді де, күрделі әрі бүгінгі жаһандану заманында кезек күттірмейтін өзекті тақырыптардың бірі. Қазақ тілінің ұлттық корпусы тиімді құрастырылып, өз дәрежесінде қолданысқа енгізілуі үшін әлемдік тәжірибедегі метабелгіленімдер қоюдың ортақ ұстанымдары мен әдіс-тәсілдерін зерттеу қажет.

Метабелгіленімдер – ғылыми-зерттеу жұмыстарында белгілі бір кезеңге, стильге, авторға, тақырыпқа т.б. қатысты материалдар жинаудың таптырмас дереккөзі. Метабелгіленімдер қойылған корпустардан зерттеушілер өзіне қажетті стиль, кезең, автор т.б. ақпараттар бойынша мәлімметтерді тез тауып алуына мүмкіндік алады. Ал бұл жетістік қазіргі дамыған еліміздің ғылыми-зерттеушілік

әлеуетін, ғылым мен білімді жан-жақты дамытатын бірден-бір күш екендігі сөзсіз.

Корпустың қайсібір түрінде болмасын, ең алдымен жинақталған мәтін туралы ақпарат беретін метамәтіндік параметрлер компьютерлік базаға салынуы қажет. Сондықтан корпус құрастыруда метабелгіленімдер (шығарма атауы, авторы, жылы, шыққан орны, стилі, жанры, көлемі т.б.) жасаудың әдіс-тәсілдерін, түрлерін, компьютерлік бағдарламаларын т.б. зерттеудің маңызы зор.

Әлемдік тіл білімінде метабелгіленім қоюдың теориялық ұстанымдары мен практикалық әдіс-тәсілдері бар. Орыс тілінің ұлттық корпусындағы метабелгіленімдер қоюдың теориялық, практикалық мәселелерін С.О.Савчук қарастырған [61].

Корпусқа метабелгіленімдер енгізу жұмысынан бұрын атқарылатын жұмыс – әр түрлі стильге тән мәтіндердің электронды нұсқасын жинақтау болып табылады. Мәтіндер жинақтау жұмысын әртүрлі сипатта жүзеге асыруға болады. Олар: 1) интернет сайттарына салынған мәтіндерді қолдап іздеу арқылы WORD бағдарламасына көшіріп алу немесе PDF бағдарламасындағы мәтіндерді арнайы бағдарламалар арқылы WORD бағдарламасына көшіру; 2) интернет сайттарындағы мәтіндерді автоматты түрде пайдалана алатын бағдарламалар арқылы мәтін жинау; 3) кітап, газет мәтіндерін сканирлеу арқылы мәтін жинақтау; 4) баспалармен келісіп, әртүрлі мәтіндердің (көркем проза, поэзия, драматургия, газет-журнал, ғылыми-техникалық мәтіндер т.б.) электронды нұсқаларын алу; 5) ауызекі сөйлеу корпустары үшін таспаға жазылып алынған видео, аудио таспаларды телевидения, радио т.б. мекемелерден алу немесе ауызша тілдесімдерді таспаға арнайы жазып алу арқылы ауызша мәтіндер жинау т.б.

Корпуста алынатын мәтіндердің авторы, сол сияқты мазмұны, сапасы да ескеріледі. Қысқасы, корпуста белгілі бір кезеңдегі тіл қолданысын толықтай сипаттайтындай функционалды стильдер қамтылуы қажет.

Корпусқа белгілі бір әдеби тілдегі әртүрлі стильдер, әртүрлі авторлар шығармалары мен еңбектері енгізілетіндіктен, тілтұтынушысының оларды белгілі бір жүйемен пайдалануы үшін олар ту-

ралы егжей-тегжейлі ақпарат беріледі. Корпустың іздеу жүйесіне салынатын мәтіндер туралы ақпараттар (метабелгіленімдер) әр тілде әртүрлі болып келеді. Мысалы, Орыс тілінің ұлттық корпусында мәтін туралы ақпараттар 25 параметрге негізделген. Олардың ішінде 9-ы мәтінді сипаттайтын мәліметтер, 3-еуі автор туралы, 3-еуі мәтіндердің бағытталған аудиториясы, 4-еуі мәтіндер туралы библиографиялық мәліметтер, 5-еуі қосымша ақпараттар.

Нақтырақ айтсақ:

Бірінші блок:

1. *мәтін авторы*: аты-жөні, жынысы, туған уақыты (немесе шамамен есептегендегі жасы);

2. *мәтіннің атауы;*

3. *мәтіннің құрастырылу уақыты мен орны* (дәл немесе шамамен көрсетуге болады);

4. *мәтін көлемі:* көркем шығармалар үшін мынадай көлемдер қабылданған: әңгіме 5 мың сөзден көп емес; әдеттегі повестер ұзындығы – 5 мың мен 15 мың сөз аралығы; романның әдеттегі ұзындығы – 15 мың сөзден асады.

Екінші блок: мәтіндер корпусының негізгі үш *массив* параметрлерінің метасипаттамасы – көркем мәтіндердің; көркем емес мәтіндердің; драмалық шығармалардың. Мысалы, ОТҰК-да көркем әдебиет мәтіндері үшін мына деректер көрсетілген:

1. мәтін жанры: жанрға жатпайтын проза, өмірбаяндық проза, детектив, балалар әдебиеті, тарихи проза, криминал әдебиеті, оқиғалар, фантастика (қияли шығарма), әзіл-сықақ (юмор, сатира);

2. мәтін типі: өмірбаяндық проза, анекдот, ассоциативтік проза, боевик, детектив, очерк, әдеби хат, повесть, нақыл (притча), пьеса, әңгіме, роман, ертегі, триллер, эпопея, эссе және т.б.;

3. мәтін хронотопы: мәтінде сипатталатын оқиғалардың шамамен алғандағы орны мен мезгілі.

Нақты алғанда, мыналар ұсынылады: көне Шығыс; XVII ғасыр Ресей; XIX ғасыр Ресей; Ресей/СССР: толығымен кеңес кезеңі; Ресей, кеңес кезеңі – Германия 1920-1940-шы жылдар; Ресей/СССР – Еуропа 1960-1980-шы жылдар; Ресей/СССР: қайтақұру; Ресей/СССР: кеңестік және одан кейінгі кезең; Америка: қазір-

гі өмір; Израиль: қазіргі өмір; Орта Азия: қазіргі өмір; ирреалды әлем және т.б. Сонымен бірге «хронотоп анықталмаған» деген тэг те кездесуі мүмкін.

ОТҰК-дағы имплициттік метабелгіленімге жататындар:

1. «мәтін-стиль», мұндай жағдайда академиялық, ғылыми-көпшілікке, ресми-іскерлік, бейтарап, төмендетілген (сниженный), өрескел қарапайым тілдің элементтерімен және жаргон, архаизм, жекеше-авторлық, диалекті және т.б. (барлығы 21) бөлініп қарастырылады;

2. тыңдаушылар – жасы;

3. тыңдаушылар – білім деңгейі;

4. тыңдаушылар – өлшем (размер)

(толығырақ: http://ruscorpora.ru/corpora-parameter.html)

Қайсыбір мәтін болмасын оның *авторы* болады. Олар:

а) мәтіннің нақты авторы болған жағдайда оның аты-жөні толық көрсетіледі; ә) мәтін авторлары бірнешеу болған жағдайда ұжымдық авторлардың аты-жөні беріледі. Олар мәселен, ұжымдық монографиялар, бірлесіп жазылған мақалалар т.б.; б) жалпылама автор, мұндай мәтіндер жеке адамның емес, ұжымның, мекеменің атынан кететін мәтіндер (яғни құжаттар, хаттар т.б. мәтіндері); в) кейбір мәтіндер авторлары белгісіз де болуы мүмкін. Бұл әсіресе газет-журналдар мәтіндерінде көп кездеседі. Мұндай мәтіндер авторлары кейде шартты есімдермен де көрсетіледі. Мысалы, орыс тілінде «Иван Иванов», «Аноним», қазақ тілінде де мұндай шартты атаулар кездеседі. Мұндай авторы анық емес мәтіндердің метабелгіленімдерінде автор деген ұяшық толтырылмай бос қалдырылады.

Метабелгіленімдерде кейде *авторлардың жынысына* қатысты да ақпарат беріледі. Автордың әйел адам екендігі немесе ер адам екендігі немесе жынысы анық көрсетілмеуі де мүмкін. Әдетте автордың жынысы мәтін авторы біреу болған жағдайда көрсетіледі, ал ұжымдық мәтіндерде автордың жынысы көрсетілмейді. Алайда Британ ұлттық корпусында ұжымдық мәтіннің авторлары әртүрлі жынысты болған жағдайда «mixed» деген белгіленім қойылады.

Метабелгіленімнің авторға қатысты тағы бір түрі – *автордың жас ерекшелігінің де көрсетілуі.* Кейбір корпустарда автордың шығарманы жазған кездегі жас шамасы көрсетілсе (Британ, Чех),

кейбір корпустарда автордың туған жылы, күні туралы нақты, дәл мәліметтер беріледі немесе шамамен көрсетіледі (Орыс тілінің ұлттық корпусы). Ал автордың жасын анықтау қиын болған жағдайда «белгісіз» екендігі туралы белгі қойылады. Ал ұжымдық, жалпылама, белгісіз авторлар болған жағдайда жас ерекшелігі берілмейді. Ал кейде күнделік, жеке хаттар сияқты жеке басқа тән мәтіндер авторлары белгілі болғанмен, олардың аты-жөндері берілмей, шартты атпен беріліп, бірақ жынысы мен жасы көрсетіле береді.

Башқұрт тілінің зерттеушісі З.А.Сиразитдинов автордың, информанттың *ұлтын* да көрсеткен [62, 32 б.].

Корпусқа енгізілген *мәтіннің атауы* да – негізгі метабелгіленімдердің бірі. Корпусқа енгізілген мәтіндердің атауының, яғни тақырыптарының бәрі болмауы мүмкін. Егер мәтінде тақырыптар атауы берілсе, олар метабелгіленімдер жүйесіне салынады, ал тақырыптары берілмеген мәтіндердің атаулары көрсетілмейді. Бұлар әдетте газеттер мен журналдардағы бір рубрика ішінде берілетін қысқа мәтіндер, демек, корпусқа салынған мәтіндердің барлығы да табиғи тіл қолданысын сипаттайтындықтан, тақырыбы жоқ болса да алына береді, бірақ метабелгіленімдер жүйесінде көрсетілмейді.

Мәтін туралы метабелгіленімдердің бірі – *мәтіннің жазылу уақыты.* Әдетте мұндай белгіленімдер автордың шығарманы жазу барысында мәтіннің соңында қалдырған мәліметтерінен алынады. Көбінесе мәтіннің жазылу уақыты библиографиялық, өмірбаяндық зерттеулерден анықталады. Ал мәтіннің жазылу уақыты туралы нақты ақпарат болмаған жағдайда оның мерзімі 5-10 жылы аралағында шамамен алынады. Кейде мәтіндердің жазылу уақыты туралы нақты мәлімет болмаған жағдайда корпусқа салынған мерзімі алынады.

Метабелгіленімдердің бірі – *корпусқа енгізілген мәтіндердің әрбіріндегі сөзқолданыс саны туралы ақпарат.* Корпусқа мәтін енгізуде стильдер арасалмағы негізінен теңгерімді болғанмен, кейбір жанрларға қатысты мәтіндер көлемі әртүрлі болып келеді. Мысалы хабарландырулар, құттықтаулар, жаңалықтар т.б. өте қысқа, олардағы сөзқолданыс саны да он шақты сөзден тұруы мүмкін. Сондықтан корпусқа енгізілген метабелгіленімдердегі мәтіндер-

дегі сөзқолданыс саны оннан он мыңдаған сөзқолданысқа жетеді. Сөзқолданыс саны өте үлкен болып келетіндері – көбінесе корпусқа тұтастай енгізілген романдар немесе монографиялар сияқты көлемді еңбектер. Бірақ кейде мұндай шығармалар мен монографиялардың бәрі алынбай, белгілі бір тарауы ғана алынуы да кездеседі. Мәтіндердегі сөзқолданыс саны оларға берілген метабелгіленімдер жүйесінде көрсетіледі. Сөзқолданыс санын мәтінді енгізу барысында арнайы компьютерлік бағдарламалар арқылы анықтайды. Кейбір корпустарда мәтіндердегі сөйлемдер саны туралы да ақпараттар берілген. Мысалы, Браун корпусы.

Метабелгіленімдердің енді бір түрі мәтіннің *қолданылу саласы* болып табылады, ол мәтіннің ең жалпы типологиялық сипаттамасы. Мәселен, Орыс тілінің ұлттық корпусында 8 функционалды қолданыс саласы көрсетілген. Олар: *оқу-ғылыми, өндірістік-техникалық, ресми-іскери, публицистика, жарнама, діни, әдеби, тұрмыстық.*

Оқу-ғылыми қолданыс аясы ғылым мен білімнің әртүрлі салаларына жататын ғылыми және ғылыми-методикалық мазмұнды мәтіндерді біріктіріп қарайды. Ондай мәтіндерде осы қолданыс саласына сай тіл ақпарат беру қызметін атқарады.

Өндірістік-техникалық қолданыс аясы техникалық құрылғыларды және өндірістік үдерістерді сипаттайтын мәтіндердің қолданыс аясы болып табылады. Ол бір жағы оқу-ғылыми саламен шектессе, екінші жағынан іскери саламен де байланысты. Өндірістік-техникалық қолданыс саласында тіл ақпараттық және ықпал ету (нұсқаулықтар) қызметін атқарады.

Ресми-іскери салаға өндірістік, ауылшаруашылық, заңнамалық т.б. істер бойынша мемлекет пен жеке адамдар арасындағы, ұйымдар мен басқа мемлекеттер арасындағы, сондай-ақ ұйымаралық немесе ұйым ішіндегі немесе ұйым мүшелерінің арасындағы т.б. қатынастар сипатталатын мәтіндер жатады. Бұл салаға қатысты мәтіндер ықпал ету және ақпараттық қызмет атқарады.

Публицистика саласы халықты ақпараттандыру және саясат, экономика, өнер, ғылым, ахлақ және т.б. салалардағы қоғамдық маңызды мәселелер бойынша қоғамдық пікірлерді қалыптастастыруды мақсат ететін мәтіндерді біріктіреді. Бұл салада тіл

ақпараттық, ықпал ету және біршама эстетикалық та функция атқарады.

Жарнама саласында негізінен материалдық қажеттілікті қалыптастыруға бағытталған мәтіндер алынады. Ондай мәтіндердің мақсаты – қандай да бір тауарды сату үшін тұтынушыларға ол тауар туралы жағымды ақпарат беру, сатып алуға шақыру. Мұндай мәтіндерде тіл ақпараттық және ықпал етуші қызмет атқарады.

Діни сала діни мазмұнда жазылған мәтіндерді біріктіреді. Мұндай мәтіндерде тіл ақпараттық және ықпал етуші функция атқарады.

Тұрмыстық сала адамдар арасындағы күнделікті қарым-қатынастағы бейресми түрдегі мәтіндерді негізгі алады, көбінесе ауызша формада болып келеді. Тұрмыстық саладағы мәтіндердің жазбаша формалары да бар. Олар: жеке хаттар, күнделіктер, электрондық хаттар, телефондағы хаттар, құттықтаулар т.б.

Әдеби сала – автордың өмір туралы ой-пайымдаулары көрініс табатын көркем шығармашылық сала. Мұндай мәтіндерде тіл ықпал етуші ғана емес, эстетикалық та қызмет атқарады.

Мәтіндерді белгілі бір *тақырыптық топтарға* бөліп беру де метамәтіндік белгіленімдердің біріне жатады. Мысалы, *қоғамдық ғылымдар, физика, биология, саяхат, спорт, табиғат, өнер, саясат* т.б.

Алайда бұлай тақырыптық топтарға бөлу кейде шартты болып келеді. Өйткені кейбір мәтіндер бірнеше тақырып аясында, бірнеше салаға қатысты қарастырыла береді. Сондықтан белгілі бір мәтіннің тақырыптық топқа қатысын көрсеткенде, олар бір салаға ғана емес, бірнеше салаға ортақ болып та келеді. Әлемдік корпустарда (Браун корпусы, Орыс тілінің ұлттық корпусы т.б.) көркем әдебиеттерде көбінесе тақырыптық топ көрсетілмейді.

Метабелгіленімдердің енді бірі – *хронотоп.* Кейбір мәтіндерді тақырыптық топтар бойынша бөлу қиындық тудыратындықтан, кейбір корпустарда мәтіннің жазылған уақыты мен жазылған жерін көрсету де қажет болған, яғни мәтіннің *жазылған жері* мен белгілі бір *кезеңге қатысы* көрсетіледі. Мысалы, Алматы, 1998 жыл немесе Ресей, 1945-50 ж.ж. немесе Қазақстан, Кеңес өкіметі жылдары т.б.

Келесі бір метабелгіленімдердің бір түрі – *мәтін типі.* Мұнда мәтіннің белгілі бір жанрға қатысы көрсетіледі. Мысалы оқу-ғылыми сала жанрлары: мақала, монография, оқулық, реферат т.б., публицистикада: күнделік, репортаж, интервью т.б., ресми-іскери салада: заң, қара, бұйрық, акт т.б., әдеби салада: роман, повесть, әңгіме т.б. Алайда жанр деген термин әдебиет саласында да өзіндік мәнге ие болғандықтан, онымен шатастырмас үшін корпус құрастыруда кейде оны «мәтін типі» деген терминмен де қолданады.

Браун корпусында мәтіндердің мұндай жанрға қатыстылығы туралы метабелгіленім қойылмаған. Мәтіннің белгілі бір салаға қатыстылығы, мәтіннің жазылу уақыты және алынған дереккөзі көрсетіледі. Алайда Браун корпусын құрастырушылар Д.Лидің мақаласында мәтіндердің жанрға қатысына қарай жаңа база жасалғанын айтады. Д.Ли 70 жанр мен субжанрларды көрсетеді. Алайда оның классификациясындағы жанрлық жіктемеде нақты жанрлармен бірге (өмірбаян, мектеп шығармалары, жеке хаттар, іскери хаттар) тақырыптық топтар да жанр ретінде беріліп, араласып кеткендігі айтылады.

Браун корпусындағы мәтіннің типологиялық сипаттамасы төмендегідей:

Мәтін типі: әдеби, ақпараттық, ауыспалы типтер т.б.

Жанр типі (60-қа жуық): драма, роман, музыка, философия, спорт, өндіріс т.б.

Субжанр типі: оқулық, мақала, энциклопедия т.б.

Мәтін типі: өлендер немесе проза.

Мәтін стилін көрсету арқылы мәтіннің тілдік формасы, әсіресе мәтіннің лексикалық құрамы анықталады. Олар: әдеби, әдеби емес стиль, бейтарап стиль, ресми стиль, арнайы (ғылыми) стиль т.б. Көркем прозада мынадай стильдер: бейтарап, аймақтық, қарапайым, жеке-авторлық стильдерді беруге болады.

Мәтіндерге осылайша стилистикалық сипаттама жасау корпустарда толық шешімін таппаған, сондықтан стилистикалық метабелгіленімдер жасау өзекті мәселе.

Метабелгіленімдердің келесі түрі – *аудиторияның жасын көрсету.* Мәтіннің кімге арналғанын білу мәтіннің мазмұны мен онда

қолданылатын тілдік құралдарды да айқындайды. Мәтіндерге мұндай жас ерекшелігіне қарай белгіленім қою балалар әдебиетін, белгілі бір жас кезеңдеріне арналған оқулықтарды табуға мүмкіндік береді. Балалар әдебиеті әдетте 1-10 жас, жасөспірімдік кезең 11-17, жастар әдебиеті 18-34 жас аралығы болып келеді. Немесе тек үлкен адамдарға арналғанын көрсететін белгіленім қойылады. Кейде мәтіндер жас ерекшелігіне қатысты көрсетілмейді, яғни аудиториясы бейтарап болып келеді.

Метабелгіленімдердің келесі түрі – мәтіннің кімге, бұл жерде *аудиторияның білім дәрежесіне* қарайғы ерекшелігін көрсету. Бұл жерде мәтін аудиториясының жалпы білімі немесе арнайы білімі, сондай-ақ жоғары білімі немесе білімінің төмендігі сияқты сипаттары негізге алынады. Себебі арнайы кәсіби салада жазылған мәтіндердің өзіне тән терминологиясы болады. Ешқандай кәсіби білімсіз жалпыға ортақ мәтіндер де болады. Сондықтан аудиторияның білім дәрежесі ескеріліп, пайдаланушы ортаны көрсететін метабелгіленімдер қоюға болады. *а) жоғары білімді ортаға арналған, ә) кәсіби білімді қажет ететін, б) кәсіби білімі жоқ, жоғары білімі жоқ ортаға арналған, в) бейтарап орта мәтіндері.*

Метабелгіленімдердің келісі түрі – *мәтіндерді пайдаланушы ортаның саны, көлемі.* Кейбір мәтіндер жалпы көпшілікке арналып, яғни мыңдаған, миллиондаған адамға арналса, отыз шақты адамнан тұратын топқа немесе бір ғана адамға арналған болуы да мүмкін. Жалпы көпшілікке арналған мәтіндер көбінесе баспа беттеріне шыққан мәтіндер, электронды қарым-қатынасқа арналған мәтіндер болса, топқа арналған мәтіндер оқу лекциялары, кеңсе құжаттары т.б., ал жеке аудиторияға арналған мәтіндер көбінесе жеке хаттар болып келеді.

Метабелгіленімдердің келесі түрі – мәтін алынған *дереккөздер* болып табылады. Мәтін жинаудың түрлі әдіс-тәсілдері бар. Электронды кітапханалардан, интернетке салынған сайттардан, газет-журналдар, кітаптар шығарылатын баспалардан, жеке адамдардан алуға болады немесе қолдап сканир жасалады немесе қолдан теріледі. Жарық көрмеген мәтіндер қолжазба ретінде көрсетіледі. Интернеттен алынған мәтіндерде сайт аты беріледі. Жарық көрген газет-журналдардың шыққан уақыты алынады.

Мәтіндер жарық көрген баспа атын да көрсету – метабелгіленімнің бір түрі. Бұл әсіресе кітаптарға арналған.

Метабелгінімнің келісі түрі – *кітаптың немесе газет-журналдардың шыққан, жарық көрген жылын көрсету.*

Корпусқа енгізілген мәтіннің корпусқа енгізуге дейін қандай формада (электронды, кітап, газет-журнал, іскери құжат т.б.) болғанын көрсету де қажет.

Метабелгіленімдердің тағы бір белгіленімі есебінде *қосымша ақпараттарды* көрсетуге болады. Олар мәселен: а) Мәтіннің электронды нұсқасының сапасы жайлы ақпарат; ә) Подкорпустардың шартты атаулары туралы ақпарат (омонимдері ажыратылған немесе ажыратылмаған корпус, ауызша корпус, жергілікті корпус т.б.; б) Корпус туралы комментарийлер (мұнда корпус туралы қосымша ақпарттар); в) Корпусты құрастыруда мәтіндер алуға көмектескен, атсалысқан ұйымдар туралы ақпарат т.б.; г) Корпусты құрастырушылар мен жауапты жетекшілер туралы ақпарат т.б.

Башқұрт тілінің корпусын құрастырушылар мынадай метабелгіленімдер енгізілгендігін көрсетеді. Олар:

Мәтіннің паспорты (барлық мәтіндерге арналған):

а) мәтін авторы;

ә) мәтін атауы;

б) мәтін көлемі (сөзқолданыс саны);

г) мәтіннің жазылған уақыты.

Көркем мәтіндер үшін а) мәтін типі (повесть, әңгіме, роман, ертегі, триллер, эпопея, эссе т.б.)

Публицистикалық мәтіндер үшін:

а) *мәтін типі (журнал, газет);*

ә) мәтін тақырыбы, саласы:

– саяси және әлеуметтік өмір;

– философия;

– экономика;

– ауылшаруашылық;

– өнер, мәдениет, әдебиет;

– ғылым және техника;

– білім;

– табиғат және саяхат;
– күнделікті өмір;
– спорт;
– дін;
– психология;
– медицина;
– сұлулық және денсаулық т.б.

б) мәтін жанры:
– интервью, сұхбат;
– мақала, очерк, репортаж, шолу;
– кеңестер;
– хаттар;
– құттықтаулар;
– көркем жанрлар; рецензия т.б.

Сонымен, жоғарыда метабелгіленімдердің түрлері туралы қысқаша сипаттама бердік. Бұлар алдағы уақытта Қазақ тілінің ұлттық корпусын құрастыру барысында негізге, басшылыққа алатын мәселелер болып табылады. Корпус жасауда метабелгіленімдерді енгізу қажеттігі дау тудырмайды. Олар корпусты пайдаланушыға, әсіресе тілші-зерттеушілерге кез келген қажетті ақпаратын тез әрі оңай тауып алуға мүмкіндік береді. Сондықтан метабелгінемдер енгізудің әлемдік тәжірибесін енгізу үшін оның теориялық мәселелерін айқындап алу қажет.

Мәтіндерге метасипаттау жүргізуге мүмкіндік беретін компьютерлік бірнеше лингвистикалық бағдарламалар бар. Ол, ең алдымен, Systematic Coder және UAM Corpus. Бағдарламалардың деректеріне еркін қол жеткізуге, тегін пайдалануға болады және корпустық лингвистикада қабылданған стандарттарға сай келеді. Осы бағдарламаның негізінде әртүрлі тілдердің пайдалануға берілген корпустарында мәтіндердің метасипаттау архитектурасы жасалған.

Енді метабелгіленім қоюдың бағдарламалық жолдары төменде сөз болады. Бұл бағдарлама қазақ тілінің ұлттық корпусының бес миллион сөзқолданысты қамтитын мәтіндері бойынша қойылған метабелгіленімнің бастапқы нұсқасы бойынша баяндалады.

2.8. Метабелгіленім енгізілген мәтіндерді арнайы компьютерлік бағдарлама бойынша өңдеу

Қазақ тілінің Ұлттық корпусын арнайы құрастырылған компьютерлік бағдарламалармен қамтамасыз ету жұмысы Visual-Studio 2010 аспаптар ортада C# бағдарламалық тілінде жүргізілді. Ал морфологиялық белгіленім жүргізілген мәтіндер корпусын, сөйлемдерді, сөздерді MSSQLServer 2008 атты жадында сақтау қорында жүзеге асты.

Мәтіндер корпусы дегеніміз – әрбір автордың немесе стильдердің (жанрлардың) MSWORD форматындағы мәтіндер топтамасы. Әрбір шығарманың басында метабелгіленім орын алады, мысалы:

<Жанр> Проза<\Жанр>

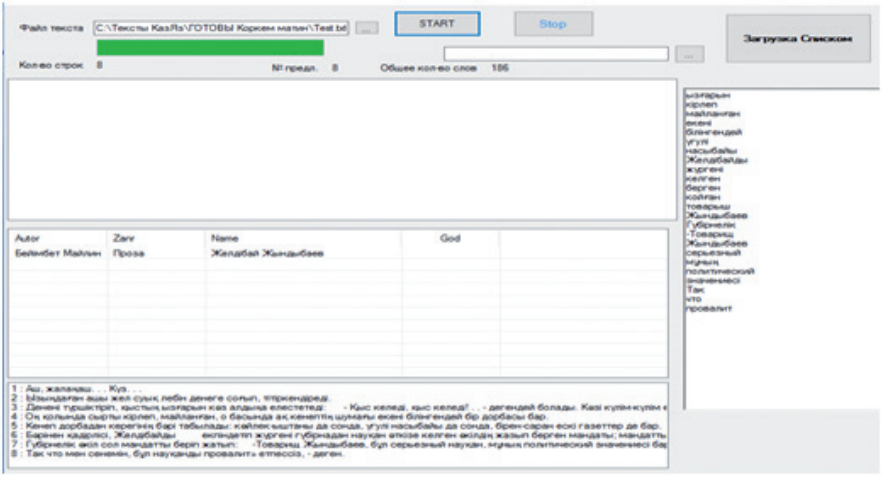
<Автор>Әбіш Кекілбаев<\Автор>

< Тақырыбы>Аңыздың ақыры<\ Тақырыбы >

< Тақырыпша > Қызыл алма<\ Тақырыпша >

<Жылы>1983<\Жылы>

Тегтер түріндегі метабелгіленімдер бойынша шығарма жан-



5-сурет. Метабелгіленімдер енгізілген файлдарды өңдеу бағдарламасы

ры, шығарма авторы, шығарма аты, тақырыпша аты және баспадан шыққан жылы.

Мәтіндер файлдарын, олардың тізімі бойынша немесе жекелік сипатын автоматты түрде өңдеудің компьютерлік бағдарламасы C# тілінде жазылды.

Корпус жадына енетін мәтін рет-ретімен келесі аталған өңдеу модульдерінен өтеді:

– Метабелгіленім тегтерін өңдеу модулі;

– Мәтін ішінен сөйлемдерді жекелеп бөліп алу модулі;

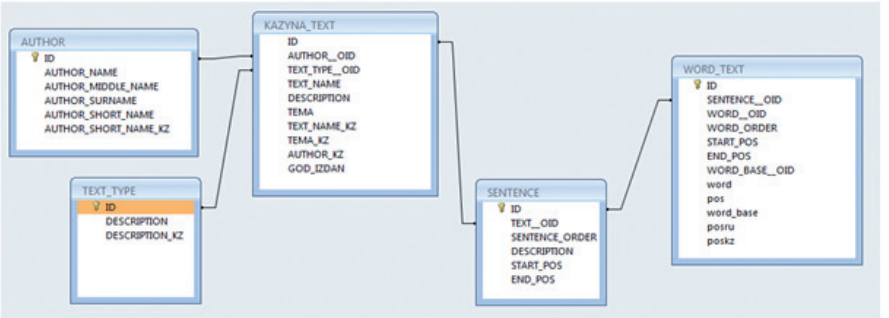
– Сөйлемдер ішінен сөздерді жекелеп бөліп алу модулі;

– Морфологиялық талдау (анализдеу) модулі;

– Деректер базасындағы өңделген ақпаратты корпуста сақтау модулі.

Аталған бағдарламаның мақсаты – ақпаратты іздестіру жүйесінің деректер базасын (қорын) қалыптастыру. Мысалы, корпус мәтіні ішінен кез келген сөз бойынша және қажетті деген шығарма авторы бойынша осы сөз орын алатын барлық сөйлемдерді іздеп табу. Осы мақсатқа жету үшін әрбір модуль әрекетінің нәтижесі сақталатын деректер базасын жобалау қажет (6-сурет).

Тегтерді өңдеу модулінің әрекеті нәтижесінде осы модульдің шыға берісінде шығарманың авторы, шығарманың аты және баспадан шыққан жылы жайлы ақпарат аламыз. Бұл мәліметтер үш түрлі кестелерде: ***Author, Kazyna_Text, Text_Type*** сақталады.



6-сурет. Деректер базасының сызбасы

18-кесте

Author кестенің құрылымы

Өріс, түрі	Өріс сипаты
Id	Идентификатордың кілті
AUTHOR_NAME	Автордың аты
AUTHOR_MIDDLE_NAME	Фамилиясы
AUTHOR_SURNAME	Әкесінің аты
AUTHOR_SHORT_NAME	Автордың фамилиясы орыс. инициалымен
AUTHOR_SHORT_NAME_KZ	Автордың фамилиясы қаз. инициалымен

Авторлар кестесінен негізгі өріс Kazyna_Text кестесінің AUTHOR__OID өрісімен байланысты. Kazyna_Text кестесінде шығарма аты, шығарманың тақырыбы, жанрдың аты сақталады. Сонымен, арнайы кілт бойынша кез келген автордың барлық шығармаларын тауып алуға болады.

19-кесте

Kazyna_Text кестесінің құрылымы

Өріс, түрі	Өріс сипаты
Id	Идентификатордың кілті
AUTHOR__OID	Авторға сілтеме
TEXT_TYPE__OID	Шығарма жанры
TEXT_NAME	Шығарма аты /орыс.
DESCRIPTION	Сипаттама
TEMA	Шығарма тақырыбы
TEXT_NAME_KZ	Шығарма аты /қаз.
TEMA_KZ	Шығарма тақырыбы /қаз.
GOD_IZDAN	Баспадан шығу жылы

Kazyna_Text кестесінде TEXT_TYPE__OID шығармасының жанрына қатысты сілтеме бар. Бұл кілт 6-суретке сәйкес ID өрісі бойынша Text_Type кестесімен байланысты. Text_Type кестесінде шығармалардың барлық жанрлары сақталады.

20-кесте

Text_Type кестесінің құрылымы

Өріс, түрі	Өріс сипаты
Id	Идентификатордың кілті
DESCRIPTION	Шығарма жанрының аты

Мәтіннен сөйлемдерді бөліп алу модулінде белгілі символдарға аяқталатын (!, ?. , ,) барлық сөйлемдер орын алады. Сөйлем соңын бейнелейтін осы символдарға дейінгі барлық сөздер толық сөйлем деп саналады да Sentence кестесіне жазылады.

21-кесте

Sentence кестесінің құрылымы

Өріс, түрі	Өріс сипаты
Id	Идентификатордың кілті
TEXT__OID	Шығарма атына сілтеме
SENTENCE_ORDER	Сөйлемнің рет санының номері
DESCRIPTION	Сөйлем мәтіні
START_POS	Мәтіндегі бастапқы позиция
END_POS	Мәтіндегі соңғы позиция

Sentence кестесінде TEXT__OID өрісі бойынша Kazyna_Text кестесіне сілтеме бар. Мұнлай байланыс кез келген таңдалып алынған шығарма авторы бойынша барлық сөйлемдерді тауып алуға мүмкіндік жасайды. Морфологиялық талдаудың соңғы модулінде сөйлемдердің барлық сөздерін тауып алып, оларды морфологиялық өңдеуден өткізіп, нәтижесінде сөйлемнің әр сөзіне морфологиялық белгіленім тіркеліп жазылады. Бұл модуль әре-

кетінің соңында біздер морфологиялық белгіленімнен өткен сөздер тізімін және сәйкес келетін мәтіндегі сөйлемнің реттік номері тіркеледі. Аталған ақпараттың барлығы да Word_Text кестесінде сақталады (22-кесте).

Word_Text кестесінде Sentence кестесіне сілтеме бар. Сонымен, әрбір сөз бойынша оның мәтіндегі сөйлемдермен байланысы бар екендігін, яғни шығарма аты мен шығарма авторы бойынша оларды қадағалап отыруға болады.

22-кесте

Word_Text кестесінің құрылымы

Өріс, түрі	Өріс сипаты
Id	Идентификатордың кілті
SENTENCE__OID	Сөйлемдер кестесіне сілтеме
START_POS	Сөйлемдегі бастапқы позиция
END_POS	Сөйлемдегі соңғы позиция
Word	Сөзформа
Pos	Морфологиялық белгіленім
word_base	Негіз сөз
Posru	Морфологиялық белгіленімнің қысқаша формасы /рус.
Poskz	Морфологиялық белгіленімнің қысқаша формасы /каз.

2.9. Корпустық лингвистикадағы стандарттау

Корпустар, әдетте, көпшілік пайдаланушылардың бірнеше рет қолдануына арналғандықтан, олардың белгіленімі мен лингвистикалық қамдалуы белгілі дәрежеде бірыңғайланған (универсал) болуы қажет. Корпусқа қатысты стандарттау мәселесін сөз қылғанда белгіленім түрлерінің үйлесімділігі жайында сөз болады. Оларды кейбір жағдайда «кодтаудың стандарты» деп те атайды. Сол сияқты

әртүрлі корпустарды бір-бірімен салыстыру, яғни корпустардың тілдік мәселелерге қатысты әртүрлі тапсырмаларды орындай алу мүмкіндіктеріне баға беру де маңызды мәселеге жатады. Оларды «бағалау стандарттары» деп атайды [63].

Белгіленімдер жайлы айтатын болсақ, онда лингвистикалық та, экстралингвистикалық та белгіленімдер түрлері кең тараған және көпшілік мақұлдаған мәтіндер мен тілдік бірліктерді сипаттау принциптеріне негізделуі қажет. Белгіленім параметрлері және оның мағыналары жеткілікті түрде «табиғи» болулары, яғни олар көпшілік мақұлдаған ғылыми классификацияға сәйкес келулері керек. Корпус-менеджерлердің лингвистикалық және программалық қамтамасыз етілу жағдайы типтік сұраныстарды өңдеу мен типтік мәселелердің шешімін табуда қолдау көрсететін болуы қажет.

Корпуста өңделетін мәліметтердің (деректерді) бірдей форматта дайындалуы, көпшілік жағдайда, бірдей программалық қамтамасыз ету бойынша пайдалануға және корпустық мәліметтермен айырбас жасауға мүмкіндік береді. Сондықтан, бір жағынан, мәліметтерді бірдей форматта дайындау, яғни оларды стандарттау туралы, екінші жағынан, мәліметтер құрылымын стандарттау жайында да сөз етуге болады.

Ең бір қиындық тудыратын мәселе, ол ауызша тілді транскрипциялауды стандарттау. Ауызша тілді графикалық белгілеу аясында бірыңғай және міндетті түрдегі стандарттың болмауына қарамастан, біршама жетістіктерге қол жетті деуге болады, ал табиғи тілдік коммуникацияның бейвербалды (ауызша емес) құрамын сипаттауда стандарттардың әлі қалыптаса қоймауы бұл саланың әрі қарайғы дамуына қиындық тудыруда [8].

Корпустарға қатысты стандарттау, деректер типтерінің үйлесімділік мәселелері әртүрлі корпустарды өзара салыстыра қарастыру үшін аса маңызды. Себебі корпустар сандық сипаты бойынша ғана емес, сапалық жақтары бойынша да бағаланады. Корпустар бойынша сандық мәліметтер олардың көлемі, әртүрлі критерийлер бойынша толықтырылуы, корпустардың (немесе субкорпустардың) лингвостатистикалық параметрлері бойынша бағалауға мүмкіндік береді. Ал бұл жердегі сапалық баға беру ұғымын

корпустардан алынатын мәліметтер негізінде жүзеге асатын бағалау мен салыстыруларды талдау нәтижелері деп түсінген жөн.

Корпустардың әрқилы лингвистикалық мәселелерді шешуге жарамдылық мәселесі де «бағалау стандартының» болуын қажет етеді. Қазіргі кезде халықаралық тәжірибенің негізінде Text Encoding Initiative (TEI) атты жобасын және EAGLES (Expert Advisory Group on Language Engineering Standards) ұсыныстарын басшылыққа алып, мәтіндерді сипаттайтын метадеректерді бейнелейтін де-факто стандарттар жүзеге асып отыр.

Дж. Синклердің классификациясы бойынша, мәтін тіліне әсер ететін екі түрлі фактор ажыратылады: сыртқы тілден тыс факторлар (E – external) және ішкі факторлар (I – internal). Бұл стандарттар орыс тілі үшін С.А.Шаровтың басшылығымен жасалып (http://ruscorpora.ru/sbornik2005/05savchuk.pdf), орыс тілінің ұлттық корпусын құрастыруда қолданыс тапқан болатын.

Мәтіндерге белгіленім жүргізудің формалды тілі ретінде корпуста SGML және XML программалық тілдер кең қолданыс табады. EAGLES стандарттар XML тілінің технологиялық ортасына тікелей қосылады (мысалы, қара: стандарт Corpus Encoding Standard for XML (XCES).

Бұдан басқа әрбір сөз және оған қатысты барлық ақпарат жеке жолда арнайы берілетін «вертикал формат» деп аталатын белгіленім тәсілі де кең қолданылады. Бұл форматта лингвистикалық параметрлер кодтаудың позициялық жүйесінде жиі беріледі және ондай жүйенің әрбір позициясына белгілі грамматикалық мағына сәйкес келеді. Мұндай белгіленімге Чех ұлттық корпусы мысал бола алады.

Морфологиялық белгіленім стандарттарының бірі ретінде көптілді морфосинтаксистік спецификацияны атауға болады (multilingual morphosyntactic specifications) MULTEXT-East Version 4 (http://nl.ijs.si/ME/V4/).

3. КОРПУСТАРДЫ ПАЙДАЛАНУ

3.1. Корпустық менеджерлер

Корпус – ізденіс жүйесі. Қолжетімді ашық жүйедегі мәтінді өңдеуге арналған үлкен көлемді құралдарды (компьютерлік бағдарламалар) қолдану мәтіндер жиынын электрондық өнімге айналдырып, олардағы лингвистикалық ақпаратты жинақтайды және зерттеушінің мақсатына қарай, оларды қажеттілікке сай өңдейді.

«Мәтіндер корпусы» деген ұғымның ажырамас бөлігі ретінде мәтіндік және тілдік деректерді басқару жүйесі болып табылады. Бұл жүйе соңғы кезде ***корпустық менеджер*** (немесе корпус-менеджер) деп жиі аталып жүр. Корпустық менеджер – корпустағы деректерді іздеп-табуға арналған программалық құралдардың жиынын құрайтын мамандандырылған ізденіс жүйесі. Бұл жүйе пайдаланушыға белгілі бір қажетті нысан туралы статистикалық ақпарат бере алады және пайдаланушыға қажетті деген нәтижелерді ыңғайлы пішінде тауып береді.

Корпустық менеджер міндеті:

– KWIC-ті (Key Word In Context) құрастырумен бірге, толық конкорданстық тізімді де жасау;

– жеке сөзді ғана іздемей, сонымен бірге, сөзтіркестерді де іздестіру;

– үлгі (шаблондар) бойынша іздеуді жүзеге асыру (күрделді сауалдарды да);

– пайдаланушының таңдауы бойынша, тізімдерді бірнеше критерийлер бойынша сұрыптау;

– табылған сөзформаларды шексіз контекстерден көрсетуге мүмкіндік туғызу;

– корпустың жеке элементтері бойынша статистикалық ақпарат беру;

– корпустың белгілену дәрежесіне қарай, леммаларды, сөзформалардың морфологиялық сипаттамасын және метадеректерді (библиографиялық, типологиялық) бейнелеп көрсету;

– нәтижелерді компьютер жадында сақтау және қағазға басып шығару;

– жеке файлдармен әрекет етумен бірге, шексіз көлемді корпустармен де жұмыс істеу;

– сауалдарды тез арада өңдеу және нәтижелерді пайдаланушыға ұсыну;

– мәтіндік деректердің әртүрлі форматтарын қолдау (txt, doc, rtf, html, xml және т.б.);

– корпусты тәжірибелі тұтынушыға да және жаңадан бастаған пайдаланушыға да бірдей дәрежеде қолдануға жеңіл (интуитивті ұғымды) болуын қамтамасыз ету.

Мынадай әмбебап корпустық менеджерлер SARA, XAIRA (BNC), Manatee/Bonito, CQP, DDC көпшілік пайдаланушыға айырықша таныс деуге болады. Корпустық деректерді өңдеу үшін де деректер базасын басқару жүйесі (орыс. СУБД, қаз. ДББЖ) немесе ізденіс жүйесі негізінде менеджерлердің жаңа үлгілерін зерттеп, өндіріске енгізуге болады.

Мысалы, орыс тілінің Ұлттық корпусы бойынша қажетті деректерді іздестіру әрекеті Яndex.Server 3.8 Professional ізденіс жүйесі арқылы жүзеге асады. Грамматикалық және метамәтіндік ақпаратқа ізденіс жасау үшін құжаттардың жасырын сипаттары мен мәтін фрагменттеріне ізденіс жасайтын Яndex.Server атты ізденіс жүйесі іске қосылған. Бұл ізденіс жүйесі веб-серверде немесе корпоративтік желіде орыс тілі морфологиясы туралы қоса ақпарат беретін толықмәтінді іздеп-табуды жүзеге асырады. Интернеттегі Яндекс ізденіс жүйесі қалай істесе, дәл осылай аталған жүйе орыс, ағылшын, украин тілдерінің морфологиясы бойынша іздестіру әрекетін орындайды. Мысалы, егер орысша «идти» сөзін табуға сауал түссе, онда ізденіс жүйесі жұмысының нәтижесі бойынша «идти», «идет», «шел», «шла» және т.б. сол сөздің мағынасын ашатын сөздер ұсынылады. Іздестіру әрекетінің нәтижесі релеванттық *(ақпараттық сауал* мен *алынған хабар мазмұнының сай келуі)* бойынша реттелген құжаттар тізімі болып табылады. Ақпараттың релеванттық қасиеті арқылы тек табылған құжаттар саны ғана ескерілмейді, сонымен бірге сөздердің контрастылығы, яғни олардың қолдану жиілігі мен сөздер арақашықтығы да анықталады [10].

3.2. Сауалдар тілі

Ақпараттық сауал – ол, белгілі ақпараттық сұраныстың сөздік көрінісі (бейнесі). Сауалдар пәндік және формалдық мазмұндары бойынша талданады және корпуспен жұмыс істейтін қолданбалы бағдарламаның сауалдар тілінің сөздігі терминдерінде сипатталады. Іздеу рәсімі сауалдың іздестіру бейнесін корпустың жеке элементтерімен салыстыра отырып, олардың сәйкестігін есепке алып отырады. Егер ондай сәйкестік кездесіп жатса, корпус элементтері релевантты *(ақпараттық сауал* мен *алынған хабар мазмұнының сай келуі)* деп саналады да және оларды пайдаланушыға ұсынуға болады.

Жалпы алғанда, сауалдар тілі моделі мынадай элементтерден тұрады:

1. Тікелей ізденіске қатысы бар элементтер (ақпараттық сұранысты білдіретін терминдер және т.б.);

2. Сауалдың мәтіндік элементтерін морфологиялық нормалау құралдары;

3. Булевтік операторлар (конъюнкция, дизъюнкция, отрицание);

4. Сызықтық грамматика құралдары/амалдары (қашықтық операторлары, позициялық операторлар);

5. Ізденістің қосымша шарты:

– корпустың белгілі өрістерінен (жерінен) іздестіру (мысалы, тэгтер ішінен);

– ізденіс өрісін шектеу (белгілі авторлардың шығармаларынан, құжаттардың жазылған датасы бойынша, олардың типтері бойынша және т.б. іздеу);

6. Шығарылатын нәтижелерді сұрыптауға (саралау) талап қою;

7. Ізденіс нәтижелерін белгілі формаға келтіруді талап ету:

– шығарылатын нәтижелердің көрінісі;

– шығарылатын нәтижелердің саны.

Енді ең бір нәтижелі корпустық менеджер – **Bonito/Manatee** сауалдар тілін қарастырайық. Мұндағы, Bonito – менеджердің аты, Manatee – корпустық қамтамасыз етудің барлық программалық жүйесі. Осы ізденіс жүйесінің негізінде көптеген мәтіндер корпус-

тарына сауалдар тілінің негізгі элементтерінің жұмысы пайдаланушыға көрсетіледі және корпустарға қойылған сауалдар тапсырмалары мысалдармен түсіндіріледі.

Bonito деп аталатын корпустық менеджер дегеніміз – мәтіндер корпустарымен жұмыс істеуге арналған программалық қамтамасыз ету жүйесі. Бұл жүйе екі бөліктен тұрады: серверден (Bonitosrv) және Bonito – графикалық пайдаланушы интерфейсінен (GUI – graphical user interface). Жүйе ғалым П. Рыхлидің басшылығымен Чехиядағы Масарики Университетінде құрастырылған болатын.

Аталған жүйенің жұмысымен толық танысу үшін ағылшын тілінің SUSANNE (Surface and Underlying Structural Analysis of Natural English) мәтіндер корпусы қолданылған болатын (http://www.grsampson.net/). SUSANNE корпусы Ұлыбританиядағы Сассекса Университетінде жасалып, SUSANNE сызбасына сәйкес аннотациялаудан өткен 130 мың сөзді қамтитын американдық ағылшын тілінің Браун корпусына негізделген.

Bonito жүйесінің негізгі ерекшеліктері

Сауалдар тілі

– жеке атрибуттарды іздеу (сөзформа, лемма, тэг);
– тұрақты айтылымды қолдану;
– логикалық операторлар;
– құрылымды бейнелейтін құралдар (сөйлем шекарасы және т.б.);
– күрделі сауалдарды тез өңдеу;
– үлгілер (шаблондар).

Конкорданстық тізімдер

– пайдаланушының сауалдар тарихы;
– сөзформалардың морфологиялық сипаттамаларын қарау;
– лемманы бейнелеу.

Конкорданса арналған амалдар

– тізімдерді файлға сақтау;
– тізімдерді басып шығару;
– негізгі сөздер бойынша, контекст бойынша сұрыптау;

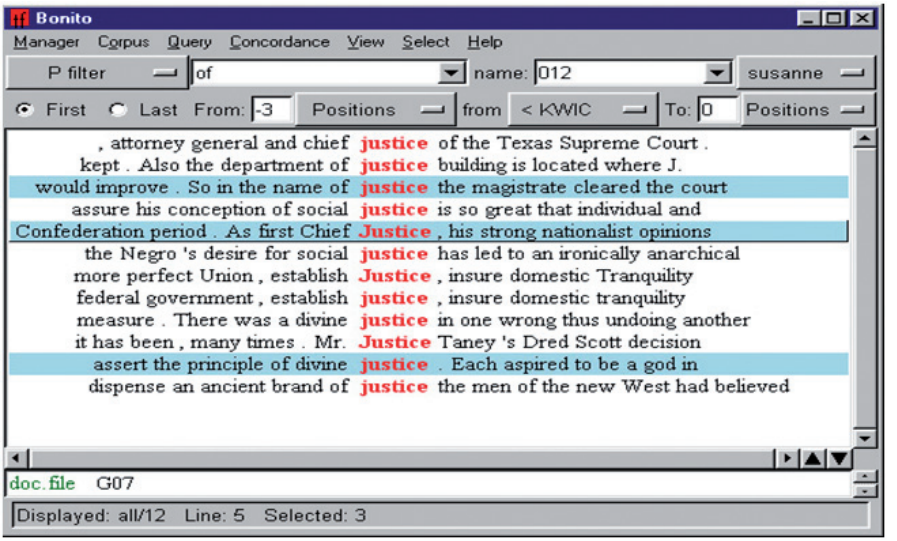
– контексті интерактивті шексіз кеңейту;
– фильтрлеу (құрылған конкорданстар бөлігін жою);
– қайталауларды жою.

Жиілікті үлестірім

– корпустағы, контекстегі сөздердің және басқа атрибуттардың жиілігі;
– топтастыру деңгейлерінің шексіз саны

Басқа да ерекшеліктер

– кодтауды таңдау;
– пайдаланушылардың субкорпусын құрастыру;
– тэгтердің еркін жиынтығы;
– басқа тілдерді қосуға мүмкіндік.



7-сурет. «Justice» сөзформасы үшін конкордансы бар Bonito корпус-менеджерінің терезесі

Сауалдар

Сауалдар жүйесі тілінің ережесі немесе арнайы үлгі (пайдаланушы ұсынған) бойынша тұжырымдалған сауалды корпус-менед-

жерінің терезесіне пайдаланушының өзі енгізуіне болады (7-сур. қара).

Сауалдар түрлері:

Оң сүзгі (P-filter) – сауалдар жолындағымен бірдей келгендер конкорданстік тізімінде беріледі;

Теріс сүзгі (N-filter) – сауалдар жолындағымен бірдей келгендер конкорданстік тізімінен алынып тасталады;

Сөзтіркестер (Collocations) – сауал позициясына жауап беретін (берілген интервалдағы нақты сөзформа) конкорданста түсті бояумен ерекшелінеді.

Оң, теріс сүзгілерге және сөзтіркестерге конкорданстың әр жолы позицияға тура келетін аралықта іздейтіндей интервал белгіленуі қажет. Пайдаланушы интервалдың шегін өзі таңдайды (енгізу терезесі «**From:**» және «**To:**»). Егер мәні оң болса, онда іздеу бастапқы позициядан оңға қарай ұйымдастырылады, ал егер мәні теріс болса – солға. Бастапқы позиция ретінде сөзформаның басы да, соңы (N-дік позицияның басы, N-дік позицияның соңы) да бола алады. Барлық енгізілген сауалдар «Сауалдар тарихы» (Query History) деп аталатын корпус жадында сақталатындығы аса маңызды мәселе, бірақ егер қойылған сауал бұрынғылардың біреуімен сәйкес келсе, онда ол «Сауалдар тарихына» екінші рет жазылмайды. Барлық тарихты қадағалау қажет болса, сауалдар терезесіндегі «төмен» деген нұсқағышты басса жеткілікті, қажеттікке қарай, «Сауалдар тарихының» алдыңғы (жоғарғы) жағындағы ақпаратқа қайта оралуға да болады. Егер «**name:**» терезесіне сауалдың атын енгізсе, онда сауал «аталғандар» (ат қойылғандар) сауалдар тізімінде (named queries) сақталады.

Үлгі (шаблон) – ол біртиптес сауалдарды енгізуді жеңілдететін сауал түрі. Яғни күрделі сауалдарды бір рет қана құрастырып, оны үлгі (шаблон) ретінде сақтап қойып, сосын тек оның мағынасын (мәнін) ғана өзгертіп отырса жеткілікті.

Сауалдарға мысалдар:

Төменде келтірілген мысалдарда Bonito корпустік менеджердің сауалдар тілінің элементтері көрнекі түрде берілді.

1-мысал. Нақты сөзформаны іздеу

Сауал терезесіне «run» сөзформасы енгізіледі.

Шығарылды:

announced that he would not <**run**> for reelection. Georgia

medical benefits paid out would <**run**>1 billion or more in the

May, said today Jones will <**run**> well ahead of his GOP opponents

reports that he had decided to <**run**> and wanted Mr. Screvane,

investigation Street car tracks <**run**> down the center of Pennsylvania

Bonito корпустік менеджер мәтіннен «run» сөзін іздеп, ол сөздің басқа сөзформаларының жоқ екендігін хабарлайды. Осы сияқты Қазақ тілі ұлттық корпусы қанатқақты жобасы бойынша «дала» сөзін іздегенде 400-ден аса мысал табылған.

Іздеу нәтижелері
Табылған құжат саны 464

1. **Әбіш Кекілбаев. Аңыздың ақыры//Проза, Қызыл алма,**
Намаз оқып отырып майдан **даласына** көзі түсіп еді , алдыңғы шеп ауыр пілдер мен қарша бораған қара тасқа қайрат қыла алмай , мысық табандап шегініп келеді екен

2. **Әбіш Кекілбаев. Аңыздың ақыры//Проза, Мұнара,**
Арқасын шуаққа тосқан **дала** тышқандары өн бойларына шымырлап түскен күн нұрына масаттанып , шықылықтап береді

3. **Бейімбет Майлин. Желдібай Жындыбаев//Проза,**
« Көк мойынның « қалғанымен басын жазып **далаға** шыға келгенде , күн сәскелікке келіп қалып еді

4. **Бейімбет Майлин.«Көктеректің» бауырында//Проза,**
Атын жегемін деп **далаға** шықты

2-мысал. Синтагманы іздеу

Сауал терезесіне «run in» енгізіледі.

Шығарылды:

contest . The Orioles got a <**run in**> the first inning when Breeding

record in the 600 - yard **<run in>** the Knights of Columbus track
The Bears added their last **<run in>** the sixth on Alusik ‘s double
for the third Indianapolis **<run in>** the ninth . Despite the 45
‘s first major league home **<run in>** the fifth put the Sox back

Сөзформаны іздеу қатаң түрде алдын ала белгіленген (сызықтық) тәртіппен тұрақты синтагма ретінде жүзеге асады.

Жоғарыда ағылшын тілінен тұрақты синтагмаларды іздеуге қатысты табылған мысалдар берілген. Қазақ тілі корпустарында да іздеуді осылай тұрақты синтагмалар бойынша да жүзеге асыруға болады.

Тұрақты емес синтагманы да («take (smth) out») іздеп-табуға болады.

Сауал терезесіне «take» сөзі енгізіледі. Осы нақты сөзформа үшін конкорданс құрылады. Сауал типі таңдалады: оң сүзгі (P-filter). «From:» және «To:» екі терезеге де «2» мәні енгізіледі. Бұл табылған сөздің (синтагманың «бөліп алынған» бөлігі үшін) оң жақтағы екінші позициясына сәйкес келеді (біздің мысалда «out»).

Шығарылды:

for governor would force it to **<take>** petitions **out** into voting
the peasant. Nonetheless , they **<take>** time **out** -- much time –
Mister McBride. You do that or **<take>** you **out** a permit right now

Әрине, мұндай сауалдардың оң сүзгіні бірнеше рет қолдануға болатын күрделі варианттарын да ойлап табуға болар еді.

3-мысал. Сөздің әртүрлі формаларын іздеу

Сауалдар терезесіне «runs? in»сөзін енгіземіз.

Бұл сауалда «?» *басқару символы* қолданылады. Аталған символдың мағынасы мынада: одан бұрын тұрған «s» әріпі нөл рет немесе бір рет дегенді білдіреді. Алынған нәтиже осының куәсі.

Шығарылады:

tied the game, and single **<runs in>** the eighth and ninth gave
record in the 600 - yard **<runs in>** the Knights of Columbus track
their eight hits for two **<runs in>** the sixth. Chuck Hinton
The Bears added their last **<runs in>** the sixth on Alusik ‘s double
‘s first major league home **<runs in>** the fifth put the Sox back

Қазақ тілі корпусы бойынша да түбір сөздерді ғана емес, оның түрлену формаларын да іздеп табу мүмкіндігі қарастырылған.

4-мысал. Сөздің әртүрлі формаларын бірден іздеуге де болады.

Сауалдар терезесіне «run(|s|ning)» енгізіледі.

Мұнда топтастырушы жақша және альтернатив операторы (|) (логикалық «немесе»). Сауалдар жүйесіне нақты сөзформаларды «run» немесе «runs» немесе «running» тап деген команда беріледі.

Шығарылады:

announced that he would not **<run>** for reelection . Georgia
medical benefits paid out would **<run>** 1 billion or more in the
the group are interested in **<running>** on the required non –
lawyer and former FBI man is **<running>** against the Republican
tied the game , and single **<runs>** in the eighth and ninth gave

Бұл үлгі бойынша қазақ тілі корпустарының сауалдар терезесіне «бала (сы/ға/дан)» сөзформасын берсек, «баласы», «балаға», «баладан» сөзформаларының әрқайсысына мысалдар конкордансы шығады.

5-мысал. Сөзформаларын морфологиялық белгілері арқылы іздеу

Сауалдар терезесіне «[tag=»VVZv»]» енгізіледі.

Шығарылады:

charge of the election, **<deserves>** « the praise and thanks of the
However , the jury said it **<believes>**» these two offices should be
of Fulton County , which **<receives>** none of this money «. The
when the new management **<takes>** charge Jan. 1 the airport be
face is a state law which **<says>** that before making a first

Бұл мысал корпустық менеджердің сөзформалардың белгілері бойынша іздеудің керемет мүмкіндігін ұсынады. «VVZv» түрдегі код (VV) мағыналы етістіктің (Zv) үшінші жақтағы жекеше формасы екендігін айқындайды. Мұндай кодтау SUSANNE аннотациялау схемасы арқылы ұсынылған. Яғни, мұндай мүмкіндікті, ең алдымен, аннотациялау схемасын кім жақсы меңгерсе, сол адам аса тиімді қолдана алады. Қазақ тілі корпусы бойынша да морфологиялық белгіленімдер бойынша іздеуді енгізуге болады. Мұндайда компьютер бағдарламасына енгізілген морфологиялық белгіленімдер кестесі мен реестр сөздер тізімі пайдаланылады.

Берілген мәтінге автоматты морфологиялық белгіленімдер қойып алып, бағдарлама, керісінше, сауалдар терезесіне жазылған морфологиялық белгіленім (шартты белгі) бойынша сол белгіленім сипаттайтын сөздер тізімін мысалдарымен бірге шығарып береді.

3.3. Шығарылым интерфейстері

Корпустық менеджерлерде іздестіру нәтижелері (шығаруға арналған), әдетте, конкорданс түрінде көрініс табады.

Шетел сөздерінің түсіндірме сөздігі бойынша ***конкорданс*** дегеніміз – мағынасы мен берілген мәтіндегі барлық қолданыстары көрсетілген сөздер мен ұғымдардың әліпбилі тізбесі.

Төменде А.С. Пушкиннің «Дубровский» атты шығарма мәтінінен алынған «имение» сөзінің конкордансынан үзінді келтірілді.

«Имение» сөзі бойынша жасалған конкорданс KWIC

с в губерниях, где находилось его	**имение.**	Соседи рады были угождать малейшим его
грубиян; я хочу взять у него	**имение,**	как ты про то думаешь? – Ваше
чтобы безо всякого права отнять	**имение.**	Постой однако ж. Это имение
имение. Постой однако ж. Это	**имение**	принадлежало некогда нам, было

Қандай мақсатта және қандай программаның қолданыста болуына қарай, KWIC конкорданс та және аса толық контекст те пайдаға асулары мүмкін.

Қарапайым конкордансер арқылы жеке сөздердің, сөзтіркестердің, тыныс белгілердің және т.б. контекстік қоршауындағы конкорданстарын құрастыруға болады. Ал аса күрделі программалар (корпустық менеджерлер) тек сөздердің ғана емес, сонымен бірге корпустың басқа элементтерінің де қамтитын толық конкорданстарын құрастыруға мүмкіндігі бар. Шынында да, корпустан алуға тиісті көптеген параметарлер бар, олар: *лемма* және *сөздердің морфологиялық сипаттамалары; сөздердің сөйлемдегі* және *белгіленім*

жасалған мәтін құрылымындағы (HTML, XML) орын тәртібі (позициясы); контекстер іріктеліп алынған құжаттардың библиографиялық және типологиялық белгілері (автор, шығарма аты, дереккөзі (источник), баспадан шыққан жылы, мәтін типі және т.б.); статистикалық деректер және бақа да мәліметтер.

Қазақ тілінің ұлттық корпусы бойынша *орман* сөзінің конкордансы былайша көрініс табады:

Іздеу нәтижелері
Табылған құжат саны 45

1. Бейімбет Майлин. Дәмештің хаты//Проза,1927
Бірер адам жанын қиып өліп те кетер ау , бірақ үй орманымен өлу қиын екен

2. Бейімбет Майлин. Сахарада//Проза,1924
деп, Орман таяғы мұның шекесін жарған болатын Ғибадатың былай тұрсын , қызулау кезінде құдайды боқтап жіберемін

3. Бейімбет Майлин. Сахарада//Проза,1924
Содан бері Орман картпен удай өш

4. Бейімбет Майлин. Қысқы ауылда//Проза,
– Әй, Орман, тұрсайшы, кісінің сұлап жатып алғаның не

5. Бейімбет Майлин. Құла жорға//Проза,
Орыс орман қылышын сүйретіп үйге кіріп келгенде, осы қалмайды

6. Ғабит Мүсірепов. Өлімді жеңген ана//Проза,
Арада тау бар, тас бар, өзен бар, теңіз бар, ит мұрны өтпейтін қалың орман бар, олардан қалай өттің

Корпустың лексикалық бірліктерінің статистикалық мәліметтерін ескере отырып, конкорданстық деректерді өңдеу лексемалар арасындағы синтагматикалық байланыс дәрежесін есептеп шығаруға мүмкіндік тудырады.

3.4. Лингвистикалық емес корпустардың корпустық менеджерлері (WWW)

Интернет желісінің ақпараттық толығуы (веб-кеңістік) аса ірі көптілді корпус ретінде қарастырылуы мүмкін. Тілдік талдаудың басты материалы – сөйлеу тілі мәтіндері түрінде жазылып алынған тіл, – Интернетте де үлкен көлемде және әртүрлілік сипатта орын алады және мәшинелік өңдеуге тікелей қолжетімді. Мұндай жағдай тілшілер үшін аса құнды деуге болады, себебі тілдік материалдарды мәшинелік формаға келтіруге уақыт пен материалдық шығын кететіні мәлім.

Интернеттегі мәтіндік массивтер корпус қалыптастыруда да кең қолданылады. Сол сияқты Интернеттегі мәтіндер әртүрлі тілдік талдау бағдарламалары үшін және мәтіндік ақпараттарды өңдеуде (әсіресе, статистикалық және ықтималдықтар әдістеріне сүйену қажет болғанда) кең түрде қолданыс табады.

Сонымен бірге веб-кеңістік тікелей корпус ретінде де қарастырылуы мүмкін. Бұл мәселе, әсіресе, А. Килгариффтың 2001 жылы жасаған баяндамасынан кейін белсенді түрде талқылана бастады [64]. Тілдік материалдардың репрезентативтілігі жағынан еш корпус веб-кеңістік материалдарымен салыстыруға келмейді деуге болады.

Веб-кеңістікті корпус ретінде қолдану кезінде корпустық менеджерлер рөлін сауалдар жүйесі орындайды. Интернетте кітапханалардағы каталогтарды еске салатын (directories, орысша «каталоги-справочники») классификациялау типіндегі жүйелер әрекет етеді. Бұл жүйелердің деректер базасы, қайсыбір мағынада, семантикалық типтегі корпустар деп қарастырылуы мүмкін, бірақ желідегі ақпарат іздеудің негізгі құралы ретінде вербалды типтегі (поисковые машины – search engines) барлық Интернет-кеңістікті индекстейтін ғаламдық ақпараттық сауалдар жүйесі болып саналады. Осыған орай, вербалды жүйелердің аталған индекстерінің қалай құрылатынын білу аса маңызды және ондай ерекшеліктерді тілдік зерттеулер жүргізу материалдары ретінде сауалдар жүйесінің деректер базасына қолдануда ескеру қажет.

Бір-бірінен тек сауалдар тілімен, дизайнмен, сервиспен және басқа да ерекшеліктерімен айырым табатын осындай көптеген жүйелер бар. Вербалды типтегі (ең бірінші, деректер базасының көлемі бойынша) ізденіс жүйесінің ең бастыларына мыналарды жатқызуға болады: Google, Fast Search (AllTheWeb), AltaVista, WiseNut, HotBot, MSN Search, Teoma. Ресейлік жүйелер ішінен ең бастыларға мына үшеуі жатады: Яндекс (Яndex, Yandex), Рамблер (Rambler), Апорт! (Aport).

Кез келген ізденіс жүйесінің құрамындағы келесі үш негізгі бөлікті айырықша бөліп қарастыруға болады:

1. ***Робот*** – Интернетті қарауды (сканирлеуді) қамтамасыз ету және инверттелген файлды (индекстік деректер базаның) актуалды жағдайда ұстауға арналған ішкі жүйе (подсистема). Бұл программалық кешен желідегі ақпараттық қордың болуы мен жағдайы жайлы ақпарат жинаудың негізгі құралы болып табылады.

2. ***Ізденіс деректер базасы*** – бұл басқаша, ***индекс*** деп те аталады – арнайы түрде ұйымдастырылған деректер құрылымы (*ағыл.* index database) және оның құрамында, ең алдымен, индекстелген веб-құжаттардан алынған лексикалық бірліктерден тұратын инверттелген файл, ол бұл бірліктер туралы неше түрлі ақпараттарды өз бойында сақтайды (мысалы, құжаттардағы олардың позициясы), сонымен бірге құжаттар мен сайттар жайында да толық ақпараттар бар.

3. ***Ізденіс жүйесі*** – пайдаланушының сауалын өңдеуді, деректер базасынан ақпаратты іздеуді және табылған нәтижені пайдаланушыға ұсынуды қамтамасыз ететін ізденістің ішкі жүйесі (подсистема поиска). Ізденіс жүйесі пайдаланушымен ***пайдаланушы интерфейсі*** арқылы қатынас жасайды – браузер-программалардың экрандық формалары: сауалдарды қалыптастыру интерфейсі және іздеу нәтижесін қарау интерфейсі.

Индекстік файл (немесе тек индекс) дегеніміз – сауал бойынша іздеген деректерді тез табуға бағытталған өзара байланыста тұратын файлдардың жиыны. Индекстің негізінде әр уақытта инверттелген файл жатады. Ауқымды (массивті) іздеуді ұйымдастырудың ***инверттелген сызбасы*** құжаттарға кіруді (қолжеткізуді) қамтамасыз етуді мазмұндық идентификаторлар арқылы жүзеге

асыруға негізделген. Мұндай сызбаны арнайы түрдегі қосымша инверттелген файлдарды, яғни қолжетім нүктелерін құру мақсатында құжаттардың реттелген массивін өңдеу арқылы алады.

Осындай әрбір қосымша ауқымды (массивті) жазу оған сәйкес келетін мазмұнды теңдестіргішпен теңдестіріледі (дескриптор, тірек сөз, термин, автордың аты, мекеме аты және т.б.) және өздері жазылған барлық құжаттың аттары да осында сақталады (адрестері). Әрбір мазмұнды теңдестіргіш (деректердің ізденіс элементі) инверттелген ауқымда (массивте) құжаттың адресімен бірге қосымша ақпарат та сақталуы мүмкін, оларға жататындар: өрістің аты, берілген құжат құрамында осы элемент кездескен сөйлемнің номері, сөйлемдегі сөздің номері және т.б. Мәтіндегі сөздің қалпын (орнын) сөйлемнің реттік номеріне дейін және сөйлемдегі сөздің номеріне дейін белгілеу құжаттағы сөзаралық қашықтыққа тапсырыс беруге арналған сауалдардың аса икемді тілін құруға мүмкіндік береді. Позициялық сипаттамалар релеванттық коэфициентті есептеуде және құжаттардың шығарылуын саралауда қолданыс табады.

Жоғарыда айтылғандай, ізденіс жүйесінің индекстері (инверттелген файлдар) – ол, шындығына келгенде, мәтіндердің виртуалды конкордансы. Оған қоса айтсақ, ақпараттық ізденіс жүйелеріндегі құжаттардың қысқаша сипаты түріндегі іздеу нәтижелері, әдетте, табылған құжаттардағы іздеген сөз кездесетін контекстердің көрінісі деуге болады. Оның айырмашылығы мынада, әдетте, конкорданстар нақты шығарма немесе шығармалар тобы (мысалы, бір автордың барлық шығармалары) бойынша құрылады да, ал Интернеттің ақпараттық ізденіс жүйесі барлық қолжетімді электрондық құжаттар жиынын индекстейді.

Веб-сайттарды индекстеу кезіндегі ең бір маңызды мәселе – құжат ішінен алынатын қай терминдер құжатқа тіркеліп жазылатынын білу болып табылады. Себебі құжаттардағы терминдердің барлығы бірдей индекс аясына ене бермейді. Мысалы, индекс өрісіне енбейтін рұқсат етілмейтін сөздер тізімі (stop-words) белсенді түрде қолданыс табады – олар, жалпы, қызметтік лексика (көмекші сөздер, жалғаулықтар және т.б.) және мағынасыз сөздер. Көптеген жүйелер құжаттың тек бөлігін ғана (әдетте бас жағын)

индекстейді, сол сияқты, бір ғана сайттың веб-бетінің тек бөлігін ғана өңдейтін роботтар да бар. Роботтардың қалай жұмыс атқаратынын білу, олардың техникалық сипаттамаларымен таныс болу веб-құжаттарды құрушыларға да және іздеу кезіндегі сауалдарды құрастырушыларға да аса пайдалы. Роботтардың жұмыстарын сипаттауды толығырақ мына сайттан қараңыз: http://www.searchengineworld.com/robots/norobots.htm. Ақпараттық ізденіс жүйесінің (АІЖ) шығару интерфейсінен қандай ақпаратты және қандай пішінде шығаруға болатындығын білу де маңызды саналады. Әртүрлі жүйелердегі шығару интерфейсі (нәтиже көрінісінің формасы) мынадай параметрлерден тұрады: сауалдағы сөздің статистикасы, табылған құжаттардың саны, табылған сайттардың саны, бір беттегі іздеу нәтижелеріне қатысты құжаттар саны, шығарудағы құжаттарды сұрыптауды басқару құралы, алынған құжатқа сәйкес келетін сайттың сипаттамасы, құжаттың сипаттамасы. Соңғысы өз құрамында құжаттың атын, URL (желідегі адресі), құжаттың көлемі, құрылған датасы (мезгілі), кодтау, аннотация (қысқаша мазмұны), сауалдағы сөздерді аннотацияда визуалды бөліктеу, сол сайттың басқа да релевантты веб-беттеріне сілтеме жасау, табылған құжат пен сайттың қатысы бар каталог рубрикасына сілтеме жасау, релеванттық коэфициентке, басқа іздеу мүмкіндіктеріне сілтеме жасау (ұқсас құжаттарды іздеу, табылған құжаттан іздестіру).

Аталған барлық реквизиттердің ішінен лингвистикалық зерттеулер үшін ең үлкен қызығушылық танытатын мәселе, ол – тілдік бірліктердің жиіліктік сипаттамалары және олардың контекстері. Жүйелердің ескеретін және шығарылатын жиіліктер ішінен әр сөзге және құжатқа қатысты жиіліктерін ажырата білу қажет болады. Тілдік бірліктердің сандық сипаты жайлы мәлімет әртүрлі жүйелерге және әртүрлі іздеу деңгейіне қарай сөзформаларға да және лексемаларға да қатысты болып келеді. Кейбір жүйелер сауалдарды тіркеп отыруымен бірге, қайталануын да ескеріп, сауал бойынша олардың статистиксын да шығарып бере алады. Мүмкіндіктердің тиімділігі мен қызығушылық тудыратын жақтары – құжаттарды тақырыптық кластарға бөліп қарастыру болып табылады [65].

3.5. Корпустық зерттеулер

3.5.1. Корпустарды пайдаланушылар

Корпустарды пайдаланушыларды, ең алдымен, тілшілерді, әдетте, нақты мәтіндердің мазмұнынан гөрі, олардың метамәтіндік ақпараты және қайсыбір тілдік элементтері мен олардың құрылымдық қолданыстарының мысалдары көбірек қызықтырады. Корпустар арқылы жүргізілген ең алғашқы тілдік зерттеулер әртүрлі тілдік элементтердің мәтіндегі қолдану жиіліктерін анықтаумен ғана шектелетін. Статистикалық әдістер машиналық аударма, сөйлеу тілін танып білу (разпознавание) мен оны синтездеу, орфография мен грамматиканы тексеретін құралдар және т.б. осы сияқты күрделі лингвистикалық мәселелердің шешімін табуда қолданылады. Мәселен, корпус материалында статистикалық әдістерді қолданып, қай сөздер әрдайым бірге қолданатынын білуге болады, бұл оларды тұрақты сөз тіркестеріне жатқызуға болатындығының айғағы деуге болады. Семантикалық жағынан алғанда, тұрақты сөз тіркестері тұтас сипаттағы (бөлінбейтін) мағыналық бірлік, ал мұндай жағдай лексикография саласы мен мәтінді автоматты өңдеу жүйелерінде ескерілуі өте маңызды.

Сонымен қатар лексикография мен грамматиканы зерттеуде корпустар аса бай дереккөз болып табылады. Лексикографиялық зерттеулермен 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 қажетті құрал ретінде қолданады. Қолданбалы бағытта жұмыс істейтін лингвистер (мұғалімдер, аудармашылар және т.б.) компьютерлік корпустарды тілдерді үйретуге және өздерінің кәсіби мәселелерін (есептерін) шешуге пайдаланады. Корпустарды пайдаланушылардың айрықша тобын компьютерлік лингвистер құрайды: олар компьютерлік тілдік модельдерді құрастыру үшін мәтінде орын алатын статистикалық

және лингвистикалық заңдылықтарды айқындау мен пайдалануды мақсат етеді. Тілге қатысты басқа да мамандар (әдебиетшілер, редакторлар) корпус арқылы өздерін қызықтыратын сұрақтар бойынша қанағаттандырарлық жауап ала алады. Қоғамдық ғылымдар салаларының мамандары (тарихшылар, социологтар), өздерінің зерттеу нысандарын тіл арқылы, яғни кезең, автор немесе жанр деп аталатын мәтін параметрлерін қолдану арқылы зерттеулеріне мүмкіндігі бар. Әдебиетшілер корпусты стильдердің ерекшеліктерін зерттейтін (стилеметрический) ғылыми зерттеулер үшін пайдаланады. Корпустар әртүрлі автоматтанған жүйелерді (машиналық аударма, сөйлеу тілін тану, ақпараттық іздеу) зерттеу үшін пайдаланылады.

Корпустық деректер теориялық лингвистикаға қандай септігін тигізеді деген сұраққа келсек, әрине, олар ғалымның өзінөзі бақылауын (интроспекцияны) және тілшілердің лексика мен грамматика жайындағы пайымдауын да алмастыра алмайды, бірақ олар мамандарға бай репрезентативтік эмпирикалық материал береді. Түптеп келгенде, корпустар тілдік зерттеулер барысында пайдалануға болатын үш типті деректер бере алады: эмпирикалық қолдау, жиілігі бойынша ақпарат, экстралингвистикалық ақпарат (метаақпарат). Осы аталған деректер типтерін мүмкіндігінше толығырақ қарастырайық.

3.5.2. Корпустарды пайдалану тәсілдері

1. Эмпирикалық қолдау. Көптеген тілшілер корпусты «мысалдар банкі» ретінде пайдаланады, яғни корпустан өз зерттеулеріндегі болжамдарына, қағидаттары мен ережелеріне эмпирикалық қолдау табуға ұмтылады. Табылған мысалдар, әрине, ойдан шығарылуы немесе кездейсоқ табылуы мүмкін, бірақ корпустық лингвистика тәсілі тілдік материалдың репрезентаттығы мен теңгерімділігін, сондай-ақ қандай да бір корпусты пайдаланушының таңдау мүмкіндігін туғызатын іздеу құралын қамтамасыз етеді.

Сондай-ақ кейбір ғылыми гипотезалардың керағар екенін дәлелдейтін материалдар алуға мүмкіндік береді, яғни ғылыми теориялардың шынайылық сипатын дәйектейді.

Орыс тілінің ұлттық корпусын пайдалана отырып, ғалым Н.В. Перцов орыс тілінің беделді тілшілерінің ой-пікірлерін былайша теріске шығарады: «...Следует признать, что возможности корпусов все-таки еще недостаточно усвоены лингвистической общественностью вообще и лингвистами в частности. Обращение к корпусным данным еще не стало столь же привычным и обязательным при формулировке и проверке тех или иных утверждений относительно фактов языка, как обращение к грамматикам и словарям, к работам коллег» [66, 318 б.]. Тіпті корпустарды кең қолданып жүрген авторлардың жұмыстарының өзінде тілдік деректер жайындағы тұжырымдар корпустық деректерге қайшы келіп жатады.

Әрбір тілдік деңгейдегі, сөз дыбыстарынан бастап, тұтас әңгіме мен мәтінге дейінгі жасалған болжамдарды анықтайтын дәлелдер корпустардан табылуы мүмкін. Өзін-өзі бақылау кезінде мүмкін болмайтын жағдайларды корпус материалдары бойынша құрылым ішінен қайта талдап, нәтижелерді қайта жаңғыртуға болады [1].

2. Жиілігі бойынша ақпарат. Корпусты қолданудың сапалы әдісі эмпирикалық қолдау түрінде көрініс табады да және сонымен бірге корпустар сандық зерттеулер жүргізу үшін сөздердің, фразалар мен сөз тіркестердің қолдану жиілігі туралы ақпарат бере алады. Сандық зерттеулер (әрине олар көбінде сапалық талдауға негізделеді) теориялық және компьютерлік лингвистиканың көптеген аясында қолданыс табады. Олар әртүрлі сөйлеушілер топтарының немесе әртүрлі мәтіндер типтерінің арасындағы ұқсастық пен айырмашылықтарды көрсетеді, психолингвистикалық және тағы да басқа зерттеулер үшін деректердің қолдану жиілігін айқындауға мүмкіндік береді.

3. Метаақпарат. Тілдік контекстке қосымша ретінде, мәтіндер корпусы сөйлеушінің не жазушының жасы немесе жынысы, мәтін жанры, мәтіннің пайда болуы жайындағы уақыттық немесе кеңістіктік және т.б. туралы экстралингвистикалық ақпарат немесе метаақпарат береді. Мұндай метаақпарат әртүрлі мәтіндер типтері мен әртүрлі сөйлеушілер топтарын салыстыруға мүмкіндік жасайды.

Көптеген ғалымдардың пайымдауынша, корпустық лингвистика – лингвистиканың жеке парадигмасы дегеннен гөрі, оның

әдіснамасы (әдістемесі) деген дұрыстау келеді. Мәселен, ағылшын тілінің көптеген белгілі корпустары тіл ғылымының әртүрлі бағыттарының өкілдері жүргізетін арнайы зерттеулер үшін құрастырылған және қолданылған. Мысалы, *CHILDES* корпусы әртүрлі коммуникативтік жағдаяттағы балалардың ауызша тілі транскрипті арқылы балалардың тілді игеру қабілетіне қызығушылық білдіретін ғалымдардың психолингвистика саласындағы зерттеулерінде кең қолданыс табады [67]. Ағылшын тілінің алғашқы кезеңдерінен бастау алатын жазба мәтіндерінің әртүрлі типтерінің *Хельсинкалық корпусы* тіл тарихының даму үдерісін зерттеу үшін қолданылады [68]. Лондондық жасөспірімдердің *ағылшын тілінің Бергендік корпусы* COLT (The Bergen Corpus of London Teenage Language) 13-17 жастағы жасөспірімдердің сөйлеу тілі мәтіндерін жинақтаған және ол әлеуметтік лингвистика саласы бойынша белгілі жастар топтарының тілдерін зерттеу үшін пайдаланылады [69]. Өз зерттеулерін сапалы нәтижелер алуға мүмкіндік беретін «нақты» тілдік материалдар бойынша жүргізілетін тілдік талдаулардың ұтымдылығы корпустарды пайдаланатын тілшілердің қызығушылығын арттырады [70].

3.5.3. Корпустарға негізделген лексикографиялық зерттеулер

Лексикографиялық зерттеулер, біріншіден, сөздіктер құрастыру үшін қажет болса, екіншіден, дискриптіктік және қолданбалы лингвистика үшін де қажет. Мәселен, зерттеу жүргізудің алдында лексикографтардың қандай ақпаратқа мұқтаж екендігін біліп алу қажет. Мысалы, орыс тілінің академиялық түсіндірме сөздігін құрастырудағы авторлардың негізгі сұранысы мынадай ақпараттарды тауып алу қажеттігін тудырған:

– белгілі бір кезеңдерде пайда болған жаңа сөз;
– сөздің бастапқы формасы;
– мағыналары белгілі сөздерді ашатын мысалдар (цитаталар);
– сөздікте мысал келтірілмеген мағыналарға мысал табу;
– қайсыбір мағынаға қосымша берілетін жаңа мысалдар;
– лексикалық және синаксистік тіркесімдердің жаңа түрлері;
– жаңа фразеологизмдер;

– арнайы терминдердің қазіргі кезге сай жаңа ғылыми түсініктемесі [9].

Грамматикалық және лексикографиялық модельдер жүйелі түрде өзара байланыста болады. Мысалы, дәстүрлі әдістер синонимдік сөздер тобын анықтай алатын болса, лексикографиялық зерттеулер корпустық тәсілдерді қолдану арқылы өзара қатынастағы сөздердің әртүрлі жағдаятта және әртүрлі контекстерде қалай қолдануға болатындығын көрсету мүмкіндігі бар.

Д. Байбера, С. Конрад, Р. Реппендердің «*Corpus Linguistics*. Investigating language structure and use» атты оқулығында [40] (әрі қарай – *Corpus Linguistics*) корпустық тәсіл арқылы жұмыс істейтін зерттеуші-лексикографтардың алдарында тұрған негізгі алты мәселені жеке қарастырады:

1. Нақты сөз бойынша қандай мағыналар ассоциацияланады?

2. Белгілі бір сөздің қолдану жиілігі оған жақын тұрған басқа сөздердің жиілігіне қарағанда қандай деуге болады?

3. Берілген сөздің қандай тілдік емес модельдері бар (регистрге, тарихи кезеңдерге, диалектілерге және т.б. қатысты)?

4. Берілген сөзбен қандай сөздер әдетте бірге кездеседі және әртүрлі регистрлерде олардың тіркесіп қолдану реттілігінің үлестірімі қаншалықты?

5. Сөз қолданысының мағынасы мен типтері қалай үлестірілген?

6. Синоним сөздер қалай пайдаланылады және қалайша әртүрлі контекстерде өз орнында қолданылады? [71]

Лексикографиядағы корпустық тәсілмен зерттеудің ұтымды жағы – корпустан берілген сөз қолданылатын контекстер жиынын кең түрде алуға болатындығы. Әрі қарай конкорданстық тізімдерді (KWIC) қолдана отырып, контекстерден сол сөзді ассоциациялайтын әртүрлі мағыналарды бөліп алуға болады.

Корпустарды құрастырудың тағы бір мақсаты – зерттеушінің өз зерттеу аясындағы еңбегін жеңілдету. Мысалы, корпус тек күрделі зерттеу саласы мәтіндерінің жиынтық бөлігі ғана емес, сонымен бірге ол, мүмкіндігінше, көлем жағынан да айтарлықтай айырым табуы қажет.

Көлемді корпустардың конкорданстары да өте үлкен болады. Соған сәйкес өте көп деректер бере алады. Конкорданстар көлемі

үлкен корпустарда бірнеше мың бетке жетуі мүмкін, тіпті бір мысалға жүздеген конкорданстар сәйкес келуі мүмкін [8]. Мысалы, көлемі 8 миллиондық Лонгман-Ланкастер подкорпусында (ішкі корпусында) *deal (дело)* сөзіне 1500-ден аса қолданысты (мысалды) ұсынған, ал мұндай көп мәлімет оларды мағынасына қарай топтауды немесе маңыздылығына қарай сұрыптауды қиындатады. Мұндай жағдайда қосымша құралдар мүмкіндігі пайдаланылған. Айталық, программалық конкорданстардың көпшілігі сөздердің қолдану жиілігі бойынша сөздер тізімін жасай алады. Ондай тізімдер әдетте әліппи тәртібімен, кездесу тәртібімен немесе жиіліктерінің кему не өсу тәртібімен жасалады. Одан басқа көпшілік қауымға белгілі **Sketch Engine** жүйесі құрылымдық-синтаксистік модель бойынша реттелген шектеулі сөзтіркестерінің статистикалық жиынын (коллокацияларды) тауып бере алады. Енді *Corpus Linguistics* оқулығы бойынша *deal* сөзімен мағынасы ұқсас сөздерді табуға мысалдар келтірейік.

Сөздердің мағынасын талдауға қиындық тудыратын жайт, ол – ағылшын тілінде көптеген сөзформалардың грамматикалық қызметтерінің көп болуы. Мәселен, *deals* сөзформасының 3-ші жақтағы жекеше формадағы етістік ретінде қолданылуы және көптік формадағы зат есім түрінде де қолданылуы. *Deal* және *dealing* сөзформалары етістік ретінде де және зат есім түрінде де қолданыла алады. Аннотация жасалмаған корпус деректері бойынша құрастырылған жиілік сөздіктердің тиімділігі шектеулі болуының себебі – олардың қай сөздердің грамматикалық қолданысының жиі, ал қай қолданысының сирек екендігін ажырата бермеуінде.

Deal сөзформасының зат есім ретінде неше рет кездесетінін және неше рет етістік ретінде қолданылатынын анықтау үшін ең алдымен олардың контекстегі формасына қарай, сөзформалардың грамматикалық категорияларын анықтап алу қажет, содан кейін барып әрбіреуінің қолдану жиілігін есептеп шығаруға болады. Мұндай шешім *deal* сөзформасының LOB корпусындағы 182 рет қолданысы бойынша көп уақытты қажет етеді, сол сияқты көп таралған 1 млн. мәтін ішінде 500 ретке жуық қолданған *look* сөзформасы қолданысының грамматикалық категориясын

анықтау да қиындық тудыратыны мәлім. Мұндай жағдайдағы ең дұрыс шешім – аннотацияланған корпусты пайдалану. Ондай корпуста әрбір сөзформаға грамматикалық категориялары бойынша шартты түрдегі арнайы белгіленім жүргізілгендіктен, оларды автоматты түрде ажыратып, жиіліктерін есептеуге мүмкіндік жасалады. Мұндай әр сөз табына жататын омоним сөздердің екі түрлі позицияда қолданысын қолдан санап шығу қиын. Мысалы, қазақ тіліндегі «қара» сөзінің белгілі бір көлемдегі мәтінде неше рет етістік, неше рет сын есім мәнінде қолданылғанын анықтау үшін мәтінге алдымен автоматты морфологиялық белгіленім қою бағдарламасы іске қосылады. Табылған омонимдер қолдан ажыратылады немесе белгілі бір контекстік қоршауларға негізделіп, омоним ажыратудың формалдық межесі анықталады. Содан кейін барып жиілікті анықтауға мүмкіндік туады.

23-кесте

Аннотацияланған корпус бойынша есептелген
DEAL формасының жиілігі

Ланкастер-Осло-Берген (LOB)

deal_nn	115
Deal_np	1
deal_nns	5
deal_vb	66
dealing_vbg	51
deals_vbz	20
dealt_vbd	14
dealt_vbn	17

23-кестеде TACT (Text-Analysis Computing Tools) программасы арқылы алынған жиіліктік тізімі берілген. Ол тізімде DEAL сөзінің Ланкастер-Осло-Берген аннотацияланған корпусындағы грамматикалық формаларының үлестірімі көрсетілген. Әр сөздің грамматикалық категориясы «астын сызу» символынан кейін жазылған. Мысалы, *deal* сөзі жекеше формадағы зат есім ретінде

(граммема nn) 115 рет кездессе, етістік ретінде (граммема vb) – 66 рет, жалқы есім ретінде – 1 рет және т.б.

Қазақ тілінде де осы сияқты аннотацияланған корпус мәтіндері бойынша әртүрлі позицияда тұрғандағы грамматикалық сипатын, қолданылу жиілігін анықтауға болады.

Сонымен қатар әрбір сөздің түрленім формаларының қайсысының жиі, қайсысының сирек кездесетінін де анықтауға мүмкіндік бар. Мысалы ағылшын тілінде DEAL лексеманың Ланкастер-Осло-Берген аннотацияланған корпусындағы зат есім ретінде жеке және көптік формаларында (*deal* және *deals*) қолданылуын қарастырайық. Ланкастер-Осло-Берген корпусы әртүрлі регистрлерге қатысты мәтіндерден құрастырылғандықтан (ғылыми әдебиет, көркем проза, қызық оқиғалы әдебиет және ковбойлық романдар), *deal* және *deals* сөздердің әртүрлі регистрдегі қолдану жиіліктерін салыстыра зерттеуге мүмкіндік туады (*24-кесте*).

24-кесте

DEAL зат есім сөздің әртүрлі регистрлердегі жиілігі
(100 мың сөзге нормаланған)

Регистр	Подкорпустағы шамамен алғандағы сөз саны	DEAL сөзі бойынша тікелей (өңделмеген) есептеулер	DEAL сөзіне қатысты нормаланған есептеулер (100 мың сөзге)
Баспасөздегі репортаждар	88000	14	15,9
Баспасөз шолулары	34000	4	11,8
Бас мақала	54000	4	7,4
Діни әдебиет	34000	5	14,7
Ғылыми әдебиет	160000	16	10,0
Ғылыми-көпшілікке арналған әдебиет	88000	11	12,5
Беллетристика	154000	24	15,6
Көркем әдебиет	58000	5	8,6

24-кестеде тікелей (өңделмеген) есептеулер (raw counts) мен нормаланған есептеулер (normed counts) берілген. ***Тікелей*** есептеулер бойынша әр реестрдегі сөздердің кездесу жағдайының саны беріледі: мысалы, ***deal*** сөзінің зат есім ретіндегі формасы баспасөз репортажында – 14 рет кездескені анықталған, баспасөз шолуларында – 4, ғылыми әдебиетте – 16 және т.б.

Бірақ ол подкорпустардың көлемі бірдей емес, ондағы сөздер саны әртүрлі: баспасөз репортажында – 88 мың сөз, баспасөз шолуларында – 34 мың сөз, ал ғылыми әдебиетте – 160 мың сөз. Осы себептен әр регистрдегі тікелей есептеулер нәтижелерін салыстыра зерттегенде, оларды аз не көп жиілікті деген критерий бойынша қарастыруға болмайды. Егер бір регистр бойынша алғашқы мәтінге қарағанда соңғы алған мәтін көлемі көп болса, ізделіп отырған сөздің басқа да қолданыстары ашылуы әбден мүмкін. ***DEAL*** сөзіне қатысты баспасөз репортажы бойынша жүргізілген тікелей есептеулер нәтижелері ғылыми әдебиет бойынша жүргізілген есептеу нәтижелерімен салыстыра қарағанда төмен болады деп алдын ала күтуге болады. Бұдан ***DEAL*** сөзі баспасөз репортажында сирек кездеседі, ғылыми әдебиетте жиі кездеседі деп қорытынды жасауға болмайды.

Дұрысында, оқырманның ***DEAL*** сөзінің зат есім ретінде жиі кездестіру мүмкіндігін зерттеу (есептеу) қажет.

Мұндай есептеулер туралы «нормаланған» жиілік деп атайды. ***Нормаланған*** есептеулер сөздің кездесу санын стандартталған меже бойынша өзге түрге ауыстырады (түрлендіреді), әдетте 1 млн. сөзқолданысқа шаққанда немесе бұл жағдайда 100 мың сөзқолданыстық көлем алынып тұр. Егер есептеулерді нормалайтын болсақ, баспасөз репортажында 100 мың сөзқолданысқа шаққанда – 15,9 рет кездеседі, ал ғылыми әдебиетте 100 мың сөзқолданысқа шаққанда, бар болғаны – 10 рет қана қолданады екен және т.б. Яғни, тек нормаланған есептеулер ғана регистрлер бойынша салыстыру үшін дәйекті (сенімді) негіз болады.

25-кесте

Лонгман-Ланкастер корпусы бойынша екі подкорпусындағы deal сөзінің зат есім ретінде және етістік ретінде қолдану жиілілігі

	Подкорпустағы шамамен алғандағы сөздер саны	1 млн. сөзге (IPM) шаққандағы нормаланған есептеулер	
		Зат есім	Етістік
Барлығы	4000000	90	119
Көркем проза	2000000	107	63
Ғылыми әдебиет	2000000	74	176

DEAL сөзінің зат есім ретіндегі кездесу жиіліктері регистр бойынша үлестірілген жағдайда, лексикографиялық жұмыстар үшін корпус көлемі мәселесі бұрынғыдан да айқын бола бастайды. 24-кестедегі мәліметтер бойынша 8 субкорпустың ішінен тек 4 субкорпуста ғана 4-5 реттен қолдану жағдайы байқалады. Регистрлердің барлығында да DEAL сөзінің айтарлықтай жиі қолдануы байқалмайды, ең көбі – 24 рет (көркем прозада). Әрине, DEAL сөзін зат есім ретіндегі қолданысын талдауға Ланкастер-Осло-Берген корпусының көлемі аздық ететіні түсінікті, сондықтан әрі қарай ол сөздің регистр бойынша үлестіру моделін көлем жағынан қомақты саналатын Лонгман-Ланкастер корпусы бойынша қарастырайық.

25-кестедегі деректер бойынша Лонгман-Ланкастер корпусында басқа базаларға қарағанда ***deal*** мен ***deals*** айтарлықтай жиі кездеседі және бұл қомақты база бұл сөздердің қолданысын талдауға мүмкіндік береді. Аталған кестедегі деректер бойынша, жиілік көрсеткіштері қызықты модельдер табуға негіз болады. Біріншіден, мәтіндердің барлық мысалдары үшін нормалданған есептеулер (толық қамтитын есептеулер) бойынша, етістік ретіндегі ***deal/deals*** зат есім ретіндегі осы сөздердің қолдану жиілігінен тек аз ғана айырым табады (етістік ретінде – 1 млн.-ға шаққанда 119 сөз, ал зат есім ретінде – 1 млн.-ға 90 сөз). Егер ол сөздердің кездесу жиілігін регистр бойынша қарастыратын болсақ, басқа

көріністі байқаймыз. Мысалы, ғылыми әдебиет мәтіндерінде ***deal/deals*** сөздер етістік ретінде қолдануы олардың зат есім ретінде қолдануымен салыстырғанда екі рет жиі қолданылатынына көзімізді жеткізуімізге болады (етістік – 176, ал зат есім – 74 рет 1 млн. сөзге шаққанда). Ал көркем проза мәтіндеріндегі ***deal/deals*** сөздерінің қолданысы жоғарыда айтылғандарға қарағанда, қарама-қарсы модельдің көрінісін бейнелейді, яғни ол сөздердің зат есім ретіндегі қолданысы етістік ретіндегі қолданысына қарағанда анағұрлым жиі (зат есім – 107, ал етістік – 63 рет 1 млн. сөзге шаққанда).

DEAL сөзінің қолдану моделі корпусты құрастыруға қатысты басқа да бір маңызды жағдайды сипаттайды: тек бір ғана регистрмен шектелген корпус тілді басқа регистрлерде бейнелей алмайды. Мысалы, бір регистрдегі материалдар бойынша құрылған модель негізінде басқа регистрден табылған сәйкес материалдар үшін жалпылама қорытынды жасауға болмайды. ***DEAL*** сөзінің қолдануына қатысты мысалдар негізінде мынадай қорытынды жасауға болады: ***deal*** сөзінің зат есім және етістік ретінде ғылыми әдебиетте қолдануының қатынастық жиіліктері олардың көркем прозадағы қолдануының қатынастық жиілігіне қарағанда толығымен қарама-қарсы сипатта деуге болады. Кез келген осы регистрлермен шектелген корпус, басқа регистр бойынша табылған мәліметтерді тіпті де көрсете алмас еді және ол сөздің тілдік қолданысының моделі дұрыс құрылмаған болар еді.

Жоғарыда регистрлер деп әртүрлі стильдер мен жанрлар алынып отыр. Жоғарыдағыдай әрбір сөздің жекелеген стильдер мен жанрлардағы қолданыс жиілігін анықтауға болады. Ол үшін әртүрлі стилдер мен жанрлар мәтіндерінің жеке субкорпустары жасалуы қажет немесе белгілі бір сөзқолданысқа шаққанда әртүрлі жанр мәтіндерін жиілік бойынша салыстыруға болады.

Регистр бойынша мағыналардың (мәндерінің) үлестірімі. Корпустар конкорданстарды пайдалана отырып, сөздердің мағынасын ашуға байланысты зерттеулер жүргізуге мүмкіндік тудырады. Сөздер мағынасын ашуға байланысты зерттеуді олардың ***коллокаттарын*** (collocates) талдаудан бастауға болады, яғни талдауға тиісті сөздің жиі тіркесетін сөзінен бастайды. Әрбір коллокация (collocation) үшін бір мәнмен немесе мағынамен ассоциациялау-

дың күшті үрдісі орын алады (бірнеше сөзтіркестері бір мағынамен ассоциациялануы мүмкін). Сондықтан, сөздің ең жиі коллокациясын бөліп алып, сөз мағыналарына тиімді және дәлелді (сенімді) талдау жүргізуге болады. Әрі қарай корпустағы зат есім ретіндегі DEAL сөзінің, оның сөздіктегі дефиницияларымен (тұжырымды анықтамасымен) бірге, коллакаттарын талдау нәтижелерімен салыстыру қажет болады.

Лонгман-Ланкастер корпусындағы екі регистр ішінде кездесетін зат есім ретіндегі ***deal*** сөзінің коллокаттары *26-кестеде* берілді. Жиілігі бойынша сұрыпталған тіркесім тізімін бейнелейтін мұндай кестелерді әртүрлі конкорданстар программалары көмегімен алуға болады.

26-кесте

Лонгман-Ланкастер корпусының (5,7 млн. сөз) екі подкорпусындағы зат есім ретіндегі deal сөзінің жиі кездесетін коллокаттары

Зат есім ретіндегі *deal* сөзінің жиі кездесетін коллокаттары	1 млн. сөзге шаққандағы нормаланған есептеулер
Ғылыми әдебиет *(2,7 млн. сөзден тұратын подкорпус)*	
Сол жақты коллокаттар	
Great	45
Good	23
Оң жақты коллокаттар	
of	39
more	7
in	3
to	3
Көркем проза *(3 млн. сөзден тұратын подкорпус)*	
Сол жақты коллокаттар	
great	40
good	28
the	8

big	3
Оң жақты коллокаттар	
of	28
to	7
about	5
more	3
with	3

Сол жақты коллокаттар – тікелей зат есім ретіндегі ***deal*** сөзінен бұрын кездесетін сөздер. Мысалы, *26-кестедегі* деректер бойынша ***good*** сөзі ***deal*** сөзі үшін жиі кездесетін сол жақты коллокат болып саналады. Оң жақты коллокаттар – тікелей зат есім ***deal*** сөзінен кейін кездесетін сөздер. Аталған кестедегі тізімдер – ол ***deal*** сөзінің оң жағында және сол жағында тұратын бірінші сөздер, бірақ сол технологиялар бойынша одан да көп қашықтықтағы тіркестерді зерттеуге болады (мысалы, екі немесе үш сөз қашықтықтары). Кестеден байқағандай, ғылыми әдебиетте ең жиі кездесетін зат есім ретіндегі ***deal*** сөзінің сол жақты коллокаты ***great*** сын есім сөзі (1 млн. сөзге 45 рет кездеседі), одан кейін тұратын ***good*** сын есім сөзі (1 млн. сөзге 23 рет кездеседі). Одан кейінгі реттегі коллокаттар (***package*** және ***that***) екі реттен кездескендіктен, *26-кестеге* енгізілмеген (кестеде әр реестрге 1 млн. сөзге ең аз дегенде 3 рет кездескен коллокаттар қамтылған).

Ғылыми әдебиетте, әлбетте, ***good deal*** және ***great deal*** коллокациялары үлкен мөлшердегі нәрсені немесе бизнестегі қайсыбір операцияны білдіреді. Оң жақты коллокаттарды қарастырған кездерде ***deal*** сөзінің бір нәрсенің санына қатысты мағыналары ең көп кездесетіні анықталды. Мұны ***of*** сөзінің оң жақты коллокатының жиілігі де растайды (1 млн. сөзге 39 рет). Келесі коллокаттың жиілігі анағұрлым кіші: ***more*** (7 рет 1 млн. сөзге), ***in*** және ***to*** (3 рет 1 млн. сөзге). Сонымен, зат есім ретіндегі ***deal*** сөзі ***a good/great deal of*** сөз тіркесіндегідей сандық мағынада ең жиі қолданылады.

Сөздердің мән-мағынасын сол сөздің сөз тіркестерінде орын алатын конкорданстар тізіміне шолу жасау арқылы анықтауға болады.

26-кестеден көркем прозадағы сөз тіркестерінің ғылыми әдебиеттердегі сөз тіркестерімен ұқсас және айырым тұстарының бар екендігіне көз жеткізуге болады. Тағы да ең жиі қолданастағы сол жақты ***great*** және ***good*** коллокаттары екенін айта кетейік. Шындығында да, ***good deal*** және ***great deal*** тіркестерінің бірге кездесулері екі регистрде де бір-біріне өте ұқсас (шамамен 1млн. сөзге 68 рет). Бірақ көркем әдебиетте «good/great + deal» моделіне қатысы жоқ аса көп қолданыс жағдайы кездеседі (96 мысал). Мұнда коллокаттар аса кең түрде үлестірілген – ***the*** 1 млн. сөзге 8 рет кездескен және ***big*** сөзі 1 млн. сөзге 3 рет, ал басқа 37 коллокат 2-ден жиі кездеспеген.

Сөз тіркесетерінің моделі бойынша сандық (мөлшерлік) мағына көркем проза үшін жалғыз мәселе болмағанмен, негізгі мәселелердің бірі болып саналады. Мысалы, сол жақты ***the*** коллокаты ***DEAL*** сөзімен тіркесіп, келісу мағынасын беретін мына мысалдардағыдай, part of the deal is... және Isn't that the deal? қолданылады. Ал ***big*** коллокаты басқа мағынада қолданылады, мысалы, ол сөз мына айтылымда «***no big deal и what's the big deal?»*** «маңызды емес» дегенді білдіреді.

Одан басқа, көптеген коллокаттар, жеке алғанда жиі кездеспейтін, бизнестегі операциялармен ассоциялданады: property deal, record deal, cash deal, land deal, mining deal. Дегенмен бұл сөздердің ***DEAL*** сөзінің негізгі бес коллокаттарына қатысы жоқ, ал бірге алғанда олар маңызды мағынаны білдіреді.

Көркем прозадағы ***deal*** сөзімен тіркесіп қолданылған сөз тіркестері ғылыми әдебиеттен табылмаған мағынаны ашады. Оң жақты коллокаттардың тізімінде ағаш түрі жайлы сөз болған ***table*** сөзінің кездесуінің 4 жағдайы және ***box*** сөзінің 1 жағдайы берілген. Ол коллокаттар жиі кездеспейді, бірақ олар ***deal*** сөзін көркем прозада қолданылуының тағы бір жағдайын көрсетеді, яғни олардың әртүрлі сипаттағы 5 мәтінде кездесуі біршама кең тарағанын білдіреді.

Corpus Linguistics оқулығында корпустық әдіс арқылы алынған жиіліктік сипаты зат есім ретіндегі ***deal*** сөзінің сөздік анықтамаларымен салыстырылады. Сөздіктерге жасалған шолу бойынша мағыналардың әртүрлілігіне көз жеткізуге болады. Кейбір сөздік-

тер тек бір ғана бас мақаланы көрсетеді де, ал басқалары – төртеуін бір-ақ көрсетеді. Сөздік мақалаларында анықтамалар саны 2-3-тен 20-30-ға дейін құбылады. Мұндай әртүрлілік қиындық тудыратыны белгілі, себебі пайдаланушың ***deal*** сөзінің ең жиі қолданыстағы мағынасын аңғарып алу әрекеті айтарлықтай күрделілікке соқтыруы мүмкін.

27-кесте бойынша зат есім ретіндегі ***deal*** сөзінің 7 мағынасы берілген, ал олар бес сөздікте жиі қайталанады. Сөздіктердің көпшілігінде барлық 7 мағына берілген, бірақ олардың орналасу реті әртүрлі. Мысалы, «большое, но неопределенное количество» мағынасы *Webster's Third Dictionary* – да бірінші мына сөздік мақаласына 2-дефиницияда енген және *Random House Dictionary* – да 21-дефиницияда енген. Осы сөздіктік анықтамаларды корпус көмегімен зат есім ретіндегі ***deal*** сөзін зерттеу нәтижелерімен салыстыра отырып, бірнеше мәселені бөліп алуға болады. Зат есім ретіндегі ***deal*** сөзінің сандық шама мағынасында қолдануы, талдап отырған екі подкорпус регистрлерінде де ең жиі қолданыста екендігі даусыз. Дегенмен, бұл мағына екі сөздіктегі 16-ға немесе тіпті 23-ке дейінгі дефинициялар да ашылмайды. Сонымен бірге, коллокаттарды талдау нәтижесінде екі сөздікте де белгіленбеген (таңбаланбаған) біркелкі жиі қолданыстағы мағынаны табуға болады: «незначительность» (*қораштық, айтарлықтай емес*) мағынасындағы ***big deal*** тіркесінің қолданылуы. Соңында айтарымыз, барлық бес сөздіктегі регистрлік айырмашылықтар ескерілмейді, ал соңғы кездердегі корпустық деректер бойынша құрастырылған жиілік сөздіктерде маңызды регистрлік модельдерді ескере бастағанын аңғаруға болады.

Бірақ бұл жерде корпустық лингвистикадағы «шегініс» («отступление») деп аталатын аса маңызды мәселені атап айта кеткен жөн болады. Корпустық лингвистика корпуста көрініс таппаған тілдік деректердің құндылығы мен қажеттігін терістемейді және мәтіндер корпусынан барлық мүмкін болатын тілдік қорытындыларды шығарып алуға да бола бермейтінін мойындайды, яғни корпус дегеніміз, жеке тұрып, барлық мәселені жеткілікті түрде шешіп бермейді [4].

Мәселен, корпустық зерттеу нәтижесінде жасалған барлық бес сөздікте де карта ойынындағы «карта тарату» деген мағынаға қатысты деректер табылмаған. Дей тұра, бұл жайт ***deal*** зат есім сөзімен тіркесімінің ең алғашқы мағыналарының бірі болғанмен, ол (карта ойынан басқа кездерде) сирек кездеседі. Аталған кемшілік лексикографиялық жұмыстар үшін аса ірі тұлғалы корпустармен жұмыс істеудің маңыздылығын көрсетеді. Сонымен бірге сөздік жасауда ондай кемшілік жібермеу үшін корпусты қолданумен бірге, сөздікшінің интуициясы да болу қажет. Тіпті карта тарату мағынасындағы ***deal*** зат есім сөзі корпуста бір рет те кездеспесе де (әрбір ағылшын тілі иесі ол сөзді білетіндіктен), сөздік ол сөздің мағынасын қамтуы қажет. Бірақ корпустық талдауға да сүйену қажет болады.

27-кесте

Deal зат есім сөзінің сөздіктік дефинициясы
(қысқаша тұжырымды анықтамасы)

Мағынасы	Webster's Encyclo. 1989	Webster's Third 981	Chamers 1993	Random House 1993	Longman Lang. ε Culture 1992
large but indefinite mount	entry 1 sense 13	entry 3 sense 3	entry 1 sense 2	entry 1 sense 3	entry 1 sense 21
agreement/ arrangement	entry 1 sense 16	entry 1 sense 16	____	entry 1 sense 18	entry 2 sense 1
distribution of cards in a game	entry 1 sense 18	entry 1 sense 3	entry 1 sense 4	entry 1 sense 21	entry 2 sense 4
treatment received	entry 1 sense 15	entry 3 sense 2	entry 1 sense 6	entry 1 sense 6	entry 2 sense 2
act of distributing	entry 1 sense 17	____	____	entry 1 sense 23	____
pine or fir wood	entry 2 3 senses	entry 4 2 senses	entry 2 sense 1	entry 2 3 senses	entry 3 sense 1
act of buying or selling/ a business transaction	entry 1 sense 13	entry 3 sense 2	entry 1 sense 5	entry 1 sense 17	entry 2 sense 1

Сонымен, лексикографиялық жұмыс келешектегі екі мүмкіндікті де қамтуы қажет: мағыналардың барлығын да бөліп алып қарастыру, бірақ олардың регистрлік қатыстылығына қарай, ең жиілеріне немесе аса маңыздыларына сілтеме жасап отыру.

Қазақ тілінің онтомдық түсіндірме сөздігін жасауда алғашқыда картотекалық қорда жинақталған 5 миллион кәртішке пайдаланылған. Кейін 15 томдық сөздікте бұрынғы қол жұмысы жартылай автоматтандырылып, сөздік жасау жұмысын әлдеқайда тездеткен болатын. Онтомдық сөздікте 2 мағынасы ғана көрсетілген көптеген сөздің 15 томдық сөздікте 10-нан аса мағынасы ашылып беріледі. Мұндай мағыналық толықтырулар мысалдар контекстері арқылы ашылған. Енді корпус материалдары бойынша жинақталатын конкорданстар түсіндірме сөздіктердегі сөз саны мен сөз мағыналарын ашуда теңдесі жоқ қуатты құрал болатыны даусыз.

DEAL сөзі етістік ретінде. *Corpus Linguistics* оқулығында етістік ретіндегі *DEAL* сөзінің сөз тіркестері ішіндегі мағыналары жоғарыда аталған технологияны қолдану арқылы қаралады. *Deal with* сөз тіркесінің коллокациясы әртүрлі мағынада қолданылады.

Екі сөзден тұратын *deal with* сөзтіркесі ғылыми әдебиетте де және көркем прозада да басқа коллокацияларға қарағанда жиі кездеседі. Лонгман-Ланкастер корпусында бұл сөз тіркесі шамамен алғанда 1 млн. сөзге шаққанда ғылыми әдебиетте – 157 рет және көркем прозада – 58 рет кездескен. Салыстыру үшін, келесі ең көп кездесетін *deal* етістігінің оң жақтық коллокаты ғылыми әдебиетте – ол *only* 1 млн. сөзге шаққанда – 2,6 жағдайда кездескен. Көркем прозада келесі коллокация – *deal in* 1 млн. сөзге шаққанда – 4,6 жағдайда кездескен.

Deal with коллокациясы үшін жасалған конкорданстар тізімі әртүрлі мағыналарға ие екендігін көрсетеді, солардың ішінен ең жиі қолданыстағы мағына – «кітапта, мақалада, зерттеулерде және т.б. сөз болатындар». Конкорданстар тізімін толық қарап шығу нәтижесінде байқалатыны, ол осы мағынаның ғылыми әдебиетте жиі кездесетіндігі және осы регистр түрінде *deal with* коллокацияның ең көп жиілікке ие екендігі: The second controversy <u>*dealt with*</u> the source of nitrogen in plants; An important point to note is that the

preceding discussion has *dealt with* thermodynamic acidity...; Other environmental effects are *dealt with* in other chapters.

Deal with коллокациясының тағы бір мағынасы – «мәселені шешу»: When they had *dealt with* the fire another crisis arose; Moreover many losses are due to chilling and crushing, both factors that can be *dealt with* by good environmental control and housing.

Басқа стильдерге қарағанда көркем прозада «жағдайдың шешімін табу» мағынасында қолданылады: He didn't have the right temperament to *deal with* the Hennigs of this world; She would have rather there been a fight, anger – or even tears and pleadings. These she could *deal with,* not this deadly coldness exhibited by Alice; But suffering is also a fact. It has to be *dealt with.*

Ақырында, *deal with* коллокациясы «адаммен қарым-қатынас жасау» деген мағынаны білдіреді, әсіресе, бизнестегі серіктесімен қатынас жасау деген мағынаға ие болады: Hansie De Beer runs the farm, he's the one Mehring usually *deals with*; The son handed me a small suitcase with the distant eyes of a man *dealing with* a chauffeur.

Шындығында, *deal with* коллокациясының аталған мағыналары арнайы талдауды қажет етеді. Корпус бойынша *deal with* тіркесінің мағыналарын әртүрлі семантикалық категориялар бойынша (әңгіме тақырыбы, мәселелер бойынша және т.б.) шартты түрде белгілеп алып (кодтап), *deal with* тіркесінен кейін кездесетін сөздерді іздестіруге қатысты талдау жүргізуге болады. Мұндай жағдайда, кеңейтілген тіркесім шеңберін осындай мағыналардың шегін табу үшін қолдануға болады. Бірақ, мұнда коллокациялардың үнемі тек бір ғана мағынаны ғана беруі міндетті емес екенін көрсету маңызды болып саналады. Керісінше, кейбір жағдайларда тек бір ғана коллокация әртүрлі мағынада қолданатындығы көптеген контексте байқалады.

Жоғарыда ағылшын тіліндегі әртүрлі стильдер бойынша ізделген сөздердің оң жақ және сол жақ контекстік қоршаулары бойынша тіркесімділігі анықталып, мағыналары айқындалған. Осы сияқты қазақ тілінде Ұлттық корпус материалдары негізінде әрбір стиль бойынша сөз мағынасын ашудың әдіс-тәсілдерін табуға болады. Ал бұл болашақ лексикографиялық жұмыстарды автоматтандыру дегенге саяды.

Коллокацияларды статистикалық әдістермен бөліп алу. Лексикалық тіркесімді талдау үшін корпустық әдістерді қолдану сөздіктердің жаңа түрлерін құрастыруға, соның ішінде тұрақты сөз тіркестерінің сөздігін құрастыруға мүмкіндік береді. Корпустарды пайдалану лексикалық бірліктердің тіркесіп қолдануы жайлы деректерді, яғни олардың тіркесімдік, меңгерімдік және т.б. ерекшеліктерін алуға мүмкіндік жасайды. Қазіргі кездегі қолданыста бар тұрақты сөз тіркестерінің сөздіктерінде, біріншіден, сөз тіркестері толық түрде қамтылмаған, екіншіден, оларды айтарлықтай жүйелі түрде жасалған деуге келмейді. Сол себепті қазірге кезде сөздіктің жаңа типін, яғни тұрақты сөз тіркестері сөздіктерінің біріктірілген немесе коллокация сөздігі деп аталатын сөздігін жасауға қажеттілік туындайды. Мұндай сөздік, шындығында да, әртүрлі тұрақты сөз тіркестерін қамтиды (фраземалар, немесе түсіндірме сөздіктерде ромбадан кейін берілетін сөз тіркестері) [72].

Қазіргі кезде тіл білімінде коллокацияның қайсыбір бөліктерінің байланыстылық дәрежесін анықтайтын бірнеше тәсілі бар. Ондай статистикалық өлшем ретінде ассоциация өлшемін (*MI*, *t-score*, *log-likelihood*) таңдап алуға болады. Олар көбінде корпустағы сөз тіркестері құрамдарының өзара жақындық дәрежесін есептеу кезінде қолданылады. **Биграмма** (берілген сөздің сол жақ немесе оң жағындағы сөздермен тіркесімі) бөліктерінің сызықтық жақындығы тұрақты тіркесімді табу әрекетінде, сол сияқты мәтіндегі коллокацияларды және сөз тіркестерінің басқа да түрлерін іздеп табу кезінде аса маңызды алғышарттық рөл атқаруы мүмкін.

Орыс тілі үшін статистикалық әдістерді қолдануға болатындығын тексеру үшін және жоғарыда көрсетілген шаралар негізінде коллокацияларды бөліп алу мүмкіндігін айқындау мақсатымен бірнеше тәжірибе жүргізілген (әрі қарай [72] жұмыс бойынша жүргізілген тәжірибелер келтірілді). Зерттеу жұмысы CQP (http://corpus1.leeds.ac.uk/ ruscorpora.html) корпус-менеджердің көмегімен орыс тілінің газеттер мәтіні (The corpus of Russian newspapers) корпусы базасы негізінде жүзеге асқан. Аталған корпус С.А.Шаровтың ғылыми жетекшілігімен Лидса университетінде (Ұлыбритания) 2001–2004 жылдары жарық көрген және 78 млн. сөзқолда-

ныстан тұратын көлемдегі газеттер мәтіні негізінде құрастырылды. Зерттеу материалына 19 зат есім сөздерден тұратын коллокация негіз болды және олар келесі аталатын қағидат бойынша таңдалып алынады. Ең алдымен С.А.Шаровтың орыс тілінің электрондық жиілік сөздігінен бірінші мыңдыққа енетін ең жиі жиіліктегі зат есім сөздер таңдап алынады. Әрі қарай, Шағын (кіші) академиялық сөздік (Малый академический словарь – МАС) (Словарь русского языка 1981-1984) бойынша таңдалып алынған сөздердің қолдану жиіліктерін бұрмалайтын омонимдік сыңарларының бар-жоғы тексеріледі (мысалы, екі мағынадағы *брак* сөзі; *друг друга* деген сияқты сөздер леммаға келтіру кезінде бір лемма түрінде көрініс табады). Омонимдік сыңарлары бар сөздер тізімнен алынып тасталып, тәжірибеде қарастырылмаған. Сосын тізімде қалған зат есім сөздер Е.Г.Борисованың орыс тілінің коллокациялар сөздігімен [6] салыстырылып қаралады. Егер қарастырылып отырған сөзге қатысты оның тіркесімдігі жайлы сөздік мақаласында ақпараттың жоқтығы немесе шектеулігі байқалса, ондай сөздер де тізімнен алынып тасталған. Сонымен, тірек сөздердің мынадай тізімі алынған: *власть*, *внимание*, *возможность*, *война*, *вопрос*, *дождь*, *жизнь*, *закон*, *любовь*, *место*, *мнение*, *мысль*, *ночь*, *ответ*, *помощь*, *радость*, *слово*, *случай*, *смысл*. Төменде 28-кестеде *война* тірек сөзімен бірге алғашқы 10 коллокацияның (106-дан) деректері келтірілген. *Война* тірек сөзі MI өлшемінің мағынасы (ортақ ақпарат көлемі) бойынша сұрыпталған. Мұндағы *Joint* – корпустағы берілген коллокацияның абсолютті жиілігі; *Freq1* – биграмманың бірінші сөзінің абсолютті жиілігі, яғни *война* сөзі үшін сол жақтық коллокат; *LL score, MI, T-score* – log-likelihood өлшем мәні, берілген коллокация үшін MI және t-score. Тірек сөздердің тізімінен мынандай жайтты байқауға болады, бір жағынан, тізімде тұрақты сөз тіркестері орын алса, екіншіден, барынша жоғары MI өлшемінің көрсеткішіне де ие екендігіне көз жеткізуге болады. Зерттеу нәтижелері бойынша, MI өлшемінің мәндері 0-ден 1-ге дейінгі аралықта тұрақты сөз тіркестеріне жатқызуға болатын тіркестер кездеспеген. Бұдан мынадай қорытынды жасауға болады, MI ассоциациясының өлшемдік мәні аталған интервал арасына түсетін тіркестер статистикалық тұрғыда онша мәнді емес деуге болады. Барлық жинақталған

тіркестер бойынша бәріне бірдей мынадай үрдіс байқалады: өлшемнің мәні неғұрлым аз болса, соншалықты ықтималдық көп болады және орыс тілі сөздіктерінде мұндай сөз тіркестері тұрақты деп бекітілмеген. Сондықтан, сөздіктерде берілген тіркесімдік туралы деректер ассоциация өлшемі негізінде алынған деректермен сәйкес келеді. Сөздіктерде келтірілген көптеген коллокациялар (фраземалар) ассоциация өлшемі негізінде түзілген тізімнің жоғарғы жағында орналасады екен. Ал бұл жайт – коллокация деректерінің байланыстылық көрсеткіші мәнінің жоғары екендігінің айғағы.

28-кесте

Война сөзіне қатысты ассоциация өлшемінің мәндері
(сол жақ контекст)

Коллокация	*Joint*	*Freq1*	*LL score*	*MI*	*T-score*
необъявленный война	9	76	30,19	11,03	3,00
междоусобный война	4	54	12,43	10,35	2,00
партизанский война	45	728	135,77	10,09	6,70
рельсовый война	6	100	18,00	10,05	2,45
победоносный война	9	174	26,31	9,84	3,00
вялотекущий война	6	142	16,92	9,54	2,45
позиционный война	5	128	13,90	9,43	2,23
холодный война	171	4747	469,90	9,31	13,06
грянуть война	14	457	37,19	9,08	3,73
финляндский война		148	10,37	8,90	2,00

Жүргізілген тәжірибенің нәтижесінде бірде-бір сөздікте кездеспеген сөз тіркестердің мәтіндерден бөлініп алынғаны белгілі болды, ал бұл жайт аса маңызды екендігін айта кетпекпіз. Мұндай тіркестерді талдаудың нәтижесінде байқалғаны, ол тізімнің бас жағында орналасқан биграммалар (қайсыбір өлшемнің кему тәртібі бойынша сұрыпталған), біршама ықтималдықтар үлесін құрайтын тіркесім сөздер тұрақты тіркестерге жататындығы анықталды, сол себепті оларды сөздікке енгізуге болады. Сөз тіркестері тізімінің төменгі жағында орналасқан сөз тіркестері көпшілік жағдайда,

еркін сөз тіркестері қатарына жататындығы белгілі болды. Түсіндірме сөздіктерде ромбадан кейін келтірілген сөз тіркестер қайсыбір дәрежеде тұрақтылыққа ие болғанымен, оларды толыққанды тұрақты сөз тіркестері қатарына жатқызуға болмайды. Жүргізілген тәжірибе нәтижесі бойынша шығатын қорытынды, ол біріншіден, сипатталған статистикалық тәсілдерді лексикография практикасында қолдануға болады, екіншіден, олар қолданыстағы сөздіктердің тұрақты сөз тіркестерін толық қамтымайтынын көрсетеді.

Мамандандырылған корпустарда коллокацияларды айқындау аса үлкен практикалық маңыздылыққа ие болуы мүмкін. Мысалы, корпус негізінде алынған Н.В. Гогольдің жазған хаттар корпусы негізіндегі деректерді жалпы тілдік корпустар деректерімен салыстырылады, нәтижесінде көптеген жағдайда, авторлық сөзқолданыстың ерекшеліктерін бейнелейтін сөз тіркестерінің елеулі айырмашылықтарын байқауға болады. Сонымен, жоғарыда сипатталған әдіс-тәсілдер жазушы тілін зерттеу мен оның сөздігін құрастыруда және әртүрлі стильдер шеңберінде немесе кезеңдік шығармалардағы тіркесімдіктің ерекшеліктерін айқындауда тиімді әдіс болып табылады.

Бірнеше тілдің корпустарындағы биграммаларды демоверсия режимінде (тәртібінде) іздеуді *http://www.aot.ru/cgi-bin/bigrams.cgi* сайты бойынша жүзеге асыруға болады.

Қазақ тілінің 15 томдық сөздігін құрастыру кезінде де көптеген мысалдар контекстері арқылы жаңа тіркестер анықталып, сөздікке енгізілгендігі белгілі. Болашақта қазақ тілінде де лексикографиялық жұмыстарға корпустың берері мол. Корпус арқылы жаңа сөздер, сөз мағыналарын, жаңа тіркестер, оккозионал қолданыстарды жүздеп табуға болады.

3.5.4. Корпустарға негізделген грамматикалық зерттеулер

Грамматиканы зерттеу тілдің құрылымын түсінумен байланысты. Лексикографияға қарағанда грамматиканы эмпирикалық зерттеулер бойынша қарастыру дәстүрі жоқ. Тіл иелерінің өз грамматикалық қорын барынша пайдалана бастағанына көп уақыт өте қойған жоқ. Дәстүрлі грамматикалық зерттеулердің корпус-

тық деректерге негізделіп жүргізілуі, яғни сөз, сөйлем, дискурс деңгейлерінде жүргізілуі олардың айтарлықтай ерекшелігіне айнала бастады.

Бұл тараушада морфологиялық сипаттаманың қолданылым проблемасы мен функциялары және олардың регистр бойынша үлестірілімін талдау мәселесі қарастырылады. Корпус арқылы морфологиялық сипаттаманың контекстік қолданысына қатысты үлестірілуін анықтап, оның атқаратын функциясын нақтырақ ұғынуға мүмкіндік туады. *Corpus Linguistics* оқулығында осы мәселенің шешімін табу жолдары номинативтену (зат есімдерден жасалған сөздер) үлестірілімі арқылы үш регистр бойынша көрсетілген.

Корпустағы морфологиялық сипаттаманы зерттеу арқылы жиілік пен сипаттаманың үлестірілуін және жекелеген варианттар функцияларының айырмашылықтарын көрсетуге болады. Корпус деректеріне негізделген басқа грамматикалық сипаттамаларды талдаумен салыстырғанда, морфологиялық сипаттамаларды талдау айтарлықтай қарапайым. Оның себебі, морфологиялық сипаттамаларды аннотацияланбаған корпусқа іздеу функциясын қолдану арқылы конкордастарды құрастыру программасы негізінде анықтауға болатындығы.

Номинативтенулердің үлестірілуі мен функциясы. Отандық тіл білімінде номинатив (субстантив) терминін, әдетте, етістіктен абстракті зат есімнің жасалуы деп түсіндіреді және сонымен бірге осы жолмен жасалған зат есімді де осылайша атайды. Еуропалық тіл білімінде бұл ұғымның түсінігі кең, себебі номинативтену ұғымына сын есімнен жасалған зат есімдерді де қосады. Мысалы, *civilization* сөзі де *civilize* етістіктен туындаған номинатив, ал *kindness* сөзі – *kind* сын есім сөзінен туындаған номинатив деп саналады.

Corpus Linguistics оқулығында төрт өнімді суффикстер (tion/-sion, -ness, -ment, -ity және олардың көпше түрі) талданған. Корпус мәтіндерін арнайы программалар бойынша автоматты өңдеуден басқа, мәтінді қолмен өңдеу жұмысы да атқарылады. Мәселен, мұндай өңдеу түр жағынан іздеу шаблонына (қалыбына) сәйкес келетін (search template), бірақ номинативке жатпайтын (*mansion, nation,* c*ity*) бірліктерді бөліп алу қажеттігінен туындады.

Номинативтерді талдау үш регистр (үш стиль) бойынша жүргізілген. Алдыңғы екеуі – ғылыми әдебиет пен көркем проза – Лонгман-Ланкастер корпусының субкорпустары арқылы жүргізіледі. Ал үшінші регистр – ауызша тіл – Лондон-Лунд корпусы арқылы (көлемі 500 мың сөзқолданыс) бейнеленді. Мұндағы жиілік көрсеткіштері мәтіннің 1 млн. сөзіне сәйкестендірілді.

29-кесте

Номинативтердің үш регистр бойынша жиіліктік үлестірілуі

	Ғылыми әдебиет (2,7 млн. с/қ)	Көркем проза (3 млн. с/қ)	Ауызша тіл (0,5 млн. с/қ)
1 млн. с/қ шаққандағы номинатив саны	44 000	11 200	11 300

29-кестеде үш регистр бойынша жүргізілген номинативтенуге қатысты жиіліктік көрсеткіштердің үлестірілулері берілген. Көркем проза мен ауызша тіл мәтіндеріндегі номинативтердің жиіліктері бір-біріне жақын екендігі, ал ғылыми әдебиет мәтіндерінде олардың жиіліктері төрт рет көп екендігі белгілі болды. Контекстегі ең жиі формаларды зерттей келе, не себепті регистрлер ондай әртүрлі үлестірілімге ие екендігін түсіндіруге әрекет жасап көруге болады. Осы ретте *Corpus Linguistics* оқулығының авторлары арнайы программа құрастырып шығарды және ол бойынша әрбір жеке номинативтену есептеліп, олардың әрқайсысы үшін конкорданстар тізімін шығарып беруге болады. Сонымен бірге ол программа арқылы әр регистрдегі номинативтің жеке типтері үшін жалпы жиілікті де есептеуге болады, ал мұндай деректер бойынша контекстегі номинативтің әр түрін жеке зерттеуге мүмкіндік туады.

Регистрде жиі кездесетін айрықша номинативтер корпус мәтіндеріндегі сөз болатын тақырыптарға байланысты болып келеді. Мысалы, ғылыми әдебиетте кездесетін алты номинатив басқаларға қарағанда айтарлықтай жиі, яғни 1 млн. сөзге шаққанда 500-ден арта жиі қолданылады: *movement* (1 млн. сөзге шаққанда 900-ға жуық жағдайда кездеседі), *activity, information, development, relation* және *equation.* Керісінше, номинативтердің ешқайсысы

көркем прозада да және ауызша тілде де жиі қолданбайтындығы байқалған. Мәселен, *movement* сөзі көркем прозада 1 млн. сөзге шаққанда 100-ге жуық кездессе, ауызша тілде 60 рет қана қолданыс тапқан. Мысалы, *development* сөзі көркем прозада 1 млн. сөзге шаққанда бар болғаны 10 рет кездессе, ал ауызша тілде тіпті де қолданыс таппаған.

Аталған 6 номинативтегі конкорданстарды талдау нәтижесі бойынша ғылыми әдебиеттердегі номинатив құбылысы әрекеттер мен адамның қатысынан бөлініп алынған абстрактыланған нысандардың үдерісін (процесін) сипаттайды. Бұл модель movement сөзінің номинативтену мысалынан көрініс табады: The legs and hips, or arms and shoulders, may be used to initiate *movement* in any direction.

Аталған контекстегі қозғалыс – күрделі сөйлемдердің құрамында бастауыш немесе толықтауыш ретінде қолдануға болатындай зат есім арқылы көрініс тапқан үдеріс. Ғылыми әдебиет саласына қатысты мәтіндерде қайсыбір субъектінің әрекеті емес, көбінесе жалпыланған әрекеттің өзі талқыға салынады. Ал көркем проза мен ауызша тіл көбінде адамға қатысты болып келеді, сондықтан оларда адам қылығын сипаттау үшін регистрлерінде етістік пен сын есім сөздер жиі қолданыс тапқан. Бұл регистрлерде әрекет субъектісі ретінде нақты адамдардың әрекеті жиі орын алады және сол себепті оларда ***move*** етістігі жиі қолданады: Garth whistled breathily to himself and *moved* his hand crabwise along the table (*fiction*); It's how much they *move* it that counts (*spoken*).

Осы модельдің *activity, development* және *information* сияқты номинативтерде де қолданыс табатындығын көруге болады: The third Important aspect of *information* is speed; Sometimes algae can stop the *development* and growth of these plants; The experimental results can be described quantitatively by defining the size and *activity* of the shoot and root systems. Көркем проза мен ауызша тілде белгілі адамдардың әрекеттерін сипаттайтын үдерістер пен әрекеттер етістіктің немесе сын есімдердің көмегімен көрініс тапқан: I do hope you know that never in this country do we *develop* the sort of mob war that makes a protest against something however unjust *develop* into an organized riot (spoken); I've *informed* the Soviet government of that visit (spoken); «Aye, the big fellow is *active* again you'll be pleased to know.» (fiction).

Бұл жерде мұндай деректердің келтірілуіне регистрлер мен номинативтер үлестірілімі арасында ассоциативті байланыстың болатындығын айту себеп болып отыр. Көркем проза мен ауызша тілде етістіктер мен сын есімдер нақты адамдардың әрекеттерін сипаттайтын болса, ал ғылыми әдебиеттерде статикалық номинативтер туралы көбірек сөз болады. Әрбір регитрге шаққанда осы сияқты грамматикалық құбылыстар заңдылығын қазақ тілі корпустары бойынша да анықтауға болады. Мұндай зерттеулер әрине бұған дейін де жүргізілген. Алайда корпус массиві мұндай зерттеулердің мүмкіндігін бұрынғыдан мыңдаған есе арттырады. Көлемді регистрлер бойынша анализ жасауға мүмкіндік береді.

Номинативтердің үлестірімі және жұрнақтардың қызметі (функциясы). Номинативтерді топ ретінде зерттегеннен кейін, әрбір жұрнақтың қалай үлестірілетіні жайлы білу аса қажет, демек, номинативтер түрлерінің қызметі мен олардың әртүрлі регистрлерде атқаратын рөлін білу де аса маңызды деген сөз.

24-кестеде нақты жиіліктер көрініс таппаса да, онда әрбір регистрдегі әрбір жұрнақтың номинативтерінің қатынастық пропорциясы көрсетілген.

1. -tion/-sion жұрнақтары номинативтердің көбінде, яғни барлық үш регистрінде де кездессе де, оның пропорциясы ғылыми әдебиеттерде барынша көп (68 %).

2. -ment жұрнағы, ғылыми әдебиетке қарағанда, ауызша тіл мен көркем әдебиетте жиі кездеседі.

3. -ness жұрнағы екі регистрдің әрбіреуінде қолдануына қарағанда, көркем прозада жиі кездеседі.

30-кесте

Әрбір жұрнаққа қатысты номинативтер пропорциясы

Жұрнақ	Ғылыми әдебиет	көркем проза	Ауызша тіл
-tion/-sion	68%	51%	56%
-ment	15%	21%	24%
-ness	2%	13%	5%
-ity	15%	15%	15%

Номинативтер жұрнақтарының ішінен таңдау жүргізуге жауап беретін нақты ережелердің жоқтығына қарамай, олардың әрбір түрін жете талдау арқылы мағынаның белгілі дәрежедегі жүйелік айырмашылығын айқындауға мүмкіндік туады. Осының нәтижесінде *30-кестедеде* көрініс тапқан үлестірім моделіне түсініктеме беруге болады. Мысалы, әдетте жалпыланған үдеріс пен күйді бідіретін (relate/rela*tion* және educate/educa*tion*) *-tion/-sion* жұрнақтары, етістіктерден зат есім тудыру үшін қолданылады, сондықтан да ғылыми әдебиетте бұл жұрнақтардың аса жоғары номинативтер пайызы орын алған. Сол сияқты *-ment* жұрнағы да етістік сөздерді зат есім сөздерге айналдыру үшін қолданылған. *-Ment* жұрнағы арқылы жасалған зат есімдер бір нәрсені өндіру үдерісін немесе белсенділігін белгілеу үшін жиі қолданылған. Бұл номинативтердің көпшілігі барлық қарастырылған үш регистрде де кездеседі, мысалы, movement, government, achievement, agreement, argument. Дегенмен, *-ment* жұрнағы арқылы жасалған номинативтер менталды күйді білдіреді: amazement, agreement, astonishment, disappointment, embarrassment, excitement. Номинативтердің бұл түрлері ғылыми әдебиет пен ауызша тілде сирек кездеседі де, ал көркем прозада кейіпкерлердің ментальды күйін сипаттау үшін айтарлықтай жиі қолданылады: Patrick shrugged in *embarrassment*; The assembly cried out savagely and Ralph stood up in *amazement*.

Аталған номинативтер кейбір жағдайларда ауызша тілде де қолданыс табады: I can quite see there's cause for *disappointment*. Бірақ, көбінде ауызша тілде ментальды жайттар етістіктер мен сын есімдер арқылы беріледі: you'll be *amazed*; are you *disappointed* by not getting honors yourself?

Көркем прозада номинативтердің ең аз саны кездеседі дегенмен, мысалы, ***-ness*** жұрнағы арқылы номинативтеу үшін нормалданған санақ бойынша бұл регистрде оның ең көп саны айқындалған (1 млн. сөзге шаққанда 1430 сөз, ал ғылыми әдебиетте 890 және ауызша тілде 480 сөз). Әдетте ***-ness*** жұрнағы арқылы сын есімдер жекелік сапаны белгілейтін зат есімдерге өзгертіледі. Көркем прозада ***-ness*** жұрнағына аяқталатын зат есім сөздер басқа регистрлерге қарағанда көп кездеседі: aware*ness*, bitter*ness*, good*ness*, happi*ness*, polite*ness*, weak*ness*. Бұл сөздер көркем прозаға

тән сөздерді жан-жақты сипаттауда аса маңызды: The *bitterness* in his heart was now mixed with a kind of childlike excitement; He could see Phyllis's face in profile, and it radiated energy and *happiness*.

Ауызша тілде қолданыс табатын сын есім сөздер жеке басқа (субъектіге) тән және оның сезіміне қатысты болғандықтан, сөйлеушіні немесе адресатты сипаттайтын зат есім сөздердің анықтауышы ретінде жиі қолданады: We feel frustrated and bitter and *annoyed*; I'm not too *happy* about this lauding of language...

Номинативтердің қолданысы жайлы, әрине, бұдан көп айтуға болар еді. Мәселен, әрбір жұрнақпен қолданылатын әр сөздің жиілігін қарастыра отырып, олардың өнімділік сипатын да анықтауға болады; сондай-ақ диахрондық зерттеулер арқылы номинативтер қолданысының даму жолдарын бақылауға мүмкіндік туады. Бірақ жоғарыда келтірілген мысалдың өзі корпустық деректер бойынша морфологиялық зерттеу жүргізудің ұтымдылығының айғағы бола алады. Демек, деривациялық жұрнақтар белгілі регистрлермен ассоциативті байланыста болады және олар регистрлердің бірінші дәрежелі коммуникативті қызметін көрсететіні сөзсіз.

Грамматикалық категориялардың үлестірімі. Ғылыми әдебиет абстрактылы жағдайларды, үдерістер мен нысандарды сипаттауға назар аударса, көркем проза нақты адамдар орындайтын жеке сипаттау мен әрекеттерге көбірек көңіл бөледі. Сондықтан зат есімдер мен етістік сөздердің үш регистрде де (ғылыми әдебиет, көркем проза, ауызша тіл) қолдану жиіліктерінде айырмашылық болуы керек. Шындығында, зерттеушілер белгілі авторлардың стилін сипаттаудың кейбір жағдайларында зат есімдер мен етістіктердің қолдану жиіліктерінің салыстырмалы шамаларын пайдаланады. Мұндай санақтар, *Corpus Linguistics* оқулығы авторларының пікірінше, регистр бойынша сөз таптарына қатысты қолдану айырмашылықтарын көрсете алады.

Грамматикалық вариативтілік барлық халықтар тіліне тән құбылыс және тілдік қолданыстардың эмпирикалық зерттеулері мұндай құрылымдық варианттардың функционалдық жасырын сырын ашып береді. Мұндай зерттеулер дидактикалық мақсатқа маңызды болса, сол сияқты таза ғылыми зерттеулер үшін де маңыздылығы жоғары. Тіл ұстарманның интуициясы бойынша, белгілі

бір варианттың басқаларға қарағанда артық деп санауы әр кезде сенімді бола бермейді. Тек қана корпустық деректерге негізделген нақты мәтіндерді зерттеу нәтижелерінде ғана осындай модельдердің сырын ашуға болады.

3.5.5. Корпустарға негізделген дискурстық зерттеу

Көптеген лексикалық және грамматикалық құбылыстардың тілдегі қолданылу заңдылығын толық түсіну тек көлемді дискурсивті контекстегі олардың атқаратын қызметін талдау (анализдеу) арқылы ғана мүмкін болмақ.

Дискурс дегеніміз – экстралингвистикалық, яғни оқиғалық аспектіде алынған прагматикалық, әлеуметтік-мәдени, психологиялық және тағы да басқа факторлармен қоса есептегендегі жүйелі (қисынды) түрдегі мәтін; мақсатты бағыттағы әлеуметтік әрекет кезінде қарастырылатын сөйлеу, адамдар арасындағы әрекет кезінде және олардың санасындағы механизмдерінде (когнитивтік процестерде) орын алатын компонент [73]. Дискурстың элементтері: баяндалатын оқиғалар, оларға қатынасушылар, перформативтік ақпарат және «оқиға-еместер», яғни а) оқиғаға себепші жағдайлар; ә) оқиғаға түсінік беретін фон; б) оқиғаға қатысушылардың бағасы; в) дискурстың оқиғаларға қатысын айқындайтын ақпарат [74]. Корпустық лингвистика әдістері дискурстық талдаудың дамуына айтарлықтай үлес қосуы мүмкін. Өйткені корпус дискурстың белгілі бір типтерінің ерекшеліктерін және жеке мәтіннің қарастырып отырған регистрдегі дискурс моделіне қаншалықты сәйкес келетіндігін де сипаттап бере алады. Корпустарға негізделген дискурстарды зерттеу мәселесі мынадай 4 салаға бөлінуі мүмкін:

1) дискурсты құрастыру және мәтін құрылымы;

2) қарым-қатынас жасаудың дискурсивті-прагматикалық аспектілері;

3) мәтіндік және прагматикалық коллокациялар;

4) мәтін мен дискурстегі вариативтілік.

Ғалым Т. Виртаненнің пікірі бойынша, ең соңғы екі саланың корпус арқылы жүргізілетін кең ауқымды зерттеулердің келешегі зор.

Corpus Linguistics оқулығының авторларының жорамалдауынша, дискурстың дифференциалдық белгілерін зерттеуде корпустық тәсілді қолданудың негізгі екі әдісі бар:

1. Дискурстың сипаттамаларын талдауда (анализ жасауда) интерактивті компьютерлік программаларды құрастырып (орфографияны тексеру программаларына ұқсас), оларды орынды қолданған жөн. Программаның жұмысы адамға қарағанда дискурстың қасиеттерін айқындауда әрі жылдам әрекет етеді, әрі сенімді нәтижеге қол жеткізеді, сонымен бірге егер дискурсты автоматты талдау мүмкін болмай жатқан жағдайларда зерттеуші өз еркімен шешім қабылдауына мүмкіндігі бар.

2. Барлық мәтіндегі үстірт-грамматикалық дифференциалдық белгілердің қолдану жағдайын қадағалап отыру үшін автоматты талдау тәсілі қолданылуы мүмкін. Талдаудың бұл типтері шындығында да дискурсивті модельдердің дамуын барлық мәтіндер бойынша жүйеге түсіреді, оларды мәтіндерді салыстыру үшін және белгілі регистрлерге тән ерекше модельдерді анықтау үшін пайдалануға болады, сонымен қатар нақты мәтіннің жалпы регистрлік модельдермен қаншалықты сәйкес келетіндігін де байқауға болады.

Дискурсқа қатысты мәселелерді корпустық негізде жоғарыда аталған әдістермен шешуді нақты түрде қарастыру үшін алдын-ала бірнеше мәселелердің шешімін тауып алу қажет болады, мысалы, әртүрлі мәтіндер типтерінде референция сияқты құбылыс белгіленімі (разметканың) қалайша жүзеге асады? *Corpus Linguistics* оқулығында төрт регистрдегі (Лондон-Лунд корпусындағы формалды емес сұхбат және Лондон-Лунд корпусындағы бұқаралық (көпшілік алдындағы) сөйлеу (публичная речь), Ланкастер-Осло-Берген корпусындағы жаңалықтар репортажы мен ғылыми әдебиет) зат есім мен есімдіктердің қолданыстарына зерттеу жүргізілген. Аталған зерттеу келесі сұрақтарға жауап алуды мақсат етеді:

1) мәтін ішіндегі зат есімдер мен есімдіктерді таңдауға қандай факторлар әсер етеді?

2) зат есімдердің ішінен қайсысы «белгілі» ақпаратты, ал қайсысы «жаңа» ақпаратты білдіреді?

3) белгілі референттер мен жаңа референттер мәтін ішінде қалайша үлестіріледі?

Сөйлемдер арасындағы грамматикалық байланыстар арқылы жүзеге асатын ауызша және жазбаша дискурс мынадай үш типке бөлініп қарастырылуы мүмкін: референция, эллипсис және жалғаулықтар. Ағылшын тілінде референцияға жататындар: жіктеу есімдіктері (he she it wi they және т.б.), сілтеу есімдіктері (*this, that, these, those* және т.т.), белгілілік артиклі *the* және *such a* тіркесі. Зат есімді есімдікпен алмастыру прономинализация (pronominalization) деп аталады. Бұл есімдік мәтінде бұрын, кейін аты аталған немесе мәтіннен тыс жатқан зат есімге қатысты болуы мүмкін, бірақ дискурс контексіне енуі қажет [75]. Егер ақпарат сипаты «таныс» деп анықталатын болса, референцияның үш түрлі типін бөліп қарастырады: анафоралық (анафорическая), экзофоралық (экзофорическая) және шығарылымдық (выводимая). Анафоралық референция жағдайында есімдік мәтінде бұрын аты аталған зат есімді алмастырады, мысалы: *The room is large. It is light and clean.* Экзофоралық референция жағдайында есімдік мәтіннен тыс жатқан зат есімге сілтеме жасайды, бірақ мәтінде сөз болып жатқан жағдаятқа қатысты әрекеттің бөлігі ретінде жорамалданып, дискурс контекстіне енеді. Экзофоралық референция әр кезде эксплицитті бола бермейді, яғни оны дұрыс түсіну үшін барлық өтіп жатқан оқиғалардан хабардар болып отыру керек, мысалы, *That winter. It was awful.* Ал шығарылымдық (выводимая) референция жағдайында референттік құралды түсіндіруді талап ететін ақпарат мәтіннің өзінде орын алады.

Сөйлемдегі есімді тіркестер адамдарға, мәтінде орын алатын нысандарға және басқа да заттық ұғымдарға сілтеме жасайтын негізгі грамматикалық құралдар болып саналады. Бірақ әртүрлі регистрлерге қатысты мәтіндер осындай «сілтеме сөйлемшелерді» қолдануда жиі ерекшеленеді. *Corpus Linguistics* оқулығында жаңалық репортажы мен бейформалды сұхбаттардан алынған екі мысал қарастырылған, есімді тіркестер курсивпен бөлініп көрсетілген:

4-мысал. Жаңалық репортажы

Thortec International Inc. said *it* reached *agreements* with *an investor group* and *Wells Fargo Bank* under which *it* will receive loans

and *an equity infusion* in return for *stock* that will reduce *the number of shares in public hands* by as much as 85 percent. *The engineering and consulting firm,* which has been plagued by *losses for five years,* said *the restructuring* is required to relieve *its debt burden* and *«acute shortage of cash.»*

5-мысал. Бейформалды сұхбат

A: Right, *I*'m ready. Have *you* locked *the back door?* [pause] *I* thought *we*

were walking.

B: Well do *you* want to walk or do *you* want to go in *the car?*

A: Well *I* have to go to *the paper shop*.

B: Well *I*'ll drop *you* at *the paper shop* while I go round.

A: Oh *that's a good idea.*

Қарастырылып отырған мәтіндегі бұл мысалдар арасындағы жақсы байқалатын елеулі бір айырмашылық есімді конструкциялардың тұрпатына (формасына) қатысты. Жаңалық репортажы мысалында негізінен толық түрдегі есімді конструкциялар *(Thortec International Inc., agreements, an investor group* және т.б.) қолданылса, бейформалды сұхбаттарда есімдіктер айтарлықтай жиі қолданылады (***I, you, we, that***). Тағы да айтарымыз, бұл мысалдарда референцияның әртүрлі типтерінің қолданылатындығы айқын байқалады. Мәселен, бейформалды сұхбатқа қатысты мысалдарда бұрын мәтінде кездескен басқа нысандарға қатысты емес, сөйлеуші мен адресатқа тікелей байланысты ***I*** және ***you*** есімдіктерімен бірге қолданылатын экзофоралық референция жоғары пайызда орын алған. Жаңалықтар сұхбаты бойынша алынған мысалда мұндай типтегі референциялар жоқ. Сонымен бірге, экзофоралық референцияға көптеп сүйенгендіктен, бейформалды сұхбатқа қатысты мысалда референттердің көпшілігі екі жақтың қатысушыларына таныс, мысалы: *I*, *you, the back door, the paper shop*. Ал жаңалықтар сұхбаты бойынша референттердің көбі басынан-ақ таныс емес деуге болады, мысалы: *agreements, an investor group*.

Сонымен, корпустық деректерге сүйенген талдау нәтижелері референциалды сөйлемшелердің сипаттамаларын зерттеуде және оларды әртүрлі регистрлерде қолданудың дәрежесін анықтауда да қолданылуы мүмкін.

Референциалды сөйлемшелердің сипаттамалары. Референциалды сөйлемшелердің көптеген сипаттамаларын олардың әртүрлі мәтіндер мен регистрлерде қолдануын зерттеу арқылы жақсырақ ұғынуға болады. *Corpus Linguistics* оқулығында мысал ретінде 4 параметрі талданады:

1. ақпарат ахуалы: белгілі :: жаңа;

2. белгілі ақпарат үшін референция типі: анафоралық, экзофоралық немесе эндофоралық;

3. анафоралық референция үшін сөйлемшенің тұрпаты (формасы): есімдік, синоним немесе қайталау;

4. анафоралық сөйлемше мен анафоралық референцияның антецеденті арасындағы арақашықтық.

Мәтіндегі есімді тіркестердің әрбіреуі ондағы ұсынылған ақпарат типіне (белгілі немесе жаңа) сәйкес жүйеленуі (топтастырылуы) мүмкін. Мәселен, жаңалық сұхбатына қатысты мысалда (4-мысал) көптеген есімді конструкциялар адамға немесе қайсыбір нысанға сілтеме жасай отырып, бұрын мәтінде аталмаған жаңа ақпаратты ұсынады. Мұндай типтегі есімді тіркестер келесілерді қамтиды: *Thortec International Inc., an investor group, Wells Fargo Bank, loans, an equity infusion, stock.* Басқа референциалды сөйлемшелер алдында аталған белгілі ақпаратты білдіреді. Сонымен, бірінші сөйлемде *it* есімдігі белгілі референтті, яғни: *Thortec International Inc* атты компанияны белгілеу үшін екі рет қолданылған.

Белгілі ақпаратты енгізуге арналған сөйлемше үш типті референциалды қатынастан тұрады. Мұндай сөйлемшелердің көпшілігі анафоралық болып келеді, яғни олар мәтінде аталып кеткен адамға немесе нысанға – антецедентке – қатысты болады. Сонымен, бірінші сөйлемдегі *it* есімдігінің антецеденті ретінде *Thortec International Inc* сөйлемшесі алынады. Бірақ басқа референттер сыртқы контекстегі адамға немесе нысанға қатысты болғандықтан, белгілі ақпаратты білдіреді.

Бейформалды сұхбаттан алынған мысалда (5-мысал) *I* және *you* есімдіктері тікелей сөйлеуші мен тыңдаушыны (адресатқа) дәлдеп көрсетіп тұр. *The back door, the car* және *the paper shop* сөйлемшелер кеңейтілген физикалық жағдаятта орын алатын физикалық нысандарға қатысты және олар сұхбатқа қатысушы-

лардың екеуіне де түсінікті деуге болады. Мұндай референттерді экзофоралық деп атайды. Оларды сәйкестендіру әрекеті физикалық жағдаятқа байланысты болғандықтан, аталған референттерді белгілі деуге болады. Керісінше, анафоралық референттердің белгілі болуының себебі, оларды сәйкестендіру (идентификация) бұдан бұрынғы *мәтіндік референцияның* арқасында ғана мүмкін болады.

Жүйелеуге аса қиындық тудыратытын бұрыннан белгілі ақпаратты түсінікті етіп жеткізудің де әдіс-тәсілдері бар. Соның ішінде жаңалық сұхбаты бойынша келтірілген мысалда (4-мысал) екінші сөйлемде көңіл бөлініп, бірінші сөйлемде бейнеленген *restructuring* сөзі оқиға ішінен «шығарылымдық» (выводимый) типке жатады және бұл зат есім анофоралық тұрғыда ешбір алдында кездескен есімді тіркестерге және сыртқы контекске де жатпайды. Осылайша *debt burden* тіркесінің орын алуын «plagued by losses» компаниясының өмірде бар болу фактысынан шығаруға болар еді, бірақ бұл да анафоралық қатынас емес екендігі белгілі. Сонымен, «шығарылымдық» категориясы да референттерді жүйелеу ісінде басқалармен қатар маңызды болып табылады.

Үшінші параметр анафоралық референттердің әртүрлі формаларының көрінісін бейнелеуге қатысты. Олар есімдіктермен жиі көрініс табады, бірақ синонимдік сөйлемшемен де берілуі мүмкін, мысалы, *the engineering and consulting firm* екінші сөйлемде мынадай *Thortec International* синонимдік тіркеске қатысты. Тағы да анафоралық референттер алғашқы сөйлемшенің тікелей қайталануы арқылы да берілуі мүмкін.

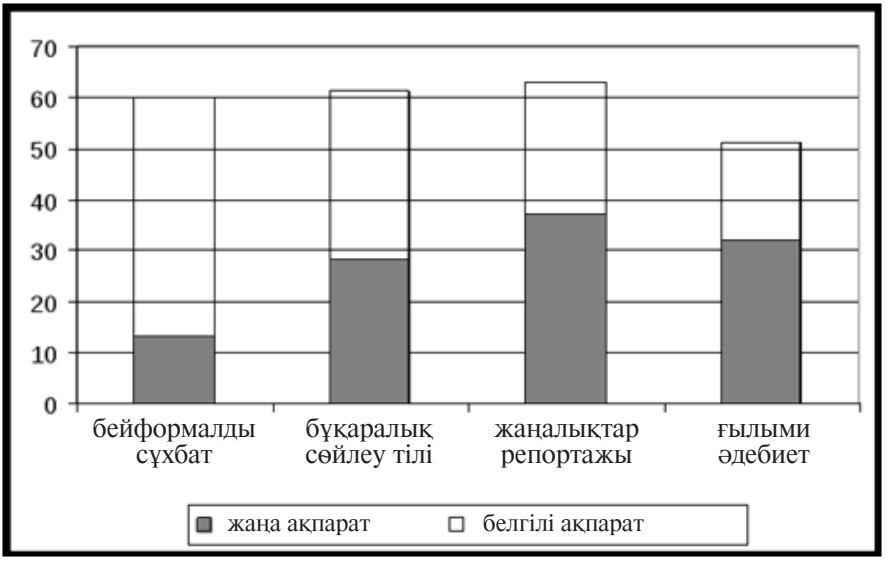
Төртінші параметр референциалды сөйлемше мен оның антецеденті арасындағы арақашықтыққа қатысты. Мәселен, жаңалық сұхбатындағы ***it*** есімдігі басқа тіркестерге қарағанда *Thortec International Inc* антецедентіне жақындау екені байқалады. *The engineering and consulting firm* барынша толық түрдегі синонимдік сөйлемшесі бұл компания атының алғашқы аталуынан айтарлықтай қашықтықта орналасқан.

Жоғарыда аталған барлық төрт параметрді бірге алып қарастыратын болсақ, олардың референцияның әртүрлі регистрлерде қолданылуының көптеген модельдерін ашуы мүмкін деуге әбден

болады. Бірнеше мың сөзден тұратын мәтінді ғана талдау ісіне көп уақыт жұмсалуы мүмкін, сондықтан референциалды сөйлемшелердің сипаттамаларын зерттеу үшін корпустық әдіс-тәсілді қолданған тиімді.

Ауызша және жазбаша регистрлерде референциалды сөйлемшелерді қолданудың модельдері. *Corpus Linguistics* оқулығы авторларының есімді конструкциялардың сипаттамаларына жүргізген статистикалық зерттеулері әртүрлі регистрлердегі *референциалды сөйлемшелер* типтерінің ерекшеліктерін айқындайды. *8-суретте* белгілі және жаңа *референциалды сөйлемшелердің* төрт регистр бойынша үлестірімдері көрсетілген.

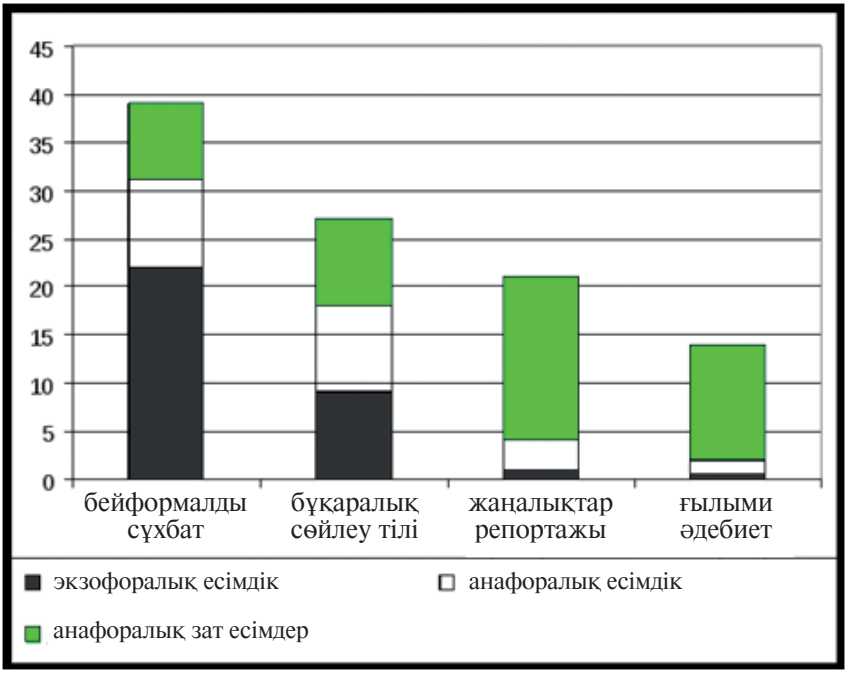
Бұл нәтижелер бойынша күтпеген ерекшеліктерді байқауға болады. Жаңалықтар сұхбатында *референциалды сөйлемшелер* өте көп, ал ғылыми әдебиетте өте аз шамада кездескен. Әдетте ауызша тілде жазбаша тілге қарағанда етістіктердің көп қолдануына қарамай, бейформалды сұхбат пен бұқаралық сөйлеу (публичная речь) тілінде *референциалды сөйлемшелер* саны айтарлықтай көп



8-сурет. Референциалды сөйлемшелердің қолдану жиілігі: белгілі :: жаңа

кездескен. Белгілі және жаңа ақпараттар сәйкестігі тұрғысынан қарағанда, барлық қарастырылған төрт регистр бір-бірінен мүлде ерекшелінеді. Бір жағынан қарағанда, бейформалды сұхбатта *референциалды сөйлемшелердің* 70 пайызынан көбі «белгілі ақпараттар» типіне жатады, ал, екінші жағынан, ғылыми әдебиеттердегі барлық *референциалды сөйлемшелердің* 65 пайызынан көбі «жаңа ақпараттар» типіне қатысты екен. Жаңалықтар сұхбатында да «жаңа ақпараттар» типінің ең жиі кездесетіні байқалады. Бейформалды сұхбатта белгілі ақпараттың ең көп кездесуін референциалды сөйлемшелердің әртүрлі типтерін қарастыра отырып жақсы ұғынуға болады.

9-суретте референциалды сөйлемшелердің кездесу жиілігі көрсетілген. Бұл кестеде экзофоралық есімдіктер, анафоралық есімдіктер мен анафоралық зат есімдердің қолдану жиіліктерінің



9-сурет. Экзофоралық және анафорикалық референциалды сөйлемшелердің жиілігі

көрінісіне сипаттама жасалған. Бейформалды сұхбатта белгілі ақпаратты бейнелейтін референциалды сөйлемшелердің жартысынан көбі экзофоралық есімдіктердің үлесінде екені байқалады. Бұл есімдіктердің түгелге жуығы сөйлеуші мен адресаттың орындарына қолданылады, дей тұра, экзофоралық референцияның контекстегі ситуациялық жағдаятқа қарай (*she, he* немесе *it*), үшінші жақтағы адамдар мен нысандардың орнына жүретіні де кездеседі.

6-мысал бейформалды сұхбатқа тән экзофоралық референцияның ең жиі қолданылатын жағдайларына сипаттама жасайды. Референцияның сөйлеуші мен тыңдаушыға қатысты жақтары курсивпен бөлектеліп көрсетілген, ал экзофоралық референцияның нысанға қатысты жайттары квадрат жақша ішінде қалың қаріппен бөлектеліп берілген (сол сияқты экзофорикалық адвербиалды зат есімдер де бөлектеліп көрсетілген).

6-мысал. Бейформалды сұхбат
A: What are *you* doing **[this afternoon]?**
B: *I*'m going **[home].** *I*'ve got to teach about half past one.
Can *you* pick
[your own trousers] up?
A: no *I* don't think it'll be likely. *I*'ve got **[this meeting at three thirty]**
and**…**
B: well what are *we* doing **[this weekend]? [Tomorrow]**
I've got **[dancing**
class] in **[the morning]**
A: well I've nothing down anyway at all.
B: well **[they]**'re open **[tomorrow afternoon]** up till three o'clock, *you*
remember **[last time]** *we* went, and **[they]** were closed.

Жазбаша регистрлер көбінде анафоралық референцияға сенім артады. Анафоралық референция үшін қолданылатын есімді формалар артықшылыққа ие бола келе, мынадай өзіне тән ерекшеліктері бар: жазбаша регистрлерде әрқашан толық түрдегі зат есімдер қолданыс табады, ал бейформалды сұхбат пен бұқаралық ауызекі

сөйлеуде анафоралық есімдіктер мен толық түрдегі зат есімдер, шамамен алғанда, сан жағынан бірдей қолданылады.

Corpus Linguistics оқулығының авторлары зерттеп дайындаған компьютерлік бағдарлама анафоралық референциалды сөйлемшелері және олардың мәтіндегі антецеденттері арасындағы арақашықтықтарды кодтау әрекетін, яғни ондай деректерге шартты белгілер енгізу жұмысын жүзеге асырады. Мұндағы арақашықтық есімді конструкциялар санымен анықталды. Әрбір анафоралық референциалды сөйлемшелерге қатысты арақашықтық өлшемін негіз ете отырып, әрбір мәтін мен регистрге қатысты орташа арақашықтықты есептеп шығаруға болады. ***31-кестеде*** көрсетілгендей, барлық төрт регистр үшін орташа арақашықтық әртүрлі. Бейформалды сұхбат пен бұқаралық сөйлеу (публичная речь) тілдерінде орташа арақашықтық жазбаша регистрлерге қарағанда өте төмен. Егер жазбаша және ауызша регистрлердегі продуцирлеу мен қабылдау жағдайларындағы айырмашылықтарды ескеретін болсақ, бұл аса маңызды. Бейформалды сұхбат пен бұқаралық сөйлеу тілдері нақты мезгілде (уақытта) қабылдануы және сонымен бірге бейформалды сұхбат та нақты мезгілде продуцирленуі қажет.

31-кесте

Төрт регистр үшін орташа арақашықтық

Регистр	Орташа арақашықтық
Бейформалды сұхбат	4,5
Бұқаралық сөйлеу тілі	5,5
Жаңалықтар репортажы	11,0
Ғылыми әдебиет	9,0

Осындай жағдайларда қысқа анафоралық арақашықтықтағы кореференцияны ұғыну жеңілге түседі. Керісінше, жазбаша мәтіндерді туындату мен қабылдап алу жағдайында оны жазатын және оқитын адамдар уақытты өз қажеттіктеріне қарай жұмсай алады

және осының нәтижесі ретінде, мұндай мәтіндердегі кореферентті сөйлемшелер арасындағы арақашықтық көбірек болады.

Бейформалды сұхбат кезінде сөйлеушіге немесе тыңдаушыға қатысты төменгі орташа қашықтықта байқалатын экзофоралық есімдіктерді жиі қолдану жайттары да маңызды фактор болып саналады. Мәселен, төменде 5-мысал курсивпен жазылып ерекшеленген «А» сөйлеушіге қатысты барлық экзофоралық жағдайлармен және қалың бас әріппен жазылып ерекшелінген «В» сөйлеушіге қатысты барлық экзофоралық жағдайлармен қайталанып берілген. Референциалды сөйлемшелердің барлық басқа түрлері квадратты жақшалар ішінде жазылған.

A: Right, *I*'m ready. Have **YOU** locked [the back door]?
I thought *we* were walking.
B: Well do *you* want to walk or do *you* want to go in [the car]?
A: Well *I* have to go to [the paper shop].
B: Well **I**'ll drop *you* at [the paper shop] while **I** go round.

Бұл қысқаша үзіндіде сөйлеуші «А»-ға қатысты референция жағдайында орташа арақашықтық 1-ден аз, себебі олардың көпшілігі араларында есімді конструкцияларсыз қатар қолданылған. Керісінше, сөйлеуші «В»-ға қатысты референция жағдайларының орташа арақашықтығы 5-ке жуық қолданылған.

Интерактивті компьютерлік бағдарламалар және мәтінді өңдеудегі автоматтандырылған әдістер көмегімен қолжеткен мұндай нәтижелер дискурсивті жүйелерді талдап зерттеуде аса маңызды. Әсіресе регистр түрлеріне тән дискурстың сипаттамаларын айқындауда корпустық әдіс-тәсілдерді қолдану мақсатқа сай келеді деуге болады [71, 34].

Бейформалды сұхбаттағы қаратпа сөздердің үлестірімі. Дж. Лич жасаған дискурсты зерттеу жұмысы ағылшын тілінің американдық және британдық варианттары бойынша сұхбаттағы қаратпа сөздердің үлестірімі мен қызметіне арналған [76]. Корпусқа негізделген зерттеудің нысаны ретінде қаратпа сөздердің грамматикасы алынған болатын. Қаратпалар – өзі кездесетін сөйленім адресатына қатысты, сөйлем мүшесі болмайтын, еркін орналаса-

тын субстантивті элементтер. Корпустан жинақталған деректерді зерттей келе, автор бірнеше семантикалық кішігірім (ішкі) категорияларды бөліп қарастырады. Мысалы, жылы сөйлеу: *Honey*, can I use that ashtray, please; туыстарға арналған сөйлеу: Thanks, *mom*, ok, talk to you later; дөрекі сөйлеу (familiariser): Got a ticket, *mate*?

Автор барлық кішігірім (ішкі) категориялар бойынша мүмкін болатын қолданыстардың кездесу санын есептеп, ағылшын тілінің американдық және британдық варианттары бойынша мынадай айырмашылықтарды айқындайды: британдық вариантқа қарағанда американдық нұсқада қаратпа сөздер 25% жиі қолданған, ал туыстық терминдер британдық вариантта жиірек кездеседі, американдық вариантта дөрекі сөйлеу (фамильяризующие обращения) жиі қолданылған.

Зерттеу жұмысында қаратпа сөздер олардың сөйлемдегі орнына қатысты қарастырылған (*32-кесте*), сонымен бірге қызмет типтері арасындағы әртүрлі қатынастар (көңіл бөлдіру, адресатқа сілтеме жасау, әлеуметтік өзара әрекет) мен сөйлемдегі қаратпа сөздердің сөйлемдегі позициялары да зерттеледі.

Корпустың айтарлықтай репрезентативті еместігі сілтемеде көрсетіледі.

32-кесте

Қаратпа сөздің сөйлемдегі орны

Сөйлемдегі орны	Пайызы	Мысал
Сөйлем соңы	68%	Come on, *Sam*.
Сөйлем басы	11,5%	*Doug*, do you want some more ice-cream?
Жеке орналасу	11,25%	*Mom*!
Сөйлем ортасы	9, 25%	What have we lost at home, *Paulie*, this season?

Автор зерттеу нәтижесінде жаңа күтпеген нәтижелер алады:

1. Адамдар даяршымен (официантпен) сөйлескенде ***sir*** қаратпа сөзін жиі қолданады, ал мұны, мүмкін, қоғамның демократизация бағытына қарай жылжуымен түсіндіруге болатын шығар.

2. Қаратпа сөздер сөйлем соңында жиі кездеседі, ал бұл, мүмкін, біреудің көңілін аударуды мақсат еткендіктен болар және басқалармен салыстырғанда, әлеуметтік қызметінің маңыздылығымен түсіндіруге болады.

Дыбыстық корпус материалы бойынша дискурстық зерттеу. Орыс тілінің материалы бойынша энантиосемиді зерттеу жүргізілген, яғни сөздегі қарама-қарсы мағыналардың бірге келуін немесе «ішкі антонимия» мәселесін зерттеу [77]. Бұл құбылысты күнделікте тілдесім көрініс тапқан орыс тілінің дыбыстық корпусы материалы арқылы (*Бір сөйлесу күні* – БСК; Один речевой день – ОРД) (3.2.2.2.2 тараушаны қара) зерттеу тілшілердің бұрын ескере қоймаған энантиосеми түрінің (типінің) бар екендігін дәлелдеуге, сөйлеу тілінде жиі қолдананатын риторикалық энантиосемияны зерттеуге мүмкіндік береді.

ОРД (немесе БСК) корпусындағы белгілі фрагменттерінің лексикасын зерттеу мынадай жайтқа әкелді, И19 арқылы шартты белгіленген ақпарат беруші (30-35 жастағы әйел адам) көңіл қалдырар ештеңе айтпайды, тек сұхбаттасушыға табыс тілейді, тіпті мақтайды (*азамат, тамаша, жақсы, өте жақсы)*. Бірақ баланың аты және басқа да жақсы көру, еркелету маркерлері (орыс тіліндегі *солнышко, заинька, зайка, котик, умница, дорогой, милый, миленький және т.б.*) балаларға айтылатын жағымды дыбыстармен, қажетті эмоциямен айтылмайды.

Келесі мәтін үзінділерінде адресант жақын адамдармен сөйлесудегі дәстүрлі қаратпа сөздерді қолданады (*солнышко, (мое) солнце, (моя) радость*) және баланың атын кішірейту, еркелету жұрнақтарын жалғап (***оньк, очк)*** айтады (мәтін үзінділері орыс тілінде):

1) *Але-о!/ Привет / что все?/ закончилось / у вас закончилось / что случилось?/ почему? / ты где? то есть она сейчас уже ушла? а спроси у Иры / да-а/ я же тебе сказала / подойти к Ирине и спросить/ попросить помощи / Да / дойди / дойди* ***солнышко****/ давай/ удачи/ ну тебе/ у тебя все хорошо/ ну замечательно/ ну /угу / хорошо /* ***солнышко*** */ давай / не теряй времени / подойти к Ирине / попроси помощи найти Нину Филипповну/ и отзвонись мне/ пока //*

2) *Але /да* ***солнышко*** */ну так /ну замечательно/ отлично /поздравляю тебя /* ***молодец*** */все не так страшно /готовься /что делать/ Держись /держись мое* ***солнце*** */ну давай / держись / пока //*

3) *Ты выпила* ***водичку****? Отлично! <...> Что случилось? Что именно ты забыла, моя* ***радость****? Можно поподробнее? Что ты забыла? Что сегодня у вас был английский/ ты забыла. <...>Я не поняла, что да? Зачет?<...>проверка по словам<...> Какие слова? Что/ и сейчас забыла? И сейчас забыла / какие слова? <...> Из какой лексики/* ***Лизонька****?*

4) ***Лизонька*** */ это Марина Викторовна ваша / такое дурацкое слово употребляет «лексика» /оно дурацкое /* ***Лизочка*** */ Лексика это всего-навсего слова <...>Лексику вы учите / бред какой /*

5) ***Моя девочка****/ <...>ну что делать-то с этой двойкой /* ***моя хорошая***

Сенімділік маркерлерін интонациялау ерекшеліктері тыңдаушының күту шегі мен сөйлеушінің интенцияларына шиеленіс тудырады, ал сөздің семантикасы мен бейтарап интонация арасындағы диссонанс (*үндестіктің бұзылуы*) тілдесу кезіндегі бейнормалық (аномальды) эмоционалдық фонға негіз болады деген болжам (гипотеза) ұсынылды. Осы болжамды тексеру үшін ғылыми тәжірибе өткізілген, ал оған қатысушыларға (30 студент пен оқушылардан тұратын топ) расшифровканы (мағынасын ашу) оқып шығу және мынандай сұрақтарға жауап беру қажет болды: «ата-аналардың өз балаларымен осылайша сөйлесуі жиі кездесе ме? Сөйлесудің мазмұны типтік пе? Осындай әйелге қатысты не айтуға болады?» Ақпарат берушілер (информанты) мәтінді дөрекі (банальный) деп қабылдайды, ал көпшілігі кішкентай бала өзі қорқатын бір емтихан немесе зачетты тапсыруы қажет, ал анасы баласы үшін қатты уайымдап, қызына көмектесуге тырысатыны туралы ақпарат берілгендігін айтады.

Жазып алынған ақпаратты *тыңдап болғаннан* кейін мынадай сұрақтарға жауап беруге ұсыныс жасалады: «Лиза туралы сіз қандай жаңалық білдіңіз?» Ақпарат берушілер (информанттар) өте бір қызбалықпен, эмоциямен қабылдап, тіпті қойылған сұраққа да жауап бермей, тек «И19»-дың айтқандарына өз пікірлерін білдіру-

мен болады: «Неге ол сондай тексеру (прокурорлық) тонмен сөйлейді?» «Енді ол бұлайша сөйлемеуі үшін, сіз оған әсер ете аласыз ба?» «Егер сіздің ұлыңыз мектепке барса, мүмкін, сіз де солайша сөйлеп тұрарсыз!» «Бұл әйел, мүмкін, жай шаршаған болуы керек, ал оның қызының әлі қандай екені белгісіз». Лиза мен оның анасының тәртібіне баға берудің әртүрлілігіне қарамай, айтылған сөздер тек мынаны байқатады: тыңдалған үзінділерден тәжірибеге қатысушылар әлдебір жағымсыз әсер мен жиіркенішті жайттарды сезінеді, ал мұндай сезімді мәтіндердің жазба түрі таралған кезде байқау мүмкін емес.

Коммуникативті актілерге қатысушылар сөздерінің риторикалық энантиосемдік мағыналары ешқандай трансформацияға (өзгеріске) ұшырамайды. Сонымен бірге, мұнда белгілер немесе фактілер деп аталатын интонациялық терістеу де жоқ. Тіпті қауіп-қатерден немесе мысқылдан айырылған интонация, риторикалық энантиосемияны өзінде сақтайтын фраза өте ауыр түрдегі эмоционалды фон жасайды. Ұзақтығы 22 мин. 36 сек. (анасымен сөйлесу ұзақтығы) болатын дыбыстық файлды тыңдаған кезде кішкентай қыздың жеті рет жылай жаздаған әрекеті кездейсоқ емес. Жазбада эмоционалды боялған интонацияны қажет ететін жағымды коннотацияға ие сөздер бейтарап үнмен айтылады. Жылы сөздерді өте көп қолдану, осындай үнмен қайталап айту (5, 6, 7-мысалдыр) дыбыстауды күштеп үдетуге (қатты дауыспен айтуға) себепкер болады, ал сосын үздіксіз сұрақтарға немесе наразылық шағымдарға жалғасады.

3-7-мысалдар шеттетушілікті, жатырқаушылықты білдіретін ерекше интонацияны сипаттайды, ал бұларды автор өгейлік, бөтендік маркеріне жатқызады. Солай бола тұра, жазбада сөйлеуші мінез-құлығында сұхбаттасына қысым білдіру мақсатындағы агрессиялық қылықтар немесе үннің (дауыстың) төмендетілуі мен көтеріңкілігі байқалмайды [78]. Сұхбаттасушылар өздерін тең құқықты деп санайтын қалыпты (агрессиялық емес) сөйлеу әдебіне қарағанда, сөйлеудің агрессиялық жағдайында диалогқа (сұхбатқа) қатысушылар арасында тең құқықтық шарт есепке алынбайды, сондықтан олардың біреуі – күш көрсетуші (агрессор) бол-

са, ал екіншісі оның «құрбаны» болады. Мүмкін, «И19» қақпасөзге қатысушылар арасында агрессияға қарсы агрессия пайда болады, алайда жақын адамдар болғандықтан, сөйлеушінің эмоциялық ширығуы төмендейтін болу керек [77].

Зерттеуге қатысты бұл мысал бойынша, дыбыстық корпус материалдары негізінде маңызды теориялық қорытынды жасауға болатындығын байқауға болады. Авторға энантиосемидің (бір сөздің бір-біріне қарама-қарсы мағынаны білдіруі) жаңа типін – риторикалық энантиосемияны айқындауға мүмкіндік туады.

Қазақ тілінің ұлттық корпусында осындай ауызша корпустар жасалған жағдайда дискурсты зерттеудің жаңа мүмкіндіктері туатындығы даусыз.

4. ҚАЗАҚ ТІЛІНІҢ ҰЛТТЫҚ КОРПУСЫН ҚҰРАСТЫРУ ТӘЖІРИБЕСІНЕН

4.1 Сөздердің морфологиялық құрамын автоматты бөлшектеуге қажетті реестр сөздерді өңдеу

Қазақ тіл білімінде тілді автоматтандыру мәселесі алғашында жиілік сөздіктер құрастыру ісінен бастау алды. Тілдік жүйені сапа мен санның арақатынасы тұрғысынан қарастыру тілдік бірліктердің табиғатын кеңінен тануға мүмкіндік беретінін ғалымдар ертеден-ақ байқап, сандық деректерге ерекше мән берген. Қазақ тілінде бұл игілікті істің нәтижесі ретінде М.Әуезовтің 20 томдық шығармаларының жиілік сөздігі, бұдан басқа ертегілер тілінің жиілік сөздігі, ғылыми, публицистика тілінің жиілік сөздіктері т.б. жарық көрді.

Қолданбалы лингвистикада 70-жылдары тілдік материалдарды автоматты өңдеу мәселесімен байланысты математикалық, формалды-логикалық, статистика-ықтималдық әдістерді игеру қажеттігі туды. Соның нәтижесінде математикалық лингвистика саласы дүниеге келді. Осы ретте қазақ тіл білімінде 2002 жылы тілдік бірліктерді модельдеуге қатысты А.Жұбанов «Основные принципы формализации содержания казахского текста» атты докторлық диссертациясын қорғады және ғалымның осы жылы көлемді монографиясы жарияланды.

А.Жұбанов білімді (мазмұнды) модельдеуге (формалдауға) қатысты мынадай тұжырымдар жасайды: «Көптеген ғылымдардың осы мәселеге қызығушылығының негізгі себебі біреу-ақ, ол автоматты жасанды интеллект жүйесін құру. Солай бола тұра, «білім» бізге «тіл» арқылы берілетіні анық. «Білім» сөйлеу мәтінінде де (монолог, диалог, әр жағдайдағы реплика) және сонымен бірге жазба мәтіндерде де (ескі жазба ескерткіш, көркем әдебиет, ғылыми-техникалық әдебиет) көрініс табады. Тіл арқылы «білімді» біз ұрпақтан ұрпаққа жеткіземіз. Сондықтан «тіл» – «білімді» сақтау формасы және оны тарату құралы десек те болады, ал ғылыми-техникалық мәтіннің семантикасын модельдеу дегеніміз – ол осы салалардың білім жүйесін модельдеу. Осылайша, біртіндеп келіп,

білімді модельдеу проблемасы мәтін мазмұнын модельдеумен ұштасып жатқанына көз жеткізуге болады. Бұл жерде ең негізгі шешімін табуды қажет ететін мәселе – мәтіннің семантикалық көрінісін құру» [79].

Қолданбалы лингвистикада тілдік бірліктерді модельдеуге қатысты әжептеуір тәжірибе жинақталған. Атап айтқанда, Н.Хомский, С.Маркус, А.В.Гладкий, В.Ингвре, Г.Г.Белоногов, Р.Г.Котов, Р.Г.Пиатровский, Қ.Бектаев, И.Н.Ревзин, Р.Шенк, Ю.П.Шабанов-Кушнаренко, А.А.Ляпунов, В.А.Вудс т.б. ғалымдардың зерттеулерінде.

Әсіресе, бұл мәселеде Н.Хомскийдің туындатушы грамматика теориясы ерекше орын алады. Н.Хомскийдің теориясы бойынша, кез келген сөзден белгілі бір модельдер бойынша сөзформаларды туындатуға болады. Бұл тәсіл теориялық тұрғыдан қызығушылық тудырғанмен, практикалық қолданыс үшін күрделі болды [80]. Осы сияқты Ю.П.Кушнаренконың математикалық модельдеу тәсілі де 80 жылдары белсенді қолданыста болғанымен, бұл әдіс қиындығына байланысты қазіргі кезде мүлде қолданылмайды деуге болады.

Модельдеуге қатысты Г.Г.Белоноговтың алгоритмдер теориясына негізделген морфологиялық модельдеу әдісі 1983 жылдары кеңінен қолданыс тапты. Бұл модель бойынша, морфологиялық талдау (анализ) мен синтез негіздер сөздігі мен көмекшілер кестесіне негізделіп жасалды [81].

Тілдің морфологиялық жүйесін модельдеу нәтижелерін автоматты басқару орталықтарына бағындыру, сөйтіп, оларды пайдаланудың компьютерлік бағдарламасын жасап, ғаламтор жүйесіне енгізу қажеттігі туды. Осындай тілдік бірліктерді модельдеудің нәтижесінде қолданбалы лингвистиканың бір кешенді жеке тармағы ретінде корпустық лингвистика саласы дүниеге келді. Корпустық лингвистика дербес ғылым саласы ретінде зерттеу объектісін кеңейте түсті. Ол енді морфологиялық модельдеумен ғана шектелмей, тілдік талдаулардың басқа да түрлеріне қарай модельдеудің түр-түрін ойлап тапты. Морфологиялық модельдеудің алғаш қолға алынуымен байланысты, тілдік корпустарға белгіленім қою ісі де, барлық тілдер корпустары бойынша, морфологиялық белгіленім қоюдан бастау алады.

Түркі тілдері бойынша, корпустар жасау ісі де соңғы кездері ғана қолға алынып келеді. Мәселен, башқұрт тілінде жоғарыдағы Г.Г.Белоноговтың модельдеу әдісі бойынша, автоматты түрде морфологиялық белгіленім қойылған тілдік корпустары жасалған.

Қазақ тіл білімінде тілдік корпус мәселесі 2009 жылдан бастап ғылыми зерттеу жобаларының тақырыбы ретінде зерттеліп жатыр. Бұл ретте қазіргі кезде негіз сөздер сөздігі мен сөзөзгерім сөздігін жасау арқылы мәтіндерге автоматты жолмен морфологиялық белгіленімдер қоюдың алғышарттары жасалу үстінде. Осы зерттеу жұмысымызда автоматты түрде морфологиялық белгіленім қоюға қажетті негіз сөздер сөздігін компьютер арқылы алудың практикалық әдістерін баяндауды мақсат етіп отырмыз.

Қайсібір тілде болмасын әр тілдің толық көптомдық түсіндірме сөздіктері сол тілдік сөздік қорындағы барлық сөзді қамтуды мақсат етеді. Сондықтан негіздер сөздігін жасауда ең алдымен сол тілдің түсіндірме сөздіктерінде берілген реестр сөздер алынады. Алайда түсіндірме сөздіктердің өзінде де сөздік қордағы сөздердің барлығы тұтастай қамтылмай қалуы мүмкін. Сондықтан да түсіндірме сөздіктер белгілі бір уақыт аралығында қайта жаңғыртылып отырады. Өйткені тіл үнемі дамып отыратын құбылыс болғандықтан, қоғам дамуымен бірге жаңа сөздер, әрбір ғылым, кәсіп салаларына қатысты жаңа терминдер жасалып жатады. Сол себепті кез келген тілдің сөздік қорын түсіндірме сөздіктерде берілген реестр сөздер санымен өлшеуге болмайды. Негіз сөздер сөздігін жасауға қажетті түсіндірме сөздіктердің реестрлерінен басқа да тілімізде, ауызша сөйлеуде немесе жазба мәтіндерде болсын сөздікке енбей қалған сөздер кездесуі әбден мүмкін. Тілдік корпустар осындай таңдама жазба мәтіндерді морфологиялық бөлшектерге бөліп, яғни негіз сөздер мен грамматикалық формаларға ажыратындықтан, танылмаған, яғни негіз сөздер сөздігінде кездеспеген сөздерді проблемалық аймаққа шығарады, яғни «түсініксіз» сөздер ретінде бөліп алады. Бұл ретте тілдік корпустардың түсіндірме сөздіктер үшін жаңа немесе сөздікке енбей қалған сөздерді нақты тауып берерлік мүмкіндігі зор. Бұл ғана емес, жалпы тілдік корпустардың келешек ғылыми-теориялық, сонымен қатар қолданбалы бағыттағы жұмыстардың автоматты түрде орындалуына

тигізер үлесі мол болмақ. Негіз сөздер сөздігін жасау үшін қазақ тілінің онбестомдық түсіндірме сөздігіндегі реестр сөздерді автоматты түрде бөліп алудың ғылыми-практикалық жолдарын білу қажет. Себебі түсіндірме сөздіктерден реестрді қолдап бөліп алу көп жұмысты қажет етеді. Сондықтан реестр алудың ұтымды жолы оны компьютер арқылы автоматты түрде бөлшектеу болып табылады.

Компьютер жұмыс істеу үшін оған арнайы бағдарламалар салынатыны белгілі. Негіз сөздерді бөліп алуда көп қолданылатын word және excel бағдарламалары пайдаланылды. Ең алдымен word бағдарламасында жұмыс істейтін сөздік мәтінді алып, оған қажетіне қарай бірнеше бағытта өзгертулер енгізілді.

Біріншіден, негіз сөздер сөздігі тек қана сөздік бірліктерді ғана қамтымайды. Олардың міндетті түрде сөз таптарына қатысы көрсетіліп отыруы қажет. Өйткені әрбір сөздің сөз табына қатыстылығы ол сөздердің категориялық грамматикалық мағынасы емес, белгілі ғалым С.Исаевтың сөзімен айтқанда, жалпы грамматикалық мағынасы болып табылады. Жиілік сөздіктерде де әрбір сөздің тұсына олардың сөз таптарына қатысты белгілері қойылып отырады. Сондықтан сөздердің сөз табына қатысты мағынасы грамматикалық мағына болғанмен, олар лексикалық бірліктердің қатарында қамтылады.

Сол себептен біздер реестр сөздерді алуда сөз таптарына арнайы белгі-код қоюдан бастадық. Мысалы, 15 томдық сөздікте реестр сөздердің тұсына сөз таптары қойылған, ал біз оны өзіміздің құрастырылған бағдарламамызға ыңғайлап өзгертулер жасадық. Мысалы, зат – зт, ет – ет, сын – сн, сан – са, үст – үс, елік – ел, од – од, шыл – шл, есім – ес, мод – мд деген секілді шартты белгілермен өзгертулер енгіздік. Мұндай грамматикалық шартты белгілерді тілдік корпустарда «разметка» терминімен атайды. Біз белгіленім немесе белгі-код терминдерімен баламалап отырмыз. Жоғарыдағыдай сөз таптарына қатысты өзгертулер енгізудің екінші бір себебі бұрынғы жасалған жиілік сөздіктерде берілген сөз таптары біздің шартты қысқартуларымызбен сәйкес келеді. Сондықтан кейінгі жиілік сөздіктер алу жұмысына да сөз таптарының белгіленімдерінің бірізділігі қажет. Көптомдық түсіндірме сөздік-

тегі мұндай өзгертулер автоматты түрде «ауыстыру» (заменить) «бұйрығы» бойынша іске асырылады.

Word бағдарламасында жұмыс істейтін 15 томдық сөздікте реестр сөздермен бірге олардың ұясындағы тіркестер де қарамен боялған (жирный). Ал қалғандары – түсіндірмелері мен иллюстрация қарайтылмаған. Осы ерекшелігіне қарай екінші кезекте арнайы «бұйрықпен» қарайтылған сөздер бөлініп алынады. Осы әмір орындалғаннан кейін word бағдарламасындағы сөздік мәтінде реестр сөздер мен сөз тіркестері және қарамен боялған сөздердің көпмағыналығын саралап көрсететін рим цифрлары ғана қалады. Бұдан кейін тағы да арнайы бұйрықпен рим цифрлары да алынып тасталады.

Ал осы сатыдан өткеннен сон Excel бағдарламасының бұйрықтары орындалады. Аталған бағдарлама бойынша, қарамен жазылған реестр сөздерді жинақтау жүргізіледі. Реестр сөздердің алдына бұған дейін «|» таяқша қойылған болатын. Міне, осы белгі арқылы арнайы бағдарлама құрылып, Excel-де жүзеге асырылады. Оны Excel-дің сұрыптау және сүзу (сортировка и фильтр) бұйрықтары негізінде, сонымен қатар ішкі әртүрлі кішігірім бұйрықтар арқылы іске асыруға болады. Осы белгінің («|») негізінде реестр сөздер жинақталады.

Корпусқа қажетті негіз сөздер сөз табына толығымен ажыратылуы қажет болғандықтан, сөз табы қойылмаған сөздерге арнайы сөз табына қатысты белгіленімдер қойып шығу қажет. Сондай-ақ негіз сөздер мен грамматикалық формаларды ажыратудағы проблемалар тағы бар. Өйткені 15 томдық түсіндірме сөздіктерді қимыл атау формасы (у) реестрде беріледі. Сондай-ақ шырай, оның ішінде салыстырмалы шырай тұлғалары да түсіндірме сөздікте қамтылған. Ал біздің жобамызда салыстырмалы шырай тұлғалары мен қимыл атаулары сөзформалар сөздігіне қосылып отыр. Осы сияқты түсіндірме сөздік негізінде алынған реестрлерді тілдік корпустар жасауға қажетті негіз сөздер сөздігі үшін өзіміздің ұстанымызға қарай өңдеу қажет. Сондай-ақ негіз сөздер мен грамматикалық формалардың аражігін ажыратуда да көптеген қиындықтар бар.

4.2. Сөзформалар (түрленім) сөздігі – морфологиялық белгіленім қоюдың құралы

Қазақ тіл білімінде жекелеген сөздердің аударма, түсіндірме, терминологиялық т.б. сөздіктері үздіксіз жарық көріп жатқаны белгілі. Өйткені тілдің сөздік қоры осындай лексикалық бірліктердің санымен өлшенеді. Алайда тіліміздің байлығы тек түбір сөздермен ғана емес, олардың сан мыңдаған түрленісінен де көрінеді. Сөздер жеке-жеке тұрғанда сөйлеу бірлігі болып танылмайды. Тіліміздегі сөздер қолданыста ғана өз мүмкіндіктерін толық ашып бере алады. Көпмағыналылық құбылысы да сөздердің басқа сөздермен байланысқа түсуі кезінде айқын көрінеді. Сондықтан сөздердің түрленісі тіл табиғатының бір көрінісі болып табылады.

Түркі тілдерінде мәтінде сөздер түбір формадан гөрі түрленген формада жиі кездеседі. Қазақ тілінде белгілі ғалым Қ.Бектаев сөздер мәтінде түрленген формада жиі қолданылатынын айтады [82]. Сол сияқты башқұрт тілінің зерттеушісі З.Сиразитдиновтың көрсетуінше, башқұрт тілінің әртүрлі стилінен алынған 600 000 сөзқолданыстан тұратын мәтінде зат есімдер мен етістіктер бірге бүкіл мәтіннің 58,69 пайызын құрайды екен [83]. Оның ішінде түбір тұлғадағы зат есімдер барлық зат есімдердің 35,5 пайызын, ал етістіктер 3,73 пайызын ғана құрайды. Сонда олардың қалған пайызы зат есімдер мен етістіктердің түрленген тұлғалары болып табылады. Сөздердің түрлену мүмкіндіктерін анықтау, бұл ретте оларды тізімдеп, сөзформалар сөздігін құрастыру да бүгінгі күн тәртібінде тұрған өзекті мәселе.

Орыс тіл білімінде мұндай сөзформалар сөздігі қатарында А.А.Зализняктың грамматикалық сөздігін атап айтуға болады [84].

Сөз таптарының ішінде түрленуге бейім сөз табы – зат есім мен етістіктер болып табылады. Ғалымдар сөздердің түрлену жүйесін есепке алғанда осы екі сөз табын негізге алып отырады. Айталық, татар тілінің зерттеушісі Х.Ф.Исхакова сөздердің түрлену мүмкіндігіне қатысты мынадай фактілерді келтіреді. Бір етістік түбірден, барлық етіс формаларын, оның түрленуін қосып есептегенде татар тілінде 17947 форма, түрік тілінде 11390 форма, өзбек тілінде 13592 форма өрбитінін айтады [85]. Қазақ тілінде профессор С.Мырза-

беков бір етістік сөзден 1150, ал егер бұлардан өзге варианттар, тағы да басқа форма жасауға қатысатын демеулік шылаулар (барған-дай-ақ, бар-а-тұғын-мын) санын қосқанда жобамен олардың санының 1500 ден асатынын болжамдайды [86]. А.Жаңабекованың зерттеуі бойынша, бір етістік сөзден өрбитін формалар санының 15 000-нан асып кететіндігі туралы болжам жасалады [87].

Қазіргі тіл сұранысы тұрғысынан үлкен көлемді сөзформалар сөздігі қазақ тіл білімінде әлі күнге дейін жарық көрген жоқ. Алайда жоғарыда сөз болғандай бір сөзден қанша форма өрбиді деген мәселе ғалымдар назарынан тыс қалмаған. Бұл орайда лингвостатист мамандардың бірі профессор А.Жұбанов еңбектерінде де бірнеше кестеге түсірілген сөзформалар тізімі зат есім мен етістік сөз табына қатысты жасалған. Қосымшалардың сөздердің соңғы дыбысының фонетикалық өзгерісіне қарай жалғанатын варианттары да кестеде берілген [88].

Автоматты морфологиялық талдау жүйесін жасау үшін сөздердің тілдік жүйедегі морфологиялық құрылысы анық болу қажет. Тілдік жүйеде сөздер ең алдымен түбір және қосымша морфемаларға бөлінетіні сияқты сөзді автоматты түрде морфологиялық бөліктерге бөлшектегенде компьютер жадына сөздердің осы екі бөлігінің тізімі салыну қажет. Яғни автомат осы екі морфологиялық жіктемеге сүйеніп жұмыс істейтін болады. Олар: біріншіден, негіз сөздер тізімі, біз оны негіз сөздер сөздігі деп атап отырмыз, екіншісі осы негіз сөздерден тарайтын, өрбитін грамматикалық формалар, яғни грамматикалық қосымшалар сөздігі немесе түрленім сөздігі болып табылады.

Түркі тілдері аглютинативтік тіл болғандықтан, сөз түрлендіруші формаларының негіз сөздерге жалғану реті айқын, шегі тұрақты болып келеді. Сондықтан олар бір ғана сөздің ғана емес, егер зат есім сөз болса, тілдегі барлық зат есімдердің түрлену жүйесі, етістік болса, бүкіл етістік сөздердің түрлену жүйесі болып табылады. Түркі тілдерінде мұндай түрлену жүйесі бар, сөзтүрленім мүмкіндігі үлкен сөз таптары – етістік пен зат есім. Бұл екі сөз табының грамматикалық категория құрау мүмкіндіктерімен де байланысты. Тілдік жүйеде грамматикалық категориялар осындай парадигмалық сипаттағы түрлену жүйесі бар ортақ грамматика-

лық мағынаға біріккен топтық бірліктерден құралады. Қазақ тілінде де сөз таптарының ішінде етістік грамматикалық категорияға бай сөз таптарының бірі. Зат есім екінші орында, ал одан кейінгі түрлену жүйесі бар сөз табы сын есім, сын есімнің шырай категориясы болып табылады. Алайда сөзтүрленім сөздігіне шырайдың синтетикалық формалары ғана енеді. Ал күшейтпелі шырай тұлғалы сөздер негіз сөздер тізімінде беріледі. Сондықтан түрленім сөздігінде тізімделетін грамматикалық формалар негізінен зат есім мен етістіктің түрленімі болып келеді. Бұл туралы башқұрт тілінің зерттеушісі З.Сиразитдинов былай деп тұжырымдайды: «В системах автоматического анализа, орфографической проверки текстов башкирского языка невозможно обойтись лишь словарем слов в основной форме. Для создания эффективных лингвистических систем необходим максимально полный список всех возможных форм словоизменений. Единственным выходом из данной ситуации является построение алгоритмической модели образования текстовой производной словоформы из корня и основы. И в первую очередь такая модель должна быть построена для существительных и глаголов» [83]. Қазақ тілінде осы мәселемен айналысқан ғалымдар еңбектерінің барлығында да түрленім сөздігі осы зат есім мен етістік сөз табына негізделіп жасалған. Алайда бұл басқа сөз таптарына бұл қосымшалар жалғанбайды деген сөз емес. Зат есімнің түрленім формалары сын есімдерге, сан есімдерге, есімдіктерге, тіпті түрленбейтін сөз табы болып табылатын үстеу сөздерге де жалғанып жұмсала береді. Бұл жағдайда зат есімнің түрленім қосымшалары басқа сөз таптарын заттандырып жібереді. Яғни сын есімге зат есімнің қосымшалары жалғанғанда ол заттық мәнде жұмсалып, субстантивтенеді. Мысалы, *Жақсымен дос бол, жақсыға айтқан сөз* т.б. қолданыстардағы «жақсы» сөзі көмектес және барыс септік формаларымен түрленіп, *жақсы адам* деген мағынада заттанып қолданылып тұр. Демек, түрленім сөздігі үшін ең қажеттісі осы аталған екі сөз табының грамматикалық қосымшалары болып табылады. Енді А.Жұбановтың еңбегіндегі осы екі сөз табынан тарайтын формалар кестесінен үзінді бергенді жөн көріп отырмыз [88].

33-кесте

Бірнеше фонетикалық вариантта жалғанатын зат есімдердің түрленім кестесінің алғашқы бетінен үзінді

п/п	1-тип: (у, и дыбысынан басқа) дауысты дыбысқа біткен	2-тип: у, и, й, р	3-тип: м, н, ң	4-тип: б, в, г, д, п, т, к, ф, ш, с, қ, х, ч, ц, h	5-тип: ж, з , л
	Әке, келі, мәсі, мекеме	Үй, белбеу, темір, сәби	Кілем, кескін, өлең	Шөп, сүт, күрек, жүріс	Тәж, жез, киіз, көз, ел
1	ге	ге	ге	ке	ге
2	де	де	де	те	де
3	ден	ден	нен	тен	ден
4	лер	лер	дер	тер	дер
5	лерге	лерге	дерге	терге	дерге
6	лерде	лерде	дерде	терде	дерде
7	лерден	лерден	дерден	терден	дерден
8	лерді	лерді	дерді	терді	дерді
9	лердің	лердің	дердің	тердің	дердің
10	лермен	лермен	дермен	термен	дермен
11	лерменен	лерменен	дерменен	терменен	дерменен
12	лері	лері	дері	тері	дері
13	лерім	лерім	дерім	терім	дерім
14	лерімде	лерімде	дерімде	терімде	дерімде
15	лерімді	лерімді	дерімді	терімді	дерімді
16	леріме	леріме	деріме	теріме	деріме
17	лерімен	лерімен	дерімен	терімен	дерімен
18	леріменен	леріменен	деріменен	теріменен	деріменен
19	леріммен	леріммен	деріммен	теріммен	деріммен
20	лерімменен	лерімменен	дерімменен	терімменен	дерімменен

Осы сияқты етістіктің де түрленімі ғалым еңбегінде арнайы кестеде беріледі.

Сөздер мәтінде, яғни сөйлеуде қолданылғанда қаншама түрлі формаларға ие болып, құбылып отырады. Сондықтан компьютер жадына алдын ала дайындалған түрленім тізімін салғанмен, ол мәтін бойында басқа да сан алуан тілдік қолданыстарды табуы мүмкін. Негіз сөздер мен түрленім сөздігінде жоқ сөзформаларды компьютер арнайы бағдарлама арқылы проблемалық аймаққа шығарып отырады. Бұл формалар өңдеу барысында тағы да алдын ала жадқа енгізілетін түрленім формаларының тізіміне немесе жаңа сөз болса негіз сөздер сөздігінде тізімге қосылады. Осылайша бұл процесс үнемі жаңартылып, өңделіп отырады. Компьютер жадына енгізілген мұндай ақпараттардың компьютерлік бағдарламасы да осы процеске сәйкес қайта өңдеу мүмкіндігін ескеріп жасалады.

Жоғарыда сөзді автоматты бөліктеудің механизміне қатысты мәселелер сөз болды. Қазіргі кезде жасалып жатқан қазақ тілінің корпустары сөздерді автоматты түрде морфологиялық бөліктерге бөлшектеумен ғана шектелмейді. Ол морфологиялық жүйенің кодталған талданымын беруді де мақсат етеді. Яғни әрбір сөзформаның сөз табына қатысын, сол сияқты оның морфемалық бөліктері мен олардың әрқайсысының грамматикалық белгісін түсіндіріп тұратын болады. Мұны корпустық лингвистикада морфологиялық белгіленім (разметка) деп атайды.

4.3. Тілдің морфологиялық жүйесін автоматты түрде бөлшектеуге қажетті негіз сөздер сөздігін жасаудың лингвистикалық проблемалары

Қазіргі жаһандану тұсында ғылым салаларының барлығында дерлік жаңа технологияны пайдалану арқылы тез әрі сапалы түрде ғылыми нәтижелер алу қажеттігі туып отыр. Тіл ғылымы саласында да қазіргі кезде ғылыми-зерттеу жұмыстарын автоматтандыру ісіне үлкен мән берілуде. Осы орайда тіл ғылымын одан әрі дамыту үшін оны зерттеудің жаңа технологияларын қарастыру қажеттігі туды. Бүгінгі тіл зерттеушісі үшін ғылыми-зерттеу жұмыстарын қысқа уақыт ішінде орындап шығуға мүмкіндік беретін дайын ақпараттық құрал қызметін атқаратын тілдік корпустар жасаудың маңыздылығы дау тудырмайды.

Мұндай аса ауқымды тілдік ақпараттармен қамтамасыз етілген корпустар жасау жұмысының барысы бірнеше сатыдан тұрады. Компьютер жадына қаншалықты ақпарат енгізілсе, компьютердің жадында арнайы компьютерлік бағдарлама бойынша өңдеп шығарып беретін ақпарат та соншалық мол болады. Сондықтан корпус құрастырушылар компьютер жадына енгізілетін мәліметтерді өз сұраныстарына қарай икемдейді. Демек, компьютер әрекеті адамның саналы ой-әрекетінің нәтижесі болып табылады.

Компьютерге арнайы тапсырма беру үшін, компьютерге сол тапсырманы орындауға қажетті шарттар толық қамтылуы қажет. Компьютер арнайы программалар арқылы жадына енгізілген формулалар, модельдер, шарттар бойынша жұмыс істейді. Сондықтан тілдік корпустар құрастыруда да тілдік материалдар арнайы модельдер бойынша бейнеленеді.

Корпусқа енгізілетін лингвистикалық ақпараттар тілдік жүйедегі тұрақты деңгейлер бірліктеріне негізделеді, яғни жүйелі тілдік қағидаттарға бағынады. Корпус құрастырушылардың компьютер жадына енгізілетін шарттары да осы тіл болмысынан туындайды. Сондықтан тілдік корпустардың барлық әрекеті лингвистикалық талаптарға байланысты болады.

Корпустан сұралатын лингвистикалық ақпараттың бірі – тілдің морфологиялық құрылысына қатысты мәліметтер. Мұндай мәліметтер алу үшін корпус жадына осы морфологиялық талдаулар туралы ақпараттар енгізіледі. Олар белгілі бір модель, формулаға түсіріледі. Корпустық лингвистикада мұндай ақпараттандырылған модельдерді (белгі-код) «разметка» деп атайды. Тіл деңгейлеріне қарай мұндай «белгіленімдер» түр-түрге бөлінеді. Морфологиялық деңгей бірліктерін таңбалаушы – морфологиялық белгіленім, синтаксис бірліктерін таңбалаушы – синтаксистік белгіленім т.б. аталады. Демек, корпусқа енгізілетін белгіленімдер де тілдік жүйеге ауысады.

Тілдің морфологиялық жүйесін автоматты түрде бөлшектеу компьютер жадында бірнеше сатылы шарттардың орындалуы арқылы іске асады. Олар:

1) Сөздердің морфологиялық құрылымы екі бөліктен тұрады: біреуі – түбір сөздер және оған жалғанатын қосымшалар. Сол

сияқты түбір сөздер де екіге бөлінеді: негізгі түбір және туынды түбірлер. Олардың екеуі біріктіріліп, негіз сөздер деп аталады. Компьютер ең алдымен мәтінде кездескен жекелеген сөздерді морфологиялық тұрғыдан бөлшектегенде негіз сөздерді ажыратып алуы қажет. Ол үшін компьютер жадына негіз сөздердің тізімі берілуі керек.

2) Екіншіден, сөздердің морфологиялық құрылымының екінші бөлігі – оған жалғанатын қосымшалар. Компьютер жадына қосымшаларды бөлшектеп, олардың тілдік жүйедегі граммтикалық мағынасын білдіретін шартты белгіленімдер енгізілуі қажет. Жадқа осылай алдын ала дайындалған қосымшалардың тізімі мен шартты белгілері салынады.

3) Осы жұмыстармен қатар корпустардың қай түрі болмасын мәтіндер жиынтығынан тұратындықтан, тілдің әртүрлі стилін қамтитын электронды пішіндегі мәтіндер енгізіледі. Өйткені жоғарыдағы компьютер жадына енгізілген шарттарға сәйкес программа жүзеге асырылуы үшін сөздердің морфологиялық құрамын талдауға қажетті мәтіндер қажет. Жалпы корпустардың қуаттылығы да осы үлкен массивтерді қамтумен де байланысты болып келеді. Орыс және шетел тіл білімінде жасалған корпустар саны миллиондаған сөзқолданысты қамтиды. Көлемді массивті корпустар сөздердің морфологиялық құрамын бөлшектеу жұмысын компьютер көмегімен автоматты түрде атқарады. Ал мұндай программалар жұмысы үшін компьютер жадына жоғарыда аталған негіз сөздер тізімі мен олардың грамматикалық түрленісі болып табылатын қосымша морфемалардың кодталған тізімі енгізілуі қажет.

Біздің бұдан бұрынғы тақырыптарымызда тілдің сөздік құрамын құрайтын негіз сөздер сөздігін жасаудың, яғни тілдегі сөздерді тізімдеп, түсіндіретін түсіндірме сөздік реестрлерін автоматты түрде теріп алудың механизмі қарастырылды. Бұл тақырыпта соның жалғасы ретінде сөз таптарына ажыратылған негіз сөздерді тізімдеудегі проблемалық мәселелер туралы сөз болмақ.

Тілдердің барлығында негіз сөздер сол тілдің әртүрлі сөздіктерінде тізімделіп беріледі. Сөздіктердің ішінде сөзді қамтуы жағынан түсіндірме сөздіктердің мүмкіндігі үлкен болып келеді.

Сондықтан кез келген тілде жинақталған сөздік қор түсіндірме сөздіктерде көрініс табады. Осымен байланысты корпусқа морфологиялық белгіленім енгізуде сол тілдің түсіндірме сөздіктерінде берілген реестр сөздер негізге алынады. Қазақ тілінде үлкен көптомдық түсіндірме сөздік: онтомдық және соңғы жарық көрген онбестомдық түсіндірме сөздіктер болып табылады. Осы орайда корпусқа қажетті морфологиялық белгіленім енгізу үшін арнайы компьютерлік бағдарламалар көмегімен жүз мыңға жуық сөзді қамтыған онбестомдық «Қазақ әдеби тілінің сөздігінен» реестр сөздерді бөліп алу жұмысы орындалды [89].

Негіз сөздер сөздігінің жұмысы 15 томдық түсіндірме сөздіктің реестрлері автоматты жолмен алынғаннан кейін де жалғасады. Олар:

а) Корпусқа қажетті негіз сөздер сөз табына толығымен ажыратылуы қажет болғандықтан, сөз табы қойылмаған сөздерге арнайы сөз табына қатысты белгіленімдер қойып шығу қажет. Яғни реестр сөздерді өңдеу жұмысы жүргізіледі. Мысалы, 15 томдық сөздікте еліктеуіш сөздердің бірсыпырасы жеке реестр ретінде мағыналары ашылып беріледі де, кейбіреулері қос нүкте қойылып, олардың етістіктерімен тіркесі бірден беріліп кетеді. (Жарқ: жарқ етті.). Ал негіз сөздер сөздігінде сөздер сөз табына ажыратылып тұруы қажет. Сондықтан еліктеуіш сөздердің барлығында қос нүктелер алынып тасталып, еліктеуіш (ел. немесе елік) деген белгіленім қойылды. Осы сияқты түсіндірме сөздікте мағынасы тіркес құрамында ғана ашылған біраз сөздер кездеседі. Олардың барлығы мақсатқа сәйкес, мүмкіндігінше белгілі бір сөз табына телініп берілді.

ә) Реестр сөздерді тізімдеу, өңдеу барысында оларға сөз таптарын қоюда бірлі-жарым қате кеткен тұстар кездеседі. Мұндай қате көбінесе сын есімнің де зат есімнің де қосымшасы болып табылатын *-лық/лік* қосымшалы сөздерде жиі кездеседі. Мысалы: *реттілік, түрпілік, өршімелік* сөздеріне *сын.* деген белгіленім қойылады. Алайда бұл сөздерге берілген дефиниция заттық мағына болып ашылады, сонымен қатар иллюстрациялық материалдар да зат есім сөзді дәйектейді. Сол сияқты *типологиялық* сөзіне *зат* деген белгіленім қойылады. Бұл сөздің еш мысалсыз-ақ сын есім екендігі

белгілі. Осы сипаттас біршама сөздер өңдеу барысында түзетіліп отырды.

б) Негіз сөздерді өңдеуде атқарылатын тағы бір әрекет – реестр сөздер тізіміндегі бір сөз табына қатысты омоним сөздердің бірін ғана қалдыру. Себебі компьютер бағдарламасы мұндай бір сөз табына қатысты омоним сөздерді олардың мағыналық ерекшелігіне қарамай, бірдей таңбалайды.

в) Реестрді өңдеудің келесі бір сатысы – сөзжасамдық жұрнақтар мен форма тудырушы жұрнақтардың аражігін айырып алу болып табылады. Түсіндірме сөздіктер сол тілдің сөз байлығы болғандықтан, олар негізінен тілдегі негізгі және туынды түбір сөздерді ғана қамтуы тиіс. Ал сөздерге грамматикалық мағына үстейтін қосымшалар түсіндірме сөздіктердің сөзтізбесіне енбеуі қажет. Алайда қазақ тілінің түсіндірме сөздігінде кейбір грамматикалық тұлғалы сөздер сөзтізбеден орын алады. Айталық, қимыл атау тұлғалары, салыстырмалы шырай тұлғалары. Бұл грамматикалық формалар түсіндірме сөздіктерде де шартты түрде алынып отыр. Корпусқа морфологиялық белгіленім енгізу үшін құрастырушылар сөздердің морфологиялық шегін компьютер жадына нақты айырып көрсету қажет. Егер де түсіндірме сөздік реестрінде берілгендіктен, қимыл атау тұлғалы етістік сөздерді негіздер қатарында қалдырсақ, онда компьютер корпусқа енгізілген мәтіндердегі сөзформаларды морфологиялық бөліктерге бөлшектегенде, ***у*** тұлғалы етістіктердің барлығын түбір сөз деп көрсетеді де, бұл қосымша сөзжасамдық жұрнақ ретінде танылады. Алайда ***у*** қимыл атау формасының грамматикалық бірлік екендігі белгілі. Сондықтан біздің өңдеу жұмысымызда қимыл атау тұлғалы етістіктердің барлығы дерлік сөзтізбеден алынып тасталды. Мұны компьютер автоматты түрде атқарады. Алайда тағы бір ескеретін мәселе кейбір қимыл атау формасындағы етістіктердің заттық мәнді иеленуі. Мысалы, *қамсыздандыру, ақпараттандыру, әртараптандыру, хабарландыру* сияқты сөздер. Компьютер мұндай сөздерді оларға қойылған *зат есім* деген белгіленіміне қарай ажыратады да, оларды сөзтізбеде қалдырады. Сол сияқты түсіндірме сөздікте күшейтпелі шырайда тұрған сын есім сөздер түгелге жуық беріледі. Мысалы: *қып-қызыл, жап-жақын, ап-алыс* т.б. Біздің

негіз сөздер тізімінде де бұл сөздер негіздер қатарына енгізілді. Өйткені сөздерді автоматты түрде бөлшектегенде күшейтпелі шырайда тұрған сөздерді компьютер арқылы ажырату қиынға соғады. Ал салыстырмалы шырай тұлғалы сөздер түсіндірме сөздікте біршама қамтылған. Алайда біздің негіз сөздер тізімінен *-рақ/рек* пен *-лау/леу* тұлғалы салыстырмалы шырай тұлғалы сын есімдер алынып тасталды. Өйткені олар сын есімнің шырай категориясын құрайтын категориялық мағынаға ие грамматикалық бірліктер болып табылады. Сондықтан бұл формалар сөздердің түрленісіне арналған грамматикалық формалардың тізіміне кетеді, грамматикалық мағына ретінде кодталады.

Онбестомдық түсіндірме сөздікте етіс тұлғалы етістіктер де сөзтізбеде беріліп, мағыналары ашылады. Етіс жұрнақтары тілдік жүйеде етістіктің грамматикалық категориясы бола тұра, онтомдыққа да, онбестомдыққа да енгізілген. Өйткені тілімізде етіс тұлғалы етістіктер өте жиі кездеседі. Етістіктің статистикасын зерттеген ғалым, профессор С.Мырзабековтің көрсетуінше, онтомдық сөздіктің жартысынан көбі тек етіс тұлғалы етістіктерден тұрады [90]. Сондықтан мұндай көлемдегі сөздік бірліктерді сөзтізбеден алып тастау үлкен ойсыраушылық туғызатыны анық. Екіншіден, жалпы түркологиялық еңбектердің көпшілігінде, соның ішінде етіс қосымшалары сөзжасамдық жұрнақтар ретінде танылады. Осы критерийлер етіс тұлғалы етістіктерді түсіндірме сөздіктердің сөзтізбесіне енгізілуіне негіз болған. Алайда етіс қосымшаларын таза тілдік теория тұрғысынан сөзжасамдық жұрнақ па, әлде форма тудырушы жұрнақ па деген мәселелер әлі де дау туғызып келе жатыр. Біздің негіз сөздер тізіміне етіс тұлғалы сөздердің барлығы енгізілді. (Яғни *азайт, азайтқыз, азайттыр* сияқты етіс тұлғалы етістіктердің барлығы негіздер тізімінде берілді). Өйткені етіс қосымшаларының таза сөзжасайтын қызметі де баршылық. Ал бұл олардың морфологиялық аражігін ажыратуда көп қиындық келтіреді. Сол себепті морфологиялық белгіленім енгізудің алғашқы сатысында әзірге етіс тұлғаларын негіздер қатарына жіберуді жөн көріп отырмыз. Тілдік корпустар құрастыруда оларға келешек түзету енгізу мүмкіндігі ескерілетіндіктен, етіс мәселесі кейінге қалдырылды.

Негіздер тізімін алуда жоғарыда аталғандардан басқа да проблемалық мәселелер бар. Олар негізінен тілдік жүйенің өзіндегі аралық қабаттардан туындайды, яғни кейбір тілдік бірліктердің көпқызметтілігіне байланысты болып келеді. Айталық, *-дай/дей* қосымшалы сөздердің сөз түрлендіруші жұрнақтардан кейін жалғануы, сол сияқты *-сыз/сіз* тұлғалы сөздердің сөйлеу бірлігі ретінде қолданылуы, *-лық/лік* жұрнағының есімше тұлғаларынан кейін жалғануы т.б. Бұлардың барлығы тілдік жүйеде сөзжасамдық қызмет атқарғанмен, қолданыста сөйлеу бірлігі болатындығын ешкім жоққа шығармайды. Башқұрт тілінің зерттеушісі, компьютерлік лингвистика маманы З.Сиразитдинов бұл қосымшаларды сөзжасамдық жұрнақтар деген атаумен грамматикалық сөздікке енетіндігін айтады [83]. Демек, З.Сиразитдиновтың пікірінше де, мұндай көпфункциялы қосымшалар, атауына қарамастан, грамматикалық тұлға болып табылады. Алдағы тақырыптарымызда осындай аралық қабаттағы тілдік бірліктерді біз жасап отырған негіз сөздер сөздігі мен грамматикалық формалар (сөзформалар) сөздігінде қалай ажыратқанымыз туралы сөз ететін боламыз. Өйткені тілдік жүйедегі мұндай жайттардың табиғатын толық ажыратып алмай, сөздердің морфологиялық құрылымын автоматты түрде бөлшектей алмаймыз.

4.4. Тілдің морфологиялық жүйесін модельдеудегі функционалды қосымшалардың проблемалық мәселелері

Корпустар құрылымы екі бөліктен тұрады деуге болады. Оның біріншісі компьютерлік бағдарламаға салынған лингвистикалық кодталған модельдер болса, екіншісі сол бағдарламалардың жұмыс істеу мүмкіндіктерін ашатын, бағдарлама жұмысына арналған мәтіндер болып табылады. Соған сәйкес корпус құрастыру жұмысы да осы екі бағытқа сәйкес жүргізіледі.

Сөздердің морфологиялық құрамы анық, айқын болғандықтан, морфологиялық белгіленім мәселесінде корпуспен айналысқан тілдер едәуір нәтижеге қол жеткізген. Нәтижесінде көлемді таңдама мәтіндер бойынша сөздердің морфологиялық құрылымына қатысты ғылыми-зерттеу жұмыстарына қажетті қаншама

мәлімет алып отыр. Қазақ тіл білімінде де осы әлемдік тіл білімінің нәтижелерін қолданысқа енгізу мәселесі бүгінгі күн тәртібінде тұрған өзекті мәселе.

Морфологиялық тұрғыдан кодталған шартты белгілер моделін жасау сөздердің морфологиялық құрамын ажыратып, айырудан басталады. Тілдік жүйеде сөздердің морфологиялық құрамы түбір сөздер мен олардан өрбитін қосымшалардан тұрады. Соған сәйкес сөздердің морфологиялық құрамын автоматты бөлшектегенде сөздердің осы екі құрам бөліктерінің аражігін нақты ажыратып алу қажет. Бұл ретте компьютер жадына сөздік бірліктер мен қосымша бірліктердің тізімі салынады. Біздің бұған дейінгі жазған тақырыбымызда сөздердің тізімін түсіндірме сөздік реестрлерінен алудың технологиясы мен сол тізім негізінде жасалатын негіз сөздер сөздігі туралы мәселені сөз еткен болатынбыз [91].

Қазіргі кезде жасалып жатқан қазақ тілінің корпустары сөздерді автоматты түрде морфологиялық бөліктерге бөлшектеумен ғана шектелмейді. Ол морфологиялық жүйенің кодталған талданымын беруді де мақсат етеді. Яғни әрбір сөзформаның сөз табына қатысын, сол сияқты оның морфемалық бөліктері мен олардың әрқайсысының грамматикалық белгісін түсіндіріп тұратын болады.

Сөздер формальдық тұрғыдан негізгі форма және туынды форма түрінде тұрады. Негізгі формалар сөздің лексикалық қорына кіретін негізгі түбір және туынды түбір сөздер. Ал туынды формалар дегеніміз – сол негізгі формадан тарайтын сөздің түрленімдері. Өйткені автомат сөздердің барлығын форма түрінде ғана таниды. Сөздік бірліктердің барлығын негізгі форма деп, олардың түрленім тұлғаларын туынды форма деп ұғынады. Негізгі форма және туынды форма термині қазақ тіл білімінде лингвостатистика саласын зерттеуші ғалымдардың бірі профессор С.Мырзабековтің зерттеу жұмыстарында қолданылады [86].

Түркі тілдерінде осы негіз формалар мен туынды формаларды ажыратуда дау туғызып жүрген мәселе – сөзтудырушы қосымшалар мен сөз түрлендіруші немесе форма тудырушы қосымшалардың аражігін айыруға байланысты қайшылықтар. Сөздердің морфологиялық құрамын автоматты бөлшектеу мәселесінде де осы қайшылықтың шешімін табу аса қажет. Өйткені тілдік жүйе-

де шешімін таппаған тілдік құбылыстар оларды компьютер арқылы автоматты бөлшектеу ісінде де алдымыздан шығып отырады. Компьютер адамның ми қызметі арқылы салынған бағдарламаға сүйенеді, яғни адам тілдік бірліктерді қалай кодтайды, компьютерлік бағдарламаны жүзеге асыратын аймақты (мәтінді) сол кодтар бойынша бөлшектеп, талдап шығады.

Сөздің морфологиялық құрамын бөлшектеу үшін сөзжасамдық қосымшалар мен форма тудырушы қосымшалардың аражігін нақты ажыратып алу қажет. Түркі тілдері жалғамалы тілдерге жататындықтан, сөздің құрам бөліктері флективті тілдерге қарағанда анағұрлым айқын. Солай бола тұра, түркі тілдерінің барлығына ортақ сипат – кейбір қосымшалардың сөзжасамдық және сөзтүрлендірушілік қызметтерінің айқын ажыратылмауы.

Сөздердің морфологиялық құрамын автоматты бөлшектеуге қатысты қиындық тудыратын мәселе түркі тілдерінің барлығында бір шешімін таппаған аралық қабатта тұрған бір топ қосымшалардың (*-дай/дей, -сыз/сіз, -лық/лік, -дағы/дегі, -ша/ше* т.б.) тілдік табиғатынан туындайды. Олар қазақ тіл білімінде Қ.Шаяхметовтің зерттеуінде екі функциялы, қос функциялы қосымшалар деген атаумен белгілі [92]. Татар тілінде мұндай қосымшаларды В.М.Насилов «кірістіруші аффикстер» (аффиксы включения) деп атаса [93], олардың бірсыпырасын татар ғалымы Ф.Ганиев септіктер парадигмасына қосады [94]. Бұлардан басқаша түркі тілдері бойынша оларды «промежуточные аффиксы», «көпфункциялы аффикстер», «функционалды жұрнақтар» т.б. атаулармен атайды.

Аралық қабатта тұрған аталмыш қосымшалар сөздердің морфологиялық құрамын автоматты бөлшектеу кезінде қиындық тудырмау үшін олардың сөзжасамдық және сөзтүрлендірушілік немесе форма жасамдық қызметтерін нақты ажыратып алу қажет. Себебі автоматты бөлшектеуде сөздер осы екі дереккөз бойынша ажыратылады. Бірі – негіз сөздер, екіншісі – сол сөздерден өрбитін грамматикалық формалар, яғни сөз түрлендіруші қосымшалар. Автоматты жіктеме үшін компьютер жадына салынған сөздің осы екі құрам бөліктерінің шекарасы айқын болуы керек. Өйткені компьютер тек қана формалдық таңбаға ғана сүйенеді. Бұл ретте аталған «қос функциялы» қосымшалар арқылы жасалған дербес

лексикалық бірліктерді бір жағынан негіз сөздер тізімінде көрсетіп, сол сияқты олардың сөз түрлендіруші жұрнақтардан кейін жалғанғандағы позицияларын олардың тілдік табиғатына, мазмұнына сәйкес атау беріп, шартты белгілер ойлап табуымыз қажет.

Тілімізде осы аталған қосымшалар арқылы жасалған және дербес лексикалық бірлік болғандықтан, түсіндірме сөздіктерде реестр сөздер ретінде берілетін, біздің зерттеу нысанымыз бойынша негіз сөздер тізіміне енетін сөздер де бар, сол сияқты олардың зат есімнің немесе етістіктің сөз түрлендіруші жұрнақтарынан кейін келіп, форма тудырушы қызмет атқаратын кездері де бар. Мәселен, *-лық/лік* жұрнағы зат есім түбірлеріне жалғанып бірде зат есім тудырушы, бірде өлшемдік, қатыстық мән үстеп, сын есім тудырушы қызмет атқарады (адамдық, жақсылық, типографиялық, сындық т.б.). Сол сияқты бұл жұрнақ есімше тұлғаларынан кейін де жалғанып жұмсалады. Мысалы: *сөйлегендік, сатқандық, жазғандық, айтарлық, берерлік* т.б қолданыстарда сөйлеу бірлігі болып қолданылады. Өйткені бұл сөздер дербес лексикалық бірлік ретінде түсіндірме сөздіктердің реестрлік қатарына енбейді. Бұл олардың әмбебаптық сипатымен де байланысты. Яғни кез келген есімше формалы етістікке *-лық/лік* қосымшасын жалғап уақытша субстантивтерді туындата беруге болады. Статистикалық мәліметтер бойынша, басқа қос функциялы қосымшаларға қарағанда *-лық/лік* қосымшасының етістік пен зат есімдерге жалғанып, сөз түрлендіруші/форма тудырушы қызмет атқаруы өте жиі кездеседі. Есімше тұлғаларынан кейін жалғанып жұмсалуына қатысты кезінде түркітанушы ғалым Н.А.Баскаков оны субстантив форма деп атаған болатын [95]. Біз сөздерді автоматты бөлшектеу мен оларды кодтауға қатысты жұмысымызда бұл қосымшаның негіз сөздер сөздігіне енбейтін, сөз түрлендіруші қосымшалардан кейін жалғанатын позицияларында оны осы өз мағыналық реңкіне сәйкес арнайы терминмен атауды жөн көріп отырмыз.

Қазақ тіл білімінде осы даулы қосымшалардың тілдік табиғатын сол кездің өзінде-ақ терең тани білген ғалым Ы.Маманов Н.А.Баскаковтың пікіріне сәйкес *-лық/лік* қосымшасының зат есімдерге жалғанып, арнаулық, өлшемдік мағына білдіретін жағдайында (*көйлектік мата, терезелік ағаш* т.б.) зат есімнің атри-

бут анықтауыш формасы деп атайды. Ал есімше (-ған/ген), қимыл атауларынан кейін жалғанғандағы жағдайын субстантив форма дейді. Сөздің морфологиялық құрамын автоматты бөлшектегенде *-лық/лік* қосымшасы үш түрлі позицияда тұрады. Біріншісі негіз сөздер тізіміне енген лексикалық бірліктер құрамында тұрады, бұл жағдайда компьютер тізімде бар *-лық/лік* формалы сөздерді дербес лексикалық единица ретінде танып, негіз сөздер орналасқан аймаққа қарай жібереді. Екінші жағдайда тізімде жоқ *-лық/лік* формалы сөздердің түбірлерін негіз сөздерге жібереді де, *-лық/лік* қосымшасының зат есім түбірлеріне жалғанған кездегі қолданысын арнаулық, өлшемдік мәнді білдіретін атрибут анықтауыш форма деп таниды. Үшінші жағдайдағы *-ған/ген* есімше формаларынан кейін жалғанатын күрделенген формадағы қолданысында (-ғандық/гендік) оларды етістіктерді заттандырушы субстантив форма деген таңбамен белгілейді. *-Лық/лік* қосымшасы *-ар/ер* есімше формасынан кейін жалғанғанда (барарлық, көрерлік, айтарлық) зат есімдерді субстантивтендірмей, керісінше зат есім түбірлеріне жалғанғандай арнаулық, өлшемдік мағына туғызып, атрибут анықтауыш форма туғызады.

Сол сияқты *-сыз/сіз* қосымшасының тұйық етістіктен кейін жалғанып жұмсалатын немесе зат есім түбірлеріне жалғанған жағдайдың өзінде дербес лексикалық бірлік тудырмай, жоқтық, болымсыздық мәнді уақытша формалар тудыратын кезі жиі кездеседі. Мысалы: *барусыз қал, кітапсыз, қаламсыз қалу, оқусыз қалу* т.б. Мұндай қолданыстардағы *-сыз/сіз* қосымшасын түркі тілдерінде М.А.Хабичев жоқтық, болымсыздық форма деп атайды [96]. Қазақ тіл білімінде Қ.Шаяхметов аталған қосымшаның зат есімдердің түбіріне тікелей жалғанғандағы қызметін – сөз тудырушылық, ал форма тудырушы жұрнақтардан жалғанған кезде форма тудырушы қызмет атқарады дейді. Зерттеуші өзі зерттеген екі функциялы аффикстердің барлығын дерлік осы жүйемен бірізді етіп береді. Алайда *-сыз/сіз* қосымшасы зат есім түбірлеріне жалғанғанда әрқашан сөз тудыра бермейді. Ол жоқтық, болымсыздық мәнді сөйлеу бірліктерін де тудырады. Яғни сөйлеу бірлігі дегеніміз – ол сөздің грамматикалық формасын ғана тудырады деген сөз. Мысалы *-сыз/сіз* аффиксі *мейірімді, көңілді* сөздерінде туынды

сын есімдер тудырған болса, *кітапсыз қалу, орамалсыз келу* деген сияқты тіркес құрамында сөйлеуде ғана қолданылатын, дербес лексикалық единица ретінде түсіндірме сөздіктерге енбейтін сөз формаларын жасайды. Демек, бұл қосымша форма тудырушы немесе сөз түрлендіруші жұрнақтардан кейін жалғанғанда ғана емес, сөздің түбіріне жалғанғанда да екі түрлі функция атқарады. Бұл қызметі жағынан олар омонимдік қатар құрайды. Егер компьютер жадына кез келген зат есімге *-сыз/сіз* қосымшасын жалғап сын есім тудыр немесе жоқтық, болымсыздық мәнді форма тудыр деген әмір берсек, автомат осы қосымшаны жалғап сөзформаларды туындата бергенмен, олардың лексикалық единица болу-болмауына қарамайды. Бұл жағдайда компьютерлік бағдарлама лексикалық бірліктер тізімінде, яғни негіз сөздер сөздігінде берілген *-сыз/сіз* тұлғалы сөздерді сын есімдер деп белгілеп, оларды негіз сөздер тобына шығарады. Ал негіз сөздер сөздігінде берілмеген *-сыз/сіз* тұлғалы зат есімдерді туынды форма деп танып, аталған қосымшаға өзіміз кодтаған «жоқтық форма» (отрицательная форма) деген белгіленім қояды.

Осыған ұқсас сын есім тудырушы, оның ішінде қатыстық сын есім тудырушы *-дай/дей* қосымшасы да *-сыз/сіз* қосымшасы сияқты зат есім түбірлеріне жалғанып та, етістік формаларынан кейін жалғанып та салыстыру мәнді сөйлеу бірліктерін туғызады. Сондықтан біз аталған қосымшаның негіз сөздер тізіміне енбейтін жағдайларын түгелдей «салыстыру формасы» деп атадық. Мысалы: *-гандай/гендей, -ымдай/імдей, -дағыдай/дегідей* т.б. Жоғарыда аталған қосымшалар сияқты *-дай/дей* қосымшасы зат есім түбірлеріне тікелей жалғанған жағдайдың бәрінде сөз тудырушы қызмет атқара бермейді. Мысалы: *кітаптай, қаламдай, үстелдей, пердедей т.б.* Мұндай кез келген сөзге жалғанған *-дай/дей* тұлғалы зат есімдер сөйлеу бірліктері болғандықтан, олардың барлығын бірдей түсіндірме сөздіктердің реестрінде беру мүмкін емес. Өйткені бұл қосымша түбірге жалғанып қолданылғанмен, түбірге тіркесіп қолданылатын «сияқты, тәрізді» септеулік шылауларымен бірдей мағыналық реңкті білдіреді, яғни жалғанған сөзіне салыстыру мәнін үстейді. Бұл жағдайда да автомат *-дай/дей* тұлғалы сөздердің негіз сөздер тізімінде барларын ғана негіз сөздер аймағына

жатқызады да, қалған тізімде жоқ қолданыстарын екінші тізімнен, грамматикалық формалар, яғни қосымшалар тізімінен іздейді. *-Дай/дей* қосымшасы *-ган/ген, -ар/ер* қосымшаларына жалғанғанда, сондай-ақ тәуелдік, көптік формаларына жалғанып (-ымдай/імдей, -лардай/лердей) жұмсалғанда жалғанған сөзіне салыстыру мәнін үстеп, форма тудырушы қызмет атқарады. Бұл жағдайларда аталған туынды формалар сөздердің грамматикалық формалары тізімінен іздестіріліп, алдын ала дайындалған морфологиялық белгіленімдер бойынша бөлшектеледі.

Осы жоғарыда қарастырылған қосымшалар сияқты тілдің синтетикалық жүйесінде әртүрлі қызмет атқаратын біршама қосымшалар бар. Мәселен, *-дағы/дегі, -шы/ші, -ша/ше, -лы/лі* сияқты қосымшалар.

Башқұрт тілінің зерттеушісі, компьютерлік лингвистика маманы З.Сиразитдинов жоғарыда қарастырылған функционалды қосымшаларды, біздің пікірлерімізге сәйкес, грамматикалық сөздікке енгізеді. Бірақ олардың сөзжасамдық қызмет атқаратын функциясын негізге алып, «сөзжасамдық жұрнақтар» деген атаумен атайды [83]. Осы атауды қолдана тұра, оларды сөзтүрлендіру тізімінде, яғни грамматикалық формалар сөздігінде қарастыратынын атап өтеді. Ғалым грамматикалық қосымшаларды сөзжасамдық қосымшалардан айырудың бір жолы олардың сөздерге талғамай жалғану шартымен де байланысты екендігін айтады. Бұл ретте қазақ тіл білімінде Ы.Маманов форма тудырушы немесе сөз түрлендіруші қосымшалар сөзге талғамай жалғанып, грамматикалық абстракция жасайды деген тұжырым жасаған болатын. Осы критерийге сәйкес аталған функционалды қосымшалар форма тудырушы қызмет атқарғанда, сөздерге жиі жалғанып жұмсалады, сөйтіп, грамматикалық абстракция жасайды. Аталған функционалды қосымшалардың осы ерекшелігі біздің оларды грамматикалық қосымша ретінде танып, грамматикалық формалар сөздігінде қарастыруымызға себеп болды.

Демек, аталмыш қосымшалардың көпфункционалды қызметі формалдық тұрғыдан қиындық туғызғанымен, әрқайсысының тілдік табиғатына сәйкес оларды сөзтүрлендіру аясында қарастырған жөн деп санаймыз.

Түрленуге бейім сөз таптары зат есім мен етістік екендігі белгілі. Сол себепті туынды формалардың дені осы екі сөз табынан туындайды. Енді грамматикалық сөздікке енетін категорияларды санамалап көрсетсек:

I. **Зат есімнің** грамматикалық категорияларынан:

1) ***Көптік жалғау категориясы*** (КЖ) (-лар/лер, -дар/дер, -тар/тер);

2) ***Септік категориясы:***

а) атау септік (∅АС) – грамматикалық формасы жоқ болғандықтан, нөльдік формамен (∅) таңбаланады;

ә) ілік септік (ІС):

б) барыс септік (БС):

в) табыс септік (ТС):

г) жатыс септік (ЖС):

д) шығыс септік (ШС):

е) көмектес септік (КС):

3) ***Жіктік категориясы*** (ЖЖ)

4) ***Тәуелдік категориясы***

а) Тәуелдік жалғау (ТЖ):

ә) Тәуелдік форма (ТФ): -нікі/дікі/тікі

5) ***Функционалды қосымшалар:***

а) Салыстыру формасы (СФ): -дай/дей, -тай/тей;

ә) Ұқсату формасы (ҰФ): -ша/ше;

б) Болымсыздық форма: -сыз/сіз;

в) Иелік форма (ИФ): -лы/лі, -ды/ді, -ты/ті;

г) Қатыстық форма (ҚФ): (-да) + ғы/ (-де) +гі, (та) +ғы/ (те) +гі;

д) Атрибутут-анықтауыш форма (АФ): -лық/лік, -дық/дік, -тық/тік (терезелік ағаш, көйлектік мата).

Зат есімнің сөзтүрленіміне жататын аталған грамматикалық категориялардан басқа да формалар автоматты белгіленім қою кезінде табылуы мүмкін. Оларды жұмыс барысында қарастыра береміз.

Етістіктің грамматикалық категорияларынан:

1) Болымсыз етістік категориясы (БЖ): -ма/ме, -ба/бе, -па/пе;

2) Рай категориясы:

а) Ашық рай (∅АР): өткен шақ, келер шақ, осы шақ тұлғалары;
ә) Бұйрық рай (БР):
б) Қалау рай (ҚР):
в) Шартты рай (ШР):
3) Көсемше категориясы (К):
4) Есімше категориясы (Е);
5) Тұйық етістік категориясы (ТЕ);
6) Функционалды қосымшалар: -шы/ші (бар-шы, айтсын-шы); -сы/сі (көрген-сі, отырған-сы) т.б.

Сонымен, қысқасы, қарашай-балқар тілінің зерттеушісі М.А.Хабичев көрсеткендей, *-дай/-дей* қосымшасын – салыстыру формасы, -ша/ше қосымшасын – ұқсату формасы, *-сыз/сіз* қосымшасын – болымсыздық форма, -лы/лі қосымшасын – иелік форма, -да/де жатыс септігінен кейін жалғанатын -ғы/гі қосымшасын – қатыстық форма деп атауға болады. Ал -лық/лік қосымшасын Н.А.Баскаковтың ізімен – атрибут-анықтауыш форма деп атауға болар еді. Бұл функционалды қосымшалар туралы кезінде түркі тілдерінде үлкен мәселе көтерілгені белгілі.

Ал *айтшы, баршы* деген сөздердегі өтініш мәнін білдеретін -шы/ші қосымшасын әзірге жалпылама «функционалды жұрнақ» деп қана көрсеттік. Мұны біздің ойымызша «өтіну формасы» (ӨФ) деп атауға да болар еді. Сонымен қатар *көргенсі, барғансы* деген сөздердегі -сы/сі қосымшасы да етістік түбіріне жаппай жалғанатындықтан, оны да «функционалды жұрнақ» деп бердік. Бұл қосымша *менменсі, көлгірсі* сөздерінің құрамында сөзтудырушы қызмет атқарса, жаппай етістіктерге жалғанғанда форма тудырады.

Етістіктің түрленіміне қатысты функционалды қосымшаларға атау беру мәселесі әзірге шартты түрде алынып отыр. Морфологиялық белгіленімді жетілдіру барысында өзгертілуі мүмкін.

Морфологиялық белгіленімдер қою мәселесінде екі нәрсенің басын ашып алу қажет. Біріншіден, қазақ тіліндегі ғана емес, түркі тілдері бойынша сөздердің морфемалық құрамы туралы нақты бір шешімге келуіміз қажет. Екіншіден, түркі тілдері бойынша морфологиялық белгіленімдерге арналған шартты таңбаларды бірізді етіп таңбалауға ұмтылуымыз қажет. Бұл ретте кейбір тілдерде шартты белгілер *цифрмен* алынса, кейбір тілдер *латынша*

қаріптермен (тэг) таңбалайды. Ал қазақ тілінде біз әзірге, өзімізге түсінікті болу үшін, кирилл әріптерімен қысқартылған формада беріп отырмыз.

Бұл тақырыпта біз аралық қабаттағы тілдік бірліктерді негіз сөздер сөздігі мен грамматикалық формалар (сөзформалар) сөздігінде қалай ажыратқанымыз туралы сөз еттік. Өйткені тілдік жүйедегі мұндай жайттардың табиғатын толық ажыратып алмай, сөздердің морфологиялық құрылымын автоматты түрде тани алмаймыз. Сол сияқты түрленім (грамматикалық) сөздігіне енетін грамматикалық категориялар мен грамматикалық формалар тізімін жалпылама атап өттік.

4.5. Қазақ тілінің ұлттық корпусындағы лингвистикалық белгіленім қоюды автоматтандыру

Қазіргі кезде барынша адекваттық сипаттағы белгіленімдермен жабдықталған зерттеу жұмыстарына арналған корпустар құрастыруды қажет етеді. Осыған байланысты арнайы программалық жабдықтау құралдарын енгізу арқылы белгіленім (разметка) енгізудің (аннотациялаудың) жеке кезеңдерін автоматтандыру әрекеті корпус құрастыру ісінде өзекті мәселе болып табылады.

Корпус мәтіндеріне лингвистикалық белгіленім енгізу үшін біз төмендегідей жабдықтау құрамын (*принципы построения инструментария*) ұстаным етуді ұйғардық:

1) Символдарды кодтау кезінде программалық жабдықтау Unicod жүйесін ұстаным етуі қажет. Осындай ұстаным негізінде диакритика мен/немесе кирил мен латын қаріптерінен айырым табатын әліпби символдарын қолданатын тілдерді сипаттауды қамтамасыз етеді;

2) Мәтіндік топтама және оған сәйкес келетін тілдік ақпарат бір ғана дерекқорда сақталулары қажет және мәтіндер корпусымен әрекет ету программалық компоненттер жүйесі арқылы стандартталған қатынас арқылы жүзеге асады.

3) Корпус деректерін өңдеуге арналған нақты жабдықтар корпустан бөліп алуға мүмкіндігі болуы қажет. Мұндай жағдай жүйенің жалпы әмбебаптылық сипатын қамтамасыз етеді және со-

нымен бірге, басқа жүйелердің де лингвистикалық білім қорын өз қажеттігі үшін қайталай пайдалануына мүмкіндігі болады.

4) Жүйенің әрбір құрамдас бөліктері жүйенің модульдік ұйымдастыру жағын қамтамасыз ететін жеке бір тілдік мәселенің шешімін таба алуы қажет.

5) Тілдік деректер бойынша мәтіндік мәліметтерді мәлімдеу біріншілік сипатта; барлық одан туындайтын тілдік ақпарат мәтіндік корпустың позициясына тікелей байланыста. Мәтіндік ақпараттарды мәлімдеудің біріншілік қағидаты корпустағы сөзформалар мен сөзтіркестерге қатысты лексикалық және грамматикалық контекстермен тез арада танысуға мүмкіндік береді. Лексикалық бірліктердің мәтінге байланыстылығы (привязка) омонимдік формалардағы сөз бен сөзтіркестеріне қойылатын (мүмкін болатын) морфологиялық белгілер жиынтығын оңай шығарып алуға мүмкіндік туғызады. Мәтіндер корпусына белгіленім енгізуді автоматтандыру мәселесі болашақтағы тілдік бірліктер қызметінен лексика-грамматикалық ақпарат алуға бағышталған. Осыған байланысты корпусқа белгіленім енгізу мәселесі морфологиялық және синтаксистік деңгейлерді қамтиды. Мұндағы белгіленімді автоматтандыру жүйесінің мәтінді жан-жақты өңдеу мүмкіндігін қамтитынын атап айту қажет. Кейбір функциялар (қызмет атқару) әртүрлі программалық құралдармен ұсынылған (тіпті кейбір функциялар қайталануы да мүмкін), бірақ ол зерттеудің ыңғайлылығын көздейді және жалпы модульдік принцип пен әмбебаптылықты бұзбайды. Корпус мәтіндеріне белгіленім енгізу кезіндегі ақпараттық ағым мен жүйе модульдерінің арасындағы әрекеттестік сызбасы 10-суретте көрініс тапты.

Қазақ сөздері мен сөзформаларының морфемдік құрамын бөліктеу процесін автоматтандыру үшін олардың сөзжасамдық, форматудырушы, сөзтүрлендіруші қызметтерін айқындау қажет. Аффикстердің екі түрлі функцияда кездесу мүмкіндігін анықтау олардың жаңа сөз жасап, өзі жалғанған сөзбен бірге негіз сөздер сөздігіне қосуға немесе сөзді түрлендіріп, туындатылған сөзформалар сөздігіне енгізуге мүмкіндік береді. Әрі қарай сөзтүрлендіруші аффикстерді сөзформалардан бөліп алғаннан кейін оларды белгіленім қойылған сөзтүрлендіруші жалғаулар тізімінің базасына қосуға болады.

10-сурет. Ақпараттық ағым мен жүйе модульдерінің арасындағы әрекеттестік сызбасы

Аглютинативті тілдерде сөзтүрлендіруші аффикстердің негізге жалғану реті тұрақты, анық болып келетіні белгілі. Сондықтан сөзтүрлендіруші қосымшалар тұрақты ретпен бір ғана емес, бір сөз табына жататын көптеген сөздерге бірдей жалғана алады. Мұндай жүйе түркі тілдерінің барлығына тән. Бұл әсіресе өте көп кездесетін зат есім сөздер мен етістіктерден анық байқалады. Сондықтан, сөзтүрлендіруші аффикстерге негізделген сөзформалар сөздігі негізінен осы аталған екі сөз табына жатады.

Қазақ тілінің ұлттық корпусындағы морфологиялық белгіленімдерге ажыратуды автоматтандыру кезеңдері:

1-блок. Корпустық базаға ең алдымен дербес база ретінде әліпби ретімен қазақ сөздерінің құрамы мен сөз табына қарай ажыратылған негіз сөздер сөздігі енгізіледі.

Бұл сөздік база сөз табына қатысты белгіленім қойылған 15-томдық және 106 мың реестр сөзді қамтитын 1-томдық қазақ әдеби тілінің түсіндірме сөздіктерінің материалдарына негізделген. 106 мың реестр сөздерден алынған тізім құрамы бойынша да морфологиялық белгіленім қойылған. (Келешекте семантикалық белгіленімдер де қою жоспарланып отыр).

Негіз сөздердің сөз табына қоса басқа да құрамдық, яғни морфологиялық және сөзжасамдық құрылымына да сипаттама жасауға болады. Бұл ретте аннотация, яғни лингвистикалық ақпарат (талдау) *сөзжасамдық белгіленімдер* бойынша жасалады. Ол үшін сөздердің сөзжасамдық құрамы туралы базаға мына сияқты ақпараттар салынуы қажет.

34-кесте

Реестр сөздердің сөз табы мен құрамына қарай
қойылған белгіленімі

Реестр сөз	Сөз табы	Құрамы бойынша	Сөзжасамдық құрамы	Сөзжасамдық тәсілі
аба	зт.	жай	негізгі	
абай-қоқай	зт.	күрделі	туынды	аналитикалық
абайтану	зт.	күрделі	туынды	аналитикалық
абайтанушы	зт.	күрделі	туынды	аналитикалық
абайшылдық	зт.	жай	туынды	синтетитикалық

Сонымен қатар негіз сөздердің мағыналық ерекшеліктерін сипаттайтын семантикалық аннотация да жасауға болады. Мысалы, зат есімдердің деректі, дерексіздігі, жалқы есім немесе жалпы есім екендігі, сонымен қатар белгілі бір лексика-семантикалық топқа қатысы (өсімдік атауы, тамақ атауы, киім атауы, т.б.) т.б. семантикалық белгілері көрсетіледі. Мұндай аннотация түрі *семантикалық белгіленім* деп аталады.

35-кесте

Зат есімнің лексика-семантикалық топтары

1	Дене мүшелері атаулары	Бас, қол, саусақ
2	Тамақ атаулары	Сорпа, ет, қазы
3	Киім-кешек атаулары	Шапан, көйлек
4	Жануарлар атаулары	Сиыр, жолбарыс
5	Құс атаулары	Сауысқан, тырна

Зат есімнің лексика-семантикалық топтары бұл аталғандармен шектелмейді (12-кестені қараңыз). Жоғарыда әрбір топқа мысал ретінде бірлі-жарым сөз ғана берілген. Тіліміздегі барлық зат атауларын осылайша белгілі бір топқа теліп шығуға болады. Мұндай кестені етістік, сын есім т.б. сөз таптарына қатысты жасауға болады. Осындай тізім базаға салынса, компьютерлік бағдарлама мәтін бойында кездескен сөздердің семантикалық сипаттарын осы сияқты кестелерді пайдаланып, сипаттап көрсетеді.

Ал синтаксистік белгіленімдер қою мәселесі лингвистикалық аннотацияның ең күрделісі болып табылады. Синтаксистік белгіленімдер сөздердің тіркесім заңдылығына, синтаксистік қатынынастар бойынша жасалады. Орыс тіл білімінде синтаксистік белгіленімдер «деревозависимость» деп аталатын заңдылыққа негізделгендігі айтылады. Мәтіндерге мұндай синтаксистік аннотация жасау мәселесі тілдердің көпшілігінде жүзеге асырылмаған. Қазақ тілінде де корпус құрастыру ісі кейінгі жылдары ғана қолға алынып жатқандықтан, лингвистикалық аннотацияның бұл түрі әлі қарастырылған жоқ. Ол үшін құрылымдық жүйедегі синтаксистік қатынастарды тауып, оларды белгілі бір модельге салу қажет.

2-блок. Корпустық базаға дербес база ретінде қазақ тіліндегі әрбір сөз табы үшін жасалған сөзтүрлендіруші қосымшаларының тізімі енгізіледі. Сондай-ақ базадағы сөзтүрлендіруші қосымшалардың құрылымы (немесе жалғану реті) алдын ала грамматикалық деңгейде белгіленім қойылады. Соны бір мысал ретінде «зат есімдердің түбір және туынды негіздерінен туындаған соңы жіңішке буынға біткен сөзформалар жасаушы қосымшаларынан» үзінді келтірейік.

36-кесте

Зат есімдердің түбір және туынды негіздерінен туындаған соңы жіңішке буынға біткен сөзформалар жасаушы қосымшаларға мысалдар

Соңғы дыбысы дауыстыға аяқталатын (*у, и*-дан басқа) негіз сөздердің сөзтүрлендіруші қосымшалары	Сөзтүрлен-діруші қосымша-лардың құрылымы	Сөзтүрлендіруші қосымшалар құрамының белгіленімдері
Дей	*дей*	*дей/СФ*
сіз	*сіз*	*сіз/БФ*
сіздің	*сіз+дің*	*сіз/БФ+дің/ІС*
сізге	*сіз+ге*	*сіз/БФ+ге/БС*
сіздер	*сіз+дер*	*сіз/БФ+дер/КЖ*
сіздерде	*сіз+дер+де*	*сіз/БФ+дер/КЖ+де/ЖС*
сіздерден	*сіз+дер+ден*	*сіз/БФ+дер/КЖ+ден/ШС*
сіздермен	*сіз+дер+мен*	*сіз/БФ+дер/КЖ+мен/КС*
сіздерменен	*сіз+дер+менен*	*сіз/БФ+дер/КЖ+менен/КС*

Кесте 3 бағаннан тұрады. 1-ші бағанда сөзтүрлендіруші қосымшалар құрамы жағынан бөлінбеген. 2-бағанда сөзтүрлендіруші қосымшалар «+» белгісі арқылы бөлінген, 3-бағанда әрбір қосымшаға қысқартылған шартты таңба қойылған, яғни сөзтүрлендіруші қосымшалардың құрылымы грамматикалық белгіленімдер арқылы көрсетілген. Атап айтып кету керек, кестедегі сөзтүрлендіруші қосымшалар тізімі осы сияқты басқа барлық сөз таптары бойынша негіз сөздердің соңғы дыбыстарына қарай ыңғайланып жасалған.

3-блок. Жоғарыда айтылған 2-блокқа сүйене отырып, морфологиялық деңгейде белгіленім қойылатын барлық сөзформалар жасалымын көрсетуге болады. Бұл жағдайда алдымен негіз сөз бөлініп алынады да, оның сөз табына қатысы көрсетіледі, сол сияқты сөзформаның құрамы бөлшектеліп, олардың әрқайсысына морфологиялық белгіленім қойылады.

Мысалы, «әке» деген зат есім сөзді алатын болсақ, базадағы реестр сөздердің тізімінен іздеу кезінде алдымен «әке» сөзін тауып алып, оны негіз деп белгілейді, содан кейін берілген сөзтүр-

лендіруші қосымшалар арқылы оны түрлендіреміз. Мысалы: ***әкедей, әкесіз, әкесіздің, әкесізге*** т.б. Кестедегі 3-баған грамматикалық деңгейде белгіленім қойылған, яғни онда шартты белгілер қысқартылған формада енгізілген.

4-блок. Енді 3-блокта берілген сипаттамаға сәйкес қазақ тілінің мәтіндер корпусынан алынған кезекті сөйлемдердегі кез келген сөзформаға морфологиялық белгіленім қою ұстанымдарын қарастырайық.

Біріншіден, сөйлемнен алынған кезекті сөз сөзформа болмай, қазақ тіліндегі реестр сөз болуы мүмкін. Оның белгіленімін қою үшін қазақ сөздерінің негіздер базасынан алғашқы екі әрпі бойынша іздеу керек. Ал мұндай сөздер міндетті түрде әліпби ретімен тұрады және оларға сөз табы көрсетіліп, құрамына қарай (бірқұрамды, екіқұрамды, көпқұрамды) бөлшектеледі.

Сөйлемнен алынған сөз сәйкес келген жағдайда «Қазақ тілінің негіздер сөздігі» базасындағы белгіленім қойылған сөзбен ауыстырылады.

Екіншіден, егер сөйлемдегі тексерілетін сөз «Қазақ тілінің негіздер сөздігі» базасындағы белгіленім қойылған сөзбен сәйкес келмесе, онда мынадай әрекеттер орындалады:

1) берілген сөзформадағы әріптер санын анықтау керек;

2) берілген сөз бойынша алғашқы екі әрпі сәйкес келуіне қарай реестр негіз сөздер базасы бойынша сөзформалар туындату және туындатылған сөзформалардан «ұзындығы» және сөзформадағы әріп саны бойынша тауып алу. Содан кейін, ол сөз зат есімге жатса, «Зат есімге арналған сөзтүрлендіруші қосымшалар» базасының барлық фонетикалық варианттары бойынша туындату.

Осындай процесті егер сөзформа басқа сөз табына жатса, сол сөз табына қатысты сөзтүрлендіруші қосымшалардың тізімі бойынша жүргізу қажет.

Үшіншіден, әрбір туындатылған сөзформа сөйлемнен алынған тексерілген кезекті сөзформамен салыстырылу керек. Олар сәйкес келген жағдайда туындатылған сөзформаны грамматикалық белгіленім қойылған, яғни оның белгілі бір сөз табына қатысы көрсетілген және негізге жалғанған сөзтүрлендіруші қосымшаға грамматикалық белгіленім қойылған деп есептей беруге болады.

Омоним сөздерді анықтауды жартылай автоматтандыру үшін төмендегі ұсыныстар жасаймыз:

1) Мәтін ішіндегі сөйлемдер құрамында кездесетін сөздерді «Қазақ тілінің негіз сөздер сөздігіндегі» сөздермен салыстыру «басқа сәйкестіктер жоқ» деген жауап шыққанша, қайталана береді, ал бірінші салыстырудан бастап табылған сәйкес, яғни бірдей сөзформалар тізімге алынып отырады. Демек, омоним сөздер бірнеше рет кездесуі мүмкін. Мұндай сөздер кейде әр сөз табына жатуы, кейде бір сөз табына жатуы да мүмкін. Мысалы, қазақ тіліндегі мынадай омомним сөздер: қара - етістік, қара - сын есім, қара - зат есім – қарайған нәрсе және т.б. Қазақ тілінде бұл сияқты омоним сөздер өте көп болуы мүмкін.

2) Сонымен қатар сөйлемдегі сөзді туындатылған сөзформалармен салыстыруда тек бір ғана сөз табына қатысты сөзтүрлендіруші формалармен ғана салыстырып қана қоймау қажет. Мысалы, деректер базасына енгізілген «зат есім сөз табының сөзтүрлендіруші аффикстерінің түрленім формаларымен шектелмей, «басқа сөз таптарында сәйкестіктер жоқ» деген ақпарат шыққанша сын есім, сан есім т.б. сөз таптарының түрленісімен салыстыра беру қажет.

Қысқасы, қазіргі уақытта біз орыс, украин және бірнеше түркі тілдеріндегі мәтіндерге автоматты түрде морфологиялық белгіленім енгізу әдісін негізге аламыз.

Тәжірибеде қазақ мәтіндеріне автоматты (жартылай автоматты) түрде белгіленім қоюдың осындай әдісін қолдану қанағаттанарлық нәтиже беріп отыр, алайда әлі де оның кейбір тұстары жетілдіруді, практикалық жағынан өңдеуді қажет етеді.

4.6. Қазақ тілінің мәтіндер корпусындағы омонимдер мәселесі

Қайсібір тілдің аннотацияланған мәтіндер корпусын алмайық, ең алдымен, морфологиялық белгіленім қою мәселесіне назар аударған. Морфологиялық деңгейде мазмұндалған корпус сөзформаның қандай морфологиялық бөлшектерден тұратынын айқындауды мақсат етеді. Егер бірдей дыбысталатын сөз әр алуан сипаттағы морфологиялық бөлшектерден тұрса, онда ол омоним

сөз болғаны. Омоним сөздер мәтіндер корпустарын морфологиялық деңгейде аннотациялауда едәуір қиындық келтіретін «проблемалық аймаққа» жатады. Өйткені компьютер тілі формалды құрылымдарды ғана түсінеді, ал тілдік бірліктердің (сөзформалардың) мағыналық айырмашылықтары тек қана «адамның» қатысуымен атқарылады. Бұл мәселе корпустық лингвистикада автоматты түрде емес, «қолдап» жүзеге асыру арқылы орындалады. Морфологиялық белгі-кодтар қою жұмысының екінші сатысы осы омонимдерді ажырату мәселесіне арналады. Демек, сөздік қордың омонимдік қабаты құрылымдық лингвистикада ғана емес, тіл білімінің жаңа бағыттарында да (бұл жерде қолданбалы лингвистикада да) күрделі құбылысқа жатады.

Құрылымдық лингвистикада омонимдер тілдік құбылыс ретінде ғана танылып қойған жоқ, сонымен қатар сөздік қордағы белгілі бір жүйеге біріккен сөздер жиынтығы ретінде де ерекшеленді. Нәтижесінде омонимдер сөздігін құрастыру қажеттігін туғызды. Басқа да лексикографиялық еңбектер сияқты омонимдер сөздіктері кітап нұсқасында ғана емес, электронды пішінге келтіріліп, интернет жүйесіне енгізіле бастады.

Қазіргі компьютерлік лингвистика саласындағы қолданбалы зерттеулердің нәтижесі көбіне тиісті лингвистикалық ресурстардың, ең алдымен лексикографиялық еңбектердің болуына байланысты. Мысалы, соңғы кездері орыс тілінің омонимдік сөздіктері кеңінен шыға бастады. Атап айтсақ, Н.П.Колесниковтың сөздігінде лексикалық омонимдермен қатар омоформалар, омофондар, омографтар қоса берілген. О.С.Ахманованың, О.М.Кимнің арнайы омонимдік сөздіктерінде функционалды омонимдер сипатталса, Н.Г.Аношкинаның интернет-ресурсында барлық грамматикалық омонимдерді жинақтауға ұмтылыс жасалған, ал Т.Ю.Кобзарованың еңбегінде функционалды омонимдердің 58 типі жіктеліп көрсетілген және т.б.

Қазақ тіл білімінде бұл салада мектеп оқытушыларына арналған М.Белбаеваның «Қазақ тілінің омонимдер сөздігінен» басқа еңбек жоқ деуге болады [97]. Сөздік, құрастырушының айтуынша, мектеп оқушылары мен қазақ тілі пәнінің мұғалімдеріне көмекші құрал ретінде ұсынылады. Әйтсе де, бұл еңбек зерттеушілер үшін

қазірше бірден-бір омонимдік сөздік ретінде болып отырғанын айта кеткеніміз жөн.

Омонимдерді аннотацияланған мәтіндер калай таниды деген мәселеге келсек, қазақ тілінің әртүрлі жанрларынан алынған мәтіндер бойынша жасалып жатқан морфологиялық белгіленімдерді қоюда омонимдерді ажырату белгі-код қоюдың автоматты және жартылай автоматты түрлерінде әртүрлі жүргізіледі. Себебі, жартылай автоматты түрде жасалған белгіленімдер адамның тікелей қол қызметі арқылы жүзеге асады. Мұндайда адам берілген мәтінді синтагмалық қатынаста тұрған, яғни жекелеген сөйлемдерді оқып, көріп отырады. Әрбір сөздің контекстегі мағыналарын ажыратып, тиісті талдаулар жасайды. Ал автоматты белгіленімде мәтіндерді құрайтын сөзформаларға қойылатын белгіленім бұрыннан комьютер жадына салынған ақпараттар бойынша өзінде бар сөзформаның сыртқы тұрпатына сай келетінідігіне қарай белгі-кодтар қойып шығады. Мұндайда белгіленім қою жұмысы автоматты түрде жүреді. Берілген сөздің қай мағынадағы сөз екендігі ескерілмейді. Сондықтан қателесулер туғызады. Осыны ескеріп, корпус жасау мәселесінде тәжірибелі елдердің өзі омонимдерді ажырату жұмысында «қолдап» жасау әдісін оңтайлы деп тапқан. Демек, контекст омонимдерді ажыратудың негізгі шешуші факторы болып табылады. К.Аханов «Омонимдер де, көп мағыналы сөздер де контексте қолданылады да, олардың мағынасы контексте айқындалады. Осымен байланысты олардың тілде өмір сүруі ойдың түсінікті болуына кедергі жасамайды. Омонимдердің немесе көп мағыналы сөздердің қолданылуынан болуы мүмкін екіұштылық контексте жойылып кетеді», – деп полисемия мен омонимияны ажыратуда мәнмәтінге сүйену қажеттігін айтады [98, 125].

«Қара» деген сөздің әртүрлі қатынаста келуін, қай мағынада қолданылғанын, қандай-қандай мағыналық реңкі бар екенін мәнмәтінде (контексте) айқындап ажыратуға болатындығын мынадай ретпен береді:

«*Қара* – сын есім, қарсы мағынасы – ақ. Мысалы: *Әмір қара жал, қара құйрық, құрым сары ат мінген* (М.Ә.).

Қара – келтірінді мағынасында *адал емес, ақ емес, кінәлі, теріс* дегенді білдіреді. ***Қара*** *деп жазаланған күнде мұның жазасы не бол-*

мақ? (М.Ә.); *Немене, Аққа болыспай,* ***қараға*** *болысып тұрмын ба?* (Ғ.Мүс.).

Қара сөзі кейбір сөздермен тіркесіп, қолданылу ыңғайына қарай басқаша мағыналық реңк алып, «ешнәрсесіз», «жаңғыз өзі», «бос», «құр» дегенді білдіреді. Бұл мағына, «қара» деген сөздің тұрақты негізгі лексикалық мағынасы емес. Мысалы: *Қайғысыз* ***қара*** *суға семіреді* (Мақал); *Қазір аттан.* ***Қара*** *су ауыз тиместен аттан!* (М.Ә.). *Көңілсіз* ***қара*** *суық қырда жүрсең* (Абай). ***Қара*** *суық желі де, бұлыңғыр жауыны да ерте иектеп, мезгілсіз бүрсеңдетіп тұр* (М.Әуезов).

«Қара су» дегендегі «қара» сөзді «басқа қосымша жоқ, жалғыз өзі, құр» дегенді білдірсе, «қара суық» деген де осы сияқты «қарсыз, құр өзі» дегенді білдіреді. Осы мағналық реңк «қара басы» деген тіркесте де сақталған. *Өз* ***қара*** *басының сондай беделді екеніне бойы үйренген Ушаков қарсылық көрмей, беті қайтпай келіп еді* (Ғ.Мүсірепов).

Қара «алыстан, көз ұшынан қарайып көрінген нәрсе, бірдеменің қарайып көрінген бойы» дегенді де білдіреді. *Содан мен шақпақтың жотасынан, биікте тұрып,* ***қараларыңды*** *көргенмін* (М.Әуезов).

Бұл мысалдардан «қара» деген сөздің әртүрлі мағыналарда, әртүрлі мағыналық реңкте қолданылатынын көрдік. Оның қай мағынада екені жоғарыдағы сөйлемдерден, сол сөйлемдегі сөздердің тіркесінен аңғарылады. Басқаша айтқанда, оның әртүрлі мағыналары контексте айқындалады.

«Қара» сөзінің жоғарыда аталған мағыналарының бір-бірімен қатысы бір сөздің әртүрлі мағыналарының өзара қатысы ретінде қаралады. Ал «мал» дегенді білдіретін «қара» сөзі мен «назар аудару, көру» дегенді білдіретін «қара» сөзінің жоғарыда аталған мағыналарға қатысы сөздердің бір-бірімен омонимдік қатысы ретінде қаралуға тиісті.

1) *Бұл құсты осы өткен жазда Абай Сыбандағы Тулақ деген бүркітшіден он шақты* ***қара*** *беріп, сатып алған* (М.Әуезов).

2) *Бәрекелде, мына қызықты* ***қара****!*

Алдыңғы мысалда «мал» деген мағынаны білдіретін «қара» сөзі сын есімнің заттануынан барып өзгеріп, заттық мағына алған деуге

болады. Ал соңғы мысалдардағы етістік «қара» дегеннің бұған қатысы жоқ, тіптен басқа сөз. Бұлар өзара омонимдес сөздер ретінде қаралады. Олардың қай мағынада екені де контексте айқындалады. Сондықтан омонимдерді қолдап ажыратпаса, ізделген сөздің барлық омонимдерінің лингвистикалық сипаттамалары беріледі. Мәселен, қазақ тілінің ұлттық корпусы бойынша *қара* сөзінің лингвистикалық сипаттамасын келтірейік:

*Әлдеқандай долы дүлей **қара** жерді тас талқанын шығарып, күл қып уатып, сосын бір алапат табаққа салып ап, елеп, кебегін желпіп ұшырып жатқандай, жер мен кек түгел шаңға айналып кеткенді (Ә.Кекілбаев).*

Қара

	Сипаттау
Лемма	қара
Морф. белгіленім	сн
Лексик. семантика	1. Күйенің, көмірдің түсіндей (түс, бояу). 2. а у ы с. Арам, қараниет, зұлым. 3. а у ы с. Теріс, қисық. // э т н. Өлікті аза тұтудың белгісі ретінде киетін киім. // Масқара болу, абыройсыздық. // Қарапайым, халық арасынан шыққан адам, қараша, халық, бұқара.
Морфо-семант белгіленім	дара, түбір, сапалық
Мәдени семантика	

	Сипаттау
Лемма	қара
Морф. белгіленім	ет
Лексик. семантика	1. Көз жіберу, назар аудару. 2. Бағу, жаю (мал, құс). 3. Іздеу, іздестіру. 4. Көру, тексеру. 5. Тексеру, бақылау, ауруын байқап білу. 6. Тыңдау, құлақ асу. 7. Бағалау, елеу, секеру, мән беру. 8. Бағыну, мойындау.

	9. Аңғару, байқау, аңдап бағу. 10. Күту, тосу, бөгелу. 11. Бағып-қағу, күту, мәпелеу, баптау. 12. Бағыну, қарауына, қол астына өту, тәуелді болу. 13. Есептеу, санау, ойлау. 14. Ауу, беттеу, ойысу, қадам басу. 15. Арнайы сөз ету, талқылау. 16. Қарастыру, зерттеу. 17. Дәл бағытталу, тура келу. 18. Оқу, оқып шығу.
Морфо-семант белгіленім	дара, түбір, негізгі, сабақты
Мәдени семантика	

	Сипаттау
Лемма	қара
Морф. белгіленім	зт
Лексик. семантика	«1. Дәл не екені айқын емес, қарайып көрінген нәрсе; қарайған. 2. Аумақ, көлем, сан, мөлшер, шама. 3. Жобасы, нобайы, пошымы, түрі. 4. а у ы с. Ес, серік, тірек, сүйеу. 5. Ірі мал: жылқы, түйе, сиыр. 6. Қарашық, көздің бұршағы. « Өсімдік, жануар, жәндік, т.б. атаулардың құрамында кездесіп, олардың ағазсы не құрылымындағы бір бөлігінің қара түсті болып келетін түрлерін атайтын күрделі сөздердің құрамдас бөлігі.
Морфо-семант белгіленім	
Мәдени семантика	

Міне, жоғарыда «қара» сөзінің омонимдік қатарына семантикалық сипаттама берілді. *Қара* сөзі омоним болғандықтан, жоғарыда үш түрлі сөз табына қатысты лингвистикалық сипаттамалары бірдей көрініс тауып отыр.

Автоматты жолмен морфологиялық талдау жасайтын бағдарламаларға омонимдерді ажырату үшін де арнайы омонимдерді танудың формалды белгілері бойынша алгоритмедер құрастырылған, олар тілдегі сөздердің синтагматикалық байланысына негізделген.

Яғни омонимдердің қай сөз табына қатыстылығын анықтауда ол сөздің қоршаулары (дистрибуциясы) бойынша ережелер жасалған. Ол ережелер алгоритмге салынған. Мұны омонимдерді ажыратудың ережеге салынған (детерминирленген) формасы дейді.

Морфологиялық белгіленім қоюда сөздердің сөз табына қатысы негізге алынатындықтан, омонимдерді ажыратуда да ол омоним сөздің сөз табына қатысын анықтау, яғни омоним сөздің сөз табына қатысын белгілеу мақсат етіледі. Бұл ретте сөйлемдегі омоним сөздің мына белгілері негізге алынады:

а) сөздің сөйлемдегі орын тәртібі, яғни орны;

ә) сөздің қоршаудағы тіркесімділігі (дистрибуциясы);

б) сөзге жалғанатын сөзтүрлендіруші тұлға (қосымша);

в) сөздің соңында немесе алдында тұрған тыныс белгісі.

Осы белгілердің әрқайсысы бойынша талдау жасап көрейік.

Қазақ тілінде бастауыш көбінесе сөйлем басында және ортасында келеді. Ал баяндауыш өлеңдерде болмаса, негізінен сөйлем соңынан келеді. Сол сияқты сын есім көбінесе зат есімнің алдында тұрады. Ал мезгіл үстеулер көбінесе сөйлем басында тұрады. Басқа да сөз таптарының сөйлемде иеленетін белгілі бір орын тәртібі бар. Осы принцип бойынша, «қара» сөзі сөйлем аяғында тұрса, онда ол – етістік. Мысалы, *Сен балаға қара.* Омонимді анықтау моделі зт + БС + *қара* → қара/ет.

Сөйлемнің басында немесе ортасында келсе, сонымен қатар зат есімнің алдында тұрса, сын есім болады. Мысалы, ***Қара мәшине*** *тоқтады. Аяжан* ***қара көйлегін*** *киіп, қонаққа кетті.* Омонимді анықтау моделі *қара* + зт → қара/сн.

Ал етістіктің алдында келсе зат есім болады. Мысалы, *Алыстан бір* ***қара көрінді.*** Омонимді анықтау моделі *қара* + *ет.* → қара/зт.

Екінші белгі бойынша, сөздер жарыққа шығып, айтылымға айналғанда (сөйлем құралғанда) семантикалық валенттілік бойынша тіркеседі. Мысалы, *ыстық қар, сабырлы жүзім* деген сияқты қолданыстар жоқ.

Үшінші белгісі бойынша, түрленуге бейім сөз таптары негізінен етістік пен зат есімдер. Етістік зат есімнің түрлену жүйесімен түрленбейді, яғни етістікке зат есім қосымшалары жалғанбайды. Тек қана тұйық етістік формасында тұрғанда ғана зат есім

парадигмасымен түрленеді. Сондықтан, омонимдерді түрленген формада ажырату қиындық келтірмейді. Мысалы, *қара* етістігінің ашық райда жіктелген формасын етістік екендігін жалғанған қосымшаларына қарап анықтауға болады. Мысалы, *қараймын, қарайсың, қарайсыз, қарайды.* Ал бұйрық райдағы жіктелген формасы 1 және 3 жақта белгілі де (*қарайын, қарасын*), екінші жақта түбір сөзбен (*Сен балаға қара*) сәйкес келеді. Бірақ бұл жағдайда да формальді белгісін табуға болады. Ол *қара* сөзінің өзінің алдында тұрған есім сөзбен барыс септік тұлғасында (есім + -ға/ге/қа/ке (БС) + қара) тіркесуі және сөйлем соңында тұруы.

Қара сөзі етістік болғанда алдыңғы сөзбен барыс септік формасында ғана емес, табыс септікте де тіркесуі мүмкін. Мысалы, *даланы қара, суретті қара* т.б. (есім + -ны/ні/ды/ді/ты/ті (ТС) + қара). *Қара* сөзі етістік болғанда шығыс септігінде тұрған сөзбен де тіркесе алады. Мысалы, *Терезеден қарады.* (есім + -дан/ден/тан/тен/нан/нен (ШС) + қара).

Сол сияқты зат есім парадигмасымен түрленген *қара* сөзін бірде (субстантивтенген) сын есім (*Көйлектің қызылын сатып алды*), бірде зат есім (*Бір қараны жетектеп келеді*) ретінде танимыз.

Ал *қара* сөзі шырай формаларын қабылдағанда (*қарарақ, қаралау*) ол міндетті түрде сын есім болатыны белгілі.

Төртінші белгісі бойынша, яғни соңынан нүкте не леп белгісінің қойылуы. Демек, сөйлем соңында келген *қара* сөзінен кейін нүкте не леп белгісі қойылса, онда ол – етістік және бұйрық рай формасында тұр деп белгілеу керек.

Жоғарыда талданған *қара* сөзі омонимдес әртүрлі сөз табына қатысты. Омонимдер кейде бір сөз табында да болады. Бірақ бір сөз табына жататын омонимдерді ажырату морфологиялық талдау (белгіленімдер) үшін маңызды емес. Өйткені бір сөз табына жататын омоним сөздердің түрленуі де, тіркесімділігі де, сөйлемдегі орын тәртібі де негізінен бірдей болып келеді. Мысалы, *Қас* деген бір сөз табына жататын омонимдес төрт сөз бар:

1. *қас* — адамның қабағындағы қас;
2. *қас* — дұшпан, жау;
3. *қас* — бірдеменің жаны, маңы (үйдің қасы);
4. *қас* — ердің қасы

Бір сөз табына жататын немесе әр сөз табына жататын омонимдердің мағынасын түсіндіру лексикалық белгіленімге жатады. Бір сөз табына жататын омонимдерді ажыратудың формалды белгісі жоқ деуге болады. Бірақ белгілі бір тіркесімділік ретіне қарай анықтауға болады дейді кейбір зерттеушілер. Мұндайда бір сөз табына жататын омонимдердің тіркесім қоршауын тұрақты модельге сап, тіркесім сөздігін жасау қажет деп көрсетеді.

Омонимдердің лексикалық бірлік пен түрленген формадағы сөзбен омографтық жұп құрайтын кездері де бер. Мысалы, *бал* сөзіне барыс септік жалғауын жалғағанда *балға* болады. Тілімізде *балға* деген де түбір сөз бар. Морфологиялық талдауыш (анализатор) мұндайда қайсысын түбір сөз (балға), қайсысын барыс тұлғалы *бал* сөзі деп анықтайды. Мұндайда *балға* түбір сөзі өзінен кейін басқа да септік формаларын қабылдайды. Мысалы, *балғаның, балғаны, балғада, балғадан, балғамен.* Ал *бал* сөзі септелгенде барыс септігінде ғана *балға* сөзімен омонимдес болады. *Балға* сөзі табыс септігінде тәуелденіп қолданылғанда да (балғамды, балғасын) барыс септік формалы *бал* сөзімен омограф болады. Алайда бұл жағдайда талдауыш оның түбірі *балға* екенін тани алуы қажет. Себебі тәуелдік жалғауы септік жалғауынан кейін жалғанбайды. Сондай-ақ сөз соңында қайталанып септік жалғауы үстелмейді.

Сол сияқты *алды* көмекші сөзі бар да (үйдің алды), жедел өткен шақ формасында тұрған *ал* етістігі бар (Ол кітапты алды). Мұндай жағдайда морфологиялық талдауыш бұл екі омографтың екеуіне де екі түрлі белгіленім қойып көрсетеді. Мысалы:

алды/зт.

алды/ ал/ет+ды/ЖӨШ

Бұл жағдайда да сөйлемдегі омограф сөздің қай сөз табы екенін немесе морфологиялық түрленімі шығарылған екі түрлі нұсқадан адам өзі таңдайды, яғни адамның араласуымен ғана анықталады. Сондықтан осындай мәселелерде де сөздер тіркесімділігі, тыныс белгісі сияқты т.б. формалды белгілеріне қарай олардың аражігін табу қажет. Мысалы, *ал* етістігі шақ тұлғасымен түрленгенде сөйлемнің соңында тұрады. Яғни соңынан нүкте қойылады. Сонымен қатар ол - сабақты етістік. Алдыңғы сөздің табыс септігінде тұрғанын талап етеді.

Сол сияқты *алды* сөзі көмекші есім болғанда үнемі алдыңғы сөзбен матаса байланысады (үйдің алды) және көбінесе сөйлемді аяқтамайды, яғни соңынан нүкте келмейді. Демек, бағдарламаға *алды* сөзі ілік септігімен тіркесуінде (есім+ІС+алды) көмекші есім, яғни зат есім болатындығы туралы нұсқаулық салынады.

Міне, омонимдер мен омографтарды ажыратуда компьютерлік бағдарламаға осындай формалды белгілері бойынша нұсқаулықтар жасау қажет. Нұсқаулықтар көп болған сайын да морфологиялық талдауыштар (анализаторлар) жұмысы жақсара түседі, автоматты талдаудың шынайылығы арта береді.

Қазақ тілінің ұлттық корпусын құрастыру тәжірибесінде автоматты морфологиялық талдауыштар іске қосылған. Белгіленім қоюда жоғарыда қарастырылған омонимдерді ажырату мәселесі бойынша зерттеу жүргізіліп жатыр.

Мәтіндерден алынған сөзформаларды аннотациялауда омонимдерді ажыратуды автоматты жүзеге асыру әлі шешімін таппаған мәселе. Алайда, қазақ тіл білімінің қолданбалы бағыты бойынша жасалып жатқан мәтіндер корпусын аннотациялау ұлттық корпус деңгейіне жеткенде, яғни қазақ тілінің барлық стильдері бойынша алынған үлкен массивтерден тұратын корпустар жасалғанда бұл мәселе бір шешімін таппақ деген ойдамыз.

4.7. Аннотацияланған корпустардың ғылыми-практикалық маңызы

Бүгінгі таңда тіл ғылымын басқа да ғылым салаларымен кең астастықта қарастыру мәселесі өзекті мәселеге айналып отыр. Қазақ тіл білімінде 10-15 жылдың ішінде психолингвистика, лингвомәдениеттану, әлеуметтік лингвистика, когнитивтік лингвистика сияқты басқа ғылым салаларымен тоғысатын тіл білімінің жаңа бағыттары бойынша көптеген ғылыми-зерттеу жұмыстары жүргізіліп келеді. Өз бастауын сонау XX ғасырдың 70 жылдарынан бастау алатын компьютерлік лингвистика саласы да бүгінгі таңда тілді автоматтандыру мәселелері бойынша жаңа ғылыми нәтижелерге қол жеткізді. Бір тіл бойынша емес, жалпы түркі тілдері бойынша лингвистикадағы компьютерлік технологияны орта-

лықтандыру мәселесі де тіл мамандары тарапынан қолдау тауып отыр. Егер мұндай түркі тілдеріне ортақ бірнеше тілді қамтитын мәтіндер корпустары жасалатын болса, ол тілді салыстармалы зерттеу, аударма, түрлі екі, үш тілді сөздіктер құрастыру үшін де аса маңызды болмақ. Сондықтан лингвистикалық мәліметтер берілген автоматты түрде жұмыс істейтін мәтіндер корпусын құрастыру ісі бүгінгі таңда кезек күттірмейтін аса өзекті мәселеге айналып, көпшілік қызығушылығын тудыруда.

Корпустардың сапасы, біріншіден, белгіленім түріне, екіншіден, көлеміне, үшіншіден, компьютерлік бағдарламалану сипатына байланысты болып келеді. Корпустар осы үш критерийге толық жауап бере алған жағдайда бір тілдің болмысын толық бейнелейтін «тұлғалы» (репрезентативті) корпус болып саналады. Бұл үш критерийдің соңғы екеуі бағдарламаға салу және электронды пішіндегі мәтіндерді жинақтау ісі бойынша жүзеге асырылса, әрбір деңгейге сәйкес лингвистикалық белгіленімдердің моделін жазу таза ғылыми-шығармашылықты қажет етеді. Ғылыми-зерттеу жұмыстарына қажетті корпустан алынатын мәліметтер оған енгізілген лингвистикалық белгіленімдердің түрі мен сапасына байланысты. Лингвистикалық ақпараттар қаншалықты қамтылса, соғұрлым одан алынатын мәліметтер де соншалықты болады. Сондықтан корпус құрастырушылар белгіленім мәселесін бірінші орынға қояды. Корпус құрастыру барысы да осы сатымен жүргізіледі. Алдымен, белгілі бір лингвистикалық белгіленім түрі жасалады, екіншіден, ол мәтіндер жинақталып, бағдарламаға салынып, алгоритімі жазылады. Үшіншіден, электронды пішінге келтірілген мәтіндерге белгіленімдер қою жұмысы автоматты түрге көшіріледі.

Мәтіндерді жинақтап, олардың құрамындағы әрбір сөзге лингвистикалық ақпарат беру, яғни түрлі белгіленімдерді қою әсіресе тілші-мамандар, тілді зерттеушілер үшін аса қажет. Корпустағы ең маңызды рөл атқаратын лингвистикалық белгіленімдердің сапалылығы оны қолданушылардың жұмыс істеу, ғылыми материалдар жинау ісін жеделдетеді әрі мол тілдік ақпараттармен қамтамасыз етеді. Сондықтан корпусты қолданушының ғылыми-зерттеу жұмысын жеңілдету және жеделдету үшін корпустар мәтіндерін-

дегі әрбір тілдік бірлікке (сөзге немесе сөз тіркесі т.б.) лингвистикалық белгіленім қойылуы керек. Осындай лингвистикалық белгіленімдер қойылған корпустар зерттеушінің қол жұмысын жеңілдетіп, зерттеу жүргізу үшін бұрынғы карточка жинау, оларды сұрыптау сияқты жұмыстардың барлығын дерлік автоматтандырады. Бұрын зерттеушілер қол жұмысы арқылы материал жинақтау ісіне бірнеше ай, жыл уақыттарын жұмсаса, корпус арқылы мұндай көлемдегі материалдарды бірнеше минуттың ішінде тауып алуға болады.

Аннотацияланған корпустар кез келген тұтынушыға көлемді мәтіндерден кез келген сөзді, оның түрленісін немесе грамматикалық форманы тауып алуға мүмкіндік береді. Сонымен қатар мұндай тілдік мәліметтерді корпустың бәрінен емес, белгілі бір кезеңге тән мәтіндерден, зерттеп жүрген ақын-жазушы шығармаларынан немесе белгілі бір стиль (жанр) мәтіндерінен алуға болады. Сондықтан корпус құрастыру жай ғана техникаға, компьютерге ғана қатысты дүние емес, ол кәсіби мамандар (филолог-лингвистер) үшін өте қажет. Қазақ тілінде әртүрлі сипаттағы корпустар жасалатын болса, зерттеуші мамандар қажетті материалды аз ғана уақыттың ішінде жинап алуына болады. Бұл келешек зерттеу жұмыстарының жылдамдығын арттыратын болады. Сонымен қатар біртілді, екі тілді, терминологиялық т.б. сөздіктер құрастыру жұмысын жеңілдетеді т.б. қажеттіліктерді өтейді.

Корпустар бойынша алғашқы жүргізілген лингвистикалық зерттеулер әртүрлі тілдік бірліктердің (элементтердің) қолдану жиілігін есептеумен шектелетін. Негізінен, мұндай элемент ретінде ең алдымен сөздер, сөзформалар, ал басқа жағдайларда – графемалар, морфемалар, сөз тіркестері алынатын. Шындығында, егер қолданыста көлемді мәтіндер корпусы болатын болса, онда ең айқын әрекет – әртүрлі тілдік бірліктердің жиіліктерін есептеу. Статистикалық әдістемелер күрделі лингвистикалық мәселелердің шешімін табуда қолданылады. Мәселен, машиналық аударма, тілдік элементтерді автоматты тану және сөйлеу тілін синтездеу, орфография мен грамматиканы тексеру құралдары және т.б. Айталық, тұрақты сөз тіркестері семантикалық тұрғыдан тұтас мағыналық бірлік ретінде есептеледі, сондықтан мұндай жағдайды лексикография сала-

сында, мәтінді автоматты өңдеу жүйелерін де ескеру аса маңызды. Корпустың материалдары бойынша статистикалық әдістер арқылы көпшілік жағдайда қандай сөздердің тіркесіп қолданатынын анықтап, оларды тұрақты сөз тіркестері қатарына қосуға мүмкіндік туады. Уақыт өте корпустарды әртүрлі лингвистикалық бағыттар шеңберінде пайдалануға болатындай аса күшті ақпараттық ресурс екендігін пайдаланушы қауымы сезіне бастады. Мәселен, корпустар лексикография үшін аса бай ақпарат дерекнамасы болып табылады. Олардың негізінде және компьютерді қолдана отырып, сөздіктер, бұрынғыға қарағанда анағұрлым тез құрастырылады және аз мезгілде оларды қайта қарап шығуға болады. Сонымен, тілдің қазіргі (ағымдағы) күйін белгілегеннен кейін, ол ескіріп те үлгермей жатып, баспадан қайта шығаруға мүмкіндік бар.

Мәселен, Британдық ұлттық корпусты және ағылшын тілінің Банкін құрудың басты мақсаттарының бірі – оларды лексикография саласында қолдану болған. Қазіргі кездегі ағылшын тілінің сөздігін Collins COBUILD English Language Dictionary құрастыру үшін Бирмингем университетінің мәтіндер базасының мәліметтері пайдаланылған болатын. Сөздіктің кіріспесінде бұл сөздіктің тұңғыш рет көлемі бірнеше миллиондай сөзден тұратын ағылшын мәтіндерінің репрезентативті тобын зерттеу арқылы құрастырылғаны жайлы айтылған. Monitor corpora деп аталатын корпустардың ерекше түрі – жаңа мәліметтермен толықтыруға болатын «ашық» корпустар лексикографтарға неологизмдерді және бұрыннан белгілі сөздердің мағыналық өзгерістерін қадағалап тұруға мүмкіндік жасайды.

Лексикографиялық зерттеулер мен семантика саласының зерттеулері тығыз байланыста келеді. Корпустағы қайсыбір тілдік бірліктердің қоршауын бақылау негізінде, қарастырып отырған тілдік бірлікті сипаттайтын нақтылы семантикалық белгілерді анықтауға болады. Сөз бірден бірнеше семантикалық категорияға қатысты болуы жиі кездеседі, сондықтан сөзді қандай да категорияға жататындығы жайлы сөз еткенде, оның тек сол категорияға жақындық дәрежесі ғана сөз болуы керек. Ал ондай дәреже шамасын тек сөздің әртүрлі категорияларға үлестірілу жиілігін есептеу арқылы ғана анықтауға болады.

Корпус бойынша мәтінді автоматты өңдеу жүйесі тестіленеді және әртүрлі лингвистикалық теориялар да тексеріледі. Мысалы, Нидерландының Nijmegen университетінде зерттеліп жатқан формалды грамматикалар мәтіндер корпусы арқылы тексерілген. Грамматиканың негізінде синтаксистік талдаудың компьютерлік бағдарламасы жасалады. Өңдеу нәтижелері грамматиканың қаншалықты толық және дәл мәліметтерді сипаттайтынын байқатады.

Сонымен, корпустар көмегімен түрлі-түрлі тілдік құбылыстарды зерттеуге болады. Мәтіндік базаға қатысты қойылатын кейбір сауалдар мүмкіндігін В.М. Андрющенко төмендегідей келтіреді:

– белгілі бір сөз қолданысының барлық (немесе ерекше типтерінің) контекстері (сөйлем құрылыстары, сөз тіркестері, құбылыстар) қандай?

– белгілі бір автордың немесе белгілі бір жүйенің барлық сөздік қорын шығару.

– мәтіндерден белгілі бір құрылымға қатысты барлық жағдаяттарды жинақтау және т.б.

Сонымен, қазақ тілінің Ұлттық корпусы лексикография саласының дамуына, жазушылар тілін, дара тілдік шығармашылықты зерттеуге қозғаушы күш болары сөзсіз. Сонымен бірге, бұқаралық сөйлеу тілін (публичная речь), баспа тілін, қоғамның соңғы талаптарын реттеуге себепші болатын радио мен теледидар тілдеріне жүйелі түрде зерттеу жүргізуге мүмкіндік туады.

Ұлттық корпус, ең бірінші кезекте, тілші-ғалымдарға сол тілдің лексикасы мен грамматикасын жан-жақты зерттеуге мүмкіндік тудырады. Ал корпустың келесі міндеті тілдің ішкі салалық (лексика, грамматика, тіл тарихы және т.б.) аясына қатысты әртүрлі анықтағыштық рөл атқару. Егер Ұлттық корпуста тілдік бірліктердің статистикалық сипаты да берілетін болса, ондай деректермен әдебиетшілер, тарихшылар және басқа да қоғамдық ғылымдардың сала мамандары пайдалана алады. Әрине, ұлттық корпустың қолдану аясы тілдерді ана тілі немесе шет тілі ретінде оқыту кезінде көбірек байқалады. Сондықтан қазіргі кезде көптеген оқулықтар мен оқу бағдарламалары мәтіндік корпустарға бағышталып құрастырылуда. Мәселен, мағынасы күңгірт сөздер мен грамматика-

лық формалардың қолдану ерекшеліктерін белгілі авторлардың шығармалары бойынша электрондық корпус көмегімен әрі тез, әрі ұтымды тексеруді шетелдік азамат та, оқушы да, оқытушы да, журналист те және жазушы да жүзеге асыра алады.

Лингвистикалық ақпарат беру моделін жасау және оны компьютер жадына енгізу мен оны біроралықты басқару жүйесі арқылы іске қосу (сайтқа салу), соның нәтижесінде оны практикалық қажеттілікке жарату қажеттігі дау туғызбайды. Сондай-ақ теориялық жағынан да маңызы өте зор. Шетел, орыс ғалымдары еңбектерінде корпус жасалғаннан кейін тіпті академиялық сөздіктер мен грамматикаларының қайта құрастырылып, қайтадан жазылып шыққандығы туралы мәліметтер берілген.

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;

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

– автоматты түрде жиілік сөздіктер жасау мүмкіндігі;

– екі тілді (аударма) корпустар негізінде автоматты адарма 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.

Бұрын ғылыми зерттеулер мысал жинау, кәртішкі жазу, жинақтау, топтау сияқты қол жұмыстары арқылы атқарылып, ұзақ уақытқа созылған ғылыми ізденісті қажет еткен болса, тілдік корпустар мұндай жұмыстарды барынша жеңілдететіндігі және де тілші мамандар үшін ғана емес, журналистер, әдебиетшілер, тарихшылар, мәдениеттанушылар, тіпті қалың көпшілік қауым үшін де

ақпарттық-анықтамалық құрал қызметін атқару мүмкіндігі бар екендігі болжанып отыр. Яғни егер қазақ тілінің ұлттық корпусы жасалатын болса, ол дәл қазіргі уақыттағы интернет жүйесі мен ұялы телефондардың т.б. қызметімен барабар қажеттілікті өтейтін болады. Тілдік корпустарды зерттеу мен жасаудың осындай маңыздылығы сала мамандары үшін ғана емес, қоғамдық-әлеуметтік мәселе ретінде жалпы халықтың қызығушылығын тудыруы тиіс.

Ұлттық корпус ғылыми зерттеулердің түр-түрін жүргізуді қамтамасыз етеді: лексикографияға, жасанды интеллектіге, әдебиеттануға, сөйлеу тілін талдау мен жинақтауға және лингвистиканың барлық салаларына қатысты зерттеу түрлері. Сонымен бірге беделді академиялық сөздіктер мен ғылыми грамматикаларды құрастыру да корпустар негізінде жүзеге асады. Ұлттық корпусты пайдаланушылар – әртүрлі саладағы тілшілер, әдебиеттанушылар, тарихшылар және гуманитарлық білім салаларының өкілдері. Ұлттық корпустың ана тілі мен шет тілін оқытуда, оқулықтар мен бағдарламалар құрастыруда маңыздылығы аса зор деуге болады. Мұндай аннотацияланған корпустардың ғылымды дамытуға қосар үлесі аса зор және олар қазіргі жаһандану кезеңіндегі қоғам дамуының бірден бір көрінісі, нәтижесі болып табылады. Корпустық лингвистиканы қазақ тіл білімінің де ерекше саласы ретінде қалыптастырып, әрі қарай жетілдіру қазақ тілші-ғалымдарына көлемді тәжірибелік материалдарды пайдалануға, қажетті деген тілдік деректерді тауып алуға және оларға тиісті өңдеулер жүргізуге мүмкіндік туғызады. Бүгінгі заманда техниканы, компьютердің қарқынды дамуына байланысты қол еңбектерін жеңілдетіп, ондай мәліметтердің бәрін компьютерге алдын ала енгізу арқылы зерттеушілердің қажет мәліметті тауып алуы үшін қор жинақтауына мүмкіндік береді. Сондай-ақ тіл тарихын немесе қазіргі кездегі тілдің дамуын, бағытын анықтауды мақсат еткен зерттеушілер үшін де берері мол. Яғни тілдің қалай дамып жатқаны, қандай тұлғалардың жиі қолданылып жатқанын, қайсыларының қолданысы азайып, сиреп бара жатқанын анықтау үшін де бірден-бір құрал бола алады.

Қысқасы, Ұлттық корпус ғылыми зерттеулердің түр-түрін жүргізуді қамтамасыз етеді: лексикографияға, жасанды интел-

лектіге, әдебиеттануға, сөйлеу тілін талдау мен жинақтауға және лингвистиканың барлық салаларына қатысты зерттеу түрлері. Сонымен бірге беделді академиялық сөздіктер мен ғылыми грамматикаларды құрастыру да корпустар негізінде жүзеге асады. Ұлттық корпусты пайдаланушылар – әртүрлі саладағы тілшілер, әдебиеттанушылар, тарихшылар және гуманитарлық білім салаларының өкілдері. Ұлттық корпустың ана тілі мен шет тілін оқытуда, оқулықтар мен бағдарламалар құрастыруда маңыздылығы аса зор деуге болады.

ҚОРЫТЫНДЫ

Компьютерлік технология шеңберіндегі прогресс өзімен бірге мәтінді автоматты өңдеу жүйесін құру мен жетілдірудің прогресіне (алға басуына) себепші болады. Осының нәтижесінде қолданыстағы нақты компьютерлік бағдарламалар орнына барынша жетілген және көпфункционалды бағдарламалар келіп, өмірге жаңа ұлттық корпустар әкелді. Корпус дегеніміз – әр тілдегі электронды пішінге келтірілген, яғни бір басқару орталығынан автоматты түрде жұмыс істейтін, лингвистикалық ақпараттар берілген мәтіндер жинағы.

Ұлттық (мәтіндер) корпустар құрастырудың әдіс-тәсілдерін, компьютерлік технологиясын корпустық лингвистика саласы зерттейді. Корпустық лингвистика тіл ғылымында жаңа ғылыми бағыт болып саналады. Корпустық лингвистика тіл білімінің жеке саласы ретінде өзімен іргелес жатқан тіл ғылымы пәндерімен жанасып жатады, яғни математикалық лингвистика, дискурстік анализ және автоматтанған лексикография салаларымен тығыз байланысты.

Корпустық лингвистика компьютерлік технологияларды қолдана отырып, лингвистикалық корпустар (мәтіндер корпустарын) жасау мен оны пайдаланудың жалпы ұстанымдарын зерттейді. Корпустық лингвистикада қолданылатын технологиялар электрондық компьютерлерден де көп бұрын пайда болған: олардың көпшілігі тіл білімі ең алғаш «нақты ғылым саласы» немесе эмпирикалық ғылым деп жарияланған XVIII ғасырдың аяғы мен XIX ғасырлардан басталады. Ал корпустық лингвистика дербес сала ретінде У.Фрэнсистің жетекшілігімен 1963 ж. АҚШ-та өмірге келген Браун корпусынан (The Brown Standard Corpus of American English) бастама алады. Браундық корпус құрастыруға қатысты зерттеулердің кеңінен тараған (атышулы) нысаны мен стандартына айналған болатын. Қазіргі кездегі ағылшын тіліндегі корпустар ішінен ең танымалы – Британдық ұлттық корпус. Орыс тілінің бірінші корпусы *Упсальский машинный фонд русского языка* (Upsal'skij korpus russkix tekstov) деп аталып, Швецияның Упсаль университетінде құрастырылған. Орыс тілінің Ұлттық корпусы –

электронды түрдегі орыс тілінің мәтіндер жиынтығына негізделген ақпараттық-анықтамалық жүйе. Алғаш рет бұл корпус 2004 жылдың сәуір айында http://ruscorpora.ru/ атпен Интернет сайтына салынады. Орыс тілі Ұлттық корпусының даму барысы XIX ғ. басынан XXI ғ. бастапқы кезеңін қамтиды деуге болады.

Корпус ұғымы тілшілердің үнемі пайдаланған дәстүрлі картотека қорын жасау жұмысының жетілген жалғасы деуге болады. Корпусқа мәтіндер белгілі бір жүйе бойынша енгізіледі. Корпус мәтіндері жасалып отырған корпустың мақсатына сай белгілі бір стильдер мен жанрлар немесе кезеңдер бойынша теңгерімді мөлшерге сай болуы қажет. Яғни мәтінде жасалып отырған корпус түріне қарай сәйкестендіріледі. Нақтырақ айтсақ, мәселен, корпустар өз ішінде шағын корпустарға бөлінеді: *поэзия тілінің корпусы, проза тілінің корпустары, газет мәтіндерінің корпусы, ғылыми-техникалық мәтіндер корпусы, ұлттық корпус.* Сонымен қатар лингвистикалық белгіленімдер енгізілу сипатына қарай да *морфологиялық белгіленімдер енгізілген корпустар, синтаксистік белгіленімдер енгізілген корпустар* т.б. Ал ұлттық корпус дегеніміз – бір тілдің қолданылу мүмкіндігін бейнелейтіндей дәрежеде жасалған, электронды сипаттағы, түрлі компьютерлік бағдарламалармен қамтамасыз етілген көлемді корпус болып табылады.

Корпустың алға қойған мақсатқа және классификациялау белгілеріне қарай әртүрлі типтерін бөліп көрсетуге болады. Олар 2 тарауда жіктеліп берілді.

Корпус құрастырудың бірден бір шарты – *ауқымдылық* (көлемділік) (представительность) және *теңгерімділік* (репрезентативность). *Репрезентативтік* – барынша жеткілікті және әртүрлі кезеңдердегі мәтіндердің жанрларға, стильдерге, авторларға және т.б. қатысты пропорционал шамамен алынған көлемі, яғни тілдік зерттеу аясын жан-жақты ашатын мәтін көлемі. Бұл әсіресе ұлттық корпустарға қойылатын басты шарт болып табылады. Корпустар құрастыруда оның түріне қарай (ұлттық, стильдік, кезеңдік) мәтіндер таңдалып алынады. Оларды *таңдама мәтіндер* деп атайды. Корпус көлемі сөзқолданыс санымен есептеледі.

Белгіленім жүргізу дегеніміз – мәтін бірліктеріне алдын ала шартты түрде арнайы қабылданған тэгтерді «тіркеп» жазу әрекеті.

Белгіленімнің лингвистикалық типі ішінде мынадай түрлерін ажыратуға болады: морфологиялық, синтаксистік, семантикалық, анафоралық, просодиялық, дискурстік және т.б. Тілдік корпустардың мәтінтүзуші бірліктеріне берілетін лингвистикалық ақпараттар жүйесі тіл ғылымының әрбір саласы бойынша айқындалған тілдік теорияларға негізделеді. Қазақ тілі корпусы бойынша автоматты морфологиялық белгіленім жасайтын программа жасалған. Лингвистикалық белгіленімдердің ішінде компьютер жадына морфологиялық белгіленімдерді енгізу барлық тілдерде ең алдымен қолға алынған. Морфология сөз бөлшектерін қарастыратындықтан, олар тілдік жүйеге түскен. Ал синтаксистік белгіленімдерді компьютер жадына енгізу мәселесі тілдердің барлығында бірдей дәрежеде шешімін таппаған. Қазақ тіл білімінде әзірге синтаксистік белгіленім әзірлемесі жасалу үстінде.

Сөздердің семантикалық топтарға қатысын корпустар мәтіндерінде аннотациялау үшін тіл жүйесіндегі әрбір сөз табының лексика-семантикалық топтарын жіктеп, тілдегі сөздердің барлығын осы топтарға қатысына қарай сипаттап базаға салу керек. Сол үшін әрбір сөз табының лексика-семантикалық топтарын анықтап, сол сөз тобына жататын сөздердің тізімін жасау қажет. Кітаптың екінші тарауында зат есім, сын есім, етістік т.б. сөз таптарының лексика-семантикалық топтары берілді. Болашақта семантикалық белгіленім әзірлемесін жасауда бұл жіктеменің тигізер септігі мол болмақ.

Анафоралық, просодикалық деңгейлердегі белгіленімдер үшін автоматтанған жүйе құру мәселесі әзірше өте күрделі болғандықтан, негізгі белгі-код қою жұмысы «қол күшімен» атқарылады.

Белгіленім қойылған корпустар құрастыру лингвистикалық зерттеулер үшін маңызды болғанмен, ең алдымен корпусқа енгізілетін мәтіндерді сұрыптаудан өткізіп, олар туралы нақты ақпараттар беру қажет. Мұны кейде экстралингвистикалық белгіленімдер деп те атайды, метамәтіндік белгіленім (метатекстовая разметка – метаразметка) деп те қолданылады. Лингвистикалық белгіленімдер ішкі белгіленім деп аталса, метамәтіндік белгіленімдерді сыртқы белгіленім деп бөлетіндер де бар. Олар: *мәтін және автор туралы ақпарат: автор, атауы, жылы, шыққан жері, мәтіннің жанры, тақырыбы, стилі, көлемі т.б.*

Корпус кез келген тілдік деңгейдегі бірліктердің шынайы қолданысын зерттеуге мүмкіндік тудырады, яғни қандай жағдаятта қайсыбір айтылымның жүзеге асқанын ескере отырып, зерттеу жүргізуге болады. Көлемді Ұлттық корпустар және арнайы мақсатпен құрылған корпустар автоматты түрде қажетті бірлікті іздеп табуға және эмпирикалық материалдарды жүйеге келтіруге, арнайы құрастырылған компьютерлік бағдарламалар бойынша үлкен көлемді тілдік массивтерді тез арада өңдеуге мүмкіндік жасайды.

Корпустарды тілдік және сөйлеу бірліктері жайлы неше түрлі анықтамалар мен статистикалық мәліметтер алуға пайдалануға болады. Мәселен, корпустар негізінде сөзформалар, лексемалар, грамматикалық категориялардың қолдану жиілігі жайында, әртүрлі кезеңдерге қатысты олардың қолдану жиілігін және контекстердің өзгеріске ұшырауын бақылап отыруға, тағы да басқа әртүрлі деректерді алуға болады. Белгілі кезеңдегі тілдік деректердің ауқымды жиыны тілдің лексикалық құрамының өзгеріске ұшырау динамикасын зерттеуге, әртүрлі жанрдың және әр автор шығармалары тілінің лексика-грамматикалық сипатына талдау жүргізуге мүмкіндік береді. Сонымен бірге корпустар көпаспектілі лексикографиялық жұмыстар жүргізуде дереккөз ретінде де, мысалы әртүрлі тарихи және замани сөздіктерді дайындауда да зор қызмет атқарады. Корпус деректерін грамматикаларды құрастыру мен нақтылау және тілді оқыту мақсатында пайдалануға болады.

Тілші-теоретиктер корпустарды өз болжамдарын тексеруге және теорияларын дәлелдеуге қажетті құрал ретінде қолданады. Қолданбалы бағытта жұмыс істейтін лингвистер (мұғалімдер, аудармашылар және т.б.) компьютерлік корпустарды тілдерді үйретуге және өздерінің кәсіби мәселелерін шешуге пайдаланады.

Осындай қажеттілікті өтейтін әлемдік дәрежедегі корпустармен парапар Қазақ тілінің ұлттық корпусын құрастыру үшін А.Байтұрсынұлы атындағы Тіл білімі институтында 2009 жылдардан бері Қолданбалы лингвистика бөлімі қызметкерлері белгілі ғалым А.Жұбановтың жетекшілігімен Қазақ тілінің ұлттық корпусын құрастырудың теориясы мен тәжірибесін зерттеп келеді.

Сонымен қатар 2016 жылдан бастап тілші-ғалым Анар Фазылжанованың ғылыми жетекшілігімен «Қазақ тілінің ұлттық корпу-

сын әзірлеу және жасау» жобасы бойынша «Қазақ тілінің жиілік сөздіктері, «Мектепке арналған қазақ тілінің жиілік сөздігі» т.б. қажеттілікке қарай әртүрлі формадағы жиілік сөздіктер құрастыру, соның негізінде тіл үйретуге қажетті лексикалық минимумдар дайындау сияқты өзекті мәселелерді шешуге қатысты ғылыми-практикалық жұмыстар жүргізіліп жатыр.

Тілдік корпустарды құрастыру ісі – үлкен көлемді мәтіндерді қамтитын күрделі жұмыс. Тілдік корпустарды зерттеу мен нәтижесін алу ұзақ жылдарға созылуы заңды құбылыс. Сонымен қатар әрбір ұлт тілінің мұндай корпустарын құрастыру динамикалық үдеріс, ол ешқашан тоқтап қалмайды, үнемі дамытылып отырады.

Корпустарды құрастырудың және қолданудың барлық тілдерге ортақ бірыңғай әдістемесі жоқ, себебі ол тілдер бір-бірінен айырым табады және дәстүрі мен технологиялық процесі де бөлек.

Қазақ тілінің мол тілдік ақпараттар жинақталған ұлттық корпусын жасау және оны қажетке жарату мүмкіндіктерін сыннан өткізу – бүгінгі күні қолданбалы лингвистика саласы бойынша атқарылуға тиісті өзекті мәселе.

Қазақ тіл білімінде корпустық лингвистиканы дамыту қоғам қажеттілігін өтейтін, мемлекеттік тілдің толыққанды дамуы үшін қажетті негізгі талаптардың бірі болып табылады. Тілдік нормалардың сақталуын, тұрақтануын, қалыптануын қамтамасыз етумен қатар кейінгі зерттеушілер үшін осы кезеңдегі тілдің, сөздік құрамның қандай деңгейде болғанынан да, ондағы өзгерістер мен жаңалықтар туралы да мол мағлұмат берері сөзсіз.

ПАЙДАЛАНЫЛҒАН ӘДЕБИЕТТЕР ТІЗІМІ

1. *Finegan E.* LANGUAGE: its structure and use. – N.Y.: Harcourt Brace College Publishers, 2004.

2. Wikipedia – http://en.wikipedia.org/wiki/

3. *McEnery T., Wilson A*. Corpus Linguistics. – Edinburgh: Edinburgh University Press, 2001.

4. *Рыков В.В.* Корпус текстов как реализация объектно-ориентированной парадигмы // Труды Международного семинара Диалог-2002. – М.: Наука, 2002.

5. *Захаров В.П.* Корпусная лингвистика: Учебно-метод. пособие. – СПб., 2005.

6. *Баранов А.Н.* Корпусная лингвистика // Баранов А.Н. Введение в прикладную лингвистику: Учебное пособие. – М.: Едиториал УРСС, 2003. – С. 114 с.

7. *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

8. *Баранов А.Н.* Введение в прикладную лингвистику. – М., 2007.

9. *Гришина Е.А., Савчук С.О.* Национальный корпус русского языка как инструмент для изучения вариативности грамматических норм // Труды международной конференции «Корпусная лингвистика – 2008» 6-10 октября 2008 г. – СПб., 2008. – С. 161-169.

10. Национальный корпус русского языка. – http://ruscorpora.ru

11. *Sinclair J.M.* Preliminary recommendations on text typology. 1996. EAGLES Document EAG-TCWG-TTYP/P. http://www.ilc.pi.cnr.it/EAGLES96/texttyp/texttyp.html

12. *Николаева Ю.В.* Кинетические признаки структуры устного нарратива (корпусное исследование) // Проблемы компьютерной лингвистики: Сборник научных трудов / Под ред. А.А. Кретова. – Вып. 4. – Воронеж, 2010. – С. 193-200.

13. *Крейдлин Г.Е.* Невербальная семиотика. – М.: Новое литературное обозрение, 2002. – 581 с.

14. *Шерстинова Т.Ю.* «Один речевой день» на временной шкале: о перспективах исследования динамических процессов на материале звукового корпуса // Филология. Востоковедение. Журналистика. Серия 9. – СПб., 2009.

15. *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.

16. *Баранов А.Н., Плунгян В.А., Рахилина Е.В.* Путеводитель по дискурсивным словам русского языка. – М., 1993.

17. *Зубов А.В., Зубова И.И.* Информационные технологии в лингвистике: Учеб. пособие. – М.: Издательский центр «Академия», 2004. – 208 с.

18. *Беляева Л.Н.* Лексикографический потенциал параллельного корпуса текстов // Труды международной конференции «Корпусная лингвистика – 2004». – СПб., 2004. – С. 55-64.

19. *Митрофанова О.А., Грачкова М.А., Шиморина А.С.* Автоматическая классификация лексики в параллельных текстах (на материале текстов из Русско-словацкого корпуса параллельных текстов PARUS) // Материалы V Международной научно-практической конференции «Прикладная лингвистика в науке и образовании: лингвистические технологии и инновационная образовательная среда». – СПб.: «ЛЕМА», 2010. – С. 231-235.

20. *Svartvik J., Quirk R.* A corpus of English Conversation. – Lund: Gleerup, 1980.

21. *Johansson S.* Some aspects of the development of corpus linguistics in the 1970s and 1980s // Lüdeling A., Kytö M., eds. Corpus Linguistics. An International Handbook. Volume 1. – Berlin & New York: Walter de Gruyter, 2008. – Pp. 33-53.

22. *Мордовин А.Ю.* К вопросу о жанровой полноценности современных неспециализированных корпусов текстов // Вестник ИГЛУ, 2009. – № 2. – С. 48-52.

23. *Камшилова О.Н.* Лингвистический объект в интерпретации корпусных технологий: в пользу доказательной парадигмы // Материалы V Международной научно-практической конференции «Прикладная лингвистика в науке и образовании: лингвисти-

ческие технологии и инновационная образовательная среда». – СПб.: «ЛЕМА», 2010. – С. 84-96.

24. *Азарова И.В., Алексеева К.Л., Захарова Л.А*. Разметка текстовых фрагментов в корпусе агиографических текстов СКАТ // Труды международной конференции «Корпусная лингвистика – 2006». – СПб: изд-во С.-Петерб. ун-та, Изд-во РХГА, 2006. – С. 16-24.

25. *Герд А.С.* Несколько слов о Специальном корпусе текстов (СКТ) // Труды международной конференции «Корпусная лингвистика – 2006». – СПб.: Изд-во С.-Петерб. ун-та; Изд-во РХГА, 2006. – С. 92-93.

26. *Засорина Л.Н.* (ред.). Частотный словарь русского языка. – М., 1977.

27. *Гарабик Р., Захаров В.П.* Параллельный русско-словацкий корпус // Труды международной конференции «Корпусная лингвистика – 2006». – СПб., 2006. – С. 81-87.

28. *Исаев С.* Қазақ тіл білімінің мәселелері. Вопросы казахского языкознания. – Алматы: Арыс, 2008. – 624 б.

29. *Оразаханова Н.А*. Қазақ тілі (Сөзжасам, морфология). Оқу құралы. – Алматы: Арыс, 2004.

30. *Ысқақов А. Ы.* Қазіргі қазақ тілі. Морфология. – Алматы: Ана тілі, 1991. – 382 б.

31. *Головин Б.Н.* Язык и статистика. – М., 1971. – 191 с.

32. *Копотев М.В., Гурин Г. Б.* Принципы синтаксической разметки Хельсинкского аннотированного корпуса русских текстов Ханко // Интернет

33. *И.С. Чардин.* Лингвистические корпусы с разметкой на основе грамматики зависимостей и их применение при автоматическом синтаксическом анализе 1 // Интернет

34. *Ермекова Т.* Компоненттер құрылысының құрмалас сөйлемнің грамматикалық-семантикалық сипатына қатысы: филол. ғыл. докт. дисс. – Алматы – 2007. – 208 б.

35. *Нұрмұханбетова А.Қ.* Қазақ тіліндегі объектілік қатынастар: филол ғыл. Канд. Дисс. – Астана. – 2010.

36. *Құрышжанов Ә., Сауранбаев Н.Т.* – Алматы – 1974. – 202 б. – 69 б.

37. *Сауранбаев Н.* Қазақ тілі. – Алматы: ОПБ. – 1953. – 220 б. – 98 б.

38. *Әмір Р.С., Әмірова Ж.Р.* Құрмалас сөйлемдер қазақ тілінің грамматикалық функционалдық жүйесінде. – Алматы: Қазақ университеті, 2009. – 122 б. – 39 б.

39. *Нұрпейісова С.К.* Қазақ тіліндегі меншіктілік қатынасты білдіретін грамматикалық амалдар: қалыптасуы, жұмсалуы: филол ғыл. канд. дисс. – Алматы – 2006. –119 б.

40. *Оразов М.* Семантика казахского глагола. (Опыт семантической классификации.): Дис. докт. – Алма-Ата, 1983. – 415 с.

41. *Оразов М.* Қазіргі қазақ тіліндегі қалып етістіктері. – Алматы, 1980.

42. *Сулейменова Э. Д.* Семантические параметры слова в интерлингвальном исследовании (теоретические и прикладные аспекты контрастивного анализа). АДД., М.: 1992.

43. *Ахметжанова З. К.* Принципы сопоставительного исследования казахского и русского языков: Дис. док. – Алма-Ата, 1989. – 288 с.

44. *Дуйсекова К. К.* Семантика глаголов ориентирован-ного положения в пространстве в казахском языке: Автореф. дис. канд. – Алма Аты, 1996. – 26 с.

45. *Дүйсенбекова Л. С.* Қазіргі қазақ тіліндегі бұйрық райдың парадигмасы мен мағыналық құрамы: Канд. дис. автореф. – Алматы, 1998. – 24 б.

46. *Курамысова К. Н.* Семантико-грамматические особенности глаголов речи//Актуальные проблемы общественных наук. – Алма-Ата, 1986, ч. 2. – 206 с.

47. *Кильдибекова Г. А.* Функционально-семантическая категория каузативности в русском языке. Межвузов. науч. сб.: Исследования по семантике. Семантика слова и словасочетания. – Уфа: БГУ, 1984. – с. 8-19.

48. *Қасабекова Қ. Қ.* Қимыл-қозғалыс етістіктердің жасалу ерекшеліктері. – Алматы: Қазақстан мектебі, (4), 1985. – 27 б.

49. *Токтарова Т. Ж.* Семантика глаголов созидания казахского и русского языков: Автореф. дис. канд. – Алма-Аты, 1998. – 25 с.

50. *Шайкенова Ж. К.* Семантика глаголов движения в казахском

и русском языках: Автореф. дис... канд. филол. наук. – Алма-Аты, 1981. – 23 с.

51. *Васильев Л. М.* Речевой смысл предложения как результат взаимодействия языкового и неязыкового знания. – В сб.: Предложение и текст: семантика, прагматика и синтаксис. – Л.: изд. Ленингр. унив. 1988. – с. 5-10.

52. *Зубов А. В.* Вероятностно-алгоритмическая модель порождения текста (семантико-синтаксический аспект). – Минск.: 1985. – 399 с. (докторская диссертация, филолог. наук).

53. *Оразов М.* Қазақ тілінің семантикасы. – Алматы: «Рауан», 1991. – 215 б.

54. *Васильев Л. М.* Семантика русского глагола. М.: Высшая школа, 1981. – 184 с.

55. *Уфимцева А. А.* Типы словесных знаков. – М.: Наука, 1974. – 206 с.

56. *Апресян Ю. Д.* К построению языка для описания синтаксических свойств слова. – В сб.: Проблемы структурной лингвистики. М.: 1973. – с. 279-375.

57. *Вольф Е. М.* Грамматика и семантика прилагательного. – М.: Наука, 1978. – 200 с.

58. *Шрамм А. Н.* Очерки по семантике качественных прилагательных (на материале современного русского языка). – Л.: ЛГУ, 1979. – 134 с.

59. *Үйықбаев И. Қ.* Есімдіктің түрлері. «Қазақ тілі методикасының мәселелері». Алматы. 1969. – 180 б.

60. *Ахабаев Ә.* Есімдіктердің морфологилық құрылымына статистикалық талдау. – В сб.: Статистика казахского текста. Алма-Ата: Наука, 1973. – с. 88-117.

61. *Савчук С.О.* Орыс тілінің ұлттық корпусындағы метамәтіндік белгіленімдер: базалық ұстанымдары мен негізгі функциялары // Орыс тілінің ұлттық корпусы: 2003-2005. Нәтижелері және болашағы. – М., 2005.

62. *Сиразитдинов З.А., Бускунбаева Л.А., Ишмухаметова А.Ш., Ибрагимова А.Д.* Информационные системы и базы данных башкирского языка. – Уфа: Книжная палата РБ, 2013. – 116 с.

63. *Atkins S., Clear J., Ostler N.* Corpus Design Criteria. Literary and Linguistic Computing. – 1992. – Vol. 7, No. 1. – Pp. 1-16.

64. *Kilgarriff A.* Web as corpus // Proc. of Corpus Linguistics 2001 conference (Lancaster University). – Lancaster, 2001. – Pp. 342-344.

65. *Захаров В.П.* Веб-пространство как языковой корпус // Компьютерная лингвистика и интеллектуальные технологии: Труды международной конференции «Диалог-2005» (Звенигород, 1-6 июня 2005 г.) – М., 2005. – С. 166-171.

66. *Перцов Н.В.* О роли корпусов в лингвистических исследованиях // Труды международной конференции «Корпусная лингвистика – 2006». – СПб.: Изд-во С.-Петерб. ун-та; Изд-во РХГА, 2006. – С. 318-331.

67. *McWhinney B.* The CHILDES Project: Tools for Analyzing Talk. – Mahwah, NJ. Lawrence Erlbaum Associates. Third Edition, 2000. – Vol. 1.

68. *Kytö M., Rissanen M.* A language in transition: The Helsinki Corpus of English texts, ICAME Journal, 1992. – 16: 7-27.

69. *Stenström A-B., Andersen, G.* More trends in teenage talk: A corpus-based investigation of the discourse items cos and innit // C. Percy, C. Meyer & I. Lancashire (eds). Synchronic corpus linguistics. Amsterdam: Rodopi, 1996. – Pp. 189-203.

70. *Meyer Ch. F.* English Corpus Linguistics: An Introduction. Cambridge: Cambridge University Press, 2002. – Xvi + 168.

71. *Biber D., Conrad S., Reppen R.* Corpus Linguistics. Investigating language structure and use. Cambridge University Press, 1998.

72. *Хохлова М.В.* Экспериментальная проверка методов выделения коллокаций // Slavica Helsingiensia 34. Инструментарий русистики: Корпусные подходы. – Хельсинки, 2008. – С. 343-357.

73. ЛЭС – Лингвистический энциклопедический словарь / Под ред. В.Н. Ярцевой. – М.: Сов. энциклопедия, 1990.

74. *Демьянков В.З.* Англо-русские термины по прикладной лингвистике и автоматической переработке текста. Вып. 2. Методы анализа текста // Всесоюзн. центр переводов. Тетради новых терминов, 39. – М., 1982.

75. *McCarthy M.J.* Discourse Analysis for Language Teachers. – Cambridge: Cambridge UP, 1991.

76. *Leech G.* The Distribution and Function of Vocatives in American and British English Conversation // In Hasselgård and Oksefjell eds., 1999. – Pp. 107-120.

77. *Маркасова Е.В.* Риторическая энантиосемия в корпусе русского языка повседневного общения «Один речевой день» // Компьютерная лингвистика и интеллектуальные технологии. Выпуск 7 (14). По материалам ежегодной международной конференции «Диалог» (2008) / Гл. ред. А. Е. Кибрик. – М., 2008. – С. 352-355.

78. *Крейдлин Г.Е.* Голос и тон в языке и речи // Язык о языке / Отв. ред. Н.Д. Арутюнова. М.: Языки русской культуры, 2000. – С. 453-501.

79. *Жұбанов А.Қ.* Основные принципы формализации содержания казахского текста. – Алматы, 2002. – 250 с.

80. *Хомский Н.* Формальные свойства грамматики // Кибернетический сборник. – М., 1966. Вып 2.

81. *Беланогов Г.Г., Калинин Ю.П., Позняк М.В., Хорошилов А.А., Яфаева Г.М.* Алгоритмы многоступенчатого морфологического анализа русских слов // НТИ, Сер. 2. – 1983., № 1, – с.6-10.

82. *Бектаев Қ.* Статистико-информационная типология тюркского текста. – Алма-Ата: Наука, 1978.

83. *Сиразитдинов З.А.* Моделирование грамматики башкирского языка. – Уфа, 2006.

84. *Зализняк А.А.* Грамматический словарь русского языка. Словоизменение. – 1980. – 880 с.

85. *Исхакова Х.Ф.* Модель татарского спряжения // Синхронно-сопоставительный анализ языков разных систем. – М., 1971.

86. *Мырзабеков С.* Қазіргі қазақ тіліндегі етістіктердің құрылымына статистико-лингвистикалық талдау: филол.ғыл.канд. дисс. – Алматы, 1973. – 242 б.

87. *Жаңабекова А.* Сөз формаларын жасаудағы сөздердің функциялық ерекшеліктері: филол. ғыл. канд. дисс. – Алматы, 2001.

88. *Жұбанов А.* Қолданбалы тіл білімінің мәселелері. – Алматы, 2008. – 640 б.

89. Қазақ әдеби тілінің сөздігі. – Алматы, 2011.

90. Қазақ тексінің статистикасы. Жинақ. – Алматы: Ғылым, 1973. – 731 б.

91. *Жаңабекова А., Карбозова Б.* Тілдің морфологиялық жүйесін автоматты түрде бөлшектеуге қажетті негізсөздер сөздігін жасаудың лингвистикалық проблемалары // «Білім беру жүйесінде кәсіби бағытты жүзеге асырудың өзекті мәселелері» атты республикалық ғылыми-тәжірибелік конференция материалдары. – Алматы: Абай ат-ғы Қаз. ұлт. пед. университеті, 2013., 29 мамыр. – 33-37 бб.

92. *Шаяхметов Қ.* Екі функциялық аффикстер: филол. ғыл. Канд. дисс. – Алматы, 1973.

93. *Насилов В.М.* Аффиксы включения //Сб. Вопросы языка и литературы стран востока. – М. 1958.

94. *Ганиев Ф.А.* «О синтетических ианалитических падежах в татарском языке. // Сб.Вопросы тюркологии. – Казань, 1970.

95. *Баскаков Н.А.* Историко-типологическая морфология тюркских языков. – М.:Наука, 1979.

96. *Хабичев М.А.* Именное словообразование и формообразование куманских языках. – М.:Наука, 1989. – 217 с.

97. *Белбаева М.* Қазақ тілінің омонимдер сөздігі. – Алматы: Мектеп, 1988. – 192 б.

98. *Аханов К.* Тіл біліміне кіріспе. – Алматы, 1962.

ҚОСЫМША

Корпустық лингвистика мәселелері туралы А.Байтұрсынұлы атындағы Тіл білімі институты Қолданбалы тіл білімі бөліміннің ғылыми жобалары бойынша жарық көрген мақалалар тізімі:

1. *Жұбанов А*. Қазақ сөзінің «Тіл – қазына» деректер базасы және оның теориялық негіздері. База данных «Тіл – қазына» казахского слова и ее теоретические основы. – Алматы: «Арыс» баспасы, 2009. – 304 бет.

2. *Жұбанов А*. Компьютер қазақ сөзін жатсынбайды. – Алматы, «Ана тілі», №17, 30 сәуір-6 мамыр, 2009 жыл.

3. *Жұбанов А*. Корпустық лингвистика – қазақ тіл білімінің жаңа бағыты // Тілтаным. Языкознание. ҚР БжҒМ А.Байтұрсынұлы атындағы Тіл білімі институты, сәуір-мамыр-маусым. - Алматы.2009 жыл, №2 (34) - 3-11 б.

4. *Жұбанов А*. Қазақ тілі мәтіндер корпусының компьютерлік базасын құрудың алғышарттары // ҚР ҰҒА-ның академигі, ф.ғ.д,, проф. Р.С.Сыздықтың 85 жылдығына орай өткізілген «Әдеби тіл және қазақ тілінің өміршеңдігі» атты ғылыми-теориялық конференцияның материалдары. –Алматы, 2009. –175-179 бб.

5. *Жұбанов А*. Проф. Қ.Жұбанов зертттеулері – қолданбалы тіл білімінің жаңа саласына ғылыми бағыт-бағдар // «Жұбанов тағылымы VII» халықаралық ғылыми конференция материалдары (Қазақ тілінің теориялық негізін қалаған лингвист, қоғам қайраткері, профессор Қ.Қ.Жұбановтың 110 жылдық және белгілі лингвист ғалым, профессор Е.Қ.Жұбановтың 80 жылдық мерейтойларына арналған). I том. –Ақтөбе, Қ.Жұбанов атындағы АқМУ РББ, 2009. –28-31-бб.

6. *Жұбанов А*. Проф. Қ.Жұбанов және қолданбалы тіл білімінің жаңа бағыты // Қазақ тіл білімінің өзекті мәселелері / Жауапты ред. С.Құлманов. –Алматы: «Кие» лингвоелтану инновациялық орталығы», 2009. –27-32-бб.

7. *Жұбанов А*. Проф. Қ. Жұбанов және қолданбалы тіл білімі-ндегі формалды модельдер // Қ.Жұбанов тағылымы: ғылыми са-

бақтастық және жаңа бағыттар: республикалық ғылыми-практикалық конференция материалдарына арналған мақалалар жинағы. – Алматы: «Ан-Арыс», 2009. –70-76-бб.

8. *Жұбанов А*. Қазақ лексикографиясының «Тіл – қазына» атты электрондық базасы жайында // Электрондық-ақпараттық кеңістіктегі қазақ тілі: терминология және тіл мәдениеті. ҚР Мәдениет және ақпарат министрлігінің Тіл комитеті мен ҚР Білім және ғылым министрлігі Ғылым комитетінің А.Байтұрсынұлы атындағы Тіл білімі институты ұйымдастырған республикалық ғылыми-теориялық конференция материалдары. –Астана, 2009.

9. *Жұбанов А*. Қазақ әдеби тілінің электрондық корпусын түзудің теориялық бастаулары // Тіл және мәдениет: тілдің антропоөзектік парадигмасы. Профессор Ж.А.Манкееваның 60 жылдығына арналған республикалық ғылыми-теориялық конференцияның материалдары. –Алматы: А.Байтұрсынұлы атындағы Тіл білімі институты, 2010. – 191-197-бб.

10. *Жұбанов А*. Мәтінді автоматты өңдеу жүйесінің құрылымы жайында // Түркі тілдерінің грамматикасы: қазіргі ахуалы мен болашағы. Профессор ҚазКСР ҒА академигі Нығмет Тінәліұлы Сауранбаевтың туғанына 100 жылдығына арналған халықаралық ғылыми-теориялық конференция материалдары. – Алматы: А.Байтұрсынұлы атындағы Тіл білімі институты, 2010. – 159-166 бб.

11. *Жұбанов А*. Мәтінді автоматты өңдеу жүйесіндегі лингвистикалық ақпарат базасы // Қазіргі грамматикалық зерттеулердің негізгі бағыттары. Профессор ҚазКСР ҒА корр.-мүшесі Ахмеди Ысқақұлы Ысқақовтың туғанына 100 жыл толуына арналған халықаралық ғылыми-теориялық конференция материалдары. – Алматы: А.Байтұрсынұлы атындағы Тіл білімі институты, 2010 (6 комп. бет, баспада).

12. *Жұбанов А*. Мәтіндер корпусын түзудің теориялық және практикалық мәселелері жайында // «Казахский язык в контексте языковой политики стран Евразийского континента» атты Халықаралық симпозиум материалдары. – Көкшетау: Ш. Уәлихан атындағы Көкшетау мемлекеттік университеті, 2010 (С.Құлмановпен бірге, 6 комп. бет, баспада).

13. *Жұбанов А*. Тәуелсіздік құндылықтарының бірі ретіндегі қазақ тілінің мәтіндер корпусын жасаудың теориялық негіздері // Мемлекеттік тіл – тәуелсіздік кепілі. – Алматы: Дайк-Пресс, 2011.– 45-81 бб.

14. *Жұбанов А*. «Тіл – қазына» компьютерлік базасы – Тәуелсіздік Қазақстан жағдайындағы ұлттық терминқор қалыптастырудың негізі // Тәуелсіздік Қазақстан: Қазақ тілінің қоғамдық-әлеуметтік қызметі. – Алматы: А.Байтұрсынұлы атындағы Тіл білімі институты, «Нұр-Принт-75», 2011. – 207-237 бб.

15. *Жұбанов А*. Тәуелсіздік кезеңіндегі қазақ қолданбалы лингвистикасы және «Тіл – қазына» атты компьютерлік қор // Тәуелсіздік кезеңіндегі қазақ тіл білімі. – Алматы: А.Байтұрсынұлы атындағы Тіл білімі институты, «Қазақ энциклопедиясы», 2011. – Б.165-195.

16. *Жұбанов А*. Мемлекеттік тілдегі ұлттық «Тіл – қазына» деректер базасының теориялық негіздері // Мемлекеттік тіл – ұлттық бірегейліктің негізі. – Алматы: А.Байтұрсынұлы атындағы Тіл білімі институты, «Қазақ энциклопедиясы», 2011. – Б. 291-318.

17. *Жұбанов А*. Қазақ лексикографиясының «Тіл – қазына» атты электрондық базасы жайында // Қазақ тілінің тарихы мен түркітану мәселелері /Жауапты редактор ф.ғ.д., профессор М.Малбақов. – Алматы: А.Байтұрсынұлы атындағы Тіл білімі институты, 2011. – Б. 225-228.

18. *Жұбанов А*. Қазақ тілі корпусындағы зат есім сөздерге лексико-морфологиялық деңгейдегі белгіленім қою бағдарламасы // Қазақ тілінің тарихы мен түркітану мәселелері /Жауапты редактор ф.ғ.д., профессор М.Малбақов. –Алматы: А.Байтұрсынұлы атындағы Тіл білімі институты, 2011. – Б. 228-233.

19. *Жұбанов А*. А.Байтұрсынұлы атындағы Тіл білімі институтында 50 жылда қалыптасқан қазақ тілінің жаңа ғылыми-инновациялық қолданбалы саласы жайында // «Тіл білімі институты: 50 жыл» /А.Байтұрсынұлы атындағы Тіл білімі институтының 50 жылдығына арналған жинақ. – Алматы, 2011. – 19 б.

20. *Жұбанов А*. Түркі тілдерінің көпаспектілі компьютерлік қорын құру қажет // А.Байтұрсынұлы атындағы Тіл білімі институтының 50 жылдығына арналған «Қазақ тіл білімі: тарих, тәжірибе, болашақ» атты халықаралық ғылыми-теориялық конференция материалдары. – Алматы, 2011. – 10 б.

21. *Жұбанов А*. Қазақ тілінің аннотацияланған мәтіндер корпусындағы етістік сөздерге лексика-морфологиялық белгі-код қоюдың (белгіленім) алғышарттары // Тілтаным. – Алматы, 2012. – № 1. – 18-25 бб.

22. *Жұбанов А*. Семантические классы существительных использованных в модели порождения казахского текста // «XXI ғасыр ғылымы және білім берудегі жаңа ізденістер» тақырыбындағы халықаралық ғылыми-әдістемелік конференция материалдары. – Шымкент, 2012. – 112-119 бб.

23. *Жұбанов А*. Қазақ тіліндегі аннотацияланған мәтіндер корпусындағы сын есімдердің лексика-морфологиялық құрылымына белгі-код қою бағдарламасы // «Қ.Есенов және қазақ грамматикасының өзекті мәселелері» атты халықаралық дөңгелек үстел материалдары. – Алматы, 2012. – 41-49 бб.

24. *Жұбанов А*. Тұрақты сөз тіркестерін кодтау және модельдеу мәселелері «Жоғарғы оқу орнында жастарға заманауи білім ментәрбие берудің өзекті мәселелері» атты халықаралық ғылыми-тәжірибелік конференцияның материалдары. – Шымент, 2012. – 261-264 бб.

25. *Жұбанов А*. Мәтіндер корпусы – лингвистер үшін ас күшті ақпараттық ресурс //«Фонетика ғылымдарының өзекті мәселелері: даму көздері мен болашағы» атты халықаралық ғылыми-теориялық конференция материалдары. – Алматы, 2012. – 186-193 бб.

26. *Жұбанов А.Қ*. Қазақ тілінің мәтіндер корпусы – қазақ тілін зерттеудің жаңа (парадигмасы) үлгісі // Тілтаным. – Алматы, 2013. № 3.

27. *Жұбанов А.Қ*. Қолданбалы лингвистика: қазақ тілінің статистикасы // Монография. – Алматы, 2013. – 209 б.

28. *Жұбанов А.Қ.* Компьютерлік лингвистикаға кіріспе // Монография. – Алматы, 2013. – 204 б.

29. *Жұбанов А.Қ.* «Қазақ әдеби тілінің жиілік сөздігін» құрастырудың теориялық негіздері // Тілтаным. – Алматы, 2013. № 2.

30. *Жұбанов А.Қ.* Национальный компьютерный фонд казахского языка. – «Тіл-қазына» // «Актуальные проблемы диалектологии языков народов Росии // XIII международная конференция, посвященная 70-летию известного венгерского тюрколога, башкироведа, этнографа, фольклориста, дипломата иожефа Тормы //– Уфа, 2013. – 191-195-бб.

31. *Жұбанов А.Қ. Мадиев С.Ш.* Қазақ тілінің мәтіндер корпусы – қазақ тілін зерттеудің жаңа (парадигмасы) үлгісі // «Актуальные проблемы диалектологии языков народов Росии // XIII международная конференция, посвященная 70-летию известного венгерского тюрколога, башкироведа, этнографа, фольклориста, дипломата иожефа Тормы // – Уфа, 2013. - 196-199-бб.

32. *Жұбанов А.Қ., Жаңабекова А., Құлманов С.* Филологтар қауымдастығы мен корпустық лингвистика орталығын құру – қазақ тілінің ұлттық корпусын жасаудың алғышарты //»Түркі тілдерін компьютерлік өңдеу» атты 1-халықаралық конференция материалдары. – Астана, 2013., 3-4 қазан. – 60-69-бб.

33. *Жубанов А.К., Кабетенов А.Т.* Казахская компьютерная лингвистика в период независимости Республики Казахстан // Сборник:»Проблемы современной прикладной лингвистики» (ответственный – А.В. Зубов). - Минск: МГЛУ, 2013. декабря. – С. 23-30.

34. *А.К. Жубаноов, А.Фазылжанова.* О концепции по созданию национального корпуса казахского языка (НККЯ) // Материалы Международной научно-практической конференции «контрастивные исследования и прикладная лингвистика». 29-30 октября 2014 года в Минском государственном лингвистическом университете. – Минск. 5 комп. стр.

35. *А.К. Жубанов.* Принципы автоматизации морфологической разметки текстов Национального корпуса казахского языка (НККЯ) // Материалы Международной научно-практической конференции «контрастивные исследования и прикладная лингвистика». 29-30 октября 2014 года в Минском государственном лингвистическом университете. – Минск. – 4 комп. стр.

36. *Жұбанов А.Қ.* Қазақ әдеби тілінің жиілік сөздігі бойынша сөз таптарының ықтималдық моделін құру // Тілтаным. – Алматы. – 2014. № 1 (53). – 16-19 бб.

37. *Жұбанов А.Қ.* Қазақ әдеби тілінің жиілік сөздігі бірліктерінің статистикалық арақатынасы // Ә.Т.Қайдардың 90 жасқа толу мерейтойына орай 2014 жылдың 7-қарашасында А.Байтұрсынұлы атындағы Тіл білімі институтында өтетін «Ғылымдағы ғұмыр: Ана тілі және түркі әлемі» атты халықаралық ғылыми-теориялық конференция материалдары. – Алматы. – 2014 ж. – 6 комп. бет. (Өндірісте).

38. *Жұбанов А.Қ.* Тіл біліміндегі жаңа бағыттардың проф. Құдайберген Жұбановтың еңбектеріндегі көрінісі // Қ.Жұбанов атындағы Ақтөбе мемлекеттік университетінде 2013 жылдың 4-5 наурызында өткен дәстүрлі «Жұбанов тағылымы – 8: Қазақстан түркі өркениеті мен тарихы тоғысынды» атты халықаралық ғылыми конференция материалдары. – Ақтөбе. – 2014. – 6-14 бб.

39. *Жұбанов А.* Принципы автоматизации морфологической разметки текстов Национального корпуса казахского языка (НККЯ) // Контрастивные исследования и прикладная лингвистика : материалы междунар. науч. конф. –Минск: МГЛУ, 2015. – 227 с.

40. *Жұбанов А.* О концепции по созданию национального корпуса казахского языка (НККЯ) // Контрастивные исследования и прикладная лингвистика : материалы междунар. науч. конф. – Минск: МГЛУ, 2015. – 227 с.

41. *Жұбанов А.* Қазақ тілінің Ұлттық корпусының коммуникативтілік пен тұтастырушылық мүмкіндіктерін арттыру жайында // «X Всероссийский съезд востоковедов с международным участием

«Россия и Восток: взаимодействие стран и народов». – Уфа, 2015, 6-9 октябрь.

42. *Жұбанов А*. Мәдени құндылықтар ретіндегі қазақ тілінің мәтіндер корпусы және метабелгіленім үдерісі // «Х Всероссийский съезд востоковедов с международным участием «Россия и Восток: взаимодействие стран и народов». – Уфа, 2015, 6-9 октябрь.

43. *Жұбанов А*. Қазақ тілінің Ұлттық корпусындағы автоматтанған лексикографиялық базасының кейбір теориялық мәселелері // «Қазақ лексикографиясы: тарихы, тәжірибесі, болашағы» атты халықаралық ғылыми-теориялық конференция. –Алматы: «Елтаным баспасы», 2015.

44. *Жұбанов А*. Тіл мен мәдениеттің лексика-фразеологиялық жүйесін қазақ тілінің Ұлттық корпусы негізінде кешенді зерттеу/ «Қазақ лексикографиясы: тарихы, тәжірибесі, болашағы» атты халықаралық ғылыми-теориялық конференция. –Алматы: «Елтаным баспасы», 2015. – 177-181-бб.

45. *Жұбанов А*. Қазақ тілінің Ұлттық корпусы және метабелгіленім мәселесі (Национальный корпус казахского языка и проблемы метаразметки) // Тілтаным. – 2015. – №1. – 23-32-бб.

46. *Жұбанов А*. Қазақ әдеби тілінің жиілік сөздігі бірліктерінің статистикалық арақатынасы / Тілтаным. №2.– Алматы, 2015. – 38-43-бб.

47. *Жұбанов А*. Мәтін тізбесіне мән беретін кез келді // «Ана тілі» газеті. – Алматы, 2015 ж., 14-мамыр. (Серіктес автор А.Жаңабекова).

48. *Құлманов С*. Қазақ тілі мәтіндері корпусының электрондық базасы жайында // Электрондық-ақпараттық кеңістіктегі қазақ тілі: терминология және тіл мәдениеті. ҚР Мәдениет және ақпарат министрлігінің Тіл комитеті мен ҚР Білім және ғылым министрлігі Ғылым комитетінің А.Байтұрсынұлы атындағы Тіл білімі институты ұйымдастырған республикалық ғылыми-теориялық конференция материалдары. –Астана, 2009.

49. *Құлманов С*. Корпусқа енгізілетін мәтіндердегі сөздерді морфологиялық талдау және оларға шартты белгілер қою прин-

циптері // Тіл және мәдениет: тілдің антропоөзектік парадигмасы. Профессор Ж.А.Манкееваның 60 жылдығына арналған республикалық ғылыми-теориялық конференцияның материалдары. –Алматы: А. Байтұрсынов атындағы Тіл білімі институты, 2010. –217-222-бб.

50. *Құлманов С.* Мәтіндер корпусын түзудің теориялық және практикалық мәселелері жайында // «Казахский язык в контексте языковой политики стран Евразийского континента» атты Халықаралық симпозиум материалдары. – Көкшетау: Ш. Уәлихан атындағы Көкшетау мемлекеттік университеті, 2010 (А.Жұбановпен бірге, 6 комп. бет, баспада).

51. *Құлманов С.* Корпусқа енгізілетін мәтіндердегі функционалды қосымшалар жайында // Түркі тілдерінің грамматикасы: қазіргі ахуалы мен болашағы. Аса көрнекті тіл маманы, филология ғылымдарының докторы, профессор, ҚазКСР ҒА академигі Нығмет Тінәліұлы Сауранбаевтың туғанына 100 жыл толуына арналған халықаралық ғылыми-теориялық конференция материалдары // Жауапты редактор Ш.Құрманбайұлы. – Алматы, 2010. – 194-201-бб.

52. *Құлманов С.* Қазақ тілі мәтіндері корпусын түзудің теориялық және практикалық мәселелері // Қазіргі білім берудегі ізгілендіру үдерістері және жаңаша оқытудың өзекті мәселелері. ХҚТУ профессоры Ө. Танабаевтың 70 жылдығына арналған Халықаралық ғылыми-теориялық конференция материалдары. I бөлім. – Шымкент, 2010. – 220-226-беттер.

53. *Құлманов С.* Қазақ тілі мәтіндері корпусы – қазақ тілінің жиілік сөздіктерін, грамматикалық сөздігін, екітілді аударма сөздіктер жасаудың негізі // А.Байтұрсынұлы атындағы Тіл білімі институтының «Тілтаным» журналы. – Алматы, 2011. №2.

54. *Құлманов С.* Мәтіндер корпусы негізінде қазақ тілінің грамматикалық сөздігін түзудің практикалық аспектілері // ҚР ҰҒА Хабарлары. – Алматы, 2011. №2.

55. *Жұбаева О.* Қазақ тілінің ұлттық корпусын түзуде қойылатын талаптар //А.Қ.Қошановтың 75 жасқа толуына арналған

«Ақпараттық инновациялық парадигмасының қалыптасуы жағдайында Қазақстанның қоғамдық ғылымдары» атты халықаралық ғылыми конференция материалдары. – 485-491 бб.

56. *Жұбаева О.* Қазақ тілінің ұлттық корпусын түзудің міндеттері // Шымкентте өткен республикалық конференция материалдары.

57. *Жұбаева О.* Корпустық лингвистиканың теориялық және практикалық құндылығы // Профессор Қ.Жұбановтың 110 жылдығына арналған «Қазақ тілінің өзекті мәселелері» атты ғылыми-теориялық конференция материалдары.

58. *Жұбаева О.* Қазақ тілінің ұлттық корпусы: мақсаттар мен күтілетін нәтижелер // Мемлекеттік тіл және әдебиет. 2010. №5.

59. *Жаңабекова А.* Қазақ тілінің грамматикалық сөздігі туралы. – «Тілтаным», №2. Алматы, 2009.

60. *Жаңабекова А.* Тілдік корпус ұғымының мәні, сипаттамасы. – «Тілтаным». Алматы, 2009.

61. *Жаңабекова А.* Форма тудырушы қосымшалардың грамматикалық сөздікке қатысы // «Тілтаным», 2010. №2.

62. *Жаңабекова А.* Қазақ әдеби тілінің аннотацияланған мәтіндер корпусына морфологиялық білімдер жүйесін енгізудің практикалық тәжірибелері және оларды оқыту жүйесіне енгізудің ғылыми-әдістемелік негіздері // «Қазақ тіл білімі: жаңа ғылыми парадигмалар мен оқытудың инновациялық технологиялары» атты халықаралық ғылыми-теориялық конференция материалдары. – Алматы, 1912. – 197-201 бб.

63. *Жаңабекова А.* Профессор А.Қ.Жұбанов және корпустық лингвистика мәселелері // «Қазіргі қазақ тіл білімі: қолданбалы лингвистиканың өзекті мәселелері» атты халықаралық ғылыми-теориялық конференция материалдары. – Алматы, 2012. 69-73 бб.

64. *Жаңабекова А.* Қазақ тілі мәтіндеріне морфологиялық белгіленім қоюдың ғылыми-тәжірибелік негіздері // Научно-

практическая конференция, посвященная 10-летию Лаборатории лингвистики и информационных технологий «Языки меньшинств в компьютерных технологиях: опыт, задачи и перспективы». – Уфа,15 декабря, 2012 г.

65. *Жаңабекова А.Ә.* Аннотацияланған корпустардың ғылыми-практикалық маңызы // «Ахмет Байтұрсынұлы – қазақ руханиятының көсемі» атты халықаралық ғылыми-теориялық конференция материалдары. – Алматы, 2013., 19 ақпан. – 216-219-бб.

66. *Жаңабекова А.Ә. Карбозова Б.* Тілдік корпустар жасауға қажетті негізсөздер сөздігін автоматты жолмен алудың практикалық мәселелері // «Білім беру жүйесінде кәсіби бағытты жүзеге асырудың өзекті мәселелері» атты ғылыми мақалалар жинағы. – Алматы: Інжу-Маржан, 2013. – 262 б. – 33-37-бб.

67. *Жаңабекова А.Ә., Карбозова Б.* Сөздердің морфологиялық құрамын автоматты бөлшектеуге қажетті реестр сөздерді өңдеу // Қырғызстан, 2013.

68. *Жаңабекова А.Ә.* Қазақ тіліндегі сөзформалар сөздігін жасаудың ғылыми-теориялық негіздері // Материалы летней научной школы // ҚазМУ. – 2013.

69. *Жаңабекова А.Ә., Алпысбай Ж.* Тілдің морфологиялық жүйесін автоматты түрде бөлшектеуге қажетті негізсөздер сөздігін жасаудың лингвистикалық проблемалары // Тілтаным, 2013. – № 2.

70. *Жаңабекова А.Ә.* Тілдің морфологиялық жүйесін модельдеудегі функционалды қосымшалардың проблемалық мәселелері // Тілтаным, 2013. – № 3.

71. *Жанабекова А.* Роль лингвистической аннотации на опыте создания национального корпуса казахского языка // Сборник:»-Проблемы современной прикладной лингвистики» (ответственный – А.В. Зубов). – Минск: МГЛУ, 2013. декабря. – С. 212-216.

72. *Жаңабекова А.Ә.* Қазақ тілі ұлттық корпусы мәтіндеріне морфологиялық белгіленімдер енгізу талдамасын жасау тәжірибесінен // «Бектаев оқулары – 1: Ақпараттандыру – қоғам дамуы-

ның болашағы» атты халықаралық ғылыми-тәжірибелік конференция материалдары. – Шымкент. – 2014.

73. *Жаңабекова А.Ә.* Тілдік корпустардың ғылыми-зерттеу әлеуеті // «Қазақ лексикографиясы: тарихы, тәжірибесі, болашағы» атты халықаралық ғылыми-теориялық конференция. – 2015., 29 мамыр. – 189-192-бб.

74. *Жаңабекова А.Ә.* Корпустағы метабелгіленімдер талдамасын жасаудың ғылыми-практикалық негіздері // X Всероссийский съезд востоковедов с международным участием «Россия и Восток: взаимодействие стран и народов». – Уфа, 2015 г., 6-9 октября.

75. *Карбозова Б.* Қазақ мәтіндеріндегі омоним сөздерді статистикалық әдіспен зерттеу // А.Байтұрсынұлы атындағы Тіл білімі институтының 50 жылдығына арналған «Қазақ тіл білімі: тарих, тәжірибе, болашақ» атты халықаралық ғылыми-теориялық конференция материалдары. – Алматы: 2011.

76. *Карбозова Б.* Қазақ тілі мәтіндер корпусындағы омонимдер мәселесі //Тілтаным. – Алматы, 2011. – № 4.

77. *Карбозова Б.*Қазақ тіл біліміндегі ұлттық корпустың мәні мен маңызы // «Фонетика ғылымдарының өзекті мәселелері: даму көздері мен болашағы» атты халықаралық ғылыми-теориялық конференция материалдары. – Алматы, 2012. – 198-201 бб.

78. *Карбозова Б.Д.* Пути различения слов омонимов и полисемии // «Актульные проблема диалектологии» конференциясынын жинтағында. – Уфа. – 2014 ж. – 5 комп. бет.

79. Карбозова Б. Корпустық лингвистикадағы метабелгіленім ұғымы туралы түсінік // «Қазақ лексикографиясы: тарихы, тәжірибесі, болашағы» атты халықаралық ғылыми-теориялық конференция. – Алматы, 2015., 29 мамыр. – 260-263-бб. (А.Қожахметовамен бірге).

80. *Карбозова Б.* Пути различения слов омонимов и полисемии // X Всероссийский съезд востоковедов с международным участием «Россия и Восток: взаимодействие стран и народов». – Уфа, 2015 г., 6-9 октября.

81. *Карбозова Б.* **Research of structural and semantic meanings of-homonymous units** // The social sciences, year: 2015 | volume: 10 | issue: 6 | page no.: 1072-1075; doi: 10.3923/sscience. 2015.1072.1075

82. *Барменқұлова А*. Мәтіндер корпусына морфологиялық аннотация жасаудағы статистикалық мәліметтер алу мүмкіндіктері // Қазақ тілін оқутудың заманауи әдіснамасы: үдеріс, сапа, жетістік» атты ғылыми-практикалық конференция материалдары. – Алматы: Елтаным, 2012. – 232-236 бб.

83. *Барменқұлова А*. Қазақ тілі мәтіндер корпусындағы нөлдік форма мәселесі // Тілтаным.– Алматы, 2012. – №1. – 167-172 бб.

84. *Барменқұлова А*. Мәтіндер корпусына морфологиялық аннотация жасаудағы күрделі сөздердің берілуі және аннотациялауда кездесетін қиындықтар // Тілтаным. – Алматы, 2012. – №2.

85. *Бесіров Е.* Корпустық лингвистика және корпустар // Тілтаным. – Алматы, 2012. – №2. – 200-206 бб.

86. *Қожахметова А*. Мәтіндер корпусындағы нөлдік форманың тілдік көрінісі //Қазақ тілін оқытудың заманауи әдіснамасы: үдеріс, сапа, жетістік» атты ғылыми-практикалық конференция материалдары. – Алматы: Елтаным, 2012. – 236-238 бб.

87. *Қожахметова А*. Мәтіндер корпусына морфологиялық аннотация жасаудағы күрделі етістіктердің берілуі //Тілтаным. – Алматы, 2012. – №2.

88. *Қожахметова А*. Корпусқа енгізілетін мәтіндердегі салттылық сабақтылық категориясы // «Қазіргі қазақ тіл білімі: қолданбалы лингвистиканың өзекті мәселелері» атты халықаралық ғылыми-теориялық конференция материалдары. – Алматы, 2012.

МАЗМҰНЫ

Алғы сөз 3
Кіріспе 5

1. КОРПУСТЫҚ ЛИНГВИСТИКАҒА КІРІСПЕ

1.1. Корпустық лингвистиканың негізгі ұғымдары, зерттеу нысаны 12
1.2. Корпустық лингвистиканы өмірге әкелуге себепші болған бағыт: «картотекадан корпусқа»..... 24
1.3. Лингвистикалық корпустардың тарихы 28
1.4. Корпуста қамтылатын мәтіндер және корпустардың репрезентативтігі 51
1.5. Корпустық лингвистика мен дәстүрлі тіл білімінің зерттеу нысандарының айырмашылықтары 57
1.6. Корпустық лингвистиканың тіл білімі салаларымен байланысы..... 58
1.7. Корпус мәтіндерінің электрондық жинақтағы мәтіндерден айырмашылығы 61
1.8. Корпустарды әртүрлі негізде классификациялау 63
1.9. Корпустардың ерекше типтері 74
1.10. Қазақ тілінің ұлттық корпусын құрастыру мәселесі 84

2. КОРПУСТАР ҚҰРАСТЫРУ МӘСЕЛЕЛЕРІ

2.1. Корпус құрастыруды жобалау және технологиялық үдеріс 87
2.2. Дерекнаманы іріктеу..... 90

2.3. Табиғи тілді өңдеудің негізгі рәсімдері: токендеу (токенизация), леммаға ажырату (лемматизация), стемминг, парсинг 92
2.4. Белгіленім ұғымы. Белгіленім түрлері 94
2.4.1. Морфологиялық белгіленім 101
2.4.2. Корпустағы морфологиялық белгіленімдердің тілдік талдаулармен сабақтастығы 105
2.4.3. Зат есімнің лексика-морфологиялық құрылымына ақпараттық белгі-кодты сәйкестендіру бағдарламасы 109
2.4.4. Етістіктің лексика-морфологиялық құрылымына ақпараттық белгі-кодты сәйкестендіру бағдарламасы 114
2.4.5. Сын есімнің лексика-морфологиялық құрылымына ақпараттық белгі-кодты сәйкестендіру бағдарламасы 121
2.4.6. Синтаксистік белгіленім 129
2.4.7. Синтаксистік белгіленімнің синтаксистік қатынас теориясымен сабақтастығы 138
2.5. Семантикалық белгіленім 142
2.5.1. Қазақ тілі лексикасын семантикалық классификациялау ұстанымдары 144
2.5.2. Зат есімнің семантикалық классификациясы 145
2.5.3. Қазақ тіліндегі етістіктердің семантикалық классификациясы 151
2.5.4. Қазақ тілі мәтіндерін туындату моделінде қолданылатын сын есімдердің семантикалық топтары 161
2.5.5. Қазақ тілі мәтіндерін туындату моделінде қолданылатын үстеулердің семантикалық топтары 170
2.5.6. Қазақ тілі мәтіндерін туындату моделінде қолданылатын есімдіктердің семантикалық топтары 172
2.6. Анафоралық белгіленім және просодикалық белгіленім 178
2.7. Экстралингвистикалық белгіленім 179

2.8. Метабелгіленім енгізілген мәтіндерді арнайы компьютерлік бағдарлама бойынша өңдеу 191
2.9. Корпустық лингвистикадағы стандарттау 195

3. КОРПУСТАРДЫ ПАЙДАЛАНУ

3.1. Корпустық менеджерлер.....198
3.2. Сауалдар тілі 200
3.3. Шығарылым интерфейстері.....207
3.4. Лингвистикалық емес корпустардың корпустық менеджерлері (WWW) 209
3.5. Корпустық зерттеулер.....213
3.5.1. Корпустарды пайдаланушылар.....213
3.5.2. Корпустарды пайдалану тәсілдері.....214
3.5.3. Корпустарға негізделген лексикографиялық зерттеулер 216
3.5.4. Корпустарға негізделген грамматикалық зерттеулер 234
3.5.5. Корпустарға негізделген дискурстық зерттеу 241

4. ҚАЗАҚ ТІЛІНІҢ ҰЛТТЫҚ КОРПУСЫН ҚҰРАСТЫРУ ТӘЖІРИБЕСІНЕН

4.1. Сөздердің морфологиялық құрамын автоматты бөлшектеуге қажетті реестр сөздерді өңдеу 257
4.2. Сөзформалар (түрленім) сөздігі – морфологиялық белгіленім қоюдың құралы.....262
4.3. Тілдің морфологиялық жүйесін автоматты түрде бөлшектеуге қажетті негіз сөздер сөздігін жасаудың лингвистикалық проблемалары.....266
4.4. Тілдің морфологиялық жүйесін модельдеудегі функционалды қосымшалардың проблемалық мәселелері.....272

4.5. Қазақ тілінің ұлттық корпусындағы лингвистикалық белгіленім қоюды автоматтандыру.....281
4.6. Қазақ тілінің мәтіндер корпусындағы омонимдер мәселесі.....288
4.7. Аннотацияланған корпустардың ғылыми-практикалық маңызы.....297

ҚОРЫТЫНДЫ 305
ПАЙДАЛАНЫЛҒАН ӘДЕБИЕТТЕР ТІЗІМІ.....310
ҚОСЫМША 318

ЕСКЕРТУЛЕР ҮШІН

ЕСКЕРТУЛЕР ҮШІН

Ғылыми-танымдық басылым

А.Қ. Жұбанов, А.Ә. Жаңабекова

КОРПУСТЫҚ ЛИНГВИСТИКА

Оқу құралы

Компьютерде беттеген: *Сәуле Беласарова*

Басуға 00.08.2016 ж. қол қойылды.
Пішімі 60×84 1/16. Офсеттік қағаз.
Гарнитурасы «Times New Roman».
Шартты баспа табағы 21,0.
Таралымы 500 дана.