Мақала/Статья/Article
ҒТАМР 16.31.02 https://doi.org/10.55491/2411-6076-2025-2-183-193

Еркін Бекжанұлы Бесіров[1*], Айман Әбділдәқызы Жаңабекова[2]
[1*]автор-корреспондент, докторант, Әл-Фараби атындағы Қазақ ұлттық университеті,
Қазақстан, Алматы қ., ORCID: 0000-0002-2291-9654 E-mail: erkin.besirov@mail.ru
[2]филология ғылымдарының докторы, А. Байтұрсынұлы атындағы Тіл білімі институты,
Қазақстан, Алматы қ., ORCID: 0000-0002-6199-7444 E-mail: aiman_miras@mail.ru

ҚАЗАҚ ТІЛІНІҢ ҰЛТТЫҚ КОРПУСЫ МӘТІНДЕРІНІҢ ЖИІЛІК СӨЗДІГІН ҚҰРАСТЫРУДАҒЫ ОМОНИМДЕРДІ ЫҚТИМАЛДЫҚ-СТАТИСТИКАЛЫҚ ӘДІСПЕН АЖЫРАТУ ТЕХНОЛОГИЯСЫ

Аңдатпа. Мақалада қазақ тілінің ұлттық корпусына негізделген жиілік сөздігін құрастыру барысында омонимдерді ықтималдық-статистикалық әдіспен ажырату технологиясы жан-жақты сипатталады. Зерттеу мақсаты – бірдей жазылып, бірақ мағынасы мен грамматикалық қызметі әртүрлі болатын сөздерді (омонимдерді) автоматты түрде дәл ажырату жолдарын айқындау. Қазақ тілінің агглютинативті табиғатына тән морфологиялық күрделілік омонимдерді тану мен өңдеуде елеулі қиындықтар туғызады. Осы мәселені шешу мақсатында тілдік деректер автоматты морфологиялық анализатор арқылы өңделіп, сөздердің леммалық нысандары мен сөз таптары белгіленді. Алайда талдаудың дәлдігін арттыру үшін омонимдерді мәнмәтін негізінде саралап, ықтималдық үлестері арқылы нақты мағынасына сай үлестіретін арнайы кесте жасалды. Мақалада жиілік сөздікті құрастыру кезеңдері, қолданылған алгоритмдер мен мәтіндік стильдер сипатталады. Нақты мысалдар арқылы омонимдердің түрлі сөз таптарында қолданылу жиілігі салыстырылып, сандық мәліметтер негізінде мағыналық үлгілер жасалды. Бұл әдістеме арқылы жиілік реестрінде омонимдердің ықтимал мағыналық үлес салмағы есептеліп, олардың тілдік қолданыстағы үлгілері айқындалды. Жиілік көрсеткіштер мен мәнмәтіндік деректердің ұштасуы тілдік мазмұнның тереңірек танылуына ықпал етеді. Ұсынылған технология омонимдерді автоматты түрде тану мен топтастыруды жетілдіру арқылы жиілік сөздіктер сапасын арттыруға, тіл үйрету құралдарын тиімді әзірлеуге, сондай-ақ жасанды интеллектіге негізделген лингвистикалық жүйелерді құруға айтарлықтай мүмкіндік береді. Сонымен қатар бұл тәсіл лингвостатистика, семантикалық модельдеу, білім беру саласындағы цифрлық құралдар мен корпус ресурстарын интеграциялау секілді бағыттарда кеңінен қолдануға әлеуетті. Зерттеу нәтижелері қазақ тіл білімінің корпус лингвистикасы, қолданбалы лингвистика және цифрлы лексикография бағыттарында тың әдістемелік үлгі ұсынады.

Тірек сөздер: омоним; жиілік сөздік; ықтималдық-статистикалық әдіс; автоматты морфологиялық анализатор; корпус; мәнмәтін

Қаржыландыру көзі: Мақала Қазақстан Республикасы Ғылым және жоғары білім министрлігінің AP19676197 «Абай шығармаларының квантитативтік бейнесі: жиілік сөздік әзірлеу және лингвостатистикалық талдау» тақырыбындағы гранттық қаржыландыру жобасы аясында дайындалды.

Сілтеме жасау үшін: Бесіров Е.Б., Жаңабекова А.Ә. Қазақ тілінің ұлттық корпусы мәтіндерінің жиілік сөздігін құрастырудағы омонимдерді ықтималдық-статистикалық әдіспен ажырату технологиясы. *Tiltanym,* 2025. №2 (98). 183-193-бб.

DOI: https://doi.org/10.55491/2411-6076-2025-2-183-193

Еркин Бекжанович Бесиров[1*], Айман Абдилдаевна Жанабекова[2]
[1*]автор-корреспондент, докторант, Казахский национальный университет имени аль-Фараби,
Казахстан, г. Алматы, ORCID: 0000-0002-2291-9654 E-mail: erkin.besirov@mail.ru
[2]доктор филологических наук, Институт языкознания имени А. Байтурсынулы,
Казахстан, г. Алматы, ORCID: 0000-0002-6199-7444 E-mail: aiman_miras@mail.ru

ТЕХНОЛОГИЯ РАЗЛИЧЕНИЯ ОМОНИМОВ С ИСПОЛЬЗОВАНИЕМ ВЕРОЯТНОСТНО-СТАТИСТИЧЕСКИХ МЕТОДОВ ПРИ СОСТАВЛЕНИИ ЧАСТОТНОГО СЛОВАРЯ ТЕКСТОВ НАЦИОНАЛЬНОГО КОРПУСА КАЗАХСКОГО ЯЗЫКА

Аннотация. В статье подробно рассматривается технология различения омонимов с использованием вероятностно-статистических методов при составлении частотного словаря, основанного на текстах национального корпуса казахского языка. Объектом исследования являются слова, совпадающие по написанию, но различающиеся по значению и грамматической функции. Агглютинативная природа казахского языка создает определённые трудности в распознавании и обработке омонимов. Для решения этой проблемы использован автоматический морфологический анализатор, с помощью которого были лемматизированы слова и определены их части речи. В то же время, для повышения точности анализа разработана специальная таблица распределения значений омонимов на

основе контекстуального анализа и вероятностных оценок. В статье изложены этапы построения частотного словаря, алгоритмы обработки и структура корпусов по стилям. На конкретных примерах сравнивается частотность омонимов в разных грамматических категориях, и на основе статистических данных формируются семантические модели. Совмещение частотных показателей и контекстной информации обеспечивает более точную интерпретацию языковых единиц. Предложенная технология значительно улучшает качество частотных словарей, способствует разработке эффективных учебных материалов по казахскому языку, а также открывает перспективы для применения в лингвистических системах на базе искусственного интеллекта. Кроме того, данный метод актуален для лингвостатистики, семантического моделирования и цифровых образовательных ресурсов. Результаты исследования представляют собой инновационную методологию в области корпусной и прикладной лингвистики, а также цифровой лексикографии.

Ключевые слова: омоним; частотный словарь; вероятностно-статистический метод; автоматический морфологический анализатор; корпус; контекст

Источник финансирования: Статья подготовлена в рамках проекта грантового финансирования AP19676197 Министерства науки и высшего образования Республики Казахстан на тему: «Квантитативная картина произведений Абая: разработка частотных словарей и лингвостатистический анализ».

Для цитирования: Бесиров Е., Жанабекова А.А. Технология различения омонимов с использованием вероятностно-статистических методов при составлении частотного словаря текстов Национального корпуса казахского языка. *Tiltanym*, 2025. №2 (98). С. 183-193. (на каз. яз.)

DOI: https://doi.org/10.55491/2411-6076-2025-2-183-193

Yerkin Bessirov[1*], Aiman Zhanabekova[2]

[1*]Corresponding author, Doctoral student, Al-Farabi Kazakh National University, Kazakhstan, Almaty, ORCID: 0000-0002-2291-9654 E-mail: erkin.besirov@mail.ru

[2]Doctor of Philological Sciences, A. Baitursynuly Institute of Linguistics, Kazakhstan, Almaty, ORCID: 0000-0002-6199-7444 E-mail: aiman_miras@mail.ru

TECHNOLOGY OF DISTINGUISHING HOMONYMS USING PROBABILISTIC AND STATISTICAL METHODS IN COMPILING A FREQUENCY DICTIONARY OF TEXTS OF THE NATIONAL CORPUS OF THE KAZAKH LANGUAGE

Abstract. The article presents a comprehensive overview of a probabilistic-statistical method for distinguishing homonyms in the process of compiling a frequency dictionary based on the Kazakh National Corpus. The research focuses on words that share identical written forms but differ in meaning and grammatical function. The agglutinative structure of the Kazakh language introduces significant challenges in the accurate recognition and classification of homonyms. To address this, an automatic morphological analyzer was employed to lemmatize word forms and assign appropriate parts of speech. In addition, a specialized probabilistic distribution table was developed to disambiguate homonyms based on contextual analysis and frequency estimation. The article details the stages of constructing the frequency dictionary, the applied algorithms, and the stylistic diversity of the corpus texts. Comparative examples demonstrate the frequency of homonyms across different grammatical categories, enabling the creation of semantic usage models supported by statistical data. The integration of frequency indicators with contextual analysis contributes to more precise interpretation of lexical units. The proposed technology enhances the accuracy and reliability of frequency dictionaries, facilitates the development of effective Kazakh language learning materials, and provides a foundation for applications in AI-based linguistic systems. Moreover, the approach holds promise for linguostatistics, semantic modeling, and the integration of corpus resources into digital education. The results of the study offer an innovative methodological framework for corpus linguistics, applied linguistics, and digital lexicography in the context of the Kazakh language.

Keywords: homonym; frequency dictionary; probabilistic-statistical method; automatic morphological analyzer; corpus; context

Source of financing: The article was prepared within the framework of the grant project AP19676197 of the Ministry of Science and Higher Education of the Republic of Kazakhstan on the theme: “The quantitative image of Abay's works: development of frequency dictionaries and linguostatistical analysis”.

For citation: Bessirov, Ye., Zhanabekova, A. Technology of Distinguishing Homonyms Using Probabilistic and Statistical Methods in Compiling a Frequency Dictionary of Texts of the National Corpus of the Kazakh Language. *Tiltanym*, 2025. No. 2 (98). P. 183-193. (in Kazakh)

DOI: https://doi.org/10.55491/2411-6076-2025-2-183-193

Кіріспе

Қазіргі заманда тіл білімінде статистикалық және ықтималдық әдістерді пайдалану кеңінен таралып келеді. Бұл үрдіс, әсіресе, түркі тілдері мен қазақ тілін зерттеу саласында өзекті бола бастады. Үндіеуропа тіл білімінде ертеректе кең қолданыс тапқан бұл әдістер түркітануға, оның ішінде қазақ тіл біліміне XX ғасырдың 60-жылдарынан бастап ене бастады. Мұндай тәсілдер

тілдің құрылымдық ерекшеліктерін, әсіресе агглютинативті жүйенің заңдылықтарын зерттеуге мүмкіндік береді. Сонымен қатар бұл әдістер тілдің ішкі логикасын ашуда, сөздердің қолдану жиілігін есептеуде, грамматикалық құрылымдардың статистикалық үлгілерін құруда да ерекше маңызға ие. Бұл бағыттағы зерттеулер қазіргі тіл біліміндегі цифрлық ресурстарды дамыту, семантикалық өріс пен синтаксистік байланыстарды зерделеу сияқты өзекті мәселелермен тығыз байланысты болып отыр.

Қазақ тіл білімінде жиілік сөздіктерін құрастыру ісі 1970-жылдары қарқын ала бастады. Бұл бағытта Қ. Бектаевтың жетекшілігімен А.Қ. Жұбанов, С. Мырзабеков, А. Белботаев, Ә. Ахабаев сынды зерттеушілер еңбектенді. Осы кезеңде «Абай тілі сөздігінің жиілік сөздігі» (Абай тілі сөздігі, 1968), «М. Әуезовтің 20 томдық шығармалар мәтінінің жиілік сөздіктері» (Бектаев, 1995) сияқты іргелі ғылыми туындылар жарық көрді. Сонымен қатар ертегі мәтіндері, публицистикалық және ғылыми стильдегі мәтіндер, оның ішінде математика саласына қатысты терминдер негізінде жасалған жиілік сөздіктер де тіл білімінің дамуына зор үлес қосты. Алайда бұл бағыттағы жұмыс 1990-жылдардан кейін бәсеңдеп, біршама уақыт тоқтап қалды.

2016 жылы А. Байтұрсынұлы атындағы Тіл білімі институтының «Қолданбалы лингвистика» бөлімі профессор А.Қ. Жұбановтың бастамасымен бұрынғы еңбектерді біріктіріп, жанрлық-стильдік тұрғыдан кең ауқымды мәтіндерді қамтитын «Қазақ тілінің жиілік сөздігін» (Жұбанов, 2016) жарыққа шығарды. Бұл сөздік оқу-әдістемелік, тіл үйрету және лингвистикалық талдау мақсатында кеңінен қолданылып келеді. Сонымен қатар бұл еңбекте жиілік көрсеткіштер негізінде тіл үйренушіге маңызды лексикалық бірліктерді іріктеу мүмкіндігі де қарастырылған. Зерттеушілер жиілік деректерін негізге ала отырып, сөздердің тілдегі белсенділігі мен функционалдық маңызын айқындауға талпыныс жасады.

Жиілік сөздіктерді белгілі бір жазушы шығармашылығына арнап жасау да маңызды зерттеу бағыты болып саналады. Қазақ тілінде осы тұрғыдағы ең ірі еңбектердің бірі – М. Әуезовтің 20 томдық шығармалары негізінде жасалған жиілік сөздік. Сонымен қатар «Абай тілі сөздігі» (Абай тілі сөздігі, 1968) тек жиілік мәліметтермен шектелмей, түсіндірме сөздік қызметін де атқарады. Бұл сөздікте Абай шығармаларында қолданылған сөздер қанша рет кездесетіні туралы статистикалық мәліметтер қамтылған. Бұдан бөлек Қ. Бектаев, А. Белботаев, Қ. Молдабековтің авторлығымен шыққан Ғ. Мүсіреповтің «Кездеспей кеткен бір бейне» повесіне арналған жиілік сөздік (Бектаев, 1973) те қазақ тіліндегі маңызды зерттеулер қатарына жатады.

Қазақ тіліндегі жиілік сөздіктер тек ғылыми зерттеулер үшін ғана емес, тіл үйрету, оқыту әдістемелерін әзірлеу және лингвистикалық талдау жұмыстарында да маңызды рөл атқарады. Олар сөздердің қолданыс жиілігін анықтауға, мәтіндердің құрылымын талдауға, белгілі бір кезеңдегі тілдік өзгерістерді зерттеуге көмектеседі. Сондықтан жиілік сөздіктерді құрастыру ісі алдағы уақытта да өзектілігін жоғалтпай, әрі қарай жалғасын табуы тиіс.

Алайда жиілік сөздіктерді құру барысында кездесетін ең өзекті қиындықтардың бірі – омонимдерді дұрыс ажырату. Омонимдер бірдей жазылып, бірдей дыбысталатын, бірақ мүлде әртүрлі мағынаға ие болатын сөздер болғандықтан, олардың мәтіндік мәнмәтінде дұрыс интерпретациялануы аса маңызды. Омонимдерді дұрыс ажырату қамтамасыз етілмесе, бұл жиілік талдаудың дәлдігіне, лексикалық қордың статистикалық сипаттамаларына және жалпы зерттеу нәтижелерінің сапасына теріс әсер етуі мүмкін.

Заманауи технологиялар мен ықтималдық-статистикалық әдістер омонимдерді ажыратудың тиімді құралдарын ұсынады. Бұл тәсілдер сөздердің мәнмәтіндік ерекшеліктерін, грамматикалық байланыстарын және мағыналық реңктерін ескере отырып, олардың нақты мағынасын анықтауға мүмкіндік береді. Мәтіндер корпусында осы әдістерді қолдану зерттеудің дәлдігін арттырып қана қоймай, автоматтандырылған талдау жүйелерін жаңа деңгейге көтереді.

Қазіргі уақытта мемлекеттік тілдің қолдану аясын кеңейту және оның тиімділігін арттыру мақсатында қолданбалы ғылыми жобаларға ерекше назар аударылуда. Осы бағытта А.Байтұрсынұлы атындағы Тіл білімі институтының ұжымы шығарған «Жалпы білім берудегі қазақ тілінің жиілік сөздігін» (2016) әбден айтуға болады. Себебі бұл жиілік сөздік мемлекеттік тілді оқыту үдерісінде маңызды рөл атқаратын лексикалық минимумдарды қалыптастыру үшін арнайы жасалды. Сөздіктің мазмұнын негізінен мектеп оқулықтарындағы мәтіндер құрады, олардың жалпы көлемі 7 миллион сөзқолданыстан тұрды, бұл – айтарлықтай ауқымды тілдік

дереккөз. Осындай көлемге сәйкес қамтылған сөздік бірліктердің саны да едәуір артты. Жиілік сөздіктерді құрастыру барысында сөздерді түбірлік негізде жүйелеу біршама оңай болғанымен, олардың белгілі бір грамматикалық категорияларға тиесілігін анықтау маңызды лингвистикалық міндет болып табылады. Осы тұрғыда жиілік сөздікте сөз таптарын белгілеу кезінде бірқатар қиындықтар туындады. Әсіресе, бұл мәселе омоним сөздер мен омографтардың әртүрлі мағыналарына байланысты күрделене түсті. Осыған қарамастан, ғалымдар жиілік сөздікті өңдеп, оны тіл үйрету үдерісінде барынша тиімді әрі нәтижелі қолдануға тырысты.

2023-2024 жылдары Тіл білімі институтының «Қолданбалы лингвистика» бөлімінің мамандары қазақ тілінің цифрлы кеңістіктегі белсенді сөздік қорын анықтау мақсатында 30 миллионнан астам сөзқолданысты қамтитын «Қазақ тілінің ұлттық корпусы мәтіндерінің жиілік сөздігін» (https://docs.google.com, 2025) құрастырды. Бұл еңбек – отандық лингвистикадағы ең ірі сандық жоба болып табылады және осы зерттеудің негізін қалайды. Жоба шеңберінде омонимдерді ажырату үшін ықтималдық-статистикалық модель енгізіліп, олардың мәнмәтіндік сипаты негізінде саралау әдісі жасалды. Осы әдістеменің қазақ тіл білімінде алғаш рет қолданылып отырғаны – зерттеудің ғылыми жаңалығы мен өзектілігін дәлелдейді.

Материал және әдістер

Зерттеу материалы ретінде «Абай Құнанбайұлының шығармаларының академиялық толық жинағы» (Абай шығармаларының академиялық толық жинағы, 2020) алынды. Бұл жинақ үш томнан тұрады және жалпы 51207 сөзқолданысын қамтиды. Зерттеу барысында осы еңбек негізінде алынған мәтіндер цифрлық өңдеуге дайындалып, жүйелік-құрылымдық, сипаттамалы, тарихи және лингвостатистикалық әдістер қолданылды. Алғашқы кезеңде жинақтағы үш томдағы мәтіндерің сапасын қамтамасыз ету үшін OCR нәтижелері тексеріліп, UTF-8 форматында стандартизацияланды. Мәтіндер құрылымдық блоктар – өлеңдер, қара сөздер және проза фрагменттері бойынша топтастырылып, жанрлық және стилдік ерекшеліктеріне қарай іріктелді. Бұл топтастыру контекстуалдық факторларды есепке ала отырып, кейінгі талдау кезеңдерінде жан-жақты қарастыруға мүмкіндік берді.

Жүйелік-құрылымдық және сипаттамалы әдістер аясында әрбір мәтіндік блоктың лингвистикалық өлшемдері зерттелді: тақырыптық және семантикалық құрылымдар, стилистикалық элементтер мен олардың жиілік көрсеткіштері айқындалды. Өлең мен қара сөздерде кездесетін нақты форма мен мазмұн бірліктері жан-жақты сараланып, араб, парсы және шағатай кірме сөздердің терминологиялық, философиялық және мәдени-тарихи айырмашылықтары сипатталды. Бұл сатыда талдаудың сапалық аспектісі басым болатынына қарамастан, жиілік мәндерін кейінгі лингвостатистикалық талдауға енгізу мақсатында мәнмәтіндік мазмұндық ерекшеліктер жүйелі түрде тіркелді.

Тарихи әдіс қолданысы Абай шығармаларының алғашқы нұсқалары мен қазіргі академиялық редакциялары арасындағы айырмашылықтарды анықтауға бағытталды. 1909 жылы Петербургте жарық көрген тұңғыш жинақ нұсқасы мен 2020 жылғы академиялық жинақтағы мәтіндердің салыстырмалы талдауы редакциялық түзетулер мен диахроникалық өзгертулерді көрсету үшін жүзеге асырылды. Бұл сатыда мәтіндік өзгерістерді анықтау үшін әрбір бөліктің редакциялық хроникасы қаралды және ескі кітаби лексиканың қолданылу мәнмәтініне әсері бағаланды (Абай шығармаларының академиялық толық жинағы, 2020: 3). Сондай-ақ тарихи мәнмәтінді ескеру барысында сол кезеңдегі тілдік дәстүрлер мен әдеби норма жағдайындағы өзгерістер талданып, кірме сөздердің қабылдану ерекшеліктері зерттелді.

Лингвостатистикалық талдау негізгі әдіснамалық тәсіл ретінде мәтіндегі сөз формаларының жиілігін сандық тұрғыдан бағалауды қамтамасыз етті. Мәтіндер токенизацияланып, морфологиялық талдау жүргізілді. Сөз формалары леммаларға бөлініп, сөз таптары бойынша санатталды. Осы деректер негізінде жиілік кестелері құрылып, әрбір лемма мен сәйкес сөз формасының кездескен саны анықталды. Қорытынды жиілік нәтижелері Excel немесе ұқсас кестелік редакторда өңделіп, жоғары және орта деңгейдегі жиіліктерге қатысты мәнмәтіндер талдауға енгізілді.

Нәтижелерді интерпретациялау барысында лингвостатистикалық және сапалық талдау әдістері қолданылды. Интерпретациялық кезеңде алынған жиілік деректері мен мәнмәтіндік мағыналық ерекшеліктер әдеби-тарихи ортамен салыстырылды. Бұл Абай шығармаларының

тілдік жағынан инновациялық және мәдени-тарихи маңызының терең бейнесін ашады.

Әдебиетке шолу

Омоним (гр. *homos* – біркелкі, onyma – есім) – шығуы жағынан да, мағына жағынан да басқа-басқа, айтылуы және жазылуы бірдей сөздер (https://kk.wikipedia.org, 2025).

Қазіргі таңда компьютерлік лингвистика саласындағы қолданбалы зерттеулердің нәтижелілігі қолжетімді лингвистикалық ресурстарға, әсіресе лексикографиялық еңбектердің сапасы мен қолданысына тікелей байланысты. Соңғы уақытта орыс тіліндегі омонимдерге арналған сөздіктердің жарық көруі белсенді түрде артып келеді. Мысалы, Н.П. Колесниковтың (Колесников, 1980) құрастырған сөздігінде лексикалық омонимдерден бөлек, омоформалар мен омофондар, сондай-ақ омографтар жүйеленген. О.С. Ахманова (Ахманова, 1957) мен О.М. Ким (Ким, 1983) еңбектерінде функционалдық омонимдердің ерекшеліктері қарастырылса, Н.Аношкина (Аношкина, 2000) өзінің интернет-қорында грамматикалық омонимдердің кең көлемдегі жинағын жасауға тырысқан. Ал Т.Ю. Кобзарева (Кобзарева, 2002) өз зерттеулерінде функционалдық омонимдердің 58 түрін ғылыми негізде жіктеп, сипаттап көрсеткен.

Қазақ тіліндегі омонимдерді алғаш рет арнайы зерттеу нысанасына айналдырып, жан-жақты саралаған ғалым – К. Аханов. Омоним мәселесін сондай-ақ Ғ. Мұсабаев (Мұсабаев, 2014), І.Кеңесбаев (Кеңесбаев, 1975), М. Белбаева (Белбаева, 1988) т.б. ғалымдар да зерттеген. К. Аханов омонимдердің таптастырылымында олардың мағыналық жақтарын да, грамматикалық формаларын да ескеріп, құрылымдық ерекшеліктеріне қарай таптастырып, үш үлкен топқа бөледі: лексикалық омонимдер; лексика-грамматикалық омонимдер; аралас омонимдер (Аханов, 1962).

Қазақ тіл білімінде мектеп мұғалімдеріне арналған омонимдер сөздігін алғашқылардың бірі болып М. Белбаева жасап шыққан. Бүгінде «Қазақ тілінің омонимдер сөздігі» осы бағыттағы бірегей еңбектердің бірі саналады (Белбаева, 1988).

Ю.Г. Зеленков, И.В. Сегалович, В.А. Титов секілді Яндекс ғалымдары морфологиялық омонимдерді ажыратудың ықтималдылық әдісін (Зеленков, Сегалович, Титов, 2005) қарастырған.

Нәтижелер және талқылау

Жиілік сөздік қазақ тілінің әртүрлі стильдік қабаттарын толық қамтуы үшін зерттеу материалы ретінде поэзиядан бастап публицистикалық, көркем проза, ғылыми, іскери, сөйлеу және оқулық мәтіндері іріктеліп алынды. Жалпы көлемі 31105791 сөзқолданыстан тұрады. Мұндай кең ауқымды қазақ тілінің ұлттық корпусы, нақты қолданылу жағдайын сипаттауға және сөзформалардың стильдік ерекшеліктерін бір-бірімен салыстыра зерттеуге мүмкіндік береді. *Бірінші кезеңде* сандық өңдеуге дайындау барысында OCR-ден алынған нәтижелер мен электронды нұсқалар тексеріліп, UTF-8 форматында стандартизация жасалды. *Екінші кезеңде* сөзформалар тізбесін (сөздігін) түзу үшін алынған мәтіндерді цифрлардан, орыс сөздерінен, қате қолданыстардан тазарту жұмыстары атқарылды. *Үшінші кезеңде* сөзформалар тізбесі леммаға (түбірге) келтіріліп, оларға сөз таптары қойылды. *Төртінші кезеңде* әліпбилі сөзформалар жиілік сөздігі өңделіп, жиілікті-әліпбилі, кері-әліпбилі түрлері жасалды.

Әр кезеңде қолданылатын әдістер бір-бірін толықтырып, жиілік сөздіктің тілдік практикаға дәл сәйкес келуін және зерттеу нәтижелерінің сенімділігін қамтамасыз етеді. Жиілік сөздік жасауда мәтіндегі сөзформаларды морфологиялық анализатор көмегімен автоматты түрде түбірге келтіріп, яғни сөзформаларды түбір мен қосымшаға бөліп, сөз табына қойдыруға болады. Автоматты морфологиялық анализатор жұмысы екі тілдік база арқылы жұмыс істейді: бірі – қазақ тіліндегі түбір (негізгі және туынды) сөздер тізімі және түрленім (словоизменительные) қосымшалар тізімі. Осы ресурстар арқылы морфологиялық анализатор сөзформалардың түбірін тауып, оларға сөз табын қойып береді. Автоматты анализатор жұмысында да омонимдерді тануда қиындық туындайды. Анализатор да жоғарыда берілген «қара» сөзінің қарады, қараған, қараса т.б. формаларын қосымшаларына қарап, етістік екендігін ажыратады. Өйткені *-са, -ды, -ған* жұрнақтарының етістік тұлғалары екенін табады. Мұндағы *-ды* қосымшасы табыс септік жалғауымен омоформа болғанмен, қара сөзіне *-на* варианты жалғанбағандықтан, табыс септігі емес екенін біліп, бұндағы *-ды*-ның жіктік жалғауының *3-жағы* деп табады. Демек, автоматты анализатор да бағдарламаға берілген тілдік деректер арқылы сөздерге талдау жасай алады. Алайда автоматты талдауда да қолмен талдаудағыдай, омонимнің түбір тұлғада тұрғандарының қай мағынада тұрғаны мен қай сөз табы екенін ажырату қиын. Омонимдерді автоматты ажырату

мәселесі әлем тілдерінде де толыққанды жасалмаған. Қазақ тілінде қазіргі кезде А. Байтұрсынұлы атындағы Тіл білімі институты мамандары жасанды интеллектіге мәнмәтінді түсінуге оқыту үшін лингвистикалық омонимдердің мәнмәтіндік базасын жасап жатыр.

Жиілік сөздік сөзтізбесіндегі (реестрі) түбір сөздердің барлығы бірдей орфографиялық нормаға сәйкес келе бермейді. Өйткені корпусқа салынған мәтіндердің авторлық қолданыстары да кездеседі. Жергілікті сөздер, сөйлеу тілінің элементтері, орфографиялық бейнормада жазылған сөздер кездеседі. Бұндай авторлық ерекшеліктерді жиілік сөздік сөзтізбесінен алып тастамадық. Себебі бұндай ерекшеліктер қалың көпшіліктің тәжірибесінде қалай қолданыс тауып жатыр деген сұраққа жауап бермек. Мәселен, кейбір сөздерді бөлек не бірге жазу мәселесі, фонетикалық варианттарды қолданудағы әртүрлілік – бұлардың бәрі нормалану үдерісінде кездесетін қалыпты жағдайлар. Төменде жиілік сөздік реестріне қандай сөздер алынғанын көрсетпекпіз:

- мәтінде кездесетін сөзформалар (түрленген сөздер) түбір сөзге келтірілді, бұл ретте сөзформалар лемма мен қосымша шегі бойынша бөлініп, сөзтізбеге тек түбірлер (негізгі және туынды) ғана алынды. Сөзтізбедегі сөздер дені негізінен әдеби тіл нормаларына сәйкес келеді;

- қазақша баламасы бола тұра, мәтіндерде жиі кездесетін, көп қолданылатын орыс сөздері де сөзтізбеде берілді. Мысалы: *папа, мама, пиджак, пинцет, посылка, помещик, пикник* т.б.;

- қазақша баламасы жоқ немесе баламасы көп қолданылмайтын орыс / термин (техникалық, әркери) сөздер сөзтізбеге алынды. Мысалы: генерал, полковник, рейка, рулетка, рычаг т.б.;

- сөзтізбеге белгілі бір сала бойынша кең қолданыстағы термин сөздер алынды;

- сөзтізбеден жалқы есімдер (кісі есімдері, жер-су атаулары) алынып тасталды;

- сөйлеу тілінің элементтері сөзтізбеге алынды. Мысалы: *өлең-мөлең, өкпе-сөкпе, су-му, сүт-пұт* т.б.;

- түрлі дыбыстық өзгерістермен жазылып жүрген фонетикалық варианттар алынды. Мысалы: *ойбой-өйбөй-ойыбой-өйбүй; қане-кәне-қәні* т.б.;

- жергілікті сөздер де сөзтізбеде берілді;

- кейбір грамматикалық формаларда тұрғанмен, уақытша трансформацияланып (заттанып, сынданып) тұрған сөздер де сөзтізбеге алынды. Мысалы: *өзгеретіндік, өлтіруші, өскендік, пішініндей, өлердегі* т.б.;

- қазақ тіліне етене еніп кеткен немесе діни мәтіндерде, сондай-ақ көпшілік қолданысында жиі кездесетін араб-парсы тілінен енген кірме сөздер сөзтізбеге алынды;

- мәтіндерде сындырылып қолданылған сөздер сөзтізбеге сол күйінде берілді. Мысалы: *пашпырт, кепке, казет, кәменес, кәмпеске, кәртішке, кінешке, лектір, пәкет* т.б.;

- етіс тұлғалы етістіктер түбірге келтірілмеді, етіс формаларымен түрленген нұсқада алынды.

Жиілік сөздік құрастырудың үшінші кезеңінде сөзформаларды түбірге келтіру мен сөздерге сөз табын қою автоматты морфологиялық анализатор арқылы жүзеге асырылды. Автоматты талдаудан кейін сөзтізбедегі сөз түбірлері мен оларға қойылған сөз табы мұқият тексерілді.

Келесі кезеңдегі күрделі мәселе – омоним ажырату жұмысы статистикалық зерттеулердегі ықтималдық теориясына негізделді. Ізделген сөз бойынша конкорданска шыққан мысалдардағы омонимдердің семантикасына қарай (қай мағынада қолданылып тұрғанына қарай) пайызы анықталып, жиілік саны есептеліп шығарылды.

Сөзтізбені тазалау, омонимдерді ажырату мәселесі шешілген соң: 1) әліпбилі жиілік сөздіктен; 2) жиілікті-әліпбилі; 3) кері-әліпбилі түрлері алынды.

Жиілік сөздіктер жасауда реестрлік бірліктерді өңдеуде омоним сөздердің жиілігін анықтау қиындық тудырады. Омонимдердегі қиындықтықты айтпас бұрын жалпы жиілік сөздік құрастыру үдерісін сипаттап өтпекпіз. Жиілік сөздіктер жасау технологиясында үдеріс төмендегі ретпен жүреді:

1. Ең алдымен, жиілік сөздіктің әліпби ретімен тұратын Әліпбилі-жиілікті сөзформалар сөздігі жасалады. Сөзформа дегеніміз – жиілік сөздік алынатын мәтіндегі түбір және түрленген формада тұрған барлық сөздер. Сөзформа тізімі түбір тұлғада тұрған сөздерден әрине көп болады. Өйткені сөзформада сөздер түрленген тұлғаларымен бірге алынады, яғни түбір формаға келтірілмейді.

2. Сөзформалар тізімі жасалғаннан кейін сөз жиілігін (негізгі түбір және туынды түбір) анықтау үшін, сөзформалар түбірге, яғни леммаға келтірілу керек және сөздерге сөз таптары

қойылуы керек. Үлкен көлемді мәтіндерден бұндай жұмысты қолмен атқару, әрине, қиын. Солай бола тұра, ХЛ-файлдағы сөзформалар тізімінен «қолдап» бірдей сөзформалар тобын бір сөзге келтіре отырып, санын қосып отыру және сөз табын қою арқылы жүзеге асыру тәсілі бар және бұл тәсіл 2016 жылы А. Байтұрсынұлы атындағы Тіл білімі институтында «Жалпы білім берудегі жиілік сөздікті» құрастыру кезінде қолданылған болатын. Қолмен *лемматизациялау* (леммаға, түбірге келтіру) және сөз табын қоюда қолмен істеу ұзақ уақытты алғанмен, сөзформадағы бірлік түрленіп, яғни қосымшалары көрініп тұрғандықтан, олардың қандай сөздің түрленімі екенін, түбірі қай сөз табы екенін білуге болады. Бұл әсіресе омонимдерді тануда маңызды. Мысалы, *қарады, қараса, қараған* т.б. сөзформаларды «*қара*» етістігінің түрленім екенін қосымшаларына қарап анықтай аламыз. Алайда «*қара*» сөзі мәтінде бұйрық райда тұрып (сен бері қара) қолданылуы мүмкін. Мұндайда ХЛ-файлдағы сөзформалар тізімінде сөздер мәнмәтінсіз (мәтінсіз) бірінің астына бірі әліпби ретімен тізімделіп тұратындықтан, «*қара*» сөзінің етістік *қара* ма әлде сын есім *қара* ма екені түсініксіз болып, анықталмай қалады. Міне, осындағы «*қара»* сын есімі мен «*қара»* етістігі – омонимдер. Қолмен сөз табын қоюда да осындай тұлғалары бірдей, еш қосымша жалғанбаған (түрленбеген) сөздердің қай мағынада тұрғанын, қай сөз табына жататынын мәнмәтінсіз ажырата алмаймыз.

Негізгі корпус мәтіндерінің жиілік сөздігі материалы 30 миллион сөзқолданыстан тұратын көлемді мәтін болғандықтан, сөзформаларды түбірге келтіру, сөз табын қою жұмысы автоматты морфологиялық анализатор арқылы жасалды. Жоғарыда айтқанымыздай, автоматты талдауда бағдарлама түбір тұлғадағы омонимдерді танымайтын (ажыратпайтын) болғандықтан, бағдарлама бірдей тұлғада тұрған сөздердің жиілігін қосып жазады немесе екі сөз табы ретінде көрсетіп, сол санды екеуіне де жазады. Мысалы,

жүз/зт – 14259

жүз/ет – 14259

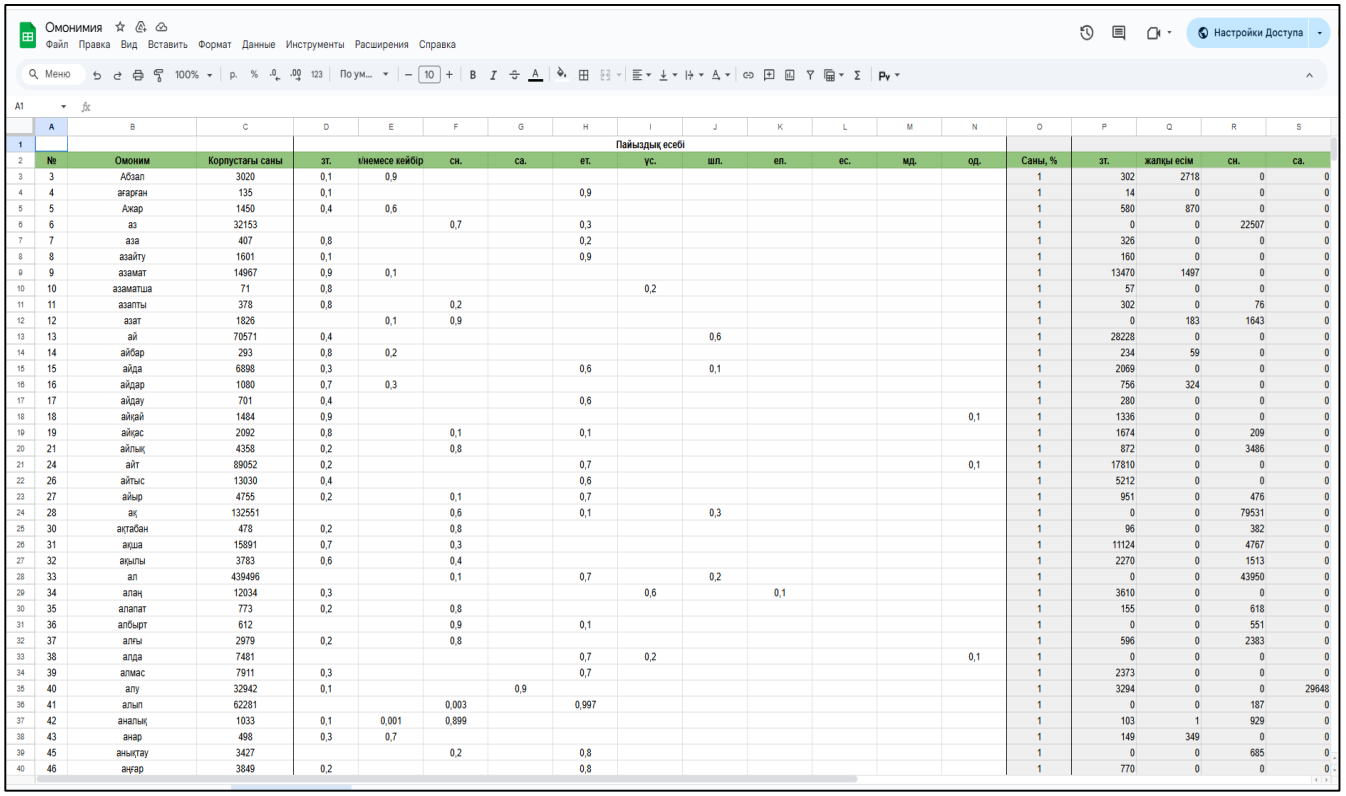
Автоматты талдау барысында бірдей тұлғада тұрған омонимдерді ажырата алмай, жиілікті екі мағынаға да бірдей санмен белгілейтін жағдайлар жиі кездеседі. Мысалы, «*жүз*» формасы зат есім (жүз/зт) және етістік (жүз/ет) ретінде 14 259 рет ретінде тіркелсе, бұл автоматты талдау жүйесінің мәнмәтіндік мағынаны ескермей, барлық «*жүз*» түрленімдерін екі мағынаға да толық қосуынан туындайды. Мұндай жағдайда жиілік сөздіктің сапасы мен дәлдігін қамтамасыз ету үшін қосымша кезеңдерде омонимдерді статистикалық және мәнмәтіндік талдау арқылы ажырату қажет.

Алдымен, әрбір омоним кандидат үшін корпус ішінде оның мәнмәтіндерін жинақтау жүзеге асырылады. Автоматты талдау нәтижесінде «*жүз*» деген түбірдің жиілігі 14 259 рет тіркелген соң, сол формаға сәйкес барлық конкорданс жолдары (мәнмәтіндік фрагменттер) шығарылады. Әр мәнмәтіндік фрагментте «*жүз*» форма қандай мағынада (зат есім ретіндегі «*жүз*» мәнінде, мысалы сандық «*100*» немесе «*бет, жүз*» мағынасы, немесе етістік ретіндегі «*жүзу*» мәнінде) қолданылғаны сараланады. Мәнмәтіндік талдау үшін бірнеше қадам ұсынылады:

Мәнмәтіндік үлгілерді іріктеу. Корпус конкордансынан алынған барлық мәнмәтіндік фрагменттерді толығымен қарап шығу практикалық тұрғыдан ауыр: олар саны өте көп болады. Сондықтан кездейсоқ таңдау немесе стратификацияланған үлгі (қабаттарға бөлінген) іріктеу әдісін қолдану қажет. Мысалы, «*жүз*» омониміндегі 14 259 мәнмәтіннің ішінен кездейсоқ 500–1000 таңбалы үлгілерді таңдап, оларды жан-жақты талдау үшін қолдануға болады. Мәтін көлемі одан да үлкен болса, стратификация (қабаттарға бөлу) мәнмәтіндердің жанрлық стильдеріне (поэзия, проза, публицистика және т.б.) қарай да жасалуы мүмкін.

Мәнмәтінді аннотациялау (мануалды белгілеу). Іріктелген үлгілерді тіл мамандары немесе сарапшылар мәнмәтінге қарай талдап, «*жүз*» формасының нақты мағынасын белгілеу қажет. Әр мәнмәтіндік фрагментте омонимнің мағынасы анықталған соң, зат есім (мысалы, «*жүз жыл*», «*жүз адам*», «*жүзгілеу*» емес) немесе етістік (мысалы, «*өзенде жүз*», «*жүзуге шығу*») ретінде қолданылғанын белгілеп отырады. Осыдан кейін әр мағынаға тиесілі үлгілердің саны есептеледі. Төменде ұсынылған кесте омонимдердің корпус ішіндегі жиілік-статистикалық талдау нәтижелерін жинақтайды. Әрбір жол бір омонимдік леммаға (түбір сөзге) сәйкес келіп, сол сөздің корпус ішіндегі жалпы кездескен санынан бастап, ықтималдық-статистикалық әдіспен сөз таптары (зат есім, сын есім, етістік және т.б.) бойынша бөлінген үлестері мен нақты жиілік

мәндерін көрсетеді.



							Пайыздық есебі											
№	Омоним	Корпустағы саны	зт.	немесе кейбір	сн.	са.	ет.	үс.	шл.	ел.	ес.	мд.	од.	Саны, %	зт.	жалқы есім	сн.	са.
3	Абзал	3020	0,1	0,9										1	302	2718	0	0
4	ағарған	135	0,1				0,9							1	14	0	0	0
5	Ажар	1450	0,4	0,6										1	580	870	0	0
6	аз	32153			0,7		0,3							1	0	0	22507	0
7	аза	407	0,8				0,2							1	326	0	0	0
8	азайту	1601	0,1				0,9							1	160	0	0	0
9	азамат	14967	0,9	0,1										1	13470	1497	0	0
10	азаматша	71	0,8					0,2						1	57	0	0	0
11	азапты	378	0,8		0,2									1	302	0	76	0
12	азат	1826		0,1	0,9									1	0	183	1643	0
13	ай	70571	0,4						0,6					1	28228	0	0	0
14	айбар	293	0,8	0,2										1	234	59	0	0
15	айда	6898	0,3				0,6		0,1					1	2069	0	0	0
16	айдар	1080	0,7	0,3										1	756	324	0	0
17	айдау	701	0,4				0,6							1	280	0	0	0
18	айқай	1484	0,9										0,1	1	1336	0	0	0
19	айқас	2092	0,8		0,1		0,1							1	1674	0	209	0
20	айлық	4358	0,2		0,8									1	872	0	3486	0
21	айт	89052	0,2				0,7						0,1	1	17810	0	0	0
22	айтыс	13030	0,4				0,6							1	5212	0	0	0
23	айыр	4755	0,2		0,1		0,7							1	951	0	476	0
24	ақ	132551			0,6		0,1		0,3					1	0	0	79531	0
25	ақтабан	478	0,2		0,8									1	96	0	382	0
26	ақша	15891	0,7		0,3									1	11124	0	4767	0
27	ақылы	3783	0,6		0,4									1	2270	0	1513	0
28	ал	439496			0,1		0,7		0,2					1	0	0	43950	0
29	алаң	12034	0,3					0,6		0,1				1	3610	0	0	0
30	алапат	773	0,2		0,8									1	155	0	618	0
31	албырт	612			0,9		0,1							1	0	0	551	0
32	алғы	2979	0,2		0,8									1	596	0	2383	0
33	алда	7481					0,7	0,2					0,1	1	0	0	0	0
34	алмас	7911	0,3				0,7							1	2373	0	0	0
35	алу	32942	0,1			0,9								1	3294	0	0	29648
36	алып	62281			0,003		0,997							1	0	0	187	0
37	аналық	1033	0,1	0,001	0,899									1	103	1	929	0
38	анар	498	0,3	0,7										1	149	349	0	0
39	анықтау	3427			0,2		0,8							1	0	0	685	0
40	аңғар	3849	0,2				0,8							1	770	0	0	0

Сурет 1 – Омонимдер жиілігін анықтаудың ықтималды-есептеуіш кестесі
Рисунок 1 – Вероятностно-вычислительная таблица определения частоты омонимов
Figure 1 – Probability-calculating table for determining the frequency of homonyms

Бұл бағдарламада омоним болып келетін немесе тұлғасы бірдей сөздердің жалпы, яғни омоним сөздердің екеуіне де ортақ саны жазылады. Келесі бағандарда қазақ тіліндегі сөз таптары жеке-жеке бағандарға беріледі. Соңында есептеуіш берілген.

Бұндағы омоним сөз Негізгі корпустан ізделіп, корпустағы саны бойынша қойылады. Саны алынғаннан кейін, орындаушы корпустан омоним сөздің қай мағынада тұрғаны жиі кездесетінін, жалпы шыққан мысалдың қанша пайызын құрайтынан конкорданска шыққан мысалдар бойынша анықтайды. Мысалы, Негізгі корпуста «*айқас*» омоним сөзі 1561 рет кездескені автоматты түрде шықты делік, осы санды омонимдерді ықтималды-статистикалық әдіспен анықтау кестесіне жазамыз. Ендігі кезекте корпустан осы сөздің қай мағынада қолданылғаны қаншалықты кездесетінін мысалдар легіне көз жүгірте отырып анықтаймыз. Біздің анықтауымызша, зат есім ретінде мысалдардың 0,49 пайызын құрады, ал сын есім 0,02 пайыз, етістік мәнінде де 0,49 пайыз құрады. Осы пайыздық көрсеткішті кестеге жазамыз. Осы пайыздық көрсеткіштер бойынша есептеуіш «*айқас*» сөзінің зат есім ретінде – 765, сын есім ретінде – 31, етістік мәнінде – 765 рет кездескенін шығарды. Сол сияқты «*ажар*» омонимі Негізгі корпуста 1450, рет кездессе, кісі атауы ретінде мәтінде көбірек қолданылып, 0,6 пайызын құрады, ал бет, әлпет мәнінде 0,4 пайызды құрады. Бұдан есептеуіш бағдарлама «*ажар*» сөзінің кісі есімі ретінде – 580 рет, бет-әлпет мәнінде – 870 рет қолданғаны туралы ықтимал санды шығарды. «*Албырт*» омоним сөзі Негізгі корпуста 404 рет кездессе, сын есім ретінде Негізгі корпустан конкорданска шыққан мысалдардың 0,9 пайызын, ал етістік мәнінде 0,1 пайызын құрады. Есептеуіш есебі бойынша, «*албырт*» сөзінің сын есім ретінде 364 рет, етістік ретінде 40 рет қолданылғаны шығарылды. Сондай-ақ «*ерін*» омоним сөзі Негізгі корпуста 3740 рет кездессе, белгілі болғандай, зат есім ретінде 0,88 пайызын құрады, ал етістік мәнінде 0,12 пайызын құрады. Есептеуіш есеп бойынша, «*ерін*» сөзінің зат есім ретінде 3291 рет, етістік ретінде 449 рет қолданылғаны белгілі болды. «*Жаз*» омоним сөзі Негізгі корпуста 41247 рет кездессе, конкорданска шыққан мысалдарда зат есім ретінде 0,1 пайызын құраса, ал

етістік мәнінде 0,9 пайызын құраған. Пайыздық есептеуіш бағдарлама бойынша, «*жаз*» сөзі зат есім ретінде 4125 рет, етістік ретінде 37122 рет қолданылғаны туралы ықтимал санды шығарды. Тағы да «*жалын*» омоним сөзі Негізгі корпуста 5221 рет кездессе, біздің анықтауымыз бойынша, зат есім ретінде мысалдардың 0,85 пайызын құрады, ал етістік мәнінде 0,15 пайыз құрады. Осы пайыздық есептеуіш көрсеткіш бойынша «*жалын*» сөзінің саны зат есім ретінде 4438 рет, етістік ретінде 783 рет қолданылғаны белгілі болды. «*Жеңіл*» омоним сөзі Негізгі корпуста 9009 рет кездессе, сын есім ретінде Негізгі корпустан конкорданска шыққан мысалдардың 0,9 пайызын, ал етістік мәнінде 0,1 пайызын құрады. Есептеуіш есебі бойынша, «*жеңіл*» сөзінің сын есім ретінде 8108 рет, етістік ретінде 901 рет қолданылғаны шығарылды. «*Жүз*» омоним сөзі Негізгі корпуста 41778 рет кездессе, белгілі болғандай, зат есім ретінде 0,57 пайызын, ал сан есім ретінде 0,1 пайыз құраса, етістік мәнінде 0,42 пайызын құрады. Бұдан есептеуіш бағдарлама «*жүз*» сөзінің зат есім ретінде 23813 рет, сан есім ретінде 418 рет, етістік мәнінде 17547 рет қолданылғаны шығарылды. Кездескен омонимдерді «Excel» бағдарламасында мынадай кестемен беруге болады:

№	Омоним	Корпустағы саны	зт.	сн.	са.	ет.	үс.	шл.	ел.	ес.	од.	Саны, %	зт.	сн.	са.	ет.	үс.	шл.	ел.	ес.	од.
1	жүз	14259	0,5		0,4	0,1						0,5	7130	0	5704	1426	0	0	0	0	0
2	айқас	1561	0,49	0,02		0,49						1	765	31	0	765	0	0	0	0	0
3	ажар	548	0,4	0,6								1	219	329	0	0	0	0	0	0	0
4	албырт	404		0,9		0,1						1	0	364	0	40	0	0	0	0	0
5	ерін	3740	0,88			0,12						1	3291	0	0	449	0	0	0	0	0
6	жаз	41247	0,1			0,9						1	4125	0	0	37122	0	0	0	0	0
7	жалын	5221	0,85			0,15						1	4438	0	0	783	0	0	0	0	0
8	жеңіл	9009		0,9		0,1						1	0	8108	0	901	0	0	0	0	0
9	жүз	41747	0,57		0,1	0,42						1	23813	0	418	17547	0	0	0	0	0

Сурет 2 – Омонимдер жиілігін анықтаудың ықтималды-статистикалық кестесі
Рисунок 2 – Вероятностно-статистическая таблица определения частоты омонимов
Figure 2 – Probabilistic-statistical table for determining the frequency of homonyms

Міне, Негізгі корпус мәтіндерінің жиілік сөздігіндегі омонимдердің ықтимал жиілігін анықтау осындай әдіспен анықталды. Омонимдерді ажырату жұмысы Әліпбилі-жиілік сөздік бойынша жасалды. Өйткені омоним сөздердің реестрде бір жерде тұрғаны ыңғайлы болды. Осы әдіспен жиілік реестрінде барлық омонимдердің ықтималды жиілігі шығарылғаннан кейін, Әліпбилі-жиілік сөздік Жиілікті-әліпбилі түрі бойынша жасалды. Бұдан кейін де жалпы сөздердің жиілігі мен оның ішінде омонимдердің жиілігі бойынша сандық мәліметтер тілші сараптамасына салынып, сандық деректердің шынайылығы тексерілді.

Омонимдерді ажыратудың жоғарыда берілген ықтималды статистикасы дәл сандық дерек бермегенмен, көлемді мәтіндер бойынша (жалпы картина) көрудің тиімді тәсілі екені анықталып отыр.

Қорытынды

Жиілік сөздіктер тіл қолдану ерекшеліктерін сандық тұрғыдан сипаттауда және практикалық міндеттерді шешуде маңызды лингвистикалық құрал болып табылады. Олар жиілік мәліметтерін жинақтау арқылы минимум сөздіктерді құрастыруға, екінші тілді меңгеру мен оқыту үдерісінде статистикалық әдістерді қолдануға мүмкіндік береді. Яғни, жиілік деректер негізінде оқулықтар мен оқу материалдарын әзірлеушілер лексикалық қорды тиімді орналастырып, грамматикалық құрылымдар мен мазмұндық категорияларды жүйелі оқытуға жәрдемдеседі. Сонымен қатар жазба және сөйлеу тілдеріндегі сөздердің қолданылу жиілігін зерттеу әдістері тіл құрылымының грамматикалық және стилистикалық ерекшеліктерін нақтыландыра отырып, мәтіндерді тілдік жағынан оңтайландыруға және аударма үдерісін жеңілдетуге ықпал етеді. Осылайша, жиілік сөздіктер тіл үйренушілерге қажетті ақпаратты сұрыптап ұсыну және сөйлеу мәнерін дамыту жөнінде практикалық көмек көрсетеді.

Лингвистикалық зерттеулер тұрғысынан жиілік сөздіктер статистикалық мәліметтер алу мен жаңа сөзжасам жүйелерін талдау үшін құнды ресурс болып табылады. Олар қазақ тіліндегі лексикалық бірліктердің таралу дәрежесін анықтап қана қоймай, басқа түркі тілдерімен немесе

шет тілдермен салыстыратын типологиялық зерттеулерде де маңызды дереккөз рөлін атқарады. Әсіресе өзге ұлт өкілдеріне арналған қазақ тілі оқулықтарын әзірлеу кезінде, А.Байтұрсынұлы атындағы Тіл білімі институтының «Қазақ тілінің ұлттық корпусы мәтіндерінің жиілік сөздігі» сияқты ресурстар оқыту үдерісіне қажетті лексикалық жиіліктерді нақты негізде ұсынады. Бұл оқулықтардағы лексикалық минимумдарды анықтау, сөйлеу және жазба дағдыларын қалыптастыру үшін бағыт-бағдар беру барысында жиілік деректер негізіндегі әдістемелерді пайдалану тиімдірек нәтижелер береді.

Жиілік сөздіктерді құрастыруда омонимдерді ықтималдық-статистикалық әдістермен ажырату технологиясы теориялық және практикалық зерттеулерде шешуші мәнге ие. Омонимдердің мәтіндегі нақты мағыналарын мәнмәтіндік үлестірімдер арқылы статистикалық тұрғыдан есептеу жиілік реестрлерді дәл әрі сапалы қалыптастыруға мүмкіндік береді. Бұл тәсіл мәтіндердегі мағыналық байланыстарды, грамматикалық және семантикалық ерекшеліктерді жан-жақты қамтып, нәтижесінде жиілік сөздіктердің ақпараттық толықтығын арттырады. Сонымен қатар алынған статистикалық үлестірулер негізінде тілдік модельдерді жетілдіру, мәтінді автоматты талдау жүйелерін дамыту және лексикалық деректерді мәшинелік оқыту әдістерімен өңдеу мүмкіндіктері кеңейеді. Демек, омонимдерді ажыратудың ықтималдық-статистикалық тәсілі жасанды интеллект саласындағы тілдік деректерді автоматты түрде талдауға негіз бола отырып, гуманитарлық зерттеулермен қатар компьютерлік лингвистикада да өзекті міндеттерді шешуге септігін тигізеді.

Жинақтай келгенде, жиілік сөздіктердің қолданылуы тіл оқыту мен лингвистикалық зерттеулерде, сондай-ақ тілдік модельдер мен автоматтандырылған жүйелерді дамытуда практикалық және теориялық маңызға ие. Омонимдерді ықтималдық-статистикалық әдістер арқылы ажырату технологиясы жиілік реестрлердің дәлдігі мен сапасын арттырып, тіл білімінің әр түрлі аспектілеріндегі статистикалық талдаулар мен модельдеуді жүйелі түрде жетілдіруге мүмкіндік тудырады.

Әдебиеттер

Абай тілі сөздігі. – Алматы, 1968. – 734 б.

Абай шығармаларының академиялық толық жинағы. Том 1, 2, 3. – Алматы: Жазушы, 2020. – 604, 524, 488 бб.

Аношкина Н.Г. Омонимия в русском языке. – Омск: Изд-во ОмГПУ, 2000. – 118 с.

Аханов К. Тіл біліміне кіріспе. – Алматы: Қазақтың мемл. оқу-педагогика баспасы, 1962. – 299 б.

Ахманова О.С. Очерки по общей и русской лексикологии. – Москва: Гос. учебно-пед. изд., 1957. – 294 с.

Бектаев Қ.Б. Ғ.Мүсіреповтің «Кездеспей кеткен бір бейне» повесі тілінің алфавитті-жиілік сөздігі/ Қ.Б. Бектаев, А. Белботаев, Қ. Молдабеков жб. // Қазақ тексінің статистикасы: «Статистика-лингвистикалық зерттеу мен автоматтандыру» тобы еңбектерінің III шығуы. – Алматы, 1973. – 519-542-бб.

Бектаев Қ.Б., Жұбанов А.Қ., Мырзабеков С., Белботаев А.Б. М.О. Әуезовтің 20 томдық шығармалар текстерінің жиілік сөздіктері. – Алматы-Түркістан, 1995. – 346 б.

Белбаева М. Қазақ тілінің омонимдер сөздігі. – Алматы: Мектеп, 1988. – 192 б.

Жалпы білім берудегі қазақ тілінің жиілік сөздігі. – Алматы: «Дәуір» баспасы, 2016. – 1472 б.

Жұбанов А., Жаңабекова А., Карбозова Б., Қожахметова А. Қазақ тілінің жиілік сөздігі. – Алматы: «Қазақ тілі» баспасы, 2016. – 665 б.

Зеленков Ю.Г., Сегалович И.В., Титов В.А. Вероятностная модель снятия морфологической омонимии на основе нормализующих подстановок и позиций соседних слов // Компьютерная лингвистика и интеллектуальные технологии. Труды международного семинара «Диалог-2005». – Москва: Наука, 2005. – 616 с.

Кеңесбаев І. Қазіргі қазақ тілі. Лексика,фонетика / Қазақ ССР Жоғары және арнаулы орта білім министрлігі бекіткен. 2-ші басылуы. – Алматы: Мектеп, 1975. – 304 б.

Ким О.М. Омонимия на уровне частей речи на современном английском языке: материалы к спецкурсу Омонимия на уровне частных речи в современном русском языке: материалы к спецкурсу. – Ташкент: Ташкентский гос. университет им. В.И. Ленина, 1983. – 68 с.

Кобзарева Т. Ю., Афанасьев Р. Н. Универсальный модуль предсинтаксического анализа омонимии частей речи в русском языке на основе словаря диагностических ситуаций // Компьютерная лингвистика и интеллектуальные технологии. Труды международного семинара «Диалог-2002». Т. 2. – Протвино: 2002. – С. 258-268.

Колесников Н.П. Словарь омонимов русского языка. – Москва: Русский язык, 1980. – 302 с.

Мұсабаев Ғ.Ғ. Қазақ тіл білімінің мәселелері. – Алматы: Абзал-Ай, 2014. – 640 б. ISBN 978-601-7172-39-8.

Google Docs (n.d.) URL: https://docs.google.com/spreadsheets/d/1ZtdK2xCBvXTXVp1ImiaKLCVh-zZ4mjJ_OSyuxrQGdtA/edit?gid=0#gid=0 [Accessed: 10 June 2025]. (online resource)

wikipedia.org/ [Қолданылуы: 2025 жылғы 10 маусым]. (интернет-ресурс)

References

Abai shygarmalarynyng akademijalyq tolyq zhinagy. Tom 1, 2, 3. Almaty: Zhazushy, 2020. 604, 524, 488 bb. [Full academic collection of Abai's works. Volume 1, 2, 3. Almaty: Zhazushy, 2020. 604, 524, 488 pp.] (in Kazakh)

Abai tili sozdigi. (1968) Almaty, 734 b. [Abai language dictionary. (1968) Almaty, 734 p.] (in Kazakh)

Ahanov, K. (1962) Til bilimine kirispe. Almaty: Qazaqtyng meml. oqu-pedagogika baspasy, 299 b. [Akhanov, K. (1962) Introduction to Linguistics. Almaty, 299 p.] (in Kazakh)

Ahmanova, O.S. (1957) Ocherki po obshchej i russkoj leksikologii. Moskva: Gos. uchebno-ped. izd., 294 s. [Akhmanova, O.S. (1957) Essays on general and Russian lexicology. Moscow: State Educational and Pedagogical Publishing House, 294 p.] (in Russian)

Anoshkina, N.G. (2000) Omonimija v russkom jazyke. Omsk: Izd-vo OmGPU, 118 s. [Anoshkina, N.G. (2000) Homonymy in the Russian language. Omsk: Publishing house of OmGPU, 118 p.] (in Russian)

Bektaev, Q.B., Belbotaev, A., Moldabekov, Q. (1973) G. Musirepovting "Kezdespei ketken bir beine" povesi tilining alfavitti-zhiilik sozdigi. Qazaq teksining statistikasy: «Statistika-lingvistikalyq zertteu men avtomattandyru» tobynyng III shyguy. Almaty. B. 519-542. [Bektayev, K.B., Belbotayev, A., Moldabekov, K. (1973) Alphabetical-frequency dictionary of the language of G.Musrepov's story "An Unmet Image". Statistics of Kazakh texts: Works of the group "Statistical-linguistic research and automation". Issue III. Almaty. P. 519-542.] (in Kazakh)

Bektaev, Q.B., Zhubanov, A.Q., Myrzabekov, S., Belbotaev, A.B. (1995) M.O. Auezovting 20 tomdyq shygarmalar teksterining zhiilik sozdikteri. Almaty-Turkistan. 346 b. [Bektayev, Q.B., Zhubanov, A.K., Myrzabekov, S., Belbotayev, A.B. (1995) Frequency dictionaries of the 20-volume works of M.O.Auezov. Almaty-Turkistan. 346 p.] (in Kazakh)

Belbaeva, M. (1988) Qazaq tilining omonimder sozdigi. Almaty: Mektep, 192 b. [Belbayeva, M. (1988) Dictionary of homonyms of the Kazakh language. Almaty: Mektep, 192 p.] (in Kazakh)

Google Docs (n.d.) URL: https://docs.google.com/spreadsheets/d/1ZtdK2xCBvXTXVp1ImiaKLCVh-zZ4mjJ_OSyuxrQGdtA/edit?gid=0#gid=0 [Accessed: 10 June 2025]. (online resource)

Kengesbaev, I. (1975) Qazirgi qazaq tili: leksika, fonetika. 2-basylymy. Almaty: Mektep, 304 p. [Kenesbayev, I. (1975) Modern Kazakh language: lexicon, phonetics. 2nd edition. Almaty: Mektep, 304 p.] (in Kazakh)

Kim, O.M. (1983) Omonimija na urovne chastej rechi na sovremennom anglijskom jazyke: materialy k speckursu. Tashkent: Tashkentskij gosudarstvennyj univ. imeni V. I. Lenina, 68 s. [Kim, O.M. (1983) Homonymy at the level of parts of speech in modern English: materials for the special course. Tashkent: Tashkent State Univ. named after V.I. Lenin, 68 p.] (in Russian)

Kobzareva, T.Yu., Afanas'ev, R.N. (2002) Universal'nyj modul' predsintaksicheskogo analiza omonimii chastej rechi v russkom jazyke na osnove slovarja diagnosticheskih situacij. Komp'juternaja lingvistika i intellektual'nye tehnologii. Trudy mezhdunarodnogo seminara «Dialog-2002». T. 2. Protivino. S. 258-268 [Kobzareva, T.Yu., Afanasyev, R.N. (2002) Universal module of pre-syntactic analysis of parts of speech homonymy in Russian based on the dictionary of diagnostic situations. Computational Linguistics and Intellectual Technologies. Proceedings of the international seminar «Dialogue-2002». V. 2. Protvino. P. 258-268.] (in Russian)

Kolesnikov, N.P. (1980) Slovar' omonimov russkogo jazyka. Moskva: Russkij jazyk, 302 s. [Kolesnikov, N.P. (1980) Dictionary of homonyms of the Russian language. Moscow: Russian Language, 302 p.] (in Russian)

Musabaev, G.G. (2014) Qazaq til bilimining maseleleri. Almaty: Abzal-Ai, 640 p. ISBN 978-601-7172-39-8. [Musabayev, G.G. (2014) Issues of Kazakh linguistics. Almaty: Abzal-Ai, 640 p.] (in Kazakh)

Wikipedia (n.d.) URL: https://kk.wikipedia.org/ [Accessed: 10 June 2025]. (online resource)

Zelenkov, Yu.G., Segalovich, I.V., Titov, V.A. (2005) Veroijatnostnaja model' snijatija morfologicheskoj omonimii na osnove normalizujushchih podstanovok i pozicij sosednih slov. Komp'juternaja lingvistika i intellektual'nye tehnologii. Trudy mezhdunarodnogo seminara «Dialog-2005». Moskva: Nauka, 616 s. [Zelenkov, Yu.G., Segalovich, I.V., Titov, V.A. (2005) Probabilistic model for resolving morphological homonymy based on normalizing substitutions and positions of neighboring words. Computational Linguistics and Intellectual Technologies. Proceedings of the international seminar “Dialogue-2005”. Moscow: Nauka, 616 p.] (in Russian)

Zhalpy bilim berudegi qazaq tilining zhiilik sozdigi. (2016) Almaty: Dauir, 1472 b. [Frequency dictionary of the Kazakh language in general education. (2016) Almaty: Dauir, 1472 p.] (in Kazakh)

Zhubanov, A., Zhangabekova, A., Karbozova, B., Qozhahmetova, A. (2016) Qazaq tilining zhiilik sozdigi. Almaty: Qazaq tili baspasy. 665 b. [Zhubanov, A., Zhanabekova, A., Karbozova, B., Kozhakhmetova, A. (2016) Frequency dictionary of the Kazakh language. Almaty: Kazakh Language Publishing House, 665 p.] (in Kazakh)

Мақала туралы ақпарат / Информация о статье / Information about the article
Редакцияға түсті / Поступила в редакцию / Entered the editorial office: 12.02.2025.
Жариялауға қабылданды / Принята к публикации / Accepted for publication: 25.06.2025.