ISSN 2518-1726 (Online),
ISSN 1991-346X (Print)

ACADEMIC SCIENTIFIC JOURNAL OF COMPUTER SCIENCE

№1
2026

ISSN 2518-1726 (Online),
ISSN 1991-346X (Print)

CENTRAL ASIAN ACADEMIC
RESEARCH CENTER

ACADEMIC SCIENTIFIC JOURNAL OF COMPUTER SCIENCE

1 (357)

JANUARY – MARCH 2026

PUBLISHED SINCE JANUARY 1963
PUBLISHED 4 TIMES A YEAR

ALMATY, NAS RK

Academic Scientific Journal of Computer Science
ISSN 2518-1726 (Online),
ISSN 1991-346X (Print)
Owner: «Central Asian Academic Research Center» LLP (Almaty).
Certificate **№ KZ77VPY00121154** on the re-registration of the periodical printed and online publication of the information agency, issued on **05.06.2025** by the Republican State Institution «Information Committee» of the Ministry of Culture and Information of the Republic of Kazakhstan
Subject area*: information and communication technologies.*
Currently: *included in the list of journals recommended by the CCSES MSHE RK in the direction of «Information and communication technologies».*
Periodicity: *4 times a year.*
http://www.physico-mathematical.kz/index.php/en/

Academic Scientific Journal of Computer Science
ISSN 2518-1726 (Online),
ISSN 1991-346X (Print)
Меншіктеуші: «Орталық Азия академиялық ғылыми орталығы» ЖШС (Алматы).
Ақпарат агенттігінің мерзімді баспасөз басылымын, ақпарат агенттігін және желілік басылымды қайта есепке қою туралы ҚР Мәдениет және Ақпарат министрлігі «Ақпарат комитеті» Республикалық мемлекеттік мекемесі **05.06.2025** ж. берген **№ KZ77VPY00121154** Куәлік.
Тақырыптық бағыты: *ақпараттық-коммуникациялық технологиялар*
Қазіргі уақытта: *«ақпараттық-коммуникациялық технологиялар» бағыты бойынша ҚР БҒМ БҒСБК ұсынған журналдар тізіміне енді.*
Мерзімділігі: *жылына 4 рет.*
http://www.physico-mathematical.kz/index.php/en/

Academic Scientific Journal of Computer Science
ISSN 2518-1726 (Online),
ISSN 1991-346X (Print)
Собственник: *ТОО «Центрально-азиатский академический научный центр» (г. Алматы).*
Свидетельство о постановке на переучет периодического печатного издания, информационного агентства и сетевого издания **№ KZ77VPY00121154.** Дата выдачи **05.06.2025**
Тематическая направленность: *информационно-коммуникационные технологии.*
В настоящее время: *вошел в список журналов, рекомендованных КОКСНВО МНВО РК по направлению «информационно-коммуникационные технологии».*
Периодичность: *4 раза в год.*
http://www.physico-mathematical.kz/index.php/en/

CONTENTS

COMPUTER SCIENCE

МАЗМҰНЫ

ИНФОРМАТИКА

СОДЕРЖАНИЕ

ИНФОРМАТИКА

ACADEMIC SCIENTIFIC JOURNAL OF COMPUTER SCIENCE
ISSN 1991-346X
Volume 1.
Namber 357 (2026). 354–372

https://doi.org/10.32014/2026.2518-1726.417

IRSTI 62.50.43
UDC 004.5

© Sarsenbayeva A.K.[1]*, Rakhimova D.R.[1], Shormakova A.N.[1], Mansurova M.E.[1], Adali E.[2], 2026.
[1]Al-Farabi Kazakh National University, Almaty, Kazakhstan;
[21]stanbul Technical University, Istanbul, Turkey.
E-mail: as.sarsenbayeva@gmail.com

APPLICATION OF SEMANTIC METHODS IN THE FIELD OF LEGISLATION: AN INTELLECTUAL SYSTEM FOR ANALYSIS OF AGGLUTINATIVE TEXTS

Sarsenbayeva Assiya — PhD student of Al-Farabi Kazakh National University, Almaty, Kazakhstan, E-mail: as.sarsenbayeva@gmail.com, https://orcid.org/0009-0008-0053-1182;
Rakhimova Diana — PhD, Associate Professor of Al-Farabi Kazakh National University, Almaty, Kazakhstan,
E-mail: drakhimova060@gmail.com, https://orcid.org/0000-0003-1427-198X;
Shormakova Assem — PhD, acting Associate Professor of Al-Farabi Kazakh National University, Almaty, Kazakhstan,
E-mail: shormakovaassem@gmail.com, https://orcid.org/0000-0002-1637-4643;
Mansurova Madina — PhD, acting Associate Professor of Al-Farabi Kazakh National University, Almaty, Kazakhstan,
E-mail: m_mansurova@kaznu.kz, https://orcid.org/0000-0001-6284-8283;
Adali Esref — PhD, Professor of Istanbul Technical University, Istanbul, Turkey,
E-mail: adali@itu.edu.tr, https://orcid.org/0000-0002-1561-8255.

Abstract. This article presents a semantic method for determining the similarity of terms, phrases, and short expressions in agglutinative languages, particularly Kazakh and Turkish. Semantic similarity plays a key role in many natural language processing tasks, including information retrieval, question–answering systems, text classification, and legal document analysis. However, the development of such systems for agglutinative and low-resource languages remains challenging due to complex morphology, limited linguistic resources, and insufficient annotated corpora. The proposed approach is based on the Ri coefficient, which evaluates semantic similarity using frequency relationships between words within their contextual environments. The method analyzes contextual sets of words that frequently co-occur with target terms and calculates normalized proximity measures between them. By integrating contextual frequency information and

synonym relations, the approach allows for a more accurate evaluation of semantic relationships between legal terms and expressions. To assess the effectiveness of the proposed method, a comparative study was conducted with widely used similarity metrics, including Jaccard similarity, Cosine similarity, and Normalized Google Distance. Experimental evaluation was carried out on a corpus of legal texts in the Kazakh language containing approximately 10,000 legal sentences and a specialized legal dictionary consisting of 8,400 terms and about 10,000 synonyms. Additional frequency data were obtained from open legal resources such as the Adilet legal information system. The findings confirm that the proposed semantic similarity approach is effective for processing legal texts in agglutinative languages and can be integrated into intelligent legal information systems, including question–answering systems and legal document retrieval platforms. Furthermore, the method has the potential to be adapted for other Turkic and low-resource languages with similar morphological characteristics.

Keywords: semantic similarity, synonyms, Kazakh language, Turkish language, legal terms, agglutinative languages

For citations: Sarsenbayeva A.K., Rakhimova D.R., Shormakova A.N., Mansurova M.E., dali E. Application of semantic methods in the field of legislation: an intellectual system for analysis of agglutinative texts. Academic Scientific Journal of Computer Science, 2026. — No.1. – P. 355–372. DOI: https://doi.org/10.32014/2026.2518-1726.417

© Сарсенбаева А.К. [1] *, Рахимова Д.Р.[1], Шормакова А.Н. [1], Мансурова М.Е.[1], Адали Э.[2], 2026.

[1] Әл-Фараби атындағы Қазақ ұлттық университеті, Алматы, Қазақстан;
[2] Стамбул Техникалық университеті, Стамбул, Түркия.
E-mail: as.sarsenbayeva@gmail.com

СЕМАНТИКАЛЫҚ ӘДІСТЕРДІ ЗАҢНАМА САЛАСЫНДА ҚОЛДАНУ: АГГЛЮТИНАТИВТІ МӘТІНДЕРДІ ТАЛДАУҒА АРНАЛҒАН ИНТЕЛЛЕКТУАЛДЫ ЖҮЙЕ

Сарсенбаева Асия — Әл-Фараби атындағы Қазақ ұлттық университетінің PhD докторанты, Алматы, Қазақстан, Email: as.sarsenbayeva@gmail.com, https://orcid.org/0009-0008-0053-1182;
Рахимова Диана — PhD, доцент, Әл-Фараби атындағы Қазақ ұлттық университетінің профессоры, Алматы, Қазақстан,
E-mail: drakhimova060@gmail.com, https://orcid.org/0000-0003-1427-198X;
Шормакова Асем — PhD, Әл-Фараби атындағы Қазақ ұлттық университетінің қауымдастырылған профессор міндетін атқарушы, Алматы, Қазақстан,
Email: shormakovaassem@gmail.com, https://orcid.org/0000-0002-1637-4643;
Мансурова Мадина — PhD, Әл-Фараби атындағы Қазақ ұлттық университетінің профессоры, Алматы, Қазақстан, Email: m_mansurova@kaznu.kz, https://orcid.org/0000-0001-6284-8283;
Адали Эшреф — PhD, Стамбул Техникалық университетінің профессоры, Стамбул, Түркия, Email: adali@itu.edu.tr, https://orcid.org/0000-0002-1561-8255.

Аннотауция. Бұл мақалада агглютинативті тілдердегі, оның ішінде қазақ және түрік тілдеріндегі терминдер, сөз тіркестері және қысқа сөз тіркестерінің семантикалық ұқсастығын анықтаудың әдісі ұсынылған. Семантикалық ұқсастық ақпаратты іздеу, сұрақ-жауап жүйелері, мәтінді жіктеу және құқықтық құжаттарды талдау сияқты көптеген табиғи тілді өңдеу тапсырмаларында маңызды рөл атқарады. Дегенмен, агглютинативті және ресурстары шектеулі тілдер үшін мұндай жүйелерді әзірлеу олардың күрделі морфологиясына, тілдік ресурстарының шектеулілігіне және аннотацияланған корпустардың аздығына байланысты қиындық тудыруда. Семантикалық ұқсастықты есептеу әдісі Ri коэффициентіне негізделген, ол терминдер мен сөз тіркестерінің контексттеріндегі жиілік қатынастарын ескере отырып есептеледі. Бұл коэффициент екі сөз арасындағы байланысты бағалауға мүмкіндік береді, және ұсынылған тәсіл сұрақ-жауап жүйелеріне интеграцияланған, мұнда бірнеше ұқсас сөз тіркестерінің нұсқалары берілгенде, ең тиісті және дұрыс жауап таңдалуы қажет. NGD, Jaccard және Cosine сияқты түрлі әдістер қарастырылып, олардың салыстырмалы зерттеуі жүргізілген. Алынған эксперименттік нәтижелер көрсеткендей, ұсынылған әдіс терминдерді анықтауда және таңдауда жоғары дәлдік көрсетеді, бұл әсіресе агглютинативті тілдердегі заңдық лексика саласында маңызды болып табылады. Эксперименттік бағалау шамамен 10 000 заң сөйлемінен тұратын қазақ тіліндегі заң мәтіндерінің корпусында және 8 400 термин мен шамамен 10 000 синонимнен тұратын мамандандырылған заң сөздігінде жүргізілді. Нәтижелер семантикалық ұқсастыққа ұсынылған тәсілдің агглютинативті тілдердегі заң мәтіндерін өңдеу үшін тиімді екенін және сұрақ-жауап жүйелері мен заң құжаттарын іздеу платформаларын қоса алғанда, интеллектуалды құқықтық ақпараттық жүйелерге біріктірілуі мүмкін екенін растайды. Сонымен қатар, әдісті ұқсас морфологиялық сипаттамалары бар басқа түркі және аз зерттелген тілдерге бейімдеу мүмкіндігі бар.

Түйін сөздер: семантикалық ұқсастық, синонимдер, қазақ тілі, түрік тілі, заң терминдері, агглютинативті тілдер

Алғыс**.** *Бұл зерттеу Қазақстан Республикасының жоғары білімі және Ғылым министрлігінің қолдауымен BR24993001 жобасымен қаржыландырылды.*

© Сарсенбаева А.К.[1]*, Рахимова Д.Р.[1], Шормакова А.Н.[1], Мансурова М.Е.[1], Адали Э.[2] , 2026.
[1]Казахский Национальный университет имени аль-Фараби, Алматы, Казахстан;
[2]Стамбульский Технический Университет, Стамбул, Турция.
E-mail: as.sarsenbayeva@gmail.com

ПРИМЕНЕНИЕ СЕМАНТИЧЕСКИХ МЕТОДОВ В ОБЛАСТИ ЗАКОНОДАТЕЛЬСТВА: ИНТЕЛЛЕКТУАЛЬНАЯ СИСТЕМА ДЛЯ АНАЛИЗА АГГЛЮТИНАТИВНЫХ ТЕКСТОВ

Сарсенбаева Асия — докторант Казахского Национального университета имени аль-Фараби, Алматы, Казахстан,
E-mail: as.sarsenbayeva@gmail.com, https://orcid.org/0009-0008-0053-1182;
Рахимова Диана — PhD, ассоциированный профессор Казахского Национального университета имени аль-Фараби, Алматы, Казахстан,
E-mail: drakhimova060@gmail.com, https://orcid.org/0000-0003-1427-198X;
Шормакова Асем — PhD, и.о. доцента Казахского Национального университета имени аль-Фараби, Алматы, Казахстан,
E-mail: shormakovaassem@gmail.com, https://orcid.org/0000-0002-1637-4643;
Мансурова Мадина — PhD, профессор Казахского Национального университета имени аль-Фараби, Алматы, Казахстан,
E-mail: m_mansurova@kaznu.kz, https://orcid.org/0000-0001-6284-8283;
Адали Эшреф — PhD, профессор в Стамбульском Техническом Университете, Стамбул, Турция,
E-mail: adali@itu.edu.tr, https://orcid.org/0000-0002-1561-8255.

Аннотация. В статье представлен семантический метод оценки сходства терминов, фраз и коротких выражений в агглютинативных языках, в частности в казахском и турецком. Семантическое сходство играет ключевую роль в задачах обработки естественного языка, включая информационный поиск, системы вопросов и ответов, классификацию текстов и анализ юридических документов. Разработка подобных систем для агглютинативных и малоресурсных языков остаётся сложной задачей из-за их морфологической сложности, ограниченности лингвистических ресурсов и недостаточного объёма аннотированных корпусов. Предложенный подход основан на коэффициенте R_i, который позволяет оценивать семантическое сходство на основе частотных соотношений слов в их контекстном окружении. Метод контекстного анализа формирует множества слов, совместно встречающихся с целевыми терминами, и вычисляет нормализованные меры близости между ними. Интеграция контекстной частотной информации и синонимических связей обеспечивает более точную оценку семантических отношений между юридическими терминами и выражениями. Для оценки эффективности метода проведено сравнительное исследование с использованием распространённых метрик сходства, включая коэффициент Жаккара, косинусное сходство и нормализованное расстояние Google. Экспериментальная проверка

выполнена на корпусе юридических текстов на казахском языке, включающем около 10 000 предложений, а также специализированном юридическом словаре, содержащем 8400 терминов и около 10 000 синонимов. Результаты показывают, что предложенный метод обеспечивает высокую точность оценки семантического сходства и может эффективно применяться для обработки юридических текстов на агглютинативных языках. Разработанный подход может быть интегрирован в интеллектуальные юридические информационные системы, включая системы вопросов и ответов и платформы поиска правовой информации. Кроме того, метод обладает потенциалом адаптации к другим тюркским и малоресурсным языкам с аналогичной морфологической структурой.

Ключевые слова: семантическое сходство, синонимы, казахский язык, турецкий язык, юридические термины, агглютинативные языки

Кіріспе. Сөздер немесе сөйлемдер арасындағы семантикалық ұқсастық олардың семантикалық байланыс дәрежесін сипаттайды. Бұл параметр табиғи тілді өңдеу саласында маңызды болып табылады, себебі ол көптеген жасанды интеллект тапсырмаларының тиімділігін анықтайды: мәтіндерді автоматты түрде санаттау және қорытындылау (Bhattacharya et al. 2020; Mandal et al., 2021), машиналық аударма (Bhattacharya et al., 2022), ақпаратты іздеу және деректерді өңдеу (Voronina, 2020), мәтінді талдау, веб-аналитика және басқа да қолданбалар (Sultanova et al., 2019). Сонымен қатар, когнитивті ғылымда семантикалық ұқсастық әртүрлі аналитикалық өлшеулер жүргізу және мағынаның қалыптасуы мен эволюциясы процестерін зерттеу үшін қолданылады.

Қысқа мәтіндердегі семантикалық ұқсастық мәселесі зерттеушілердің ерекше назарын аударды, бұл сандық ортада қысқа мәтін форматтарының кеңінен қолданылуына байланысты: өнім сипаттамалары, сурет тақырыптары мен тегтері, жаңалықтар тақырыптары және т.б. (Ayazbayev et al., 2023). Семантикалық ұқсастық білім беру технологияларында да маңызды рөл атқарады, мысалы, автоматтандырылған тестілеу мен бағалауда. Біздің зерттеуіміз аясында құқықтық мәтіндердің семантикалық ұқсастығы Қазақстан Республикасының заңнамасы үшін интеллектуалды сұрақ-жауап жүйесін әзірлеу үшін қажет.

Сот құжаттарының ұқсастық дәрежесін анықтау күрделі және маңызды міндет болып табылады, ол сот тәжірибесін талдау, ұқсас істерді анықтау және дәйексөзге сілтеме жасау бойынша ұсыныстарды тұжырымдау үшін өте маңызды. Қазіргі уақытта бұл мәселені шешудің екі негізгі тәсілі бар: желілік әдістер және мәтіндік модельдер. Олардың негізгі мақсаты - құқықтық прецеденттер арасындағы ұқсастық дәрежесін анықтау. Статистикалық ұқсастықты бағалаудың классикалық тәсілі семантикалық кеңістік құруға негізделген, мұнда сөздер мәтін корпусындағы таралуын көрсететін векторлармен көрсетіледі. Семантикалық ұқсастық осы векторлардың

салыстырмалы орналасуымен анықталады. Бұл тәсіл мағыналары ұқсас сөздер ұқсас контексттерде кездеседі деген таралу гипотезасына негізделген. Бұл гипотеза сөздер сөйлемде міндетті түрде бір-бірінің жанында емес, керісінше, олар бірдей контексттік элементтер жиынтығымен бірге кездеседі деп болжайды.

Заңды құжаттарды талдаудың заманауи әдістері әдетте мәтіндік мазмұнға немесе сілтеме желісінің құрылымына негізделген. Пахели Бхаттачарья және т.б. (Bhattacharya et al., 2022) жазған «Заңды іс құжаттарының ұқсастығы: Сізге желі де, мәтін де қажет» атты зерттеу екі тәсілдің де шектеулері бар екенін көрсетеді және заңдық ұқсастықты бағалаудың дәлдігін арттыру үшін мәтін мен желілік ақпаратты біріктіруді ұсынады. Көптеген NLP ресурстары жасалған ағылшын тілінен айырмашылығы, көптеген басқа тілдер корпустардың, сөздіктердің және құралдардың жетіспеушілігіне тап болады. Бұл әсіресе күрделі морфологиясы бар ресурстарға тапшы тілдерге қатысты. Сондықтан, көптеген семантикалық модельдер Ағылшын тілі үшін жасалған ұқсастықтар басқа тілдік жағдайларда қолданылмайды.

Біз жоғары морфологиялық өзгергіштікпен және шектеулі тілдік инфрақұрылыммен сипатталатын қазақ тілі үшін әдіс әзірлеу кезінде дәл осы мәселеге тап болдық. Сондықтан біз семантикалық ұқсастықты есептеуге негізделген тәсілді қабылдадық және тек қазақ тіліне ғана емес, сонымен қатар ұқсас тілдік ерекшеліктері бар басқа да ресурстарға тапшы тілдерге қолданылатын әдістемені әзірледік.

Әдеби шолу. Сөйлемдер арасындағы ұқсастықты анықтау тілді өңдеудің көптеген салаларында, соның ішінде мәтінді қайта пайдалануды анықтау, мәтіннің өзектілігі, парафразаны анықтау, ақпаратты алу, ақпаратты іздеу, қысқа жауаптарды бағалау, ұсыныс және бағалау сияқты қолданбалы маңызды міндет болып табылады. Ғылыми еңбектердегі мәтін ұқсастығын бағалаудың бұрынғы көптеген әдістері лексикалық тәсілдерге сүйенді, мұнда ұқсастық сәйкес келетін белгілер санымен немесе белгілер тізбегімен анықталады. Дегенмен, бұл әдістер мәтіндер синонимдік терминдерді қолданғанда немесе мағыналары ұқсас, бірақ бірдей болмаған кезде тиімсіз болады. Мысалы, «Парламент заңды қабылдады» және «Заң шығарушы жиналыс заң жобасын мақұлдады» немесе мәтіндердің мағыналары ұқсас, бірақ бірдей болмаған кезде, мысалы, «Президент заңға қол қойды» және «Мемлекет басшысы заң шығару бастамасын мақұлдады» сияқты.

Б. Бакиевтің «Қазақ тіліндегі мәтіндік құжаттардың синонимдерін ескере отырып ұқсастығын анықтау әдісі: TF-IDF кеңейтімі» (Bakiyev, 2022) еңбегінде синонимдерді ескеруді қамтитын TF-IDF әдісінің кеңейтімі ұсынылған. Ұсынылған әдістің тиімділігі қазақ тіліндегі мәтіндердің ұқсастығын бағалау үшін Cosine, Dice және Jaccard сияқты функцияларды пайдаланатын тәжірибелермен расталады. Зерттеудің мақсаты - екі мәтіндік құжаттың ұқсастығын салыстырмалы түрде талдау. Құжаттар Q құжаты - сұраныс және D құжаты ретінде белгіленеді. Бірінші кезең барлық сөздерді

ұқсастықты бағалауға дайындауды қамтиды, бұл маңызды емес қысқа сөздерді (мысалы, есімдіктер, сандар, таңбалар және т.б.) алып тастауды, содан кейін қалған сөздердің негізгі формаларын алуды қамтиды. Әдетте, бұл үшін Портер немесе Сноубол алгоритмдері қолданылады, бірақ бұл зерттеуде белгілі бір ережелерге негізделген қазақ тілінің стеминг алгоритмі қолданылды (Sultanova et al., 2019). Содан кейін сөздердің бұл негізгі формалары TF-IDF әдісін пайдаланып векторларға түрлендіріледі, бұл сандық кестелерді жасауға әкеледі. Осыдан кейін, құжаттардың ұқсастығын бағалау үшін косинус метрикасы немесе басқа да тиісті әдістер қолданылады. Ұсынылған әдіс мәтінде лексикалық элементтердің жеткілікті саны болған жағдайда екі қазақ тіліндегі құжаттың ұқсастығын бағалайды, бұл салыстыру процесін жеңілдетеді. Дегенмен, қысқа мәтіндерде семантикалық ұқсастық мәселесі күрделене түседі. Салыстыру үшін лексикалық элементтердің саны шектеулі болғандықтан, сәйкес терминдер арасындағы байланысты анықтау қажет.

«Контекст жиынтығын пайдаланып терминдердің семантикалық ұқсастығын анықтау» мақаласында Д.В. Бондарчук (Bondarchuk, 2020) семантикалық жағынан ұқсас терминдер ұқсас контексттерде қолданылады деген болжамға негізделген терминдердің семантикалық ұқсастығын есептеу әдісін ұсынды. Ұсынылған әдістің нәтижелері Жаккард қашықтығымен және 50 адамнан тұратын топ жүргізген бағалаумен салыстырылады. Нәтижелер жоғары корреляция коэффициентін көрсетті, бұл ұсынылған әдістің Жаккард қашықтығын есептеуге негізделген әдіске қарағанда объективтірек екенін көрсетеді (Bondarchuk, 2020). Дегенмен, бұл тәсілді қолдану жаңа мәселені тудырады: термин идентификаторларын таңдаудың қиындығы. Ресурстары аз тілдер үшін айқын анықталған құрылымы бар веб-ресурстарды табу қиын.

«Таратылған Apache Spark жүйесін пайдаланып, қазақ тіліндегі семантикалық жағынан ұқсас сөздерді анықтау» (Ayazbayev et al., 2023) жұмысының аясында жүргізілген қазақстандық ғалымдардың зерттеуінде семантикалық жағынан ұқсас сөздерді анықтау үшін NLPL репозиторийінен алынған векторлар және Polyglot сөз кірістірулері пайдаланылды. Косинус ұқсастығын есептеудің үлкен көлемін өңдеу үшін тапсырма Apache Spark платформасында параллель орындалды. Әрбір мақсатты сөз және сәйкес вектор үшін 176 643 косинус мәні есептелді, содан кейін олар ең қолайлы он кілт сөзді таңдау үшін сұрыпталды. Нәтижелер Polyglot және NLPL модельдері үшін сәйкесінше 88,61% және 92,2% дәлдікті көрсетті. Бұл семантикалық жағынан ұқсас сөздер Уикипедиядағы сәйкес мақалаларды табуға арналған Komekshy іздеу жүйесінде қолданылды. Дегенмен, бұл жұмыста ұқсастық тек сөздердің түбірлерімен анықталғанын, «қолды қылу» сияқты мүмкін фразеологиялық бірліктерді ескермегенін атап өткен жөн, бұл контексте бір нәрсені рұқсатсыз иеленуді білдіреді.

Кесте 1 – Семантикалық жақындық бойынша ғылыми еңбектер

Автор / Жыл	Атауы	Тіл	Әдіс	Нәтиже
Светлана Билощицкая, Арайм Тілеубаева, Александр Кучанский, Салтанат Шарипова – 2025 (Biloshchytska et al., 2025)	Агглютинативті тілдердегі мәтін ұқсастығын анықтау: гибридті N-грамм және семантикалық модельдерді қолдана отырып, қазақ тілінің жағдайын зерттеу	Қазақша	Гибридті N-Gram + TF-IDF + LSH + LSA + LDA + семантикалық модельдер	F1 ұпайы = 0,84, дәлдік = 1,00, еске түсіру = 0,73
Йесси Асри және т.б. басылымы – 2025 (Asri et al., 2025)	Косинус ұқсастығын пайдаланып контекстік ұқсастық үшін сөзді енгізу	ағылшын	Қарама-қарсы оқыту + аспектіге негізделген сөйлемдерді енгізу + косинус ұқсастығы	Ақпаратты іздеу тапсырмалары бойынша ~+3,97%-ға жақсарды
А.В.К.рюкова және т.б. – 2020 (Kryukova et al., 2020)	DKPro құралын пайдаланып мәтіндердің семантикалық ұқсастығын анықтау Ұқсастық	орыс	Семантикалық ұқсастық үшін жолдық метрикалар + ML модельдері	әртүрлі » мәтіндер үшін жіктеуіш негізіндегі тәсілдермен жақсы нәтижелер
З. Шен және З. Сяо – 2024 (Shen and Xiao, 2024)	Сөйлем деңгейіндегі және сөз тіркесі деңгейіндегі семантиканы біріктіретін қытай тіліндегі қысқа мәтіндік ұқсастық әдісі	Қытай	Сөйлем деңгейіндегі + Фразалық деңгейдегі семантика + сөз/сөз тіркестерінің ендірілуі	Қысқа мәтінді сәйкестендіру тапсырмаларының дәлдігі ≈ 90,16%

Кесте 1-де шетел тілдері мен қазақ тілінің семантикалық ұқсастығын талдау бойынша ақпараттық жүйелер саласындағы соңғы зерттеулер мен ғылыми еңбектер келтірілген. Қолданыстағы еңбектерді талдау көрсеткендей, аз ресурсты тілдерге арналған сөйлемдердегі семантикалық ұқсастықтарды анықтау барлық қолжетімді зерттеулерді шолу арқылы әдіснаманы әзірлеуді қажет ететін өзекті мәселе болып табылады. Біздің зерттеуімізде мақсат - қысқа сөйлемдермен байланысты болуы мүмкін заңдық сұраулардағы синонимдерді анықтау. Қысқа мәтіндерде семантикалық ұқсастықты анықтау міндеті салыстыруға арналған сөздердің шектеулі жиынтығымен күрделене түседі, бұл сәйкес терминдер арасындағы байланыстарды табуды маңызды етеді. Қолданыстағы әдістерді салыстырғаннан кейін, келесі тәсіл таңдалды, себебі ол агглютинативті және аз ресурсты тілдерге жарамды.

Қазақ және қырғыз тілдерінің тілдік жақындығы олардың ортақ шығу тегі мен тарихи өзара ықпалдастығымен түсіндіріледі. Екі тіл де түркі тілдер тобына кіреді және лексика, фонетика мен грамматика тұрғысынан ұқсастықтарға ие (Tüymebayev, 2024). Зерттеулер қазақ және қырғыз этностарының

қалыптасуына, сондай-ақ олардың дүниетанымы мен мәдениетінің дамуына бірдей тайпалар мен халықтар қатысқанын көрсетеді, бұл олардың тілдік туыстығын одан әрі дәлелдейді.

Мақалада (Toleush et al., 2021) көне ұйғыр тіліндегі жұп сөздер қазіргі қазақ тіліндегі баламаларымен салыстырылып, осы лексикалық бірліктердің сақталу деңгейі мен өзгерістері анықталған.

Кесте 2-де түркі тілдер тобына кіретін қазақ, қырғыз және түрік тілдері арасындағы лексикалық және фонетикалық ұқсастықтардың салыстырмалы талдауы ұсынылған. Кестеде күнделікті тұрмыста қолданылатын негізгі зат есімдер (су, бас, күн, үй және т.б.) қарастырылып, ортақ түркі түбірлерінің айқын байқалатыны көрсетілген (Bekmanova et al., 2017). Қазақ және қырғыз тілдеріндегі сөздер ресми кирилл әліпбиінде және латынша транскрипцияда берілген.

Көптеген жағдайларда мағынасы мен фонетикалық тұлғасы жағынан дерлік толық сәйкестік байқалады, мысалы: su – suu – su (су), jol – jol – yol (жол). Сондай-ақ түркі тілдеріне тән тұрақты фонетикалық сәйкестіктер көрінеді. Атап айтқанда:

Қазақ және қырғыз тілдеріндегі j дыбысы көбінесе түрік тіліндегі y дыбысына сәйкес келеді (мысалы, jol → yol, jyl → yıl);

kök – gök сияқты жұптарда k және g дауыссыздарының алмасуы сөздің позициясына тәуелді фонологиялық заңдылықты көрсетеді.

Бұл салыстырулар аталған тілдердің терең тілдік туыстығын айқындайды. Алынған нәтижелер тарихи лингвистика, салыстырмалы грамматика салалары үшін маңызды болып табылады, сондай-ақ табиғи тілдерді өңдеу (NLP) саласында, әсіресе көптілді немесе аралық тілдік мәтіндерді өңдеу модельдерін әзірлеуде практикалық қолданысқа ие (Kairat and Burkitbayeva, 2024).

Кесте 2 – Қазақ, түрік және қырғыз тілдеріндегі сөздердің салыстырмасы

Kazakh	**Turkish**	**Kyrgyz**
су (su)	su	суу (suu)
бас (bas)	baş	баш (bash)
күн (kün)	gün	күн (kün)
жыл (jyl)	yıl	жыл (jyl)
түн (tün)	tun	түн (tün)
ата (ata)	ata	ата (ata)
бала (bala)	bala	бала (bala)
жол (jol)	yol	жол (jol)
көк (kök)	gök	көк (kök)

Eşref Adalı зерттеулерінде (Adali, 2024) қазақ және түрік етістіктерінің морфологиялық ережелерін сипаттайтын онтологиялық модельдер құрастырылған. Бұл модельдер олардың морфологиялық ерекшеліктерін NLP жүйелерінде қолдануға мүмкіндік береді.

Қазақ және басқа агглютинативті тілдер арасындағы құрылымдық және

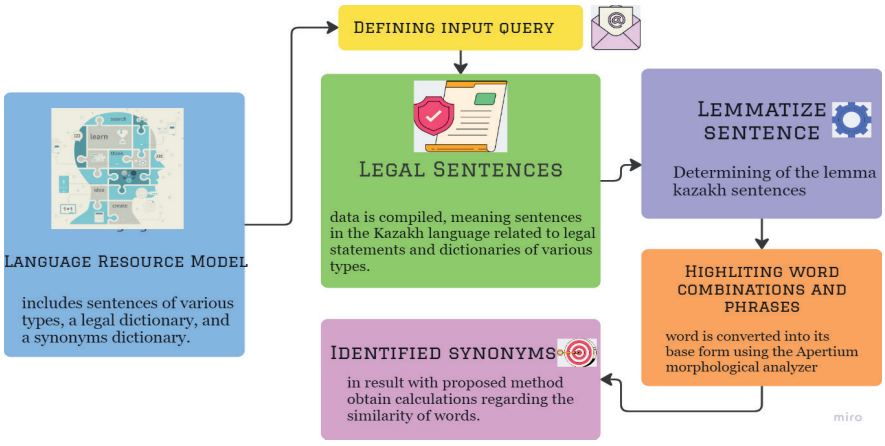
морфологиялық ұқсастықтарды ескере отырып, біз әзірлеген мәтіндерді өңдеу және талдау әдістерінің қырғыз, өзбек немесе түрік тілдеріне бейімделіп, қолданылуы мүмкін деген болжам жасауға болады.

Материалдар мен әдістер. Деректер жиынтығын жасау үшін egov.kz, online.zakon.kz және adilet.zan.kz порталдарын қоса алғанда, құқықтық құжаттарды қамтитын ашық қазақстандық ресурстар пайдаланылды. Бұл платформалар ұқсас құрылымға ие және құқықтық ақпараттың кең ауқымын ұсынады. adilet.zan.kz веб-сайты ең үлкен деректер көлеміне ие, кесте 3-те сайттағы қазақша құжаттар саны көрсетілген, 1947 жылдан 2023 жылға дейінгі құқықтық құжаттарды қамтиды және пайдаланудың қарапайымдылығымен ерекшеленеді. Дегенмен, бұл ресурстар бойынша нақты сұрақтарға жауап іздеу әдетте қолмен орындалады, бұл көп уақытты алады. Нәтижесінде 8,400 заңдық термин мен осы терминдерге арналған 10,000 синонимді қамтитын кешенді қазақ тіліндегі заңдық сөздік жасалды. Сондай-ақ 10,000 заңдық сөйлемнен тұратын корпус дайындалып, ол семантикалық талдау үшін пайдаланылды. Сөздік пен корпус үшін деректер ашық ресурстардан алынды, оның ішінде adilet.kz платформасы да бар, ол Қазақстанның қазіргі заң мәтіндерін ұсынады. Бұл деректер контексттер мен үлгі сөйлемдерді алу үшін негіз болып, заңдық лексика мен контексттердің әртүрлілігін қамтамасыз етіп, қазақ тілінде семантикалық модельдерді әрі қарай талдау және әзірлеу үшін пайдаланылды.

Кесте 3 – Adilet сайтынан алынған мәліметтер

Website	Sentences	Words	Number of documents	Size
https://adilet.zan.kz/kaz/	*506636*	*14656047*	*9575*	*233*

Алдын ала өңдеу бөлімі 1-суретте көрсетілген бірнеше тізбекті кезеңдерді қамтиды:



Сурет 1 – Тіл ресурстары модулі және алдын ала өңдеу құбыры

Тілдік ресурстар модуліне заңдық сөйлемдер, заңдық сөздік және синонимдік сөздік кіреді. 1-суретте көрсетілгендей, енгізу процесі бірнеше кезеңнен тұрады. Алдымен, қазақ тіліндегі заңдық сөйлемдер мен тиісті сөздіктерді қамтитын деректер жиынтығы дайындалды. Деректерді дайындағаннан кейін алынған сөз тіркестері мен сөз тіркестері 1-6 формулаларын пайдаланып есептеулер үшін пайдаланылды.

Нәтижелер. Терминдер мен сөз тіркестерін қамтитын деректерді дұрыс өңдеу үшін лемматизация қажет. Қазақ тілі күрделі морфологияға ие болғандықтан, сөздер мен сөз тіркестері әртүрлі формада болуы мүмкін. Оларды негізгі формаларына түрлендіру үшін Apertium морфологиялық анализаторы қолданылды. Мысалы, «Құқық қорғаушы заңды қорғады» деген сөйлемді Apertium көмегімен талдау келесі лексемаларды анықтайды:

құқық <n> <sg> <nom>
қорғаушы <n> <g> <атауы>
заңды <n> <g> <acc>
қорғадас <v> <tv> <өткен> <sg>

Деректерді дайындағаннан кейін, одан әрі есептеу үшін қазақ сөйлемдері қолданылады.

1. Сөздердің контекстік жиынтығын қалыптастыру.

C1 = {c 11,c 12,...,c1n} және C2 = {c21,c22,...,c2m} сәйкесінше w 1 және w 2 сөздерінің контекстік жиындары болсын. Бұл жиындарда w1 және w2 жиі бір контексте қолданылатын сөздер бар. Содан кейін біз сөздердің ортақ контекстік жиынын құрамыз: $C = C1UC2$

2. w_1 және w_2 сөздерінің әрқайсысы арасындағы нормаланған жақындықтарды есептеу:

$$sim(c_i, w_1) = \frac{f(c_i, w_1)}{\max f\ (w_1)} \qquad (1)$$

$$\text{sim}(c_i, w_2) = \frac{f\ (c_i, w_2)}{\max f(w_2)} \qquad (2)$$

мұндағы жақындық (c_i, w_1) - c_i және w_1 бірге кездесетін құжаттар саны , ал макс.жиілік (w_j) C алынған барлық сөздер үшін максималды жиілік ретінде есептеледі:

$$\max f\ (wj\ ) = max(f(\text{ci}\ , wj\ )), ci \in C \qquad (3)$$

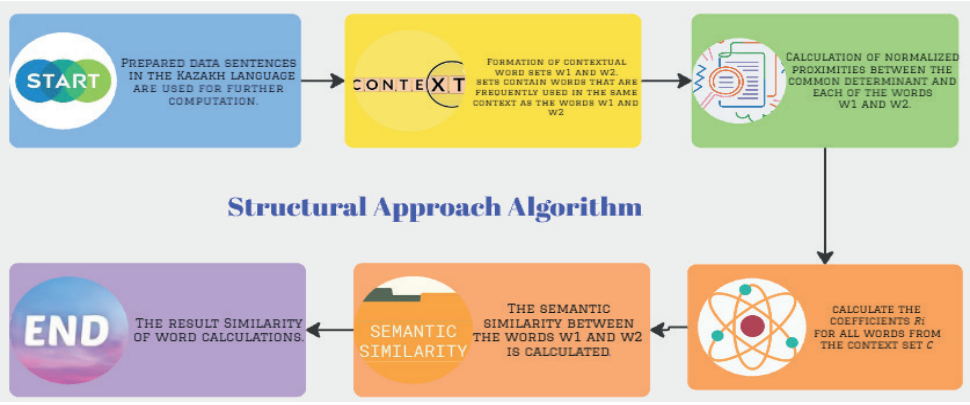
3. C контекст жиынындағы барлық сөздер үшін R_i коэффициенттерін төмендегі формуланы пайдаланып есептейміз:

$$R_i = \frac{min\{sim(c_1, w_1), sim(c_1, w_2)\}}{max\{sim(c_1, w_1), sim(c_1, w_2)\}} \qquad (4)$$

$$\text{SemSim}(w_1, w_2) = \frac{\sum_1^k \left(\frac{p_i R_i}{1+R_i} + s\right)}{1+s} \quad (5)$$

Ri – бүкіл үлгідегі w 1 және w 2 бірлесіп пайда болу коэффициенті болсын , екі сөз де бір құжатта кездескенде 2-ге тең, ал басқа жағдайда 1 болса, s – синонимдік коэффициент, егер w 1 және w 2 сөздері синонимдер болса 1-ге тең, ал басқа жағдайда 0-ге тең.

Анықтық үшін сөйлемнің семантикалық ұқсастығын есептеу алгоритмі ұсынылады. Ұсынылған семантикалық тәсілдің жалпы диаграммасы мен негізгі кезеңдері 2-суретте көрсетілген.



Сурет 2 - Ұсынылған семантикалық тәсілдің алгоритмі

4-кестеде семантикалық ұқсастықты есептеу үшін лексикалық ресурсты қалыптастыруда қолданылатын қазақ тіліндегі семантикалық жағынан жақын заң терминдері мен сөз тіркестерінің мысалдары келтірілген.

Кесте 4 – Қазақ тіліндегі мағыналық жағынан жақын заң терминдерінің мысалдары

Құқықтық ұғым санаты	**Семантикалық жағынан ұқсас терминдер**
Нормативтік құқықтық актілер	заң , қаулы , акт, құжат
Заңды әрекеттер	бас тарту , тоқтату , қысқарту
Қылмысқа қатысу	қылмас жасауға айдап салу, қылмыс жасауға азғыру
Сынақ	сот ісі , сот процесі , соттық қудалау
Заңды рөлдер	адвокат, қорғаушы
Жауапкершілікті тоқтату	ақтау , кешу , жазадан босату

Әдістің тиімділігін «заң» және «қаулы » сөздерінде тексеріп көрейік, оларды қолданыстағы қазақ корпустары мен сөздіктерін пайдаланып оқытайық. Бұл жағдайда w1 «заң», ал w2 «қаулы». Сондай-ақ, осы сөздер үшін контекст жиынтықтарын жасаймыз:

С 1 ={ зан , қабылда, шеш, бекіт, отыр, ен , өзгерт }

С 2 ={ қаулы, жанарт, бекіт, өзгерт, ауыс }

«Заң» және «қаулы» сөздері мағынасы жағынан ұқсас (синонимдер) болғандықтан, егер құжатта «заң» сөзі болса, онда «қаулы» сөзі де бар деп есептейміз. 5-кестеде 1-формуланы пайдаланып есептелген ортақ контекст жиынтығы мен сөздер арасындағы қалыпқа келтірілген жақындықтар көрсетілген.

Кесте 5 – Қалыптастырылған жақындық арасында ортақ контекстік жиынтық және заң және қаулы сөздері

	заң	*қаулы*
қабылда	0,653	0,609
шеш	0,555	0,476
бекіт	0,667	0.5001
ен	0,143	0,0224
өзгер	0,833	0.1875

Мысалы , 5-кестедегі бірінші жол үшін есептеулер келесідей орындалады:

$$sim(c_i, w_1) = \frac{f\,(c_i, w_1)}{\max f\,(w_1)} = 980/1500=0{,}653$$

$$sim(c_i, w_2) = \frac{f\,(c_i, w_2)}{\max f(w_2)} = 672/1102=0.609$$

Ri коэффициенттерін есептейміз; нәтижелері 6-кестеде көрсетілген. Есептеу жалпы контекст жиынынан алынған әрбір сөз үшін жүргізілді. Мысалы, «заң» және «қаулы» сөздері үшін *Ri* коэффициенттері келесі мәндерге ие:

$$R_i = \frac{min\{sim(c_1, w_1), sim(c_1, w_2)\}}{max\{sim(c_1, w_1), sim(c_1, w_2)\}} = 0.609/0.653=0.93$$

Кесте 6 – R_i коэффициенті (*қаулы, заң*)

заң	0,93
қабылда	0,87
шеш	0,63
бекіт	0,74
ен	0,58
өзгерт	0,61
ауыс	0,47
жаңарт	0,78
қаулы	0,93

Осы коэффициенттерді есептегеннен кейін, семантикалық ұқсастықты нақты есептеуге көшуге болады. (1) формуланы пайдаланып, « қаулы » және

«заң» сөздерінің ұқсастығын есептеп, 0,93 аламыз. Сондай-ақ, семантикалық ұқсастықты Жаккард қашықтығын (Cahyani et al., 2025) пайдаланып есептеп, 0,38 аламыз. Қосымша есептеулерсіз нәтижелердің бір-бірінен айтарлықтай ерекшеленетіні анық.

Семантикалық ұқсастықты есептеудің ұсынылған әдісінің тиімділігін бағалау үшін кеңінен қолданылатын семантикалық ұқсастық көрсеткіштерімен, атап айтқанда косинуспен салыстырмалы талдау жүргізілді. Ұқсастық, Жаккард Ұқсастық және қалыпқа келтірілген Google қашықтығы (NGD).

Косинус ұқсастығы — салыстырылатын мәтіндердің векторлық көріністері жасалады және осы векторлар арасындағы бұрыштың косинусы есептеледі (Kanishkaa and Santhi, 2025).

$$cos\theta = C(\vec{X},\vec{Y}) = C(\overrightarrow{Y},\vec{X}) = \frac{\vec{X}\times\vec{Y}}{\overrightarrow{X\vee\times\vee\vec{Y}\vee}=\frac{\Sigma_{i=1}^{n} x_i y_i}{\sqrt{\Sigma_{i=1}^{n} x_i^2}\sqrt{\Sigma_{i=1}^{n} y_i^2}}} \qquad (6)$$

Жаккард өлшемі — құжаттар n-граммға бөлінеді, ал Жаккард коэффициенті жалпы n-грамм санының n-граммның жалпы санына қатынасы ретінде есептеледі.

$$J(\vec{X},\vec{Y}) = \frac{\vec{X}\times\vec{Y}}{|\vec{X}|^2\times|\vec{Y}|^2-\vec{X}\times\vec{Y}} \qquad (7)$$

мұндағы x_i – X y_i векторының компоненті, - Y векторының компоненті, n - ұзындық X және Y векторлары.

NGD (Қалыптанған Google қашықтығын) келесі формуланы пайдаланып есептеуге болады (Ihnaini, 2024).:

$$NGD(x,y) = \frac{max\{logf(x), logf(y)\} - logf(x,y)}{logN - min\{logf(x), logf(y)\}} \qquad (8)$$

мұндағы N - веб-беттердің жалпы саны, Google іздеу жүйесімен өңделген,

f(t1) және f(t2) - t1 және t2 терминдері бөлек кездесетін беттер саны,

f (t1, t2)- t1 және t2 терминдері бірге орналасқан беттер саны.

Бұл көрсеткіштер бастапқы салыстыру әдістері ретінде пайдаланылды, бұл бізге ұсынылған тәсілдің қазақ тіліндегі заң терминдері мен сөз тіркестерін өңдеу кезіндегі артықшылықтарын сандық бағалауға мүмкіндік берді. Салыстырмалы эксперимент нәтижелері 7-кестеде келтірілген.

Кесте 7 – Сөз жұптары үшін метрикалар мен ұсынылған әдіс арасындағы салыстырмалы талдау

Бірнеше сөз	Jaccard	Cossine	NGD	Proposal method
Заң қабылда	0,282	0,44	0,83	0,96

Қаулы енгіз	0,311	0,561	0,783	0,89
Заң ен	0,572	0,697	0,802	0,901

8-кестеде ұсынылған әдістің тиімділігінің салыстырмалы нәтижесі сөз тіркестері үшін Jaccard қашықтығымен, NGD және Cossine мәндерімен салыстырғанда көрсетілген.

Кесте 8 – Салыстырмалы талдау метрика және сөз тіркестері үшін ұсынылған әдіс

Сөз тіркестері	Jaccard	Cossine	NGD	Proposal method
қамаудағы жазбаша түсінік беру	0,0482	0,066	0,545	0,632
кісі өлтіруде айыпталған тұлға	0,058	0,068	0,601	0,728
кісі өлтіруде сезікті адам	0,012	0,034	0,045	0,0521

Жүргізілген семантикалық ұқсастықты есептеу нәтижелеріне сүйене отырып, бірнеше негізгі қорытынды жасауға болады. Біріншіден, ұсынылған әдіс, контекстік талдау мен Ri коэффициентін пайдалану арқылы, Жаккард, Косинус және NGD сияқты классикалық метрикалардан жоғары дәлдік көрсетті. Мысалы, «Заң қабылда» сөз тіркесі үшін ұсынылған әдіс коэффициенті 0,96 болды, ал Жаккард пен Косинус метрикалары сәйкесінше 0,282 және 0,44 көрсетті, бұл ұсынылған әдістің елеулі артықшылығын дәлелдейді.

Екіншіден, күрделі сөз тіркестері үшін, мысалы, «қамаудағы жазбаша түсінік беру», әдіс те жақсы нәтижелер көрсетті, бірақ бұл фразалар үшін көрсеткіштер төмен болды. Бұл сөз тіркесі үшін ұсынылған әдіс 0,632 нәтижесін көрсетті, ал Жаккард пен Косинус метрикалары 0,0482 және 0,066 көрсетті. Бұл күрделі сөз тіркестері үшін контекстік деректердің аздығына байланысты семантикалық ұқсастықты есептеу қиынырақ екенін көрсетеді.

Кесте 9. Түрік және қазақ тілдеріндегі семантикалық ұқсастықтың әртүрлі әдістері бойынша F1-ұпайлары

Әдіс	F1-score (қазақ тілі)	F1-score (түрік тілі)
NGD	0,28	0,27
Jaccard	0,26	0,25
Cossine	0,24	0,23
Proposal method	0,69	0,48

Кесте 9 нәтижелер бойынша, NGD әдісінің F1-ұпайы үшін түрік тілінде 0,2764, ал қазақ тілінде 0,2824 болды. Бұл көрсеткіштердің жақын болуы, екі тілдің де агглютинативті құрылымға ие болуымен және ұқсас интернет-ресурстардан алынған деректермен қамтамасыз етілуімен түсіндіріледі. Алайда, қазақ тілінің көрсеткіші сәл жоғары, бұл қазақ тіліндегі заңды терминдердің кеңірек корпусы мен нақты терминдер арасындағы байланыстардың күшті екенін көрсетеді.

Jaccard әдісі бойынша, түрік тілінде 0,2587, ал қазақ тілінде 0,2653 F1-ұпайы есептелді. Бұл айырмашылық, қазақ тіліндегі терминдер мен фразалардың көбірек дәл сәйкес келуімен байланысты болуы мүмкін, әсіресе заң мәтіндерінде. Осылайша, қазақ тілінің терминологиялық базасы жоғары сапалы әрі көпшілікке ашық болғандықтан, Jaccard метрикасы сәл жоғары нәтиже көрсетті.

Cosine әдісінде түрік тілінде 0,2419, қазақ тілінде 0,2421 F1-ұпайы болды. Бұл әдіс бойынша нәтижелер бір-біріне өте жақын, бұл көрсеткіштердің контекстуалды ұқсастықтарды бағалаудағы тиімділігін білдіреді, бірақ толық емес контекстер мен морфологиялық ерекшеліктерді ескермеуі мүмкін.

Ұсынылған әдіс екі тіл үшін де едәуір жоғары нәтижелер көрсетті: қазақ тілінде F1-ұпайы 0,69, ал түрік тілінде 0,48 болды. Мұндай айырмашылық ұсынылған тәсілдің контекстік жиындарды және Ri коэффициентін пайдалану арқылы семантикалық байланыстарды тереңірек модельдейтінімен түсіндіріледі. Қазақ тіліндегі жоғары көрсеткіш корпус деректерінің құрылымдық бірізділігі мен контексттердің тығыздығымен байланысты болса, түрік тіліндегі салыстырмалы түрде төмен нәтиже корпус көлемі мен лексикалық таралу ерекшеліктерінің әсерінен болуы мүмкін. Жалпы алғанда, алынған нәтижелер ұсынылған әдістің агглютинативті тілдердегі семантикалық ұқсастықты бағалауда тиімді тәсіл екенін дәлелдейді.

Эксперименттік есептеулер үшін 10000 заңдық сөйлемнен тұратын монолингвальды корпус қолданылды, сондай-ақ деректер Adilet.kz платформасынан алынған. Бұл платформа Қазақстанның заң мәтіндерінің соңғы деректерін ұсынады. Осы платформадағы корпус 506636 сөйлемнен тұрады, ал жалпы көлемі 14656047 сөз болды, бұл контексттер мен сөз тіркестерін талдау кезінде жоғары дәлдікті қамтамасыз етеді. Зерттеу жұмыстары түрік тіліндегі семантикалық ұқсастықты анықтау бойынша жүргізілді, автор ұсынған әдіс жиілік талдауы мен Ri коэффициентіне негізделген. Түрік тілі үшін жиілік деректері Түрік Ұлттық Корпусының (TNC) мәтіндерінен алынды. Корпустың жалпы көлемі шамамен 10 миллион сөз тіркесінен тұрады.

Талдаудың дәлдігін қамтамасыз ету үшін барлық сөздер Zemberek ашық бастапқы морфологиялық анализаторы арқылы лемматизацияланды. Контекстуалды етістіктер (мысалы, kabul et – қабылдау, onayla – мақұлдау, değiştir – өзгерту) қолмен таңдалды, олар мақсатты есімдермен (мысалы, iş – жұмыс, görev – тапсырма) жиіліктік қатынастарына негізделіп алынды. Осы жиіліктер кейін семантикалық ұқсастықты есептеу үшін ұсынылған Ri коэффициентімен қолданылды.

Алынған деректер көрсеткендей, ұсынылған әдіс универсалдылыққа ие және қазақ тілі мен түрік тілі сияқты басқа агглютинативті тілдерге де бейімделуі мүмкін.

Осылайша, эксперименттік нәтижелер ұсынылған әдістің жоғары тиімділігін және оның қазақ тіліндегі заң мәтіндерін өңдеуде қолдануға

жарамдылығын растайды. Әдіс семантикалық ұқсастықты дәл және сенімді бағалауды қамтамасыз етеді, агглютинативті тілдер үшін бұл тәсілдің болашағы зор екені көрсетілді.

Талқылау. Эксперименттік зерттеу семантикалық ұқсастықты есептеудің ұсынылған әдісін Косинус, Жаккард және NGD сияқты классикалық семантикалық ұқсастық көрсеткіштерімен салыстырды. Эксперименттер қазақ тіліндегі заң мәтіндерінің корпусында жүргізілді және жеке терминдерді де, орфографиялық сөз тіркестерін де, қысқа сөз тіркестерін де қамтыды.

Сөз жұптарының салыстырмалы талдау нәтижелері 5-кестеде келтірілген. Алынған мәндер ұсынылған әдістің бастапқы көрсеткіштермен салыстырғанда семантикалық ұқсастықтың жоғары дәрежесін көрсететінін көрсетеді. Бұл контекст жиынтықтарын пайдалану және терминдер арасындағы байланыстарды симметриялы бағалаумен байланысты, бұл құқықтық саладағы семантикалық байланыстарды дәлірек есепке алуға мүмкіндік береді.

6-кестеде көп сөзден тұратын сөз тіркестерінің эксперименттік нәтижелері көрсетілген. Ұсынылған деректерден көрініп тұрғандай, тұтастай алғанда сөз тіркестерінің семантикалық ұқсастық мәндері жеке сөздерге қарағанда төмен. Бұл көп сөзден тұратын құрылымдарды өңдеудің күрделілігінің жоғарылығымен, сондай-ақ қазақ тілінің морфологиялық ерекшеліктерінің және үш немесе одан да көп сөздердің бірге қолданылу жиілігінің таралуының әсерімен түсіндіріледі.

Әртүрлі метрикаларды пайдаланып алынған нәтижелерді салыстыру косинус ұқсастығы және Жаккард ұқсастығы сияқты классикалық тәсілдердің, әсіресе шектеулі деректермен, заң терминдері арасындағы семантикалық қатынастарды анықтауға онша қабілетті емес екенін көрсетеді. NGD метрикасы тұрақты мәндерді көрсетеді, бірақ жиілік таралуына және деректер көздеріне сезімтал болып қалады. Дегенмен, ұсынылған әдіс контекст пен синонимдік қатынастарды ескере отырып, семантикалық ұқсастықты сенімдірек бағалауды қамтамасыз етеді.

Жалпы алғанда, эксперименттік нәтижелер қазақ тіліндегі заң терминдері мен сөз тіркестерінің семантикалық ұқсастығын талдау үшін ұсынылған тәсілдің тиімділігін растайды. Әдіс әртүрлі мәтін түрлерінде тұрақты жұмыс істеуді көрсетті және заң құжаттарын, әсіресе ресурстары аз тілдерде, ақылды іздеу және талдау үшін пайдаланылуы мүмкін.

Қорытынды. Бұл мақалада ресурстары аз тілдің нақты терминологиясы мен сипаттамаларын ескере отырып, қазақ тіліндегі заң мәтіндерінің семантикалық ұқсастығын анықтауға контекстке негізделген тәсіл ұсынылады. Жалпы мақсаттағы статистикалық өлшемдерден айырмашылығы, әзірленген әдіс мамандандырылған заң саласындағы семантикалық қатынастарды талдауға бағытталған және үлкен, белгіленген деректер корпустарын пайдалануды талап етпейді.

Жүргізілген семантикалық ұқсастықты есептеу нәтижелері бойынша ұсынылған әдіс классикалық метрикалар, атап айтқанда Жаккард және NGD

көрсеткіштерінен айтарлықтай жоғары нәтижелер көрсетті. Контекстік талдау мен Ri коэффициентін қолданатын ұсынылған әдіс, әсіресе заң мәтіндерінде, семантикалық ұқсастықты анықтауда жоғары дәлдік көрсетті. Мысалы, «Заң қабылда» сөз тіркесі үшін Ri коэффициенті 0,96 болды, бұл Жаккард пен Косинус нәтижелерімен салыстырғанда 0,282 және 0,44 айтарлықтай жоғары. Бұл ұсынылған әдістің агглютинативті тілдер, мысалы, қазақ тілінде терминдердің семантикалық ұқсастығын дәл бағалайтынын көрсетеді.

Түрік тілінде де ұсынылған әдіс жоғары дәлдікті көрсетті, бірақ жалпы F1-ұпайлары төмен болды (0,546), бұл екі тілдің агглютинативті құрылымдарының ұқсастығы мен заңдық лексикада жалпы семантикалық байланыстардың бар екенін дәлелдейді. Алайда, күрделі сөз тіркестері, мысалы, «қамаудағы жазбаша түсінік беру», нәтижелерінің дәлдігі төмендеді, өйткені бұл тіркестер корпуста жиі кездеспейді. Алайда, ұсынылған әдіс әлі де жоғары дәлдікті көрсетті, бұл Жаккард пен Косинус метрикаларына қарағанда контекст пен синонимдер арасындағы байланыстарды дәл есептеуге мүмкіндік береді. Бұл зерттеу қазақ және түрік тілдеріндегі заң мәтіндерін өңдеуге арналған ұсынылған әдістің тиімділігін дәлелдейді және оны басқа агглютинативті тілдерге бейімдеуге мүмкіндік береді.

Алайда, күрделі сөз тіркестері, мысалы, «қамаудағы жазбаша түсінік беру», нәтижелерінің дәлдігі төмендеді, өйткені бұл тіркестер корпуста жиі кездеспейді. Алайда, ұсынылған әдіс әлі де жоғары дәлдікті көрсетті, бұл Жаккард пен Косинус метрикаларына қарағанда контекст пен синонимдер арасындағы байланыстарды дәл есептеуге мүмкіндік береді. Бұл зерттеу қазақ тіліндегі заң мәтіндерін өңдеуге арналған ұсынылған әдістің тиімділігін дәлелдейді және оны басқа агглютинативті тілдерге бейімдеуге мүмкіндік береді. Ұсынылған тәсілді құқықтық құжаттарды іздеу, салыстыру және түсіндіру жүйелерін қоса алғанда, интеллектуалды құқықтық ақпаратты өңдеу жүйелерінде, сондай-ақ сұрақ-жауап жүйелерінде қолдануға болады. Оны қолдану ұқсас терминдер мен сөз тіркестерін түсіндірудің дәлдігін жақсартады, бұл құқықтық мәтіндер үшін өте маңызды. Әрі қарайғы зерттеу бағыттарына құқықтық деректер корпусын кеңейту, лексикалық ресурстарды толықтыру және әдісті ұзын мәтін фрагменттерімен және басқа да аз ресурстарды қажет ететін тілдермен жұмыс істеуге бейімдеу кіреді.

References

Adali E. (2024) The logic of Turkish language for NLP. Journal of Problems in Computer Science and Information Technologies, 2:34–46 (in Eng.).

Asri Y., Kuswardani D., Sari A.A., Ansyari A.R. (2025) Word embedding for contextual similarity using cosine similarity. Indonesian Journal of Electrical Engineering and Computer Science, 38:1170–1180. DOI: 10.11591/ijeecs.v38.i2.pp1170-1180 (in Eng.).

Ayazbayev D., Bogdanchikov A., Orynbekova K., Varlamis I. (2023) Defining semantically close words of Kazakh language with distributed system Apache Spark. Informatics, 7(4):160. DOI: 10.3390/informatics7040160 (in Eng.).

Bakiyev B. (2022) Method for determining the similarity of text documents for the Kazakh language taking into account synonyms: Extension to TF-IDF (in Eng.).

Bekmanova G., Sharipbay A., Altenbek G., Adali E., Zhetkenbay L., Kamanur U., Zulkhazhav A. (2017) A uniform morphological analyzer for the Kazakh and Turkish languages. Proceedings of AIST (Supplement):20–30 (in Eng.).

Bhattacharya P., Ghosh K., Pal A., Ghosh S. (2020) Hier-SPCNet: A legal statute hierarchy-based heterogeneous network for computing legal case document similarity. Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval:1657–1660. DOI: 10.1145/3397271.3401191 (in Eng.).

Bhattacharya P., Ghosh K., Pal A., Ghosh S. (2022) Legal case document similarity: You need both network and text. Information Processing & Management, 59(6):103069. DOI: 10.1016/j.ipm.2022.103069 (in Eng.).

Biloshchytska S., Tleubayeva A., Kuchanskyi O., Sharipova S. (2025) Text similarity detection in agglutinative languages: A case study of Kazakh using hybrid n-gram and semantic models. Applied Sciences, 15:6707. DOI: 10.3390/app15126707 (in Eng.).

Bondarchuk D.V. (2020) Opredelenie semanticheskoi blizosti termov s pomoshchyu kontekstnogo mnozhestva [Determination of semantic similarity of terms using a contextual set] (in Russ.).

Cahyani A.D., Fathoni M.W., Rachman F.H., et al. (2025) Automatic essay scoring: Leveraging Jaccard coefficient and cosine similarity with n-gram variation. arXiv preprint. DOI: 10.48550/arXiv.2501.01234 (in Eng.).

Ihnaini B. (2024) Semantic similarity on multimodal data: A comprehensive survey. Journal of Information and Data Management, 15(2):1–20. DOI: 10.5753/jidm.2024.3210 (in Eng.).

Kairat A.A., Burkitbayeva Sh.D. (2024) Reduplicated words in Old Uyghur and modern Kazakh languages. Tiltanym, 2:133–139. DOI: 10.55491/2411-6076-2024-2-133-139 (in Eng.).

Kanishkaa S., Santhi K. (2025) A comparative analysis of Jaccard and cosine similarity for plagiarism detection. International Journal of Innovative Research in Electrical, Electronics, Instrumentation and Control Engineering, 13(3). DOI: 10.17148/IJIREEICE.2025.13305 (in Eng.).

Kryukova A.V. et al. (2020) Opredelenie semanticheskih svyazei s pomoshchyu DKPro Similarity [Determination of semantic relations using DKPro Similarity] (in Russ.).

Mandal A., Ghosh K., Ghosh S., Mandal S. (2021) Unsupervised approaches for measuring textual similarity between legal court case reports. Artificial Intelligence and Law, 29:1–25. DOI: 10.1007/s10506-020-09260-2 (in Eng.).

Shen Z., Xiao Z. (2024) A Chinese short text similarity method integrating sentence-level and phrase-level semantics. Electronics, 13:4868. DOI: 10.3390/electronics13244868 (in Eng.).

Sultanova N., Kozhakhmet K., Jantayev R., Botbayeva A. (2019) Stemming algorithm for Kazakh language using rule-based approach. Proceedings of the 15th International Conference on Electronics, Computer and Computation (ICECCO):1–4. DOI: 10.1109/ICECCO48375.2019.9043249 (in Eng.).

Toleush A., Israilova N., Tukeyev U. (2021) Development of morphological segmentation for the Kyrgyz language on complete set of endings. Proceedings of the 13th Asian Conference on Intelligent Information and Database Systems (ACIIDS). DOI: 10.1007/978-3-030-73100-7_46 (in Eng.).

Tuymebayev J. (2024) Comparative grammar of the Kazakh and Kyrgyz languages (in Eng.).

Voronina L.V. (2020) Voprosy relevantnosti referenta i referentnogo vyrazheniya antecedentno-anaforicheskogo kompleksa, realizuyushchego semantiku tseli v politicheskom diskurse [Issues of relevance of the referent and referential expression of the antecedent-anaphoric complex implementing the semantics of purpose in political discourse]. Otechestvennaya filologiya, 5:16–25. DOI: https://doi.org/10.18384/2310-7278-2020-5-16-25 (in Russ.).

Publication Ethics and Publication Malpractice
in the journals of the Central Asian Academic Research Center LLP

For information on Ethics in publishing and Ethical guidelines for journal publication see http://www.elsevier.com/publishingethics and http://www.elsevier.com/journal-authors/ethics.

Submission of an article to the journals of the Central Asian Academic Research Center LLP implies that the described work has not been published previously (except in the form of an abstract or as part of a published lecture or academic thesis or as an electronic preprint, see http://www.elsevier.com/postingpolicy), that it is not under consideration for publication elsewhere, that its publication is approved by all authors and tacitly or explicitly by the responsible authorities where the work was carried out, and that, if accepted, it will not be published elsewhere in the same form, in English or in any other language, including electronically without the written consent of the copyright-holder. In particular, translations into English of papers already published in another language are not accepted.

No other forms of scientific misconduct are allowed, such as plagiarism, falsification, fraudulent data, incorrect interpretation of other works, incorrect citations, etc. The Central Asian Academic Research Center LLP follows the Code of Conduct of the Committee on Publication Ethics (COPE), and follows the COPE Flowcharts for Resolving Cases of Suspected Misconduct (http://publicationethics.org/files/u2/New_Code.pdf). To verify originality, your article may be checked by the Cross Check originality detection service http://www.elsevier.com/editors/plagdetect.

The authors are obliged to participate in peer review process and be ready to provide corrections, clarifications, retractions and apologies when needed. All authors of a paper should have significantly contributed to the research.

The reviewers should provide objective judgments and should point out relevant published works which are not yet cited. Reviewed articles should be treated confidentially. The reviewers will be chosen in such a way that there is no conflict of interests with respect to the research, the authors and/or the research funders.

The editors have complete responsibility and authority to reject or accept a paper, and they will only accept a paper when reasonably certain. They will preserve anonymity of reviewers and promote publication of corrections, clarifications, retractions and apologies when needed. The acceptance of a paper automatically implies the copyright transfer to the Central Asian Academic Research Center LLP.

The Editorial Board of the Central Asian Academic Research Center LLP will monitor and safeguard publishing ethics.

Правила оформления статьи для публикации в журнале смотреть на сайтах:
www:nauka-nanrk.kz
http://physics-mathematics.kz/index.php/en/archive
ISSN2518-1726 (Online),
ISSN 1991-346X (Print)

Ответственный редактор *А. Ботанқызы*
Редакторы: *Д.С. Аленов, Т. Апендиев*
Верстка на компьютере: *Г.Д. Жадырановой*

Подписано в печать 31.03.2026.
Формат 60x881/8.
20,0 п.л. Заказ 1.

«Central Asian Academic Research Center» LLP
Алматы, Шевченко к-сі, 28