

Мақала
FTAMP: 16.31.21

ҚАЗАҚ ТІЛІНДЕГІ СИНТЕТИКАЛЫҚ ЖӘНЕ АНАЛИТИКАЛЫҚ СЕБЕПТІК ҚҰРЫЛЫМДАРДЫҢ ТИПОЛОГИЯЛЫҚ АЙЫРМАШЫЛЫҒЫ¹

Р. Таберхан¹, М.А. Самбетбаева^{2*}

^{1,2} Л.Н. Гумилев атындағы Еуразия ұлттық университеті, Сәтпаев, 2, Астана, 010008,
Қазақстан

¹ roman.tn82@gmail.com, ² *sambetbayeva_ma_1@enu.kz

Аннотация. Мақалада KazCausal сұлбасы негізінде қазақ тіліндегі себеп-салдарлық құрылымдарды қолмен аннотациялау және құрылымдық-семантикалық тұрғыдан жүйелеу әдістемесі ұсынылады.

Зерттеудің мақсаты – қазақ тіліндегі себептілікті білдіретін морфологиялық, лексикалық және синтаксистік құралдарды формалды аннотациялық модель арқылы сипаттау және табиғи тілді өңдеу міндеттеріне бейімделген корпус құрудың ғылыми негізін көрсету.

Зерттеу нысаны ретінде ғылыми, публицистикалық және нормативтік мәтіндерден іріктелген 1250 каузалды сөйлемдер алынды. Барлық сөйлемдер себеп-салдарлық қатынасты білдіретін тілдік бірлік ретінде қарастырылып, үш негізгі құрылымдық типке бөлінді: SYNTHETIC_CAUSE, ANALYTICO_SYNTHETIC_CAUSE және ANALYTIC_CAUSE.

Нәтижелер синтетикалық құрылымдардың 431 сөйлемін (34,5%), аналитика-синтетикалық құрылымдардың 425 сөйлемін (34,0%), ал аналитикалық құрылымдардың 394 сөйлемін (31,5%) құрайтынын көрсетті. Ұсынылған сұлба себеп фрагментін, салдар фрагментін, каузалдық маркерді, құрылымдық типті, нақты морфологиялық форманы және семантикалық ішкі типті қатар белгілеуге мүмкіндік береді. Аннотация сапасы $F1=0.81$ және Cohen's $k=0.72$ көрсеткіштері арқылы бағаланды.

Зерттеу нәтижелері қазақ тіліндегі себеп-салдарлық қатынастарды автоматты анықтауға арналған типологиялық тұрғыдан бейімделген NLP ресурстарын құруға негіз болады.

Түйін сөздер: Мәтінді аннотациялау, корпус лингвистикасы, NLP, агглютинативті тілдер, KazCausal, себеп-салдарлық қатынастар, аннотациялаушылар келісімі

DOI: <https://doi.org/10.32523/bulmathenu.2026/1.2>

2000 Mathematics Subject Classification: 68T50,68T07

1. Кіріспе

Себеп-салдарлық қатынастар табиғи тілдегі түсіндірмелі, аргументативтік және нормативтік құрылымдардың негізінде жатқан іргелі семантикалық байланыстардың бірі болып табылады. Олар оқиғалардың уәждемесі, салдары, логикалық тізбегі, сондай-ақ

¹Зерттеу Қазақстан Республикасы Ғылым және жоғары білім министрлігі Ғылым комитетімен қаржыландырылған (грант №AP22686434)

эмоциялық және мінез-құлықтық реакциялар туралы ақпаратты жеткізуге мүмкіндік береді. Сондықтан себептілік мәтіннің мағыналық ұйымдасуын талдауда негізгі рөл атқарады.

Табиғи тілді өңдеудің (Natural Language Processing, NLP) қолданбалы міндеттерінде себеп-салдарлық құрылымдар білімді автоматты түрде шығару, логикалық қорытынды жасау, дәлелдеме құрылымдарын талдау, түсіндіру генерациясы, эмоция мен өшпенді риториканы анықтау, мәтінге негізделген диагностикалық жүйелер құру сияқты бағыттарда кеңінен қолданылады.

Адам қоршаған ортадағы құбылыстарды көбіне себеп пен салдар байланысы арқылы түсінеді: бір оқиға екіншісін туындатады, әсер етеді немесе шектейді. Қазақ тілі де кез келген табиғи тіл секілді себеп пен салдарды білдірудің ең бай және көпқабатты жүйесіне ие. Мұндай конструкциялар әрекеттің уәжін, эмоциялық немесе әлеуметтік реакцияны, санкциялық талапты, сондай-ақ мақсат пен шартты көрсету үшін қолданылады.

Алайда, ағылшын (Penn Discourse Treebank, BECauSE Corpus), қытай, неміс және араб тілдерінде себептілікке арналған ірі аннотацияланған корпустар жасалғанымен, қазақ тілі бұл бағытта ұзақ уақыт бойы назардан тыс қалып келді. Бұған агглютинативті морфологиялық құрылымдардың күрделілігі және осы тілге арнайы әзірленген стандартталған аннотациялық схеманың болмауы себеп.

Себептілік мағынасы қазақ тілінде жиі уақыттық, шартты, модальдылық, эмоциялық және риторикалық мағыналармен қиылысып келеді. Сондықтан оны формалдандыру тек семантиканы ғана емес, морфология-синтаксистік өзара әрекеттесуді қатар талдауды талап етеді.

Осы олқылықтың орнын толтыру мақсатында бұл жұмыста KazCausal аннотациялық схемасы – қазақ тіліндегі себеп-салдарлық құрылымдарды лингвистикалық және типологиялық тұрғыдан жүйелі сипаттауға бағытталған алғашқы кешенді тәсіл ұсынылады. Сызба агглютинативті морфологияға тән формаларды (есімше тұлғалары: -ған, -атын, морфологиялық маркерлер: -ғандықтан, -май, септік формалары: -нан, -на, аналитикалық көрсеткіштер: үшін, кейін, соң, қарамастан және т.б.) қамтып, себептік қатынастардың мағыналық жіктелімімен байланыстыра аннотациялауға мүмкіндік береді.

Бұл зерттеу бұрын ұсынылған KazCausal: Manual Annotation of Causal Constructions in the Kazakh Language атты конференциялық жұмыстың жалғасы болып табылады [1]. Алдыңғы жұмыста KazCausal сұлбасының бастапқы аннотациялық принциптері ұсынылса, осы мақалада корпус құрылымы кеңейтіліп, себептік құрылымдардың типологиялық айырмашылықтары, статистикалық таралуы және аннотация сапасын бағалау өлшемдері тереңірек қарастырылады.

2. Зерттеу мәселесі бойынша әдебиеттерге шолу

Табиғи тілдегі себеп-салдарлық қатынастарды зерттеу – лингвистика мен табиғи тілді өңдеу NLP салаларындағы негізгі бағыттардың бірі. Себептілік когнитивтік, логикалық және прагматикалық процестердің өзегінде жатыр: ол оқиғаларды түсіндіру, шешімдерді уәждеу, эмоциялар мен санкцияларды интерпретациялау және әлеуметтік салдарларды сипаттау үшін қолданылады. Алайда, оның маңыздылығына қарамастан, мәтіндегі себептілікті аннотациялау күрделі мәселе болып қала береді. Бұл күрделілік лексикалық және синтаксистік алуандық тұрғысынан, тілдерлік айырмашылықтармен және уақыт, шарттылық, корреляция, модальдық сияқты басқа да семантикалық категориялармен өзара қиылысуымен байланысты.

В тексте статьи недостаточно чётко раскрыт вопрос архитектуры обработки данных радиолокационного сканирования. Неясно, осуществляется ли обработка сигналов и принятие решений в режиме реального времени непосредственно на борту мобильного робототехнического комплекса либо вычислительная обработка выполняется на удалённом сервере с последующей передачей результатов обратно на платформу. В случае бортовой обработки целесообразно уточнить технические характеристики вычислительного модуля (тип аппаратной платформы, вычислительная производительность, быстродействие, габариты и энергопотребление). Если же обработка данных реализована на серверной стороне,

Л.Н. Гумилев атындағы ЕҰУ хабаршысы. Математика, компьютерлік ғылымдар, механика сериясы, 2026, Том 154, №1
Вестник ЕНУ им. Л.Н. Гумилева. Серия Математика, компьютерные науки, механика, 2026, Том 154, №1

представляется важным указать параметры канала связи и оценить суммарную задержку принятия решения, включая передачу данных и обратную коммуникацию с мобильным роботом.

В статье не приведено достаточного описания применяемых методов и алгоритмов обработки радиолокационных данных, используемых для обнаружения взрывных устройств. Просьба уточнить, какие именно алгоритмические подходы применялись (например, методы цифровой обработки сигналов, признаки и критерии детектирования, использование машинного обучения или нейросетевых моделей), а также обосновать их выбор и эффективность в контексте поставленной задачи. В тексте статьи недостаточно чётко раскрыт вопрос архитектуры обработки данных радиолокационного сканирования. Неясно, осуществляется ли обработка сигналов и принятие решений в режиме реального времени непосредственно на борту мобильного робототехнического комплекса либо вычислительная обработка выполняется на удалённом сервере с последующей передачей результатов обратно на платформу. В случае бортовой обработки целесообразно уточнить технические характеристики вычислительного модуля (тип аппаратной платформы, вычислительная производительность, быстродействие, габариты и энергопотребление). Если же обработка данных реализована на серверной стороне, представляется важным указать параметры канала связи и оценить суммарную задержку принятия решения, включая передачу данных и обратную коммуникацию с мобильным роботом.

В статье не приведено достаточного описания применяемых методов и алгоритмов обработки радиолокационных данных, используемых для обнаружения взрывных устройств. Просьба уточнить, какие именно алгоритмические подходы применялись (например, методы цифровой обработки сигналов, признаки и критерии детектирования, использование машинного обучения или нейросетевых моделей), а также обосновать их выбор и эффективность в контексте поставленной задачи.

Халықаралық деңгейде себептілікті аннотациялауға арналған әртүрлі аспектілерді қамтитын көптеген зерттеулер жасалған. Мысалы, VerbNet [2] және PropBank [3] себепті етістіктерді және олардың семантикалық рөлдерін жүйелейді. Бұл ресурстар семантикалық талдау міндеттері үшін пайдалы болғанымен, олар тек етістіктермен ғана шектеледі. Сонымен қатар себептілікті басқа грамматикалық тәсілдермен, атап айтқанда, шылаулар, жалғаулықтар, морфологиялық формалар арқылы анықтау жолдарын қарастырмайды. Ал FrameNet корпусы семантикалық фреймдерге сүйене отырып, лексикалық бірліктерді концептуалдық сценарийлермен байланыстырады [4]. Дегенмен, бұл ресурса себептілік CAUSATION, THWARTING, PURPOSE, EXPLANATION секілді көптеген өзге де фреймдер арқылы берілген, бірақ біртұтас себептілік моделі ұсынылмаған.

ASFALDA жобасы француз тілі үшін себептіліктің барынша жүйелі иерархиясын ұсынды [5]. Бұл зерттеу бір тіл аясында да себептіліктің толық моделін құру типологиялық бейімдеу мен лексико-грамматикалық біріздендіруді талап ететінін көрсетті.

Penn Discourse Treebank (PDTB) дискурстық байланыстарды, соның ішінде себептілікті, көпреттік таңбалау (мысалы, бір мезетте себепті және қарсы қоюшы) мүмкіндігімен аннотациялайды [6]. Алайда, бұл ресурс тек сөйлемдер деңгейіндегі дискурстық байланыстарға баса назар аударады да, етістіктік немесе морфологиялық конструкцияларды қамтымайды.

Біздің зерттеуіміз үшін ең өзекті болып табылатын аннотациялық схема BECauSE 2.0 болып табылады [7]. Бұл схема себептілікті мағыналық өзек ретінде қарастырады және оны білдіретін барлық құрылымдарды морфологиялық, синтаксистік және прагматикалық маркерлерді қамтиды. Аталған BECauSE-тің маңызды жаңалығы – қиылысып келетін семантикалық қатынастарды аннотациялау идеясы, мұнда себептілік уақыт, шарттылық, корреляция, санкция, эмоциялық реакция және басқа да мәндермен қатар жүруі мүмкін. Сонымен қатар, онда себептіліктің ішкі типтері жіктелінген. Бұл ішкі типтердің атаулары: consequence, motivation, purpose. Олардың айқындық дәрежесі әртүрлі болуы мүмкін. Алайда, BE-CauSE корпусы ағылшын тіліне арналғандықтан, ол агглютинативті тілдердің типологиялық ерекшеліктерін ескермейді. Сондықтан қазақ тілі үшін себеп-салдарлық конструкцияларды

аннотациялау жергілікті морфологиялық және синтаксистік ерекшеліктерді бейімдеуді талап етеді.

Халықаралық аннотациялық сұлбаларды қазақ тіліне тікелей қолданудың шектеулері

PDTB, FrameNet және BECauSE сияқты халықаралық аннотациялық сұлбалар себептілікті сипаттау үшін маңызды теориялық және әдістемелік негіз береді. Алайда оларды қазақ тіліне тікелей қолдану бірқатар типологиялық қиындықтарға алып келеді. Бұл қиындықтардың басты себебі – қазақ тілінің агглютинативті құрылымы және себептік мағынаның көбіне сөздің ішкі морфологиялық құрамында берілуі.

Ағылшын тілі сияқты аналитикалық тілдерде себептілік көбіне because, since, therefore, as а result сияқты дербес лексикалық немесе дискурстық маркерлер арқылы беріледі. Мұндай жағдайда себептік көрсеткішті жеке токен ретінде белгілеу салыстырмалы түрде жеңіл. Ал қазақ тілінде себептік мағына көбіне -ғандықтан, -етіндіктен, -ғаны үшін, -ған соң, -май/-мей сияқты күрделі морфологиялық және морфосинтаксистік тұлғалар арқылы беріледі. Бұл тұлғаларда шақтық, қимылдық, себептік, болымсыздық және септік мағыналар бір форма ішінде қабаттасып келеді.

Мысалы, «Кешіккендіктен, ол жиналысқа қатыса алмады» сөйлеміндегі «кешіккендіктен» тұлғасы жеке жалғаулық емес, түбір, есімше тұлғасы, номиналдану және септік мағынасының біріккен формасы болып табылады. Сондықтан мұндай бірлікті тек дискурстық маркер ретінде белгілеу жеткіліксіз. Оны морфологиялық құрылым, каузалдық маркер және себеп фрагментінің құрамдас бөлігі ретінде қатар қарастыру қажет.

Сонымен қатар қазақ тілінде кейбір формалар бір мезгілде бірнеше мағына бере алады. Мысалы, -ған соң және -ғаннан кейін тұлғалары бір контексте уақыттық ретті, екінші контексте себептік тәуелділікті, үшінші жағдайда себеп-мезгілдік қабаттасуды білдіруі мүмкін. Осыған байланысты KazCausal схемасы халықаралық аннотациялық тәжірибені қазақ тілінің морфологиялық және семантикалық ерекшеліктеріне бейімдейді. Атап айтқанда, ол себеп фрагментін, салдар фрагментін, маркерді, құрылымдық типті және семантикалық ішкі типті бір мезгілде белгілеуге мүмкіндік береді.

Соңғы жылдары табиғи тілдерді өңдеу, яғни NLP саласында себеп-салдарлық қатынастарды аннотациялау бағытында елеулі ілгерілеу байқалады. Арнайы корпустар мен аннотациялық сұлбаларды жасау себептік байланыстарды неғұрлым нақты әрі жүйелі сипаттауға мүмкіндік береді. CaTeRS [8], Richer Event Description [9] және TimeML Causal [10] сияқты жобалар осы салаға айтарлықтай үлес қосты. Алайда бұл жобалардың көпшілігі оқиға құрылымымен ғана шектеледі немесе себептілікті нақты морфосинтаксистік маркерлермен байланыстырмай, интуитивті түсініктерге сүйене отырып қарастырады.

2024 жылы ұсынылған EventFull құралы осы саладағы зерттеулер арасындағы маңызды жетістіктердің бірі болып саналады [11]. Бұл құрал мәтіндегі оқиғалар арасындағы уақыттық, себептік және кореференттік қатынастарды кешенді әрі келісімді түрде аннотациялауға мүмкіндік береді. Аннотациялау үдерісін жеңілдетіп, аннотаторлар арасындағы сәйкестікті арттыру арқылы NLP саласындағы болашақ зерттеулер мен қолданбалар үшін құнды ресурсқа айналды. Сондай-ақ, себептілікті анықтаудың заманауи әдістеріне арналған кең көлемді шолу жарық көріп, бар тәсілдерді жүйелеп, негізгі ғылыми және технологиялық қиындықтарды анықтап берді.

Халықаралық сұлбалар қазақ тіліндегі себептілікті сипаттау үшін маңызды теориялық база ұсынғанымен, оларды қазақ тіліне механикалық түрде көшіру жеткіліксіз. Қазақ тілінде себептік мағына түбір, есімше/көсемше тұлғасы, болымсыздық көрсеткіші, номиналдану, септік және шылау элементтері арқылы бір морфологиялық кешенде берілуі мүмкін. Kaz-Causal сұлбасы дәл осы ерекшелікті есепке алып, халықаралық тәжірибені қазақ тілінің типологиялық табиғатына бейімдейді.

Кесте 1-де көрсетілгендей, KazCausal сұлбасының басты ерекшелігі – себептілікті қазақ тіліндегі морфологиялық және морфосинтаксистік құрылымдармен бірге қарастыруы. Бұл

Кесте 1 – Халықаралық аннотациялық сұлбалар мен KazCausal сұлбасының салыстырмалы сипаттамасы

Критерий	PDTV	FrameNet	BECauSE 2.0	KazCausal
Негізгі объект	Дискурстық қатынас	Семантикалық фрейм	Себептік конструкция	Қазақ тіліндегі себеп-салдарлық құрылым
Аннотация деңгейі	Сөйлемаралық байланыс	Лексикалық бірлік және фрейм	Конструкция және қатынас	Морфология, маркер, себеп, салдар, семантикалық тип
Морфологиялық маркерлер	Қамтымайды	Ішінара	Шектеулі	Арнайы қамтиды
Агглютинативті тілге бейімделуі	Жоқ	Ішінара	Жоқ	Иә
Қабаттасқан қатынастар	Ішінара	Шектеулі	Иә	Иә, арнайы белгіленеді
Қазақ тіліне бейімделуі	Жоқ	Жоқ	Жоқ	Иә

тәсіл агглютинативті тілдердегі себептік маркерлерді жеке лексема ретінде ғана емес, толық формант және себеп фрагментінің құрамдас бөлігі ретінде белгілеуге мүмкіндік береді.

Аталған шолу сөйлем және мәтін деңгейінде қолданылатын әдістер таксономиясын ұсына отырып, модельдердің дәлдігі мен әмбебаптығын жетілдіру қажеттілігін анықтады.

Ал соңғы кездегі зерттеулерде себеп-салдарлық байланыстарды анықтау үшін терең нейрондық желілердің архитектураларын қолдану кеңінен таралуда. Атап айтқанда, 2025 жылғы эмпирикалық зерттеу BioBERT-BiGRU сияқты модельдердің себепті байланыстарды автоматты түрде айқындауда жоғары тиімділікке ие екенін көрсетті [12]. Бұл нәтижелер модель архитектурасы мен трансферлік оқыту стратегияларын дұрыс таңдаудың маңызды екенін дәлелдейді.

Кең таралған, кең қолданысқа ие ағылшын, француз, араб, итальян т.б. және басқа тілдер үшін аннотталған конструкцияларды анықтауға бағытталған зерттеулер де, корпусар да, сол салаға арналған ғылыми ізденістер де кездеседі. Алайда құрылымдық жүйесі жағынан агглютинативті тілдер қатарына кіретін қазақ тілінің құрылымдық конструкцияларына қатысты зерттеу тәжірибелері аз ғана деуге болады. Қазақ тілінің ұлттық корпусы Qazcorpus.kz [13] және Qazcorpora.kz [14] сияқты жекелеген бастамалар бар болғанымен, себеп-салдарлық конструкцияларға арналған арнайы корпус әлі жасалған жоқ. Ғылыми әдебиеттерде себептілікті білдіретін морфологиялық маркерлер – -ғандықтан, -май, -үшін жұрнақтары, -нан кейін, -соң сияқты постпозициялық құрылымдар, өйткені, себебі, неге десен сияқты жалғаулықтар сипатталғанымен, бұл сипаттамалар фрагментарлы сипатта және формалды аннотациялық сұлбамен толықтырылмаған. Сондықтан біздің зерттеу жұмысымыз осындай себеп-салдарлық конструкцияларды анықтауға бағытталған корпус құрудағы алғашқы талпыныс болып табылады және зерттеудің өзектілігін танытады.

Қазақ тіліндегі себеп-салдарлық конструкцияларды аннотациялау бағыты бойынша алғашқы арнайы зерттеулердің бірі ретінде KazCausal корпусына арналған жұмыс ерекше маңызға ие. Авторлар қазақ тіліндегі каузалдық конструкцияларды қолмен аннотациялау мәселесін қарастырып, KazCausal сұлбасының бастапқы қағидаларын ұсынды. Бұл зерттеуде қазақ тіліндегі себептік маркерлер, себеп және салдар фрагменттері, сондай-ақ синтетикалық, аналитика-синтетикалық және аналитикалық құрылымдар жүйеленді. Аталған жұмыс осы мақалада ұсынылып отырған KazCausal сұлбасын әрі қарай кеңейтуге, құрылымдық типтер бойынша сандық талдау жасауға және аннотация сапасын F1 мен Cohen's k көрсеткіштері арқылы бағалауға негіз болды.

3. Материалдар мен әдістер

3.1. Корпус құрылымы және дереккөздер

Зерттеу барысында корпус құрылымына қазақ тіліндегі ғылыми, публицистикалық және нормативтік сипаттағы мәтіндерден алынған 1250 каузалды сөйлем енгізілді. Бұл сөйлемдердің барлығы себеп-салдарлық қатынасқа ие тілдік бірліктер ретінде іріктелді. Сондықтан зерттеуде каузалды және каузалды емес сөйлемдердің арақатынасы емес, каузалды сөйлемдердің ішкі құрылымдық типтері, морфологиялық көрсеткіштері және семантикалық қабаттары талданды.

Корпус бес домен бойынша жүйеленді: филология, құқық, дінтану, саясат және лингвистика. Мұндай көпдоменді іріктеу себеп-салдарлық қатынастардың әртүрлі ғылыми және қоғамдық дискурстарда қалай көрінетінін салыстыруға мүмкіндік береді

Кесте 2 – Корпустағы каузалды сөйлемдердің домендер және құрылымдық типтер бойынша таралуы

Домен	SYNTHETIC CAUSE	ANALYTICO SYNTHETIC CAUSE	ANALYTIC CAUSE	Барлығы	Үлесі
Филология	123	115	103	341	27,3%
Құқық	91	97	82	270	21,6%
Дінтану	56	60	47	163	13,0%
Саясат	58	62	55	175	14,0%
Лингвистика	103	91	107	301	24,1%
Барлығы	431	425	394	1250	100%

Кесте 2-де көрсетілгендей, корпус көлемі 1250 каузалды сөйлемнен тұрады. Ең көп мысал филология және лингвистика домендерінен алынған, бұл аталған салаларда түсіндіру, негіздеу және себеп-салдарлық байланыстардың жиі қолданылуымен байланысты. Құқық доменінде санкциялық және нормативтік себептілік, саясат мәтіндерінде логикалық негіздеу, ал дінтану мәтіндерінде мотивациялық және түсіндірмелік қатынастар жиі кездеседі.

3.2. Құрылымдық типтер бойынша жіктеу

KazCausal сұлбасында барлық каузалды сөйлемдер үш негізгі құрылымдық типке бөлінді: SYNTHETIC_CAUSE, ANALYTICO_SYNTHETIC_CAUSE және ANALYTIC_CAUSE. Бұл жіктеу қазақ тіліндегі себептілік мағынасының берілу жолдарын формалды түрде сипаттауға мүмкіндік береді.

Кесте 3-те 1250 каузалды сөйлемнің құрылымдық типтер бойынша жалпы таралуы берілді. Нәтижелер синтетикалық және аналитика-синтетикалық құрылымдардың үлесі бір-біріне жақын екенін көрсетті: сәйкесінше 34,5% және 34,0%. Аналитикалық құрылымдар 31,5%-ды құрайды. Бұл қазақ тілінде себеп-салдарлық қатынастардың тек дербес жалғаулықтармен емес, морфологиялық және морфосинтаксистік тәсілдер арқылы да белсенді берілетінін дәлелдейді.

Кесте 3 - Корпустағы каузалды сөйлемдердің құрылымдық типтер бойынша жалпы үлесі

Құрылымдық тип	Саны	Үлесі
SYNTHETIC_CAUSE	431	34,5%
ANALYTICO_SYNTHETIC_CAUSE	425	34,0%
ANALYTIC_CAUSE	394	31,5%
Барлығы	1250	100%

Кесте 4-те қазақ тіліндегі каузалды құрылымдардың негізгі типтері мен маркерлері көрсетілген. Синтетикалық құрылымдарда себептік мағына морфологиялық қосымшалар

Л.Н. Гумилев атындағы ЕҰУ хабаршысы. Математика, компьютерлік ғылымдар, механика сериясы, 2026, Том 154, №1
Вестник ЕНУ им. Л.Н. Гумилева. Серия Математика, компьютерные науки, механика, 2026, Том 154, №1

арқылы берілсе, аналитика-синтетикалық құрылымдарда морфологиялық тұлға мен септеулік/шылау тіркеседі. Аналитикалық құрылымдарда себеп-салдарлық қатынас дербес жалғаулықтар мен лексикалық маркерлер арқылы беріледі. Бұл типология KazCausal аннотациялық сұлбасында себеп, салдар және маркер компоненттерін жүйелі түрде белгілеуге мүмкіндік береді.

Кесте 4 – Каузалды құрылымдардың негізгі типтері мен маркерлері

Құрылымдық тип	Сипаттамасы	Негізгі маркерлер	Мысал
SYNTHETIC CAUSE	Себептік мағына сөздің морфологиялық құрамында беріледі	-ғандықтан, -гендіктен, -атындықтан, -май/-мей	Жаңбыр жауғандықтан, жол тайғақ болды.
ANALYTICO SYNTHETIC CAUSE	Морфологиялық тұлға мен шылау/септеулік тіркеседі	-ған соң, -ғаннан кейін, -ғаны үшін	Ережені бұзғаны үшін, студентке ескерту берілді.
ANALYTIC CAUSE	Себептік қатынас дербес жалғаулықтар мен лексикалық маркерлер арқылы беріледі	себебі, өйткені, сондықтан	Ол сабаққа келмеді, себебі ауырып қалды.

3.3. KazCausal аннотациялық сұлбасы және формалды моделі

KazCausal аннотациялық сұлбасы қазақ тіліндегі себеп-салдарлық құрылымдарды көпдеңгейлі бірлік ретінде сипаттайды. Бұл зерттеуде себептілік объективті дүниедегі себеп-салдар заңы ретінде емес, мәтінде автор тарапынан қалай берілгені тұрғысынан қарастырылады. Сондықтан аннотацияның негізгі міндеті – сөйлемдегі қай фрагмент себеп, қай фрагмент салдар екенін және олардың арасындағы байланысты қандай тілдік маркер орнататынын белгілеу.

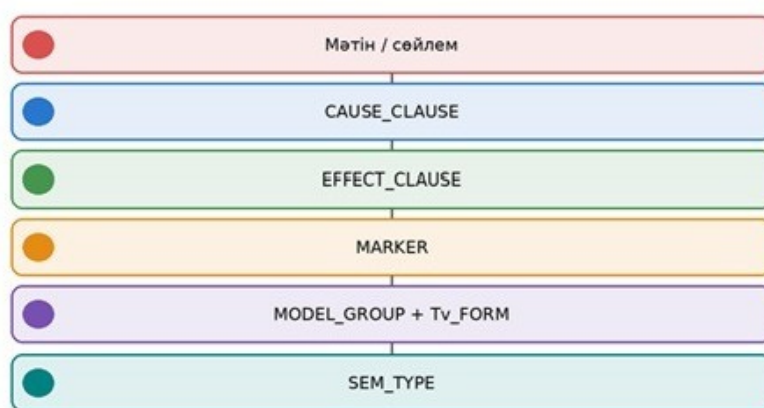
KazCausal сұлбасында әрбір каузалды құрылым төмендегі формалды модель арқылы сипатталады:

$$K = \langle C, E, M, ST, TF, G, SemT \rangle,$$

мұндағы C - себеп фрагменті, E - салдар фрагменті, M - каузалдық маркер, ST - құрылымдық тип, TF - нақты морфологиялық немесе синтаксистік модель, G - басқарылушылық сипаты, SemT - семантикалық ішкі тип. Мұндай формализация аннотацияланған деректерді кейінгі NLP модельдерін оқытуға жарамды құрылымданған ресурсқа айналдырады.

Сурет 4-те KazCausal аннотациялық моделінің көпдеңгейлі құрылымы көрсетілген. Модельде мәтін немесе сөйлем бастапқы бірлік ретінде алынып, себеп фрагменті, салдар фрагменті, каузалдық маркер, құрылымдық тип, нақты морфологиялық форма және семантикалық тип қабаттары арқылы белгіленеді. Мұндай көпқабатты құрылым қазақ тіліндегі себеп-салдарлық қатынастарды жүйелі түрде сипаттауға және оларды автоматты аннотациялау үдерісіне бейімдеуге мүмкіндік береді.

Кесте 5-те KazCausal аннотациялық сұлбасында қолданылатын негізгі қабаттар, олардың белгілері, сипаттамалары және мысалдары берілген. Бұл қабаттар себеп пен салдар фрагменттерін, каузалдық маркерді, құрылымдық типті, нақты форманы және семантикалық қатынас түрін жеке-жеке белгілеуге мүмкіндік береді. Осылайша, аннотация сұлбасы қазақ тіліндегі каузалды құрылымдарды формалды және бірізді түрде сипаттауға негіз болады.



Сурет 1 – KazCausal көпдеңгейлі аннотациялық моделі

Кесте 5 - KazCausal аннотациялық қабаттарының сипаттамасы

Қабат	Белгі	Сипаттамасы	Мысал
Себеп фрагменті	CAUSE_CLAUSE	Салдарға негіз болатын оқиға, күй немесе әрекет	Ғылыми мақала кіріспеден басталатындықтан
Салдар фрагменті	EFFECT_CLAUSE	Себептен туындайтын нәтиже	зерттеу мақсаты мен міндеттері айқындалады
Маркер	MARKER	Себеп-салдарлық байланысты білдіретін көрсеткіш	басталатындықтан
Құрылымдық тип	MODEL_GROUP	Синтетикалық, аналитика-синтетикалық немесе аналитикалық тип	SYNTHETIC
Нақты форма	Tv_FORM	Морфологиялық немесе синтаксистік үлгі	Tv=атындықтан
Семантикалық тип	SEM_TYPE	Қатынас мағынасы	DEPENDENCE

3.4. Қолмен аннотациялау үдерісі және Label Studio құралын қолдану

Зерттеу барысында қазақ тіліндегі себеп–салдарлық құрылымдарды жүйелі түрде белгілеу үшін Label Studio платформасы қолданылды. Бұл құрал мәтін ішіндегі себеп фрагментін, салдар фрагментін және каузалдық маркерді жеке span ретінде белгілеуге, сондай-ақ әрбір мысалға құрылымдық және семантикалық сипаттама беруге мүмкіндік береді. Аннотациялау бірлігі ретінде жеке сөз ғана емес, сөйлем ішіндегі мағыналық фрагмент алынды, себебі қазақ тілінде себептік мағына көбіне бір лексема арқылы емес, морфологиялық және морфосинтаксистік кешен арқылы беріледі.

Label Studio ортасында KazCausal аннотациялық сұлбасы бірнеше деңгей бойынша іске асырылды. Бірінші деңгейде сөйлем ішіндегі себеп бөлігі CAUSE_CLAUSE, салдар бөлігі EFFECT_CLAUSE, ал себеп–салдарлық байланысты білдіретін көрсеткіш MARKER ретінде белгіленді. Екінші деңгейде конструкцияның құрылымдық SYNTHETIC, ANALYTICO_SYNTHETIC немесе ANALYTIC типтері анықталды. Үшінші деңгейде нақты морфологиялық немесе синтаксистік модель, яғни Tv_FORM таңдалды. Төртінші деңгейде басқарылушылық сипаты MANAGED немесе UNMANAGED түрінде белгіленді. Бесінші деңгейде семантикалық ішкі тип анықталды: DEPENDENCE, EMOTIVE_REACTION, SANCTION, JUSTIFICATION және басқа да мағыналық қатынастар.

Аннотация үдерісі бірнеше кезеңнен тұрды. Алдымен корпус мәтіндері алдын ала өңделіп, сөйлемдерге бөлінді. Кейін себептік маркерлері бар немесе себеп–салдарлық мағынасы

контекст арқылы байқалатын сөйлемдер іріктелді. Үшінші кезеңде аннотаторлар әр сөйлемдегі себеп фрагментін, салдар фрагментін және маркерді белгіледі. Төртінші кезеңде әрбір сөйлемге құрылымдық модель және семантикалық тип тағайындалды. Соңғы кезеңде даулы мысалдар қайта қаралып, аннотация ережелері негізінде біріздендірілді.

Мысалы, «Ғылыми мақала кіріспеден басталатындықтан, зерттеу мақсаты мен міндеттері айқындалады» деген сөйлемде «Ғылыми мақала кіріспеден басталатындықтан» фрагменті себеп бөлігі ретінде, «зерттеу мақсаты мен міндеттері айқындалады» фрагменті салдар бөлігі ретінде, ал «басталатындықтан» тұлғасы каузалдық маркер ретінде белгіленеді. Құрылымдық типі – SYNTHETIC, нақты тұлғасы – Tv=атындықтан, басқарылушылық сипаты – UNMAN-AGED, семантикалық ішкі типі – DEPENDENCE ретінде анықталады.

Осылайша Label Studio негізіндегі аннотациялау тәсілі қазақ тіліндегі себеп-салдарлық құрылымдарды тек жалпы себептік қатынас ретінде емес, көпдеңгейлі құрылымдық-семантикалық бірлік ретінде сипаттауға мүмкіндік берді. KazCausal сұлбасының Label Studio платформасында іске асырылу үлгісі сурет 2-де көрсетілген.

Сурет 2 – Label Studio ортасында KazCausal сұлбасы бойынша себеп-салдарлық құрылымды көпдеңгейлі аннотациялау үлгісі

Сурет 2-де көрсетілгендей, бір сөйлем бірнеше деңгей бойынша белгіленеді: CAUSE_CLAUSE, EFFECT_CLAUSE, MARKER, MODEL_GROUP, Tv_FORM, GOVERNANCE және SEM_TYPE. Бұл KazCausal схемасының қазақ тіліндегі себептік құрылымдарды тек лексикалық деңгейде емес, морфологиялық, синтаксистік және семантикалық деңгейлерде қатар сипаттайтынын көрсетеді.

3.5. Аннотация сапасын бағалау

Аннотация сапасын бағалау үшін екі түрлі өлшем қолданылды. Біріншісі - каузалдық маркерлер мен себеп-салдар фрагменттерін анықтау сапасын бағалайтын F1 көрсеткіші, екіншісі - аннотаторлар арасындағы келісім деңгейін анықтайтын Cohen's Kappa коэффициенті.

$$\text{Precision} = \frac{TP}{TP + FP},$$

$$\text{Recall} = \frac{TP}{TP + FN},$$

$$F_1 = \frac{2 \times (\text{Precision} \times \text{Recall})}{\text{Precision} + \text{Recall}},$$

$$\kappa = \frac{P_o - P_e}{1 - P_e},$$

мұндағы TP - дұрыс белгіленген каузалдық бірліктер саны, FP - қате белгіленген бірліктер саны, FN - белгіленбей қалған каузалдық бірліктер саны. P_o - аннотаторлар арасындағы нақты сәйкестік үлесі, ал P_e - кездейсоқ сәйкестік ықтималдығы.

Зерттеу нәтижесінде себептік маркерлерді белгілеу бойынша $F1=0.81$, ал себептілік қатынастарының типтерін анықтау бойынша $k=0.72$ мәні алынды. Бұл көрсеткіштер KazCausal сұлбасының аннотациялау кезінде жеткілікті деңгейде бірізді қолданылғанын көрсетеді.



Сурет 3 – KazCausal корпусын құру және аннотациялау үдерісі

Сурет 3-те корпус құрудың негізгі кезеңдері берілген: мәтіндерді жинау, каузалды сөйлемдерді іріктеу, Label Studio арқылы қолмен аннотациялау, құрылымдық типтерді жіктеу және сапаны бағалау. Бұл жұмыс процесі корпус деректерін кейінгі NLP модельдеріне бейімдеуге мүмкіндік береді.

Аталған зерттеу жұмысында қазақ тіліндегі себеп-салдарлық құрылымдардың үш негізгі түрі айқындалды. Олар:

- синтетикалық,
- аналитикалық-синтетикалық,
- аналитикалық.

Мұндағы синтетикалық құрылымдардың мағынасын себеп мағынасын білдіретін аффикстер арқылы беріледі. Мәселен, -ғандықтан, -май, -ғаны үшін және т.б. Аналитикалық-синтетикалық құрылымдар мағынасы – синтетикалық және аналитикалық тәсілдердің үйлесуі арқылы беріледі. Мұнда -ған соң, -ғаннан кейін және т.б. аффикстер қолданылады. Аналитикалық құрылымдар мағынасы – жалғаулықтар арқылы жасалған құрылымдар болып табылады, олар өйткені, себебі, неге десең және т.б. түрінде кездеседі.

Сонымен қатар, семантикалық қатынастардың келесідегідей сыныптары белгіленеді:

- hasCause (себеп бар),
- hasConsequence (салдар бар),
- hasSanction (санкциялық себеп),
- isEmotionallyTriggeredBy (эмоциялық реакцияға итермелейді),
- isAmbiguousCause (көп мағыналы себеп),
- isNegativeCause (теріс себеп),
- isPositiveCause (оң себеп),
- performedBy (орындаушысы),
- experiencedBy (тәжірибелеуші субъект) және т.б.

Мысалы: Бұл жолдастық кездесу болғандықтан, қосалқы құрамдағы ойыншыларды тексеріп алуға жақсы мүмкіндік болды [15]. Мұндағы,

себеп: Бұл жолдастық кездесу болғандықтан.

салдар: қосалқы құрамдағы ойыншыларды тексеріп алуға жақсы мүмкіндік болды.

маркер: болғандықтан.

Осылайша, әрбір мысал құрылымдық модель мен семантикалық белгілер тұрғысынан таңбаланады. Бұл тәсіл алынған корпусқа әрі лингвистикалық талдау, әрі NLP модельдерін оқыту үшін пайдалануға төмендегідей мүмкіндіктерді береді:

1. қазақ тіліндегі себеп-салдарлық конструкциялар (ССК) үшін аннотациялық схеманы енгізеді;

2. 10-нан астам құрылым түрін жүйелейді;

3. грамматикалық форма мен семантикалық функция арасындағы байланысты жүйелейді;

4. қазақ тіліндегі себептілік онтологиясын құруға арналған жалпы базаны қалыптастырады.

Бұл зерттеу қазақ тіліндегі себеп-салдарлық байланыстарды анықтауға негізделген сапалы лингвистикалық ресурсты жасау жолындағы алғашқы қадам болып табылады. Аталған

зерттеу теориялық және қолданбалы мақсаттарға жетуге бағытталған және болашақта формалды семантикалық талдау, машиналық оқыту, автоматты аннотация, онтология құру және интеллектуалды мәтін өңдеу сияқты зерттеу салаларында кеңінен қолданыла алады.

Жоғарыда атап өткендей, ағылшын, француз сияқты кең қолданыстағы тілдер үшін тілдегі белгілі бір конструкцияларды анықтауға бағытталған аннотацияланған бай ресурстар жасалғанын байқаймыз, алайда қазіргі қазақ тілі осы бағытта айтарлықтай артта қалып отыр. Сондай терең тиянақты зерттеуді қажет ететін ғылыми зерттеу нысандарының бірі – қазіргі қазақ тіліндегі себеп-салдарлық байланыстарды анықтау және аталған конструкцияларды аннотациялау. Қазіргі қазақ тіліндегі бұл конструкция ғылыми әдебиеттерде, жаңалықтар мәтіндерінде және публицистикалық сипаттағы ресурстарда кеңінен қолданылады. Осы себеп-салдарлықты білдіретін морфологиялық маркерлер - ғандықтан, -май, -үшін жұрнақтары, -нан кейін, -соң сияқты постпозициялық құрылымдар, өйткені, себебі, неге десең сияқты жалғаулықтармен сипатталғанымен, бұл сипаттамалар фрагментарлы сипатта ғана және формалды аннотациялық сұлбамен толықтырылмағаны анықталды.

Осы олқылықтың орнын толтыру мақсатында біздің бұл зерттеу ізденісіміз қазіргі қазақ тіліндегі себеп-салдарлық конструкцияларды жүйелі түрде аннотациялаудың алғашқы талпынысын ұсынады. Ізденушілердің арнайы әзірлеген KazCausal аннотациялық сұлбасы мынадай құрамдас бөліктерден тұрады:

- формалық классификация (синтетикалық, аналитика-синтетикалық, аналитикалық конструкциялар);
- семантикалық қатынастар (hasCause, hasConsequence, hasSanction, isEmotionallyTriggeredBy, isAmbiguousCause және т.б.);
- себеп түрлерінің типологиясы: жалпы себеп, санкциялайтын себеп, эмоциялық себеп, теріс себеп, баламалы себеп.

Осылайша, KazCausal алғаш рет позитивті себеп (арқасында), баламалы мотивация (-ғандықтан ба) және қарсы қойылған себеп (қарамастан) сияқты категорияларды енгізеді. Бұл категориялар қазақ тілінің прагматикалық және риторикалық ерекшеліктерін ескеруге мүмкіндік береді.

Жалпы, KazCausal аннотациялық сұлбасы халықаралық тәжірибені морфологиялық тұрғыдан бай және агглютинативті тіл жағдайына бейімдеп қана қоймай, оны жаңа ұғымдармен толықтырады. Бұл ресурс лингвистикалық зерттеулер үшін қолданбалы міндеттерді шешуді алға қояды. Бұл міндеттерге автоматты мәтіндік талдау, себептік байланыстарды анықтау, онтологиялар құру және NLP модельдерін оқыту үшін де маңызды құрал тағайындау мәселелері жатады.

Зерттеу барысында қарастырылған түрлі тілдердегі себеп-салдарлық конструкциялардың аннотациялық сұлбаларының арасында біршама айырмашылықтар мен ерекшеліктер кездеседі. Біз зерттеу нысанына алынған аталған аннотациялық сұлбаларды өзара салыстыра отырып, ерекшеліктерін анықтадық (Кесте 6 қараңыз)

Кесте – 6. Әртүрлі тілдердегі себеп-салдарлық конструкциялардың аннотациялық сұлбаларының салыстырмалы талдауы

№	Сипаттамасы	PDTB	FrameNet	BECauSE 2.0	KazCausal
1	Негізгі тәсілі	Дискурстық байланыстардың элементтері	Семантикалық фреймдер	Конструкциялар мен мағыналар	Конструкциялар мен мағыналар
2	Негізгі бағыты	Анық және анық емес дискурстық қатынастар	Лексико-семантикалық сценарийлер	Себептік конструкциялар	Қазақ тіліндегі себеп-салдарлық конструкциялар
3	Аннотация түрі	Сөйлемдер арасындағы байланысты белгілеу	Лексикалық бірліктерді аннотациялау	Мәтіндегі конструкцияларды белгілеу	Синтетикалық және аналитикалық конструкцияларды белгілеу

4	Көп мағыналықты қолдану	Иә	Жоқ	Иә	Иә
5	Қолданылатын конструкциялар	Дискурстық маркерлер	Етістіктер, зат есімдер және т.б.	Барлық тілдік құралдар	Морфология, постпозициялар, жалғаулықтар, синтаксис
6	Қосымша семантикалық қатынастар	Шарт, қарсы қою, нәтиже	Explanation, Purpose және т.б.	Уақыт, корреляция, салыстыру	Санкция, эмоциялық себеп, альтернативалық модальдылық
7	Тілдері	Ағылшын	Көптілді (EN, FR, DE)	Ағылшын	Қазақ
8	Формализация деңгейі	Орташа, байланыстырғыш элементтер	Жоғары, фреймдер мен рөлдер	Жоғары, қатынастар	Жоғары, конструкция түрлері мен семантикасы бөлек белгіленеді
9	Агглютинативті тілдерді қолдану	Жоқ	Ішінара, локализация арқылы	Жоқ	Иә
10	Эмоциялық себеп категориясының болуы	Жоқ	Жоқ	Ішінара иә	Иә
11	Онтологиялық сәйкестік	Ішінара	Иә, фреймдік онтологиямен байланыс	Жоқ	OWL/RDF арқылы жоспарлануда
12	Корпустың болуы	Иә, PDTB 3.0	Иә	Иә	Иә, қолмен аннотацияланған синтетикалық және аналитикалық формалар

Жоғарыдағы кесте 6-ға сәйкес, қысқартылған атаулардың төмендегідей көрсеткіштері бар. Олар:

- PDTB – дискурстық маркерлерге негізделген классикалық аннотациялық сұлба.
- FrameNet – себептілікті семантикалық фреймдер арқылы сипаттаушы сұлба.
- BECauSE 2.0 – семантикалық және құрылымдық қиылыстарды белгілейтін озық сұлба.
- KazCausal – қазақ тілінің морфологиялық және синтаксистік ерекшеліктерін ескеретін алғашқы аннотациялық сұлба.

PDTB, FrameNet, BECauSE 2.0 сұлбаларынан KazCausal-дың басты ерекшеліктері ретінде қолмен аннотацияланған синтетикалық және аналитикалық формалардың болуын, агглютинативті тілдерді қолдай алатындығын, санкция, эмоциялық себеп, альтернативтілік, модалдылық сияқты қосымша семантикалық қатынасты білдіре отырып, себеп-салдарлық конструкцияларды аннотациялауға мүмкіндік беретіндігін атауға болады.

Демек, KazCausal – себептілікті білдіретін морфологиялық құрылымдар мен синтаксистік конструкциялардың ерекшеліктерін ескере отырып, қазақ тіліне арнайы бейімделген алғашқы аннотациялық сұлба болып табылады. Аталған сұлбаны өзге де түркі тілдес, морфологиялық және синтаксистік ерекшеліктері ұқсас болып келетін тілдерді жобалау үшін де қолдануға болады деп есептейміз.

4. Нәтижелер және талқылау

Зерттеу нәтижелері қазақ тіліндегі себеп-салдарлық қатынастардың тек дербес жалғаулықтар арқылы ғана емес, морфологиялық формалар, аналитика-синтетикалық

құрылымдар және контекстік семантикалық белгілер арқылы да берілетінін көрсетті. Корпустық талдау барысында каузалды құрылымдар үш негізгі құрылымдық тип бойынша жүйеленді: SYNTHETIC_CAUSE, ANALYTICO_SYNTHETIC_CAUSE және ANALYTIC_CAUSE. Бұл жіктеу қазақ тіліндегі себептілік категориясының көпқабатты табиғатын ашуға мүмкіндік берді.

Алынған нәтижелер бойынша синтетикалық және аналитика-синтетикалық құрылымдар қазақ тіліндегі каузалдық қатынастарды білдіруде маңызды орын алады. Бұл қазақ тілінің агглютинативті құрылымымен тікелей байланысты. Себептік мағына көп жағдайда жеке жалғаулық арқылы емес, етістік негізіне жалғанатын қосымша, жұрнақтар немесе есімше мен шылаулардың тіркесуі арқылы беріледі. Сондықтан қазақ тіліндегі себеп-салдарлық қатынастарды автоматты түрде анықтау үшін тек лексикалық маркерлерге сүйену жеткіліксіз. Морфологиялық көрсеткіштерді, синтаксистік позицияны және семантикалық контексті қатар ескеретін көпдеңгейлі модель қажет.

4.1. Морфологиялық және аналитика-синтетикалық құрылымдардың ерекшелігі

Зерттеу нәтижелері қазақ тілінде себептік мағынаның дербес жалғаулықтар арқылы ғана емес, сөздің морфологиялық құрамына кіріккен формалар арқылы да белсенді берілетінін көрсетті.

SYNTHETIC_CAUSE тобына жататын -ғандықтан, -гендіктен, -атындықтан, -май/-мей тұлғалары себептілікті тікелей морфологиялық деңгейде білдіреді. Бұл формаларда қимыл, шақ, болымсыздық және себептік мағына бір тілдік бірлік ішінде қабаттасады.

ANALYTICO_SYNTHETIC_CAUSE тобы қазақ тіліндегі себептіліктің аралық, яғни морфологиялық және синтаксистік тәсілдердің біріккен түрін көрсетеді. -ған соң, -ғаннан кейін, -ғаны үшін сияқты құрылымдар көбіне уақыттық реттілік, санкция немесе мақсаттық реңкпен қатар жүреді. Сондықтан оларды тек уақыттық немесе тек себептік маркер ретінде қарастыру жеткіліксіз.

ANALYTIC_CAUSE тобына себебі, өйткені, сондықтан, неге десең сияқты дербес жалғаулықтар мен лексикалық маркерлер жатады. Бұл құрылымдар халықаралық дискурстық аннотация сұлбаларына жақынырақ болғанымен, қазақ тіліндегі себептіліктің тек бір бөлігін ғана қамтиды. Корпуста аналитикалық құрылымдардың үлесі 31,5% болуы қазақ тіліндегі каузалдық қатынастардың басым бөлігі морфологиялық немесе аралас тәсілдер арқылы берілетінін көрсетеді.

Осы нәтижелер қазақ тіліндегі себептілік категориясының типологиялық тұрғыдан күрделі екенін дәлелдейді. Ағылшын немесе басқа аналитикалық тілдерде себеп-салдарлық байланыс көбіне дербес жалғаулықтар мен дискурстық маркерлер арқылы берілсе, қазақ тілінде бұл мағына сөздің морфологиялық құрамында, есімше тұлғаларында және шылаулы тіркестерде көрініс табады. Сондықтан *KazCausal* сұлбасы себеп фрагментін, салдар фрагментін және маркерді ғана емес, нақты морфологиялық форманы да жеке қабат ретінде белгілеуді ұсынады.

4.2. Қабаттасқан семантикалық қатынастар

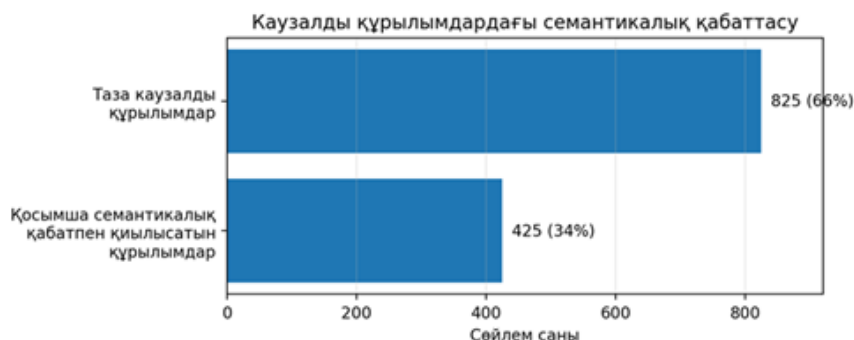
Қазақ тіліндегі кейбір себептік құрылымдар бір мезгілде бірнеше мағына бере алады. Мысалы, -ған соң және -ғаннан кейін тұлғалары кейде уақыттық реттілікті, кейде себептік тәуелділікті, ал кейбір контексте уақыт-себеп қабаттасуын білдіреді. -ғаны үшін тұлғасы себептік қатынаспен қатар санкциялық реңкті де көрсете алады. Осыған байланысты *Kaz-Causal* сұлбасы бірнеше семантикалық белгіні қатар қоюға мүмкіндік береді.

Кесте 7-де қазақ тіліндегі қабаттасқан семантикалық қатынастардың негізгі түрлері көрсетілген. Кестеде себеп-салдарлық мағынамен қатар жүретін DEPENDENCE, SANCTION, EMOTIVE_REACTION, JUSTIFICATION және TEMPORAL_CAUSE сияқты семантикалық типтер беріліп, олардың сипаттамасы, маркерлік формалары және нақты мысалдары ұсынылған. Бұл жіктеу қазақ тіліндегі себептілік қатынастарының тек бір мағынамен шектелмей, контекстке байланысты санкциялық, эмоциялық, негіздеу және уақыттық мағыналармен қабаттаса алатынын көрсетеді.

Қабаттасқан семантикалық қатынастарды жеке белгілеу аннотация сапасын арттырады. Себебі бір ғана маркер әртүрлі контексте әртүрлі қызмет атқаруы мүмкін. Мысалы, -ған соң тұлғасы бір сөйлемде уақыттық реттілікті білдірсе, екінші сөйлемде салдардың пайда болуына себеп болған алдыңғы әрекетті көрсетеді. Мұндай жағдайларда маркердің сыртқы формасына ғана сүйену жеткіліксіз, оның сөйлем ішіндегі семантикалық қызметін де анықтау қажет.

Кесте 7 - Қабаттасқан семантикалық қатынастардың негізгі түрлері

Семантикалық тип	Сипаттамасы	Маркер/форма	Мысал
DEPENDENCE	Бір оқиға екінші оқиғаның тікелей себебі болады	-ғандықтан, себебі	Күн суық болғандықтан, балалар сыртқа шықпады.
SANCTION	Әрекет санкциялық немесе нормативтік салдарға әкеледі	-ғаны үшін	Тәртіпті бұзғаны үшін, айыппұл салынды.
EMOTIVE_REACTION	Себеп эмоциялық реакцияны тудырады	-ғанына, -геніне	Жеңіске жеткеніне қуанды.
JUSTIFICATION	Пікірдің немесе шешімнің негіздемесі беріледі	себебі, өйткені	Бұл әдіс тиімді, себебі нәтиже нақты көрінеді.
TEMPORAL_CAUSE	Уақыттық және себептік мағына қатар көрінеді	-ған соң, -ғаннан кейін	Жұмыс аяқталған соң, есеп тапсырылды.



Сурет 4 – KazCausal көпдеңгейлі аннотациялық моделі

Сурет 4-те таза каузалды құрылымдар мен қосымша семантикалық қабаттармен қиылысатын құрылымдардың арақатынасын көрсетеді. Қабаттасқан құрылымдар қазақ тіліндегі себептілік категориясының контекстке тәуелді және көпмағыналы сипатын дәлелдейді. Бұл нәтиже KazCausal сұлбасында көптаңбалы семантикалық аннотацияны қолданудың қажеттілігін көрсетеді.

4.3. Себептік құрылымдарды анықтаудың алгоритмдік логикасы

KazCausal сұлбасы тек қолмен аннотациялауға арналған нұсқаулық қана емес, сонымен қатар автоматты анықтау жүйесіне бейімделген формалды модель болып табылады. Себептік құрылымдарды автоматты түрде анықтау логикасы төмендегі кезеңдер арқылы сипатталады:

- 1) мәтінді алдын ала өңдеу және сөйлемдерге бөлу;
- 2) каузалдық маркер кандидаттарын анықтау;
- 3) маркердің синтетикалық, аналитика-синтетикалық немесе аналитикалық типін белгілеу;
- 4) маркер позициясына сүйене отырып себеп фрагментінің шекарасын анықтау;
- 5) салдар фрагментін негізгі предикативтік бөлік ретінде белгілеу;
- 6) құрылымдық модель мен нақты Tv_FORM түрін таңдау;

- 7) контекстке қарай семантикалық ішкі типті анықтау;
- 8) аннотация нәтижесін құрылымданған форматта сақтау.

Бұл алгоритм ережеге негізделген маркерлік іздеу мен контекстік семантикалық талдауды біріктіреді. Мұндай гибриді тәсіл қазақ тіліндегі морфологиялық себептік тұлғаларды автоматты түрде анықтауға қолайлы, себебі себептік мағынаның бір бөлігі ашық лексикалық маркерлер арқылы, ал екінші бөлігі сөз ішіндегі форманттар арқылы беріледі.

Алгоритмдік логиканың негізгі артықшылығы – оның көпдеңгейлі аннотациялық құрылымға сүйенуі. Бірінші деңгейде маркер анықталса, келесі деңгейде оның құрылымдық типі белгіленеді. Одан кейін себеп пен салдар фрагменттерінің шекарасы көрсетіліп, соңында семантикалық ішкі тип таңдалады. Мұндай кезеңдік тәсіл NLP модельдерін оқыту үшін құрылымданған дерек дайындауға мүмкіндік береді. Әсіресе KazBERT, BiLSTM-CRF, SpERT немесе MRC негізіндегі модельдер үшін себеп, салдар, маркер және семантикалық типтің бөлек белгіленуі supervised learning міндеттерін нақтылауға жағдай жасайды.

Осылайша, ұсынылған KazCausal сұлбасы қазақ тіліндегі себеп-салдарлық қатынастарды тек лингвистикалық сипаттау құралы ғана емес, сонымен қатар автоматты ақпарат алу, мәтінді семантикалық талдау және себептік байланыстарды машиналық өңдеу міндеттеріне арналған дерекқорлық негіз ретінде қарастыруға болады.

5. Шектеулер және болашақ зерттеу бағыттары

KazCausal сұлбасы қазақ тіліндегі себеп-салдарлық құрылымдарды жүйелі аннотациялауға мүмкіндік бергенімен, зерттеу барысында бірқатар шектеулер анықталды. Біріншіден, кейбір құрылымдарда себептік мағына айқын маркер арқылы емес, контекстік немесе прагматикалық түрде беріледі. Мұндай жағдайларда аннотаторға «бұл себеп пе, әлде тек тақырыптық/бағалау қатынасы ма?» деген шешім қабылдау қажет болады.

Екіншіден, бастау көзі, бағыт, тақырыптық байланыс, әлеуметтік мәртебе, бағалау позициясы сияқты домендер ағымдағы сұлбаға енгізілмеді. Себебі оларды бірізді аннотациялау үшін арнайы онтологиялық модельдер қажет. Үшіншіден, кейбір мақсаттық құрылымдар екіжақты себептілік сипатқа ие болуы мүмкін: бір әрекет мақсатқа жетудің себебі болса, мақсаттың өзі сол әрекеттің мотивациясы ретінде түсіндіріледі.

Болашақ зерттеулерде корпус көлемін ұлғайту, әлеуметтік желі және ауызша дискурс материалдарын қосу, KazBERT негізіндегі автоматты аннотация модулін құру, сондай-ақ KazCausal сұлбасын басқа түркі тілдеріне бейімдеу жоспарланады.

Қорытынды

Бұл мақалада қазақ тіліндегі себеп-салдарлық құрылымдарды аннотациялауға арналған KazCausal сұлбасы ұсынылды. Зерттеу қазақ тіліндегі себептіліктің морфологиялық, аналитика-синтетикалық және аналитикалық тәсілдер арқылы берілетінін және бұл тәсілдердің барлығы NLP тұрғысынан формалды түрде белгіленуі қажет екенін көрсетті.

1250 каузалды сөйлемнен тұратын корпус үш құрылымдық тип бойынша жүйеленді: SYNTHETIC_CAUSE - 431 сөйлем (34,5%), ANALYTICO_SYNTHETIC_CAUSE - 425 сөйлем (34,0%), ANALYTIC_CAUSE - 394 сөйлем (31,5%). Бұл нәтижелер қазақ тіліндегі себептік қатынастардың едәуір бөлігі морфологиялық және морфосинтаксистік деңгейде берілетінін дәлелдейді.

F1=0.81 және Cohen's k=0.72 көрсеткіштері KazCausal сұлбасының қолмен аннотациялау кезінде жеткілікті дәрежеде бірізді қолданылатынын көрсетті. Ұсынылған сұлба қазақ тіліндегі себеп-салдарлық қатынастарды автоматты түрде анықтауға арналған болашақ модельдер үшін лингвистикалық және есептеу негізін қалыптастырады. Сонымен қатар бұл тәсіл агглютинативті және төмен ресурсты тілдердегі каузалдық қатынастарды зерттеу үшін типологиялық тұрғыдан маңызды үлгі бола алады.

KazCausal аннотациялық сұлбасы мен осы сұлбаға негізделген корпус – қазақ тіліндегі себеп-салдарлық конструкцияларды жүйелі сипаттауға бағытталған алғашқы қадам болып табылады. Алайда, зерттеу барысында бірқатар шешілмеген мәселелер мен басы ашық сұрақтар анықталып, олар болашақта қосымша зерттеуді қажет ететіндігіне көз жеткіздік:

1. Себептілікті анықтаудағы күмәнді шешімнің шектеулері.

KazCausal бірнеше тегті қатар белгілеуді және қабаттасқан қатынастарды ко-аннотациялауды қолдағанымен, аннотациялаушыға бәрі бірдей «бұл себеп пе, жоқ па» деген күмәнді шешім қабылдауға тура келеді. Бұл әсіресе прототиптік емес себептік құрылымдарда қиындық тудырады (арқасында, себеп болды, белгісі ретінде).

Аннотация кезінде аннотациялаушылар жиі жартылай интуитивтік таңдау жасайды, ол әсіресе төмендегі құрылымдарда байқалады:

- риторикалық себептілікті білдіретін құрылымдарда;
- шартты сөйлемдердегі логикалық тәуелділіктерде;
- айқын себептік маркері жоқ эмоциялық реакцияларда.

Мұндай жағдайларда континуалды себептілік шкаласын енгізу – яғни, себептілікті жоқтан бастап айқын формаға дейінгі спектр түрінде көрсету – тиімді шешім болар еді. Алайда, бұл тәсілді формализациялау және автоматтандыру аса күрделі.

2. Себептілікпен ішінара қабаттасатын семантикалық домендерді алып тастау.

KazCausal жобасы себептілікпен байланысты кейбір мағыналық салаларды әдейі қамтымайды, себебі оларды анық және бірізді аннотациялау мүмкін емес. Мұндай домендерге:

- бастау көзі / бағыт (жарықтан шыққан ұшқындар, мақсатқа жету жолында);
- тақырыптық қатынас (саяси шешімге риза болмау);
- бөліктік-компоненттік қатынастар (жоба аясында жүзеге асырылады);
- дәлелділік (эвиденциалдылық) (болжам бойынша, ақпарат негізінде);
- әлеуметтік/құқықтық мәртебе (азамат ретінде, оқытушы ретінде);
- бағалау позициясы (бұл қадам халықты қауіпке тіреді).

Мұндай мағыналарды формализациялау үшін арнайы онтологиялар қажет болар еді, ал бұл ағымдағы зерттеу шеңберінен тыс ғылыми нысан болып табылады. Осы аспектілерді болашақ зерттеулерде қарастыру жоспарлануда.

3. Екіжақты себептілікті білдіретін құрылымдар.

Зерттеу кезінде ұшырасатын қиындықтардың бірі – себептілікті екі бағытта имплицитті білдіретін құрылымдар. Бұл әсіресе қажетті шарт немесе міндеттілік мағынасында айқын көрін табады. Мысалы:

Табысты болу үшін, біз бәріміз бірігуіміз керек.

(Табысты болу үшін, барлығымыз бірігуіміз қажет). Мұнда: бір жағынан, бірігу – табыстың себебі; екінші жағынан, табысты болу ниеті – бірігудің мотивациясы. Мұндай жағдайлар hasPurpose немесе Motivation категорияларына жатқызылды. Дегенмен, толыққанды семантикалық сипаттама үшін анағұрлым күрделі логикалық модель қажет екені мойындалады.

4. Жетістіктер мен болашақ қолдану мүмкіндіктері.

Жоғарыда аталған шектеулерге қарамастан, KazCausal жобасы келесі маңызды ғылыми зерттеу нәтижелерін көрсетті:

- себеп-салдарлық конструкцияларды агглютинативті, морфологиялық күрделі тілде аннотациялау мүмкін екені дәлелденді;
- сызба мағыналық көпқабаттылық пен прагматикалық жүктемені ескеруге мүмкіндік береді;
- алынған деректерді лингвистикалық зерттеулерде де, NLP модельдерін (трансформерлер, sequence labeling) оқытуда да қолдануға болады;
- қабаттасқан қатынастарды аннотациялау (мысалы, hasCause + isTemporal) қазақ тіліндегі мағыналардың өзара әрекеттесуін жақсырақ түсінуге жол ашады және бұл тәсілді басқа түркі тілдеріне де бейімдеуге болады.

Авторлар үлесі. Авторлар үлесі тең.

Әдебиеттер тізімі

- 1 R. Taberkhan and M.A. Sambetbayeva, KazCausal: Manual Annotation of Causal Constructions in the Kazakh Language, International Conference on Computer Science and Engineering, UBMK 2025 (2025), 463–467. <https://doi.org/10.1109/UBMK67458.2025.11206832>
- 2 K.K. Schuler, VerbNet: A broad-coverage, comprehensive verb lexicon, PhD dissertation, University of Pennsylvania, Philadelphia, 2005. AAI3179808. Available at: <https://www.proquest.com/docview/305449413> (Қол жеткізу күні: 20.11.2025).
- 3 M. Palmer, D. Gildea, and P. Kingsbury, The Proposition Bank: An annotated corpus of semantic roles, Computational Linguistics 31(1) (2005), 71–106. <https://doi.org/10.1162/0891201053630264>
- 4 J. Ruppenhofer, M. Ellsworth, M.R.L. Petruck, C.R. Johnson, C.F. Baker, and J. Scheffczyk, FrameNet II: Extended theory and practice, International Computer Science Institute, Berkeley, 2010. <https://d-nb.info/1136959238/34> (Қол жеткізу күні: 20.11.2025).
- 5 L. Vieu, P. Muller, M. Candito, and M. Djemaa, A general framework for the annotation of causality based on FrameNet, Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16) (2016), 3807–3813. <https://aclanthology.org/L16-1603/> (Қол жеткізу күні: 25.11.2025).
- 6 R. Prasad, N. Dinesh, A. Lee, E. Miltsakaki, L. Robaldo, A. Joshi, and B. Webber, The Penn Discourse Tree-Bank 2.0, Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08), Marrakech, Morocco, 2008. <https://aclanthology.org/L08-1093/> (Қол жеткізу күні: 25.11.2025).
- 7 J. Dunietz, L. Levin, and J.G. Carbonell, The BECauSE Corpus 2.0: Annotating causality and overlapping relations, Proceedings of the 11th Linguistic Annotation Workshop (LAW-XI) (2017), 95–104. <https://doi.org/10.18653/v1/W17-0812>
- 8 N. Mostafazadeh, A. Grealish, N. Chambers, J. Allen, and L. Vanderwende, CaTeRS: Causal and temporal relation scheme for semantic annotation of event structures, Proceedings of the Fourth Workshop on Events (2016), 51–61. <https://doi.org/10.18653/v1/W16-1007>
- 9 T. O’Gorman, K. Wright-Bettner, and M. Palmer, Richer Event Description: Integrating event coreference with temporal, causal and bridging annotation, Proceedings of the 2nd Workshop on Computing News Storylines (CNS 2016) (2016), 47–56. <https://doi.org/10.18653/v1/W16-5706>
- 10 P. Mirza, Extracting temporal and causal relations between events, Proceedings of the ACL 2014 Student Research Workshop (2014), 10–17. <https://doi.org/10.3115/v1/P14-3002>
- 11 A. Eirew, E. Nachshoni, A. Slobodkin, and I. Dagan, EventFull: Complete and Consistent Event Relation Annotation, Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (System Demonstrations) (2025), 494–508. <https://doi.org/10.18653/v1/2025.naacl-demo.40>
- 12 M.A.K. Raiaan, M.A.A.S. Khan, A. Muhammad, T.T.M. Nguyen, and S. Hussain, A review on large language models: Architectures, applications, taxonomies, open issues and challenges, IEEE Access 12 (2024), 26839–26874. <https://doi.org/10.1109/ACCESS.2024.3365742>
- 13 Qazcorpus : Қазақ тілінің ұлттық корпусы. <https://qazcorpus.kz/> (Қол жеткізу күні: 23.11.2025).
- 14 Qazcorpora: Қазақ тілінің корпустары. <https://qazcorpora.kz/> (Қол жеткізу күні: 23.11.2025).
- 15 Қазақстан футбол федерациясының ресми сайты, available at <https://kff.kz/kk/news/23788> (Қол жеткізу күні: 22.11.2025).

TYPOLOGICAL DIFFERENCES BETWEEN SYNTHETIC AND ANALYTICAL CAUSAL STRUCTURES IN THE KAZAKH LANGUAGE

R. Taberkhan, M.A. Sambetbayeva

L.N. Gumilyov Eurasian National University, Satpayev str., 2, Astana, 010008, Kazakhstan

Abstract. This article presents a methodology for the manual annotation and structural-semantic systematization of causal constructions in the Kazakh language based on the KazCausal annotation scheme.

The aim of the study is to formally describe the morphological, lexical, and syntactic means of expressing causality in Kazakh through an annotation model and to provide a scientific basis for the development of a corpus adapted to natural language processing tasks.

The study is based on 1,250 causal sentences selected from scientific, journalistic, and normative texts. All sentences were considered as linguistic units expressing cause-and-effect relations and were classified into three main structural types: SYNTHETIC_CAUSE, ANALYTICO_SYNTHETIC_CAUSE, and ANALYTIC_CAUSE.

The results show that synthetic constructions account for 431 sentences (34.5%), analytico-synthetic constructions for 425 sentences (34.0%), and analytic constructions for 394 sentences (31.5%). The proposed annotation scheme makes it possible to simultaneously annotate the cause fragment, the effect fragment, the causal marker, the structural type, the specific morphological form, and the semantic subtype. The quality of annotation was evaluated using an F1 score of 0.81 and Cohen's kappa coefficient of $k=0.72$.

The findings of the study provide a foundation for the development of typologically adapted NLP resources aimed at the automatic detection of cause-and-effect relations in the Kazakh language.

Keywords: text annotation, corpus linguistics, NLP, agglutinative languages, KazCausal, cause-and-effect relations, inter-annotator agreement.

ТИПОЛОГИЧЕСКИЕ РАЗЛИЧИЯ СИНТЕТИЧЕСКИХ И АНАЛИТИЧЕСКИХ ПРИЧИННЫХ СТРУКТУР В КАЗАХСКОМ ЯЗЫКЕ

Р. Таберхан, М.А. Самбетбаева

Евразийский национальный университет им. Л.Н. Гумилева, ул. Сатпаева, 2, Астана, 010008, Казахстан

Аннотация. В статье предлагается методика ручного аннотирования и структурно-семантической систематизации причинно-следственных конструкций казахского языка на основе схемы KazCausal.

Цель исследования заключается в формальном описании морфологических, лексических и синтаксических средств выражения каузальности в казахском языке посредством аннотационной модели, а также в обосновании научных принципов создания корпуса, адаптированного к задачам обработки естественного языка.

В качестве объекта исследования были использованы 1250 каузальных предложений, отобранных из научных, публицистических и нормативных текстов. Все предложения рассматривались как языковые единицы, выражающие причинно-следственные отношения, и были классифицированы по трём основным структурным типам: SYNTHETIC_CAUSE, ANALYTICO_SYNTHETIC_CAUSE и ANALYTIC_CAUSE.

Полученные результаты показали, что синтетические конструкции составляют 431 предложение (34,5%), аналитико-синтетические конструкции — 425 предложений (34,0%), а аналитические конструкции — 394 предложения (31,5%). Предложенная схема позволяет одновременно аннотировать фрагмент причины, фрагмент следствия, каузальный маркер, структурный тип, конкретную морфологическую форму и семантический подтип. Качество аннотации было оценено с использованием показателей $F1=0,81$ и коэффициента согласия Коэна $k=0,72$.

Результаты исследования формируют основу для разработки типологически адаптированных NLP-ресурсов, предназначенных для автоматического выявления причинно-следственных отношений в казахском языке.

Ключевые слова: аннотация текста, корпусная лингвистика, NLP, агглютинативные языки, KazCausal, причинно-следственные отношения, межаннотаторское согласие.

References

- 1 R. Taberkhan and M.A. Sambetbayeva, KazCausal: Manual Annotation of Causal Constructions in the Kazakh Language, International Conference on Computer Science and Engineering, UBMK 2025 (2025), 463–467. <https://doi.org/10.1109/UBMK67458.2025.11206832>
- 2 K.K. Schuler, VerbNet: A broad-coverage, comprehensive verb lexicon, PhD dissertation, University of Pennsylvania, Philadelphia, 2005. AAI3179808. Available at: <https://www.proquest.com/docview/305449413> (accessed on November 20, 2025).
- 3 M. Palmer, D. Gildea, and P. Kingsbury, The Proposition Bank: An annotated corpus of semantic roles, Computational Linguistics 31(1) (2005), 71–106. <https://doi.org/10.1162/0891201053630264>

Л.Н. Гумилев атындағы ЕҰУ хабаршысы. Математика, компьютерлік ғылымдар, механика сериясы, 2026, Том 154, №1
Вестник ЕНУ им. Л.Н. Гумилева. Серия Математика, компьютерные науки, механика, 2026, Том 154, №1

- 4 J. Ruppenhofer, M. Ellsworth, M.R.L. Petruck, C.R. Johnson, C.F. Baker, and J. Scheffczyk, FrameNet II: Extended theory and practice, International Computer Science Institute, Berkeley, 2010. <https://d-nb.info/1136959238/34> (accessed on November 20, 2025).
- 5 L. Vieu, P. Muller, M. Candito, and M. Djemaa, A general framework for the annotation of causality based on FrameNet, Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16) (2016), 3807–3813. <https://aclanthology.org/L16-1603/> (accessed on November 25, 2025).
- 6 R. Prasad, N. Dinesh, A. Lee, E. Miltsakaki, L. Robaldo, A. Joshi, and B. Webber, The Penn Discourse Tree-Bank 2.0, Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08), Marrakech, Morocco, 2008. <https://aclanthology.org/L08-1093/> (accessed on November 25, 2025).
- 7 J. Dunietz, L. Levin, and J.G. Carbonell, The BECauSE Corpus 2.0: Annotating causality and overlapping relations, Proceedings of the 11th Linguistic Annotation Workshop (LAW-XI) (2017), 95–104. <https://doi.org/10.18653/v1/W17-0812>
- 8 N. Mostafazadeh, A. Grealish, N. Chambers, J. Allen, and L. Vanderwende, CaTeRS: Causal and temporal relation scheme for semantic annotation of event structures, Proceedings of the Fourth Workshop on Events (2016), 51–61. <https://doi.org/10.18653/v1/W16-1007>
- 9 T. O’Gorman, K. Wright-Bettner, and M. Palmer, Richer Event Description: Integrating event coreference with temporal, causal and bridging annotation, Proceedings of the 2nd Workshop on Computing News Storylines (CNS 2016) (2016), 47–56. <https://doi.org/10.18653/v1/W16-5706>
- 10 P. Mirza, Extracting temporal and causal relations between events, Proceedings of the ACL 2014 Student Research Workshop (2014), 10–17. <https://doi.org/10.3115/v1/P14-3002>
- 11 A. Eirew, E. Nachshoni, A. Slobodkin, and I. Dagan, EventFull: Complete and Consistent Event Relation Annotation, Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (System Demonstrations) (2025), 494–508. <https://doi.org/10.18653/v1/2025.naacl-demo.40>
- 12 M.A.K. Raiaan, M.A.A.S. Khan, A. Muhammad, T.T.M. Nguyen, and S. Hussain, A review on large language models: Architectures, applications, taxonomies, open issues and challenges, IEEE Access 12 (2024), 26839–26874. <https://doi.org/10.1109/ACCESS.2024.3365742>
- 13 Qazqorpus : Kazak tilinin ulttyk korpusy [Qazqorpus: National Corpus of the Kazakh Language], available at <https://qazqorpus.kz/> (accessed November 23, 2025) [in Kazakh]
- 14 Qazqorpora: Kazak tilinin korpustary [Qazqorpora: Corpora of the Kazakh Language], available at <https://qazqorpora.kz/> (accessed November 22, 2025) [in Kazakh]
- 15 Kazakhstan futbol federaciyasynyn resmi sajty [Official Website of the Kazakhstan Football Federation], available at <https://kff.kz/kk/news/23788> (accessed November 22, 2025) [in Kazakh].

Авторлар туралы мәлімет:

Таберхан Роман – Л.Н. Гумилев атындағы Еуразия ұлттық университеті, Ақпараттық жүйелер кафедрасының 3 курс PhD докторанты, Сәтпаев 2, Астана, 010008, Қазақстан.

Самбетбаева Мадина – PhD докторы, қауымдастырылған профессор, Л.Н. Гумилев атындағы Еуразия ұлттық университеті, Ақпараттық жүйелер кафедрасының профессоры, Сәтпаев 2, Астана, 010008, Қазақстан.

Information about the authors:

Taberkhan Roman – 3rd-year PhD doctoral student, Department of Information Systems, L.N. Gumilyov Eurasian National University, Satpayev str., 2, Astana, 010008, Kazakhstan.

Sambetbayeva Madina– Doctor PhD, Associate Professor, Professor of the Department of Information Systems, L.N.Gumilyov Eurasian National University, Satpayev str., 2, Astana, 010008, Kazakhstan.

Қабылданған күні: 24.11.2025. Өңдеуден кейін: 17.03.2026.

Мақұлданған күні: 31.03.2026. Онлайн қолжетімді: 17.06.2026.



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