Қ.И. Сәтбаев атындағы Қазақ ұлттық зерттеу техникалық университеті
«Ақпараттық және есептеу технологиялары институты» ҚР ҒЖБМ ҒК

ӘОЖ 004.89 Қолжазба құқығында

ЫБЫТАЕВА ГАЛИЯ СЕЙТКАЛИЕВНА

Онтологиялық тәсіл негізінде құқыққа қайшы мәтіндік ақпарат мониторингінің ақпараттық-талдамалық жүйесін әзірлеу

8D06103 – Management information systems

Философия докторы (PhD)
дәрежесін алу үшін дайындалған диссертация

Отандық ғылыми кеңесші
доктор PhD,
қауымдастырылған профессор
Мамырбаев О.Ж.

Шетелдік ғылыми кеңесші
техника ғылымдарының докторы,
профессор
Хайрова Н.Ф.
(«Харьков
политехникалық институты» ҰТУ)

Қазақстан Республикасы
Алматы, 2024

МАЗМҰНЫ

НОРМАТИВТІК СІЛТЕМЕЛЕР

Бұл диссертацияда келесі нормативтік құқықтық актілерге сілтемелер қолданылды:

ҚР МЖМБС 5.04.034 – 2011 «Қазақстан Республикасының Мемлекеттік жалпыға міндетті білім беру стандарты. Жоғары оқу орнынан кейінгі білім. Докторантура» Негізгі ережелер ҚР білім және ғылым министрімен бекітілген. «17» маусым 2011ж. №261, Астана 2011.

«PhD философия докторының диссертациясын ресімдеу жөніндегі нұсқаулық», Қ.И. Сәтбаев атындағы ҚазҰЗТУ. «18» сәуір 2023. №6, Алматы 2023.

МЕСТ 7.32-2001. Ғылыми зерттеу жұмыстары туралы есеп. Рәсімдеудің ережесі мен құрылымы. Астана, 2001.

МЕСТ 7.1-2003. Библиографиялық жазба. Библиографиялық сипаттама. Құрастырудың жалпы талаптары мен ережелері.

БЕЛГІЛЕУЛЕР МЕН ҚЫСҚАРТУЛАР

CMC	– компьютерлік-жанама коммуникация (Computer-Mediated Communications)
ПәС	– пәндік сала
RDF	– ресурсты сипаттау ортасы (Resource Description Framework)
XML	– кеңейтілетін белгілеу тілі (eXtensible Markup Language)
CREE	– криминалистік маңызды оқиғаларды шығарып алу (Crime-related Event Extraction)
CRE	– криминалистік маңызды оқиға (Crime-related Event)
LLEs	– логикалық-лингвистикалық теңдеулер (Logical-linguistic equations)
CdEE	– жабық доменнен оқиғаны шығарып алу (Closed-domain Event Extraction
CNN	– үйірткілі нейрондық желі (Convolutional Neural Network)
IE	– ақпаратты шығарып алу (Information Extraction)
k-NN	– k-жақын көршілер әдісі (k-Nearest Neighbors)
HMM	– жасырын марков моделі (Hidden Markov Model)
LSTM	– ұзақ қысқа мерзімді жады (Long Short-Term Memory)
ME	– максималды энтропия (Maximum Entropy)
ML	– машиналық оқыту (Machine Learning)
NLP	– табиғи тілді өңдеу (Natural Language Processing)
EE	– оқиғаларды шығарып алу (Event Extraction)
SVM	– тірек векторлар әдісі (Support Vector Machine)
VSM	– векторлық кеңістік моделі (Vector Space Model)
АПА	– ақырғы предикаттар алгебрасы

КІРІСПЕ

Көптеген интернет-контентпен қанықкан заманауи ақпараттық кеңістік пайдаланушыға ақпарат алудың кең мүмкіндіктерін ұсынады. Алайда, қолда бар білім көлемінің өсуімен бірге жеке пайдаланушыларға да, жалпы қоғамға да зиян келтіруі мүмкін жалған, зиянды немесе құқыққа қайшы контентті анықтау мәселесі де артып келеді. Бұл зерттеу жұмысының маңыздылығы мен өзектілігі қылмыстық құрылымдар қызметінің күшеюіне және Интернет желісінде жалған ақпараттың таралуына байланысты.

Қазіргі уақытта әртүрлі елдердің құқық қорғау және мемлекеттік органдары қылмыс жасалғаннан кейін онымен күресуден гөрі қылмыс пен терроризмнің алдын алуға көбірек көңіл бөледі [1, 2]. Қылмыстың алдын алу парадигмасын ұстану үшін мәтіндік және дауыстық [3] ақпаратты қоса алғанда, ақпараттың үлкен көлемін талдау, деректерді іздеу және мәтінді талдаудың озық технологияларын, сондай-ақ NLP құралдары мен тәсілдерін пайдалану қажет.

Жаңа ақпараттық технологиялардың әлеуетін кеңейтумен байланысты қазіргі қоғамның басты проблемаларының бірі – Интернетті деструктивті және қоғамға қарсы мақсаттары бар қылмыстың құралы ретінде пайдалану мүмкіндігі. Екінші жағынан, Facebook, Twitter, Instagram, YouTube сияқты компьютерлік-жанама коммуникациядағы (Computer-Mediated Communication (CMC)) құрылымданбаған ақпаратты талдаудың жаңа технологиялары мәтіндік деректерді алдын-ала өңдеуге және ықтимал қылмыстардың алдын алуға мүмкіндік береді. Алайда қылмысқа ниетті немесе дайындықты көрсетуі мүмкін барлық сайттардың мазмұнын қолмен табу және қадағалау мүмкін болмағандықтан, құқыққа қайшы Интернет ақпаратты анықтауды автоматтандыру қажет. Көптілді құқыққа қайшы контентті автоматты немесе автоматтандырылған іздеу және талдаудың мұндай жүйелері адамның криминалдық әрекетті жасауды жоспарлап отырғанын немесе жасағанын жоғары ықтималдықпен анықтау мүмкіндігіне ие болуы керек.

Жоғарыда айтылғандардың барлығы интернет-контенттегі ықтимал терроризмнің, экстремизмнің және басқа да құқыққа қайшы және зорлық-зомбылық әрекеттерінің цифрлық іздерін анықтау қажеттілігіне қатысты осы зерттеуге бағытталған **негізгі мәселені** анықтайды. Сонымен қатар, статистикалық әдістер мен машиналық оқыту әдістеріне негізделген тәсілдер оқытылған корпустың болмауына және әлсіз «лингвистикалық» маркерлерге ие құқыққа қайшы Интернеттің пәндік саласының (ПәС) анық еместігіне байланысты жақсы нәтиже бермейді. Демек, осы зерттеуде қарастырылатын **екінші мәселе** – мәтіннің құқыққа қайшы контентке жататынын анықтау кезінде оның семантикалық (мағыналық) құраушысын ескеру қажеттілігі. Бұл мәселе қолданылатын машиналық оқыту әдістеріне семантикалық дифференциалданатын белгілерін қосудан тұратын онтологиялық тәсілді қолдану арқылы шешілуі керек. **Үшінші мәселе** қазақ тілі үшін «Құқыққа қайшы интернет-контент» пәндік саласында онтологияның болмауы, сондай-ақ орыс және ағылшын тілдері үшін мұндай онтологиялардың ашық қолжетімділікте

болмауы болып табылады. Сонымен қатар, алдыңғысына қатысты мәселе мәтіндік корпустарға негізделген тиімді алгоритмдер мен автоматты онтологиялық генерациялау бағдарламалық құралдарының болмауы болып қала береді.

Жоғарыда айтылғандарға сүйене отырып, қазіргі уақытта көптілді құқыққа қайшы интернет-контентті автоматты түрде іздеу және талдау жүйесінің тиімді әдістері, модельдері мен бағдарламалық құралдарының қажеттілігі ерекше **өзекті** деген қорытындыға келуге болады.

Зерттеудің мақсаты Интернет желілерде қазақ және орыс тілдерінің құқыққа қайшы мәтіндерін автоматты сәйкестендіру жүйесінің ақпараттық моделін әзірлеу болып табылады.

Зерттеу міндеттері. Зерттеудің қойылған мақсаттарын іске асыру үшін мынадай мәселелер шешіледі:

1. Онтологиялық тәсіл негізінде құқыққа қайшы мәтіндік ақпарат мониторингінің заманауи әдістері мен модельдерін талдау.
2. Қазақ, орыс, украин және ағылшын тілдеріндегі Интернеттің криминалистік маңызды мәтіндерінің корпустарын әзірлеу.
3. Көптілді (қазақ, орыс және ағылшын тілдері) терминологиялық тезаурус құру.
4. «Құқыққа қайшы интернет-контент» онтологиясын құру.
5. Криминалистік маңызды мәтіндердің мамандандырылған корпустарын автоматты семантикалық белгілеу әдісі мен құралдарын әзірлеу.
6. Онтологиялық тәсіл негізінде құқыққа қайшы мәтіндік ақпарат мониторингінің ақпараттық-талдамалық жүйесін әзірлеу.

Ақпараттық-талдамалық жүйесі «Құқыққа қайшы интернет-контент» онтологиясын, мәтіндердің мамандандырылған корпустарын, криминалистік маңызды мәтіндердің мамандандырылған корпустарын автоматты семантикалық белгілеудің бағдарламалық құралын және әлеуметтік желілерде және басқа интернет көздерінде құқыққа қайшы контентті талдау мен мониторингінің интеграцияланған технологиясын қамтиды.

Қойылған мақсат аясында құқыққа қайшы көптілді интернет-контентті автоматты түрде іздеудің және талдаудың ақпараттық-талдамалық жүйесін әзірлеу міндеті шешілуде. Әзірленген жүйені пайдалану қылмыстарды ашу және құқыққа қайшы әрекеттерді болдырмау ықтималдығын арттыру есебінен құқық қорғау және арнаулы мемлекеттік ұйымдар жұмысының тиімділігін арттыруға мүмкіндік береді. Бұл зерттеудің әлеуметтік әсері құқықтық және криминогендік жағдайды жақсарту және жалпы қоғамның өмір сүру сапасын жақсарту болып табылады.

Зерттеу нысаны құқыққа қайшы көптілді мәтіндік ақпаратты автоматты түрде іздеу және талдау жүйелері болып табылады.

Зерттеу пәні онтологиялық тәсілге негізделген құқыққа қайшы көптілді мәтіндік ақпаратты іздеуге және талдауға арналған модельдер мен әдістер, бағдарламалық құралдар болып табылады.

Зерттеу әдістері. Статистикалық ықтималдық әдістері, машиналық оқыту әдістері, корпустық лингвистика әдістері, интеллект теориялары, сондай-ақ табиғи тіл мәтіндерін семантикалық және грамматикалық талдау әдістері және сараптамалық бағалау әдістері.

Зерттеу жұмысының ғылыми жаңалығы.

– интернеттің криминалистік маңызды көптілді мәтіндерінің корпустары және онтологиялық тәсілге негізделген криминалистік маңызды мәтіндердің корпустарын автоматты семантикалық белгілеу құралдары жасалды;

– көптілді терминологиялық тезаурус, «Құқыққа қайшы интернет-контент» онтологиясы және ақпараттық модель құрылды;

– онтологиялық тәсіл негізінде құқыққа қайшы мәтіндік ақпарат мониторингінің ақпараттық-талдамалық жүйесі әзірленді.

Жұмыстың ақпараттық қауіпсіздік және заңсыз контентпен күресу саласында маңызы зор, сонымен қатар мәтінді өңдеу және семантикалық талдау саласында одан әрі зерттеулер жүргізу үшін перспективалар ашылады. Жобаны іске асырудың экономикалық және индустриялық мүдделілігі алынған нәтижелерді мінез-құлықты талдау жөніндегі мамандардың, құқық қорғау органдарының және қауіпсіздік қызметтерінің ықтимал құқыққа қайшы қатерлерді бағалау рәсімдерін орындау немесе жақсарту кезінде жедел қолдануы үшін пайдалану мүмкіндігі болып табылады.

Алынған нәтижелердің **теориялық маңыздылығы** көптілді мәтіндік ақпаратты өңдеу мен талдаудың қолданыстағы модельдері мен әдістерін бейімдеу және әзірлеу болып табылады.

Алынған нәтижелердің **практикалық құндылығы** қорғауға шығарылған қағидалар негізінде ақпараттық-талдамалық жүйені әзірлеу болып табылады.

Қорғауға шығарылатын негізгі қағидалар.

1. Екі корпус және көптілді терминологиялық тезаурус құрылды:

– орыс, украин және ағылшын тілдеріндегі мәтіндерді (украин тілінде 3147 мәтін, орыс тілінде 5506 мәтін, ағылшын тілінде 300 мәтін) қамтитын көптілді корпус;

– орыс тіліндегі 3000 мәтінді және қазақ тіліндегі 3000 мәтінді, оның ішінде мағынасы жағынан тураланған қазақша-орысша сөйлемдерді құрайтын 2000 мәтінді қамтитын параллель қазақ-орыс корпусы;

– 600-ден астам негізгі сөздерді (330 зат есім, 107 сын есім және 170-ке жуық етістік) және 2500-ден астам негізгі сөз синонимдерін қамтитын тезаурус.

2. «Құқыққа қайшы интернет-контент» онтологиясы құрылды.

3. Криминалистік маңызды мәтіндердің мамандандырылған корпустарын автоматты семантикалық белгілеу әдісі мен құралдары әзірленді.

4. Онтологиялық тәсіл негізінде көптілді құқыққа қайшы интернет-контентті автоматты іздеу және талдау жүйесінің ақпараттық моделі мен бағдарламалық құралы әзірленді.

Сенімділік дәрежесі мен апробациялау нәтижелері. Диссертацияның негізгі нәтижелері халықаралық және шетелдік ғылыми конференцияларда, ғылыми семинарларда баяндалды және талқыланды:

1. Automatic Multilingual Ontology Generation Based on Texts Focused on Criminal Topic // Proceedings of the 5th International Conference on Computational Linguistics and Intelligent Systems (COLINS 2021). – 2021. – Vol. 1. – P. 108-117.

2. Қазақ тіліндегі мәтіндерде коллокацияларды анықтаудың статистикалық әдістерін талдау // «Информатика және қолданбалы математика» VI Халықаралық ғылыми конференциясының материалдары (Алматы, 2021. – Б. 256-262).

3. Параллель корпусты әзірлеу мәселелері // «Информатика және қолданбалы математика» VII Халықаралық ғылыми конференциясының материалдары (Алматы, 2022. – Б. 175-182).

4. Creating a Thesaurus "Crime-Related Web Content" Based on a Multilingual Corpus // Proceedings of the 7th International Conference on Computational Linguistics and Intelligent Systems (COLINS-2023). – 2023. – Vol. 3396. – P. 77-87.

5. Experimental Verification of Collocation Detection Methods // Proceedings of the 5th Workshop for Young Scientists in Computer Science and Software Engineering - CS&SE@SW, SciTePress. – 2023. – P. 13-18.

6. «Құқыққа қайшы интернет-контент» көптілді онтология қосымшасын құру // «Информатика және қолданбалы математика» VIII Халықаралық ғылыми-практикалық конференциясы (Алматы, 2023. – Б. 107-114).

Зерттеушінің жеке үлесі. Докторант диссертациялық жұмыстың міндеттерін өз бетінше орындап, шешті. Интернеттің криминалистік маңызды мәтіндерінің көптілді корпустарын жасады. «Құқыққа қайшы интернет-контент» онтологиясын құрды. Онтологиялық тәсіл негізінде құқыққа қайшы мәтіндік ақпарат мониторингінің ақпараттық-талдамалық жүйесін әзірледі. Әзірленген модельдер мен технологияларды эксперименттік бағалауды орындады.

Диссертация тақырыбының ғылыми-зерттеу жұмысының жоспарларымен байланысы. Диссертациялық жұмыс «Онтологиялық тәсіл негізінде көптілді құқыққа қайшы веб-контентті автоматты іздеу және талдау жүйесінің ақпараттық моделі және бағдарламалық құралы» – AP09259309 (2021-2023) ҚР ҒЖБМ гранттық зерттеулер жобасы аясында орындалды.

Диссертациялық зерттеудің негізгі нәтижелерінің жарияланымдары. Диссертациялық жұмыс тақырыбы бойынша 2 авторлық куәлік алынды және 5 жұмыс жарияланды, оның ішінде 3 мақала ҚР ҒЖБМ ғылым және жоғары білім саласындағы сапаны қамтамасыз ету Комитеті ұсынған журналдарда жарияланды, 2 мақала Scopus және Web of Science базасымен индекстелген басылымдарда жарияланды, 1 монография жарыққа шықты:

1. A Parallel Corpus-Based Approach to the Crime Event Extraction for Low-Resource Languages // in IEEE Access. – 2023. – Vol. 11. – P. 54093-54111 *(Scopus: Процентиль – 89, Q1; Web of science: IF – 0.89, Q2).*

2. Correction of Kazakh synthetic text using finite state automata // (2021) Journal of Theoretical and Applied Information Technology. – 2021. – Vol. 99, №22. – P. 5559-5570 *(Scopus: Процентиль – 30, Q3).*

3. Методы формального представления сущностей в криминальных новостях для автоматического построения онтологии преступлений // Известия НАН РК. Серия информатики. – 2022. – №3. – С. 136-152.

4. Лингвистикалық онтологияны қолдану және қалыптастыру мәселелеріне шолу // ҚР ҰҒА жаңалықтары. Информатика сериясы. – 2022. – №1. – С. 96-106.

5. Сарапшылар пікірлерінің келісім өлшемі ретінде Коэннің каппа коэффициентінің ерекшеліктері // ҚР ҰИА жаңалықтары. Ақпараттық технологиялар сериясы. – 2023. – №3(89). – С. 139-151.

6. Веб-приложение многоязычной базовой онтологии «Противоправный веб-контент». №32055 от 26.01.2023 г. (Қосымша А).

7. «Құқыққа қайшы интернет-контент» көптілді онтология қосымшасы. №38766, 29.08.2023 ж. (Қосымша А).

8. Automatic identification of illegal texts in Internet (Almaty: Institute of Information and Computational Technologies, 2023. - 151 p. (монография).

Диссертациялық жұмыстың құрылымы мен көлемі. Диссертациялық зерттеу жұмысы кіріспеден, 4 бөлімнен, қорытындыдан, 161 атаудан тұратын әдебиеттер тізімінен және 3 қосымшадан тұрады. Жұмыс 80 бетте берілген және 19 сурет, 13 кестеден тұрады.

Автор ғылыми жетекшілерге: PhD, қауымдастырылған профессор О.Ж. Мамырбаевқа қойылған міндеттері мен жұмыстағы қолдауы үшін, профессор Н.Ф. Хайроваға жемісті пікірталастары мен құнды ескертулері үшін алғысын білдіреді.

1 ҚҰҚЫҚҚА ҚАЙШЫ МӘТІНДІК АҚПАРАТТЫ ӨҢДЕУДІҢ ЗАМАНАУИ ӘДІСТЕРІ МЕН ТАЛДАУЫ

1.1 Көптілді құқыққа қайшы интернет-контентті сәйкестендірудің қазіргі мәселелері

Қазіргі уақытта криминалдық әрекеттерге және полицияның арнайы қызметтерінің жұмысына қатысты ақпарат күн сайын дүние жүзінде әртүрлі блогтарда және басқа да интернет көздерінде жарияланады. Мұндай мәтіндерден қылмысқа байланысты оқиғаларды алу қылмыстық әрекеттерді, сондай-ақ әртүрлі елдердегі немесе аймақтардағы полицияның жұмысын бақылау, талдау және салыстыру арқылы контекстік талдауға мүмкіндік береді. Интернет-контент құқық қорғау органдары үшін де, мамандандырылған күш құрылымдары үшін де маңызды ақпарат көзіне айналды [4]. Интернетте құқыққа қайшы және криминалдық ақпараттың әртүрлі түрлерін қамтитын мәтіндерді автоматты түрде сәйкестендіру мәселесі бойынша бар ғылыми зерттеулерді екі негізгі бағытқа бөлуге болады: 1) көбінесе көңіл-күйді талдау мәселелеріне негізделген психолингвистикалық тәсіл (Opinion Mining) және 2) кілттік сөздерге немесе сирек жағдайларда онтологияларға негізделген тәсіл.

Психолингвистикалық тәсіл адамның психологиялық және психикалық жағдайын сол адам жасаған мәтіндер негізінде талдауға мүмкіндік береді [5]. Осы бағыттағы зерттеулердің көпшілігі интернет-контенттегі белсенділіктің мінез-құлық маркерлерін анықтауға бағытталған. Мұндай маркерлер қарым-қатынасты, мотивтерді, ниеттерді, тіпті ықтимал қылмыстық немесе радикалды зорлық-зомбылық мүмкіндіктерін көрсететін белгілі бір лингвистикалық ерекшеліктер арқылы көрінуі мүмкін. [6] зерттеуде мұндай лингвистикалық маркерлер радикалды зорлық-зомбылық маркерлері ретінде анықталған, олардың жиынтығы саясаткерге шабуыл, террористік акт және т.б. сияқты формаларда қылмыстық әрекетке дайындықты көрсете алады. Сонымен қатар, мұндай маркерлер қаржылық алаяқтық, авторлық құқықты бұзу, балалар порнографиясын тарату және т.б. сияқты қылмыстық түрлерінде жасалған заңсыз әрекеттер туралы хабарлай алады [7, 8].

Мысалы, [9] зерттеуде фразаны қолдану мен автордың психологиялық жағдайы арасындағы корреляцияның бар екендігі туралы гипотеза негізделеді және расталады. [4, p. 246-255]-жұмыста авторлар мәтіндерде кездесетін «ескерту мінез-құлқы» деп аталатын әлеуметтік желілердегі немесе блогтардағы жазбаша мәтіндердегі радикалды зорлық-зомбылықтың мінез-құлық белгілерін бақылау мүмкіндігін зерттеді және дәлелдеді. [10] мақалада «жалғыз террорист» мінез-құлық үлгісін бағалау үшін [6, p. 78-85] авторлар зорлық-зомбылық жасау қабілетін көрсететін акторлардың әлеуметтік белсенділігінің лингвистикалық және лингвистикалық емес маркерлерімен байланысты 198 айнымалыны пайдаланды. Дегенмен, қазіргі уақытта мінез-құлықтың лингвистикалық маркерлеріне негізделген зерттеулер негізінен эксперименттік болып қала береді және Deep Web-пен жұмыс істейтін қолданбаларға қосылады.

Көбінесе психолингвистикалық тәсілмен байланысты зерттеулер әртүрлі форумдардың мәтіндерінде байқалатын ашу, жеккөрушілік және нәсілшілдік деңгейлерін салыстыру үшін көңіл-күйді талдау әдістерін (Opinion Mining) қамтиды [11]. Сонымен қатар, интернет-контенттегі радикалды және криминалдық мәтіндерді анықтау үшін көңіл-күйді талдау тәсілдерін қолдану олардың сенімділігі мен дәлдігінің төмендігіне байланысты кең таралмаған [12]. Көптеген жағдайларда мұндай зерттеулер әлі де эксперименттік сипатта болады.

Көңіл-күйді талдаумен қатар, заңсыз және құқыққа қайшы ақпараттың әртүрлі түрлерін қамтитын мәтіндерді автоматты түрде ерекшелеу үшін кейбір зерттеулер машиналық оқытуды классификациялау тәсілдерін (аңғал байес алгоритмі, логистикалық регрессия, сызықтық SVM, кездейсоқ ормандар, градиентті жоғарылату) қолданады. Сонымен қатар, көптеген мақалалар классификация сапасын жақсарту үшін дифференциалды лексикалық және семантикалық белгілерді қосымша қолдануды қарастырады [13]. Мысалы, [14] мақалада терроризмді қолдайтын Twitter хабарламаларын жіктеу үшін авторлар функционалды сөздер, жиілік сөздері, тыныс белгілері, биграммалар және т.б. сияқты стилометриялық белгілерді қолданды.

Көптілді құқыққа қайшы интернет-контентін іздеу және талдау мәселесіне қолданылатын екінші тәсілдеме кілттік сөздерді пайдалануға негізделген. Мысалы, [15, 16] мақалаларында авторлар әртүрлі экстремистік әрекеттерге тән кілттік сөздер мен сөз тіркестерін қамтитын сөздіктерді пайдаланды. [17] мақалада жеккөрушілік пен зорлық-зомбылық тақырыбына қатысты кілттік сөздер акторлар-қолданушылармен өзара байланысты картографиялық веб-сайттар құру үшін қолданылды. Бұл әдіс көбінесе лақап атпен шығатын заңсыз және экстремистік мәтіндік ақпараттың авторларын анықтау үшін қолданылды. Дегенмен, авторлар өз зерттеулерінде қарапайым кілттік сөздердің орнына пәндік саланың ұғымдары арасындағы байланыстарды көрсететін онтологияларды пайдаланған кезде әдістің тиімділігі айтарлықтай жоғарылайтынын атап көрсетті.

Көптілді құқыққа қайшы интернет-контентін автоматты түрде іздеу мен талдаудың қолданыстағы тәсілдеріне аналитикалық шолу көптеген әдістердің болуына қарамастан, криминалистік маңызды хабарламаларды немесе құжаттарды сәйкестендірудің әмбебап моделі туралы айтуға әлі ерте екенін көрсетеді. Сонымен қатар, талдау қазіргі уақытта осы мәселені шешуге бағытталған тиімді қосымшалардың жоқтығын көрсетеді. Қолданыстағы идеялар мен тәсілдердің алуан түрлілігі мен әртүрлілігі осы ғылыми салада жүргізіліп жатқан теориялық зерттеулердің көптігін және бұл мәселенің шешімін табудан гөрі оның өзектілігін көрсетеді.

1.2 Лингвистикалық онтологияны қолдану және қалыптастыру мәселелеріне шолу

Қазіргі уақытта онтология әр түрлі пәндік салалардағы білім қорын ұсыну мәселелерін шешу үшін жиі қолданылады. Онтология семантикалық желілермен қатар белгілі бір пәндік саладағы білімді көрсету үшін ыңғайлы абстракция

болып табылады [18]. Бұл міндет онтологияны толықтыру міндетімен тығыз байланысты. Онтологияны толықтыру дегеніміз – әртүрлі көздерді автоматты түрде талдау және мәліметтер базасы ПәС онтологиясына негізделген ақпараттық жүйенің контентін табылған ақпаратпен толтыру. Мұндай жүйелерде алынған ақпарат берілген онтологияның түсініктері мен қатынастарының экземплярі ретінде ұсынылады.

Ал қазіргі заманда онтология ұғымын барлық интернет желілеріне енгізу өте қажет. Онтологиялар тезаурустар мен таксономияларға көп жағынан ұқсас, бірақ олардан кеңірек, өйткені олар сипатталатын деректердің құрылымын сипаттауға арналған қосымша құралдарды ұсынады; олардың негізінде онтологиялар ақпарат туралы ақпарат болып табылады, олар метадеректер болып табылады [19].

Онтологияны концептуализация негізінде құрылған ПәС-ның формалды көрінісі ретінде анықтау оның өзара байланысты үш құрамдас бөлігін анықтауды қамтиды: терминдер таксономиясы, терминдердің мағынасын сипаттау, сондай-ақ оларды пайдалану және өңдеу ережелері. Осылайша, *O* онтология моделі туралы үштік орнатады [19, с. 10]:

$$O = (X, R, F), \tag{1.1}$$

мұнда X – онтология ұсынатын тұжырымдамалардың (ұғымдардың, терминдердің) ақырғы жиынтығы;

R – ұғымдар арасындағы қатынастардың ақырғы жиынтығы;

F – ұғымдарда және (немесе) қатынастарда берілген интерпретация функцияларының ақырғы жиынтығы. Бұл модель желілік білім моделінің бір түрі болып табылады.

Онтологияны әзірлеу мен қолдауға қойылатын талаптарды қанағаттандыруға бағытталған жобалар саны үнемі өсіп келеді.

Mikrokosmos онтологиясы (кейінірек OntoSem деп аталды) – ең танымал онтологиялық ресурстардың бірі. Бұл онтология «онтологиялық семантика» деп аталатын тәсілдің аясында жасалған [20]. Онтология мәтінді автоматты түрде өңдеу қосымшаларында қолдануға және мәтіндік сөйлемдердің мазмұнын мағыналық, тілге тәуелсіз ұсынуға арналған. Кіріс мәтін үшін алдын-ала өңдеу, морфологиялық талдау, синтаксистік талдау, семантикалық талдау жасалады, оның нәтижелері мәтіндік-семантикалық көрініс түрінде ұсынылады.

Авторлар онтологияның көлемін шектеуге көп күш жұмсағандықтан, Mikrokosmos онтологиясының мөлшері шамамен 6 мың ұғымды құрайды, олардың әрқайсысы орта есеппен 16 қасиетпен сипатталады. Жүйенің сөздік қоры бірнеше ондаған мың сөздер мен сөйлемдерді құрайды.

Қазіргі уақытта лексикалық семантиканы сипаттау саласындағы белгілі жобалардың бірі – белгілі лингвист Чарльз Филлмордың [21] басшылығымен фреймдік семантика тұжырымдамасы аясында құрылған FrameNet лингвистикалық ресурсы. Жобаның мақсаты – фреймдік семантикаға негізделген онлайн-лексикалық ресурс құру және оған мәтіндер корпусы түрінде база беру.

2009 жылы ресурста 11000-нан астам лексикалық бірліктермен байланысқан 960 иерархиялық ұйымдастырылған фреймдер болды.

Орыс тіліндегі ең үлкен тезаурус – RuThes тезаурусы. Бұл жобаны Ақпараттық зерттеулер зертханасы әзірлеуде. Ол WordNet принциптеріне негізделген, бірақ нысандарды сипаттау моделі әртүрлі. Тезаурус бірлігі – мағыналары осы тұжырымдамаға сәйкес келетін терминдер жиынтығымен жабдықталған ұғым. Қазіргі уақытта RuThes тезаурусында 55 мың ұғым, 158 мың сөз бен сөйлемше, осы ұғымдар арасындағы 210 мың байланыс бар.

Орыс тілінің тезаурусын құрудың тағы бір белсенді жобасы – Yet Another RussNet (YARN). Әзірлеушілер WordNet-ке толық сәйкес келетін модельді пайдаланады, ол синсеттерге – жалпы лексикалық мағынамен біріктірілген синонимдер мен квазисинонимдер топтарына негізделген. Синонимдер қарама-қарсы мағыналары бар терминдер арасында орнатылатын иерархиялық қатынастар мен антонимиялық қатынастармен байланысты. Сонымен қатар, сөздер арасындағы қатынастар, сондай-ақ YARN және WordNet synsets жиынтықтары арасындағы тіларалық қатынастар бар.

Бастапқы контент ретінде Уикисөздіктен алынған ақпарат пайдаланылды. Бұл жобаның айрықша ерекшелігі – краудсорсингтік тәсіл негізінде тезаурус қосуды ұйымдастыру: YARN сайтына тіркелу арқылы кез келген адам деректерді қосуға және өңдеуге қатыса алады.

Қазіргі уақытта YARN құрамында 143508 сөз, 69799 синонимдік жиынтық, 104906 өңделмеген синоним жұптары және 29764 өңделмеген жалпы қатынастар бар.

Осы контекстегі терминдер мен қатынастарды түсіндіру үшін қылмысқа байланысты онтологиялар да бар және олар кейінірек кейбір қолданыстағы жүйелердегі білімді көрсету үшін қолданылады. Мысалдарға мультимодальды жағдайды бағалау және талдау платформасының жобасы (MOSAIC) [22]; LEAs өзара әрекеттесуінің еуропалық онтологиясын және көп тілді қылмыс онтологиясын бір уақытта қолданатын CAPER [23]; және ePOOLICE жобасы [24], COPKIT жобасы [25], ASGARD жобасы [26], TENSOR жобасы [27] жатады.

Онтологияны құру және сүйемелдеу әр түрлі жобаларда танымал назарға айналды. Осындай жобалардың бірі – Интерполдың жіктемелік схемаларына сәйкес есірткі қылмыстарымен күресуге арналған CAPER [23, p. 189-197] онтологиясын әзірлеу. Бұл көптілді онтология төрт негізгі ұғымнан тұрады – «Қылмыстар», «Әдістер», «Маңызды шарттар» және «Елдер». Онтологияның қазіргі нұсқасында 346 түйін бар және қазіргі уақытта итальян, испан, ағылшын және иврит тілдеріне аудармалар, синонимдер және жаргон терминологиясын жинау жұмыстары жүргізілуде.

Сонымен қатар, құқық қорғау органдарына қылмыстық әрекеттерді талдау мен алдын-алуда қолдау көрсету үшін бірнеше басқа жобалар құрылды. Мысалы, COPKIT [25] жобасы тергеу мен стратегиялық талдауға көмектесу үшін деректерге негізделген полиция технологияларын әзірледі. Оның құралдар жинағы жедел және стратегиялық деңгейлерде ерте ескерту/алдын ала әрекет ету

парадигмасын сақтай отырып, білімді қалыптастыруға және пайдалануға мүмкіндік береді.

ASGARD [26] жобасы сонымен қатар криминалистикалық тергеу үшін киберқылмыс деректерін қоса алғанда, үлкен деректерді алуға, біріктіруге, бөлісуге және талдауға арналған кластағы ең жақсы құралдар жинағын әзірледі. Тағы бір жоба, TENSOR [27], терроризмге қатысты контентті біріктіру және интернеттегі қауіптерді анықтау үшін бірыңғай семантикалық инфрақұрылым жасады. Бұл құрылым онтологияны және құқық қорғау органдарына террористік әрекеттерді ерте анықтау, радикалдандыру және жалдау үшін жоспарлау және алдын алу функцияларын ұсынатын семантикалық негіздеудің бейімделу механизмін қамтиды. Дегенмен, бұл құралдардың әдетте екі негізгі кемшілігі бар. Біріншіден, олар көптілді емес және қолданыстағы аннотацияланған криминалдық есептер негізінен тек ағылшын тілінде қол жетімді. Екіншіден, кейбіреулерінде визуализация мүмкіндіктері шектеулі, атап айтқанда аталған субъектілер арасындағы танылған байланыстардың графикалық көрінісі.

Л.Н. Гумилев атындағы Еуразия ұлттық университетінің ғалымдары Шарипбаев А.А., Омарбекова А.С. қазақ тілінің морфологиялық ережелерінің онтологиялық модельдерін құрастырды, бұл 40000 бастапқы сөз формаларының білім қорынан 3200000-нан астам сөз формалары бар мәліметтер базасын автоматты түрде құру үшін әр сөз табының сөзжасамы мен сөзөзгерімінің формалды ережелерін жазуға мүмкіндік берді. Авторлар интеллектуалды электронды университеттің семантикалық моделін онтология түрінде құрды.

Талдау көрсеткендей, қазіргі уақытта криминалдық тақырыпқа байланысты орыс және қазақ тілдерінің жалпыға қолжетімді онтологиялары жоқ.

Пәндік саланың көптілді онтологияларының толық болмауы онтологияны жасаушылардың қазіргі жұмыс бағытын көрсетеді. Сондай-ақ, мұндай лексикалық ресурстардың да, оларды әзірлеу модельдерінің де ашықтығы онтологияның сапасын жақсартуға, сондай-ақ тиісті салалардағы мәтінді автоматты түрде өңдеу мәселелерін шешуге айтарлықтай үлес қосуға мүмкіндік беретінін атап өткен жөн.

1.3 Мәтін корпустары негізінде онтологияны автоматты түрде генерациялау саласындағы зерттеулердің қазіргі жағдайы

Онтологияны автоматты түрде құру әдістері белгіленген пәндік саланың арнайы лингвистикалық корпустарын қолдануға негізделген. Мұндай корпустарда арнайы семантикалық белгілер болуы керек немесе оқиғалардың түрлерін сипаттайтын семантикалық аннотацияланған белгілер, мысалы, әлеуметтік-саяси (SPE) және <Person>, <Organization>, <Location>, <Time>, <Vehicle>, <Weapon> және және т.б. сияқты оқиға дәлелдері болуы керек. Мысалы, Linguistic Data Consortium әзірлеген DEFT Richer Event Description Annotation Corpus аннотацияланған корпусында алдын ала оқыту жинағы ретінде 158 құжат және тестілік жинағы ретінде 202 қосымша құжат бар [28]. Авторлар ұсынған корпустың аннотация схемасы 8731 оқиға мен 10319 нысанды қамтиды және ағылшын, қытай және испан жаңалықтар мақалалары мен пікірталас

форумдарынан нысандар мен оқиғаларды шығарып алу мәселелеріне көзқарастарды формалды бағалау үшін пайдаланылуы мүмкін.

Көбінесе онтологияны құру кезінде ұғымдар мен терминдерді шығарып алу үшін аннотацияланған оқиғаларды қамтитын тар шеңберде мамандандырылған корпустар қолданылады. Ең танымал ұқсас корпустар – биомедициналық пәндік саланың тар шеңберде мамандандырылған корпустары. Мысалы, A. Ramponi және т.б. [29] жұмысында биомедициналық бағдарламада қолмен аннотацияланатын оқиғаларды қамтамасыз ететін лингвистикалық ресурстар талданды, мұндай ресурстарға GENIA оқиғалар корпусы, BioInfer (Biomedical Information Extraction Resource) корпусы, Gene Regulation Event Corpus (GREC), GeneReg корпусы және басқалары кіреді. Аннотацияланған оқиғаларды қамтитын корпустар бар, олар тек биомедицинаға ғана емес, сонымен қатар кейбір басқа бағдарламалық жасақтамаларға да қатысты. Жақында жүргізілген [30] зерттеуде X. Ding және т.б. авторлар музыкаға қатысты оқиғаларды шығарып алу үшін музыкалық бағдарламалық жасақтамаға тән корпусты пайдалануды ұсынды. Оқиға түрлерін тану үшін олар кілттік сөздер мен триггерлерді сүзуге негізделген әдісті қолданды.

Соңғы бірнеше жылда қылмысқа байланысты салаларды зерттеу үшін лингвистикалық ресурстарды пайдалану да күшейе түсті [31, 32]. Мысалы, жеккөрушілік тілін анықтау мәселесін шешу үшін онтологияларды, корпустарды, тезаурустарды және құрылымдық лексикалық базаларды қолдану әлі де өзекті болып табылады. F. Poletto және басқалардың [33] жүйелі және өзекті шолуы бүгінгі күнге дейін жеккөрушілік мәселесін зерттеуге бағытталған 64-тен астам аннотацияланған корпустар мен лексикалық ресурстар (оның 37-сі ағылшын тілінде) бар екенін көрсетті. Ç. Çöltekin [34] жұмысында Twitter микроблогының кездейсоқ жазбаларынан қорлайтын сөздерді қамтитын түрік тілінің алғашқы корпусы ұсынылған. J. Salminen және басқалары [35] интернетте жеккөрушілік тілінің түрлері мен мақсаттарын қамтитын жеккөрушілік пікірлердің егжей-тегжейлі таксономиясын құрды. Жақында жарияланған M. Sanguinetti және т.б. [36] мигранттарға қатысты жеккөрінішті сөздер үшін аннотацияланған Twitter-дегі хабарламалар корпусын сипаттайды. Ондағы әрбір твитте жеккөрушілік (иә/жоқ), агрессивтілік (күшті/әлсіз/жоқ), қорлау (күшті/әлсіз/жоқ), ирония (иә/жоқ) және стереотип (иә/жоқ) белгілері бар. Осы зерттеуді жалғастыра отырып, E. Bassignana және басқалары [37] HurtLex деп аталатын жеккөрушілік сөздерін қамтитын ағылшын және испан сөз қорын жасады. R. Kumar мен әріптестердің жұмысы [38] Twitter мен Facebook-тен хинди-ағылшын тілінің аралас деректеріне негізделген аннотацияланған корпус құру мәселесін қарастырады. Бұл корпус агрессияны көрсететін тегтер жиынтығымен түсіндіріледі. D. Battistelli-дің [39] соңғы зерттеуі француз тілінде жеккөрушілік тілінің онлайн доменінің онтологиясын құру әдістемесін ұсынды. Алайда, жұмыс тек модельді әзірлеу аспектілеріне бағытталған, ал мәтіндерге түсініктеме беру үшін онтологияны практикалық қолдану қарастырылмаған.

Зерттеудің осы бағытының танымалдығы SemEval-2019 [40] және SemEval-2020 [41] тапсырмаларындағы 6-тапсырма: «Әлеуметтік медиадағы

қорлайтын сөздерді сәйкестендіру және санаттау» (OffensEval) және 12-тапсырма: «Әлеуметтік медиадағы қорлайтын сөздерді көптілді сәйкестендіру" (OffensEval 2020) конференцияларындағы ұсыныстармен расталады. Қорлау түрлерін автоматты түрде санаттау және қорлау мақсаттарын анықтау кезінде қорлайтын сөздерді анықтаудың әртүрлі тәсілдерін бағалау үшін ағылшын [40, p. 75-85], араб, дат, ағылшын, грек және түрік [41, p. 1425-1446] тілдеріндегі твиттер жинақтары пайдаланылды, олар OLID иерархиялық таксономиялық схемасына сәйкес түсіндірілді [42]. Дұшпандық тілді анықтауға арналған жақсы құрылымдалған иерархиялық жүйе бар екені анық. Алайда, құқыққа қайшы және полиция қызметіне байланысты оқиғаларды анықтаудың мұндай жалпы схемасы әлі жоқ.

Сонымен қатар, құқықтық және сот мәтіндерінің жанрларына бағытталған бірнеше корпус бар (Cambridge Corpus of Legal English, The House of Lords Judgments Corpus, The Proceedings of the Old Bailey, JUD-GENTT, A Corpus of Malawi Criminal Cases) [43-47]. Мысалы, G. Pontrandolfo [43, p. 209-233; 44, p. 56-77] 480000-ға жуық лексемадан тұратын салыстырмалы монолингвистикалық корпусты ұсынды, оған Италияның Жоғарғы сотының, Еуропалық Одақ сотының және Еуропалық адам құқықтары сотының шешімдерінің ішкі корпустары кіреді. Сонымен қатар, әдетте, мұндай корпустар көптілді немесе параллель болып табылады және заңды мәтіндерді аудару мен түсіндірудің нақты мәселелерін шешуге арналған.

Көптеген жағдайларда қылмысқа байланысты мәтіндік талдау міндеттері газет мақалаларының корпусына негізделген. I.A. Ras [48] тезистерінде авторлар корпоративті алаяқтық туралы жаңалықтарды талдау үшін шамамен 85000 жаңалықтар мақаласы бар Британдық газет корпусын пайдаланды. S. Mukherjee мен K. Sarkar [49] бенгал тілінде жазылған газеттер корпусын қылмыс деңгейі жоғары жерлердің суретін автоматты түрде шығарып алу үшін пайдалануды ұсынды.

Алайда, тұтастай алғанда, V.D. H de Carvalho мен A.P.C.S. Costa [50] талдауы көрсеткендей, полиция есептерін қамтитын криминалдық тақырыптағы мәтіндері бар мамандандырылған домендік корпустар газет мақалаларына қарағанда мәтіндерді интеллектуалды талдау үшін аз қолданылады. G. Karystianis және басқалар [51] және A. Adily [52] Жаңа Оңтүстік Уэльс полициясы берген 492393 тұрмыстық зорлық-зомбылық оқиғасының сипаттамасын қамтитын корпусты пайдаланды [53]. M.R. Alagheband-тің соңғы жұмысы [54] медиа және академиялық ақпаратқа негізделген және морфологиялық белгілеу нәтижелерін қамтитын киберқауіпсіздікке бағытталған екі корпусты салыстырды. D. Gunawan және басқалары [55] зерттеуінде порнографиялық сайттарды бұғаттау техникасын қолдану үшін ұсынылған Индонезия тіліндегі (Bahasa) порнографиялық домендердің нақты корпусын құруға мүмкіндік берді. Өз жұмысында R.R. de Mendonça және басқалары [56] қылмыстық ниеттерді жіктеудің онтологиялық негізін ұсынды, ал [57] зерттеуде қылмысқа байланысты ықтимал Twitter хабарламаларын таңдау үшін криминалдық сөйлемше онтологиясы (OntoCexp) қолданылды.

Тұтастай алғанда, жүргізілген талдау криминалистік маңызды оқиғаларды, олардың атрибуттары мен қатысушыларын шығарып алуға байланысты соңғы онжылдықтағы зерттеулер аннотацияланған корпустарға, онтологияларға немесе мамандандырылған сөз қорларына негізделгенін көрсетеді.

Ағылшын тілін жоққа шығаратын барлық басқа тілдер үшін мұндай жағдайларды табу әлдеқайда аз. Мысалы, қазақ және орыс тілдері үшін заңсыз немесе криминалистік маңызды мәтіндік ақпаратты қамтитын корпустар ашық қолжетімділікте жоқ [58].

Болашағы зор зерттеудің авторлары [59] қазақ тіліндегі экстремистік мазмұндағы мәтіндер корпусын ұсынды. Олар TF-IDF салмақ функциясын автоматты түрде есептеуді қарастырды, ол осы корпустың кілттік сөздерінің тізімін олардың флекциялық формаларын сақтай отырып анықтайды. Алайда, өкінішке орай, іс жүзінде пайдалану үшін бұл корпус тым кішкентай. Өз зерттеуімізде біз Қазақстан Республикасының жаңалықтар сайттарының криминалистік маңызды мәтіндерін қамтитын қазақ-орыс параллель корпусын пайдалануға негізделеміз [1, p. 116-124].

Осылайша, жүргізілген шолу көрсеткендей, қазіргі уақытта құқыққа қайшы контентті іздеуге және талдауға бағытталған зерттеулер жеткілікті болғанымен, негізінен қолданыстағы әзірлемелер ағылшын, француз, қытай және басқа да еуропалық тілдердің мәтіндерін өңдеуді қамтиды.

Машиналық оқыту мен NLP саласындағы соңғы жылдардағы жетістіктер мәтіндік корпустардың құрылымданбаған білімі негізінде онтологияны автоматты түрде генерациялау бағытында жаңа зерттеулердің пайда болуына ықпал етті. Мысалы, онтологияны дамытуда синтаксистік шаблондар мен ережеге негізделген тәсілдер қолданылады [60, 61]. Алайда, бұл тәсілдер қолмен жасалған еңбекті қажет ететін ережелер жиынтығын және білімді ұсыну шаблондарын қажет етеді, бұл оларды пәнге тәуелді етеді. Сонымен, [62] және [63] авторлары медициналық тақырыпқа байланысты мәтіндерден онтологияны автоматты түрде генерациялау жүйесінде алдын-ала анықталған сөздіктерге, статистикалық әдістерге және мәтіндерді өңдеу әдістеріне негізделген.

[64] жұмыста онтологияны құру үшін Уикипедия мәтіндері ұғымдар мен олардың арасындағы қатынастарды алу үшін қолданылды. Авторлар машиналық оқытудың бақыланатын әдістерін қолданды, оларды пайдалану қолмен аннотациялау мен деректерді тексерудің үлкен күш салуын қажет етеді. [65] жұмыста Text2Onto Альцгеймер ауруы онтологиясының автоматты генерациялау жүйесі келтірілген, ол алдын-ала анықталған сөздікті, байланысты мәліметтермен және машиналық оқыту алгоритмдерімен бірге морфологиялық (POS) белгілеулерді белгілер ретінде қолданады. [66] жұмыста мәтіндік құжаттар мен XML құжаттарының дерекқор схемаларының жиынтығынан онтологияны автоматты түрде генерациялау жүйесі ұсынылған. [67, 68] соңғы зерттеулерде мәтіндік құжаттар мен WordNet лингвистикалық базасына негізделген онтология құрылды. Бұл жұмыстар, алайда, қалыптасқан онтологияның сапасы білімнің бастапқы ресурсының сапасына қатты тәуелді екенін көрсетеді, сонымен қатар сөздер мен сөз тіркестерінің ықтимал семантикалық көп мағыналылықты жою

үшін сараптамалық араласуды қажет етеді. Осыған байланысты K. Meijer және әріптестерінің [69] зерттеуі қызықты болып көрінеді, онда мәтіндік корпустардан пәндік саланың таксономиясын құруға арналған құрылым жасалды. Авторлар иерархияны құру үшін түсініксіздікті жою қадамын және бағыну әдісін қолданды. Алайда, зерттеу аясы тек ұғымдар мен қатынастардың таксономиясын қоса алғанда, класс экземпляры арасындағы қатынастарды жоққа шығарды. Сонымен қатар, авторлар жасаған жүйе қарым-қатынастың дұрыс көрсетілмеуіне байланысты өте төмен семантикалық дәлдік пен толықтыққа ие.

Жүргізілген аналитикалық шолу бүгінгі күні мәтіндердің құрылымданбаған корпусын онтологиялық форматқа түрлендіру тәсілдері әлі толық әзірленбегенін және нақты бағдарламалық қамтамасыз етуге тән екенін көрсетеді. Оның үстіне, доменге тәуелсіз онтологияны генерациялау тәсілі [70] қолданылса да, ол ережелерді, үлгілерді және кейінгі өңдеуді жасау үшін интеллектуалды араласуды қажет етеді.

Онтологиялар сияқты, білім графтары (Knowledge Graphs (KG)) нысандар және олардың байланыстары туралы құрылымдық ақпаратты графикалық түрге кодтайды және оны G = (C, R) бағытталған граф ретінде көрсетеді, мұндағы C – шыңдар немесе ұғымдар жиынтығы, ал R – графтағы екі ұғым арасындағы екілік қатынастар жиынын көрсететін жиектер жиыны [68, p. 67]. Сонымен қатар, кейбір авторлар KG өзінің визуализациясы, жеделдігі және ауқымдылығы бойынша онтологиядан асып түсуі мүмкін деп мәлімдейді [70, P. 4860-4868; 71]. Құрылымданбаған мәтіндер негізінде онтологияны автоматты түрде генерациялаудың заманауи тәсілдерін талдауға сүйене отырып, біз веб-желілердің криминалистік маңызды ақпаратының құрылған корпустарын пайдалана отырып, көптілді «Құқыққа қайшы интернет-контент» онтологиясын генерациялау технологияларын әзірледік және құрылған онтология негізінде білім графын қалыптастырдық.

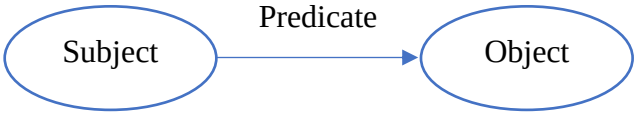
1.4 Мәтіндік корпустардан генерацияланатын онтологияның негізгі тұжырымдамаларын шығарып алу әдістерінің жай-күйі мен даму перспективалары

Онтологияның негізі деректер моделі (ДМ) ұғымы болып табылады. Resource Description Framework (RDF) негізгі онтология тілінің деректер моделі ресурс, аталған қасиет және оның мағынасын қамтитын мәлімдеме болып табылады.

RDF-де ресурстар ресурстарды анықтауға мүмкіндік беретін ресурстар идентификаторлары арқылы қамтамасыз етіледі. Ресурс кейбір идентификатор тағайындалуы мүмкін кез келген нәрсе ретінде түсініледі, ол тіпті Интернеттен тыс әлемнен ерікті нысан болуы мүмкін. Жалпы алғанда, RDF деректер моделін түйіндер әртүрлі қатынастар арқылы қосылған график ретінде қарастыруға болады.

1.1-суретте көрсетілген RDF онтологиялық графын триплет ретінде көрсетуге негізделген, онда графиктің сол жағы *Subject* және оң жағы *Object* деп аталады, олардың арасындағы қатынас субъектіден объектіге өтетін доғамен

және сәйкес *Predicate* үшін біз сөйлемдерден әрекеттерді анықтайтын сөздерді, берілген әрекетке қатысатын қатысушыларды атайтын сөздерді және оқиға атрибуттарын шығаруға мүмкіндік беретін Event Extraction (EE) тәсілін қолданамыз.



Сурет 1.1 – RDF графының триплет деректер моделінің жалпы схемасы

Талдау көрсеткендей, мәтіннен оқиғаларды шығаруға арналған EE тәсілдеріне негізделген негізгі бағдарламалық тұжырымдамаларды автоматты түрде шығарып алудың барлық қолданыстағы әдістерін төрт үлкен топқа бөлуге болатынын көрсетеді: (1) шаблонды сәйкестендіру алгоритмдерімен байланысты тәсілдер, (2) бақыланбалы машиналық оқыту әдістері, (3) терең оқыту модельдері және (4) бақыланбайтын машиналық оқыту әдістері.

Зерттеулердің бірінші тобы мәтіннен оқиғаларды шығарып алу үшін шаблонға негізделген тәсілдерді қолданады. Бұл тәсілді алғаш рет 1993 жылы E. Riloff және басқалары [72] арнайы мәтіндерден лаңкестік шабуылдарға қатысты оқиғаларды алу үшін ұсынған. Қазіргі уақытта белгілі бір салаларға бағытталған және оқиғалардың әртүрлі түрлерін шығаратын шаблонға негізделген EE жүйелері өте көп. Мысалы, TrigNER биомедициналық оқиғаларды шығару жүйелері BioNLP 2013 EE мәселелері [73] немесе Turku Event Extraction System (TEES) [74] биомедициналық мәтіндерді талдаудың әртүрлі мәселелерінде қолданылады. Көбінесе лексикалық-семантикалық шаблондарға негізделген оқиғаларды шығарып алу әдістері қаржы саласында қолданылады [75, 76].

Терроризм және криминал саласындағы соңғы EE тенденциялары біздің жобамызға үлкен қызығушылық тудырады. José A. Reyes-Ortiz [77] испан мәтіндерінен криминалдық оқиғаларды шығарып алу үшін бейнелерді тануға негізделген тәсілді енгізді. Әзірленген тәсілдің нәтижелерін бағалау үшін автор оқиғалардың нақты категориялардың белгілерін қамтитын қолмен белгіленген газет мәтіндерінің жиынтығын пайдаланды. Li Q. Zhang және т.б. [78] EE технологиясын Қытайдың заңды мәтіндеріндегі криминалдық істерді сипаттайтын мақалаларға қолданды. Мысалы, авторлар оқиғаның түрін, оқиғаның аргументтерін және ұрлық істеріндегі оқиға аргументтерінің рөлін анықтады. F. Abdelkoui [79] Араб твиттеріндегі криминалдық оқиғалардың EE-ін сипаттады. Автордың көзқарасы твиттерде пайда болатын орын атаулары мен уақыт белгілерін қамтитын әртүрлі көрсеткіштерді біріктіруге негізделген.

Оқиғаларды шығарып алу тапсырмалары бойынша соңғы жұмыс машиналық оқытуға (ML) негізделген тәсілдердің екінші тобына жатады. Бұл тәсілдер тірек векторлық әдісі (SVM), максималды энтропия (ME), жақын көршілер әдістері және т.б. сияқты дәстүрлі ML жіктеу алгоритмдерін пайдаланады. Көбінесе бұл алгоритмдер белгілер ретінде POS-тегтерді, лемматизацияланған сөздерді, триггер сөз бен тұлға арасындағы синтаксистік

тәуелділіктің түрін, сондай-ақ тәуелді сөздер мен тұлғалардың түрлерін пайдаланады.

Сонымен қатар, көбінесе ML алгоритмдеріне негізделген ЕЕ тәсілдері биомедицина, қаржы және экономика сияқты ерекше салаларда қолданылады. Кейбір авторлар ML шаблондары мен модельдеріне негізделген тәсілдерді бір уақытта қолдануды ұсынады. Pham Xuan-Huong және т.б. авторлар [80] жұмыста GENIA Event Extraction мәселесін шешуге арналған ережелерді де, ML тәсілдерін де біріктіретін гибридті тәсілді қолданатын жүйені ұсынды. Бұл зерттеуде ML сатысында N-граммдар, жиілік белгілері және тәуелділік белгілері жіктеу белгілері ретінде қолданылады. [81] жұмыста шаблондарды теңдестіру әдісі (Regularization-Based Pattern Balancing Method – (RBPB)) ұсынылады, ол оқиғаны анықтау үшін сөйлемдегі оқиға шаблондарын пайдалануды да, триггер түрін анықтау үшін SVM классификаторын да қамтиды.

Сонымен қатар, дәстүрлі бақыланатын машиналық оқыту әдістеріне негізделген жүйелер көптеген күрделі және қолмен белгіленетін белгілерді қажет етеді. Оларды қалыптастыру үшін лингвистикалық білімі бар мамандар мен домендік білімі бар сарапшылар қажет. Сонымен қатар, бұл белгілер көбінесе бір шыңды векторлармен ұсынылады, бұл деректердің сирек болуына және белгілерді таңдау мәселелеріне әкеледі [82].

Екінші жағынан, Ch.D. Manning [83] сәйкес, терең оқыту әдістері жіктеуге байланысты әртүрлі NLP тапсырмаларында, әсіресе сөйлемдерді, сөздерді немесе толық мәтіндерді жіктеуде сәтті қолданылуы мүмкін. Сондықтан, оқиғаларды шығарып алу міндеті сөйлемдер мен сөздерді жіктеу мәселесімен байланысты болғандықтан, біз жақын арада мәтіндерден оқиғаларды шығарып алу үшін терең оқыту әдістерін қолдануда прогресс күте аламыз. Соңғы бірнеше жылда ЕЕ-де үйірткілі нейрондық желілерді (convolutional neural networks – CNN) және рекуррентті нейрожелілік модельдерді пайдалану осыған байланысты. Мысалы, биомедициналық оқиғаларды шығарып алу үшін L. Li және т.б. [84] шаблондарды қолдануды қамтитын сөйлемдердің композициялық семантикалық ерекшеліктерін түсіру үшін терең оқытуды, атап айтқанда CNN моделін қолдануды ұсынды. S. Yagcioglu және т.б. [85] шулы қысқа мәтіндегі киберқауіпсіздік оқиғаларын анықтау үшін үйірткілі нейрондық желіні (CNN) және ұзақ қысқа мерзімді жады бар рекуррентті нейрондық желіні (Long short-term memory – LSTM) пайдаланды.

Кейбір зерттеушілер мәтіндерден оқиғаларды алу үшін графикалық нейрондық желілерді (graph neural networks – GNNs) пайдаланады. Мұндай GNN евклидтік емес кеңістіктерде терең білім алуға мүмкіндік беретін графикалық құрылымда жұмыс істейтін көптеген нейрондарды пайдаланады. Мысалы, [86] жұмыста авторлар зейінге негізделген графикалық үйірткілі желілерді пайдалана отырып, бірнеше оқиға триггерлері мен аргументтерін бірлесіп шығарып алуды ұсынды. [87] жұмыста авторлар үйірткілі нейрондық желіні зерттеп, оқиғаларды анықтау үшін тәуелділік ағаштарына негізделген графикті құрастырды. Олар біріктіріліп жатқан тұқыртпа векторларындағы нысан ескертулеріне негізделген біріктіру әдісін ұсынды. X. Liu және т.б. [88] нейрондық модельді оқиғаларды

шығарып алу міндеті үшін маңыздылығына сәйкес келген деректердің әр компонентіне тең емес қатынасқа бағыттау үшін нейрондық модельдерге назар аудару механизмдерін қолданды.

Алайда, іс жүзінде машиналық оқыту мен терең оқытуды қолдану әлі де үлкен қиындықтарға тап болады. Негізгі себеп – үлкен аннотацияланған корпустар негізінде модельді оқыту талабы. Әдетте, мұндай корпусты құру лингвистика саласындағы және қарастырылып отырған салалардағы көптеген кәсіби сарапшыларды тарта отырып, көп еңбекті және көп уақытты қажет ететін жұмыс болып табылады.

Белгіленген корпусты пайдалану қажеттілігін болдырмау үшін кейбір ғалымдар бақыланбайтын машиналық оқыту тәсілдерін қолданады. Бұл жағдайда оқиғаны шығарып алу міндеті Open-Domain Event Extraction (OdEE) тәсілдеріне бағытталған [89]. OdEE тәсілдері алдын-ала анықталған схемаларсыз жұмыс істейді және әдетте оқиғаны шығарып алу екі кезеңде жүзеге асырылады. Бірінші кезеңде оқиғалар сөйлемдерде немесе сөз тіркестерінде кездеседі, ал екінші кезеңде оқиғалардың кілттік сөздері немесе бақыланбайтын машиналық оқыту негізінде ұқсас оқиғалар кластерленеді. Алайда, бұл жағдайда, OdEE пайдалану кезінде мәтіннен оқиғаларды алу дәлдігі өте төмен болып шығады және оқиғалардың өзі негізінен бұлдыр және бұлыңғыр болып қалады.

Ресурстармен қамтамасыз етілмеген және аннотацияланбаған тілдерде жазылған мәтіндер үшін оқиғаларды шығарып алу міндеті одан да күрделене түседі. Осы себепті соңғы бірнеше жылда кросс-тілдік оқытуға (Cross-lingual learning – CLL) арналған EE зерттеулері пайда болды [90]. Соңғы онжылдықта біз көптілді BERT моделін [91] және үйірткілі нейрондық желіні (CNN) [92] байланыстар мен оқиғаларды тіларалық шығарып алу үшін қолдануды бақыладық. Сонымен қатар, көп жағдайда оқиғаларды кросс-лингвистикалық шығарып алу тәсілдері ML модельдерінің көптілді нұсқаларына негізделді, олар бұрын көптілді корпустарда оқытылды [93, 94], ал ресурстарға бай және жақсы түсіндірілген тіл корпустың бастапқы тілі ретінде қолданылды. Әдеттегідей бұл ағылшын тілі болды [95].

Мәтіннен оқиғаларды шығарып алу POS-тегтеу немесе Named Entities Recognition (NER) сияқты басқа NLP міндеттеріне қарағанда күрделірек міндет болып табылады. Мұның басты себебі – мәтінді өңдеудің дәстүрлі міндеттеріне қарағанда семантикалық белгілердің үлкен кеңістігінің қажеттілігі. Бұл тек бірнеше тілдегі оқиғаларды шығарып алуға арналған алтын стандарттың жалпыға қол жетімді аннотацияларының болуы мәселесін негіздейді [96].

1.5 Құрылымдалмаған мәтіндерден оқиғаларды шығарып алу data-driven және knowledge-driven тәсілдерін салыстырмалы талдау

Қазіргі қоғам бүкіл әлемде болып жатқан проблемалар мен оқиғаларды қалыптастыру және сипаттау үшін жаңалықтар сайттарын қамтитын интернет-контентті пайдаланады. Осыған байланысты, бұл контенті, соның ішінде жаңалықтар ағындарын автоматты түрде өңдеу және талдау қоғамда және бұқаралық ақпарат құралдарында қандай тақырыптар танымал екенін анықтауға

мүмкіндік беретін қоғамдық, саяси және медиа күн тәртібінің өзара әрекеттесуін зерттеу мүмкіндіктерінің бірі болып табылады. Сонымен қатар, ақпараттық күн тәртібі «Бірқатар тәуелсіз сипаттамалары бар өзекті мәселелер мен сюжеттердің жиынтығы» дегенді білдіреді [97, 98].

Жаңалықтар ағынындағы құқыққа қайшы контент туралы ақпаратты іздеуді қоса алғанда, ақпараттық күн тәртібін анықтаудың интеграцияланған технологиясын әзірлеу тіл жүйесінің әртүрлі деңгейлері элементтерінің семантикалық жақындығын (соның ішінде синонимдік коллокациялар мен толық мәтінді жаңалықтар мақалаларының семантикалық ұқсастығын) анықтау міндеті және құрылымдалмаған мәтіннен оқиғаларды шығарып алу мәселесі сияқты бағыттарды есепке алуды және әзірлеуді талап етеді. Жалпы алғанда, бұл мәселелерді шешу дистрибутивті семантиканы, тұрақты тіркестерді, грамматикалық ережелерді және бақыланатын және бақыланбайтын машиналық оқыту әдістерін қолдануға негізделген.

Сонымен қатар, құрылымдалмаған мәтіндерден оқиғаларды (Event Extraction (EE)) шығарып алу – бұл орын алған немесе болған оқиғалар туралы негізгі ақпаратты автоматты түрде шығарып алуға бағытталған онтологияны автоматты түрде құру міндетінің күрделі кезеңдерінің бірі. Құқыққа қайшы әрекеттерге қатысты мәтіндер жағдайында мұндай ақпарат, мысалы, не болды және кіммен (оқиғаға қатысушыларды анықтау), немесе оқиға қашан және қайда болды (оқиға атрибуттарын анықтау) деген сұраққа жауап болуы мүмкін. Бұл ақпарат әртүрлі көздерден, соның ішінде радио, теледидар және газет сияқты дәстүрлі бұқаралық ақпарат құралдарынан немесе әлеуметтік желілерден (Facebook, Twitter және т.б.) келуі мүмкін.

Жаңалықтар хабарламалары сияқты құрылымдалмаған деректерден оқиғаларды шығарып алу онтологияны автоматты түрде құру тәсілдерімен қатар, ақпаратты өңдеудің басқа міндеттері үшін де пайдалы болуы мүмкін. Мысалы, оқиғаларды тәуекелдерді талдау, бақылау және құқыққа қайшы әрекеттердің алдын алуға байланысты шешімдерді қолдау жүйелерінде қолдануға болады.

Жаңалықтар мәтіндерінен оқиғаларды автоматты түрде шығарып алумен байланысты әлемдік зерттеулердің болашағы зор нәтижелеріне қарамастан, бұл мәселе әлі де күрделі болып қала береді, өйткені оқиғалар әртүрлі құрылымдар мен әртүрлі компоненттерге ие. Табиғи тілдің семантикалық түсініксіздігі мәселесі де жиі шешілмейді. Барлық аталған мәселелерді шешу үшін оқиға триггерлерін, олардың түрлерін және дәлелдерін белгілеуді қамтитын үлкен аннотацияланған мәтіндік корпустар қажет. Алайда, мұндай жағдайлардың қолмен аннотациясы көп уақыт пен ақшаны қажет етеді. Демек, құқыққа қайшы әрекеттер туралы ақпаратты қамтитын жаңалықтар деректерінен тиісті оқиғаларды шығарып алудың жаңа технологияларын қолданудың шұғыл қажеттілігі туындады.

EE әдістері мен алгоритмдері құрылымдалмаған деректерді (мәтіндерді) құрылымдық нысандарға, атап айтқанда оқиғаларға түрлендіруге бағытталған. Білімге негізделген (knowledge-driven methods) әдістер бар [99, 100], олар білімді сараптамалық білімді жинақтайтын ережелерді білдіретін шаблондар арқылы

шығарып алады. Бұл әдістер алдын-ала анықталған лингвистикалық үлгілерге негізделген корпус мәтіндерінен ақпарат алуға мүмкіндік береді.

Білімге негізделген әдістерге қарағанда, деректерге негізделген әдістер (data-driven method) статистикалық машиналық оқыту тәсілдерін (иерархиялық кластерлеу [101], терең нейрондық желілер [102], бустрэп әдістері [103], қашықтан бақылау [104] және т.б.) пайдалана отырып, деректерді білімге айналдырады. Бұл тәсілдер оқиғаларды анықтаудың сандық әдістеріне негізделген және лингвистикалық құбылыстарды жақындататын модельдер жасау үшін мәтіндердің үлкен корпустарын қажет етеді.

Әдетте, деректер ағынынан статистикалық маңызды заңдылықтарды шығарып алу үшін ассоциация өлшемдері қолданылады. Байланыстың беріктігін бағалау үшін барлығы 80-нен астам өлшем бар. Олар мәтіндік корпустан алынған сөздердің пайда болу және бірлескен пайда болу статистикасы туралы болжамдар ретінде тұжырымдалған бірнеше түрлі шығарып алу принциптеріне негізделген. Математикалық тұрғыдан бұл өлшемдер сөздер арасындағы коллокацияның ассоциация дәрежесін анықтайтын формулалар түрінде көрінеді.

Жалпы алғанда, ассоциация өлшемі коллокация ассоциациясын анықтаудың үш негізгі тәсілінің негізінде анықталады: потенциалды коллоканттар арасындағы статистикалық байланысты өлшеу, коллокант кандидаттарының контекстінің сапасын өлшеу және коллокант кандидаттары мен олардың компоненттер контексттерінің айырмашылығын өлшеу арқылы [105].

Қолданылатын негізгі метрикалардың бірі – өзара ақпарат (Mutual Information (MI)), ол контекстке байланысты тәуелді жиіліктердің сөздердің тәуелсіз пайда болу жиіліктеріне қатынасы ретінде есептеледі (сөздер контексте кездейсоқ, тәуелсіз түрде пайда болған кезде). Басқа метрика – нүктелі өзара ақпарат (Pointwise Mutual Information (PMI)) бірлескен үлестірімдегі олардың тәуелсіз жеке үлестірімдерімен үйлесімділік ықтималдығы арасындағы сәйкессіздікті сандық түрде анықтайды. Мәтіндегі сөздердің өзара тәуелділігін анықтайтын тағы бір метрика – бұл коллокациялар арасындағы байланыс күшін бағалауға мүмкіндік беретін T-score [106].

Осылайша, data-driven әдістері үлкен деректерді және ПәС туралы аз білімді қажет етеді, нәтижесінде интерпретациясы төмен, ал білімге негізделген оқиғаларды (knowledge-based) шығарып алу аз деректерді қажет етеді. Сонымен қатар, статистикалық әдістер қосымша шулы деректерді шығарып алады және алыс қашықтықтағы сөздер арасындағы синтаксистік байланыстарды елемейді.

Екі тәсілдің де кемшіліктері болғандықтан, біз өз жобамызда құқыққа қайшы әрекеттерге қатысты жаңалықтар мәтіндерінің корпусынан оқиғаларды шығарып алу үшін білімге (лексика-синтаксистік шаблондар) және деректерге (машиналық оқытудың статистикалық әдістері) негізделген әдістердің комбинациясын қолданамыз.

Сондай-ақ, жеке сөздермен емес, тұтас сөз тіркестерімен көрсетілген аргументтер мен оқиғаның триггерін анықтау үшін коллокацияларды ерекшелеуге негізделу керек. Коллокация – екі немесе одан да көп

лексемалардың кездейсоқ емес синтаксистік және семантикалық тіркесімі, негізгі және тәуелді компоненттен тұратын, жеке сөздердің мағыналарының жай қосындысынан гөрі нақтырақ семантикалық ақпаратты білдіреді. Сондықтан, жаңалықтар ағынындағы мағынасы жақын коллокацияларды автоматты түрде анықтау қазіргі күн тәртібін көрсететін жалпы жаңалықтар контентін анықтауға мүмкіндік береді.

Берілген лексикалық ішкі жүйенің мүшелері арасындағы семантикалық байланыстың күшін, яғни коллокацияны анықтау үшін дистрибутивті-статистикалық әдістемені қолдану ұсынылады, оның мәні үлкен деректер ағындарында лексемалардың бірлескен пайда болуын статистикалық талдау болып табылады. Әдетте, талдау алгоритмдері [107, 108] математикалық статистика мен алгебра құралдарын пайдаланады, ал лингвистикалық ақпарат тек коллокацияның морфологиялық белгілеуінде болады.

Статистикалық көрсеткіштер немесе ассоциация өлшемдері лексикалық бірліктерге тән тұрақтылықты есептеу үшін коллокация жиіліктеріне және олардың құрамына кіретін жеке коллокаттарға (сөздерге) негізделген. Байланыс күшін бағалау үшін барлығы 80-нен астам өлшем бар [109-111]. Басқаларына қарағанда MI, PMI, t-score, log-likelihood, ықтималдық коэффициенті, Пирсон квадраты – критерийі және басқалары сияқты өлшемдер жиі қолданылады.

Дегенмен, статистикалық әдістер шығарып алынған деректерді шулы етеді және ұзақ қашықтықтағы сөздер арасындағы синтаксистік байланыстарды елемейді. Сонымен қатар, коллокацияны анықтау морфологиялық және синтаксистік құралдарды қолданудың статистикалық талдау модельдеріне қосымша қажет. Бұл жағдайда семантикалық жақын коллокацияларды анықтау коллокация компоненттерінің бірлескен пайда болу ықтималдығын ескеріп қана қоймай, негізгі және тәуелді компоненттер арасындағы грамматикалық тәуелділіктерді ресімдеуге мүмкіндік береді [112].

Жалпы, сөздер арасындағы семантикалық байланыстарды анықтау мәселесін шешудегі дистрибутивті семантиканың статистикалық әдістері мен модельдерін талдау статистикалық өлшем (атап айтқанда, тәуелді контексттік байланысты жиіліктерді тәуелсіздермен салыстыратын өзара ақпарат коэффициенті) мен коллокация компоненттері арасындағы қатынасты сипаттайтын ақырлы предикаттар алгебрасына негізделген семантикалық ұқсастықтың дамыған моделінің үйлесімі ең нәтижелі екенін көрсетті [113].

1.6 Құқыққа қайшы әрекетке байланысты мәтіндік деректерді іздеу және талдау үшін Text Mining құралын пайдалану мүмкіндіктеріне аналитикалық шолу

Соңғы онжылдықта құқыққа қайшы әрекеттерді талдау үшін Text Mining тәсілдерін қолдануға қатысты көптеген зерттеулер пайда болды. Барлық қол жетімді тәсілдерді келесі алты бағытқа бөлуге болады:

1. Қылмыстық әрекетке байланысты мәтіндерді сәйкестендіруді білдіретін қылмыстарды анықтау.

2. Осы мәтінмен байланысты құқыққа қайшы оқиға түрлерінің жіктелуі, бұл мәтінмен байланысты қылмыстардың әртүрлі түрлерін анықтау.

3. Ықтимал қылмыстарды болжауға болатын қылмыстарға қатысты мәтін шаблондарын әзірлеуді білдіретін қылмыстарды болжау үшін мәтін шаблондарын жасау.

4. Жеккөрушілік және/немесе кемсітушілікке байланысты сөйлемдері бар мәтіндерді анықтауды білдіретін жеккөрушілік тілін анықтау.

5. Құқыққа қайшы әрекеттерге байланысты ақпарат шығарып алу. Бұл бағыт жасалған немесе ықтимал қылмысқа байланысты нақты нысандарды анықтауды қамтиды.

6. Криминалистік маңызды оқиғалар туралы ақпаратты шығарып алуды қамтитын қылмыстық маңызы бар оқиғаларды (Crime Related Event, CRE) шығарып алу.

Зерттеулердің көп бөлігі құқыққа қайшы әрекеттермен байланысты болуы мүмкін мәтіндік ақпаратты іздеу және анықтау бағытына бағытталған. Бұл бағыттағы жұмыстар әдетте кластерлеу тәсілдерін және тақырыптық модельдеуді қолданады [114, 115]. Соңғы жылдары көптеген зерттеушілер құқыққа қайшы әрекеттерге қатысты мәтіндерді анықтаудан басқа, криминал немесе қылмыспен байланысты құқыққа қайшы оқиғаның түрін анықтауға баса назар аударды [116-118].

[118, p. 120-131]-жұмыста аннотацияланған Corpus Anotado de Delitos мамандандырылған корпусының бір мың жаңалықтарында шабуыл, кісі өлтіру, ұрлау және жыныстық зорлық-зомбылық сияқты қылмыстардың түрлерін анықтады. Сәл кейінірек Salas әріптестерімен бірге [116, p. 725-740] испан тілді корпуста криминалдық жаңалықтар түрлерін жіктеу үшін SVM және нейрондық желілер қолданылды. Жақында Santhiya [117, p. 2133-2151] контекстке негізделген шешімдерді қолдау жүйесін пайдалана отырып, ағылшын твиттерін жыныстық құдалау, зорлау, өзін-өзі өлтіру және т.б. қылмыстардың әртүрлі түрлеріне біріктірді. Гибридті өңдеу үшін мәтінді өңдеу тәсілдерімен бірге машиналық оқыту әдістері қолданылды. Сонымен бірге, Hassani және әріптестері [119] өз талдауында қылмыспен байланысты мәтіндерді жіктеу үшін көбінесе тірек векторлық әдістер мен нейрондық желілер қолданылатынын байқады.

Зерттеулердің айтарлықтай аз болуы қылмыстарды болжау үшін мәтіндік шаблондарды әзірлеуге байланысты. Бұл бағыттағы жұмыстар көбінесе полиция құжаттарында тіркелген қылмыстардың сыртқы деректері мен атрибуттарын пайдаланады. Мысалы, P. Chen [120] жұмысында уақытты, жағдайды және әрекет ету тәсілін слоттар ретінде қолдана отырып, ықтимал сериялық қылмыстардың модельдері қалыптастырды. N. Joseph [121] ықтимал қылмыстық аймақтардың шаблондарын және қылмыстың өзекті түрлерін жасау үшін K-Nearest Neighbors Algorithm алгоритмін қолданды. Белгілер ретінде US Arrests dataset-тен алынған уақытша тәуелділік қылмыстарының әртүрлі түрлері бойынша аймақтағы қамауға алу саны туралы ақпарат пайдаланылды. Сол сияқты, Y.L. Lin және әріптестері [122] қылмысты болжаудың графикалық

моделін құру үшін уақыт, қылмыс түрі және географиялық сипаттамалар белгілер ретінде пайдаланылды.

Бүгінгі таңда көптеген сәтті зерттеулер интернет желілерінде, блогтарда және твиттерде жеккөрушілік тілін табу бағытымен байланысты [123-129]. Hate Speech ұғымы нәсіліне, түсіне, этникалық тегіне, жынысына, жыныстық бағдарына, ұлтына, дініне және басқа белгілеріне негізделген адамның қадір-қасиетін төмендететін мәтіндерді қамтитын әлеуметтік веб-медианың [130] қорлайтын пайдаланушы контентін қамтиды [131]. Жеккөрушілік тілін анықтау бағыты кең таралған, оның маңыздылығы CLEF 2021 конференциясының PAN21 байқауына осы мәселенің қосылғанын көрсетеді, онда жеккөрушілік тілін анықтау мәселесін шешудің алпыс алтыдан астам алгоритмдері бағаланды [122, p. 298-1-298-15]. Көп жағдайда бақыланатын машиналық оқыту [126, p. 9503413-19503413-8; 127, p. 233-237; 128, p. 1145-1154] мен рекурренті және үйірткілі нейрондық желілер [125, p. 353-370; 126, p. 9503413-1-9503413-8], сондай-ақ трансформерлі модельдер және BERT моделі [128, p. 1145-1153; 129, p. 928-939] қолданылды.

Әдетте, мәселені шешу үшін құқыққа қайшы әрекеттерге қатысты ақпаратты шығарып алу, полиция есептерінен немесе куәгерлердің сипаттамаларынан білім тарылады [132, 133]. C.H. Ku және әріптестері сөз қоры және грамматикалық ережелер мен шаблондарды қолданды. Karystianis және әріптестері [81, p. 1224-1233] NLP алгоритмдерін Оңтүстік Уэльстегі тұрмыстық зорлық-зомбылыққа байланысты оқиғалардың корпусында зорлық-зомбылық түрлері мен зардап шеккендердің жарақаттарын анықтау үшін қолданды. P. Das және A. K. Das [133, p. 55-75] АҚШ, БАӘ және Үндістаннан келген полиция есептерін талдай отырып, қылмыстар туралы хабарлайтын сөйлемдерді аннотациялау үшін графикалық кластерлеуді қолданды.

[134, 135] еңбектерінде ашық көздерден құқыққа қайшы әрекеттер туралы хабарлайтын мәтіндер талданды. P. Das және әріптестері [134, p. 101269-101281] графикалық кластерлеу әдістерін қолдана отырып, Үндістан штаттарының газет мақалаларының мәтіндер корпусының субъектілері арасындағы байланысты зерттеді. T. Dasgupta әріптестерімен [135, p. 541-548] есептеу лингвистикасының әдістерін қылмыскерлер мен олардың құрбандарының аттарын, қылмыс түрін, оның географиялық орналасуын, жасалу уақытын, ықтималдық классификаторын және домендік онтологияны қолдана отырып, нысандарды алу дәлдігін арттыру үшін қолданды. Жеке зерттеулер газет мақалалары мен жаңалықтарынан басқа полиция есептерін, құрбандар мен қылмыс куәгерлерінің есептерін қолданды. Алайда, ең сәтті нәтижелер тар шеңберде мамандандырылған мәтіндерді, атап айтқанда полиция мен куәгерлердің есептерін талдау кезінде алынды [136].

Құқыққа қайшы әрекеттерді талдауда Text Mining тәсілдерін қолданудың ең қиын бағыты криминалға қатысатын нысандар мен осы нысандар арасындағы қатынастар туралы ақпаратты алуды қамтитын криминалистік маңызды оқиғаларды (CRE) шығарып алу болып табылады [137-140]. F. Rahma мен A. Romadhony [137, p. 10-14] Индонезия тіліндегі мәтіндерде жәбірленушіні,

қылмыскерді, қылмыстың орнын, түрін және уақытын табу үшін онтология мен ережелерге негізделген әдістерді қолданды. J.G. Joseph пен әріптестері [138, p. 516-519] жаңалықтардан есірткіге қатысты қылмыстардың мәнін, атап айтқанда, қылмыс жасалған жерді, сондай-ақ қылмысқа қатысатын есірткінің түрі мен мөлшерін анықтады. Олар стандартты NLP конвейерінен NER әдістерін қолданды. Өз кезегінде, P. Das және A. K. Das [140, p. 1-4] әйелдерге бағытталған зорлық-зомбылық туралы ақпаратты жариялайтын жаңалықтар веб-сайттарының мәтіндерінен штаттардың, көшелердің, қалалардың, ауылдардың атауларын, сондай-ақ олардың гендерлік жіктелуін жүзеге асыратын фамилияларды шығарды.

Оқиғаны құқыққа қайшы әрекетке байланысты мәтіндерден шығарып алу мәселесін шешудің алғашқы әрекеттері DARPA Message Understanding Conferences (MUCs) алғашқы конференцияларында жүзеге асырылды. Алайда, Crime Related Event Extraction (CREE) мәселесі әлі де өзекті болып табылады. Сонымен қатар, MUC-3 және MUC-4 конференцияларында Латын Америкасындағы терроризм туралы мәтіндерге назар аударылды [141] және алынған оқиғалар нақты террористік актілермен байланысты болды. Кейінгі зерттеулер құқыққа қайшы әрекеттерге байланысты оқиғалардың басқа түрлерін анықтады. Бұл ретте, әдетте, терроризм, киберқылмыс, тұлғаға қарсы немесе көлікпен байланысты қылмыстар сияқты әртүрлі қылмыстар бір-бірінен оқшауланып қаралады. Мысалы, [138, p. 516-519; 142] зерттеулер есірткі қылмысы мен нашақорлыққа байланысты оқиғаларды шығарып алуды қарастырды. K.R. Rahem мен N. Omar [142, p. 250-253] есірткі сатушылардың ұлтына, атауына, нарықтық бағасына және есірткі санына қатысты деректерді шығарып алды. Басым көпшілігінде CREE-ге қатысты зерттеулер сыртқы деректерді қамтитын грамматикалық және эвристикалық ережелер мен шаблондарға негізделген, мысалы, Малайзия ұлттық агенттігінің (BERNAMA) [142, p. 250-253] ақпарат жұмысында.

S. Yagcioglu және әріптестерініің зерттеуі [143] нөлдік күндік эксплуатациялар, бопсалаушы бағдарламалары, деректердің шығып кетуі, қауіпсіздіктің бұзылуы және осалдық сияқты киберқауіпсіздікке қатысты оқиғаларды анықтауға бағытталған. X. Wang пен әріптестердің жұмысы [144] оқиғаларды твиттерден жол-көлік оқиғалары туралы мәтіндермен шығаруды қарастырады. Зерттеу ықтимал криминалдық жол-көлік оқиғаларын болжауға бағытталған. K.T. Hossain мен әріптестердің жұмысы [114, p. 269-282] ықтимал зорлық-зомбылыққа байланысты ықтимал оқиғаларды болжауға бағытталған. Мұндай оқиғалар MANSA оқиғалары деп аталатын әскери әрекеттер немесе террористік шабуылдар болуы мүмкін. Әдісті бағалау қалалар мен елдердің географиялық орындарында болған оқиғалар туралы қолмен алынған құрылымдық есептерді қолдануға негізделген. Зерттеулердің үлкен тобы FBI's UCR Program анықтамасы бойынша нәсіліне, дініне, мүгедектігіне, жыныстық бағдарына, этникалық тегіне, жынысына немесе гендерлік сәйкестігіне қарсы қылмыстық құқық бұзушылық болып табылатын жеккөрушілік қылмыс оқиғаларын шығарумен байланысты. A. M. Davani мен әріптестерінің жұмысы

[145] кісі өлтіру мен ұрлауға байланысты оқиғалардың жаңалықтар мақалаларынан табылуы мен шығарып алынуын сипаттайды. Авторлар оқиғаның түрін анықтайды және оның мақсаты мен қылмыс түрі сияқты атрибуттарын шығарып алады. Алайда, олардың эксперименттері үшін олар, бұрын айтылған авторларға ұқсас, жергілікті жаңалықтар мақалаларының мәтінінің белгіленбеген корпусының ішкі жиынын қолмен аннотациялады.

Көбінесе CREE-ге қатысты зерттеулер қылмыстарды болжауды жақсарту үшін алынған оқиғаларды пайдаланады. S. Han және оның әріптестері [146] New York Times газетінен жеккөрушілік оқиғаларды шығарып алуға, штаттық және жалпы мемлекеттік деңгейдегі қылмыстың өзгеруінің жалпы тенденцияларын анықтау үшін алынған оқиғаларды қолдануға назар аударды. Олар өз зерттеулерінде deep learning әдістерін қолданды.

1.1-кестеде зерттеу мысалдары мен қолданылатын әдістермен құқыққа қайшы әрекеттерді талдау үшін Text Mining тәсілдерін қолдану бағыттары туралы жиынтық ақпарат берілген.

Осылайша, жүргізілген әдеби дереккөздерді талдау, заңсыз әрекеттерге байланысты мәтіндерді іздеудің және осы мәтіндерден құрылымдық ақпаратты алудың ең көп қолданылатын тәсілдері машиналық оқыту және терең оқыту әдістері болып табылатындығын растайды, іздеу мен талдаудың дәлдігін жақсарту үшін қосымша білім ретінде [125, p. 353-370; 126, p. 9503413-9503413-8; 127, p. 233-237; 146] қолданылатын онтологияларды қосады. Мұндай әдістердің басты кемшілігі – нақты ережелерді қолдануға негізделген және тілдік ерекшеліктерді ескеретін кәсіби сарапшылар үшін көп уақытты қажет ететін, аннотацияланған теңдестірілген корпустардың қажеттілігі [126, p. 9503413-9503413-8; 128, p. 1145-1153; 147].

Кесте 1.1 – Құқыққа қайшы қызметке қолданылатын Text Mining бағыттарының жиынтық кестесі

Text Mining пайдалану бағыттары	Зерттеу мысалдары	Қолданылатын әдістер	Қолданылатын корпустар	Тиімділік
1	2	3	4	5
Қылмыстық әрекетке байланысты мәтіндерді сәйкестендіру	Salas A.H. және басқ., Santhiya K. және басқ.	Кластерлеу әдістері (Grid-based, constraint-based, k-means cluste ring algorithm және т.б.)	Әлеуметтік желі мәтіндік корпусы	F_1 >= 87%
Құқыққа қайшы оқиға түрлерінің жіктелуі	Campos D. және басқ., Third Message Understanding Conference (MUC-3), Han S. және басқ., Mullah N.S. және басқ.	Машиналық оқыту әдістерімен жіктеу (тірек векторлар әдісі, нейрондық желілер және т.б.)	Annotated Crimes Corpus (Corpus Anotado de Delitos), Spanish corpus of Peruvian news, English tweets dataset, Mexico	77,9% <= F_1 <= 84%
Қылмыстарды болжау үшін мәтіндік шаблондарды құру	Nockleby J.T. және басқ., Ku C.H. және басқ., Das P. және басқ.	Жіктеу және кластерлеу әдістері + уақытша және кеңістіктік сипаттамалар	Полиция есептері мәтіндерінің корпустары	Precision <= 50%, Recall<=16%
Құқыққа қайшы оқиғаларға байланысты нысандарды шығарып алу	Sha L. және басқ., Das P. және басқ., Ku C.H. және басқ., Joseph J.G. және басқ., Sotomayor M., және басқ., Third Message Understanding Conference (MUC-3), Wang X., және басқ., Davani A.M. және басқ., Das P. және басқ.	Ережелерге негізделген тілдік сөйлемшелердің шаблондары+ лексикологиялық қорлар. Графикалық кластерлеу әдістері	Полицияның қылмыс туралы есептері мен куәгерлердің айғақтарының мамандандырылған мәліметтер қорлары. Құқыққа қайшы әрекеттер туралы веб-жаңалықтар мақалалары, испан және ағылшын тілдеріндегі қылмыстық оқиғалар туралы кеңістіктік-уақыттық белгіленген твиттер және т.б.	Полиция есептерінен мамандандырылған нысандарды алу (мысалы, қарулар) Precision <= 96%, Recall >= 90%. Қоғамдық жаңалықтар мақалаларынан, әлеуметтік желілерден немесе твиттерден мамандандырылған нысандарды алу: 61% <=F_1<= 71%

1.1-кестенің жалғасы

1	2	3	4	5
Жек көрушілік тілін анықтау	Hassani H. және басқ., Joseph N., Lin Y.L. және басқ., Rangel F. және басқ., Siino M. және басқ., Qureshi K.A. және басқ., Alfina I. және басқ., Schmidt A. және басқ.	Мұғаліммен машиналық оқыту әдістерімен жіктеу, рекуррентті және үйірткілі нейрондық желілер, лексикалық қорларды қосымша қолдана отырып, BERT моделі	Аннотацияланған әлеуметтік желі, твиттер мен блогтар мәтіндері. Корпус тілдері: ағылшын, испан, голланд, итальян, португал, араб және басқа тілдер.	75% <= precision <= 84 F_1<= 90% твиттер үшін
Құқыққа қайшы оқиғаларды шығарып алу (CRE)	Қарастырылып отырған құқыққа қайшы әрекет түрі: - есірткі қылмысы (Rahma F. және басқ., Joseph J.G. және басқ.). Киберқауіпсіздік (Yagcioglu S. және басқ.). жол-көлік қылмыстары (Karimi M. және басқ.). соғыс немесе террористік актілер оқиғалары (Moreno-Jimenez L-G. және басқ.). жеккөрушілік қылмыстары (Hossain K.T. және басқ., Mozafari M. және басқ.)	Machine Learning және Deep Learning грамматикалық және эвристикалық ережелерін қолдануға негізделген тәсілдер, қайталанатын және үйірткілі нейрондық желілер, НММ.	Веб-жаңалықтар мақалаларының қолмен аннотацияланған корпустары; уақытша және жергілікті сипаттамалары бар полиция есептері мен куәгерлердің сауалнамаларының дерекқорлары. Аннотацияланған твиттер; ұлттық ақпарат агенттіктерінің мәліметтер қорлары. Мысалы, BERNAMA – Малайзияның Ұлттық ақпарат агенттігі.	60% (киберқауіпсіздік оқиғалары үшін <=F_1<= 64% (жеккөрушілік қылмыстарымен байланысты оқиғалар). Precision <= 83 жақсы аннотацияланған оқу корпусы үшін.
Ескерту – Әдебиет негізінде құралған [73, p. 1-12; 81, p. 1224-1233; 114, p. 269-282; 115, p. 61-81; 116, p. 725-740; 117, p. 2133-2150; 118, p. 120-131; 119, p. 139-153; 121; 122, p. 298-1-298-15; 123, p. 1-17; 124, p. 1-10; 126, p. 9503413-1-9503413-8; 127, p. 233-237; 129, p. 928-939; 130, p. 1-9; 131, p. 1277-1278; 132, p. 193-197; 133, p. 55-75; 134, p. 101269-101281; 136, p. 162-169; 137, p. 10-14; 138, p. 516-519; 139, p. 1-4; 140, p. 1-4; 141;, 142, p. 250-243; 143, p. 1366-1371; 144, p. 231-237; 145, p. 5753-5756; 146; 147, p. 88364-88375]				

1-бөлімнің қорытындысы

Бұл тарауда көптілді құқыққа қайшы интернет-контентті автоматты түрде іздеу және талдау саласындағы бар мәселелерге аналитикалық шолу берілген, ол қазіргі уақытта құқыққа қайшы контентті іздеуге және талдауға бағытталған зерттеулер жеткілікті болғанымен, негізінен қолданыстағы әзірлемелер ағылшын, қытай, француз және басқа да еуропалық тілдердің мәтіндерін өңдеуді қамтиды.

Семантикалық талдау және семантикалық корпусты белгілеу саласындағы негізгі мәселелер сипатталған. Сондай-ақ, лингвистикалық онтологияларды қолдану және пішімдеу мәселелеріне шолу жасалды.

Онтологияны автоматты түрде генерациялауда қолданылатын негізгі ұғымдарды мәтіндерден шығарып алу әдістерінің күйі мен даму перспективалары қарастырылады. Құрылымдалмаған мәтіндерден data-driven және knowledge-driven оқиғаларды шығарып алу тәсілдерінің салыстырмалы талдауы берілген. Бұл тәсілдер оқиғаларды анықтаудың сандық әдістеріне негізделген және лингвистикалық құбылыстарды жақындататын модельдер жасау үшін мәтіндердің үлкен корпустарын қажет етеді.

Корпустарды талдау және семантикалық белгілеу үшін Text Mining әдістері мен құралдарын қолданудың бар мүмкіндіктеріне шолу жасалды. Талдау негізінде тар шеңберде мамандандырылған тақырыптағы құрылымдалмаған ақпараттың контенті негізінде онтологияны автоматты түрде құруға жалпы тәсілдеме жасалды.

2 КРИМИНАЛДЫҚ ЛЕКСИКАНЫҢ МӘТІНДІК КОРПУСТАРЫ МЕН ТЕРМИНОЛОГИЯЛЫҚ ТЕЗАУРУСЫ НЕГІЗІНДЕ КӨПТІЛДІ ОНТОЛОГИЯНЫ ӘЗІРЛЕУ

2.1 Әртүрлі деректер көздерінен криминалистік маңызды мәтіндердің мәтіндік корпустарын кеңейту және толықтыру

Әлбетте, тар шеңберде мамандандырылған онтологияның автоматты және автоматтандырылған генерациясы пәнге бағытталған мәтіндік корпустарға негізделуі керек. Бұл зерттеуде құрылған көптілді онтологияға арналған мәтіндерден нақты лексикалық ресурстарды шығарып алу үшін криминалдық тақырыптарға арналған екі корпус қолданылды.

Бірінші көптілді корпусқа орыс, украин және ағылшын тілдерінің мәтіндері кіреді. Оны толтыру үшін ақпарат 2018 жылдың маусымынан 2020 жылдың қазанына дейін Python BeautifulSoup кітапханасының талдаушысын пайдалана отырып, жаңалықтар интернет-сайттарынан алынды. Корпустағы әрбір мәтін криминалдық тақырыппен байланысты және Ukr_texts, Eng_tests, Ru_tests үш каталогының біріне орналастырылған.

Украин тіліндегі мәтіндер автоматты түрде «Украинская правда» ресми сайтынан, сондай-ақ «Главком» сайтынан жүктелді. Украинаның ішкі корпусында 3147 мәтін бар.

Орыс тіліндегі мәтіндер «Редпост» жаңалықтар сайтынан, Харьковтың қоғамдық-саяси аймақтық басылымынан, атап айтқанда «Преступность и происшествия» бөлімінен алынды. Корпустың бұл бөлігінде 5506 мәтін бар. Ағылшын тіліндегі мәтіндер Корпус Кристи, Техас штатындағы «Caller Times» газетінің «Crime» бөлімінен алынған. Қазіргі уақытта бұл ішкі корпуста 300 мәтін бар.

Онтологияны генерациялау үшін қор ретінде пайдаланылатын екінші көптілді корпус – үш жылдан астам уақыт бойы дамып келе жатқан параллель қазақ-орыс корпусы [1, p. 116-124]. Осыған байланысты мәтіндердің сапалы параллельді көптілді корпустарын құру қазіргі лингвистиканың ең өзекті және прогрессивті бағыттарының бірі екенін атап өткен жөн.

Осы параллель корпусты кеңейту және толықтыру 2021 жылдың наурыз-қараша айлары аралығында Қазақстанның zakon.kz, caravan.kz, lenta.kz, nur.kz ақпараттық интернет-кеңістігінің төрт ақпараттық сайттарын және Жетісу облысының Полиция департаментінің https://www.gov.kz/memleket/entities/mvd-zhetysu сайтының «Криминалдық полиция» қосымша бетін талдау арқылы жүзеге асырылды. Бұл сайттарда қылмыстық ақпаратқа, соның ішінде тонау, кісі өлтіру, жол-көлік оқиғалары және т. б. сияқты криминалға қатысты көптеген жаңалықтар мақалалары бар. Қазіргі уақытта параллель қазақ-орыс корпусының көлемі орыс тіліндегі 3000 мәтінді және қазақ тіліндегі 3000 мәтінді, оның ішінде мағынасы бойынша тураланған қазақ-орыс сөйлемдерін қамтитын 2000 мәтінді құрайды.

Мұндай туралау үшін аудармалар сөздігіне негізделген мәтінді автоматты түрде туралау қосымшасы қолданылды [1, p. 116-124]. Параллель корпуста

сөйлемдердің семантикалық тураланuының дұрыстығын бағалауды мамандар-лингвистер, қазақ және орыс тілдерін меңгерушілер жүзеге асырды. Сарапшыларға қазақ және орыс тілдерінде автоматты түрде жұптасып тураланған сөйлемдер ұсынылды. Әрбір сарапшыға әр сөйлемді шкала бойынша туралаудың дұрыстығын бағалау қажет болды: 0 – автоматты туралау дұрыс орындалмады, 1 – сөйлемдерді автоматты түрде туралау дұрыс орындалды. 2.1-суретте параллель қазақ-орыс корпусын автоматты туралау процесін сараптамалық бағалау нәтижесінің үзіндісі көрсетілген.

	Sentence_ID	Ru	Kz	Resul	Expert 1	Expert 2
2	1_zakon_20.07.2018_ru_raw.01	Глава государства поручил Касымову и Кожамжарову взять на контроль дело Дениса Тена.	Мемлекет басшысы Қасымов пен Қожамжаровқа Денис Теннің ісін бақылауға алуды тапсырды.	=	1	1
3	1_zakon_20.07.2018_ru_raw.02	Руководству Администрации Президента было поручено держать	Президент Әкімшілігінің Басшылығына тергеу барысын үнемі бақылауда ұстау	=	1	1
4	1_zakon_20.07.2018_ru_raw.03	генеральному прокурору Кайрату Кожамжарову и министру внутренних	қызметінің ақпаратына сүйене отырып хабарлауы бойынша, Мемлекет басшысы	=	1	1
5	1_zakon_20.07.2018_ru_raw.04	Президента было поручено держать ход расследования на постоянном	тергеу барысын үнемі бақылауда ұстау тапсырылды.	=	1	1
6	1_zakon_20.07.2018_ru_raw.05	Для расследования уголовного дела создана следственно-оперативная группа из числа наиболее опытных	Қылмыстық іс бойынша тергеу жүргізу үшін Алматы қаласы ІІМ және ІІД тәжірибелі қызметкерлерінен	=	1	1
7	1_zakon_20.07.2018_ru_raw.06	Убийцам Дениса Тена грозит пожизненное заключение.	бостандығынан айыру жазасы берілуі мүмкін.	≠	0	0
8	2_zakon_20.07.2018_ru_raw.01	За совершение убийства разыскивается Кудайбергенов Арман Бурибаевич.	Кісі өлтіргені үшін Құдайбергенов Арман Бөрібаев іздестірілуде.	≠	0	0
9	2_zakon_20.07.2018_ru_raw.02	МВД РК20 июля 2018, 11:00 Фотографию второго подозреваемого в убийстве Дениса Тена распространило	ҚР ІІМ 2018 жыл, 20 шілде 11:00 Zakon.kz ақпарат көзінің хабарлауы бойынша, ҚР ІІМ Денис Теннің өліміне	=	1	1
10	2_zakon_20.07.2018_ru_raw.03	За совершение убийства разыскивается Кудайбергенов Арман Бурибаевич, 1994 года рождения, уроженец	Кісі өлтіргені үшін Қызылорда облысының тумасы - 1994 жылғы Құдайбергенов Арман Бөрібаев іздестірілуде.	=	1	1

Сурет 2.1 – Криминалистік маңызды мәтіндердің параллель қазақ-орыс корпусының сөйлемдерін автоматты түрде туралау процесін сараптамалық бағалау нәтижесінің үзіндісі

Сарапшылардың пікірлерінің сәйкес келу дәрежесі Коэннің каппа статистикалық коэффициентін қолдану арқылы анықталды. Автоматты туралау нәтижесімен сарапшылардың пікірлерінің келісушілік коэффициенті 1-ге жақын (agreement ≈ 0.98). Коэффициенттің бұл мәні әзірленген қазақ-орыс мәтінінің сөйлемдерін мағынасы бойынша тураланған деп есептеуге мүмкіндік береді.

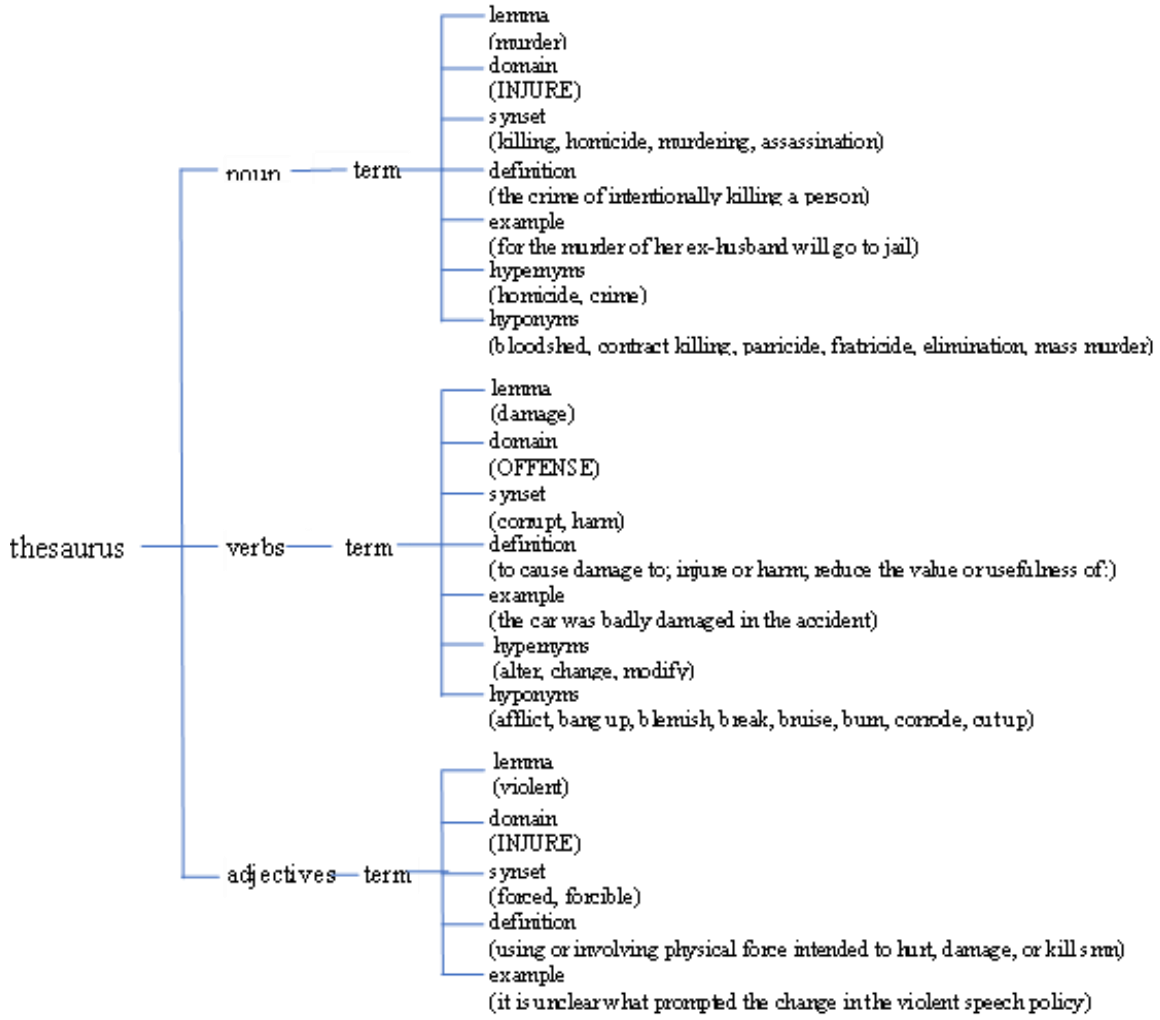
2.2 Көптілді терминологиялық тезаурус құру

«Құқыққа қайшы интернет-контент» онтологиясын генерациялаудың негізгі көздері криминалдық жаңалықтарды қамтитын мәтіндерден тұратын әзірленген параллель қазақ-орыс корпусы [1, p. 116-124] және XML форматында әзірленген криминалдық лексиканың көптілді терминологиялық тезаурусы [148] болып табылады.

Тезаурустың негізгі лексикасы ағылшын, украин, қазақ және орыс тілдеріндегі криминалдық тақырыптағы мәтіндерден қолмен алынды. Жеті негізгі тақырыптық категория бөлінді: «*Movement*» (Жол қозғалысы), «*Traffic Accident*» (Жол-көлік оқиғасы), «*Injure*» (Зиян келтіру), «*Offense*» (Құқық бұзушылық), «*Arrest*» (Тұтқынға алу), «*Trial*» (Соттың істі қарауы) және «*Police Department*» (Полиция қызметі), бұл тезаурусты тар тақырыптық етуге мүмкіндік

берді. Кластардың бұл таңдауы тезаурусты толтыру үшін қолданылатын ақпараттық ресурстардың тақырыбына байланысты. Зерттелген қылмыстық жаңалықтар үш криминалдық бағытқа қатысты: «*Police*» (Полиция), «*Transfer*» (Қозғалыс), «*Crime*» (Криминал) және олардың жоғарыда аталған ішкі түрлері.

Тезаурус сөз таптарының зат есімдерін, етістіктерін және сын есімдерін қамтиды. 2.2-суретте әзірленген тезаурустың құрылымдық схемасы көрсетілген, оның XML құжаты үш негізгі элементті қамтиды: <*nouns*>, <*verbs*> және <*adjectives*>, олар өз кезегінде <*term*> еншілес элементтерді қамтиды. Әрбір <*term*> элементі сөздің берілген табының сөзін XML еншілес элементтерімен ұсынылған оның синонимикалық қатарымен *(synsets)*, анықтамасымен *(definition)*, мысалымен *(example)*, гипонимдерімен *(hyponym)* және гиперонимдерімен *(hypernyms)* білдіреді. Сөздіктің <*domain*> элементі қылмыс пен құқыққа қайшы әрекеттерге қатысты жоғарыда аталған жеті тақырыптық категорияның бірін білдіреді. Бұл тақырыптық категориялар, яғни оқиғалардың түрлері мен ішкі түрлері 3.1-бөлімде, ал оқиғалар аргументтерінің рөлін анықтаудың ақпараттық моделі 3.2-бөлімде толық сипатталады.



Сурет 2.2 – Құқыққа қайшы бағыттағы лексиканың көптілді тезаурусының құрылымдық схемасы

2.3-суретте қазіргі уақытта 600-ден астам негізгі сөздерді (330 зат есім, 107 сын есім және 170-ке жуық етістік) және 2500-ден астам негізгі сөз синонимдерін қамтитын тезаурустың үзіндісі көрсетілген. Болашақта тезаурусты биграммаларды қолдану арқылы кеңейту жоспарлануда [149].

```
<vocabulary>
    <nouns>
        <term id="1">
            <lemma lang="ru">стрельба</lemma>
            <domain>OFFENSE</domain>
            <synset lang="ru">обстрел, выстрел</synset>
            <definition lang="ru">учебные занятия по ведению огня из различных видов оружия;  ведение огня,
применение огнестрельного оружия для выполнения поставленной задачи (Пулевая с., стендовая с., с. из
пистолета, с. из лука)</definition>
            <example lang="ru">Два человека получили ранения при стрельбе в Таразе</example>
            <hypernims lang="ru">['приведение в действие', 'движение']</hypernims>
            <hyponims lang="ru">['контрвыстрел', 'разряд', 'отстрел', 'выстрел', 'выстрел из пистолета',
'выстрел в голову', 'выстрел из снаряда', 'перестрелка']</hyponims>
            <lemma lang="en">shooting</lemma>
            <synset lang="en">firing, fire, gunfire</synset>
            <definition lang="en">the act of firing a projectile</definition>
            <example lang="en">his shooting was slow but accurate</example>
            <hypernims lang="en">['actuation', 'propulsion']</hypernims>
            <hyponims lang="en">['countershot', 'discharge', 'firing', 'firing off', 'gunfire', 'gunshot',
'headshot', 'potshot', 'shellfire', 'shoot']</hyponims>
            <lemma lang="ka">атыс</lemma>
            <synset lang="ka">ату, оқ жаудыру, атылыс</synset>
            <definition lang="ka">оқ атылғанда шығатын дыбыс, тарсыл; көздесу, нысанаға алып, оқ тигізу
</definition>
            <example lang="ka">Алматының Ақбұлақ мөлтекауданында атыс болып, бес адам қаза тапты</example>
            <hypernims lang="ka">['іске қосу','қозғаушы күш']</hypernims>
            <hyponims lang="ka">['қарсы атыс', 'ату', 'басына ату', 'снарядтан ату', 'атыспақ', 'атысу',
'атысып қалу']</hyponims>
            <lemma lang="ua">стрілянина</lemma>
            <synset lang="ua">стрільба, пальба</synset>
        </term>
```

Сурет 2.3 – Криминалдық лексиканың көптілді тезаурусының үзіндісі

Әзірленген көптілді тезаурусты пайдалану мен толтыруды жеңілдету үшін XML файлының өзін ашпай-ақ сөздікке жаңа терминдерді жылдам қосуға және іздеуге мүмкіндік беретін арнайы қосымша әзірленді.

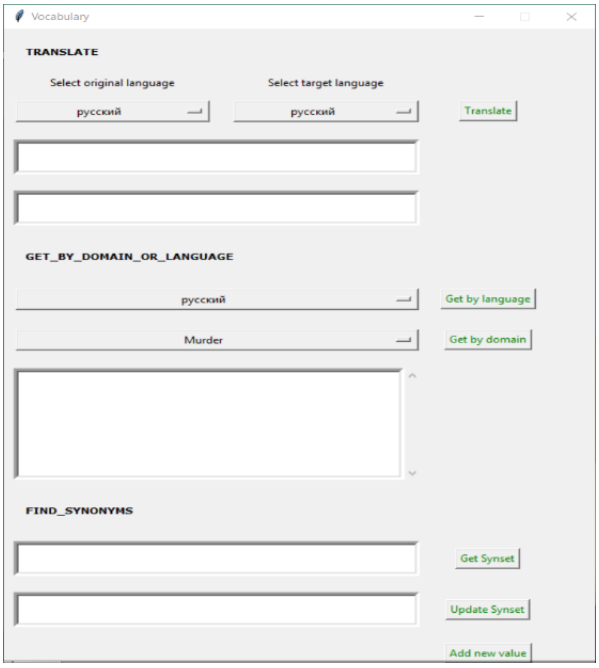
Қосымша:

1) терминдерді енгізу және өңдеу тілін өзгертуге;

2) сөз таптарының үшеуінің біріне жаңа сөз қосуға (зат есім, етістік, сын есім);

3) аудармалар мен сөздің синонимдерін қосу/өзгерту;

4) мүмкін болатын жеті доменнің бірін таңдау арқылы енгізілген сөздің доменін анықтауға мүмкіндік береді. Осылайша, 2.4-суретте көрсетілген әзірленген қосымшада тезаурустың мазмұнын толық басқаруға мүмкіндік беретін интерфейс бар.



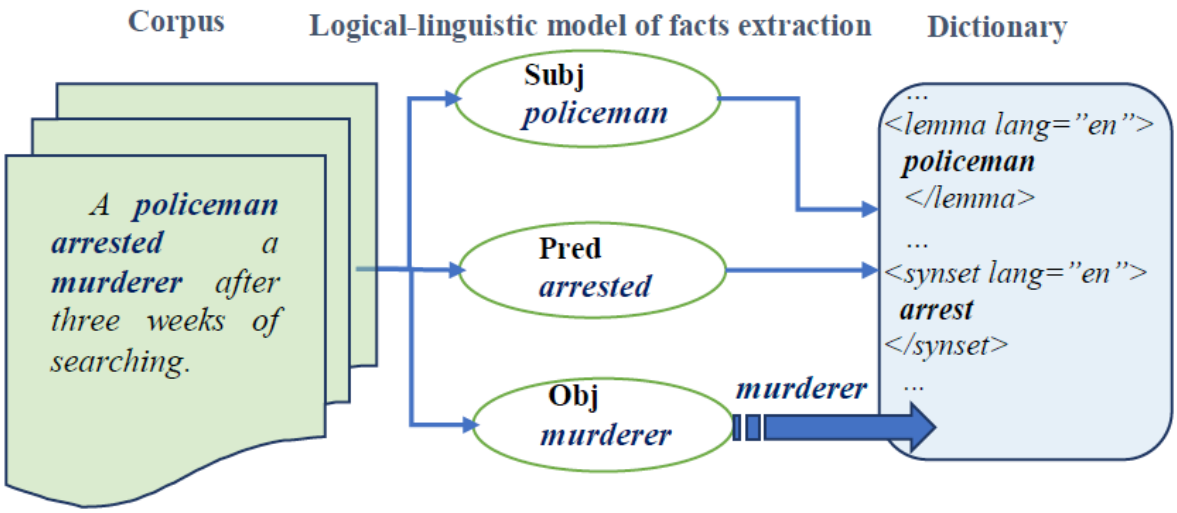
Сурет 2.4 – Криминалистік маңызды лексиканың көптілді тезаурусының бағдарламалық интерфейсі

2.3 Криминалистік маңызды мәтіндердің корпустары негізінде көптілді тезаурусты автоматтандырылған толтыру және кеңейту үшін қолданылатын тәсілдеме

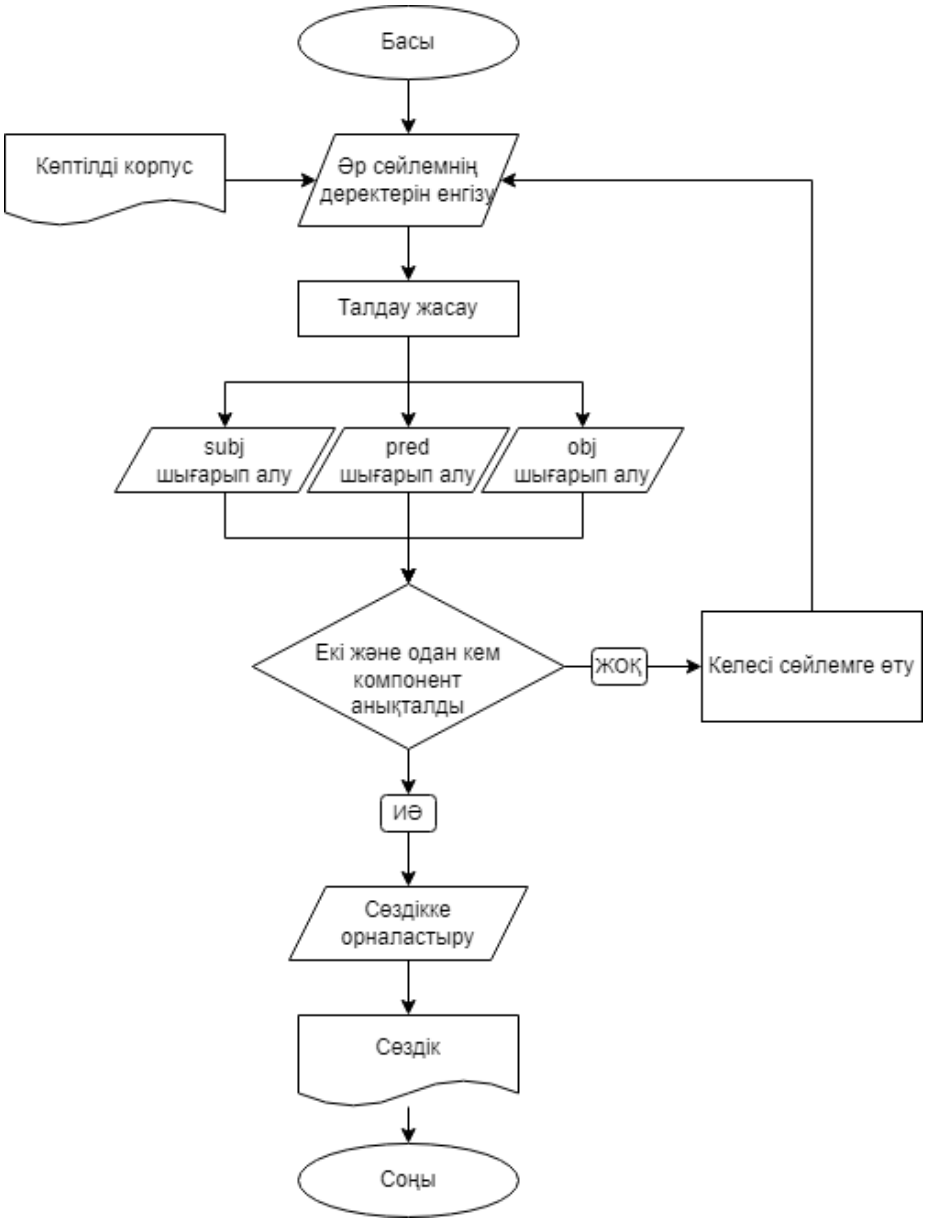
Біз ұсынған онтологияны автоматтандырылған толтыру және кеңейту тәсілдемесі бірнеше белгілі гипотезаларға негізделген. Біріншісі – статистикалық семантиканың негізгі гипотезасы, ол адам сөздерін қолданудағы статистикалық заңдылықтарды адамдардың нақты нені білдіретінін анықтау үшін пайдалануға болатынын айтады [150]. Басқаша айтқанда, адамның интеллектісі қоршаған ортаға байланысты сөздерді түсіне алады. Бұл жалпы гипотеза лингвистикадағы нақты үлестіру гипотезасының негізінде жатыр. [151] сәйкес, бұл гипотеза ұқсас контексте кездесетін сөздердің әдетте ұқсас мағынасы бар екенін айтады.

Іс жүзінде, логикалық-лингвистикалық модельге сүйене отырып [152], біз оқиғаны сөйлемнен шығарып аламыз. Ең көп таралған жағдайда мұндай оқиға семантикалық категорияларды білдіретін Субъект – Объект – Предикат триплетін білдіреді [153]. *Субъект* сөйлемде сипатталған әрекеттің орындаушысы мен бастамашысын атайды. *Объект* әрекет бағытталған объектіні немесе адамды атайды. Ал *Предикат* өз кезегінде сөйлемде сипатталған әрекетті немесе оқиғаның триггерін атайды.

Тезаурустың кеңеюі криминалдық тақырыпқа бағытталған құрылған көптілді корпустар мен негізгі көптілді криминалдық лексикамен қолмен толтырылған тезаурустың негізінде жүзеге асырылады [154]. Корпуста белгіленген оқиғалар мен 2.5-суретте көрсетілген алгоритмді пайдалану тезаурустың автоматтандырылған кеңеюіне мүмкіндік берді.



Сурет 2.5 – Көптілді тезаурусты толтыру мен кеңейтудің жалпы схемасы



Сурет 2.6 – Тезаурусты автоматты түрде толтыру

Әрекеттің *Агентін*, *Объектісін* және *Предикатын* (триггерін) қамтитын аннотацияланған корпус оқиғаларын пайдалану тезаурусты автоматты түрде толтыру үшін келесі алгоритмді (2.6-сурет) жасауға мүмкіндік береді:

1. Бірінші кезеңде оқиғаны қамтитын мәтіннің әрбір жеке сөйлемі талданады және предикатпен, субъектімен және объектімен ұсынылған RDF фактісі анықталады.

2. Келесі кезеңде негізгі тезауруста әрбір нақты тіл үшін триплеттің үш элементінің болуы тексеріледі. Егер триплеттің екі компоненті <*lemma*> немесе <*sunset*> XML құжатының тег элементтері түрінде табылса және бір компонент сөздікте табылмаса, соңғысы автоматты түрде сөздікке орналастырылады.

2.1-кестеде *Субъект*, *Объект* және *Предикат*, сондай-ақ тезаурусқа бұрын болмаған терминдерді қосу нәтижесі бар ағылшын тіліндегі сөйлемдерден автоматты түрде алынған оқиғалардың мысалдары келтірілген.

Кесте 2.1 – Сөйлемдерден автоматты түрде шығарып алынған оқиғалардың мысалдары, тезаурустағы терминдерді іздеу нәтижесі және тезаурусқа термин қосу

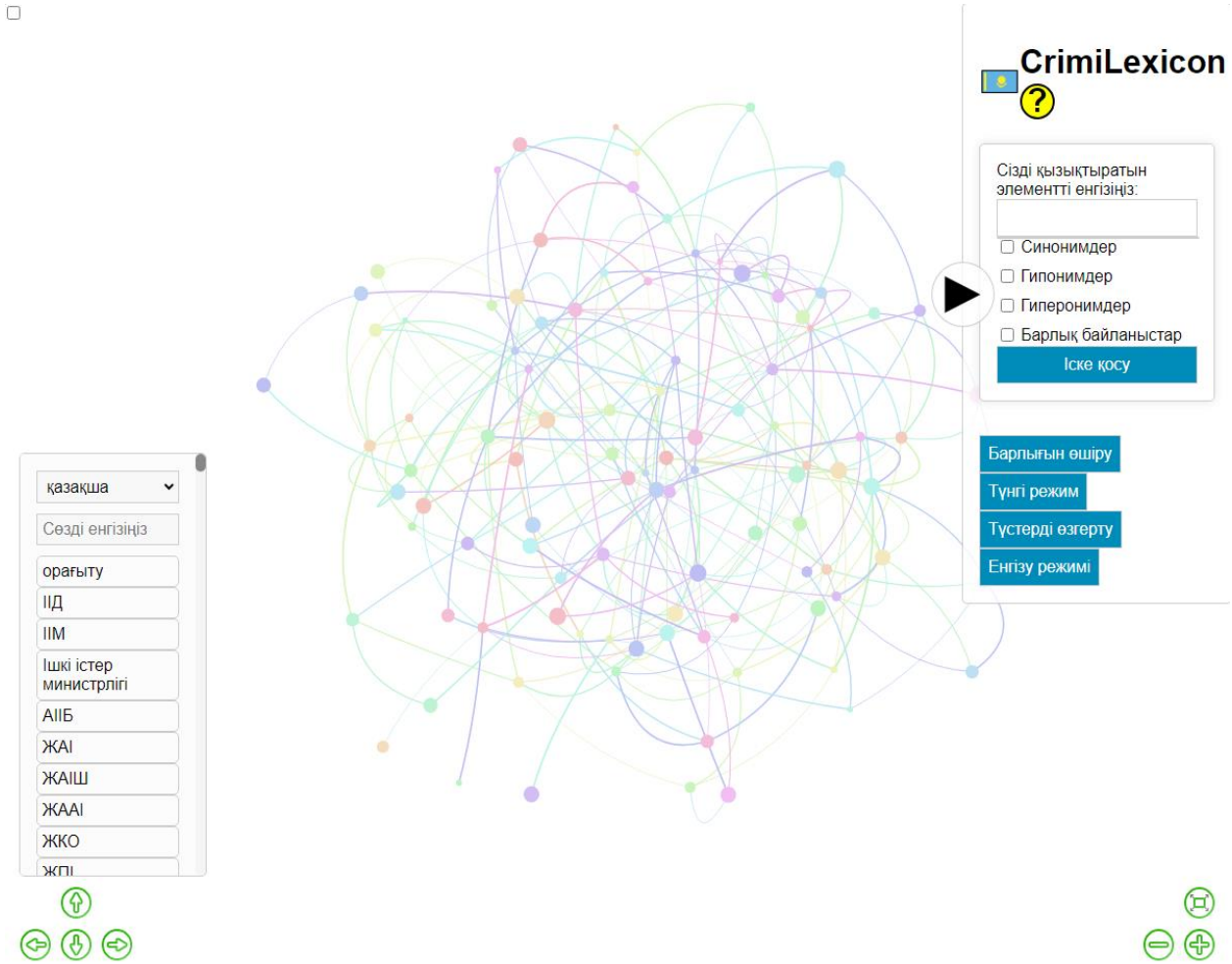
Сөйлем	Автоматты түрде шығарылып алынған оқиға	Тезаурус терминдері	Тезаурусқа қосылған терминдер
The governor stormed into the hospital and demanded to know how many children died	*Subj*: governor *Obj*: hospital *Pred*: demanded	hospital demand	governor
Police are searching for a person of interest in the murder of an 18-year-old woman in November.	*Subj*: police *Obj*: person of interest *Pred*: searching	police search	person of interest
Police encountered a distraught woman crying that her baby had died.	*Subj*: police *Obj*: distraught woman *Pred*: encountered	police encountered	distraught woman

Соңғы қадамда ана тілінде сөйлеуші езаурустың автоматты түрде аяқталу нәтижесін тексереді, оның тақырыптық бағытына, атап айтқанда, құқыққа қайшы әрекеттерге толық сәйкес келмейтін терминдерді жояды. Осылайша, мамандандырылған сөздік корпустың кеңеюімен қатар автоматты түрде кеңейеді және онтологияны құруға негіз болады.

2.4 «Құқыққа қайшы интернет-контент» көптілді онтологиясын құру

Құқыққа қайшы интернет-контентті іздеу мен талдаудың, криминалистік маңызды оқиғаның триггерін анықтаудың және триггердің және/немесе оқиғаның түрін анықтаудың интеграцияланған технологиясында «Құқыққа қайшы интернет-контент» көптілді онтологиясы негізделген [148, p. 108-116]. Онтология құқыққа қайшы әрекеттер мен қылмысқа қатысты терминдердің синонимдік сөздігіне негізделген.

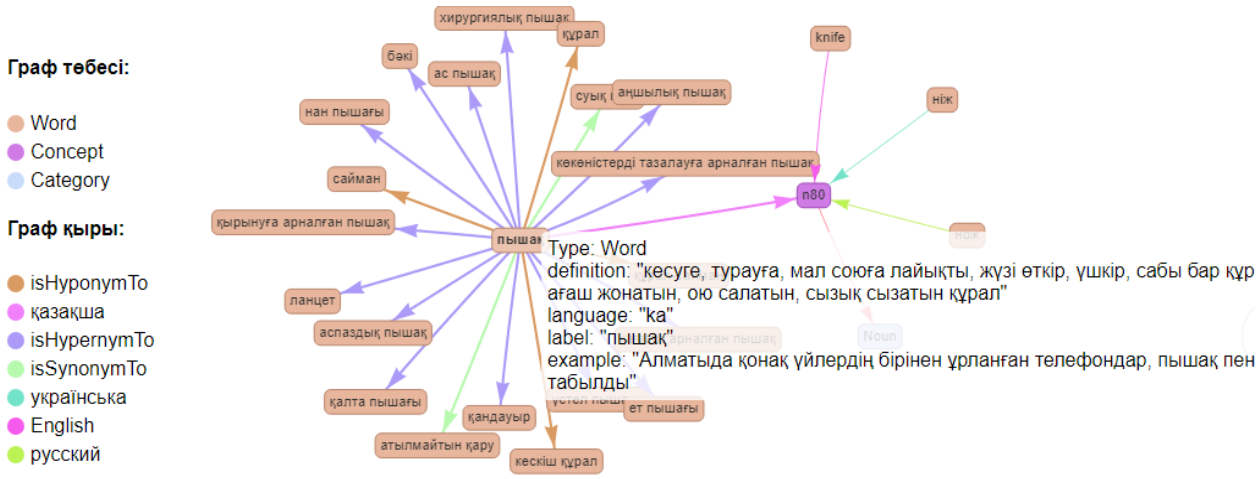
Онтологиялық визуализация динамикалық және интерактивті веб-қосымша ретінде жасалған және 2.7-суретте көрсетілген.



Сурет 2.7 – «Құқыққа қайшы интернет-контент» көптілді онтологиясының визуализациясы

Тезаурус файлынан алынған онтологияның ағымдағы нұсқасы 12885 объектіні қамтиды, олардың әрқайсысында бірегей 10 таңбалы идентификатор бар. Онтология төрт негізгі класты қамтиды: *Category*, *Domain*, *Term* және *Word*. *Category* класы терминнің сөз табын анықтауға арналған. *Domain* класы құқыққа қайшы әрекеттің лексикасын жіктеудің семантикалық кластарының ішкі түрі болып табылады. *Term* класы онтологиямен анықталған ұғымдарды қалыптастыру үшін қолданылады. Бұл класс қазақ, орыс, ағылшын және украин тілдеріндегі тұжырымдаманың әртүрлі лексикалық сөйлемдерін қамтиды. Мысалы, мұндай n80 концепті орыс тілінде *«нож»* сөзін, ағылшын тілінде *«knife»* сөзін, қазақ тілінде *«пышақ»* сөзін білдіреді. *Word* класы онтологиядағы жеке сөздерді білдіреді және оның тілді анықтайтын екі әріптен тұратын атрибуттары, анықтамасы (definition), мысалы (example) бар.

2.8-суретте онтология субъектілер және олардың байланыстары туралы құрылымдық ақпаратты графикалық түрде кодтайтын білім графигі (KG) түрінде берілген.

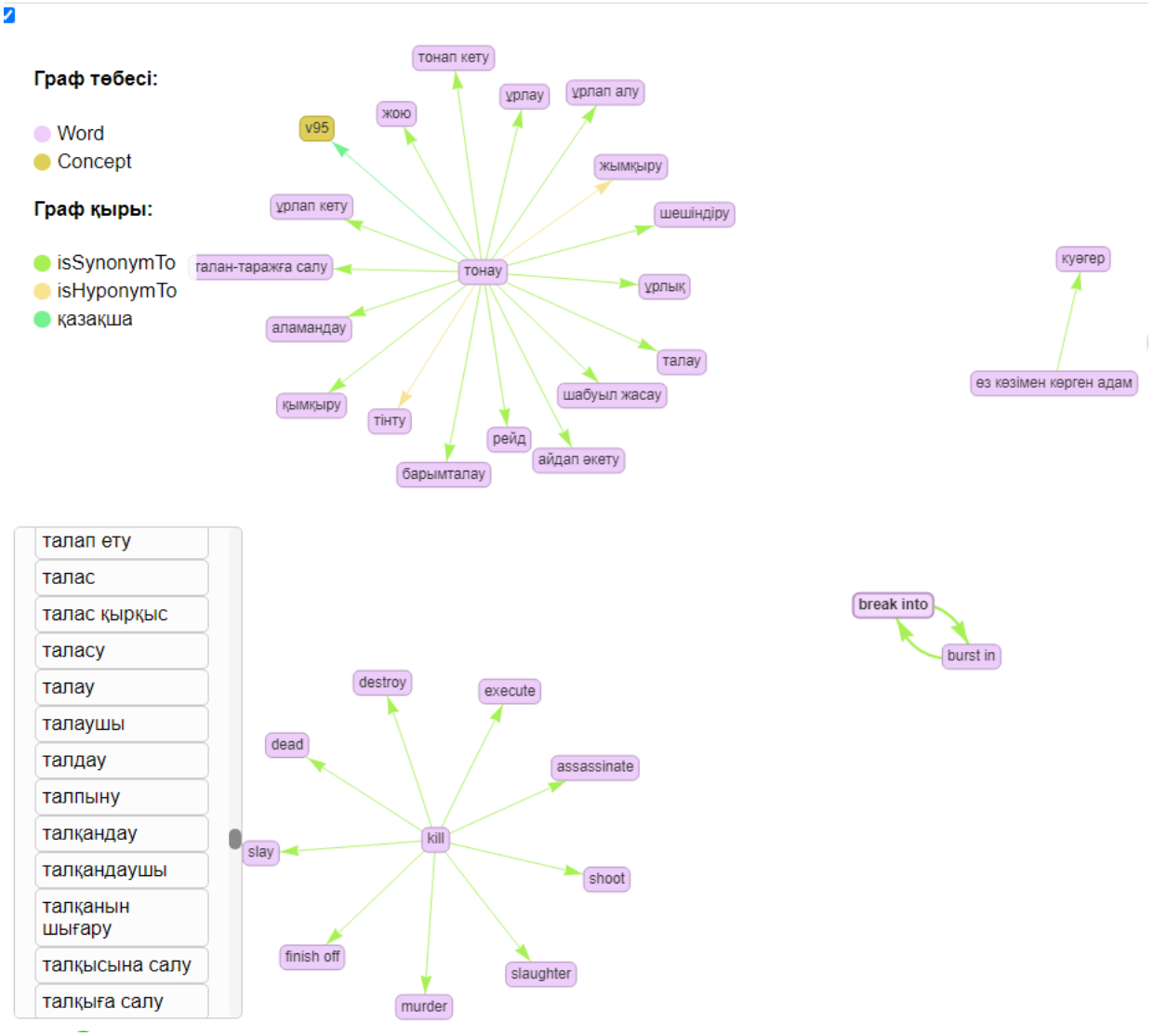


Сурет 2.8 – Көптілді онтологияның білім графигі түріндегі визуализациясы

Онтологияны жақсы құрылымдалған көптілді лингвистикалық ресурс ретінде қалыптастыру үшін біз кластардан басқа семантикалық қатынастардың келесі түрлерін қолданамыз. *BelongsToDomain* қатынасы жеке терминдер мен кеңірек тақырыптар арасындағы байланысты көрсете отырып, ұғымның құқыққа қайшы әрекеттер доменінің белгілі бір класына немесе ішкі класына жататындығын анықтайды. *BelongsToTerm* қатынасы белгілі бір тілдің сөзі мен оның ұғымы арасында байланыс орнатады. *GrammaticalCategory* қатынасы сөзді белгілі бір грамматикалық категориямен байланыстырады. Келесі екі қатынас ұғымдар арасындағы семантикалық иерархияны көрсетеді. *IsHypernymTo* қатынасы гипероним кең категория ұғымын білдіретін жалпы субъектімен байланысты білдіреді. *IsHyponymTo* қатынасы гипонимияның кері қатынасын білдіреді, ол жалпы субъектіден тар немесе нақты ұғымға қатысты орнатылады. Сонымен қатар, онтология лингвистикалық мәліметтер қорының дәстүрлі қатынасын – синонимділік қатынасын қамтиды. *IsSynonymTo* қатынасы бірдей немесе бірдей дерлік мағынаға ие ұғымдар арасында орнатылады.

Пайдаланушы интерфейсін жақсарту және пайдаланушының онтологиямен өзара әрекеттесуін жеңілдету үшін қосымша Query of Crime (QC) әзірленген сұратым тілімен толықтырылған, бұл пайдаланушыға күрделі сұратымдарды оңай және дәл тұжырымдауға мүмкіндік береді. QC синтаксисі синонимдерді (*syn*()), гипонимдерді (*hypo*()), гиперонимдерді (*hyper*()), терминдерді (*term*()) және домендерді (*domain*()) іздеу құрылымдарын қамтиды. *All()* – жақшадағы сөз/ұғым/терминмен барлық байланыстарды визуализациялайды. {Cypher-ге сұратым} – егер QC мүмкіндіктері қажеттіліктерді қанағаттандырмаса және күрделі сұратымды орындау қажет болса, онда сұратымды бұйра жақшаға (n (бірінші нүкте), r (қатынас) және m (екінші нүкте)) сұрау операндтары ретінде енгізу керек. Барлық қатынастарды визуализациялауға, сондай-ақ мәліметтер қорының бірнеше қосымша элементтерін көрсетуге мүмкіндік беретін QC тіліндегі сұратым мысалы 2.9-суретте көрсетілген.

Сұратым мысалы; all(тонау)+syn(kill)+{MATCH (n)-[r]-(m) WITH n, r, m LIMIT 3 RETURN n, r, m} – бұл «тонау» сөзінің барлық байланыстарын, «kill» сөзінің синонимдерін және мәліметтер қорының 3 кездейсоқ элементтерін шығарады. Бұл әдеттегі енгізу режимінен тыс ерекше жағдайларды зерттеуге мүмкіндік береді.



Сурет 2.9 – «Құқыққа қайшы интернет-контент» онтологиясына күрделі сұратымды жүзеге асыру

Жүйе клиент-сервер моделінде жұмыс істейді. Клиент жағында пайдаланушыларға белгілі бір домен арқылы қосымшаға кіруге мүмкіндік беретін веб-шолғыш пайдаланылады. Сервер компоненті ретінде RDF/XML Protege файлдарынан деректерді өңдейтін Neo4J мәліметтер қоры қолданылады. Интерфейс vis.js, сонымен қатар жергілікті JS компоненттері, CSS және HTML арқылы жасалған. Protege файлы Owlready2 кітапханасымен бірге Python бағдарламалау тілін қолдану арқылы жасалған. Архитектура және интерфейс дизайны компоненттері онтологиялық деректерді пайдалануды жеңілдету үшін

vis.js, CSS және HTML сияқты құралдарды пайдаланады. Әзірленген онтологияның бағдарламалық кодының үзіндісі (Қосымша Ә)-да келітірлген.

Көптілді онтологияның қолданылу саласы – бұл құқық қорғау органдарының интеграцияланған ақпараттық-криминалистикалық жүйелері, сондай-ақ интернет-контенттің ашық бөлігіндегі құқыққа қайшы ақпаратты ақпараттық іздеу және талдау үшін қолданылатын басқа мемлекеттік органдардың интеграцияланған ақпараттық-талдамалық жүйелері.

«Құқыққа қайшы интернет-контент» көптілді онтологиясы http://multilingual.ontology.iict.kz:38000/crimeontoscope/ мекенжайы бойынша еркін және жалпыға қол жетімді. Онтологияның құрылымы оны мәтіндік құжаттардың мағынасын модельдеу кезінде компьютерлік лингвистика мен жасанды интеллект тәсілдерін дамытуға, сондай-ақ мәтіндерді семантикалық талдау әдістерін жасауға, құрылымдалмаған мәтіндерге негізделген онтологияларды автоматты түрде құруға және ақпарат іздеуге пайдалы құрал етеді.

2-бөлімнің қорытындысы

Екінші тарауда криминалистік маңызды СМС ақпаратының құрылған мәтіндік корпустары және криминалдық лексикамен жасалған көптілді терминологиялық тезаурус кіретін құқыққа қайшы контенттің онтологиясын толтырудың негізгі көздері қарастырылады. Бірінші көптілді корпусқа орыс, украин және ағылшын тілдерінің мәтіндері кіреді. Украин тіліндегі ішкі корпусында 3147 мәтін, орыс тілінде – 5506, ағылшын – 300 мәтін бар.

Онтологияны қалыптастыру үшін қор ретінде пайдаланылатын екінші көптілді корпус – параллель қазақ-орыс корпусы. Бұл корпустың көлемі орыс тіліндегі 3000 мәтінді және қазақ тіліндегі 3000 мәтінді, оның ішінде мағынасы бойынша тураланған қазақ-орыс сөйлемдерін қамтитын 2000 мәтінді құрайды.

Көптілді терминологиялық тезаурус 600-ден астам негізгі сөздерді (330 зат есім, 107 сын есім және 170-ке жуық етістік) және 2500-ден астам негізгі сөз синонимдерін қамтиды.

Тарауда сонымен қатар құрылымдалмаған мәтіндерден фактілерді шығарып алудың қол жетімді моделіне негізделген тезаурусты автоматтандырылған толтыру және кеңейту үшін жасалған тәсілдің сипаттамасы бар.

«Құқыққа қайшы интернет-контент» көптілді онтологиясын әзірлеу, оны визуализациялау және пайдалану шектеулері ұсынылған. Онтология құқыққа қайшы әрекеттер мен қылмысқа қатысты терминдердің синонимдік сөздігіне негізделген.

3 КРИМИНАЛИСТІК МАҢЫЗДЫ МӘТІНДЕРДІҢ МАМАНДАНДЫРЫЛҒАН КОРПУСТАРЫН АВТОМАТТЫ СЕМАНТИКАЛЫҚ БЕЛГІЛЕУ ӘДІСІ МЕН ҚҰРАЛДАРЫН ӘЗІРЛЕУ

3.1 Құқыққа қайшы контенттің лингвистикалық және лексикалық маркерлерін ерекшелеу әдісі

Құрылымдалмаған мәтіндерден оқиғаларды шығарып алудың әртүрлі тәсілдерінің болуына қарамастан, ЕЕ мәтінді өңдеудің өте күрделі мәселесі болып қала береді, бұл әсіресе қазақ тілін жатқызуға болатын лингвистикалық ресурстары шектеулі тілдер үшін өзекті болып табылады.

Бұл зерттеуде қазақ-орыс корпусынан қылмысқа байланысты жаңалықтар мақалаларының мәтіндерінен құрылымдық ақпаратты шығарып алу кезінде біз *Оқиғаларды аннотациялау жөніндегі нұсқаулықта (Automatic Content Extraction2005 English*) келтірілген оқиғаны анықтауға негізделдік [155]. Осы нұсқаулыққа сәйкес, оқиға белгілі бір уақытта және белгілі бір жерде, объектілердің (нысандардың) бір немесе бірнеше әрекет етушілерінің қатысуымен болатын нақты әрекетті білдіреді. Алайда, біздің зерттеуімізде біз тек қылмысқа байланысты жабық пәндік саланың оқиғаларын қарастырамыз, яғни құқыққа қайшы әрекеттерге немесе полиция қызметтерінің әрекеттеріне байланысты оқиғаларды шығарып аламыз.

Компьютерлік-жанама коммуникация (Computer-Mediated Communications (CMC)), мысалы, әлеуметтік медиа және жаңалықтар арналары, әдетте, ең соңғы ақпаратты ұсынады, көптеген қылмыстық және зорлық-зомбылық әрекеттерін сипаттайды және олардың қатысушылары, орны, уақыты, құралдары және кейде қылмыстық әрекеттің себептері туралы күн сайын есеп береді. Осылайша, тиісті онтологияға оқиғалар туралы құрылымдық ақпаратты алуға және қосуға мүмкіндік береді.

Біз әзірлеген криминалистік маңызды мәтіндердің қазақ-орыс корпусына төрт қос тілді *zakon.kz, caravan.kz, lenta.kz, nur.kz* және Жетісу облысының Полиция департаменті сайттарының жаңалықтар мақалалары кірді. Таңдалған сайттар Қазақстан Республикасының танымал және сенімді порталдары болып табылады, олардың жаңалықтар бағыттарының бірі криминалдық жаңалықтар болып табылады. Корпусқа автоматты түрде енгізілген жаңалықтарда тонау, көлік ұрлау, кісі өлтіру, жол-көлік оқиғалары және т.б. сияқты криминалдық әрекеттер туралы ақпарат бар. [138, p. 516-519; 142, p. 250-253; 145, p. 5753-5756; 146] зерттеулерге сүйене отырып, біз CRE-ні полиция мен криминалға қатысты жаңалықтар корпусынан анықтаймыз және шығарып аламыз. Алайда, алдыңғы зерттеулерден айырмашылығы, біз қылмыстың белгілі бір түрін емес (тек есірткі қылмысы немесе тек жол-көлік оқиғалары сияқты) емес, құқыққа қайшы әрекеттерге қатысты барлық оқиғалардың үлкен тобын қарастырамыз. Бұл жол-көлік оқиғалары (TRANSFER), криминалдық құқық бұзушылықтар (CRIME) және полиция қызметі (POLICE) сияқты оқиғалардың түрлері. Өз кезегінде осы үш түрдің әрқайсысы ішкі түрлерге бөлінеді. 3.1-кестеде қарастырылып отырған оқиғалардың түрлері мен ішкі түрлері келтірілген.

Кесте 3.1 – Аннотацияланған оқиғалардың түрлері мен ішкі түрлері

Оқиға түрі	Оқиғаның ішкі түрі
TRANSFER	Movement, Traffic Accident
CRIME	Injure, Offense
POLICE	Arrest, Trial, PD

Оқиға түрлерін, оның сөз тіркесінің немесе сөйлемнің шекарасы бойынша шекарасын анықтап, белгілей отырып, біз оқиғаларға қатысушыларды және оқиғаның атрибуттарын ерекшелеп, оқиғаның аргументтерін анықтаймыз. Жалпы, жоғарыда анықталған CRE барлық түрлері оқиғаның екі қатысушысын қамтиды және әрекеттің немесе оқиғаның бірнеше атрибуттарын анықтайды. Оқиғаның бірінші қатысушысы – оқиғаның бастамашысын атайтын *Агент*. Оқиғалардың барлық осы түрлерінің екінші қатысушысы әрекет бағытталған адам, ұйым немесе көлік құралымен ұсынылған *Объект* болып табылады. Coplink жобасы [156] негізінде CRE қатысушыларын анықтау үшін қылмыстық әрекетке қатысы болуы мүмкін субъектілердің үш түрін анықтаймыз. Мұндай тұлғалар семантикалық кластар болып табылады: <Person>, <Organization> және <Vehicle>. Оған қоса, біз қарастыратын барлық оқиға түрлері мен ішкі түрлері дәстүрлі TIME-ARG және PLACE-ARG атрибуттарын қамтиды.

Тұжырымдамалық тұрғыдан алғанда, CRIME түріндегі оқиға адам немесе ұйым қандай да бір криминалдық әрекетті жасаған кезде орын алады. Біз оқиғаның осы түріндегі оқиғаның екі ішкі түрін бөліп көрсетеміз. Бұл INJURE және OFFENSE. Құқыққа қайшы әрекеттің объектісі <Person> болып табылған кезде, жеке тұлғаға қарсы қылмыстар деп аталатын INJURE ішкі түрді анықтай аламыз. Қатысқан адам дене жарақатын (өлтіру, жарақат алу) алуы мүмкін немесе басқа қылмыстық әрекеттерден (тонау, алаяқтық) зардап шегуі мүмкін. INJURE ішкі түрінің *Объектісі* тек зардап шеккен адам (адамдар) бола алады, ал құқыққа қайшы әрекеттің бастамашысы болып табылатын осы ішкі түрдің *Агенті* физикалық зиян келтіретін адам да, ұйым да бола алады.

OFFENSE ішкі түрі қылмыстық әрекеттің *Объектісі* тікелей адам болып табылмайтын жағдайларда пайда болады. OFFENSE оқиғасының *Агенті* құқық бұзушылықтың бастамашысы, қандай да бір зиян келтіретін немесе құқыққа қарсы қызметті жүзеге асыратын тұлға немесе ұйым болып табылады. *Агент* оқиғаның қажетті қатысушысы болып табылады. Дегенмен, осы ішкі түрдің екінші қатысушысы болып табылатын жансыз *Объект* белгілі бір фразада немесе сөйлемде болуы немесе болмауы мүмкін.

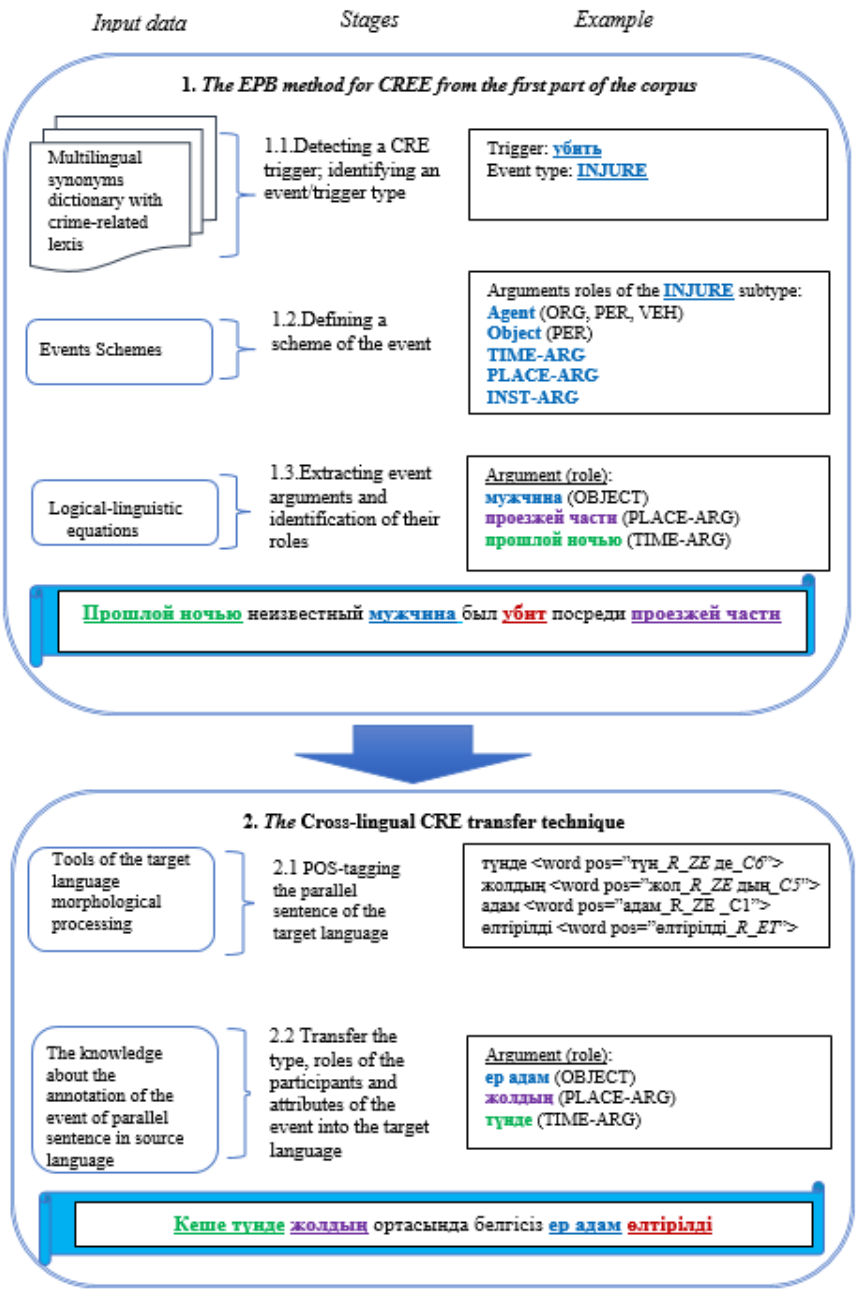
TRANSFER түріндегі криминалдық оқиға екі ішкі түрді қамтиды, атап айтқанда MOVEMENT және TRAFFIC ACCIDENT. MOVEMENT оқиғасының TRANSFER ішкі түрі жансыз зат немесе PERSON бір жерден екінші орынға ауысқан кезде пайда болады. Сонымен қатар, біз ұрлық немесе ұрлық үшін бір нәрсені жылжыту MOVEMENT ішкі түріне жатпайды, бірақ CRIME түріндегі оқиғаға қатысты деп болжадық. TRANSFER оқиғасының криминалдық түрінің тағы бір ішкі түрі – TRAFFIC ACCIDENT. Көлік апатқа ұшыраған кезде пайда

болатын оқиға. Бұл жағдайда *Агент* жол-көлік оқиғасын тудырған PERSON немесе VEHICLE класының нысаны болуы керек.

Біз қарастырған криминалдық оқиғаның соңғы түрі – полиция немесе шенеунік жасаған кезде пайда болатын POLICE оқиғасы. POLICE криминалдық оқиғасының түріне үш ішкі тип кіреді, атап айтқанда ARREST, TRIAL және PD. ARREST – бұл адамның қозғалысы мемлекеттік субъектімен (мысалы, полиция немесе сот төрелігі) шектелген кезде анықталатын ішкі тип. ARREST ішкі типі жағдайында *Агент* басқа адамды ұстаудың бастамашысы болып табылатын PERSON немесе ORGANIZATION ретінде нақты анықталуы мүмкін, ал *Объект* тек PERSON семантикалық класындағы ұсталған адам болуы мүмкін. TRIAL – бұл сот немесе қандай да бір үкіметтік ұйым адамды немесе ұйымды қылмыс жасады деп айыптаған кезде пайда болатын POLICE түріндегі криминалдық оқиғаның ішкі түрі. Оқиғаның осы ішкі түріндегі *Агент* әрқашан ORGANIZATION болып табылады, ал *Объект* тек PERSON семантикалық класына жататын нысан бола алады. PD (Police Department) оқиғасының ішкі түрі полиция қызметкері немесе полиция департаменті ресми міндеттерді орындаған кезде пайда болады. Мұндай оқиғаның *Агенті* тек адам ретінде полиция қызметкері немесе ұйым ретінде полиция бөлімі болуы керек. Сонымен қатар, криминалдық оқиғаның осы ішкі түрінің *Объектісі* тек PERSON класының нысаны бола алады.

Осылайша, алдыңғы зерттеулерге [138, p. 516-519; 142, p. 250-253; 143, p. 1366-1371; 144, p. 231-237; 145, p. 5753-5756] сүйене отырып, жобаның алдыңғы кезеңдерінде құрылған қазақ-орыс жаңалықтар корпусындағы криминалистік және полиция қызметіне қатысты оқиғаларға аннотациялау үшін біз криминалистік маңызды оқиғалардың үш түрін және жеті ішкі түрін бөліп көрсетеміз және шығарып аламыз. Оқиғалар аргументтерінің аннотациясының екі кезеңдік тәсілі [157] екі кезеңнен тұрады: (1) орыс тіліндегі мәтіндер корпусының кіріс бөлігі үшін шаблондарды қолдануға негізделген Enhanced Pattern-Based (EPB) әдісі және (2) қазақ тіліндегі мәтіндер корпусының екінші бөлігіне арналған криминалистік маңызды оқиғаларды тіл аралық тасымалдау әдісі (Cross-lingual CRE transfer method). 3.1-суретте қазақ-орыс параллель корпусының мәтіндерінде криминалистік маңызды оқиғаларды (CRE) семантикалық түсіндіруге қолданылатын екі сатылы тәсілдің жалпы схемасы көрсетілген.

CdEE зерттеу тәсілдеріне сүйене отырып [131, 51, p. e13067; 131, p. 1277-1278], біз оқиғаның триггерін оқиғаны, оқиға/триггер түрін сипаттайтын фразада дәйекті түрде анықтаймыз, оқиғаның аргументтерін және олардың рөлдерін сәйкестендіреміз. Бұл кезең келесі үш қадамды жүзеге асыруды қамтиды. Кезеңнің бірінші қадамында CRE триггерін бір уақытта анықтау және құқыққа қайшы интернет-контенттің көптілді онтологиясына негізделген оқиға/триггер түрін анықтау әдісі қолданылады [148, p. 108-116]. Мысалы, «убить» етістігі үшін <DOMAIN> элементінің мағынасы «INJURE» болып табылады. Тиісінше, *«Прошлой ночью неизвестный мужчина был убит посреди проезжей части»* деген сөйлеммен сипатталған оқиға түрі «INJURE» ретінде анықталады.



Сурет 3.1 – Қазақ-орыс параллель корпусының мәтіндеріндегі криминалистік маңызды оқиғаларды (CRE) семантикалық түсіндіруге қолданылатын екі сатылы тәсілдің схемасы

Алайда, корпуста криминалдық жаңалықтардың мәтіндері болғандықтан, көптеген жағдайларда құқыққа қайшы әрекетті сипаттайтын және атайтын сөйлемдер мен сөз тіркестеріндегі негізгі етістіктер семантикалық тұрғыдан «жеңіл» етістік болып табылады, мысалы, *«сенің тапсыру»*, *«айтып жеткізу»*, *«болжау»*, *«беру»* және т. б. Мұндай сөйлемдерді есепке алу үшін қолда бар тезаурустан мыңнан астам зат есімдер мен олардың синонимдерін қамтитын тізім триггер ретінде қарастырылды. Бұл тізімге, мысалы, *«өлтіруші»*, *«қорлық»*, *«атыс»*, *«детонация»* және т.б. сияқты зат есімдер кіреді.

Екінші қадамда жоғарыда сипатталған әрбір құқыққа қайшы ішкі түр үшін оқиға аргументтерінің схемасы анықталады. Схема *Агенттер* немесе *Объектілер* сияқты белгілі бір ішкі түрдегі оқиғаларға қатысушыларды көрсететін онтологияларды қамтиды. Сонымен қатар, біз полиция немесе криминалдық жаңалықтарын қарастыратындықтан, бізді орын мен уақыт (PLACE-ARG және TIME-ARG) сияқты оқиға атрибуттары әрқашан қызықтырады. Қатысушылардың рөлдерін және оқиғалар аргументтерін сипаттау үшін сөйлемдегі немесе сөз тіркесіндегі сөздердің грамматикалық және семантикалық сипаттамалары арасындағы байланыс арқылы оқиға аргументтерін сипаттайтын логикалық-лингвистикалық теңдеулер қолданылады. Тұтастай алғанда, бұл тәсіл шаблонды сәйкестендіру әдісіне (pattern matching technique) жақын, онда алдымен белгілі бір оқиға шаблондары жасалады, содан кейін бұл шаблондар өңделмеген немесе аннотацияланған мәтінмен салыстырылады. Алайда, әдетте, осы тәсілге негізделген ЕЕ қосымшалары оқиғалар шаблондарының салыстырмалы түрде аз санын пайдалануға мүмкіндік береді. Біздің әдіс логикалық-лингвистикалық теңдеулерді қолдана отырып, талданатын тілдің белгілі бір саласында бар атрибуттың әрбір мүмкін рөлін сипаттауға мүмкіндік береді.

Өңдеудің бірінші кезеңінің соңғы қадамында оқиғалар аргументтерін шығарып алу және олардың рөлдерін сипаттау үшін біздің [157, p. 54093-54110] жұмысымызда жақсы сипатталған дамыған логикалық-лингвистикалық теңдеулер (LLE) қолданылады. LLE қолдану сөйлемдегі сөздердің грамматикалық және семантикалық сипаттамаларының қатынастары арқылы қатысушылардың рөлдері мен оқиға атрибуттарын анықтауға мүмкіндік береді.

Екінші кезеңде біз қазақ тіліндегі сөйлемдерден оқиғаларды шығарып алу үшін CRE тіларалық тасымалдау әдісін қолданамыз. Бұл әдіс бір оқиғаны бастапқы тілдің белгіленген сөйлемінде де, параллель корпустың мақсатты тілінің тураланған сөйлемінде де сипаттауға болады деген гипотезаға негізделген [158]. Бұл әдісті іске асыру аннотацияланған оқиға туралы білімді, атап айтқанда оның түрін, триггерін, қатысушылардың рөлі мен оқиға атрибуттарын бастапқы тілдің параллель сөйлемінен мақсатты тіл сөйлеміне жеткізу үшін мақсатты тілдің POS-тегтеу белгілеуін пайдаланады. Орыс және қазақ тілдерінің тураланған сөйлемдерінің жалпы семантикалық кеңістігін анықтау үшін қазақ тілінің сөйлемінің POS-тегтері мен орыс тілінің тураланған сөйлемінен оқиғаға қатысушылардың/атрибуттардың ықтимал рөлдері арасындағы сәйкестік үлгілері пайдаланылады.

3.2 Мәтіндегі оқиғалар аргументтерінің рөлін анықтаудың ақпараттық моделі

Біз қолданатын оқиғалар атрибуттары мен олардың рөлдерін мәтіндерден шығарып алу әдісі фактілерді алудың логикалық-лингвистикалық моделіне негізделген [152, p. 1714829], оның негізгі математикалық құралы ақырғы предикаттар алгебрасы (АПА) болып табылады. Сөйлем сөздерінің бар грамматикалық және семантикалық сипаттамаларын сипаттау үшін АПА x_i^a

предикаттық айнымалысын пайдаланады, мұнда *a* – сөздің белгісі немесе сипаттамасы.

$$x_i^a = \begin{Bmatrix} 1, if \;\; x_i = a \\ 0, if \;\; x_i \neq a \end{Bmatrix} (1 \le i \le n), \qquad (3.1$$

Мысалы, егер орыс тіліндегі сөйлемдегі *i-ші* сөз зат есім болса және ілік септігінде болса, x_i^{gen} предикат айнымалысы бірлік мәніне ие болады, ал $x_i^{gen} \vee x_i^{nom} = 1$ логикалық теңдеудегі дизъюнкция орыс тіліндегі сөйлемнің *i-ші* сөзі ілік немесе атау септігінде зат есім болуы мүмкін екенін білдіреді.

Оқиға аргументтерінің рөлдерін білдіретін орысша сөйлемдеріндегі зат есімдердің мүмкін грамматикалық және семантикалық сипаттамалары ретінде біз мыналарды ерекшелеп көрсетеміз:

– грамматикалық септік;

– жанды немесе жансыз;

– объектінің семантикалық класы;

– орыс тіліндегі пассив етісті формалдайтын зат есім және бірнеше белгілер (фразада «*н*» жұрнағы бар етістіктің, «*-ся*» шылауының және «*быть*» көмекші етістігінің болуы).

Осылайша, біз $M = \{x, y, z, m, l, f\}$ алты предикаттық айнымалылардың ақырғы жиынтығын енгіземіз, олар біз енгізген грамматикалық және семантикалық белгілердің соңғы жиынтығын сипаттайды, қатысушыларды және орыс сөйлемдеріндегі оқиғалардың енгізілген түрлерінің атрибуттарын атайды.

Модельдің келесі кезеңінде *S* предикат жүйесі енгізіледі. Егер x_i сипаттамаларының мәндері оқиғаның аргументтерін атайтын сөздің грамматикалық немесе семантикалық мағынасына сәйкес келсе, $P_i(x_i) \in S$ предикаты бірге тең болады. Әйтпесе, $P_i(x) = 0$ предикаты жалған болады. Орыс сөйлемдерінің зат есімдерінің мүмкін болатын грамматикалық және семантикалық сипаттамаларының ақырғы жиынын сипаттайтын көптеген предикаттар төмендегі алты модельде келесі предикаттармен сипатталады.

Зат есімнің грамматикалық септіктері *z* предикат айнымалысы арқылы беріледі:

$$P(z) = z^{nom} \vee z^{gen} \vee z^{dat} \vee z^{acc} \vee z^{ins} \vee z^{loc}, \qquad (3.2)$$

мұнда *nom, gen, dat, acc, ins, loc* предикаттық айнымалысының белгілері тиісінше атау, ілік, барыс, табыс, көмектес және жатыс септіктері болып табылады.

X предикаттық айнымалысы арқылы зат есімнің жанды-жансыз сияқты семантикалық белгілері көрсетіледі.

$$P(x) = x^{anim} \vee x^{inan}, \qquad (3.3)$$

мұнда *anim* предикаттық айнымалысының белгісі жанды зат есімді, ал *inan* жансыз зат есімді білдіреді.

Зат есімнің семантикалық категориялары, оны нысанды тану сатысында анықтауға болады, *y* предикат айнымалысы арқылы енгізіледі:

$$P(y)= y^{ORG} \vee y^{PER} \vee y^{LOC} \vee y^{VEH} \vee y^{TIME} \vee y^{TOOL} \vee y^{Others} , \quad (3.4)$$

мұнда *ORG, PER, LOC* және *VEH* белгілері сәйкесінше ұйымның, тұлғаның, жер атауының немесе көлік құралының семантикалық белгісінің болуын білдіреді; *TIME* және *TOOL* оқиғаларды зерттеу процесінде анықталған атрибуттарда әдетте болатын күннің және/немесе уақыттың және құралдың семантикалық белгісінің болуын білдіреді; *Others* нысанды ерекшелеу сатысында сөздің семантикалық атрибутын анықтау мүмкін болмаса пайдаланылады.

Оқиғаға қатысушыларды анықтаған кезде, *Агентті* іс-әрекеттің бастамашысы және әрекет бағытталған нысан ретінде *Объектті* дұрыс анықтау үшін, біз модельге орыс тілінің сөйлемдерінде кездесетін пассив етісті ресімдейтін грамматикалық белгілерді сипаттайтын үш қосымша айнымалыны {*m, f, l*} енгіземіз.

m предикат айнымалысын енгізу әрекетті тікелей шақыратын сөйлемнің негізгі етістігінде «-ся» бөлшегінің болуын сипаттауға мүмкіндік береді:

$$P(m)= m^{Part} \vee m^{NOT\_Part} , \quad (3.5)$$

f предикат айнымалысы сөз тіркесінде «*был*», «*была*», «*было*», «*были*» көмекші етістіктің болуын немесе болмауын көрсетеді:

$$P(f)= f^{aux} \vee f^{NOT\_aux} , \quad (3.6)$$

Ал *l* айнымалысы негізгі етістіктен шыққан «н» есімше жұрнағының болуын немесе болмауын көрсетеді:

$$P(l)= l^{suff} \vee l^{NOT\_suff} , \quad (3.7)$$

P (x, y, z, m, l, f) көп өлшемді предикаты сөздердің грамматикалық және семантикалық сипаттамаларын сипаттайтын енгізілген алты айнымалы арқылы оқиға атрибуттарының рөлін анықтайды:

$$P(x,y,z,m,l,f) \rightarrow P(x) \wedge P(y) \wedge P(z) \wedge P(m) \wedge P(l) \wedge P(f) , \quad (3.8)$$

Егер предикатпен сипатталған грамматикалық және семантикалық сипаттамалардың тіркесімі (3.8) сәйкес сөйлеммен немесе сөз тіркесімен сипатталатын *Агент*, *Объект* немесе оқиға атрибуттарының бірі болса, онда

предикат $P(x,y,z,m,l,f)=1$ болады. Сипаттамалар арасындағы барлық қатынастар нақты сөздерге, оқиғаларға немесе сөйлемдерге тәуелді емес екені анық.

Іс жүзінде оқиғалар аргументтерінің келісілген грамматикалық және семантикалық сипаттамаларының жиынтығы барлық сипаттамалар жиынтығының декарттық көбейтіндісіне тең емес. Сондықтан, біз $S \times S$ декарттық көбейтіндісінде сипаттамалардың бар қатынастарының предикатын келесідей қоя аламыз:

$$P(x,y,z,m,l,f) = \gamma_k(x,y,z,m,l,f) \times P(x) \times P(y) \times P(z) \times P(m) \times P(l) \times P(f), \qquad (3.9)$$

Берілген $k \in [1,h]$ теңдігінде, мұндағы h = 6 – модельде қарастырылатын оқиға дәлелдерінің рөлдерінің саны болып табылады, атап айтқанда: Агент, Объект, PLACE-ARG, TIME-ARG, INSTRUMENT-ARG, REASON-ARG. Егер белгілі бір оқиғаны атайтын сөз тіркесіндегі сөздердің аталған сипаттамалары жоғарыда аталған рөлдердің бірін анықтаса, онда предикат $\gamma_k(x,y,z,m,l,f)=1$ болады және керісінше, егер сөздің грамматикалық және семантикалық сипаттамаларының тіркесімі жоғарыда аталған рөлдердің ешқайсысына сәйкес келмесе, онда $\gamma_k(x,y,z,m,l,f)=0$ болады. Соңғы жағдайда, оқиғаның рөлін сипаттамайтын сөйлемдегі сөздердің сипаттамалары арасындағы қатынастар (3.9) формуласынан предикатпен алынып тасталады.

Біз оқиға *Агентінің* рөлін γ_1 предикаты арқылы анықтай аламыз. Бұл предикат аргументті қарастырылып отырған оқиғаның *Агенті* деп аталатын орыс сөйлемдеріндегі сөздің грамматикалық және семантикалық сипаттамалары арасындағы байланысты сипаттайтынын білдіреді:

$$\gamma_1(x,y,z,m,l,f) = (y^{ORG} \vee y^{PER} \vee y^{VEH} \vee y^{Others})(x^{anim} \vee x^{inan}) \wedge$$
$$\wedge (z^{nom}(f^{NOT\_aux} l^{NOT\_suff} \vee m^{NOT\_Part}) \vee z^{ins}(f^{aux} l^{suff} \vee m^{Part}))^{\cdot} \qquad (3.10)$$

Біз сондай ақ γ_2 предикаты арқылы оқиға *Объектісінің* рөлін нақты анықтаймыз:

$$\gamma_2(x,y,z,m,l,f) = (y^{ORG} \vee y^{PER} \vee y^{VEH} \vee y^{Others})(x^{anim} \vee x^{inan}) \wedge$$
$$(z^{acc} \vee z^{dat})(f^{NOT\_aux} l^{NOT\_suff} \vee m^{NOT\_Part}) \vee z^{nom}(f^{aux} l^{suff} \vee m^{Part}))^{\cdot} \qquad (3.11)$$

Объект оқиғаның негізгі қатысушысы болып табылады, маңыздылығы бойынша *Агенттен* кейінгі екінші орында. Әдетте, дәстүрлі грамматикада *Объект* зат есім немесе атаулы фраза бола алады, ол әрекет ететін немесе күйі немесе қозғалысы өзгеретін нысанды атайды. Біздің құқыққа қайшы әрекеттердің нақты саласында *Объект* көбінесе зардап шеккен адам, бір жерден екінші жерге ауыстырылатын көлік құралы, алаяқтыққа ұшыраған ұйым болып табылады.

Оқиғаға қатысушылардың рөлдерінен басқа, модель логикалық-лингвистикалық теңдеулер арқылы дәлелдердің басқа рөлдерін анықтауға

мүмкіндік береді. Мысалы, біз сәйкесінше γ_3 және γ_4 предикаттары арқылы PLACE-ARG және TIME-ARG әрекет атрибуттарын ажырата аламыз:

$$\gamma_3(x,y,z) = (y^{LOC} \vee y^{Others}) x^{inan} z^{loc} . \quad (3.12)$$

$$\gamma_4(x,y,z) = y^{Time} x^{inan} (z^{loc} \vee z^{acc}) . \quad (3.13)$$

3.3 Веб-желілердің криминалистік маңызды ақпараттың параллель қазақ-орыс корпусының орыс бөлігіне негізделген онтологиялық нысандар мен қатынастарды автоматты түрде қалыптастыру

Тақырыптық тезауруста (2.2-бөлім) және корпустың сөз тіркестерінде бір уақытта ұсынылған етістіктерді қолдана отырып, параллель корпустың орыс бөлігінде қылмысқа байланысты 30 мыңнан астам оқиға таңдалды. Мысалы, тезаурустағы берілген сөздің <*domain*> элементіне сәйкес *«ұрлады»* етістігі CRIME оқиғасының триггері болып табылады.

3.2-кестеде корпустың орыс бөлігіндегі оқиғалардың жеті ішкі түрге бөлінуі көрсетілген. Бұл үлестірімді алу үшін мәтіндегі етістіктердің түпнұсқа формалары, лемматизацияланған етістіктер және алдын-ала өңдеу сатысында стемминг кезеңінен өткен етістіктер триггер ретінде қарастырылды.

Кесте 3.2 – Корпустың орыс бөлігінде табылған оқиғалардың жеті ішкі түр бойынша үлестірілуі (етістіктер оқиғаның триггерлері ретінде қарастырылды)

Оқиға түрлері	Оқиға ішкі түрлері	Етістіктің түпнұсқалық түрі	Лемматизация-ланған етістік	Стеммингтен өткен етістік
CRIME	Injure	75	3984	3542
	Offense	366	5178	3909
TRANSFER	Movement	9	507	461
	Traffic Accident	139	2351	2909
POLICE	Arrest	239	9035	8221
	Trial	231	4250	3804
	PD	294	7433	6723

Тек етістік қана емес, зат есім де триггер бола алатындығын ескере отырып, тезаурустың 500-ге жуық зат есімдері қарастырылды, олар <*domain*> элементінің мәнімен ерекшеленетін жеті тақырыптық категорияға бөлінеді.

3.2-кестеде етістікті оқиға триггері ретінде қолданған жағдайда мәтіндегі етістіктің сөздік формасының мәтін етістік леммасымен сәйкестіктері қарастырылған жағдайда мәтіндегі оқиғаларды анықтаудың толықтығы жоғары болатыны көрсетілген.

Зерттеудің келесі кезеңінде ұғымдар арасындағы ұғымдар мен қатынастарды бөліп көрсету үшін «зат есім + етістік» жұбы триггер ретінде қолданылды. Мысалы, «сот» + «үкім шығарылды» деген екі сөзді триггер ретінде қолдану тек «үкім шығарылды» етістігін триггер ретінде қолданумен салыстырғанда TRIAL оқиғасының ішкі түрін анықтаудың дәлдігін арттырады.

3.2-суретте корпустың орыс бөлігіндегі оқиғалардың ішкі түрлерге үлестірулерін алу үшін қолданылатын «етістік + зат есім» ретінде пайдаланылған тізімнің үзіндісі көрсетілген.

file	sent_numbe	Event type	Subtype of event	Trigger_noun	Trigger_verb
7670_ru_parsed	2		TRIAI	приговор	осуждены
7670_ru_parsed	2		TRIAI	УК	осуждены
1312_ru_parsed	13		TRIAI	УК	возбуждено
7887_ru_parsed	10		ARREST	показания	задержан
3708_ru_parsed	5		ARREST	задержанный	установлено
2596_ru_parsed	4		PD	полиция	обратились
1575_ru_parsed	9		INJURE	травма	умер
7854_ru_parsed	8		INJURE	ранение	нанесено
8704_ru_parsed	1		INJURE	убийство	подозревается
7991_ru_parsed	2		TRAFFIC _ACCIDENT	обгон	вылетел
9146_ru_parsed	11		TRAFFIC _ACCIDENT	наезд	совершил

Сурет 3.2 – «Етістік + зат есім» жұбы болып табылатын оқиғалардың үлестірілу фрагменті

Оқиға триггерлерінің барлық үш тізімін (зат есімдер, етістіктер және «етістік + зат есім») бір уақытта қолдана отырып, біз орыс бөлігіндегі оқиғалардың түрлері мен ішкі түрлері бойынша үлестірілуін құрдық. Оқиғаларды анықтайтын триггерлердің түрлері бөлек ерекшеленген. 3.3-кестеде корпустың орыс бөлігінде табылған оқиғалардың жеті ішкі түрге бөлінуі көрсетілген. Оқиғаның триггері ретінде криминалдық лексикамен көптілді тезаурустағы етістік пен зат есімнің терминдері мен жұптары қолданылады.

Кесте 3.3 – Корпустың орыс бөлігінде табылған оқиғаларды алдын-ала анықталған жеті ішкі түрге үлестіру (оқиғалардың триггерлері етістіктер, зат есімдер және «зат есімдер + етістіктер» болып табылады)

Оқиға түрі	Оқиғаның ішкі түрі	Зат есім триггері	Етістік триггері	«Етістік + зат есім» триггері
CRIME	Injure	456	3984	298
	Offense	1972	5178	495
TRANSFER	Movement	132	507	69
	Traffic Accident	611	2351	104
POLICE	Arrest	947	9035	498
	Trial	1363	4250	1212
	PD	2217	7433	1653

Келесі кезеңде, корпустағы оқиғаларды сәйкестендіру және олардың түрлері мен ішкі түрлерін анықтағаннан кейін, 3.1-бөлімде сипатталған қатысушыларды және оқиғалардың атрибуттарын қамтитын оқиғалар атрибуттары анықталады. (3.10)-(3.13) логикалық-лингвистикалық теңдеулерін қолдана отырып, әр нақты оқиға үшін *Агент*, *Объект* және атрибуттардың рөлдері анықталады.

3.3-суретте оқиғаның әртүрлі ішкі түрлерінің тізімінің үзіндісі және олардың аргументтері көрсетілген.

File	Sentence	Trigger of event	POS of trigger	Subtype of Event	Action	Agent	Object	PLACE-ARG	TIME-ARG	1st expert	2nd expert
8188_ru_parsed	0	убить	VERB	INJURE	убить	муж	женщина		с сентября	2	1
8704_ru_parsed	1	убить	VERB	INJURE	убить	сотрудник	жена	уральске		2	2
2555_ru_parsed	2	застрелить	VERB	INJURE	застрелить	полиция	подозреваемого	нападение		1	1
2018_ru_parsed	2	застрелить	VERB	INJURE	застрелить	оппонент			3 апреля 1998 года	2	0
2485_ru_parsed	2	сбить	VERB	INJURE	сбить	автомашина	ребенок	зебра		2	1
6507_ru_parsed	3	сбить	VERB	INJURE	сбить	машина	супруг		в 9:00 утра 12 ноября	2	2
3402_ru_parsed	0	грабить	VERB	INJURE	грабить	мужчина	несовершеннолетних	талдыкорган е		2	2

Сурет 3.3 – Корпустың орыс бөлігінде табылған оқиғалар түрлері мен ішкі түрлері, сондай-ақ қатысушылар мен атрибуттар көрсетілген тізімінің фрагменті

Корпустың орыс бөлігіндегі оқиғалардың қатысушылары мен атрибуттық рөлдерін анықтаудың сенімділігі мен дәлдігі сараптамалық тексерумен расталды. Әрбір нақты оқиға үшін сарапшы: 1) оқиға түрін анықтаудың дұрыстығын; 2) оқиғаның *Агентін*, *Объектісін* және атрибуттарын сәйкестендірудің дұрыстығын анықтады.

Оқиғаны шығарып алудың дұрыстығын бағалау үш балдық шкала бойынша анықталды: 1 – оқиғаның *Агенті, Объектісі* және ішкі түрі дұрыс анықталды; 2 – алдыңғы пунктке қосымша оқиға атрибуттарының рөлдері дұрыс анықталды; 0 – *Агент* немесе *Объект* немесе оқиғаның ішкі түрі дұрыс анықталмады.

Корпустың орыс бөлігінен құқыққа қайшы оқиғаларды шығарып алу дәлдігі 73%-дан асады және 4.1-бөлімде егжей-тегжейлі сипатталған. Алынған оқиғалардың нәтижелері бойынша пайда болған онтологияға қосылатын ұғымдар және ұғымдар арасындағы қатынастар анықталды.

3.4 Параллель корпустың қазақ бөлігі негізінде нысандар мен қатынастарды автоматты түрде генерациялау әдісі

Қазақ тілін формализациялау, демек, автоматты түрде өңдеу өте қиын. Оның басты себебі түркі тілдерінің агглютинативтілігі мен жоғары флективтілігі екені сөзсіз. Бұл дегеніміз, қазақ тілінде бір түбір жүздеген сөз формаларын бере алады және әрбір сөзжасамдық морфеманың өзіндік морфологиялық немесе семантикалық мағынасы бар (мысалы, жақ, септік, жалғау, шақ, рай және т.б.). Осыған байланысты, дәстүрлі түрде белгілі ЕЕ тәсілдері негізінде оқыту корпусының қазақ бөлігіндегі оқиғалардың жеткілікті санын белгілеу өте қиын болып көрінеді.

Осы себепті корпустың қазақ бөлігіндегі оқиғаларды анықтау үшін корпустың туралanған сөйлемдері туралы білімге негізделген әдіс пайдаланылды. Әдістің дамуы бір оқиғаны әртүрлі тілдерде сипаттауға болады деген гипотезаға негізделген, бір тілдің белгіленген деректері басқа тілдегі оқиғалар, оның триггерлері, *Агент, Объект* және атрибуттары туралы ұқсас ақпаратты береді. Осылайша, біз осы белгілерді корпустың қазақ бөлігіне беру

үшін параллель корпустың орыс бөлігіндегі оқиғаның барлық элементтерінің белгілерін қолдандық.

Корпустың қазақ бөлігін өңдеудің бірінші кезеңінде О. Махамбетов пен бірлескен авторлары [159] ұсынған морфологиялық өңдеу құралдарын пайдалана отырып, бастапқы мәтіндерді POS-тегтеу жүзеге асырылды. Олар қолданатын әдіс флекциялық және деривациялық морфологияны қарастырды. Осы POS-тегтеуде қазақ тілі морфологиясының көп мағыналылығының бір бөлігін жою үшін HMM (Hidden Markov Model) негізіндегі стандартты тәсіл қолданылды.

Осындай жүргізілген морфологиялық белгілеудің нәтижесінде қазақ мәтініне күрделі морфологиялық ақпараты бар тегтер қосылды, оған сөз түбірінің POS-тегі де, сөздің әрбір морфемасымен ұсынылған морфологиялық ақпарат белгілері де енгізілді. Мысалы, «қызметкерлерінен» сөзінде <*word pos=«қызметкер_R_ZE лер_N1 і_S3 нен_C6»*> тегінде *R_ZE* белгісі жалпы зат есімді білдіреді *(Noun, common)*; *N1* белгісі көптік жалғаудың мәніне ие *(plural)*; ал *S3* белгісі жекеше түрдің үшінші жағының тәуелдік септігін білдіреді *(possessive, third singular/plural)* және *C6* көмектес септікті білдіреді *(ablative case).*

Келесі кезеңде белгіленген оқиғалар мен олардың құрамдас бөліктерін корпустың орыс бөлігінен негізге ала отырып, оқиғалардың түрлері, оқиғаларға қатысушылардың рөлдері, сондай-ақ корпустың қазақ бөлігіндегі оқиғалардың атрибуттары белгіленді. Бұл белгілерді тасымалдау екі қадамды қамтыды. Бірінші қадамда корпустың екі бөлігінде орналасқан екі тілдің мағынасы бойынша тураланған сөйлемдерін іздеу жүргізілді. Екінші қадамда, табылған әрбір қазақ тілінің сөйлемі орыс тілінің тиісті сөйлемімен бірдей қатысушылар мен дәлелдерге тән оқиғаны сипаттайтынын ескере отырып, оқиғаның ықтимал дәлелдерін іздеу үшін қазақ сөйлемінің морфологиялық белгілері мен орыс тілінің сөйлемінен алынған оқиғаның кейіпкерлері мен атрибуттарының ықтимал рөлдері арасындағы сәйкестік шаблондары қолданылды. 3.4-кестеде қазақ мәтінінің морфологиялық белгілері мен оқиға атрибуттарының тиісті рөлдері арасындағы сәйкестік шаблондары көрсетілген. Құрастырылған шаблон тегтер жиынтығына негізделген. Мұнда *R_ZE, R_ZEQ, R_BOS* және *R_ET* белгілері сәйкесінше POS-тегтерді: жалпы зат есім; жеке зат есім; шетелдік сөз және етістікті білдіреді. *C2, C3, C4, C5, C6, C7* тегтері зат есімнің ілік, барыс, табыс, жатыс, көмектес және шығыс септіктеріне сәйкес келеді; және *S** белгісі тәуелдік септігіне сәйкес келеді.

Корпустың қазақ бөлігіне белгілерді көшіру үшін қолданылатын алгоритмнің нәтижесінде криминалдық жаңалықтарға немесе полиция жұмысына қатысты оқиғалар, осы оқиғалардың триггерлері, олардың қатысушылары мен атрибуттары анықталды. Корпустың қазақ бөлігінен барлығы 443-тен астам оқиға ерекшеленді (3.5-кесте).

Кесте 3.4 – Қазақ мәтінінің морфологиялық белгілерінің және оқиға аргументтерінің рөлдерінің сәйкестік шаблондары

Оқиғалар аргументтерінің рөлдері	POS-tags	Септік тегтері	Тәуелдік жалғаудың тегтері
Agent	R_ZE, R_ZEQ, R_BOS	-	-
Object	R_ZE, R_ZEQ, R_BOS	_C4, _C2, _C3	_S*
PLACE-ARG	R_ZE, R_ZEQ	_C5, _C6, _C3	-
TIME-ARG	R_ZE	_C6	-
INSTRUMENT-ARG	R_ZE	_C7, _C3	
Trigger of Event	POS tags	Қосымша морфологиялық ақпарат	
Action	R_ET	Not ET_KSE and not ET_ESM and not ET_ETU and not ET_ETB	

Кесте 3.5 – Корпустың қазақ бөлігінен табылған оқиғалардың жеті ішкі түрге үлестірілуі

Оқиға түрі	Оқиғаның ішкі түрі	Оқиға саны
CRIME	Injure	91
	Offense	86
TRANSFER	Movement	1
	Traffic Accident	31
POLICE	Arrest	66
	Trial	69
	PD	99

3.4-суретте корпустың қазақ бөлігіндегі оқиғалар тізімінің фрагменті көрсетілген.

File	Sentence	Event	Subtype of Event	Action	Agent	Object	PLACE-ARG	TIME-ARG	1 expert results	2 expert results
2728_kz_parsed	40	POLICE	TRIAL	танысуға	құқығыңыз	хаттамамен			1	0
8542_kz_parsed	4	POLICE	PD	деп хабарлайды	қызметкерлері	қаласының			1	1
7995_kz_parsed	8	POLICE	PD	іздеуге кіріседі	полицейлер	хабарламаны			1	1
8188_kz_parsed	0	CRIME	INJURE	іздеуде болған	күйеуі	әйелді		Қыркүйектен	1	0
4650_kz_parsed	0	CRIME	INJURE	қағып кетті	мас жүргізуші	әйелді	Талдықорғанда		2	2
3706_kz_parsed	2	POLICE	TRIAL	алдады	«Әкімдік қызметкері»	тұрғындарын			2	2
8188_kz_parsed	16	CRIME	INJURE	қол жұмсап		өзіне		желтоқсан айында	0	0

Сурет 3.4 – Корпустың қазақ бөлігінде табылған оқиғалар тізімінің және олардың түрлері мен ішкі түрлері, сондай-ақ қатысушылары мен атрибуттары көрсетілген фрагменті

Оқиғаларды корпустың қазақ бөлімінде белгілеп шықпас бұрын алынған нәтижелердің дәлдігіне сараптамалық бағалау жүргізілді. Корпустың қазақ бөлігін сараптамалық бағалау корпустың орыс бөлігін бағалауға ұқсас жүргізілді және оқиғаның түрі мен шегін, оқиғаның триггерін, оның қатысушылары мен

атрибуттарын анықтау дәлдігін бағалауды қамтыды. Жоғарыда қате анықталған элементтері бар оқиға 0 деп белгіленді, дұрыс анықталған түрдегі оқиға және *Агент* пен *Объект* дұрыс анықталған оқиға 1 деп белгіленді, ал одан әрі дұрыс анықталған әрекет атрибуттары бар оқиға 2 деп бағаланды. Нәтижесінде қазақ мәтіндерінен құқыққа қайшы және криминалдық әрекеттерге байланысты оқиғаларды шығарып алудың алынған дәлдігі 55,76 құрайды (4.1-бөлімінде егжей-тегжейлі сипатталған).

3-бөлімнің қорытындысы

Үшінші бөлімде криминалистік маңызды мәтіндердің мамандандырылған корпустарын автоматты семантикалық белгілеу әдісі мен құралдарын әзірлеу қарастырылады. «Құқыққа қайшы интернет-контент» онтология кластарының экземплярларын шығарып алу үшін лингвистикалық корпустарды пайдалану ұсынылады және корпус мәтіндерінде құқыққа қайшы контенттің лингвистикалық және лексикалық маркерлерін ерекшелеп көрсету әдісі енгізіледі.

Фактілерді алудың логикалық-лингвистикалық моделіне негізделген мәтіндегі оқиғалар аргументтерінің рөлін анықтау моделі келтірілген.

Бөлімде оқиғаларды триггер түрі бойынша анықтауға және оқиғаға қатысушыларды, оқиға атрибуттарын және қатысушылардың рөлдерін алу үшін онтологияны автоматты түрде генерациялау әдістері сипатталған. Оқиғаның мұндай анықтамасы осы нысандар арасындағы қатынастың мәнін анықтауға, оларды одан әрі онтологияға жазуға мүмкіндік береді.

4 ҚҰҚЫҚҚА ҚАЙШЫ МӘТІНДІК АҚПАРАТ МОНИТОРИНГІНІҢ АҚПАРАТТЫҚ-ТАЛДАМАЛЫҚ ЖҮЙЕСІН ӘЗІРЛЕУ

4.1 Заманауи Интернеттің криминалистік маңызды мәтіндерінің корпустарынан тұжырымдамаларды автоматты түрде шығарып алу нәтижелерінің дәлдігін бағалау үшін Коэннің каппа метрикасын қолдану

Эксперимент нәтижелерін бағалауды екі сарапшы параллель корпус тілдерінің әрқайсысы үшін жүргізді. Триггер, түр/ішкі түр, *Агент* немесе *Объект* дұрыс анықталған оқиға қысқартылған CRE ретінде анықталды және сарапшы 1 деп белгіледі. Сарапшы дұрыс анықталған триггерге, түрге/ішкі түрге және оқиғаға қатысушыларға (Агент және Объект) қосымша дұрыс анықталған атрибуттар рөлдері бар оқиғаны 2 деп белгілеуі керек. Мұндай оқиға толық CRE ретінде анықталды. Ақырында, триггер, түр/ішкі түр немесе оқиғаға қатысушы сияқты негізгі элементтердің кем дегенде біреуі дұрыс анықталмаған оқиға 0 деп белгіленді. 4.1 және 4.2-суреттерде сарапшылардың орыс және қазақ тілдеріндегі оқиғаларды автоматты түрде анықтау нәтижелерін бағалау кестелерінің үзінділері көрсетілген.

File	Sentence	Trigger of event	POS of trigger	Event	Subtype of Event	Action	Agent	Object	PLACE-ARG	TIME-ARG	1 expert results	2 expert results
8188_ru_parsed	0	убить	VERB	CRIME	INJURE	убить	муж	женщина		с сентября	2	1
8704_ru_parsed	1	убить	VERB	CRIME	INJURE	убить	сотрудник	жена	уральске		2	2
2555_ru_parsed	2	застрелить	VERB	CRIME	INJURE	застрелить	полиция	подозреваемого	нападение		1	1
2018_ru_parsed	2	застрелить	VERB	CRIME	INJURE	застрелить	оппонент			3 апреля 1998 года	2	0
2485_ru_parsed	2	сбить	VERB	CRIME	INJURE	сбить	автомашина	ребенок	зебра		2	1
6507_ru_parsed	3	сбить	VERB	CRIME	INJURE	сбить	машина	супруг		в 9:00 утра 12 ноября	2	2
3402_ru_parsed	0	грабить	VERB	CRIME	INJURE	грабить	мужчина	несовершеннолетних	талдыкоргане		2	2
8801_ru_parsed	9	раскрыть	VERB	POLICE	ARREST	раскрыть	полицейский	похищение		6 февраля 2021 года	2	2
4916_ru_parsed	4	иметь	VERB	POLICE	ARREST	иметь	происшествие	место	пересечение	сегодня в 11 часов	2	1
8040_ru_parsed	0	подозревать	VERB	POLICE	ARREST	подозревать		муж	похищение		1	1
1486_ru_parsed	5	подозревать	VERB	POLICE	ARREST	подозревать	полиция		убийство		1	0
7500_ru_parsed	5	задержать	VERB	POLICE	ARREST	задержать	страж	подозреваемого	результат	5 часов	1	0
4647_ru_parsed	2	столкновение	NOUN	TRANSFER	TRAFFIC ACCIDENT	произойти	столкновение			19 февраля	2	2
6250_ru_parsed	2	наезд	NOUN	TRANSFER	TRAFFIC ACCIDENT	допустить	водитель	наезд		14:45 20 сентября	2	2
1356_ru_parsed	3	столкновение	NOUN	TRANSFER	TRAFFIC ACCIDENT	совершить	водитель	столкновение	км	3 сентября	1	1
1145_ru_parsed	4	скончаться	NOUN	CRIME	INJURE	скончаться	женщина		уральске	14:40 часов 23 октября	2	2

Сурет 4.1 – Екі сарапшының корпустың орыс бөлігінен автоматты түрде алынған қысқартылған және толық оқиғаларды бағалауының мысалы

Пайда болған онтология тұжырымдамаларын қамтитын оқиғаларды шығарып алудың дәлдігі келесі дәстүрлі (4.7) формула бойынша анықталды

$$precision = tp/(tp + fp) \quad (4.7)$$

Қысқартылған криминалистік маңызды оқиғаны ақиқат оң оқиға ретінде (true positive – TP) алу дәлдігін есептеу кезінде триггер, ішкі түр/түр, *Агент* және *Объект* дұрыс анықталған оқиға қарастырылды. Бұл жағдайда сарапшылар 1 деп белгілеген оқиғалар ескерілді.

File	Sentence	Event	Subtype of Event	Action	Agent	Object	PLACE-ARG	TIME-ARG	1 expert results	2 expert results
2728_kz_parsed	40	POLICE	TRIAL	танысуға	құқығыңыз	хаттамамен			1	0
8542_kz_parsed	4	POLICE	PD	деп хабарлайды	қызметкерлері	қаласының			1	1
7995_kz_parsed	8	POLICE	PD	іздеуге кіріседі	полицейлер	хабарламаны			1	1
8188_kz_parsed	0	CRIME	INJURE	іздеуде болған	күйеуі	әйелді		Қыркүйектен	1	0
4650_kz_parsed	0	CRIME	INJURE	қағып кетті	мас жүргізуші	әйелді	Талдықорғанда		2	2
3706_kz_parsed	2	POLICE	TRIAL	алдады	«Әкімдік қызметкері»	тұрғындарын			2	2
8188_kz_parsed	16	CRIME	INJURE	қол жұмсап		өзіне		желтоқсан айында	0	0

Сурет 4.2 – Сарапшының параллель корпустың қазақ бөлігінен автоматты түрде алынған қысқартылған және толық оқиғаларды бағалауының мысалы

Криминалистік маңызды толық оқиғаның дәлдігін есептеу кезінде біз *tp*-ді оқиғаның триггері, ішкі түрі/түрі, *Агенті* және *Объектісі* дұрыс анықталғаннан басқа, оның атрибуттары мен олардың рөлдері дұрыс анықталған оқиға деп есептедік. Бұл жағдайда біз *tp* тобының оқиғаларын, сарапшылар 2 деп белгілеген оқиғалар ретінде қарастырдық.

4.1-кестеде сәйкесінше орыс және қазақ болып табылатын параллель корпустың кіріс және шығыс тілдерінің мәтіндерінен криминалистік маңызды оқиғаларды алудың дәлдігі көрсетілген.

Кесте 4.1 – Криминалистік маңызды мәтіндердің параллель корпусының орыс және қазақ тілдері үшін алынған қысқартылған және толық оқиғалардың дәлдігі

Оқиғалар	Корпустың кіріс тілі (орыс), %	Корпустың шығыс тілі (қазақ), %
Қысқартылған криминалистік маңызды оқиғаларды шығарып алу	76.3	61.5
Толық криминалистік маңызды оқиғаларды шығарып алу	73	55.76

Тәжірибе көрсеткендей, қысқартылған CRE алу дәлдігі триггерден, оның түрінен және оқиғаға қатысушылардан басқа оқиғаның атрибуттарын қамтитын толық оқиғаларды алу дәлдігінен жоғары. Алайда, егер құрылымдық оқиғаны тек оның қатысушылары, түрі және триггері ретінде қарастыратын болсақ, дәлдік айтарлықтай өспейтіні байқалады.

Сонымен қатар, біздің эксперимент корпустың кіріс тілінің мәтіндерімен салыстырғанда корпустың шығыс тілінің мәтіндерінен оқиғаларды шығарудың дәлдігі мен толықтығының төмендеуін көрсетті. Мұның басты себебі қазақ тілінің морфологиялық белгілеуінің агглютинативтілігі мен көп мағыналылығы болып табылады.

Алынған дәлдікті алдыңғы зерттеулердің [143, p. 1366-1371; 145, p. 5753-5756; 146] нәтижелерімен салыстыра отырып, біз өте жоғары емес дәлдік мәндеріне қарамастан, біздің нәтижелер шығыс тілі үшін (PR=0.556) басқа тәсілдермен салыстырмалы және кіріс тілі үшін (PR=0.73) жақсырақ деп айта аламыз. Сонымен қатар, біздің жобада біз CRE туралы барлық ақпаратты қамтитын оқиғаларды шығарамыз, атап айтқанда: оқиға түрі/ішкі түрі, триггер, Агент, Объект, Time-ARG, PLACE-ARG және INSTRUMENT-ARG.

Сонымен қатар, алдыңғы зерттеулерден [143, 114, p. 269-282; 138, p. 516-519; 142, p. 250-253; 143, p. 1366-1371; 144, p. 231-237] айырмашылығы, оқиғалардың тек бір түрін қарастырған және осы оқиғаға қатысты тұжырымдамаларды шығарған (мысалы, жол апаттары [144, p. 231-237], жеккөрушілік қылмыстары [145, p. 5753-5756; 146] және т.б., біз CRE түрлері мен ішкі түрлерінің кең ауқымын қарастырамыз және бір уақытта алынатын қылмысқа байланысты оқиғалардың барлық түрлері үшін шығарып алу дәлдігін есептейміз. Біздің тәсілдеменің қосымша артықшылығы – ресурстар деңгейі төмен және аннотацияланған корпустар саны жеткіліксіз тілдердегі мәтіндерден оқиғаларды шығарып алу мүмкіндігі.

Алайда, болашақта эксперименттердің нәтижелеріне сілтеме жасау үшін олар мүмкіндігінше объективті және дәл болуы керек. Әдетте, белгілі бір корпуста жүргізілген эксперименттердің дұрыстығын бағалау сарапшылардың қатысуымен немесе «алтын стандарт» деп аталатын, яғни алдын-ала аннотацияланған корпуспен салыстыру арқылы жүзеге асырылады. Екі жағдайда да *recall, precision* және *F-measure* сияқты әмбебап метрикалар жиі қолданылады.

Қазіргі уақытта криминалдық тақырыпқа байланысты оқиғалардың аннотациясын қамтитын орыс және қазақ тілдерінің корпустары ашық қолжетімділікте болмағандықтан, біз эксперименттер нәтижелерінің дұрыстығын «алтын стандартқа», яғни алдын ала семантикалық аннотацияланған корпустарға сүйене отырып тексере алмаймыз. Осы себепті зерттеу нәтижелерін бағалау үшін біз сарапшылардың пікірін тарттық.

Дегенмен, сараптамалық бағалаудың дұрыстығын арттыру үшін біз осы саладағы сарапшылардың бірнеше тәуелсіз пікірлерін қолдандық, содан кейін олардың пікірлерінің сәйкестік дәрежесін тексердік. Мұндай тексеру сарапшылардың жұмысының жеткіліктілік деңгейін, сондай-ақ оларды бағалаудың сенімділігі мен объективтілігін анықтауға мүмкіндік берді. Сарапшылардың пікірлерінің әлсіз келісушілігі сарапшылардың жұмысының қисынсыздығын көрсетіп, сәйкесінше бүкіл зерттеудің нәтижелеріне күмән келтіруі мүмкін.

Зерттеуіміздің дұрыстығын арттыру үшін біз сарапшылардың келісімін Коэннің каппа коэффициенті арқылы есептедік [160, 161]. Бұл коэффициент кездейсоқ алынған екі сарапшы арасындағы номиналды келісім шкаласын өлшеу үшін қолданылады.

Біз орыс және қазақ тілдеріндегі ішкі корпустардың мәтіндерінен бөлек алынатын оқиғалардың екі деңгейі (қысқартылған және толық CRE) үшін Коэннің каппа коэффициенттерін есептедік. Жоғарыда айтылғандай,

сарапшылардан оқиғаны анықтау нәтижелерін үш ықтимал нұсқаны ескеретін рейтингтік шкаланың көмегімен бағалау ұсынылды: 1 – қысқартылған CRE дұрыс анықталуы, 2 – толық CRE анықталуы және 0 – егер оқиғаға қатысушылардың кем дегенде біреуі, триггер немесе оқиға ішкі түрі дұрыс анықталмауы.

4.2-кестеде сарапшылардың қазақ-орыс параллель корпусынан криминалистік маңызды оқиғаларды шығарып алудың дұрыстығын бағалаудың шатасу матрицасы келтірілген. Кесте жолдарында параллель корпустың кіріс және шығыс бөліктерінің қысқартылған және толық CRE үшін бірінші сарапшының пікірі, ал бағандарда екінші сарапшының пікірі берілген.

Кесте 4.2 – Қазақ-орыс параллель корпусынан қысқартылған және толық криминалистік маңызды оқиғалар үшін сарапшылардың пікірлерін бағалаудың шатасу матрицасы

Қысқартылған CRE шығарып алу				
бірінші сарапшы	кіріс тілі (орыс)		шығыс тілі (қазақ)	
	екінші сарапшы			
	“0”	“1”	“0”	“1”
“0”	93	36	164	72
“1”	15	256	7	257
Толық CRE шығарып алу				
бірінші сарапшы	кіріс тілі (орыс)		шығыс тілі (қазақ)	
	екінші сарапшы			
	“0”	“2”	“0”	“2”
“0”	117	8	192	11
“2”	28	347	48	249

4.3-кестеде осы матрица негізінде есептелген екі тілдің қысқартылған және толық CRE үшін Коэннің каппа коэффициенттері берілген.

Кесте 4.3 – Коэннің каппа келісім коэффициенттері

Оқиғалар	Кіріс тілі (орыс), %	Шығыс тілі (қазақ), %
Қысқартылған CRE	89.8	84.2
Толық CRE	92.8	88.2

Коэннің каппа коэффициенттерін бағалау үшін қолданылатын жалпы қабылданған шкала бойынша [101, p. 362-371]:

– 0,81-нен 0,99-ға дейінгі коэффициенттің мәні сарапшылардың пікірлерінің мінсіз келісімділігін көрсетеді;

– 0,60-тан 0,80-ге дейінгі коэффициенттің мәні пікірлердің айтарлықтай келісімділігін көрсетеді;

– 0,41-ден 0,60-қа дейінгі коэффициенттің мәні сарапшылардың пікірлерінің орташа келісімділігін көрсетеді;

– 0,21-ден 0,40-қа дейінгі коэффициенттің мәні сараптамалық пікірлердің қанағаттанарлық келісімділігін көрсетеді.

Берілген шкалаға сүйене отырып, біз Коэннің каппа коэффициенттерінің алынған мәндері біздің зерттеуіміздің сенімділігін арттыруға ықпал етеді деп айта аламыз. Дегенмен, сарапшылардың пікірінің келісімділік коэффициенттерінің мәндері көп үміт күтерлік болып көрінгенімен, оқиғалардың семантикалық аннотациясын қамтитын корпустар негізінде сенімділікті тексеру мүмкіндігінің артуымен жұмыстың келесі кезеңінде одан әрі зерттеу жалғасуы керек екені анық.

4.2 Әзірленген технологияның тиімділігін эксперименттік дәлелдеу

Құқыққа қайшы әрекетке байланысты оқиғалар корпусындағы семантикалық белгілеу экспериментінің нәтижелерін бағалауды әр тіл үшін екі сарапшы жүргізді. Триггер, түр/ішкі түр, *Агент* немесе *Объект* дұрыс анықталған оқиға *short CRE* ретінде анықталды және сарапшы «1» деп белгіледі. Осы элементтерден басқа атрибуттардың рөлдері дұрыс анықталған оқиға сарапшы «2» деп белгіледі және *complete CRE* ретінде анықтады. Элементтердің кем дегенде біреуі анықталған оқиғаны сарапшы «0» деп белгіледі.

Құрылған онтология тұжырымдамаларын қамтитын оқиғалардың белгілеу дәлдігі келесі (4.7) дәстүрлі формула бойынша анықталды. *Complete CRE* үшін шығарып алу дәлдігін есептей отырып, біз «2» деп белгіленген ақиқат оң (tp) оқиғалар ретінде қарастырдық. Сарапшылар «0» немесе «1» деп белгілеген оқиғаларды false positive (fp) деп белгіледік. Біздің *complete CRE* анықтауына сәйкес, *short CRE* және оқиға атрибуттарын қамтиды. Онда *short CRE* true positive (tp) үшін сарапшылар «2» немесе «1» деп белгілеген оқиғалардың саны болып табылады. Ал false positive (fp) сарапшылар «0»деп белгілеген оқиғалардың саны ретінде анықталады. 4.4-кестеде орыс және қазақ тілдерінің корпустары үшін CRE анықтамасының дәлдігі (precision) көрсетілген.

Кесте 4.4 – Қазақ және орыс тілдері мәтіндерінің корпустарында CRE анықтау дәлдігі (precision)

Оқиға	Орыс тіліндегі мәтіндер корпусы, %	Қазақ тіліндегі мәтіндер корпусы, %
short CRE	76.30	61.50
complete CRE	73.00	55.76

Толықтықты (recall) анықтау құқыққа қайшы әрекетке байланысты оқиғаларды шығарып алу және белгілеу, егер сөйлемде «Құқыққа қайшы интернет-контент» онтологиясынан алынған оқиға триггері болса, ол құқыққа қайшы әрекетке байланысты кейбір оқиғаны сипаттауы керек деген гипотезаға негізделген. Осылайша, біз қазақ және орыс тілдерінің корпустарынан криминалдық сипаттағы триггер етістіктері бар кездейсоқ таңдалған 500 сөйлемнен қанша оқиға алынғанын тексердік және келесі дәстүрлі теңдеуді қолдана отырып толықтығын есептедік:

$$recall = \frac{tp}{tp + fn} \quad (4.8)$$

Бұл жағдайда біз *short CRE* және *complete CRE* оқиғаларының екі түрін де ақиқат оң деп санадық (true positive - TP). 4.5-кестеде сәйкесінше орыс және қазақ тіліндегі корпустағы шығарып алу және белгілеудің *recall* және *F1-measure* көрсетілген.

Кесте 4.5 – Қазақ және орыс мәтіндік корпусындағы CRE анықтауының толықтығы (recall) және F_1-measure

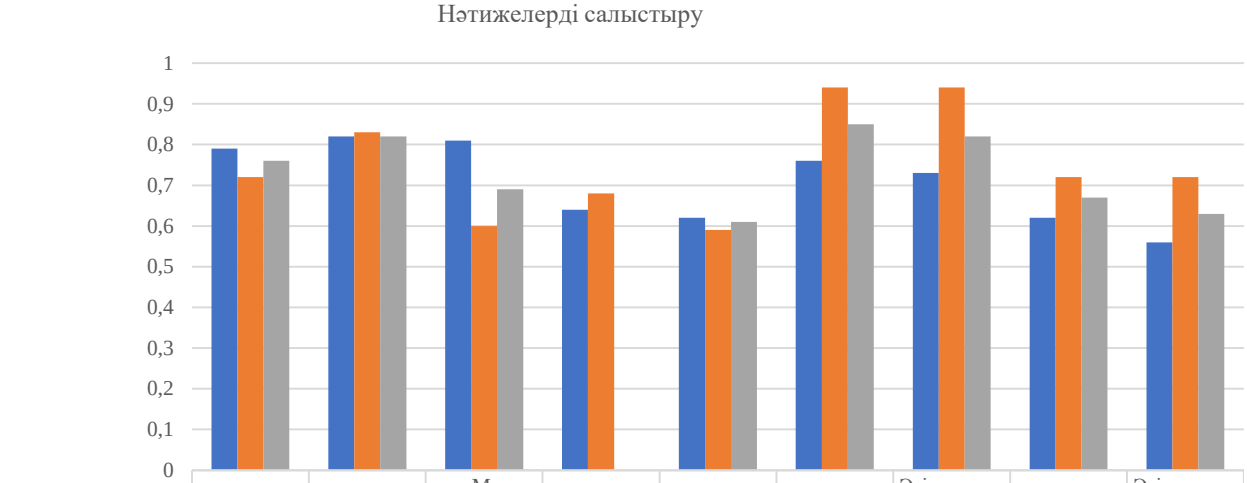
Метрика	Орыс тіліндегі мәтіндер корпусы, %	Қазақ тіліндегі мәтіндер корпусы, %
Толықтық (recall)	94.80	72.40
F_1 short CRE	84.55	66.51
F_1 complete CRE	82.48	63.00

Жүргізілген талдау орыс тілімен салыстырғанда қазақ тілін өңдеудің дәлдігі мен толықтығының төмендегенін көрсетті. Мұның басты себебі қазақ тілі морфологиясының агглютинативтілігі мен көп мағыналылығы болып табылады. Сонымен қатар, *short* CRE шығарып алу дәлдігі мен F1-өлшемі *complete* CRE алуына қарағанда жоғары, бұл қысқа оқиғаны көрсетуге қарағанда *complete CRE-де* элементтер санының көбірек болуына байланысты екені анық.

Кесте 4.6 – Әзірленген технологияның тиімділігін оқиғаларды бақылаудың басқа тәсілдемелерімен салыстыру

Тәсілдемелер мен технологиялар	Дәлдік (precision)	Толықтық (recall)	F_1-measore	Тіл	Оқиға түрлері	Оқиға аргументтері
S. Yagcioglu et al.	0.79	0.72	0.76	Ағылшын	Cyber-security	Оқиғаның ішкі түрлері
P. Das and A. K. Das,	0.82	0.83	0.82	Ағылшын	Hate crime	Тек жеке
M. Sotomayor, and F. Veloz	0.81	0.60	0.69	Испан	Crime	Тек бірнеше
K.T. Hossain et al.	0.64	0.68	-	Араб	MANSA	Агент және объект
F Rahma and A Romadhony	0.62	0.59	0.61	Индонезия	Әртүрлі	Complete CRE
Әзірленген технология	0.76/0.73 0.62/0.56	0.94 0.72	0.85/0.82 0.67/0.63	Орыс Қазақ	Әртүрлі	Short CRE/complete CRE Short CRE/ complete CRE
Ескерту – Әдебиет негізінде құралған [114, p. 269-282; 133, p. 55-75; 137, p. 10-14; 139, p. 1-4; 143, p. 1366-1371]						

4.6-кестеде және 4.3-суретте әзірленген технологияның тиімділігін құқыққа қайшы әрекетке байланысты оқиғаларды бақылаудың басқа тәсілдерімен салыстыру келтірілген. Алынған толықтықты, дәлдікті және F1 көрсеткішті алдыңғы зерттеулермен [143, p. 1366-1371; 145, p. 5753-5756] салыстыра отырып, коэффициенттердің өте жоғары емес мәндерін алғанымен, біздің нәтижелеріміз қазақ тілі үшін, ал кейде орыс тілі үшін жақсырақ деп айтуға болады. Алайда, біздің жағдайда біз құқыққа қайшы әрекетке байланысты оқиға туралы барлық мүмкін ақпаратты қамтитын оқиғаларды (*complete CRE*) шығарып аламыз, атап айтқанда түр/ішкі түр, триггер, *Агент*, *Объект*, Time-ARG, PLACE-ARG және INSTRUMENT-ARG аргументтері.



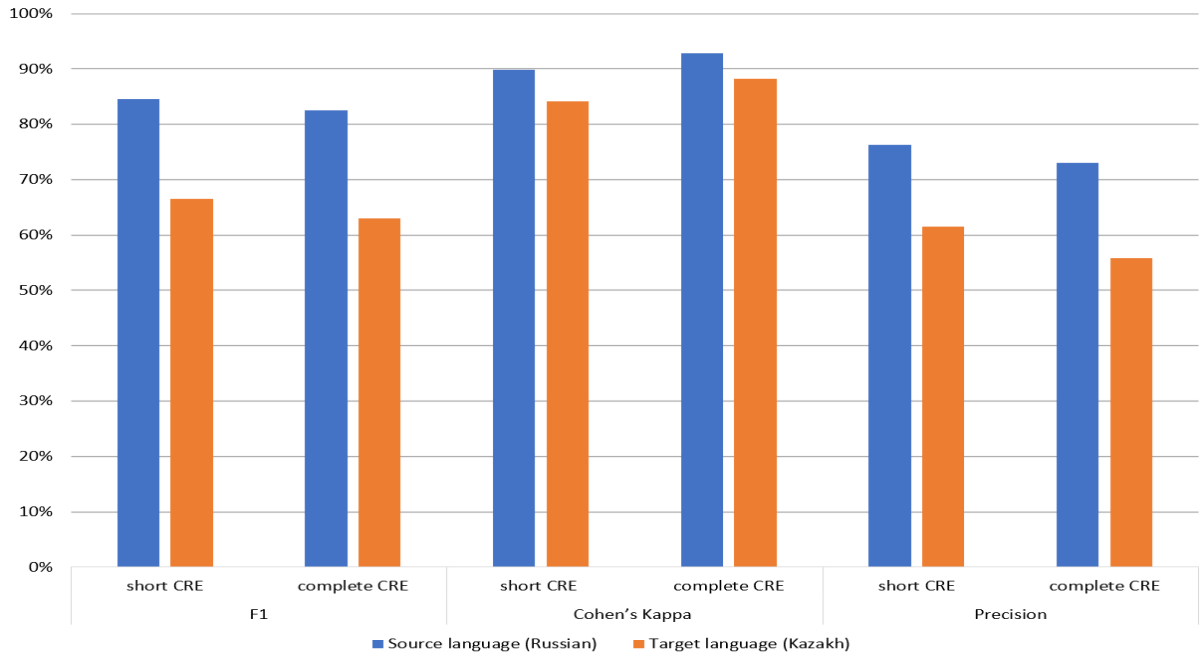
	S. Yagcioglu (2019 ж.)	P. Das and A.K. Das (2019 ж.)	M. Sotomayor and F. Veloz. (2017 ж.)	K.T. Hossain және басқ. (2020 ж.)	F. Rahma and A. Romadhony (2021 ж.)	Әзірленген технология (орыс) Short CRE	Әзірленген технология (орыс) Complete CRE	Әзірленген технология (қазақ) Short CRE	Әзірленген технология (қазақ) Complete CRE
Дәлдік (precision)	0,79	0,82	0,81	0,64	0,62	0,76	0,73	0,62	0,56
Толықтық (recall)	0,72	0,83	0,6	0,68	0,59	0,94	0,94	0,72	0,72
F1-measore	0,76	0,82	0,69	0	0,61	0,85	0,82	0,67	0,63

Сурет 4.3 – Әзірленген технологияның тиімділігін құқыққа қайшы әрекетке байланысты оқиғаларды бақылаудың басқа тәсілдерімен салыстыру

Бұған қоса, жек көрушілік қылмыстары [90, p. 313-324] және т.б. сияқты оқиғаның бір түрін ғана қарастырған алдыңғы зерттеулерден [55, p. 1-4; 114, p. 269-282; 137, p. 10-14; 142, p. 250-253] айырмашылығы, біз құқыққа қайшы әрекеттермен байланысты оқиға түрлері мен ішкі түрлерінің кең ауқымын қарастырамыз және қылмыспен байланысты оқиғаның барлық түрлерін шығарып алудың жалпы дәлдігін есептейміз. Біздің тәсілдеменің қосымша артықшылығы – оқиғаларды төмен ресурстық және жеткілікті аннотацияланбаған тілдердегі мәтіндерден шығарып алу мүмкіндігі.

Біздің зерттеуіміз осындай тілдерге қатысты болғандықтан және біз «алтын стандартқа» негізделген эксперименттердің нәтижелерін растай алмағандықтан, зерттеу нәтижелерін бағалау үшін екі сарапшының пікірлері пайдаланылды, содан кейін пікірлердің келісімділік деңгейін бағалау Коэннің каппа коэффициентімен тексерілді [126, p. 9503413-1-9503413-8]. 4.4-суреттің

диаграммасы қазақ және орыс тілдеріндегі *short* және *complete CRE* үшін шығарып алудың Коэннің каппа коэффициенттерінің мәндерін көрсетеді.



Сурет 4.4 – Эксперименттердің дұрыстығын бағалау өлшемдерінің жинақтамасы

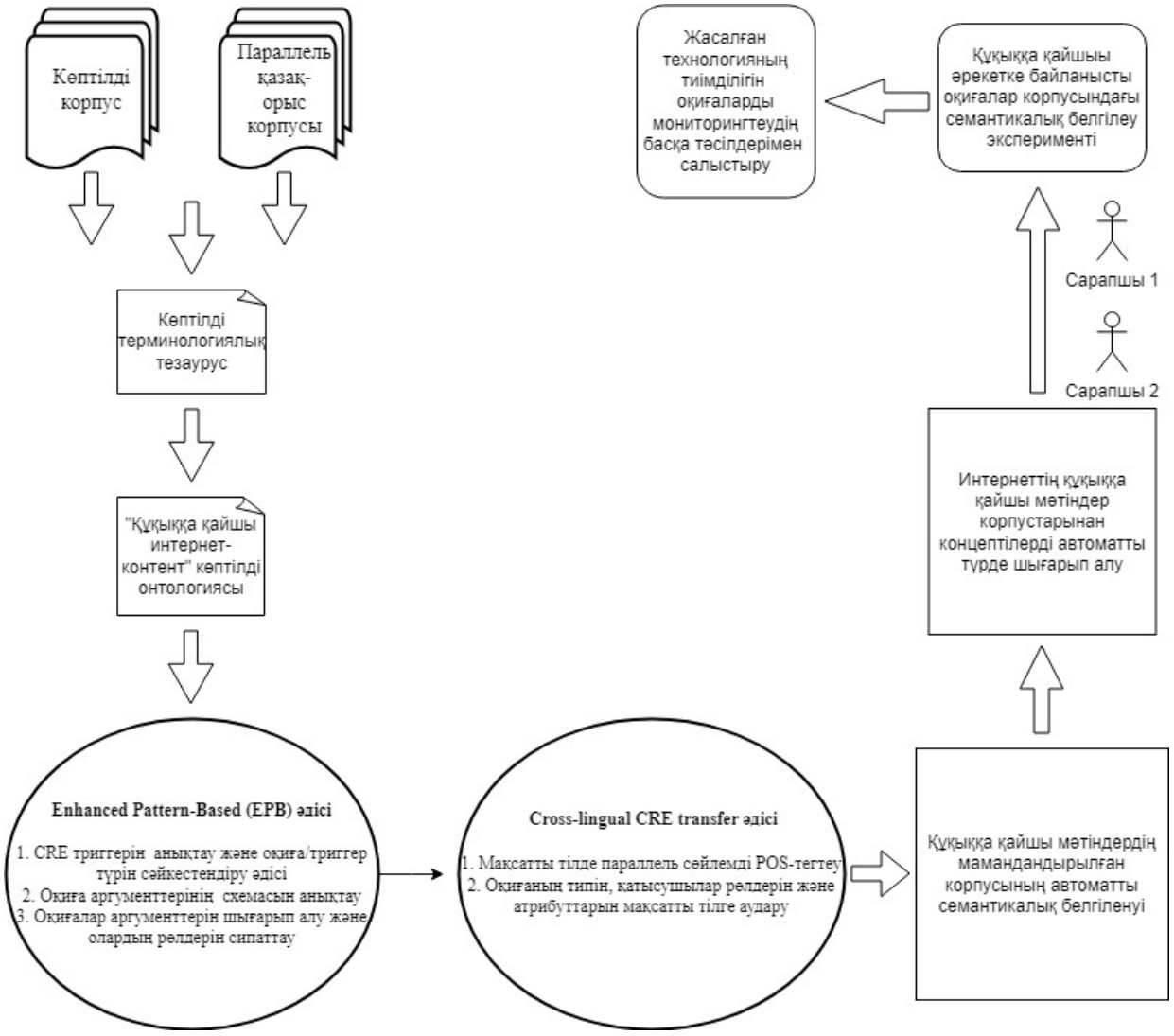
4.5-суретте Интернет желілерде қазақ және орыс тілдерінің құқыққа қайшы мәтіндерін автоматты сәйкестендіру жүйесінің ақпараттық моделінің жалпы схемасы көрсетілген. Оған мыналар кіреді:

– криминалдық жаңалықтарды қамтитын мәтіндерден тұратын арнайы әзірленген мәтіндер корпустары. Бұл орыс, украин, ағылшын тілдерінің мәтіндерін қамтитын көптілді корпусы және параллель қазақ-орыс корпусы [1, p. 116-124];

– XML форматында әзірленген криминалдық лексикасы бар көптілді терминологиялық тезаурус [148, p. 108-116] және динамикалық және интерактивті веб-қосымша ретінде әзірленген «Құқыққа қайшы интернет-контент» онтологиясы;

– криминалистік маңызды мәтіндердің мамандандырылған корпустарын автоматты семантикалық белгілеудің бағдарламалық құралы;

– әлеуметтік желілерде және басқа интернет көздерінде құқыққа қайшы контентті талдау мен мониторингтеудің интеграцияланған технологиясын әзірлеу.



Сурет 4.5 – Интернет желілерде қазақ және орыс тілдерінің құқыққа қайшы мәтіндерін автоматты сәйкестендіру жүйесінің ақпараттық моделінің жалпы схемасы

4-бөлімнің қорытындысы

Төртінші тарауда машиналық оқыту әдістері мен онтологиялық тәсілді қамтитын әлеуметтік желілердегі және басқа интернет көздеріндегі құқыққа қайшы контентті іздеу мен талдаудың интеграцияланған технологиясы қарастырылады. Жұмыста қолданылатын онтологиялық тәсіл деректерді тиімді құрылымдау мен классификациялауды қамтамасыз ететін семантикалық желілер түрінде құқыққа қайшы контент туралы білімді формалдауға мүмкіндік береді, ал жұмыста машиналық оқыту әдістерін қолдану үлкен көлемдегі ақпаратты өңдеуге, құқыққа қайшы контенттің жаңа түрлерін анықтауға және өзгеретін жағдайларға бейімделуге мүмкіндік беретін модельдің мүмкіндіктерін толықтырады және кеңейтеді. Бұл тәсілдің артықшылықтары: (1) мәтіндік ақпаратты талдаудың жоғары дәлдігінде, (2) құқыққа қайшы әрекетке байланысты ақпаратты іздеу және талдау процестерін автоматтандыруда және (3) жүйенің өзгеретін белгілер мен сипаттамаларға бейімделуінде.

Сонымен қатар, әзірленген ақпараттық-талдамалық жүйесінің бірнеше тілдегі, соның ішінде қазақ, орыс және ағылшын тілдеріндегі мәтіндерді талдау мүмкіндігі жүйені әмбебап етеді және оның тілі мен локализациясына қарамастан, құқыққа қайшы мазмұнды анықтауға мүмкіндік беретін әртүрлі ақпарат көздерімен жұмыс істеу үшін қолданылады.

Бұл тарауда қазіргі Интернеттің криминалистік маңызды мәтіндерінің корпустарынан тұжырымдамаларды автоматты түрде шығарып алу нәтижелерінің дәлдігін бағалау үшін Коэннің каппа метрикасын қолдану көрсетілген.

Әзірленген технологияның тиімділігінің эксперименттік дәлелі де сипатталған. Әзірленген технологияның тиімділігін оқиғаларды бақылаудың басқа тәсілдерімен салыстыру келтірілген. Ағылшын және испан тілдері үшін көрсеткіштер precision – 0.79, 0.82, 0.81, recall – 0.72, 0,83, 0.60, F_1-measure – 0.76, 0.82, 0.69 аралықтарында болса, араб және индонезия тілдері үшін precision – 0.64, 0.62, recall – 0.68, 0.59, F_1-measure – 0.61 мәндерін көрсетсе, біздің әзірлеген технологиямен алынған көрсеткіштер орыс тілі үшін precision – 0.73 және 0,76 recall – 0.94, F_1-measure – 0.82 және 0.85, ал қазақ тілі үшін precision – 0.56 және 0.62, recall – 0.72. F_1-measure – 0.67 және 0.63 мәндерін береді.

ҚОРЫТЫНДЫ

Онтологиялық тәсіл негізінде құқыққа қайшы мәтіндік ақпарат мониторингінің ақпараттық-талдамалық жүйесін әзірлеу бойынша диссертациялық жұмысты орындау барысында келесі ғылыми және практикалық нәтижелер алынды:

1. Онтологиялық тәсіл негізінде құқыққа қайшы мәтіндік ақпарат мониторингінің заманауи әдістері мен модельдеріне талдау жүргізілді.

2. Екі корпус және көптілді терминологиялық тезаурус құрылды:

– орыс, украин және ағылшын тілдеріндегі мәтіндерді (украин тілінде 3147 мәтін, орыс тілінде 5506 мәтін, ағылшын тілінде 300 мәтін) қамтитын көптілді корпус;

– орыс тіліндегі 3000 мәтінді және қазақ тіліндегі 3000 мәтінді, оның ішінде мағынасы жағынан тураланған қазақша-орысша сөйлемдерді құрайтын 2000 мәтінді қамтитын параллель қазақ-орыс корпусы;

– 600-ден астам негізгі сөздерді (330 зат есім, 107 сын есім және 170-ке жуық етістік) және 2500-ден астам негізгі сөз синонимдерін қамтитын тезаурус.

3. «Құқыққа қайшы интернет-контент» онтологиясы құрылды.

4. Криминалистік маңызды мәтіндердің мамандандырылған корпустарын автоматты семантикалық белгілеу әдісі мен құралдары әзірленді.

5. Онтологиялық тәсіл негізінде көптілді құқыққа қайшы интернет-контентті автоматты іздеу және талдау жүйесінің ақпараттық моделі мен бағдарламалық құралы әзірленді.

Бұл диссертацияда онтологиялық тәсіл мен машиналық оқыту әдістері негізінде ақпараттық-талдамалық жүйе жасалды. Білімді онтология түрінде формалдауға негізделген онтологиялық тәсіл деректерді құрылымдау мен классификациялаудың қуатты құралы болып табылады. Онтология мәтіндік контентті талдау және берілген ақпараттық домен мәтінінің әртүрлі элементтері арасындағы байланыстарды анықтау үшін тиімді пайдалануға болатын білімнің семантикалық графиктері түріндегі құқыққа қайшы әрекеттер туралы білімді ұсынуға мүмкіндік береді. «Құқыққа қайшы интернет-контент» онтологиясын құру Интернет ортасында таратылатын көптілді құқыққа қайшы мәтіндік ақпаратты автоматты түрде іздеу мен талдаудың дәлдігін арттыруға мүмкіндік береді. Интернеттегі ақпараттың үлкен көлемін өңдеу және құқыққа қайшы контенттің жаңа және күрделі түрлерін анықтау үшін машиналық оқыту әдістерін қолдану жүйеге үлкен көлемдегі мәліметтер негізінде «үйренуге», күрделі заңдылықтарды анықтауға және мазмұнның өзгеруіне бейімделуге мүмкіндік береді.

Алынған нәтижелер ақпараттық қауіпсіздік саласында маңызды мәнге ие және мүмкін болатын құқыққа қайшы қауіптерді бағалау саласында, оның ішінде мінез-құлықты талдау бойынша мамандармен де қолданылуы мүмкін. Жұмыс сәтті жүзеге асырылды, табиғи тілді түсіну саласында маңызды үлес қосады және одан әрі зерттеу мен практикалық міндеттерде қолдану перспективаларын ашады.

Зерттеу нәтижелері бойынша барлығы 11 жұмыс жарияланды, оның ішінде 3 мақала ҚР ҒЖБМ Ғылым және жоғары білім саласындағы сапаны қамтамасыз ету Комитеті ұсынған журналдарда, 2 мақала Scopus және Web of Science базасымен индекстелген басылымдарда, 6 – халықаралық конференцияларда жарияланды және 1 монография жарыққа шықты. Сондай-ақ, 2 авторлық куәлік және диссертациялық зерттеу нәтижелерін енгізу туралы акт (Қосымша Б) алынды.

Осылайша, диссертацияда қойылған зерттеу міндеттері толығымен шешілді. Мәтіннен фактілер триплеті түрінде білім алатын жүйенің және диссертациялық зерттеуде әзірленген модельдер мен алгоритмдерге негізделген мәтіннің тар шеңберде мамандандырылған тақырыпқа семантикалық жақындығын анықтайтын жүйенің жұмыс сапасын бағдарламалық қамтамасыз ету және бағалау жақсы нәтиже көрсетеді.

ПАЙДАЛАНЫЛҒАН ӘДЕБИЕТТЕР ТІЗІМІ

1 Khairova N., Kolesnyk A., Mamyrbayev O. et al. The aligned Kazakh-Russian parallel corpus focused on the criminal theme // CEUR Workshop Proceedings. – Lviv, 2019. – P. 116-125.

2 Mena J. Machine Learning Forensics for Law Enforcement, Security, and Intelligence. – Boca Raton, 2011. – 350 p.

3 Mamyrbayev O., Kydyrbekova A., Alimhan K. et al. Development of security systems using DNN and i & x-vector classifiers // Eastern-European Journal of Enterprise Technologies. – 2021. – Vol. 4/9, Issue 112. – P. 32-45.

4 Cohen K., Johansson F., Kaati L. et al. Clausen Mork. Detecting Linguistic Markers for Radical Violence in Social Media // Terrorism and Political Violence. – 2014. – Vol. 26, Issue 1. – P. 246-256.

5 Pennebaker W.J., Boyd R.L., Jordan K. et al. The development and psychometric properties of LIWC2015. – Austin, TX, 2015. – 27 p.

6 Hamm M.S., Spaaij R.FJ., Cottee S. The age of lone wolf terrorism. – NY: Columbia University Press, 2017. – 322 p.

7 Cohen K., Kaati L., Shrestha A. et al. Linguistic markers of a radicalized mind-set among extreme adopters // Procced. 10st internat. conf. on Web Search and Data Mining. – NY., 2017. – 823-824.

8 Kaufhold M. A., Reuter C. Cultural Violence and Peace in Social Media // Information Technology for Peace and Security. – 2019. – P. 361-381.

9 Goldberg Y. et al. word2vec Explained: deriving Mikolov et al.'s negative-sampling word-embedding method // https://arxiv.org/abs/1402.3722. 25.01.2021.

10 Schuurman B., Bakker E., Gill P. et al. Lone Actor Terrorist Attack Planning and Preparation: A Data-Driven Analysis // Journal of Forensic Sciences. – 2018. – Vol. 63, Issue 4. – P. 1191-1120.

11 Fu T., Abbasi A., Zeng D. et al. Sentimental spidering: leveraging opinion information in focused crawlers // ACM Transactions on Information Systems (TOIS). – 2012. – Vol. 30, Issue 4. – P. 1-30.

12 Scrivens R., Davies G., Frank R. Searching for signs of extremism on the web: an introduction to Sentiment-based Identification of Radical Authors // Behavioral sciences of terrorism and political aggression. – 2018. – Vol. 10, Issue 1. – P. 39-59.

13 Nouh M., Nurse RC.J., Goldsmith M. Understanding the radical mind: Identifying signals to detect extremist content on twitter // Procced. 2019 IEEE internat. conf. on Intelligence and Security Informatics (ISI). – Shenzhen, 2019. – P. 98-103.

14 Ashcroft M., Fisher A., Kaati L. et al. Detecting jihadist messages on twitter // Intelligence and Security Informatics Conference (EISIC). – Manchester, 2015. – P. 161-164.

15 Finlayson M.A., Halverson J.R., Corman S.R. The N2 Corpus: A Semantically Annotated Collection of Islamist Extremist Stories // Proceed. of the Ninth internat. conf. on Language Resources and Evaluation. – Reykjavik, 2014. – P. 896-902.

16 Wadhwa P., Bhatia M.P.S. Classification of radical messages in Twitter using security associations // In book: Case studies in secure computing: Achievements and trends. – NY., 2014. – P. 273-294.

17 Brynielsson J., Horndahl A., Johansson F. et al. Harvesting and analysis of weak signals for detecting lone wolf terrorists // Security Informatics. – 2013. – Vol. 2, Issue 1. – P. 11-26.

18 Bessmertny I. Knowledge visualization based on semantic networks // Program Comp. Soft. – 2010. – Vol. 36, Issue 4. – P. 197-205.

19 Добров Б.В., Иванов В.В., Лукашевич Н.В. и др. Онтологии и тезаурусы: модели, иснтрументы, приложения: учеб. пос. – М., 2009. – 173 с.

20 Nirenburg S., Raskin V. Ontological Semantics. – M.: MIT Press, 2004. – 82 p.

21 Fillmore C.J., Miriam R.L, Petruck J.R. et al. Framenet in Action: The Case of Attaching // International Journal of Lexicography. – 2003. – Vol. 16.3. – P. 297-332.

22 Adderley R., Seidler P., Badii A. et al. Semantic Mining and Analysis of Heterogeneous Data for Novel Intelligence Insights // Procced. 4th internat. conf. on Advances in Information Mining and Management. – Paris, 2014. – P. 36-40.

23 Casanovas P., Arraiza J., Melero F. et al. Fighting Organized Crime Through OpenSource Intelligence: Regulatory Strategies of the CAPER Project // Front. Artif. Intell. Appl. – 2014. – Vol. 271. – P. 189-198.

24 Brewster B., Andrews S., Polovina S. et al. Environmental scanning and knowledge representation for the detection of organised crime threats // Procced. 21st internat. conf. on Conceptual Structures (ICCS 2014). – Iași, 2014. – P. 275-280.

25 COPKIT project // https://copkit.eu/. 15.05.2021.

26 ASGARD project // https://www.asgard-project.eu. 15.05.2021.

27 TENSOR project // https://tensor-project.eu. 15.05.2021.

28 Rich ERE Annotation Guidelines Overview, Linguistic Data Consortium // https://catalog.ldc.upenn.edu/LDC2016T23. 20.05.2021.

29 Ramponi A., Plank B., Lombardo R. Cross-Domain Evaluation of Edge Detection for Biomedical Event Extraction // Procced. International Conference on Language Resources and Evaluation (LREC). – Marseille, 2020. – P. 1982-1989

30 Ding X. et al. Research on typical event extraction method in the field of music // Journal of Chinese Information Processing. – 2011. – Vol. 25, Issue 2. – P.15-20.

31 Osathitporn P., Soonthornphisaj N., Vatanawood W. A scheme of criminal law knowledge acquisition using ontology // Procced. of the 18th IEEE/ACIS internat. conf. on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing (SNPD). – Kanazawa, 2017. – P. 29-34.

32 Rosa F.F., Jino M., Bonacin R. Towards an ontology of security assessment: a Core model proposal // Information Technology-New Generations. – 2018. – Vol. 738. – P. 75-80.

33 Poletto F. et al. Resources and benchmark corpora for hate speech detection: a systematic review // Language Resources and Evaluation. – 2021. – Vol. 55, Issue 2. – P. 477-523.

34 Çöltekin Ç. A Corpus of Turkish Offensive Language on Social Media // in Proc. of the 12th Conference on Language Resources and Evaluation (LREC). – Marseille, 2020. – P. 6174-6184.

35 Salminen J. et al. Anatomy of online hate: developing a taxonomy and machine learning models for identifying and classifying hate in online news media // Procced. of the Twelfth internat. AAAI conf. on Web and Social Media. – Stanford, 2018. – P. 330-339.

36 Sanguinetti M. et al. An italian Twitter corpus of hate speech against immigrants // Procced. of the 11th internat. conf. on Language Resources nd Evaluation (LREC). – Miyazaki, 2018. – P. 2798-2805.

37 Bassignana E., Basile V., Patti V. Hurtlex: A Multilingual Lexicon of Words to Hurt // Procced. of 5th Italian conf. on Computational Linguistics (CLiC-it). – Torino, 2018. – P. 1-6.

38 Kumar R., Reganti A.N., Bhatia A. et al. Aggression-annotated Corpus of Hindi-English Code-mixed Data // Procced of the 11th Language Resources and Evaluation conf. (LREC). – Miyazaki, 2018. – P.1425-1431.

39 Battistelli D., Bruneau C., Dragos V. Building a formal model for hate detection in french corpora // Procedia Computer Science. – 2020. – Vol. 176. – P. 2358-2365.

40 Zampieri M. et al. SemEval-2019 Task 6: Identifying and categorizing offensive language in social media (OffensEval) // Procced. of the 13th internat. workshop on semantic evaluation (SemEval-2019). – Minneapolis, 2019. – P. 75-86.

41 Zampieri M. et al. SemEval-2020 Task 12: Multilingual Offensive Language Identification in Social Media (OffensEval 2020) // in Procced. of the 14th internat. workshop on semantic evaluation. – Barcelona, 2020. – P. 1425-1447.

42 Zampieri M. et al. Predicting the type and target of offensive posts in social media // Procced of the Annual conf. of the North American Chapter of the Association for Computational Linguistics: Human Language Technology (NAACL-HLT). – Minneapolis, 2019. – P. 1415-1420.

43 Pontrandolfo G. Phraseology in criminal judgments: A corpus study of original vs. translated Italian // Sendebar. – 2011. – Vol. 22. – P. 209-234.

44 Pontrandolfo G. Corpus Methods in Legal Translation Studies // In book: Research Methods in Legal Translation and Interpreting. – London, 2019. – 232 p.

45 Goźdź-Roszkowski S. Corpus linguistics in legal discourse // International Journal for the Semiotics of Law-Revue internationale de Smiotique juridique. – 2021. – Vol. 34, Issue 5. – P. 1515-1540.

46 Ramos F.P. The use of corpora in legal and institutional translation studies: Directions and applications // Translation Spaces. – 2019. – Vol. 8, Issue 1. – P. 1-11.

47 Taylor A.V. MWCC: A Corpus of Malawi Criminal Cases // Procced. conf.: Natural Legal Language Processing Workshop 2020KDD. – San Diego, 2020. – P. 39-47.

48 Ras I.A. A Corpus-Assisted Critical Discourse Analysis of the Reporting on Corporate Fraud by UK Newspapers 2004-2014: thes. … doc. PhD. – Leeds: University of Leeds, 2017. – 255 p.

49 Mukherjee S., Sarkar K. Analyzing Large News Corpus Using Text Mining Techniques for Recognizing High Crime Prone Areas // Procced. of the 2020 IEEE Calcutta conf. (CALCON). – Kolkata, 2020. – P. 444-450.

50 de Carvalho V.D.H., Costa A.P.C.S. Exploring Text Mining and Analytics for Applications in Public Security: An in-depth dive into a systematic literature review // https://preprints.scielo.org/index.php/scielo/preprint/view. 01.06.2021.

51 Karystianis G. et al. Automated analysis of domestic violence police reports to explore abuse types and victim injuries: Text mining study // Journal of Medical Internet Research. – 2019. – Vol. 21, Issue 3. – P. e13067.

52 Adily A., Karystianis G., Butler T. Text mining police narratives for mentions of mental disorders in family and domestic violence events // Trends and Issues in Crime and Criminal Justice. – 2021. – Vol. 7. – P. 629-1-629-6.

53 Karystianis G. et al. Automatic Extraction of Mental Health Disorders From Domestic Violence Police Narratives: Text Mining Study // Journal of Medical Internet Research. – 2018. – Vol. 20, Issue 9. – P. e11548.

54 Alagheband M.R., Mashatan A. et al. Time-based gap analysis of cybersecurity trends in academic and digital media // ACM Transactions on Management Information Systems (TMIS). – 2020. – Vol. 11, Issue 4. – P. 1-20.

55 Gunawan D. et al. Building the Pornography Corpus for Bahasa Indonesia Based on TRUST+™ Positif Database // Procced. 2019 internat. conf. on ICT for Smart Society (ICISS). – Bandung, 2019. – P. 1-5.

56 de Mendonça R.R. et al. A framework for detecting intentions of criminal acts in social media: A case study on Twitter // Information. – 2020. – Vol. 11, Issue 3. – P. 154-1-154-40.

57 de Mendonça R. et al. OntoCexp: a proposal for conceptual formalization of criminal expressions // Procced. 16th internat. conf. on Information Technology-New Generations (ITNG 2019). – Cham, 2019. – P. 43-48.

58 Devyatkin D., Smirnov I., Ananyeva M. et al. Exploring linguistic features for extremist texts detection (on the material of Russian-speaking illegal texts) // Procced. IEEE internat. conf. on Intelligence and Security Informatics. – Beijing, 2017. – P. 188-190.

59 Bolatbek M.A., Mussiraliyeva Sh.Zh., Tukeyev U.A. Creating the dataset of keywords for detecting an extremist orientation in web-resources in the Kazakh language // Journal of Mathematics, Mechanics and Computer Science. – 2018. – Vol. 97, Issue 1. – P. 134-142.

60 Maynard D., Funk A., Peters W. Using lexico-syntactic ontology design patterns for ontology creation and population // Proc. of the Workshop on Ontology Pattern. – 2009. – Vol. 516. – P. 39-52.

61 Alobaidi M., Malik K.M., Hussain M. Automated ontology generation framework powered by linked biomedical ontologies for disease-drug domain // Computer methods and programs in biomedicine. – 2018. – Vol. 165. – P. 117-128.

62 Doing-Harris K., Livnat Y. et al. Automated concept and relationship extraction for the semi-automated ontology management (SEAM) system // Journal of biomedical semantics. – 2015. – Vol. 6, Issue 1. – P. 15-1-15-15.

63 Kumar N., Kumar M. et al. Automated ontology generation from a plain text using statistical and NLP techniques // Int. Journal of System Assurance Engineering and Management. – 2016. – Vol. 7, Issue 1. – P. 282-293.

64 Huang J.X., Lee K.S., Choi K.S. et al. Extract reliable relations from Wikipedia texts for practical ontology construction // Computación y Sistemas. – 2016. – Vol. 20, Issue 3. – P. 467-476.

65 Cahyani D.E., Wasito I. Automatic ontology construction using text corpora and ontology design patterns (ODPs) in Alzheimer's disease // Jurnal Ilmu Komputer dan Informasi. – 2017. – Vol. 10, Issue 2. – P. 59-66.

66 Yehia N., Mokhtar S.A. et al. Automatic generation of OWL ontology from XML data source // International Journal of Computer Science Issues. – 2012. – Vol. 9, Issue 2. – P. 1-8.

67 Hu H., Liu D.Y. Learning OWL ontologies from free texts // Proceed. of 2004 internat. conf. on Machine Learning and Cybernetics (IEEE Cat. No.04EX826). – Shanghai, 2004. – P. 1233-1237.

68 West D.B. Introduction to graph theory. – New Jersey: Prentice Hall., 2001. – 588 p.

69 Meijer K., Frasincar F. et al. A semantic approach for extracting domain taxonomies from text // Decision Support Systems. – 2014. – Vol. 62. – P. 78-93.

70 Elnagar S., Yoon V., Thomas M.A. An automatic ontology generation framework with an organizational perspective // Proceed. of the 53rd Hawaii International conf. on System Sciences. – Honolulu, 2022. – P. 4860-4869.

71 Ehrlinger L., Wöß W. Towards a Definition of Knowledge Graphs // SEMANTiCS 2016: Posters and Demos. – Leipzig, 2016. – P. 1-4.

72 Riloff E. Automatically constructing a dictionary for information extraction tasks // Proceed. of the 11th nat. conf. on Artificial Intelligence. – Cambridge, 1993. – P. 811-816.

73 Campos D. et al. TrigNER: automatically optimized biomedical event trigger recognition on scientific documents // Source code for biology and medicine. – 2014. – Vol. 9, Issue 1. – P. 1-13.

74 Björne J., Ginter F., Salakoski T. et al. EPE 2017: The Biomedical Event Extraction Downstream Application // Proceed. of the 2017 Shared Task on Extrinsic Parser Evaluation (EPE 2017) at the 4th internat. conf. on Dependency Linguistics (Depling 2017) and the 15th internat. conf. on Parsing Technologies (IWPT 2017). – Pisa, 2017. – P. 17-24.

75 Hogenboom F. Automated Detection of Financial Events in News Text: thes. … doc. – Dordrecht, 2014. – 250 p.

76 Arendarenko E., Kakkonen T. Ontology-based information and event extraction for business intelligence // Proceed. of the 15th internat. conf. on Artificial Intelligence: methodology, systems, and applications. – Varna, 2012. – P. 89-102.

77 Reyes-Ortiz J.A. Criminal Event Ontology Population and Enrichment using Patterns Recognition from Text // International Journal of Pattern Recognition and Artificial Intelligence. – 2019. – Vol. 33, Issue 66. – P. 1940014.

78 Li Q., Yao Z.J., Zhang Y. Event Extraction for Criminal Legal Text // Procced. IEEE internat. conf. on Knowledge Graph (ICKG). – Nanjing, 2020. – P. 573-580.

79 Abdelkoui F., Kholladi M.-Kh. Extracting criminal-related events from Arabic tweets: A spatio-temporal approach // Journal of Information Technology Research. – 2017. – Vol. 10. Issue 3. – P. 34-47.

80 Pham X.-Q., HO B.-Q. Combination of Rule-based and Machine Learning for Biomedical Event Extraction // Procced. internat. conf. on Information Resources Management: Big Data - Revolutionizing How We Live, Work, and Think. – Ho Chi Minh, 2014. – P. 18-1-18-9.

81 Sha L. et al. RBPB: Regularization-Based Pattern Balancing Method for Event Extraction // Procced. 54th Annual Meeting of the Association for Computational Linguistics. – Berlin, 2016. – P. 1224-1234.

82 Xiang W., Wang B. A survey of event extraction from text // IEEE Access. – 2019. – Vol. 7. – P. 173111-173137.

83 Manning Ch.D. Computational linguistics and deep learning // Computational Linguistics. – 2015. – Vol. 41, Issue 4. – P. 701-707.

84 Li L., Liu Y., Qin M. Extracting Biomedical Events with Parallel Multi-Pooling Convolutional Neural Networks // IEEE/ACM Trans Comput Biol Bioinform. – 2020. – Vol. 17, Issue 2. – P. 599-607.

85 Yagcioglu S. et al. Detecting Cybersecurity Events from Noisy Short Text // https://www.researchgate.net/publication/332342241_Detecting. 25.06.2021.

86 Liu X. et al. Jointly Multiple Events Extraction via Attention-based Graph Information Aggregation // Procced. the 2018 conf. on Empirical Methods in Natural Language Processing. – Brussels, 2018. – P. 1247-1256.

87 Nguyen T.H., Grishman R. Graph convolutional networks with argument-aware pooling for event detection // Procced. 32nd AAAI conf. Artificial Intelligence. – New Orleans, 2018. – P. 5900-5907.

88 Liu X., Chen Y., Liu K. et al. Event detection via gated multilingual attention mechanism // Procced. 323nd AAAI conf. on Artificial Intelligence. – New Orleans, 2018. – P. 4865-4872.

89 Allan J. Topic Detection and Tracking: Event-Based Information Organization // Proceed. of the DARPA Broadcast News Transcription and Understanding Workshop. – Lansdowne, 2012. – P. 194-218.

90 Subburathinam A. et al. Cross-lingual Structure Transfer for Relation and Event Extraction // Proceed. of the 2019 conf. on Empirical Methods in Natural Language Processing and the 9th internat. joint conf. on Natural Language Processing (EMNLP-IJCNLP). – Hong Kong, 2019. – P. 313-325.

91 Fincke S., Agarwal S., Miller S. et al. Language Model Priming for Cross-Lingual Event Extractio // https://arxiv.org/abs/2109.12383. 28.06.2021.

92 Lin Y., Liu Z., Sun M. Neural relation extraction with multi-lingual attention // Procced. of the 55th Annual Meeting of the Association for Computational Linguistics. – Vancouver, 2017. – P. 34-43.

93 Pelicon A. et al. Investigating cross-lingual training for offensive language detection // Peer J Comput Sci. – 2021. – Vol. 7. – P. e559-1-e559-39.

94 Ahmad W.U., Peng N., Chang K.-W. GATE: Graph Attention Transformer Encoder for Cross-lingual Relation and Event Extraction // Procced. 32nd AAAI conf. Artificial Intelligence. – New Orleans, 2021. – P. 74-75.

95 Taghizadeh N., Faili H. Cross-lingual Adaptation Using Universal Dependencies // ACM Transactions on Asian and Low-Resource Language Information Processing. – 2021. – Vol. 20, Issue 4. – P. 1-23.

96 Getman J., Ellis J., Strassel S. et al. Laying the ground- work for knowledge base population: Nine years of linguistic resources for TAC KBP // Procced. of the 11h internat. conf. on Language Resources and Evaluation (LREC-2018). – Miyazaki, 2018. – P. 1552-1558.

97 Каминченко Д.И. Информационная повестка дня современных сетевых СМИ: политический аспект // Via in tempore. История. Политология. – 2019. – №46(3). – С. 576-584.

98 Adams A., Harf A. et al. Agenda Setting Theory: A Critique of Maxwell McCombs & Donald Shaw's Theory In Em Griffin's A First Look at Communication Theory // Meta-communicate. – 2014. – Vol. 4, Issue 1. – P. 176-187.

99 Pan X., Zhang B., May J. et al. Crosslingual name tagging and linking for 282 languages // Proceed. of the 55th Annual Meeting of the Association for Computational Linguistics. – Vancouver, 2017. – P. 1946-1958.

100 Peng H., Song Y., Roth D. Event detection and co-reference with minimal supervision // Proceed. of the 2016 conf. on Empirical Methods in Natural Language Processing. – Austin, 2016. – P. 392-402.

101 Zhang B., Lu D., Pan X. et al. Embracing non-traditional linguistic resources for low-resource language name tagging // Proceed. of the 8th internat. Joint conf. on Natural Language Processing. – Taipei, 2017. – P. 362-372.

102 Ferguson J., Lockard C., Weld D.S. et al. Semi-supervised event extraction with paraphrase clusters // Proceed. of the 2018 conf. of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. – New Orleans, 2018. – P. 359-364.

103 Lample G., Ballesteros M., Kawakami K. et al. Neural architectures for named entity recognition // Proceed. of the 2016 conf. of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. – San Diego, 2016. – P. 260-270.

104 Zhang T. et al. Joint entity and event extraction with generative adversarial imitation learning // Data Intelligence. – 2019. – Vol. 1. – P. 99-120.

105 Pecina P. Lexical association measures and collocation extraction // Lang Resources & Evaluation. – 2010. – Vol. 44. – P. 137-158.

106 Petrasova S., Khairova N. Information technology for extraction of coherent text fragments from scientometric databases // Procced. 2nd internat.

Workshop on Intelligent Information Technologies and Systems of Information Security. – Khmelnytskyi, 2021. – P. 151-157.

107 Lenci A. Distributional Models of Word Meaning // Annual Review of Linguistics. – 2018. – Vol. 4. – P. 151-171.

108 Dinu A., Dinu L., Sorodoc I. Aggregation methods for efficient collocation detection //. Proceedings of the Ninth International Conference on Language Resources and Evaluation. – Reykjavik, 2014. – P. 4041-4045.

109 Хохлова М.В. Сопоставительный анализ статистических мер на примере частеречных преференцій сочетаемости существительных // Компьютерная лингвистика и вычислительные онтологии. – 2017. – №1. – С. 166-171.

110 Liu X., Huang D., Yin Zh. et al. Recognition of Collocation Frames from Sentences // IEICE Trans. Inf. Syst. – 2019. – Vol. 102-D. – P. 620-627.

111 Хохлова М.В. К вопросу о сходстве мер ассоциации применительно к задаче автоматического извлечения глагольных коллокаций // Компьютерная лингвистика и вычислительные онтологии. – 2019. – №3. – С. 9-18.

112 Petrasova S., Khairova N., Lewoniewski W. et al. Similar Text Fragments Extraction for Identifying Common Wikipedia Communities // Data. Stream Mining and Processing. – 2018. – Vol. 3, Issue 4. – P. 66-1-66-9.

113 Бондаренко М.Ф., Шабанов-Кушнаренко Ю.П. Теория интеллекта. – Харьков: СМИТ, 2007. – 576 с.

114 Hossain K.T. et al. Forecasting violent events in the Middle East and North Africa using the Hidden Markov Model and regularized autoregressive models // The Journal of Defense Modeling and Simulation. – 2020. – Vol. 17, Issue 3. – P. 269-283.

115 Karimi M., Gharehchopogh F.S. An Improved K-Means with Artificial Bee Colony Algorithm for Clustering Crimes // Journal of Advances in Computer Research Quarterly. – 2020. – Vol. 11, Issue 3. – P. 61-82.

116 Salas A.H., Morzan-Samam J., Nunez-del-Prado M. Crime alert! crime typification in news based on text mining // Lecture Notes in Networks and Systems. – 2020. – Vol. 69. – P. 725-741.

117 Santhiya K., Bhuvaneswari V., Murugesh V. Automated Crime Tweets Classification and Geo-location Prediction using Big Data Framework // Turkish Journal of Computer and Mathematics Education. – 2021. – Vol. 12, Issue 14. – P. 2133-2152.

118 Moreno-Jimenez L-G. et al. Criminal events detection in news stories using intuitive classification // Procced. of Mexican internat. conf. on Artificial Intelligence (MICAI). – Enseneda, 2017. – P. 120-132.

119 Hassani H., Huang X., Silva E.S. et al. A review of data mining applications in crime // Statistical Analysis and Data Mining: The ASA Data Science Journal. – 2016. – Vol. 9, Issue 3. – P. 139-154.

120 Chen P., Kurland J. Time, place, and modus operandi: a simple apriori algorithm experiment for crime pattern detection // Procced. of the 9th internat. conf. on Information, Intelligence, Systems and Applications. – Zakynthos, 2018. – P. 1-3.

121 Joseph N. Crime Analysis Based on K-Means Clustering // https://easychair.org/publications/preprint/GVWM.

122 Lin Y.L., Yen M.F., Yu L.C. Grid-based crime prediction using geographical features // ISPRS International Journal of Geo-Information. – 2018. – Vol. 7. – P. 298-1-298-16.

123 Rangel F. et al. Profiling hate speech spreaders on twitter task at PAN 2021 // Procced. CLEF 2021 conf. and Labs of the Evaluation forum. – Bucharest, 2021. – P. 1-18.

124 Siino M. et al. Detection of hate speech spreaders using convolutional neural networks // Procced. CLEF 2021 conf. and Labs of the Evaluation forum. – Bucharest, 2021. – P. 1-11.

125 Miok K., Skrlj B., Zaharie D. et al. To BAN or Not to BAN: Bayesian Attention Networks for Reliable Hate Speech Detection // Cognitive Computation. – 2021. – Vol. 14, Issue 6. – P. 353-371.

126 Qureshi K.A., Sabih M. Un-Compromised Credibility: Social Media Based Multi-Class Hate Speech Classification for Text // IEEE Access. – 2021. – Vol. 9. – P. 9503413-1-9503413-9.

127 Alfina I., Mulia R., Fanany M.I. et al. Hate speech detection in the Indonesian language: A dataset and preliminary study // Procced. internat. conf. on Advanced Computer Science and Information Systems (ICACSIS). – Bali, 2017. – P. 233-238.

128 Mou G., Ye P., Lee K. SWE2: SubWord Enriched and Significant Word Emphasized Framework for Hate Speech Detection // Procced. of the internat. conf. on Information and Knowledge Management. – NY., 2020. – P. 1145-1154.

129 Mozafari M., Farahbakhsh R., Crespi N. A BERT-Based Transfer Learning Approach for Hate Speech Detection in Online Social Media // Procced. of the internat. conf. on Complex Networks and Their Applications. – Lisbon, 2019. – P. 928-940.

130 Schmidt A., Wiegand M. A survey on hate speech detection using natural language processing // Procced. of the Fifth International Workshop on Natural Language Processing for Social Media. – Valencia, 2017. – P. 1-10.

131 Nockleby J.T. Hate speech // In book: Encyclopedia of the American Constitution. – NY.: Macmillan, 2000. – P. 1277-1279.

132 Ku C.H., Iriberri A., Leroy G. Crime Information Extraction from Police and Witness Narrative Reports // Procced. of the 2008 IEEE conf. on Technologies for Homeland Security. – Waltham, 2008. – P. 193-198.

133 Das P., Das A.K. Graph-based clustering of extracted paraphrases for labelling crime reports // Knowledge Based Systems. – 2019. – Vol. 179. – P. 55-76.

134 Das P. et al. A graph based clustering approach for relation extraction from crime data // IEEE Access. – 2019. – Vol. 7, Issue 1. – P. 101269-101282.

135 Dasgupta T. et al. CrimeProfiler: crime information extraction and visualization from news media // Procced. of the internat. conf. on Web Intelligence, ACM. – Leipzig, 2017. – P. 541-549.

136 Ku C.H., Iriberri A., Leroy G. Natural Language Processing and e-Government: Crime Information Extraction from Heterogeneous Data Sources //

Procced. of the 9th Annual internat. Digital Government Research conf. – Montreal, 2008. – P. 162-170.

137 Rahma F., Romadhony A. Rule-Based Crime Information Extraction on Indonesian Digital News // Procced. of the 2021 internat. conf. on Data Science and Its Applications (ICoDSA). – Bandung, 2021. – P. 10-15.

138 Joseph J.G. et al. Automatic Information Extraction and Inferencing System from Online News Sources for Substance Abuse Cases // Proceed. of the internat. Semantic Intelligence conf. (ISIC). – New Delhi, 2021. – P. 516-520.

139 Sotomayor M., Veloz F. Thesaurus-based named entity recognition system for detecting spatio-temporal crime events in Spanish language from Twitter // Procced. IEEE 2nd Ecuador Technical Chapters Meeting (ETCM). – Salinas, 2017. – P. 1-5.

140 Das P., Das A.K. A two-stage approach of named-entity recognition for crime analysis // Procced. of internat. conf. on Computing, Communication and Networking Technologies. – Delhi, 2017. – P. 1-5.

141 Procced. 3rd Message Understanding conf. (MUC-3) / Defense Advanced Research Projects Agency. – San Diego, 1991. – 363 p.

142 Rahem K.R., Omar N. Drug-Related crime information extraction and analysis // Procced. of internat. conf. on Information Technology and Multimedia (ICIMU). – Putrajaya, 2014. – P. 250-254.

143 Yagcioglu S. et al. Detecting Cybersecurity Events from Noisy Short Text // Proceed. of the 2019 conf. of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. – Minneapolis, 2019. – P. 1366-1372.

144 Wang X., Gerber M.S., Brown D.E, "Automatic crime prediction using events extracted from Twitter posts // Procced. internat. conf. on Social Computing, Behavioral - Cultural Modeling, and Prediction. – Berlin, 2012. – P. 231-238.

145 Davani A.M. et al. Reporting the Unreported: Event Extraction for Analyzing the Local Representation of Hate Crimes // Procced. and the 9th internat. Joint conf. on Natural Language Processing (EMNLP-IJCNLP). – Hong Kong, 2019. – P. 5753-5757.

146 Han S. et al. American hate crime trends prediction with vent extraction // https://arxiv.org/abs/2111.04951. 05.11.2021.

147 Mullah N.S., Zainon W.M.N.W. Advances in Machine Learning Algorithms for Hate Speech Detection in Social Media: A Review // IEEE Access. – 2021. – Vol. 9. – P. 88364-88376.

148 Khairova N., Kolesnyk A., Ybytayeva G. et al. Automatic Multilingual Ontology Generation Based on Texts Focused on Criminal Topic // Proceed. of the 5th internat. conf. on Computational Linguistics and Intelligent Systems. – Kharkiv, 2021. – P. 108-117.

149 Ыбытаева Г.С., Мамырбаев О.Ж., Хайрова Н.Ф. және т.б. Қазақ тіліндегі мәтіндерде коллокацияларды анықтаудың статистикалық әдістерін талдау // Матер. 6-й междунар. науч. конф. «Информатика и прикладная математика». – Алматы, 2021. – С. 256-262.

150 Turnay P.D., Pantel P. From frequency to meaning: Vector Space Models of Semantics // Journal of Artificial Intelligence Research. – 2010. – Vol. 37. – P. 141-188.

151 Harris Z. Distributional structure // Word. – 1954. – Vol. 10, Issue 23. – P. 146-162.

152 Khairova N., Mamyrbayev O., Mukhsina K. et al. Logical-Linguistic model for multilingual open information extraction // Cogent Engineering. – 2020. – Vol. 7, Issue 1. – P. 1714829.

153 Khairova N.F., Petrasova S., Gautam A.P. The logical-linguistic model of fact extraction from English texts // Information and Software Technologies: procced. internat. conf. – Cham: Springer, 2016. – P. 625-635.

154 Ybytayeva G., Mamyrbayev O., Khairova N. et al. Creating a Thesaurus" Crime-Related Web Content" Based on a Multilingual Corpus // Procced. 7th internat. conf. on Computational Linguistics and Intelligent Systems. – Kharkiv, 2023. – P. 77-87.

155 Ace (Automatic Content Extraction) English Annotation Guidelines for Events // https://www.bibsonomy.org/bibtex. 15.02.2022.

156 Chen H. et al. COPLINK connect: information and knowledge management for law enforcement // Decision support systems. – 2003. – Vol. 34, Issue 3. – P. 271-285.

157 Khairova N., Mamyrbayev O., Ybytayeva G. et al. A Parallel Corpus-Based Approach to the Crime Event Extraction for Low-Resource Languages // IEEE Access. – 2023. – Vol. 11. – P. 54093-54111.

158 Petrasova S., Khairova N. Event Extraction from News Media Based on Knowledge and Data-Driven Methods // Procced. 2021 IEEE 16th internat. conf. on Computer Sciences and Information Technologies (CSIT). – Lviv, 2021. – P. 60-63.

159 Makhambetov O., Makazhanov A. et al. Data-driven morphological analysis and disambiguation for Kazakh // Procced. internat. conf. on In-telligent Text Processing and Computational Linguistics. – Cairo, 2015. – P. 151-163.

160 Banerjee M., Capozzoli M., McSweeney L. et al. Beyond kappa: A review of interrater agreement measures // The Canadian J. of Statistics. – 2008. – Vol. 27, Issue 1. – P. 3-23.

161 Kolesnyk A.S., Khairova N.F. Justification for the Use of Cohen's Kappa Statistic in Experimental Studies of NLP and Text Mining // Cybernetics and Systems Analysis. – 2022. – Vol. 58. – P. 280-288.

ҚОСЫМША А

Авторлық куәліктері

ҚАЗАҚСТАН РЕСПУБЛИКАСЫ

РЕСПУБЛИКА КАЗАХСТАН

АВТОРЛЫҚ ҚҰҚЫҚПЕН ҚОРҒАЛАТЫН ОБЪЕКТІЛЕРГЕ ҚҰҚЫҚТАРДЫҢ МЕМЛЕКЕТТІК ТІЗІЛІМГЕ МӘЛІМЕТТЕРДІ ЕНГІЗУ ТУРАЛЫ

КУӘЛІК

2023 жылғы «26» қаңтар № 32055

Автордың (лардың) жөні, аты, әкесінің аты (егер ол жеке басын куәландыратын құжатта көрсетілсе):
Мамырбаев Оркен Жумажанович, Ыбытаева Галия Сейткалиевна, Санжарсултан Бердали Хамитулы

Авторлық құқық объектісі: ЭЕМ-ге арналған бағдарлама

Объектінің атауы: Веб-приложение многоязычной базовой онтологии «Противоправный веб-контент»

Объектіні жасаған күні: 24.01.2023

Құжат түпнұсқалығын http://www.kazpatent.kz/ru сайтының "Авторлық құқық" бөлімінде тексеруге болады https://copyright.kazpatent.kz

Подлинность документа возможно проверить на сайте kazpatent.kz в разделе «Авторское право» https://copyright.kazpatent.kz

ЭЦҚ қол қойылды

Н. Абулкаиров

ҚАЗАҚСТАН РЕСПУБЛИКАСЫ

РЕСПУБЛИКА КАЗАХСТАН

АВТОРЛЫҚ ҚҰҚЫҚПЕН ҚОРҒАЛАТЫН ОБЪЕКТІЛЕРГЕ ҚҰҚЫҚТАРДЫҢ МЕМЛЕКЕТТІК ТІЗІЛІМГЕ МӘЛІМЕТТЕРДІ ЕНГІЗУ ТУРАЛЫ

КУӘЛІК

2023 жылғы «29» тамыз № 38766

Автордың (лардың) жөні, аты, әкесінің аты (егер ол жеке басын куәландыратын құжатта көрсетілсе):
Ыбытаева Галия Сейткалиевна, Мамырбаев Оркен Жумажанович

Авторлық құқық объектісі: ЭЕМ-ге арналған бағдарлама

Объектінің атауы: «Құқыққа қайшы интернет-контент» көптілді онтология қосымшасы

Объектіні жасаған күні: 25.08.2023

Құжат түпнұсқалығын http://www.kazpatent.kz/ru сайтының "Авторлық құқық" бөлімінде тексеруге болады https://copyright.kazpatent.kz

Подлинность документа возможно проверить на сайте kazpatent.kz в разделе «Авторское право» https://copyright.kazpatent.kz

ЭЦҚ қол қойылды

Е. Осланов

ҚОСЫМША Ә

Әзірленген онтологияның бағдарламалық кодының үзіндісі

```
let currentLanguage = 'kz'; let buttons = [];
function replaceStringWithDictionaryValue(inputString) { if
(dictionary.hasOwnProperty(inputString)) {
return dictionary[inputString]; } else {
return inputString; }
}
document.querySelector('.form-arrow').addEventListener('click', function()
{ const formCheckbox = document.getElementById('toggle-form');
// If user manually shows the
if (!formCheckbox.checked) { manuallyShown = true;
}
// If user manually hides the
if (formCheckbox.checked) { manuallyShown = false;
} });
form, set manuallyHidden to true
form, set manuallyHidden to false
function toggleFormCheckbox() {
const formCheckbox = document.getElementById('toggle-form');
// Only show the form if it was not manually shown by the user
if (!manuallyShown) {
if (!formCheckbox.checked) {
formCheckbox.click(); }
} }
document.querySelector('.form-container').addEventListener('click',
function(event) {
// Get the height of the form container
const formContainerHeight = this.offsetHeight; // "this" refers to the element
that the event listener is attached to
// Get the position of the click relative to the form container
// event.clientY provides the position relative to the viewport, not the form-
container
// event.offsetY provides the position relative to the closest positioned ancestor
(form-container) const clickPosition = event.offsetY;
// Calculate the top 67% of the form container
const threshold = formContainerHeight * 0.67;
// If the click position is above the threshold (meaning in the bottom 33%),
trigger the checkbox
if (clickPosition >= threshold) { toggleFormCheckbox();
} });
```

```
document.getElementById("question-mark").addEventListener("click",
function() { document.getElementById("help-dialog").classList.remove("hidden");
});
document.getElementById("close-dialog").addEventListener("click",
function() { document.getElementById("help-dialog").classList.add("hidden");
});
// Make the dialog draggable
var dialog = document.getElementById("help-dialog"); var isMouseDown =
false;
var startX, startY;
dialog.onmousedown = function(e) {
isMouseDown = true;
startX = e.clientX - parseInt(window.getComputedStyle(dialog).left); startY =
e.clientY - parseInt(window.getComputedStyle(dialog).top); // Prevent text selection
during dragging
dialog.style.userSelect = "none";
// Change cursor to move
document.body.style.cursor = "move";
};
document.onmousemove = function(e) { if (isMouseDown) {
dialog.style.left = (e.clientX - startX) + 'px';
dialog.style.top = (e.clientY - startY) + 'px'; }
};
document.onmouseup = function(e) { isMouseDown = false;
// Allow text selection after dragging dialog.style.userSelect = "";
// Change cursor back to default document.body.style.cursor = "default";
};
dialog.ondragstart = function() { return false;
};
function destroyGraph() {
// Check if the graph has already been destroyed if (!isGraphDestroyed) {
// Retrieve the container element
var container = document.getElementById('graph-container');
// Clear the container by setting its innerHTML to an empty string
container.innerHTML = '';
// Set the flag to indicate that the graph has been destroyed
isGraphDestroyed = true; }
function resetGraph(network, nodes)
{ document.getElementById('query').value = "";
// Clear the graph data
network.setData({ nodes: [], edges: [] });
// Reset the legend
const legendContainer = document.getElementById('legend-container');
legendContainer.innerHTML = ''; legendContainer.classList.add('hidden');
```

```
tooltip.classList.add('hidden');
// Remove event listeners for node label dropdowns
const dropdowns = document.querySelectorAll(".node-label-dropdown");
dropdowns.forEach(dropdown => {
dropdown.removeEventListener("change"); });
// Remove hoverNode, hoverEdge, blurNode, blurEdge event listeners from the
network
network.off("hoverNode"); network.off("hoverEdge");
network.off("blurNode"); network.off("blurEdge");
}
function assignNodeLabels(node) {
let labels = Object.keys(node.properties); if (node.properties.name) {
return node.properties.name;
} else if (node.properties.title) {
return node.properties.title;
} else if (node.properties.label) {
return node.properties.label;
} else if (labels.length > 0) {
return node.properties[labels[0]];
} else {
return "Unknown"; }
}
function generateLegendHTML(nodes, edges) {
const key_node = 'graph_node';
const key_edges = 'graph_edges';
let nodeLegendHTML = '<h4 data-translate-key="graph_node">' +
translations[currentLanguage][key_node] + '</h4>'; let edgeLegendHTML = '<h4
data-translate-key="graph_edges">' + translations[currentLanguage][key_edges] +
'</h4>';
const uniqueNodeLabels = [...new Set(nodes.map(node =>
replaceStringWithDictionaryValue(node.title)))].filter(label => label.trim()); const
uniqueEdgeLabels = [...new Set(edges.map(edge =>
replaceStringWithDictionaryValue(edge.label)))].filter(label => label.trim());
uniqueNodeLabels.forEach(label => {
const originalLabel = nodes.find(node =>
replaceStringWithDictionaryValue(node.title) === label).title;
nodeLegendHTML += `<div class="legend-item"><span class="legend-color"
style="background-color:${labelToColorMap[originalLabel]}"></span> ${label
});
uniqueEdgeLabels.forEach(label => {
const originalLabel = edges.find(edge =>
replaceStringWithDictionaryValue(edge.label) === label).label;
edgeLegendHTML += `<div class="legend-item"><span class="legend-color"
style="background-color:${typeToColorMap[originalLabel]}"></span> ${label}
```

```
        });
        return `<div class="legend-section">${nodeLegendHTML}</div><div
class="legend-section">${edgeLegendHTML}</div>`; }
        function showLegend(nodes, edges) {
        const legendContainer = document.getElementById('legend-container');
legendContainer.innerHTML = generateLegendHTML(nodes, edges);
legendContainer.classList.remove('hidden');
        }
        let let let
        }
        driver; labelToColorMap = {}; typeToColorMap = {};
        const driverTwo = neo4j.driver("bolt://89.250.84.93:37687",
neo4j.auth.basic("readonly", "crimeontoscope"));
        async function fetchLanguages() {
        let languages = localStorage.getItem('languages'); if (languages) {
        return JSON.parse(languages); }
        const session = driverTwo.session();
        const result = await session.run('MATCH (w:Word) RETURN DISTINCT
w.language as language ORDER BY language');
        session.close();
        languages = result.records.map(record => record.get('language'));
        localStorage.setItem('languages', JSON.stringify(languages));
        return languages; }
        async function displayLanguages() {
        const languages = await fetchLanguages();
        const languageSelect = document.getElementById('language-select');
        languageSelect.innerHTML = languages.map(language => `<option
value="${language}">${getLanguageName(language)}</option>`).join('');
        }
        // Download word arrays for all languages in the background
        languages.forEach(language => { displayWords(language);
        });
        languageSelect.value = "ka";
        // Fetch and display words for "ka" by default displayWords("ka");
        }
        window.addEventListener('load', displayLanguages);
        let wordsCache = JSON.parse(localStorage.getItem('wordsCache')) || {}; let
searchTerm = '';
        async function filterAndDisplayWords(language) {
        let words = wordsCache[language];
        if (searchTerm !== '') {
        words = words.filter(word =>
word.toLowerCase().startsWith(searchTerm.toLowerCase()));
        }
```

```
const wordList = document.getElementById('word-list'); wordList.innerHTML = '';
words.forEach(word => {
let wordItem = document.createElement('li'); wordItem.textContent = word; wordItem.classList.add('word-item'); wordItem.addEventListener('click', () => {
document.getElementById('query').value = `all(${word})`; const submitEvent = new Event('submit', { cancelable: true }); queryForm.dispatchEvent(submitEvent);
});
wordList.appendChild(wordItem); });
}
// Modify existing function to use the new function
async function displayWords(language) { if (wordsCache[language]) {
filterAndDisplayWords(language);
return; }
const session = driverTwo.session();
const result = await session.run('MATCH (w:Word) WHERE w.language = $language RETURN w.label as label ORDER BY label', { language }); session.close();
wordsCache[language] = result.records .map(record => record.get('label'))
.filter(word => word && !/^['()"].*/.test(word));
localStorage.setItem('wordsCache', JSON.stringify(wordsCache)); filterAndDisplayWords(language); // call the function again to render words with click events
}
window.addEventListener('load', () => {
// Moved the event listener inside the window.load event handler to ensure the DOM is ready document.getElementById('language-select').addEventListener('change', (event) => {
displayWords(event.target.value); });
// Add an event listener for the search input
document.getElementById('search-input').addEventListener('input', (event) => { searchTerm = event.target.value;
const languageSelect = document.getElementById('language-select'); filterAndDisplayWords(languageSelect.value);
}); });
// Function to convert HSL to RGB
function hslToRgb(h, s, l){
// Convert saturation and lightness to a fraction of 1 s /= 100;
l /= 100;
let c = (1 - Math.abs(2 * l - 1)) * s;
let x = c * (1 - Math.abs((h / 60) % 2 - 1));
let m = l - c/2;
let r = 0;
let g = 0;
```

```
let b = 0;
if (0 <= h && h < 60) {
r = c; g = x; b = 0;
} else if (60 <= h && h < 120) {
r = x; g = c; b = 0;
} else if (120 <= h && h r = 0; g = c; b = x;
} else if (180 <= h && h r = 0; g = x; b = c;
} else if (240 <= h && h r = x; g = 0; b = c;
} else if (300 <= h && h r = c; g = 0; b = x;
}
< 180) { < 240) { < 300) { < 360) {
// Convert RGB to 0-255 range and add the lightness value
r = Math.round((r + m) * 255); g = Math.round((g + m) * 255); b =
Math.round((b + m) * 255); return [r, g, b];
}
let hue = Math.random() * 360; function generateRandomColor() {

// Increment hue by golden ratio (approximately 137.5 degrees)
// This helps in distributing the colors evenly around the color wheel, ensuring
diversity hue = (hue + 137.508) % 360;
// Define saturation variable - a number between 50 and 100
// Keeping saturation relatively high to get pastel colors
let saturation = Math.floor(Math.random() * (100 - 50 + 1)) + 50;
// Define lightness variable - a number between 60 and 90
// Keeping lightness high to ensure colors are light (pastel) and will work well
with black text let lightness = Math.floor(Math.random() * (90 - 60 + 1)) + 60;
// Convert HSL to RGB
let rgbColor = hslToRgb(hue, saturation, lightness);
// Convert RGB to hexadecimal
let color = '#';
for (let i = 0; i < 3; i++) {
let hex = rgbColor[i].toString(16); if (hex.length < 2) {
hex = '0' + hex; }
color += hex; }
// Return the generated pastel color
return color; }
function getLanguageName(code) { const languages = {
'ru': 'русский', 'ka': 'қазақша', 'en': 'English', 'ua': 'українська',
};
return languages[code]; }
function assignNodeTitles(value) {
for (let i = 0; i < value.labels.length; i++) { // iterating over each label
if (value.labels[i] !== "NamedIndividual" && value.labels[i] !== "Resource")
{
```

```
        return value.labels[i]; // if label is not "NamedIndividual" or "Resource", assign
it as a title
        } }
        return value.labels[0]; // default to first label if no suitable labels found }
        function assignNodeGroupColors(nodes, regenerate = false)
{ nodes.forEach(node => {
        if (regenerate || !labelToColorMap[node.title]) { labelToColorMap[node.title] =
generateRandomColor();
        }
        node.color = labelToColorMap[node.title]; });
        // Check if the "belongsToTerm" edge label is being assigned
        const termEdges = edges.filter(edge => edge.label === "belongsToTerm");
        if (termEdges.length > 0) { termEdges.forEach(edge => {
        const sourceNode = nodes.find(node => node.id === edge.from); const
targetNode = nodes.find(node => node.id === edge.to);
        // Check if either the source or target node has a "language" property
        if (sourceNode && sourceNode.properties.language) {
        const language = getLanguageName(sourceNode.properties.language);
edge.label = `${language}`;
        } else if (targetNode && targetNode.properties.language) {
        const language = getLanguageName(targetNode.properties.language);
edge.label = `${language}`;
        } });
        } }
        function assignEdgeGroupColors(edges, regenerate = false) {
        edges.forEach(edge => {
        // If the label is "GrammaticalCategory", replace it with an empty string
        if (edge.label === "GrammaticalCategory" || edge.label ===
"belongsToDomain" || edge.label === "belongsToTerm") { edge.label = " ";
        }
        if (regenerate || !typeToColorMap[edge.label]) { typeToColorMap[edge.label]
= generateRandomColor();
        }
        edge.color = {
        color: typeToColorMap[edge.label], highlight: typeToColorMap[edge.label],
hover: typeToColorMap[edge.label], inherit: false
        }; });
        }
        function regenerateColors() {
        const queryValue = document.getElementById('query').value; if (queryValue
=== " || queryValue === null) {
        return; // if the 'query' value is empty or null, do nothing and return }
        assignNodeGroupColors(nodes, true);
         assignEdgeGroupColors(edges, true);
```

```
        regenerateColorsFlag = true;
        const submitEvent = new Event('submit', { cancelable: true });
queryForm.dispatchEvent(submitEvent);
        function initializeAndAnimateGraph() {
        // Initialize an empty network
        var container = document.getElementById('graph-container');
        // Create an array for nodes and edges
        var nodes = new vis.DataSet([]); var edges = new vis.DataSet([]);
        // Populate nodes with colorful and unique elements
        for(let i = 0; i < 100; i++) {
        let color = 'hsl(' + 360 * (i/100) + ', 70%, 85%)'; // Changed color for a pastel
look let size = Math.floor(Math.random() * 10) + 5;
        nodes.add({
        id: i, color: color, size: size
        }); }
        // Populate edges with connections between nodes
        for(let i = 0; i < 100; i++) {
        for(let j = i + 1; j < 100; j++) {
        if(Math.random() > 0.95) {
        let width = Math.random() * 2 + 1;
        let color = 'hsl(' + 360 * ((i+j)/200) + ', 70%, 85%)'; // Changed color for a
pastel look edges.add({
        from: i, to: j,
        }
        }); }
        } }
        // Provide the data in
        var data = { nodes: nodes,
        edges: edges };
        width: color:
        width, color
        the vis format
        // Initialize your network with options
        var
        options = { nodes: {
        shape: 'dot',
        borderWidth: 0 // Remove border for a softer look },
        edges: { smooth: {
        type: 'dynamic', forceDirection: 'none', roundness: 0.5
        } },
        physics: {
        solver: 'repulsion', // Using the repulsion solver for a more free, dreamlike
motion repulsion: {
```

```
        nodeDistance: 80, // Decreased node distance for a more interconnected
network
        centralGravity: 0.05 // Added a slight pull to the center for an appealing
clustering effect },
        stabilization: false // Disable stabilization for a more free, dreamlike motion },
        interaction: {
        tooltipDelay: 200,
        hideEdgesOnDrag: false,
        dragNodes: true,
        hover: true,
        navigationButtons: true, // Add navigation buttons for user to change view
keyboard: {
        enabled: true, // Enable keyboard shortcuts for user to change view
bindToWindow: false
        } }
        };
        // Creating the network
        var network = new vis.Network(container, data, options); }
        connectionForm.addEventListener('submit', async (event) =>
{ event.preventDefault();
        const url = document.getElementById('url').value;
        const username = document.getElementById('username').value; const
password = document.getElementById('password').value;
        driver = neo4j.driver(url, neo4j.auth.basic(username, password));
        try {
        await driver.verifyConnectivity(); connectionForm.style.display = 'none';
queryConstructorForm.style.display = 'block';
        } catch (error) {
        alert('Ошибка при подключении к Neo4j: ' + error.message);
        }
        /* await displayWords(); */
        });
        // Function to create a button
        function createButton(translateKey, clickHandler) {
```

ҚОСЫМША Б

Енгізу актісі

УТВЕРЖДАЮ
Начальник УКП
ДП области Жетісу
полковник полиции
Журунов К.М.
_____
«15» 09 2023 г.

АКТ
о внедрении результатов работы по проекту
ГФ AP09259309 «Информационная модель и программный инструментарий системы автоматического поиска и анализа многоязычного противоправного веб-контента на базе онтологического подхода»

Настоящий акт свидетельствует, что информационная-аналитическая система, разработанная Мамырбаевым Оркеном Жумажановичем, Хайровой Ниной Феликсовной и Ыбытаевой Галией Сейткалиевной, внедрена в Управление криминальной полиции ДП области Жетісу.

В ходе эксплуатации информационной-аналитической системы подтверждено, что она обладает заявленными возможностями и позволяет облегчить работу Управления. Использование разработанной системы позволяет повысить эффективность работы правоохранительных органов за счет повышения вероятности раскрытия преступлений и предотвращения противоправных действий.

Оперуполномоченный УКП
ДП области Жетісу
майор полиции **Нургалымов М.К.**