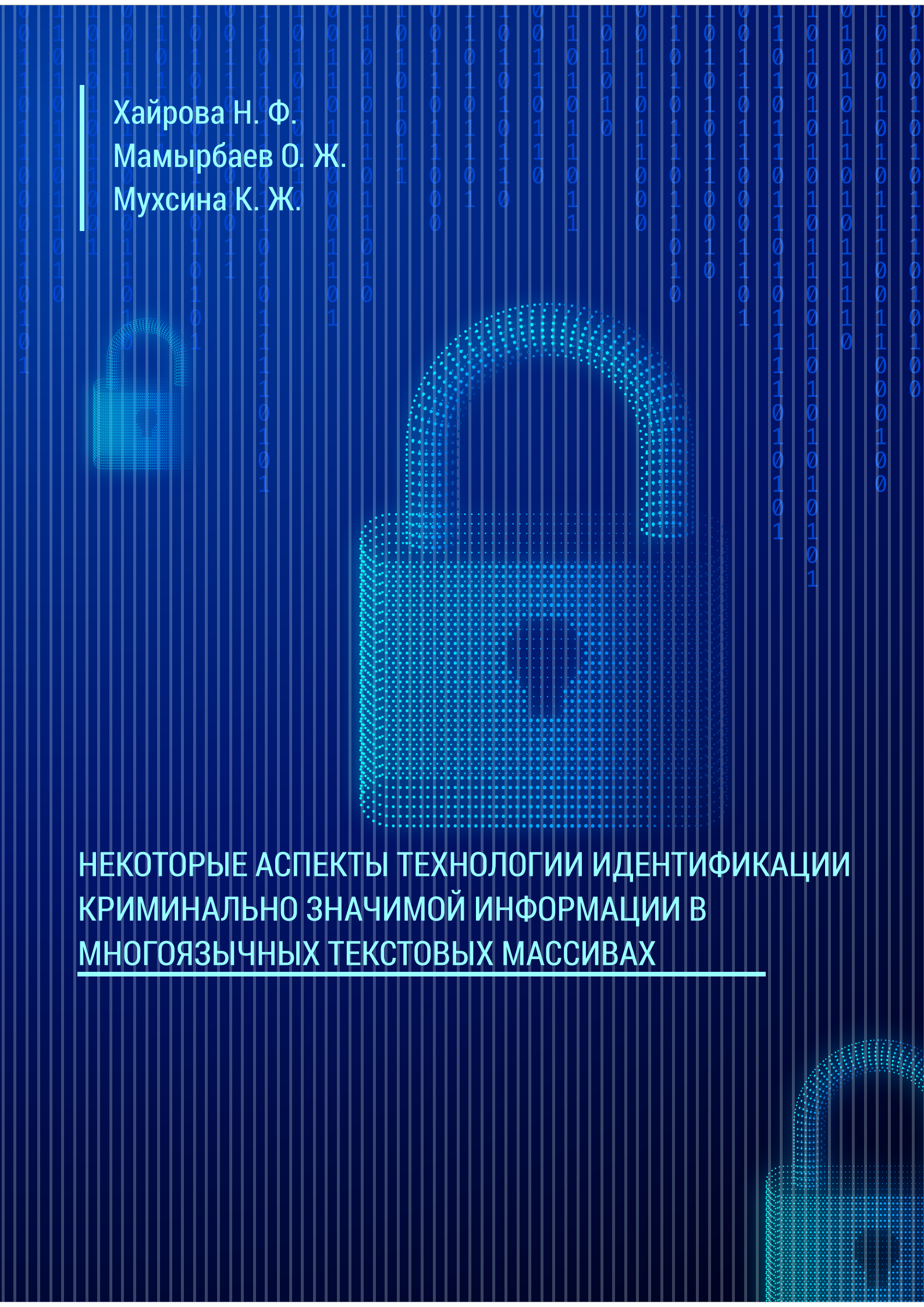
Хайрова Н. Ф.
Мамырбаев О. Ж.
Мухсина К. Ж.



НЕКОТОРЫЕ АСПЕКТЫ ТЕХНОЛОГИИ ИДЕНТИФИКАЦИИ КРИМИНАЛЬНО ЗНАЧИМОЙ ИНФОРМАЦИИ В МНОГОЯЗЫЧНЫХ ТЕКСТОВЫХ МАССИВАХ

Хайрова Н. Ф., Мамырбаев О. Ж., Мухсина К. Ж.

НЕКОТОРЫЕ АСПЕКТЫ ТЕХНОЛОГИИ ИДЕНТИФИКАЦИИ КРИМИНАЛЬНО ЗНАЧИМОЙ ИНФОРМАЦИИ В МНОГОЯЗЫЧНЫХ ТЕКСТОВЫХ МАССИВАХ

Алматы, 2020

УДК 80/81:004
ББК 81.2 Рус-5
Х15

РЕЦЕНЗЕНТЫ:
Мекебаев Н.О. – PhD, кафедра «Информатика и прикладная математика», Казахский национальный женский педагогический университет
Калижанова А.У. – асс. профессор, к.ф-м.н., РГП «Институт информационных и вычислительных технологий»

Публикуется по решению ученого совета университета,

протокол № 10 от 11 августа 2020 г.

Х 15 Хайрова Н. Ф., Мамырбаев О. Ж., Мухсина К. Ж. Некоторые аспекты технологии идентификации криминально значимой информации в многоязычных текстовых массивах / Хайрова Н. Ф., Мамырбаев О. Ж., Мухсина К. Ж. – Алматы: Институт информационных и вычислительных технологий, 2020. – 92 с.; Ил. 18; Табл. 12; Библиогр. 109 наим.

ISBN 978-601-332-917-8

В монографии рассмотрены проблемы в области поиска криминально окрашенных текстов и вопросы зависимости между лингвистическими формализмами текстов веб-контента и реальной сущностью общественно значимого события. Описана логико-лингвистическая модель извлечения фактов из текстовых массивов казахского, русского и английского языков. Приведены особенности формирования выравненного казахско-русского паралельного корпуса текстов криминальной тематики.

ISBN 978-601-332-917-8

УДК 80/81:004
ББК 81.2 Рус-5

СОДЕРЖАНИЕ

ОБОЗНАЧЕНИЯ И СОКРАЩЕНИЯ

IDF – Inverse Document Frequency
IE – Information Extraction
IR – Information Retrieval
KLC – Kazakh Language Corpus
LSA – Latent Semantic Analysis
ML – Machine Learning
MT – Machine Translation
NER – Named Entity Recognition
NLP – Natural Language Processing
NLTK – Natural Language Toolkit
PMI – Pointwise mutual information
PPMI – Positive Pointwise Mutual Information
POS-tagging – Part of Speech tagging
RDF – Resource Description Framework
RE – Relationship Extraction
RSS – Rich Site Summary
SD – Stanford Dependencies
SL – Supervised Learning
SSL – Semi-Supervised Learning
TF – Term Frequency
UD – Universal Dependencies
VSM – Vector Space Model
АККЯ – Алматинский Корпус Казахского Языка
АКП – алгебра конечных предикатов
БД – База Данных
ДТП – дорожно-транспортное происшествие
ИАС – информационно-аналитические системы
КЗИ – криминально-значимая информация
ПО – предметная область
СДНФ – совершенная дизъюнктивная нормальная форма
СМИ – средства массовой информации

ВВЕДЕНИЕ

В последние десятилетия, в связи с распространением сетевых компьютерных технологий, мобильной связи и Интернета, информационные ресурсы современного общества подвергаются растущему числу угроз, чреватых экономическим ущербом, и увеличивающих опасность национальной информационной инфраструктуры. Подобным атакам подвергаются как государственные, так и коммерческие системы. Рост криминальной активности в глобальных сетях (в таких формах, как финансовые мошенничества, нарушения авторского права, распространение детской порнографии, хакерство и т.д.) создает угрозы безопасности, как личности, так и общества в целом [1].

Чем больше расширяется Интернет, тем больше сетевых преступлений регистрируется. Благодаря компьютерным сетям насильственный экстремизм может глобально распространятся, сохраняя низкую стоимость и высокую скорость. Таким образом, открытость глобальной сети обусловливает ее большую уязвимость от преступных посягательств.

Одновременно, открытость и глобальность Интернета, представляющего собой всемирную телекоммуникационную сеть, создают огромные потенциальные возможности для криминалистов и работников правоохранительных органов. Существующие к настоящему времени технологии обработки текстов позволяют специалистам по анализу разведывательных данных и полиции осуществлять превентивную обработку текстовых данных компьютерной сети, собирая, соединяя и анализируя ‘слабые сигналы’ или ‘цифровые следы’ огромных текстовых массивов, которые присутствуют в Интернете. В некоторых случаях такой анализ может помочь обнаружить потенциал противоправного действия прежде, чем оно будет осуществлено.

Для этого, наряду с существующими традиционными способами борьбы с преступлениями в сфере безопасности обращения компьютерной информации, должны быть использованы практические достижения искусственного интеллекта и математической лингвистики, связанные с проблемами Natural Language Processing (NLP). При этом, одной из главных проблем определения криминального значения текстов Интернета, наряду с громадным объемом информации, подлежащей анализу [2], остается проблема слабой «окрашенности» криминальных текстов для использования традиционно принятых подходов классификации, кластеризации и выделения шаблонов NLP.

Традиционные подходы обработки языка, использующиеся в рамках решения задачи выделения криминально значимой информации и потенциально связанных с терроризмом текстов, базируются на анализе стиля текста и определе-

нии его эмоциональной составляющей, связанной с неявно выраженным намерением, но не учитывают тему и содержание текста [3].

Данная монография, рассматривает информационно-лингвистическую технологию автоматического определения, выделения, поиска и анализа криминально значимой составляющей в неструктурированных и слабоструктурированных тестовых массивах различных языков, фокусируясь на содержании текста и выделении фактов.

В работе расматриваются:

- основные проблемы в области существующих технологий поиска противоправной информации в текстах;
- существующие проблемы формализации и автоматической обработки казахского языка.

В работе предлагаются:

- логико-лингвистическая модель извлечения фактов из текстовых массивов;
- имплементация данной модели для казахского, русского и английского языков;
- информационная технология идентификации криминально-значимой информации в текстовых корпусах казахского языка;
- описание созданного выравненного паралельного казахско-русского корпуса криминально окрашенных текстов.

Монография выполнена в рамках грантового проекта «Методы и модели поиска и анализа криминально значимой информации в неструктурированных и слабоструктурированных текстовых массивах», ИРН проекта AP05131073

1. ОСНОВНЫЕ ПРОБЛЕМЫ В ОБЛАСТИ ТЕХНОЛОГИИ ПОИСКА ПРОТИВОПРАВНОЙ ИНФОРМАЦИИ В ТЕКСТОВЫХ МАССИВАХ

1.1.Состояние и перспективы развития методов формализации и поиска криминальной информации в неструктурированных текстах

Большая часть исследований, связанных с предотвращением террористических атак, направлена на анализ использования террористами и террористическими организациями Интернета и социальных сетей [4], [5], [6], [7]. Так, одно из научных направлений, посвященных обнаружению “лингвистических маркеров насильственного экстремизма в онлайн среде” [8] фокусируется именно на идентификации цифровых следов, которые имеют отношение к потенциальному «террористу-одиночке» [9]; другие исследования рассматривают потенциальные виды Интернет-насилия [10]. Такие направления информатики, как поиск противоправной информации в текстовых данных, обнаружение шаблонов преступления и оценки риска возможных киберпреступлений, становятся одними из самых популярных исследований NLP. Всё больше исследователей фокусируются на способах и формах применения технологий обработки естественного языка в рамках широкого спектра видов деятельности, имеющих отношение к предотвращению террористической активности.

Например, для обнаружения лингвистических маркеров, которые свидетельствуют о потенциальном «предупреждающем» поведении, в работе [8] предложено использовать списки слов насильственных действий, подготовка и поиск которых базируется на стандартных подходах обработки текстов, таких как лемматизация и Part of Speech tagging (POS-тегирование), а так же на использовании лексических баз данных, подобных WordNet [11]. Однако, такие лингвистические маркеры, использующиеся в качестве дополнения к стандартным алгоритмам обработки текстов, могут распознать потенциальные признаки оговоренного, заранее предполагаемого радикального насилия. Они не могут принять автоматизированные решения по любым видам преступлений. Кроме того, если отдельные этапы обработки естественного языка будут неточными, точность выделения лингвистических маркеров значительно уменьшится, и количество ошибок увеличится.

Еще одним направлением NLP, используемом в рамках решения задачи выделения криминально значимой информации и потенциально связанных с терроризмом текстов, является анализ стиля текста и выявление его эмоциональной составляющей, связанной с неявным выражением намерения. Такая эмоциональная составляющая может включать хвастовство, идеологические за-

явления или восхищение террористическими лидерами [12]. Текстовый анализ стиля, в этом случае, позволяет обнаружить шаблоны фраз, связанных с такими эмоциональными мотивациями как гнев, унижение или позор. В данном контексте следует подчеркнуть, что стиль общения не зависит от определенной темы или от содержания. Добавление для его анализа более глубокой психологической обработки, использование лингво-психологии речевой деятельности и социолингвистики, позволяет не только идентифицировать «предупреждающее» преступление поведение, но и раскрыть, в некоторых случаях, корпоративное мошенничество [13].

Для анализа больших объемов разнородных текстов, тематика которых не известна заранее, используются методы классификации и кластеризации NLP. Например, объединение в кластеры может выделить такие темы как оружие, тактика или цели [12]. В этом случае, дополнительное использование технологий распознавания речи и машинного перевода могут значительно увеличить объем текста, доступного для анализа.

Одной из разновидностей классификации является сентимент анализ (Sentiment Analysis). Различные формы онлайнового выражения авторского мнения (например, обзоры, личные мнения, рейтинги и рекомендации) стали основными источниками информации как для компаний, надеющихся продавать свои продукты и управлять своей репутацией [14], так и для СМИ, определяющих отношение общества к реальным событиям. Например, в работе [7] сентимент анализ используется при анализе твитов для определения мнений авторов по отношению к определенным зонам преступлений в режиме реального времени. Кроме того, многие исследования, которые сосредоточились на обнаружении шаблона преступления, используют методы сбора данных в их временном изменении. Такие исследования, кроме твитов, блогов и социальных сетей, используют информацию СМИ для обнаружения преступления в каждой определенной области [15].

Методы классификации NLP довольно хорошо разработаны и отлажены. В то же время, их использование при анализе эмоциональной составляющей текста или выявлении намерения, не всегда дает хорошие результаты. Основным недостатком подобных подходов является: не специфичность выделеных закономерностей, когда выявленные закономерности (даже если они явно угрожающие), могут быть не связаны с угрозами, и их интерпретация часто зависит от культурных и отдельных особенностей человека. Обычно, анализ и классификация текстов исключает непрямую терминологию, которая явно не называет оружие или насильственное действие и не включает угрожающую лексику, т.е. терминологию строго не относящуюся к криминалу.

В рассмотренных выше статьях и подходах, при семантическом анализе эмоциональной составляющей текста, параграфы, которые представляют факты, как правило, удаляются, и исследователи сосредотачиваются на параграфах, в которых автор выражает свое мнение, используя распространенные классификаторы — наивный Байесовский метод, максимальную энтропию или метод опорных векторов. В нашем исследовании мы, наоборот, предлагаем фокусироваться именно на фактическом материале текстов, и использовать технологии, базирующиеся на подходах Open Information Extraction (IE), в частности на методах извлечении фактов из слабоструктурированных текстов [16].

1.2. Общий подход к формализации и идентификации криминально значимой информации

Весомый вклад в формирование научных основ информатизации криминалистической деятельности в разное время внесли такие известные ученые-криминалисты, как Т. В. Аверьянова (исследование автоматизации получения и использования информации в криминалистической деятельности) [17]; Л. Е. Ароцкер (определение идентификационных и неидентификационных методов работы с криминалистической текстовой информацией) [18]; А. В. Астахова (возможности использования компьютерных экспертных систем в исследованиях криминалистики) [19]; Д. Д. Бегов (проблемы автоматизации криминалистических экспертиз, пути создания технических систем для определения эмоционального состояния лица в криминалистических целях) [20]; Р. С. Белкин (философская теория отражения как гносеологическая основа криминалистической науки) [21] и другие.

Исследования ученых в основном проводились в четырех направлениях: (1) выяснение возможности использования математических методов и средств вычислительной техники в экспертной практике; (2) решение проблемы, связанной с удобными формами поиска, хранения, обработки и передачи криминалистической информации; (3) определение роли и места отдельных компьютерных технологий в информационном обеспечении судебной экспертизы; (4) изучение и оценка типичных трудностей, имеющих организационный, правовой, научный и психологический характер на этапе автоматизации отдельных видов судебной экспертизы.

Сегодня, современное интенсивное наполнение информационного пространства и наличие в нем информации, имеющей значение для оперативно-служебной деятельности правоохранительных органов, формирует ряд новых вызовов, связанных с возможностью поиска и автоматического извлечения

криминально значимой информации. И хотя, по-прежнему, деятельность любого правоохранительного подразделения направлена, прежде всего, на раскрытие, предупреждение и пресечение преступлений и правонарушений, появились новые направления и возможности данной деятельности, связанные с доступом к огромным информационным потокам.

При этом, всю информацию, которая представляет интерес для правоохранительных органов, можно разделить на два уровня: уровень отдельного самостоятельного структурного подразделения, интерес которого определяется контрольным перечнем необходимой информации; и макро- или общий уровень, представляющий любую информацию, имеющую признаки криминальной среды. Именно информация макро-уровня и будет интересовать большинство правоохранительных структур.

Поиск информации в новых источниках – это только одна из задач, которая решается при выполнении оперативно-служебной деятельности. Другой важной задачей является ретроспективный поиск латентных закономерностей в неструктурированных массивах, подобных суточным сведениям событий в правоохранительных органах. Это задача состоит в поиске событий схожих по каким-то параметрам (место совершения, вид, механизм, участники и т.д.), зарегистрированным в суточных ведомостях либо в несвязанных между собой источниках.

Индикативным признаком криминально значимой информации, который отличает ее от обычной информации, является понятие состава преступления. Состав преступления представляет собой предусмотренную действующим законодательством систему объективных и субъективных элементов, характеризующих определенное общественно опасное деяние – т. е. конкретное преступление [22]. Взаимосвязь компонентов, являющихся первичными составляющими системы «состав преступления» (объект, объективная сторона, субъект, субъективная сторона) показана на рисунке 1.1.

Информационное обеспечение информационно-аналитической системы криминалиста может учитывать только три элемента состава преступления: объект, субъект и объективная сторона, поскольку субъективная сторона не несет необходимой смысловой нагрузки.

Конкретные элементы состава преступления определяются контрольным перечнем нужной информации для конкретного подразделения правоохранительных органов в соответствии с его юрисдикцией. Формирование непосредственно самого контрольного перечня нужной информации осуществляется исходя из диспозиции нормы права, и подпадает под юрисдикцию подразделения.

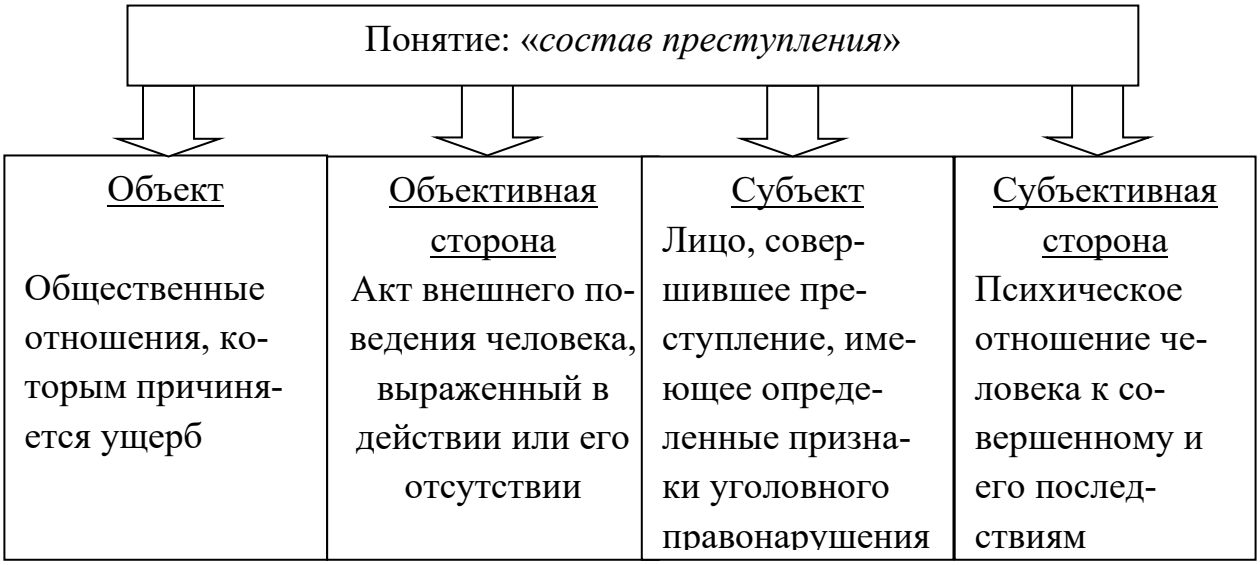


Рисунок 1.1 – Структура индикативных признаков криминально значимой информации

При прогнозировании преступлений, выявлении признаков скрытых преступлений, установлении зависимости между личными качествами преступников и выбором места совершения преступления, а также другой аналитической следственно-розыскной деятельности, следователю (или другому процессуальному лицу) необходимо обработать большое количество электронных текстовых документов, извлекая из них криминально значащую информацию. Этими электронными текстами могут быть как документы, имеющие электронную форму: объяснительные, служебные записки, отчеты, словесные портреты фигурантов, протоколы, которые накопились в результате расследования, так и электронные коллекции интернет публикаций, RSS - рассылок и социальных сетей.

Все подобные электронные документы представлены в виде слабоструктурированной текстовой информации, под которой понимается — текстовый электронный документ, имеющий высокую степень вариативности контента, меняющегося в зависимости от конкретной ситуации. В целом, эти документы представляют доступный репозиторий криминалистических знаний [23].

При этом, главное качество криминально значимой информации заключается в содержании информации, способствующей поиску доказательств и закономерностей, присущих именно криминалистическим аспектам преступной деятельности. Иными словами, криминалистическая характеристика преступления, как средства оптимизации расследования, должна представлять собой со-

вокупность информации, имеющей не квалифицирующее или процедурное и предупреждающее, а именно поисково-познавательное значение [24].

В общем случае, информационные процессы, связанные с расследованием преступления, получением криминально значимой информации (КЗИ), а также данных и фактов из массивов электронных текстовых документов и электронных ресурсов, можно представить в виде схемы, показанной на рисунке 1.2.

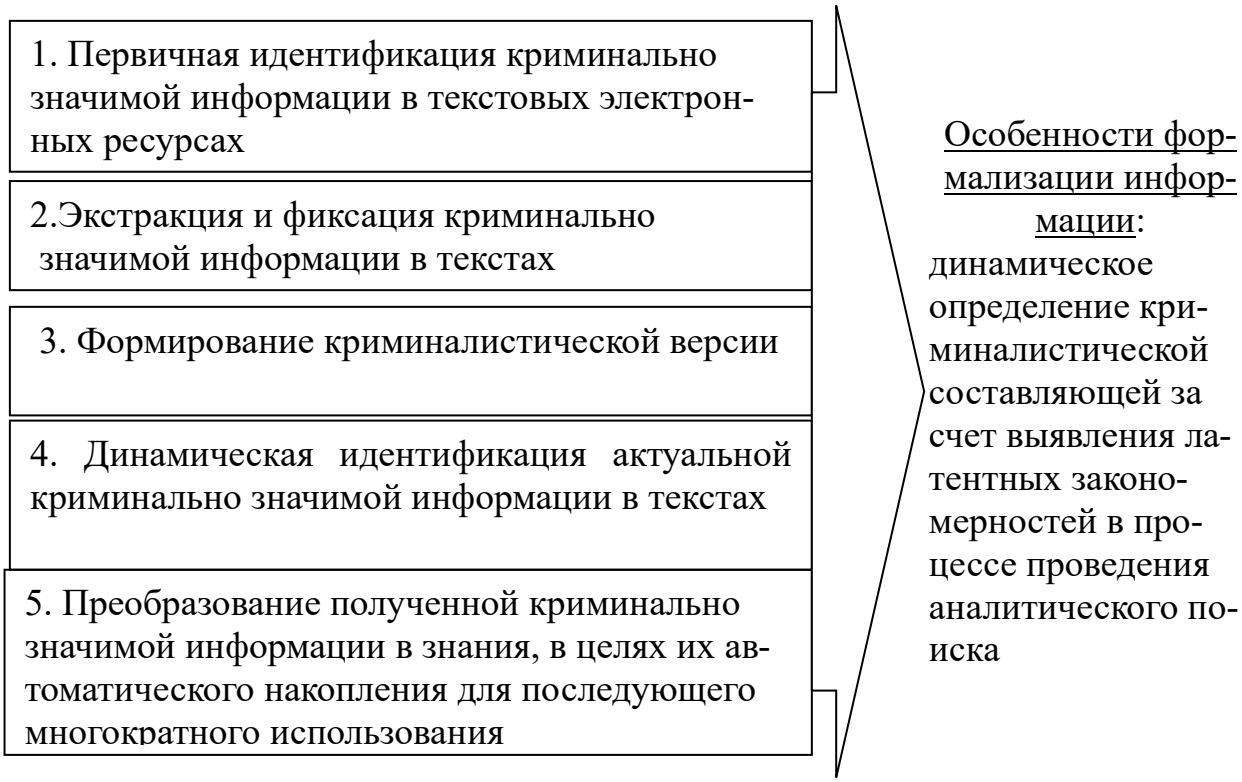


Рисунок 1.2 – Общая схема информационных процессов, связанных с получением криминально значимой информации из электронных текстовых ресурсов

Актуальная КЗИ, которая часто не имеет причинно-следственной связи с событием преступления, но имеет потенциальное криминалистическое значение, не позволяет при ее поиске использовать предварительно разработанный тезаурус, заранее известной предметной области. Такая информация, с одной стороны, характеризуется дефицитом выходных признаков, а с другой стороны, не позволяет использовать для ее идентификации только ключевые слова, описывающие преступные деяния, которые, будучи своего рода индикативным признаком, обычно имеют свою специфику.

Для последующего долгосрочного использования актуальную КЗИ необходимо трансформировать в знания, извлекая новые понятия, которые не всегда являются идентификаторами криминальности, и осуществляя их систематизацию. Таким образом, необходимо осуществлять динамичное распространение и

накопление криминалистических знаний за счет обработки новой текстовой информации корпусов документов и ссылок.

1.3. Обзор существующих возможностей использования методов Information Extraction для извлечения криминально значимой информации

В последние годы резко повысился интерес к такому направлению исследований искусственного интеллекта и математической лингвистики, как поиск информации (Information Extraction), и связанным поиском фактов или фактографической информации. Общая цель исследований в области IE заключается в возможности извлечения информации из ранее неструктурированных данных. Более конкретная цель, связанная с исследованиями именно криминально значимых текстов, заключается в возможности получения фактов, на основе которых можно провести логические рассуждения и сделать выводы о криминальной окрашенности текста.

Факт, в общем случае, представляет собой зафиксированное, классифицированное событие, которое произошло. В информатике факт явным формальным образом представляют в виде триплета: *Субъект ->Предикат -> Объект* [16]. Субъектами фактов, как правило, являются сущности, свойства которых (временные, пространственные, качественные, количественные и т.д.) выделяются дополнительно в виде атрибутов факта. При этом, факт может быть извлечен из текстовой информации (как слабо структурированной, так и не структурированной), и может определять как свойства объекта, так и связь объекта с другими объектами.

Как правило, на первом этапе задачи поиска фактов осуществляется распознавание именованных сущностей (Named Entity Recognition (NER)), включающее в себя - определение известных имен сущностей (для людей и организаций), географических названий, временных выражений и некоторых типов многочисленных выражений, которые используют существующие знания о домене или информацию, полученную из других предложений. Как правило, задача распознавания включает в себя назначение уникального идентификатора для извлечения сущности.

При идентификации криминально значимых фактов задача распознавания сущностей направленна на выявление лиц, при отсутствии каких-либо знаний о конкретном экземпляре сущностей. Например, при обработке предложения "*М. Смит принял участие в организации встречи*", распознавание сущности означает понимание того, что наименование "*М. Смит*" действительно относится к интересующему нас человеку. При этом, не обязательно знания о некоем *М.*

Смите, о котором идет речь в этом предложении, существовали до начала его анализа.

Следующий этап формирования факта представляет собой экстракцию отношений между субъектами (Relationship Extraction (RE)), определяемых фактом.

При этом, наибольшие трудности при выделении фактографической информации из неструктурированных текстов возникают при экстракции знаний из открытых областей, а также при обработке «временных» знаний [25], к которым и относится криминально значимая информация. Особая сложность заключается в том, что один и тот же факт может быть выражен различными грамматическими конструкциями и различными словами, определяющими сущности и отношения, устанавливаемые между ними.

На момент проведения исследования вопрос извлечения фактов из текстов широкой тематической направленности остается открытым. Существующие общие модели и подходы напрямую зависят от степени специфичности и структурированности текста. Несмотря на то, что существует довольно много методов извлечения фактов из структурированных текстовых данных [26], [27], [28], надежной технологии извлечения фактов из полу-структурированных и неструктурированных текстовых данных на рынке пока нет [29], [30], [31].

В то же время, интересующие нас тексты социальных сетей, средств массовой информации и других интернет источников представлены именно в неструктурированном виде, именно, факты, извлеченные из них, должны составлять базу криминально значимой информации для последующей аналитики.

Текущие исследования в области интеллектуального анализа текста, в основном, базируются на статистических моделях, использующих машинное обучение с учителем (Supervised Learning (SL)), частичное машинное обучение (Semi-Supervised Learning (SSL)) и открытое извлечение информации (Open IE). Основная сложность экстракции фактов в заранее не определенных предметных областях методами машинного обучения с учителем и полуавтоматическом обучении заключается в необходимости размеченной обучающей выборки [32]. Открытое извлечение информации (Text Runner system), как правило, позволяет извлекать только бинарные отношения из простого текста и имеет не очень высокую полноту и точность [33].

В целом, использование статистических методов для извлечения информации и, в частности, для экстракции фактов, малоэффективно. Прежде всего, это связано с тем, что статистические методы рассматривают документы как неупорядоченный «мешок слов», что хорошо реализуется в задачах информационного поиска и классификации текстов, но не может быть использовано при

извлечении фактов, когда единицей обработки становится предложение, а не корпус текстов [34].

Еще одной причиной низкой эффективности статистических методов извлечения фактов является невозможность учета такими методами синтаксиса и семантики предложений, а также омонимичности, синонимичности и многозначности естественного языка. В то время, когда *Предикат*, *Субъект* и *Объект* факта могут быть представлены различными словами и даже различными частями речи. Например, английские предложения "*The company management sold a part of share*", "*Management of Apple Inc. sold their share*", и "*They marketed its*" репрезентуют один и тот же факт (рис. 1.3).

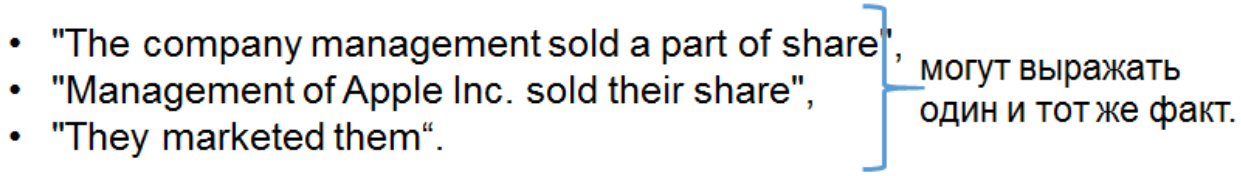


Рисунок 1.3 – Пример представления одного факта
различными синтаксическими и лексическими структурами

Опираясь на проведенный анализ, в данном исследовании для извлечения фактографической информации из слабоструктурированных и неструктурированных текстов мы предлагаем использовать:

1) логико-лингвистические модели извлечения фактов из слабоструктурированных и неструктурированных текстов;

2) формализацию грамматических способов выражения одного и того же факта в предложениях;

3) модель семантической близости коротких фрагментов текста.

2. ЗАВИСИМОСТЬ МЕЖДУ ЛИНГВИСТИЧЕСКИМИ ФОРМАЛИЗМАМИ ТЕКСТОВ ВЕБ-КОНТЕНТА И РЕАЛЬНОЙ СУЩНОСТЬЮ ОБЩЕСТВЕННО ЗНАЧИМОГО СОБЫТИЯ

2.1. Существующие методы генерация структурированной машинно-читаемой информации из неструктурированных текстов

К настоящему времени проблема извлечения информации и фактов из неструктурированных текстов окончательно не решена. Существующие модели и алгоритмы извлечения фактографической информации зависят от степени структурированности анализируемого документа [35]. Подобно общей классификации степени формализации информации, мы можем разделить текстовые документы по степени структурированности на: (1) хорошо структурированные тексты, часто представленные табличными данными; (2) полуструктурированные текстовые документы, описывающие конкретный домен (например, патенты, справки, отчеты и т.д.), и (3) неструктурированные тексты любой предметной области (например, тексты веб-медиа) [36].

Для извлечения фактов, представленных в структурированных текстовых документах, существуют достаточно надежные алгоритмы [37, 38, 39]. В то же время, несмотря на постоянный рост интереса к исследованиям, направленным на поиск способов выявления и извлечения фактов из текстовых корпусов и веб-контента, в настоящее время, не существует общего достоверного метода излечения структурированной информации из неструктурированных разнородных текстов [40, 41]. Рост интереса к данному направлению исследований связан, прежде всего, с огромными объемами текстовой информации в корпоративных и Интернет сетях, представленной в неструктурированном и слабоструктурированном виде (по некоторым источникам, такой информации более 85%). Кроме того, растущий интерес к идентификации и генерации фактов на базе текстовой информации обусловлен расширением областей использования такой структурированной информации.

Например, извлечение фактов из неструктурированных текстов может стать серьезным дополнительным источником для создания онтологий на основе знаний веб-контента. Недавние подходы Open IE извлекают факт в виде триплета *Субъект -> Предикат -> Объект*, где *Объект* и *Субъект*, обычно, представлены существительными или именными фразами, тогда как *Предикат* в основном выражается глаголом. Такой подход соответствует представленному на рисунке 2.1 RDF-графу, структурно отображающему некоторый элемент знания.

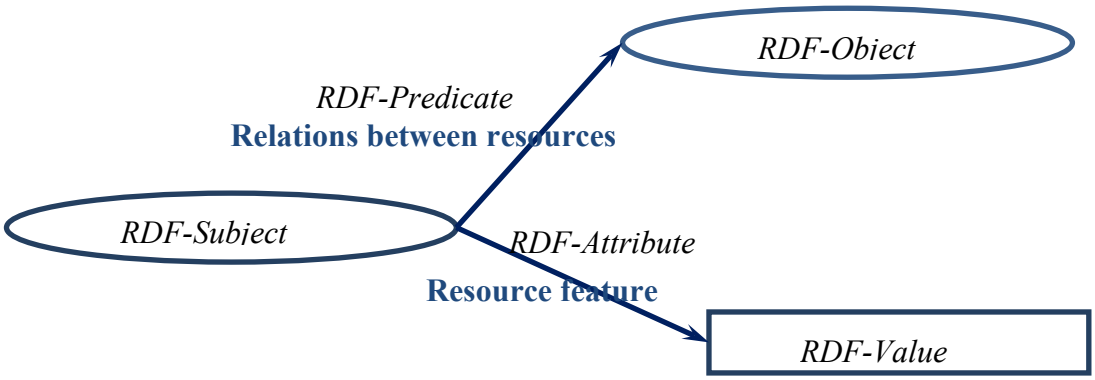


Рисунок 2.1 — Представление RDF-триплета,
соответствующее концепту факта в моделях Open IE

Сегодня существует два основных подхода в направлении извлечения информации из неструктурированных текстов: IE и Open IE. Оба этих направления позволяют обрабатывать большие объемы текстов, содержащих относительно небольшое количество фактической информации. При этом, методы IE можно рассматривать как особый вид Information Retrieval (IR), когда запрос к поисковой базе сформулирован заранее. Однако, результатом IE являются структурированные данные, описывающие факты из набора документов, в то время как результатом IR является набор ссылок на документы, соответствующие запросу.

Первые IE-системы были, в основном, ориентированы на конкретную предметную область и основывались на знаниях, полученных в процессе предварительной разработки. Примером такого подхода является одна из первых систем IE, работающая с текстами, посвященными латино-американскому терроризму, которая использовала заранее разработанные морфосемантические шаблоны [42, 43]. Современные системы IE так же используют предопределенный набор правил, которые позволяют идентифицировать информацию, определяющую тот или иной конкретный факт [44]. Большинство систем IE извлекают и представляют информацию в форме кортежей из двух объектов, с заранее предопределенным типом отношений между ними [45]. Таким образом, подходы IE, направленные на создание предопределенных структур знаний, не позволяют работать с произвольным веб-контентом неограниченного объема знаний, где целевые отношения не могут быть заданы заранее.

Технологии IE обычно используют статистические методы, а также методы контролируемого (supervised) и неконтролируемого (unsupervised) машинного обучения Machine Learning (ML) [46]. Дополнительно используются методы распознавания специфических доменных объектов (лица, названия компаний и т.д.), базирующиеся на традиционных подходах NER; синтаксический парсер и семантическое тегирование [47, 48].

Новая, появившаяся в 2007 году, парадигма извлечения знаний — Open IE [43], позволяет идентифицировать неограниченное число отношений и, следовательно, не зависит от доменной ориентации. Open IE включает широкий спектр задач: (1) идентификация и отслеживание сущностей, (2) идентификации отношений и атрибутов этих сущностей, (3) определение и характеристика событий.

Большинство приложений Open IE включают такие инструменты NLP как POS-tagging, а так же синтаксический анализ зависимостей (Dependency parsing) [49, 50]. Кроме того, данные приложения используют лексические ограничения [51] или семантические аннотации [52] для ограничения количества возможных специфических отношений [53].

Основные причины неэффективности использования статистических методов при решении задач Open IE заключаются в следующем. Прежде всего, статистический подход, используемый в задачах IR, классификации или кластеризации текстов, рассматривает документ как неупорядоченный «мешок слов» (bag of words) [54]. При этом знания, связанные с грамматикой и семантикой во многом теряются.

Вторая причина, не позволяющая использовать подход “мешка слов” в задачах Open IE, связана с очевидной необходимостью извлекать факты из предложений, а не из полного текста (full text) [55]. Такой подход связан с ранее упоминавшемся представлением факта в форме триплета: *Субъект –> Отношение –> Объект*. При данном подходе знания об объектах/субъектах некоторой ПрО, их свойствах и отношениях представляют собой совокупность сведений, выражаемых в изолированных предложениях.

Третья причина низкой эффективности использования статистических методов в задачах Open IE связана с синонимией и неоднозначностью языковых единиц, что приводит к появлению скрытых фактов в тексте [56]. Одной из таких проблем является разрешение кореференции, когда одни и те же сущности или действия представлены разными словами (иногда, разными частями речи).

На сегодняшний день проблема автоматического извлечения фактов исследуется для всех языков, и имеет высокий уровень реализации не только для текстов английского языка, но и для многих других.

Например, в работе [57] был проведен эксперимент для оценки адекватности использования фактической плотности и информативности 50 случайно выбранных документов испанского языка в CommonCrawl корпусе. В недавнем исследовании [58] плотность простых и сложных фактов рассматривалась как характеристики измерения качества статей в русской Википедии. В работе [59]

была представлена первая система Open IE, которая способна извлекать триплеты фактов из произвольных текстов китайского языка.

Однако, несмотря на достигнутые результаты, сегодня не существует мультиязычных стандартных методик и подходов Open IE [57], в частности, для языков с ограниченными лингвистическими ресурсами, к которым относится и казахский язык.

2.2. Гносеологические аспекты информационных процессов идентификации некоторых семантических/лексических и грамматических идентификаторов криминальности

Расследование преступлений представляет собой динамическую систему, основная функция которой заключается в эффективном противостоянии преступной деятельности. Такое расследование можно рассматривать как вид познавательной деятельности, имеющей специфические черты. Уголовно-процессуальное законодательство республики Казахстан определяет формы, средства и сроки деятельности, осуществляемой органами досудебного следствия и дознания при расследовании преступлений. Содержание этой деятельности составляют процессы обнаружения, фиксации, изъятия, хранения и использования информации, имеющей отношение к расследуемому событию и установлению истины по делу. Указанные процессы называются информационными и образуют в познавательной деятельности гносеологический ряд: *Факт* $\rightarrow$ *Отражение* $\rightarrow$ *Информация* $\rightarrow$ *Знания*.

Терабайты базовой текстовой информации данной познавательной деятельности хранятся в информационных сетях Казахстана, ежедневно пополняясь. Все информационные ресурсы, используемые правоохранительными органами, можно разделить на два типа: внутренние и внешние.

Для внутренних информационных ресурсов государственных и правоохранительных органов характерно наличие больших массивов данных, которые представляются в виде различных текстовых файлов: неструктурированные данные, производимые в процессе административной, оперативно-розыскной, следственной, аналитической и иной деятельности.

Но, кроме внутренних информационных ресурсов, оперативным подразделениям часто нужны такие данные, как сведения о конфликтах (уголовные, экономические, политические, бытовые, религиозные, семейные и т.д.); данные о деяниях с признаками противоправности (незаконная производственная и коммерческая деятельность, завладение движимым и недвижимым имуществом); разбойные нападения, мошеннические действия, пожары с признаками поджогов, массовые драки, массовые протесты и другие нарушения обществен-

ного порядка. Такие данные, в основном, содержатся в различных текстовых массивах, заранее не определенных как база оперативно-розыскной деятельности, и не являются четко выраженной криминальной информацией. Это могут быть, например, социальные сети, справочники, каталоги, форумы, которые могут содержать данные о фигурантах уголовного дела и, при этом, не иметь уголовной окраски. Так же, это могут быть рекламные объявления, содержащие данные о мошенниках, о нелегальной экономической деятельности в сфере производства и финансов, и, при этом, ничем не отличающиеся от обычных объявлений для потребителя.

В общем случае, внешний ресурс частично состоит из массивов файлов аналогичных внутренним, но, обычно, представленных в Веб-форматах. К внешним источникам, в частности, относятся: средства массовой информации (СМИ); данные различных учреждений, организаций, предприятий (картотеки, архивы, библиотеки, электронные массивы); данные других государственных органов; сеть Internet со всеми ресурсами; корпоративные сети, социальные сети.

Таким образом, особенность извлечения КЗИ, в основном, определяется тем, что криминальная значимость некоторого множества данных будет определяться только множеством метаданных. Множество метаданных формируется в результате обработки массива определенных криминально окрашенных и неокрашенных текстов, содержащих некоторые семантические/лексические или грамматические идентификаторы криминальности.

Таким образом, структурно, всю информацию, представляющую некоторый интерес для правоохранительных и других заинтересованных государственных органов, можно представить двумя уровнями:

– уровень отдельного структурного подразделения, который определяется контрольным перечнем необходимой информации;

– макро или общий уровень, включающий любую информацию с признаками криминальной среды.

Мировые спецслужбы пришли к выводу, что, на макро-уровне или в открытых источниках информации, может содержаться большое количество знаний, представляющих интерес для государственных и правоохранительных органов [60]. Однако, для того, чтобы получать необходимую информацию из неструктурированных данных и проанализировать ее, необходимо иметь специальный инструментарий, в основе которого находится определенная информационно-лингвистическая технология.

Одной из таких наиболее известных технологий является технология Text Mining, представляющая алгоритмическое выявление неизвестных связей и

корреляций в имеющихся текстовых данных. В то же время, существующие на сегодня традиционные подходы Text Mining (реферирование, машинный перевод, классификация, кластеризация, диалоговые системы, тематическое индексирование, средства поддержки и создания таксономий и тезаурусов, и поиск по ключевым словам) не позволяют получить лингвистические маркеры КЗИ.

Анализ КЗИ, представленной в открытых внешних для правоохранительных органов дополнительных источниках информации, должен включать следующие необходимые этапы:

1) определение источников, содержащих не только криминально окрашенную информацию, но и криминально значимую информацию;

2) исследование возможности извлечения такой информации и ее структурирования, на основе алгоритмов, составленных с учетом особенностей предметной области (ПО).

Как правило, при определении части общего информационного пространства, интересующего правоохранительные органы, прежде всего, необходимо выделить информацию о преступлении или о его потенциальной возможности. Обычно формирование контрольного перечня необходимой информации осуществляется из диспозиции нормы права, попадающей под юрисдикцию того или иного подразделения. С точки зрения информатики, диспозиция является фильтром для объективной стороны состава преступления. В тоже время, при анализе текстовой информации, находящейся во внешних для правоохранительных органах источниках, обычно, не представляется возможным выделить и формализовать такой контрольный перечень необходимой информации.

Получения криминально значимых данных является нетривиальной процедурой обработки текста, зависящей от глубины анализа и задач, стоящих перед специалистом или автоматической аналитической системой, которая осуществляет анализ.

В основу разрабатываемого подхода выделения лингвистических маркеров КЗИ могут быть положены следующие имеющиеся разработки:

– классификация текстов с использованием статистических критериев для построения правил отнесения документов к определенным категориям;

– кластеризация, основанная на признаках документов, осуществляемая без выделения определенных категорий и использующая лингвистические и математические методы с возможным применением таксономии и онтологии, обеспечивающих эффективный охват больших объемов данных;

– построение семантических сетей для анализа связей, которые определяют появление дескрипторов (ключевых фраз) в документе при осуществлении поиска;

– экстракция фактов — извлечение фактов из текста, с целью улучшения классификации, поиска и кластеризации.

По сути, выполнение перечисленных задач в представленной последовательности и является процессом осмысления текстовой информации с целью выявления новых знаний.

2.3. Методы выявления семантических идентификаторов КЗИ в корпусе текстов

Корпуса криминально окрашенных текстов должны, наряду с морфологической разметкой, содержать элементы семантического аннотирования. Семантическая разметка важна не только для будущих исследований языка, вопросов сочетаемости лексических единиц, разработки семантического словаря криминально окрашенной лексики, но так же для выделения лингвистических идентификаторов КЗИ.

Существует несколько основных подходов семантической обработки текстов, ориентированных на конкретный домен:

1) ручное (интеллектуальное) наделение объекта некоторыми атрибутами, и обработка именно этих атрибутов;
2) использование частотных словарей;
3) метод латентного семантического анализа – Latent Semantic Analysis (LSA).

К первому подходу, включающему большую часть ручного (интеллектуального) труда, можно отнести: семантическое тегирование, ручную каталогизацию, использование онтологий и концепцию Веб 3.0. При этом создается база знаний, представляющих RDF-триплеты, созданные вручную или автоматически полученные из обрабатываемых текстов.

Второй подход, позволяющий обрабатывать семантику и находить общие смысловые элементы в текстах, базируется на использовании частотных словарей. Для учета разных размеров/объемов корпусов обычно учитывают относительную частоту слов в корпусе (instances per million words) [61]. Словари можно создать на базе имеющихся корпусов, классифицированным по разным темам. Например, слово, “*shotgun*” может встречаться в корпусе новостных текстов, связанных с криминальной информацией, во много раз чаще, чем в корпусе новостных текстов, имеющих отношение к экономике. Однако, когда речь идет об узкой специализации конкретного домена, использование словарей, обычно, дает менее значимый эффект.

Третий подход использует статистические вычисления, методы machine и deep leaning, которые базируются на гипотезе о том, что близкие по смыслу слова встречаются в подобных контекстах, а близкие по смыслу тексты содержат семантически подобные слова. Информацию о совместной встречаемости можно формально представить в виде матрицы или в виде набора векторов в многомерном пространстве векторов Vector Space Model (VSM). Модель векторного пространства имеет ряд базовых преимуществ перед стандартной булевой моделью. Прежде всего, VSM — это простая модель, построенная на основе линейной алгебры; кроме того, данный подход позволяет вычислять непрерывную степень сходства между терминами и документами.

В нашем исследовании в качестве исходной информации метода LSA используется векторная модель документа, описывающая набор данных обученного корпуса. При этом не учитывается порядок слов в документе и их морфологические формы, а учитывается только количество вхождений конкретной леммы [62] в текст. При таком подходе строки матрицы «термин-документ» соответствуют леммам (где T — общее количество слов или лемм в корпусе), а столбцы — текстам нашего корпуса, где D — общее количество текстов или документов.

Такая матрица может представлять собой матрицу инцидентности; тогда в ее ячейках содержатся нули и единицы: 1, если слово есть в документе, и 0, если данного термина нет в данном документе. В более сложном случае, ячейки матрицы могут содержать количество появлений того или иного термина в документе, представленного весом термина, учитывающего частоту использования каждого терма в каждом документе и встречаемость терма во всех документах (*TF-IDF*). Для того чтобы сравнить семантику двух документов, нужно определить степень сходства двух столбцов таблицы или косинусное сходство векторов в векторной модели документа:

$$\text{Tf-idf}\,(t,d,D) = \text{tf}\,(t,d) \times \text{idf}\,(t,D), \qquad (2.1)$$

где *Tf* — частота термина; *idf* — инвертированная частота документов, вычисляемая как частное количества текстов, в которых данный термин встречается, деленое на общее количество текстов в корпусе; t — анализируемый термин; d — текст, в котором обнаружен данный термин; D — общее количество текстов в корпусе.

В нашем исследовании в качестве значения вектора мы используем величину Positive Pointwise Mutual Information (PPMI). Метрика Pointwise mutual information (PMI) была предложена как вероятностная величина, определяющая насколько чаще события x и y происходят одновременно, по сравнению с тем, как они бы происходили, если бы они были совсем независимы. PMI между

двумя событиями определяется как вероятность того, что эти два события происходят совместно, деленная на произведение вероятностей двух независимых событий, с взятием логарифма от данного деления.

Применяя эту формулу для проверки совпадения контекстных векторов определяем PMI между целевым словом *w* и контекстным словом *c* как логарифм вероятности одновременного проявления двух слов совместно, деленной на произведение вероятности появления каждого из двух слов отдельно:

$$PMI(w,c) = \log_2 \frac{P(w,c)}{P(w)P(c)} \tag{2.2}$$

Область значений PMI в диапазоне от минус бесконечности до плюс бесконечности. Но, так как отрицательные значения обозначают, что слова (целевое и контекстное) появляются совместно реже, чем они бы появлялись, даже если бы были совсем независимыми, рассматриваются только положительные значения PMI, все отрицательные значения заменяются на нуль.

$$PPMI(w,c) = \begin{cases} PMI(w,c), & \text{если } PMI(w,c) > 0 \\ 0, & \text{противном случае} \end{cases} \tag{2.3}$$

Для того чтобы учесть возникающую при расчете PMI проблему редких слов, т.е. слов, которые не встретились в созданных корпусах и, следовательно, вероятность совместного появления данных слов равна нулю, используется сглаживание Лапласа. Идея сглаживания Лапласа заключающееся в следующем: мы считаем, что каждое слово появлялось в тексте на два раза больше, чем оно появлялось на самом деле, т. е. изначально при формировании вектора прибавляем два к частоте появления каждого слова.

2.4. Технология поиска семантически близких коротких фрагментов текста

Слова, описывающие преступные деяния, имеют свою специфику, и часто именно они являются индикативным признаком, по которому осуществляется отбор документов, предназначенных для последующей аналитической обработки. Специалисту понятны такие словосочетания как *ножевое ранение*, *признаки насилия, огнестрельное ранение, взрывчатое вещество, наркотическое вещество, угон автомобиля, завладение имуществом, умышленный поджог, кража денег* и т.п. Однако, иногда, интересно выявить менее привычные, но более эффективные сочетания слов для поиска криминально значимой информации, например, "*винт солянка*". У профессиональных работников правоохранительной системы и общепонятные и профессиональные словосочетания вызывают

ассоциации с определенным видом преступления, а, следовательно, их наличие в тексте требует, по крайней мере, глубокого изучения этого текста.

В связи с этим, на первом этапе обработки массива криминально окрашенных текстов необходимо выделить именные словосочетания или коллокации, используемые в качестве объектов или характеристик данных объектов, которые определяются через взаимное информационное влияние слов в предложении. В рамках семантико-синтаксического подхода, коллокации (устойчивые словосочетания) рассматриваются как синтаксически связанные, лексически определённые элементы грамматических структур, которые характеризуются семантической, синтаксической и дистрибутивной регулярностью.

При выделении коллокаций рассматриваются только предложения, подчиняющиеся закону проективности, то есть предложения «делового стиля». Содержательный смысл условия проективности предложения состоит в том, что синтаксически и семантически связанные слова близки друг к другу и по положению в предложении. Например, именная группа может быть образована только из смежных слов. Проективность не допускает разрыва именной группы [63].

При решении задач семантического анализа слабоструктурированных и неструктурированных текстов, интерес представляет не только выделение коллокаций, но и поиск синонимичных коллокаций, обозначающих близкие по смыслу понятия.

В последнее время число исследований, связанных с семантическим подобием различных по уровню текстовых элементов (слов, словосочетаний, коллокаций, коротких текстовых фрагментов различной длины) постоянно увеличивается. Это связано, прежде всего, с расширением границ использования семантически близких фрагментов текста в различных NLP-приложениях. Второй причиной роста интереса к идентификации семантически сходных элементов в текстах является ежедневная публикация в социальных сетях миллиардов небольших текстовых сообщений, каждое из которых состоит из 30-40 слов, в то время как традиционные популярные алгоритмы, такие, как, например, Tf-Idf не работают на текстах столь малого размера [64]. Для текстов такой длины часто необходимы новые алгоритмы, отличающиеся от статистических.

В тоже время существует достаточное количество методов поиска близких по смыслу слов, но нет достаточно надежных алгоритмов по определению семантически близких предложений или словосочетаний (коллокаций), и это связано, прежде всего, со сложностью формализации значения короткого текстового фрагмента.

Для определения смысловой близости коротких фрагментов текста сегодня применяются такие подходы как:

– использование двуязычного корпуса [65],

– выравнивание двух фрагментов предложений для извлечения небольших фраз с одинаковым значением [66];

– использование Machine Translation (MT) для получения нескольких переводов одной и той же фразы [67],

– использование латентного семантического анализа (LSA) [68] и другие.

Однако, данные подходы не являются универсальными для любых языков и предметных областей и пока не позволяют получить достаточно высокие показатели точности и полноты поиска семантически близких коллокаций в тексте.

В нашем исследовании для поиска семантически близких коллокаций с последующим выделением их криминальной значимости мы используем логико-лингвистические уравнения. Уравнения представляют собой конъюнкцию морфологических и семантических характеристик слов, составляющих коллокации [69]. Для того чтобы правильно идентифицировать грамматические характеристики слов предложения используются Stanford POS-tagger и Stanford Universal Dependencies (UD) парсер [70]. Дополнительно, для того чтобы найти синонимы для слов, входящих в найденные коллокации, используется библиотека обработки синсетов WordNet [71].

На рисунке 2.2 показана структурная схема технологии поиска семантически близких коллокаций, включающая несколько шагов. На первом этапе для того, чтобы правильно разметить обрабатываемые тексты, применяется POS-tagging и UD парсер. Основной причиной использования парсера UD является то, что его древовидные структуры централизованно организованы вокруг понятий субъекта, объекта, клаузального дополнения, определителя существительного, модификатора существительного и т. д. [72]. Поэтому синтаксические отношения, соединяющие слова предложения друг с другом, определяемые UD парсером, могут выражать семантический контент, необходимый для получения семантических характеристик коллокантов.

Для выделения направленных отношений между двумя существительными, глаголом и существительным, и существительным и прилагательным мы используем шесть типов синтаксически меток парсера UD: *compound*, *nmod*, *nmod*:*possobj*, obj (*dobj*), *amod* и *nsubj*.

На следующем этапе для формализации семантически сходных фрагментов текста посредством конъюнкции грамматических и семантических характеристик коллокантов используется разработанная логико-лингвистическая мо-

дель [69]. Семантико-грамматические характеристики определяют роль слов в субстантивных, атрибутивных и вербальных коллокациях.

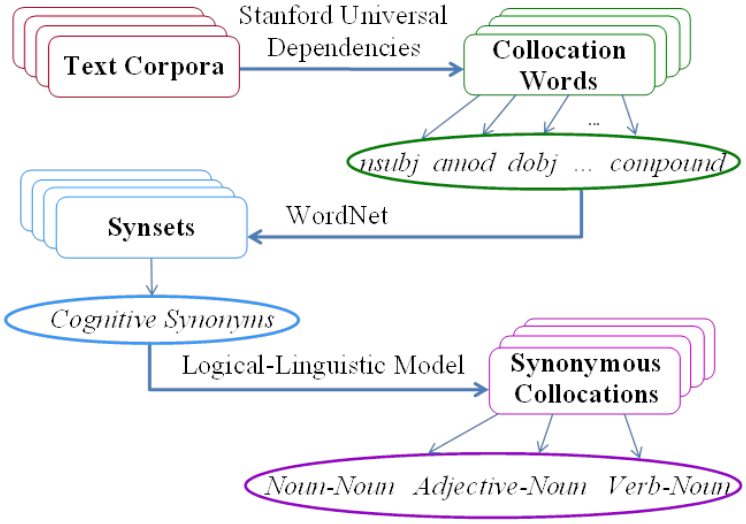


Рисунок 2.2 – Структурная схема технологии поиска семантически близких коллокаций

В модели множество грамматических и семантических характеристик слов коллокаций определяется двумя предметными переменными a^i и c^i. Во всех трех типах коллокаций возможные грамматические и семантические характеристики для главного слова коллокации определяются через предикат *P(x)*, а возможные грамматические и семантические характеристики зависимого слова коллокации определяется предикатом *P(y)*.

Двухместный предикат *P(x,y)* описывает бинарное отношение, определенное на декартовом произведении *P(x)• P(y)* и определяет корреляцию семантической и грамматической информации первого *x* и второго *y* слов коллокации:

$$P(x,y) = (x^{NSubAg} \vee x^{NSubOfAg} \vee x^{VTr})(y^{NObjAtt} \vee \vee y^{NObjPac} \vee y^{AAtt} \vee y^{A\Pr}) \quad (2.4)$$

Используя данное уравнение, определяем предикат семантической эквивалентности между двумя двух словными коллокациями как:

$$P(x_1,y_1) \times P(x_2,y_2) = \gamma_i(x_1,y_1,x_2,y_2) * * P(x_1,y_1) * P(x_2,y_2) \quad (2.5)$$

где

× обозначает семантическое сходство двух коллокаций,

* – декартовое произведение,

а предикат γ_i исключает коллокации, между которыми семантическая эквивалентность не может быть идентифицирована. Значения предикатов для трех основных типов коллокаций показаны в таблице 2.1

Таблица 2.1 – Предикаты семантической близости субстантивных, атрибутивных и вербальных коллокаций

Тип коллокации	Вид предиката γ_i	Пример семантически близких словосочетаний
Атрибутивный (Adjective-Noun)	$\gamma_1(x_1, y_1, x_2, y_2) = y_1^{AAtt}x_1^{NSubAg} \wedge$ $\wedge x_2^{NSubAg}y_2^{A\Pr} \vee y_1^{AAtt}x_1^{NSubAg} \wedge$ $\wedge y_2^{AAtt}x_2^{NSubAg} \vee$ $\vee x_1^{NSubAg}y_1^{A\Pr}x_2^{NSubAg}y_2^{A\Pr}$	guaranteed outcome ~ assured result
Субстантивный (Noun-Noun)	$\gamma_2(x_1, y_1, x_2, y_2) = x_1^{NSubOfAg}y_1^{NObjAtt} \wedge$ $\wedge y_2^{NObjAtt}x_2^{NSubAg} \vee$ $x_1^{NSubOfAg}y_1^{NObjAtt}x_2^{NSubOfAg}y_2^{NObjAtt}$ $\vee y_1^{NObjAtt}x_1^{NSubAg}y_2^{NObjAtt}x_2^{NSubAg}$	access control ~ admission monitoring
Вербальный (Verb-Noun)	$\gamma_3(x_1, y_1, x_2, y_2) = x_1^{VTr}y_1^{NObjPac} \wedge$ $\wedge x_2^{VTr}y_2^{NObjPac}$	receive commands ~ obtain instructions

На следующем этапе, для того, чтобы получить синонимы слов, входящих в заданые типы коллокаций, используется WordNet. Для каждого типа коллокации (субстантивного, атрибутивного и вербального) осуществляется поиск WordNet сенсета. Если синонимичное слово найдено, соответствие грамматических и семантических характеристик коллокаций для потенциальной синонимичной комбинации слов проверяется с помощью разработанных логико-лингвистических равенств. В таблице 2.2 приведены примеры идентифицированных синонимичных коллокаций.

Таблица 2.2 – Примеры найденных в корпусе английских текстов синонимичных коллокаций

Коллокации	Tags синтаксических	Синонимичные коллокации	Tags синтаксич	Типы коллокаций

	связей		еских отношений	й
history of land	nmod:of	nation’s story	nmod:poss	substantive
soul power	compound	ability of person	nmod:of	substantive
spectacular progression	amod	outstanding advance	amod	attributive
restoration is incompetent	nsubj:cop	restitution is incapable	nsubj:cop	attributive
qualify place	dobj	modify position	dobj	verbal
preserve fire	dobj	maintain flame	dobj	verbal

3. ЛОГИКО-ЛИНГВИСТИЧЕСКАЯ МОДЕЛЬ ИЗВЛЕЧЕНИЯ ФАКТОВ ИЗ ТЕКСТОВЫХ МАССИВОВ

3.1. Базовые математические средства модели извлечения фактов из неструктурированных текстов

Знания о некоторой предметной области представляют собой совокупность сведений об объектах/субъектах данной ПО, их существенных свойствах, связывающих отношениях и фактах, описывающих действия или свойства данных объектов/субъектов. То есть, запись фактографической информации должна включать указатель на агента действия, на атрибут или предикат этого объекта, и давать конкретное значение этого атрибута. Такая запись позволяет извлекать понятия из слабоструктурированных текстовых источников информации и представлять отношения между ними в структурированном виде. Получаемая структура представляет собой факты, как в виде достаточно простых понятий: ключевых слов, персоналий, организаций, географических названий, — так и в более сложном виде, например, имя персоналии с ее должностью и родом деятельности.

В криминально окрашенных текстах информация о компонентных элементах состава преступления может быть представлена в виде слабоструктурированных фактов, которые семантически объединяют партиципанты предметной области и их отношения в триаду *Субъект* → *Атрибут* → *Значение* (или *Субъект* → *Отношение*→*Объект*).

Базируясь на имеющихся грамматических типах казахских, русских и английских предложений, мы выделяем четыре типа структурированного факта (Рисунок 3.1). Первый выделяемый нами тип факта *subj-fact*, выражается простейшим грамматическим предложением, включающим действие, называемое глаголом, и *Субъект* действия, назвываемый существительным.

Второй аналогичный тип факта *obj-fact*, так же выражается простейшей наименьшей грамматической формой предложения, включающей глагол и существительное. Существительное в данном типе факта определяет *Объект* действия, то есть, предмет или персоналию, на которые направлено действие.

Третий выделяемый нами тип факта – *subj-obj fact*, выражается простым предложением, включающим действие (глагол) и два существительных (*Субъект* и *Объект* действия).

И четвертый тип факта – *complex fact*, выражается простым предложением, состоящим из глагола, называющего действие, и нескольких существительных (или личных местоимений). При этом, одно из существительных называет *Субъект* действия, второе существительное называет *Объект* действия, остальные существительные определяют атрибуты названного действия. Это

могут быть атрибуты времени, места, инструмента, длительности действия и так далее.

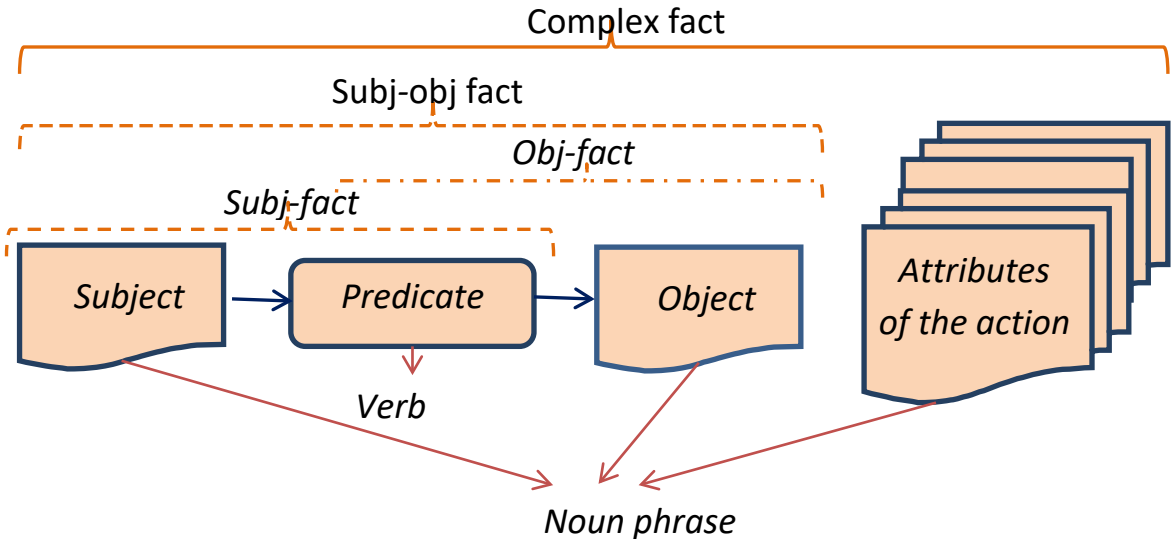


Рисунок 3.1 – Структурная схема формализации четырех семантических типов фактов в неструктурированном тексте.

В общем случае выделение фактов из слабоструктурированной текстовой информации включает следующие этапы:

1) *Entity Extraction* или *Named Entity Recognition* — извлечение слов или словосочетаний, важных для описания смысла текста (списки терминов предметной области, персоналий, организаций, географических названий и т.д);

2) *Feature Association Extraction* – исследование связей между извлеченными понятиями;

3) *Event and Fact Extraction* – извлечение сущностей, распознавание фактов и действий.

При этом центральной и, к настоящему моменту не до конца решенной задачей получения фактографической информации, остается второй этап обработки, представляющий извлечение отношений между сущностями. Для идентификаций таких смысловых отношений предлагается использовать грамматику семантических падежей. Для чего разработана строгая модель, связывающая информацию, содержащуюся в определении семантических ролей, с элементами поверхностной структуры предложений естественного языка. Такой подход рассматривается в рамках падежной грамматики и основывается на понятии глубинных падежей, введенных Ч. Филлмором [73]. Филмор выделял пропозицию (или основной смысл предложения), включающую предикат, выражаемый в поверхностной структуре чаще глаголом, и партиципантов (или участников данного действия), выражаемых чаще существительными или именными груп-

пами, которые связываются с предикатом с помощью определенных глубинных падежей.

Поскольку слабоструктурированный факт обычно выражается различными нерегламентированными конструкциями естественного языка, то для его идентификации необходимо извлечь из предложения некоторый предикат, выражаемый определенными глаголами, и определить партиципантов отношения или действия, называемого данным предикатом.

В предлагаемой модели для задания смысловых связей предлагается использовать семантические функции, явным образом определяющие отношения морфологических и семантических категорий партиципантов предложения. Такие отношения морфологических и семантических признаков участников действия могут быть описаны средствами алгебры конечных предикатов (АКП) [74].

В роли базисных элементов АКП используются предикаты 0 и 1, а также предикаты x_i^a узнавания предмета a по переменной x_i, $i = \overline{1, m}$, $a \in A_i$, где

$$x_i^a = \begin{cases} 1, если\ x_i = a, \\ 0, если\ x_i \neq a. \end{cases} \tag{3.1}$$

В роли базисных операций в дизъюнктивно-конъюнктивной алгебре предикатов используются дизъюнкция и конъюнкция предикатов. Любой предикат $P(x_1, x_2, ..., x_m)$ в этой алгебре можно записать формулой в виде его совершенной дизъюнктивной нормальной формы (СДНФ):

$$P(x_1, x_2, ..., x_m) = \bigvee_{(a_1, a_2, ..., a_m) \in P} x_1^{a_1} x_2^{a_2} \dots x_m^{a_m}. \tag{3.2}$$

Используя АКП в качестве базового математического аппарата, вводим универсум элементов U, отображающий специфику данной предметной области. В исследуемой ПО слабоструктурированных и неструктурированных текстов универсум U включает все возможные характеристики объектов языковой системы: лексемы, коллокации, грамматические, семантические характеристики слов, синтаксические характеристики словосочетаний и предложений и т.д.

Из элементов универсума образуется конечное подмножество грамматических и семантических характеристик партиципантив предложения $M = \{m_1, \dots, m_n\}$, где n – количество указанных характеристик. Отношение между характеристиками может быть представлено в виде $m_i \bullet m_j \bullet \dots \bullet m_k$, где $m_i, m_j, \dots, m_k \in M$, а знак • – обозначает, что данные характеристики соответствуют существительному, который выполняет определенную семантическую функцию.

Множество всех n-арных предикатов, заданных на универсуме U, на котором определены операции дизъюнкции, конъюнкции и отрицания, называется алгеброй n-арных предикатов на U. При этом, операции дизъюнкции, конъюнкции и отрицания являются базисными для алгебры предикатов. Алгебра предикатов при любом значении n является разновидностью булевой алгебры; в ней выполняются все основные тождества булевой алгебры [75].

На множестве M вводится система предикатов S так, чтобы любой предикат $P(q_{\text{m}}) \in S$ был равен 1 на множестве существительных с грамматико-семантической информацией, соответствующей определенной семантической функции, и был равен 0 в противном случае. Предикатом P, заданным на U, называется любая функция $\varepsilon = P(x_1, x_2, \ldots, x_n)$, отображающая множество U в множество $\sum = \{0,1\}$. Переменные $x_1, x_2, \ldots, x_n$, называются предметными переменными, а их значения предметами (3.1).

N-мерный предикат $P(x_1,\ldots,x_n)$ определяет семантическую роль участника действия через предметные переменные, называющие грамматические и семантические характеристики слова предложения:

$$P(x_1,\ldots,x_n) \to P(x_1) \wedge \ldots \wedge P(x_n) \tag{3.3}$$

Предикат $P(x_1,\ldots,x_n) = 1,$ если анализируемое слово, выполняющее некоторую семантическую функцию, обладает определенными морфологическими и семантическими характеристиками языка. Очевидно, что описанные уравнением отношения грамматических характеристик не зависят от конкретного слова.

На практике подмножество согласованных морфологических, синтаксических и семантических признаков участников действия не совпадает с декартовым произведением совокупности всех признаков. Исходя из этого, мы можем определить предикат на декартовом произведении $S \times \mathrm{S}$:

$$P(x_1,\ldots,x_n) = \gamma_k(x_1,\ldots,x_n) \times P_1(x_1) \times \ldots. \times P_n(x_n), \tag{3.4}$$

где $k \in [1,h]$, здесь h — число рассматриваемых в модели участников и атрибутов действия. Предикат $\gamma_k(x_1,\ldots,x_n)$=1, если конъюнкция грамматических характеристик слов предложения показывает некую семантическую роль участников (*Субъект*, *Объект*) или атрибутов действия; и $\gamma_k(x_1,\ldots,x_n) = 0$ в противном случае. Таким образом, если отношения между грамматическими характеристиками слов предложения не выражают любой составляющий элемента факта, они исключаются из формулы (3.4) предикатом $\gamma_k(x_1,\ldots,x_n)$.

Таким образом, семантические функции участников и атрибутов действия явным образом выражаются отношением грамматических характеристик поверхностной структуры естественных языков. Однако, в связи с существующими различиями в грамматике, а иногда и семантике, возможны особенности реализации модели для каждого конкретного языка [76].

В связи с тем, что в разных естественных языках глубинные семантические отношения выражаются различными поверхностными характеристиками и структурами, очевидно, что данную логико-лингвистическую модель необходимо отдельно реализовывать для разных естественных языков. Количество и состав семантических ролей и, следовательно, предметных переменных, выделяемых при описании языка, в каждой имплементации модели могут существенно различаться в зависимости от задач описания, языка и его степени детализации.

Мы рассматриваем имплементацию нашей логико-лингвистической модели Open IE для английского [77], русского [78] и казахского [79] языков.

3.2. Логико-лингвистическая модель извлечения фактов из слабоструктурированных текстов русского языка

Как для казахского, так и для русского языков, семантические роли или функции партиципантов предложения определяются, в своем большинстве, грамматическими падежами. Для формального определения семантических падежей русского языка выделим достаточно четко сформированное множество семантико-грамматических признаков, используя несократимый набор трех переменных:

– X – признак одушевленности (со значениями x^o — предметная переменная, характеризующая семантический признак живого, $x^н$ — предметная переменная характеризующая семантический признак неживого;);

– Y – элемент семантического значения существительного ($y^м$ — механизм, $y^с$ — имя собственное, $y^и$ — инструмент, $y^ч$ — часть тела, $y^т$ — плоскость/точка, $y^о$ — объемное пространство, $y^в$ — определенное время, $y^п$ — период, $y^ц$ — пункт назначения);

– Z — грамматический падеж существительного ($z^и$, $z^р$, $z^д$, $z^в$, $z^т$, $z^п$— предметные переменные, описывающие свойства существительных иметь тот или иной грамматический падеж).

Область изменения введенных переменных формально задается следующим образом:

$$\begin{gathered} x^{о} \vee x^{н} = 1, \\ z^{и} \vee z^{р} \vee z^{д} \vee z^{в} \vee z^{т} \vee z^{п} = 1, \\ y^{м} \vee y^{с} \vee y^{и} \vee y^{ч} \vee y^{т} \vee y^{о} \vee y^{в} \vee y^{п} \vee y^{ц} = 1, \end{gathered} \tag{3.5}$$

Семантическая функция существительного — партиципанта предложения — описывается предикатом $P(x, y, z) = 1$, связывающим элементы семантического значения существительного x и y с его грамматическими значениями z [80]. Тогда, используя конъюнкцию предикатов, можно записать:

$$P(x, y, z) \rightarrow P(x) \bullet P(y) \bullet P(z), \tag{3.6}$$

где • — операция конъюнкции.

Так как возможность согласования грамматической и семантической информации не зависит от того, к какой конкретно словоформе она относится, на декартовом квадрате множества $S * S$ можно задать предикат $\gamma(x_n, y_n, z_n)$, принимающий значение *1*, если морфосемантическая информация словоформы n формирует некоторый семантический падеж лексемы, и значение *0* в противном случае.

Таким образом, отношения морфосемантических признаков существительных предложения, выражающих семантические падежи, требуемые валентностью глагола, можно задать формулой:

$$P(x_n) * P(y_n) * P(z_n) = \gamma_k(x_n, y_n, z_n) \bullet P(x_n) \bullet P(y_n) \bullet P(z_n). \tag{3.7}$$

Практически никогда подмножество согласующейся морфосемантической информации, выражающей семантические падежи, не совпадает с декартовым произведением на множестве морфологических и семантических признаков. Те морфосемантические признаки, которые в своем согласовании не формируют семантический падеж существительного, должны исключаться из формулы (3.7) множителями $\gamma_k(x_n, y_n, z_n)$, $k \in [1;m]$, где m — количество принятых к рассмотрению в системе семантических падежей. Предикат γ_k принимает значение 1, если морфосемантическая информация словоформы n формирует некоторую семантическую функцию лексемы, и значение 0 в противном случае.

Семантическая функция *агенс*, представляющая *Субъект* действия, обычно определяет инициатора действия, лицо или предмет, имеющие потенцию на осуществление действий, и выражается предикатом:

$$\gamma_A(x_n, y_n, z_n) = x_n^{о} z_n^{и} \vee z^{и} x_n^{н} y_n^{м} \vee z^{и} x_n^{о} y_n^{с}. \tag{3.8}$$

Семантическая функция *объектив*, определяющая *Объект*, над которым непосредственно осуществляется действие, выражается предикатом:

$$\gamma_{О}(x_n, y_n, z_n) = z^{в}\, x_n{}^{н} \vee z^{в}\, x_n{}^{о}. \quad (3.9)$$

Семантическая функция *инструменталис*, определяющая непосредственную причину действия, играющую определенную роль в совершении процесса, выражается предикатом:

$$\gamma_{И}(x_n, y_n, z_n) = z_n{}^{т}\, x_n{}^{н}\, y_n{}^{и} \vee z_n{}^{т}\, x_n{}^{н} y_n{}^{ч}. \quad (3.10)$$

Семантическая функцию *локатив*, выражающая характеристики месторасположения, пространственной ориентации действия или состояния, выражается предикатом:

$$\gamma_{Л}(x_n, y_n, z_n) = z^{п}\, x_n{}^{н}\, y_n{}^{т} \vee z^{п}\, x_n{}^{н} y_n{}^{м} \vee z^{п}\, x_n{}^{н} y_n{}^{ц} \vee z^{п}\, x_n{}^{н} y_n{}^{о}. \quad (3.11)$$

Множество возможных связей грамматической и семантической информации существительного семантического падежа *темпоралис* задается предикатом $\gamma_{Т}(x_n, y_n, z_n)$:

$$\gamma_{Т}(x_n, y_n, z_n) = z^{в}\, x_n{}^{н}\, y_n{}^{в} \vee z^{п}\, x_n{}^{н}\, y_n{}^{п}. \quad (3.12)$$

Применении данной модели позволяет следователю (или иному процессуально-должностному лицу) извлечь факты конкретного уголовного дела из огромных информационных потоков полнотекстовой информации, обрабатываемых в процессе оперативно-служебной деятельности (сводки, объяснительные/служебные записки, отчеты, газетные и интернет публикации, словесные портреты фигурантов и т. п.). В подавляющем большинстве случаев к таким фактам относятся: сведения о фигурантах, сведения об объекте посягательства, сведения о механизме и способе совершения преступления.

Таким образом, при извлечении из неструктурированной текстовой информации фактов даты, места Субъекта, и Объекта некоторого противоправного действия используются следующие семантические функции, выраженные выше приведенными предикатами:

агенс – семантическая функция, представляющая *Субъект* действия, обычно выступающего инициатором действия (в нашем случае: лицо/субъект противоправного действия);

объектив – функция, определяющая сферы и продукты деятельности человека (в данном случае: сведения об *Объекте* посягательства);

темпоралис – временная характеристика события, позволяющая определить дату (в нашем случае: дату рождения человека, или некоторого противоправного деяния);

локатив – функция, характеризующая местонахождение, положение или состояние *Объекта* или *Субъекта*, определяет место (в нашем примере, место рождения человека или некоторого преступного события).

На рисунке 3.2 показана структурная схем идентификации факта, связанного с криминальным событием.

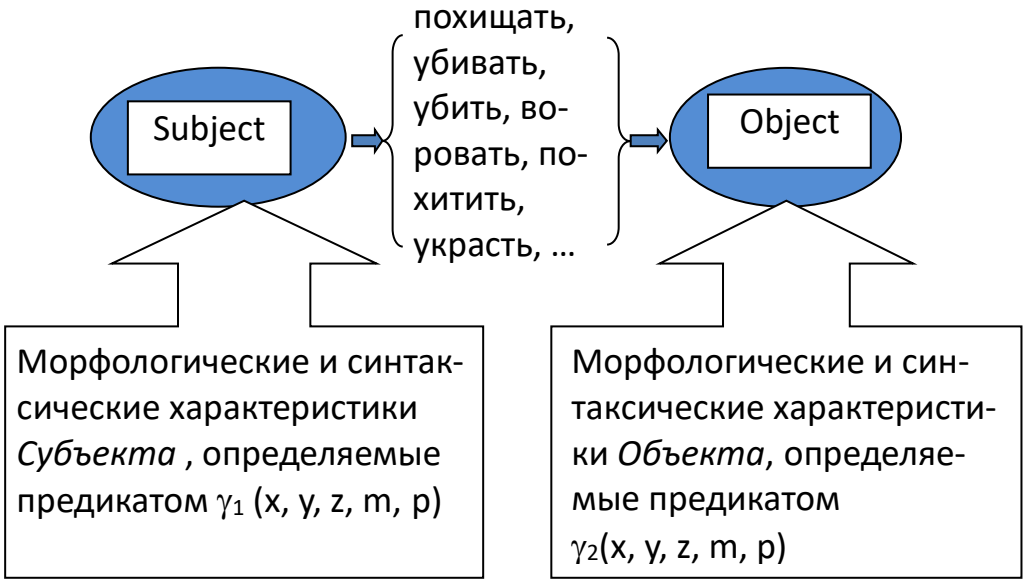


Рисунок 3.2 – Структурная схема идентификация криминально-значимого факта

В таблице 3.1 приведены семантические функции, соответствующие фактам преступного действия и идентификации личности; и определены соответствующие им предикаты (формулы (3.8 – 3.12)), описывающие отношения морфологических и семантических категорий существительных партиципантов данных фактов.

Таблица 3.1 – Формальная структура фактов преступного действия и идентификации личности

Действие, определяемое глаголом	Основные семантические функции	Предикаты, реализующие семантические функции
родиться, похищать, похитить, убивать, убить, красть, выкрасть, украсть, обмануть, обманывать, грабить, ограбить и др.	Темпоралис (*whenacted*)	$\gamma_{Т}\,(x_n, y_n, z_n)$ – формула 3.12
	Локатив (*whereacted*)	$\gamma_{Л}(x_n, y_n, z_n)$ – формула 3.11
	Объектив (*toactsmth*)	$\gamma_{О}(x_n, y_n, z_n)$ – формула 3.9
	Агенс (*tobeactedbysmth*)	$\gamma_{А}\,(x_n, y_n, z_n)$ – формула 3.8
	Инструменталис (*bysmth*)	$\gamma_{И}(x_n, y_n, z_n)$ – формула 3.10

3.3. Информационная технология извлечения фактов из слабоструктурированных английских текстов

Для формализации семантических функций предложений английского языка и их явного представления средствами поверхностной структуры, были выделены и описаны следующие синтаксические и морфологические категории:

$$
\begin{gathered}
z^{to} \vee z^{by} \vee z^{with} \vee z^{about} \vee z^{of} \vee z^{on} \vee z^{at} \vee z^{in} \vee z^{out} = 1, \\
y^{ap} \vee y^{aps} \vee y^{out} = 1, \\
x^{f} \vee x^{l} \vee x^{kos} = 1, \\
m^{is} \vee m^{are} \vee m^{havb} \vee m^{hasb} \vee m^{hadb} \vee m^{was} \vee m^{were} \vee m^{out} = 1, \\
p^{III} \vee p^{ed} \vee p^{I} \vee p^{ing} \vee p^{II} = 1, \\
f^{can} \vee f^{may} \vee f^{must} \vee f^{should} \vee f^{could} \vee f^{need} \vee f^{might} \vee f^{would} \vee f^{out} = 1, \\
n^{not} \vee n^{out} = 1,
\end{gathered}
\tag{3.13}
$$

где использованы предметные переменные, характеризующие следующие категории:

– предметная переменная *z* характеризует наличие предлога *to*, *by*, *with*, *about*, *of* , *on*, *at*, *in* после предиката триплета, или его отсутствие – *out*;

– предметная переменная *y* характеризует наличие или отсутствие апострофа в конце слова, определяющего притяжательный падеж у *Субъекта* триплета – *ap*, *aps*, *out*;

– предметная переменная *x* характеризует расположение существительного, определяющего сущность: перед глаголом в личной форме– *f*, после глагола в личной форме – *l* или после косвенного дополнения – *kos*;

– предметная переменная *m* характеризует наличие любой формы глагол *to be — is*, *are*, *havb*, *hasb*, *hadb*, *was*, *were* или его отсутствие *out*;

– предметная переменная *f* характеризует наличие в простом предложении модального глагола – *can*, *may*, *must*, *should*, *could* , *need*, *might* , *would* или его отсутствие – *out*;

– предметная переменная *n* характеризует наличие – *not* или отсутствие– *out* в предложении отрицания;

– предметная переменная *p* характеризует форму основного глагола предложения: первая, вторая/третья и четвертая форма правильного глагола – *I*, *II*, *III*, *ing*; и третья форма неправильного основного глагола – *ed*.

Семантические связи партиципантов действия в простом предложении английского языка определяются через предикаты P_k, связывающие категории наличия предлога после предиката; существование апострофа, определяющего

притяжательный падеж; расположение существительного в предложении; наличие отрицания; наличие модального глагола; а так же наличие глагола *to be* и формы основного глагола:

$$P(x, y, z, m, p, n, f) \rightarrow P(x) \wedge P(y) \wedge P(z) \wedge P(m) \wedge P(p) \wedge P(n) \wedge P(f) \qquad (3.14)$$

Можно записать предикаты $P_k(x, y, z, m, p, n, f)$, явным образом определяющие отношения предметных переменных x, y, z, m, p, n и f для каждой семантической функции:

$$\begin{gathered} P_{\mathrm{k}}(x, y, z, m, p, n, f) = \gamma_{\mathrm{k}}(x, y, z, m, p, n, f) \wedge \\ \wedge P(x) \wedge P(y) \wedge P(z) \wedge P(m) \wedge P(p) \wedge P(n) \wedge P(f), \end{gathered} \qquad (3.15)$$

где предикат $\gamma_{\mathrm{k}}(x, y, z, m, p, n, f)$ принимает значение 1 или 0.

Практически никогда подмножество согласующихся грамматических и семантических категорий слова, являющегося элементом факта, не совпадает с декартовым произведением на множестве признаков. Грамматические категории, которые в своей конъюнкции не формируют семантические связи понятий триплета, и, соответственно, семантические падежи патисипанта некоторого факта, исключаются из формулы (3.15) множителем $\gamma_{\mathrm{k}}(x, y, z, m, p, n, f)$, $k \in [1;h]$, где h — количество, принятых к рассмотрению в системе типов сематических падежей или семантических функций участников действия.

Согласно полученной модели извлечения фактов из английских предложений; семантическое отношение, определяющее *Субъект* действия в таких типах фактов как *subj-fact*, *subj-obj fact* и *complex fact*, может быть явно определено через следующее логико-лингвистическое уравнение:

$$\begin{gathered} \gamma_1(z, y, x, m, p, f, n) = y^{\mathrm{out}} ((f^{\mathrm{can}} \vee f^{\mathrm{may}} \vee f^{\mathrm{must}} \vee f^{\mathrm{should}} \vee fc^{\mathrm{ould}} \vee f^{\mathrm{need}} \vee \\ \vee f^{\mathrm{might}} \vee f^{\mathrm{would}} \vee f^{\mathrm{out}}) (n^{\mathrm{not}} \vee n^{\mathrm{out}}) (p^{\mathrm{I}} \vee p^{\mathrm{ed}} \vee p^{\mathrm{III}}) x^{\mathrm{f}} m^{\mathrm{out}} \vee \\ (x^{\mathrm{l}} (m^{\mathrm{is}} \vee m^{\mathrm{are}} \vee m^{\mathrm{havb}} m^{\mathrm{hasb}} \vee m^{\mathrm{hadb}} \vee m^{\mathrm{was}} \vee m^{\mathrm{were}} \vee m^{\mathrm{be}} \vee m^{\mathrm{out}}) z^{\mathrm{by}}) \end{gathered} \qquad (3.16)$$

Объект действия является вторым наиболее важным аргументом глагола (действия) после *Субъекта* действия. Мы определяем грамматические характеристики *Объекта* действия в *obj -fact*, *subj-obj fact* и *complex fact* фактах английских предложений следующим логико-лингвистическим уравнением:

$$\begin{gathered} \gamma_2(z, y, x, m, p, f, n) = y^{\mathrm{out}} (n^{\mathrm{not}} \vee n^{\mathrm{out}}) (f^{\mathrm{can}} \vee f^{\mathrm{may}} \vee f^{\mathrm{must}} \vee f^{\mathrm{should}} \vee f^{\mathrm{could}} \vee f^{\mathrm{need}} \vee \\ \vee f^{\mathrm{might}} \vee f^{\mathrm{would}} \vee f^{\mathrm{out}}) (z^{\mathrm{out}} x^{\mathrm{l}} m^{\mathrm{out}} (p^{\mathrm{I}} \vee p^{\mathrm{ed}} \vee p^{\mathrm{III}}) \vee x^{\mathrm{f}} (z^{\mathrm{out}} \vee z^{\mathrm{by}}) (m^{\mathrm{is}} \vee m^{\mathrm{are}} \vee \\ \vee m^{\mathrm{havb}} \vee m^{\mathrm{hasb}} \vee m^{\mathrm{hadb}} \vee m^{\mathrm{was}} \vee m^{\mathrm{were}} \vee m^{\mathrm{be}} \vee m^{\mathrm{out}}) (p^{\mathrm{ed}} \vee p^{\mathrm{III}}) \end{gathered} \qquad (3.17)$$

Аналогичным образом, с помощью логико-лингвистических уравнений определяются такие атрибуты действия как время, место, тип действия, принадлежность *Субъекта* или *Объекта* действия, инструмент действия и другие.

Например, мы можем определить семантическую функцию времени действия как дизъюнкцию следующих грамматических характеристик:

$$\gamma_3\,(x, y, z, m, p, n, f) = (z^{\text{on}}x^{\text{kos}}\,y^{\text{out}} \vee z^{\text{in}}x^{\text{kos}}\,y^{\text{out}} \vee z^{\text{at}}x^{\text{kos}})\,(p^{\text{III}} \vee p^{\text{ed}} \vee p^{\text{I}} \vee \vee p^{\text{ing}} \vee p^{\text{II}})\,(m^{\text{is}} \vee m^{\text{are}} \vee m^{\text{havb}} \vee m^{\text{hasb}} \vee m^{\text{hadb}} \vee m^{\text{was}} \vee m^{\text{were}} \vee m^{\text{out}})\,(n^{\text{not}} \vee \vee n^{\text{out}})\,(f^{\text{can}} \vee f^{\text{may}} \vee\vee f^{\text{must}} \vee \text{fs}^{\text{hould}} \vee f^{\text{could}} \vee f^{\text{need}} \vee f^{\text{might}} \vee f^{\text{would}} \vee f^{\text{out}}) \quad (3.18)$$

Для правильной идентификации грамматических и семантических категорий слов при обработке англоязычных предложений корпуса текстов мы используем как POS-tagging, так и синтаксический Parser. Выбор в качестве синтаксического парсера UD parser обосновывается его возможностью правильно анализировать синтаксические глагольные группы, подчинительные предложения и многословные словосочетания для большой группы языков. UD парсер представляет собой продолжение развития Stanford Dependencies (SD) парсера, базирующегося на грамматических отношениях, явно определенных во многих лингвистических корпусах и представляющих отношения, группирующиеся вокруг понятий субъекта, объекта, клаузального дополнения, определения, модификатора существительного и т.д. [72]. Глагол является структурным центром грамматики деревьев синтаксической зависимости, а все другие слова предложения прямо или косвенно зависят от глагола.

Синтаксические отношения, связывающие слова друг с другом в предложении и определяемые парсером, часто выражают некоторый семантический контент. Аналогично структурной схеме триплета факта (*Субъект →Предикат→Объект*) в грамматике зависимостей глагол является центральным элементом и все участники действия (партиципанты) зависят от него прямо или опосредованно. Например, на рисунке 2.3 показано графическое представление универсальных зависимостей для предложения "*The Marines reported that ten Marines and 139 insurgents died in the offensive*", полученное с использованием специального инструмента визуализации UD парсера – DependenSee [70].

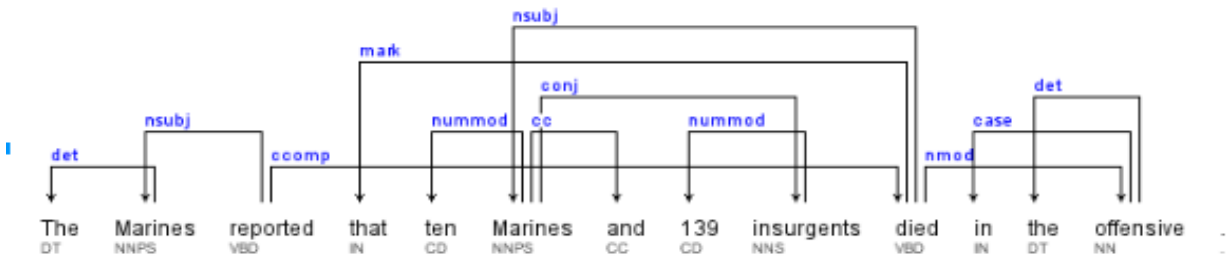


Рисунок 3.3 – Графическое представление UD парсера для предложения "*The Marines reported that ten Marines and 139 insurgents died in the offensive*", полученное с использованием DependenSee

Для нашего анализа мы используем 7 из 40 грамматических отношений между словами в английских предложениях, которые включает UD версии 1. Например, для того, чтобы определить *subj-fact* мы выделяем три типа зависимостей: *nsubj*, *nsubjpass* and *csubj*. Метка <*nsubj*> показывает синтаксическую зависимость субъекта действия от глагола, обозначенного меткой <*Root*>. Метка <*csubj*> показывает клаузуальный синтаксический *Субъект* предложения, а метка <*nsubjpass*> обозначает отношение между глаголом, обозначенным меткой <*Root*>, и *Субъектом* в английском предложении пассивного залога.

Для того, чтобы определить *obj_fact* факт мы выделяем четыре типа зависимостей UD парсера: *obj*, *iobj*, *dobj* и *ccomp*. Метка < *obj* > указывает на сущность, на которую оказывается действие, или состояние которой изменяется или перемещается. Метки <*iobj*>, <*dobj*> и <*ccomp*> используются для обозначения более точных типов зависимостей объектов действия от глаголов.

В таблице 3.2 показан пример результата работы программы автоматического выделения фактов из английских текстов [81], базирующийся на использовании разработанной логико-лингвистической модели извлечения фактов, POS-tagging и UD-parser.

Таблица 3.2 – Фрагмент результата работы программы автоматического выделения фактов из английских предложений

Предложения	Предикат, глагол	Субъект действия	Тип отношения						Root
			nsubj	iobj	Advcl	dobj	ccomp (object)	xcomp (object)	
1	consisted	War	nsubj		fighting				root
2	lasted, took	War, majority	nsubj			place			root
3	focused	insurgents	nsubj		featured, ambushing				root
3	featured	fighting	nsubj			warfare			
4	killed	Iraqis	nsubjpass					many	root
5	saw	Anbar	nsubj			fighting			root
6	occupied	it	nsubjpass						root
7	killed	Iraqis	nsubjpass						
7	began	Violence	nsubj						root
8	relinquished	Army	nsubj			command			root
10	occurred	fighting	nsubj						root
11	struggled	sides	nsubj					secure	
11	secure					Valley			
11	escalated	Violence	nsubj		struggled				root
12	became, turned	Qaeda	nsubj			capital		group	root
13	issued	Corps	nsubj		declaring	report			root
13	lost	province	nsubjpass						
14	become	what	nsubj					Awakening	
14	form						become		

3.4. Формализация грамматических способов выражения факта побуждения к действию в английском языке

Можно выделить пять основных способов выражения одного и того же смысла в коротком фрагменте текста: (1) использования разных типов лексики одного и того же значения; (2) замена порядка слов; (3) использование разных типов грамматики; (4) замена текста определениями; (5) объединение предложений. Наиболее используемым является первый способ, который обычно реализуется через использование множества синонимов (синсетов).

Для определения идентичности фактов, передаваемых разными предложениями, используются тезаурусы или словари синонимов предметной области [82]. Для английского языка в качестве словаря синонимов широких областей знаний используется тезаурус понятийной системы английского языка WordNet. Хотя в синонимичных рядах (синсетах) тезауруса понятия связаны различными парадигматическими и синтагматическими отношениями (отношения гипо-гиперонимии, холонимо-меронимии и др.), базовым лексическим отношением WordNet является отношения синонимии, а главным логическим отношением является иерархическая подчиненность слов [83]. Причем, отношения в WordNet связывают именно понятия, а не слова.

При этом, хотя слова часто и имеют множество синонимов, однако синонимы могут довольно сильно отличаться друг от друга по значению. Следовательно, существование множества пар синонимов в предложениях может привести к изменению смысла выражаемого факта, т.е. проще говоря, к иному факту.

Наиболее полно смысл факта сохраняется при изменении порядка слов в предложении, или изменении грамматической конструкции. Изменение порядка слов в цепочке является наиболее простым способом выражения одного и тоже же смысла (факта), так как слова, включенные в предложение, остаются неизменными. Однако данный способ не очень просто применить для английского языка, в котором порядок слов в предложении жестко определен грамматикой языка.

Не смотря на кажущуюся сложность использования различных грамматик для выражения одного и того же факта, данный способ представляется более простым, чем замена словарного запаса. Кроме того, при изменении грамматики смысл факта меняется редко, тогда как, ошибки при замене словарного запаса могут привести к искажению смысла факта.

Мы рассматриваем различные грамматические конструкции представления одного и того же факта побуждения к действию с использованием лексических синонимов выражения *Предиката* и участников действия (*Субъекта* и *Объек-*

ma). Выбор в качестве исследования факта побуждения к действию связан с широкой возможностью использования предложений побуждающего действия в текстах, имеющих криминальную окрашенность.

В своем исследовании мы формализуем такие наиболее часто встречающиеся грамматические конструкции побуждения к действию в английском языке, как: повелительное наклонение, предложения с герундием, предложения с модальными глаголами, пассивный залог.

Для анализа синтаксической структуры предложения в работе использовался парсер, представляющий процесс сопоставления линейной последовательности лексем естественного языка с его формальной грамматикой. В результате формируется синтаксическое дерево разбора. Мы используем UD парсер английского языка [70], базирующийся на зависимостях, представляющих межъязыковые соответствия наиболее известных пользователю понятий и существующих стандартов разметки.

Полученные формальные схемы грамматических конструкций представляют собой регулярные выражения, использующие POS-тегинговую разметку в качестве алфавита.

Полученная формальная схема грамматической модели, использующей модальные глаголы, будет выглядеть следующим образом:

$$\text{TO-VB-[JJ*]-NN*-NN|NNS1-MD-VB-[JJ*]-NN|NNS2}, \qquad (3.19)$$

где *MD* = {should, have to, need to, must, may} модальный глагол, использующийся для выражения повелительного наклонения,
VB — основной глагол в первой форме,
NN|NNS2 — дополнение,
NN|NNS1 = {User, Customer, Operator, Worker, Employer, Manipulator, Handler, Manager} — *Субъект* действия такого предложения,
*NN** — дополнение инфинитива цели,
TO-VB — инфинитив цели.

Примеры предложений, отвечающих данной схеме, приведены в таблице 3.3

Таблица 3.3 – Примеры предложений, разбираемых по формальной модели (3.19)

TO-VB	NN*	NN|NNS[1]	MD	VB	NN|NNS[2]
To add	2 pin con tact at center of connector	user	should	update	figures and text 95mm to 25mm
to	(no) part of	user	have to	give	he written

reproduce	this material				permission of the copyright owner
to use	the power cable	user	must	switch on	power and fan module

Формальная схема грамматической модели повелительного наклонения будет представлять следующее выражение:

$$(\text{V1 obj } V_{\text{purpose}} \text{ object}_{\text{purpose}} !), \quad (3.20)$$

где *V1*— глагол в первой форме, занимающий первое место в предложении,
obj — второстепенный член предложения, представляющий собой объект, на который направлено, или с которым связано действие,
$V_{purpose}$ — глагол первой формы, обозначающий инфинитив цели,
$object_{purpose}$ — дополнение инфинитива цели.

Данной схеме будут соответствовать предложения, представленные в таблице 3.4.

Таблица 3.4 – Примеры предложений, разбираемых по формальной модели (3.20)

VV	NN|NNS[1]	TO-VV	NN|NNS[2]
Update	figures and text from 95mm to 25mm	to add	2-pin connector contact center
Give	the written permission of the copyright owner	to reproduce	(no) part of this material
Switch on	power and fan module	to use	the power cable

Кроме приведенных выше способов побуждения к действию, определенная степень побуждения в английском языке может быть выражена активной и пассивной формой герундийного предложения.

Формальная схема предложения, использующего герундий, представлена следующим образом:

$$\text{VBG-[JJ*]- NN|NNS1-VB- NN|NNS2}, \quad (3.21)$$

где *VBG*— инфинитив глагола,
NN|NNS1 — дополнение,
VB —смысловой глагол первой формы (в некоторых случаях с окончанием s),
NN|NNS2 — дополнение инфинитива цели.
Данной схеме будут соответствовать предложения, показанные в таблице 3.5.

Таблица 3.5 – Примеры предложений,
разбираемых по формальной модели (3.21)

VBG	NN|NNS[1]	VB	NN|NNS[2]
updating	figures and text 95mm to 25mm	add	2 pin contact at center of connector
giving	the written permission of the copyright owner	reproduce	(no) part of this material
the switching-on	power and fan module	use	the power cable

Формальная схема предложения, побуждающего к действию в пассивном залоге, будет выглядеть следующим образом:

[JJ]- NN|NNS1-VB-VBD|VBN-RP-VBG-[JJ*]- NN|NNS2,* (3.22)

где

NN|NNS1 — дополнение инфинитива цели,

VB-VBD|VBN-RP — грамматическая структура, включающая форму вспомогательного глагола *to Be (am, is, are, was, were, been)*, глагола в третьей неправильной форме или глагола с окончанием *–ed*, и предлога, формирующего творительный падеж *by* или *with*,

VBG-[JJ]- NN|NNS2* — герундийная грамматическая структура, которая включает динамический или производный глагол, с окончанием ing и приложение.

Данной схеме будут соответствовать предложения, представленные в таблице 6

Таблица 3.6 – Примеры предложений,
разбираемых по формальной модели (3.22)

NN|NNS[1]	VB-VBD|VBN	RP-VBG	NN|NNS[2]
2 pin contact at center of connector	is added	with updating	figures and text 95mm to 25mm
no part of this material	may be reproduced	without giving	the written permission of the copyright owner
the power cable	is used	with the switching-on	power and fan module

4. ИНФОРМАЦИОННАЯ ТЕХНОЛОГИЯ ИДЕНТИФИКАЦИИИ КРИМИНАЛЬНО-ЗНАЧИМОЙ ИНФОРМАЦИИ В ТЕКСТОВЫХ КОРПУСАХ КАЗАХСКОГО ЯЗЫКА

4.1. Анализ существующих проблем формализации казахского языка

Казахский язык относится к кыпчакской ветви тюркской группы алтайской языковой семьи. Наиболее близки к нему ногайский и каракалпакский языки. Всего на казахском языке говорят в мире около 12 млн. человек, из них в Казахстане — 8 млн. человек, 2 млн. в остальных странах СНГ, 1.5 млн. в Китае. Кроме того этот язык распространен в Монголии, Афганистане, Пакистане, Иране, Турции, Германии. С 1991 года казахский язык является государственным языком Республики Казахстан. Анализируя Казахский язык с точки зрения возможности его формализации и автоматической обработки, можно выделить следующие его основные особенности.

Прежде всего, в казахском языке четко определен строгий порядок слов: на первом месте стоит подлежащее, потом дополнение, завершает предложение сказуемое. Кроме того, четкий порядок требует располагать определение перед определяемым, а обстоятельства места и времени, обычно, перед подлежащим или перед прямым дополнением, но не после сказуемого

Кроме того, казахский язык является агглютинативным языком, где слово, обычно, состоит из основы и целого ряда морфем, следующих за ней, каждая из которых имеет определенное значение. Например: “*отырғанмын*”, “*киындықтарға*”.

Особую роль для построения модели извлечения фактической информации из текстов играет сказуемое, называющее действие, описываемое фразой или предложением. В казахском языке сказуемое может быть выражено глаголом, существительным, прилагательным, причастием, деепричастием, вспомогательными словами (“*бар*"; “*жоқ*”; “*көп*”; “ *керек*” и другими). Однако, в подавляющем большинстве случаев, сказуемое в казахском языке выражается глаголом. В связи с особой ролью глагола в представлении произошедшего факта в казахском предложении, рассмотрим возможность его формального описания более подробно.

Глаголам казахского языка присущи два типа словообразования. Глагол может быть образован: (1) синтетически (путем аффиксации), образовывая производные глагольные основы; (2) аналитически, образуя составные и сложные глаголы.

При синтетическом словообразовании казахских глаголов определяют более восьмидесяти глаголо-образующих аффиксов. Вместе с фонетическими вариантами число их составляет около двухсот, при этом залоговые аффиксы не

учитываются [84]. При синтаксическом способе словообразования глаголов, первый компонент отвечает за семантическое значение, а вспомогательный глагол полностью утрачивает свое первоначальное лексическое значение, и превращается в носителя грамматического формата.

В современно казахском языке различают две группы глаголов, полученных синтаксическим способом словообразования:

1) сложные глагольные основы, состоящие из имени, или звукоподражательного слова, плюс вспомогательный глагол;
2) сложные глагольные основы, состоящие из глагола в деепричастной форме плюс вспомогательный глагол

Совершенный вид глагола образуется посредством сочетания основного глагола в форме деепричастия, заканчивающегося на –n, используемого вместе со специальными вспомогательными глаголами: *бол*, *бітір*, *біт*, *кет*, *қой*, *жібер*, *шық*, *шығ*, *сал*, *таста*, *қал*. Несовершенный вид глагола образуется сочетанием основного глагола в форме деепричастия, заканчивающегося на –п (-ып, -іп), со специальным функционально-вспомогательными глаголами : *отыр*, *жат*, *тұр*, *жат*, *жүр*, *бер* (с деепричастием, заканчивающимся на–а [-е, -й]).

Еще одной грамматической категорией словообразования глагола казахского языка, распознаваемой по семантике, морфологическому оформлению и синтаксическим функциям, является залог. Залог формируется прибавлением к глагольным основам положительной формы аффиксов, передающих отношения между действием и субъектом или объектом. В казахском языке выделяют пять залогов, разделяемых по грамматическому оформлению, семантическому значению и синтаксическим функциям: возвратный, страдательный, совместно-взаимный, понудительный и основной или исходный. При этом, первые четыре оформляются аффиксами.

Все остальные грамматические значения (наклонения, времени, лица и т.д.) вносятся соответствующими морфологическими формантами. Например, спряжение глаголов в казахском языке, так и в других тюркских языках, оформляется при помощи аффиксов сказуемости. То есть, любая часть речи, выступающая в предложении предикатом, может принимать личные окончания, определяющие субъект действия. На рисунке 4.1 показана структура формирования глагола в личной форме.

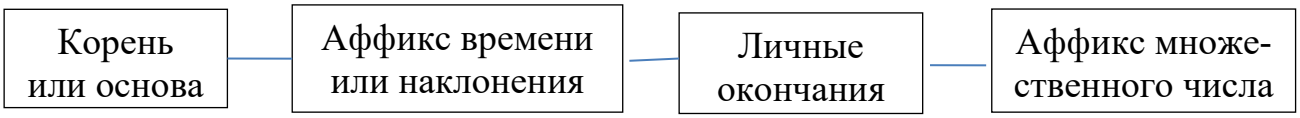
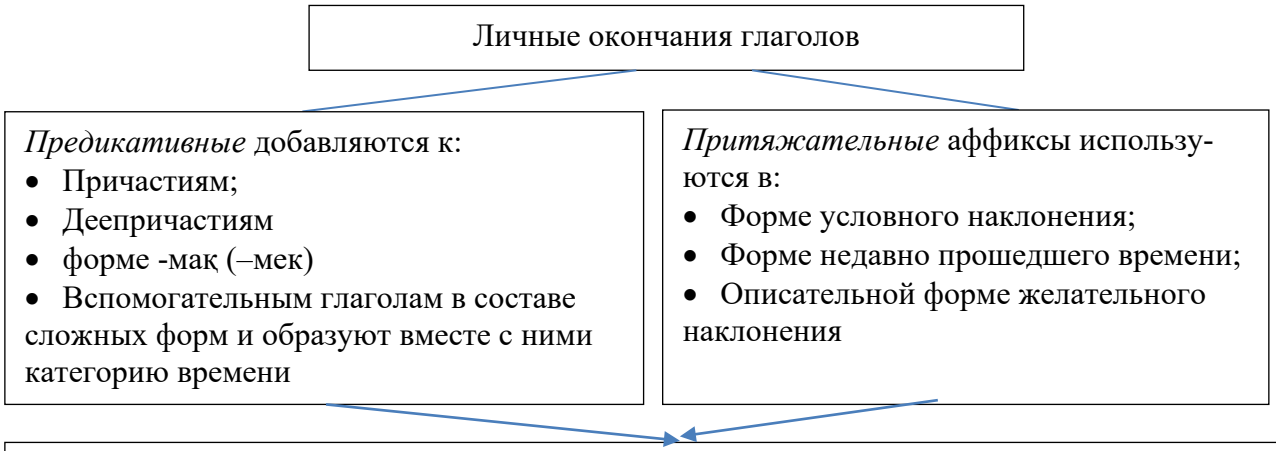


Рисунок 4.1 — Схема формирования личной формы глагола казахского языка

При этом личные окончания глаголов делятся на предикативные и притяжательные. На рисунке 4.2 показана схема формирования предикативных и притяжательных окончаний глаголов.



• Предикативные личные окончания

(не добавляются к форме недавно прошедшего (окончательного) времени; формам условного наклонения, которые принимают притяжательные аффиксы 1 и 2 лица ед и мн.числа)Окончания	Лицо, число	примеры
-сың , -сыз -сің, -сіз	2, ед.	аласын (ты берешь), аласыз (вы берете) айт-а-сың айт-а-сыз бара- сың келе-сің
-сыңдар , -сыздар, сіндер, сіздер	2, мн.	айт-а-сың-дар айт-а-сыз-дар
-мын, -мін, -пын, -пін, -бын, -бін	1, ед.	айт-а-мын быры-мын келе-мін
-мыз (міз, -пыз, -піз, -быз, біз)	1, мн.	айт-а-мыз, бара-мыз
-ды	3, ед.	айт-а-ды
-ды	3, мн.	айт-а-ды
• *Притяжательные* личные окончания		
-ң -ңыз –ңіз -сың, -сің	2, ед.	сент қайт-са-ң, сіз қайт-са-ң-ыз, ал-а-сың (берешь), бер-е-сің (даешь)
-дар, -сыңдар, -сіңдер, -сыздар,-сіздер	2, мн.	сендер қайт-са-ң-дар, сіздер қайт-са-ң-ыз-дар, жина-са-ң-дар (если соберете)
-м	1 , ед.	мен қайт-са-м, сана-са-м (когда я считаю), терле-се-м (когда я потею)
-қ, -к	1 , мн.	біз қайт-са-қ, жаз са-қ (если будем писать)

Рисунок 4.2 — Структурная схема формирования предикативных и притяжательных окончаний глаголов казахского языка

Наклонение глаголов казахского языка включает в себя такие грамматиче ские категории, как время, лицо и число. В современном казахском языке различают пять наклонений: повелительное, изъявительное, желательное, условное и отглагольно-именное. Изъявительное наклонение является наиболее употребительным наклонением. В нем представлены грамматические формы глагола, выражающие временные отношения. При этом, категория времени формально выражается морфологически и синтаксически. Морфологически категория времени формируется прибавлением к причастным и деепричастным формам личных окончаний, в свою очередь, синтаксически форма времени оформляется сочетанием глагольных имен с соответствующими вспомогательными глаголами.

При определении логико-лингвистического уравнения формального действия в казахской фразе мы основываемся на гипотезе, что факт – это реальное событие, действие, которое действительно произошло или произойдет. Исходя из этого, мы определяем изъявительное наклонение глаголов и не учитываем такие существующие в казахском языке наклонения как повелительное, желательное и условное.

Специфическая форма глагола "*тұйық рай*" (неопределенное наклонение) не является инфинитивом, а выступает как имя или название действия. Она образуется путем присоединения к основе глагола аффикса — *у*. Например, *тап-сыр-у*, *шақыр-у*.

Неопределенная форма глагола в лексическом отношении ближе к именам существительным — она не спрягается, а склоняется, принимая аффиксы принадлежности*: у-дың*, *-у-ды*, *-у-ға*, *-у-дан*, *-у-да*, *-у-мен*, *-у*, *-у-ім* , *-у-ің*, *-у-і*, *-у-іміз*, *-у-і-ңіз*, *-у-дің, -у-ге*, *-у-ді*, *-у-де*, *-у-ден* (*уы*, *уым*, *уымыз*, *у-ың*, *у-ы-ңыз*). Глаголы неопределенной формы образуют изафетную конструкцию, требуя впереди себя личное местоимение или существительное в родительном падеже.

Многие глаголы в неопределенной форме переходят в отглагольные существительные (*жаз-у* (*письмо*), *ойла-у* (*мышление*)). Далее, по словообразовательной цепочке от отглагольных имен при помощи аффикса *–шы* часто образуются производные имена существительные (*жаз-у-шы* (*писатель*), *сайла-у-шы* (*избиратель*)).

В созданной нами логико-лингвистической модели (см. подраздел 4.2) особое значение имею существительные, которые, с точки зрения семантики, выступают как *Субъект*, *Объект* и атрибуты действия. В казахском языке, по сравнению с русским, границы между частями речи несколько размыты: имена существительные могут выступать в предложении в функции определения, подлежащего, дополнения, обстоятельства и именного сказуемого.

В синтаксической связи двух имен, представляющих примыкание, первый компонент всегда выступает как определение, второй — как определяемое. Имя существительное, выступающее в качестве определения (стоит первым) может быть оформлено аффиксом родительного падежа или аффиксом принадлежности. Имена существительные изменяются по числам, падежам, лицам, а так же принимают аффиксы принадлежности. Существует два типа склонения: простое и притяжательное склонение. На рисунке 4.3 показана структурная схема простого склонения существительных

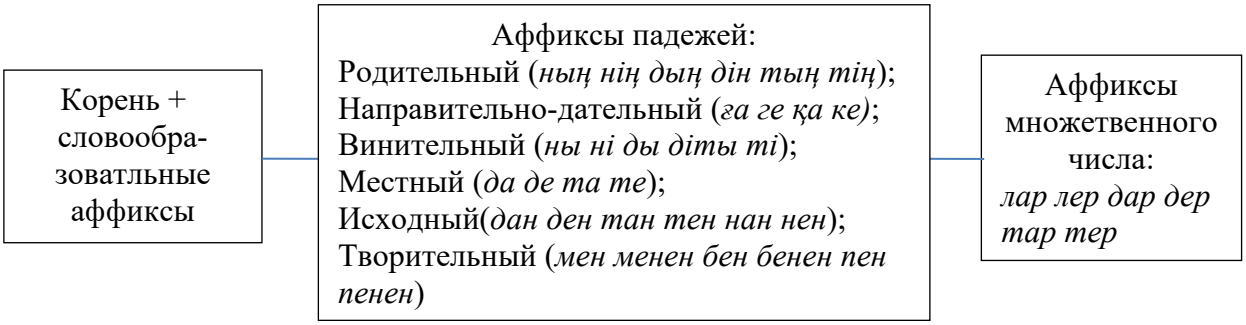


Рисунок 4.3 — Структурная схема простого склонения существительных

При притяжательном склонении существительные содержат указание на обладателя, принадлежность предмета кому/чему –нибудь, которое выражается посредством присоединения к основе слова аффиксов принадлежности. Форма принадлежности указывает одновременно и на предмет обладания, и на имя обладателя. В таблице 4.1 показаны особенности аффиксного формирования притяжательности и в казахском языке.

Таблица 4.1 – Аффиксное формирование притяжательности казахского языка

Имя обладателя стоит в единственном числе		
Лицо	аффиксы	Примеры
1	*м*, *ым*, *ім*	ана-м, калам-ым, дәптер-ім
2	*ң*, *ың*, *ің*	ана-ң, қалам-ың, дәптер-ің
3	*сы*, *ы*, *і*, *сі*	ана-сы, қалам-ы, дәптер-і
Имя обладателя и предмет обладания стоят во множественном числе		
1	*ымыз*, *іміз*	қаламдар-ымыз, дәптерлер-іміз
2	*ыңыз*, *ныз*, *іңіз*, *ңіз*	қаламдар-ыңыз, дәптерлер-іңіз
3	*ы і*	қаламдар-ы, дәптерлер-і

В таблице 4.2 показаны падежные аффиксы единственного и множественного числа притяжательного склонения казахского языка.

Таблица 4.2 – Падежные аффиксы единственного и множественного числа притяжательного склонения

Падежи	1 лицо	2 лицо	3 лицо
Родительный	*ның*, *нің*	*ның*, *нің*	*ның*, *нің*
Напр.-дательный	*а*, *е*	*а*, *е*	*на*, *не*
Винительный	*ды*, *ді*	*ды*, *ді*	*н*
Местный	*да*, *де*	*да*, *де*	*нда*, *нде*
Исходный	*нан*, *нен*	*нан*, *нен*	*нан*, *нен*
Творительный	*мен*, *менен*	*мен*, *менен*	*мен*, *менен*

4.2. Реализация логико-лингвистической модели Open IE для казахского языка

Как упоминалось ранее, казахский язык, в отличие от русского или английского, является агглютинативным языком. Это означает, что слово строится из морфем, каждая из которых имеет определенное морфологическое или семантическое значение (см. подраздел 4.1). Такое словообразование противоположно флективному языку, где каждая морфема имеет несколько неразделимых значений одновременно (например, падеж, род, число и т. д.) и аналитическому языку, в котором почти нет флексий.

При адаптации разработанной модели извлечения фактов из слабоструктурированных текстов к казахскому языку введено достаточно четко сформированное несократимое множество M из десяти грамматических и семантических признаков, влияющих на семантическую роль партиципантов казахского предложения [85, 86]. Большинство из данных признаков представляют собой морфологические или семантические характеристики, выраженные в поверхностной структуре языка с помощью аффиксов. Это такие характеристики как: положение анализируемого слова во фразе; наличие вспомогательного глагола во фразе; грамматический падеж анализируемого существительного; суффиксы множественного числа и лица; аффиксы предопределенного действия, и другие морфологические и семантические характеристики.

Наличие в модели казахского языка большого числа предметных переменных обусловлено, прежде всего, агглютинативностью языка, когда каждая грамматическая особенность выражается определенным аффиксом, а так же необходимостью выделения в казахском языке не только участников действия и их атрибутов, но и самих типов действия.

Предикат $P_{\text{x}}(x)$ определяет местоположение анализируемого слова во фразе. Выбор признака местоположения слова в предложении предопределен зна-

нием о строгом порядке слов в казахской фразе, когда на первом месте стоит подлежащее, затем дополнение и завершает предложение сказуемое, при этом определение всегда стоит перед определяемым.

$$P_x(x) = x^1 \vee x^2 \vee x^3 \vee x^{-1} \vee x^{-2} \vee x^{-3} \vee x^0, \quad (4.1)$$

где 1, 2, 3, -1, -2, -3 показывают смещение слова во фразе, «минус» обозначает начало отсчета с конца фразы; 0 показывает любую другую позицию слова, кроме первых трех и последних трех слов в предложении.

Предикат $P_f(f)$ определяет, есть ли во фразе вспомогательный глагол:

$$P_f(f) = f^{aux} \vee f^0, \quad (4.2)$$

где *aux* показывает признак существования любого глагола из списка 35 вспомогательных глаголов казахского языка в анализируемой фразе [*ал, бар, біт, бітір, бол, ғой, де, деген, дейтін, деп, е, еді, екен, емес, ер,ет, жазда, жат, жатыр, жет, жібер, жүр, кел, келеді, кет, кір, көр, қал, қой, сал, отыр, түс, тұр, шығ, шығар*].

Предикат $P_z(z)$ идентифицирует семь грамматических падежей казахского языка: именительный, родительный, дательно-направленный, винительный, местный, творительный и исходный:

$$P_z(z) = z^{Nom} \vee z^{Gen} \vee z^{Dat} \vee z^{Acc} \vee z^{Ela} \vee z^{Ins} \vee z^{Abl}, \quad (4.3)$$

где *Nom* — именительный падеж (атау септік); *Gen* — родительный падеж (ілік септік), определяемый с помощью списка падежных аффиксов [*ның*, *нің*, *дың*, *дің*, *тың*, *тің*]; *Dat* – направительно-дательный падеж (барыс септік) простого склонения, определяемый с помощью списка падежных аффиксов [*ға*, *ге*, *қа*, *ке*, *а*, *е*, *на*, *не*]; *Acc* — винительный падеж (табыс септік), определяемый с помощью списка падежных аффиксов [*ны*, *н*, *ні*, *ды*, *ді*, *ты*, *ті*]; *Loc* — местный падеж (жатыс септік), определяемый с помощью списка падежных аффиксов [*да*, *де*, *нда*, *нде*, *та*, *те*]; *Abl* — исходный падеж (шығыс септік), определяемый с помощью списка падежных аффиксов [*дан*, *ден*, *тан*, *тен*, *нан*, *нен*]; *Ins*— творительный падеж (көмектес септік), определяемый с помощью списка падежных аффиксов [*мен*, *менен*, *бен*, *бенен*, *пен*, *пенен*].

В связи с тем, что в казахском языке существует два типа склонения существительных: простое (безотносительно к обладателю) и притяжательное (с указанием обладателя), вводим предикат $P_a(a)$, определяющий два возможных типа склонения казахских существительных:

$$P_a(a) = a^{\mathrm{NSim}} \vee a^{\mathrm{NPos}}, \quad (4.4)$$

где *NSim* — признак простого склонения существительного, *а NPos* – признак притяжательного склонения существительного, определяемого наличием аффиксов [*м*, *ым*, *ім*, *ң*, *ың*, *ің*, *сы*, *ы*, *і*, *сі*, *ымыз*, *іміз*, *ыңыз*, *ныз*, *іңіз*, *ңіз*, *ы*, *і*]. Суффиксы простого склонения соответствуют суффиксам соответствующих падежей, определяемых формулой (4.3).

Предикат $P_n(n)$ идентифицирует особенности отрицания в предложении казахского языка:

$$P_n(n) = n^{me} \vee n^{\text{emes}} \vee n^{\text{joq}} \vee n^0, \tag{4.5}$$

где *me* - признак отрицательного предложения, который представлен наличием частицы из списка [*ma, me, ba, be, pa, pe*], *emes* и *joq* – признак отрицательного предложения, который представлен наличием в предложении слов «*emes*» и «*joq*», соответственно; 0 показывает отсутствие какого-либо признака отрицания в предложении.

Предикат P_c(c*)* определяет наличие или отсутствие множественных суффиксов:

$$P_c(c) = c^{\text{tar}} \vee c^{\text{ter}} \vee c^{\text{dar}} \vee c^{\text{der}} \vee c^{\text{lar}} \vee c^{\text{ler}} \vee c^0, \tag{4.6}$$

где 0 показывает, что слово употребляется в единственном числе, т.е. у слова отсутствует аффикс множественности, а значения *tar, ter, dar, der, lar, ler* показывают наличие множественных аффиксов *тар*, *тер*, *дар*, *дер*, *лар*, *лер* соответственно.

Предикат $P_y(y)$ идентифицирует признак наличия словообразовательного аффикса конкретной части речи — глагола, причастия, деепричастия и существительного:

$$P_y(y) = y^{\text{ParP}} \vee y^{\text{Vpas}} \vee y^{VaP} \vee y^{UnFu} \vee y^{FuCo} \vee y^{VAd} \vee y^{OAd} \vee y^{\text{Psuf}} \vee y^{Usuf} \vee \vee y^{Part} \vee y^{NoV} \vee y^{NoN} \vee y^{NCom} \vee y^{NDer} \vee y^{\text{y}} \vee y^0, \tag{4.7}$$

где:

– *0* определяет глагольную основу в чистом виде, в форме второго лица единственного числа будущего времени повелительного наклонения при обращении на “ты”;

– *y* определяет признак наличия аффикса инфинитивной формы;

– *Vad*, *Oad*, *VaP* – признаки деепричастия: *Vad* определяет признак деепричастия через аффиксы [*п*, *ып, ін*]; *Oad* определяет деепричастия на [*а*, *е*, *й*, *и*], *VaP* определяет деепричастия на [*ға*, *ғалы* , *ге*, *гелі*, *ғі*, *ғы*, *ке*, *келі*, *қа*, *қалы* , *кі*, *қон*, *қы*];

– *FuCo*,*UnF* признаки будущего времени изъявительного наклонения глагола: *FuCo* определяет список аффиксов глаголов будущего предположительного времени [*ар*, *ер*, *ыр*, *ір*], *UnFu* определяет аффиксы неопределенного будущего времени [*мақ*, *мек*, *пақ*, *пек*, *бақ*, *бек*, *пақшы*, *мақшы*, *мекшы*, *пекшы*, *бақшы*, *бекшы*, пақші, *мақші*, *мекші*, *пекші*, *бақші*, *бекші*];

– *Part*, *ParP* признаки словообразования причастий: *Part* определяет аффиксы словообразования причастия из списка [*ған*, *ген*, *қан*, *кен*, *қон*, *ға*, *ге*, *қа*, *ке*], *ParP* определяет аффиксы словообразования причастий из списка [*атын*, *етін*, *йтын*, *йтін*]-;

– *Vpas* определяет 20 специальных словообразовательных аффиксов глагола [*ді*, *дік*, *дік*, *дім*, *дің*, *ды*, *дык*, *дық*, *дым*, дың, қ, *ті*, *тік*, тім, *тің*, *ты*, *тык*, *тық*, *тым*, *тың*];

– *Psuf* определяет 189 продуктивных аффиксов словообразования глаголов (ПРИЛОЖЕНИЕ А), в том числе — залоговые аффиксы (ПРИЛОЖЕНИЕ Б);

— *Usuf* определяет 65 малопродуктивных аффиксов словообразования глаголов [*азы*, *ақта*, *ал*, *ала*, *аңғыра*, *аура*, *бала*, *бе*, *беле*, *би*, *бі*, *бы*, *дала*, *ди*, *ді*, *ды*, *екте*, *ел*, *еңгіре*, *еуре*, *жи*, *жіре*, *жыра*, *зы*, *і*, *ін*, *ірей*, *іс*, *ім*, *қи*, *лі*, *лы*, *ма*, *мала*, *меле*, *ми*, *мсіре*, *мсыра*, *ңра*, *ңре*, *палапеле*, *пи*, *пі*, *пы*, *ра*, *ре*, *си*, *сіре*, *сый*, *сыра*, *т*, *ти*, *ті*, *ты*, *усіре*, *усыра*, *ши*, *ші*, *шы*, *ы*, *ын*, *ыра*, *ырай*, *ыс*, *ыт*] (ПРИЛОЖЕНИЕ А);

Четыре начения *NoN, NoV, Ncom, Nder* предметной переменной у определяют признак принадлежности токена к существительному через списки конкретных аффиксов:

NoN – определяет наличие аффикса именного образования существительного [*ғай*, *гей*, *гер*, *ғи*, *ғой*, *дас*, *дес*, *дік*, *дық*, *кер*, *кес*, *қай*, *қар*, *қи*, *қой*, *қор*, *лас*, *лес*, *лік*, *лық*, *ман*, *паз*, *пана*, *сақ*, *тас*, *тес*, *тік*, *тық*, *хана*, *ша*, *шақ*, *ше*, *шек*, *ші*, *шік*, *шы*, *шық*];

NoV – определяет наличие аффикса отглагольного образования существительного [*ақ*, *ба*, *бе*, *ғақ*, *ғаш*, *гек* , *гі*, *ғіш*, *ғы*, *ғыш*, *дақ*, *дек*, *ек*, *ік*, *ім*, *іс*, *іш*, *к*, *кі*, *кіш*, *қ*, *қаш*, *қы*, *қыш*, *лақ*, *лек*, *м*, *ма*, *мақ*, *ме*, *мек*, *па*, *пақ*, *пе*, *пек*, *с*, *тақ*, *тек*, *уік*, *уық*, *ш*, *ық*, *ым*, *ыс*, *ыш*];

Ncom – определяет наличие сложного аффикса образования существительного [*герлік*, *гіштік*, *ғыштық*, *дастық*, *дестік*, *ділік*, *дылық*, *кеәтік*, *қорлық*, *ластық*, *лестік*, *лілік*, *лылық*, *паздық*, *сақтық*, *сіздік*, *сыздық*, *тастық*, *тестік*, *тілік*, *тылық*, *шақтық*, *шілдік*, *шілік*, *шылдық*, *шылық*];

Nder – определяет наличие аффикса экспрессивной оценки (уменьшительно-ласкательные и уничижительные оттенки) образования существительного [*жан*, *ке*, *қан*, *сымақ*, *тай*, *ш*, *ша*, *шақ*, *ше*, *шік*, *шық*].

Предикат P_d(d) определяет наличие признака сослагательности действия:

$$P_d(d) = d^{\text{shi}} \vee d^0, \tag{4.8}$$

где значение предметной переменной *shi* определяет наличие у анализируемого слова сослагательных суффиксов *ші* или *ши*, а значение 0 определяет признак отсутствия сослагательных суффиксов

Предикат $P_m(m)$ определяет наличие личного предикативного или притяжательного окончания глагола или отглагольных форм:

$$P_m(m) = m^{\text{PrFl}} \vee m^{\text{PoFl}} \vee m^0, \tag{4.9}$$

PrFl (personal predicative flexion) – определяет наличие личного предикативного окончания причастий, деепричастий, основных и вспомогательных глаголов [*біз*, *бін*, *быз*, *бын*, *ды*, *міз*, *мін*, *мыз*, *мын*, *піз*, *пің*, *пыз*, *пын*, *сіз*, *сіздер*, *сіндер*, *сің*, *сыз*, *сыздар*, *сың*, *сыңдар*, *ті*, *ты*];

PoFl (personal possessive flexion) – определяет признак наличия личного притяжательного окончания некоторых глагольных форм [*дар*, *йік*, *йін*, *йық*, *йын*, *іздар*, *к*, *қ*, *м*, *ндар*, *ң*, *ңдер*, *ңіз*, *ңіздер*, *ңыз*, *сіздер*, *сің*, *сіңдер*, *сыздар*, *сың*, *сыңдар*, *ыздар*];

0 – определяет отсутствие личного окончания глагола.

Предикат $P_b(b)$ определяет наличие некоторой дополнительной семантики или значения анализируемого глагола:

$$P_b(b) = b^{\text{se}} \vee b^{\text{mic}} \vee b^0, \tag{4.10}$$

где значение *mic* показывает предположительность действия, определяемую через наличие суффиксов [*мыс*, *міс*]; значение *se* показывает существование условного наклонения, определяемое суффиксами [*са*, *се*].

В таблице 4.4 представлены ранее определенные в логико-лингвистической модели Open IE казахского языка предметные переменные и их области изменения.

Полученные уравнения (4.1) - (4.10) позволяют преобразовать предикат согласованности грамматических и семантических особенностей слов, являющихся элементами факта (3.4) для казахского языка к следующему виду:

$$P() = \gamma_k \times P_x(x\ ) \times P_y(y\ ) \times P_z(z\ ) \times P_f(f\ ) \times P_m(m\ ) \times P_n(n\ ) \times P_a(a\ ) \times P_b(b\ ) \times P_c(c\ ). \times P_{\text{d}}(d). \tag{4.11}$$

Предикат инициатора действия или *Subject* факта определяется через γ_{1K}:

$$\gamma_{1K} = (x^1 \vee x^2 \vee x^3) z^{Nom} (c^{\text{tar}} \vee c^{\text{ter}} \vee c^{\text{dar}} \vee c^{\text{der}} \vee c^{\text{lar}} \vee c^{\text{ler}} \vee c^0\ ) \tag{4.12}$$

Семантическую роль *Объекта* факта в казахской фразе, т.е. лицо или

$$\gamma_{2K} = (x^0 \vee x^2 \vee x^3)(z^{Gen} \vee z^{Acc})(y^{NoV} \vee y^{NoN} \vee y^{NCom} \vee y^{NDer} \vee y^0) \wedge \\ \wedge (c^{\text{tar}} \vee c^{\text{ter}} \vee c^{\text{dar}} \vee c^{\text{der}} \vee c^{\text{lar}} \vee c^{\text{ler}} \vee c^0)a^{\text{NSim}} \quad (4.13)$$

предмет, на которое направлено действие, определяем с помощью γ_{2K}:

Таблица 4.4 – Предметные переменные и их области значений логико-лингвистической модели Open IE казахского языка

Предметная переменная	Грамматический/семантический признак языка, описываемый переменной	Области изменения предметной переменной
x	месторасположение анализируемого слова в предложении	(4.1)
f	наличие или отсутствие вспомогательного глагола во фразе	(4.2)
z	грамматический падеж казахского существительного	(4.3)
a	склонения существительного	(4.4)
n	наличие отрицания во фразе	(4.5)
c	наличие аффикса множественности	(4.6)
y	словообразовательный аффикс конкретной части речи (глагола, причастия, деепричастия, существительного)	(4.7)
d	сослагательность действия	(4.8)
m	наличие личного предикативного или притяжательного окончания глагола и отглагольных форм	(4.9)
b	дополнительная определяемая семантика действия	(4.10)

Формализация логико-лингвистического уравнения *Предиката действия* в казахской фразе основывается на определении факта. Согласно “Новой философской энциклопедии” [87], факт представляет собой реальное, конкретное единичное событие или результат действия, которые произошли или произойдут. Таким образом, уравнение *Предиката действия* триплета учитывает только изъявительное наклонение казахского языка, оставляя повелительное, желательное и условное наклонения за границами исследования.

Предикат γ_{VK} определяет комбинацию семантических и грамматических признаков центральной части триплета факта, а именно Действия или *Предиката* факта:

$$\gamma_{VK}= (x^{-1}\vee x^{-2}\vee x^{-3})\,((f^{tur} \vee f^{otur}\vee f^{jatyr}\vee f^{júr})\, m^{PrF} z^{Vad} \vee (y^{Oad}\vee y^{FuCo})m^{PrFl}\vee$$
$$y^{FuCo}\,(m^{PrFl}\vee (m^{PrFl} f^{edi}))\vee y^{y}\,(f^{edi}\vee f^{eken})\vee (y^{Vad}\, m^{PrFl}\,(p^{mic}\vee p^{0}\,))\;\vee \qquad (4.14)$$
$$\vee m^{PoFl}\,((y^{Vart}\vee y^{Vpa}\vee y^{Vpas})\vee f^{edi}(n^{joq}\vee n^{emes}\vee n^{me}\vee n^{0})\wedge$$
$$\wedge (y^{Part}\vee y^{Vad}\vee f^{otur}\vee f^{tur}\vee f^{jatyr}\vee f^{júr}\vee f^{ParP}\vee f^{UnFu})))$$

На рисунке 4.4 показан пример реализации модели для предложения казахского языка. В казахской фразе «*Операторлар үйде мылтық тапты*», согласно формуле (4.14) глагол «*тапты*» представляет действие (давно прошедшее время). Согласно уравнению (4.12), существительное «*Операторлар*» идентифицируется как *Субъект* действия или Субъект факта. Предикат γ_{2K} (4.13) идентифицирует существительное «*мылтық*», как *Объект*, факта, называемого данной фразой.

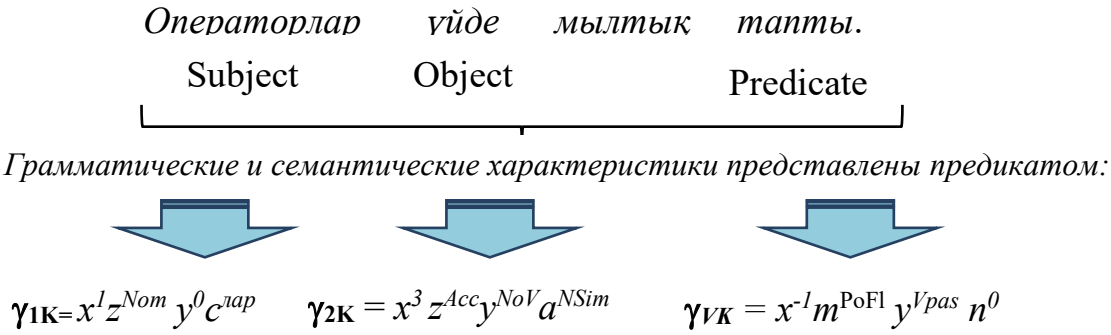


Рисунок 4.4 — Пример идентификации факта во фразе казахского языка. Предикат γ_{1K} определяет грамматические особенности *Subject* действия; предикат γ_{2K} определяет *Object*; и γ_{VK} — это *Predicate* факта.

5. ОСОБЕННОСТИ ФОРМИРОВАНИЯ КАЗАХСКО-РУССКОГО ПАРАЛЛЕЛЬНОГО КОРПУСА ТЕКСТОВ КРИМИНАЛЬНОЙ ТЕМАТИКИ

5.1.Проблемы формирования параллельных корпусов

На сегодняшний день лингвистические ресурсы являются не только неотъемлемой частью любого лингвистического исследования, но и важной базой разработки любых NLP приложений, таких как системы Machine Translation, Information Retrieval, Text Summarization, Human-Machine Dialogue и др. Такого рода лингвистические ресурсы, обычно, включают в себя словари, тезаурусы, лингвистические онтологии, одноязычные, двуязычные и многоязычные корпуса. Процесс их разработки включает словарные исследования, анализ лексической структуры языка, рассмотрение текстовых характеристик и изучение похожих работ.

При этом, одним из самых актуальных и прогрессивных направлений разработки лингвистических ресурсов является проектирование, создание и развитие высококачественных текстовых корпусов [88]. Обработанный и систематизированный с помощью конкордансера корпус позволяет хранить огромное количество лингвистической информации, необходимой для статистического анализа; диахронических изменений и других исследований в разговорном и письменном языках.

Существующие на сегодня корпуса можно разделить на специализированные (жанр, время, место), общие, мульти-язычные, обучающие, исторические или диахронические, мониторинговые и другие. Исследуемые нами мультиязычные корпуса, в свою очередь, делятся на сопоставительные (сравнительные) и параллельные или переводческие корпуса. Как правило, параллельные корпуса остаются наиболее важными при изучении языка и особенностей перевода, разработке синтаксического парсера, задач по распознаванию речи и т.д.

В частности, концепт параллельного корпуса является составной частью такого более широкого и сложного понятия, как машинный перевод. Качество машинного перевода во многом зависит от количества параллельных предложений, использованных при обучении. При этом, несмотря на стремительный рост количества разнообразных программ и практических приложений, машинный перевод, по-прежнему, остается нерешенной задачей компьютерной лингвистики.

За последнее десятилетие в мире было создано множество двуязычных и мульти-язычных корпусов, среди которых, с нашей точки зрения, самыми интересными являются: EUROPARL — корпус Европейского Парламента, включа-

ющий 20.000.000 слов на 11 языках[1]; CHEMNITZ GERMAN-ENGLISH TRANSLATION CORPUS, включающий 1.000.000 слов[2]; KACENKA — англо-чешский корпус, содержащий 3.000.000 слов[3]; English-French Canadian Hansard — англо-французский параллельный корпус [89].

В тоже время существует достаточно много лингвистических корпусов казахского языка, среди которых наиболее известными являются:

– Алматинский корпус казахского языка (АККЯ), содержащий более 40 миллионов словоупотреблений, 86% которых имеют грамматический разбор;
– Almaty Corpus of Kazakh [90];
– Kazakh text corpora on Sketch Engine [91];
– Open-Source-Kazakh-Corpus, созданный с использованием инструмента Wikipedia dump и включающий коллекцию из 20 миллионов слов (из которых 600 тысяч уникальных) [92];
– Корпус Казахского языка или Kazakh Language Corpus (KLC) [93].

В то же время, несмотря на существование большого количества параллельных мульти-язычных корпусов, для казахского языка задание по созданию параллельных корпусов продолжает быть достаточно актуальным. Данная задача существенно усложняется, если мы говорим о разработке параллельного казахско-русского корпуса, входной язык которого принадлежит к тюркской, а выходной к индоевропейской языковым семьям.

Для того чтобы реализовывать свой потенциал, современные параллельные корпуса должны быть выравнены. Выравнивание подразумевает совпадение определенных фрагментов оригинального текста с соответствующими фрагментами текста перевода.

В большей части работ, посвященных параллельным корпусам, прямо или косвенно выделяются два уровня выравнивания: выравнивание предложений и лексическое выравнивание. Обычно задача автоматического сравнения предложений, включающая сопоставление слов оригинала с их эквивалентами в переводе, очень сложна и трудоемка, так как для многих языков предложения или слова могут не сопоставляться «один к одному». Например, несколько абзацев в исходном языке могут соответствовать только одному абзацу на языке перевода; кроме того, при переводе некоторые слова могут быть удалены или заменены отдаленными синонимами или устойчивыми выражениями, которые могут быть абсолютно разными для разных языков, и т.д.

Существующие методы выравнивания предложений мы можем разделить на 3 категории. Методы первой категории основаны на определении длин пред-

[1] https://www.isi.edu/~koehn/publications/europarl/
[2] http://www.tu-chemnitz.de/phil/InternetGrammar
[3] http://www.phil.muni.cz/angl/kacenka/kachna.html

ложений и абзацев [89]. Данный подход базируется на гипотезе, утверждающей, что длины предложений в оригинале и переводе приблизительно совпадают.

Вторая группа методов выравнивания использует лексическую информацию, полученную из корпуса [94]. Методы данной группы используются крайне редко, что обусловлено труднодоступностью двуязычных словарей и сложностью автоматического морфологического анализа, используемого для идентификации слов в текстах. На сегодняшний день, большая часть программ, основанных на этой группе методов, использует только тексты специализированных тем, например, тексты речей парламента и правовые тексты [95].

Третья группа алгоритмов выравнивания текстов параллельных корпусов основана на POS-tagging или морфологической разметке, содержащейся в аннотированных корпусах [96].

Однако, реализация любого метода из этих трех групп связана с определенным количеством неточностей, в связи с чем, непрерывно растет интерес к созданию систем, использующих совокупность всех трех методов. В частности, в работе [97] описывается гибридный метод выравнивания параллельных текстов, сочетающий зависимости длин фрагментов и элементы перевода. Основой исследования послужили венгерский, румынский и словенский языки.

Авторы исследования [98] показали, что выравнивания текста можно достигнуть без использования дополнительных ресурсов конкретных языков. Они использовали алгоритм выравнивания, основанный на длине предложений, и тренировали систему МТ на текстах, нуждающихся в выравнивании. Система МТ использовалась для того, чтобы перевести ресурсы параллельного тренировочного корпуса; после чего на полученном автоматическом переводе осуществлялось выравнивание.

Другой подход выравнивания предложений описан в статье [99]. В данной работе, авторы предложили Fast-Champollion алгоритм выравнивания текстов, который применяет комбинацию методов, основанных на длине и на лексиконе, полученном с помощью словаря. Алгоритм получил эпитет «быстрого», поскольку он оптимизировал процесс деления введенного двуязычного текста на маленькие фрагменты выравнивания. В работе [100] осуществлен обзор специального инструментария InterText, применяемого для выравнивания параллельных корпусов. Работа приложения основана на гибридном методе выравнивания. Это же приложение использовалось для создания казахско-английского корпуса в исследовании [101]. Авторы исследования использовали инструмент Bitextor для генерации корпуса на базе многоязычных веб-сайтов. Они загру-

жали весь веб-сайт и применяли набор правил, основанных главным образом на HTML-структуре текста и длине текстовых блоков [102] .

Авторы статьи [103] выравнивали свои тексты на уровне предложений, используя знаки пунктуации для сегментации. При этом, подход нуждался в ручной отладке. Основанные на данном подходе корпуса финского и русского языков выравнены достаточно удачно [104].

Дополнительной сложностью при создании выравненного бинарного параллельного корпуса является выбор соответствующего контента наполнения корпуса. В настоящее время существует большое количество исследований, описывающих получение параллельных предложений из непараллельной или сопоставимой информации. К примеру, такую информацию можно получить с помощью хорошо известного ресурса Википедия [105], включающего похожие статьи на различных языках. Кроме того, такие статьи могут быть связаны через ссылки “ interwiki”, аннотируемые в Википедии пользователями. Однако, возможности по созданию параллельных корпусов даже такого глобального ресурса к настоящему моменту исследованы не до конца [106].

Очевидно, что проблема создания параллельных выравненных корпусов к настоящему моменту до конца не решена, и универсальные методы выравнивания не определены. Более того, можно сказать, что на сегодняшний день, в большинстве подходов, выбор метода выравнивания напрямую зависит от исследуемой пары языков, тематического направления и типа документов, представленных в корпусе [107].

5.2. Разработка и аннотирование корпусов текстов казахского и русского языков криминальных текстов

Разрабатываемый корпус казахских и русских криминально окрашенных текстов представляет собой файловую структуру, показанную на рисунке 5.1.

Имя текстового файла должно соответствовать шаблону:

порядковыйНомер_названиеАгенства_дата_язык_tag|row.txt

Например, размеченный текстовый файл текста порядкового номера 49 на казахском языке, полученный на сайте информационного агентства *partrul* седьмого сентября 2018 года должен иметь имя: 49*_partrul_07.09.18_Kz_tag.txt*

Критериями оценки текстового корпуса, помимо его репрезентативности и размера, является используемая система разметки и правильность кодирования метаданных корпуса. Разметка представляет собой добавление в текстовый корпус некоторой дополнительной лингвистической мета-информации. Это

может быть морфологическая (POS-tagging), синтаксическая, семантическая и другая информация. Для созданных корпусов криминально значимой информации мы используем морфологическую, семантическую и темпоральную разметку.

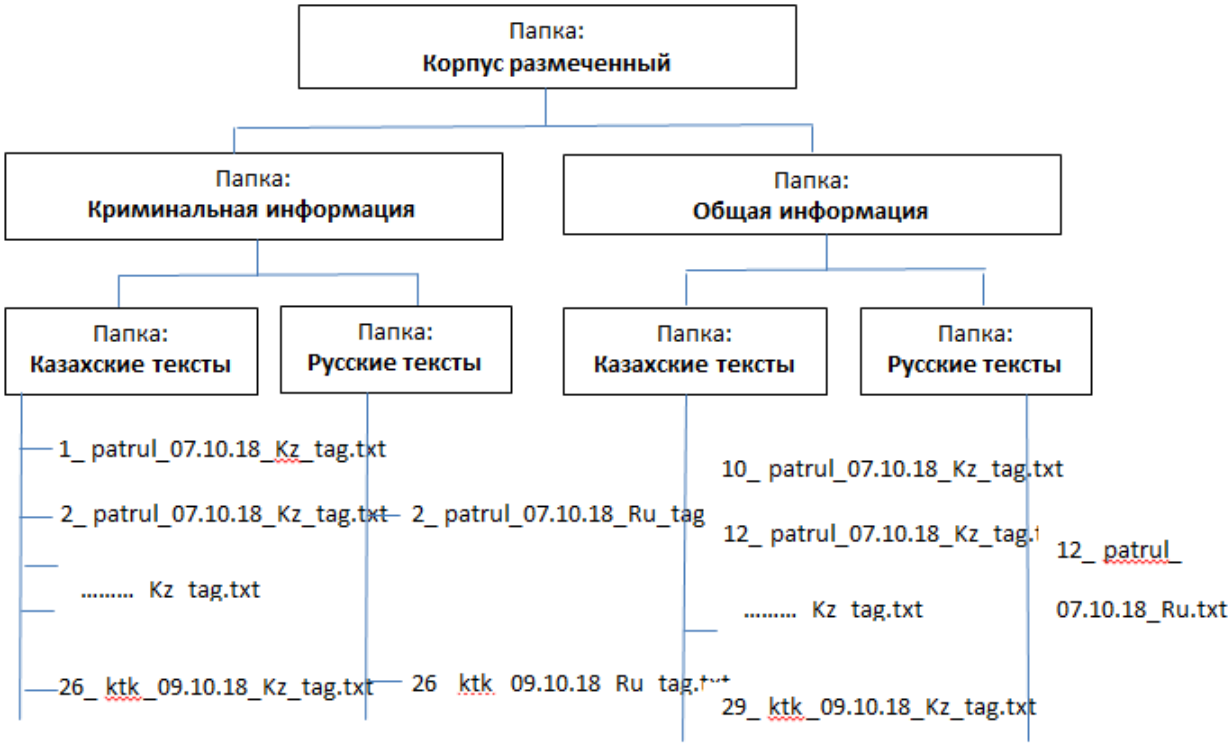


Рисунок 5.1 – Файловая структура казахско-русского корпуса криминально окрашенных текстов

Использование ручного метода разметки повышает ее точность, в то же время, требует много трудозатрат и не позволяет размечать корпуса большого объема. И наоборот, использование автоматической разметки увеличивает количество ошибок, уменьшает точность, но позволяет достаточно быстро осуществлять разметку больших корпусов.

Кроме того, в некоторых случаях достаточно сложно создать алгоритмы семантической, стилистической и некоторых других видов разметки. По этой причине, в нашем исследовании часть информации, в частности, POS-tagging может быть добавлена в корпус автоматически, а семантическая информация добавляется вручную. В итоге, процент ошибки и многозначности аннотирования разрабатываемых корпусов не должен превышать 5%.

Структурная разметка разрабатываемого корпуса включает следующие теги:

а) заголовок текста выделяется тегом:

<head type=main> заголовок</head>

b) если в тексте есть подзаголовки они выделяются тегами:

<head type=h1> заголовок</head>

c) дата публикации выделяется:

<date> </date>

d) сайт информационного Веб-агенства, откуда взят текст, выделяется тегом:

<site> </site>

e) если есть автор, то он размечается тегом:

<author> автор < /author>

Одним из основных этапов грамматического и семантического тегирования текстов корпуса является выбор множества тегов или множество категорий слов, которые будут применяться к токенам (tagset). *Tagset* представляет собой множество тегов или категорий слов, используемых для данной задачи грамматического тегирования, от выбора которых зависит скорость и точность автоматической обработки. При разработке tagset необходимо учитывать следующие критерии выбора меток:

1) краткость (*conciseness*) – короткие метки более удобны, чем более подробные и, соответственно, длинные;
2) понятность (*perspicuity*) – легко интерпретируемые метки;
3) анализируемость (*analysability*) – метки должны легко декомпозироваться на логические части, как легко читаемые при машинной обработке, так и легко понимаемые человеком.

Основываясь на вышеперечисленных критериях, были выбраны следующие метки:

– тегами *<s type=crim> предложение </s>* выделяются предложения, имеющие криминальную окраску;

– тегом *<v type=crim > глагол </v>* выделяются глаголы, имеющие криминальную окраску;

– тегом *<n type=crim > существительное </n>* выделяются существительные, в значении которых есть определенная криминальная окраска;

– тегом * прилагательное и причастие * выделяются прилагательные и причастия, в значении которых есть определенная криминальная окраска;

В таком обозначении имена тегов (типы элементов) *<s>*, *<v>*, *<n>*, *<a>* определяют грамматическую информацию части речи и предложения как синтаксической единицы. А значения атрибута *type=crim* определяет семантическую информацию криминальной окрашенности.

На рисунке 5.2 показан фрагмент размеченного файла корпуса.

<head type=main>Полицейские Алматинской области изъяли у мужчины более 5 кг наркотиков
</head>¶

<date> 06.10.2018 </date>¶

<site> patrul.kz</site>¶

<s type=crim>Задержание подозреваемого <v type=crim> произведено</v>
<n type=crim> полицейскими</n> в ночное время на контрольном посту, действующем на 59 км
трассы «Алматы – Бишкек».</s>¶

<s type=crim>В автомашине мужчины специально<v type=crim> обученная</v> на поиск<n
type=crim> наркотиков</n>служебно-розыскная собака по кличке Таймас<
type=crim> обнаружила</v> тайник с <n type=crim> наркотиками</n>.</s>¶

<s type=crim>-«<n type=crim>Тайник </n> был установлен под одним из сидений автомашины
Mercedes.</s><s type=crim><n type=crim>Полицейскими</n> было<v type=crim> изъято</v>
5 килограмм 360 грамм <n type=crim> гашиша</n>.</s><s type=crim>
Задержанный мужчина</v> помещен</v> под «<n type=crim> стражу</n> в Жамбылское
районное подразделение <n type=crim> полиции</n>».</s>¶

Рисунок 5.2 – Фрагмент размеченного файла корпуса

5.3. Алгоритм семантической разметки казахского корпуса текстов, включающих криминальное значение

POS-тегирование казахских текстов осуществлялось с использованием разработанного тегера, базирующегося на *RegexpTagger* классе пакета *NLTK Python*. На рисунке 5.3 показан фрагмент регулярного выражения, позволяющего идентифицировать некоторые формы существительных в казахских предложениях.

```
patterns=[(r'.*бен$','NN'), ('r.* пенен$','NN'),
('r.* басшылық$','NN'), (r'.* ықпону$','NN'), (r'.* тар-
мен$','NN'), (r'.* герлермен$','NN'), (r'.* здар$','NN')]
```

Рисунок 5.3 – Фрагмент регулярного выражения, позволяющего идентифицировать некоторые формы существительных в казахских предложениях.

Семантическая разметка корпуса текстов казахского языка, содержащего криминальнально значимую информацию, заключается в выделении и обозначении триплета факта: *Субъект* → *Предикат*→*Объект*. Корпус имеет горизонтальный формат разметки. Использование полученных имен тегов морфологической разметки и некоторых синтаксических характеристик слов в предложе-

нии в качестве значений предметных переменных уравнений (4.12 – 4.14) позволяет извлекать *Subject*, *Object* и *Predicate* факта из предложений казахского языка

Субъект действия, обозначаемый меткой "*_Sub*" представляет собой персоналию или предмет, являющийся инициатором действия. Предикат определяется на базе формулы (4.12). *Объект* действия, обозначаемый меткой "*_Ob*", представляет собой персоналию или предмет, на который действие направлено, и определяется на базе формулы (4.13). Ядро триплета факта представляет собой *Предикат*, обозначаемый меткой "*_Pred*", называет действие факта и определяется на базе логико-лингвистического уравнения (4.14). При разметке использовался нижеследующий алгоритм дерева принятия решений.

1. К слову добавляется метка "*_Sub*", если первое, второе или третье слово в предложении, которое имеет суффикс множественного числа из списка [*мар*, *тер*, *дар*, *дер*, *лар*, *лер*], перед которым не стоят суффиксы:
   - - родительного падежа из списка [ның, нің, дың, дің, тың, тің],
   - - направительно-дательного падежа [*ға*, *ге*, *қа*, *ке*, *а*, *е*, *на*, *не*],
   - - винительного падежа [*ны*, *н*, *ні*, *ды*, *ді*, *ты*, *ті*],
   - - местного падежа [*да*, *де*, *нда*, *нде*, *та*, *те*]
   - -исходного падежа [*дан*, *ден*, *тан*, *тен*, *нан*, *нен*]
   - -творительного падежа [*мен*, *менен*, *бен*, *бенен*, *пен*, *пенен*]

2. Если слова, отвечающего условию 1 в предложении нет, метка "*_Sub*" добавляется к слову, для которого выполняются условия 2 а) и 2 б) одновременно:

   а) есть суффикс словообразования существительного из списка:
      – именного образования существительного [*ғай*, *гей*, *гер*, *ғи*, *ғой*, *дас*, *дес*, *дік*, *дық*, *кер*, *кес*, *қай*, *қар*, *қи*, *қой*, *қор*, *лас*, *лес*, лік, *лық*, *ман*, *паз*, *пана*, *сақ*, *тас*, *тес*, *тік*, *тық*, *хана*, *ша*, *шақ*, *ше*, *шек*, *ші*, *шік*, *шы*, *шық*];
      – отглагольного образования существительного [*ақ*, *ба*, бе, *ғақ*, ғаш, *гек* , *гі*, *ғіш*, *ғы*, *ғыш*, *дақ*, дек, *ек*, *ік*, *ім*, *іс*, *іш*, *к*, *кі*, *кіш*, *қ*, *қаш*, *қы*, *қыш*, *лақ*, *лек*, *м*, *ма*, *мақ*, *ме*, *мек*, *па*, *пақ*, *пе*, *пек*, *с*, *тақ*, *тек*, *уік*, *уық*, *ш*, *ық*, *ым*, *ыс*, *ыш*];
      – сложного аффикса образования существительного [*герлік*, *гіштік*, *ғыштық*, *дастық*, *дестік*, *ділік*, *дылық*, *кеәтік*, *қорлық*, *ластық*, *лестік*, *лілік*, *лылық*, *паздық*, *сақтық*, *сіздік*, *сыздық*, *тастық*, *тестік*, *тілік*, *тылық*, *шақтық*, *шілдік*, *шілік*, *шылдық*, шылық];
      –экспрессивной оценки [*жан*, *ке*, *қан*, *сымақ*, *тай*, *ш*, *ша*, *шақ*, *ше*, *шік*, шық].

b) после которого не стоит падежный суффикс:

- родительного падежа из списка [*ның*, *нің*, дың, *дің*, *тың*, *тің*],
- направительно-дательного падежа [*ға*, *ге*, *қа*, *ке*, *а*, *е*, *на*, *не*],
- винительного падежа [*ны*, *н*, *ні*, *ды*, *ді*, *ты*, *ті*],
- местного падежа [*да*, *де*, *нда*, *нде*, *та*, *те*]
- исходного падежа [*дан*, *ден*, *тан*, *тен*, *нан*, *нен*]
- творительного падежа [*мен*, *менен*, *бен*, *бенен*, *пен*, *пенен*]

3. К слову добавляется метка "_ Obj ", если это первое слово от начала предложения, имеющее суффикс:

- направительно-дательного падежа [*ға*, *ге*, *қа*, *ке*, *а*, *е*, *на*, *не*],
- винительного падежа [*ны*, *н*, *ні*, *ды*, *ді*, *ты*, *ті*],
- после которого может стоять окончание множественности [*тар*, *тер*, *дар*, *дер*, *лар*, *лер*],

4. К слову добавляется метка "*_Pred*", если это— последнее слово предложения, с суффиксами [*п*, *ып, ін*] и в предложении есть слово, начинающееся с [*тұр*, *отыр*, *жатыр*, *жүр*], после которых стоит суффикс [*біз*, *бін*, *быз*, *бын*, *ды*, *міз*, *мін*, *мыз*, *мын*, *піз*, *пің*, *пыз*, *пын*, *сіз*, *сіздер*, *сіндер*, *сің*, *сыз*, *сыздар*, *сың*, *сыңдар*, *ті*, *ты*].

5. К слову добавляется метка "*_Pred*", если это — последнее слово предложения, которое имеет суффикс будущего предположительного времени [*ар*, *ер*, *ыр*, *ір*] или суффикс деепричастия [*а*, *е*, *й*, *и*], после которого стоит личный предикативный суффикс [*біз*, *бін*, *быз*, *бын*, *ды*, *міз*, *мін*, *мыз*, *мын*, *піз*, *пің*, *пыз*, *пын*, *сіз*, *сіздер*, *сіндер*, *сің*, *сыз*, *сыздар*, *сың*, *сыңдар*, *ті*, *ты*]

6. К слову добавляется метка "*_Pred*", если это — последнее слово предложения, которое имеет суффикс будущего предположительного времени [*ар*, *ер*, *ыр*, *ір*], и в предложении есть вспомогательный глагол [*еді, е*], после которого может стоять личный предикативный суффикс [*біз*, *бін*, *быз*, *бын*, *ды*, *міз*, *мін*, *мыз*, *мын*, *піз*, *пің*, *пыз*, *пын*, *сіз*, *сіздер*, *сіндер*, *сің*, *сыз*, *сыздар*, *сың*, *сыңдар*, *ті*, *ты*]

7. К слову добавляется метка "*_Pred*", если в предложении есть вспомогательный глагол [*еді*, *екен*], после которого может стоять личный предикативный суффикс [*біз*, *бін*, *быз*, *бын*, *ды*, *міз*, *мін*, *мыз*, *мын*, *піз*, *пің*, *пыз*, *пын*, *сіз*, *сіздер*, *сіндер*, *сің*, *сыз*, *сыздар*, *сың*, *сыңдар*, *ті*, *ты*], а данное слово является последним словом предложения, которое имеет один из следующих суффиксов

– [*ді*, *дік*, *дің*, *дім*, *дің*, *ды*, *дык*, *дық*, *дым*, *дың*, *қ*, *ті*, *тік*, *тім*, *тің*, *ты*, *тык*, *тық*, *тым*, *тың*];

– [*а*, *ай*, *ал*, *ан*, *ар*, *арыс*, *ға*, *ғал*, *ғар*, *ғе*, *ге*, *гер*, *гі*, *гіз*, *гіздір*, *гіле*, *гір*, *гім*, *ғы*, *ғыз*, *ғыздыр*, *ғызыл*, *ғыла*, *ғыр*, *ғыт*, *да*, *дан*, *дар*, *дас*, *дастыр*, *де*, *ден*, *дендір*, *дес*, *діг*, *дік*, *дір*, *діргіз*, *дық*, *дыр*, *дырғыз*, *дырыл*, *е*, *ей*, *ел*, ен, *ер*, *й*, *іг*, *іғ*, *ік*, *ікіс*, *іл*, *іла*, *ілде*, *ілу*, *імсіре*, *ін*, *індір*, *ініс*, *іну*, *іңкіре*, *ір*, *ірде*, *іре*, *ірей*, *іріс*, *ірке*, *іркен*, *ірқе*, *іс*, *ісу*, *ім*, *ке*, *кер* , кіз, *кіле*, *кір* , *қа*, *қал*, *қан*, *қар*, *қе*, *қур*, *қыз*, *қыла*, *қыла*, *қыр*, *л*, *ла*, *лан*, *ландыр*, *лас*, *ластыр*, *лат*, *ле*, *лен*, *лендір*, *лес*, *лестір*, *лет*, *ліг*, *лік*, *лікіс*, *лін*, *ліс*, *лқа*, *лу*, *лығ*, *лық*, *лын*, *лыс*, *мала*, *меле*, *мсіре*, *мсыра*, *н*, *ні*, *ніл*, *ніс*, *ныл*, *ныс*, *ңгіре*, *ңғыра*, *ңкіре*, *ңқыра*, *ңра*, *ңре*, *р*, *ра*, *ре*, *с*, *са*, *сан*, се, *сен*, *сет*, *сетіл*, *сі*, *сін*, *сіре*, *стір*, *стыр*, *сы*, *сын*, *сыра*, *т*, *та*, *тан*, *тандыр*, *тас*, *те*, *тен*, *тендір*, *тес*, *тік*, *ттыр*, *тығ*, *тығс*, *тығыс*, *тық*, *тыр*, *тырыл*, *ура*, *ші*, *шы*, *ығ*, *ығыс*, *ық*, *ықыс*, *ыл*, *ыла*, *ылда*, *ылу*, *ылыс*, *ымсыра*, *ын*, *ындыр*, *ыну*, *ыныс*, *ыр*, *ыра*, *ырай*, *ырқа*, *ырқан*, *ырла*, *ыс*, *ысу*, *ыт*];

– [*азы*, *ақта*, *ал*, *ала*, *аңғыра*, *аура*, *бала*, *бе*, *беле*, *би*, *бі*, *бы*, *дала*, *ди*, *ді*, *ды*, *екте*, *ел*, *еңгіре*, *еуре*, *жи*, *жіре*, *жыра*, *зы*, *і*, *ін*, *ірей*, *іс*, іт, *қи*, *лі*, лы, *ма*, *мала*, *меле*, *ми*, *мсіре*, *мсыра*, *ңра*, *ңре*, *палапеле*, *пи*, *пі*, *пы*, *ра*, *ре*, *си*, *сіре*, *сый*, *сыра*, *т*, *ти*, *ті*, *ты*, *усіре*, *усыра*, *ши*, *ші*, *шы*, *ы*, *ын*, *ыра*, *ырай*, *ыс*, *ыт*].

8. К слову добавляется метка "*_Pred*", если это — последнее слово предложения, которое имеет суффикс [*п*, *ып*, *ін*], после которого стоит личный предикативный суффикс [*біз*, *бін*, *быз*, *бын*, *ды*, *міз*, *мін*, *мыз*, *мын*, *піз*, *пің*, *пыз*, *пын*, *сіз*, *сіздер*, *сіндер*, *сің*, *сыз*, *сыздар*, *сың*, *сыңдар*, *ті*, *ты*]; в слове так же могут присутствовать суффиксы [*мыс*, *міс*].

9. К слову добавляется метка *"_Pred*", если это — последнее слово предложения, которое имеет один из следующих суффиксов:

– [*ді*, *дік*, *дің*, *дім*, *дің*, *ды*, *дык*, *дық*, *дым*, *дың*, *қ*, *ті*, *тік*, *тім*, *тің*, *ты*, *тык*, *тық*, *тым*, *тың*];

– *[а*, *ай*, *ал*, *ан*, *ар*, *арыс*, *ға*, *ғал*, *ғар*, *ғе*, *ге*, *гер* , *гі*, *гіз*, *гіздір*, *гіле*, *гір*, *гім*, *ғы*, ғыз, *ғыздыр*, *ғызыл*, *ғыла*, *ғыр*, *ғыт*, *да*, *дан* , *дар*, *дас*, *дастыр*, *де*, *ден*, *дендір*, *дес*, *діг*, *дік* , дір, *діргіз*, *дық*, *дыр*, *дырғыз*, *дырыл*, *е*, *ей*, *ел*, *ен*, *ер*, *й*, *іг*, *іғ*, *ік*, *ікіс*, *іл*, *іла*, *ілде*, *ілу*, *імсіре*, *ін*, *індір*, *ініс*, іну, *іңкіре*, *ір*, *ірде*, *іре*, *ірей*, *іріс*, *ірке*, *іркен*, *ірқе*, *іс*, *ісу* , *ім*, *ке* , *кер* , *кіз*, *кіле*, *кір*, *қа*, *қал*, *қан*, *қар*, *қе*, *қур*, *қыз*, *қыла*, *қыла*, *қыр*, *л*, ла , *лан*, *ландыр*, *лас*, *ластыр*, *лат*, *ле*, *лен*, *лендір*, *лес*, *лестір*, *лет*, *ліг*, *лік*, *лікіс*, *лін*, *ліс*, *лқа*,

лу, *лығ*, *лық*, *лын*, *лыс*, *мала*, *меле*, *мсіре*, *мсыра*, *н*, *ні*, *ніл*, *ніс*, *ныл*, *ныс*, *ңгіре*, *ңғыра*, *ңкіре*, *ңқыра*, *ңра*, *ңре*, *р*, *ра*, *ре*, *с*, *са*, *сан*, *се*, *сен*, *сет*, *сетіл*, *сі*, *сін*, *сіре*, *стір*, *стыр*, *сы*, *сын*, *сыра*, *т*, *та*, *тан*, *тандыр*, *тас*, *те*, *тен*, *тендір*, *тес*, *тік*, *ттыр*, *тығ*, *тығс*, *тығыс*, *тық*, *тыр*, *тырыл*, *ура*, *ші*, *шы*, *ығ* , *ығыс*, *ық*, *ықыс*, *ыл*, *ыла*, *ылда*, ылу, *ылыс*, *ымсыра*, *ын*, *ындыр*, *ыну*, *ыныс*, *ыр*, *ыра*, *ырай*, *ырқа*, *ырқан*, *ырла*, *ыс*, *ысу*, *ыт*];

– [*ған*, *ген*, *қан*, *кен*, *қон*, *ға*, *ге*, *қа*, ке], [*атын*, *етін*, *йтын*, *йтін*].

После данных суффиксов могут стоять личные притяжательные окончания некоторых глагольных форм [*дар*, *йік*, *йін*, *йық*, *йын*, *іздар*, *к*, *қ*, *м*, *ндар*, *ң*, *ңдер*, *ңіз*, *ңіздер*, *ңыз*, *сіздер*, *сің*, *сіңдер*, *сыздар*, *сың*, *сыңдар*, *ыздар*];

10. К слову добавляется метка "*_Pred*", если это — последнее слово предложения, которое имеет один из суффиксов:

– [*ді*, *дік*, *дік*, *дім*, *дің*, *ды*, *дык*, *дық*, *дым*, *дың*, *қ*, *ті*, *тік*, *тім*, *тің*, *ты*, *тык*, *тық*, *тым*, *тың*];

– – [*а*, *ай*, *ал*, *ан*, *ар*, *арыс*, *ға*, *ғал*, *ғар*, *ғе* , *ге* , *гер* , *гі*, *гіз*, *гіздір*, *гіле*, *гір* , *гіт*, *ғы*, *ғыз*, *ғыздыр*, *ғызыл*, *ғыла*, *ғыр*, *ғыт*, *да*, *дан* , *дар*, *дас*, *дастыр*, *де*, *ден*, *дендір*, *дес*, *діг*, *дік*, *дір*, *діргіз*, *дық*, *дыр*, *дырғыз*, *дырыл*, *е*, *ей*, *ел*, *ен*, *ер*, *й*, *іг*, *іғ*, *ік*, *ікіс*, *іл*, *іла*, *ілде*, *ілу*, імсіре, *ін*, *індір*, *ініс*, іну, *іңкіре*, *ір*, *ірде*, *іре*, *ірей*, *іріс*, *ірке*, *іркен* , *ірқе*, *іс*, *ісу*, *ім*, *ке*, *кер* , *кіз*, *кіле*, *кір* , *қа*, *қал*, *қан*, *қар*, *қе*, *құр*, *қыз*, *қыла*, *қыла*, *қыр*, *л*, *ла*, *лан*, *ландыр*, *лас*, *ластыр*, *лат*, *ле*, *лен*, *лендір*, *лес*, *лестір*, *лет*, *ліг*, *лік*, *лікіс*, *лін*, *ліс*, *лқа*, *лу*, *лығ*, *лық*, *лын*, *лыс*, *мала*, *меле*, *мсіре*, *мсыра*, *н*, *ні*, *ніл*, *ніс*, *ныл*, *ныс*, *ңгіре*, *ңғыра*, *ңкіре*, *ңқыра*, *ңра*, *ңре*, *р*, *ра*, *ре*, *с*, *са*, *сан*, *се*, *сен*, *сет*, *сетіл*, *сі*, *сін*, *сіре*, *стір*, *стыр*, *сы*, *сын*, *сыра*, *т*, *та*, *тан*, *тандыр*, *тас*, *те*, *тен*, *тендір*, *тес*, *тік*, *ттыр*, *тығ*, *тығс*, *тығыс*, *тық*, *тыр*, *тырыл*, *ура*, *ші*, *шы*, *ығ* , *ығыс*, *ық*, *ықыс*, *ыл*, *ыла*, *ылда*, *ылу*, *ылыс*, *ымсыра*, *ын*, *ындыр*, *ыну*, *ыныс*, *ыр*, *ыра*, *ырай*, *ырқа*, *ырқан*, *ырла*, *ыс*, *ысу*, *ыт*];

– — [*азы*, *ақта*, *ал*, *ала*, *аңғыра*, *аура*, *бала*, *бе*, *беле*, *би*, *бі*, *бы*, *дала*, *ди*, *ді*, *ды*, *екте*, *ел*, *еңгіре*, *еуре*, *жи*, *жіре*, *жыра*, *зы*, *і*, *ін*, *ірей*, *іс*, *ім*, *қи*, *лі*, *лы*, *ма*, *мала*, *меле*, *ми*, *мсіре*, *мсыра*, *ңра*, *ңре*, *палапеле*, *пи*, *пі*, *пы*, *ра*, *ре*, *си*, *сіре*, *сый*, *сыра*, *т*, *ти*, *ті*, *ты*, *усіре*, *усыра*, *ши*, *ші*, *шы*, *ы*, ын, *ыра*, *ырай*, *ыс*, *ыт*].

После данных суффиксов могыт стоять личные притяжательные окончания [*дар*, *йік*, *йін*, *йық*, *йын*, *іздар*, *к*, *қ*, *м*, *ндар*, *ң*, *ңдер*, *ңіз*, *ңіздер*, *ңыз*, *сіздер*, *сің*, *сіңдер*, *сыздар*, *сың*, *сыңдар*, *ыздар*];

11. Если последнее слово предложения не удовлетворяет ни одному из условий пунктов 3-9 данного алгоритма, то на эти же условия проверяется предпоследнее слово и затем третье от конца предложения слово.

Для оценки результатов автоматического извлечения фактов из текстов, содержащих криминально значащую информацию, использовалось следующая методика экспертной оценки. Из автоматически извлеченных фактов произвольным образом было выбрано около тысячи фактов и представлено для оценки эксперту. Эксперт оценивал извлеченный факт как 1, если триплет факта идентифицирован правильно. Т.е. правильно определены все три элемента факта: инициатор действия — Subject, предмет или лицо, на которого направленно действие, — Object, действие, которое объединяет всех участников, — Prediate. Если же хотя бы один из трех элементов факта был выявлен не верно, эксперт оценивал данный факт как 0 —неверно определенный и извлеченный факт. На рисунке 5.4 показан интерфейс приложения, используемого для оценки правильности работы вышеприведенного алгоритма. Экспертное оценивание осуществлялось двумя экспертами.

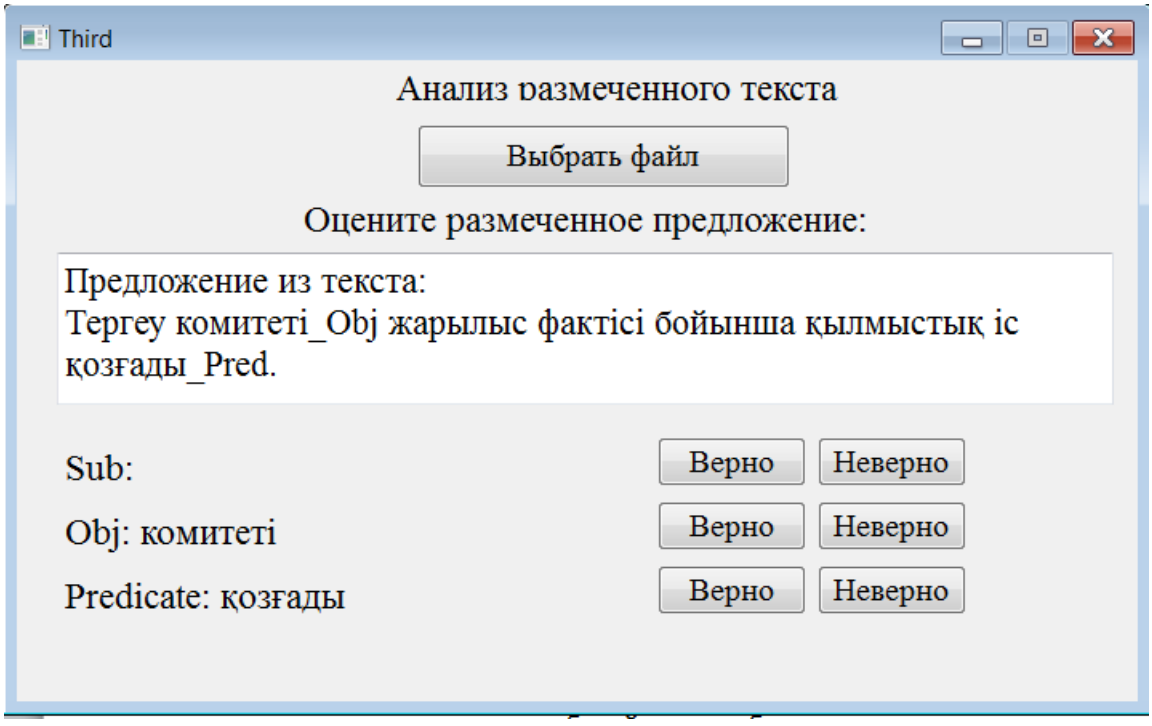


Рисунок 5.3 – Интерфейс приложения, позволяющего эксперту оценить правильность работы приложения.

В таблице 5.1 показаны полученные коэффициенты точности и согласованности (agreement) разработанной модели для корпуса криминально-окрашенных текстов казахского языка.

Таблица 5.1. Точность и согласованность (agreement) разработанной модели для корпуса криминально-окрашенных текстов казахского языка.

Язык корпуса	Размер корпуса (тыс. слов)	precision	agreement
казахский	225	71,0 %	0,72

5.4. Информационная технология выравнивания созданного корпуса текстов криминальной тематики.

Задача создания параллельного корпуса включает в себя несколько этапов. Первый базовый этап требует использования специализированных программных инструментов и техник для сбора текстового материала корпуса. При этом, несмотря на то, что интернет содержит огромное количество двуязычных и мульти-язычных веб-сайтов, выбор нужных двуязычных ресурсов представляет собой важную часть разработки параллельного корпуса. Это задание усложняется в связи с тем, что обрабатываются два таких разных языка, как казахский и русский.

Для сбора текстов были выбраны четыре двуязычных веб-сайта: *zakon.kz*, *caravan.kz*, *lenta.kz* и *nur.kz* [108]. Выбранные сайты представляют собой известные и надежные порталы Республики Казахстан, одним из новостных направлений которых являются криминальные новости. Порталы могут содержать новости о таких криминальных деяниях как грабежи, угоны машин, убийства, ДТП и другие. Тексты именно данной предметной области и представляют базовый ресурс создаваемого корпуса. Кроме того, данные порталы являются двуязычными, и часто содержат информационно близкие новостные публикации на казахском и русском языках. В результате осуществленного скрапинга вышеперечисленных сайтов были получены 3000 текстов на двух языках: русском и казахском. Для автоматического сбора (scraping) текстов сайтов разработана программа, осуществляющая разбор сайтов заданной структуры и требуемого контекста.

На следующем этапе вручную осуществлялась корректировка текстов криминальной окрашенности. Таким образом, был получен корпус, размером более 50410 слов (около 25600 слов принадлежат казахской половине корпуса и около 24800 слов — русской).

На следующем этапе определяется структурная организация корпуса. На сегодняшний день существует три базовых формата корпусов, определяемые в зависимости от прагматических целей создателей или пользователей: (1) тради-

ционный текстовый формат с отсылками к переводу; (2) представление текстов в табличной «зеркальной» форме, более удобной для восприятия и сравнения, (3) организация параллельного корпуса в форме базы данных.

Для созданного корпуса в качестве формата хранения данных была определена третья возможная структура — база данных (БД). БД является наиболее удобной структурой хранения большого количества данных, представляющих небольшие текстовые фрагменты, с возможностью последующего перманентного расширения и дополнения базы. Фрагмент таблицы созданной базы данных, включающей ID, название, адрес сайта, и текст новостной статьи, показан на рисунке 5.4.

id	head	url	text
3256	Месть за сестру:	https://www.inform.kz/	АЛМАТЫ. КАЗИНФОРМ - В специализи
3257	Чем чаще всего	https://www.inform.kz/	АСТАНА. КАЗИНФОРМ - Депутат Сенат
3258	Сенаторы ратиф	https://www.inform.kz/	АСТАНА. КАЗИНФОРМ - Депутаты Сен
3259	115 тысяч кварти	https://www.inform.kz/	АСТАНА. КАЗИНФОРМ - 115 тысяч ква
3260	Сенатор выразил	https://www.inform.kz/	АСТАНА. КАЗИНФОРМ - Депутат Сенат
3261	820 камер видеои	https://www.inform.kz/	КОКШЕТАУ. КАЗИНФОРМ - 820 камер в
3262	Почему приоста	https://www.inform.kz/	АСТАНА. КАЗИНФОРМ - Вице-министр
3263	Задержан подозр	https://www.inform.kz/	ТУРКЕСТАН. КАЗИНФОРМ - Полицейск
3264	Сенатор рассказ	https://www.inform.kz/	АСТАНА. КАЗИНФОРМ - В Казахстане р

Рисунок 5.4 — Фрагмент БД криминально окрашенных текстов новостных вебсайтов на казахском и русском языках.

Для осуществления POS-тегирования русских текстов корпуса ипользовался пакет *pymorphy2*[4] Python, разработанный специально для морфологического анализа русскоязычных текстов. Библиотеки пакета используют словарь *OpenCorpora*[5] и делают гипотетические выводы по нераспознаваемым словам.

Сложность структурного и типового аннотирования текстов казахского языка связана с его принадлежностью к агглютинативным языкам. Агглютинативный формат, в котором каждая агглютинация (суффикс или окончание) несет только одно семантическое или морфологическое значение, противоположен флективному, в котором каждая морфема имеет несколько неделимых значений сразу (например, падежа, рода и числа).

POS-тегирование казахских текстов осуществлялось с помощью регулярных выражений, основанных на классе *RegexpTagger* библиотеки *nltk Python* и ряда синтаксических правил [109]. Например, мы можем идентифицировать некоторые типы существительных казахского языка с помощью списка регулярных выражений, показанных на рисунке 5.6. Здесь, тег "*_NNat*" определяет

[4] https://nlpub.ru/Pymorphy
[5] https://www.pydoc.io/pypi/gensim-3.2.0/autoapi/corpora/dictionary/index.html

именительный падеж (атау) существительного, тег "*_NNil*" определяет родительный падеж (ілік), а тег "*_NNba*" — дательно-направленный падеж (барыс) (см. также пример рис. 5.3).

```
patterns=[(r'.*(шык|шы|пыр|мпыр|алар|ашыщ|лар|елер|ды) $',
'NNat'), (‘r.*(мның|енің|рдың|дың,)$','NNil'), (r'.*(
да|те|та|нда|нде|ға|ге|қа|ке|на|не|тік|еге|ырға|рға}йға|ыға|аға|ша
ға|сіз|мға|ға)$','NNba')]
```

Рисунок 5.6 — Пример регулярного выражения, позволяющего идентифицировать некоторые существительные именительного, родительного и дательно-направленного падежа.

Для того чтобы увеличить точность POS-тегирования дополнительно использовались семь синтаксических правил. Базой разработки таких правил являлся строгий порядок слов в предложениях казахского языка. Например, "Если токен следует за словами из специального списка – токен отмечается как глагол":

$$[\text{list\_1 of words}]\ \text{tokeni} \Rightarrow \text{tagtokeni='\_VV'}, \qquad (5.1)$$

где list_1 of words = [*қойды*, *қой*, *қалды*, *қал*, *салдым*, *салып*, *кетті*, *кетсеңші*, *бару*, *келу*, *шығу*, *жүру*, *түру*, *бар*, *қел*, *шық*, *жүр*, *қайт*, шыққан, *барған*, *түсті*, түс, *тұрыңдар*, *тұсын*, *көрме*, …]

Далее можно выбрать несколько подходов выравнивания предложений. Первый подход, основанный на совпадении длин предложений выравниваемых текстов, обеспечивает более высокую продуктивность. Однако, в нашем исследовании, не смотря на преимущества, данный подход не может принести точных и объективных результатов, так как в казахском языке часто для выражения некоторой смысловой и морфологической информации используются дополнительные слова, коренным образом изменяющие длину предложения. Именно по этой причине разницы в организации грамматики и семантики флективных и агглютинативных языков использование выравнивания по длинам предложений в нашем параллельном казахско-русском корпусе не эффективно.

Второй, выбранный нами, более ресурсоемкий подход, использует лексическое выравнивание слов. В качестве «лексического инструмента выравнивания» был использован созданный казахско-русский словарь, основанный на англо-казахско-русском словаре, включающем примерно 50 000 элементов. Фрагменты данных словарей показаны на рисунках 5.5. и 5.6. соответственно.

kz	ru
қылмыстық	криминал
мәйіті	трупы
мәліметтерге	данные
министрлігімен	с министерством
Оңтүстік	юг
орындарын	мест
өзара	взаимное
өзінің	его собственный
пиғылды	недобросовестный
Рудныйда	в Рудном
сату	продажа
сәттері	моменты

Рисунок 5.5 — Фрагмент созданного казахско-русского словаря

```
native###_туасынан, ·жаратылысынан, ·тумысынан###_прирождённый·
(прирожденный)¶
native###_табиғи, ·жаратынды###_природный¶
native###_тұнық, ·ұқыпты, ·кірсіз, ·пәк, ·мөлдір, ·кіршіксіз, ·бейкүнә, ·
әйнектей, ·ақ, ·мұнтаздай, ·таза, ·шайдай·ашық, ·мөлдір·бұлақ, ·дақсыз, ·саф, ·
ашық, ·адал, ·шын, ·нағыз, ·айна·дай, ·бәкізе, ·ғилман, ·жазықсыз###_чистый¶
native###_жергілікті·адам, ·туып-өскен, ·туған###_уроженец¶
native###_туған, ·бір·туған, ·тіған-туысқандар, ·туыстық, ·туып·өскен, ·
қарағым, ·шырағым###_родной¶
native·of·a·country·(·aborigine·)###_абориген###_абориген¶
nativity·(the·Nativity)###_рождество###_рождество¶
nativity###_дүниеге·келу, ·жаралу###_рождение¶
north·atlantic·treaty·organization·(NATO)###_°###_североатлантический·
союз·(НАТО)¶
North·Atlantic·Treaty·Organization·(NATO)###_°###_Североатлантический·
союз·(НАТО)¶
nato·(NATO;·North·Atlantic·Treaty·Organization)###_°###_нато·(НАТО;·
```

Рисунок 5.6 — Фрагмент используемого базового англо-казахско-русского словаря

Для полноценного использования данного словарного метода автоматического выравнивания предложений мы используем предварительно полученные результаты POS-тегирования текстов двух языков [109]. Использование корректной морфологической разметки позволяет правильно определить соответствие между словами, выделяя опорные токены выравненных предложений.

Созданный выравненный параллельный казахско-русский корпус состоит из текстов криминальной окрашенности, собранных с четырех казахских новостных сайтов за период май – декабрь 2018 года. Корпус включает около 50410 слов.

Для оценки точности проведенного автоматического выравнивания предложений корпуса было использовано экспертное оценивание трех экспертов, которые применяли специально разработанное приложение. Приложение позволяет экспертам выбрать текст на любом (русском или казахском) языке и автоматически загружает параллельный файл с текстом на противоположном языке (казахском или русском соответственно). Жирным шрифтом выделяются предложения, не получившие параллельный эквивалент на противоположном языке после автоматического выравнивания. Работая с корпусом, эксперты мо-

гут отметить тексты, сохранить их с отметками, или корректировать выровненные предложения вручную. На рисунке 5.7 показан пользовательский интерфейс приложения, используемого для работы с выровненным параллельным корпусом.

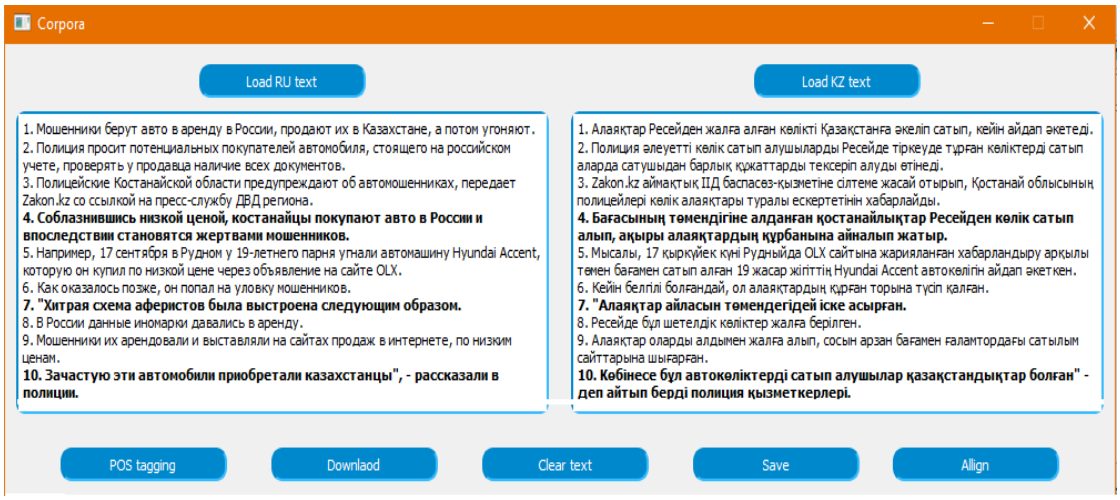


Рисунок 5.7 — Пользовательский интерфейс приложения, используемого для работы с выровненным параллельным корпусом

Проведенное экспертное оценивание показало, что точность автоматически выравненных предложений, созданного параллельного казахско-русского корпуса, составляет около 60%, при коэффициенте согласованности (agreement) равном 0.83. Остальные предложения выравнены вручную.

Проанализировав полученные результаты выравнивания, можно сделать следующие выводы о причинах ошибок.

– Наибольшее влияние на относительно низкую точность выравнивания созданного корпуса оказывает большая разница синтаксических структур казахского и русского языков, что глобально приводит к несовпадению количества предложений в двух частях корпуса. Некоторым предложениям русского текста соответствует несколько предложений казахского текста.

– Результат словарного метода выравнивания во многом зависит от качества используемого переводного словаря. Однако, в связи с тем, что русский и казахский языки находятся в отдаленных группах, при создании словарей возможны некоторые ошибки многозначности.

– Сложность и ограниченность использования сопоставимой грамматики для казахского и русского языков требует дальнейшей работы в направлении контрастивной лингвистики.

– Выравнивание текстов криминальной тематики требует учета имен собственных, званий, должностей и некоторых паттернов семантических классов слов (валют, дат и др.), особенно в новостных заголовках.

Все эти причины должны быть рассмотрены и учтены при дальнейшей работе с выровненным параллельным казахско-русским корпусом криминально-окрашенных текстов.

ЗАКЛЮЧЕНИЕ

В первом разделе монографии приведен аналитический обзор существующих проблем в области технологии поиска противоправной информации в текстовых данных. Рассмотрено современное состояние и перспективы развития методов формализации и поиска информации в неструктурированных и слабоструктурированных текстовых массивах, а так же проанализированы существующие возможности использования методов IE для извлечения криминально значимой информации. На основании проведенного анализа разработан общий подход к формализации и идентификации КЗИ.

Во втором разделе исследуется зависимость между лингвистическими формализмами в естественно-языковых текстах и реальной сущностью криминально или общественно значимого события в обществе. Рассматриваются гносеологические аспекты информационных процессов идентификации семантических (лексических) и грамматических идентификаторов криминальности. На основе проведенного анализа предлагается метод генерации структурированной машино-читаемой информации на базе неструктурированного текста. Также, во втором разделе рассматриваются специфические особенности извлечения КЗИ из текстов. В частности, рассмотрена технология поиска семантически близких коротких фрагментов текста, имплементация которой позволяет повысить полноту выдачи системы информационного поиска КЗИ.

В третьем разделе рассматривается математическое описание разработанной логико-лингвистической модели извлечения фактов из массивов слабоструктурированных текстов, и показаны особенности реализации данной модели для текстов русского и английского языков. Также, приведен метод формализации грамматических способов выражения факта побуждения к действию в английском языке, использование которого позволит идентифицировать тексты определенной агрессивно-побуждающей направленности.

В четвертом разделе проведен анализ существующих проблем автоматической обработки казахского языка и рассмотрены особенности его формализации. На основе анализа возможностей формализации представления фактической информации в текстах казахского языка, разработана логико-лингвистическая модель Open IE для казахского языка. Использование разработанной модели позволяет извлекать элементы триплета факта из предложений казахского языка на основе отношений грамматических и семантических категорий слов предложения.

В пятом разделе рассматриваются элементы информационной технологии идентификации и анализа криминально-значимой информации в текстовых

корпусах. В частности, приводится описание технологии формирования казахско-русского параллельного корпуса текстов криминальной тематики. Рассматривается метод выравнивания созданного корпуса текстов криминальной тематики, базирующийся на идентификации фактов, и приводится описание разработанного приложения, позволяющего работать с корпусом. Кроме того, в разделе приводится структура и tagset создаваемых корпусов казахского и русского языков.

В целом, использование технологии идентификации криминально значимой информации в многоязычных текстовых массивах, аспекты которой приведенны в монографии, позволят повысить эффективность семантического анализа текстов казахского языка, и естественного языка в целом.

Дальнейшая работа по практической реализации разрабатываемого комплекса моделей, методов и технологий позволит автоматизировать извлечение государственными органами информации, имеющей элементы криминальной значимости из внешних текстовых источников данных. Таких как, социальные сети, электронные СМИ, форумы, блоги и другие электронные ресурсы.

СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ

1. Бондарева Л. В., Борисенко Т. И., Валентей Т. В. и др. Современный терроризм: сущность, причины, модели и механизмы противодействия. – М.: Импульс, 2013. – 252 с.
2. Meloy, J. R., Hoffmann, J., Guldimann, A., & James, D. The role of warning behaviors in threat assessment: An exploration and suggested typology // Behavioral Sciences & the Law. – 2012. - № 30(3). – P. 256–279.
3. Мамырбаев О. Ж., Мухсина К. Ж., Хайрова Н. Ф., Колесник А. С. Лингвистические инструменты выявления криминально окрашенной текстовой информации веб-контента // Вестник казахстанско-британского технического университета. – 2018. – № 3(46). – С. 112-117
4. Cohen, K., Johansson, F., Kaati, L., & Mork, J. C. Detecting linguistic markers for radical violence in social media // Terrorism and Political Violence. – 2014. - № 26(1). – P. 246–256.
5. Cohn, M. A., Mehl, M. R., & Pennebaker, J. W. Linguistic markers of psychological change surrounding // Psychological Science. - 2001. - № 15(10). – P. 687–693.
6. TE-SAT 2012: EU Terrorism situation and trend report // https://www.europol.europa.eu/activities-services/main-reports/te-sat-2012-eu-terrorism-situation-and-trend-report : 25.05.2018
7. Shyam Varan Nath. Crime Pattern Detection Using Data Mining // IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology Workshops. - 2006. - №4. – P. 41–44.
8. Meloy, J. R., Hoffmann, J., Guldimann, A., & James, D. The role of warning behaviors in threat assessment: An exploration and suggested typology // Behavioral Sciences & the Law. – 2012. - № 30(3). - P. 256–279.
9. Meloy, J. R. Approaching and attacking public figures: A contemporary analysis of communications and behavior // Threatening communications and behaviour: Perspectives on the pursuit of public figures. - Washington, DC: The National Academies Press, 2011. - P. 75–101.
10. Meloy, J. R., Mohandie, K., Knoll, J. L., & Hoffmann, J.. The concept of identification in threat assessment // Behavioral Sciences & the Law. – 2015. - № 33(2-3). – P. 213–237.
11. Cohen, K., Johansson, F., Kaati, L., & Mork, J. C.. Detecting linguistic markers for radical violence in social media // Terrorism and Political Violence. – 2014. - № 26(1). – P. 246–256.

12. Paul K. Davis et al. Using Behavioral Indicators to Help Detect Potential Violent Acts: A Review of the Science Base. - RAND Corporation, 2013. – 258 p.
13. Meloy, J. R., Hoffmann, J., Roshdi, K., & Guldimann, A.. Some warning behaviors discriminate between school shooters and other students of concern // Journal of Threat Assessment and Management. – 2014. - № 1(3). - P. 203–211.
14. Hsinchun Chen et al. Crime Data Mining: An Overview and Case Studies // IEEE Computer Society Press Los Alamitos. - CA, USA, 2004. – Vol. 37, Is. 4. - P. 50–56.
15. Bolla, Raja Ashok. Crime pattern detection using online social media // Masters Theses. – 2014. – P. 7321.
16. Хайрова Н., Шаронова Н. Логико-лингвистическая модель извлечения фактов из слабоструктурированной текстовой информациию // International Journal "Information Models and Analyses". – Varna, Bulgaria, 2013. - Vol.2, № 2. – С. 167 – 175.
17. Аверьянова Т. В. Интеграция и дифференциация научных знаний как источники и основы научных методов судебной экспертизы. – М. : Академия МВД РФ, 1994. – 123 с.
18. Ароцкер Л. Е. Об использовании лингвистической статистики для установления автора анонимного текста // Криминалистика и судебная экспертиза.– К. : РИО МООП УССР, 1966. – Вып. 3. – С. 141-151.
19. Астахова А. В. Возможности использования компьютерных экспертных систем при изучении криминалистики // Вестник Алтайской науки. Юриспруденция.– Барнаул: Издательство ААЭП, 2004. – Вып. 1. – С. 154-160.
20. Бєгов Д. Д. Сучасні технології в судовій акустиці (проблеми автоматизації експертних досліджень): дис. … к. ю. н.: 12.00.09. – К. : Національна академія внутрішніх справ, 2002. – 217 с.
21. Белкин Р. С. Криминалистика: проблемы, тенденции, перспективы. Общая и частные теории. – М. : Юридическая литература, 1987. – 272 с.
22. Уголовный кодекс Республики Казахстан от 3 июля 2014 года №226-V с изменениями и дополнениями. - Алматы: LEM (Лем), 2018. – 268 с.
23. Князьков А. С. О критериях значимости криминалистической характеристики преступления // Вестник Томского государственного университета. – Томск : ТГУ, 2007. – № 304. – С. 122-128.
24. Westphal C. Data Mining for Intelligence, Fraud and Criminal Detection. Advanced Analytic & Information Sharing Technologies. – NY.: CRC Press Taylor & Francis Group, 2009. – 440 p.

25. Ермакова Л. М. Методы извлечения информации из текста. // Вестник Пермского Университета. Серия: Математика. Механика. Информатика. - 2012. - Вып.1(9). – С. 77 – 84.
26. Crestan, E., Pantel P. Web-Scale Knowledge Extraction from Semi-Structured Tables // In: WWW '10 Proceedings of the 19th international conference on World wide web. – 2010. – P. 1081 – 1082.
27. Gatterbauer, W.; Bohunsky, P.; Herzog, M.; Krupl, B.; and Pollak, B. Towards Domain-Independent Information Extraction from Web Tables // In Proceedings WWW-07. – Banff, Canada, 2007. - P. 71–80.
28. Wong; Y. W., Widdows, D.; Lokovic T.; Nigam K. Scalable Attribute-Value Extraction from Semi-structured Text // In: 2009 IEEE International Conference on Data Mining Workshops. – 2009. – P. 302 –307.
29. Phillips, W., Riloff, E. Exploiting Strong Syntactic Heuristics and Co-Training to Learn Semantic Lexicons // In: Proceedings of the conference on Empirical Methods in Natural Language Processing (EMNLP) . – 2002. – P. 22 –32.
30. Jones, R., Ghani, R., Mitchell, T., Riloff, E. Active Learning with Multiple View Feature Sets // Workshop on Adaptive Text Extraction and Mining. - 2003. – P. 203 –233.
31. Agichtein, E., Gravano, L. Snowball: Extracting Relations from Large Plaintext Collections // In: Proceedings of the 5th ACM International Conference on Digital Libraries. - San Antonio, Texas, 2000. – P. 85–94.
32. Mooney, R. J., Bunescu R. Mining Knowledge from Text Using Information Extraction // In: Newsletter. ACM SIGKDD Explorations Newsletter - Natural language processing and text mining. – 2005. - Vol.7, issue 1. – P. 3–10.
33. Yahya, M., Whang, E. S., Gupta R., Halevy A. ReNoun: Fact Extraction for Nominal Attributes // In: Proceedings of the Conference on Empirical Methods in Natural Language (EMNLP). – 2014. – P. 325 – 335.
34. Luckicgev, S. Graphical Notations for Rule Modeling. In: Giurca, A., Gašević, D., Taveter, K. Handbook of Research on Emerging Rule-Based Languages and Technologies // Open Solutions and Approaches. - Hershey, NY., 2009. - Vol.1. – P. 76– 98.
35. Ландэ Д. В. Интернетика: Навигация в сложных сетях: модели и алгоритмы: моногр. / Д. В. Ландэ, А. А. Снарский , И. В. Безсуднов — М.: Либроком (Editorial URSS), 2009. 264 с.
36. Sint, R., Schaffert, S., Stroka, S., Ferstl, R.: Combining Unstructured, Fully Structured and Semi-Structured Information in Semantic Wikis. In: Proceedings of the 4th Semantic Wiki WorkShop (SemWiki) at the 6th European Semantic Web Conference, ESWC (2009)

37. Crestan, E., Pantel P. Web-Scale Knowledge Extraction from Semi-Structured Tables. In: WWW '10 Proceedings of the 19th international conference on World wide web, 1081 – 1082 (2010)
38. Gatterbauer, W.; Bohunsky, P.; Herzog, M.; Krupl, B.; and Pollak, B. Towards Domain-Independent Information Extraction from Web Tables. In Proceedings WWW-07. pp. 71–80. Banff, Canada. (2007).
39. Wong; Y. W., Widdows, D.; Lokovic T.; Nigam K. Scalable Attribute-Value Extraction from Semi-structured Text. In: 2009 IEEE International Conference on Data Mining Workshops, 302 –307 (2009)
40. Phillips, W., Riloff, E.: Exploiting Strong Syntactic Heuristics and Co-Training to Learn Semantic Lexicons. In: Proceedings of the conference on Empirical Methods in Natural Language Processing (EMNLP) (2002).
41. Jones, R., Ghani, R., Mitchell, T., Riloff, E.: Active Learning with Multiple View Feature Sets. In: ECML 2003 Workshop on Adaptive Text Extraction and Mining (2003).
42. ARPA. Proceedings of the 3rd Message Undestanding Conference. – 1991.
43. Etzioni, O., Banko, M., Soderland, S., Weld, D. (2008). Open information extraction from the web. Communications of the ACM, Vol. 51 No. 12 (pp. 68-74). New York, NY, USA.
44. Fader, A., Soderland, S., Etzioni, O. (2011). Identifying relations for open information extraction. Proceedings of the conference on empirical methods in natural language processing (pp. 1535-1545). Edinburgh, Scotland, UK.
45. Duc-Thuan Vo, Ebrahim Bagheri. Open information extraction. Encyclopedia with Semantic Computing and Robotic intelligence. – Vol. 1, No. 1 (2016).
46. Shinzato, K., Sekine, S.: Unsupervised extraction of attributes and their values from product description. In: Sixth International Joint Conference on Natural Language Processing, IJCNLP 2013, pp. 1339–1347 (2013).
47. Liyuan, Liu, Xiang, Ren, Qi, Zhu, et al. : Heterogeneous Supervision for Relation Extraction: A Representation Learning Approach. In: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, pp. 46–56 (2017).
48. Wang, Xuan, Zhang, Yu, Chen, Yinyin: A Survey of Truth Discovery in Information Extraction (2018).
49. Gamallo, P., Garcia, M., Fernandez-Lanza, S.: Dependency-based open information extraction. In: Proceedings of the Joint Workshop on Unsupervised and Semi-Supervised Learning in NLP, pp. 10–18 (2012).

50. Akbik, A., Loser, A.: Kraken: N-ary facts in open information extraction. In: Proceedings of the Joint Workshop on Automatic Knowledge Base Construction and Web-scale Knowledge Extraction, pp. 52–56 (2012).
51. Fader, A., Soderland, S., Etzioni, O.: Identifying relations for open information extraction. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing, pp. 1535–1545 (2011).
52. Angeli, G., Premkumar, M.J., Manning, C.D.: Leveraging linguistic structure for open domain information extraction. In: Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics, pp. 344–354 (2015).
53. Gashteovsk, K., Gemulla, R., Del Corro, L.: MinIE: Minimizing Facts in Open Information Extraction. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing, pp. 2630–2640 (2017).
54. Mooney, R. J., Bunescu R.: Mining Knowledge from Text Using Information Extraction. In: Newsletter. ACM SIGKDD Explorations Newsletter - Natural language processing and text mining 7(1), 3–10 (2005)
55. Joakim Nivre et al. Universal Dependencies v1: A Multilingual Treebank Collection. In Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016), Paris, France, May. European Language Resources Association (ELRA)
56. Agichtein, E., Gravano, L. Snowball: Extracting Relations from Large Plaintext Collections. In: Proceedings of the 5th ACM International Conference on Digital Libraries, 85–94, San Antonio, Texas, (2000).
57. Gamallo, P., Garcia, M.: Multilingual Open Information Extraction. In: Portuguese Conference on Artificial Intelligence, pp. 711–722 (2015).
58. Khairova, N., Lewoniewski, W., Wecel, K.: Estimating the Quality of Articles in Russian Wikipedia Using the Logical-Linguistic Model of Fact Extraction. In: Conference proceedings of BIS 2017. Part of the Lecture Notes in Business Information Processing book series, pp. 28-40. Poland, Poznan (2017).
59. Yuen-Hsien Tseng, Lung-Hao Lee, Shu-Yen Lin, Bo-Shun Liao, Mei-Jun Liu, Hsin-Hsi Chen, Oren Etzioni, Anthony Fader. Chinese open relation extraction for knowledge acquisition. In Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics, volume 2: Short Papers, 2014. – p. 12-16.
60. Полевой Н. С. Криминалистическая кибернетика. Теория и практика математизации и автоматизации информационных процессов и систем в криминалистике : учеб. пособ. / Н. С. Полевой. – М. : Издательство МГУ, 1989. – 328 с.
61. Захаров В., Богданова С. Корпусная лингвистика. С.-П., 2013, 148 с.

62. Зубов А.В., Зубова И.И. Информационные технологии в лингвистике. М.: Академия, 2004. — 208 с.
63. Гладкий А. В. Грамматики деревьев: опыт формализации преобразований синтаксических структур естественного языка // Информ. вопросы семиотики, лингвистики и автоматического перевода.— 1971. — Вып. 1. — С. 16—41.
64. C. De Boom, S. V.Canneyt, S. Bohez, T. Demeester, B. Dhoedt, “Learning Semantic Similarity for Very Short Texts,” // Pattern Recognition Letters. – 2016. - Vol. 80. – P. 150–156.
65. H. Wu, M. Zhou, “Synonymous Collocation Extraction Using Translation Information” // in Proc. of the 41st Annu. Meeting on Association for Computational Linguistics. - Stroudsburg, PA, USA, 2003. - Vol.1. - P. 120–127.
66. M. Pasca, P. Dienes, Aligning Needles in a Haystack: Paraphrase Acquisition Across the Web // in Proc. of the Second Int. Joint Conf.: Natural Language Processing. - Korea, 2005. - P. 119–130.
67. R. Barzilay, Kathleen R. McKeown, “Extracting Paraphrases from a Parallel Corpus,”// in Proc. of the 39th Annu. Meeting on Association for Computational Linguistics. - Stroudsburg, PA, USA, 2001. - P. 50–57.
68. L. Han, A. Kashyap, T. Finin, J. Mayfield, J. Weese, “UMBC EBIQUITY-CORE: Semantic Textual Similarity Systems,” // in Proc. of the Second Joint Conf. on Lexical and Computational Semantics. – 2013. - Vol. 1. - P. 44–52.
69. N. Khairova, S. Petrasova, W. Lewoniewski, O. Mamyrbayev, K. Mukhsina. Automatic Extraction of Synonymous Collocation Pairs from a Text Corpus. FedCSIS // Proceedings of the Federated Conf. on Computer Science and Information Systems. – 2018. - Vol. 15. - P. 485–488.
70. DependenSee: A Dependency Parse Visualisation /Visualization Tool // http://chaoticity.com/dependensee-a-dependency-parse-visualisation-tool/: 15.04.18
71. NLTK 3.3 documentation: Source code for nltk.stem.wordnet // http://www.nltk.org/_modules/nltk/stem/wordnet.html :27.06.2018
72. Joakim Nivre. A Multilingual Treebank Collection // In Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC). - Paris, France, 2016. – P. 62-69.
73. Филлмор Ч. Дело о падеже открывается вновь // Новое в зарубежной лингвистике. – М.: Изд. иностр.лит., 1981. - Вып. 10. – С. 496-530.
74. Бондаренко М. Ф. Шабанов-Кушнаренко Ю. П. Теория интеллекта: учебник. - Харьков: Комп. СМИТ, 2007. — 576 с.
75. Бондаренко, М. Ф. Мозгоподобные структуры: Справочное пособие. – Київ : Наукова думка, 2011. – 460 с.

76. Nina Khairova, Orken Mamyrbayev, Kuralay Mukhsina, Anastasiia Kolesnyk. Logical-Linguistic model for multilingual open information extraction // Cogent Engineering, 2020, 7:1, 1714829
77. Khairova, N.F.,Petrasova, S.,Gautam, A.P. The logical-linguistic model of fact extraction from English texts . Information and Software Technologies. Volume 639 of the series Communications in Computer and Information Science, Springer, ISBN: 978-3-319-46253-0, 2016, pp. 625-635. doi> 10.1007/978-3-319-46254-7_51
78. Nina Khairova, Włodzimierz Lewoniewski, Krzysztof Wecel. Estimating the Quality of Articles in Russian Wikipedia Using the Logical-Linguistic Model of Fact Extraction. Conference proceedings. BIS 2017. Part of the Lecture Notes in Business Information Processing book series (LNBIP, volume 288). Pages 28-40. 20th International Conference, BIS 2017, Poznan, Poland, June 28–30, 2017
79. Nina Khairova, Svitlana Petrasova, Orken Mamyrbayev and Kuralay Mukhsina. Open Information Extraction as Additional Source for Kazakh Ontology Generation / ACIIDS 2020, Lecture Notes in Artificial Intelligence 12033, pp.86-96. (Intelligent Information and Database Systems – 12th Asian Conference, ACIIDS 2020, Phuket, Thailand, March 23-26, 2020, Proceedings, Part I)
80. Хайрова Н. Ф. Використання логіко-алгебраїчної моделі семантичних відмінків для семантичного аналізу речення // Зб. наук. пр. Військового ін-ту Київ. нац. ун-ту. – К.: ВІКНУ, 2012. – Вип. № 38. – С. 239– 245.
81. Khairova N., Lewoniewski W., Węcel K., Mamyrbayev O., Mukhsina K. Comparative Analysis of the Informativeness and Encyclopedic Style of the Popular Web Information Sources // Business Information Systems. - Springer, Cham, 2018. - Vol 320. – P. 333-347.
82. Гендина Н. И., Информационно-поисковые тезаурусы: основные виды и области применения // Научные и технические библиотеки. – М.: Государственная публичная научно-техническая библиотека России, 2008 – С. 5-14.
83. T. Pedersen, S. Patwardhan, and J. Michelizzi, “WordNet: Similarity - Measuring the Relatedness of Concepts” // Proceedings of Demonstration Papers at HLT-NAACL. – 2004. - P. 38–41..
84. Современный казахский язык. Фонетика и морфология [Текст] / АН Казах. ССР. Ин-т языкознания ; редкол.: М. Б. Балакаев, Н. А. Баскаков, С. К. Кенесбаев. - Алма-Ата : Изд-во АН Казах. ССР, 1962. - 453 с.
85. Fillmore, C. J. Verbs of Judging: An Exercise in Semantic Description // Studies in Linguistic Semantics. N.Y. etc., 1971. ~ P. 273-290
86. Филлмор Ч Фреймы и семантика понимания. - В кн.: Новое в зарубежной лингвистике, вып. XXIII. - М., 1988

87. Новая философская энциклопедия: в 4 т. / Институт философии РАН; Национальный общественно-научный фонд. — М.: Мысль, 2010. — ISBN 978-5-244-01115-9
88. Rizun, N., Waloszek, W.: Methodology for Text Classification using Manually Created Corpora-based Sentiment Dictionary. In: Proceedings of the 10th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management (IC3K 2018) – Volume 1: KDIR, pages 212-220. Poland (2018)..
89. Gale, W.A., Church, K.W.: A program for aligning sentences in bilingual corpora. In: ACL'93 29th Annual Meeting, vol. 19(1), pp.75–102. USA (NJ) (1993)
90. 82 80 40-2 Мадиева Г. Б. , Уматова Ж. М. Об алматинском корпусе казахского языка. // Вестник КазНУ. Серия филологическая. – 2015. №5 (157). – С. 98 -103
91. Kilgarriff, A., Baisa, V., Bušta, J., Jakubíček, M., Kovář, V., Michelfeit, J., Rychlý, P., VíSuchomel, V.: The Sketch Engine: Ten Years On. In: Lexicography, pp. 7-36. Springer, Berlin, Heidelberg (2014).
92. Chapaev, D., Turapbekov, B.: Building Kazakh language open source corpora using wikipedia resources. In: Suleyman Demirel University Bulletin, pp. 153- 160. Kaskelen, (2018)
93. Makhambetov, O., Makazhanov, A., Yessenbayev, Z., Matkarimov, B., Sabyrgaliyev, I., Sharafudinov, A.: Assembling the Kazakh Language Corpus. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing, pp. 1022-1033. Kazakhstan (2013)
94. Kay, M., Roscheisen, M.: Text translation alignment. Computational Linguistics, 19(1), 121–142 (1993)
95. Fung, P., McKeown, K.: Aligning noisy parallel corpora across language groups: word pair feature matching by dynamic time warping. In: Proceedings of the First Conference of the Association for Machine Translation in the Americas (AMTA-94), pp. 81–88. Columbia, Maryland, USA (1994)
96. Simard, M., Foster G.F., Isabelle, P., Using cognates to align sentences in bilingual corpora. In: Proceedings of the Fourth International conference on theoretical and methodological issues in Machine translation (TMI 1992), pp. 67–81. Montreal, Canada (1992)
97. Varga, D., Halacsy, P., Kornai, A., Nagy, V., Nemeth, L., Tron, V.: Parallel corpora for medium density languages. Amsterdam Studies In: The Theory And History Of Linguistic Science Series 4(292), 247 (2007)

98. Sennrich, P., Volk, M.: Iterative, MT-based Sentence Alignment of Parallel Texts. In: Proceedings of the 18th Nordic Conference of Computational Linguistics (NODALIDA 2011), pp. 175-182. Switzerland (2011)
99. Li, P., Sun, M., Xue, P.:Fast-Champollion: A Fast and Robust Sentence Alignment Algorithm. In: Proceedings of the 23rd International Conference on Computational Linguistics: Posters, pp. 710-718. Beijing, China (2010).
100. Vondricka, P. Aligning parallel texts with InterText. In: Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14). pp. 1875-1879. European Language Resources Association (ELRA) (2014).
101. Zhumanov, Z., Madiyeva, A., Rakhimova, D.: New Kazakh parallel text corpora with online access. In: Conference on Computational Collective Intelligence Technologies and Applications, pp.501-508. Almaty, Kazakhstan (2017).
102. Rakhimova, D., Zhumanov, Z.:.Complex Technology of Machine Translation Resources Extension for the Kazakh Language. Advanced Topics in Intelligent Information and Database Systems. Springer International Publishing, Almaty, Kazakhstan. (2017).
103. Grabar, N., Kanishcheva, O., Hamon, T.: Multilingual aligned corpus with Ukrainian as the target language. In: SLAVICORP, pp. 53-57. Prague (2018).
104. Harme, J.: Last year but not yesterday? Explaining differences in the locations of Finnish and Russian time adverbials using comparable corpora. In: SLAVICORP, pp. 60-63. Prague (2018).
105. Smith, J. R., Quirk, C., Toutanova, K.: Extracting Parallel Sentences from Comparable Corpora using Document Level Alignment. In: Proceedings of the Human language Technologies/North American Assosiation for Computational Linguistics, pp. 403-411 (2010)
106. Lewoniewski, W., Węcel, K., Abramowicz, W. Quality and importance of Wikipedia articles in different languages. In International Conference on Information and Software Technologies, pp. 613 – 624. Poznan, Poland (2016)
107. Rosen, A.: In search of the best method for sentence alignment in parallel texts. In Computer treatment of Slavic and East European languages. Third international seminar, pp. 174-185. Bratislava, Slovakia (2005)
108. Khairova, N., Kolesnyk, A., Mamyrbayev, O., Mukhsina, K. The aligned Kazakh-Russian parallel corpus focused on the criminal theme .//CEUR Workshop Proceedings. – 2019, p 116-125
109. Хаирова Н., Колесник А., Мамырбаев О., Мухсина К. Выровненный казахско-русский параллельный корпус, ориентированный на криминальную тематику / Вестник Алматинского университета энергетики и связи № 1 (48) 2020. – с.84-92.

ПРИЛОЖЕНИЕ А

Классификация основных словообразовательных глагольных аффиксов казахского языка

Суффиксы	Описание	Примеры
-қыла/ -кіле, ғыла/-гіле, ңқыра/-ңкіре, -іңкіре ңғыра/-ңгіре, мсыра/- мсіре, ымсыра/- імсіре (и их фонетические варианты), ылда/-ілде, ырла/-ірде	Аффиксы глагольного вида, прибавляемые к глагольным основам	Ес- ңгіре-у, са- ңғыра-у
-ла /-ле (-да /-де, -та /-те во всех фонетических вариантах), -лан/-лен (дан /-ден, -тан /-тен, лат/-лет), -лас/-лес, -ландыр/-лендір, -ластыр/-лестір (во всех фонетических вариантах)	Наиболее продуктивные аффиксы Основные глаголообразующие суффиксы от других частей речи	Көңілсіз-де-н-у Әлсіз-де-н-у
-ла /-ле, -а/-е, -лық/-лік,- ық/-ік, -шы/-ші (во всех фонетических вариантах –іг, -ліг, -тығ, -діг, -ре), -ыра, -іре, -ыла, -іла	Именные и отглагольные	Үр-ле-у, тісте-ле-у, түй-ре-у
ай/-ей, й, ар/-ер, р, ра/-ре, ыр/-ір,		Көг-ер-т-у, үлк-ей-т-у , кішір-ей-т-у
-зы, -азы, -ма, -бе, -ды/-ді, -ы/-і, ты/-ті, лы/-лі, ра/-ре,-шы/-ші, ын, ін, ал, ел	Мало продуктивные аффиксы	
мала /-меле, -палапеле, -бала/-беле, -ақта/-екте, -дала/-ала		
сыра/-сіре, мсыра/-мсіре, усыра/-усіре, жыра/-жіре, аңғыра/-еңгіре, ңра/-ңре ыра/-сіре, аура/-еуре,		
сы/-сі , сын/-сін, са/-се, сан/-сен	*н*-формальный показатель возвратного залога	

-лық/-лік ық/-ік дық/-дік тық/-тік,-тығ – лығ -ығ іғ		Соқ-тық, соқ-тығ-у
қар/-кер қыр/-кір ғар/-гер ғыр/-гір қа/-ке ға/-ге қа/-ке, қан/-кен ан/-ен, ғал/-қал	Добавляются аффиксы *–л* и *–н* возвратного залога	
ырқа /-ірке лқа ырқан /-іркен, -т, ыт/-іт, бы, бі, пы, пі,		
ғы/-гі ғыт/-гіт ға/-ге, қы/-кі	Аффикс – *т* принудительного залога	
сый, си, ырай, ірей, ыс, іс, жи, ши, ди, ти, ми, пи, би, қи ки ыс іс	Образные глаголы	

ПРИЛОЖЕНИЕ Б

Основные словообразовательные залоговые суффиксы глаголов казахского языка

Суффикс	Залог	Пример
–н, -ын, -ін	возвратный	*жу-ын-у*, *ора-н-у* *көр-ін*, *көр-ін-ді*
-лан, -лен, -дан, -ден, -тан, -тең	возвратный	*намыс-тан-у*, *шат-тан-у*
-сын, -сін, -қан, -кен	Возвратный	
-л, -ыл, -іл	возвратный, страдателный	*бу-ыл-у*, *түй-іл-у*
-лык, -лік, -дық, -дік, -тік, -ық, -ік, -лығ, -іг, -ліг	возвратный	*Бу-лығ-у*, *іл-іг-у*
-с, -лас, -лес	Возвратный	*орна-лас-у, қате-лес-у*
-л, -н	Страдательный	*жина-л-ды*
-ыл, -іл, -ын, -ін	Страдательный	*Оқ-ыл-ды*
-лын, -лін, -нып, -ніл	Страдательный	*же-лін-у*, *қолда-нып-у*, *пайдала-нып-у*
-с, ыс, -іс	совместно-взаимный	
-лас, -лес, -дас, -дес, -тас, -тес	совместно-взаимный	*мұн-дас-у, сыр-лас-у, қас-тас-у*
-ылыс, -ныс, -ыныс, -ініс -тығыс и др.	совместно-взаимный, возвратный	*сапыр-ыл-ыс-у*, *байла-н-ыс-у*), *ұғ-ын-ыс-у, соқ-тығ-ыс-у*
-стыр, -стір, -ластыр, -лестір	совместно-взаимный, прину-	*жара-с-тыр-у*, *таны-с-тыр-у*

	дительный	
-ландыр, -тендір, -лендір	возвратный и понудительный	*Индустрия-лан-дыр, коллектив-тен-дір, ірі-лендір*
-дастыр	принудительный	*Колхоз-дас, колхоз-дастыр, колхоз-дас-тыр-ыл*
-т, -ыт, -іт, -дыр, -дір, -тыр, -ғыз, -гіз, -қыз, -кіз, -ар, -ер, -ыр, -ір, -қар, -кер, -ғыр, -дар, -сет	побудительный	
-ғыздыр, -гіздір, -дырғыз, -діргіз -ындыр, -індір, -ландыр, -лендір, -тандыр, -дендір и	сложные аффиксы к возвратному залогу + понудительный	
-ылыс, -лыс, -ініс, -ліс , -ығыс, -тығс, -тығыс, -лікіс, -ықыс, -ікіс, -ныс, -ніс, -ыныс, -ініс	К возвратному залогу + совместно-взаимный залог	
-лін, -лын, -нілл	К возвратному + страдательный	
-тырыл, дырыл, -ғызыл, -сетіл -ттыр	Понудительный + срадательный Понудительный + совместно взаимный	
-арыс, -іріс -стыр, -стір, -ластыр, -лестір	понудительный Совместно-взаимный+ понудительный	