На правах рукописи

ЗИНОВЬЕВА Анастасия Юрьевна

МОДЕЛЬ МНОГОЯЗЫЧНОГО ИНТЕЛЛЕКТУАЛЬНОГО КОНТЕНТ-АНАЛИЗА (НА МАТЕРИАЛЕ АНГЛО-, ФРАНКО- И РУССКОЯЗЫЧНЫХ НОВОСТНЫХ СООБЩЕНИЙ О ТЕРРОРИСТИЧЕСКОЙ ДЕЯТЕЛЬНОСТИ)

Специальность 10.02.21 – Прикладная и математическая лингвистика

АВТОРЕФЕРАТ
диссертации на соискание ученой степени
кандидата филологических наук

Челябинск
2022

Работа выполнена в федеральном государственном автономном образовательном учреждении высшего образования «Южно-Уральский государственный университет (национальный исследовательский университет)» на кафедре лингвистики и перевода Института лингвистики и международных коммуникаций.

Научный руководитель:	**Шереметьева Светлана Олеговна** доктор филологических наук, доцент, профессор кафедры лингвистики и перевода ФГАОУ ВО «Южно-Уральский государственный университет (национальный исследовательский университет)»
Официальные оппоненты:	**Андреев Вадим Сергеевич** доктор филологических наук, профессор, заведующий кафедрой иностранных языков ФГБОУ ВО «Смоленский государственный университет»
	Захаров Виктор Павлович кандидат филологических наук, доцент, доцент кафедры математической лингвистики ФГБОУ ВО «Санкт-Петербургский государственный университет»
Ведущая организация:	**ФГБУН «Институт проблем передачи информации им. А. А. Харкевича Российской академии наук»**

Защита диссертации состоится «21» июня 2022 г. в 12:00 на заседании диссертационного совета Д 212.274.15 по защите диссертаций на соискание ученой степени кандидата наук, на соискание ученой степени доктора наук при ФГАОУ ВО «Тюменский государственный университет» по адресу: ул. Ленина, 23, ауд. 516.

С диссертацией и авторефератом можно ознакомиться в Библиотечно-музейном комплексе ФГАОУ ВО «Тюменский государственный университет» по адресу: 625003, г. Тюмень, ул. Семакова, 18, а также на официальном сайте ТюмГУ: diss.utmn.ru.

Автореферат разослан «20» апреля 2022 г.

Ученый секретарь
диссертационного совета,
кандидат филологических наук, доцент *Д. В. Шапочкин*

ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

Реферируемая диссертация посвящена методологическим и практическим аспектам моделирования интеллектуального контент-анализа (ИКА) неструктурированной текстовой информации на примере новостных сообщений предметной области (ПО) «Терроризм» на английском, французском и русском языках.

Актуальность исследования обусловлена тем, что наблюдаемый в настоящее время постоянный рост объема неструктурированной информации на различных языках требует систем автоматизации ее анализа, среди которых первостепенное значение имеют системы интеллектуального контент-анализа, которые могли бы обеспечить высокое качество извлечения из текстов соответствующего информационному запросу пользователя контента. Создание таких систем невозможно без решения проблем моделирования процессов интеллектуального контент-анализа, которые, несмотря на то что в нашей стране и за рубежом исследованиям в этой области уделяется достаточно серьезное внимание, еще не решены.

Настоящее исследование актуально и в том отношении, что в центре его внимания находится многоязычность, поскольку многоязычная модель интеллектуального контент-анализа может быть использована повторно для обработки текстов на различных языках с минимизацией времени, затрат и усилий разработчиков. Актуальность исследования определяется также тем, что оно охватывает не только английский и французский языки, на которые ориентированы многие исследования в данной области, но и русский, интеллектуальному контент-анализу которого до настоящего времени еще не уделяется достаточно внимания, что отражено в значительно меньшем количестве работ по теме.

Разрабатываемая модель многоязычного ИКА ориентирована на ПО «Терроризм», что также обеспечивает актуальность настоящей работы, поскольку борьба с терроризмом является одной из главных задач современности, и контент-анализ текстов данной предметной области имеет большое значение для аналитической деятельности в сфере политического прогнозирования и контртерроризма.

Степень разработанности проблемы. Исследования по разработке моделей и систем ИКА можно разделить на два крупных класса: исследования, ориентированные на обработку текстов любой тематики, в результате чего страдает глубина извлечения информации из узконаправленных текстов; и исследования, посвященные решению точечных задач ИКА в конкретной предметной области и на конкретном языке, отсюда невозможность применения таких систем контент-анализа к текстам других предметных областей и на других языках. На данный момент наиболее реалистичным и поэтому наиболее разрабатываемым направлением исследований

является построение моделей ИКА для конкретных областей на материале конкретных языков. Исследований по разработке многоязычных моделей интеллектуального контент-анализа значительно меньше, и подавляющее большинство также ориентировано на конкретный подъязык. Ключевым этапом интеллектуального контент-анализа является категоризация текстовых единиц и формализация экспертных знаний. Наиболее распространенным и перспективным подходом к реализации этого этапа является подход на основе онтологий, позволяющий проводить глубокий семантический анализ. Несмотря на большое количество исследований по рассматриваемой теме, проблема моделирования интеллектуального контент-анализа не может считаться решенной в полной мере ввиду комплексности, неоднозначности естественного языка и сложности его формализации.

Цель исследования – разработать модель многоязычного интеллектуального контент-анализа на основе онтологических знаний на примере англо-, франко- и русскоязычных новостных сообщений предметной области «Терроризм» с возможностью дальнейшего ее применения к другим языкам и предметным областям.

Поставленной целью продиктованы следующие **задачи исследования**:

1) уточнить понятие интеллектуального контент-анализа и наиболее перспективные пути его реализации;

2) разработать методику создания модели многоязычного интеллектуального контент-анализа ограниченной предметной области;

3) выявить ограничения и закономерности подъязыка новостных сообщений предметной области «Терроризм» на уровне структуры релевантности и суперструктуры, морфосинтаксическом и лексико-семантическом уровнях на материале английского, французского и русского языков;

4) построить онтологическую базу знаний модели многоязычного интеллектуального контент-анализа на основе данных анализа подъязыка;

5) разработать алгоритм интеллектуального контент-анализа;

6) апробировать модель на материале новых текстов английского, французского и русского языков, не относящихся к тренировочному корпусу.

Объект исследования – подъязык англо-, франко- и русскоязычных новостных сообщений предметной области «Терроризм»; **предмет** – моделирование концептуальной структуры рассматриваемого подъязыка и формализация извлечения проблемно-ориентированного контента из текстов предметной области.

В качестве **материала** исследования использованы три псевдопараллельных корпуса новостных сообщений о терроризме за 2014–2020 гг. равного объема общим объемом более 600 000 словоупотреблений, для доработки модели – корпусы текстов за 2019–2020 гг. объемом 20–40 тыс. с. у. Все тексты получены методом

целевой выборки из интернет-источников по ключевым словам. В качестве источников выбраны сайты англо-, франко- и русскоязычных информационных агентств и газет, например BBC News, Bloomberg, Reuters, 20 Minutes, AFP, L'Express, Le Figaro, «Вести», «Лента», «ТАСС» и др. На отдельных этапах работы использовались случайные выборки из указанных корпусов, концептуально размеченные сотрудниками НОЦ «Лингво-инновационные технологии» ЮУрГУ, в том числе автором диссертации. Дополнительным материалом для проверки работы модели послужили новые корпусы текстов на трех языках разного объема за 2019–2020 гг.

Материал исследования представлен корпусами текстов на трех языках как по лингвистическим, так и по экстралингвистическим причинам. Во-первых, в настоящее время английский, французский и русский являются одними из самых распространенных языков в мире, официальными языками ООН. Во-вторых, эти языки принадлежат к разным языковым группам и, как следствие, имеют разные характеристики. В частности, различия наблюдаются на морфосинтаксическом уровне: английский язык является в большей степени аналитическим, русский язык – преимущественно синтетическим, французский язык занимает промежуточное положение, сохраняя черты синтетизма в своей письменной форме. Наконец, использование в исследовании трех языков обусловлено необходимостью создания независимой от конкретного естественного языка онтологии и, следовательно, модели интеллектуального контент-анализа, что позволит в будущем с минимальными усилиями разработчиков распространить ее на другие языки.

Теоретико-методологическую базу исследования составили труды отечественных и зарубежных ученых в области *лингвистического моделирования* (Ю. Д. Апресян, И. И. Ревзин, С. О. Шереметьева и др.), *корпусной лингвистики* (В. П. Захаров), *контент-анализа* (А. А. Бызов, Ю. П. Воронов, В. И. Шалак, B. Berelson, C. Cioffi-Revilla, L. A. Kort-Butler, K. Krippendorff и др.), в том числе *интеллектуального контент-анализа* (С. О. Шереметьева, A. C. M. Fong, Ph. Mayring, L. Weber и др.), *неоднозначности* (Л. Н. Иорданская, В. Н. Поляков, Е. В. Рахилина и др.), *построению одноязычных и многоязычных баз знаний для автоматической обработки текстов* (Е. Б. Козеренко, С. О. Шереметьева), *онтологической семантике* (K. Mahesh, S. Nirenburg, B. Onyshkevych, V. Raskin), *построению онтологий* (И. М. Богуславский, Н. В. Лукашевич, О. А. Митрофанова, T. Gruber, N. Guarino. O. Lassila, D. McGuinness и др.), *извлечению информации на основе онтологических знаний* (С. О. Шереметьева, A. Konys, E. Iosif), а также прикладные исследования по *разработке онтологий терроризма* (С. Л. Мишланова, A. Mannes, M. D. Turner, L. Wendelberg и др.), работы в области исследования *подъязыков* (Н. Д. Андреев, С. О. Шереметьева, Z. Harris, J. Lehrberger и др.), в том числе

подъязыка новостных текстов (Т. Г. Добросклонская, M. K. Ozguven, T. A. van Dijk и др.). В целях автоматизации исследования применялись экстрактор лексики LanaKey, разработанный С. О. Шереметьевой, и корпус-менеджер L. Anthony.

В работе использованы следующие **методы исследования**: метод моделирования, методы целевой и случайной выборки для создания исходных и тестовых корпусов, статистический и сопоставительный анализ, компонентный анализ, дистрибутивный анализ, метод оппозиций, контекстный метод.

Основные положения, выносимые на защиту:

1. Интеллектуальный контент-анализ представляет собой извлечение из корпусов текстов соответствующего информационному запросу пользователя контента, его интерпретацию и представление в удобной для пользователя форме; включает качественный этап категоризации текстовых единиц и количественный этап подсчета выделенных категорий.

2. Создание модели многоязычного интеллектуального контент-анализа текстов ограниченной предметной области для обеспечения высокого качества результатов осуществляется на основе онтологической базы знаний.

3. Построение онтологической базы знаний осуществляется на основе анализа подъязыка соответствующей ПО на материале многоязычных текстов и предусматривает последовательное выполнение трех стадий разработки, которые включают в себя циклические этапы и подэтапы: анализ лингвистического материала рассматриваемой ПО, создание исходной модели многоязычного ИКА на основе результатов анализа лингвистического материала и совершенствование созданной модели с использованием дополнительного текстового материала ПО.

4. Подъязык новостных сообщений предметной области «Терроризм» ограничен, его ограничения имеют двойственную природу, что обусловлено жанром новостных сообщений и предметной областью «Терроризм».

5. База знаний модели многоязычного интеллектуального контент-анализа состоит из концептуальных (онтология предметной области), эпизодических (база экземпляров) и лингвистических (онтолексиконы, ономастиконы) знаний, правил онтологического анализа и логического вывода, динамических концептуально-лексических фреймов для представления результатов контент-анализа.

6. Алгоритм модели многоязычного интеллектуального контент-анализа включает этап постановки задачи контент-анализа и набор процедур для ее последовательного решения на основе сочетания автоматических и ручных методов, а именно отбор текстового материала, онтологический анализ, определение релевантных концептуальных тегов, заполнение концептуально-лексических фреймов, представление результатов интеллектуального контент-анализа.

Научная новизна работы заключается в том, что впервые разработаны методологические и практические аспекты моделирования интеллектуального контент-анализа неструктурированной информации на примере новостных сообщений ПО «Терроризм» на английском, французском и русском языках, что обусловлено применением совокупности современных лингвистических и компьютерных методов к анализу языкового материала; выявлением и уточнением ограничений и закономерностей подъязыка новостных сообщений ПО «Терроризм» на уровне структуры релевантности и суперструктуры, морфосинтаксическом и лексико-семантическом уровнях; исследованием проблемы концептуальной неоднозначности; разработкой предметно-ориентированной модели контент-анализа с акцентом на интеллектуальность и многоязычность; построением многоязычной онтологии ПО «Терроризм»; созданием онтолексиконов для английского, французского и русского языков; разработкой правил онтологического анализа, логического вывода и формирования динамических концептуально-лексических фреймов для представления результатов контент-анализа; а также разработкой алгоритма ИКА, который позволяет извлекать не только явно выраженную, но и имплицитную информацию.

Теоретическая значимость исследования состоит в развитии методологических аспектов многоязычного интеллектуального контент-анализа, в том числе в уточнении определения интеллектуального контент-анализа; ввиду чего результаты исследования способны внести вклад как в данное направление научных исследований, так и в смежные с ним направления: лексикографию, лексикологию и терминоведение, компьютерную и корпусную лингвистику, информационный поиск, автоматическое реферирование и аннотирование, машинный перевод.

Практическая ценность результатов работы заключается в возможности создания системы многоязычного интеллектуального контент-анализа на базе разработанной модели. Потенциальная область применения модели не ограничивается новостными интернет-сообщениями о террористической деятельности; лежащие в ее основе принципы могут быть экстраполированы на другие языки и предметные области. Отдельные положения могут быть включены в курсы по прикладной лингвистике, функциональной стилистике, общественно-политическому переводу.

Достоверность и научная обоснованность результатов исследования обеспечивается опорой на авторитетные положения работ отечественных и зарубежных ученых, применением современных методов исследования, адекватных его цели и задачам; использованием в качестве фактического материала англо-, франко- и русскоязычных корпусов текстов значительного объема, согласованностью теоретических выводов с результатами практического исследования. Научные положения и

выводы, сформулированные в работе, подкреплены фактическими данными, представленными в приведенных в диссертации таблицах и рисунках.

Апробация работы. Основные положения исследования обсуждались на заседаниях кафедры лингвистики и перевода ФГАОУ ВО «Южно-Уральский государственный университет (национальный исследовательский университет)», были представлены на конференции аспирантов и докторантов ЮУрГУ (Челябинск, 2019–2021), 64-й Всероссийской научной конференции МФТИ (Москва, 2021), международных конференциях «Цифровые трансформации и глобальное общество» (Санкт-Петербург, 2018), «Интерактивные системы: проблемы человеко-компьютерного взаимодействия» (Ульяновск, 2019), «Интернет и современное общество» (Санкт-Петербург, 2020), «Язык. Культура. Перевод: межкультурная коммуникация в цифровую эпоху» (Одинцово, 2022).

По теме исследования опубликовано 10 печатных работ в рецензируемых журналах и сборниках научных трудов, из них 4 статьи в журналах, рекомендованных ВАК, 3 статьи в изданиях, индексируемых в базе данных Scopus.

Структура работы. Диссертация состоит из введения, трех глав, заключения, списка литературы, включающего 175 наименований, из них 109 – на иностранных языках, и шести приложений.

ОСНОВНОЕ СОДЕРЖАНИЕ РАБОТЫ

Во **введении** обоснована актуальность и новизна работы, описана степень разработанности проблемы, поставлена цель, обозначены задачи, установлены объект и предмет исследования, представлена теоретико-методологическая база исследования, перечислены методы, приводятся основные результаты исследования, отраженные в положениях на защиту, и сведения об апробации и структуре работы.

Первая глава «Интеллектуальный контент-анализ как объект моделирования» посвящена общим методологическим аспектам моделирования многоязычного интеллектуального контент-анализа.

Моделирование – это создание модели объекта познания для изучения его свойств и характеристик. Несмотря на частое применение моделирования в лингвистических исследованиях, термин «модель» трудно определим (Лосев, 2004; Ревзин, 1962), ввиду чего ученые предпочитают характеризовать модель через ее свойства, отмечая, в частности, ее упрощенный характер (Ревзин, 1977: 62; Комарова, 2014: 314). Особой спецификой обладают **лингвистические модели**, под которыми понимается «искусственно созданное лингвистом реальное или мысленное устройство, воспроизводящее или имитирующее своим поведением […] поведение

какого-либо другого устройства […] в лингвистических целях» (Лингвист. энцикл. словарь, 1990: 304). Среди лингвистических моделей выделяют модели речевой деятельности, модели лингвистического исследования и метамодели (Апресян, 1966: 99). В последние десятилетия набирает популярность построение **многоязычных лингвистических моделей** (Козеренко, 1995; Sheremetyeva, 2000, 2017, 2018, 2020). Ориентация на многоязычные модели позволяет экономить время, затраты и исследовательские усилия при создании новых ресурсов (Sheremetyeva, 2017).

Контент-анализ (КА) – это широко используемый метод анализа содержания текста в соответствии с определенной информационной задачей. Существующие подходы к определению данного термина в основном различаются взглядами на его **квантитативность**: в работах (Berelson, 1952; Lasswell, 1949) она считается обязательным атрибутом КА в отличие от работ (Holsti, 1969; Krippendorff, 1980).

Типологизация КА проводится на основе ряда параметров: **количественная мера** (Пашинян, 2012), **объект** (Kort-Butler, 2016), **предмет** (Rich и др., 2018), **выбор элементов кодирования** (Janis, 1965), **степень учета контекста** (Altheide, 1987), **степень автоматизации** (Олейник, 2021; Krippendorff, 2018). На практике дифференциация контент-анализа производится в основном в рамках **дихотомии «количественный – качественный»**: количественный контент-анализ предполагает подсчет слов и словосочетаний в тексте; качественный, напротив, основан на категоризации текстовых единиц и кодировании их метками категорий. Эта дихотомия часто подвергается обоснованной критике. Например, авторы (Gauch, 2005; Mayring, 2014) определяют качественный контент-анализ как подход смешанных методов и выступают за применение к ним общих исследовательских критериев. В частности, отмечается, что качественный контент-анализ включает присвоение категорий текстовым элементам (кодирование) как качественный этап, а обработку множества текстов и анализ частот категорий – как количественный этап.

В литературе отсутствует исчерпывающее определение интеллектуального КА, однако отмечается, что он представляет собой попытку «достичь семантического понимания контекста, в котором встречаются определенные ключевые слова» (Fong, 2006: 172), а его функцией является категоризация содержимого текста и последующее принятие решений на основе имеющихся знаний (Лазарев, 2013). Изучение немногочисленных определений ИКА и практических исследований по данной теме позволяет сформулировать следующее определение: **интеллектуальный контент-анализ** – это извлечение из неструктурированных, возможно, разноязычных текстов ограниченной ПО контента, соответствующего информационному запросу пользователя, числовая обработка такого контента, его интерпретация и представление в удобной для пользователя форме.

Исследования, выполненные в русле интеллектуального контент-анализа, могут быть разделены на два крупных класса: исследования, ориентированные на обработку текстов любой тематики, и исследования, посвященные решению точечных задач ИКА ПО на конкретном языке. Построение моделей и систем ИКА для конкретной ПО на данный момент наиболее реализуемо и поэтому представляет собой наиболее разрабатываемое направление исследований. Большинство работ нацелено на обработку одного языка, как правило, английского или иного национального (не русского) языка, например (Streibel, 2010; Weber, 2013); исследований в области разработки многоязычных моделей и систем значительно меньше, например (Chaves, 2010; Flaounas и др., 2010) и др. Среди исследований, ориентированных, среди прочего, на русский язык, представлены работы (Добров и др., 2015; Лукашевич и др., 2018; Efimenko, 2004; Manicheva и др., 2012).

ИКА предполагает ряд процедур, не все из которых реализуются в отдельных исследованиях в полном объеме: сбор текстов анализируемой предметной области; предварительную обработку текстов (устранение опечаток, извлечение релевантной лексики); формализацию экспертных знаний; обработку размеченных текстов; интерпретацию извлеченных данных и формирование аналитического отчета.

Формализация экспертных знаний – один из самых трудоемких этапов, куда входят такие подэтапы, как определение категорий (категоризация) и единиц анализа, построение баз знаний и разметка текстов на основе выделенных категорий. Категоризация и разметка могут выполняться в следующем порядке: **дескриптивным способом**, когда категории, релевантные поставленной задаче, определяются в ходе анализа текста (Воронов, 2005); **прескриптивным**, когда категории заданы заранее (Бызов, 2019; Kort-Butler, 2016); **суммативным**, когда часть категорий задана заранее, остальные же добавляются в ходе анализа (Hsieh, 2005).

Знания могут быть формализованы разными способами: от самых простых (например, облако ключевых слов) до более глубоких с упором на семантические, в частности **онтологические**, базы знаний. Обработка неструктурированной информации с помощью онтологий является перспективным вариантом реализации ИКА, поскольку онтологический анализ позволяет аннотировать релевантные для КА элементы текста семантическими тегами и таким образом формализовать семантические признаки лексем, обеспечивая их измеримость, что в сочетании с традиционными поверхностно-статистическими характеристиками способно повысить качество извлекаемой информации (Konys, 2015; Iosif, 2012; Sheremetyeva, 2020; Streibel, 2010) и, соответственно, интеллектуального контент-анализа.

Ресурсы интеллектуального контент-анализа могут быть разделены на статические и динамические. Статические ресурсы – это ресурсы, не изменяемые

во времени и содержащие данные в той или иной форме (корпусы текстов, лексиконы, онтологии); динамические ресурсы – инструменты, обеспечивающие создание и последующую обработку новых данных (инструменты разметки текста, морфологические анализаторы и генераторы) (Witt и др., 2009).

Особый интерес представляет **онтология** как основной ресурс модели интеллектуального контент-анализа. В современной науке данный термин представлен в нескольких значениях: в частности, онтология рассматривается как подраздел философии, учение о бытии (Nickles и др., 2007) и как артефакт, описывающий значения и взаимосвязи элементов некой системы (Gruber, 1993; Nirenburg, 2004).

Согласно классическому определению онтологии во втором значении, **онтология** – это «явная спецификация концептуализации», где «концептуализация» – «упрощенная модель мира, создаваемая нами для определенных целей» (Gruber, 1993: 199). В прикладной лингвистике концептуализация обычно интерпретируется как представление знаний об объектах и явлениях действительности через описание множества взаимосвязанных концептов, ориентированное на решение конкретной задачи, и формализуется в виде графа, в узлах которого находятся концепты, а дуги представляют отношения между ними (Nirenburg, 2004).

При этом интерпретации этого понятия в современных лингвистических исследованиях значительно различаются. Наиболее существенное различие отражено в отношении исследователей **к степени зависимости онтологии от конкретного естественного языка**. Одни ученые считают онтологию независимой от лексики и грамматики естественного языка (Nirenburg и др., 1996); такими онтологиями являются, например, MikroKosmos (Nirenburg, 2004); Suggested Upper Merged Ontology (Niles, 2003); Basic Formal Ontology (Arp, 2015). Другие ученые полагают, что независимость от естественного языка не является определяющим фактором для отнесения ресурса к онтологиям (Brewster и др., 2005). Наиболее зависимым от конкретного языка ресурсом, который часто относят к онтологическим, является тезаурус WordNet (Miller и др., 1990). Промежуточное положение занимают такие ресурсы, как Cyc (Foxvog, 2010) и модель «Смысл↔Текст» (Мельчук, 1999). Существует концепция, по которой онтологии могут быть распределены по спектру зависимости от естественного языка (Onyshkevych, 1997). В современной отечественной лингвистике наиболее многообещающее исследование по созданию независимого от естественного языка онтологического ресурса для задач анализа текста и машинного перевода связано с разработкой системы ЭТАП (Богуславский, 2012).

Онтологии существенно различаются и по ряду иных параметров, например, **степени формальности** (Gómez-Pérez, 2010; Lassila, 2001; Uschold, 1996), **содержанию** (Митрофанова, 2015; Hois, 2013; Mizoguchi, 1995; Swartout и др., 1997),

наличию экземпляров (Коваль, 2007), **коммуникативным возможностям** (Рубашкин, 2013), **способам разработки и исследования** (Митрофанова, 2015) и др.

Из всего многообразия типов онтологий, как правило, выделяют онтологии верхнего уровня и предметной области. **Онтологии верхнего уровня** включают в себя общие знания о мире (например, MikroKosmos, SUMO и BFO). **Онтологии предметной области** содержат знания о конкретных предметных областях.

Далее, выделяют **лингвистические онтологии**, которые в свою очередь имеют следующие интерпретации: **модели представления лингвистических знаний** (Соколова, 2008; Farrar, 2003); **онтологии для обработки естественного языка** (Mahesh, 1996; Nirenburg, 2004); **лексические онтологии**, созданные по wordnet-модели (Данченков, 2010; Лукашевич, 2011; Madsen и др., 2010).

Многоязычность онтологий понимается в двух значениях: как возможность применения одной онтологии к обработке текстов на различных языках вне зависимости от того, какой язык используется для обозначения названий концептов (Nirenburg и др., 1996), а также как адаптация названий концептов для пользователей, являющихся носителями различных языков (Chaves, 2010; Espinoza, 2008; Montiel-Ponsoda и др., 2008). Онтологии, не зависящие от конкретного естественного языка, являются по определению многоязычными в первом значении. В рамках направления, в котором онтологии трактуются как зависимые от языка ресурсы, вопрос многоязычности онтологий рассматривается иначе. Например, предлагаются универсальные инструменты для полуавтоматического создания онтологий на разных языках (Alatrish, 2014) или разрабатываются методики объединения онтологий, изначально созданных для разных языков (Embley и др., 2011).

В результате анализа работ, посвященным типологизации онтологий и подходам к их разработке, определено, что для решения поставленной задачи моделирования многоязычного интеллектуального контент-анализа онтология должна быть независимой от конкретного естественного языка и многоязычной (т. е. ориентированной на обработку нескольких языков). Кроме того, онтология должна описывать предметную область «Терроризм», должна быть связана с какой-либо существующей онтологией верхнего уровня и представлена в ее формализме.

В результате анализа литературы по теме исследования, проведенного в первой главе, определена **методика построения модели** многоязычного интеллектуального контента-анализа, этапы которой подробно изложены в следующих главах.

Вторая глава «Корпусный анализ подъязыка новостных сообщений предметной области „Терроризм“ (на материале русского, английского и французского языков)» посвящена анализу подъязыка предметной области на уровне структуры релевантности и суперструктуры, морфосинтаксическом и

лексико-семантическом уровнях, сравнительно-сопоставительному анализу характеристик русского, английского и французского корпусов предметной области.

Все корпусы последовательно проанализированы по трехэтапной методике с элементами автоматической и ручной обработки материала: 1) **анализ структуры релевантности и суперструктуры** (van Dijk, 1988) каждого новостного сообщения; 2) **морфосинтаксический анализ**, в ходе которого из корпуса автоматически с помощью экстрактора LanaKey (Sheremetyeva, 2012) были извлечены лексические группы длиной от 1 до 4 компонентов (программное ограничение экстрактора), затем с помощью функции поиска из корпуса были извлечены релевантные для предметной области лексические группы длиной от 5 до 10 компонентов; проведена лемматизация, подсчет частоты встречаемости лексических групп, анализ морфосинтаксических свойств лексики, выявлены лексико-грамматические корреляции; 3) **лексико-семантический анализ** и распределение выделенных лексических единиц по концептуальным классам, релевантным для предметной области, на основании общих сем с помощью прескриптивно-дескриптивной методики.

Проведенный корпусный анализ демонстрирует **двойственную природу ограничений** исследуемого подъязыка: часть ограничений обусловлена жанром новостных сообщений, часть – предметной областью «Терроризм».

Ограничения, обусловленные жанром, проявляются в структуре текста. Рассмотренные новостные сообщения в большинстве случаев вне зависимости от языка имеют **структуру релевантности «перевернутая пирамида»** и обладают характерной **суперструктурой**, включающей информационные части «Заголовок», «Вводка», «Основная часть», «Фоновая информация» и «Реакция общественности»; последние две части дискретны и факультативны. Полученные результаты согласуются с положениями работ (Добросклонская, 2008; van Dijk, 1988).

К жанровым ограничениям подъязыка также относятся ограничения глагольной лексики на морфологическом уровне. Поскольку новостные сообщения описывают прошедшие события, бóльшая часть глаголов реализуется в прошедшем времени. Глаголы и глагольные группы имеют ограниченное число морфологических репрезентаций на всех трех языках, что отражено в приложениях к диссертации.

В ходе морфосинтаксического анализа в русском и французском языках выявлены **лексико-грамматические корреляции** функционирования отдельных глаголов (в английском корпусе таких корреляций не обнаружено).

Ограничения, обусловленные предметной областью, проявляются на лексико-семантическом уровне. Бóльшая часть контента во всех трех языках отражена в существительных, в меньшей степени – в глаголах, прилагательных и наречиях. Результаты анализа показывают, что частотное распределение лексем по частям

речи во всех трех языках почти совпадает (рис. 1), при этом частотное распределение словоупотреблений различается (рис. 2).

Рис. 1. Относительная частота лексем

Рис. 2. Относительная частота словоупотреблений

Из общего набора лексических единиц выделены единицы, отражающие контент предметной области, и отнесены к 25 базовым **концептуальным классам**, набор которых одинаков для каждого из языков исследования (например, «Контртерроризм», «Объект теракта», «Последствия», «Средства теракта», «Террорист», «Террористическая организация», «Тип теракта»; полный список концептуальных классов с лексическими примерами на русском, английском и французском языках представлен в приложениях к диссертации). При этом выявлено, что отдельные единицы, в общем употреблении не имеющие связанного с терроризмом значения, приобретают его в контексте данной предметной области (например, английская лексема *militant* ‘боец’ в корпусе имеет значение ‘боевик, террорист’). В таблице 1 представлен фрагмент списка лексических единиц русского, английского и французского языков, отнесенных к одному концептуальному классу «Террорист».

Таблица 1

Фрагмент лексического наполнения концептуального класса «Террорист»

Русский язык	Английский язык	Французский язык
боевик	militant	combattant
джихадист	jihadi	djihadiste
смертник	suicide bomber	bombe humaine
террорист	terrorist	terroriste
террористка-смертница	female suicide bomber	femme kamikaze
террорист-одиночка	lone-wolf terrorist	loup solitaire
шахид	shaheed	chahid

Отдельные лексические единицы предметной области обнаруживают свойства **концептуального синкретизма** и **концептуальной неоднозначности** и поэтому отнесены более чем к одному концептуальному классу. Концептуально синкретичными являются единицы, реализующие *не противоречащие друг другу* концептуальные значения; концептуально неоднозначными – единицы, которые в зависимости от контекста могут иметь *противоречащие друг другу* значения и в каждом конкретном случае *реализуют только одно* из них. **Причины** концептуальной неоднозначности могут быть **лингвистическими** (частеречная омонимия, лексическая и синтаксическая неоднозначность) и **экстралингвистическими** (множественность концептуальных значений и экстралингвистический контекст).

Результаты, полученные в ходе корпусного анализа подъязыка новостных сообщений ПО «Терроризм», были применены для создания базы знаний модели интеллектуального контент-анализа и базы знаний платформы концептуального аннотирования. Выявленные ограничения подъязыка могут быть при необходимости использованы для снятия неоднозначности в процессе концептуальной разметки.

В третьей главе «Построение модели многоязычного интеллектуального контент-анализа новостных сообщений предметной области „Терроризм“» описана модель многоязычного ИКА и ее компоненты – база знаний и алгоритм, а также разработка вспомогательного инструмента концептуальной разметки и его настройка на обработку франкоязычных новостных сообщений ПО «Терроризм». В главе также приведено три примера использования модели для интеллектуального контент-анализа новостных сообщений о терроризме на примере русского, английского и смешанного русско-англо-французского корпуса.

Основными компонентами разработанной базы знаний модели являются: онтология предметной области «Терроризм» и база экземпляров, онтолексиконы и ономастиконы, правила онтологического анализа и логического вывода, правила представления знаний в динамических концептуально-лексических фреймах.

При создании онтологии соблюдались **три методологических принципа**:

1. Онтология – это независимый от конкретного языка ресурс, который может быть использован многократно для обработки текстов на различных языках.

2. «Знания о предметной области не изолированы от общих знаний о мире» (Moreno, 2011: 233), поэтому онтология терроризма связана с **онтологией верхнего уровня MikroKosmos** (Nirenburg, 2004) для повторного использования имеющихся знаний и представлена в ее формализме. Все концепты онтологии терроризма, как и онтологии MikroKosmos, подразделяются на объекты (OBJECT), события (EVENT) и свойства (PROPERTY), которые в свою очередь разделены на бинарные отношения (RELATION) и унарные атрибуты (ATTRIBUTE); в качестве названий концептов онтологии (для совместимости с MikroKosmos) использованы слова английского языка, значение концептов определяется их дефинициями (см. таблицу 2).

3. Знания ПО получены на основе анализа многоязычных псевдопараллельных корпусов текстов с использованием комбинированной методики, включающей методы корпусной лингвистики и неформальные эвристические методы.

Разработанная в рамках настоящего исследования онтология ПО «Терроризм» содержит 113 концептов типов OBJECT и EVENT на трех уровнях иерархии, 28 отношений, 5 атрибутов и 32 аксиомы. Онтология представлена в виде запутанного дерева концептов (см. фрагмент дерева онтологии на рис. 3), обладающих фреймовой структурой (см. пример фрейма в таблице 2). Полная иерархическая структура концептов, фреймы концептов и отдельных экземпляров онтологии, а также список аксиом представлены в приложениях к диссертации.

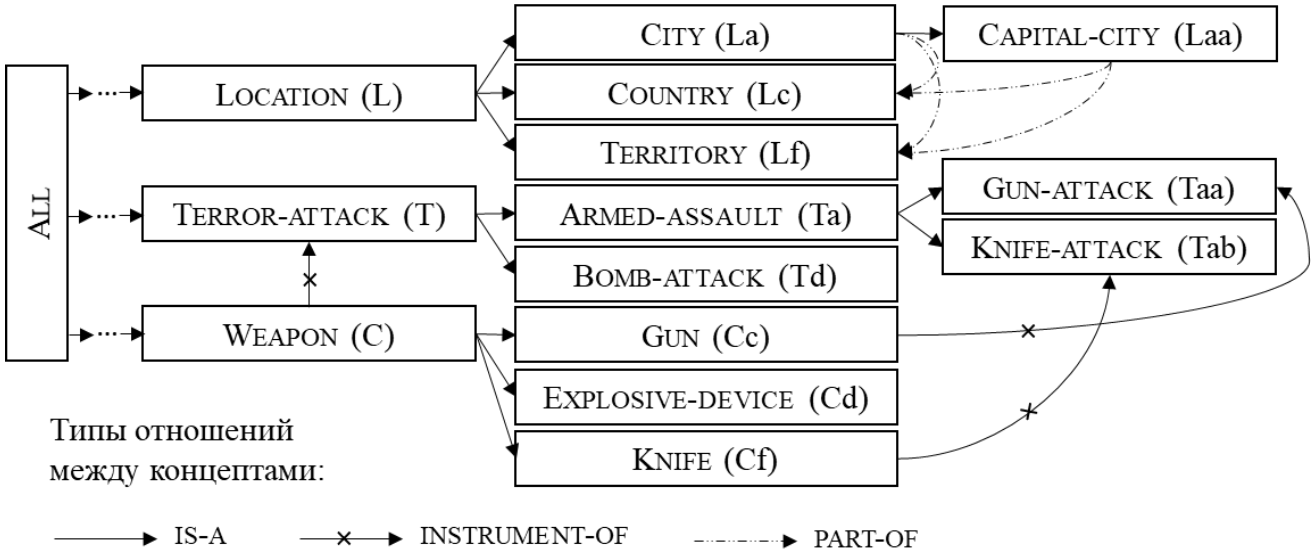

Рис. 3. Фрагмент онтологии терроризма (для наименования концептов использован английский язык, в скобках указаны теги концептов)

Таблица 2.

Фрейм концепта WEAPON

LABEL	WEAPON	VAL
DEFINITION	Оружие или подобные объекты (например, грузовик), используемые для совершения террористического акта, а также функциональные элементы такого оружия	VAL
IS-A	ARTIFACT	VAL
SUBCLASSES	BIOLOGICAL-WEAPON, CHEMICAL-WEAPON, GUN, EXPLOSIVE-DEVICE, INCENDIARY-WEAPON, KNIFE, RAMMING-VEHICLE, ROCKET	VAL
INSTRUMENT-OF	TERROR-ATTACK	SEM
WEAPON-OF	TERRORISM-AGENT, ARTIFACT	SEM
DEFAULT-NAME-RU	оружие	VAL
DEFAULT-NAME-EN	weapon	VAL
DEFAULT-NAME-FR	arme	VAL

В качестве базовых **концептов** онтологии были использованы выявленные ранее концептуальные классы, при этом отдельные классы для более точного отражения ПО были преобразованы в несколько концептов онтологии: так, широкий класс «Последствия», выявленный при анализе подъязыка, был преобразован в три концепта типа OBJECT, обозначающие жертв, террористов и неодушевленные объекты, пострадавшие в результате теракта, и три концепта типа EVENT, обозначающие последствия для жертв, террористов и неодушевленных объектов.

Для определения **отношений** между концептами по методике (Moreno, 2011) были проанализированы связи между лексическими единицами, репрезентующими различные концепты и расположенными в коллокационном диапазоне. В частности, так были выявлены отношения INSTRUMENT, LOCATED, AGENT, DIRECTION, представляющие, по существу, семантические валентности концепта TERROR-ATTACK. Некоторые отношения были выявлены посредством компонентного анализа: например, компонентный анализ именных групп *исламистский терроризм, fundamentalist terrorism, racist terrorism, terrorisme d'extrême-gauche* <терроризм + идеология> позволил выделить отношение HAS-IDEOLOGY. Отдельные отношения были взяты из онтологии верхнего уровня MikroKosmos (CAPITAL-OF, PRECONDITION-OF).

Атрибуты концептов были выявлены посредством компонентного анализа лексических единиц и метода оппозиций: так, оппозиции *девочка / girl / fille – женщина / woman / femme*, *ребенок / child / enfant – подросток / teenager / adolescent* свидетельствуют о наличии у концепта HUMAN наследуемого атрибута AGE, который может быть заполнен численной величиной больше нуля. Кроме того, введен атрибут NEGATION для корректного извлечения информации, поскольку в текстах нередко встречаются формулировки вида *«**Никакая** организация **не взяла** на себя*

ответственности за теракт». В случае отсутствия атрибута NEGATION подобные формулировки могли бы привести к искажению извлеченной информации.

Полученные по результатам описанных этапов концепты типов OBJECT, EVENT, RELATION и ATTRIBUTE были объединены в группы на основе общих признаков и соотнесены с верхними уровнями онтологии MikroKosmos.

Далее базовые концепты онтологии были детализированы с помощью **текстовых паттернов**, представляющих определенные отношения между концептами (Hearst, 1992). Например, отношение DIRECTION, установленное между концептом TERROR-ATTACK и концептами, обозначающими объекты теракта, в русском корпусе представлено паттернами X *против* Y, X *направлен* на* Y, X *нацелен* на* Y, Y *подверг*с** X; в английском: X *target** Y, X *is directed at* Y; во французском: X *cibl** Y, где X – тип теракта, Y – объект теракта. С использованием данных паттернов было установлено, что целями террористов могут быть, например, военные (MILITARY-TARGET), полицейские (POLICE-TARGET), медицинская инфраструктура (MEDICAL-INSTITUTION), транспортные объекты (TRANSPORT-FACILITY) и пр., исходя из чего в онтологию терроризма были введены соответствующие подконцепты.

В онтологию также введен ряд **аксиом**, связывающих элементы онтологии и позволяющих делать различные умозаключения. Пример аксиомы приведен ниже:

$$\forall\, T\ located(T, La) \cap part\ of(La, Lc) \rightarrow located(T, Lc),$$

где T – TERROR-ATTACK, La – CITY, Lc – COUNTRY; т. е. для всех терактов верно, что, если теракт произошел в некоем городе и нам известно, что этот город находится в некой стране, то, следовательно, теракт произошел в этой стране.

Эпизодические знания представлены в онтологии **базой экземпляров**, репрезентующих ветви NATION, LOCATION, TERRORIST-ORGANIZATION и SOURCE. Например, TERRORIST-ORGANIZATION-3 является экземпляром концепта TERRORIST-ORGANIZATION и обозначает организацию «Ястребы свободы Курдистана».

Вторым элементом базы знаний являются англо-, франко- и русскоязычные лексиконы, содержащие релевантные для предметной области единицы длиной от 1 до 10 компонентов, соотнесенные с концептами онтологии, названные **онтолексиконами**. Каждый онтолексикон содержит свыше 20 000 единиц, снабженных концептуальными и морфосинтаксическими признаками, а также перекрестными ссылками на эквиваленты на других языках. Между единицами онтолексиконов и концептами установлены отношения N:1 и 1:N (где $N \geq 1$), что обусловлено возможной концептуальной однозначностью, концептуальной неоднозначностью и концептуальным синкретизмом лексических единиц.

Поскольку языковые репрезентации многих экземпляров онтологических концептов имеют несколько вариантов наименования и написания (например, экземпляр TERRORIST-ORGANIZATION-3 связан с синонимичными лексическими единицами *ТАК, Ястребы свободы Курдистана, Соколы свободы Курдистана, Kurdistan Freedom Hawks, Kurdistan Freedom Falcons, Faucons de la liberté du Kurdistan*, а репрезентация экземпляра TERRORIST-ORGANIZATION-8 в английском ономастиконе имеет варианты написания *Hizbollah, Hezbolla, Hezbollah, Hisbollah, Hizbu'llah, Hizb Allah*), эти варианты отнесены в специальный блок онтолексикона – **ономастикон**.

Правила онтологического анализа включают правила разметки текстов тегами онтологических концептов.

Правила логического вывода основаны на аксиомах онтологии и используются при отсутствии в тексте лексических единиц, размеченных релевантными для поставленной задачи интеллектуального КА концептуальными тегами.

Динамические концептуально-лексические фреймы, генерируемые автоматически на основе исходных пользовательских данных, необходимы для первичного представления и подсчета извлеченной информации.

Концептуально-лексический фрейм содержит слоты, соответствующие выявленным в тексте релевантным концептам и заполняющиеся фрагментами текста, размеченными тегами данных концептов, с указанием частоты встречаемости. При генерации фрейма могут использоваться символы «*», «?» и «!»: символом «*» обозначаются концепты, полученные путем логического вывода, «?» – предположительные концепты (фрагмент текста помечен тегом I атрибута ASSUMPTION), «!» – концепты, наличие которых в тексте отрицается (фрагмент помечен тегом NEG).

Приведем пример фрейма для следующего фрагмента: «*В результате {теракта}~T {в Лондоне}~Laa-186 {никто}~Ha~NEG {не погиб}~Paa~NEG. {Подозреваемый}~A~I~Hb {арестован}~Pbd»:*

AFFECTED-TERRORIST {1}	Подозреваемый {1}
!PEOPLE-DYING {1}	не погиб {1}
*COUNTRY-186 {1}	–
CAPITAL-CITY-186 {1}	в Лондоне {1}
TERRORIST-CAPTURE {1}	арестован {1}
TERROR-ATTACK {1}	теракта {1}
?TERRORISM-AGENT {1}	Подозреваемый {1}
!VICTIM {1}	никто {1}

При формировании конечных результатов ИКА условные названия концептов для удобства пользователя заменяются на слова русского языка, содержащиеся

в слоте DEFAULT-NAME-RU. Подвергнутые числовой обработке результаты оформляются в виде таблиц или графиков, например, средствами программы MS Excel.

Алгоритм интеллектуального контент-анализа, представленный в центральной части блок-схемы на рис. 4, включает в себя этап определения задачи интеллектуального контент-анализа и процедуры ее решения, выполняемые вручную или автоматически на основе авторских программ. На вход каждой из процедур подается информация, полученная на выходе предыдущей процедуры алгоритма.

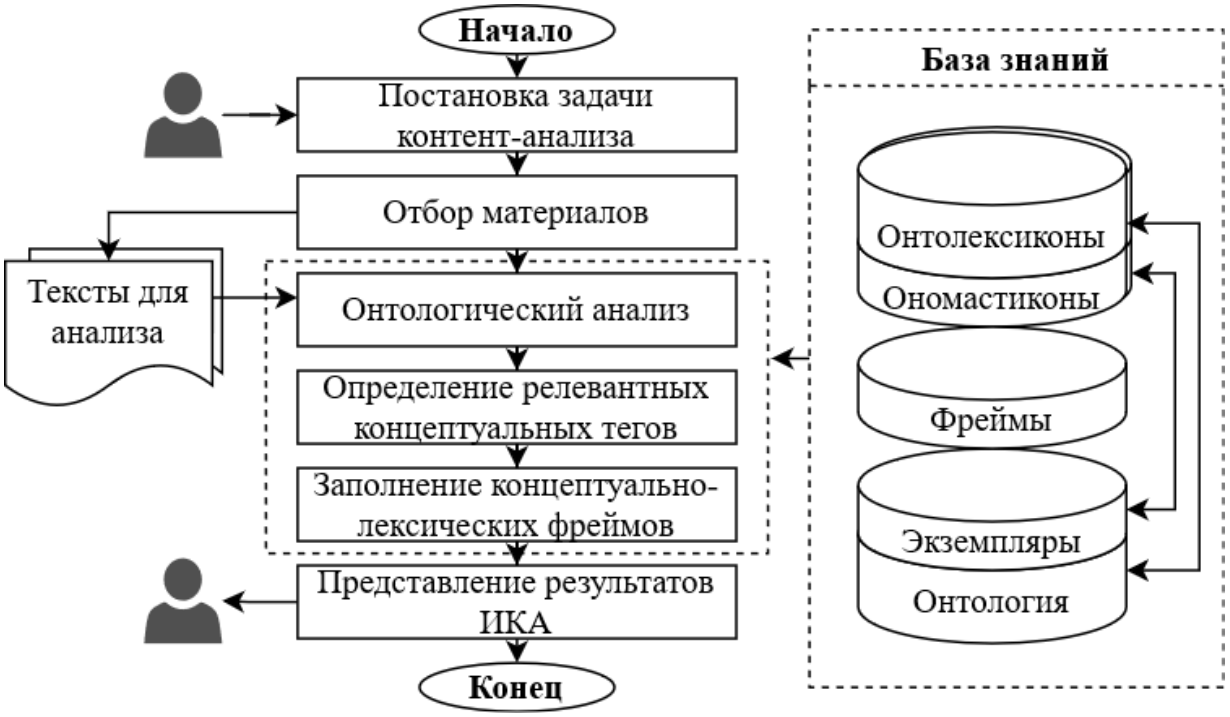

Рис. 4. Архитектура модели многоязычного интеллектуального контент-анализа

Постановка задачи контент-анализа требует четкой формулировки, поскольку именно на этом этапе определяются предпосылки для отбора текстового материала и идентификации релевантных концептуальных тегов.

После формулирования задачи контент-анализа выполняется процедура отбора релевантного текстового материала (отдельных текстов или корпусов) на одном или нескольких языках в зависимости от поставленной задачи. Отбор материала производится методом критериальной выборки через интернет-агрегатор новостей по ключевым словам, т. е. словам, входящим в формулировку задачи. Репрезентативность выборки обеспечивается путем фильтрации материала по следующим критериям: 1) текст опубликован в открытом доступе и не требует платной подписки; 2) текст не является републикацией уже отобранного в корпус текста; 3) текст может содержать информацию о нескольких релевантных запросу событиях, только если не было найдено отдельного текста для описания каждого их событий.

Следующие процедуры алгоритма применяются последовательно к каждому тексту собранного корпуса. Сначала выполняется онтологический анализ, состоящий из подпроцедуры формальной разметки текста тегами концептов на основе онтолексиконов, выполняемой автоматически, и подпроцедуры снятия концептуальной неоднозначности, в настоящем исследовании выполняемой вручную; при этом мультитеги концептуально синкретичных лексем не требуют разрешения.

В ходе следующей процедуры определяется набор релевантных для задачи КА концептуальных тегов; эта процедура включает в себя подпроцедуру определения фрагмента онтологии, содержащего релевантные для поставленной задачи концепты (и их теги), и подпроцедуру логического вывода, которая актуализируется только в том случае, если в тексте нет лексем, релевантных для поставленной задачи, т. е. если релевантная информация не представлена в тексте эксплицитно.

Подпроцедура логического вывода основана на анализе связей онтологических концептов и позволяет получить на выходе дополнительные релевантные для поставленной задачи теги, а также размеченные ими лексические единицы анализируемого текста. Например, концепт GUN (дочерний концепт WEAPON, связанного отношением INSTRUMENT-OF с концептом TERROR-ATTACK) связан отношением INSTRUMENT-OF с концептом GUN-ATTACK и может быть использован для логического вывода информации о типе теракта, если в тексте таковая отсутствует.

После идентификации релевантных тегов выполняется процедура извлечения информации и заполнения концептуально-лексического фрейма, формируемого динамически на основе вводных данных. Процедура выполняется автоматически с помощью разработанного автором экстрактора. Затем данные в заполненных фреймах подвергаются числовой обработке и выдаются пользователю. Представление результатов исследования выполняется в форме таблиц или графиков.

С целью автоматизации процедуры онтологического анализа сотрудниками НОЦ «ЛИнТ» ЮУрГУ при участии автора диссертации разработана платформа концептуального аннотирования, состоящая из модуля сбора и хранения знаний и концептуального теггера, для обработки текстов ПО «Терроризм» на русском и английском языках (Шереметьева, 2020; Sheremetyeva, 2020). Автором диссертации платформа настроена также на обработку французского языка: в частности, были изменены набор лексико-грамматических классов, морфологическая зона словарной статьи, содержащая иконическое представление словоформ, и морфологический генератор для заполнения полей морфологической зоны глаголов. Платформа может быть использована для полной автоматизации формальной концептуальной разметки и частичной автоматизации снятия концептуальной неоднозначности.

Приведем **пример использования** модели многоязычного интеллектуального контент-анализа для анализа тенденций в части количества терактов и способов их совершения на материале англоязычного корпуса.

Постановка задачи: сравнить количество терактов, освященных в англоязычных СМИ в первом квартале 2019 и 2020 гг., в странах Евразии, и способы совершения терактов во всем мире за указанный период.

Отбор материала: с учетом критериев отбора через агрегатор «Google Новости» по ключевым словам *terror attack*, *terrorist attack*, *act of terrorism* с ограничением по датам публикации 01.01.2019–31.03.2019 и 01.01.2020–31.03.2020 собрано два корпуса англоязычных сообщений о терактах.

Онтологический анализ: на рис. 5 показан один из текстов на выходе процедуры онтоанализа, включающей разметку и снятие концептуальной неоднозначности (жирным выделены единицы, размеченные релевантными тегами); причем мультитеги концептуально синкретичных лексических единиц не устраняются.

{Attempted stabbing terror attack}~Tab~K **{near Hebron}~La-86**, {terrorist}~Hb~Aa {killed}~Pba~Rb

{An}~DEF **{attempted stabbing terror attack}~Tab**~K occurred **{near Hebron}~La-86** […] {the}~DEF {terrorist}~A {attempted}~K to **{stab}~Tab** {soldiers}~Zab stationed at {an}~DEF **{IDF post}~Le {in Kiryat Arba}~La-102**.

Рис. 5. Фрагмент текста на выходе процедуры онтоанализа

Определение релевантных концептуальных тегов: в соответствии с указанной выше задачей интеллектуального контент-анализа требуется выделить фрагмент онтологии с вершинами LOCATION и TERROR-ATTACK и теги узлов этого дерева считать релевантными. В размеченном тексте примера на рис. 5 отсутствуют лексические единицы, эксплицитно репрезентующие концепты COUNTRY и (или) TERRITORY (т. е. в размеченном тексте нет тегов Lc и Lf), однако присутствуют единицы, репрезентующие концепт CITY (La), который отношением PART-OF связан с концептами CITY и (или) TERRITORY. Это предложные группы *in Kiryat Arba* ‘в городе Кирьят-Арба’ и *near Hebron* ‘недалеко от Хеврона’, включенные в базу знаний как экземпляры CITY-102 и CITY-86 соответственно и связанные отношением PART-OF с экземпляром TERRITORY-2 (*West Bank* ‘Западный берег реки Иордан’). Следовательно, на основании аксиомы «Если теракт произошел в некоем городе и нам известно, что этот город находится на некой территории, то, следовательно, теракт произошел на этой территории» можно заключить, что территорией, на которой был совершен описываемый теракт, является Западный берег реки Иордан.

Формирование концептуально-лексического фрейма: фрейм, сформированный на основе автоматической экстракции релевантной информации из текста на рис. 5, имеет вид (результат логического вывода отмечен символом «*»):

*TERRITORY-2 {3}	–
CITY-86 {2}	near Hebron {2}
CITY-102 {1}	in Kiryat Arba {1}
SPECIFIC-LOCATION {2}	IDF post {1}, building {1}
KNIFE-ATTACK {3}	attempted stabbing terror attack {2}, stab {1}

Представление результатов: результаты представлены средствами MS Excel. На рис. 6 показаны картограммы распределения терактов по странам и территориям в первом квартале 2019 и 2020 гг. На рис. 7 представлена сравнительная информация о типах терактов, совершенных в первом квартале 2019 и 2020 гг. Языковые репрезентации на русском языке для представления концептов на графиках взяты из слотов DEFAULT-NAME-RU соответствующих концептов.

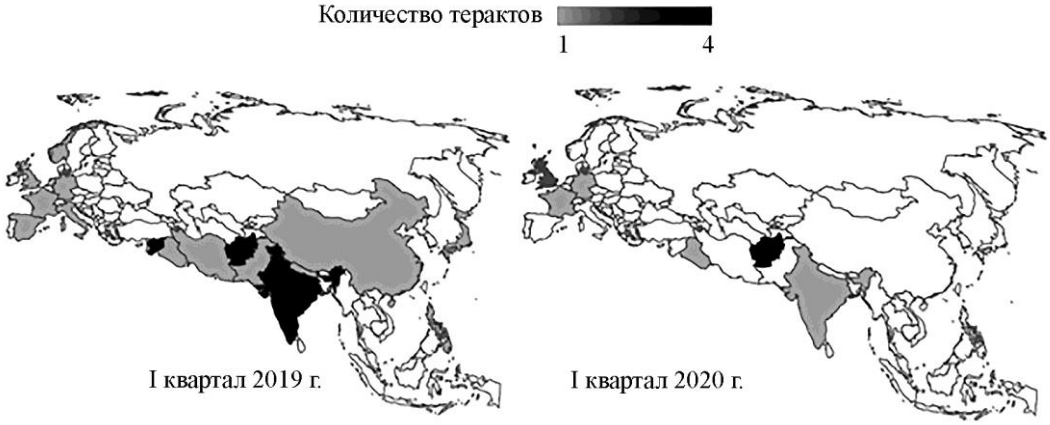

Рис. 6. Количество терактов в Евразии в I квартале 2019 и 2020 гг.

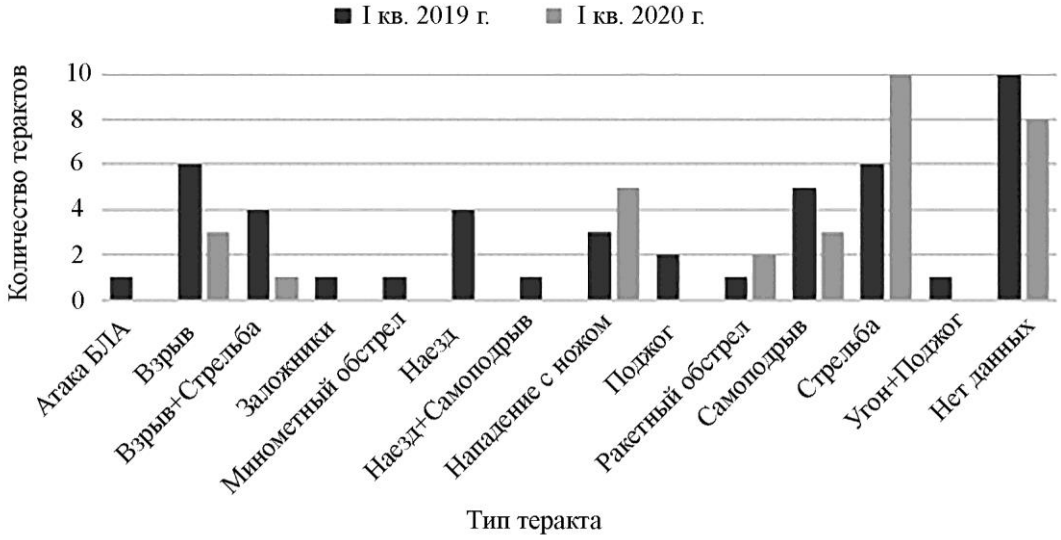

Рис. 7. Типы терактов в мире в I квартале 2019 и 2020 гг.

Данный пример демонстрируют возможность успешного извлечения из текстов релевантного для информационного запроса пользователя контента, в том числе выраженного имплицитно, и представления его в графической форме.

В **заключении** приводятся общие итоги и выводы по проведенному исследованию, обозначаются перспективы дальнейших исследований по теме.

Основным итогом исследования является методика разработки модели многоязычного интеллектуального контент-анализа на основе многоязычного корпуса текстов предметной области, а также сама модель. В ходе исследования проведен анализ подъязыка новостных сообщений предметной области «Терроризм» на материале русского, английского и французского языков; на основе результатов анализа были разработаны база знаний и алгоритм многоязычного интеллектуального контент-анализа. На базе разработанной модели возможно создание системы многоязычного интеллектуального контент-анализа. Сфера применения модели может быть расширена посредством включения новых концептов. Использование независимой от конкретного языка онтологии в основе базы знаний делает возможным применение данной модели к другим языкам при условии присоединения к ней новых лексиконов. Кроме того, при замене блока концептуальных и эпизодических знаний модель может быть использована для анализа других предметных областей.

Перспективным является дальнейшее исследование проблемы концептуальной неоднозначности при разметке текстов ограниченной предметной области и методов ее разрешения.

Основные положения диссертации отряжены в следующих публикациях:

В изданиях, рекомендованных ВАК РФ:

1. Шереметьева, С. О. К вопросу о разработке онтологических ресурсов предметной области «Терроризм» / С. О. Шереметьева, А. Ю. Зиновьева // Вестник Южно-Урал. гос. ун-та. Сер.: Лингвистика. – 2017. – Т. 14, № 4. – С. 48–54.

2. Анализ англоязычных интернет-сообщений о террористических актах на основе многоязычной онтологии / С. О. Шереметьева, О. И. Бабина, А. Ю. Зиновьева, Е. Д. Неручева // Вестник Южно-Урал. гос. ун-та. Сер.: Лингвистика. – 2020. – Т. 17, № 1. – С. 30–36.

3. Об использовании метода кейс-стади для создания универсальных ресурсов концептуального аннотирования многоязычных текстов / С. О. Шереметьева, О. И. Бабина, А. Ю. Зиновьева, Е. Д. Неручева // Вестник Южно-Урал. гос. ун-та. Сер.: Лингвистика. – 2020. – Т. 17, № 4. – С. 46–52.

4. Шереметьева, С. О. Моделирование многоязычного интеллектуального контент-анализа / С. О. Шереметьева, А. Ю. Зиновьева // Вестник Южно-Урал. гос. ун-та. Сер.: Лингвистика. – 2021. – Т. 18, № 2. – С. 52–58.

Прочие публикации по теме исследования:

5. Sheremetyeva, S. On Modelling Domain Ontology Knowledge for Processing Multilingual Texts of Terroristic Content / S. Sheremetyeva, A. Zinoveva // Communications in Computer and Information Science. – 2018. – Vol. 859. – P. 368—379 (Scopus).

6. Sheremetyeva, S. Ontological Analysis of e-News: A Case for Terrorism Domain / S. Sheremetyeva, A. Zinoveva // CEUR Workshop Proceedings. IS 2019 – Proc. of the 14th International Conference on Interactive Systems: Problems of Human-Computer Interaction. – 2019. – P. 130–141 (Scopus).

7. Зиновьева, А. Ю. Источники концептуальной неоднозначности во франкоязычном корпусе новостей о терроризме и методы ее разрешения / А. Ю. Зиновьева // Научный поиск. Материалы 12-й науч. конф. аспирантов и докторантов. – 2020. – С. 79–87.

8. Зиновьева, А. Ю. Концептуальная неоднозначность в англоязычных текстах о терроризме: причины возникновения и методы разрешения / А. Ю. Зиновьева // INJOIT. – 2020. – Т. 8, № 11. – С. 64–72.

9. Зиновьева, А. Ю. Анализ неоднозначности концептуальной разметки русскоязычного текста / А. Ю. Зиновьева, С. О. Шереметьева, Е. Д. Неручева // Вестник Тюмен. гос. ун-та. Гуманитарные исследования. Humanitates. – 2020. – Т. 6, № 3 (23). – С. 38–60.

10. Zinoveva, A. On Resolving Conceptual Ambiguity in an English Terrorism e-News Corpus / A. Zinoveva // CEUR Workshop Proceedings. International Conference “Internet and Modern Society” (IMS-2020). – 2021. – Vol. 2813. – P. 288–299 (Scopus).