УДК 004.82 + 81’32 + 811.1 + 323.28 DOI: 10.14529/ling200408

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
Южно-Уральский государственный университет, г. Челябинск, Россия

Создание аннотированных корпусов текстов имеет критически важное значение для разработки компьютерных технологий обработки неструктурированной информации (автоматической классификации, интеллектуального контент- и тренд-анализов, машинного обучения, машинного перевода и др.) и находится в центре внимания международных теоретических и прикладных лингвистических исследований. При этом ключевым аспектом этих исследований является автоматизация аннотационных процедур, что, в свою очередь, требует статических (лингвистических) и динамических (программных) ресурсов, с возможностью их полного или частичного многократного применения для аннотирования многоязычных текстов различных предметных областей. В настоящей статье представлен опыт применения метода кейс-стади для создания ресурсов автоматизации концептуального аннотирования, одного из самых востребованных и проблематичных видов аннотаций. Под концептуальной аннотацией понимается тип семантической аннотации, ориентированный на решение конкретных информационных задач в рамках определенной предметной области. Методология и конкретные результаты исследования представлены на основе кейс-стади корпусов текстов предметной области «Терроризм» на русском, английском и французском языках. Ресурсы, созданные в ходе исследования, включают в себя как методику их разработки, так и конкретный программный инструментарий и лингвистический материал (многоязычную онтологию и концептуально аннотированные корпусы текстов предметной области «Терроризм» на трех языках). Результаты исследования можно напрямую использовать для увеличения объема концептуально аннотированных корпусов предметной области «Терроризм», разработки метрик разрешения концептуальной многозначности, а также для автоматизации аннотирования текстов других предметных областей и языков. Результаты настоящего исследования представляют интерес и для сравнительных лингвистических исследований.

Ключевые слова: концептуальное аннотирование, статические и динамические ресурсы, предметная область, онтология, многоязычность, независимость от естественного языка, терроризм.

Введение

Создание лингвистически аннотированных корпусов текстов имеет критически важное значение для разработки компьютерных технологий обработки неструктурированной информации и находится в центре внимания международных теоретических и прикладных лингвистических исследований [6, 9]. Современные исследования в этой области свидетельствуют о тенденции к семантизации, в частности, к разработке семантических аннотаций, отражающих контент корпусов текстов ограниченных предметных областей, что является наиболее реалистичным способом решения таких задач, как интеллектуальный контент-анализ, машинное обучение, автоматическая классификация, машинный перевод и др. В настоящей работе такой тип семантической аннотации называется концептуальной аннотацией. Основным инструментом концептуального аннотирования является разметка текстов на основе предметно-ориентированных онтологий, которые создаются для решения конкретных задач обработки информации, например, для автоматического анализа историй болезней [10], персонализированной классификации новостей [14], отслеживания и предотвращения террористической деятельности [4, 7, 8]. Большая часть таких исследований выполняется для английского языка. Один из немногих примеров работ по концептуальному аннотированию русских текстов описан в статье [1]. Поскольку создание больших объемов аннотированных текстов, что необходимо для информационных разработок, сопряжено с большими затратами труда, средств и времени, на первый план выдвигаются задачи автоматизации аннотационных процедур и обеспечения возможности их многократного применения к многоязычным текстам различных предметных областей. В настоящее время прилагается много усилий для создания различных ресурсов, направленных на решение этих задач, в частности, создаются стандартизированные языки семантического (в том числе концептуального) аннотирования, например, XML, SGML и др., разрабатываются приложения для автоматической аннотации лекси-

ческих единиц текста концептами онтологии и/или поддержки постредактирования результатов автоматического аннотирования [2, 13]. Тем не менее проблема разработки аннотационных ресурсов еще далека от решения, и трудно найти систему, которая бы точно отвечала требованиям исследователей.

Целью настоящей работы является создание методологии и собственно ресурсов концептуального аннотирования многоязычных текстов с возможностью их многократного применения к различным предметным областям. Методология и разработанные на ее основе конкретные ресурсы концептуального аннотирования получены в ходе применения метода кейс-стади к корпусам текстов предметной области «Терроризм» на русском, английском и французском языках. Метод кейс-стади, или просто кейс-стади, как эмпирическая стратегия исследований предполагает качественные и количественные методы анализа конкретного материала с последующими выводами дедуктивного и/или индуктивного характера. В отечественной литературе в основном описывается его применение в образовательной, социальной и бизнес-сферах. Однако в международном масштабе этот метод широко используется и в области компьютерной лингвистики [3].

Статья организована следующим образом. В разделе 1 определяются задачи исследования и приводятся данные, лежащие в основе кейс-стади. Раздел 2 дает описание конкретных шагов кейс-стади и полученных на его основе результатов. Раздел 3 представляет процедуру аннотирования «сначала машина – потом человек» на основе построенных ресурсов. В заключении рассмотрены итоги исследования.

1. Постановка задачи и данные для кейс-стади

Разработка многоязычных ресурсов для аннотирования должна быть тесно связана с самим процессом аннотирования, что в настоящем исследовании определяется пересечением следующих критериев: 1) методология на основе кейс-стади корпусных данных; 2) ориентация на предметную область; 3) многоязычность и независимость от естественного языка; 4) автоматизация процесса аннотирования; 5) возможность многократного использования ресурсов.

Мы утверждаем, что при создании многоязычного концептуально-аннотационного ресурса необходимо строго разделять лексические данные, зависящие от конкретного языка, и не зависящие от отдельного языка концептуальные данные, которые наилучшим образом могут быть представлены в предметно-ориентированной онтологии, построенной для выполнения конкретной информационной задачи (в нашем случае – многоязычного концептуального аннотирования). При этом онтологический анализ на основе такой онтологии является основным, не зависящим от естественного языка инструментом концептуального аннотирования текста онтологическими концептами. С целью сократить затраты усилий и времени на «ручное» аннотирование мы предлагаем сначала создать ресурсы, автоматизирующие этот процесс, и затем совершенствовать их в процессе аннотирования. Разработка таких ресурсов, среди которых мы, следуя классификации, предложенной в [16], выделяем статические (лингвистические) и динамические (программные), определяет основное содержание настоящей статьи. Динамические ресурсы – это инструменты для автоматизации как создания статических ресурсов, так и собственно процедуры аннотирования. Программный инструментарий будет описан ниже, по мере его применения. Что касается статических ресурсов, то в нашем исследовании выделяются две их составляющие. К первой составляющей статических ресурсов относятся исходные данные кейс-стади (сопоставимые корпусы текстов интернет-новостей о террористических актах на трех языках – английском, французском и русском), универсальная схема концептуальной аннотации, многоязычная онтология предметной области «Терроризм», одноязычные лексиконы каждого из языков, содержащие релевантные для предметной области лексические единицы, концептуальный статус каждой из которых определен онтологическим концептом. Ко второй составляющей относятся полученные в результате применения автоматизированного инструментария концептуально аннотированные русские, английские и французские корпусы текстов, которые могут использоваться как для дальнейшего увеличения и совершенствования статических ресурсов, так и для разработки приложений обработки информации.

Исходные данные кейс-стади можно разделить на три части:

(1) русскоязычный корпус интернет-сообщений о террористических актах за 2016–2017 годы объемом 400 000 словоупотреблений;

(2) русско-английский лексикон, состоящий из многокомпонентных лексических единиц, созданный профессиональными переводчиками на основе русского вокабуляра указанного корпуса;

(3) три корпуса текстов интернет-сообщений о террористических актах за 2016–2019 годы на русском, английском и французском языках объемом в 500 000 словоупотреблений каждый.

Первые две части данных были созданы в рамках предыдущих проектов и повторно использованы в качестве базы знаний автоматического инструмента (веб-кроулера) для сбора из сети Интернет новых текстов предметной области «Терроризм» (части 3 перечисленных данных). Русская часть знаний кроулера включает многокомпонентные именные ключевые фразы, наиболее близко отражающие содержание текста [5, 15], которые были автоматически извлечены из корпуса (1) с помощью автоматического экстрактора [11].

Английская часть знаний кроулера включает английские эквиваленты русских единиц из лексикона (2); французская – французские эквиваленты русских ключевых фраз, переведенные профессиональными переводчиками.

Полученные с помощью веб-кроулера три корпуса разноязычных текстов интернет-сообщений предметной области «Терроризм» (3) были использованы в качестве основных данных кейс-стади, в процессе применения которого были разработаны статические и динамические ресурсы автоматизации концептуального аннотирования, представленные в следующих разделах статьи.

2. Статические ресурсы

В этом разделе описывается процесс создания первой составляющей статических ресурсов концептуального аннотирования многоязычных текстов на этапе, предшествующем аннотационному процессу. Результаты этого этапа исследования были использованы в качестве исходной базы знаний при разработке автоматической платформы многоязычного концептуального аннотирования (см. раздел 3 данной статьи) и периодически обновлялись в течение всего периода исследования.

В каждом из корпусов текстов (3) на русском, английском и французском языках были выделены лексические единицы, отражающие контент предметной области «Терроризм», и проведена их классификация по концептуальным классам (категориям). Исходный набор концептуальных классов (категорий) был задан интуитивно-прескриптивно. Значения концептуальных категорий заданы дефинициями на русском языке и обозначены простыми английскими словами для универсальности и совместимости с релевантными для настоящей работы зарубежными исследованиями. Фрагменты использованных концептуальных категорий приведены в табл. 1 и 2. В качестве единиц анализа приняты одно- и многокомпонентные лексемы (именные, глагольные, предложные, наречные группы и т. д.). Использование многокомпонентных лексических единиц для концептуального анализа текста, с одной стороны, повышает точность анализа, а с другой – позволяет избежать многих проблем, связанных с многозначностью естественного языка [15]. Концептуальная классификация лексического состава каждого из корпусов (3) осуществлялась параллельно с помощью одинаковой последовательности процедур. Первоначальный список (всех, не только ключевых единиц анализа) был получен из корпусов (3) с помощью автоматического экстрактора [11] и состоял из одно–четырехкомпонентных фраз (что обусловлено техническим ограничением инструмента на длину фразы). Затем в каждом одноязычном корпусе проведен поиск более длинных именных фраз с использованием функции текстового редактора «Найти» и первоначального списка автоматически извлеченных четырехкомпонентных лексических единиц. Из исходного списка были отобраны лексические единицы, отражающие контент предметной области, которые, в свою очередь, были разнесены по прескриптивным концептуальным категориям. Важный компонент этого предварительного исследования – заранее разработанная четкая инструкция выполнения концептуальной классификации, которая на этом этапе

Таблица 1

Концепты второго уровня для концептов верхнего уровня COUNTER-TERRORISM и CONSEQUENCES

COUNTER-TERRORISM: люди, оказывающие сопротивление терроризму, и антитеррористические действия:
COUNTER-TERRORISM AGENT: люди, оказывающие сопротивление терроризму
COUNTER-TERRORISM MEASURES: действия по борьбе с терроризмом
CONSEQUENCES: все последствия теракта
PUBLIC LOSS: погибшие, раненые, заложники, не пострадавшие
DESTRUCTION: объекты, которым был нанесен урон или полностью уничтоженные
TERRORISTS' LOSS: покончил с собой, убит, ранен, сбежал
TERRORISTS' GAIN: ответ террористов
PUBLIC REACTION: проявление поддержки населения
RECONSTRUCTION: восстановление разрушенных объектов

Таблица 2

Фрагменты частотных списков лексических единиц, отнесенных к концептуальной категории AGENT–TERRORIST

Язык	Концептуальная категория: "AGENT–TERRORIST"
Английский	terrorist, militant, fighter, gunman, suicide bomber, jihadi, female suicide bomber, female terrorist, lone-wolf terrorist, ISIS terrorist
Французский	terroriste, kamikaze, combattant, femme kamikaze, djihadiste. loup solitaire, terroriste de l'EI, combattant terroriste, femme terroriste
Русский	террорист, боевик, смертник, террорист-смертник, террористка-смертница, игиловец, террористка, джихадист, террорист-одиночка

участниками проекта выполнялась вручную. В течение всего периода концептуальной классификации проводились еженедельные обсуждения, в процессе которых согласовывался и дорабатывался список концептов. В результате этого этапа кейс-стади был получен обновленный набор концептуальных категорий предметной области «Терроризм» в виде трехуровневой древовидной структуры, содержащий 117 категорий предметной области, представленных во всех трех корпусах текстов (русском, английском и французском), из которых 20 категорий отнесено к верхнему уровню и 97 – ко второму и третьему уровням. В табл. 1 приведены примеры категорий первого и второго уровня и их определения. Далее для каждого языка получены частотные списки лексем, отнесенных к выделенным концептуальным категориям. Табл. 2 приводит фрагмент самых частотных лексем на трех языках, отнесенных к концепту AGENT–TERRORIST (исполнитель террористических актов).

Как показал кейс-стади, в лексике каждого языка, несмотря на ограничения предметной области, наблюдается явление концептуальной многозначности, т. е. в корпусах предметной области на всех трех языках имеются лексемы, которые могут функционировать в *различных взаимоисключающих* концептуальных значениях и поэтому одновременно заносятся в лексиконы различных концептуальных категорий. Так, как показано в приведенных ниже примерах, одна и та же английская лексема «military» в англоязычном корпусе предметной области встречается в 3 концептуальных значениях и поэтому занесена в 3 лексикона следующих категорий:

SOURCE: *"Our forces killed 15 terrorists," the **military** said in a statement.*

COUNTER-TERRORISM: *A police intelligence officer said it could be a diversionary tactic as the **military** launched air and ground assaults against Maute terrorist group in Lanao del Sur.*

OBJECT (TARGET): *The attacks targeting police and **military** increased after the ouster of Islamist ex-president Mohamed Morsi in 2013 by military following massive protests against his rule*.

Далее в связи с тем, что единицами кейс-стади являются не только однокомпонентные, но и многокомпонентные лексические единицы, относительно их категориальной классификации нами было принято специфическое, не общепринятое решение. Его суть заключается в том, что в многокомпонентной лексической единице каждый ее компонент может иметь индивидуальные концептуальные *не противоречащие друг другу* значения. Например, в английской именной группе «airport shooting suspect» слово «shooting» содержит в себе информацию о типе атаке, слово «airport» указывает на место, где произошла атака, слово «suspect» имеет сразу два концептуальных значения – «предположение» и «исполнитель террористического акта». Поэтому данная многокомпонентная лексема заносится в лексиконы четырех концептуальных категорий: AGENT-TERRORIST, ASSUMPTION, TYPE OF ATTACK и LOCATION. Аналогично фраза «Algerian terrorist» соотносится с концептами AGENT–TERRORIST и NATION. В данном случае в отличие от лингвистического явления концептуальной многозначности, проиллюстрированного на примере лексической единицы «military», мы имеем дело с синкретичностью нескольких концептуальных значений, выраженных одной лексической единицей. Это явление выявлено в корпусах на всех трех языках.

Таким образом, построенные на этом этапе исследования лексиконы, соотнесенные с концептуальными категориями, содержат пересекающиеся множества лексических единиц, что обусловлено двумя кардинально различными лингвистическими явлениями – концептуальной многозначностью и концептуальной синкретичностью.

На следующем этапе кейс-стади была построена исходная многоязычная, т. е. не зависящая от конкретного языка, онтология предметной области «Терроризм», в которой набор концептов соответствовал выделенным концептуальным категориям одноязычных лексем. Исходная онтология путем дополнительного анализа корпусного материала с помощью специализированных методик анализа текстов (например, метода шаблонов) была уточнена и расширена. Детали построения онтологии приведены в [12]. Все лексемы, отражающие контент предметной области, были соотнесены с онтологическими концептами. В нашем случае в связи с тем, что составленные на предыдущем этапе категориально классифицированные одноязычные лексиконы состояли из пересекающихся множеств лексем, кроме однозначного соотнесения лексем и онтологических концептов, были разрешены следующие отношения между текстовыми лексическими единицами и концептами онтологии: «один-ко-многим», «многие-к-одному» и «многие-ко-многим».

Создание многоязычной онтологии предметной области «Терроризм» и одноязычных (английского, французского и русского) лексиконов, единицы которых соотнесены с концептами онтологии, завершило создание первого компонента статических ресурсов концептуального аннотирования, которые, как указывалось выше, использованы для создания первой версии автоматической платформы аннотирования – основного компонента динамических аннотационных ресурсов.

3. Динамические ресурсы процесса аннотирования

Процесс концептуального аннотирования основан на онтологическом анализе текстов и на практике состоит в разметке корпусных лексем специальными кодами – метками концептов онтологии. Задачей динамического ресурса концепту-

ального аннотирования, который в нашем исследовании называется *платформой аннотирования*, является автоматизация этого процесса. Разработка платформы аннотирования проходила с учетом следующих факторов. Во-первых, платформа позволяет аннотирование текстов на разных языках с возможностью перенастройки ее параметров для разных типов лингвистической информации. Во-вторых, в программе предусмотрены функции сбора и управления знаниями. В-третьих, учитывая, что между концептами онтологии и лексическими единицами текста, как указано в предыдущем разделе, не во всех случаях имеется однозначное соответствие, т. е. одной лексеме может быть присвоено несколько концептуальных меток, желательно, чтобы в базе знаний платформы хранились знания, которые можно использовать для разрешения концептуальной неоднозначности. Объем настоящей статьи не позволяет дать детальное описание разработанной платформы концептуального аннотирования – это тема отдельной статьи. Здесь же отметим, что платформа состоит из двух основных блоков: 1) электронной базы знаний, фиксирующей отображения каждого из предметно-релевантных русских, английских и французских лексиконов на концепты онтологии в соответствии с процедурой, описанной в предыдущем разделе, 2) работающего на этих знаниях концептуального теггера. Теггер позволяет осуществлять простую и развернутую разметку корпуса. Простая разметка представляет собой автоматическое присвоение лексическим единицам только концептуальных тегов. Этого может быть достаточно для выполнения определенных задач обработки информации. При развернутой разметке корпусные лексемы, в дополнение к концептуальным, содержат морфосинтаксические теги. Каждый из блоков платформы снабжен пользовательским интерфейсом. Интерфейс базы знаний позволяет разработчикам автоматизированно собирать, обновлять и кодировать многоязычные лингвистические знания. Интерфейс теггера служит для контроля корректности разметки и снабжен компиляторами правил разрешения многозначности разметки. Построенная таким образом платформа концептуального аннотирования использована для реализации нашего подхода к созданию концептуально аннотированных корпусов текстов на различных языках «сначала машина – затем человек».

В связи с большим объемом аннотационных работ подход был протестирован при создании «золотых стандартов» концептуально-аннотированных одноязычных (русских, английских и французских) корпусов текстов новостных сообщений о террористических атаках, объемом в 20 000 словоупотреблений каждый с использованием онтологических концептов верхнего уровня. Процесс аннотирования происходил в несколько этапов итеративным способом. Сначала небольшая часть необработанного текста, автоматически тегировалась нашей компьютерной платформой, после чего постредактировалась (корректировалась) вручную. В случае если лексема, относящаяся к предметной области, оставалась неразмеченной или была размечена неправильно, то исследователь добавлял эту лексему со всей необходимой языковой информацией в базу знаний платформы концептуального аннотирования. Таким образом, знания платформы пополнялись, и обновленную платформу использовали для автоматического аннотирования следующих фрагментов текстов, и т. д. Знания обновлялись регулярно, и точность автоматического аннотирования повышалась с каждым циклом. Точность оценивалась на основе отчетов исследователей о времени, потраченном на постредактирование каждого нового фрагмента корпуса, и количестве новых лексических единиц, добавленных в базу знаний платформы после каждого цикла. Наш эксперимент показал правомерность предложенного нами подхода концептуального аннотирования на основе предварительного создания статических и динамических ресурсов для автоматизации этого процесса.

Заключение

В данной статье была предложена методология по созданию статических и динамических ресурсов для осуществления многоязычного концептуального аннотирования корпусов текстов предметной области и были представлены сами аннотационные ресурсы, созданные по предложенной методологии для английского, французского и русского корпусов текстов интернет-сообщений о террористических актах. В число ресурсов входят универсальная схема концептуального аннотирования, многоязычная предметная онтология, компьютерная платформа с гибкими настройками и «золотые стандарты» концептуально аннотированных корпусов текстов на трех языках. Данное исследование является одной из основных частей проекта аннотирования. Оно значительно отличается от других подобных исследований, которые фокусируют свое внимание на морфологических, синтаксических или общих типах семантического аннотирования. Качественное и количественное исследование созданных нами аннотационных ресурсов открывает ряд новых возможностей для исследований, связанных, например, с теоретическими аспектами социолингвистики и сравнительного языкознания, а также с развитием технологий обработки естественного языка.

Литература/References

1. Добров А.В., Доброва Н.Л., Сомс Н.Л., Чугунов А.В. Семантический анализ новостных сообщений по теме «Электронные услуги»: опыт применения методов онтологической семантики. Труды XVIII объединенной конференции «Интернет и современное общество», Санкт-Петербург, 23–25 июня 2015 г. СПб., 2015. С. 120–125. [Dobrov A.V., Dobrova N.L., Soms N.L., Chugunov A.V.

[Semantic Analysis of News Items on 'Electronic Services' Subject Domain: Experience of Applying Methods of Ontological Semantics]. *Trudy 18 ob'edinennoj konferencii "Internet i sovremennoe obshchestvo"* [Proceedings of the 18th United Conference "Internet and Modern Society", Saint Petersburg, June 23–25, 2015]. Saint Petersburg, 2015, pp. 120–125. (in Russ.).]

2. Загорулько М.Ю., Кононенко И.С., Сидорова Е.А. Система семантической разметки корпуса текстов в ограниченной предметной области. Материалы международной конференции «Компьютерная лингвистика и интеллектуальные технологии», Бекасово, 30 мая – 3 июня 2012. М., РГГУ, 2012, Вып. 11(18), с. 674–683. [Zagorul'ko M.Yu., Kononenko I.S., Sidorova E.A. [System for Semantic Annotation of Domain-Specific Text Corpora]. *Materialy mezhdunarodnoy konferentsii "Komp'yuternaya lingvistika i intellektual'nye tekhnologii", Bekasovo, 30 maya – 3 iyunya 2012*. [*Proceeding of the International Conference* "Computational linguistics and intelligent technologies", Bekasovo, May 30–June 3, 2012]. Moscow, RSUH, 2012, vol. 11(18), pp. 674–683. (in Russ.).]

3. Hao Wu, Jun He, Yijian Pei. Scientific Impact at the Topic Level: A Case Study in Computational Linguistics. *Journal of the American Society for Information Science and Technology*. 2010, November, vol. 61, issue 11, pp. 2274–2287.

4. Inyaem U., Haruechaiyasak Ch., Meesad Ph., Tran D. Ontology-Based Terrorism Event Extraction. *Proceedings of the 1st International Conference on Information Science and* Engineering (ICISE 2009), December 26-28, 2009. Nanjing, China, 2009, pp. 912–915.

5. Lefever E., Macken L., Hoste V. Language-Independent Bilingual Terminology Extraction from a Multilingual Parallel Corpus. *Proceedings of the 12th Conference of the European Chapter of the ACL*. Athens, Greece, 2009, pp. 496–504.

6. Mair C. The Corpus-based Study of Language Change in Progress: The Extra Value of Tagged Corpora. *The AAACL/ICAME Conference, May 11–15*. Ann Arbor, MI, 2005.

7. Mannes A., Golbeck J. Building a Terrorism Ontology. *Proceedings of the ISWC Workshop on Ontology Patterns for the Semantic Web, 36*. 2005. http://goo.gl/WXeVVv (23.05.2020).

8. Najgebauer A., Antkiewicz R., Chmielewski M., Kasprzyk R., Prediction of Terrorist Threat on the basis of Semantic Association Acquisition and Complex Network Evolution. *The Journal of Telecommunications and Information Technology*. 2008, vol. 2, pp. 14–20.

9. Pustejovsky J. *Natural Language Annotation for Machine Learning*. 1st ed. O'Reilly Media, 2012, 342 p.

10. Roberts A., Gaizauskas R., Hepple M., Demetriou G., Guo Y., Roberts A., Setzer A. Building a Semantically Annotated Corpus of Clinical Texts. *Journal of Biomedical Informatics*. 2009, vol. 42 (5), pp. 950–966.

11. Sheremetyeva S. Automatic Extraction of Linguistic Resources in Multiple Languages. *Proceedings of NLPCS 2012, 9th International Workshop on Natural Language Processing and Cognitive Science in conjunction with ICEIS 2012*. Wroclaw, Poland, 2012, pp. 44–52.

12. Sheremetyeva S., Zinovyeva A. On Modelling Domain Ontology Knowledge for Processing Multilingual Texts of Terroristic Content. *Communications in Computer and Information Science*, 859. Springer, Cham, 2018, pp. 368–379.

13. Stenetorp P., Pyysalo S., Topic G., Ohta T., Ananiadou S., Jun'ichi Tsujii J. BRAT: a Web-based Tool for NLP-Assisted Text Annotation. *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, *April 23–27, 2012*. Avignon, France, 2012, pp. 102–107.

14. Tenenboim L., Shapira B, Shoval P. Ontology-based Classification of News in an Electronic Newspaper. *Advanced Research in Artificial Intelligence:* ed. by K. Markov, K. Ivanova, I. Mitov. International Book Series "Information Science and Computing", vol. 2. Sofia, Bulgaria, 2008, pp. 89–97.

15. Witschel H.F. Terminology Extraction and Automatic Indexing – Comparison and Qualitative Evaluation of Methods. *Terminology and Content Development – TKE 2005: 7th International Conference on Terminology and Knowledge Engineering*. Copenhagen, Denmark, 2005, pp. 363–374.

16. Witt A., Heid, U., Sasaki, F., Gilles Sérasset. Multilingual Language Resources and Interoperability. *Language Resources & Evaluation*. 2009, March, vol. 43, issue 1, pp. 1–14. DOI: 10.1007/s10579-009-9088-x

Шереметьева Светлана Олеговна, доктор филологических наук, доцент, профессор кафедры лингвистики и перевода, Южно-Уральский государственный университет (Челябинск), sheremetevaso@susu.ru

Бабина Ольга Ивановна, кандидат филологических наук, доцент, доцент кафедры лингвистики и перевода, Южно-Уральский государственный университет (Челябинск), babinaoi@susu.ru

Зиновьева Анастасия Юрьевна, аспирант кафедры лингвистики и перевода, Южно-Уральский государственный университет (Челябинск), bihcwd@bk.ru

Неручева Екатерина Дмитриевна, лаборант НОЦ «Лингво-инновационные технологии» института лингвистики и международных коммуникаций, Южно-Уральский государственный университет (Челябинск), neruchevaekaterina@mail.ru

Поступила в редакцию 22 июня 2020 г.

DOI: 10.14529/ling200408

ON USING THE CASE STUDY METHOD TO CREATE UNIVERSAL RESOURCES FOR CONCEPTUAL ANNOTATION OF MULTILINGUAL CORPORA

S.O. Sheremetyeva, *sheremetevaso@susu.ru*
O.I. Babina, *babinaoi@susu.ru*
A.Yu. Zinoveva, *bihcwd@bk.ru*
E.D. Nerucheva, *neruchevaekaterina@mail.ru*
South Ural State University, Chelyabinsk, Russian Federation

The development of annotated corpora is crucial for the computer technologies meant to process unstructured information (automatic classification, intellectual content and trend analysis, machine learning, machine translation, etc.). It is therefore one of the focuses of international theoretical and applied linguistic research. The key aspect here is the automation of annotation procedures, which, in turn, requires static (linguistic) and dynamic (software) resources that could be reused, at least partially, for annotating multilingual texts of various domains. This paper presents an effort to create such resources for the conceptual type of annotation, one of the most popular and problematic annotation levels, by using the case study method. Conceptual annotation is understood as a kind of semantic annotation focused on solving specific information problems within specific domains. The methodology and results of the study are worked out by applying the case study method to the "Terrorism" domain texts in Russian, English and French. The resources created during the research thus include a universal methodology for the resource development, as well as domain oriented software and linguistic material (multilingual ontology and conceptually annotated corpora in three languages), which can directly be used for augmenting the coverage of annotated corpora in the "Terrorism" domain, developing metrics to resolve conceptual ambiguity, as well as for automating text annotation in other domains and languages. The results of the current research are also of interest for contrastive linguistic studies.

Keywords: conceptual annotation, static and dynamic resources, domain, ontology, multilingualism, terrorism.

Svetlana O. Sheremetyeva, PhD (Habilitation), professor of the Department of Linguistics and Translation Studies, South Ural State University (Chelyabinsk), sheremetevaso@susu.ru

Olga I. Babina, PhD, associate professor of the Department of Linguistics and Translation Studies, South Ural State University (Chelyabinsk), babinaoi@susu.ru

Anastasia Yu. Zinoveva, post-graduate student of the Department of Linguistics and Translation Studies, South Ural State University (Chelyabinsk), bihcwd@bk.ru

Ekaterina D. Nerucheva, laboratory assistant, Research and Education Centre of Innovative Linguistic Technologies, Institute of Linguistics and International Communications, South Ural State University (Chelyabinsk), neruchevaekaterina@mail.ru

Received 22 June 2020

ОБРАЗЕЦ ЦИТИРОВАНИЯ

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 ЮУрГУ. Серия «Лингвистика». – 2020. – Т. 17, № 4. – С. 46–52. DOI: 10.14529/ling200408

FOR CITATION

Sheremetyeva S.O., Babina O.I., Zinoveva A.Yu., Nerucheva E.D. On Using the Case Study Method to Create Universal Resources for Conceptual Annotation of Multilingual Corpora. *Bulletin of the South Ural State University. Ser. Linguistics*. 2020, vol. 17, no. 4, pp. 46–52. (in Russ.). DOI: 10.14529/ling200408