

ПЛАТФОРМА ДЛЯ КОНЦЕПТУАЛЬНОГО АННОТИРОВАНИЯ МНОГОЯЗЫЧНЫХ ТЕКСТОВ

С.О. Шереметьева, О.И. Бабина

Южно-Уральский государственный университет, г. Челябинск, Россия

В настоящей статье представлен опыт разработки программного инструмента ПАНТ – многофункциональной платформы для создания концептуально аннотированных корпусов текстов, необходимого ресурса компьютерных технологий обработки неструктурированной информации. Платформа ПАНТ разработана на основе лингвистической базы знаний, что допускает различные уровни аннотирования. Инструмент многофункционален и имеет гибкие настройки, которые обеспечивают ее адаптацию к многоязычным текстам различных предметных областей и возможность использования как для процесса аннотирования, так и для совершенствования базы знаний. В настоящей статье методология разработки платформы представлена на примере ее версии, ориентированной на концептуальное аннотирование текстов предметной области «Терроризм» на русском, английском и французском языках. Обсуждается применение альфа-версии ПАНТ в рамках реального проекта по созданию аннотационных ресурсов указанной предметной области; при этом демонстрируется сокращение общего времени на создание качественно аннотированных корпусов текстов («золотого стандарта»).

Ключевые слова: концептуальное аннотирование, автоматизация, онтология, многоязычность, предметная область, терроризм.

Введение

Создание корректно аннотированных корпусов текстов является необходимым условием для решения большинства задач обработки языка на основе информационных технологий. При этом аннотирование является одним из наиболее трудоемких и финансово дорогостоящих компонентов многих лингво-информационных исследований [4, 5], поскольку создание аннотаций «золотого стандарта», как правило, осуществляется вручную. Современные инструменты, в той или иной степени автоматизирующие аннотационный процесс, как правило, требуют технических знаний и в основном предлагают поддержку, не выходящую за рамки минимально необходимой функциональности. Поэтому все больше исследователей подчеркивают необходимость создания интуитивно понятных, удобных интерфейсов и разумного применения технологий обработки естественного языка для поддержки, а не подмены человеческих суждений, которые могут сделать использование программного обеспечения более доступными для пользователей, не обладающих техническими навыками (например, экспертов предметных областей) [3, 8]. Снижение финансовых и временных затрат может быть достигнуто и возможностью повторного использования аннотационного инструментария для обработки текстов на разных языках.

В настоящей статье на примере платформы концептуального аннотирования текстов ПАНТ описывается наша попытка реализовать эти возможности. Под концептуальным аннотированием мы понимаем разметку корпусов текстов определенной предметной области тегами контентно

релевантных концептов (например, слово «полицейский» в корпусе предметной области «Терроризм» может иметь концептуальные теги «контрагент теракта», «агент теракта», «жертва теракта» и т. д.). Таким образом, мы, понимая их взаимосвязь, отличаем концептуальное аннотирование от семантического, при котором этому слову приписываются универсальные семантические признаки («одушевленный», «лицо», «профессия» и т. д.). Методология разработки платформы представлена на примере ее альфа-версии, которая ориентирована на концептуальное аннотирование текстов предметной области «Терроризм» на русском, английском и французском языках.

Статья организована следующим образом. В разделе 1 определяются конкретные задачи исследования и общая архитектура платформы концептуального аннотирования ПАНТ. Раздел 2 дает описание многоязычной базы знаний инструмента, настроенной на аннотирование новостных сообщений о террористических актах на русском, английском и французском языках. Раздел 3 представляет программные модули платформы и процедуры ее использования и совершенствования. В заключении рассмотрены итоги и дальнейшие этапы исследования.

1. Постановка задачи

При разработке платформы концептуального аннотирования ПАНТ мы исходили из следующих общих соображений: инструмент должен автоматизировать процесс концептуального аннотирования, допускать повторное применение к корпусам текстов различных предметных областей на различных языках и быть удобным в использовании,

что, в свою очередь, требует решения следующих конкретных задач:

- определение методологии концептуального аннотирования:
 - построение базы знаний, обеспечивающей концептуальную разметку многоязычных текстов набором не зависящих от конкретного языка концептуальных тегов предметной области и поддерживающей решение проблем возможной концептуальной многозначности;
 - алгоритмизация процедуры аннотирования;
- программная реализация методологии в виде платформы концептуального аннотирования с возможностью:
 - настройки ее параметров на различные языки и типы лингвистической информации;
 - администрирования знаний.

При этом архитектура платформы обеспечивает: а) легкий в использовании интерфейс разработчика, выполняющего автоматизированный сбор данных для концептуального аннотирования каждого конкретного языка, и б) легкий в использовании интерфейс аннотатора.

В настоящем исследовании методология концептуального аннотирования основана на онтологическом анализе корпусов текстов с использованием предметно-ориентированной многоязычной онтологии. Это определяет содержание и способы построения базы знаний, также алгоритм реализации аннотационного процесса и архитектуру инструмента ПАНТ, представленную на рис. 1.

2. База знаний

База знаний платформы концептуального аннотирования состоит из трех основных компонентов:

- не зависящие от конкретного естественного языка концептуальные знания предметной области «Терроризм»;

- зависящие от русского, английского и французского языков лингвистические знания предметной области «Терроризм»;
- знания об отражении лингвистических данных каждого их одноязычных корпусов текстов на концепты онтологии.

Не зависящие от естественного языка концептуальные знания области «Терроризм» представлены в предметно-ориентированной многоязычной онтологии, которая содержит 117 универсальных концептов (категорий) предметной области, представленных в корпусах текстов на всех трех языках (русском, английском и французском), из которых 20 категорий отнесено к верхнему уровню и 97 – ко второму и третьему уровням. Детальное описание этой онтологии дано в работе [7].

Лингвистические данные, зависящие от конкретного языка, представляют собой русский, английский и французский лексиконы одно- и многокомпонентных текстовых единиц с морфологической информацией о релевантных для каждого языка и части речи морфологических признаках. Например, число, род, падеж для русских существительных; число и род для французских существительных; число для английских существительных.

Знания об отражении лингвистических данных каждого из одноязычных корпусов текстов на концепты онтологии представлены в виде категоризации контентно-релевантных для области «Терроризм» лексем русского, английского и французского корпусов по классам, соответствующим концептам онтологии. При этом лексиконы, соотнесенные с концептами онтологии, могут содержать пересекающиеся множества лексических единиц, что обусловлено двумя различными лингвистическими явлениями – концептуальной многозначностью и концептуальной синкретичностью. Концептуально многозначной и занесенной в два категориальных класса является, например,

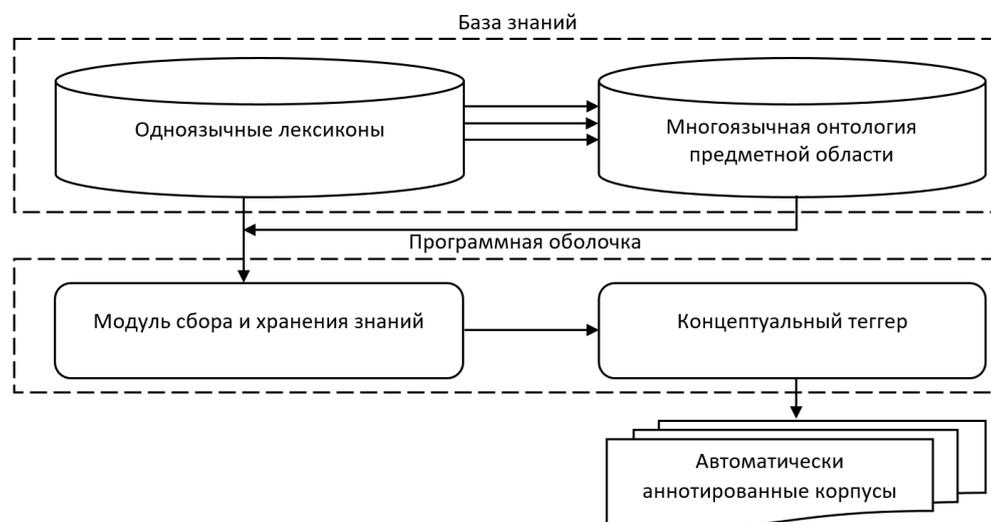


Рис. 1. Архитектура платформы концептуального аннотирования ПАНТ

русскоязычная лексема «автомобиль», которая в корпусе предметной области встречается в трех концептуальных значениях MEANS OF ATTACK (средства атаки), OBJECT OF ATTACK (объект атаки) и CONSEQUENCES. DAMAGE (ущерб от теракта), и при этом в каждом конкретном случае реализуется только в одном из этих значений. Концептуальная синкретичность может иметь место в том случае, если лексическая единица одновременно реализует несколько не противоречащих друг другу концептуальных значений. Чаще всего (но не исключительно) явление концептуальной синкретичности наблюдается у многокомпонентных лексем. Например, русскоязычная лексема «албанский террорист-смертник» сочетает в себе концептуальные значения NATION (национальность), AGENT-TERRORIST (террорист) и TYPE OF ATTACK (тип атаки: теракт через самоубийство) и поэтому тоже занесена в несколько (три) лексиконов, соответствующих указанным категориям.

В связи с наличием явлений концептуальной многозначности и концептуальной синкретичности кроме однозначного соотнесения лексем с онтологическими концептами имеют место следующие отношения между лексическими единицами каждого из языков и концептами онтологии: «один-комногим», «многие-к-одному» и «многие-комногим». Поскольку наша методология аннотирования основана на онтологическом анализе, на практике состоящем в разметке текстов тегами концептов онтологии, на которые отражены лексеммы, то очевидно, что при автоматизированном аннотировании появятся аннотации, состоящие из нескольких концептуальных тегов, и потребует решения проблема устранения концептуальной неоднозначности, что будет темой следующей статьи.

3. Программная реализация ПАНТ

Программная реализация платформы концептуального аннотирования многоязычных текстов определенной предметной области состоит из двух основных блоков (см. рис. 1): 1) модуля сбора и хранения знаний, т. е. 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. Каждый из модулей платформы снабжен пользовательским интерфейсом на английском языке, поскольку он считается международным языком делового общения.

Модуль сбора и хранения знаний использует адаптированные к настоящему исследованию элементы оболочки электронного словаря TransDict [6] и представляет собой ряд электронных одноязычных словарей с перекрестными ссылками и настраиваемой на различные языки конфигурацией, что позволяет кодировать большое количество релевантной для каждого языка и целей исследо-

вания лингвистической информации. Словарная статья каждого одноязычного лексикона предназначена для одного концептуального значения лексеммы в соответствующем корпусе текстов предметной области. Это значит, что, если лексическая единица концептуально многозначна или синкретична и поэтому отображена на несколько концептов онтологии, модуль администрирования знаний содержит для этой единицы столько словарных статей, сколько у лексеммы концептуальных значений. Информация об онтологическом концепте находится в семантической зоне словарной статьи. Также словарная статья имеет морфосинтаксическую зону, где иконически перечислены все входящие в парадигму лексеммы словоформы, каждая из которых сопровождается детальной информацией о части речи и соотнесенных с ней в конкретном языке грамматических признаках: число, род, падеж, залог и т. д. Каждой словоформе присваивается супертег (анкод), который кодирует всю относящуюся к ней концептуальную и морфологическую информацию в виде дерева признаков/тегов. Модуль администрирования знаний имеет многофункциональный и удобный интерфейс, который позволяет контролировать, увеличивать и корректировать базу знаний. Фрагмент основного окна интерфейса, отражающий знания русскоязычного корпуса, показан на рис. 2. В левой части окна (слева направо) даны фрагмент русского списка лексем, коды соответствующих им онтологических концептов: UW – TERRORIST ORGANIZATION (террористическая организация); A – TERRORIST-AGENT (террорист); I – ASSUMPTION (предположение); RW – COUNTER-TERRORISM (контртерроризм) и части речи. Дублирование лексических единиц в левом столбце интерфейса отражает способ представления их концептуальной многозначности в электронной базе знаний. Например, двухкомпонентная лексема «подозреваемый террорист» указана дважды, поскольку она отображена на онтологические концепты A и I. Правая часть окна интерфейса предназначена для показа словарной статьи, соответствующей высвеченной в списке лексемме, с детальной лингвистической информацией и ее кодировкой в электронной базе знаний. На рис. 2 показана словарная статья русской лексеммы «подозреваемый террорист», привязанной к онтологическому концепту ASSUMPTION (предположение). Сверху дана информация о языке, семантике (концепте) и части речи. Ниже перечислены все словоформы парадигмы лексеммы с соответствующими им супертегами, кодирующими информацию о концепте, части речи, падеже и числе члена парадигмы лексической единицы. Супертеги позиционны, концептуальный код – второй по порядку (этот формат наследуется от родительского приложения).

Модуль электронной базы знаний платформы ПАНТ имеет мощную среду для автоматизирован-

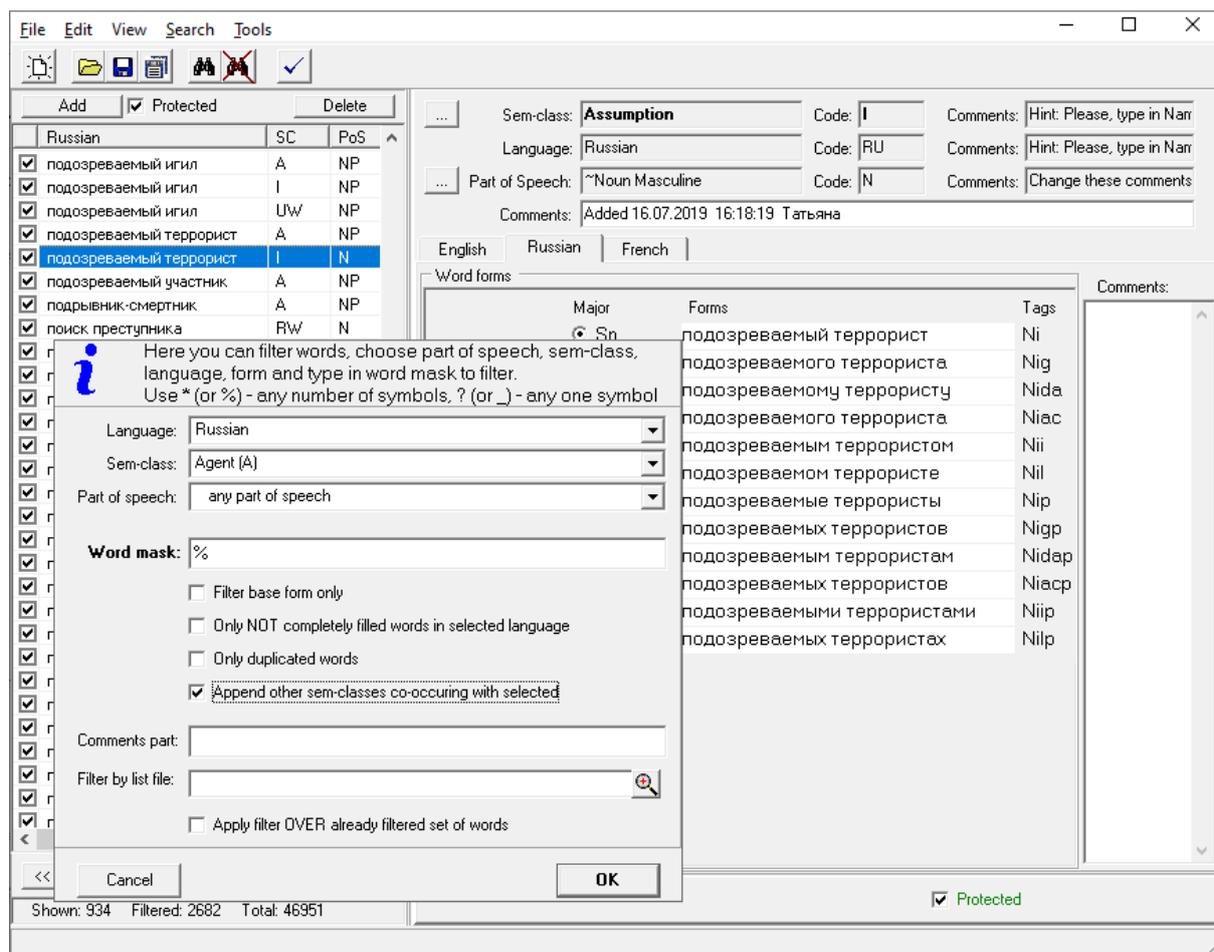


Рис. 2. Фрагмент основного интерфейса электронной базы знаний с всплывающим окном модуля поиска (в центре)

ного сбора, администрирования и кодирования многоязычных лингво-онтологических знаний, доступ в которую обеспечивается через меню интерфейса модуля. Основные блоки этой среды и их функции описаны далее.

Блок настройки конфигурации: а) настраивает и редактирует конфигурацию интерфейса для одноязычного лексикона – можно задать любое количество семантических классов (концептов), языков, частей речи, словоформ и их тегов; любое изменение в настройках автоматически распространяется на соответствующие словарные статьи; б) редактирует значения элементов в словарных статьях.

Импортер данных импортирует списки лексем и/или значения их лингвистических характеристик из внешних файлов, содержащих списки лексем с одинаковыми концептуальными и грамматическими признаками, и/или из приложений. Одним из таких приложений является теггер платформы, из которого в базу знаний ПАНТ можно автоматически импортировать неизвестные лексемы, выдаваемые теггером из аннотируемого текста, с присвоением им соответствующих кодов (тегов).

Еще одним приложением, специально разработанным для пополнения базы знаний платформы, является вспомогательный модуль СЛЕЙВ, который представляет собой пустую программную оболочку, сконфигурированную так же, как соответствующая версия описанной электронной базы знаний, которую мы называем МАСТЕР-базой. СЛЕЙВ может заполняться новыми знаниями в процессе отладки платформы участниками проекта в индивидуальном режиме. СЛЕЙВ функционально отличается от МАСТЕР-базы только тем, что пользователь не может менять настройки конфигурации (наборы языков, концептуальные классы, структуры словарных статей и теги). Знания, введенные в СЛЕЙВы, регулярно импортируются в МАСТЕР-базу, при этом настройки последней устраняют повторяющиеся лексемы. Таким образом, новые знания могут вводиться в электронную базу ПАНТ как в пакетном режиме, так и индивидуально. Всплывающие окна для осуществления этих процедур показаны на рис. 3.

Для введения новой лексической единицы на закладке соответствующего языка через кнопку «Add» основного интерфейса вызывается всплывающее окно с меню концептов для «привязыва-

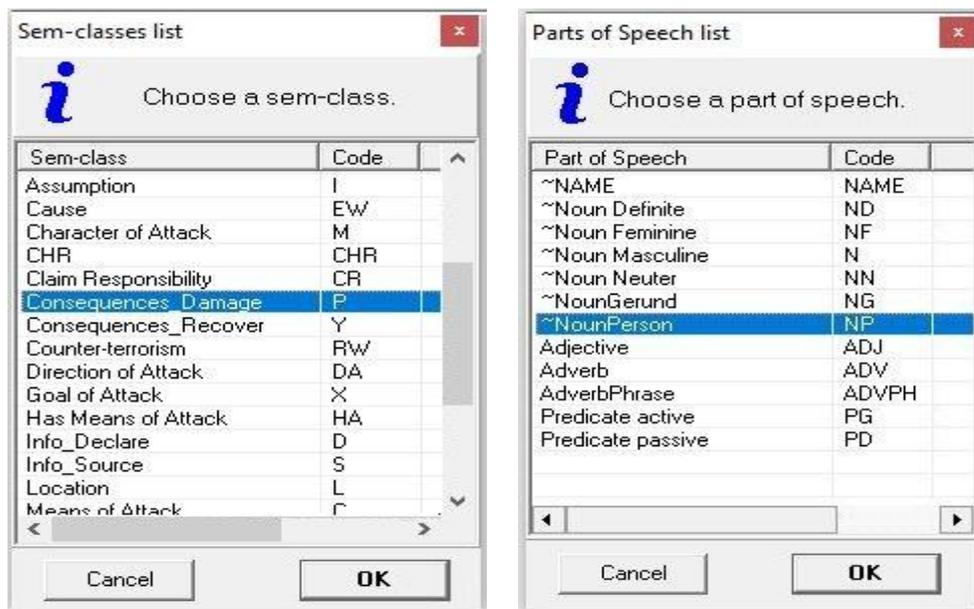


Рис. 3. Всплывающие окна для отнесения новой лексики к концептуальному классу и части речи в конкретном языке

ния» лексики к концептуальному классу, выбор которого вызывает окно с меню частей речи, релевантных для конкретного концепта. Выбор части речи, в свою очередь, вызывает соответствующую введенной информации форму словарной статьи с заполненными полями, передающими информацию о языке, онтологическом концепте, пустыми полями морфологической зоны с уже сгенерированными для каждого поля супертегами, после чего поля морфологической зоны заполняются с помощью морфологического генератора.

Дефолтер автоматически присваивает структуры новых словарных статей и некоторые значения лингвистических признаков новым лексемам в зависимости от выбранных концептуального класса, языка и части речи.

Морфологический генератор автоматически генерирует словоформы парадигмы лексики и заполняет поля морфологической зоны с возможностью их редактирования. Морфологическая генерация в настоящий момент реализована для частей речи с развернутой морфологической парадигмой на английском и русском языках [1, 2].

Блок копирования записей копирует заполнение всех или отдельных полей из одной словарной статьи в другую.

Блок поиска с настраиваемыми по многим параметрам масками позволяет отфильтровать фрагменты базы знаний по заданным маскам. Маски поиска могут учитывать один или сочетание нескольких параметров, например, язык, концепт, часть речи, цепочку букв, списки лексем, комментарии и т. д. На рис. 2 слева показан фрагмент списка русскоязычных лексем, соотнесенных с онтологическим концептом А (террорист) и одновременно с любыми другими концептами в

соответствии с маской, заданной во всплывающем окне поисковика (см. на рис. 2 в центре). Можно также отфильтровывать не полностью заполненные словарные статьи, словарные статьи повторяющихся лексем и т. д. Использование этого блока позволяет выявить лакуны в знаниях, которые необходимо заполнить, откорректировать знания платформы, и дает большие возможности для анализа концептуальных характеристик предметной области. Особенно значимую информацию дает фильтрация знаний платформы по концептуальному классу (в сочетании с другими параметрами поиска или без них). Анализ отфильтрованных по различным маскам списков лексем с сопровождающей их концептуально-лингвистической информацией позволяет получать различные качественные и количественно-статистические данные об универсальных и зависящих от конкретного языка лексико-онтологических характеристиках определенных предметных областей и, например, судить об уровне концептуальной многозначности отдельных лексем или всего корпуса/языка, что, в свою очередь, может использоваться для разработки метрик автоматического снятия концептуальной многозначности при аннотировании и открывает широкие возможности для контрастивных лингвистических исследований.

Экспортер данных экспортирует выгрузки из базы данных во внешние приложения, в частности, в теггер для осуществления тегирования. Кроме этого, знания базы ПАНТ можно экспортировать во внешние файлы, что бывает необходимо для ее анализа.

Теггер на входе получает «сырой» текст на определенном языке и позволяет осуществлять разметку на двух уровнях: концептуальном и концептуально-грамматическом. При концептуальной

размеченной, то человек-аннотатор в первом случае корректировал словарную статью лексемы, а во втором – добавлял лексему со всей необходимой языковой информацией в базу знаний платформы концептуального аннотирования как описано выше (индивидуально или путем заполнения и импортирования СЛЕЙВОВ). Обновленная таким образом платформа ПАНТ использовалась для автоматического аннотирования следующих корпусных фрагментов и т. д. Таким образом, знания обновлялись регулярно и точность автоматического аннотирования повышалась с каждым циклом работы. Точность оценивалась на основе отчетов исследователей о времени, потраченном на постредактирование каждого нового фрагмента корпусов текстов, и количестве новых лексических единиц, добавленных в базу знаний платформы после каждого цикла автоматического аннотирования.

Заключение

В настоящей статье представлена методология разработки и программная реализация компьютерной платформы ПАНТ для концептуального аннотирования текстов предметных областей на различных языках. Платформа ПАНТ имеет мощную программную среду для автоматизированного сбора, администрирования и кодирования многоязычных лингво-онтологических знаний, что обеспечивает широкие возможности для контроля за покрываемостью и корректностью базы знаний, а также качественного и количественно-статистического анализа корпусных лексико-онтологических характеристик. Применение альфа-версии ПАНТ, ориентированной на концептуальное аннотирование текстов предметной области «Терроризм» на русском, английском и французском языках в рамках реального проекта по созданию аннотационных ресурсов указанной предметной области, показало ее эффективность для сокращения трудоемкости, а также временных и финансовых затрат аннотационных работ и показало широкие возможности использования ПАНТ как для продолжения настоящего исследования, одной из основных задач следующего этапа которого является разработка метрик для автоматизированного разрешения концептуальной многозначности, так и для исследований в области социолингвистики и сравнительного языкознания. Многоязычная база

знаний платформы ПАНТ может напрямую использоваться для переводческих работ.

Литература/References

1. Бабина О.И., Дюмин Н.Ю. Автоматизация заполнения морфологической зоны машиночитаемого словаря. Наука ЮУрГУ: Материалы 65-й научной конференции. Челябинск, 15–17 апреля 2013 г. Челябинск, 2013, с. 227–230. [Babina O.I., Dyumin N.Yu. Avtomatizatsiya zapolneniya morfologicheskoy zony mashinochitaemogo slovarya [Automation of the Morphological Zone Compilation in a Machine-Tractable Dictionary]. *Nauka SUSU: materialy 65-y nauchnoy konferentsii*, Chelyabinsk, April 15–17, 2013 [Science of SUSU: Proceedings of the 65th scientific conference. Chelyabinsk, April 15–17, 2013]. Chelyabinsk, 2013, pp. 227–230. (in Russ.)]
2. Babina O.I., Osminin P.G. A Model for Automating the Morphological Generation of Russian Predicates for a Lexicon. *Proceedings of the 10th International Technology, Education and Development Conference*, Valencia, Spain, 7–9 March, 2016, pp. 1320–1328.
3. Druskat S., Bierkandt L., Gast V., Rzymiski Ch., Zipser F. Atomic: an open-source software platform for multi-layer corpus annotation. *Proceedings of the 12th Konferenz zur Verarbeitung natürlicher Sprache (KONVENS 2014)*, Hildesheim, October 2014, pp. 228–234.
4. Pustejovsky J. *Natural Language Annotation for Machine Learning*. 1st edition. O'Reilly Media, 2012, 342 p.
5. 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.
6. Sheremetyeva S.O. Linguistic Models and Tools for Processing Patent Claims. Chelyabinsk, SUSU Publishing center, 2017, 157 p.
7. 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.
8. 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

Поступила в редакцию 22 июня 2020 г.

A PLATFORM FOR KNOWLEDGE ASSISTED CONCEPTUAL ANNOTATION OF MULTILINGUAL TEXTS

S.O. Sheremetyeva, *sheremetevaso@susu.ru*

O.I. Babina, *babinaoi@susu.ru*

South Ural State University, Chelyabinsk, Russian Federation

This paper presents an experience of creating the PAnT software tool, a platform for conceptual annotation of multilingual texts, a prerequisite for automated processing of unstructured information. The PAnT platform is supported by linguistic knowledge, which allows various levels of annotation. The tool is multifunctional and has flexible settings that ensure its adaptation to multilingual texts of various domains. PAnT can be used both for the automation of the annotation process as such and for linguistic knowledge acquisition and administration. In this paper, the platform and the methodology of its development are presented on the example of its version, tuned for the conceptual annotation of the “Terrorism” domain corpora in Russian, English and French. The use of PAnT as part of a real annotation project is discussed demonstrating a reduction in the total time and human effort necessary for annotating texts up to the gold standard.

Keywords: conceptual annotation, automation, ontology, multilingualism, domain, 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*

Received 22 June 2020

ОБРАЗЕЦ ЦИТИРОВАНИЯ

Шереметьева, С.О. Платформа для концептуального аннотирования многоязычных текстов / С.О. Шереметьева, О.И. Бабина // Вестник ЮУрГУ. Серия «Лингвистика». – 2020. – Т. 17, № 4. – С. 53–60. DOI: 10.14529/ling200409

FOR CITATION

Sheremetyeva S.O., Babina O.I. A Platform for Knowledge Assisted Conceptual Annotation of Multilingual Texts. *Bulletin of the South Ural State University. Ser. Linguistics*. 2020, vol. 17, no. 4, pp. 53–60. (in Russ.). DOI: 10.14529/ling200409
