Русско-татарский общественно-политический тезаурус: публикация в облаке лингвистических открытых связанных данных

А.М. Галиева, А.В. Кириллович, Н.В. Лукашевич, О.А. Невзорова, Д.Ш. Сулейманов, Д.Д. Якубова

***Аннотация*—В статье описываются основные принципы и практические аспекты разработки нового двуязычного лексического ресурса – русско-татарского общественно-политического тезауруса и его публикация в облаке Открытых лингвистических связанных данных. Тезаурус создан на основе русскоязычного тезауруса РуТез и представляет собой иерархическую сеть концептов. Каждый концепт имеет уникальное имя и набор текстовых входов, характеризующих его текстовую реализацию. Обсуждается общая методология перевода имен концептов и текстовых входов, учитывающая специфику татарской лексико-семантической системы. Рассматриваются модели и онтологии облака Открытых лингвистических связанных данных и описание построенного ресурса как набора связанных данных в формате LLOD.**

***Ключевые слова*—татарский язык; общественно-политическая терминология; концепт; текстовый вход; открытые связанные данные; лингвистический ресурс; публикация лингвистических данных.**

I. Введение

Формализованное описание лексической системы языка предполагает на определенном этапе создание предметно-ориентированных тезаурусов. В статье описывается проект по разработке двуязычного русско-татарского тезауруса в общественно-политической сфере. Общественно-политическая сфера представляет собой широкую область современных социальных отношений [1], включающую политику и международные отношения, экономические и финансовые аспекты, промышленность, военную сферу, искусство, спорт и т.д.

Создание общественно-политических тезаурусов представляет для языка большой интерес по ряду причин. Во-первых, общественно-политическая сфера помимо терминов соответствующей предметной области включает общеупотребительную лексику, которая может встречаться в текстах любой области, обычно неоднозначна и представляет сложность для описания. Логично, что в отличие от общеупотребительной лексики терминам общественно-политической сферы гораздо в меньшей степени свойственна многозначность. Во-вторых, общественно-политические тезаурусы могут, при необходимости, быть расширены и включать в себя лексику других предметных областей. В-третьих, общественно-политический тезаурус может использоваться как формализованный лингвистический ресурс в различных приложениях для автоматической обработки новостных документов, правовых актов или сообщений в социальных сетях.

В начале XXI века появилась тенденция к публикации лингвистических данных в соответствии с принципами Linked Open Data - LOD (Открытые связанные данные), что привело к возникновению быстрорастущего облака Лингвистических открытых связанных данных - Linguistic Linked Open Data (LLOD) [2-3].

Размещение лингвистических ресурсов в LLOD имеет множество преимуществ, среди которых:

1) Структурная и концептуальная интероперабельность.
2) Возможность совместного использования нескольких лингвистических ресурсов для решения общей задачи. Примером такой задачи может быть задача нахождения упоминаний населенных пунктов в корпусе текстов, которая требует совместного использования корпуса и тезауруса.
3) Возможность совместного использования лингвистических ресурсов вместе с экстралингвистическими ресурсами и программными инструментами из облака Открытых связанных данных (LOD). Примерами такого совместного использования могут быть: (1) Вербализация фактов из баз знаний и онтологий в

Статья получена 23 октября 2017. Данное исследование было проведено при поддержке Российского научного фонда, проект № 16-18-02074.

А.М. Галиева, Академия наук Республики Татарстан (e-mail: amgalieva@gmail.com)

А.В. Кириллович, Казанский федеральный университет (e-mail: alik.kirillovich@gmail.com)

Н.В. Лукашевич, Московский государственный университет им. Ломоносова (e-mail: louk_nat@mail.ru)

О.А. Невзорова, Академия наук Республики Татарстан, Казанский федеральный университет (e-mail: onevzoro@gmail.com)

Д.Ш. Сулейманов, Академия наук Республики Татарстан, Казанский федеральный университет (e-mail: dvdt.slt@gmail.com)

Д.Д. Якубова, Казанский федеральный университет, Академия наук Республики Татарстан (e-mail: suleymanovad@gmail.com)

виде текста на естественном языке. (2) Генерация SPARQL-запроса к набору данных на основе вопроса на естественном языке. (3) Семантическая разметка и семантический поиск документов, опубликованных в облаке LOD.

4) Поддержка мощной экосистемы Semantic Web, которая включает в себя хранилища данных, логические системы вывода и различные приложения.

II. Обзор близких исследований

Существует несколько направлений в разработке двуязычных и многоязычных тезаурусов. Первое из них связано с созданием многоязычных поисковых тезаурусов, предназначенных для международного использования. Такие тезаурусы обычно создаются международными организациями, располагающими большими коллекциями документов на нескольких языках, такими как Парламент Европейского союза или Организация Объединенных Наций. Разработка многоязычных поисковых тезаурусов регулируется национальными и международными стандартами [4, 5, 6].

Многоязычный тезаурус Организации Объединенных Наций ЮНБИС [7] и Тезаурус по ядерной безопасности [8] имеют схожие принципы построения. При создании обоих ресурсах сначала вводится дескриптор на английском языке, затем добавляются соответствующие ему уникальные дескрипторы других языковых вариантов тезауруса, и далее для каждого дескриптора могут быть указаны менее предпочтительные варианты наименования.

Второе направление в разработке многоязычных тезаурусов связано с развитием лексических систем естественных языков и созданием многоязычной модели представления и применения языковых ресурсов. Один из популярных типов многоязычных ресурсов построен на принципах тезауруса Princeton WordNet (далее подобные ресурсы называются wordnet) [9-11].

Тезаурус РуТез [12] можно причислить к лингвистическим онтологиям. Основными его единицами являются концепты, которые вводятся в тезаурус на основе значений существующих слов и выражений. Каждому концепту соответствует уникальное имя, а также набор текстовых входов, характеризующих его текстовую реализацию.

При переводе тезаурусов важно обеспечить сохранение эквивалентности отношений между единицами тезауруса. Двуязычные словари, используемые для перевода, должны быть дополнительно проверены на достоверность значений с помощью толковых словарей на целевом языке, поскольку двуязычными словарями могут даваться варианты перевода с более широким или узким значением, чем требуемое.

Тезаурус РуТез был взят нами за основу для создания татарского общественно-политического тезауруса по ряду причин. Во-первых, РуТез содержит различные языковые единицы (термины, состоящие из одного или нескольких слов, вербализующие соответствующие концепты) для русского языка, что предоставляет богатый материал для создания двуязычного ресурса. Кроме того, русские и татары сосуществуют в одном и том же геополитическом пространстве, что определяет взаимное проникновение лексических систем и существование общих черт в ментальных пространствах этих языков.

Современный подход к разработке лексических ресурсов предполагает открытость ресурсов и их присутствие в облаке связанных данных (LOD). Лексический ресурс в LOD представляется как набор структурированных данных, описанных в специальном формате, который допускает взаимное использование данных и создание перекрестных ссылок. Публикация русско-татарского тезауруса в облаке LOD является для нас одной из актуальных задач.

В настоящее время в облаке LLOD уже представлены лингвистические ресурсы для многих языков, среди которых: тезаурус WordNet [13-15, 32], английская и немецкая версии Викисловаря [14, 16], FrameNet [14], VerbNet [14], англоязычный корпус BROWN, а также некоторые тезаурусы для узкоспециализированных областей (EuroVoc [17], AgroVoc [18-19], TheSoz [20], Library of Congress Subject Headings [21]).

Лингвистические ресурсы на русском языке и языках народов России в облаке LLOD практически не представлены. Существует несколько проектов в этой области, но их результаты крайне ограничены. Во-первых, можно отметить проект RTLOD [22], в рамках которого с использованием некоторых технологий LLOD были опубликованы тезаурусы русского языка РуТез-Lite, Universal Dictionary of Concepts и YARN. Однако опубликованные в рамках этого проекта ресурсы не соответствуют стандартам LLOD: во-первых, ресурсы не содержат разрешимые URL, во-вторых, данные ресурсы не связаны с внешними ресурсами из облака LLOD. Во-вторых, опубликованы в облаке LLOD отдельные небольшие многоязычные лексические ресурсы для узких предметных областей, такие как сельскохозяйственный тезаурус AgroVoc, тезаурус социальных наук TheSoz или онтология математического знания OntoMathPRO [23-24]. Большинство данных ресурсов полностью соответствуют стандартам LLOD, но покрывают лишь ограниченные предметные области. В третьих, проект BabelNet [25-26], представляющий собой сверхбольшой многоязычный тезаурус, автоматически сгенерированный на базе WordNet и Википедии. Данный ресурс покрывает общую предметную область и полностью соответствует стандартам LLOD. Однако русские лексические единицы BabelNet были получены с помощью системы машинного перевода и не были проверены вручную. Корпуса и грамматические словари для русского языка в облаке LLOD не представлены ни в каком виде; лингвистические ресурсы для татарского языка в облаке LLOD не представлены вообще.

III. Принципы моделирования и методология построения русско-татарского тезауруса

Для установления принципов моделирования

двуязычного ресурса необходимо определить: 1) формат ресурса; 2) объем и качество лексического материала; 3) методы выявления и разрешения конфликтных ситуаций в процессе перевода [27].

Поскольку русско-татарский тезаурус разрабатывается в формате тезауруса РуТез, структура татарской части концептуально повторяет структуру РуТез, т.е. в ее основу был положен перечень концептов русского тезауруса. Концептуальная структура РуТез также определяет направление для работы с концептами татарской компоненты. Так, структура отношений между концептами в татарской части повторяет русскую; однако в тех случаях, когда семантические отношения между татарскими контекстами не соответствуют отношениям между единицами РуТез, они пересматриваются и могут быть установлены иначе.

Важной задачей при разработке татарской части двуязычного тезауруса является передача специфики татарской лексико-семантической системы. Это достигается тремя основными способами:

1. Поиск эквивалентов (соответствующих терминов), которые в реальности используются в татарском языке в качестве переводов русских терминов.

Поиск эквивалентов перевода и адекватных соответствий, фактически используемых в текстах, часто является трудоемкой и затратной по времени задачей, поскольку существующие двуязычные русско-татарские словари общеупотребительной и специализированной лексики явно устарели (например, [1, 12, 13]), прежде всего, по следующим двум причинам:

1) в словарях отсутствуют современные общественно-политические термины, широко используемые в России;
2) словари содержат не актуальные варианты перевода терминов на татарский язык, а использовавшиеся в советское время.

Таким образом, существующие двуязычные татарско-русские словари, в том числе специализированные общественно-политические словари, могут быть использованы в процессе разработки русско-татарского тезауруса с определенной оговоркой из-за наличия устаревших лексических данных и отсутствия лексических единиц, являющихся актуальными в настоящее время. Это обстоятельство значительно замедляет процесс перевода, так как разработчики вынуждены прорабатывать вручную большое количество медиатекстов и текстов официальных документов в поисках подходящих татарских терминов.

2. Добавление новых концептов, отражающих важную для общественно-культурной жизни татарского общества тематику, которые не представлены или недостаточно представлены в тезаурусе РуТез (например, концепты, связанные с исламом, социальной иерархией в восточных странах, татарскими этнокультурные явления и т.д.).

3. Внесение в тезаурус концептов, вербализованных в татарском, но не в русском языке. Например, в русском языке существуют лексемы, соответствующие концептам ПАСЫНОК и ПАДЧЕРИЦА, однако отсутствует лексема, называющая неродного ребенка одного из супругов независимо от пола (как stepchild в английском).

Методология создания татарской части тезауруса включает следующие этапы.

На первом этапе имена концептов и текстовые входы тезауруса РуТез переводятся на татарский язык. Источниками перевода являются двуязычные словари общего назначения, специальные словари [28-30], татарские тексты соответствующих тем и русско-татарские параллельные тексты.

При выборе имени концепта основными являются следующие критерии:

1) ясность значения имени;
2) использование нейтральных и общеупотребительных лексических единиц;
3) при наличии нескольких вариантов выбирается наиболее точно соответствующий исходному наименованию, и, по возможности, наиболее частотный и официально используемый вариант.

В ряде случаев в качестве имени русского или татарского концепта используется многозначное слово в конкретном значении, в таком случае в скобках или через запятую дается необходимое уточнение в соответствующем языке (примеры в Таблице 1).

Таблица 1. Имена концептов для многозначных слов

Пример русского имени концепта	Пример татарского имени концепта	Комментарий
Руководство, начальство	җитәкчелек	В русской части представлен синоним для пояснения значения многозначного слова.
Обучение, учебная деятельность	Уку/укыту	В русской части представлен синоним для пояснения значения многозначного слова; татарский перевод представляет собой конверсивы: учится – обучать.
руководство (учебное пособие)	кулланмалык	В скобках представлено пояснение, уточняющее значение русского многозначного слова
население	халык (бер территориядә яшәүчеләр)	В скобках представлено пояснение, уточняющее значение татарского многозначного слова

Когда в качестве имени концепта выбирается многозначное слово, мы указываем в скобках, какое именно значение является актуальным в данном случае. Из-за различий в семантической структуре русских и татарских слов эти комментарии и уточнения, как правило, различаются в русской и татарской частях тезауруса.

Перевод текстовых входов предполагает приведение максимального количества возможных вариантов наименования концепта в современном татарском языке. Неустоявшийся характер татарской терминологии обусловливает наличие большого

количество вариантов терминов в татарском языке. Сосуществование разных центров разработки татарских терминов и отсутствие координации между ними приводит к нестабильности и дисбалансу терминологии, когда одновременно используются различные терминологические варианты разного происхождения и структуры для наименования одного концепта.

Соответственно, важной задачей становится поиск всех синонимичных наименований и вариантов с целью их внесения в тезаурус в качестве текстовых входов соответствующего концепта.

В таблице 2 представлены татарские синонимичные единицы разной структуры, соответствующие русскому термину *теневая экономика*.

Таблица 2. Особенности синонимии текстовых входов концептов (на примере)

Концепт	
Обозначение концепта на русском языке	Обозначение концепта на татарском языке
ТЕНЕВАЯ ЭКОНОМИКА	КҮЛӘГӘДӘГЕ ИКЪТИСАД
Текстовые входы	
Текстовые входы на русском языке	Текстовые входы на татарском языке
Теневая экономика	күләгәдәге икътисад, күләгәле икътисад, күләгә икътисады, күләгә икътисад, күләгәдәге бизнес, күләгәдәге сектор, күләгә секторы

При приведении текстовых входов мы стремимся к максимальному охвату существующих вариантов, т.е. параллельных наименований. Поскольку языковая ситуация в Республике Татарстан нестабильна, параллельные деноминации могут использоваться для широкого круга явлений.

Структура текстовых входов в двух языках для значительного числа концептов не совпадает, так, если в русском языке в качестве синонимической номинации часто используются различные случаи аббревиации, то в татарском языке используется значительное количество синонимов (см. таблицу 3).

Таблица 3. Аббревиация текстовых входов концептов (на примере)

Концепт	
Обозначение концепта на русском языке	Обозначение концепта на татарском языке
МАТЕРИАЛЬНАЯ ПОМОЩЬ	МАТЕРИАЛЬ ЯРДӘМ
Текстовые входы	
Текстовые входы на русском языке	Текстовые входы на татарском языке
Материальная помощь, матпомощь	материаль ярдәм, матди ярдәм

IV. Модели и онтологии облака Открытых лингвистических связанных данных

Для представления простейших лексических ресурсов используется онтология SKOS [31-32]. Данная онтология позволяет описывать набор понятий, связанных иерархическими отношениями. Каждое понятие может иметь текстовое представление на разных языках.

Для представления сложных лексических ресурсов предназначена онтология Lemon [33-35]. Данная онтология основана на международном стандарте ISO 24613:2008 «Language resource management - Lexical markup framework (LMF)» [36]. Базовыми элементами онтологии являются: лексикон, лексическая единица, форма лексической единицы, смысл лексической единицы и понятия из онтологии предметных областей. Данная онтология позволяет описывать:

1) Формы слова с лингвистической информацией об этой форме (род, число, падеж, время и т.д.).
2) Декомпозицию сложной лексической единицы (например, словосочетания) на простые единицы.
3) Синтаксическое поведение лексической единицы.
4) Привязку лексической единицы к обозначаемому ей понятию из внешней онтологии.
5) Отношения между разными лексическими единицами / лексическими формами (например, этимологические отношения, отношение между полной формой слова и аббревиатурой и т.д.).
6) Семантические отношения (антоним, перевод, мероним, гипоним, гипероним и т.д.) между смыслами
7) Морфологические шаблоны для лексической единицы.
8) Для описания языковых категорий (род, число, падеж, время, прямой объект, косвенный объект, синоним, антоним и т.д.) в Lemon используется внешняя онтология LexInfo.

Для описания языковых категорий в LLOD используются различные онтологии, привязанные к регистру языковых категорий IsoCat [37-39], который является реализацией международного стандарта ISO 12620:2009 «Terminology and other language and content resources — Specification of data categories and management of a Data Category Registry for language resources» [40]. Примерами данных онтологий являются LexInfo [41] и ubyCat [14].

OLiA [42-43] − это семейство онтологий, предназначенных для того, чтобы связать друг с другом множество различных систем языковых категорий и аннотационных схем с помощью промежуточной онтологии. OLiA содержит следующее части:

1) OLiA Reference Model − таксономия категорий данных, на которую отображаются все остальные аннотационные схемы.
2) Annotation Models − онтологическое представление разных систем языковых категорий и аннотационных схем (в том числе аннотационные схемы для разных корпусов, например, корпуса BROWN).
3) OLiA Linking Models − отображение аннотационных схем на таксономию Reference Model.
4) External Reference Models − отображение внешних терминологических репозиториев, которые уже имеют онтологическое представление, на Reference Model. Примеры: GOLD, ISOcat.

Для разметки корпусов используются онтологии NLP Interchange Format (NIF) [44] и Web Annotation [45].

V. Описание LLOD-ресурса

A. *Общий обзор*

Опишем построенный ресурс (русско-татарский общественно-политический тезаурус) как набор связанных данных в формате LLOD.

Текущая версия ресурса включает в себя: (1) Иерархию независимых от языка концептов. (2) Лексические единицы на татарском языке, привязанные к обозначаемым ими концептам и (3) Канонические формы лексических единиц.

Ресурс представлен на базе онтологий SKOS [31-32], Lemon[33], LexInfo [41] и PROV [49]. Кроме того, мы создали собственную онтологию, которая описывает те классы и отношения, которые не могут быть представлены с использованием существующих онтологий. При этом, классы не определялись с нуля, а были определены на основе существующих онтологий. Такой подход облегчает понимание семантики онтологии и позволяет на данном этапе использовать ресурс неподготовленному пользователю. В тех случаях, когда это возможно, мы привязали единицы нашей онтологии к регистру языковых категорий IsoCat [39].

B. *Иерархия концептов*

Иерархия концептов нового тезауруса отражает иерархию концептов тезауруса РуТез. Концепты рассматриваются как «единицы» мысли и не зависят от конкретного языка. Описание концепта содержит его однозначное название, глоссы, а также ссылку на ресурс-источник (откуда этот концепт был получен). В текущей версии ресурсом-источником является тезаурус РуТез. Для связи концептов определены следующие типы отношений:

1) Отношение гиперонимии, которое является объединением традиционных онтологических отношений isa и instanceOf. Примером отношения isa является отношение между концептами *Европейское государство* и *Государство*, примером отношения instanceOf − отношение между концептами *Польша* и *Европейское государство*.
2) Отношение гипонимии (обратное к отношению гиперонимии). Концепт может иметь несколько гипонимов и гиперонимов.
3) Отношения холонимии и меронимии (целое и часть), которые являются обратными друг к другу.
4) Отношения направленной ассоциации: ассоциация1 и ассоциация2, которые выражают отношение внешней онтологической зависимости между концептами [50]. Отношение ассоциация1 связывает концепт с1 и концепт с2, если с1 онтологически зависит от с2 и не является его частью. Примером такого отношения является отношение между концептами Автогонки и Автомобиль. Отношение ассоциация2 является обратным к отношению ассоциация1.
5) Отношение ненаправленной ассоциации.

Концепты ресурса представлены как экземпляры класса `skos:Concept`. Название концепта выражается с помощью свойства `rdfs:label`, а глосс с помощью свойства `skos:definition`.

Отношение между концептом и ресурсом-источником выражается с помощью свойства `prov:wasDerivedFrom` онтологии PROV.

Обычно в тезаурусах для выражения отношения гиперонимии и гипонемии используется свойства `skos:broader` и `skos:narrower`. Однако это отношение имеет недоопределенную семантику и может также выражать отношение часть и целое. В связи с этим для отношений гипероними и гипонемии мы определили собственные свойства `ruthes-ontology:hypernym` и `ruthes-ontology:hyponym`, которые являются подсвойствами свойств `skos:broader` и `skos:narrower`.

Для выражения отношений холонимии и меронимии между концептами в тезаурусах BabelNet [26], а также в RDF-представлении РуТеза из проекта RTLOD [22] использовались свойства `lexinfo:meronymTerm` и `lexinfo:holonymTerm` (их подсвойства). Однако это не совсем корректно, так как данные свойства являются подсвойством свойства `lemon:senseRelation`, чьей областью определения и областью значения являются не концепты, а лексические смыслы. В официальном LOD-представлении Wordnet [15], а также в RDF-представлении Wordnet от W3C [13] отношения холонимии и меронимии определялись с нуля. Недостатком такого подхода является проблема интероперабельности. В нашем ресурсе выбран промежуточный вариант, и для выражения этих отношений определены собственные свойства `ruthes-ontology:holonym` и `ruthes-ontology:meronym`, которые, однако, являются подсвойствами стандартных свойств `skos:broader` и `skos:narrower`.

То, что свойства `ruthes-ontology:hyponym` и `ruthes-ontology:meronym` являются подсвойствами общего свойства `skos:narrower` является удобным при решении задачи расширения поискового запроса, где требуется для заданного концепта найти все нижестоящие концепты, как гипонимы, так и меронимы.

Для двух отношений направленной ассоциации и отношения ненаправленной ассоциации мы определили собственные свойства `ruthes-ontology:association1`, `ruthes-ontology:association2` и `ruthes-ontology:association` соответственно, которые являются подсвойствами стандартного свойства `skos:related`.

Фрагмент сети иерархии концептов представлен на рис. 1.

C. *Лексические единицы, формы и смыслы*

Лексические единицы бывают двух видов: единичные слова и многословные выражения.

Описание лексической единицы содержит ее лемму, часть речи, а также ссылку на ресурс, из которого эта лексическая единица была получена.

В текущей версии лексическая единица связана только со своей канонической формой. В дальнейшем планируется добавление грамматических форм с другими значениями грамматических показателей.

Лексические единицы обозначают концепты. Одна и та же лексическая единица может обозначать несколько

концептов, и один и тот же концепт может обозначаться несколькими лексическими единицами. Лексические единицы связаны с концептами через лексические смыслы, которые можно рассматривать как вариант лексического разрешения этой единицы.

```
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#>.
@prefix skos: <http://www.w3.org/2004/02/skos/core#>.
@prefix lemon: <http://www.lemon-model.net/lemon#>.
@prefix lexinfo: <http://www.lexinfo.net/ontology/2.0/lexinfo#>.
@prefix prov: <http://www.w3.org/ns/prov#>.
@prefix ruthes-ontology: <http://lod.ruthes.org/ontology#>.

<http://lod.ruthes.org/resource/concept/445> #Концепт Государство
  a skos:Concept;
  rdfs:label "ГОСУДАРСТВО"@ru;
  rdfs:label "ДӘҮЛӘТ"@tt;
  ruthes-ontology:hypernym
    <http://lod.ruthes.org/resource/concept/149182>;
  ruthes-ontology:hyponym
    <http://lod.ruthes.org/resource/concept/2193>, ...;
  ruthes-ontology:meronym
    <http://lod.ruthes.org/resource/concept/2052>, ...;
  ruthes-ontology:association2
    <http://lod.ruthes.org/resource/concept/105371>, ... ;
  lemon:isReferenceOf
    <http://lod.ruthes.org/resource/sense/445-RU-государство-n>,
    <http://lod.ruthes.org/resource/sense/445-RU-держава-n>,
    <http://lod.ruthes.org/resource/sense/445-TT-дәүләт-n>, ... .

<http://lod.ruthes.org/resource/sense/445-RU-государство-n> #Смысл, связывающий концепт и лекс. единицу
  a lemon:LexicalSense;
  lemon:reference <http://lod.ruthes.org/resource/concept/445>;
  lemon:isSenseOf <http://lod.ruthes.org/resource/entry/RU-государство-n>.

<http://lod.ruthes.org/resource/entry/RU-государство-n>  #лексическая единица “Государство”
  a lemon:Word;
  rdfs:label "ГОСУДАРСТВО"@ru;
  lexinfo:partOfSpeech lexinfo:noun;
  lemon:sense <http://lod.ruthes.org/resource/sense/445-RU-государство-n>;
  lemon:canonicalForm
    <http://lod.ruthes.org/resource/form/RU-государство-n-neut-sg-nom>;
  lemon:otherForm
    <http://lod.ruthes.org/resource/form/RU-государства-n-neut-sg-gen>,
    <http://lod.ruthes.org/resource/form/RU-государств-n-neut-pl-gen>, ... .

<http://lod.ruthes.org/resource/form/RU-государство-n-neut-sg-nom> # каноническая форма
  a lemon:Form;
  rdfs:label "государство"@ru;
  lemon:writtenRep "ГОСУДАРСТВО"@ru;
  lexinfo:partOfSpeech lexinfo:noun;
  lexinfo:gender lexinfo:neuter;
  lexinfo:case lexinfo:nominative;
  lexinfo:number lexinfo:singular.

<http://lod.ruthes.org/resource/form/RU-государств-n-neut-pl-gen> #Форма во мн. числе и род. падеже
  a lemon:Form;
  rdfs:label "государств"@ru;
  lemon:writtenRep "ГОСУДАРСТВ"@ru;
  lexinfo:partOfSpeech lexinfo:noun;
  lexinfo:gender lexinfo:neuter;
  lexinfo:case lexinfo:genitive;
  lexinfo:number lexinfo:plural.

<http://lod.ruthes.org/resource/sense/445-TT-дәүләт-n> #Смысл, связывающий концепт с лекс. единицей “Дәүләт”
  a lemon:LexicalSense;
  lemon:reference <http://lod.ruthes.org/resource/concept/445>;
  lemon:isSenseOf <http://lod.ruthes.org/resource/entry/TT-дәүләт-n>.

<http://lod.ruthes.org/resource/entry/TT-дәүләт-n> #лексическая единица “Дәүләт”
  a lemon:Word;
  rdfs:label "ДӘҮЛӘТ"@tt;
  lexinfo:partOfSpeech lexinfo:noun;
  lemon:sense <http://lod.ruthes.org/resource/sense/445-TT-дәүләт-n>;
  lemon:canonicalForm <http://lod.ruthes.org/resource/form/TT-дәүләт-n-sg-nom>.

<http://lod.ruthes.org/resource/form/TT-дәүләт-n-sg-nom> #каноническая форма
  a lemon:Form;
  rdfs:label "дәүләт"@tt;
  lemon:writtenRep "ДӘҮЛӘТ"@tt;
  lexinfo:partOfSpeech lexinfo:noun;
  lexinfo:case lexinfo:nominative;
  lexinfo:number lexinfo:singular.
```

Рис. 1. Концепт *Государство* и его смыслы, лексические единицы и их формы на русском и татарском языках

Единичные слова в данном ресурсе представлены как экземпляры класса `lemon:Word`. Многословные выражения представлены как экземпляры класса `lemon:Phrase`. Лемма лексической единицы выражается как свойство `rdfs:label`, часть речи выражается с помощью свойства `lexinfo:partOfSpeech`. Отношение между лексической единицей и ресурсом-источником выражается с помощью свойства `prov:wasDerivedFrom`.

Лексическая единица связана с главной формой свойством `lemon:canonicalForm`, и с неглавными – свойством `lemon:otherForm`.

Формы лексической единицы в данном ресурсе представлены как экземпляры класса `lemon:Form`.

Смыслы лексической единицы в данном ресурсе представлены как экземпляры класса `lemon:LexicalSense`. Лексическая единица связана со смыслом с помощью свойства `lemon:sense`, смысл связан с концептом с помощью свойства `lemon:reference`.

Пример лексической единицы и ее связей с концептом и формами изображен на рис. 1.

D. Публикация в Вебе

Ресурс опубликован в Вебе и доступен через:

1) разрешимые ссылки по адресу: http://lod.ruthes.org;
2) SPARQL-точку доступа по адресу: http://lod.ruthes.org/sparql;
3) RDF-файл по адресу: http://lod.ruthes.org/download.

Доступ к ресурсу через разрешимые ссылки осуществляется с поддержкой принципов content negotiation. Когда URL запрашивает web-браузер, происходит перенаправление на web-страницу с HTML-представлением соответствующего элемента, а когда Semantic Web агент, то на страницу с RDF-представлением.

VI. Заключение

В статье описаны основные подходы и решения, применяемые при разработке нового лексического ресурса – русско-татарского общественно-политического тезауруса. Новый тезаурус разработан на основе тезауруса РуТез и, в общем, повторяет его концептуальную структуру, однако в ряде случаев допускаются структурные изменения, связанные с добавлением отдельных концептов и новых тематических веток, отражающих национальную специфику татарского языка.

В настоящее время размер русско-татарского общественно-политического тезауруса составляет 6000 концептов, и ведется работа над расширением объема тезауруса как в части концептов, так и в части текстовых входов. При разработке тезауруса решаются такие задачи, как перевод имен концептов и текстовых входов РуТез на татарский язык, пересмотр перечней концептов и отношений между ними с учетом лексических особенностей татарского языка, добавление новых концептов и установление связей между ними и фиксирование всех фактически используемых текстовых входов на татарском языке.

Построение татарской компоненты тезауруса позволяет уточнить (и в какой-то мере создать) новый слой терминологии в общественно-политической области, а также изучить влияние русского языка на процессы формирования терминов на татарском языке.

Анализ лексического материала тезауруса позволяет сделать вывод о том, что современная татарская терминология находится в процессе развития. При этом различные терминологические центры предлагают разные терминологические решения, что приводит к сосуществованию многочисленных синонимичных деноминаций, обозначающих актуальные реалии современной общественно-политической сферы. Стандартизация татарской терминологии позволит обеспечить системность, и важным шагом на пути к стандартизации является создание и публикация в открытом доступе терминологических баз данных. Эта задача решена на основе моделей и технологий публикации лингвистических данных в облаке Открытых связанных данных. Русско-татарский тезаурус в общественно-политической сфере представлен в формате LLOD как набор связанных данных и опубликован в облаке Открытых лингвистических связанных данных. Данный результат является первым для двуязычных ресурсов языков народов РФ и отработанная технология подготовки к публикации в LLOD лингвистических ресурсов позволяет применять ее для различных лексикографических ресурсов, разработанных для языков РФ.

Библиография

[1] Loukachevitch, N. and Dobrov B. 2015. The Sociopolitical Thesaurus as a resource for automatic document processing in Russian. In Terminology. Vol. 21, 2, 237-262.

[2] Chiarcos, C., McCrae, J., Cimiano, P., Fellbaum, C.: Towards open data for linguistics: Linguistic Linked Data. In: Oltramari, A., Vossen, P., Qin, L., Hovy, E. (eds.) New Trends of Research in Ontologies and Lexical Resources. Theory and Applications of Natural Language Processing, pp. 7–25. Springer, Heidelberg (2013). doi:10.1007/978-3-642- 31782-8_2

[3] McCrae, J.P., et al.: The open linguistics working group: developing the Linguistic Linked Open Data cloud. In: Calzolari, N., et al. (eds.) Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC 2016), pp. 2435–2441 (2016)

[4] ГОСТ 7.24-2007 Межгосударственный стандарт. Тезаурус информационно-поисковый, многоязычный. – М.: Стандартинформ, 2007.

[5] ISO 25964-1:2011. Information and documentation – Thesauri and interoperability with other vocabularies – Part 1: Thesauri for information retrieval. Available: http://www.iso.org/iso/catalogue_detail.htm?csnumber=53657

[6] ISO 5964-1985. Documentation – Guidelines for the establishment and development of multilingual thesauri. Available: http://www.iso.org/iso/catalogue_detail.htm?csnumber=12159

[7] United Nations. UNBIS Thesaurus. English version. 1986. New York, Dag Hammarskjold Library of the United Nations.

[8] INIS Thesaurus. English version. IAEA-INIS Reference Series. 2016.

[9] Miller, G.A. 1995. WordNet: A lexical database for English. In Communications of the ACM. Vol. 38, 11, 39-41.

[10] MultiWordNet. 2004. Available: http://multiwordnet.fbk.eu/english/home.php

[11] WordNet: An Electronic Lexical Database. 1998. ed. Fellbaum, C. Cambridge, MIT Press.

[12] Loukachevitch, N. and Dobrov B. 2015. The Sociopolitical Thesaurus as a resource for automatic document processing in Russian. In Terminology. Vol. 21, 2, 237-262.

[13] van Assem, M., Gangemi, A., Schreiber, G.: Conversion of WordNet to a standard RDF/OWL representation. In: Calzolari, N., et al. (eds.)

Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC 2006), pp. 237–242 (2006).
[14] Eckle-Kohler, J., McCrae, J.P., Chiarcos, C.: lemonUby - a large, interlinked, syntactically-rich lexical resource for ontologies. Semant. Web 6(4), 371–378 (2015). doi:10.3233/SW-140159
[15] McCrae, J.P., Fellbaum, C., Cimiano, P.: Publishing and linking WordNet using Lemon and RDF. In: Chiarcos, C., et al. (eds.) Proceedings of the 3rd Workshop on Linked Data in Linguistics (LDL-2014) (2014).
[16] Sérasset, G.: DBnary: Wiktionary as a Lemon-based multilingual lexical resource in RDF. Semant. Web 6(4), 355–361 (2015). doi:10.3233/SW-140147
[17] Paredes, L.P., AÍvarez Rodriǵuez, J.M., Azcona, E.R.: Promoting government controlled vocabularies for the Semantic Web: the EUROVOC thesaurus and the CPV product classification system. In: Kollias, S., Cousins, J. (eds.) Proceedings of the 1st International Workshop on Semantic Interoperability in the European Digital Library (SIEDL 2008), pp. 111–122 (2008).
[18] Caracciolo, C., Stellato, A.: Thesaurus maintenance, alignment and publication as Linked Data: the AGROVOC use case. Int. J. Metadata Semant. Ontol. 7(1), 65–75 (2012). doi:10. 1504/IJMSO.2012.048511
[19] Caracciolo, C., Stellato, A., Morshed, A., Johannsen, G., Rajbhandari, S., Jaques, Y., Keizer, J.: The AGROVOC linked dataset. Semant. Web 4(3), 341–348 (2013). doi:10.3233/SW- 130106
[20] Zapilko, B., Schaible, J., Mayr, P., Mathiak, B.: TheSoz: a SKOS representation of the thesaurus for the social sciences. Semant. Web 4(3), 257–263 (2013). doi:10.3233/SW- 2012-0081
[21] Summers, E., Isaac, A., Redding, C., Krech, D.: LCSH, SKOS and Linked Data. In: Greenberg, J., Klas, W. (eds.) Proceedings of the 2008 International Conference on Dublin Core and Metadata Applications (DC 2008), pp. 25–33 (2008).
[22] Ustalov, D.: Russian thesauri as Linked Open Data. In: Computational Linguistics and Intellectual Technologies: papers from the Annual conference “Dialogue”, vol. 1, pp. 616– 625. RGGU (2015).
[23] Nevzorova, O., Zhiltsov, N., Kirillovich, A., Lipachev, E.: OntoMathPro ontology: a Linked Data hub for mathematics. In: Klinov, P., Mouromtsev, D. (eds.) KESW 2014. CCIS, vol 468, pp. 105–119. Springer, Cham (2014). doi:10.1007/978-3-319-11716-4_9
[24] Elizarov, A.M., Kirillovich, A.V., Lipachev, E.K., Nevzorova, O.A., Solovyev, V.D., Zhiltsov, N.G.: Mathematical knowledge representation: semantic models and formalisms. Lobachevskii J. Math. 35(4), 348–354 (2014). doi:10.1134/S1995080214040143
[25] Navigli, R., Ponzetto, S.P.: BabelNet: the automatic construction, evaluation and application of a wide-coverage multilingual semantic network. Artif. Intell. 193, 217–250 (2012). doi:10.1016/j.artint.2012.07.001
[26] Ehrmann, M., Cecconi, F., Vannella, D., McCrae, J., Cimiano, P., Navigli, R.: Representing multilingual data as Linked Data: the case of BabelNet 2.0. In: Calzolari, N., et al. (eds.) Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC 2014), pp. 401–408 (2014).
[27] Галиева А.М. Создание русско-татарского тезауруса по общественно-политической тематике: общие принципы и аспекты реализации / А.М. Галиева, А.В. Кириллович, Н.В. Лукашевич, О.А. Невзорова, Д.Ш. Сулейманов // Научно-техническая информация. Серия 2: Информационные процессы и системы. – 2017. – №2. – С. 20-28.
[28] Амиров Ф.К. Русско-татарский юридический словарь. – Казань: 1996.
[29] Низамов И.М. Краткий русско-татарский общественно-политический словарь. – Казань: 1995.
[30] Низамов И.М. Русско-татарский общественно-политический словарь. – Казань: 1997.
[31] Antoine Isaac, Ed Summers. SKOS Simple Knowledge Organization System Primer. W3C Working Group Note 18 August 2009. Available: https://www.w3.org/TR/skos-primer/
[32] Baker, T., et al.: Key choices in the design of Simple Knowledge Organization System (SKOS). J. Web Semant. 20, 35–49 (2013). doi:10.1016/j.websem.2013.05.001
[33] McCrae, J., Spohr, D., Cimiano, P.: Linking lexical resources and ontologies on the Semantic Web with Lemon. In: Antoniou, G., et al. (eds.) ESWC 2011. Part I, LNCS, vol. 6643, pp. 245–259. Springer, Heidelberg (2011). doi:10.1007/978-3-642-21034-1_17
[34] McCrae, J., et al.: The Lemon cookbook. Available: http://lemon-model.net/lemon-cookbook.pdf
[35] Cimiano, P., McCrae, J.P., Buitelaar, P.: Lexicon model for ontologies. Final community group report, 10 May 2016. Available: https://www.w3.org/2016/05/ontolex/
[36] ISO 24613:2008: Language resource management - Lexical markup framework (LMF).
[37] Kemps-Snijders, M., Windhouwer, M., Wittenburg, P., Wright, S.E.: ISOcat: remodelling metadata for language resources. Int. J. Metadata Semant. Ontol. 4(4), 261–276 (2009). doi:10.1504/IJMSO.2009.029230
[38] Kemps-Snijders, M., Windhouwer, M., Wittenburg, P., Wright, S.E.: ISOcat: corralling data categories in the wild. In: Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC 2008), pp. 887–891 (2008).
[39] Windhouwer, M., Wright, S.E.: Linking to linguistic data categories in ISOcat. In: Chiarcos, C., Nordhoff, S., Hellmann, S. (eds.) Linked Data in Linguistics, pp. 99–107. Springer, Heidelberg (2012). doi:10.1007/978-3-642-28249-2_10
[40] ISO 12620:2009: Terminology and other language and content resources—Specification of data categories and management of a Data Category Registry for language resources.
[41] LexInfo. Available: http://www.lexinfo.net/
[42] Chiarcos, C.: OLiA – Ontologies of Linguistic Annotation. Semant. Web 6(4), 379–386 (2015). doi:10.3233/SW-140167
[43] Chiarcos, C.: Ontologies of linguistic annotation: survey and perspectives. In: Calzolari, N., et al. (eds.) Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC 2012), pp. 303–310 (2012).
[44] Hellmann, S., Lehmann, J., Auer, S., Brümmer, M.: Integrating NLP using Linked Data. In: Alani, H., et al. (eds.) ISWC 2013, Part II. LNCS, vol 8219, pp. 98–113. Springer, Heidelberg (2013). doi:10.1007/978-3-642-41338-4_7
[45] Sanderson, R., Ciccarese, P., Young, B.: Web annotation data model. W3C Recommendation, 23 February 2017. Available: https://www.w3.org/TR/annotation-model/
[46] Nevzorova, O., Nevzorov, V.: The Development Support System “OntoIntegrator” for Linguistic Applications. Information Science and Computing, vol. 13, Intelligent Information and Engineering Systems, vol. 3, pp. 78–84. ITHEA, Rzeszow-Sofia (2009).
[47] Loukachevitch, N., Dobrov, B., Chetviorkin, I.: RuThes-Lite, a publicly available version of thesaurus of Russian language RuThes. In: Computational Linguistics and Intellectual Technologies: Papers from the Annual International Conference “Dialogue”, pp. 340–349. RGGU (2014).
[48] Loukachevitch, N., Dobrov, B.: Development of ontologies with minimal set of conceptual relations. In: Lino, M.T., et al. (eds.) Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC 2004), pp. 1889–1892 (2004).
[49] Gil, Y., Miles, S.: PROV Model Primer. W3C Working Group Note, 30 April 2013. Available: https://www.w3.org/TR/prov-primer/
[50] Guarino, N., Welty, C.A.: A Formal ontology of properties. In: Dieng, R., Corby, O. (eds.) EKAW 2000. LNCS, vol. 1937, pp. 97–112. Springer, Heidelberg (2000). doi:10.1007/3- 540-39967-4_8.

Russian-Tatar Socio-political Thesaurus: Publishing in the Linguistic Linked Open Data Cloud

A.Galieva, A.Kirillivich, N.Loukachevitch, O.Nevzorova, D.Suleymanov, D.Yakubova

***Abstract*—The paper discusses the main principles and practical aspects of implementing a new bilingual lexical resource - the Russian-Tatar socio-political thesaurus and its publishing in the Linguistic Linked Open Data Cloud (LLOD). This thesaurus is developed on the basis of the Russian RuThes thesaurus format which is built as a hierarchy of concepts. Each concept has a unique name and a set of language expressions that refer to it in texts. The authors discuss the general methodology of translating concept names and their text entries, as well as ways of reflecting the specificity of the Tatar lexical-semantic system. The models and ontologies of the Linguistic Linked Open Data Cloud are considered. The paper gives the description of the constructed resource as a set of linked data in the LLOD format.**

***Keywords* — concept; linguistic resource; linked open data; publishing in the Linguistic linked open data cloud; socio-political terminology; the Tatar language; thesaurus; text entry.**