УДК 81.33
DOI: 10.15827/0236-235X.121.112-120

Дата подачи статьи: 18.09.17
2018. Т. 31. № 1. С. 112–120

ПРОГРАММНАЯ СИСТЕМА ДЛЯ РАЗРАБОТКИ МНОГОЯЗЫЧНОГО ТЕЗАУРУСА

А.В. Кириллович^{1, 2}, младший научный сотрудник, ассистент, alik.kirillovich@gmail.com

А.М. Баширов^{1, 3}, инженер-программист, зам. директора, a.basheerov@gmail.com

А.Р. Гатиатуллин¹, к.т.н., зав. отделом, доцент, agat1972@mail.ru

¹ Академия наук Республики Татарстан, ул. Баумана, 20, г. Казань, 420111, Россия

² Казанский (Приволжский) федеральный университет,
ул. Кремлевская, 18, г. Казань, 420008, Россия

³ Компания «ТемирТех», ул. Кулахметова, 10/29, г. Казань, 420033, Россия

Работа посвящена описанию нового инструмента для разработки двуязычного русско-татарского тезауруса.

В статье перечислены требования, предъявляемые к реализации данного инструмента: поддержка выбранного формализма представления тезаурусных знаний, многоязычных тезаурусов, коллективной работы, возможность публикации тезауруса в облаке Linking Open Data, кроссплатформенность и свободная лицензия. Приводятся результаты анализа существующих редакторов онтологий общего назначения и специализированных редакторов тезаурусов, основанных на моделях SKOS, WordNet и PyTез: Protégé/WebProtégé, VocBench, PoolParty, iQvoc, SKOSEd, OMWEdit, WordNetLoom, GernEtiT, DEBVisDic, WordNet Editor, редактор тезауруса YARN и редактор тезаурусов PyTез. Показаны преимущества и недостатки этих редакторов. На основании этого объясняется выбор модели PyTез, а также разъясняется необходимость разработки собственного инструмента, удовлетворяющего всем предъявляемым требованиям.

Предложен новый инструмент для разработки двуязычного тезауруса, реализованный в виде web-приложения, который имеет поддержку коллективной работы и функцию публикации тезауруса в облаке Linking Open Data (LOD). Описаны пользовательский интерфейс инструмента, а также его использование в режиме редактирования тезауруса. Отмечается возможность использования программного продукта для разработки не только русско-татарского тезауруса, но и тезаурусов других языков.

Поставлены задачи для дальнейшего развития инструмента, среди которых доработка пользовательского интерфейса, включая расширенные возможности визуализации в виде графа, интеграция с электронными корпусами и дополнительными словарями, доработка модуля публикации в облаке LOD, создание универсального инструмента для разработки лексических ресурсов, основанных на онтологии Lemon, разработка средств для детального описания отдельных лексических единиц.

Ключевые слова: информационно-поисковый тезаурус, skos, wordnet, pytez, linking open data, татарский язык.

В области компьютерной лингвистики и искусственного интеллекта существует множество задач, для решения которых активно используются электронные тезаурусы и онтологии. Это задачи информационного поиска и обработки естественного языка, среди которых расширение поискового запроса, автоматическая рубрикация документов, вычисление семантической близости, разрешение лексической многозначности, разрешение кореференции, суммаризация текста, вопросно-ответные системы. Электронные многоязычные тезаурусы используются также для кросс-языкового поиска и машинного перевода.

Электронные тезаурусы разработаны для многих языков мира, однако для татарского языка до сих пор такого тезауруса не существует. В связи с этим задача разработки тезауруса для татарского языка, а также двуязычного русско-татарского тезауруса весьма актуальна. Разработка подобного ресурса требует специализированных программных средств, к которым предъявляются следующие требования:

- поддержка выбранного формализма представления тезаурусных знаний;
- поддержка многоязычных тезаурусов;

- поддержка коллективной работы;
- возможность публикации тезауруса в облаке Linking Open Data;
- кроссплатформенность;
- свободная лицензия.

Настоящая работа имеет две цели: провести развернутый обзор программных инструментов для разработки тезаурусов и представить новый программный инструмент, отвечающий всем заявленным выше требованиям (эта цель возникла из анализа результатов обзора).

Обзор существующих инструментов для работы с тезаурусами

Рассмотрим существующие инструменты для редактирования тезаурусов: редакторы онтологий общего назначения и специализированные редакторы тезаурусов, основанные на моделях SKOS/SKOS-XL [1, 2], WordNet/EuroWordNet [3, 4] и PyTез [5].

Модели тезаурусов

Модель SKOS/SKOS-XL. В соответствии с моделью SKOS тезаурус организован в виде сети не-

зависимых от языка концептов. Концепты связаны друг с другом обобщенным иерархическим отношением выше/ниже и отношением ассоциации. Концепт имеет ярлык – текстовое обозначение имени концепта. Несколько концептов могут иметь один и тот же ярлык. Ярлыки концепта могут относиться к разным языкам. В модели SKOS ярлыки являются свойствами концепта, а не отдельными сущностями, поэтому они не могут иметь собственные свойства. Данное ограничение помогает устранить специальное расширение для модели SKOS – SKOS-XL. В SKOS-XL ярлыки могут быть представлены как отдельные объекты и иметь собственные свойства и отношения.

Модель WordNet/EuroWordNet. В соответствии с этой моделью тезаурус организован как сеть синсетов, где синсет – это набор слов, взаимозаменяемых в некотором контексте. Синсеты служат для представления понятий реального мира, хотя между синсетами и понятиями нет взаимно-однозначного соответствия. Если слово имеет несколько значений, то оно одновременно принадлежит нескольким синсетам. Синсеты связаны между собой семантическими отношениями: родовидовое отношение, отношение часть/целое, отношение следования (для глаголов) и др.

Использование синсетов делает структуру тезауруса сильно зависимой от конкретного языка. Одному понятию реального мира могут соответствовать несколько разных синсетов. Например, к разным синсетам относятся слова, обозначающие одно понятие, но являющиеся разными частями речи («двигаться» и «движение»).

Модель Рутез. В соответствии с этой моделью тезаурус организован в виде сети независимых от языка концептов. Концепты связаны друг с другом отношениями: выше/ниже, часть/целое, направленная ассоциация и ненаправленная ассоциация. С концептом связаны текстовые входы, которые представляют собой выражения концепта на естественном языке. Один и тот же концепт может быть связан с несколькими текстовыми входами, а один и тот же текстовый вход – с несколькими концептами. Текстовые входы могут относиться к разным языкам.

Инструменты для редактирования тезаурусов и онтологий

Одним из наиболее известных инструментов для создания и редактирования онтологий общего назначения является система **Protégé** (<http://protege.stanford.edu>) [6, 7]. Данный продукт разработан Центром биомедицинской информатики Стэнфордского университета и распространяется под свободной лицензией. Программа имеет две реализации: в виде кроссплатформенного десктопного приложения и в виде web-приложения **WebProtégé**. Внешний вид программного продукта

представлен на рисунке (см. http://www.swsys.ru/uploaded/image/2018_1/2018-1-dop/2.jpg).

Данный инструмент имеет широкий набор возможностей. В то же время он является слишком универсальным и плохо подходит для редактирования тезаурусов.

Следующую группу инструментов образуют системы, предназначенные для редактирования тезаурусов, основанных на модели SKOS/SKOS-XL.

Система VocBench (<http://vocbench.uniroma2.it>) является разработкой Продовольственной и сельскохозяйственной организации ООН и Научно-исследовательской группы ART при Римском университете Тор Вергата [8, 9]. Продукт доступен для скачивания и распространяется под свободной лицензией. Система реализована в виде web-приложения. Изначально продукт создавался для редактирования сельскохозяйственного тезауруса AGROVOC, но в данный момент используется для работы над тезаурусами произвольной тематики.

В системе реализована поддержка совместной работы с гибким управлением правами доступа пользователей, которое осуществляется на основе системы ролей: администратор, редактор онтологии, редактор терминов (терминологист), проверяющий, публикатор. Администратор регулирует права доступа для каждой роли, а также может создать новые роли с заданным набором прав. Редактор вносит правки, которые затем одобряет или отклоняет модератор, после чего публикатор принимает окончательное решение о том, должны ли эти правки войти в публичную версию тезауруса.

В программе имеется система отслеживания правок, которая помогает понять, над какими концептами сейчас идет работа, и оценить вклад каждого сотрудника.

Интерфейс программы (см. http://www.swsys.ru/uploaded/image/2018_1/2018-1-dop/3.jpg) организован в виде набора вкладок, каждая из которых предоставляет доступ к той или иной информации о концепте (терминология, определения, отношения, атрибуты и т.д.). Концепты тезауруса визуализированы в виде графа.

Имеется функция проверки согласованности, которая автоматически предотвращает нарушения правил построения SKOS-тезауруса при его редактировании и исправляет уже существующие нарушения при импорте из внешних источников.

Система позволяет подключать внешние словари и онтологии, а также импортировать другие онтологии, благодаря чему элементы тезауруса могут иметь произвольные свойства и отношения. Кроме того, благодаря поддержке логического вывода в системе отображаются свойства элементов тезауруса, не только заданные в явном виде, но и выведенные логически из других элементов.

Пользователи программы имеют возможность связывать элементы тезауруса с ресурсами в облаке LOD.

PoolParty (<http://www.poolparty.biz>) [10] – разработка компании Semantic Web Company, распространяемая под коммерческой лицензией. Программа реализована в виде web-приложения (см. http://www.swsys.ru/uploaded/image/2018_1/2018-1-dop/4.jpg).

В системе присутствуют сервисы для полуавтоматического пополнения тезауруса с помощью извлечения терминов из документов, для семантического аннотирования документов терминами тезауруса и для семантического поиска документов. Как и в VocBench, возможно подключение внешних словарей. Имеется функция связывания концептов тезауруса с внешними ресурсами в облаке LOD и публикации в облаке самого тезауруса.

Среди недостатков программы следует отметить ее цену и отсутствие встроенной поддержки SKOS-XL (только через специальный плагин).

iQvoc (<http://iqvoc.net>) [11–13] – разработка Federal Environment Agency (Германия) и innoQ Deutschland GmbH, распространяемая по свободной лицензии. Программа реализована в виде web-приложения. Данный редактор изначально разрабатывался для работы над немецкими тезаурусами в области защиты окружающей среды. Рабочий процесс аналогичен рабочему процессу в VocBench и поддерживает совместную работу пользователей с разграничением прав доступа на основе ролей. Имеется функция публикации тезауруса в облаке LOD. Среди недостатков данной системы следует отметить отсутствие встроенной поддержки SKOS-XL и встроенной поддержки полииерархии: концепт тезауруса не может иметь больше одного вышестоящего концепта. Поддержка данных функций возможна только с помощью специального плагина.

SKOSED (<https://code.google.com/archive/p/skoseditor/>) [14] – разработка Школы компьютерных наук Манчестерского университета, распространяемая под свободной лицензией. Программа реализована в виде специального плагина для редактора онтологий Protégé и представляет собой дополнительную вкладку в этом редакторе. С помощью данного плагина пользователь получает интерфейс, специально предназначенный для редактирования тезаурусов, но при этом может воспользоваться всеми функциями Protégé: расширение онтологии SKOS, подключение внешних онтологий, логический вывод, запрос с использованием логического вывода и т.д. Насколько известно авторам, плагин предназначен только для десктопной версии Protégé и не поддерживает коллективную работу.

В отдельную группу можно выделить редакторы тезаурусов типа WordNet и Euro WordNet. Рассмотрим программы для работы с тезаурусами такого типа.

OMWEdit (<http://compling.hss.ntu.edu.sg/omw/>) является разработкой Отделения лингвистики и

межъязыковых исследований Наньянского технологического университета, распространяемой под свободной лицензией [15]. Программа реализована в виде web-приложения (см. http://www.swsys.ru/uploaded/image/2018_1/2018-1-dop/5.jpg), предназначена для работы над тезаурусами из проекта Open Multilingual Wordnet и интегрирована с NTU Multilingual Corpus. В системе реализована поддержка коллективной работы и есть система отслеживания правок. Интерфейс системы имеет две отличительные особенности. Во-первых, редактирование элементов тезауруса происходит полностью на одном экране. Во-вторых, пользователь одновременно работает над несколькими версиями одного и того же концепта для разных языков. В программе присутствует сервис проверки согласованности и автоматического исправления ошибок.

WordNetLoom (<http://nlp.pwr.wroc.pl/en/tools-and-resources/tools/wordnetloom>) [16] – разработка Группы языковых технологий G4.19 при Вроцлавском политехническом университете. Лицензию, под которой распространяется продукт, установить не удалось. Программа реализована в виде десктопного приложения и предназначена для работы с польским ворднетом p1WN.

В системе реализована поддержка коллективной работы. Особенностью интерфейса является возможность редактирования тезауруса в виде графа (рис. 1). В редакторе присутствует специальный сервис, облегчающий пополнение тезауруса: когда пользователь добавляет в тезаурус новое слово, сервис подсказывает, к каким синсетам оно может относиться. Имеется отдельное web-приложение Management tools, предназначенное для отслеживания правок и проверки согласованности. Поддержка многоязычности отсутствует.

GernEdiT (<http://www.sfs.uni-tuebingen.de/lst/tools.shtml>) [17] – разработка Группы общей и вычислительной лингвистики при Тюбингенском университете. Продукт доступен для свободного скачивания; лицензия, под которой он распространяется, авторам не известна. Программа реализована в виде десктопного приложения и предназначена для работы над немецким ворднетом GermaNet. Интерфейс программы позволяет отображать синсеты тезауруса в виде графа, однако в этом виде синсет доступен только для просмотра и не доступен для редактирования. Имеется возможность отслеживания правок и проверки согласованности. Редактор жестко завязан на работу с немецким ворднетом и не имеет поддержки многоязычности.

DEBVisDic (<https://deb.fi.muni.cz/clients-debvisdic.php>) [18, 19] – разработка Центра обработки естественного языка при факультете информатики Масарикова университета. Продукт доступен для свободного скачивания, но предназначен для использования только в некоммерческих целях. Старая версия программы реализована в виде плагина

для браузера Firefox, новая – в виде отдельного web-приложения (работает крайне нестабильно). Редактор изначально создавался для работы над тезаурусами из проекта BalkaNet. Присутствует поддержка многоязычности: когда пользователь редактирует синсет в ворднете некоторого языка, он может открыть связанный с ним синсет в ворднете другого языка. Имеется возможность отслеживания правок и подключения внешних ресурсов: морфологического анализатора, корпусов и электронных словарей.

WordNet Editor (<http://wordventure.eti.pg.gda.pl>) [20] – разработка факультета электроники, телекоммуникаций и информатики Гданьского политехнического университета. Лицензия, под которой распространяется программа, авторам не известна, а код программы предоставляется по запросу. Редактор является частью проекта по пополнению Принстонского ворднета силами добровольцев. В рамках данного проекта любой пользователь Интернета может вносить правки, которые затем принимаются или отклоняются модераторами. Инструмент позволяет редактировать не все элементы тезауруса, а только самые важные с точки зрения разработчиков: слова, синсеты, смыслы и отношения. Редактор жестко завязан на работу с Принстонским ворднетом и не имеет поддержки многоязычности.

Редактор тезауруса YARN (<https://russianword.net>) [21, 22] – разработка Уральского федерального

университета, распространяемая под свободной лицензией и реализованная в виде web-приложения. Инструмент не является редактором тезауросов общего назначения и предназначен для решения единственной задачи – сборки синсетов тезауруса YARN (открытого тезауруса русского языка, основанного на принципах краудсорсинга, разрабатываемого силами нескольких сотен добровольцев). Синсеты строятся на основе сырых данных, автоматически собранных из внешних ресурсов и содержащих список слов, их определения, примеры использования и возможные синонимы. С помощью данного инструмента пользователи объединяют слова в синсеты, подбирают примеры использования слова в синсете и определение синсета. Внесенные пользователями правки принимают или отклоняют модераторы. Редактор жестко завязан на работу с тезаурусом YARN и не имеет поддержки многоязычности.

Для работы с тезаурусами типа **PyТез** [5] существует закрытая разработка авторов этих тезауросов (рис. 2). Данный редактор полностью соответствует выбранной модели тезауруса, однако он не опубликован и не доступен для свободного скачивания. Судя по доступной информации, программа реализована в виде десктопного приложения для Windows и не является кроссплатформенной. В системе отсутствуют поддержка коллективной работы и возможность публикации тезауруса в облаке LOD.

Рис. 2. Редактор тезаурусов типа RuTез

Fig. 2. A thesaurus editor of the RuTез type

Выводы по результатам обзора

Проведенный анализ существующих моделей тезаурусов и инструментов для работы с этими моделями показал, что наиболее предпочтительной для двуязычного русско-татарского тезауруса является модель RuTез.

Ни один из рассмотренных инструментов полностью не удовлетворяет предъявляемым требованиям. В наибольшей степени им удовлетворяет редактор VocBench. Однако он основан на модели SKOS, которая предлагает слишком бедный набор отношений: в этой модели имеется только обобщенное отношение выше/ниже, которое не разделяется на отношения класс/подкласс, класс/экземпляр и часть/целое. Редактор позволяет расширять модель SKOS, добавляя в нее новые отношения, но эти отношения не отображаются в дереве концептов. Как было сказано, редактор для тезаурусов типа RuTез почти полностью соответствует выбранной модели, однако он не доступен для свободного доступа и, кроме того, не удовлетворяет большинству других требований.

В связи с этим принято решение разработать собственный инструмент, удовлетворяющий всем требованиям.

Редактор русско-татарского тезауруса

На основе проведенного анализа существующих систем принято решение о создании специального программного инструментария, позволяющего хотя бы частично автоматизировать рутинный процесс создания двуязычного русско-татарского тезауруса. В результате такое ПО было разработано в виде веб-приложения, которое позволяет работать пользователям с общей БД. Про-

грамма предназначена для работы пользователей трех типов. Первый тип – пользователи, которые могут работать только в режиме просмотра и поиска информации по тезаурусу. Второй тип – эксперты, которые могут работать в режиме заполнения и редактирования. Для доступа к этому режиму необходимо пройти авторизацию. Третий тип пользователей – администратор, который дополнительно к перечисленным возможностям имеет возможность регистрации новых пользователей системы, блокировки существующих, а также управления структурой тезауруса.

В программе тезаурус подразделяется на русскую и татарскую части. Внешний вид страницы сайта с русской частью тезауруса в режиме просмотра и поиска представлен на рисунке (http://www.swsys.ru/uploaded/image/2018_1/2018-1-dop/6.jpg).

Программа предоставляет пользователям следующий набор сервисов:

- алфавитный указатель текстовых входов;
- поиск текстовых входов;
- алфавитный указатель концептов (смысловых значений текстовых входов);
- просмотр определенного домена тезауруса (например, общественно-политический, географический и т.п.).

Доступ к этим сервисам осуществляется с помощью главного меню.

Пользователи первого типа могут работать со всеми предоставляемыми ресурсами в режиме просмотра, а эксперты в режиме редактирования концептов, текстовых входов, отношений между концептами и текстовыми входами.

Текстовые входы, а также концепты можно искать по алфавитному указателю, выбрав из алфавитного списка интересующую букву. Поскольку объем тезауруса достаточно большой, для отображения на одной странице всех записей применяется деление списков текстовых входов на страницы по 100 записей. Для удобства страницы не пронумерованы, а обозначены первыми тремя буквами из названия первой записи на каждой из страниц.

Кроме алфавитного доступа, в программе реализована функция быстрого поиска. Быстрый поиск возможен как по концептам, так и по текстовым входам. Во время ввода поискового запроса функционирует система подсказок на основе списка концептов и текстовых входов, имеющих в БД.

Режим редактирования

Для работы в режиме редактирования необходимо пройти авторизацию в системе. Редактор может работать в тех же разделах сайта, что и незарегистрированный пользователь, только с возможностью редактирования. Страница со списком концептов в режиме редактирования представлена на рисунке 3.

В режиме редактирования на страницах текстовых входов и концептов рядом с названием сущностей располагаются кнопки «Правка», «Просмотр», «Удаление» с соответствующими значками. Наверху страницы также присутствует кнопка «Добавить текстовый вход» или «Добавить концепт» (в зависимости от открытого раздела).

Концепты или текстовые входы можно редактировать и на отдельных страницах, где также присутствуют кнопки «Правка» и «Удаление», позволяющие производить соответствующие действия над активной записью. На детальной странице концепта показываются текстовые входы, связанные с этим значением, а также взаимоотношения с другими концептами.

Страницы редактирования концептов и текстовых входов имеют аналогичный внешний вид.

На странице редактирования текстового входа содержатся поля Текст (текстового входа) и Язык (к которому относится текстовый вход). Ниже расположен блок связанных концептов, возле каждого из которых доступны кнопки «Открепить» и «Редактировать». Кнопкой «Прикрепить концепт» на той же странице открывается диалоговое окно с по-

иском концепта в системе (с помощью поисковой строки). Интерфейс позволяет привязать к текстовому входу сразу несколько найденных концептов. При нажатии кнопки «Создать концепт» создается новый концепт (отношения которого с другими концептами первоначально не заданы), который связывается с текущим текстовым входом. При этом требуется ввести следующие данные: название и описание концепта, по крайней мере, на одном из языков системы. Поле «Описание» является необязательным, однако его заполнение позволяет уточнить смысл концепта. Таким образом, добавление концептов в систему возможно не только на странице концептов, но и при правке (создании) текстового входа.

Все текстовые входы с синонимами и концепты являются гиперссылками, что позволяет быстро перемещаться по содержимому тезауруса. Типы связей между концептами также являются гиперссылками. При нажатии на такую ссылку на странице отображается всплывающее окно с деревом концептов для данного типа связи. Страница редактирования концептов отличается лишь тем, что выводится только один выбранный концепт. Для этого концепта показаны связанные с ним текстовые входы, а также отношения с другими концептами и типы отношений. При нажатии на заголовок детальной страницы концепта отображается дерево концептов с возможностью просмотра всех отношений в любом направлении.

В программе для удобства эксперта реализована система подсказок, которые отображают информацию о состоянии концептов и наличии нере-

Рис. 3. Страница тезауруса в режиме редактирования

Fig. 3. Thesaurus page in the edit mode

шенных задач. Определять наличие нерешенных задач эксперт может также по цвету гиперссылок концептов: синий цвет означает, что информация о концепте введена на всех языках системы, а красный показывает наличие проблем, которые нужно исправить. При отсутствии у концепта отношений с другими концептами выводится отметка «Нет отношений». Если связанные текстовые входы или название концепта указаны только на одном языке, система добавляет значок «Только один язык». Если же с концептом вообще не связан ни один текстовый вход, около концепта в списке будет указано «Нет текстовых входов».

Еще одним инструментом, помогающим в работе эксперта-переводчика, является встроенный русско-татарский словарь. Его можно разместить в любом удобном месте страницы и скрывать при необходимости. Словарь содержит поисковую строку, ниже которой выводится список найденных результатов. При создании и редактировании концептов в блоке текстовых входов записи на русском языке всегда выводятся в виде гиперссылок. При нажатии на них автоматически появляется окно русско-татарского словаря с введенным запросом и списком найденных в словаре результатов.

В программе реализована возможность отслеживания истории изменений. В подразделе «История операций» в табличном виде отображается список всех действий зарегистрированных пользователей по отношению к БД тезауруса. Список изменений отображается постранично, записи выстраиваются по убыванию даты. Каждая запись содержит информацию об авторе и времени изменения по отношению к определенному справочнику БД тезауруса. В верхней части списка доступны фильтры по именам пользователей и по справочникам. Изменения сохраняются в отдельной БД в виде json-документов, поэтому можно просматривать, кто, когда и какие именно изменения вносил в ту или иную запись системы, независимо от того, чем является эта запись: текстовым входом, концептом или учетной записью пользователя. Данный инструмент позволяет восстанавливать удаленные записи или производить откат изменений по состоянию на определенную дату.

Особенности программной реализации

Система разработана с использованием инструментов разработки ПО с открытым исходным кодом. Код программы написан на языке PHP с применением библиотек фреймворка Yii 2. Раздел администратора вынесен в отдельную часть кода. Основные данные хранятся в СУБД MySQL. Также используются хранилище Redis (для размещения сессий пользователей) и БД MongoDB (для хранения истории действий авторизованных пользователей и всех изменений в тезаурусе). Тезаурус размещается на сервере с операционной системой семей-

ства Linux, при необходимости он может быть установлен в любых операционных системах, поддерживающих указанные технологии, в том числе и на платформе Microsoft Windows.

Заключение

В статье сделан анализ моделей тезаурусов и инструментов для редактирования тезаурусов, описан новый инструмент для создания русско-татарского тезауруса, раскрыты его функциональные возможности. В настоящее время данный инструмент находится на стадии внутреннего тестирования разработчиками русско-татарского тезауруса, после которого планируется опубликовать исходный код продукта под свободной лицензией на репозитории GitHub.

Созданный программный продукт может быть использован для разработки не только русско-татарского тезауруса, но и тезаурусов других языков.

План дальнейшего развития программного продукта предусматривает первоочередные и долгосрочные задачи. К первоочередным задачам относятся следующие:

- доработка пользовательского интерфейса, включая расширенные возможности визуализации тезауруса в виде графа;
- интеграция с электронными корпусами и дополнительными словарями, которая должна помочь разработчикам тезауруса находить новые концепты и текстовые входы, а также подбирать примеры их использования;
- доработка модуля публикации тезауруса в облаке LOD.

Долгосрочная задача состоит в том, чтобы создать универсальный инструмент для разработки лексических ресурсов, основанных на онтологии Lemon. Помимо существующих средств для описания концептов, необходимы средства для детального описания отдельных лексических единиц (текстовых входов), например, декомпозиция сложной лексической единицы, парадигматическое, морфологическое и синтаксическое описания (например, модель управления глагола, его синтаксические аргументы и их характеристики).

Данный инструмент в перспективе может стать стандартом де-факто для разработки тезаурусов, основанных на модели RuTез, а также лексических ресурсов, основанных на ISO-стандарте LMF и онтологии Lemon. Разработка новых тезаурусов и публикация этих данных в вебе будут способствовать развитию глобального облака связанных данных (Linguistic Linking Open Data Cloud).

Исследование выполнено при финансовой поддержке Российского научного фонда, проект № 16-18-02074 «Разработка моделей связывания терминологии в разных языках (на материале русского и татарского языков)».

Литература

1. Isaac A., Summers E. SKOS simple knowledge organization system. Primer, W3C, 2009. URL: <https://www.w3.org/TR/skos-primer/> (дата обращения: 17.09.2017).
2. Miles A., Bechhofer S. (eds.). SKOS simple knowledge organization system eXtension for labels (SKOS-XL). Namespace document. 2009. URL: <https://www.w3.org/TR/skos-reference/skos-xl.html> (дата обращения: 17.09.2017).
3. Fellbaum C. (ed.) WordNet: an electronic lexical database. MIT Press, 1998, 449 p.
4. Vossen P. (ed.). EuroWordNet: a multilingual database with lexical semantic networks. Kluwer Acad. Publ., 1998, 179 p.
5. Лукашевич Н.В. Тезаурусы в задачах информационного поиска. М.: Изд-во МГУ, 2011. 512 с.
6. Tudorache T., Noy N.F., Tu S., Musen M.A. Supporting collaborative ontology development in Protégé. Proc. 7th Intern. Semantic Web Conf. (ISWC 2008), pp. 17–32.
7. Tudorache T., Vendetti J., Noy N.F. Web-Protégé: a light-weight OWL ontology editor for the Web. Proc. 5th OWL: Experiences and Directions Workshop (OWLED 2008), CEUR-WS.org, 2008, vol. 8, iss. 10, p. 569.
8. Stellato A., Rajbhandari S., Turbati A., Fiorelli M., Caracciolo C., Lorenzetti T., Keizer J., Pazienza M.T. VocBench: a web application for collaborative development of multilingual thesauri. Proc. 12th Europ. Semantic Web Conf. (ESWC 2015), LNCS, 2015, vol. 9088, pp. 38–53.
9. Caracciolo C., Dister S., Rajbhandari S., Stellato A. VocBench v2.3 User Manual. November, 2015. URL: <http://aims.fao.org/activity/blog/user-manual-vocbench-version-23-available> (дата обращения: 17.09.2017).
10. Schandl T., Blumauer A. PoolParty: SKOS thesaurus management utilizing linked data. Proc. 7th ESWC, 2010.
11. Bandholtz T., Schulte-Coerne T., Glaser R., Fock J., Keller T. iQvoc – open source SKOS(XL) maintenance and publishing tool. Proc. 6th Workshop on Scripting and Development for the Semantic Web (SFSW 2010), 2010.
12. Bandholtz T., Fock J., Wolff A., Schentz H. LOD-ready environmental terminology with iQvoc. Proc. 25th EnviroInfo Conf. Innovations in Sharing Environmental Observations and Information (EnviroInfo Ispra 2011), Shaker Verlag, Aachen, 2011, part 1, pp. 343–352.
13. iQvoc Editorial Team Tutorial. 2015. URL: <http://iqvoc.net/EditorialTeamTutorial.pdf> (дата обращения: 17.09.2017).
14. Jupp S., Bechhofer S., Stevens R. A flexible API and editor for SKOS. Proc. 6th Europ. Semantic Web Conf. (ESWC 2009), LNCS, vol. 5554, pp. 506–520.
15. Morgado da Costa L., Bond F. OMWEdit – the integrated open multilingual wordnet editing system. Proc. ACL-IJCNLP 2015, pp. 73–78.
16. Piasecki M., Marcińczuk M., Ramocki R. and Maziarz M. WordNetLoom: a WordNet development system integrating form-based and graph-based perspectives. Intern. Jour. of Data Mining, Modelling and Management, 2013, vol. 5, no. 3, pp. 210–232.
17. Henrich V., Hinrichs E. GernEdit – the GermaNet editing tool. Proc. 7th Intern. Conf. on Language Resources and Evaluation (LREC 2010), 2010, pp. 2228–2235.
18. Horák A., Pala K., Rambousek A., Povoln M. DEBVisDic – first version of new client-server WordNet browsing and editing tool. Proc. 3rd Intern. WordNet Conf. (GWC 2006), 2006, pp. 325–328.
19. Horák A. and Rambousek A. DEB Platform Deployment: Current Applications. Proc. 1st Workshop on Recent Advances in Slavonic Natural Language Processing (RASLAN 2007), 2007, pp. 3–11.
20. Szymański J. Cooperative WordNet editor for lexical semantic acquisition. Proc. 1st Intern. Joint Conf. on Knowledge Discovery, Knowledge Engineering and Knowledge Management (IC3K 2009), 2009, pp. 187–196.
21. Braslavski P., Ustalov D., Mukhin M. A spinning wheel for YARN: user interface for a crowdsourced thesaurus. Proc. 14th Conf. Europ. Chapter of the Association for Computational Linguistics (EACL 2014), 2014, pp. 101–104.
22. Braslavski P., Ustalov D., Mukhin M., Kiselev Yu. YARN: spinning-in-progress. Proc. 8th Global WordNet Conf. (GWC 2016), 2016, pp. 161–167.

Software & Systems

DOI: 10.15827/0236-235X.121.112-120

Received 18.09.17

2018, vol. 31, no. 1, pp. 112–120

MULTILINGUAL THESAURUS DEVELOPMENT FRAMEWORK

A.V. Kirillovich^{1,2}, Junior Researcher, Assistant, alik.kirillovich@gmail.com**A.M. Bashirov**^{1,3}, Engineer-Programmer, Deputy Director, a.basheerov@gmail.com**A.R. Gatiatullin**^{1,4}, Ph.D. (Engineering), Head of Department, agat1972@mail.ru¹ Academy of Sciences of the Republic of Tatarstan, Bauman St. 20, Kazan, 42008, Russian Federation² Kazan (Volga region) Federal University, Kremlevskaya St. 18, Kazan, 42008, Russian Federation³ TemirTech LLC, Kulakhmetova St. 10/29, Kazan, 420033, Russian Federation⁴ Institute of Applied Semiotics of the Academy of Sciences of the Republic of Tatarstan, Bauman St. 20, Kazan, 420111, Russian Federation

Abstract. The paper describes a new tool for creating a bilingual Russian-Tatar thesaurus. The paper lists the requirements for the implementation of the tool: support for selected formalism of thesaurus knowledge representation, support for multilingual thesauri, teamwork support; publishing a thesaurus in the Linking Open Data cloud (LOD); cross-platform and free licensing.

The paper gives a detailed analysis of the existing editors of general purpose ontologies and specialized thesaurus editors based on SKOS, WordNet and RuThes models, which include: Protégé/WebProtégé, VocBench, PoolParty, iQvoc, SKOSEd, OMWEdit, WordNetLoom, GernEdit, DEBVisDic, WordNet Editor, YARN and RuThes editors. It also describes the advantages and disadvantages of these editors and, therefore, it explains the choice of the RuThes model, and the reasons of developing an own tool that meets all the initial requirements. The new tool for developing a bilingual thesaurus is implemented

as a web application with support for teamwork and publishing the thesaurus data in LOD. The paper describes the user interface (UI) both for the end-user, and the database editor modes. It is stressed that the software can be used not only for developing a Russian-Tatar thesaurus, but also for other languages. The goals for the further development are: UI refining, including advanced graphical visualization; integration with electronic cases and additional dictionaries; LOD publication module finalization; a universal tool for Lemon ontology-based lexical resources development; individual lexical units detailed description module development.

Keywords: information retrieval thesaurus, SKOS, WordNet, RuThes, Linking Open Data, Tatar language.

Acknowledgements. *The research has been done with the financial support of the Russian Science Foundation (Project no. 16-18-02074, "Development of models for linking terminology in different languages (on the Russian and Tatar language material)").*

References

1. Isaac A., Summers E. *SKOS Simple Knowledge Organization System Primer*. W3C Working Group Note 18 August 2009. Available at: <https://www.w3.org/TR/skos-primer/> (accessed September 17, 2017).
2. Miles A., Bechhofer S. *SKOS Simple Knowledge Organization System eXtension for Labels (SKOS-XL)*. Namespace Document. 2009. Available at: <https://www.w3.org/TR/skos-reference/skos-xl.html> (accessed September 17, 2017).
3. Fellbaum Ch. *WordNet: An Electronic Lexical Database*. MIT Press, 1998.
4. Vossen P. (Ed). *EuroWordNet: A multilingual database with lexical semantic networks*. Springer Publ., 1998.
5. Lukashovich N.V. *Tezaurusy v zadachakh informatsionnogo poiska* [Thesauri in Information Retrieval Problems]. Moscow, Lomonosov MSU Publ., 2011, 512 p.
6. Tudorache T., Noy N.F., Tu S., Musen M.A. Supporting Collaborative Ontology Development in Protégé. *Proc. 7th Intern. Semantic Web Conf. (ISWC 2008)*. Germany, 2008.
7. Tudorache T., Vendetti J., Noy N.F. Web-Protégé: A Lightweight OWL Ontology Editor for the Web. *Proc. 5th OWL: Experiences and Directions Workshop (OWLED 2008)*. 2008.
8. Stellato A., Rajbhandari S., Turbati A., Fiorelli M., Caracciolo C., Lorenzetti T., Keizer J., Pazienza M.T. VocBench: A Web Application for Collaborative Development of Multilingual Thesauri. *Proc. 12th European Semantic Web Conf. (ESWC 2015)*. 2015.
9. Caracciolo C., Dister S., Rajbhandari S., Stellato A. *VocBench v2.3 User Manual*. 2015. Available at: <http://aims.fao.org/activity/blog/user-manual-vocbench-version-23-available> (accessed September 17, 2017).
10. Schandl T., Blumauer A. PoolParty: SKOS Thesaurus Management Utilizing Linked Data. *Proc. 7th Extended Semantic Web Conf. (ESWC 2010)*. 2010.
11. Bandholtz T., Schulte-Coerne T., Glaser R., Fock J., Keller T. iQvoc – Open Source SKOS(XL) Maintenance and Publishing Tool. *Proc. 6th Workshop on Scripting and Development for the Semantic Web (SFSW 2010)*. 2010.
12. Bandholtz T., Fock J., Wolff A., Schentz H. LOD-ready Environmental Terminology with iQvoc. *Proc. 25th EnviroInfo Conf. "Innovations in Sharing Environmental Observations and Information" (EnviroInfo 2011)*. 2011.
13. iQvoc Editorial Team Tutorial. 2015. Available at: <http://iqvoc.net/EditorialTeamTutorial.pdf> (accessed September 17, 2017).
14. Jupp S., Bechhofer S., Stevens R. A Flexible API and Editor for SKOS. *Proc. 6th European Semantic Web Conf. (ESWC 2009)*. 2009.
15. Morgado da Costa L., Bond F. OMWEdit – The Integrated Open Multilingual Wordnet Editing System. *Proc. 53rd Annual Meeting of the Association for Computational Linguistics and 7th Intern. Joint Conf. on Natural Language Processing of the Asian Federation of Natural Language Processing (ACL-IJCNLP 2015)*. 2015.
16. Piasecki M., Marcińczuk M., Ramocki R., Maziarz M. WordNetLoom: a WordNet development system integrating form-based and graph-based perspectives. *Int. Jour. of Data Mining, Modelling and Management*. 2013, vol. 5, no. 3, pp. 210–232.
17. Henrich V., Hinrichs E. GernEdit – The GermaNet Editing Tool. *Proc. 7th Int. Conf. on Language Resources and Evaluation (LREC 2010)*. 2010.
18. Horák A., Pala K., Rambousek A., Povoln M. DEBVisDic – 1st Version of New Client-Server Wordnet Browsing and Editing Tool. *Proc. 3rd Int. WordNet Conf. (GWC 2006)*. 2006.
19. Horák A., Rambousek A. DEB Platform Deployment: Current Applications. *Proc. 1st Workshop on Recent Advances in Slavonic Natural Language Processing (RASLAN 2007)*. 2007.
20. Szymański J. Cooperative WordNet Editor for Lexical Semantic Acquisition. *Proc. 1st Int. Joint Conf. on Knowledge Discovery, Knowledge Engineering and Knowledge Management (IC3K 2009)*. 2009.
21. Braslavski P., Ustalov D., Mukhin M. A Spinning Wheel for YARN: User Interface for a Crowdsourced Thesaurus. *Proc. 14th Conf. of the European Chapter of the Association for Computational Linguistics (EACL 2014)*. 1014.
22. Braslavski P., Ustalov D., Mukhin M., Kiselev Yu. YARN: Spinning-in-Progress. *Proc. 8th Global WordNet Conf. (GWC 2016)*. 2016.