УДК 004.82

ПОДХОД К АВТОМАТИЗИРОВАННОМУ ПОСТРОЕНИЮ ОБЩЕЦЕЛЕВОЙ ЛЕКСИЧЕСКОЙ ОНТОЛОГИИ НА ОСНОВЕ ДАННЫХ ВИКИСЛОВАРЯ[*1]

© 2013 г. **А.А. Крижановский, А.В. Смирнов**

Санкт-Петербург, Федеральное государственное бюджетное учреждение науки Санкт-Петербургский ин-т информатики и автоматизации РАН

Поступила в редакцию 16.04.2012 г.

Предложен подход и рассмотрена архитектура системы автоматизированного построения общецелевой лексической онтологии. В качестве онлайн-словаря был выбран викисловарь, поскольку он имеет большую базу данных из слов с переводами на многие языки. На примере Русского Викисловаря рассмотрена структура словарной статьи, на основе которой спроектирована база данных для хранения извлечённой информации. В системах управления знаниями важной составляющей частью являются онтологии, для работы с которыми требуется разработка подходов и алгоритмов для их построения. В результате построены лексические онтологии и выполнено сравнение основных показателей двух баз данных онтологий, созданных на основе Русского и Английского Викисловарей. Выполнен анализ динамики изменения численных параметров Викисловарей и построенных авторами на их основе общецелевых лексических онтологий за 2010-2012 гг.

Введение. В компьютерной лексикологии (направление вычислительной лингвистики) можно видеть последовательный переход (и в терминологии, и в смысловом наполнении) от машиночитаемых словарей к лексическим базам знаний и затем к лексическим онтологиям. Машиночитаемый словарь [1] представляет данные бумажного словаря в электронном виде с возможностью обработки этих данных на компьютере. Лексическая база знаний (lexical knowledge base) отличается от машиночитаемого словаря тем, что в ней явно выделены значения слов и указаны связи между соответствующими значениями этих слов, что позволяет использовать эти данные для логического вывода [2].

В данной работе представлен подход к построению общецелевой лексической онтологии, интегрирующей лексическую и семантическую информацию.

Лексическая онтология (lexical ontology) содержит структурированную информацию о словах и включает семантические отношения (например, синонимия, гиперонимия, холонимия) между значениями слов [3]. Под словом «*общецелевая*» в названии онтологии подразумевается отсутствие привязанности к конкретной предметной области, т.е. в словарь онтологии пытаются включить все слова данного языка. Однако значительная часть прикладных онтологий строится для конкретной предметной области с указанием отношений между концептами данной области [4]. Существует направление автоматического построения «специализированных лексических онтологий», где аргументом для их создания служит то, что такая специализация «значительно уменьшает

[*] Работа выполнена при финансовой поддержке РФФИ (проект № 11-01-00251, 12-01-00481, 12-07-00070), РГНФ (проект № 12-04-12062), проекта № 213 Программы фундаментальных исследований Президиума РАН «Интеллектуальные информационные технологии, математическое моделирование, системный анализ и автоматизация» и проекта № 2.2 Программы ОНИТ РАН «Интеллектуальные информационные технологии, системный анализ и автоматизация»

[1] Это препринт материалов, принятых для публикации в журнал «Известия РАН. Теория и системы управления», © 2013 Компания «Pleiades Publishing, Ltd.» http://www.maik.ru .

размер онтологии и соответственно сокращает время ее обработки» [5]. Однако в настоящее время в прикладных задачах бóльшую и труднопреодолимую проблему представляет именно недостаточный объём словарей, тезаурусов и онтологий, а не их избыток [6].

Таким образом, *общецелевая лексическая онтология* содержит структурированную информацию о словах и включает семантические отношения, при этом отсутствует привязанность к конкретной предметной области. Одним из наиболее успешных проектов подобного рода считается WordNet.

WordNet – это толковый словарь и тезаурус английского языка в машиночитаемой форме. В основе словаря лежат психолингвистические теории, с учётом которых были определены значения слов и связи между словами и значениями, а также связи между самими значениями [7]. Данные WordNet используются для решения многих задач, например, определения значения слова [8–10], вычисления логичности и связности предложений в тексте [11, 12]. Многие онтологии и базы знаний включают данные WordNet, либо связаны со списками синонимов WordNet, например: OpenCyc [13], DBPedia [14]. Существует несколько баз знаний, включающих не только WordNet, но и Викисловарь, обсуждаемый далее. Это лексико-семантический ресурс UBY [15] для английского и немецкого языка и система Lexvo.org [16], содержащая отношения в виде RDF-троек между словами около 7000 языков.

При выборе источника данных для построения общецелевой лексической онтологии (далее – *онтологии*) был выбран викисловарь[2] по нескольким причинам. Викисловарь – это свободно пополняемый многофункциональный многоязычный онлайн-словарь и тезаурус. В Викисловаре содержатся толкования и переводы слов, описание фонетических и морфологических свойств, семантические отношения. Кроме того – произношение слов (транскрипция и аудиофайлы), правила разбиения слов на слоги, ударения в словах, информация об этимологии слов, а также цитаты из литературных произведений, иллюстрирующие употребление слов, и даже видео и фотографии, иллюстрирующие значения слов в прямом смысле. Достоинствами викисловаря является большой объем и разнообразие лексикографических данных. В работах [17–18] показано, что по объёму информации Немецкий Викисловарь сопоставим с тезаурусами GermaNet и OpenThesaurus, а Английский Викисловарь даже превосходит объём данных WordNet.

Научная значимость многофункциональных онлайн-словарей (викисловарей) подтверждается и тем, что викисловарь и родственный проект – википедия [19] активно используются в научных экспериментах. С помощью викисловаря решаются самые разные задачи, связанных с обработкой текста и речи:

- ✓ в машинном переводе между нидерландским и бурским языками [20];
- ✓ для автоматического определения части речи слов с помощью скрытой марковской модели для трёх языков: английского, вьетнамского и корейского [21];
- ✓ в обработке текста парсером NULEX, где используется интеграция части данных Викисловаря (времена глаголов) с базой данных WordNet и VerbNet [22];
- ✓ в системе распознавания и синтеза речи, где викисловарь – основа для быстрого создания словаря произношений [23];
- ✓ для построения онтологий [6];
- ✓ при отображении онтологий [26].

Далее в статье дается краткий обзор структуры словарной статьи Русского Викисловаря (на примере статьи для слова «танцевать»). Рассмотрены подход и архитектура системы построения онтологии. На основе лексикографических данных викисловарей построены онтологии, что позволило провести анализ и сравнить лексику английского языка в многоязычных словарях (Английский и Русский Викисловари) и WordNet.

[2] Здесь и далее название конкретного проекта (Английский Викисловарь, Русский Викисловарь) пишется с заглавной буквы, название вообще словарей данного типа, т.е. викисловарей, пишется с маленькой буквы.

1. Викисловарь и структура его словарной статьи. В викисловаре содержатся не только толкования и переводы слов, но в том числе в словарных статьях описываются фонетические и морфологические свойства слов, указываются семантические отношения. Для задания семантических свойств в викисловаре используется несколько взаимодополняющих информационных структур: семантические категории, контекстные пометы (задают стиль, предметную область, языковую принадлежность)**.**

Структура словарной статьи викисловаря достаточно жёстко и однозначно задаётся правилами. Такие правила есть в Английском Викисловаре,[3] в Русском Викисловаре[4] и, вероятно, в остальных 170 викисловарях[5]. Наличие структуры и правил форматирования словарных статей позволяет взглянуть на статью как на интереснейший объект с точки зрения автоматического извлечения данных, например с помощью регулярных выражений [25]. Такое автоматическое извлечение позволит преобразовать «неявную» структуру, т.е. структуру, понятную только читателю словаря, в явную, «понятную» компьютерным программам форму, чтобы обеспечить в дальнейшем успешное использование данных Викисловаря в различных проектах, связанных с обработкой текста.

Рассмотрим структуру викисловарей на примерах из Русского Викисловаря. В словарной статье можно выделить следующие разделы: морфологический и синтаксический, фонетический, семантический, этимологический, а также разделы родственных слов, фразеологизмов и переводов. Проиллюстрируем с помощью фрагментов словарных статей все эти разделы.

Морфологический и синтаксический раздел. В нём указаны морфологические свойства (часть речи, для существительных – род, склонение и тип склонения по классификации А.А. Зализняка [24] и т.д.). Указано членение слова на морфемы, например приставка, корень (рис. 1).

Фонетический раздел содержит произношение в транскрипции международного фонетического алфавита и звуковой файл, озвученный носителем языка (рис. 2).

Семантический раздел включает толкования и цитаты, иллюстрирующие каждое из значений слова. Особенностью толкований является наличие ссылок на словарные статьи в этом же словаре. Цитаты сопровождаются библиографической информацией: автор, название произведения, год издания, источник (корпус текстов или электронная библиотека).

С помощью помет определяется сфера использования слова (литературная, диалектная, терминологическая, жаргонная лексика), предметная область (авиационная, автомобильная, альпинистская и т.д.). Также в этом разделе описываются семантические отношения (синонимы, гиперонимы и т.д.) отдельно для каждого из значений слова. На рис. 3 для синонимов указаны такие пометы, как высокий стиль, разговорное, просторечное.

[3] См. http://en.wiktionary.org/wiki/Wiktionary:ELE.
[4] См. http://ru.wiktionary.org/wiki/Викисловарь:Правила_оформления_статей.
[5] См. http://meta.wikimedia.org/wiki/Wiktionary/Table.

Морфологические и синтаксические свойства

тан-це-ва́ть

Глагол, несовершенный вид, переходный, тип спряжения по классификации А. Зализняка — 2а.

Корень: **-танц-**; суффикс: **-ева-**; глагольное окончание: **-ть**.

	наст.	прош.	повелит.
Я	танцу́ю	танцева́л, танцева́ла	—
Ты	танцу́ешь	танцева́л, танцева́ла	танцу́й
Он она оно	танцу́ет	танцева́л танцева́ла танцева́ло	—
Мы	танцу́ем	танцева́ли	—
Вы	танцу́ете	танцева́ли	танцу́йте
Они	танцу́ют	танцева́ли	—
Пр. действ. наст.	танцу́ющий		
Пр. действ. прош.	танцева́вший		
Деепр. наст.	танцу́я		
Пр. страд. наст.	танцу́емый		
Будущее	буду/будешь… танцева́ть		

Рис. 1. Фрагмент словарной статьи «танцевать» с таблицей форм глагола

Произношение

МФА: [tənt͡si'vatʲ] Пример произношения

Рис. 2. Фрагмент словарной статьи «танцевать» со ссылкой на аудиофайл (в виде графического значка нот и динамика) и транскрипцией

Семантические свойства

Значение

1. *что* плясать, ритмично двигаться (как правило, под музыку), исполнять какой-либо танец ◆ Она **танцевала** страстно, с увлечением и вальс, и польку, и кадриль, переходя с рук на руки, угорая от музыки и шума, мешая русский язык с французским, картавя, смеясь и не думая ни о муже, ни о ком и ни о чём. *А. П. Чехов, «Анна на шее», 1895 г.*

Синонимы

1. высок.: исполнять танец; разг., прост.: плясать, выплясывать

Гиперонимы

1. двигаться

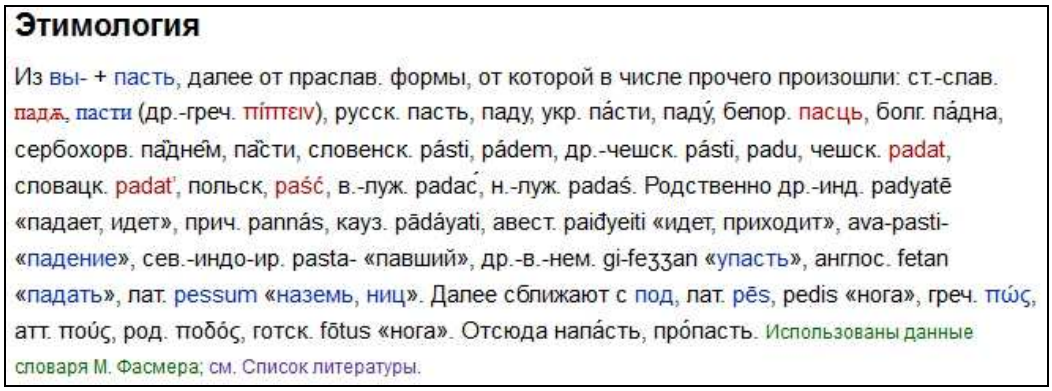
Гипонимы

1. приплясывать, вытанцовывать, вальсировать, бить чечётку

Рис. 3. Фрагмент словарной статьи «танцевать», приведено только первое значения и семантические отношения для него

Этимологический раздел. В этом разделе (рис. 4) содержится информация об истории слова, т.е. описываются фонетические и семантические изменения, которые претерпело слово. Могут быть зафиксированы различные точки зрения со ссылками на соответствующую литературу. В конце раздела указывается источник этимологической информации и даётся ссылка на «*Список литературы*», который представляет собой

страницу Русского Викисловаря, включающую огромный список корпусов и словарей, данные которых используются для создания данного словаря. Аналогично тексту толкований лексикографы (редакторы викисловаря) связывают ссылками текст этимологии и соответствующие словарные статьи.



Этимология

Из вы- + пасть, далее от праслав. формы, от которой в числе прочего произошли: ст.-слав. падѫ, пасти (др.-греч. πίπτειν), русск. пасть, паду, укр. па́сти, паду́, белор. пасць, болг. па́дна, сербохорв. па̏днем, па̏сти, словенск. pásti, pádem, др.-чешск. pásti, padu, чешск. padat, словацк. padať, польск, paść, в.-луж. padać, н.-луж. padaś. Родственно др.-инд. padyatē «падает, идет», прич. pannás, кауз. pādáyati, авест. paiδyeiti «идет, приходит», ava-pasti- «падение», сев.-индо-ир. pasta- «павший», др.-в.-нем. gi-fe3an «упасть», англос. fetan «падать», лат. pessum «наземь, ниц». Далее сближают с под, лат. pēs, pedis «нога», греч. πώς, атт. πούς, род. ποδός, готск. fōtus «нога». Отсюда напа́сть, про́пасть. Использованы данные словаря М. Фасмера; см. Список литературы.

Рис. 4. Фрагмент словарной статьи «выпасть»

Раздел родственных слов группирует однокоренные слова с разбиением по частям речи (на рис. 5 – существительные, прилагательные, глаголы и наречия, отдельно – уменьшительно-ласкательные формы существительного). Таким образом, помимо того, что в морфологическом и синтаксическом разделе указано разбиение слова на морфемы, данный раздел содержит совокупность слов (словообразовательное гнездо) с корневой морфемой. Список слов представляет собой гиперссылки на соответствующие словарные статьи.

Родственные слова [править]

Ближайшее родство: [скрыть]

- уменьш.-ласк. формы: танцулька
- пр. существительные: танец, танцор, танцорка, танцовщик, танцовщица, танцевание, подтанцовка, танцзал, танцплощадка, танцпол, танцкласс, танцмейстер, подтанцовывание, пританцовывание
- прилагательные: танцевальный, станцованный, истанцованный
- глаголы: подтанцовывать, пританцовывать, станцевать, натанцеваться, дотанцеваться, затанцевать, потанцевать, оттанцевать, протанцевать, растанцеваться, недотанцевать, перетанцевать
- наречия: танцевально, станцованно

Рис. 5. Фрагмент словарной статьи «танцевать» со списком однокоренных слов, разбитым по частям речи

Раздел устойчивых сочетаний и фразеологизмов с участием данного слова (рис. 6). Кроме фразеологизмов данный раздел может включать пословицы и поговорки, содержащие данное слово.

Фразеологизмы и устойчивые сочетания

- бальный танец
- белый танец
- танец живота
- танцевать до упаду
- танцевать от печки
- танцевать под дудку

Рис. 6. Фрагмент словарной статьи «танцевать» со списком устойчивых сочетаний

Раздел переводов содержит переводы каждого из значений слова на иностранные языки (рис. 7). Переводы представляют собой ссылки на соответствующие словарные статьи об иностранных словах. В статье Русского Викисловаря в разделе переводов в конце названия каждого языка (рис. 7) идет двух- или трехбуквенный код языка. Множество этих кодов является частью базы лексикографических констант, подробно представленной в следующем разделе.

Специальные значки, идущие после названия языков в разделе переводов, позволяют выделить в Русском Викисловаре отдельные группы языков: восстановленные, искусственные, реконструированные, фантастические и мертвые языки. На рис. 7 представлено два языка из этих групп: это древнегреческий язык (мертвый) и эсперанто (искусственный).

Перевод

исполнять танец [скрыть]

- Абхазский$_{ab}$: акәашара
- Аймарский$_{ay}$: thoqoña
- Албанский$_{sq}$: vallëzoj
- Английский$_{en}$: dance
- Армянский$_{hy}$: պարել (parel)
- Астурийский$_{ast}$: baillar
- Африкаанс$_{af}$: dans
- Баскский$_{eu}$: dantza egin; dantzatu
- Белорусский$_{be}$: танцаваць, танчыць
- Бенгальский$_{bn}$: নাচ (naca)
- Болгарский$_{bg}$: танцувам
- Бретонский$_{br}$: dañsal; koroll; korolliñ
- Валлийский$_{cy}$: dawnsio
- Венгерский$_{hu}$: táncol
- Верхнелужицкий$_{hsb}$: rejowaś
- Греческий$_{el}$: χορεύω
- Гэльский$_{gd}$: danns
- Датский$_{da}$: danse
- Древнегреческий$_{grc}$†: ἐμπαίζω; χορεύω
- Удмуртский$_{udm}$: эктыны
- Узбекский$_{uz}$: o`ynamoq (ўйнамоқ)
- Украинский$_{uk}$: танцювати
- Фарерский$_{fo}$: dansa
- Финский$_{fi}$: tanssia
- Французский$_{fr}$: dancer
- Фризский$_{fy}$: dûnsje
- Фриульский$_{fur}$: balâ
- Хорватский$_{hr}$: plésati
- Цыганский$_{rom}$: юхэлэс про гэра
- Чешский$_{cs}$: tancovat
- Шведский$_{sv}$: dansa
- Шорский$_{cjs}$: серги
- Эве$_{ewe}$: ɖu ɣe
- Эрзянский$_{myv}$: киштемс
- Эсперанто$_{eo}$[и]: danci
- Эстонский$_{et}$: tantsima
- Якутский$_{sah}$: үҥкүүлээ
- Японский$_{ja}$: 踊る (おどる, odoru), ダンスする (dansusuru)

заниматься балетом [показать]

мерно двигаться [показать]

Рис. 7. Фрагмент словарной статьи «танцевать» с переводами, приведена часть переводов для первого значения слова «танцевать» (языки от «А» до «Д», абхазский – древнегреческий, и от «У» до «Я», удмуртский – японский)

2. Подход и архитектура системы построения общецелевой лексической онтологии. Предлагаемый автоматизированный подход состоит из двух этапов: 1) первичного анализа данных онлайн-словаря экспертами с целью определения особенностей его структуры и 2) последующего автоматического построения общецелевой лексической онтологии с помощью разработанной компьютерной системы.

В викисловаре (www.wiktionary.org) специальные теги и другие маркеры словаря помечают семантические элементы и определяют иерархическую структуру словарной статьи. Тем не менее без специальной обработки словаря можно делать только полнотекстовый поиск по текстам словарных статей или навигацию по гиперссылкам статей. Для реализации сложных поисковых запросов (например, получить список всех синонимических рядов, содержащих данное слово) онлайн-словарь должен быть преобразован в формат, удобный для машинной обработки. Поэтому на первом этапе подхода работает команда из лингвистов-лексикографов (экспертами по исследуемому онлайн-словарю) и инженеров (экспертами по базам данных и программистов).

Первый этап включает следующие подзадачи:

1) анализ статьи онлайн-словаря;
2) определение структуры словарной статьи;
3) выявление «опорных» элементов словарной статьи (ключевые слова, элементы гипертекстовой разметки) с помощью регулярных выражений [25];
4) проектирование реляционной базы данных на основе модели структуры словарной статьи для сохранения извлечённых данных;
5) настройка экспертами системы построения онтологии.

При проектировании реляционной базы данных онтологии определяются разделы словарной статьи онлайн-словаря, которые будут обрабатываться. На данный момент из словарной статьи системой извлекаются следующие лексикографические данные: язык, часть речи, значение слова (толкование), цитата (иллюстрирующая значение слова), семантические отношения и перевод (табл. 1). В дальнейшем планируется учитывать данные всех разделов.

Таблица 1. Разделы и данные словарной статьи, включаемые в онтологию

Раздел словарной статьи	**Данные раздела**	**Общецелевая лексическая онтология**
Морфологический и синтаксический раздел	Язык, часть речи, род, склонение и тип склонения по классификации А.А. Зализняка, членение слова на морфемы, ударение, разбиение на слоги	Язык, часть речи
Фонетический раздел	Транскрипция, аудиофайл	–
Семантический раздел	Помета (стилевая, предметная), толкование, цитата синонимы, антонимы, гипонимы, гиперонимы, меронимы, холонимы	Толкование, цитата, синонимы, антонимы, гипонимы, гиперонимы, меронимы, холонимы
Этимологический раздел	Этимология слова	–
Раздел родственных слов	Список однокоренных слов	–
Раздел устойчивых сочетаний и фразеологизмов	Устойчивые сочетания, фразеологизмы, пословицы	–
Раздел переводов	Переводы	Переводы

Далее с учётом структуры словарной статьи и ее обрабатываемых разделов настраивается компьютерная система для автоматического построения онтологии на основе данных Викисловаря.

Второй этап построения онтологии выполняется полностью в автоматическом режиме (рис. 8). Задачами второго этапа являются: построение онтологии (главная задача этапа); автоматический сбор отладочной информации.

Отладочная информация необходима разработчикам системы на первом этапе построения онтологии на следующей итерации (например, для расширения базы лексикографических констант при добавлении кода языка, так как данные извлекаются из многоязычного словаря). Другими словами, разработка системы ведется итеративно, в систему постепенно добавляются модули для извлечения лексикографических данных (семантический модуль, модуль переводов и т. д.). Разработанная архитектура системы построения онтологий является модульной и расширяемой. В дальнейшем планируется создать модули для каждого из разделов словарной статьи. На рис. 8 показано взаимодействие основных частей данной системы.

Программной системе требуется задать следующие входные параметры. Во-первых, язык викисловаря (русский или английский), поскольку есть отличия в структуре словарных статей Русского и Английского Викисловарей. Для каждого из этих викисловарей будет запускаться свой анализатор. Во-вторых, адрес баз данных онлайн-словарей и онтологии, т.е. параметры для подключения к удалённым базам данных: IP-адрес, имя базы данных, имя и пароль пользователя. Источником данных является база данных онлайн-словарей, т.е. корпус текстов словарных статей.

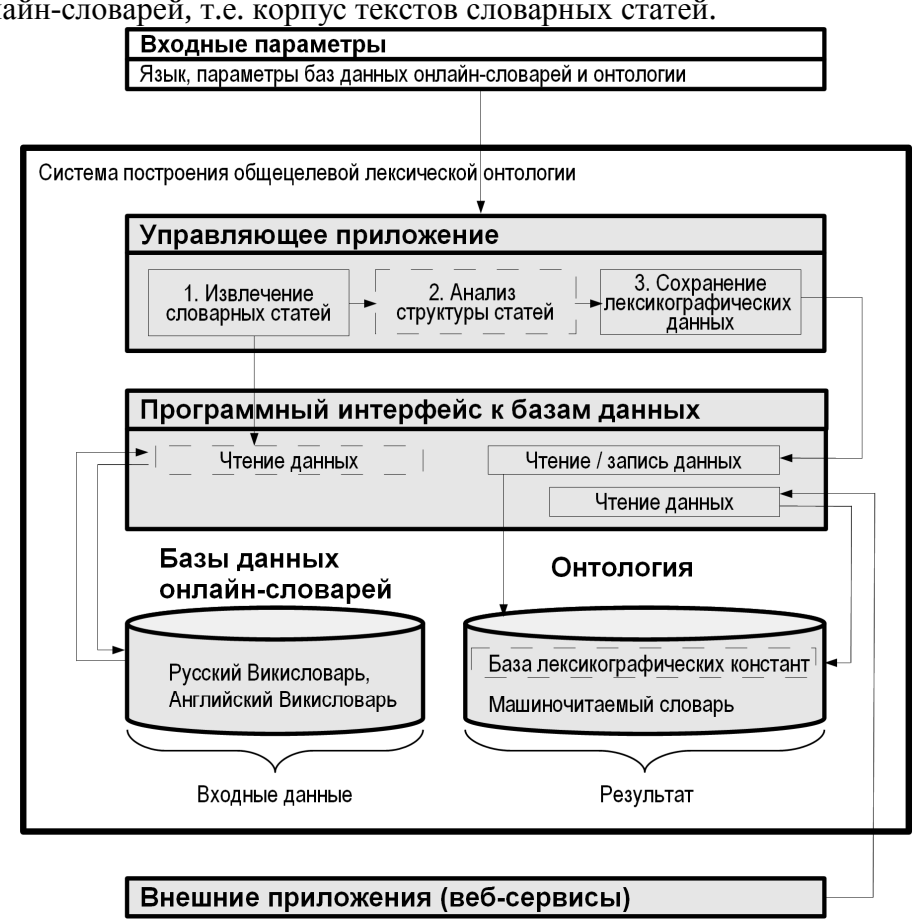


Рис. 8. Архитектура системы автоматического построения общецелевой лексической онтологии

Управляющее приложение выполняет последовательно три шага для каждой словарной статьи (вики-текста), извлекаемой из базы данных Викисловаря (что и составляет первый шаг). На втором шаге проводится анализ – с помощью регулярных выражений выполняется поиск «опорных» элементов статьи (ключевые слова, элементы гипертекстовой разметки), указывающих на ее подразделы. «Опорные» элементы

задаются экспертами на основе анализа правил викисловаря по оформлению статей. Анализ выполняется «от общего к частному», статья сначала разбивается на крупные части, затем эти части анализатор «сканирует» повторно и разбивает их на более мелкие.

На третьем шаге полученные данные сохраняются в базе данных онтологии. Результатом работы системы будет онтология, состоящая из двух частей: базы лексикографических констант и машиночитаемого словаря. На рис. 8 пунктиром выделены те части системы, которые требуют предварительной настройки: 1) база лексикографических констант, 2) модуль анализа в управляющем приложении, 3) модуль чтения данных программного интерфейса к базам данных. Дело в том, что сообщество редакторов викисловарей по мере развития проекта год от года вносит различные улучшения и уточнения в структуру словаря. Эти изменения должны быть своевременно учтены разработчиками системы. Настройка выполняется на первом этапе построения онтологии в ходе совместной работы лексикографов и инженеров.

Таким образом, *база лексикографических констант* создается экспертами вручную и задает соответствие между идентификатором и некоторой лексикографической информацией. База лексикографических констант содержит следующие списки, необходимые для анализа словарных статей и поиска по базе данных машиночитаемого словаря:

- ✓ коды языков в соответствии с международным стандартом сокращения названий языков “ISO 639”, самоназвания языков, названия языков на английском и на русском (370 языков и их кодов в Русском Викисловаре, 274 – в Английском Викисловаре);
- ✓ «заголовки третьего уровня» в терминологии Викисловаря, включающие названия частей речи и такие заголовки, как: *имя собственное*, *артикль*, *приставка*, *суффикс*, *акроним*, *аббревиатура* и т.п. (58 заголовков в Английском Викисловаре, 25 – в Русском Викисловаре);
- ✓ названия семантических отношений (*синонимия*, *антонимия* и т.д.).

База лексикографических констант заполняется экспертами и на этапе извлечения данных из Викисловаря работает только в режиме «чтения».

Программный интерфейс к базе данных представляет собой набор функций для чтения данных Викисловаря, для чтения и записи данных онтологии. Функции, связанные с обработкой данных онтологии, можно разделить на *функции низкого уровня*, где выполняется поиск, удаление, добавление или обновление записи для одной таблицы и *функции высокого уровня*, задействующие сразу много таблиц. Например, высокоуровневые функции программного интерфейса позволяют для заданного языка и части речи получить: (i) список толкований, (ii) список синонимов, антонимов и т.д., (iii) список переводов с исходного на целевой язык для заданного слова.

Внешние приложения (например, веб-сервисы), зная параметры базы данных онтологии и программный интерфейс для доступа к данным, могут работать с онтологией удаленно. Если необходима высокая скорость обработки данных, то базу данных онтологии и программную систему, обеспечивающую функционал онтологии, можно скачать с сайта авторов статьи (http://code.google.com/p/wikokit/) и установить локально.

3. Результаты работы системы автоматического построения общецелевой лексической онтологии. В ходе экспериментов были построены две базы данных лексических онтологий, что позволило сравнить их и викисловари по различным параметрам (табл. 2-4). В качестве исходных данных были взяты данные Английского Викисловаря от 8 октября 2011 г. (так называемый «дамп» базы данных) и данные Русского Викисловаря от 21 мая 2011 г. Реляционная база данных онтологии состоит из множества взаимосвязанных таблиц, поэтому для каждой таблицы во втором столбце табл. 3 приводится ее имя на английском языке и после имени в скобках дается пояснение значений следующих столбцов. В следующих столбцах табл. 3 указана размерность таблиц, т.е. число строк в этих таблицах. В предпоследнем столбце этих трех таблиц дано

отношение предыдущих двух столбцов, т.е. параметров онтологии Английского Викисловаря к онтологии Русского Викисловаря за 2011-2012 гг. Для анализа динамики роста викисловарей и онтологий в последнем столбце даны аналогичные отношения по данным работы[6] за 2010 г., где учитывались данные Английского Викисловаря от 6 января 2010 г. и данные Русского Викисловаря от 5 апреля 2010 г.

Таблица 2. Основные показатели Английского и Русского Викисловарей

№ п.п.	Свойство	Викисловарь		en / ru	
		Английский (en)	Русский (ru)	2011-2012 гг.	2010 г.
	Версия викисловаря (дата)	*08.10.2011 г.*	*21.05.2011 г.*	*2011 г.*	*2010 г.*
1	Число страниц	2 936 016	325 128	9.03	7.13
2	Число правок с момента установки викисловаря	16 971 706	3 183 428	5.33	3.65
3	**Среднее число правок на страницу**	5.35	5.12	**1.04**	**1.03**
4	Число активных редакторов словаря	1060	176	6.02	7.17

Таблица 3. Основные показатели баз данных онтологий, построенных на основе Английского и Русского Викисловарей

№ п.п.	Название таблицы в базе данных онтологии (и комментарий)[7]	Онтология		en / ru	
		Английского Викисловаря (en)	Русского Викисловаря (ru)	2011-2012 гг.	2010 г.
1	page (число словарных статей)	1 283 011	532 024	2.41	3.77
2	**relation** (число семантических отношений)	227 430	144 675	**1.57**	**1.57**
3	lang_pos (число частей речи по всем языкам)	1 219 090	427 876	2.85	4.63
4	wiki_text (число фрагментов текста – в толкованиях, переводах, семантических отношениях)	1 641 186	375 787	4.37	7.81
5	wiki_text_words (число слов-гиперссылок)	2 304 041	427 116	5.39	10.81
6	meaning (число значений)	1 634 749	248 497	6.58	12.68
7	inflection (число словоформ)	102 322	28 582	3.58	8.84
8	translation (число блоков с переводами, т.е. число переведённых значений слов)	88 450	29 856	2.96	1.55
9	translation_entry (суммарное число пар переводов на все языки)	801 943	228 805	3.50	1.96
10	Число статей (число пар: язык – часть речи) с семантическими отношениями	144 530	53 554	**2.70**	**3.89**
11	Число статей на родном языке (подсчет пар: язык – часть речи)	282 281	135 396	**2.08**	**2.43**
12	Число семантических отношений для слов на родном языке	**60 844**	**104 513**	**0.58**	**0.52**

[6] Крижановский А. Сравнение тезаурусов Русского и Английского Викисловарей, преобразованных в машиночитаемый формат. Препринт. 2010. http://scipeople.com/publication/99331/ .

[7] Эти или более свежие данные по Английскому Викисловарю доступны здесь: http://en.wiktionary.org/wiki/User:AKA_MBG/Statistics:Parameters_of_the_database_created_by_the_Wiktionary_parser, по Русскому Викисловарю – здесь: http://ru.wiktionary.org/wiki/Участник:AKA MBG/Статистика:Размеры базы данных, созданной парсером Викисловаря.

Сравнение самих викисловарей (табл. 2) показывает, что Английский Викисловарь больше Русского примерно в 9 раз по числу страниц и в 6 раз – по числу активных участников. Среди удивительно стабильных параметров, значения которых сохраняются в 2010-2011 гг., можно указать соотношение среднего числа правок (т.е. редактирований) 1.03-1.04 (строка 3 в табл. 2), приходящихся на словарную статью: 4.8-4.96 в 2010 г. и 5.12-5.35 в 2011 г. в Русском и Английском Викисловарях соответственно.

И ещё один стабильный параметр – соотношение числа семантических отношений (строка 2, табл. 3) – в Английском Викисловаре их больше в 1.57 раза, чем в Русском Викисловаре, в 2010 г. их было 157 и 100 тыс., а в 2011 г. – 227 и 145 тыс.

Под понятием «родной язык» (строки 1 и 3 в табл. 4) подразумевается основной, главный язык Викисловаря, т.е. русский язык в Русском Викисловаре, английский – в Английском Викисловаре. Главное различие всех 170 викисловарей в том, что:

толкования всех слов, в том числе иностранных (строка 6 в табл. 3), даются только на родном языке;

переводы на все языки мира (строки 8, 9 в табл. 3) приводятся только в словарных статьях для слов родного языка.

Выполнено сравнение для статей о словах на родном языке, т.е. сравнили статьи об английских словах в Английском Викисловаре со статьями о русских словах в Русском Викисловаре.

1. Семантических отношений в Русском Викисловаре между русскими словами (строка 12 табл. 3) почти в 1.7 раза больше (104.5 тыс.), чем между английскими словами в Английском Викисловаре (60.8 тыс.), в 2010 г. их было в 2 раза больше (84 и 44 тыс.).
2. В Английском Викисловаре словарные статьи о словах родного языка (строка 1 табл. 4) составляют менее десятой части (9.6%) всех статей (в 2010 г. составляли пятую часть – 18.3 %). В Русском Викисловаре процент словарных статей о словах родного языка значительно выше, хотя и снизился – 41% (в 2010 г. было больше половины статей о русских словах – 53.7 %). Таким образом, несмотря на общую цель обоих Викисловарей – описание всех словарных единиц всех языков, Русский Викисловарь продолжает оставаться более моноязычным.
3. Среднее число семантических отношений на словарную статью (строка 3 табл. 4) в Русском Викисловаре больше, чем в Английском в 3.5 раза, и они составляют соответственно 0.77 и 0.22 (в 2010 г. было 0.65 и 0.14).

Таблица 4. Статистика по словарям, полученная на основе анализа онтологий

	Свойство	Викисловарь		en / ru	
		Английский (en)	Русский (ru)	2011-2012 гг.	2010 г.
1	Число статей на родном языке (language & POS) к общему числу страниц (отношение значений в строке 11 в табл. 3 к значению в строке 1 в табл. 2), %	9.61	41.64	0.23	0.34
2	Число статей с семантическими отношениями к общему числу страниц (отношение значений в строке 10 в табл. 3 к значению в строке 1 в табл. 2), %	4.92	16.47	0.30	0.55
3	Среднее число семантических отношений для статей о словах на родном языке (отношение значений в строке 12 к значению в строке 11 в табл. 3)	0.22	0.77	0.29	0.21

Построение онтологий позволило выполнить численный анализ и сравнить лексику английского языка в Английском Викисловаре, в Русском Викисловаре и в WordNet. Для сравнения с базой данных WordNet в многоязычных викисловарях учитывались только

статьи об английских словах, поскольку WordNet включает только их. В результате сравнения получено:

- ✓ число английских слов и значений в словарях: больше всего английских слов (276470) и значений (369778) содержится в Английском Викисловаре по данным на 2011 г. В нём статей больше чем в WordNet в 1.78 раза, а значений – больше в 1.79 раза;
- ✓ распределение слов английского языка по частям речи (в Английском Викисловаре, WordNet, Русском Викисловаре), в %: существительные 52, 71, 83 (занимают наибольшую долю во всех словарях); прилагательные 20, 14, 6; глаголы 15, 12, 8; наречия 4, 3, 1;
- ✓ число слов с одним значением (81% в Английском Викисловаре и 88% в WordNet) и многозначных слов: оба словаря (WordNet и Английский Викисловарь) содержат больше слов с одним значением, чем с несколькими. В Английском Викисловаре для существительных, глаголов и прилагательных многозначным является практически каждое пятое слово (17-22%), только для наречий – многозначных слов всего 12%;
- ✓ среднее число значений для слов, принадлежащих разным частям речи: во всех трех словарях наиболее многозначными оказались глаголы, среднее число значений у глаголов (без учета слов с одним значением) больше трех (диапазон 3.08-3.57). Меньше всего значений (без учета слов с одним значением) у наречий (диапазон 2.35-2.88).

Более подробный анализ и сравнение словарей (на основе построенных онтологий) и WordNet представлены в [17].

Заключение. В работе предложен подход и рассмотрена архитектура системы автоматизированного построения общецелевой лексической онтологии. На первом этапе команда из лингвистов-лексикографов (экспертов по исследуемому онлайн-словарю) и инженеров (экспертов по базам данных и программистов) выполняет анализ словарной статьи онлайн-словаря с целью проектирования и разработки системы. На втором этапе система в автоматическом режиме обходит все словарные статьи, записывает извлечённую информацию в базу данных общецелевой лексической онтологии.

Разработанная система может быть использована: 1) в научных прототипах и приложениях (автоматизация построения онтологий и баз знаний, распознавание значения слова), 2) в офисных приложениях (проверка правописания, перевод [26]), 3) в промышленных системах (выявление интересов пользователей, построение профиля клиента), 4) в системах мониторинга и кластеризации текстовых потоков, выявление текстов заданной тематики.

СПИСОК ЛИТЕРАТУРЫ

1. *Amsler R. A.* Machine readable dictionaries. // Annual Review of information Science and technology. V.19 / Ed. M. E Williams. Knowledge Industry Publication, Inc., White Plains, NY, USA, 1984. P.161-209.
2. *Calzolari N.* Lexical Databases and Textual Corpora: Perspectives of integration for a Lexical Knowledge Base / Ed. U. Zernik, Lexical Acquisition: Exploiting On-Line Resources to Build a Lexicon, Hillsdale, New Jersey, USA, 1991. P.191-208.
3. *Wandmacher T., Ovchinnikova E., Krumnack U. et al.*. Extraction, evaluation and integration of lexical-semantic relations for the automated construction of a lexical ontology. // Third Australasian Ontology Workshop (AOW), V.85 of CRPIT, Gold Coast, Australia, 2007. P.61-69.
4. *Болотникова Е.С., Гаврилова Т.А., Горовой В.А.* Об одном методе оценки онтологий // РАН. ТИСУ, 2011. N3. С.98-110. http://scipeople.com/publication/107484/ .
5. *Лесько О., Рогушина Ю.* Использование специализированной лексической онтологии для автоматизации формирования онтологии предметной области по естественно-языковым текстам / Знание - Диалог - Решение (KDS-2010), Киев, Украина, ITHEA, 2010. С.93-100. http://www.foibg.com/ibs_isc/ibs-19/ibs-19-p10.pdf .
6. *Meyer C. M., Gurevych I.* OntoWiktionary – Constructing an Ontology from the Collaborative Online Dictionary Wiktionary / Eds. M. T. Pazienza and A. Stellato, Semi-Automatic Ontology Development: Processes and Resources. Hershey, Pennsylvania, USA, IGI Global, 2012. P.131-161. http://www.ukp.tu-darmstadt.de/data/lexical-resources/ontowiktionary/ .
7. *Ferrer-i-Cancho R.* The structure of syntactic dependency networks: insights from recent advances in network theory // Eds. V. Levickij and G. Altmann, Problems of quantitative linguistics, Chernivtsi, Ukraine, Ruta, 2005. P.60-75.
8. *Montoyo A., Palomar M., Rigau G.* Method for WordNet enrichment using WSD // Proc. 4th International Conf. on Text Speech and Dialogue TSD'2001. Železná Ruda – Spieak. Czech Republic: Published in Lecture Notes in Artificial Intelligence 2166. Springer-Verlag, 2001.
9. *Resnik P., Yarowsky D.* Distinguishing systems and distinguishing senses: new evaluation methods for word sense disambiguation // Natural Language Engineering. 2000. V.5, №2. P.113-133.
10. *Yarowsky D.* Unsupervised word sense disambiguation rivaling supervised methods // 33rd Annual Meeting of the Association for Computational Linguistics. Cambridge, MA, 1995. P.189-196.
11. *Harabagiu S., Moldovan D.* A marker-propagation algorithm for text coherence // Working Notes of the Workshop on Parallel Processing at the 14th Intern. Joint Conf. on Artificial Intelligence. Montreal, 1995. P.76-86.
12. *Teich E., Fankhauser P.* WordNet for lexical cohesion analysis // Proc. Second Global WordNet Conf. Brno, Czech Republic, 2004. P.326-331.
13. *Reed S., Lenat D.* Mapping Ontologies into Cyc // Proc. of AAAI 2002 Conf. Workshop on Ontologies For The Semantic Web, Edmonton. Canada, 2002.
14. *Bizer C., Lehmann J., Kobilarov G. et al.* DBpedia - A crystallization point for the Web of Data // Web Semantics: Science, Services and Agents on the World Wide Web. 2009. V.7, №3. P. 154-165. ISSN 1570-8268. http://www.wiwiss.fu-berlin.de/en/institute/pwo/bizer/research/publications/Bizer-etal-DBpedia-CrystallizationPoint-JWS-Preprint.pdf .
15. *Gurevych I., Eckle-Kohler J., Hartmann S. et al.* Uby – A Large-Scale Unified Lexical-Semantic Resource // Proc. 13th Conf. of the European Chapter of the Association for

Computational Linguistics, 2012. Avignon, France, 2012. P.580-590. http://www.ukp.tu-darmstadt.de/fileadmin/user_upload/Group_UKP/publikationen/2012/uby_eacl2012_camera ready.pdf .
16. *Melo G., Weikum G.* Language as a Foundation of the Semantic Web // Proc. Poster And Demonstration Session of the 7th Intern. Semantic Web Conf. (ISWC 2008), CEUR. V. 401. Karlsruhe, Germany, 2008.
17. *Крижановский А.А.* Количественный анализ лексики английского языка в викисловарях и WordNet // Тр. СПИИРАН. 2011. Вып. 19. С.87–101.
18. *Meyer C. M., Gurevych I.* Wiktionary: a new rival for expert-built lexicons? Exploring the possibilities of collaborative lexicography // Electronic Lexicography. Oxford: Oxford University Press. 2012. (preprint). http://www.informatik.tu-darmstadt.de/fileadmin/user_upload/Group_UKP/publikationen/2011/oup-elex2012-meyer-wiktionary.pdf .
19. *Крижановский А.А., Смирнов А.В.* К вопросу об индексировании вики-текстов // Изв. РАН. ТИСУ.2009. №4. С.121-129.
20. *Otte P., Tyers F. M.* Rapid rule-based machine translation between Dutch and Afrikaans // 16th Annual Conf. of the European Association of Machine Translation, EAMT11, Leuven, Belgium, 2011.
21. *Kiem-Hieu Nguyen, Cheol-Young Ock.* Using Wiktionary to Improve Lexical Disambiguation in Multiple Languages // CICLing, V.7181 of Lecture Notes in Computer Science. 2012. P.238-248.
22. *McFate C., Forbus K.* NULEX: An Open-License Broad Coverage Lexicon // The 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies. Portland, Oregon, USA, 2011. P.363-367.
23. *Qingyue He.* Automatic Pronunciation Dictionary Generation from Wiktionary and Wikipedia // Thesis. Karlsruhe Institute of Technology. 2009.
24. *Зализняк А. А.* Грамматический словарь русского языка. М.: Русский язык, 2010. 720 с.
25. Фридл Дж. Регулярные выражения. Библиотека программиста. СПб.: Питер, 2001. 352 с.
26. *Lin F., Krizhanovsky A.* Multilingual ontology matching based on Wiktionary data accessible via SPARQL endpoint // Proc. 13th Russian Conf. on Digital Libraries RCDL'2011. Voronezh, Russia. 2011. P.19-26. http://arxiv.org/abs/1109.0732 .