# Сравнение тезаурусов Русского и Английского Викисловарей, преобразованных в машинно-читаемый формат[1]

Крижановский Андрей

Санкт-Петербургский институт информатики и автоматизации РАН
Санкт-Петербург, 14 линия д.39, 199178
+7 (812) 328-80-71

andrew dot krizhanovsky@gmail.com

АННОТАЦИЯ

Викисловарь – это уникальный, значимый и богатый ресурс для автоматической обработки текста (NLP). В статье вслед за особенностями Викисловаря рассматривается архитектура парсера Викисловаря, в котором учтены эти особенности. Не оставлены без внимания открытые вопросы Викисловаря и сложности в реализации парсера. Построенный парсер извлекает значения слова, семантические отношения и переводы из Английского и Русского Викисловарей. Статья может быть интересна учёным и программистам, которые хотят использовать построенный машинный словарь для решения NLP задач либо желают построить парсер на основе данного проекта для обработки ещё одного из оставшихся неохваченными 170 Викисловарей. Выполнено сравнение словарных статей Английского и Русского Викисловарей, а именно были сравнены количество и тип семантических отношений, число значений слов, число переводов. Английский Викисловарь оказался больше по числу семантических отношений в полтора раза (157 и 100 тыс), однако в Русском Викисловаре больше слов «богатых» на отношения (например, в полтора раза больше словарных статей с числом семантических отношений больше трёх). Сравнение позволило выявить некоторые методологические недостатки викисловарей.

Ключевые слова: Викисловарь, словарь, тезаурус, лексикография, машинно-читаемый словарь, парсер.

Keywords: Wiktionary, Dictionary, Thesaurus, Lexicography, Machine-readable dictionary, Parser.

## 1. ВВЕДЕНИЕ

Викисловарь – это уникальный ресурс, который востребован во многих NLP задачах. Однако напрямую он использоваться не может. Он должен быть преобразован в формат, удобный для машинной обработки, т.е. в машинно-читаемый словарь (MRD). Этот «преобразователь» – парсер Викисловаря и описан в данной работе.

В статье представлена архитектура парсера. Описана база данных MRD, которая заполняется в ходе работы парсера. Созданное программное обеспечение находится в свободном доступе с открытой лицензией (http://code.google.com/p/wikokit) с тем, чтобы привлечь учёных и программистов к использованию построенного машинного словаря и развитию парсера.

Особенность словаря в том, что его создаёт сообщество энтузиастов, причём далеко не все из них являются профессиональными лингвистами лексикографами. Структура словаря постепенно, но постоянно меняется, т.к. сообщество постоянно обсуждает и вырабатывает новые правила оформления статей, может изменяться структура статей, не говоря уже о том, что постоянно растёт сам словарь, в него добавляются новые словарные статьи, и более того – добавляются новые языки. Сейчас в Английском Викисловаре примерно 740 языков, а парсер распознаёт языковые коды 540 языков.

Итак, обработка такого ресурса предъявляет серьёзные и специфические требования к парсеру, которые будут описаны в следующей главе. В третьей главе описана архитектура парсера Викисловаря. Сравнение основных характеристик Викисловарей и их тезаурусов представлено в четвёртой главе. Завершают статью обсуждение, обзор современных работ и заключение.

## 2. ТРЕБОВАНИЯ И РЕШЕНИЯ

До создания парсера Викисловаря уже были получены навыки при разработке компьютерных программ для поиска семантически близких слов в статьях Википедии [5], для построения индексной базы данных по текстам Википедии [6]. Закономерное желание использовать наработанный программный код (например, набор функций на языке Java для

[1] See English version of this paper: http://arxiv.org/find/cs/1/au:+Krizhanovsky_A/0/1/0/all/0/1

извлечения текстов из базы данных MediaWiki) привело к следующим требования при разработке парсера:

- программный код пишется на языке Java;
- для обработки берётся дамп базы даных Викисловаря, который загружается в MySQL.

*Требование*. Приведём требования, которыми должен обладать создаваемый парсер, чтобы он был надёжным, полезным и успешным в разработке.

> *Решение*. А так же приведём то решение, которые было выбрано как оптимальное для выполнения данного требования, вместе с замечаниями по реализации.

Итак, далее перечислены требования к программному коду парсера, структуре БД и процессу кодирования.

*Надёжность и устойчивость*. На данном этапе развития Викисловарей нет специальных средств, контролирующих ввод данных, поэтому участник-редактор может ввести код языка, которого не существует или дать бесконечно длинное толкование слова. Поэтому парсер должен корректно обрабатывать ошибки и неправильности.

> Данные требования достигаются за счёт тестирования и визуализации результата (см. ниже). Используется методика экстремального программирования, а именно: *Unit-тестирование*. Каждую нетривиальную функцию сопровождает один или несколько тестов. Кроме того, тесты несут дополнительную нагрузку, документируя работу функции на рабочих примерах.
>
> На июнь 2010 г. код содержит 233 успешно отрабатываемых тестов и 29 – неуспешно. Дамп Русского Викисловаря (2010 г., 300 тыс. статей) уже обрабатывается без фатальных ошибок. При обработке Английского Викисловаря (2010 г., 1.5 млн статей) было около 10 «сложных» словарных статей, которые вызвали аварийный останов парсера.

*Гибкость*. Изменение и улучшение оформления словарных статей, а точнее – правил оформления, к сожалению, не приводит к тому, что все статьи разом меняют свою структуру, даже несмотря на автоматизацию работ, достигаемую благодаря использованию механизма ботов. Новой структурой обладают в полной мере только те статьи, которые редактировались уже после принятия правила. Большой ряд статей отвечает предыдущим, устаревшим правилам оформления, что может длиться годами, до тех пор, пока до них не дойдут руки редактора-волонтёра. [2] Итак, парсер должен обладать значительной гибкостью, чтобы всё-таки извлекать данные из статей, отвечающим несколько разным форматам оформления.

> Выручает снова тестирование. Входом для Unit-тестов служат части словарной статьи в разных форматах.

*Визуализация*. Визуальная проверка результата парсинга позволит увидеть, какие поля БД для данной словарной статьи были успешно заполнены информацией из Викисловаря. Т. е. нужна возможность быстро отобразить все поля словарной статьи, сохранённой в БД машинно-читаемого словаря, либо убедиться, что эти данные не получилось распознать и сохранить. Визуализация позволит избежать утомительного формулирования рутинных низкоуровневых SQL-запросов.

> Параллельно с парсером разрабатывается приложение Wiwordik – визуальная оболочка (на языке JavaFX) машинно-читаемого словаря. Приложение позволяет осуществлять поиск по словарной статье, по переводу. Для выбранного слова отображается вся информация, которая была извлечена из соответствующей словарной статьи Викисловаря (Рис. 1).

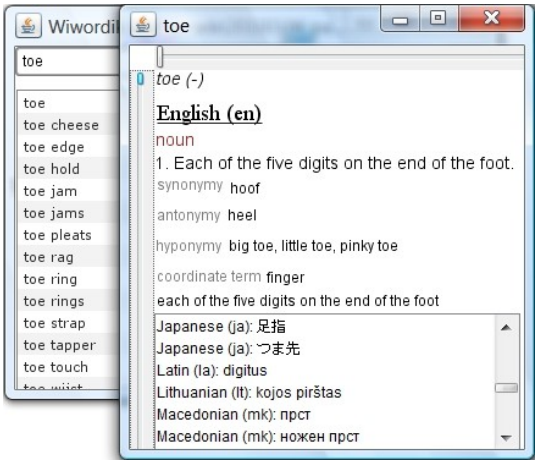


**Рис. 1. Данные об английском слове “toe", хранимые в машинно-читаемом словаре**

[2] Статьи, созданные в разные года, настолько сильно внешне отличаются друг от друга, что профессиональный редактор Викисловаря (так же как геолог по срезу горной породы) определяет примерный возраст статьи, которой не касалась рука человека. Определяет и приводит в божеский, т.е. надлежащий вид.

*Викисловарь ++*. (*Рост в ширину*). Безболезненное добавление модулей для парсинга новых Викисловарей. Для такого роста в ширину (с минимумом повторного излишнего кодирования) необходимо чёткое разделение кода парсера Викисловаря на две части: ядро, т.е. часть, не зависимая от языка, и часть, зависимая от языка, которая пишется заново для каждого добавляемого Викисловаря.

Текущая реализация парсера работает уже с двумя викисловарями: русским и английским. Язык (русский или английский) является одним из входных параметров парсера (Рис. 2, см. "Input").

*Инкрементальный подход (Рост в глубину)*. Безболезненное (т.е. без тотального переписывания кода) добавление модулей для парсинга новых подразделов словарной статьи в Викисловаре.[3]

Глупо пытаться извлечь сразу из Викисловаря все данные, которые в нём есть. На данный момент из Викисловаря извлекаются только: значение слова, семантические отношения и переводы слова.

По-видимому, инкрементальный подход уже заложен в разрабатываемый парсер, поскольку викитекст словарной статьи (в соответствии со структурой статьи [4], см. WT:ELE) делится сначала на самые крупные подразделы – *языки*, затем – подраздел *этимология*, после – *часть речи* (POS) и т.д. Поэтому, например, если потребуется извлечь данные из раздела "Pronunciation" (произношение), то найти ту часть программного кода в парсере, которая должна быть расширена – достаточно просто.

*Интеграция*. Важной задачей является интеграция данных, извлечённых из разных Викисловарей, в единую БД. Объединение необходимо, так как каждый Викисловарь составляется вручную и содержит уникальные данные, отсутствующие в других Викисловарях.

Полуоткрытый вопрос. С одной стороны, потенциал для решения этой задачи заложен в архитектуру парсера, т.к. база данных MRD имеет единую структуру и для Русского, и для Английского Викисловаря (Рис. 3). С другой стороны, практической вопрос – как объединить две БД в одну – остаётся открытым. Крайне возможно, что какая-то информация будет дублироваться, и более того – быть противоречивой, что потребует нескучных разработки нетривиальных алгоритмов по слиянию этих данных.

*Скорость*. Необходима достаточно быстрая работа парсера для приемлемого времени обработки дампа Викисловаря. Поскольку Викисловарь постоянно пополняется, необходимо регулярная обработка дампа для поддержания MRD в актуальном состоянии.

За сутки обрабатывается примерно 100-150 тыс. словарных статей. Обработка полуторамиллионного Английского Викисловаря занимает примерно 10 суток [5]. Работа парсера многократно ускоряется, если включены индексы у тех таблиц, по которым ведётся поиск. А поиск ведётся, поскольку, например, перед добавлением выполняется проверка – есть ли такое слово в базе MRD.

## 3. АРХИТЕКТУРА

Три части, изображённые в виде прямоугольников на рис. (Рис. 2), представляют собой данные, с которыми работает парсер, входные параметры и модули компьютерной программы.

*Данные*. Первопричиной всего является Викисловарь. В нём работают как проклятые десятки (в Русском) и сотни (в Английском Викисловаре) добровольных редакторов, благодаря чему Викисловарь постепенно растёт и развивается по каким-то своим внутренним законам. Регулярно, автоматически и по очереди создаются дампы («слепки») всех викисловарей и википедий и выкладываются на известные страницы интернета. Именно такой дамп, загруженный в локальную БД MySQL, и является тем сырцом, исходным материалом, который парсеру нужно перепахать и перелопатить, чтобы получить искомый результат в виде базы данных машинно-читаемого словаря, в которой части словарных статей будут разложены по полочкам (см. "Parsed Wiktionary DB" на Рис. 2).

[3] См. возможные подразделы в описании структуры статьи в Русском Викисловаре: http://ru.wiktionary.org/wiki/Викисловарь:Правила_оформления_статей

[4] См. структуру статьи, определённую в правилах Английского Викисловаря: http://en.wiktionary.org/wiki/Wiktionary:ELE

[5] Параметры компьютера: 3 Гб ОЗУ, 2.4 GHz Core 2 DUO.

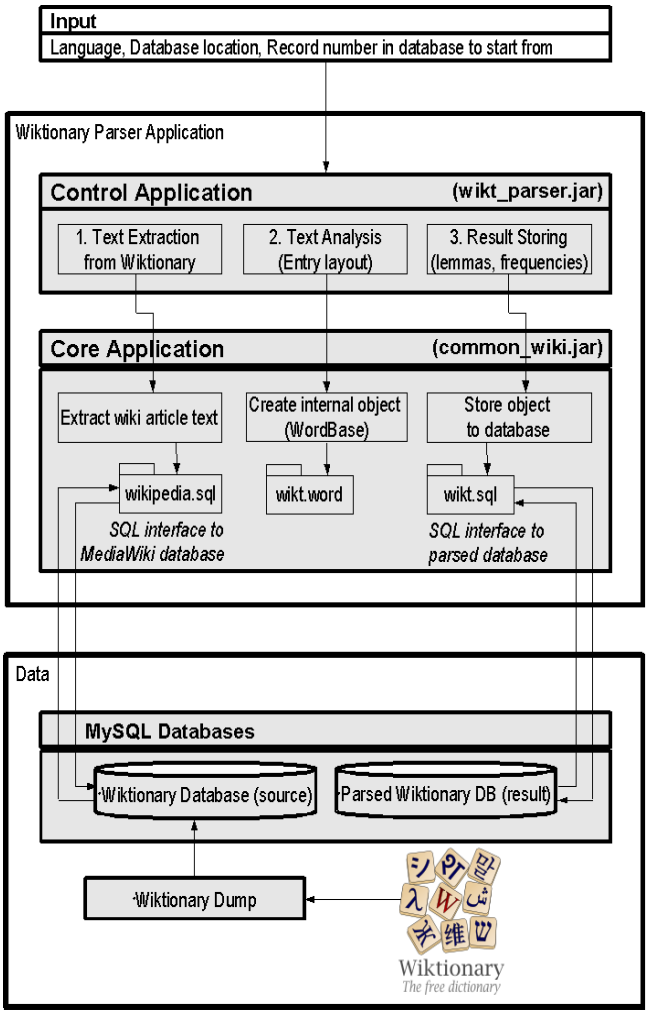


**Рис. 2. Архитектура парсера Викисловаря**

*Входные данные* (в верхней части рисунка) задают:

- native language – «родной» язык для данного Викисловаря, например, русский для Русского Викисловаря;
- параметры доступа к обеим БД (источника и приёмника);
- номер записи исходной БД, с которой нужно начать чтение и обработку данных парсером. Это позволяет безболезненно прервать обработку данных и затем продолжить её.

*Компьютерная программа* выполняет обработку словарной статьи в три шага:

1) Извлечение названия и текста словарной статьи из исходной БД Викисловаря. За эту работу отвечают классы пакета *wikipedia.sql*.
2) *Анализ текста* словарной статьи. Многажды используются разнообразные регулярные выражения для вычленения из текста словарной статьи интересующей информации.[6] Такое вычленение и анализ страницы возможен только благодаря известной структуре статьи и применению шаблонов внутри статьи. И чем более жёсткая структура статьи принята сообществом данного Викисловаря, тем проще и надёжнее алгоритмы парсера. Чем больше шаблонов, широко используемых в Викисловаре, тем легче извлечь информацию, структурированную с их помощью. (Известный проект DBpedia, извлекающий данные из Википедии, так же в большой степени полагается на шаблоны.)
   При анализе статьи одновременно создаётся промежуточный временный объект языка Java, а именно: класс *WordBase* пакета *wikt.word*. Иерархия его подклассов соответствует и структуре словарной статьи в Викисловаре (источник), и таблицам в создаваемой БД (приёмник), см. таблицу 3. Успешное наполнение информацией данного объекта является предпосылкой для следующего шага.
3) *Сохранение* созданного объекта класса WordBase (со всеми полями и подклассами) в БД машинно-читаемого словаря (Parsed Wiktionary DB на Рис. 2)

Для полноты картины необходимо привести схему базы данных машинно-читаемого словаря, которая часто упоминается в статье (Рис. 3).

По сравнению с предыдущей публикацией [7] на этой схеме появились новая таблица *index_native*, которая содержит список слов на родном для данного Викисловаря языке.

Индексные таблицы для прочих языков (не поместились на схеме) имеют вид *index_XX*, где XX – код языка. На данный момент автоматически создаются и заполняются 540 индексных таблиц для тех языков Английского Викисловаря, которые уже «известны» парсеру. Ещё необходимо добавить в код парсера несколько сотен кодов языков.

Эти индексные таблицы были добавлены для ускорения поиска слов на конкретном языке в программе *wiwordik* (Рис. 1). Раньше, без этих вспомогательных таблиц, поиск слова производился в огромной таблице *page*, в которой хранятся заголовки

[6] См. http://ru.wiktionary.org/wiki/Викисловарь:Шаблоны

всех словарных статей Викисловаря. Теперь, за счёт разделения слов по таблицам-языкам, в программе *wiwordik* можно оперативно получить список слов для заданного языка.

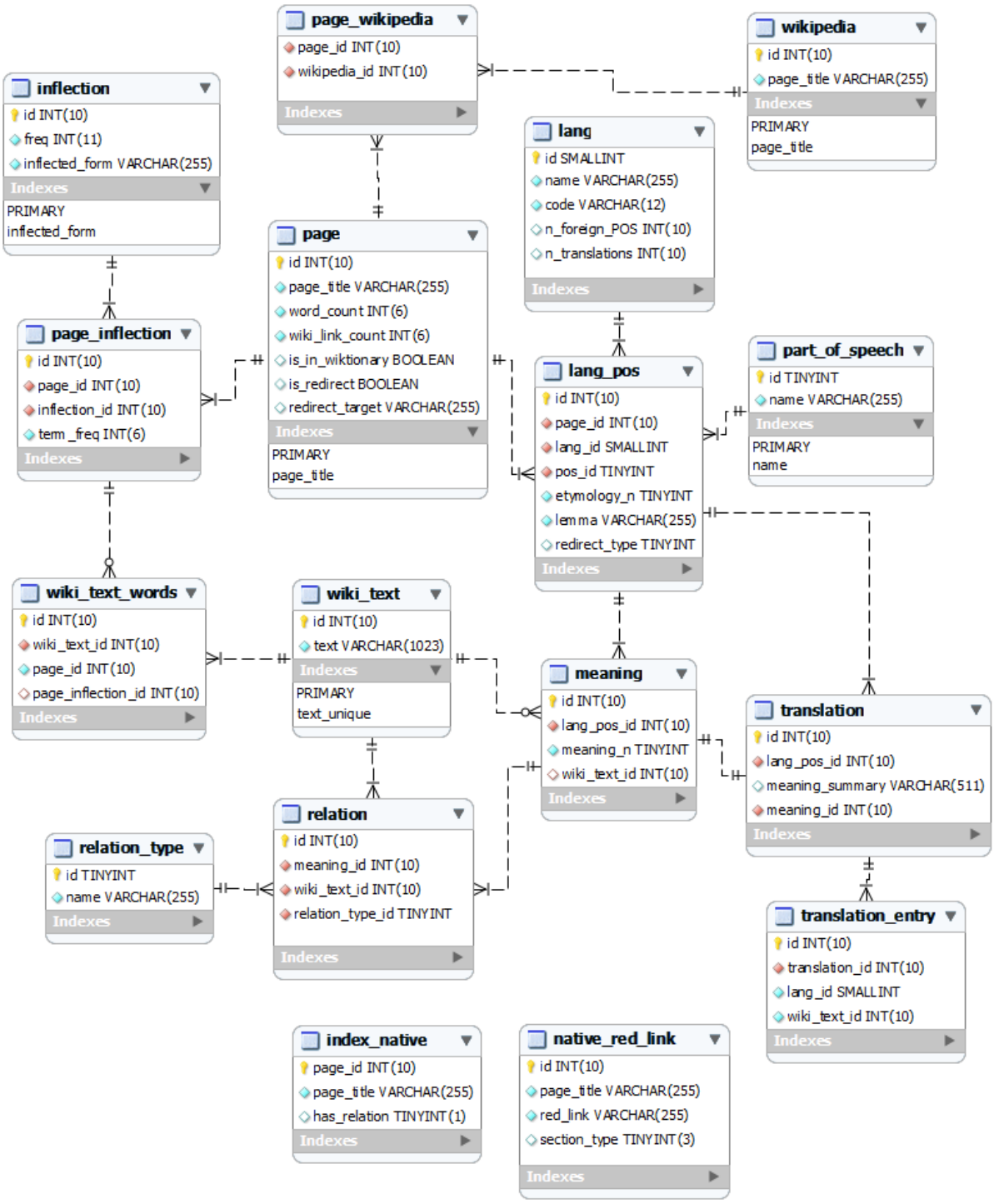


**Рис. 3. Таблицы и отношения в базе данных машинно-читаемого словаря**

## 4. ЭКСПЕРИМЕНТЫ

В ходе экспериментов были построены базы данных машинно-читаемые словари, что позволило сравнить Викисловари по различным параметрам. В качестве исходных данных были взяты дамп Английского Викисловаря (далее enwikt) от 6 января 2010 и дамп Русского Викисловаря (далее ruwikt) от 5 апреля 2010.

### 4.1 Сравнение основных характеристик Викисловарей

Были сравнены основные (с нашей точки зрения) показатели Викисловарей и размеры построенных по ним баз данных машинно-читаемых словарей (Таблица 1). Сравнение самих викисловарей (раздел А в таблице) показывает, что по числу страниц и числу активных участников Английский Викисловарь больше Русского примерно в семь раз. Однако в обоих словарях на каждую страницу приходится в среднем поровну, по пять правок редакторов и ботов.

Процентное соотношение числа страниц, имеющих хотя бы одно семантическое отношение в описании словарной статьи, к общему числу страниц в основном пространстве [7] (Таблица 1, строка 18) в Русском Викисловаре (10.7 %) почти в два раза больше, чем в Английсков Викисловаре (5.8 %).

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

1. Семантических отношений в Русском Викисловаре между русскими словами (16 строка таблицы) в два раза больше (84 тыс), чем между английскими словами в Английском Викисловаре (44 тыс).
2. В Английском Викисловаре словарные статьи об английских словах (строка 17) составляют пятую часть статей (18.3 %). В Русском Викисловаре процент словарных статей о словах родного языка значительно выше – больше половины (53.7 %). Таким образом, несмотря на общую цель обоих Викисловарей – описание всех словарных единиц всех языков, Русский Викисловарь является более моноязычным на данный момент.
3. Среднее число семантических отношений на словарную статью (строка 19) в Русском Викисловаре больше, чем в Английском почти в пять раз, и составляют соответственно 0.65 и 0.14.

По размерам таблиц БД построенных MRD (строки 5-13 в таблице) Английский Викисловарь превосходит Русский по всем параметрам в основном в 3-8 раз. Значительно больше, в 12.7 раза, различаются размеры таблицы “*meaning*” (см. обсуждение в разделе *5.2 Число значений, словоформ, лемм и жулики-боты*).

С другой стороны, интересны таблицы, размеры которых значительно меньше различаются, а именно: таблица “*relation*” (разница в 1.57 раза) и таблица “*translation*” (1.55). Посмотрим пристальнее на таблицу “*relation*” в следующем подразделе.

В данной статье не проводится сравнение данных, извлечённых из раздела «Перевод» словарной статьи и сохраняемых в таблице “translation”. Дело в том, что необходимо добавить в код парсера ещё значительное число языковых кодов тех языков, которые представлены в Английском Викисловаре. Несмотря на это доступна онлайн предварительная информация о числе переводов, извлечённых парсером из Английского[8] и Русского Викисловарей[9].

[7] В основное пространство входят словарные статьи, без учёта страниц помощи, шаблонов, категорий, страниц пользователей и т.п.

[8] См. http://en.wiktionary.org/wiki/User:AKA_MBG/Statistics:Translations

[9] См. http://ru.wiktionary.org/wiki/Участник:AKA_MBG/Статистика:Переводы

**Таблица 1. Основные показатели Английского Викисловаря, Русского Викисловаря и построенных на их основе баз данных машинно-читаемого словаря**

| N | Свойство | English (en) | Russian (ru) | en / ru |
|---|---|---|---|---|
| *(А) Статистика самих викисловарей (на 13 мая 2010)* | | | | |
| 1 | Число страниц | 1 721 584 | 241 573 | 7.13 |
| 2 | Число правок с момента установки Викисловаря | 9 230 581 | 2 529 788 | 3.65 |
| 3 | Среднее число правок на страницу | 4,96 | 4,8 | 1.03 |
| 4 | Активные участники | 1082 | 151 | 7.17 |
| *(Б) Статистика по базе данных, заполненной парсером викисловаря, размеры таблиц MRD (Рис. 3)* | | | | |
| | Дата дампа Викисловаря | Jan 6 2010 | April 5 2010 | – |
| | Название таблицы БД MRD (и комментарий)[10] | | | |
| 5 | page | 1 721 798 | 456 138 | 3.77 |
| 6 | **relation** (число семантических отношений) | 157 198 | 100 121 | **1.57** |
| 7 | lang_pos | 1 732 162 | 374 257 | 4.63 |
| 8 | wiki_text | 2 151 393 | 275 530 | 7.81 |
| 9 | wiki_text_words | 3 356 231 | 310 398 | 10.81 |
| 10 | *meaning* | *2 158 845* | *170 313* | *12.68* |
| 11 | inflection | 205 219 | 23 208 | 8.84 |
| 12 | **translation** (число блоков с переводами, т.е. число значений слов с переводами) | 59 321 | 38 306 | **1.55** |
| 13 | translation_entry | 373 008 | 189 844 | 1.96 |
| 14 | Число статей (число пар: язык – часть речи) с семантическими отношениями | 100 268 | 25 747 | 3.89 |
| 15 | Число статей на родном языке (те же пары: language & POS) | 315 343 | 129 669 | 2.43 |
| 16 | Число сем. отн. для слов на родном языке | **43 814** | **83 968** | **0.52** |
| *(В) Статистика, полученная из (А) и (Б)* | | | | |
| 17 | Число статей на родном языке (language & POS) к общему числу страниц [(15) / (1)], % | 18.32 | 53.68 | 0.34 |
| 18 | Статьи с сем. отн. к общему числу страниц [(14) / (1)], % | 5.82 | 10.66 | 0.55 |
| 19 | Среднее число сем. отн. для статей о словах на родном языке [(16) / (15)] | 0.14 | 0.65 | 0.21 |

[10] 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

## 4.2 Сравнение семантических отношений Викисловарей по числу и типу

Таблица "*relation*" машинно-читаемого словаря связывает тип семантического отношения (таблица "*relation_type*"), конкретное значение словарной статьи (таблица "*meaning*") и тот викитекст, который указан в словарной статье в разделе семантических отношений (таблица "*wiki_text*"), см. Рис. 2.

Размер таблицы "*relation*" равен числу семантических отношений, извлечённых из Викисловаря. Из Английского Викисловаря получилось извлечь 157 тыс таких отношений, из Русского Викисловаря – 100 тыс (Таблица 1).

С помощью данных MRD были получены следующие интересные цифры. Было посчитано – какое количество *семантических отношений приходится на одну словарную статью*. Результат представлен на Рис. 4. Например, словарная статья «выпь»[11] содержит 14 семантических отношений и 2 типа семантических отношений (гиперонимы, гипонимы).

Отметим однако, что число семантических отношений на Рис. 4 считается раздельно для омонимов, хотя они и включены в одну словарную статью Викисловаря. Например, первый омоним статьи «граф»[12] содержит всего 6, зато второй – уже 15 слов в разделе, посвящённом описанию семантических отношений.

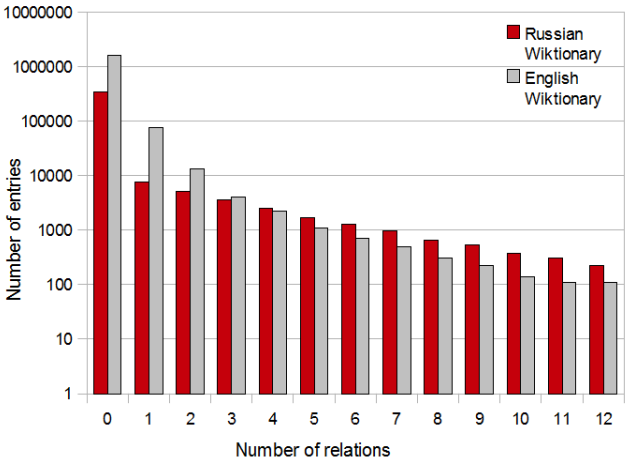


**Рис. 4. Сравнение числа словарных статей с разным числом семантических отношений (от 0 до 12) в Русском и Английском Викисловарях**

[11] См. http://ru.wiktionary.org/wiki/выпь

[12] См. http://ru.wiktionary.org/wiki/граф

Из исходных данных для Рис. 4 следует, что в Русском Викисловаре в 1.65 раза больше словарных статей с числом семантических отношений больше трёх. И наоборот, в Английском Викисловаре значительно больше статей с одним семантическим отношением (в 10 раз), с двумя (в 2.5 раза) и чуть больше с тремя (в 1.12 раза) чем в Русском Викисловаре. Итак, в Русском Викисловаре больше слов «богатых» на семантические отношения.

Таблица 2 содержит данные для ещё одного сравнения семантических отношений – по числу типов. Например, существительное "iron"[13] содержит 6 типов семантических отношений (Synonyms, Hypernyms, Hyponyms, Meronyms, Holonyms, Coordinate terms) из 9 возможных[14].

Из таблицы видно, что число тех слов в Русском Викисловаре, которые имеют по три или четыре типа семантических отношения (не говоря уже о пяти), значительно превышает эти значения в Английском Викисловаре (в 60-170 раз, Таблица 2). Гипотезу, объясняющую этот феномен см. в разделе «*5.3 Почему Русский Викисловарь побил Английский по числу семантических отношений?*»

**Таблица 2. Число слов с разным числом типов семантических отношений в Русском Викисловаре (ru) и Английском Викисловаре (en)**

| Types | Number of words | | ru / en | en / ru |
|---|---|---|---|---|
| | ru | en | | |
| 1 | 6254 | 16907 | 0.37 | **2.7** |
| 2 | 8167 | 3750 | 2.18 | 0.46 |
| **3** | 3215 | 53 | **60.66** | 0.02 |
| **4** | 844 | 5 | **168.8** | 0.01 |
| 5 | 45 | 1 | 45 | 0.02 |
| 6 | 6 | 1 | 6 | 0.17 |

Исходные табличные данные этого раздела и список слов, словарные статьи которых содержат большое количество семантических отношений, представлены онлайн.[15]

[13] См. http://en.wiktionary.org/wiki/iron#Noun

[14] См. http://en.wiktionary.org/wiki/Wiktionary:Semantic_relations

[15] См. http://ru.wiktionary.org/wiki/User:AKA_MBG/Статистика:Се

Следует признать, что сравнение семантических отношений не является полным, т.к. не были учтены данные Викизауруса в Английском Викисловаре. В своё оправдание можно сказать, что Викизаурус не получил развитие в других Викисловарях, и более того – Викизаурус описывает только английские слова многоязычного Английского Викисловаря. По этим причинам не планируется когда-нибудь в обозримом будущем «учить» парсер извлекать из него данные.

## 5. ОБСУЖДЕНИЕ

В разделе рассматриваются особенности Викисловаря как лингвистического ресурса, и то – как эти особенности должны быть учтены при разработке парсера. Также даны замечания по текущему состоянию реализации парсера, указаны его недостатки.

### 5.1 Проблемы и недостатки парсера

Сложностью обработки данных Викисловаря, является то, что задачей Викисловаря является «описание всех лексических единиц *всех* языков». Из 974 языков Английского Викисловаря[16] на данный момент в парсер вручную добавлено 479 соответствий: код языка – название языка.

Парсер пропускает словарные статьи, написанные на неизвестных ему языках. Поэтому число слов, значений и переводов, указанных в первой таблице, после добавления недостающих кодов языков должно увеличиться.

### 5.2 Число значений, словоформ, лемм и жулики-боты

В некоторых Викисловарях словоформы (например, «*хитрю*», «*хитришь*», «*хитрят*»,) описываются как отдельные лексические единицы. Такие словарные статьи–словоформы создаются ботами, т.е. автоматически. Несоблюдение требования лемматизации при подсчёте словарных статей, вероятно, и приводит к такому странному разбросу (Таблица 1) отношений размеров Викисловарей по разным параметрам при сравнении Английского Викисловаря и Русского.

В Русском Викисловаре боты, конечно, применяются для создания статей-заготовок для лемм (так называемые болванки или пустышки), в которых сразу присутствует структура статьи (см. следующий подраздел). Однако автоматическое создание статей–словоформ не поощряется сообществом.

В Английском Викисловаре с точностью до наоборот: боты создают статьи-словоформы, структура (пустые заголовки) новых статей автоматически не заполняется.

Поэтому одна из приоритетных задач при разработке парсера – научиться различать статьи, созданные вручную, где есть толкование, и статьи, сгенерированные ботом, где вместо значения идёт отсылка к основной статье – лемме. Такое различение позволит более точно сравнивать реальный размер словарей, без жульничества с ботами.

### 5.3 Почему Русский Викисловарь побил Английский по числу семантических отношений?

По многим параметрам Английский Викисловарь превосходит Русский в 3-8 раз (Таблица 1). Однако по числу семантических отношений разница всего в 1.57 раза в пользу Английского Викисловаря. Более того, по процентному соотношению числа страниц, имеющих хотя бы одно семантическое отношение в описании словарной статьи, к общему числу страниц Русский Викисловарь превосходит в два раза Английский.

Предложим две гипотезы для объяснения этого обстоятельства.

*Первая гипотеза:* наличие пустых полей (подзаголовков) в разделах "семантические отношения" (т.е. так называемые «болванки» или «стабы») увеличивают заполняемость данного раздела в целом по словарю.

Этой политики «пустых заголовков» придерживаются в Русском Викисловаре (см. фрагмент словарной статьи «*собака*» [17] на Рис. 6), но иначе считает сообщество Английского Викисловаря (см. фрагмент статьи "*dog*"[18] на Рис. 5).

---

мантические_отношения и http://en.wiktionary.org/wiki/User:AKA_MBG/Statistics:Semantic_relations

[16] См. полный список языков: http://en.wiktionary.org/wiki/Wiktionary:Index_to_templates/languages#Template_table

[17] См. http://ru.wiktionary.org/wiki/собака

[18] См. http://en.wiktionary.org/wiki/dog

**Synonyms**

- (*previous scientific names*): *Canis familiaris*, *Canis domesticus*
- (*animal*): domestic dog, hound, canine
- (*man*): bloke (*British*), chap (*British*), dude, fellow, guy, man
- (*morally reprehensible person*): cad, bounder, blackguard, fool, hound, heel, scoundrel
- (*mechanical device*): click, detent, pawl
- (*metal support for logs*): andiron, firedog, dogiron
- See also Wikisaurus:dog
- See also Wikisaurus:man

A dog (a Labrador)

**Hyponyms**

- (*animal*): sighthound, retriever, pointer, setter, shepherd, war dog, lapdog, guard dog, terrier

**Рис. 5. Пример описания семантических отношений в Английском Викисловаре, пустые подразделы отсутствуют**

**Синонимы**

1. пёс, псина; шутл.: друг человека, четвероногий друг
2. кобяк, скотина
3. хищник, насильник
4. собачка; перен., жарг.: обезьяна, обезьянка; офиц.: коммерческое at

**Антонимы**

1. -
2. -
3. -

Собака [1].

**Гиперонимы**

1. псовое, хищник, млекопитающее, позвоночное, животное, существо
2. -
3. -
4. символ, знак

**Гипонимы**

1. бульдог, дворняга, дог, овчарка, пудель, терьер, шнауцер; щенок
2. -
3. -

**Рис. 6. Пример описания семантических отношений в Русском Викисловаре, где один из подразделов пустой (Антонимы)**

По нашему глубокому убеждению наличие пустых заголовков, во-первых, провоцирует пользователя-редактора к их заполнению, во-вторых, упрощает работу. Упрощает, поскольку в отсутствии данных подзаголовков пользователь должен сначала создать их и только затем заполнять данными. При этом при создании подзаголовка нужно сообразить, в какое место словарной статьи заголовок должен быть добавлен по правилам словаря, а также нужно не ошибиться с числом знаков «=», которые указывает в вике-разметке уровень заголовка (для семантических отношений – это обычно 3, для омографов 4 уровень[19]).

*Вторую гипотезу* приведём в виде утверждения: информация, имеющая отношение к словарной статье, не должна «расползаться» по другим страницам проекта

Обсудим ещё раз идею проекта Викизаурус. Это специальный проект внутри enwikt, как раз и предназначен для удобной работы с семантическими отношениями. Однако «вынесение» семантических отношений на отдельную страницу с одной стороны не «провоцирует» пользователей расширять эти данные в словарной статье, с другой стороны – требует дополнительных усилий по работе над отдельной страницей Викизауруса.

Недостаток идеи Викизауруса в том, что одна и та же работа, например, указание синонимов слова – может выполняться и в Викизаурусе, и в словарной статье. Автоматическая синхронизация между ними отсутствует, поэтому информация может быть уникальной, может дублироваться или даже противоречить друг другу. По-видимому, на данный момент Викизаурус – это распыление небольших сил редакторов Викисловаря.

Таким образом, решением проблемы Английского Викисловаря (малое число семантических отношений относительно других параметров Викисловаря) может быть автоматическое создание пустых разделов семантических отношений при создании новых словарных статей, т.е. предлагается перенять успешный опыт Русского Викисловаря.

## 5.4 Явное указание языков как источник ошибок

С точки зрения разработчика парсера, явное написание *вручную* названий языков (в заголовках статей, в разделе переводов) является несомненным злом, т.к. легко позволяет совершить ошибку, опечатку. При этом пользователь никак не будет проинформирован об ошибке.

[19] См. http://en.wiktionary.org/wiki/Wiktionary:ELE#Etymology

Рассмотрим примеры из словарных статей “bush”[20] и «ангел»[21] (Таблица 3). Для указания языка, к которому принадлежит слово, описываемое в словарной статье, в enwikt явно пишется название языка, в ruwikt указывается специальный шаблон (строки 1-2 в таблице).

При указании языка в разделе переводов в enwikt применяется уже не настолько детский подход. Используется специальный шаблон {{t|}} (строка 3). Однако при этом происходит ненужное вредное дублирование, а именно: пользователь должен написать явно название языка (Finnish) и ещё указать его код (fi). В ruwikt эта проблема изящно решена с помощью огромного шаблона {{перев-блок|}} [22], в котором достаточно указать параметр – код языка, например “fi” для финского или “ko” для корейского языков (строки 3-4). Пока что в enwikt встречается и старый вариант перевода – вовсе без шаблонов, голым текстом (строка 4 в таблице).

**Таблица 3. Сравнение указания языков (явно или с помощью шаблонов) в Английском и Русском Викисловарях**

| N | English Wiktionary | Russian Wiktionary |
|---|---|---|
| Language of the entry | | |
| 1 | ==English== | = {{-en-}} = |
| 2 | ==Albanian== | = {{-sq-}} = |
| Translation section | | |
| 3 | * Finnish: {{t+\|fi\|pensas}} | \|fi=[[enkeli]] |
| 4 | * Korean: [[수풀]] (supul) | \|ko=[[천사]] |

Итак, из амбразуры парсера видится, что в этом плане русская редакция Викисловаря более продвинута, чем английская, поскольку (i) и для названий языков в заголовках словарных статей, (ii) и для указания языков в переводах – везде используется жёсткий, требующий дополнительных усилий (нужно выучить коды языков), но более надёжный механизм шаблонов. Тогда, если пользователь ошибся и ввёл, например, вместо “et” код языка “es”, то он сразу увидит надпись «Испанский» вместо искомого «Эстонского». Увидит и сразу исправит.

[20] См. http://en.wiktionary.org/wiki/bush

[21] См. http://ru.wiktionary.org/wiki/ангел

[22] См. http://ru.wiktionary.org/wiki/Шаблон:перев-блок

### 5.5 Уникальность данных каждого викисловаря

В статье уже неоднократно утверждалось, что каждый Викисловарь содержит уникальные данные, отсутствующие в других Викисловарях.

Эти слова может подтвердить, например такой факт. Между русскими словами в Английском Викисловаре указано всего лишь 3.9 тыс семантических отношений, зато в Русском Викисловаре – целых 84 тыс. И наоборот, между английскими словами в Русском Викисловаре – 3.4 тыс, в Английском Викисловаре – 44 тыс.

Оставим на будущее (себе или читателю) такие задачи, которые более основательно подтвердят или опровергнут уникальность данных Английского и Русского Викисловарей, сравнивая построенные MRD словари. Итак предлагается:

- указать степень покрытия / пересекаемости слов разных языков, слов со значениями, наличия семантических отношений;
- составить список уникальных и почти уникальных языков, представленных только в одном словаре (красная книга языков);
- построить два упорядоченных списка: языки, лучше представленные в одном и в другом Викисловаре по разным параметрам (число значений, семантических отношений).

## 6. ИССЛЕДОВАНИЯ В ДАННОЙ ОБЛАСТИ

Несомненно, задача преобразования бумажных и электронных словарей в машинно-читаемый формат стояла задолго до появления Викисловарей [8], [14]. Однако только сейчас появился такой удивительный ресурс, предоставляющий беспрецендентный объем лексикографических данных на всех языках мира.

При решении задач автоматической обработки текста огромной популярностью пользуются тезаурусы, созданные вручную (например, WordNet). В практи–ческих приложениях также активно пользуются тезаурусы, сгенерированные автоматические, например по данным Википедии или Веб. Тезарус, пред–став–ленный в данной работе и являющийся частью MRD, занимает, по-видимому, промежуточное положение.

Не только Викисловарь, но и Википедию можно рассматривать, как тезаурус. Разрабатываются специальные алгоритмы для извлечения семантических отношений из Википедии. Например, из Японской Википедии [12] извлекают гипонимы и гиперонимы. Из текстов статей биологической тематики Английской

Википедии извлекают и другие таксономические отношения для построения онтологий [4].

Тем не менее, исследований, посвящённых непосредственно Викисловарю, крайне мало. В качестве примера можно привести работу [15], в которой представлен программный интерфейс (API) к Википедии и Викисловарю (немецкая и английская редакция викисловарей).

Есть ряд работ, посвящённых сравнению Викисловарей и других тезаурусов. В нашей предыдущей работе [7] мы сравнили поиск семантически близких слов на основе Русского Викисловаря и WordNet в пользу WordNet. В работе [9] выполнено сравнение трёх ресурсов: Викисловарь, OpenThesaurus, and GermaNet. Оказалось, что Немецкий Викисловарь содержит меньше всего семантических отношений (157 тысяч на июнь 2009).

Викисловарь, в свою очередь, может быть источником данных для построения других тезаурусов. Так в работе [3] описано построение французского и словенского WordNet'ов на основе данных, извлечённых из французского, словенского и английского викисловарей.

## 7. ЗАКЛЮЧЕНИЕ

Разработана архитектура модульного и расширяемого парсера Викисловаря. Реализованы модули для извлечения трёх блоков данных из словарных статей, а именно: значение слова, семантические отношения и перевод. Модули адаптированы к особенностям и разнице в оформлении и структуре статей Английского и Русского Викисловарей. Расширяемость модульной архитектуры парсера заключается в том, что без значительного переписывания программного кода возможны:

- добавление в парсер модулей для извлечения других подразделов словарных статей (например, произношение, этимология и т.д.);
- адаптация существующих модулей для извлечения данных из других Викисловарей (французский, китайский и т.д.).

Выполнено сравнение словарных статей Английского и Русского Викисловарей, а именно: сравнили количество и тип семантических отношений, число значений слов, число переводов. Получены следующие интересные результаты:

(1) Английский Викисловарь содержит больше семантических отношений в полтора раза (157 и 100 тыс).

(2) Процентное соотношение числа страниц, имеющих хотя бы одно семантическое отношение в описании словарной статьи, к общему числу страниц в основном пространстве в Русском Викисловаре (10.7 %) почти в два раза больше, чем в Английсков Викисловаре (5.8 %).

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

   a. Семантических отношений в Русском Викисловаре между русскими словами в два раза больше (84 тыс), чем между английскими словами в Английском Викисловаре (44 тыс).

   b. В Английском Викисловаре словарные статьи об английских словах составляют пятую часть статей (18.3 %). В Русском Викисловаре процент словарных статей о словах родного языка значительно выше – больше половины (53.7 %). Таким образом, несмотря на общую цель обоих Викисловарей – описание всех словарных единиц *всех языков*, Русский Викисловарь является более моноязычным на данный момент.

   c. Среднее число семантических отношений на словарную статью в Русском Викисловаре больше, чем в Английском почти в пять раз, и составляют соответственно 0.65 и 0.14.

Создание машинно-читаемых словарей является важным кирпичиков в основании небоскрёба автоматической обработки текста. MRD словари, и в том числе данные Википедии и Викисловарей, используются при построении онтологий [13], [16]; в машинном переводе [2], [10], при автоматическом упрощении текста [11], при поиске изображений [2], при разрешение лексической многозначности [8].

Видно много заманчивых дорог, пойдя по которым можно развивать и парсер, и приложения на его основе. Однако в первую очередь нужно создать графический интерфейс к машинно-читаемому словарю, построенному по данным Английского Викисловаря. Для Русского Викисловаря такая оболочка уже создана и доступна онлайн.[23]

[23] См. программу *wiwordik* , построенную на основе данных Русского Викисловаря: http://code.google.com/p/wikokit/wiki/wiwordikRu

## ССЫЛКИ


[1] Bizer, C.; Lehmann, J.; Kobilarov, G.; Auer, S.; Becker, C.; Cyganiak, R.; Hellmann, S. 2009. "DBpedia - A crystallization point for the Web of Data". Web Semantics: Science, Services and Agents on the World Wide Web 7 (3): 154-165. ISSN 1570-8268. http://www.wiwiss.fu-berlin.de/en/institute/pwo/bizer/research/publications/Bizer-etal-DBpedia-CrystallizationPoint-JWS-Preprint.pdf

[2] Etzioni, O., Reiter, K., Soderland, S., and Sammer, M. (2007). Lexical Translation with Application to Image Search on the Web. In the proceedings of MT Summit XI. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.73.7536&rep=rep1&type=pdf

[3] Fiser D., Sagot B. (2008). Combining multiple resources to build reliable wordnets. In TSD 2008, Brno, Czech Republic. http://alpage.inria.fr/~sagot/pub-en.html

[4] Herbelot A., Copestake A. (2006). Acquiring ontological relationships from wikipedia using rmrs. In Proceedings of the ISWC 2006 Workshop on Web Content Mining with Human Language Technologies. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.132.8469&rep=rep1&type=pdf

[5] Krizhanovsky, A. A. 2006. Synonym search in Wikipedia: Synarcher. In Proceedings of the 11-th International Conference "Speech and Computer" (The St. Petersburg, Russia, June 26 - 29, 2006). SPECOM '06. 474-477. http://arxiv.org/abs/cs/0606097

[6] Krizhanovsky, A. A. 2008. Index wiki database: design and experiments. In Proceedings of the Corpus Linguistics (The St. Petersburg, The Russia, October 6 - 10, 2008) CORPORA '08. http://arxiv.org/abs/0808.1753.

[7] Krizhanovsky, A. A.; Feiyu Lin. Related terms search based on WordNet / Wiktionary and its application in Ontology Matching. In Proceedings of the 11th Russian Conference on Digital Libraries RCDL'2009. September 17-21, Petrozavodsk, Russia. 363-369. http://arxiv.org/abs/0907.2209

[8] Krovetz R., Croft W. B. Word sense disambiguation using machine-readable dictionaries. In Proceedings of the 12th annual international ACM SIGIR conference on Research and development in information retrieval, p.127-136, June 25-28, 1989, Cambridge, Massachusetts, United States. http://elvis.slis.indiana.edu/irpub/SIGIR/1989/pdf14.pdf

[9] Meyer C. M., Gurevych I. Worth its Weight in Gold or Yet Another Resource — A Comparative Study of Wiktionary, OpenThesaurus and GermaNet. In Proceedings of the 11th International Conference on Intelligent Text Processing and Computational Linguistics, p. 38-49. Iasi, Romania, 2010. http://www.informatik.tu-darmstadt.de/fileadmin/user_upload/Group_UKP/publikationen/2010/cicling2010-meyer-lsrcomparison.pdf

[10] Muller, C., Gurevych, I. (2008). Using Wikipedia and Wiktionary in Domain-Specific Information Retrieval. In: Working Notes for the CLEF 2008 Workshop, Aarhus, Denmark, September 17-19. http://www.clef-campaign.org/2008/working_notes/mueller-paperCLEF2008.pdf

[11] Napoles C., Dredze M. (2010) Learning Simple Wikipedia: A Cogitation in Ascertaining Abecedarian Language. Workshop on Computational Linguistics and Writing: Writing Processes and Authoring Aids at NAACL-HLT. http://www.cs.jhu.edu/~mdredze/

[12] Sumida A., Torisawa K. (2008). Hacking Wikipedia for Hyponymy Relation Acquisition. In Proceedings of International Joint Conference on NLP (IJCNLP'08). http://acl.eldoc.ub.rug.nl/mirror/I/I08/I08-2126.pdf

[13] Wandmacher, T., Ovchinnikova, E., Krumnack, U. and Dittmann, H. (2007). Extraction, evaluation and integration of lexical-semantic relations for the automated construction of a lexical ontology. In Proc. Third Australasian Ontology Workshop (AOW 2007), Gold Coast, Australia. CRPIT, 85. Meyer, T. and Nayak, A. C., Eds. ACS. 61-69. http://crpit.com/abstracts/CRPITV85Wandmacher.html

[14] Wilms G. J. (1990). Computerizing a Machine Readable Dictionary. In Proceedings of the 28th annual Southeast regional conference, ACM Press. 306–313. http://computerscience.uu.edu/faculty/jwilms/papers/acm90/acm90.pdf

[15] Zesch T., Mueller C., Gu¬re¬vych I. (2008). Extracting lexical semantic knowledge from Wikipedia and Wiktionary. In Proceedings of the Conference on Language Resources and Evaluation (LREC). http://elara.tk.informatik.tu-darmstadt.de/publications/2008/lrec08_camera_ready.pdf

[16] Рубашкин, В.Ш.; Бочаров, В.В.; Пивоварова, Л.М.; Чуприн Б.Ю. Опыт автоматизированного пополнения онтологий с использованием машиночитаемых словарей. Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной Международной конференции «Диалог» (Бекасово, 26-30 мая 2010

г.). Вып. 9 (16). - М.: Изд-во РГГУ, 2010. http://www.dialog-21.ru/dialog2010/materials/html/63.htm