Evaluation of the corpora and digital libraries used in Russian Wiktionary

Krizhanovsky A. A.
St.Petersburg Institute for Informatics and Automation of RAS

An electronic corpus is an essential tool for a lexicographer for creating dictionary. An amount of using of corpora and digital libraries in illustrating definitions in Wiktionary is estimated in this paper. The Wiktionary is a multilingual and multifunctional dictionary and thesaurus. 51.5 thousands of quotations were automatically extracted. It was found that 17 000 of quotations refer to a corpus or a digital library as a source of the quotation. The main source is the Russian National Corpus, which is referred by 16 158 quotations (or 95% of quotations with references to sources) in the Russian Wiktionary.

Крижановский А.А.
Санкт-Петербургский институт информатики и автоматизации Российской академии наук

Оценка использования корпусов и электронных библиотек в Русском Викисловаре[1]

1. Введение

Корпус является важным инструментом лексикографов при создании словарей. В этой работе оценивается – в каком объёме используются различные корпуса и электронные библиотеки в

[1] В статье представлены результаты, полученные в рамках проектов РФФИ 09-07-00066, 09-07-00436, 11-01-00251 и проекта Программы фундаментальных исследований Президиума РАН "Интеллектуальные информационные технологии, математическое моделирование, системный анализ и автоматизация".

Русском Викисловаре для иллюстрации значений и употреблений слов и словосочетаний.

Всё большую популярность приобретает Викисловарь, свободно пополняемый многофункциональный многоязычный словарь и тезаурус. В Викисловаре содержатся толкования и переводы слов, описание фонетических и морфологических свойств, семантические отношения. На начало 2011 г. Русский Викисловарь (всего существует порядка 170 викисловарей) содержал больше 280 тысяч словарных статей.

Причиной популярности Словаря является его постоянное пополнение новыми данными и серьёзное отношение к работе над содержанием, в том числе внимательный подбор иллюстраций для значений слов. Для единообразного представления и единого формата добавления цитат в Викисловаре используется механизм шаблонов. Сейчас в Русском Викисловаре используется около 30 шаблонов для указания источника данных для цитат (например, Национальный корпус русского языка, Фундаментальная электронная библиотека и др.).

Для анализа и оценки использования в Викисловаре данных корпусов и электронных библиотек была доработана система автоматического извлечения данных из Викисловаря (парсер), был расширен машинно-читаемый словарь, который строится этой системой. Модульная структура парсера Викисловаря позволила добавить программный блок для обработки цитат. В базу данных машинно-читаемого словаря были добавлены таблицы, в которых теперь хранится информация о цитатах Викисловаря. Наличие этих таблиц (т.е. структурирование информации о цитатах) позволило построить SQL-запросы и численно оценить объём словарных единиц, снабжённых цитатами, количество обращений к корпусам и т. д.

2. Особенности оформления цитат в Русском Викисловаре

На данный момент из Русского Викисловаря извлекается и сохраняется в базу данных машинно-читаемого словаря только часть словарной информации: толкование, семантические отношения (синонимы, гиперонимы и т. д.), переводы и цитаты. По правилам Викисловаря каждое толкование должно сопровождаться одной или несколькими цитатами, иллюстрирующими употребление слова.

Ниже представлены фрагменты словарных статей с толкованиями и цитатами.

БУД**И**РОВАТЬ

1. *устар.* дуться, сердиться на кого-либо; выражать недовольство кем-либо; быть настроенным против кого-либо, быть в конфликте с кем-либо ◆ Анна Нерягина появилась в московском обществе пятнадцать лет тому назад, когда она вышла замуж за состоятельного эксдипломата, **буди́ровавшего** правительство и потому покинувшего Петербург. *Валерий Брюсов, «Через пятнадцать лет», 1909 г.* (цитата из Национального корпуса русского языка, см. Список литературы) ◆ А мы с одним министерством **буди́ровали**, ну, нас и по шапке. *Д. Н. Мамин-Сибиряк, «Черты из жизни Пепко», 1894 г.* (цитата из Национального корпуса русского языка, см. Список литературы)
2. *разг.* возбуждать, будоражить; поднимать (вопросы и т. п.) ◆ Они **буди́ровали** крестьян против неё, и это обстоятельство толкало меня на более решительные действия. *Н. И. Махно, «Воспоминания», 1929 г.* (цитата из Национального корпуса русского языка, см. Список литературы)

ОК**О**ШЕЧКО

1. *уменьш.* к окошко, окно; маленькое окно ◆ В доме Гаврилы Афанасьевича из сене́й на право находилась тесная кама́рка с одним **око́шечком**. *А. С. Пушкин, «Арап Петра Великого», 1828 г.* (цитата из Национального корпуса русского языка, см. Список литературы)

2. *перен.* место в помещении организации, предназначенное для общения клиента с её представителем один на один ◆ Подал бланк, облокотился возле **окóшечка** и стал считать деньги. *В. М. Шукшин, «Калина красная», 1973 г.* (цитата из Национального корпуса русского языка, см. Список литературы)

Иллюстрирующие примеры сопровождаются следующей информацией: автор цитаты, название и дата создания произведения, источник цитаты.

Для удобства читателя слово, иллюстрируемое цитатой, выделено визуально. Несколько цитат для одного толкования упорядочены по времени написания. Цитаты разделены знаком ромбика ◆ благодаря использованию специального шаблона для цитат. Наличие этого шаблона позволило извлечь цитаты из текстов словарных статей с помощью компьютерной программы и сохранить в базу данных машинно-читаемого словаря. В следующем разделе представлены оценки, полученные при анализе этой базы.

3. Эксперименты

Далее представлены оценки, характеризующие цитаты, извлечённые из версии Русского Викисловаря от 13 апреля 2011 г. Построение базы данных машинно-читаемого словаря по данным Викисловаря описано в предыдущей работе[2].

3.1. Языки цитат

С помощью программы, извлекающей данные из Викисловаря, получилось подсчитать, что Русский Викисловарь содержит всего 51.5 тысячу цитат. При этом 42 тысячи (82 % от всего числа цитат) иллюстрируют русские слова.

[2] *Крижановский А. А.* Преобразование структуры словарной статьи Викисловаря в таблицы и отношения реляционной базы данных. 2010. **http://scipeople.com/publication/100231/**

Второй язык по числу цитат — немецкий – содержит 1522 цитаты (3 % от всех цитат). В десятку языков с наибольшим числом цитат в Русском Викисловаре входят также английский, латинский, татарский, сербский, французский, украинский, калмыцкий и польский. 46 языков имеют больше 10 цитат, 15 языков — больше ста цитат. Словарные статьи с цитатами представлены для 108 языков.

3.2. Корпуса текстов и электронные библиотеки

В Русском Викисловаре для 17 тысяч цитат (33 % цитат) указан источник, из которого получена данная цитата. Главным источником является Национальный корпус русского языка[3], на который ссылается 95 % цитат с источниками. Далеко позади, на втором месте, оказалась электронная Библиотека Максима Мошкова, на которую ссылается 215 цитат (1.3 %). Менее процента, в порядке убывания, получили следующие пять ресурсов: Oxford Latin Dictionary, Викитека, Толковый словарь живого великорусского языка В. И. Даля, словарь Ушакова[4], BYU Corpus of American English[5].

В Викисловаре разработаны специальные шаблоны, упрощающие указание источников данных. На 2011 год есть возможность выбрать один из 20 источников. Список рекомендованной литературы (включающий корпуса текстов, бумажные и электронные словари) содержит более полутора сотен наименований[6].

---

[3] *Гришина Е. А., Плунгян В. А.* Перспективы развития Национального корпуса русского языка // Национальный корпус русского языка. М.: Индрик, 2005. **http://www.ruscorpora.ru/corpora-biblio.html**

[4] Толковый словарь русского языка: В 4-х т. / Под ред. Д. Н. Ушакова. — М.: Сов. энцикл.: ОГИЗ, 1935—1940.

[5] См. **http://www.americancorpus.org/**

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

3.3. Авторы цитат

Одним из параметров, указываемых в шаблоне цитат, является имя автора произведения. Для 23.5 тысяч цитат автор указан, что составляет 46% от всего числа цитат. В таблице 1 приведены десять писателей, чьи произведения наиболее часто цитируются в Русском Викисловаре. В таблице 1 в колонке «имя» приведено имя автора, наиболее часто используемое в Викисловаре (см. далее о кластеризации).

Таблица 1. Самые популярные авторы цитат в Викисловаре

Автор	**Число цитат**
Чехов	716
Л. Н. Толстой	529
Пушкин	520
Достоевский	500
Тургенев	457
Гоголь	321
Стругацкие	171
Лесков	245
Булгаков	207
Виктор Астафьев	142

Словарь содержит 5300 уникальных имён авторов, что, однако, превышает их реальное количество. Дело в том, что при редактировании словарных статей неизбежны опечатки в именах авторов, это и увеличивает число авторов.

Чтобы найти эти опечатки был реализован алгоритм кластеризации на основе вычисления между словами расстояния Jaro-Winkler. Для вычислений была использована открытая Java библиотека SimMetrics.

В таблице 2 представлено несколько небольших кластеров, группирующих близкие (в смысле метрики Jaro-Winkler) имена. Полный список кластеров представлен на странице проекта[7].

Таблица 2. Примеры кластеров имён авторов

Автор	**Число цитат**
Cluster 13	
Ю. К. Олеша	10
Ю. Олеша.	1
Cluster 25	
Уладзімір Арлоў	1
Уладізір Арлоў	1
Cluster 304	
О. Генри	4
О'Генри	1
Cluster 304	
Горький	105
Горький.	1
М.Горький	9
М. Горький	7

Заключение

В предыдущей работе [8] сравнивались тезаурусы Русского Викисловаря и Английского Викисловаря. Было получено, что в Английском Викисловаре словарные статьи об английских словах составляют пятую часть от всех 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 –

[7] **http://ru.wiktionary.org/wiki/Участник:AKA_MBG/Статистика:Цитаты_(авторы)**

[8] *Крижановский А. А.* Сравнение тезаурусов Русского и Английского Викисловарей, преобразованных в машинно-читаемый формат. 2010. **http://scipeople.com/publication/99331/**

описание всех словарных единиц всех языков, Русский Викисловарь оказался более моноязычным по данным на 2010 г.

В данной работе получены дополнительные данные, подтверждающие моноязычность Русского Викисловаря, а именно: 42 тысячи цитат (82 % от всего числа цитат в 51.5 тысячу) иллюстрируют русские слова. На немецком языке (втором по числу цитат) представлено 1522 цитаты, т.е. только 3 % от всех цитат.

Анализ Русского Викисловаря показал, что для 17 тысяч цитат (33 % от всего числа цитат) указан корпус текстов или электронная библиотека, из которого получена цитата. Главным источником оказался Национальный корпус русского языка, на который ссылается 95 % цитат с источниками.

Из 5300 авторов цитат у редакторов Викисловаря наибольшей популярностью пользуются А. П. Чехов, Л. Н. Толстой, А. С. Пушкин, Ф. М. Достоевский, И. С. Тургенев, Н. В. Гоголь, Аркадий и Борис Стругацкие, Н. С. Лесков, М. А. Булгаков и В. П. Астафьев.

Численная оценка цитат Русского Викисловаря получена с помощью разработанной компьютерной системы автоматического извлечения данных из Викисловаря.

Поскольку Викисловарь создаётся сообществом редакторов на добровольной основе, поэтому правила в нём (в особенности выбор источников для иллюстрирования значений слов) носят скорее рекомендательный, нежели жёсткий ограничительный характер. Поэтому такой инструмент (в виде компьютерной программы), дающий численную оценку материалу, включённому в Словарь, будет интересен и редакторам, и читателям. Редакторам он поможет скорректировать их деятельность. Читатели смогут лучше себе представлять возможности Словаря. Косвенно, результаты работы позволяют сравнить популярность корпусов и электронных библиотек.