Извлечение и анализ дат произведений в корпусе цитат онлайн-словаря

Retrieval and analysis of quotes' dates in corpus of quotes of online dictionary

А. А. Крижановский[1], Н. Б. Луговая[2], В. М. Круглов[3]
Andrew.Krizhanovsky@gmail.com, Nataly@krc.karelia.ru, VMKruglov@yandex.ru

[1] Санкт-Петербургский институт информатики и автоматизации РАН, Санкт-Петербург, Россия
[2] Институт прикладных математических исследований КарНЦ РАН, Петрозаводск, Россия
[3] Институт лингвистических исследований РАН, Санкт-Петербург, Россия

Ключевые слова: лексикография, машиночитаемый словарь, Викисловарь, корпусная лингвистика

Quantitative evaluation of quotations in the Russian Wiktionary was performed with the use of the developed Wiktionary parser. It was found that the number of quotes in the dictionary grows fast (51.5 thousands in 2011, 62 thousands in 2012). These quotes were extracted and stored to the database of the machine-readable dictionary. The tables of the relational database of the machine-readable dictionary related to the quotations were designed. The histogram of distribution of quotations of literary works created in different years was built.

Введение. Развитие компьютерных технологий обеспечило появление словарей нового типа – онлайн-справочников, в создании которых может принять участие широкий круг будущих заинтересованных пользователей. По сравнению с традиционной лексикографией подобная организация работы, с одной стороны, обеспечивает несомненные преимущества (высокий темп

работы, возможность обсуждения и редактирования словарных статей на любом этапе составления), с другой – обнаруживает существенный недостаток, который заключается в высокой вероятности появления лакун как в используемом материале, так и в тексте самого словаря. Представляется, что данная проблема в силу своей актуальности заслуживает особого внимания и может быть решена путем создания специальных программ, осуществляющих автоматический анализ онлайн-словаря на любом этапе его создания. В настоящей статье на примере анализа иллюстративного материала, представленного в «Русском Викисловаре», будут намечены и продемонстрированы некоторые возможные подходы к решению названной проблемы.

Викисловарю – это многофункциональный словарь, сочетающий тезаурус, толковый и фразеологических словарь. В Викисловаре содержатся переводы слов, описание фонетических и морфологических свойств, семантические (парадигматические) отношения, этимология слов.

Достоинствами Викисловаря является большой объём и разнообразие лексикографических данных. В работах [Крижановский 2011а], [Meyer 2012] показано, что по объёму информации Немецкий Викисловарь сопоставим с тезаурусами *GermaNet* и *OpenThesaurus*, а Английский Викисловарь даже превосходит объём данных *WordNet*. Словаря, сравнимого по объёму и разнообразию лексикографических данных с Русским Викисловарём и находящегося в открытом доступе, на данный момент, по-видимому, нет.

Значения слов в онлайн-словаре сопровождаются цитатами из литературных произведений. В этом смысле Викисловарь продолжает традиции толковых словарей [Леденёва 2008].

Целью статьи является анализ дат литературных произведений, использующихся в качестве источников цитат. Эксперименты проводились на основе корпуса цитат, построенного по данным Русского Викисловаря. База данных корпуса цитат является частью машиночитаемого Викисловаря, находящегося в открытом доступе.[1]

[1] См. http://code.google.com/p/wikokit/

Архитектура базы данных корпуса цитат. База данных корпуса цитат является частью реляционной базы данных машиночитаемого словаря, представленной в работе [Крижановский 2010].

При извлечении слабоструктурированных данных Викисловаря распознаются такие поля цитат (см. рис. 1), как:

- текст цитаты,
- перевод на русский язык;
- транскрипция цитаты;
- информация о цитате:
    - название произведения;
    - автор произведения;
    - издание для цитат из журнальных, газетных статей;
    - дата создания произведения;
    - корпус текстов, откуда взята цитата.

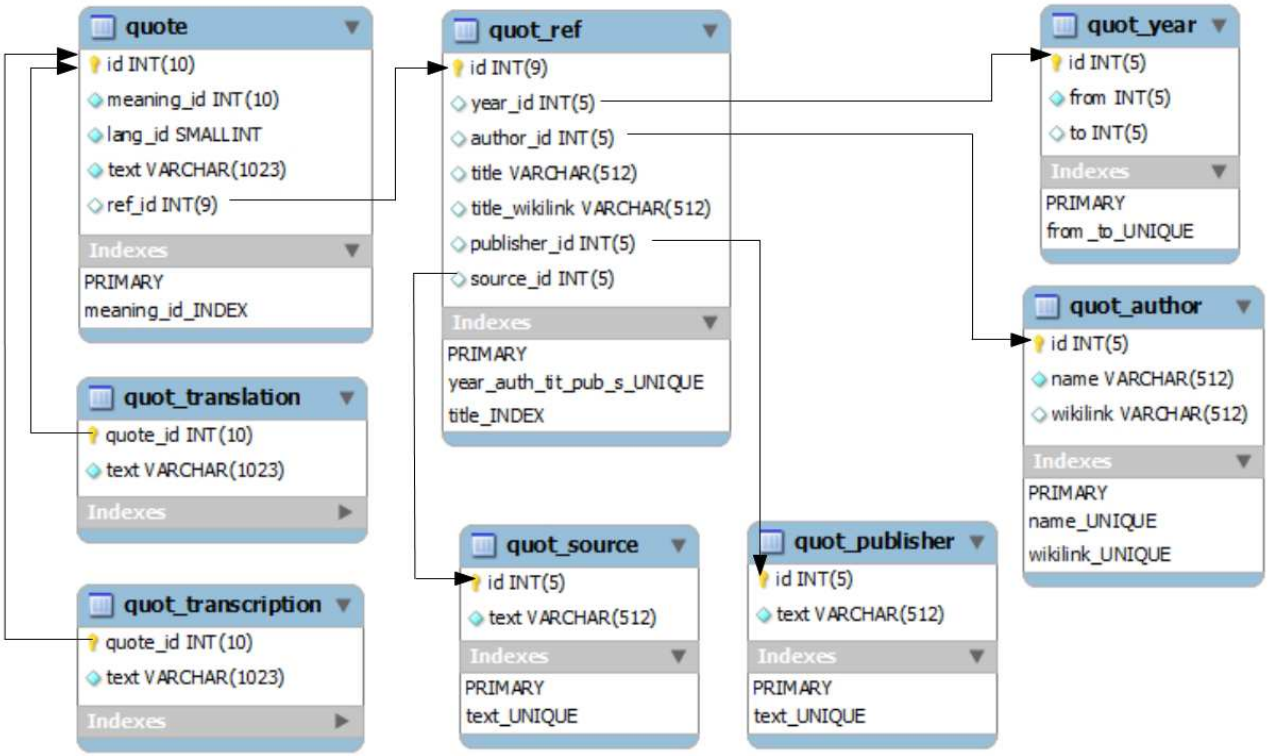

Рис. 1. Таблицы и отношения в базе данных корпуса цитат машиночитаемого словаря

Эксперименты. В предыдущей работе [Крижановский 2011б] в ходе экспериментов были проанализированы данные имён авторов цитат. В этой работе эксперименты связаны с датами произведений. В таблицу с годами (*quot_year*) сохраняются два года – начала и окончания создания литературного произведения. Если произведение было создано в течение одного года, то оба поля содержат одинаковое число. Число уникальных пар (год начала – год окончания) составило 862.

Для вычисления числа цитат по годам выполнялся обход всех цитат и если в цитате указан год создания произведения XXXX, либо диапазон в несколько лет XXXX-YYYY, то число цитат для этих лет увеличивается на единицу. Например, в Русском Викисловаре в словарных статьях «Возрождение» и «танцевать» в цитатах указаны 1927 г. и 1945-1955 гг. соответственно:

- Она бежала мимо парчовых кресел итальянского **Возрождения**, мимо голландских шкафов, мимо большой готической кровати с балдахином на чёрных витых колоннах. *Илья Ильф, Евгений Петров, «Двенадцать стульев», **1927** г.*
- Она никогда не могла предположить, что он так хорошо **танцует**. *Б. Л. Пастернак, «Доктор Живаго», **1945—1955** г.*

Обход 26596 цитат, в которых указана дата, позволил построить следующую гистограмму (рис. 2), показывающую зависимость числа цитат в онлайн-словаре от года произведения в диапазоне 1750-2012 гг.

Пик числа цитат в 2000-е г., вероятно, можно объяснить относительно большим количеством газет и журналов, доступных в Национальном корпусе русского языка (НКРЯ) за этот период. В данном случае следует ориентироваться на корпус НКРЯ, т.к. подавляющее число цитат Русского Викисловаря взято из него [Крижановский 2011б].

Для понимания относительно большого числа цитат на рис. 2 в 1830-е – 1880-е попробуем проанализировать вклад самых часто цитируемых в Викисловаре писателей.

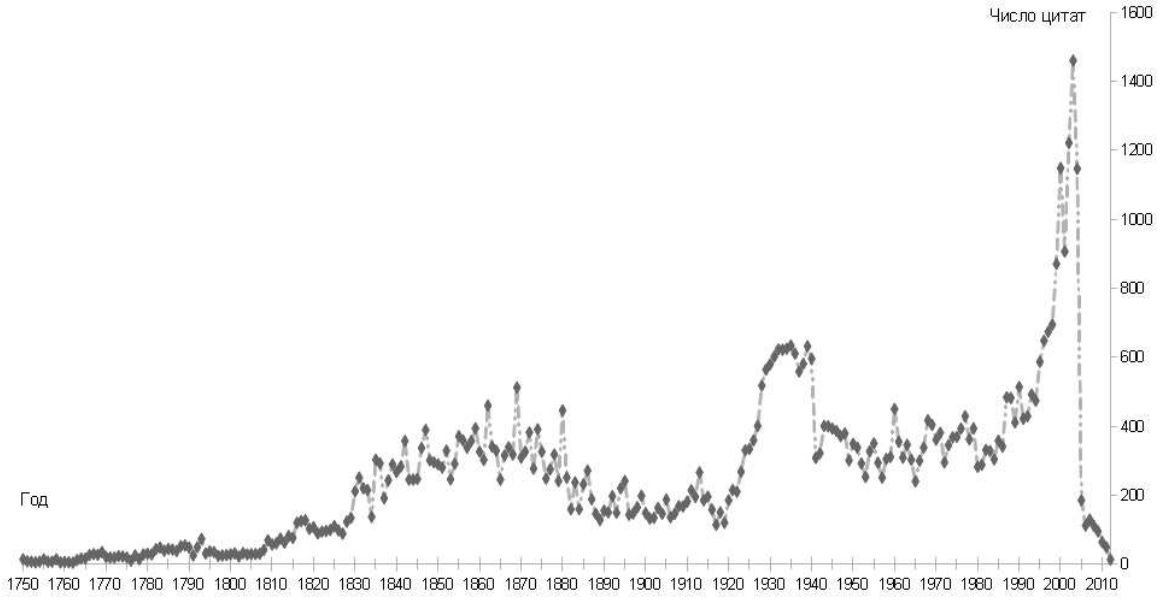

Рис. 2. Зависимость числа цитат в онлайн-словаре от года произведения

В табл. 1 в колонке «Автор» приведено имя писателя с максимальным числом цитат в Викисловаре. Вторая и третья колонки показывают стремительный рост размера словаря по числу цитат для этих писателей за 2011 и 2012 года.

Основной источник цитат для Викисловаря – НКРЯ, отсюда столбец «Публикации в НКРЯ» (годы первой и последней публикации данного автора, существующей в корпусе). И для этого же периода подсчитано общее число цитат в Викисловаре. Последний столбец показывает процентное отношение числа цитат автора (третий столбец) к числу всех цитат в Викисловаре за тот период, за который публикации автора доступны в НКРЯ (предпоследний столбец).

Суммарное число цитат семи первых писателей за 1815-1910 гг. равно 22429, что составляет 20.5%, т.е. одну пятую от всех цитат Викисловаря за этот период. Вероятно, высокая цитируемость этих писателей и объясняет пик на рис. 2 в 1830-е – 1880-е годы.

Таблица 1. Самые популярные авторы цитат в Викисловаре

№ о	**Автор**	**Число цитат (2011)**	**Число цитат (2012)**	**Публика-ции в НКРЯ**	**Всего цитат в Викисловаре (за этот период)**	**% (2012)**
1	А.П. Чехов	716	931	**1880**-1904	4704	19,8%
2	Л.Н. Толстой	529	710	1852-**1910**	14954	4,8%
3	А.С. Пушкин	520	627	1815-1836	3217	19,5%
4	Ф.М. Достоевский	500	776	1846-1881	11853	6,6%
5	И.С. Тургенев	457	697	1846-1882	12012	5,8%
6	Н.В. Гоголь	321	473	1831-1847	4511	10,5%
7	Н.С. Лесков	245	386	1862-1894	9039	4,3%
8	М.А. Булгаков	207	267	1920-1940	10049	2,7%
9	Братья Стругацкие	171	225	1964-1979	5699	4,0%
10	В.П. Астафьев	142	199	1967-2001	16327	1,2%

Рис. 3. Гистограмма числа цитат в онлайн-словаре по годам
и годы творческой активности наиболее цитируемых в Русском Викисловаре писателей

Остаётся открытым вопрос – за счёт каких авторов получился такой пик на рис. 2, длящийся с середины 1920-х до 1940 года?

Распределение по векам. В ходе экспериментов было получено распределение числа цитат в словарных статьях по векам – с 17 по 21, где под 21 веком имеются в виду года с 2000 по 2012 включительно (рис. 4).

Можно отметить, что каждый следующий век представлен в цитатах более полно чем предыдущий. Вероятно, эта тенденция сохранится, т.к. первые 12 лет этого века уже дают 10% от всего числа цитат в словаре.

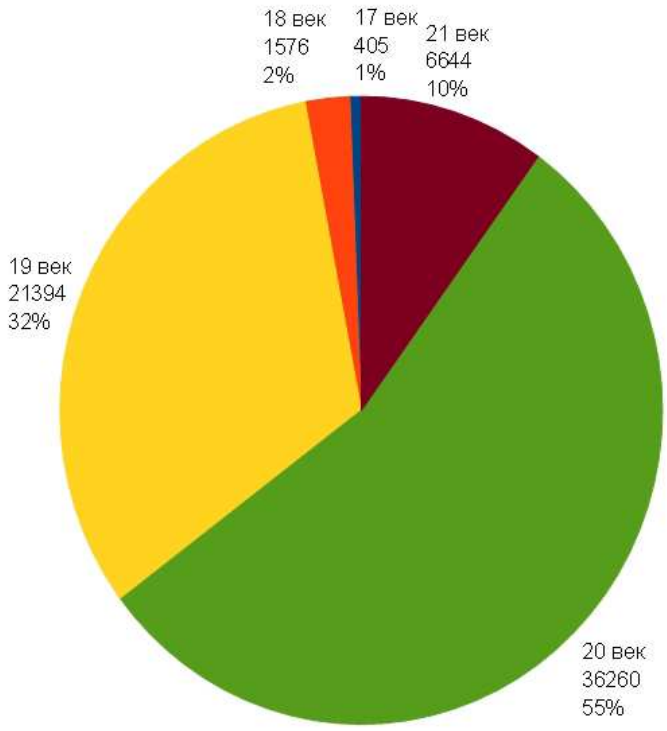

Рис. 4. Распределение числа цитат в словарных статьях по векам с 17 до 21 вв.
(21 век соответствует 2000-2012 гг.)

Объём корпуса цитат. С помощью программы, извлекающей данные из Русского Викисловаря [Крижановский 2010], по данным Викисловаря за 25.03.2012 был построен корпус из 62 тысяч цитат (в 2011 г. было 51.5 тысяч). При этом 52 тысячи цитат (84 % от всего числа цитат) иллюстрируют русские слова (в 2011 было 82 %).

В Русском Викисловаре для 23.8 тысяч цитат (38.35 % от всех цитат) был указан источник, из которого получена данная цитата (в 2011 г. было 17 тысяч цитат с источниками, т.е. 33 %). Главным источником является Национальный корпус русского языка, на который ссылается 94.15 % цитат с источниками.

Заключение. В представленной работе спроектирована архитектура базы данных корпуса цитат. Построена гистограмма распределения числа цитат в онлайн-словаре по годам создания произведений. Выполнена попытка объяснить характер гистограммы, увязав её особенности с годами творчества наиболее популярных в Викисловаре писателей.

Благодарности

Работа выполнена при финансовой поддержке РФФИ (проект № 11-01-00251, № 12-01-00481, № 12-07-00070), поддержке 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». Спасибо Николаю Тесля за плодотворное обсуждение экспериментальной части работы.

Литература

Гришина и др. 2005 – *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

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

Крижановский 2011а – *Крижановский, А. А.* Количественный анализ лексики английского языка в викисловарях и Wordnet // Труды СПИИРАН. 2011. Вып. 19. – С. 87–101. http://scipeople.com/publication/106012/

Крижановский 2011б – *Крижановский, А. А.* Оценка использования корпусов и электронных библиотек в Русском Викисловаре // Труды международной конференции «Корпусная лингвистика–2011». – СПб.: С.-Петербургский гос. университет, Филологический факультет, 2011. – С. 217—222. http://scipeople.com/publication/102432/

Леденёва 2008 – *Леденёва, В.В*. Лексикография современного русского языка. Практикум: Учеб. пособие. - М.: Высшая школа, 2008. – 648 С.

Meyer 2012 – *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. (в печати). http://www.informatik.tu-darmstadt.de/fileadmin/user_upload/Group_UKP/publikationen/2011/oup-elex2012-meyer-wiktionary.pdf