Analysis of the Quotation Corpus of the Russian Wiktionary

Alexander Smirnov[1], Tatiana Levashova[1], Alexey Karpov[1,2], Irina Kipyatkova[1,2], Andrey Ronzhin[1,2], Andrew Krizhanovsky[3], and Nataly Krizhanovsky[3]

[1] St.Petersburg Institute for Informatics and Automation of the Russian Academy of Sciences, Russia,

[2] Saint-Petersburg State University, Department of Phonetics, Russia

[3] Institute of Applied Mathematical Research of the Karelian Research Centre of the Russian Academy of Sciences, Russia

{smir, tatiana.levashova}@iias.spb.su,
{karpov, kipyatkova}@iias.spb.su
andrew.krizhanovsky@gmail.com, nataly@krc.karelia.ru

Abstract. The quantitative evaluation of quotations in the Russian Wiktionary was performed using the developed Wiktionary parser. It was found that the number of quotations in the dictionary is growing fast (51.5 thousands in 2011, 62 thousands in 2012). These quotations were extracted and saved in the relational database of a machine-readable dictionary. For this database, tables related to the quotations were designed. A histogram of distribution of quotations of literary works written in different years was built. It was made an attempt to explain the characteristics of the histogram by associating it with the years of the most popular and cited (in the Russian Wiktionary) writers of the nineteenth century. It was found that more than one-third of all the quotations (the example sentences) contained in the Russian Wiktionary are taken by the editors of a Wiktionary entry from the Russian National Corpus.

Keywords. Wiktionary, corpus, quotations, literary works.

1 Introduction

The progress of computer technologies provides a basis for a new type of dictionaries. This type is an online dictionary, where any interested person can take part in the dictionary development. On the one hand, this way of organizing collective work provides obvious advantages (high intensities of the work, the possibility of online discussion and correction of the articles at any stage of the work); on the other hand, there is a high possibility that some gaps can be presented in the source material and some gaps can be found in the dictionary itself.

One of the possible solutions to the problem of bridging the gaps is to develop a special software tool that can analyze the online dictionary at any stage of the development. Some possible solutions to this problem will be presented in this paper

based on an analysis of the quotations of some literary works contained in the Russian Wiktionary that is a good example of an online dictionary.

The research presented in this paper aims at two purposes: (1) construction of a quotation corpus from the online dictionary, (2) an analysis of the chronological distribution of the quotation corpus within 1750–2012 years. The choice of this period is caused by the fact that the period includes years with more than 10 quotations which refer to this year in the dictionary.

The Wiktionary is a multilingual and multifunctional computer dictionary which combines thesaurus, lexicon and phraseological dictionary. The Wiktionary combines a glossary and explanatory, grammatical, etymological, and translation dictionaries. The Wiktionary contains not only concepts' definitions, semantically related concepts (synonyms, hypernyms, etc.), and multilingual translations, but also the pronunciations (phonetic transcriptions, audio files), hyphenations, etymologies, quotations, parallel texts (quotations with translations), and illustrations (to illustrate meaning of the words).

The Wiktionary data are used:

- In translation:
  - *Dictionary-based* machine translation implemented in the Pandictionary and the Panlingual Translator system [14]. Pandictionary is a sense-distinguished translation dictionary compiled from Wiktionaries and more than 600 machine-readable bilingual dictionaries. Panlingual Translator system uses a lemmatic encoding of the original text as a form of human-aid translation;
  - Translation of the taxonomically organized labels in web-based multilingual resources (the folktale domain); translation is based on multilingual lexical entries and semantic categories of English, German and Hungarian language versions of Wiktionary [2];
  - *Machine translation* between Dutch and Afrikaans [10].
- In the text parsing system NULEX, where some Wiktionary data (verb tense) were integrated with WordNet and VerbNet [8];
- In a *speech recognition and speech synthesis* where the Wiktionary is a basis for the rapid pronunciation dictionary creation [12], [13];
- In *ontology matching* [7];
- For extraction of semantic relations [11];
- For knowledge base construction, e.g. Concept Net[1].

The paper [2] discussed several potential shortcomings of the Wiktionary. The Wiktionary may lack basic information or be of poor quality as the result of collaborative work performed by volunteers. The Wiktionary is formatted in a lightweight mark-up language and wiki format is often applied in an inconsistent manner within one dictionary or across different language versions of Wiktionary, which makes the extraction of structured lexical information a challenging task [3].

The advantages of the Wiktionary for the present research are a huge volume of data and a wide variety of the lexicographical material. An analysis of the German and English language editions of Wiktionary [6], [9] has shown that the sizes of these

[1] See http://conceptnet5.media.mit.edu.

dictionaries are close to or exceeds thesauri in corresponding languages. For example, the sizes of the German Wiktionary and the German thesauri *GermaNet* and *OpenThesaurus* are comparable, whereas the size of the English Wiktionary exceeds the size of the English thesaurus *WordNet*. Any freely available dictionaries in Russian (in the public domain) have not been found. It can be suggested, that the size of Russian Wiktionary is enough for the purposes of the present research.

A word's definition in the Wiktionary is accompanied by quotations, which illustrate the meaning by the surrounding context. The quotations can include references to a source (book, newspaper, blog) with the date of its publication or writing.

The analysis of the dates of literary works, which are used as a source of quotations, is the goal of the paper. The experiments were carried out on the corpus of quotations build on the basis of the Russian Wiktionary. The database of quotation corpus is a part of the machine-readable Wiktionary, which is an open-source project.[2]

2 Framework of Machine-Readable Wiktionary

The conception of the machine-readable Wiktionary is flexible in relation to input data, but it is strict and formal to output data.

Input data. This conception suggests that different wiktionaries can have different article structures (e.g. different names of the article sections and different order of sections), which must be taken into account by a wiktionary-parser. Moreover, even within one Wiktionary, the structure of an article can change with time as new sections appear and templates vary and change. Therefore there is need for a flexible and modular framework in order to parse so much "live" and various wiktionaries (Fig. 1). The specific properties of different wiktionaries are taken into account in the submodules "*ruwikt*" and "*enwikt*" in the module "*Data extraction*" in Fig. 1.

Output data. The data extracted from a Wiktionary are stored in the database of the machine-readable dictionary. The result databases filled by the parser have identical structure independent on the source wiktionary. This ensures compatibility of different machine-readable wiktionaries with external applications.

The framework can be extended with new wiktionaries since many parts of the parser have been already developed and do not depend on a specific wiktionary. These parts are as follows:

- Common application programming interface (API) to the source databases (*input data)*.
- "Common part" of the module "Data extraction" (Fig. 1). It contains (1) language codes in accordance with the international standard for language codes ISO 639, (2) names of languages in English and in Russian; now the parser recognizes 370 languages and codes in the English Wiktionary and 274 in the Russian Wiktionary.

[2] The machine-readable Wiktionary is available at http://code.google.com/p/wikokit/.

– Common API to the result databases of the machine-readable wiktionaries (*output data*).

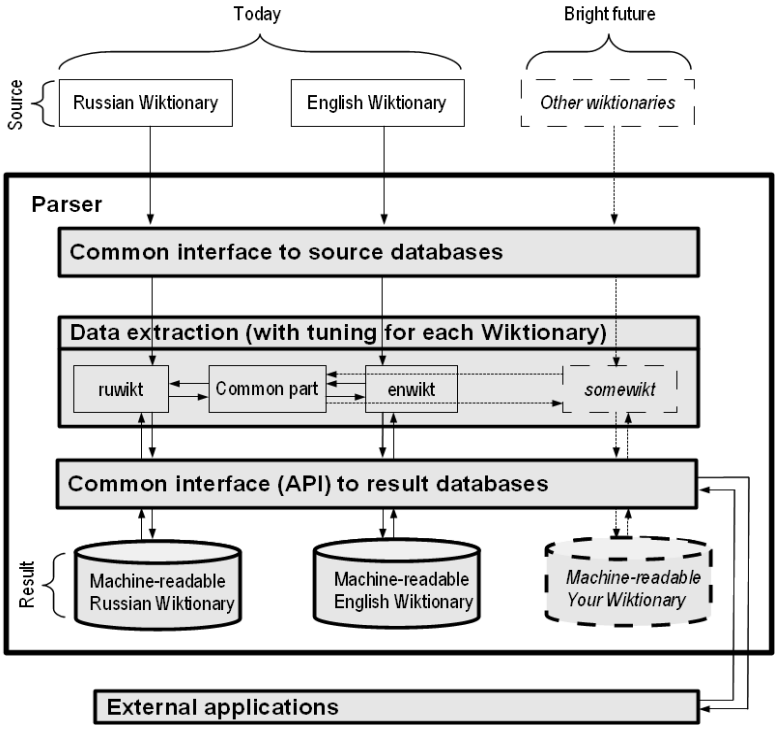

Fig. 1. Machine-readable Wiktionary: the framework.

3 Architecture of the Database of Quotation Corpus

The database of the quotation corpus is a part of the relational database of the machine-readable Wiktionary presented in the paper [5].

The following fields of the quotation template are recognized and added to the database during the extraction of semistructured data from the Wiktionary by the parser (Fig. 2):

- The text of the quotation (stored in the field *text* of the table *quote*).
- The translation into Russian (the table *quot_translation*).
- The transcription of the quotation (the table *quot_transcription* is reserved for the English Wiktionary, it is not used in the Russian Wiktionary).
- Information about a quotation reference is collected in the table *quot_ref*. This table comprises the following fields:
  - Title of the source (the field *title* of the table *quot_ref*).
  - Author of the source (the table *quot_author*).
  - Publisher (the table *quot_publisher*).

- Publication date (the table *quot_year*).
- Name of the resource or corpus, where the quotation is taken from (the table *quot_source*).

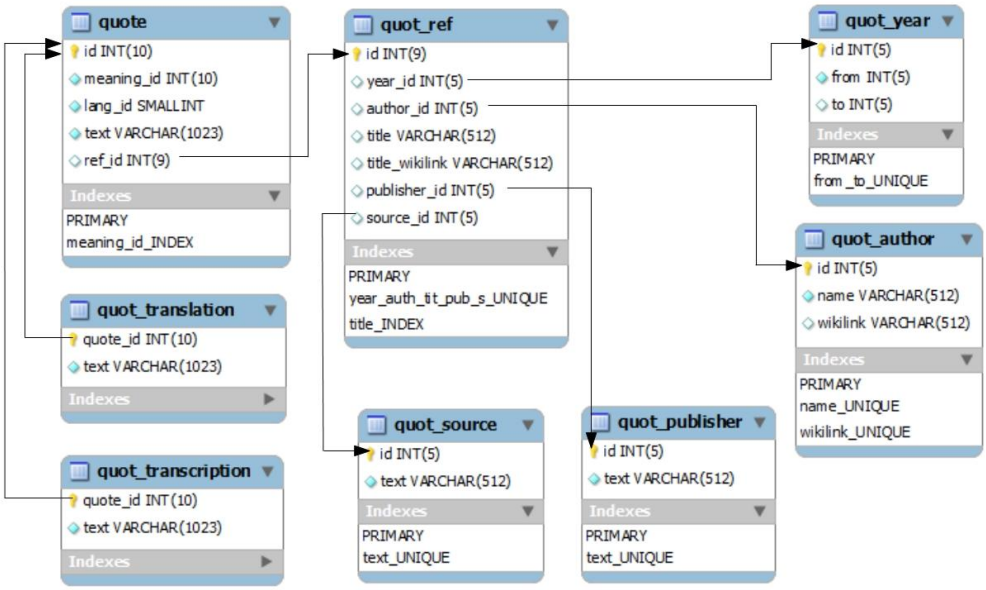

Fig. 2. Tables and relations relevant to quotations in the database of the machine-readable Wiktionary.

4 Database Queries

Various SQL-queries to the database are supported (Fig. 2). For example, using a few queries, one can solve the following search task: *get a list of quotations in English, which refer to books written during more than one year*.[3]

This task is solved step-by-step:

1) Get a list of quotations in English. As of March 2012, there are 1355 quotes in English in the Russian Wiktionary.
2) Get a sublist of quotations with non-empty reference (a source). There are 222 quotations where "*ref_id"* is not NULL in the table "*quote"*.
3) Get a subsublist of quotations, which contain a date in the reference. There are 123 quotations with years.

[3] See these queries: http://code.google.com/p/wikokit/wiki/MRDQuote#SQL_queries

4) Get a list of quotations, which contain a range of years in the reference. I.e., the value of the field "*to"* is greater than "*from*" in the table *quot_year* (Fig. **2**). As the result, seven quotes (Table 1) were found.

The column "*entry*" in Table 1 contains the headword of a Wiktionary article. The quotation is placed in the row below and the word in question is marked by **bold** font in the quote. If there is a translation of this quote into the Russian then it is presented in the next row below. The author name, the title of the source book and the publication (or writing) date (in years) are given in the columns "*author"*, "*title"*, "*from"*, "*to"*, respectively.

Table 1. English quotations from the Russian Wiktionary, which refer to books written during more than one year.

N	Entry	Author	Title	From	To
1	Moscow[4]	Andrei Platonov	The Ethereal Tract	1926	1927
	Moscow awakened and screamed with trams. … The summer sun rejoiced over the full-blooded land, and two men appeared before the gaze of a new **Moscow** — a wonderful city of powerful culture, stubborn labor and intelligent happiness.				
	Москва проснулась и завизжала трамваями. … Летнее солнце ликовало над полнокровной землёй, и взорам двух людей предстала новая **Москва** — чудесный город могущественной культуры, упрямого труда и умного счастья.				
2	cacophony	H. P. Lovecraft	Herbert West: Reanimator	1921	1922
	Not more unutterable could have been the chaos of hellish sound if the pit itself had opened to release the agony of the damned, for in one inconceivable **cacophony** was centered all the supernal terror and unnatural despair of animate nature.				
3	hoarder	–	[Central News autocue data.] 3623 s-units.	1985	1994
	The picture was owned by antiques **hoarder** Ronnie Summerfield who died three years ago leaving a collection valued at millions of pounds.				
4	hoarder	–	The Economist. 3341 s-units.	1985	1994
	The biggest official gold **hoarder** by far is America, which holds 27,9 % of the world's central-bank gold reserves.				
5	order	Charles Dickens	Oliver Twist	1837	1839
	In pursuance of this determination, little Oliver, to his excessive astonishment, was released from bondage, and **ordered** to put himself into a clean shirt.				
6	order	Charles Dickens	Oliver Twist	1837	1839
	Oliver was **ordered** into instant confinement; and a bill was next morning pasted on the outside of the gate, offering a reward of five pounds to anybody who would take Oliver Twist off the hands of the parish.				
7	practitioner	Charles Dickens	The Posthumous Papers of the Pickwick Club	1836	1837
	These sequestered nooks are the public offices of the legal profession, where writs are issued, judgments signed, declarations filed, and numerous other ingenious machines put in motion for the torture and torment of His Majesty's liege subjects, and the comfort and emolument of the **practitioners** of the law.				

[4] See entry "Moscow" in the Russian Wiktionary: http://ru.wiktionary.org/wiki/Moscow

5 Experiments

5.1 Corpus of Quotations

The corpus of quotations was built on the basis of the Russian edition of Wiktionary as of March 25, 2012. It was constructed by means of the developed Wiktionary parser [5]. The corpus includes 62 thousand quotations (51.5 thousand in 2011). It is important that 52 thousand quotations (84% of the whole number of quotations) are occurred in the explanations for Russian words (82% in 2011).

In the Russian Wiktionary, 23.8 thousand quotations (38.35% of the whole number) have a reference to the source (17 thousand quotations with references in 2011, i.e. 33%). The main source of quotations in the Russian Wiktionary is the *Russian National Corpus* [1]. There are 94.15% quotations (of the whole number of quotations with references) which refer to the Russian National Corpus.

5.2 Publication Date: Analysis and Hypothesis

In this study publication dates indicated in the sources of quotations are under investigation. These dates are stored in the table *quot_year*, discussed in Section 3. The table *quot_year* contains two fields "*from"* and "*to"* of integer type (Fig. 2) indicating the years of source publication. If the work was published or written during one year, then both fields have equal values. The number of unique pairs (start year "*from*", finish year "*to*") is 862 in the Russian Wiktionary (it equals to the number of records in the table *quot_year*).

In order to calculate a number of quotations for each year the algorithm similar to the well-known game "Tetris" was used (Fig. 3). The algorithm traverses all quotations; if a quotation contains a year or a range of years then the number of quotations for years in this range are incremented.

For example, there are years 1926–1927 and 1929 in the following quotations of entries "Moscow" and "medal" in the Russian Wiktionary:

> –**Moscow** awakened and screamed with trams. … The summer sun rejoiced over the full-blooded land, and two men appeared before the gaze of a new **Moscow —** a wonderful city of powerful culture, stubborn labor and intelligent happiness. *Andrei Platonov*. "The Ethereal Tract". **1926-1927**
>
> –This is a very brave young man. He has been proposed for the silver **medal** of valor. *Ernest Miller Hemingway*. "Farewell to Arms". **1929**

These quotations are presented in Fig. 3a in the form of bricks on the abscissa in 1929 and in the range 1926–1927. Let's suppose that during the traversal a quotation from the source written in 1924–1926 has been found. Then the value of the histogram at 1926 in Fig. 3b is two (quotations) and at 1924–1925 is one (quotation).

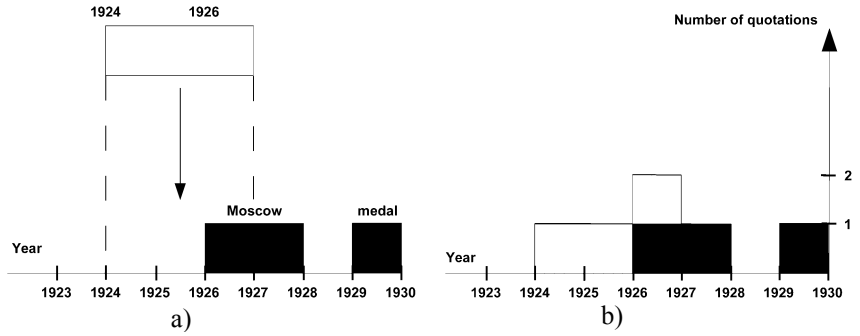

Fig. 3. An idea of calculation of number of quotes for each year in online dictionary (histogram construction).

The traversal of 26,596 quotations (which contains date) makes possible to build the following histogram (Fig. 4), which relates the number of quotations and the source's publication date in the range 1750-2012.[5]

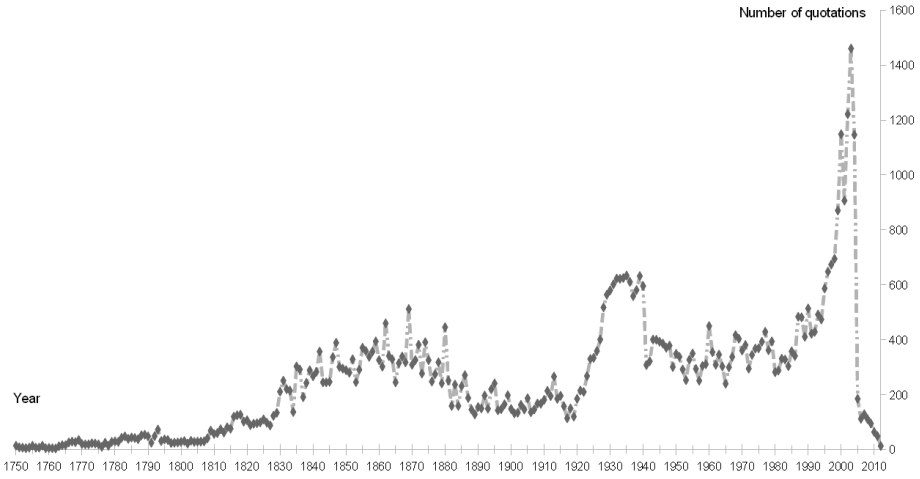

Fig. 4. The dependence of number of quotations with respect to the source's publication date.

The peak number of quotations in the 2000s might be explained by a relatively high number of newspapers and journals available at the Russian National Corpus within this time range since this Corpus is the main source of quotations for the Russian Wiktionary (see above).

[5] The source data for Fig. 4 is available at http://ru.wiktionary.org/Участник:AKA%20MBG/Статистика:Цитаты%20(дата)

In order to understand the relatively high number of quotations in Fig. 5 in the time range from the 1830s to the 1880s, the contribution of the most cited in the Russian Wiktionary writers is analyzed.

The writers with the highest number of quotations in the Russian Wiktionary are listed in the column "Author" in Table 2. The second and third columns demonstrate the fast growing size of the dictionary in a number of quotations for these writers in 2011 and 2012.

The main source of quotations in the Russian Wiktionary is the Russian National Corpus hence there is a column labeled "Publication in Russian National Corpus", which provides the years of the first and last publications of the author presented in the corpus. For the same periods the total numbers of quotations in the Wiktionary were counted (column "Total quotes…"). The last column is the ratio of the number of quotations of the author (third column, 2012) to the total numbers of quotations for the periods, when the publications of the author is presented in the corpus (next to last column).

Table 2. The most popular authors in the Russian Wiktionary.

Author	Number of quotes		Publication in Russian National Corpus	Total quotes in Wiktionary (within this time range)	Contribution (%) 2012
	2011	2012			
Anton Chekhov	716	931	1880-1904	4,704	19,8%
Leo Tolstoy	529	710	1852-**1910**	14,954	4,8%
Alexander Pushkin	520	627	**1815**-1836	3,217	19,5%
Fyodor Dostoyevsky	500	776	1846-1881	11,853	6,6%
Ivan Turgenev	457	697	1846-1882	12,012	5,8%
Nikolai Gogol	321	473	1831-1847	4,511	10,5%
Nikolai Leskov	245	386	1862-1894	9,039	4,3%
Mikhail Bulgakov	207	267	1920-1940	10,049	2,7%
Arkady and Boris Strugatskye	171	225	1964-1979	5,699	4,0%
Viktor Astafyev	142	199	1967-2001	16,327	1,2%

The total number of quotations of the first seven authors in the period 1815-1910 (*Chekhov – Leskov* in Table 2) is 22429, it is 20.5%, i.e. one fifth part of the whole amount of quotations in this period in the Russian Wiktionary. Most probably, the high citations of these writers is the reason of the peak in Fig. 5 in the time range from the 1830s to the 1880s.

The open question remains, what Russian authors contribute to the peak in Fig. 5 in the period 1920s – 1940s before the World War II?

Fig. 5. The dependence of the number of quotations with respect the source's publication date and the years of literary activity of the most cited in the Russian Wiktionary writers.

5.3 Distribution for Centuries

The analysis revealed that the earliest quotations in the Russia Wiktionary are dated:

- **70** BC, Cicero, "Against Verres", Latin, the entry "asylum".
- **1076**, «Изборник Святослава» (Svyatoslav's Miscellanies), Old East Slavic, the entry «воинъ».
- **1364**, Guillaume de Machaut, "Dit de la Marguerite", Old French, the entry "chançon"

In the course of experiments the distribution of quotes from the Russian Wiktionary dating from 17th to 21st century was made (Table 3). The 21st century corresponds to the range 2000-2012, inclusively.

Table 3. The distribution of Wiktionary quotes dating from 17th to 21st century.

Century	Quotes	%
17th	405	1
18th	1 576	2
19th	21 394	32
20th	36 260	55
2000-2012	6 644	10

It could be seen that each subsequent century contains more quotations than the previous one. Probably, this tendency will remain, since the first 12 years of this century already have given 10% of the whole number of quotations in the dictionary.

6 Conclusion

In this paper a framework of the machine-readable Wiktionary was designed, which emphasizes the possibility to add new wiktionaries to the parser modular architecture.

The architecture of the database of quotation corpus was described. An exemplary search task (to get a list of quotations in English, which refer to books written during more than one year) was solved.

The characteristics of corpus of quotations constructed based on the Russian Wiktionary were investigated. It was found that the number of quotations in the dictionary grows fast (51.5 thousands in 2011, 62 thousands in 2012).

The interesting statement is made in the paper [4] that "*the example sentences contained in Wiktionary are often artificially constructed by the authors of a Wiktionary entry and are, thus, not authentic materials taken from actual text corpora*". Now it is possible to estimate the percentage of quotations taken from literary works (at least for the Russian Wiktionary). In the Russian Wiktionary, 23.8 thousand quotations (38.35% of the whole number) have a reference to the source in 2012 (17 thousand quotations with references in 2011, i.e. 33%). The percentage of quotations with references is growing.

The main source of quotations in the Russian Wiktionary is the *Russian National Corpus*. There are 94.15% quotations (of the whole number of quotations with references) which refer to the Russian National Corpus. Thus, more than one-third of all the quotations (36.1%) are authentic materials taken from the actual text corpus.

The following shortcomings and drawbacks of the quotation corpus of the Russian Wiktionary were revealed. There are a few quotations with references to texts dated by 17 and 18 centuries (3% of quotations only). Almost there are no quotations dated before the 17th century (Table 3).

The histogram which relates the number of quotations and the source's publication date in the range 1750–2012 was created. It was made an attempt to explain the characteristics of the histogram by associating it with the years of the most popular and cited (in the Russian Wiktionary) writers of the nineteenth century: Anton Chekhov, Leo Tolstoy, Alexander Pushkin, Fyodor Dostoyevsky, Ivan Turgenev, Nikolai Gogol, and Nikolai Leskov.

Acknowledgments. Some parts of the research were carried out under projects funded by grants # 11-01-00251, # 12-01-00481 and # 12-07-00070 of the Russian Foundation for Basic Research, grant # 12-04-12062 of the Russian Foundation for Humanities and project of the research program "Intelligent information technologies, mathematical modeling, system analysis and automation" of the Russian Academy of Sciences. Some parts of this work were supported by the Ministry of Education and Science of Russian Federation (The Russian Federal Targeted Program "R&D in Priority Fields of S&T Complex of Russia for 2007-2013", Contract No. 07.514.11.4139). The authors are grateful to Nickolay Teslya for his insightful comments.

References

1. Apresjan, Ju., Boguslavsky, I., Iomdin, B., Iomdin, L., Sannikov, A., Sizov, V. A: Syntactically and Semantically Tagged Corpus of Russian: State of the Art and Prospects. In: Proceedings of LREC. Genova, Italy, pp. 1378–1381 (2006)
2. Declerck, T., Morth, K., Lendvai, P.: Accessing and standardizing Wiktionary lexical entries for the translation of labels in Cultural Heritage taxonomies. In: Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12). Istanbul, Turkey (2012)
3. Hellmann, S., Auer, S.: Towards Web-Scale Collaborative Knowledge Extraction. Theory and Applications of Natural Language Processing pp. 1–27. (preprint) (2012)
4. Henrich, V., Hinrichs, E., Suttner, K.: Automatically Linking GermaNet to Wikipedia for Harvesting Corpus Examples for GermaNet Senses. Journal for Language Technology and Computational Linguistics (JLCL), Vol. 27, Number 1, pp. 1–19 (2012)
5. Krizhanovsky, A.A.: Transformation of Wiktionary entry structure into tables and relations in a relational database schema. Preprint. (2010)
6. Krizhanovsky, A.A.: A quantitative analysis of the English lexicon in Wiktionaries and WordNet. Int. J. of Intelligent Information Technologies (IJIIT), accepted. Preprint (2013)
7. Lin, F., Krizhanovsky, A.: Multilingual ontology matching based on Wiktionary data accessible via SPARQL endpoint. In: Proceedings of the 13th Russian Conference on Digital Libraries RCDL'2011. Voronezh, Russia, pp. 19–26 (2011)
8. McFate, C., Forbus, K.: NULEX: An Open-License Broad Coverage Lexicon. In: The 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies. Portland, Oregon, USA. Vol. 2, pp. 363–367 (2011)
9. 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, pp. 259–291 (2012)
10. Otte, P., Tyers, F.M.: Rapid rule-based machine translation between Dutch and Afrikaans. In: 16th Annual Conference of the European Association of Machine Translation, EAMT11 (2011)
11. Panchenko, A., Adeykin, S., Romanov, P., Romanov, A.: Extraction of Semantic Relations between Concepts with KNN Algorithms on Wikipedia. In: Concept Discovery in Unstructured Data Workshop (CDUD) of International Conference On Formal Concept Analysis, Belgium, pp. 78–88 (2012)
12. Qingyue, He: Automatic Pronunciation Dictionary Generation from Wiktionary and Wikipedia. Thesis. Karlsruhe Institute of Technology (2009)
13. Schlippe, T., Ochs, S., Schultz, T.: Wiktionary as a Source for Automatic Pronunciation Extraction. In: Proceedings of the 11th Annual Conference of the International Speech Communication Association, Makuhari, Japan, pp. 2290–2293 (2010)
14. Soderland, S., Lim, C., Mausam, Bo Qin, Etzioni, O., Pool, J.: Lemmatic machine translation. In: Proceedings of Machine Translation Summit XII, Ottawa, Canada (2009)