Retrieval and Analysis of Quotations’ Dates in Corpus of Quotations of Online Dictionary

1 Institute of Applied Mathematical Research of the Karelian Research Centre of the RAS
2 Institute for Linguistic Studies

1 Andrew.Krizhanovsky gmail.com
1 Nataly Lugovaya (Nataly krc.karelia.ru)
2 Vasily Kruglov (VMKruglov yandex.ru)
Contents

- Wiktionary
- Framework of the MRD Wiktionary
- Quotation corpus analysis
- Results
Goal

To estimate the quality of the quotation corpus

First task

To extract the quotes from the Russian Wiktionary.

Second task

To analyse the chronological distribution of the quotation corpus extracted from the online dictionary in the time period 1750-2012 (> 10 quote).
Wiktionary is a multilingual and multifunctional dictionary and thesaurus.

Glossary + Defining, Grammatical, Etymological, and Translation Dictionary.
Wiktionary Applications

**Problems:**
- Word-sense disambiguation
- Ontology matching
- Automatic thesauri construction
- Machine translation (lemmatic translation)

**Software:**
- Text search systems
  - query reformulation based on thesaurus data
- Question answering systems
- Speech recognition and synthesis systems
  - based on the International Phonetic Alphabet from Wiktionary entries
Largest Eight Wiktionary Editions (2003-2010)
First Task

Present:
• To extract the data (+quotes) from the Russian Wiktionary.

Future:
• To extract the data from other Wiktionary.
Architectural diagram showing a system for data extraction from source databases to result databases. The diagram includes:

- **Today** and **Bright future** sections.
- **Source** with options for Russian Wiktionary, English Wiktionary, and Other wiktionaries.
- **Parser**
  - Common interface to source databases
  - Data extraction (with tuning for each Wiktionary)
    - ruwikt
    - Common part
    - enwikt
    - somewikt
- **Common interface (API) to result databases**
  - Machine-readable Russian Wiktionary
  - Machine-readable English Wiktionary
  - Machine-readable Your Wiktionary
- **External applications**
**Quote Example (Source Data )**

<table>
<thead>
<tr>
<th>Entry</th>
<th>Author</th>
<th>Title</th>
<th>From</th>
<th>To</th>
</tr>
</thead>
<tbody>
<tr>
<td>Moscow</td>
<td>Андрей Платонов</td>
<td>Эфирный тракт</td>
<td>1926</td>
<td>1927</td>
</tr>
</tbody>
</table>

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.
Number of Quotes per Year

0. Lets our corpus has three quotes
1. E.g. there are two quotes: 1927 and 1944-1945
2. + the quote with the reference 1950-1957
The dependence of the number of quotations with respect to the source’s publication date
The most popular authors in the Russian Wiktionary

<table>
<thead>
<tr>
<th>Author</th>
<th>Number of quotes (2011)</th>
<th>Number of quotes (2012)</th>
<th>Publication in Russian National Corpus</th>
<th>Total quotes in Wiktionary (within this time range)</th>
<th>Contribution % (2012)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Anton Chekhov</td>
<td>716</td>
<td>931</td>
<td>1880-1904</td>
<td>4704</td>
<td>19,8%</td>
</tr>
<tr>
<td>Leo Tolstoy</td>
<td>529</td>
<td>710</td>
<td>1852-1910</td>
<td>14954</td>
<td>4,8%</td>
</tr>
<tr>
<td>Alexander Pushkin</td>
<td>520</td>
<td>627</td>
<td>1815-1836</td>
<td>3217</td>
<td>19,5%</td>
</tr>
<tr>
<td>Fyodor Dostoyevsky</td>
<td>500</td>
<td>776</td>
<td>1846-1881</td>
<td>11853</td>
<td>6,6%</td>
</tr>
<tr>
<td>Ivan Turgenev</td>
<td>457</td>
<td>697</td>
<td>1846-1882</td>
<td>12012</td>
<td>5,8%</td>
</tr>
<tr>
<td>Nikolai Gogol</td>
<td>321</td>
<td>473</td>
<td>1831-1847</td>
<td>4511</td>
<td>10,5%</td>
</tr>
<tr>
<td>Nikolai Leskov</td>
<td>245</td>
<td>386</td>
<td>1862-1894</td>
<td>9039</td>
<td>4,3%</td>
</tr>
<tr>
<td>Mikhail Bulgakov</td>
<td>207</td>
<td>267</td>
<td>1920-1940</td>
<td>10049</td>
<td>2,7%</td>
</tr>
<tr>
<td>A. and B. Strugatskye</td>
<td>171</td>
<td>225</td>
<td>1964-1979</td>
<td>5699</td>
<td>4,0%</td>
</tr>
<tr>
<td>Viktor Astafyev</td>
<td>142</td>
<td>199</td>
<td>1967-2001</td>
<td>16327</td>
<td>1,2%</td>
</tr>
</tbody>
</table>
The dependence of the number of quotations with respect to the source’s publication date and the years of literary activity of the most cited in the Russian Wiktionary writers.
The distribution of Wiktionary quotes (17-21 c.)

- 20th century: 36,260 (55%)
- 19th century: 21,394 (32%)
- 18th century: 1,576 (2%)
- 17th century: 405 (1%)
- 21st century: 6,644 (10%)
Results (1)

- Corpus of quotations (from the Russian Wiktionary) analysed:
  - number of quotations in the dictionary grows fast (51.5K in 2011, 62K in 2012);
  - most popular writers found;
  - histogram (the number of quotations & date) created and bound with years of the most popular (in the Russian Wiktionary) writers.
Results (2)

- Framework and architecture of the machine-readable Wiktionary were designed.

- Mobile application (Android) developed (multilingual offline dictionary)
  - kiwidict-ru (+ quotations)

- Project site
  - http://code.google.com/p/wikokit/