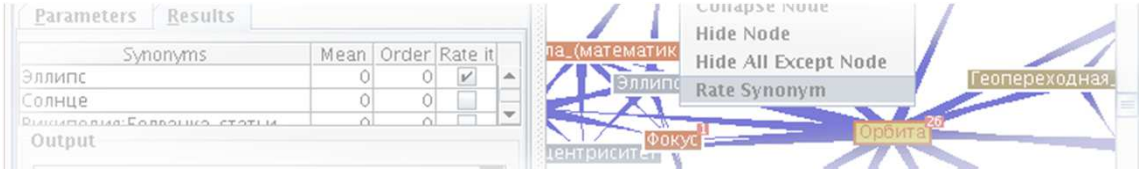


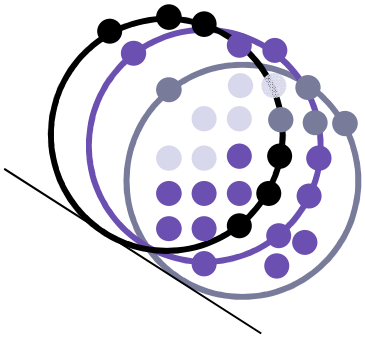

# Retrieval and Analysis of Quotations' Dates in Corpus of Quotations of Online Dictionary

[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

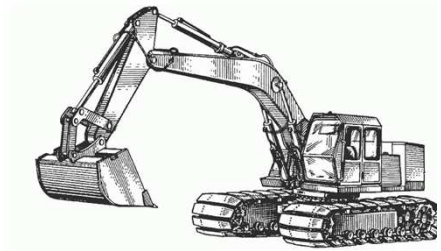
## Goal

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.

---

**100 000+**

Ελληνικά • English • Français • Ido • Русский • தமிழ் • Türkçe • Tiếng Việt • 中文

**10 000+**

Afrikaans • العربية • Български • Brezhoneg • Deutsch • Eesti • Español • فارسی • Galego • 한국어 / 조선어 • Bahasa Indonesia • Íslenska • Italiano • Kurdî / كوردی • Lietuvių • Limburgs • Magyar • 日本語 • Nederlands • Polski • Português • Română • Sicilianu • Српски / Srpski • Suomi • Svenska • తెలుగు • Volapük

**1000+**

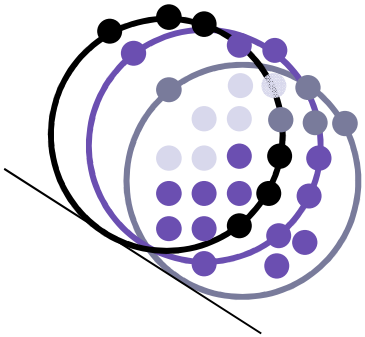
Asturianu • Bân-lâm-gú / Hō-ló-oē • Català • Corsu • Česky • Dansk • Englisc • Esperanto • Frysk • Gaeilge • Հայերեն • हिन्दी • Hornjoserbsce • Hrvatski • Interlingua • עברית • Kalaallisut • Kaszëbsczi • ພາສາລາວ • Latina • മലയാളം • Bahasa Melayu • Norsk (bokmål) • Occitan • Қазақша • Sesotho • Shqip • Simple English • Slovenčina • Slovenščina • Kiswahili • Tatarça / татарча • ไทย • Українська • اردو

**100+**

አማርኛ • Aragonés • Avañe'ẽ • Azərbaycan • Беларуская • Bosanski • Cymraeg • Euskara • Føroyskt • Gàidhlig • ગુજરાતી • Interlingue • ភាសាខ្មែរ • ಕನ್ನಡ • Kinyarwanda • Кыргызча • Latviešu • Македонски • मराठी • монгол • Nāhuatlahtōlli • ਪੰਜਾਬੀ • Plattdüütsch • Runa Simi • سنڌي • Basa Sunda • Tagalog • ትግርኛ • GWY • Xitsonga • ئۇيغۇرچە • Wolof • יידיש • isiZulu

**Other languages**

# Wiktionary Applications

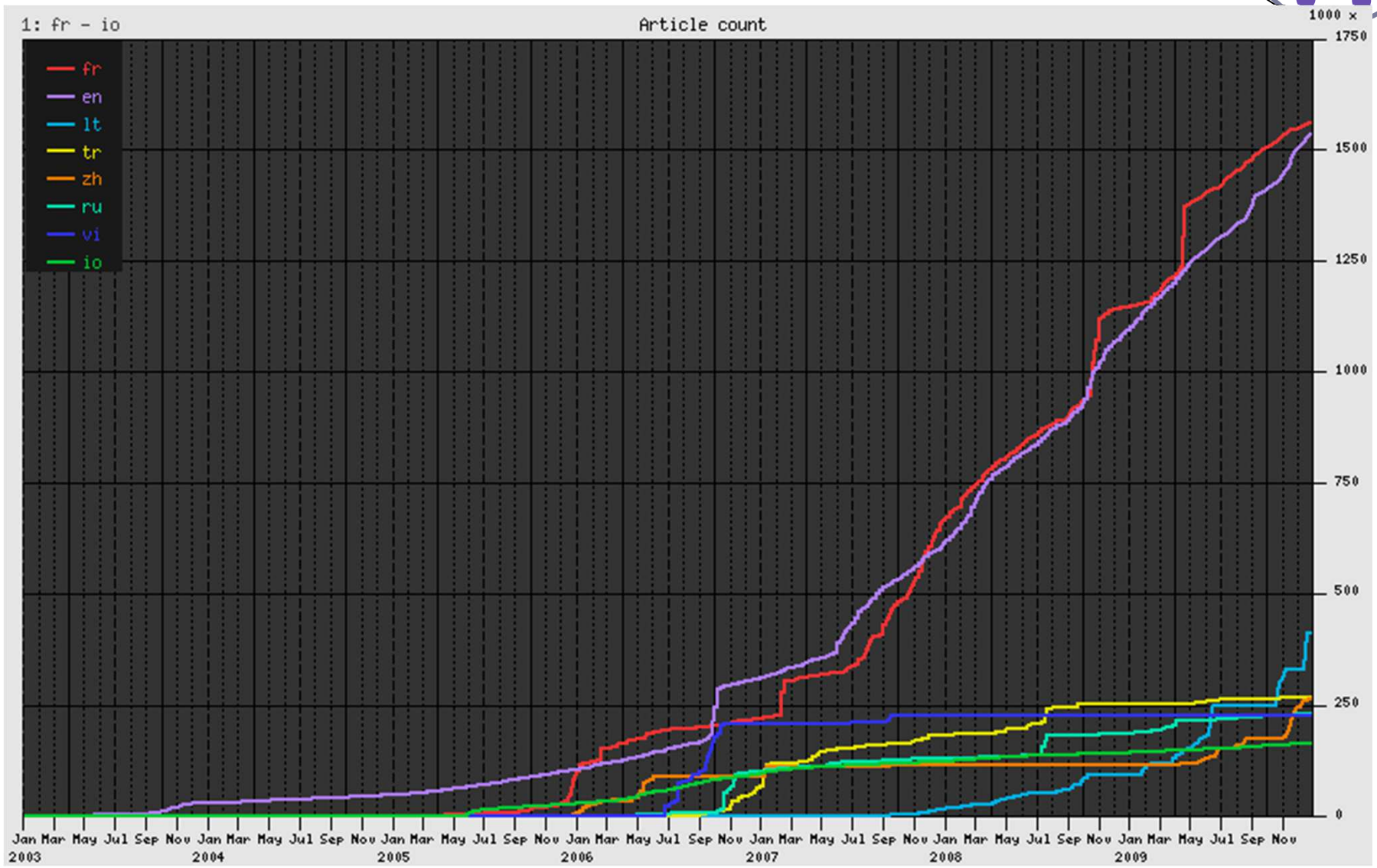


- Problems:
  - Word-sense disambiguation
  - Ontology matching
  - Automatic thesauri construction
  - Machine translation (lemmatic translation)
- Software:
  - Text search systems

    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

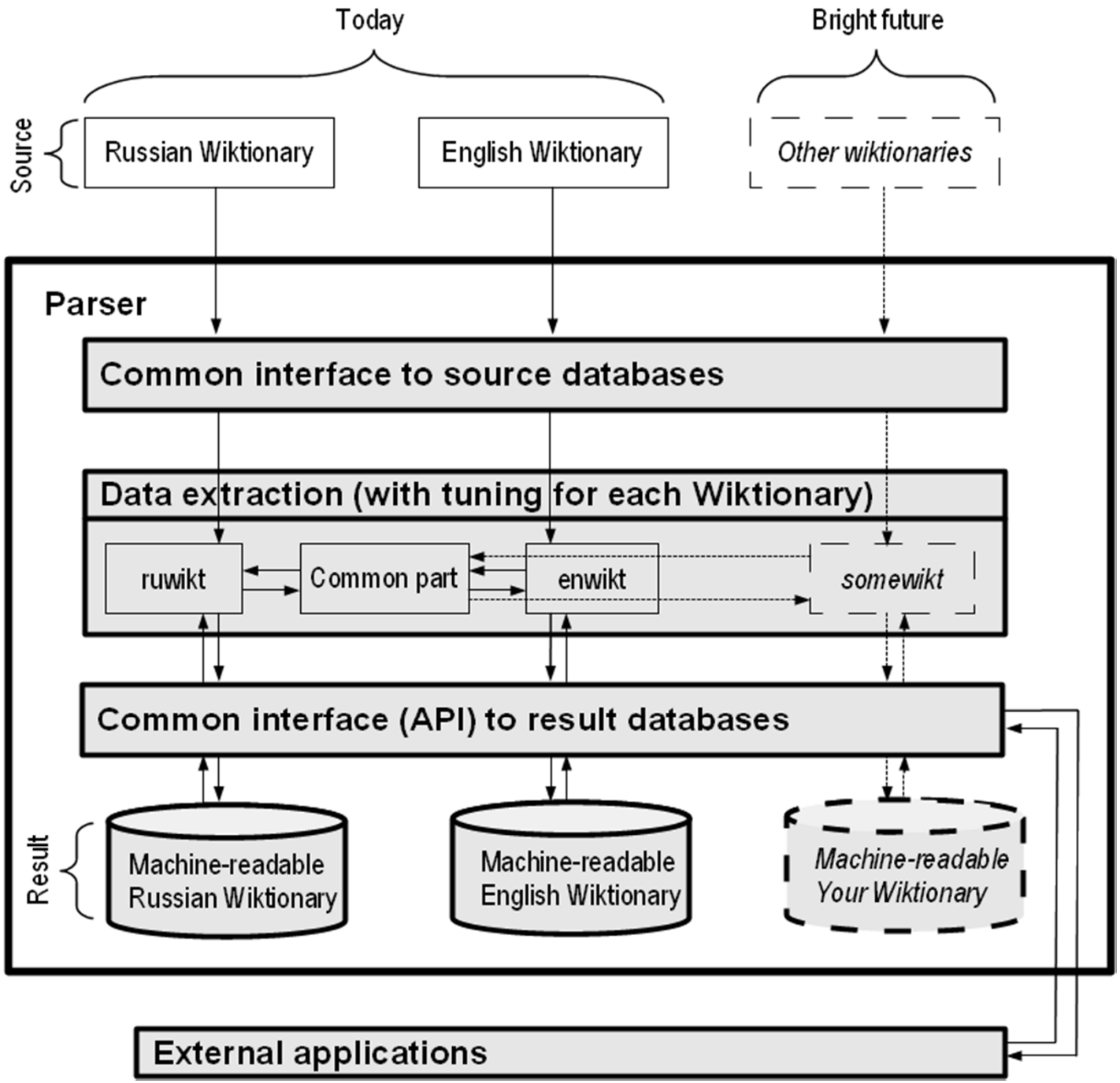

*Present*:

- To extract the data (+quotes) from the Russian Wiktionary.

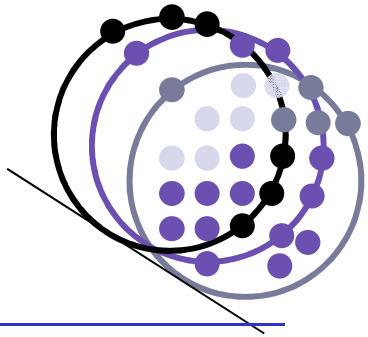
*Future*:

- To extract the data from other Wiktionary.

# Architecture



# Quote Example (Source Data )



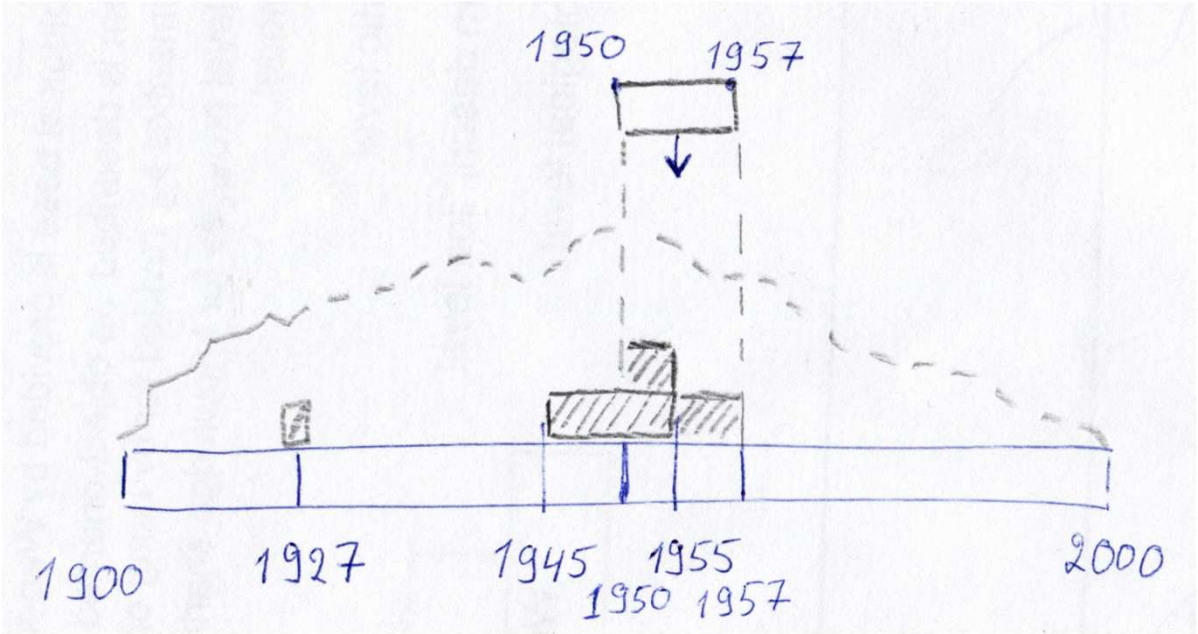
| Entry | Author | Title | From | To |
|---|---|---|---|---|
| **Moscow** | Андрей Платонов | Эфирный тракт | 1926 | 1927 |

**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.**

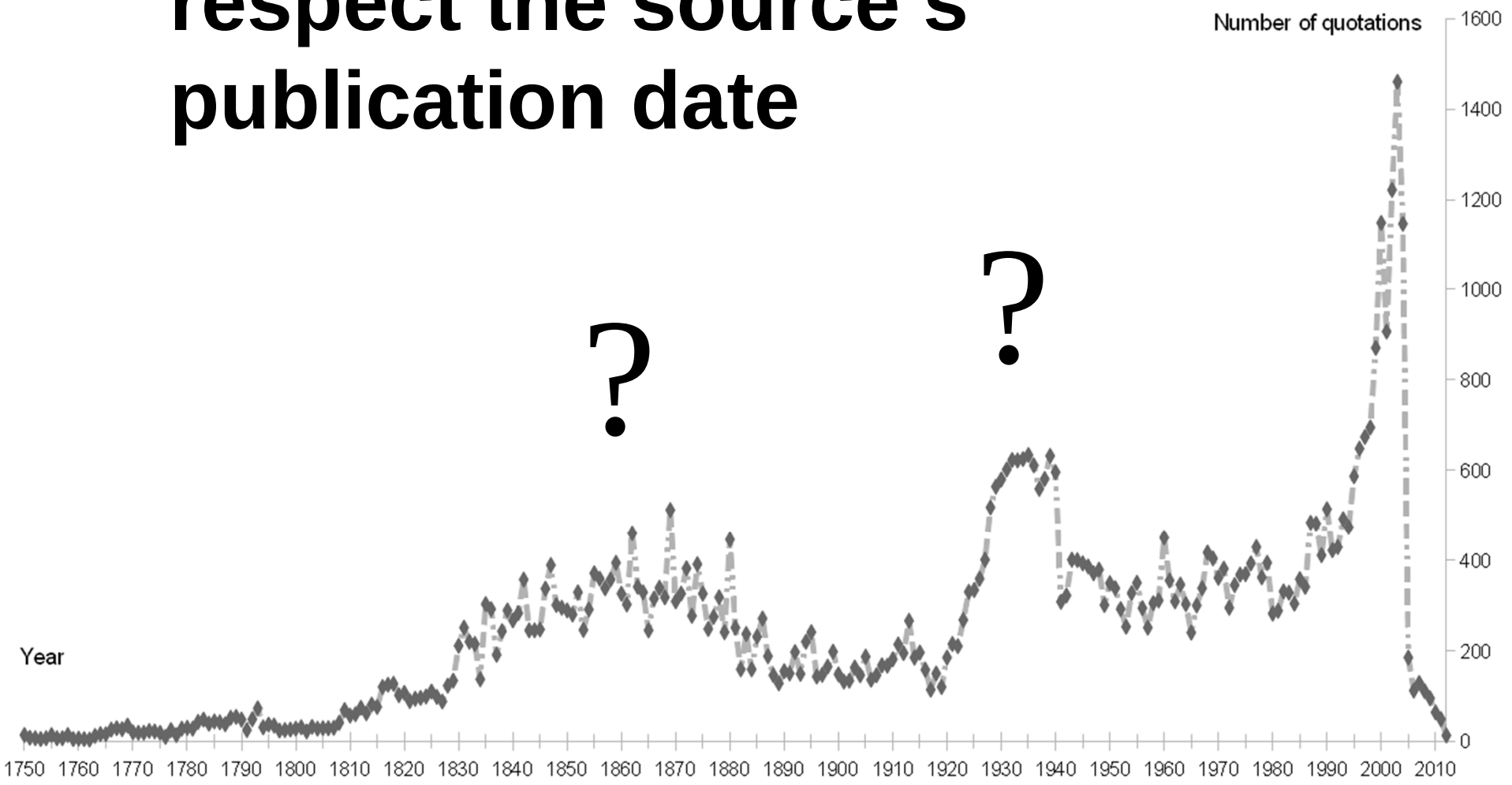
**Москва проснулась и завизжала трамваями. … Летнее солнце ликовало над полнокровной землёй, и взорам двух людей предстала новая Москва — чудесный город могущественной культуры, упрямого труда и умного счастья.**

# Number of Quotes per Year

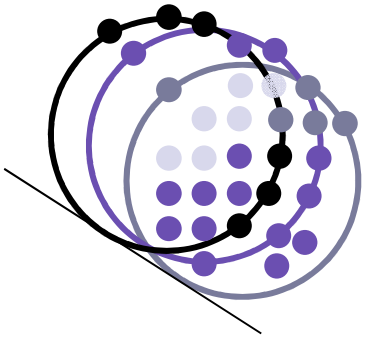
0. Lets our corpus has three quotes
1. E.g. there are two quotes: 1927 and 1944-1945
2. + the quote with the reference 1950-1957



# The dependence of the number of quotations with respect the source's publication date



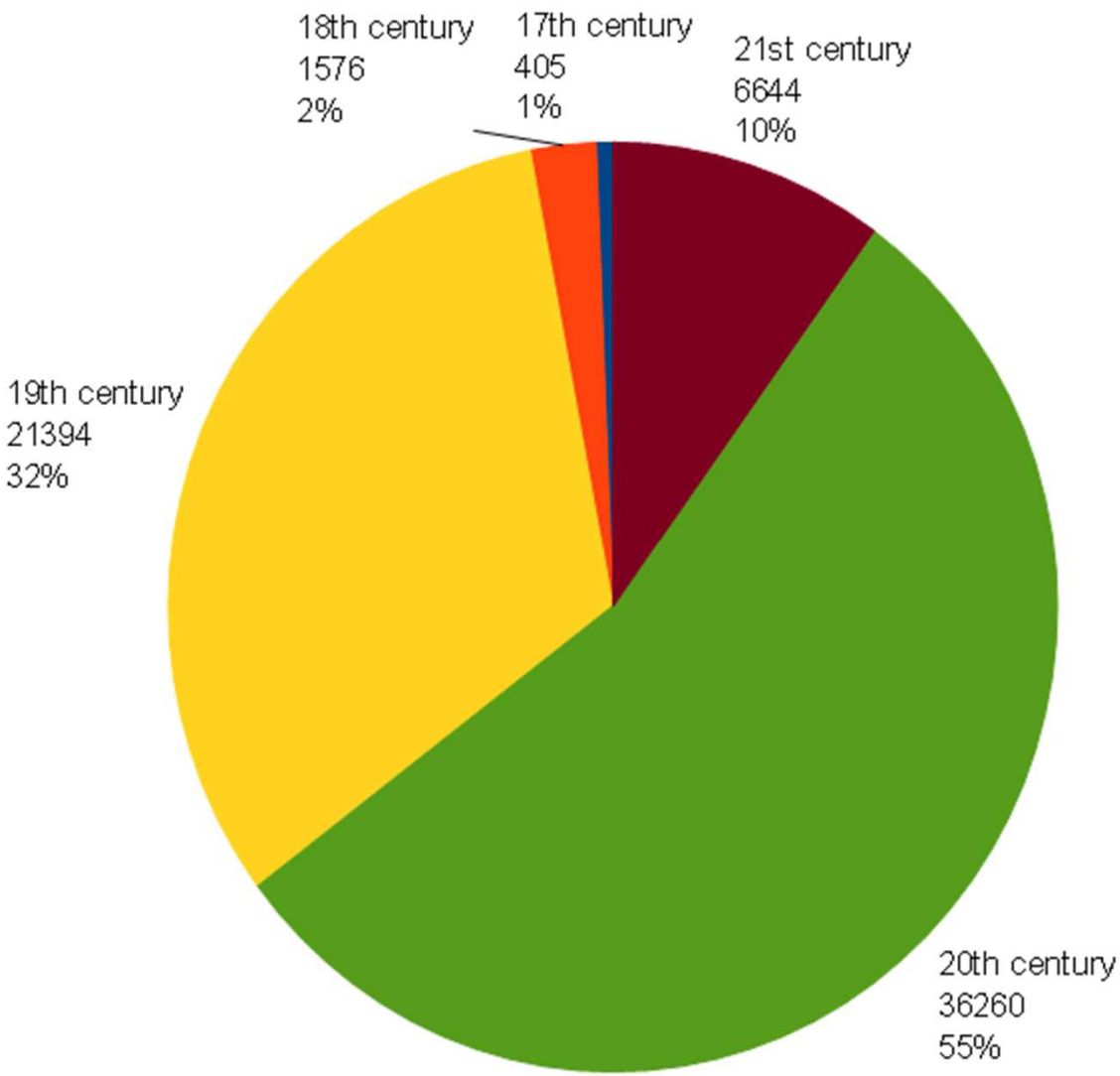
# The most popular authors in the Russian Wiktionary



| Author | Number of quotes (2011) | Number of quotes (2012) | Publication in Russian National Corpus | Total quotes in Wiktionary (within this time range) | Contri-bution % (2012) |
|---|---|---|---|---|---|
| Anton Chekhov | 716 | 931 | **1880**-1904 | 4704 | 19,8% |
| Leo Tolstoy | 529 | 710 | 1852-**1910** | 14954 | 4,8% |
| Alexander Pushkin | 520 | 627 | 1815-1836 | 3217 | 19,5% |
| Fyodor Dostoyevsky | 500 | 776 | 1846-1881 | 11853 | 6,6% |
| Ivan Turgenev | 457 | 697 | 1846-1882 | 12012 | 5,8% |
| Nikolai Gogol | 321 | 473 | 1831-1847 | 4511 | 10,5% |
| Nikolai Leskov | 245 | 386 | 1862-1894 | 9039 | 4,3% |
| Mikhail Bulgakov | 207 | 267 | 1920-1940 | 10049 | 2,7% |
| A. and B. Strugatskye | 171 | 225 | 1964-1979 | 5699 | 4,0% |
| Viktor Astafyev | 142 | 199 | 1967-2001 | 16327 | 1,2% |

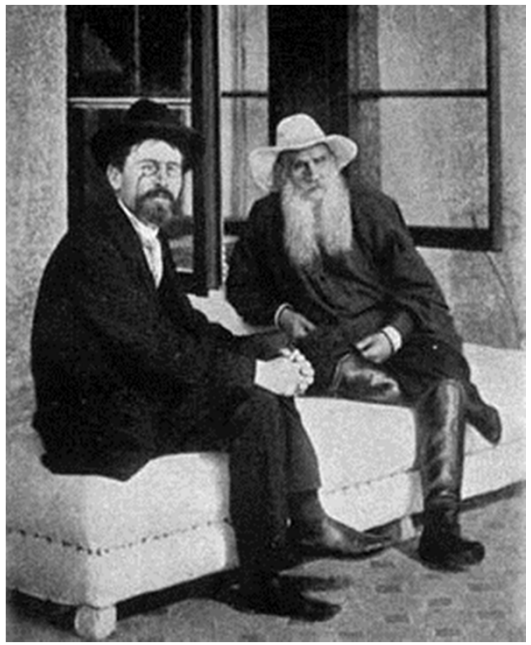
# The dependence of the number of quotations with respect the source's publication date and the years of literary activity of the most cited in the Russian Wiktionary writers

# The distribution of Wiktionary quotes (17-21 c.)

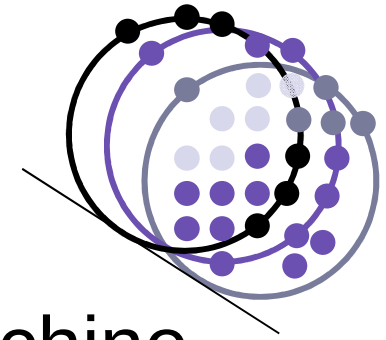


# Results (1)

- 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)

Google play

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