

Geography of Russian poetry: countries and cities inside the poetic world

Elizaveta Kuzmenko, Boris Orekhov

ekuzmenko_2@edu.hse.ru, borekhov@hse.ru

National Research University Higher School of Economics

Our paper is dedicated to two major problems: the first problem is the digital one and the second problem is a humane one. The digital problem involves automatic extraction of named entities, and the humane problem is connected to the usage of toponyms in poetic texts. Correspondingly, our research comprises two parts: automatic processing of a huge amount of texts from the corpus of Russian poetry and revealing major trends in the functioning of toponyms during the history of the Russian poetry from XVIII to XX centuries.

Our research is based on the data from the poetic corpus which is a part of Russian National Corpus¹. This corpus includes the main texts belonging to the Russian poetry from all the periods of its history, up to the XX century. The principles of text selection in the poetic corpus are described by its creators (Grishina et al. 2009; <http://ruscorpora.ru/search-poetic.html>). The size of the corpus is approximately 11 million word tokens.

Up to the present moment, research papers considering toponyms in Russian poetry described a concrete toponym from the perspective of an isolated text or a particular author (see, for example, Mednis 1999). Our approach is quite different: we describe the geography of Russian poetry as a whole, consistently to the framework of distant reading (Moretti 2005, Moretti 2013). Thus, the result demonstrates global trends in the usage of toponyms in Russian poetry as a system.

We used two different technologies to extract geographic entities from poetic texts, and the comparison of these two approaches is one of the results of our research. The first technology is a proprietary commercial software Textocat², which is based on machine learning with the use of nonfictional texts as a training sample. The creators of this software claim that the F1-measure for the retrieval of named entities is 0.75. However, it is expected that the performance would be much lower in the case of poetic texts, because the language of poetry differs radically from the language of prose.

The second approach we use is a self-made tool for the extraction of toponyms based on the dictionary of geographical names. We are forced to create such a tool because there is no open-source software for the extraction of toponyms for Russian. As a basis for our dictionary of geographical names, we use the list of toponyms from Wikipedia. We compared the figures retrieved with our approach to the ones resulting from Textocat. We used for evaluation a sample of toponyms consisting of countries and cities. The comparison showed that Textocat retrieves only 25.7% of country names and 19.3% of city names that are found with our tool. In addition, Textocat makes a lot of mistakes; for example, locative pronouns *там* 'there' and *где* 'where' are retrieved among geographical entities. The words *страна* 'country' and *город* 'city' are also included by Textocat in the list of found toponyms.

As we can see, the dictionary-based approach proves to be more efficient, and in further results we consider only the data extracted with this method.

First, we will look in detail on the names of countries extracted from poetic texts. The distribution of mentioning countries is presented in Table 1 (six most popular countries are taken):

1 <http://ruscorpora.ru/en/>

2 <http://textocat.ru/>

Table 1. The most frequently mentioned countries.

Country	Number of times it is mentioned
Russia	2744
France	283
Italy	241
Poland	201
Lithuania	160
Greece	159
Egypt	151

It is not surprising that Russia takes the first place on this list. The top of the list is occupied mainly by European countries. The second place goes to France, because its influence on Russian culture was immense: in the XIX centuries French was even the main language of communication for Russian aristocracy. Italy can be found on the third place, despite the fact that it is very important for the poetic mythology in the XIX century, and it was the main geographical location for the poetry of eminent Russian poets Batyushkov and Baratynsky.

It should also be mentioned that Russian poetry does not favor exotic countries and is primarily occupied with the European neighbors of Russia (Poland, Lithuania, Greece). The only African country in the top of «poetic» countries is Egypt, which is renowned for its ancient mythology and pyramids with considerable poetic potential.

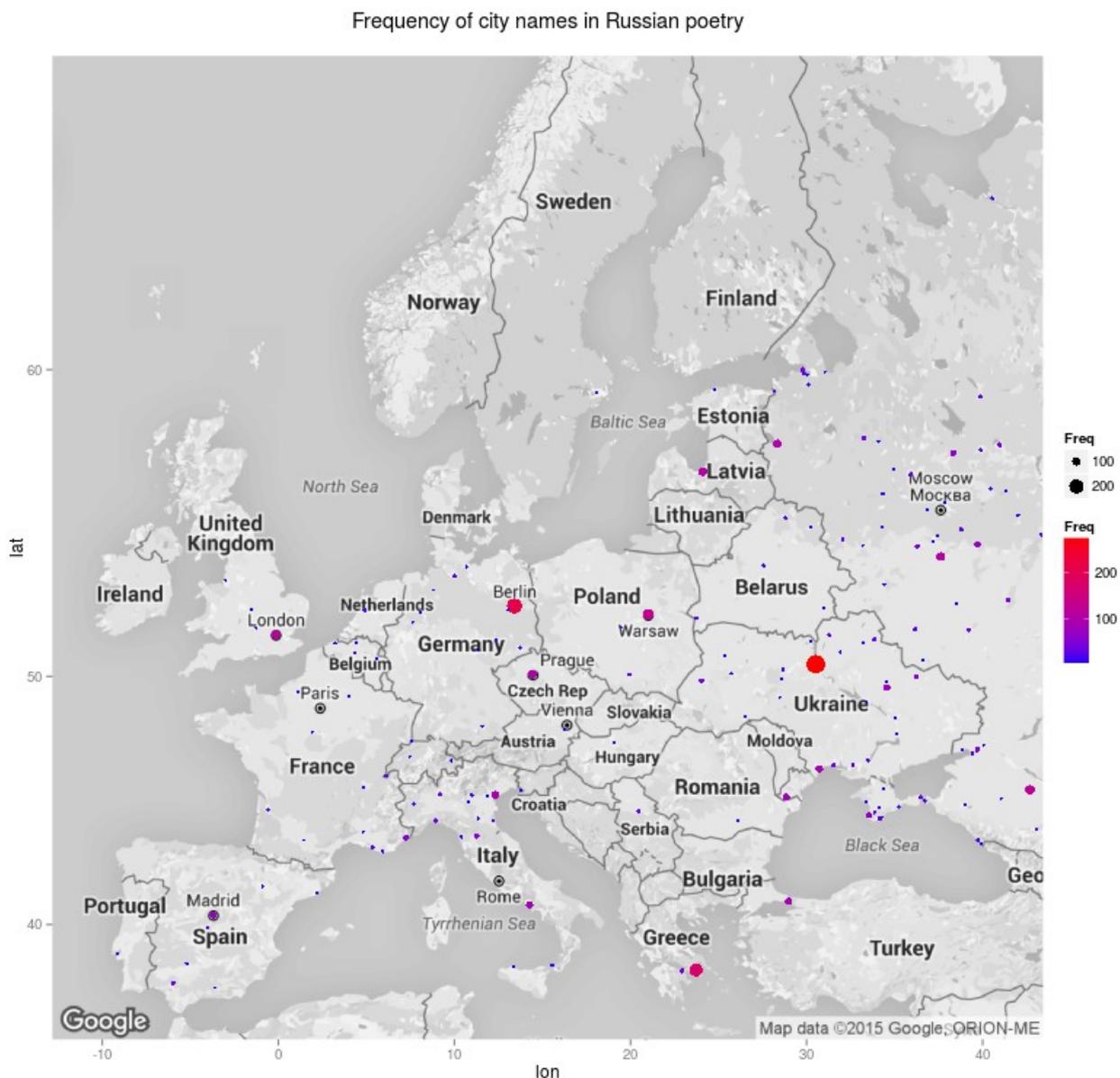
The second exotic country in our list is India, and it is followed by countries of specific «Russian East», which includes Georgia and Iran. The most popular country from Middle East is Lebanon. If we take a look at the contexts in which Lebanon is used, we see that this country is mostly mentioned due to the cedars of Lebanon. It is unexpected that Lebanon appeared to be more frequently mentioned than the Northern European countries (Finland, Norway, Denmark).

Now we will consider mentionings of city names in Russian poetry. The frequency of names for European cities can be seen on Figure 1. This map reflects mentioning of cities with frequency lower than 690. Thus, we drop such cities as Moscow (with frequency of 2470), Rome (with frequency of 1135), Paris (771), and Saint Petersburg (695). Let us note that Rome is more frequent in Russian poetry than Paris, although France dominates Italy in the list of countries' mentionings. Also, we do not mark on the map those cities whose frequency is lower than 4.

As we can see from the visualization, the most «poetic» cities from the point of view of Russian poets are concentrated near Moscow and Saint Petersburg, and also in Ukraine and Northern Italy. Ukraine was a part of Russia during the history of Russian poetry, but the specifically Russian territory demonstrates uneven distribution of mentionings, whereas Ukraine is densely populated with poetic cities. The Crimea draws attention as the most «charged» with poetic cities, though it is not a big territory itself. Sea coasts of Southern Europe generally provide us with cities popular among Russian poets.

Continental Europe is not frequently mentioned in the poetry, with an exception of the capital cities of Prague and Warsaw. The blank spaces can be found throughout the territories of France and Germany. Scandinavian cities also don't have considerable amount of mentionings within the history of Russian poetry.

Figure 1. Frequency of city names in Russian poetry.



Another interesting opposition lies in the distribution of mentionings for Romanic and Germanic cities. As we can see, Russian poets prefer the cities of Italy, France, Spain and Belgium, whereas cities of Britain, Germany and Netherlands appear to be less poetic. Probably, the reason underlying this distribution is not the language, but the confession. Apparently, Russian poets prefer catholic countries to the protestant ones, and Vatican itself can be found on the 17th place judging by the frequency of mentioning countries in the Russian poetry.

To sum up, mentioning of toponyms presents interesting data which can be used to reveal trends in the poetry, and those trends help to describe Russian poetry as a system. It is impossible to notice such patterns with manual or visual analysis of poetic texts, but it can be achieved through digitalization of poetry and computational analysis of the texts.

Grishina E., Korchagin K., Plungian V., Sitchinava D. *Poeticheskij korpus v ramkah NKRYa: obschaja struktura i perspektivy ispol'zovanija*. 'Natsional'nyj korpus russkogo jazyka: 2006-2008. Noveye rezul'taty I perspektivy'. Saint Petersburg, 2009. P. 71-113. [Poetic Corpus in RNC: general

structure and using perspectives]

Mednis N.E. *Venecija v ruskoj literature*. Novosibirsk, 1999 [Venice in Russian Literature].

Moretti, Franco. *Graphs, Maps, Trees: Abstract models for a literary history*. Verso, 2005.

Moretti, Franco. *Distant Reading*. Verso, 2013