Proceedings of the Second Workshop on

# Corpus-Based Research in the Humanities

# CRH-2

25-26 January 2018 Vienna, Austria

Editors:
Andrew U. Frank
Christine Ivanovic
Francesco Mambrini
Marco Passarotti
Caroline Sporleder

Proposed BibTeX entries:

```
@Proceedings{crh-2,
  title  = {Proceedings of the
       Second Workshop on        Corpus-
Based Research            in the
Humanities {CRH-2}},
  year   = {2018},
  editor = {Andrew U. Frank and
       Christine Ivanovic and
       Francesco Mambrini and
       Marco Passarotti and
       Caroline Sporleder},
  volume = {1},
  series = {Gerastree proceedings},
  isbn = {978-3-901716-43-0},
}

@InProceedings{crh2intro2018,
  author    = {Francesco Mambrini and Marco Passarotti
               and Caroline Sporleder},
  title     = {Preface},
  booktitle = {Proceedings of the Second Workshop on
                     Corpus-Based Research in the
                     Humanities {CRH-2}},
  year      = {2018},
  editor    = {Andrew U. Frank and Christine Ivanovic
               and Francesco Mambrini and Marco
               Passarotti and Caroline Sporleder},
  volume    = {1},
  series    = {Gerastree proceedings},
  pages     = {I-IV},
  isbn      = {978-3-901716-43-0},
}
```

Copyright ©2018 by the individual authors.
All rights reserved.

ISBN 978-3-901716-43-0
Published by Gerastree Proceedings, GTP 1.
Dept. of Geoinformation, TU Wien, Austria.

Cover:
Les Fourches, seen from the Bretagne coast near Plouarzel (France). Photo by Andrew U. Frank.

# The Poetic Corpus of Russian: Where the Poems are Written

Dmitri Sitchinava* and Boris Orekhov[+]
*,[+]School of Linguistics, Higher School of Economics, Moscow
*Institute of the Russian language, Russian Academy of Sciences
Emails: *mitrius@gmail.com, [+]nevmenandr@gmail.com

The paper discusses the marking of the composition location in the Poetic Corpus of Russian that enables customizing subcorpora by these locations and subsequent search by this parameter. The place names indicated by the authors are extracted, tagged and "normalized", that is, all the different versions of names and minor locations are boiled down to a narrower range of locations put on an interactive map. This enables a study of lexical and other means used in the texts with regard to the location where the text is composed.

## 1. The Poetic Corpus of Russian: an introduction

The Poetic Corpus of Russian was launched in 2006 (see Grishina et al. [3]) and is available for online searching within the Russian National Corpus (ruscorpora.ru). It counts currently 11 million tokens, encompassing the period since the beginning of the 18th century until the end of the 20th century and has been further expanded every year. The texts include short and long poems, as well as drama in verse. The texts feature the word-by-word morphological markup common for the whole Russian National Corpus, to which another layer of information is added, viz. the poetic structure proper: its metric scheme (iambic, trochaic, free verse etc.), information on stanzas, the number of metric feet in each line, the type of rhyme and other parameters. A lemmatized word or a word combination is searchable within the subcorpus consisting of lines with customized poetic parameters, including the rhyming position. All the poems meeting a certain criterion can be also browsed without searching within their text a specific linguistic expression. The strong syllables (ictuses) in the traditional syllable-tonic or tonic (non-free) verse are systematically marked up, making the corpus a linguistic source in the history of the Russian stress in its own right (see eg. Grishina [3], Sitchinava [7] and others).

## 2. The text proper vs. Title-final Complex (TFC)

A poem as it is published by the author or in a critical edition often does not consist only of poetic (metric) lines, but includes also what is called sometimes Title-Final Complex (henceforth TFC; in Russian see eg the

Handbook of Poetry [1] on the topic). It may include all or some of the following elements: the title of the poem, subtitles, dedications, epigraphs (in the beginning of the text), the author's notes, the author's date of the composition of the poem, the place of its composition (in the end of the text) and possibly other elements. They are a part of the author's text and they are necessary for the analysis of a piece of poetry without instantiating verse in its proper meaning. The Corpus should provide the possibility to search key words within the TFC and the lines separately (see eg. Leibov [6]). For example, a search by the months' and seasons' names in Russian verse may not yield words like "summer" or "July" if they occur in the date in the end of the text (such as "July 23, 1856"). They are marked separately by the means of XML tags and they are separated from the metric lines. A regular search by Poetic corpus will not normally find them by default. The same should apply for the epigraphs (that are, in a vast majority of cases, quotations from other texts) and titles (the word usage in a poem's title can be a research topic in its own right apart from the study of the text of the poem's body, and the possibility of such a search query is to be provided; cf. Grishina [2]).

## 3. Marking up "composition locations"

The toponyms occurring in poetical texts have been already studied on corpora, for example by Leibov [5]: this author explores the "rhyming potential" of such place names as *Moskva* 'Moscow', *Varšava* 'Warsaw' and *Poltava* (Ukrainian town famous as the place of the 1709 battle fought by Peter the Great against the Swedish Caroline army and its Cossack allies); each of them has a range of political associations that are activated through rhyming words (such as *golova* 'head' or *slava* 'glory'). Different place names occur with different frequency in the rhyming position. A difficulty for this study was related to the fact that at that time the place markers included into the TFC were also marked up and searched alongside with the bulk of the text, which was the only search option and obfuscated the raw search results and statistics. In the work by Kuzmenko and Orekhov [4], the space of Russian poetry is analyzed from the point of view of toponyms (countries and cities) mentioned in the texts. Now, thanks to the markup of the TFC, we are able to compare the places mentioned in the texts with the location from the TFC. The map shows that in both cases Russian poetry is more European than American poetry. It has very few American toponyms. And the places mentioned by the authors and the places in which the poems were written are mostly the same.

A project of our team undertaken in 2016 consisted in separate marking of "composition locations" specified by the authors to make them a searchable field of the corpus's database and to put them into the (modern) geographical map, showing the spatial domain of the Russia poetry. This parameter, however, could be marked if and only if it was specified in the known author's text. Very often the poets failed to do so, either if they never did it (or neglected the TFC at all) or when a poem was composed in a "default", unmarked location. It was unusual (although not unknown) for a poet living in Moscow to specify the city in all the texts created there; it was more natural for a denizen of Saint Petersburg or Odessa who visited the city. Alongside with this general challenge, some other problems occur.

The authors naturally specify the place names that were official or commonplace at the time of composition; these names could have graphic and other variants (*Sankt Peterburg – Petrograd – Leningrad; Peterburg, SPb., Pburg* etc.), including even mistakes (for example, spelling of foreign place names with different order of letters or without some diacritics); all these variants were to be merged. These names could have been borrowed from different languages than it is the standard practice now (for examples, Estonian and Flemish place names were, before 1917, taken from German and French respectively, like *Hungerburg*, now *Narva-Jõesuu* in Estonia or names in *Saint-* instead of *Sint-* in Flanders). The authors often specified well-known or obscure microtoponyms beyond the city/town level (streets, neighborhoods, hospitals, hotels, restaurants etc.), whereas all the locations within the same city were to be kept together in order to make possible building of a subcorpus of this city. Some specifications within the TFC locations contained more than one location or additional information that was not toponymic or even altogether locative in nature ("train", "asleep" etc.).

The corpus yielded 672 different locations that were later "normalized", that is, for all of them a string of unified modern geographical names was proposed, with some localities included into other wider ones (for example: *Lubyanka IN Moscow; Boulevard Raspail IN Paris*). This boils down to about 400 "normalized" locations.

The metatextual markup of the Russian National Corpus was expanded by two additional fields: "location" and "normalized location". The first one allows for an exact search as it was put but the author (*Leningrad* but not *Sankt-Peterburg*), whereas the second one consists of normalized locations that merge under one label all the alternative names.

## 4. Search according to the geographical markup and case studies

The Russian National Corpus (ruscorpora,ru/search-poetic.html) provides now customization of subcorpora according to the both additional metatextual fields. The normalized location can be also specified with means of an interactive map based on Yandex Maps where the locations are marked according to the current Russian transliteration (http://ruscorpora.ru/saas/poetry_map.html). It is possible to search within these customized subcorpora some lexical items that are characteristic for the texts created in a given location.

For example, it is possible to create a subcorpus of Russian poetic texts marked by Parisian locations; these poems are normally written by the Russian authors visiting Paris (not living there, as, for example, it was the case after the post-revolutionary emigration).

It is possible to use the subcorpora customized in such a way for studying some markers that correlate with the place where the text is created. The "Parisian texts" are predictably marked by a higher frequency of lexical markers that create the (stereotyped) image of the city: *bul'var* 'boulevard' (460 instances per million vs. 120 in the Muscovite corpus), *kaštan* 'chestnut' (250 instances per million in the Parisian corpus, not a single example in the Muscovite corpus), whereas *fonar'* 'lantern' is not characteristic for either city and even is slightly more frequent (per million) in the Muscovite texts. Prosodic factors can also be statistically studied on these subcorpora, for example to define whether there is any statistically significant difference between the "Muscovite" and "Saint Petersburg" verse cultures (as this is sometimes claimed).

## References


[1] Azarova, Natalija, et al. (2016). *Poezija: Učebnik* [*Poetry: a handbook*]. Moscow: BSG Press.

[2] Grishina, Elena (2005). Dva novyx proekta dlja Nacional'nogo korpusa: mul'timedijnyj podkorpus i podkorpus nazvanij [Two new projects for the Russian National Corpus: Multimedia Corpus and Titles' Corpus]. In: *Nacional'nyj korpus russkogo jazyka: 2003—2005*. Moscow: Indrik.

[3] Grishina, Elena, Korchagin, Kirill, Plungian, Vladimir, Sitchinava, Dmitri (2009). Poeticheskij korpus v ramkax NKRJa: obščaja struktura i perspektivy ispol'zovanija [Poetic Corpus within the RNC: general structure and applications]. In: *Nacional'nyj korpus russkogo jazyka: 2006—2008. Novye rezul'taty i perspektivy*. SPb.: Nestor-Istorija.

[4] Kuzmenko, Elizaveta, Orekhov, Boris (2016). Geography Of Russian Poetry: Countries And Cities Inside The Poetic World. In *Digital Humanities 2016: Conference Abstracts*. Jagiellonian University & Pedagogical University, Kraków. http://dh2016.adho.org/abstracts/3

[5] Leibov, Roman (2012). Russkaja slava i pol'skaja stolica: k istorii odnogo rifmennogo kliše. [The Russian glory and the Polish capital: on the history of one rhyming cliché]. In: *Istorija literatury. Poètika. Kino: sbornik v čest' Mariètty Omarovny Čudakovoj*. Moscow: NLO.

[6] Leibov, Roman (2014). Neblagodarnyj pajščik: opyt korpusnogo analiza teksta [The Ungrateful Shareholder: an experience in corpus analysis of a text]. In: *Korpusnyj analiz russkogo stixa: vyp.2.* M.: Azbukovnik.

[7] Sitchinava, Dmitri (2014). Akcentuacija glagola *byt'* v russkom stixe [Accentuation of the verb **byt'** in the Russian verse]. In: *Korpusnyj analiz russkogo stixa, vyp. 2*. M.: Azbukovnik.