NATIONAL RESEARCH UNIVERSITY
HIGHER SCHOOL OF ECONOMICS

Olga Lyashevskaya, Kristina Litvintseva,
Ekaterina Vlasova, Eugenia Sechina

A DATA ANALYSIS TOOL FOR THE CORPUS OF RUSSIAN POETRY

BASIC RESEARCH PROGRAM

WORKING PAPERS

SERIES: LINGUISTICS
WP BRP 77/LNG/2018

SERIES: *LINGUISTICS*

Olga Lyashevskaya[1], Kristina Litvintseva[2], Ekaterina Vlasova[3], Eugenia Sechina[4]

A Data Analysis Tool for the Corpus of Russian Poetry[5]

A data analysis tool of the Corpus of Russian Poetry (a part of the Russian National Corpus) is designed for quantitative research in various areas of versology and linguistics aspects of the poetic texts. The core part, a frequency database of the corpus, includes annotation at the level of texts, verses, words as well as patterns of words, letters, and stress. The tool allows a user to study certain properties (e. g. rhyming patterns, lexical co-occurrence) taken alone and in their interaction, both in the whole corpus and in subcorpora. Besides that, it facilitates the contrastive studies of two chosen subcorpora. The paper reports a few case studies demonstrating applicable descriptive and exploratory methods and potential for further research in the field of the digital literary studies.

JEL Classification: Z.

Keywords: poetic corpora, quantitative linguistics, lexical markers, lexical diversity, rhyme, linguistic poetics, versology, Russian language, Russian National Corpus

1. Introduction

Russian versology has always heavily relied on statistics data as the basis for predictions and generalizations on meter, rhyme, and other formal and linguistic features of poetic language (see Gasparov 2005, Taranovsky 2010, Jakobson et al. 1973, Jarkho 2006, to name only a few; see also overviews in Semyonov 2009, Kizhner et al. 2018). This gets support from methods employed in the Slavic quantitative corpus linguistics (Kopotev et al. 2018, Divjak et al. 2017) as well as from the formal methods in poetry in general (Scherr et al. 2011).

As quantitative analysis requires processing a big collection of texts, the language technologies responded to this challenge by creating the Poetry Corpus as a part of the Russian National Corpus. The Russian Poetry Corpus is a digital resource provided with the standard morphological and lexico-semantic tagging and a number of specific tags particularly suited for poetic language. For example, the search options offer possibilities to

[1] National Research University Higher School of Economics, Moscow, Russia; Vinogradov Institute of the Russian Language RAS, Moscow, Russia; olesar@yandex.ru
[2] National Research University Higher School of Economics, Moscow, Russia; E-mail: tinalitvina@gmail.com
[3] National Research University Higher School of Economics, Moscow, Russia; E-mail: evlasova@hse.ru
[4] National Research University Higher School of Economics, Moscow, Russia; E-mail: grunwaldus@gmail.com
[5] The research was prepared within the framework of the Academic Fund Program at the National Research University Higher School of Economics (HSE) in 2018 (grant #18-05-0047) and by the Russian Academic Excellence Project «5-100».

collect texts written in various poetic metres, genres, certain patterns of rhyme, verse forms and even graphical shapes. For more information about preparation of text collection included in the Poetry Corpus and the principles of its annotation see (Grishina et al. 2009).

The Poetry Corpus has been proven as an effective tool for fast extraction of raw and normalized frequencies required for stylistic and diachronic research on poetic language. As a digital resource it provides additional large-scale data for verification and support of traditional close reading methods. However, the state-of-arts and emerging field of Digital humanities, computational stylistics, and neurocognitive poetics develop methodology that requires more sophisticated statistic data and correlations (Jacobs 2018). Comprising more than ten million tokens with multilevel annotation, the Russian Poetry Corpus is already a large representative resource for sophisticated quantitative studies. However, for revealing the patterns within the data, the text collection requires additional annotation of poetic and linguistic features as well as tagging of relevant historical background information essential for observation of cultural trends supported by quantitative data.

This article describes a new resource assembled from the data and annotation of the Russian Poetry Corpus, henceforth called a frequency database of the corpus. The new resource is designed by a interdisciplinary research group from the Higher School of Economics (Moscow). This project aims to design a database with elaborated annotation and an open-access web application which provide statistical tools for data summarising, filtering, and pattern structuring. The remaining text is structured as follows. Section 2 provides a detailed overview of the Russian Poetry Corpus that we employed while designing the frequency database. Section 3 describes the structure and improved features of the frequency database. Section 3 features a number of case studies in which the frequency database was used for the cross-disciplinary exploratory analysis of the Russian poetry. In conclusions, we discuss the potential and limitations of the designed tool.

2. The Russian Poetry Corpus

The Russian Poetry Corpus is a large representative collection that comprises poetic texts since the 18th century till the present - from the early verses of Stephen Yavorski and Mikhail Maksimovich seen as forerunners of the Russian poetry tradition to the poetry of modern authors, such as Mikhail Aizenberg, Sergei Gandlevski, and so on. As an integral part

of the Russian National Corpus, the Russian Poetry Corpus eventually inherits the morphological and semantic annotations applied in other subcorpora, such as the main search, diachronic, newspaper subcorpus and the smaller ones. However, the poetic language primarily differs from the other registers, since it prescribes poetic speech to be specifically harmonized and structured according to certain patterns of rhythm. The Russian Poetry Corpus eventually contains additional layer of annotation with literary significant tags, such as rhyme, metre, stanza (onegin stanza, otta rima), foot, graph shape, including both regular patterns and their innovative variations documented and systematized in profound monographs of Mikhail Gasparov and Aleksandr Kvyatkovski.

The fossilisation and innovations of main poetic parameters, as well as synchronic and diachronic trends in poetic language, depend to a wide extend on cultural conventions and historical context. Therefore, the core element of the Russian Poetry Corpus is meta-textual annotation by literary parametres, such as a genre, an author, his or her lifespan, a gender of a poet, a date of poems, an original / translation and some other available information. Due to the set of morphological, semantic, poetic, and metatextual annotation layers, the search toolkit provides wide scope of options for extracting raw and normalized frequency lists corresponding with different linguistic and literary parameters.

However, the Russian Poetry Corpus was primarily a tool for linguistic and formal studies, and the original rationale did not include elaborated metatextual annotation, which primarily reflects information included by Soviet editors of poetry series that constitute the main text collection. Although the academic series named as Biblioteka poeta (Poet's library) has been seen as a well-recognized edition, many volumes do not provide much metatextual information about poets or poems or many judgements have been later revised. Eventually, the Russian Poetry Corpus inherits the metatextual annotation from the soviet academic editions and lacks some of the necessary information, even after preparatory revision by the corpus developers. Due to this, the research potential of the Russian Poetry Corpus is fairly limited for Digital literary studies.

As Grishina et al. (2009: 71–113) note, the Russian Poetry Corpus allows to reduce significantly amount of manual work while extracting basic statistical data relevant for the versology studies. Meanwhile, at present, most of the data summarizing, filtering,

normalizing and preprocessing for the follow-up quantitative analysis is done on a user's side. To make the Russian Poetry Corpus more efficient and user-friendly, the is a need in the new digital tool to maximize the output of the corpus and provide the pre-processed datasets for lexemes, syntactic units, and other tags available in the corpus.

3. A frequency database of the Corpus of Russian Poetry

The database is compiled using the materials of the Russian Poetry Corpus. The objective is to assemble the source corpus data with more accurate and elaborated metatextual, versological and linguistic annotation which is validated by experts in modern literary studies and NLP. Another objective is to develop a web-application with an incorporated tool for data analysis that computes basic statistical information, frequency lists, and powerful illustrative visualisation for different parameters in the whole corpus or in a subcorpora specified by a user.

The database with enriched annotation consists of five sections with the following types of information:

1) about lemmas and relevant syntactic units;
2) about poetic strings;
3) about a text and its historical context;
4) about rhymed units;
5) about letter combinations and stress patterns.

Every lemma has several tags indicating its position in a poetic string, ictus structure, PoS, dependencies between syntactic units. The text collection was annotated by the means of the Ru-Syntax (Medyankin, Droganova 2016), and the morphological disambiguation of lemmas and PoS was done manually. For semantic annotation, we used the taxonomy of the Russian National Corpus developed by (Kustova et al. 2005). The semantic classification relies on the first (basic) meaning, and this approach enables clusterization of lexemes into bigger groups, such as names of plants, sound verbs, color adjectives, compounds, diminutives, and so on. The new database also contains information about lexical collocations seen as 2-grams and 3-grams with syntactic dependencies between them. Due to the syntactic annotation, a user can also find a phrase constituent in distant position from its controller and

therefore study a group of phenomena specific for poetic syntax, such as atypical word order, rhythmic and repetitive syntactic patterns, as well as enjambment, or incomplete syntactic units in the end of a line.

Every line contains tags indicating metre and foot, a number of words and syllables, ictuses and information about rhyming segment. We also marked up the end of a sentence in the middle of the string, in the beginning, and in the end. At the next stage, we will add information about rhythmic forms (Taranovski 1971, Lyapin 1997).

We also enhanced original metatextual annotation about poetic texts and designed tables with core background information about authors, dates of poems, poetic features and so on. The criteria suited for cluster analysis are as follows: gender, age at the moment when a poem was created, place and date of a poem. For distributional analysis, some features and nuances were unified. We also added new metatextual parameters, such as decade and poetic school, original or translated texts, extended authorship, which comprises texts written with another poet or associated with a certain poet.

The rhyme annotation comprises rhymed chains, an order of rhyming elements, a number of elements in a chain (usually two, less often three, four, or longer chain rhymes, e.g. monorim). Since the corpus does not provide information on rhyming pairs, we retrieved this data automatically based on the information about elements of rhyming units and rhyming schemas[6]. We also marked up word-long rhymes and larger rhyming units, stress and ictus patterns, patterns of vowels and syllable structure, PoS and other grammatical features of rhyming elements.

Finally, elaborated annotation comprises char-grams and combinations of letters, their classification, combinations of vowels and consonants, as well as stress patterns. This information is based on the graphic elements. In a further perspective, we plan to add morpheme-specific annotation.

The database includes corpus texts dated from 1800 to the present. The enhanced corpus comprises about 80 000 poetic texts both of short and long genres. This comprises more that 2 million verses and about 10 million words. Poetic metatexts, such as headings,

[6] At the moment, we have processed 60% verses taking only data in which the rhyming schema does not change within the whole document.

dates, epigraphs, prosaic comments made by an author, editors' notes belong to a separate part of the database and are subject to a separate investigation, see for example (Kuzmenko, Orekhov 2016).

4. Case studies

The new frequency database provides data on the occurrences and distribution of the linguistic and versological elements within the corpus. This includes the frequency lists of words, lemmas, PoS, collocations, char-grams, metre and size properties, rhyming schemas and rhyming chains. The search toolkit gives a user an option to choose subcorpora and to conduct contrastive research based on their comparison, for example, a user can determine the key lemmas of the subcorpus. In addition to the raw frequencies provided by the database, the statistical tool calculates the relative frequencies, metrics of variation and keyness and allows one to visualize the distributional data. What is more important, it is possible to look at the data at the intersection of different dimensions such as metric properties, rhyming properties, position within the verse, grammatical properties, stress pattern, etc. Therefore, the statistical tool facilitates the research in the field of poetic stylometry, lexicology, collocations, morphology and syntax, as well as diachronic and synchronic studies of the poetic tradition and language.

The following sections report on a few case studies which illustrate an interdisciplinary research potential of the database as a tool for the digital literary studies.

4.1. Distribution of basic features

Figure 1 demonstrates how the size of the corpus measured in tokens varies if we group the texts by (a) decades or (b) authors. A user can gather information about the proportions of subcorpora and normalize the raw frequencies with regards to the size of the specified subcorpus. A user can also determine the period of poetic tradition to be included in their research. For example, the size of the subcorpus comprising the period after 1980 is rather small and cannot provide reliable statistical data in many cases. Furthermore, for the contributors and maintainers of the corpus, these graphs would also suggest which parts of the corpus need balancing and adding new data.

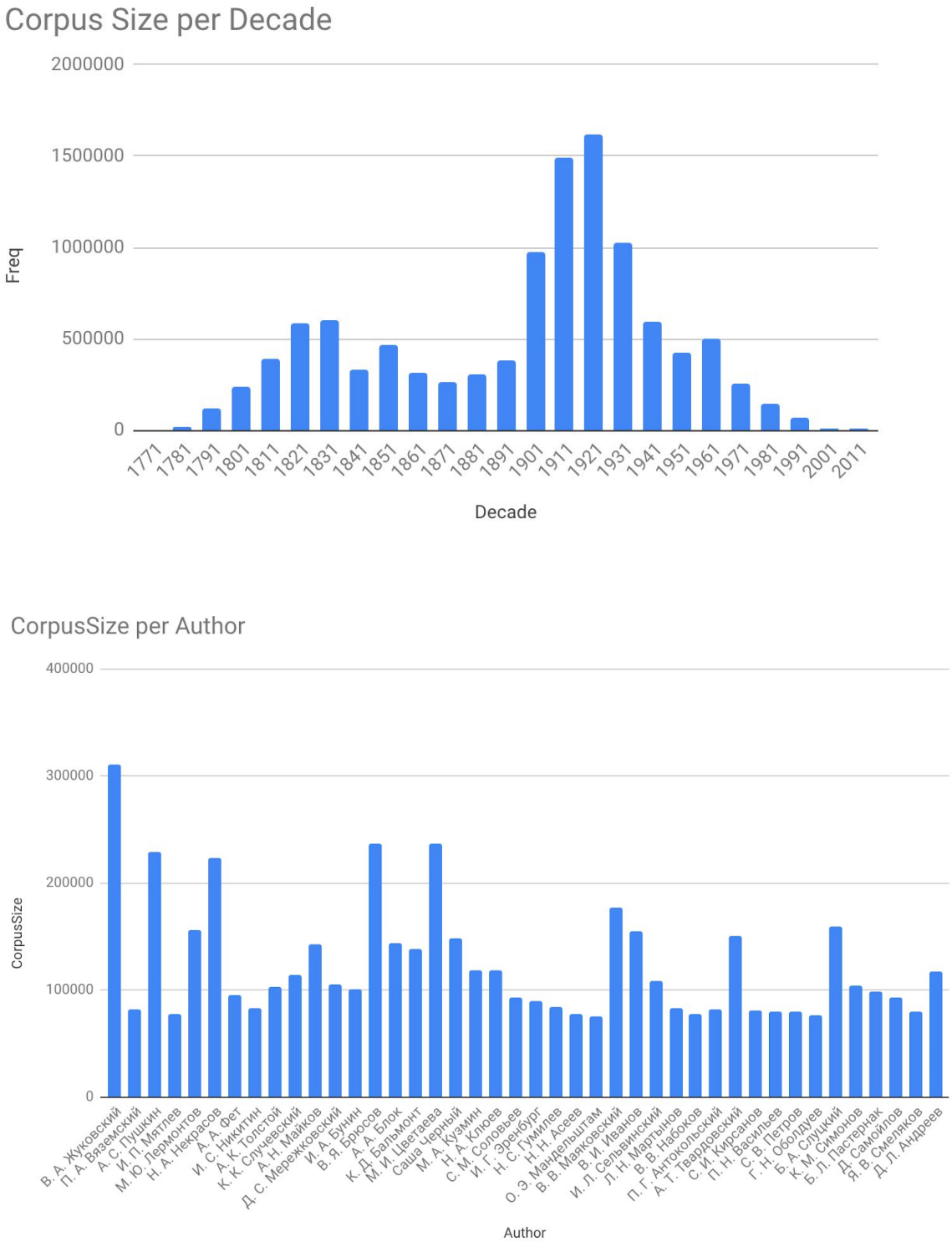

Fig 1. Varying of corpus size (a) by decade и (b) by author

Table 1 illustrates a key lemma list of a user-specified subcorpus. Here, we present the key verb lexemes in the poetry of the 1920s compared to the whole corpus. The words in the frequency list are filtered by the part of speech and ranked using the `delta` metric[7].

Table 1. Key verbs of the poetry in the 1920s.

Lemma	F subcorpus	F corpus	Delta
плыть 'float'	672	3530	221
встать 'rise (pf.)'	644	3416	208
вставать 'rise (impf.)'	546	2853	182
бить 'beat'	627	3492	181
петь 'sing'	1476	10241	168
звенеть 'ring'	493	2557	166
лечь 'lie down'	456	2362	154
гудеть 'buzz, drone'	330	1419	149

Eight most key verbs consists mostly of actional predicates referring to ascent (*встать*, *вставать* 'to rise') and sound (*петь* 'sing', *звенеть* 'ring', *гудеть* 'buzz, drone'). The follow-up qualitative corpus study would suggest that they catch up the cheerful spirit of the post-revolution era as well as the nostalgic overtones of the immigrants' lyrics. The next step would be to analyse why the verbs *плыть* 'float' and *лечь* 'lie down' also were on the list, which poets used these lexemes frequently, and which meaning did they imply. Compare the examples from the corpus.

Косяк овец вдали // *Плывет гурьбой волнистой* 'A shoal of sheep in the distance // Floating as wavy crowd' (Sasha Cherny);

Рука, на приклад ляг! 'Hand, lie down the (rifle) butt!' (Vladimir Mayakovsky);

Рядышком ляжь 'lie down next (to me)' (Marina Tsvetaeva).

Further we will discuss a few other case studies related to the word frequencies.

4.2. Lexical diversity of adjectives across time periods

Using the tool, a user can extract the frequency lists of certain parts of speech. As an example, we will compare the frequencies of different adjectives within two periods of

[7] Delta is calculated as the difference between the observed and expected frequencies divided by 2. The other metrics provided by the statistical tool include keyness add-N (Kilgarriff 2009), TF-ICTF (modified according to Baranov 2018), log-likelihood ratio and chi-squared score (Scott, Tribble 2006).

Russian poetry which are traditionally called ‘the Golden Age’ (1811–1840) and ‘the Silver Age’ (1901–1917). The size of the Golden Age subcorpus is 1, 270, 000 tokens and the size of the Silver Age subcorpus is 1, 619, 228 tokens. The first prominent difference between these two periods becomes apparent at the stage of frequency extraction. Although the proportion of adjectives related to other PoS does not change dramatically (10.8% in the Golden Age and 12.6% in the Silver Age), the number and diversity of lexemes has increased by the 20th century. While the Golden Age subcorpus contains 6, 421 unique adjectives, in the Silver Age there are 11, 777 lexemes.

We examined the adjectives that belong to the most frequent “top” at least in one of the periods — namely, all adjectives which have a frequency higher than 0,1% at least in one period. The number of such adjectives is not large: the Golden Age “top” by frequency contains 226 adjectives, whereas the Silver Age “top” includes 210 lexemes, the total amount of different adjectives excluding repetitions being 278. Most of the adjectives that occur in the poetic texts of the two periods belong to the “top” list: 60,9% and 54,8% in the Golden and the Silver Age respectively.

For the further comparison of the two periods, we have selected the adjectives whose frequencies in the Silver Age differs significantly from the Golden Age numbers. The first list (Appendix, Table A) contains the lexemes which are considerably more frequent in the Silver Age rather than in the Golden Age, namely whose frequencies more than doubled in comparison to the beginning of the 19th century. The second list (Appendix, Table B) presents the adjectives occurrences of which in the Silver Age decreased by half or more in comparison to the Golden Age data.

The first list includes 54 lexemes whose popularity increased in the Silver Age; 40 of them do not belong to the “top” of the Golden Age. Notably, many of the adjectives from this list belong to one of the two semantic groups. The first group consists of color adjectives. The colors *белый* ‘white’, *чёрный* ‘black’, *золотой* ‘golden’, *красный* ‘red’, *синий* ‘blue’, *зелёный* ‘green’, *голубой* ‘light blue’ have already belonged to the “top” in the Golden Age and their frequencies have increased even more. In the Silver Age, these words occur 2–3.6 times more often than in the Golden Age. The color terms *алый* ‘scarlet’, *серый* ‘grey’, *жёлтый* ‘yellow’, *розовый* ‘pink’, *серебряный* ‘silver’, which did not previously belong to

the “top”, in the 20th century become much more frequent and occur 2.3–6.7 times more often than in the Golden Age. Although we can explain the frequency of the most popular of these words (*золотой*, *чёрный*) due to their polysemy, a significant part of these adjectives (*синий*, *зелёный*, *голубой*, *жёлтый*, *серый*, *розовый*) denotes only color terms. The group of color words also include adjectives referring to color saturation and brightness — *тёмный* ‘dark’, *пёстрый* ‘motley’, *тусклый* ‘dull’ (the first one has already been in the “top” before, and all the three are also used in a figurative meaning). The last word associated with visual qualities and colors is *прозрачный* ‘transparent’, meaning the absence of any color.

The second group of adjectives refer to nature objects or elements, for example, *лунный* ‘moon (attr)’, *звездный* ‘star (attr)’, *солнечный* ‘sun (attr)’, *снежный* ‘snowy’, *огненный* ‘flame (attr)’, *лесной* ‘forest (attr)’, *горный* ‘mountain (attr)’, *весенний* ‘spring (attr)’, *осенний* ‘autumnal’, *зимний* ‘wintry’, *вешний* ‘vernal’, *вечерний* ‘evening (attr)’. The adjective *вечерний* has already been in the “top” in the Golden Age and its frequency demonstrated the least growth in this group (2.3 times); the most drastic growth is displayed by the adjective *лунный* (8.8 times).

The other words from the list #1 establish small thematic groups of 2–4 lexemes each. Some of the adjectives describe the size of an object — remarkably, tending to the smaller sizes: *тонкий* ‘thin’, *маленький* ‘small’, *узкий* ‘narrow’. The temperature and humidity are described by *сухой* ‘dry’, *тёплый* ‘warm’, *горячий* ‘hot’, *жгучий* ‘burning’ (all these four are often used in the figurative sense). The material of an object is meant by *каменный* ‘stone (attr)’, *медный* ‘copper (attr)’ — this is an example of overlapping groups, as this word also describes a reddish color; on the contrary, the color adjectives *золотой* and *серебряный* also mean ‘made of gold’ and ‘made of silver’. *Девичий* ‘maidenly’ and *людской* ‘human’ are connected to people; *ласковый* ‘affectionate’ and *влюбленный* ‘enamoured’ belong to the emotional sphere; *мудрый* ‘wise’ and *вещий* ‘prophetic’ describe human experience. *Старый* ‘old’, *былой* ‘bygone’ and *далёкий* ‘remote’ describe temporal and spatial distance.

Another six adjectives are not that close thematically; however, all of them describe some “abnormal” state of a person (*усталый* ‘weary’, *пьяный* ‘drunken’), of the perceived world (*пыльный* ‘dusty’, *душный* ‘stuffy’, *зыбкий* ‘unsteady’) or of both (*странный* ‘strange’). Finally, the adjectives *певучий* ‘melodious’ и *загробный* ‘beyond the grave’ do

not demonstrate any thematic relations to the other adjectives in this list; remarkably, it is precisely these two lexemes which display the sharpest rise in frequency as compared with the Golden Age: 15.8 times and 25.8 times, respectively.

The list #2 consists of 63 words which are less popular in the Silver Age than before.[8] Some of them leave the "top" positions because of their archaic form (*златой* 'golden', *хладный* 'cold', *младой* 'young' which have stylistically neutral forms *золотой*, *холодный*, *молодой*) or because the shift in the meaning of the word (*бранный* 'martial' > 'abusive'). Some adjectives are replaced by a synonym: for example, the frequency of *прежний* 'former' and *минувший* 'past' decreases (though *прежний* stays in the "top") — but the decline is partially compensated by the newfound popularity of *былой* 'bygone' as stated above. However, the majority of the adjectives does not have such obvious explanations of the decrease in frequency. Still, some tendencies can be noticed.

The largest thematic group of the lexemes whose frequency decreased in the Silver Age consists of the adjectives with vague but distinctly positive meaning: *прекрасный* 'beautiful, excellent', *прелестный* 'charming', *приятный* 'nice', *отрадный* 'pleasant', *пленительный* 'captivating', *благой* 'good', *возвышенный* 'sublime'. This group is ajoined by *волшебный* 'magical', *чудный* 'wonderful' и *чудесный* 'miraculous', which are used mostly in the figurative positive meaning. The only adjectives from this group staying in the "top" in the Silver Age are *прекрасный* and *волшебный*.

A number of adjectives positively characterizes the human character and deeds; they belong mostly to the sphere of emotions and/or ethics: *добрый* 'kind', *сердечный* 'warm-hearted', *любезный* 'obliging'; *достойный* 'respectable', *благородный* 'noble'; *удалой* 'daring', *храбрый* 'brave', *отважный* 'courageous'; *вдохновенный* 'inspired', *пылкий* 'passionate'; *резвый* 'vivacious', *беспечный* 'carefree' (only the last one still staying in the "top"). Positive characteristics belonging to another spheres are *умный* 'intelligent' and *величавый* 'stately'; a negative one connected to the emotions and ethics — *коварный* 'insidious'.

[8] Two of them are particular cases which will not be discussed below: firstly, the pronominal adjective *многий* 'many'; secondly, the adjective *готовый* 'ready' which is used mostly in its short form *готов* (557 usages in the 1811–1840 subcorpora, whereas the full form occurs there only 87 times).

Another lexemes can refer to a person as well as to the world around. There are, for instance, the polysemantic adjectives which describe among other things some positive personal traits: *славный* ('prepossessing' or 'famous'), *прямой* ('truthful' or 'genuine'), *твёрдый* (as in *твёрдое решение* 'a firm decision' or as in *твёрдый дуб* 'a strong oak'), *важный* ('high-ranking', or 'imposing', or 'significant'). The idyllic mood is created by *мирный* 'peaceful', *невинный* 'innocent', *смиренный* 'humble', *скромный* 'modest' (for example, *смиренный уголок* 'a humble nook' or *смиренный рыбарь* 'a humble fisherman'). They are contrasting with *шумный* 'noisy', *бурный* 'violent', *громкий* 'loud', *мятежный* 'restless' which describe the eventful life of society (*шумный свет* 'the noisy society', *бурные речи* 'a passionate speech') as well as the violence of the nature (*бурный океан* 'the restless ocean', *шумный лес* 'the restless forest'). Another row of adjectives describing both human beings and nature in Russian poetry refers to danger and negative emotions: *ужасный* 'horrible', *мрачный* 'gloomy', *унылый* 'cheerless', *опасный* 'perilous', *грозный* 'terrible', *свирепый* 'ferocious'.

The other adjectives refer to the age (*молодой* 'young', *юный* 'youthful'); to the prosperity (*богатый* 'rich', *роскошный* 'luxurious'); explaining the connections between events (*роковой* 'fatal', *напрасный* 'vain'); the antonyms *счастливый* 'fortunate' and *несчастный* 'unfortunate'. Finally, these adjectives are not included in any thematic groups: *русский* 'Russian', *гробовой* 'sepulchral', *звучный* 'sonorous', *летучий* 'flying'.

By contrasting the two lists retrieved from the frequency database, we can describe the main differences between the Golden and the Silver Ages regarding the usage of adjectives. In the Golden Age, the lexemes which openly name the mood, the emotions (*мирный*, *мрачный*...) or evaluate some characters, deeds and objects (*добрый*, *приятный*...) are much more popular — and more than ⅔ of these lexemes have a vague or a definite positive meaning. On the contrary, the poetry of the Silver Age names the feelings rarely; it describes instead of evaluating and relies above all on the visual component in this description. The adjectives referring to human qualities decrease in frequency; the lexemes referring to the nature become more popular; it can also be noticed that the Silver Age is interested in anything marginal or strange (unusual dimensions, unusual states). Of course, the poetry of

both periods appeals to the reader's emotions; however, the means of the influence upon the emotions have changed.

Many further directions of the study are possible. It is worth noticing that not only the frequency of adjectives differs between periods, but the combinatory power as well. For instance, the lexeme *усталый* 'weary' belongs to the "top" both in the Golden and in the Silver Age, but the number of possible collocations with this adjective grows significantly. In the earlier period, it can describe a person both in the spiritual and in the physical aspect; some other living creatures (*конь* 'a horse', *вол* 'an ox', *стада* 'herds') and objects of nature (*облак* 'a cloud') are called *усталый* as well. However, in the 1901–1917 subcorpus there are mentions of 'weary' movements (*поступь* 'tread', *взмах* 'a wave', *прикосновенье* 'a touch') and even everyday inanimate objects (*паровик* 'a steam engine', *шлейф* 'a train of a dress'), which are not found in the poetry of the Golden Age.

Furthermore, the overview above states the differences between the two periods, but does not describe the individual trajectories of the lexemes which are quite diverse. Table C in the Appendix shows two contrasting examples, the words *больной* 'sick' and *безумный* 'insane', the former one being on the peak of its popularity before the Silver Age in 1880s, the latter experiencing a decline at the same time.

Last but not least, some individual preferences of different authors can be described — an example of such study will be discussed in the next section.

4.3. Authors' use of the color hues in the Silver Age Poetry

The following example illustrates how the frequency database can be used in the analysis of the lexical diversity and the author's word usage. During the Silver Age, multiple poetic schools manifested a new aesthetics and art syncretism trying to combine painting and poetry. By exploring color hue adjectives, this case study aims to reveal how the aesthetic rationale influenced the poetic lexicon. This is done by applying several methods. The first methods concerns a small-scale diachronic analysis of word frequencies through 19th-20th centuries. At the next stage, we apply a method of correspondence analysis (CA) to define frequency-based associations between color hue adjectives and certain poets. The CA method also involves clustering the poets based on the contingency between words and authors.

At the preparatory stage, we extracted a frequency lists of the color adjectives using the lexico-semantic annotation of the frequency database. Then we compiled a list of adjective for color hues by filtering out the most frequent lexemes (such as *красный* 'red', *синий* 'blue') and hapax legomena (such as *алмазно-рубиновый* 'diamond ruby'). The middle part of the frequency list consists of the following lexemes subject to the further analysis: *фиолетовый* 'violet', *лиловый* 'lilac', *лазурный* 'azure', *багряный* 'blood-red', *пурпурный* 'tyrian purple', *белоснежный* 'snow-white', *изумрудный* 'emerald', *лазоревый* 'azure', *бирюзовый* 'turquoise', *златой* 'golden-yellow', *сумрачный* 'murky'.

These lexemes occur more than 100 times each and, apart from the most frequent color hue adjectives (*белый* 'white', *черный* 'black' , *темный* 'dark', *светлый* 'light', *красный* 'red'), do not constitute idiomatic collocations (*красная армия* 'the red army', *белое вино* 'white wine'). The most frequent lexemes usually spread equally in texts regardless individual and genre variation. The less frequent words have potential to become a stylistic feature of a personal style as well as epoch.

The diachronic research comprises the period from 1801 to 1970, including several decades before and after the Silver Age. The graph shows that the frequency of the adjective фиолетовый 'violet' starts rapidly increasing since 1880s to the 1920s. During this time, its frequency increased from ~3 ipm to ~21 ipm and then stays at this rate. The first poet who brought the adjective фиолетовый into poetry was Vasili Zhukovski. However, the frequency of this lexeme reached the peak in the Silver Age. Apart from Zhukovski, фиолетовый occurred in poetry only three times. In 1895, Valeri Bryusov used a collocation *фиолетовые руки* 'violet hands', later the adjective *фиолетовый* occurs in poems of Maksimilian Voloshin, Andrei Belyi, Vyacheslav Ivanov, Ivan Bunin, Aleksandr Blok and many others. The small-scale diachronic frequency analysis has demonstrated that *фиолетовый is* a specific stylistic feature of the Silver Age.

Alongside with the interest towards the violet color, poets employ adjectives denoting its hues. For example, the adjective *лиловый* has the same diachronic graph as *фиолетовый*.

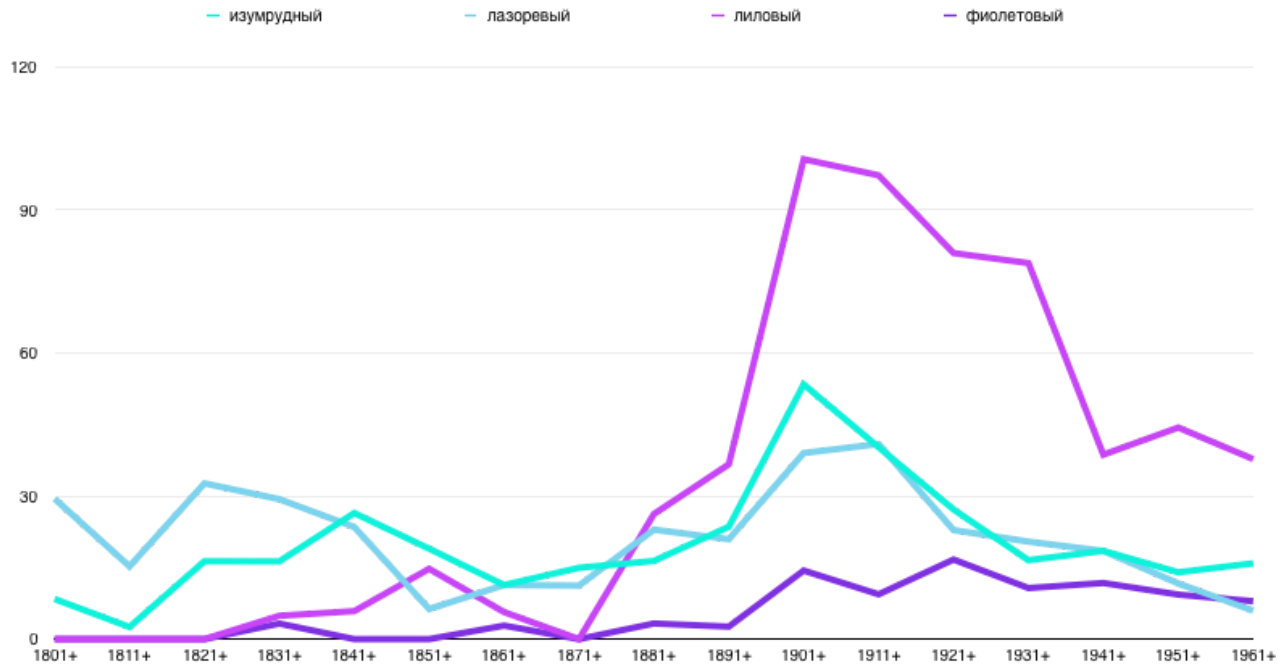

Fig 2. Occurrences of the adjectives лиловый, фиолетовый изумрудный, лазоревый by decade, in ipm.

As the Russian Poetry Corpus shows, the frequency of *лиловый* fluctuates at the rate from ~1 ipm to ~10 ipm. In 50 years from 1970s to 1920s, its frequency increased up to ~120 ipm, stayed at this rate for about 10 years, and started gradually declining in the 20th century.

How does this fashion for certain words emerge and do certain poets play a role in this process? Why does the frequency of some color hue adjectives increase and decrease rapidly? The toolkit of the frequency database allows one not only extract generalized frequency data across decades, but also explore frequency distributions within the corpora of certain poets. For example, the search results show that *лиловый* is regularly attested in the poems of Ivan Bunin, Vyacheslav Ivanov, Mirra Lokhvitskaya, Boris Pasternak, and Igor Severyanin. These poets mostly contribute to the high frequencies of *лиловый* in the Silver Age.

Another two highly frequent color hue adjectives of the Silver Age are *лазоревый* and *изумрудный.* The frequency of *лазоревый* during the period of 1890-1930 do not decrease below 34 ipm, and the average frequency of this adjective is two times higher than beyond the Silver Age. The adjective *изумрудный* has the similar diachronic distribution. Its lowest

frequency within this period is about 50 ipm, and this is twice more that beyond the Silver Age.

At the next stage, we visualized distributional data drawn from the database using the method of Correspondence Analysis (CA, Levshina 2015, Kassambara 2017) as applied to the use of the color adjectives in focus by individual authors.

For a case study, we took ten subcorpora written by Valery Bryusov, Alexander Blok, Konstantin Bal'mont, Igor Severyanin, Nilolay Gumilev, Anna Akhmatova, Marina Tsvetaeva, Osip Mandel'shtam, Boris Pasternak, the choice of texts is not limited by the time of creation.

As a source data, CA takes a contingency table which shows how the linguistic units (9 adjectives of color hues, in our case) are distributed across the subcorpora (10 authors, in our case). The distribution of each adjective across the subcorpora we call a color profile, and the distribution of the uses of each author with respect to the adjectives we call an author profile. Firstly, we calculated an average profile for both adjectives and authors. Secondly, we computed the distance between each pair of the colors profiles and from each color profile to the average color profile. The distances for the author profiles are calculated the same way. Further, a matrix of distances is plotted onto the 2D space using a method of multidimensional reduction. The closer the data points are on the horizontal or the vertical axes, the closer are their profiles. The closer they are to the origin (0,0), the closer their profiles to the average profile.

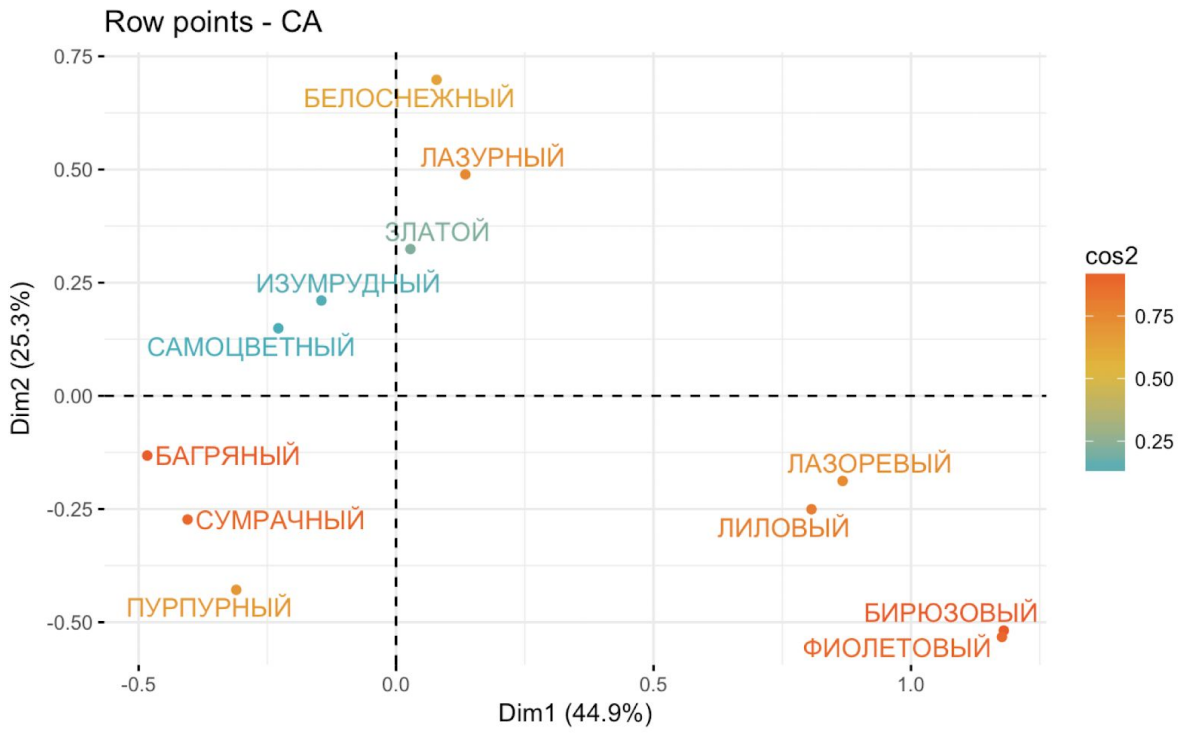

Fig 3. Correspondence analysis plot: adjectives of colors (distances are defined by the authors' use profiles).

Figure 3 visualizes the similarity among the adjectives of color assessed by their frequency distributions in subcorpora. The profiles of белоснежный, лазурный, златой, изумрудный, самоцветный can be considered as opposed to the profiles of пурпурный, бирюзовый, фиолетовый, багряный, сумрачный, лазоревый, лиловый (top vs. bottom part of the plot), and the profiles of багряный, сумрачный and the profiles of бирюзовый, фиолетовый are two poles on the horizontal axis (left vs. right part of the plot). Furthermore, the profiles of изумрудный, самоцветный are much closer to the origin than the profile of белоснежный. This can be interpreted in such a way that изумрудный, самоцветный are used roughly evenly by different authors whereas белоснежный is used considerably more frequently in one or several subcorpora than in others. The axis labels provide information to what extent the variance in the frequency profiles is explained by the 2D visualization, in other words, how much information was lost when the multidimensional space was reduced to two dimensions (100% - 44.9% - 25.3% = 29.8%). It is notable that not all the data points are displayed equally well in the 2D space. The color of the data points shows the quality of their representation on the map (calculated using the squared cosine (cos2) metric), ranging from red (high quality, see фиолетовый, бирюзовый, багряный) to green (low quality, see самоцветный, изумрудный). The latter means that the proximity of the points for

самоцветный and изумрудный on the map can be misleading, and another visualization (a map of the the 2rd and 3rd dimensions) is needed.

The authors' profiles can be plotted the same way. Figure 4 demonstrates a global pattern within the data (symmetric biplot), the colors' profiles (blue points) and the authors' profiles (red triangles) being plotted simultaneously.

The plots on Figures 3 and 4 were built using the subcorpora of nine authors (all except Pasternak). The reason is that his profiles differs a lot from all other poets, and a user would see a dense cloud of points in the center and an outlier. A technique of supplementary points allows one to plot the outlier's point over the plot created for the remaining part of data. To put it differently, the color profiles of Pasternak and Severyanin are not particularly similar, but they are more similar than the profiles of Pasternak and Bryusov, or Pasternak and Tsvetaeva.

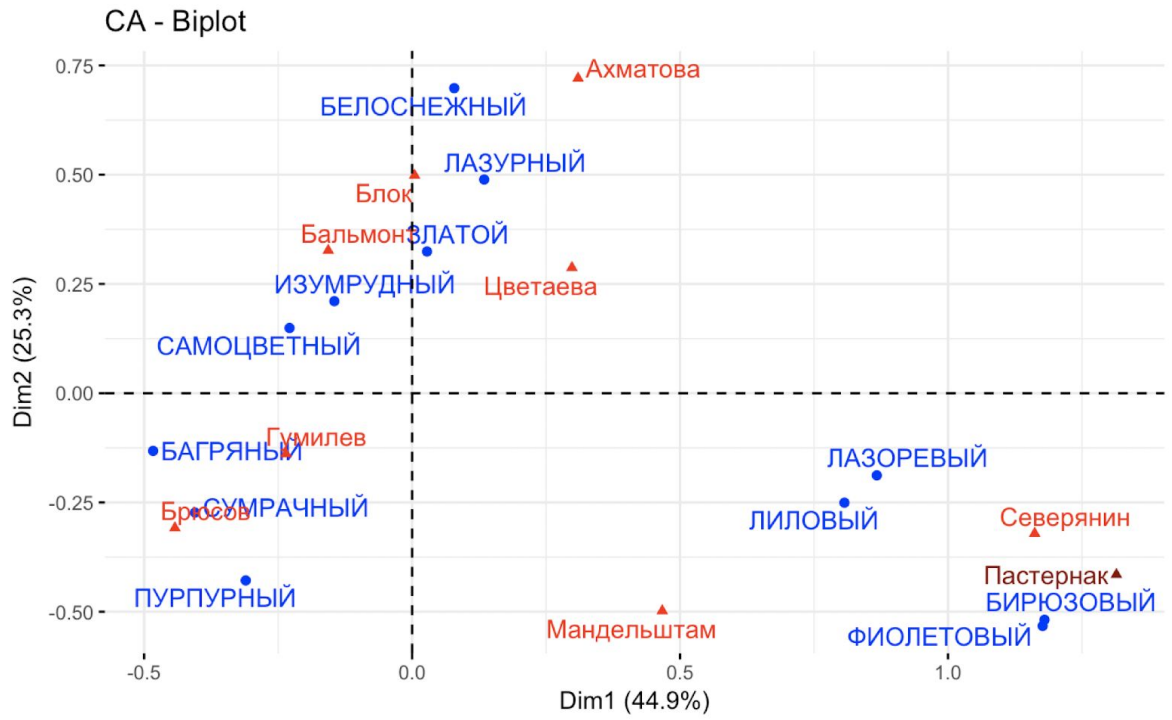

Fig 4. Correspondence analysis plot: authors and adjectives of colors. A supplementary point: Pasternak.

As expected, the graph shows that both Boris Pasternak and Igor Severyanin use the adjective *лазоревый* and *лиловый* frequently. The visualization also illustrates that the

lexemes *бирюзовый* and *фиолетовый* also belong to their poetic lexicon. Despite the low average frequency of color adjectives in poetry of Nikolay Gumilev and Valeri Bryusov, the graph demonstrates that their poems have similar lexical features, such as color terms of *багряный, пурпурный* и *сумрачный*. Anna Ahmatova stands out from the other poets as her poetry is imbued with color adjectives. She also used the most frequent words, such as *белоснежный, лазурный*.

The latter adjective, *лазурный,* alongside with *златой,* are the distinct feature of Aleksandr Blok and Konstantin Balmont. Meanwhile, the lexemes *златой* and *изумрудный* belongs to the poetic lexicon of Marina Tsvetaeva. The adjective *самоцветный* is placed on the scale, so that it can be attributed to the lexicon of Aleksandr Blok and Konstantin Balmont. However, this adjective is not a prominent feature of their poems.

The Correspondence Analysis showed that certain poets prefer different colors and hues. In some cases, their preferences are very explicit (see лиловый in Pasternak's poetry). Meanwhile, some poets turn out to be neutral with respect to the use of the given set of color adjectives (the case of Nikolay Gumilev). It is notable that poets from the same poetic school do not necessarily favour the same color hues. For example, despite of the same aesthetic framework, the acmeists Anna Akhmatova, Nikolay Gumilev, Osip Mandel'shtam are rather distant from each other on the graph. Conversely, the subcorpora of authors belonging to different poetic schools can demonstrate similar distribution of color adjectives. This data supports the conclusion that despite the inner influences within poetic schools, poets' color preferences can be very different. However, the analysis only comprises lexemes with quite moderate frequencies. Due to this, the statistical validity of our observation needs to be proven with additional tests.

4.4. Verb rhymes and verb forms

The dispute about how 'good' the words of the same grammatical form — and especially verb forms — rhyming with each other are, started in the 18th century by Antiochus Cantemir, 'the father of Russian poetry'. He condemned the infinitive forms on *-ati* rhyming with each other as being 'vile', but allowed them to rhyme with other parts of speech, for example, *мати* 'a mother' — *спати* 'to sleep' (Gasparov 2000: 53). Later on, the use of verbal rhymes became a reason to accuse one to be poorly mastering the poetic form

(Samoilov 2005: 341), and there were authors which were known to be consistently avoiding this type of rhyming (for example, in Vladimir Mayakovsky, the verb rhymes are found in only 1% poems with female endings and 2% with male endings (Gasparov 2000: 321)). In the 20th century, however, some authors intentionally played with homonymous and tautological rhymes, and among them with the verb rhymes.

Taking the quantitative corpus data, we can study whether the authors follow Cantemir's recommendation and if the trend changes over time. We propose two hypotheses: (i) the authors seek to avoid the verb-to-verb rhyme in the beginning of the 1800s, but the rule is less strictly observed in the later period; (2) there are periods in which the authors follow the recommendations, but they alternate with periods in which the rule is less strictly observed. In both cases, we need to identify in which period(s) the rule is violated most and under which conditions.

In order to put the analysis in the broader perspective, we retrieved data on the use of the verb rhymes, both in the pairs 'verb form - verb form' (V-V) and 'verb form - non-verb form' (V-non-V) (in any order). Rhymes consisting of more than one word in any rhyming unit (e.g. verbs followed by particles in pairs like *дотяну ли* - *Калигуле* 'if (I) reach - Caligula') are excluded; this also excludes pairs with the subjunctive forms with the particle *бы, б* 'would'. In order to simplify the calculations, rhyming chains which include more than two elements are decomposed into simple pairs in which the lines always rhyme with the first line of the chain. The data is limited by the time of creation from 1801 to 1960, the timeline is binned by 20 years. Texts with the longer timespan of creation which lies astride the bin boundaries, are also excluded from examination. Since not all rhyming pairs have been annotated in the current database (see Footnote 6 above), the normalized frequencies are calculated[9] taking into account the size of subcorpora which only include texts annotated with regard to the rhyming pairs. Lastly, we set a threshold of three or more occurences of a particular rhyming pair in the corpus and two or more authors using the same pair to exclude author's individual choice (cf. the rhyme *обманут* - *устанут* 'deceived - get tired' used only by Bryusov) and possible errors of the automatic rhyme identification. The resulting

[9] in ipm (items per million words). Note that yet another natural way to perform normalization for the poetry data is to weight the occurrences of units per the number of verses (lines) rather than words or tokens. However, the rhymed chains can include three, four, even more than 10 rhyming lines, so weighting these data method per line would be difficult.

dataset consists of 39 319 rhyming pairs, of which 9 172 are of the V-V type and 30 147 are of the V-non-V type.

Figure 5 shows a relatively complicated structure of the verb rhyme distribution over the period of 1801-1960, with peaks at 1841-1860, 1881-1920, and 1941-1960. Should we interpret this as evidence against hypothesis (i)? This is not necessarily so: if we take the verb-to-verb rhymes, there is a general decrease in their normalized frequencies. Three periods can be distinguished: 1801-1860 - 'higher', 1861-1920 - 'middle', and 1921-1960 - 'lower' proportions of the V-V rhyming pairs. At the same time, the number of the V-non-V pairs increases consistently, and it appears to be three other stages in this case: 1801-1820 - 'lower', 1821-1880 - 'middle', and 1881-1960 - 'higher' proportions of the V-non-V rhyming pairs.

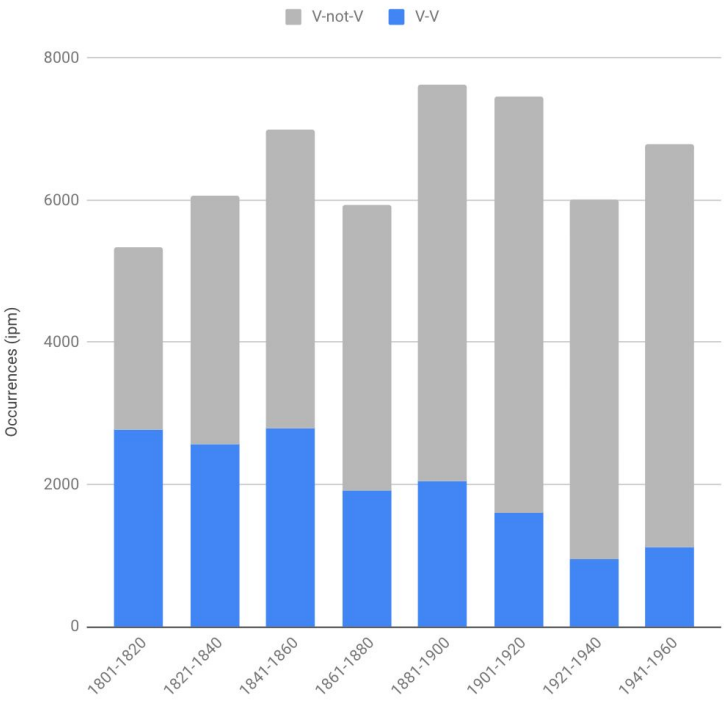

Fig. 5. Distribution of verb rhymes in 1801-1960: all verb forms, verbs rhyming with verbs (V-V) and verbs rhyming with other parts of speech (V-non-V).

Is an instance, the authors of 1841-1860 are rather conservative in terms of the use of the V-V rhymes compared against the previous periods. This is in line with the increased interest in the use of the folk and vernacular motifs in this period, for which the simpler the

more stylised the rhyme is. The V-V rhyme is most actively used by I. S. Nikitin (150 pairs, 5009 ipm[10]), I. P. Myatlev (22 pairs, 4360 ipm), A. A. Fet (116 pairs, 3451 ipm), L. A. Mej (47 pairs, 3123 ipm), N. F. Scherbina (72 pairs, 2873 ipm), A. A. Grigoriev (32 pairs, 2611 ipm). Besides that, the authors of 1841-1860 follow the trend of the previous time period to rhyme more actively verbs with non-verb elements. In addition to those mentioned, the most noticeable authors in this respect include A. N. Apukhtin (97 pairs, 6469 ipm), N. A. Nekrasov (98 pairs, 4052 ipm), A. N. Plescheev (87 pairs, 5896 ipm), and V. G. Benediktov (56 pairs, 4794 ipm).

In the beginning of the Soviet era, in 1921-1940, there is a sharp decrease in the use of the V-V pairs. Still, we can identify authors which use them comparatively frequently: V. V. Nabokov (48 pairs, 1818 ipm), B. Ju. Poplavsky (48 pairs, 1631 ipm), and A. T. Tvardovsky (34 pairs, 1401 ipm), and examine if this pattern correlates with the active use of the V-non-V combinations within the period and the same authors (Nabokov and Tvardovsky are among the top-5 in this respect). The data analysis tool also allows one to follow the distribution of the rhymes of one author across different time periods.

Figure 6 illustrates the contribution of various grammatical forms to the V-V distributional pattern presented in Figure 5. Interestingly, the infinitive forms are underused in 1821-1860 and 1901-1920, i.e. in the periods before the sharp drop of the V-V rhymes[11]. When the drop happens (1861-1880 and 1921-1940), it is followed by decrease in the ratio of the non-past indicative forms. The contribution of the past indicative forms is the largest in the period of 1941-1960, associated with the war and post-war poetic narrative. It would also be useful to compare the form distributions observed within specific time periods to the overall distribution of the grammatical forms in the V-V pairs and in the rhyming zone, in general.

[10] The absolute frequencies are weighted by the size of the corpus of a given author in a given time period.
[11] Note that whereas A. Cantemir cited the feminine infinitive rhymes like *спати* ‘to sleep’ as ‘vile’, by the 19th century the *-ти*(*сь*) forms were replaced by forms with the endings *-ть*(*ся*), *-чь*(*ся*), consistent with the masculine rhyme.

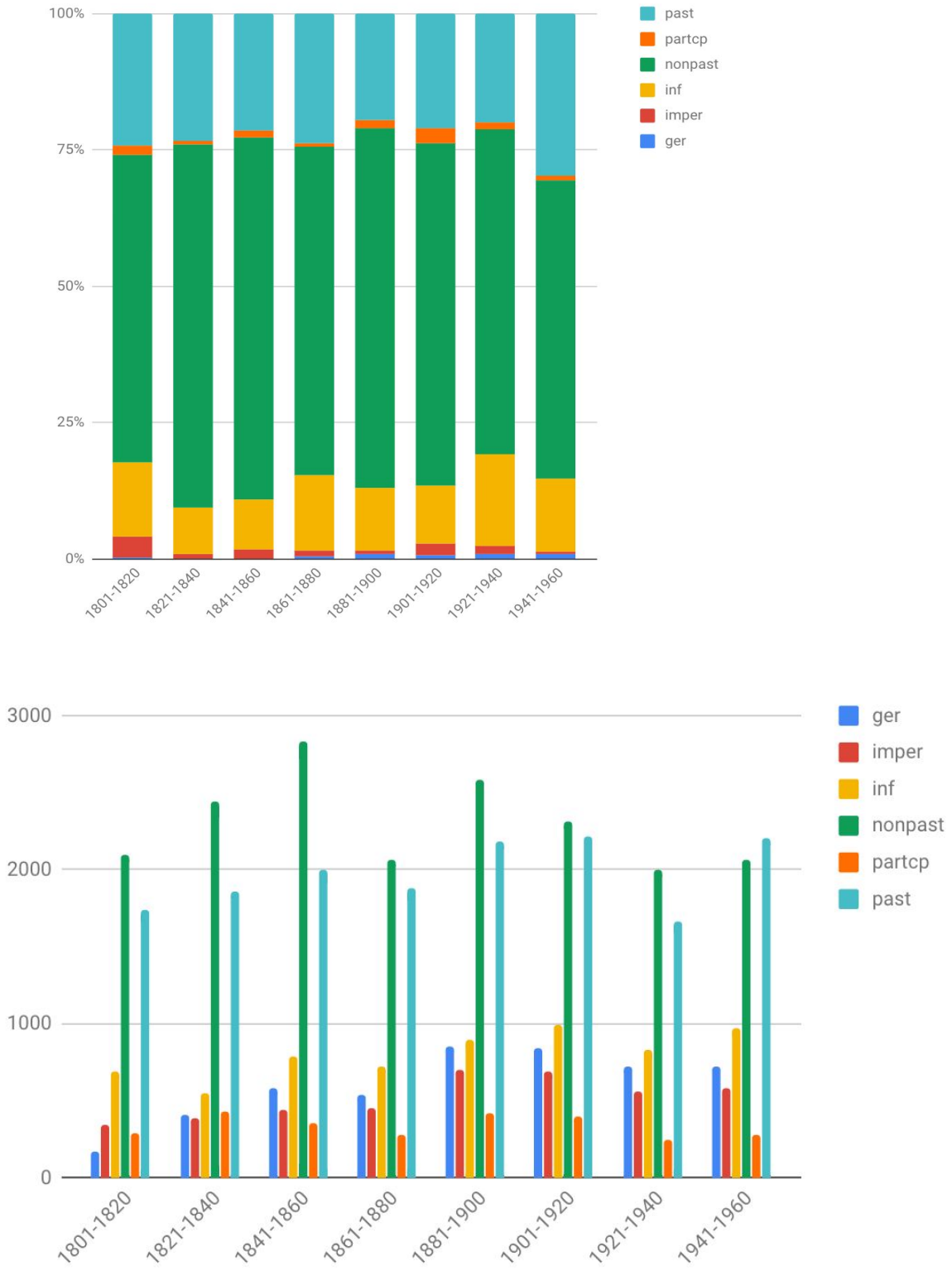

Fig. 6. Distribution of V-V rhymes in 1801-1960: by grammatical forms: (a) 100% stacked column chart and (b) barplot.

In the remaining part of the section we examine word lists and focus more on the V-non-V rhymes. The data analysis tool of the corpus gives a user an option to retrieve the frequency word lists for the category in question, in our case, for the rhyming counterparts of the particular grammatical forms. The list of rhymes with the infinitive includes the pairs *опять — спать* 'again — to sleep' (36 occurrences), *опять — понять* 'again — to comprehend' (25 occurrences). The following rhymes top the list for the non-past forms: *пою — свою* '(I) sing — oneself' (35 occurrences) and *поют — приют* '(they) sing — a refuge' (34 occurrences). Among the rhymes with the gerundive the most frequent forms are *любя — тебя* 'loving — you' (267 occurrences) and *любя - себя* 'loving - myself' (117 occurrences). Among the rhymes with the past indicative forms the most frequent forms are *был — любил* 'was — loved' (45 occurrences) и *любил — сил* 'loved' — 'force' (38 occurrences). Among the rhymes with the imperative the most frequent forms are *прости — пути* 'forgive — paths' (75 occurrences) и *живи — любви* 'live — love' (61 occurrences). The most frequent rhymes with the participle include *давно — дано* 'long ago — given' (35 occurrences) and *дано — одно* 'given — the one' (30 occurrences). Moreover, only the participle *дано* 'given' is attested in this type of rhymes.

It can be seen that verbs which are most frequent in such word lists are those with the 'poetic' meaning related to such topics like life and love (cf. *понять* 'understand', *петь* 'sing', *любить* 'love', *простить* 'forgive', *жить* 'live'). Besides that, the verb *быть* 'be' is frequently used, both as a common verb and as an auxiliary. The top list of the frequent V-V pairs includes a lot of similar verbs (see Table 2). Verbs such as *быть* 'to be', *любить* 'to love', *петь* 'to sing' are the most frequent verbs in both types of rhymes: V-V and V-non-V.

Tables 3 and 4 illustrate the distribution in pairs in which one element is represented by a particular grammatical form of the verb and another - by a particular part of speech. All in all, the verbs rhyme most often with nouns (23594 rhymes (60%), of which 3624 are unique), with verbs themselves (9172 rhymes (23%); 1639 unique), with adverbials (2442 rhymes (6%); 370), and with the nominal pronouns (1704 rhymes (4%); 164 unique). The remaining part-of-speech combinations make up 6% of the data. Thus, the verb-to-verb rhymes are the second among most used and productive forms.

Table 2. Top frequent rhyming verb pairs: V-V (right) and V-non-V (left).

F	V1	V2	F	V	non-V
87	ды̀шит ‘breathe’	слы̀шит ‘hear’	271	любя́ ‘love’	тебя́ ‘you’
66	бу̀дет ‘be’	забу̀дет ‘forget’	187	дыша́ ‘breathe’	душа́ ‘soul’
62	го̀ворѝт ‘say’	горѝт ‘burn’	169	помо̀чь ‘help’	но̀чь ‘night’
60	бы̀ть ‘be’	любѝть ‘love’	147	идти́ ‘go’	пути́ ‘way, path’
56	мо̀жет ‘can’	трево̀жит ‘be disturbing’	137	найти́ ‘find’	пути́ ‘way, path’
49	жѝть ‘live’	любѝть ‘love’	124	спеша́ ‘be in a hurry’	душа́ ‘soul’
49	блѐщет ‘blister’	трепѐщет ‘flutter’	117	любя́ ‘love’	себя́ ‘oneself’
47	бы̀ть ‘be’	забы̀ть ‘forget’	110	звеня́ ‘ring’	меня́ ‘I’
46	бы̀л ‘be’	любѝл ‘love’	104	ѐсть ‘be, eat’	чѐсть ‘honor’
44	бы̀ть ‘be’	жѝть ‘live’	90	зовѝ ‘call’	любвѝ ‘love’
43	ды̀шит ‘breathe’	колы̀шет ‘flutter’	85	о̀тдохну̀ть ‘relax’	пу̀ть ‘way, path’
39	ловлю̀ ‘catch’	люблю̀ ‘love’	83	шутя́ ‘joke’	дитя́ ‘child’
30	бу̀ду ‘be’	забу̀ду ‘forget’	82	мо̀г ‘can’	бо̀г ‘God’
30	бы̀л ‘be’	забы̀л ‘forget’	78	моглѝ ‘can’	землѝ ‘Earth, land, soil’

Table 3. Rhyming pairs: verb forms and part of speech matches (the number of occurrences).[12]

Form	A	ADV*	APRO	CONJ	INTJ	NUM	PART	PR	S	SPRO	V
ger	297	105	108			4		6	2357	1108	51
imper	64	214	63			25	6		2729	54	101
inf	4	347	1		1	40	1		3629	2	1163
nonpast	224	1003	723	19	23	20	76		6818	158	6186
partcp	194	154	55	7			12	2	1043	273	119
past	364	619	48	10		8			7018	109	1552
Total	1147	2442	998	36	24	97	95	8	23594	1704	9172

Table 4. Rhyming pairs: verb forms and part of speech matches (the number of unique pairs).

Form	A	ADV*	APRO	CONJ	INTJ	NUM	PART	PR	S	SPRO	V
ger	66	11	15			1		1	333	55	9
imper	15	35	15			3	2		361	11	20
inf	1	52	1			8			465		208
nonpast	47	135	83	3	5	3	12		1100	30	1054
partcp	47	25	7	1			2		210	55	28
past	66	112	8	3		1			1155	13	320
Total	242	370	129	7	5	16	16	1	3624	164	1639

[12] The category of ADV* includes adverbs, predicatives, and parentheticals (cf. Lyashevskaya, Sharoff 2009).

The infinitives often make up a rhyme with nouns (3629 occurrences; 465 unique pairs), verbs (1163 occurrences; 208 unique pairs), and adverbs (347 occurrences; 52 unique pairs). The non-past forms demonstrate the similar distribution of pairs: most common combinations are those with nouns (6818 occurrences, of which 1100 are unique), verbs (6186 occurrences; 1054 unique), adverbs (1003 occurrences; 135 are unique), and adjectival pronouns (723 occurrences; 83 are unique). The past tense forms most often rhyme with nouns (7010 occurrences; 1155 unique), verbs (1152 occurrences; 320 unique), adverbials (619 occurrences; 112 unique). However, there is a larger proportion of the past forms rhyming with adjectives (364 occurrences, 4% of all past forms). It is likely that the rhyming potential of of adjectives is connected with the combination of the short *l*-adjectives and *l*-forms of the verb (52 of 66 unique forms: *было - мило* 'was - sweet', *была — мила* 'was - sweet', *была — светла* 'was - bright', *заметил - светел* 'noticed - light'). The same applies to the combinations of the short adjectives and short participles. Besides that, the infinitives rhyme with numerals, mostly with *шесть* '6', *пять* '5', and *десять* '10', cf. *есть - шесть* 'eat - six' and *сесть - шесть* 'sit down - six', *спать - пять* 'sleep - five', *повесить - десять* 'hang - 10'). The non-past forms also rhyme with conjunctions and particles.

To sup up, a relatively uniform hierarchy of the rhyming groups is observed, with nouns, verbs, and adverbs being among the most frequent non-verb elements of the pair. In contrast, the gerundives stand out among the rhyming verb forms since they frequently rhyme with the nominal pronouns (*тая - моя* 'harboring - my', *вися - вся* 'hanging - all') and adjectives, but rarely combine with verbs. In general, the observed distribution indicates a high activity of the verb in the rhyming zone and, accordingly, there are no particular limitations on the verb combining with other parts of speech. Obviously, some frequent and stable cliché such as *идти - пути* 'walking - paths' (147 occurrences) and *зови - любви* 'call - love' (92 occurrences) may affect (both positively and negatively) the use of particular grammatical forms and lexemes.

Our study suggests that the implied restriction on the use of the verb rhymes in the Russian poetry has not been supported by the corpus data. Moreover, in spite of Cantemir's interdictions, rhymes that include infinitives are the third most frequent after the non-past and past forms. The fall in the frequency of use of verb rhymes in the beginning of the 20th century could be associated with the language experiments of avant-garde poets rather than

with the influence of the ‘rules’ of the poetry mastering. Going to the hypotheses put forward in the beginning of this section, hypothesis (i) has not been confirmed for the V-V rhymes but it is true for the V-non-V rhymes. The authors avoid the verb-to-verb rhyme more in the later periods than in the beginning of the 1800s, however, they experiment with the V-non-V rhymes more actively over the time. Hypothesis (ii) holds true with respect to the verb rhyme in general: there are periods with relatively more limited use of the verb rhymes which alternate with the periods of their expansion. The detailed corpus statistics allows a user to identify authors which use such rhymes relatively more frequently compared to the others, and to examine grammatical forms, parts of speech and words which are more actively engaged into the verb rhyming.

5. Conclusions

This paper describes a new database assembled from the Russian Poetry Corpus, which is a part of the Russian National Corpus. The database contains more than 13 million tokens with a few layers of linguistic, versological, and metatextual annotation to facilitate interdisciplinary research of the Russian poetry. The designers of the database also developed tagging of sublexical (phonological) and supralexical syntactis units for the quantitative analysis.

As a demonstration of the research potential of the database, we presented a few case studies illustrating the suitable techniques and methodology of the computer-assisted analysis. Using the flexible toolkit of the frequency database, a scholar can define subcorpora for a wide range of research goals and support qualitative and contrastive analysis with quantitative data drawn from a large-scale corpus.

As an exemplary study of diachronic constastive research, we compared the use of adjectives in the Golden Age (1811–1840, the size of the subcorpus is over 1,270,000 tokens) and the Silver Age (1901–1917, the size of the subcorpus is about 1, 619, 228 tokens). Although the proportions of adjectives within each subcorpus does not demonstrate significant discrepancy (10.8% in the Golden Age and 12.6% in the Silver Age), the adjectival lexicon of the Silver Age is almost twice as large than that of the Golden Age: cf. 6,421 unique lexemes in the Golden Age and 11,777 lexemes in the Silver Age. For this comparative analysis, we compiled a list of adjectives that belong to the top of the word

frequency lists within each period. The difference between the two periods is noticeable. While the most frequent adjectives of the Golden Age are the lexemes referring to mood, emotions, and feelings ((*мирный* 'peaceful', *мрачный* 'grumpy'), as well as judgments about behavior (*добрый* 'kind', *приятный* 'pleasant'). Meanwhile, the Silver Age poets favor adjectives denoting nature elements and objects, they tend to explicit strange and unusual states and qualities.

As another exemplary study of lexical diversity, we explored the use of color adjectives in the Silver Age applying method of Correspondence Analysis which offers visualization of multidimensional frequency associations of lexemes and authors. This method supports contrastive stylistic analysis and identifies similarities between different poets. In this case, we defined subcorpora of authors traditionally seen as key figures of the Silver Age such as Aleksandr Blok, Konstantin Balmont, Anna Akhmatova, Nikolay Gumilev, Marina Tsvetaeva. This study has revealed the stylistic differences in individual poetic lexicon and demonstrated that despite the inner influences within poetic schools, poets' color preferences can be very different.

The third case presented a multi-dimensional study at the intersection of the rhyme (versological annotation), parts of speech, grammatical forms, words (linguistic annotation), authorship and time of creation (metatextual annotation of the corpus). We examined a mutual distribution of such categories using contingency tables and barplots. While studying the distribution over the time, we made systematic use of the methods of data aggregation and normalization against the size of subcorpora and the method to deal with the partial coverage of some data in the database. The study has demonstrated the variation in the distribution of the verb rhymes across different time periods (binned per 20 years) which is followed by the minor variation in the distribution of grammatical verb forms and by the individual author's preferences.

Currently, the data analysis tool allows a user to retrieve various type of datasets (frequency lists with raw and normalized frequencies, contingency tables, corpus datasets in the long format) defined by the combination of one to four chosen parameters (annotation categories). In addition, it provides basic descriptive statistics on the data a number of visualizations (charts) for the exploratory analysis. The further line of development is to add

more sophisticated functional techniques to perform statistical tests, make data validation and to add a number of scenarios for some standard data analysis pipelines.

References

Baranov, V. A. (2018). Statistical analysis of the Slavonic paraenesis by Ephrem the Syrian (on three electronic copies of the 13–14th centuries from the Manuscript corpus). Journal of Siberian Federal University. Humanities & Social Sciences 8 (11), 1211-1228

Bonch-Osmolovskaya A., Orekhov B. Nekotorye primenenija korpusnykh metodov k naivnoj poezii [Corpus-based Approaches to Naive Poetry]. Statji na sluchaj: sbornik v chest' 50-letija R. G. Lejbova [Articles on Case: Collection in Honor of the 50th Anniversary of R. G. Leibov]. Electronic publication (http://www.ruthenia.ru/leibov_50/article_b-osm_orexov.html).

Gasparov M. Ocherk istorii russkogo stikha [A History of the Russian Verse]. Moscow: Fortuna Limited, 2000.

Gasparov M. (2005). Sovremennyj russkij stikh: metrika i ritmika [Contemporary Russian Verse: Metrics and Rhythmics]. Moscow: Nauka.

Grishina E., Korchagin K., Plungian V., Sichinava D. (2009). Poeticheskij korpus v ramkakh Nalsional'nogo korpusa russkogo yazyka: obshchaya struktura i perspektivy ispol'zovaniya [The corpus of poetry within the Russian National Corpus: a general outline and perspectives of use]. Natsional'nyj korpus russkogo yazyka: 2006–2008. Novye rezul'taty i perspektivy [The Russian National Corpus: 2006–2008. New results and prospects]. St. Petersburg: Nestor-Istoriya Publ.

Jacobs A. M. (2018). The Gutenberg English Poetry Corpus: Exemplary Quantitative Narrative Analyses. In: Frontiers in Digital Humanities (5).

Kilgarriff A. (2009). Simple maths for keywords. In: Proceedings of the Corpus Linguistics Conference. Liverpool, UK, 2009.

Kizhner I., Terras M., Manovich L., Orekhov B., Bonch-Osmolovskaya A., Rumyantsev M. (2018). 'The history and context of the Digital Humanities in Russia'. In: Proceedings of DH 2018, Mexico City, June 26-29, 2018.

Kustova G. I., Lashevskaja O. N., Paducheva E. V., Rakhilina E. V. (2005). 'Semanticheskaja razmetka leksiki v Nacional'nom Korpuse Russkogo Yazyka: principy, problemy, perspektivy [Lexico-semantic annotation in the National Corpus of Russian: Principles, problems, prospects].' In: Nacional'nyj korpus russkogo jazyka: 2003-2005 [Russian National Corpus: 2003-2005]. Moscow. Pp.155-174.

Kuzmenko E., Orekhov B. (2016) 'Geography of Russian Poetry: Countries and Cities Inside the Poetic World'. In: Digital Humanities 2016: Conference Abstracts. Jagiellonian University & Pedagogical University, Kraków, pp. 830-832.

Lyapin S. E. (1997). ‘K demifologizacii ritmiki russkigo 4-stopnogo jamba (preimuschestvenno na materiale odicheskogo stikha Derzhavina) [To the demythologization of the rhythms of the Russian 4-foot iamb (mainly on the material of Derzhavin's odic verse)].’ Philologica 4.

Lyashevskaya O., Sharoff S. (2009). Chastotnyj slovar' sovremennogo russkogo jazyka (na materialakh Nacional'nogo Korpusa Russkogo Jazyka) [A Frequency Dictionary of Contemporary Russian based on the Russian National Corpus data]. Moscow: Azbukovnik.

Medyankin N., Droganova K. (2016). Building NLP Pipeline for Russian with a Handful of Linguistic Knowledge. Proceedings of the Workshop “Computational linguistics and language science”(CLLS) Moscow 2016, CEUR Workshop Proceedings (in print, manuscript is available at http://web-corpora.net/wsgi3/ru-syntax/static/downloads/Medyankin_Droganova_CLLS_2016 .pdf).

Orekhov B. (2015). ‘Eshche raz ob issledovatel’skom potenciale poeticheskogo korpusa: metr, leksika, formula [Once again about the Research Potential of the Poetic Corpus: Meter, Vocabulary, Formula].’ Trudy instituta russkogo jazyka im. Vinogradova, No. 6: 449–463.

Orekhov B. (2016). ‘Stikhi i proza cherez prizmu distributivnoj semantiki [Poetry and Prose in the Light of Distributional Semantics].’ In: Ostrova ljubvi BorFeda: Sbornik v chest’ 90-letija Borisa Fedorovicha Egorova [A Collection of Articles in Honor of the 90th Anniversary of Boris Fedorovich Egorov]. St. Petersburg. Rostok.

Orekhov B. ‘Objasnimy li oshibki V. A. Zhukovskogo s pomoshchju analiza dannykh? [Can we explain V. A. Zhukovskij’s Mistakes by Means of Data Analysis?].’ In: Cifrovaja gumanitaristika: resursy, metody, issledovaniya: Materialy Mezhdunar. nauch. konf. (Perm’, 16–18 maja 2017). Perm’: Izdatel’stvo permskogo gosudarstvennogo issledovatel’skogo universiteta, 2017.

Samojlov D. (2005). Kniga o russkoj rifme [A Book on the Russian Rhyme]. Moscow: Vremya.

Scott M., Tribble C. (2006). Textual patterns: Key words and corpus analysis in language education. Amsterdam: Benjamins.

Taranovsky K. F. (1971). ‘On the rhythmic structure of Russian double-complex sizes’, Poetics and stylistics of Russian literature: In memory of Academician Viktor Vladimirovich Vinogradov, Leningrad, 420-429.

Appendix

Table A. Adjectives more frequent in the Silver Age than in Pushkin’s Age

Adjective	F_1 1811–1840	F_2 1901–1917	F_2/F_1	Adjective	F_1 1811–1840	F_2 1901–1917	F_2/F_1
загробный	0,000044	0,001131	25,8	*влюбленный*	0,000680	0,001893	2,8
певучий	0,000066	0,001042	15,8	*серебряный*	0,000724	0,002001	2,8
лунный	0,000263	0,002316	8,8	*каменный*	0,000519	0,001426	2,7
жгучий	0,000154	0,001116	7,3	*красный*	0,002275	0,006181	2,7
желтый	0,000439	0,002955	6,7	*зимний*	0,000578	0,001569	2,7
серый	0,000615	0,003634	5,9	*прозрачный*	0,000885	0,002400	2,7
алый	0,000717	0,003747	5,2	*пьяный*	0,000827	0,002198	2,7
звездный	0,000402	0,002055	5,1	*сухой*	0,000702	0,001839	2,6
тонкий	0,000790	0,003486	4,4	*людской*	0,000607	0,001564	2,6
узкий	0,000322	0,001318	4,1	*голубой*	0,001836	0,004647	2,5
солнечный	0,000505	0,001996	4,0	*душный*	0,000432	0,001087	2,5
осенний	0,000732	0,002857	3,9	*медный*	0,000468	0,001156	2,5
розовый	0,000549	0,002124	3,9	*усталый*	0,001054	0,002596	2,5
весенний	0,000849	0,003167	3,7	*зыбкий*	0,000468	0,001136	2,4
пыльный	0,000293	0,001062	3,6	*горячий*	0,000819	0,001947	2,4
белый	0,004053	0,014510	3,6	*вечерний*	0,001829	0,004288	2,3
синий	0,001749	0,005925	3,4	*пестрый*	0,000571	0,001308	2,3
маленький	0,000746	0,002503	3,4	*странный*	0,001193	0,002670	2,2
снежный	0,000585	0,001918	3,3	*мудрый*	0,000746	0,001662	2,2
тусклый	0,000388	0,001264	3,3	*теплый*	0,000739	0,001559	2,1
девичий	0,000337	0,001077	3,2	*черный*	0,004990	0,010419	2,1
ласковый	0,000593	0,001883	3,2	*темный*	0,004968	0,010297	2,1
горный	0,000461	0,001372	3,0	*далекий*	0,002334	0,004824	2,1
огненный	0,000593	0,001755	3,0	*вещий*	0,000680	0,001396	2,1
зеленый	0,001983	0,005704	2,9	*былой*	0,001346	0,002700	2,0
лесной	0,000644	0,001824	2,8	*золотой*	0,003702	0,007400	2,0
вешний	0,000527	0,001470	2,8	*старый*	0,003431	0,006692	2,0

Table B. Adjectives less frequent in the Silver Age than in the Golden Age

Прилагательное	**Частота 1: 1811–1840**	**Частота 2: 1901–1917**	**$Ч_2/Ч_1$**	**Прилагательное**	**Частота 1811–1840**	**Частота 1901–1917**	**$Ч_2/Ч_1$**
напрасный	0,001068	0,000526	0,5	*опасный*	0,001258	0,000384	0,3
прямой	0,001266	0,000620	0,5	*роскошный*	0,001492	0,000447	0,3
мятежный	0,001602	0,000777	0,5	*отрадный*	0,001427	0,000423	0,3
готовый	0,004316	0,002080	0,5	*благородный*	0,001083	0,000320	0,3
скромный	0,001310	0,000629	0,5	*несчастный*	0,002502	0,000723	0,3
твердый	0,001097	0,000521	0,5	*достойный*	0,001953	0,000561	0,3
величавый	0,001222	0,000575	0,5	*гробовой*	0,001273	0,000359	0,3
невинный	0,001990	0,000929	0,5	*златой*	0,003460	0,000959	0,3
летучий	0,001024	0,000477	0,5	*роковой*	0,005428	0,001495	0,3
благой	0,001156	0,000536	0,5	*свирепый*	0,001002	0,000270	0,3
пленительный	0,001295	0,000595	0,5	*резвый*	0,001529	0,000403	0,3
юный	0,005275	0,002331	0,4	*коварный*	0,001478	0,000384	0,3
прекрасный	0,008223	0,003501	0,4	*славный*	0,003263	0,000806	0,2
прежний	0,003343	0,001401	0,4	*бурный*	0,002524	0,000600	0,2
шумный	0,002707	0,001116	0,4	*вдохновенный*	0,001690	0,000379	0,2
умный	0,001346	0,000546	0,4	*мрачный*	0,003877	0,000846	0,2
многий	0,001141	0,000462	0,4	*русский*	0,005363	0,001156	0,2
волшебный	0,002956	0,001175	0,4	*бранный*	0,001083	0,000221	0,2
грозный	0,004975	0,001913	0,4	*отважный*	0,001097	0,000221	0,2
молодой	0,007916	0,003009	0,4	*пылкий*	0,001661	0,000315	0,2
смиренный	0,001544	0,000580	0,4	*удалой*	0,001346	0,000246	0,2
унылый	0,003212	0,001165	0,4	*храбрый*	0,001105	0,000197	0,2
звучный	0,001039	0,000374	0,4	*ужасный*	0,004082	0,000683	0,2
громкий	0,001251	0,000447	0,4	*приятный*	0,001588	0,000226	0,1
богатый	0,002590	0,000920	0,4	*прелестный*	0,003234	0,000428	0,1
чудесный	0,002136	0,000738	0,3	*любезный*	0,002158	0,000266	0,1
добрый	0,004726	0,001559	0,3	*сердечный*	0,002305	0,000270	0,1
беспечный	0,001492	0,000492	0,3	*хладный*	0,003358	0,000364	0,1
чудный	0,002422	0,000787	0,3	*возвышенный*	0,001266	0,000093	0,1
счастливый	0,006555	0,002124	0,3	*младой*	0,012159	0,000315	0,0

Table C. The frequency of the adjectives *больной* ‘sick’ and *безумный* ‘insane’ by decades (delta)

	1800-e	1810-e	1820-e	1830-e	1840-e	1850-e	1860-e	1870-e	1880-e	1890-e	1900-e
больной	–24	–32	–50	–26	+83	+77	+39	+56	+112	+72	+71
безумный	–27	–33	+44	+16	+40	+29	–16	0	+79	+76	+139

Contact details:

Olga Lyashevskaya
National Research University Higher School of Economics (Moscow, Russia), School of linguistics, Professor;
E-mail: olesar@yandex.ru

Any opinions or claims contained in this Working Paper do not necessarily reflect the views of HSE.