Linguistic and Stylistic Parameters for the Study of Literary Language in the Corpus of Russian Short Stories of the First Third of the 20th Century*

Tatiana Sherstinova[1,2] tsherstinova@hse.ru

Gregory Martynenko [1] g.martynenko@spbu.ru

[1] National Research University Higher School of Economics,
[2]Saint Petersburg State University,
Saint Petersburg, Russian Federation

Abstract

One of the important tasks of creating the Corpus of Russian Short Stories of the first third of the 20th century is to identify and describe the changes that took place in the Russian language and in stylistics of Russian literature in the chain of dramatic events of the World War I, the February and October Revolutions, and the Civil War. The essential principle for creating the corpus is an attempt to include in the database literary texts of the maximum number of authors who wrote stories in 1900–1930. The article describes the principles of writers and text selection for the annotated subcorpus containing stories of 300 Russian prose writers and considers the list of linguistic and stylistic parameters proposed for studying the language of literary texts in synchrony and diachrony.

Keywords: *stylometrics, Russian literature, short story, frequency lists, POS, literary corpus, literary system*

1 Introduction

The Corpus of Russian Short Stories of the first third of the 20th century is currently being developed in St. Petersburg State University in cooperation with National Research University Higher School of Economics, St. Petersburg [Martynenko et al., 2018a; 2018b]. When developing corpus conceptions, we relied on the notion of literary system proposed by an outstanding representative of the Russian formal school Yury N. Tynyanov. Tynyanov suggested the idea of systemic nature of fiction and proposed to distinguish synchronic and diachronic literary systems. By synchronic systems he understood the collection of works of a given literary

period, and by diachronic systems — the sequence of successive synchronic ones [Tynyanov, 1929]. If to consider texts as a statistical population [Čebanov, Martynenko, 1998], each literary system can be characterized through a set of statistical variables, which are used for two purposes: to solve taxonomic tasks (diagnostics, clustering, typology, etc.) and to build theoretical models describing a distribution of texts and their units through a system of stylistic variables [Martynenko, Sherstinova, 2000].

Tynyanov's approach implies the necessity to study literary texts of the maximum number of writers who wrote in a given historical period. Compiling the Corpus of Russian short stories, we try to adhere to this principle to the extent possible [Martynenko, Sherstinova, 2018]. This method makes the research more objective, as apart from well-known and outstanding writers, the large number of second-rate authors are also involved into consideration. Thereby we achieve better literary representation of different aspects of social and cultural life, as well as of language and stylistics diversity [Martynenko, 2019a].

Besides implementation of Tynyanov's idea, our other important goal is to study the changes that occur in the language at the crucial historical moments. Thus, this research is made in the framework of the project aimed at studying the complex of linguistic and stylistic variables in dynamics during the first three decades of the 20th century in order to identify and describe the changes that occurred in Russian in the chain of dramatic events of the World War I, the February and October Revolutions, and the Civil War. To fulfill this task we consider Russian literature in the period of 1990–1930.

Obviously, any revolutionary period cause changes in the usual way of life and existing social relations; it leads to transformations both of behavior stereotypes and of common system of values. Any significant social turbulence cannot but affect language changes. The revolutionary events that took place in Russia over 100 years ago dramatically influenced the Russian language. Thus, a huge amount of out-of-date vocabulary was replaced by new words reflecting new concepts and ideas, many words from past acquired either new meanings or new connotations. Significant changes took place in stylistics too, revealing themselves by transformation of generally accepted speech structures — the functional frequencies of many lexical units were changed, the set of frequency collocations were revised, new syntactic patterns appeared, etc.

Some of the linguistic shifts that occurred in the revolutionary period are noticeable by the unaided eye, other changes are more latent in nature. However, in order to become aware of the magnitude of language transformations of different kinds, it is necessary to apply strict quantitative methods, to process the representative volume of language data at several linguistic levels, and to compare the adjacent chronological periods in dynamics. Only in the result of such an analysis will it be possible to say with certainty to what extent one of the most dramatic periods of Russian history influenced the transformation of Russian, what language levels were affected in the first place, and what is the real share of the changes that took place [Martynenko et al., 2018a].

In this paper we briefly discuss our approach for text selection for the annotated subcorpus and consider the list of appropriate linguistic and stylistic parameters which will be used for text analysis.

2 The Corpus of Russian Short Stories: the Principles of Writers and Texts Selection

In recent years, more and more studies of literary texts using computer or corpus methods have appeared and increasing numbers of literary corpora have become available [Fischer-Starcke, 2010; Balossi, 2014; Mahlberg et al., 2016; etc.]. Literary texts are necessarily included in all major national corpora, as fiction prose is traditionally used for linguistic studies [Sinclair, 2004]. However, it should be mentioned that in the most of computational studies of literature, long literary forms (novels) prevail; dramatic pieces are frequently considered as well [Zyngier, 2008; Archer et al., 2009; Fischer et al., 2018; Skorinkin, Fischer, 2018], while research on small literary forms is less common.

The Corpus of Russian Short Stories — as the name suggests — is designed to be homogeneous in genre. We decided to choose this short literary form, because it is the story is the most common genre of fiction. Being very popular among prose writers, it covers almost all literary schools and involves almost all writers. Besides, short stories go through the publishing cycle much faster than the larger prosaic works. Therefore, we can say that the story, as a genre, performs a prospecting function and even is working in a proactive manner, sensitively capturing the changes in public consciousness and responding to them [Martynenko et al., 2018b].

For creating the corpus, the first important task was to obtain the most complete list of authors who wrote stories in 1900–1930. It was decided to include in the list the writers whose literary heritage is represented by at least one story. At the same time, the goal was to involve in the corpus not only capital writers living in Moscow and St. Petersburg (Petrograd, Leningrad), but also regional writers who wrote in Russian and lived anywhere on the territory of the Russian Empire (until 1917), and later on the territory of the Russian SFSR and the USSR. However, taking into account the project's aim of studying language changes, it was decided that writers who emigrated after Revolution should not be included in the systems analysis after their emigration. Also, at this stage it is not planned to involve in research literary works by children's writers [ibid.].

Firstly, the list of writers was being formed on the basis of literary encyclopedias and bibliographic dictionaries [KLE; Muratova, 1963; Russian Writers; etc.]. Then it was decided to catalogue the works presented in electronic catalogues of Russian National Library, which are tagged as a story. As a result, a database of all stories published as separate brochures in 1900–1930 was formed. In addition, the famous literary magazines and anthologies of this historical period (Apollon, Shipovnik, Niva, Ogonyok, Novy Mir, and many others) as well as some other Internet resources were selectively catalogued [Sovlit; 3500 Texts; etc.]. As a result, a preliminary list of writers' names for the given time period was obtained, which currently comprises an impressive number of 2800 personalities. However, this list is still not complete, and we proceed the cataloging process. In parallel with compilation of authors' list, we search for electronic versions of texts in open sources (if available) and we have to digitize texts for less-known and forgotten authors. Currently, the digital collection of short stories comprises more than 4500 texts.

Since the number of Russian writers in the first third of the 20th century turned out to be much bigger than it was expected, in the first place we decided to annotate the subcorpus for 300 writers (one story for each writer). The selection of 300 stories was made to present evenly three historical periods:

1) **Early 20th century (1900–1913): 100 stories.** The writers's list includes: *Vladimir Korolenko, Leo Tolstoy, Anton Chekhov, Alexandr Kuprin, Valery Bryusov, Zinaida Gippius, Evgeny Chirikov, Boris Zaytsev, Ignaty Potapenko, Aleksey Remizov, Lydia Zinovieva-Annibal, Leonid Andreyev, Konstantin Balmont, Boris Savinkov, Sergey Gusev-Orenburgsky, Fedor Sologub, Arkady Averchenko, Alexander Grin, Andrei Bely, Teffi, Yevgeny Zamyatin and others.*

2) **World War I and Revolutions (1914–1922): 100 stories**

2.1. World War I before revolution (1914–1916): 50 stories

2.2. Revolutions and the Civil War (1917–1922): 50 stories

The writers's list includes: *Mikhail Kuzmin, Boris Pilnyak, Ivan Bunin, Konstantin Trenev, Alexander Blok, Panteleimon Romanov, Lydiia Seifullina, Vikenty Veresaev, Mikhail Zoshchenko, Veniamin Kaverin, Valentin Kataev, Alexander Kuprin, Alexandra Kollontai, Ivan Bunin, Maxim Gorky, Alexander Serafimovich and others.*

3) **New time, or the Early Soviet Period (1923–1930): 100 stories.**

The writers's list includes: *Alexander Belyaev, Arkady Gaidar, Leonid Dobychin, Panteleimon Romanov, Aleksey Chapygin, Sergey Sergeyev-Tsensky, Vyacheslav Shishkov, Alexander Fadeyev, Mikhail Bulgakov, Konstantin Paustovsky, Leonid Leonov, Mikhail Prishvin, Aleksey Tolstoy, Mikhail Sholokhov, Olga Forsh, Marietta Shaginyan and others.*

As for the choice of a story for each of the writers, texts were randomly selected, however we tried to balance both well-known short stories, which are included in different anthologies — like Mikhail Bulgakov's A towel with a rooster (Notes of a young doctor), Easy Breathing by Ivan Bunin, Alexander Malyshkin's A Train to the South, The Viper by Aleksey Tolstoy, Yuri Slezkin's The Cucumber Queen, etc. — as well as less known stories.

In addition, we include into selection 10 additional stories by Alexander Serafimovich, Yevgeny Zamyatin, Alexander Kuprin, Ivan Bunin, Maxim Gorky, Aleksey Chapygin, Panteleimon Romanov, Yuri Slezkin, and Vikenty Veresaev as these authors were actively writing during all three historical periods in concern, so it worth seeing how their style and language have changed through years.

In the next sections we consider the list of linguistic and stylistic parameters for synchronic and diachronic study of literary texts that are supposed to be appropriate for the given research.

3 The Approaches for Creating the List of Parameters

In linguistics, there are two main approaches to the selection of essential variable used for solving various tasks associated with processing of large text data. The first one is a technocratic approach, when an extended list of variables (or attributes) is input into the system. Then, one of statistical algorithms for reducing the number of variables is used with the aim to reveal the most essential features of the sample. The second approach may be called a thoughtful one and it is based on the knowledge of the basic mechanisms of language and text formation. The second method allows to identify structurally significant features on the base of expert estimates. However, it does require a high level of philological competence. Data processing of the Corpus of Russian Short Stories assume the implementation of both approaches with an assessment of the effectiveness of each of them.

Taking into account the important goal to track the changes of the language during 30 years under study, it seems obvious to start with the parameters, which may refer to the

feature list of the language of revolutionary prose, as it is the revolution, that caused the most of language dramatic changes. Therefore, first, on the basis of numerous publications devoted to a qualitative analysis of the features of the "language of revolutionary era" and the influence of revolutionary events on the language [Jacobson, 1921; Selishchev, 1928; Polivanov, 1931; etc.], a list of linguistic and stylistic parameters was formed, according to which the most significant transformations of the Russian language in the revolutionary and post-revolutionary period were noted.

Then, based on the current achievements of computational and mathematical linguistics and on that of the corpus developers (e. g., [Martynenko, 1988; 2019a; Martynenko, Fomin, 1989; Martynenko, Sherstinova, 2000; Martynenko, Martinovich, 2003; Grebennikov, 1998; 2007; Kuznetsov, Skrebtsova et al., 2019]), a wider list of linguistic and stylistic parameters aimed at describing the fiction language in synchrony and diachrony was created, and a list of working statistics in nominal, quantitative and ordinal scales was determined. The list presented in this article would be considered as preliminary. Some of the variable are already being calculated, the others are still in a planning state. At the initial stage of project implementation, it is supposed to use first those algorithms that have been tested in previous studies. These are primarily the methods of statistical lexicography (associated with building frequency dictionaries and studying their properties) and syntactic studies, which are considered to be the most effective in the study of fiction [Martynenko, 2019a].

4 Traditional Conceptual Approach: Some Features of Revolutionary and Early Soviet Prose Language

Numerous publications concerning peculiarities of the language in the revolutionary historical period [Barannikov, 1921; Jacobson, 1921; Rempel, 1921; Kartsevsky, 1923; Vinokur, 1923; Selishchev, 1928; Polivanov, 1931; Granovskaya, 2005; etc.] note primarily lexical changes. The other linguistic levels — phonetic, morphological and syntactic — are usually considered to be more conservative and not amenable to the influence of organized management [Polivanov, 1931], therefore, they commonly remain without any detailed analysis. On the lexical level, the following lexical features are normally noted: 1) abbreviations, 2) toponymic changes, 3) neologisms, 4) bureaucratese, formulaicity, and phraseology, 5) vulgar tongue, popular speech and slang words. Most of these phenomena can be studied by means of frequency list analyses (see below). Let us make some comments in concern:

1. Abbreviations. It is planned to obtain a list of the most frequent abbreviations, to determine the share of abbreviations in texts, as well as the number of them, which are still understandable and relevant so far.

2.Toponymic changes. Renaming geographical names is aimed at creating a new reality. From the point of view of language development, it is interesting to see how quickly the language responds to changes initiated by the authorities. Knowing the date of the official toponymic change, it is worth seeing how fast the new name takes root in literary works.

3. Neologisms and new word meanings. During periods of significant social changes, the emergence of a huge mass of neologisms is quite understandable, as new words are required to label new realities. For the same reason, in such time periods one can observe the development of new word meanings. Both of these phenomena can be analyzed with the use of frequency lists.

As for neologisms, it seems relevant to trace the most common word-forming patterns.

Currently, no continuous morphemic annotation of texts is planned. Nevertheless, a selective analysis of individual word formation constructions and their shares can be calculated basing on word frequency lists, and the identification of the most frequent models of neologisms formation in the revolutionary and early Soviet era will be done.

4. Bureaucratese, Formulaicity, and Phraseology. Set phrases consisting of two or more words usually fall outside the scope of traditional frequency lists analysis. However, these typical features of revolutionary and post-revolutionary language can be studied by means of n-grams calculation [Sherstinova, 2018].

5. Vulgar Tongue, Popular Speech and Slang Words. Due to the activation of broad masses of people and the cardinal change of the role, which people from the lowest social classes started to play in society, vulgarisms, slang words and vernacular expressions began to penetrate the language. Certainly, these are reflected in literature. Thus, the evident feature of the literature of this period is its ever-growing democracy, caused by involving huge human masses to literary work. In stories and novels, multivarious folk speech can be heard — albeit being literary, and albeit being stylized — it was very unusual for the reader brought up on classical literature [Martynenko, 2019b, p. 396]. As in a number of cases described above, frequency word lists will be used for their analyses.

5 Lexical Analysis and Frequency Word Lists

In the study of individual writers' lexical systems — as well as of the totality of works belonging to a particular historical period — the theory of distributions, mainly the rank ones, will be used. Then, these distribution are subject to their subsequent parameterization using rank-frequency (for example, rank averages [Martynenko, Fomin, 1989]) and ordinal statistics (e. g., quantiles), which have shown good effectiveness in earlier studies [Martynenko 2019a]. In the previous section, the importance of lexical analysis for solving the given task was already noted. The main approach for lexical analysis in this research is the method of compiling and measuring frequency dictionaries [Tuldava, 1986; Alekseev, 2001; Grebennikov, 2007; Popescu, 2009; Shaykevich, 2015; etc.].

Statistical measures of frequency word lists are the following:

- Number of units (*tokens*)
- Number of units (*types*)
- Number of single-used words (*hapax, lexical richness*)
- Lexical density
- Lexical diversity coefficient [Voronchak, 1972; Tuldava, 1977]
- Standardized diversity index (TTR_{St}*: Type / Token Ratio*) [Baker et al., 2006].

In addition to this traditional set of variables we plan to introduce an extended list of statistics. Thus, in [Martynenko, Martinovich, 2003], a set of parameters suitable for studying communicative-thematic fields displayed in the form of frequency word lists is discussed and a complex of statistical parameters is formed that would reflect the system properties of these frequency lists in a concise form, for example: diversity vs. limitation of diversity, concentration vs. scattering, stability vs. instability, homogeneity vs. heterogeneity, etc. For detailed description of each parameter see [ibid]:

Variables in the nominal scale:

- Mode (M_o)
- Dictionary Volume (n)

• Maximum Frequency (F_{max})
• Entropy (E)
• Maximum Entropy (E_{max})
• Degree of order ($O = E \,/\, E_{max}$)

Variables in the quantitative scale:

• Arithmetic mean (F_{ave})
• Geometric mean (F_{geom})
• Standard Deviation (σ)
• Mean linear deviation (D_f)
• Variation coefficient of standard deviation (V)
• Variation coefficient of the mean linear deviation (V_{Df})
• Diversity coefficient (K).

Variables in the ordinal (rank) scale [Martynenko, 2017]:

• Rank mean (R_{ave}) [Martynenko, Fomin, 1989]
• Standard Deviation (σ)
• Coefficient of variation (V_r)
• Median (M_{er})
• Golden Ratio (G_r)
• Mean Deviation (D_r)
• Coefficient of variation for Dr (V_{Dr})
• Concentration Index (γ).

Besides word frequency lists it is worth compiling that of POS (see Fig. 1) and grammar word forms, that may be also processed in the similar way.

6 Other Linguistic and Stylistic Variables and Statistics

In this section we list the variables and statistics which are planned to be used in the Corpus of Russian Short Stories.

General Quantitative Text Variables

• The number of characters.
• The number of words.
• The number of sentences.
• The number of paragraphs.
• The number and the share of punctuation marks. POS Distribution.
• Frequency list of POS and their shares (an example is shown on Fig. 1).
• The number and the share of prepositions.
• The number and the share of prepositions in the initial position of the sentence.
• The number and the share of co-ordinate conjunctions.
• The number and the share of co-ordinate conjunctions in the initial position of the sentence.
• The number and the share of subordinative conjunctions.
• The number and the share of subordinative conjunction in the initial position of the sentence.
• The number and the share of modal words (*mozhno can, nuzhno need, dolzhen should/must*, etc.).
• The number and the share of particles.

Word sizes and positions

• The average size of words in text.

• The average word size in the first, second, third position, etc. in the sentence

• The average word size in the last position in one-word, two-word, etc. sentences.

• The dynamic rank average of word size in a sentence (here, the independent variable is the number of words in a sentence, and the dependent variable is the average word size in sentences of different length).

Sentences and paragraphes

• The average number of words in a sentence.

• The average number of sentences in a paragraph.

• The average length of the sentence in the first, second, third, etc. position in the paragraph consisting of one, two, three, etc. sentences.

• The dynamic rank average length of sentences in paragraphs of different sizes.

• The dynamic rank average size of paragraphs. Measures of syntactic complexity.

• The width of the tree in a root node (i. e., the number of subordinate members).

• The number of left and right subordinate members in a root node.

• Tree symmetry index — the ratio of the left subordinate members to the right ones (Symmetry I).

• The ratio of the left subordinate members to the right ones relatively to the root node measured in word numbers (Symmetry II).

• The height of the tree (i. e., the maximum number of sequentially subordinate nodes).

• The maximum length of left-branching branches in a sentence.

• The maximum length of right-branching branches in a sentence.

• The ratio of left-branching subordinates' length to the right one.

• The distance degree — the maximum number of nodes between two syntactically related

POS	Abs. Freq.	%
V	245	20,6
S	235	19,7
SPRO	185	15,5
CONJ	111	9,3
PR	109	9,1
APRO	78	6,5
A	75	6,3
ADV	62	5,2
PART	61	5,1
ADVPRO	25	2,1
NUM	3	0,3
INTJ	2	0,2
ANUM	1	0,1

Figure 1: Frequency list of POS and their shares for the story "The Marble head" *(Mramornaya golovka)* by Valery Bryusov

words (Distance I).

• The distance degree — the maximum number of embedded words between two syntactically related words (Distance II).

• The number of homogeneous parts (groups) in a sentence.

• The number of elements in an enumeration sequence.

• Unprojectivity.

Measures of syntactic complexity is to be calculated on syntax annotation, which is made by ETAP-4 [ETAP-4]. An example of tree graph is shown on Fig. 2. (visualization utility was developed by Alexey Melnik).

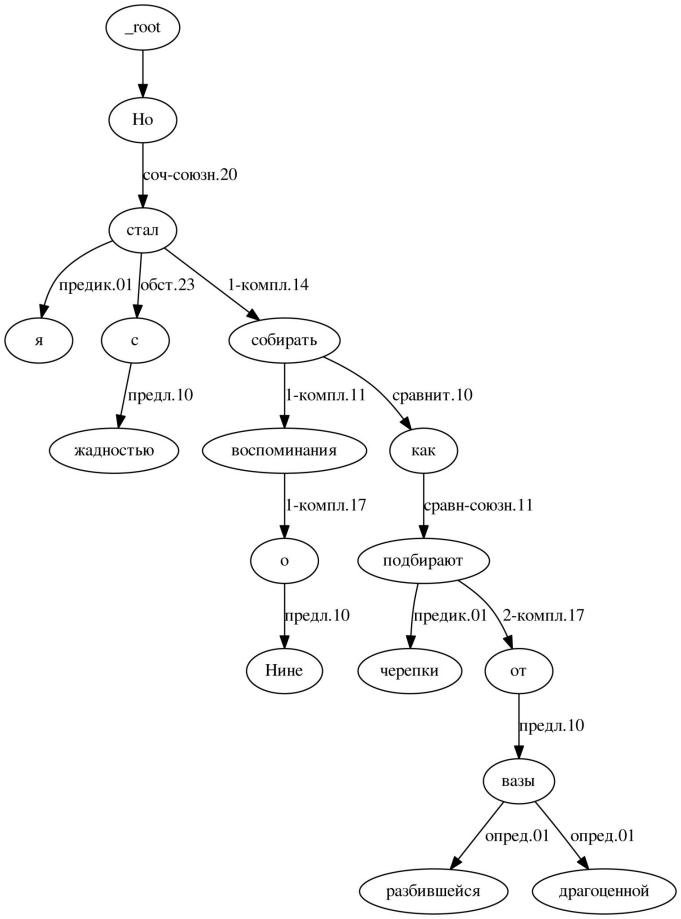

Figure 2: Syntactic tree for sentence 80 in the story "The Marble head" (*Mramornaya golovka*) by Valery Bryusov "*Но я с жадностью стал собирать воспоминания о Нине, как подбирают черепки от разбившейся драгоценной вазы.*" [*But I eagerly began to collect memories of Nina, how they pick up the shards from a broken precious vase.*] by Valery Bryusov

Other stylistic features

• The number and the share of the first person pronouns.

• The number and the share of the third person pronouns.

• The number and the share of indefinite pronouns.

• The number and the share of participles, ending in *–vshij, –shchij.*

• The number and the share of nouns, ending in *–ost, –stvo, –anie, –enie* [Martynenko, 1988, p. 90]. Phonetic Variables

• Consonant coefficient (i. e., the ratio of the number of consonants to the number of vowels);

• Stress index (characterizes the accent structure);

• Rhythmic structure of the word;

• The rhythmic dictionary and the language model [Kazartsev, 2015; 2017].

Each statistic listed in this and previous section will be calculated for individual texts and for four historical periods in the whole: 1) the early 20th century (1900–1913), 2) the World War I and prerevolutionary years (1914–1916), 3) the revolutionary years — the February and October Revolution — and the Civil War (1917–1922), and 4) the postwar years and the early Soviet period (from the end of the Civil War until 1930). Thereby individual and generalized story profiles for each analyzed parameter will be built.

This will allow to analyze these linguistic and stylistic variables in diachrony using the time series method, special attention being paid to the behavior of variables during the revolutionary years. As a result, the variables that have undergone the maximum transformation will be revealed, and a quantitative description of these changes will be obtained.

7 Conclusion

The paper discussed the principles of text selection for the annotated subcorpus containing stories of 300 Russian prose writers and considers the list of linguistic and stylistic parameters proposed for studying the language of literary texts in synchrony and diachrony. Currently we are working on the development of methods, which should allow automatic text processing, and make text annotations required for obtaining the proposed diagnostic parameters, using both automatic natural language processing methods (POS markup, syntactic structures, etc.) and expert methods of manual annotation.

The proposed list of variables should not be considered final. Thus, it turned out to be efficient to use some additional statistical parameters. For example, the effectiveness of specificity as a measure of subcorpus lexical distinction from the whole corpus is currently being evaluated [Lafon, 1980; Lavrentiev et al., 2018], and it seems interesting to compare the results to be obtained with those that provide the measures of determining keywordness, as well as those that are used for dynamic thematic modeling (e. g., see the article by Ekaterina Zamirailova in this volume).

Moreover, already in the process of creating the corpus, the need for additional annotation, in particular thematic text tagging and some elements of literary annotation [Skrebtsova, 2019; Rogova, 2018] became evident, so it is planned to include these types of tagging too.

The approbation of the proposed variables on corpus data will allow to evaluate their effectiveness and deficiencies. It is expected that the technique developed in the framework of this study can be successfully applied in research of diachronic and evolutionary changes in texts of other prose genres — publicistic and scientific literature, special texts, transcripts of oral speech, etc. — and can be applied to study linguistic changes traced in texts of any historical period, including the modern linguistic trends. Moreover, in the future it can be adapted to be used for texts in any language.

Acknowledgements

The research is supported by the Russian Foundation for Basic Research, project 17-29-09173 "The Russian language on the edge of radical historical changes: the study of language and style in prerevolutionary, revolutionary and post-revolutionary artistic prose by the methods of mathematical and computer linguistics (a corpus-based research on Russian short stories)".

References

[Alekseev, 2001] Alekseev P. M. (2001) Chastotnyye slovari: Uchebnoye posobiye [Frequency Dictionaries: Textbook]. St. Petersburg: Publishing House of St. Petersburg University, 2001 – 156 p.

[Archer et al., 2009] Archer, D., Culpeper, J., and Rayson, P. (2009) Love — 'a familiar or a devil'? An Exploration of Key Domains in Shakespeare's Comedies and Tragedies, Word frequency and keyword extraction, AHRC ICT Methods Network Expert Seminar on Linguistics, 8 September 2006, Lancaster University. Available at .

[Baker et al., 2006] Baker P. et al. (2006) Glossary of Corpus Linguistics, Edinburgh University Press.

[Balossi, 2014] Balossi G. (2014) A Corpus Linguistic Approach to Literary Language and Characterization: Virginia Woolf's The Waves. Amsterdam; Philadelphia: John Benjamins Publishing Company.

[Barannikov, 1921] Barannikov A.P. (1921) Iz nablyudenij nad razvitiem russkogo yazyka v poslednie gody [From observations of the development of the Russian language in recent years].

[Čebanov, Martynenko 1998] Chebanov S., Martynenko G. (1998) Text as Real Population in A.A.Čuprov's Sense, Journal of Quantitative Linguistics, Volume 5, Number 3. December 1998. Pp. 163–166.

[ETAP-4] http://cl.iitp.ru/ru/etap4download (last accessed on 31.08.2019).

[3500 Texts] "3,500 Russian prose works (1800–1940)", (accessed: 19.11.2019).

[Fischer et al., 2018] Fischer F., Trilcke P., Kittel C., Milling C., Skorinkin D. (2018) To Catch a Protagonist: Quantitative Dominance Relations in German Language Drama (1730–1930), in: Digital Humanities 2018: Book of Abstracts / Libro de resúmenes. Mexico: Red de Humanidades Digitales A. C., Pp. 193–201.

[Fischer-Starcke, 2010] Fischer-Starcke B. (2010) Corpus Linguistics in Literary Analysis: Jane Austen and Her Contemporaries. London; New York: Continuum.

[Granovskaya, 2005] Granovskaya L.M. (2005) Russkij literaturnyj yazyk v konce XIX i XX veke [Russian literary language at the end of the 19th and 20th centuries]. Moscow: Elpis. [Grebennikov, 1998] Grebennikov A.O. (1998) O sostoyatel'nosti statistik chastotnogo slovarya khudozhestvennoy prozy [On the consistency of statistics of the frequency dictionary of fiction], Structural and Applied Linguistics. Vol. 5. St. Petersburg.

[Grebennikov, 2007] Grebennikov, A.O. (2017) K voprosu o merakh leksicheskogo skhodstva chastotnykh slovarey [On the measures of lexical similarity between frequency dictionaries], Advances in Social Science Education and Humanities Research, 75019, Paris: Atlantis Press, Том 122. Pp. 256–259.

[Jacobson, 1921] Jacobson R. (1921) Vliyanie revolyucii na russkij yazyk [The influence of the revolution on the Russian language]. Prague.

[Kartsevsky, 1923] Kartsevsky S.O. (1923) Yazyk, vojna i revolyuciya [Language, war and revolution]. Berlin.

[Kazartsev, 2015] Kazartsev E.V. (2015) The Rhythmic Structure of Tales of Belkin and the Peculiarities of a Poet's Prose, A Convenient Territory. Russian Literature at the Edge of Modernity. Essays in Honor of Barry Scherr. Edit. J. Kopper M. Wachtel. Columbus: Slavica. Pp. 55–65.

[Kazartsev, 2017] Kazartsev E.V. (2017) Stikhopodobnyye fragmenty prozy A. S. Pushkina, A. K. Tolstogo, F. K. Sologuba i B. L. Pasternaka v kontekste evolyutsii russkogo stikha [The Verse-Similar Fragments in the Prose by A. S. Pushkin, A. K. Tolstoy, F. Sologub, and B. L. Parternak], Trudy instituta russkogo yazyka im. V.V. Vinogradova [Proceedings of the Russian language Institute]. V.V. Vinogradova. XI. Pp. 235–244.

[KLE] Concise literary encyclopedia, in 9 volumes (1962–1978) M.: Soviet Encyclopedia, (last accessed: 19.11.2019).

[Lafon, 1980] Lafon P. Sur la variabilité de la fréquence des formes dans un corpus. Mots, 1980, No. 1. Pp. 127–165.

[Lavrentiev et al., 2018] Lavrentiev A.M., Solovyev F.N., Suvorova (Ananyeva) M.I., Fokina A.I., Chepovskiy A.M. (2018) Novyy kompleks instrumentov avtomaticheskoy obrabotki teksta dlya platformy TXM i yego aprobatsiya na korpuse dlya analiza ekstremistskikh tekstov [A New Toolkit for Natural Text Processing with the TXM Platform and its Application to a Corpus for Analysis of Texts Propagating Extremist Views]. Vestnik NSU. Series: Linguistics and Intercultural Communication, 16(3). Pp. 19–31.

[Mahlberg et al., 2016] Mahlberg, M., Stockwell, P., de Joode, J., Smith, C., O'Donnell, M. B. (2016) CLiC Dickens: Novel uses of concordances for the integration of corpus stylistics and cognitive poetics. Corpora, 11(3). Pp. 433–463.

[Martynenko, 1988] Martynenko, G.Ya. (1988) Osnovy stilemetrii [The Foundation of Stylometics]. St. Petersburg State University, St. Petersburg.

[Martynenko, 2019a] Martynenko, G. Ya. (2019) Metody matematicheskoj lingvistiki v stilisticheskikh issledovanijakh [Methods of mathematical linguistics in stylistic studies], St. Petersburg: Nestor-Istoriya.

[Martynenko, 2019b] Martynenko G.Ya. (2019) Stilizovannyye sintaksicheskiye triady v russkom rasskaze pervoy treti XX veka [Stylized syntactic triads in Russian short story of the first third of the 20th century], Proc. of the Int. Conf. Corpus Linguistics – 2019. St. Petersburg State University. Pp. 395–404.

[Martynenko et al., 2018a] Martynenko, G.Ya., Sherstinova, T.Yu., Melnik, A.G., Popova, T.I. (2018) Metodolo-gicheskie problemy sozdaniya Komp'yuternoj antologii russkogo rasskaza kak yazykovogo resursa dlya issledovaniya yazyka i stilya russkoj khudozhestvennoj prozy v ehpokhu revolyucionnykh peremen (pervoj treti XX veka) [Methodological problems of creating a Computer Anthology of the Russian story as a language resource for the study of the language and style of Russian artistic prose in the era revolutionary changes (first third of the 20th century)]. In: Computational linguistics and computational ontologies. Issue 2 (Proceedings of the XXI International United Conference The Internet and Modern Society, IMS-2018, St. Petersburg, May 30 - June 2, 2018 Collection of scientific articles), ITMO University, St. Petersburg. Pp. 99–104.

[Martynenko et al., 2018b] Martynenko G.Ya., Sherstinova T.Yu., Popova T.I., Melnik A.G., Zamirajlova E.V. (2018) O printsipakh sozdaniya korpusa russkogo rasskaza pervoy treti XX veka [On the principles of creation of the Russian short stories corpus of the first third of the 20th century]. Proceedings of the XV International Conference on Computer and Cognitive Linguistics TEL 2018, Kazan. Pp. 180–197.

[Martynenko, 2017] Martynenko, G. Order Statistics as a Tool for Comparing Frequency Dictionaries (2017) ACM International Conference Proceeding Series, Volume Part F133135, 2017 International Conference on Internet and Modern Society, IMS 2017; St. Petersburg. Pp. 116–119.

[Martynenko, Fomin, 1989] Martynenko G.Y., Fomin S.V. (1989) Ranking moments, Nauchno-tekhnicheskaya informatsiya, Seriya 2 – informatsionnye protsessy i sistemy, Issue: 8. Pp. 9–14.

[Martynenko, Martinovich, 2003] Martynenko G.Ya., Martinovich G.A. (2003) Mnogoparametricheskiy statisticheskiy analiz rezul'tatov assotsiativnogo eksperimenta [Multiparametric statistical analysis of the results of an associative experiment] St. Petersburg State University. Scientific reports. St. Petersburg State University. — 28 pp.

[Martynenko, Sherstinova, 2000] Martynenko, G., Sherstinova, T. (2000) Statistical Parametrization of Text Corpora, P. Sojka, I. Kopeček, and K. Pala (eds.) TSD 2000, LNAI 1902, Springer: Berlin-Heidelberg. Pp. 99–102.

[Martynenko, Sherstinova, 2018] Martynenko G., Sherstinova T. (2018) Emotional Waves of a Plot in Literary Texts: New Approaches for Investigation of the Dynamics in Digital Culture. In: Alexandrov D.et al. (eds.) Digital Transformation and Global Society. DTGS 2018. Communications in Computer and Information Science, vol 859. Springer, Cham. Pp. 299–309.

[Muratova, 1963] Muratova K.D. (1963) Istoriya russkoy literatury kontsa XIX – nachala XX veka, Bibliograficheskiy ukazatel' [The History of Russian literature of the late XIX – early XX century, Bibliographic index], Moscow: Academy of Science of SSSR.

[Polivanov, 1931] Polivanov E.D. (1931) Revolyuciya i literaturnye yazyki soyuza SSR [The revolution and the literary languages of the USSR], Polivanov E.D. Za marksistskoye yazykoznaniye. Moscow: Federatsiya. Pp. 73–94.

[Popescu, 2009] Popescu, I.-I. (2009) Quantitative Linguistics: Word Frequency Studies. Berlin-New-York: Mouton de Gruyter.

[Rempel, 1921] Rempel E. (1921) Yazyk revolyucii i revolyuciya yazyka [The language of revolution and the revolution of language] Riga.

[Rogova, 2018] Rogova K. A., ed. (2018) Analiz khudozhestvennogo teksta. Russkaya literatura XX veka: 20-ye gody: uchebnoye posobiye [Analysis of the literary text. Russian literature of the XX century: the 20s: a textbook], St. Petersburg: Publishing House of St. Petersburg University. — 286 p.

[Russian Writers] Russian Writers of 1800–1917, Biographical Dictionary. In 7 voluems. Ed. by Nikolaev P.A. 1992–2000. Moscow: Scientific publishing house Big Russian Encyclopedia.

[Selishchev, 1928] Selishchev A.M. (1928) Yazyk revolyutsionnoy epokhi: Iz nablyudeniy nad russkim yazykom poslednikh let (1917–1926) [The language of the revolutionary era: From observations of the Russian language of recent years. (1917–1926)], Moscow: Rabotnik prosveshcheniya.

[Shaykevich, 2015] Shaykevich A.Ya. (2015) Mery leksicheskogo skhodstva chastotnyh slovarej [Measures of lexical similarity between frequency dictionaries], Proc. of the Int. Conference Corpus linguistics-2015 [Trudy mezhd. konf. Korpusnaya linguistica-2015]. Pp. 422–429.

[Sherstinova, 2018] Sherstinova, T. (2018) Quantitative Data on POS Distribution in the Beginnings and the Ends of Utterances in Everyday Russian Speech. In: Potapova R., Jokisch O., Karpov A. (eds.) Speech and Computer (SPECOM 2018), Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), LNAI 11096, Springer Verlag. Pp. 596–605.

[Sinclair, 2004] inclair, J. (2004) Trust the Text: Language, Corpus and Discourse, Routledge, 224 p.

[Skorinkin, Fischer 2018] Skorinkin D., Fischer F. (2018) Measuring the 'Epification' of Drama, in: Applications in Cultural Evolution: Arts, Languages, Technologies. University of Tartu. Pp. 46–48.

[Skrebtsova, 2019] Skrebtsova T. G. (2019) Struktura narrativa v russkom rasskaze nachala XX veka [Narrative structure of the Russian short story in the early XX century], Proceedings of the International Conference Corpus Linguistics-2019. St. Petersburg: Publishing House of St. Petersburg University. Pp. 426–431.

[Sovlit] roject Sovlit,(last accessed: 19.11.2019).

[Tuldava, 1977] Tuldava Yu.A. (1977) O kvantitativnykh kharakteristikakh bogatstva leksicheskogo sostava khudozhestvennykh tekstov [On the quantitative characteristics of the richness of the lexical composition of literary texts] Linguistica. Tartu. IX. Pp. 159–175.

[Tuldava, 1986] Tuldava Yu.A. (1986) O chastotnom spektre leksiki teksta [On the frequency spectrum of text vocabulary], Scientific Notes of Tartu University, Vol. 745. Quantitative linguistics and automatic text analysis. Tartu. Pp. 139–162.

[Tynyanov, 1929] Tynyanov, Yu.N. (1929) Arkhaisty i novatory [Archaists and Innovators]. Leningrad: Priboj.

[Vinokur, 1923] Vinokur G.O. (1923) O revolyucionnoj frazeologii (Odin iz voprosov yazykovoj politiki) [On revolutionary phraseology (One of the issues of language policy)]. LEF. No 2. Moscow–Petrograd.

[Voronchak, 1972] Voronchak E. (1972) Metody vychisleniya pokazateley leksicheskogo bogatstva tekstov [Methods of calculating indicators of the lexical richness of texts], Semiotics and artmetry. Moscow. Pp. 232–250.

[Zyngier, 2008] Zyngier, S. (2008) Macbeth through the computer: Literary evaluation and pedagogical implications, The Quality of Literature: Linguistic studies in literary evaluation, edited by Willie van Peer, Linguistic Approaches to Literature, 4. John Benjamins Publishing Company. Pp. 169–190.