Terra Linguistica. 2023, Vol. 14, No. 1. Pp. 21–29.
Terra Linguistica. 2023, Том 14, № 1. С. 21–29.

Research article
UDC 81'33
DOI: https://doi.org/10.18721/JHSS.14103

MAPPING WORD FREQUENCIES IN FICTION ON SOCIOPOLITICAL CONTEXT: THE CASE OF EARLY 20TH CENTURY RUSSIAN SHORT STORIES

A.O. Grebennikov ✉ , **N.M. Marusenko** , **T.G. Skrebtsova**

St. Petersburg State University,
St. Petersburg, Russian Federation

✉ a.grebennikov@spbu.ru

Abstract. The paper deals with the language of Russian short stories written in the period from 1900–1930. It is based on the Russian Short Stories Corpus, an ongoing research project aimed to collect, digitally process, and present the Russian literature of the early 20th century in an electronic form. The Corpus contains the stories written by thousands of Russian authors, both well-known and almost forgotten ones. From the corpus, a sample was taken to serve as a testbed for linguists, lexicographers and literary scholars, enabling them to check their intuitions concerning the language and style of the epoch. The sample has been divided into three subsamples along the lines set by the dramatic turns of Russian history. The first subsample contains the stories produced from the onset of the 20th century up to WWI (1900–1913), the second one refers to the tumultuous period of wars and revolutions (1914–1922), and the third accounts for the stories written in the Soviet Union (1923–1930). The Corpus has proved instrumental in detecting manifold changes in language use, including grammar, vocabulary, syntactic patterns, collocations, and stylistics. In the present paper, frequency-sorted word lists are used to bring out relevant changes in Russian vocabulary, linking them to the sociopolitical context. The results obtained will provide valuable data for the lexicographers compiling Russian dictionaries of the above-mentioned period.

Keywords: Russian short stories, text corpus, frequency dictionary, Russian lexicography, stylometry.

Citation: A.O. Grebennikov, N.M. Marusenko, T.G. Skrebtsova, Mapping word frequencies in fiction on sociopolitical context: the case of early 20th century Russian short stories, Terra Linguistica, 14 (1) (2023) 21–29. DOI: 10.18721/JHSS.14103

Научная статья
DOI: https://doi.org/10.18721/JHSS.14103

ЧАСТОТНЫЙ СЛОВАРЬ ХУДОЖЕСТВЕННОЙ ПРОЗЫ В КОНТЕКСТЕ СОЦИОПОЛИТИКИ (НА МАТЕРИАЛЕ «КОРПУСА РУССКОГО РАССКАЗА 1900–1930 ГГ.»)

А.О. Гребенников ✉ , **Н.М. Марусенко** , **Т.Г. Скребцова**

Санкт-Петербургский государственный университет,
Санкт-Петербург, Российская Федерация

✉ a.grebennikov@spbu.ru

Аннотация. Работа выполнена на материале «Корпуса русского рассказа 1900–1930 гг.» – масштабного проекта, направленного на сбор, цифровую обработку, анализ и представление произведений русской литературы начала XX века в электронном виде. Корпус содержит рассказы нескольких тысяч авторов, как признанных мастеров художественного слова, так и практически неизвестных. Исследование проведено на материале аннотированной выборки из Корпуса, которая служит «полигоном» для проверки гипотез, касающихся стиля языка эпохи, лингвистами, литературоведами и лексикографами. Выборка разделена на три подвыборки, отражающие основные этапы русской истории начала XX века: 1) довоенный период (1900–1913), 2) военно-революционные годы (1914–1922) и 3) советский период (1923–1930). Установлено, что анализ корпусного материала показателен при прослеживании различный изменений в использовании языка, включая грамматику, лексику, синтаксические модели, коллокации и стилистику. В настоящем исследовании построенные по выборкам частотные словари используются для выявления значимых изменений в лексическом составе, которые рассматриваются в социополитическом контексте. Полученные результаты представляют интерес для специалистов в области русского языка, стилеметрии и лексикографии.

Ключевые слова: Русский рассказ, корпус текстов, частотный словарь, лексикография, стилеметрия.

Для цитирования: Гребенников А.О., Марусенко Н.М., Скребцова Т.Г. Частотный словарь художественной прозы в контексте социополитики (на материале «Корпуса русского рассказа 1900–1930 гг.») // Terra Linguistica. 2023. Т. 14. № 1. С. 21–29. DOI: 10.18721/JHSS.14103

The Russian Short Stories Corpus (1900–1930)

The present paper draws on the Russian Short Stories Corpus (1900–1930), an ongoing project aimed to collect, digitally process, and present the Russian literary heritage of the early 20th century in an electronic form, thus making it available to a wide range of users. In particular, the Corpus is supposed to become a major resource for the linguists and literary scholars, enabling them to research into the language and style of the pre-revolutionary, revolutionary and post-revolutionary prose [1, 2]. For lexicographers, it contains valuable information on the way Russian grammar, vocabulary, phraseology, and stylistics kept changing over this tumultuous period of Russian history and will be instrumental in compiling Russian dictionaries of this period.

The Corpus currently contains a few thousand stories written in Russia, and later the Soviet Union, and published in literary journals or anthologies. It seeks to include works by a maximal number of writers, not only the famous ones but also the lesser-known or almost forgotten authors, thus maintaining a well-balanced and representative collection. The whole Corpus is divided into three chronological subcorpora, their boundaries marked by significant historical events. Thus, the first subcorpus (1900–1913) refers to a

pre-war period, the second one (1914–1922) covers a series of dramatic events (WWI, the February and October revolutions, the Civil War) that resulted in an overall radical change in the Russian political landscape and social life, and the third one (1923–1930) accounts for the post-war socialist period.

Each author can be represented by a single story per period. Stories written in emigration are not included in the Corpus. Thus, the Corpus features two stories by Ivan Bunin, one written in the 1st period and the other in the 2nd one. As the writer left Russia shortly after the revolution, his 3rd period stories are not accounted for.

Besides Ivan Bunin, the Corpus, thus, contains stories by such prominent Russian authors as Leo Tolstoy, Leonid Andreev, Arkady Averchenko, Alexander Blok, Sergey Esenin, Konstantin Balmont, Andrey Belyj, Anton Chekhov, Maxim Gorky, Zinaida Gippius, Nadezhda Teffi, Alexander Kuprin, Mikhail Zoschenko, Evgeny Zamyatin, Ivan Schmelev, Valentin Kataev, Veniamin Kaverin, Mikhail Kuzmin, Isaac Babel, Mikhail Bulgakov, Yuri Olesha, Arkady Gaydar, Konsyantin Paustovsky, Andrey Platonov, Mikhail Sholokhov, Alexey Tolstoy, etc.

From the text corpus, a random sample was taken containing 310 stories by 300 authors, ca. 100 stories per period (the slight discrepancy in numbers is due to the fact that some writers feature in more than one period). This sample serves as an initial testbed enabling scholars to put forward and prove or refute the hypotheses bearing on Russian language and literature of the given period [3–5]. The present research is also based on this sample.

Word Frequency Distribution as a Window on the Sociopolitical Context

The underlying idea of any research on the Russian Short Stories Corpus (1900–1930) is that language use cannot help being affected by the sociopolitical processes, hence the division of the whole Corpus into subcorpora in accordance with the major milestones in Russian history (see above). The present paper is no exception. Focusing on the short stories vocabulary and, more specifically, on frequency-sorted word lists, it aims to find significant variation across the periods and account for it in terms of political events and social developments.

Word frequency analysis has proved instrumental in language studies, in general, and in corpus linguistics, in particular [6–8]. With respect to the project concerned, it has been used to explore the major statistical parameters of the Corpus, including the words' absolute and relative frequency, rankings, part-of-speech distribution, rank mean, keyness, lexical specificity, as well as to perform cluster analysis both for each individual period and for the whole corpus [9].

Apart from this, word frequency analysis is helpful in bringing out lexical features characteristic of a writer's individual style. Thus, a comparison of word frequency ranks drawn from the works of a few writers may provide an insight into their individual world views and priorities. It has been shown, in particular, that Ivan Bunin's stories are primarily about the rural life and the beauty of nature, whereas Anton Chekhov focused mainly on social life and human relationships. Word frequency analysis of the stories by Leonid Andreev has convincingly demonstrated his obsession with the tragic aspects of life, including loneliness and fear of death [10].

In the present research, word frequency ranks have been calculated over a variegated collection of stories by a few hundred authors. Individual differences are thus neutralized, and the results obtained may be said to reveal a certain flavour of the epoch. The frequency dictionaries under investigation have been complied using UNILEX-T software [11].

The technique is not ideal in that certain errors can be made in automatically deriving lemmas from the word forms, due to the homonymy. This is often the case with the highly inflected languages, such as Russian. Still, the ratio of such errors is quite small and can usually be neglected.

Another minor problem may arise from the polysemous words being treated as a single unit. Semantic tagging has always been a challenge to automatic processing, and the current state of the project does not provide such an option. Therefore, strictly speaking, one cannot check which of the individual word senses

suddenly got activated and brought about an increase in the overall frequency. But plausible conjectures can still be made, drawing on the previously detected dynamics of change in the stories' thematic content [5, 12, 13] and the political context in which the stories were produced. This is the case, in particular, with the words *tovarishch* ('comrade') and *krasnyj* ('red'), whose frequency drastically rose in the socialist period, obviously due to the activation of the new, ideological, senses.

In what follows, "upper zones" of the frequency lists of all the three periods are analysed, each containing content words with frequency over 100. For each period, the number of such lexical units is well over 200. Taken together, they amount to ca. 800. Table 1 provides data on some of the words discussed below.

Interestingly, the highest six ranks of all the frequency lists are filled by the same six content words (though, with varying order), namely *govorit'* ('to say'), *skazat'* ('to tell'), *odin* ('one'), *glaz* ('eye'), *ruka* ('hand, arm'), and *moch* ('can, may, be able'). Below these top ranks, the frequency distributions display quite a few noteworthy differences.

By comparing an upper zone of a later period with that/those of the earlier one(s), the following terms are identified:

1. words previously unfound in the upper zone;
2. words that, by contrast, are no longer present in the upper zone;
3. words demonstrating sharp drops and rises within the upper zone across periods.

In all the three types of cases, an attempt is made to interpret our findings in light of the relevant sociopolitical context and link them to the previously detected dynamics of change in the thematic content of the Russian short stories [12, 13].

Tracing Word Frequency Change across the Periods

1. The second (wartime) vs. the first (pre-war) period

In the wartime period (1914–1922) the words *ofitser* ('officer'), *russkij* ('Russian'), *dyakon* ('deacon') and *pisat'* ('write') made their way into the upper zone of the frequency list. This obviously resulted from the very character of the epoch. A long chain of wars and revolutions brought about, among other things, the separation of families, anxiety, distress, and sorrow, the need to keep in touch and pray. The rankings of the words *soldat* ('soldier'), *Bog* ('God') and *pis'mo* ('letter'), already present in the upper zone in the pre-war period, also went up. These facts are in accord with the increased activation of the relevant themes.

The war issues pushed down themes bearing on the regular work and study, so the words *barin* ('master'), *khozyain* ('employer'), *rabotat'* ('to work'), *rabochij* ('worker'), *student* ('student') left the upper zone.

A tougher time demanded a tougher modality, with the word *mozhno* ('one may') leaving the upper zone and the words *dolzhnyj* ('one must') and *nel'zya* ('one should not'), by contrast, entering it. Thus, permission was replaced by compulsion and prohibition.

Another conspicuous fact indicative of a difficult time is a significant drop in frequency of a wide range of terms carrying positive connotations, cf. *prazdnik* ('feast'), *dobryj* ('kind'), *svetlyj* ('bright'), *krasivyj* ('beautiful'), *vesyolyj* ('merry'), *smekh* ('laughter'), *schast'je* ('happiness'), *ulybka* ('smile'), *ulybatsya* ('to smile'), *vera* ('faith'), *tikhij* ('silent'), *tishina* ('silence'). All of these left the upper zone in the 2nd period, with only a few to return in the 3rd one (see below). Accordingly, topics bearing on love, family life, charity, magnanimity, etc. show decreasing frequencies.

The words *pit'* ('to drink') and *pjanyj* ('drunken') also went well below the upper zone, evidently due to prohibition enforced in Russia at the beginning of WWI. It continued through the turmoil of the revolutions and the Civil War until 1925.

2. The third (post-war, socialist) period vs. the preceding ones

Perhaps, the most remarkable feature of the word frequency distribution in the 3rd period is the upward movement of a vast number of concrete nouns associated, firstly, with rural life and peasantry, and secondly, with technical progress. Thus, the upper zone was enriched by such words as *ded* ('grandfather, old man'), *starukha* ('old woman'), *rebyata* ('children'), *pole* ('field'), *khleb* ('bread'), *kust* ('shrub'), *trava*

(‘grass’), *sobaka* (‘dog’), *kon*’ (‘horse’), *ptitsa* (‘bird’), *mashina* (‘machine’), *poezd* (‘train’), *vagon* (‘railway carriage’), *khod* (‘motion’). The frequency of the corresponding themes enjoyed a sharp rise, too. Abstract nouns, by contrast, yielded, many of them leaving the upper zone.

The list of the body-part names steadily featuring in the upper zone – *ruka* (‘hand, arm’), *glaz* (‘eye’), *golova* (‘head’), *litso* (‘face’), *guba* (‘lip’), *zub* (‘tooth’), *noga* (‘leg, foot’), *telo* (‘body’), *plecho* (‘shoulder’), *palets* (‘finger’), *volosy* (‘hair’) – in the 3rd period was almost doubled by the adding of *nos* (‘nose’), *ukho* (‘ear’), *yazyk* (‘tongue’), *sheya* (‘neck’), *shcheka* (‘cheek’), *boroda* (‘beard’), *bok* (‘side’), *koleno* (‘knee’). There are more numerals to be found in the top ranks, too.

The permission modality (*mozhno*) is back, with prohibition (*nel’zya*) gone and compulsion (*dolzhnyj*) remaining. Some words that left the upper zone in the 2nd period are back, too, cf. *vesyolyj* (‘merry’), *smekh* (‘laughter’), *tikhij* (‘silent’), *tishina* (‘silence’), *igrat*’ (‘play’), *razgovor* (‘talk’), *rabotat*’ (‘to work’), *rabochij* (‘worker’), which must be due to the beginning of peace. Accordingly, the words *ofitser* (‘officer’) and *soldat* (‘soldier’) left the upper zone.

Social relations in the 3rd period center primarily on work and family, hence a drop in the frequency of the words *gost*’ (‘guest’) and *znakomyj* (‘acquaintance’). Family relations, though, are also fading, cf. the falling frequency of *muzh* (‘husband’), *zhena* (‘wife’), *deti* (‘children’). *Rebyonok* (‘child’) already left the upper zone in the 2nd period and failed to re-appear. These lexical trends are corroborated by a similar dynamics of change in the stories’ thematic component.

Many words remain in the upper zone throughout all the three periods. Some of them hold a more or less stable position in frequency rankings, while others demonstrate a progressive upward or downward movement pattern.

The rising pattern is particularly characteristic of the words tovarishch (‘comrade’) and krasnyj (‘red’). The opposite trend can be observed in words referring to the family life and those denoting emotional and spiritual life aspects, cf. *lyubit*’ (‘to love’), *chuvstvovat*’ (‘to feel’), *smeyatsya* (‘to laugh’), *dusha* (‘soul’), *mysl*’ (‘thought’), *Bog* (‘God’).

Discussion

Above, the most spectacular word frequency changes have been mentioned that can be easily accounted for in terms of the relevant sociopolitical context. However, with other words, the dynamics of frequency change is at least not so understandable and may even seem counter-intuitive. Thus, the words *strashnyj* (‘dreadful’), *strakh* (‘fear’), *uzhas* (‘horror’), *drozhat*’ (‘tremble’), *umeret*’ (‘die’), *toska* (‘anguish’), *bol’noj* (‘sick’), present in the upper zone in the 1st (pre-war) period, left it in the 2nd (wartime) period, although it would look more natural the other way round.

It may also seem strange that the military terms *ruzhjo* (‘gun’) and *rota* (‘company as a military unit’), together with *krov*’ (‘blood’), were absent from the upper zone in the 2nd period but did enter it in the 3rd one. One would expect them, instead, to show higher frequency in the stories of 1914–1922. This fact, though, nicely fits with our previous finding concerning the stories themes, as there proved to be twice as many stories about the Civil War in the 3rd period as in the 2nd one [13]. Such postponed effect, in general, is typical of the decisive events affecting the very course of a nation’s history. They retain significance for many decades, being evoked in scholarship, literature, and art.

Other cases defying a rough and ready explanation are the words with a broken-line pattern of frequency dynamics, reaching a local maximum or minimum in the 2nd period. The whole lot looks rather heterogeneous and inconclusive, so they are not considered in detail. Perhaps, such patterns would become revealing if a larger upper zone of the frequency distribution were examined.

Conclusion

In the present paper, the upper zones of the word frequency distribution in early 20th century Russian short stories have been analysed. The cases well-marked by a progressive dynamics of change have been

Table 1. Frequency ranks of selected words across the three periods

LEMMA	PERIOD 1 (1900–1913)	PERIOD 2 (1914–1922)	PERIOD 3 (1923–1930)
skazat' (to say)	1	2	3
odin (one)	2	3	4
glaz (eye)	3	4	2
govorit' (to tell)	4	1	5
ruka (hand, arm)	5	5	1
moch (may, can, be able)	6	6	6
dusha (soul)	42	35	151
zhena (wife)	52	84	139
chuvstvovat' (to feel)	54	152	197
deti (children)	69	107	198
mozhno (one may)	80		95
bog (god)	89	65	211
vera (faith)	114		
soldat (soldier)	116	62	
milyj (gentle, nice)	122	176	
chuvstvo (feeling)	126		
lyubov' (love)	134	112	
strashnyj (horrible)	135		241
krasivyj (beautiful)	143		
muzh (husband)	146	166	236
veselyj (merry)	152		232
krasnyj (red)	163	110	54
uzhas (horror)	166		
ulybatsya (to smile)	173		
znakomyj (acquaintance)	179	193	
drozhat' (to tremble)	185		276
pis'mo (letter)	190	117	
tovarisch (comrade)	198	105	138
rabochij (worker)	202		112
rebyonok (infant, child)	213		
student (student)	217		
rabotat' (to work)	218		109
schastje (happiness)	228		
pyanyj (drunken)	241		
derevnya (village)	250	192	150
muzhik (peasant man)	252		
smekh (laughter)	257		234
izba (peasant hut)	266		138
prazdnik (feast)	268		
gost' (guest)	272	165	
dyakon (deacon)		83	
baba (peasant woman)		143	81
nel'zya (one should not)		147	
ofitser (officer)		169	

End of table 1

pisat' (to write)		181	
russkij (Russian)		182	
krov' (blood)			105
vagon (railway carriage)			173
rota (company as a military unit)			219
ruzhjo (gun)			278

specifically focused on. Most of them can be accounted for in terms of the relevant sociopolitical situation and the previously detected changes in the stories' thematic content. Yet, there are words whose frequency change pattern remains not quite clear. An extension of the upper zone may help as it will bring into light a larger number of similar cases.

Also, beyond our present study are words demonstrating a steady position in frequency rankings regardless of the historical context and thus representing a kind of distribution invariants. The top six ranks being filled by the same set of words is, perhaps, the brightest example, but certainly not the only one. It would be of interest to examine how far such invariance stretches by extending the temporal boundaries of the stories beyond the given timespan.

The perspectives of our future work are manifold. Firstly, the frequency-sorted word lists of our sample can be set against the frequency distributions drawn from the stories by a particular author (see [13–17]). A pilot study [18] has shown a remarkable discrepancy between the two data sets testifying to the significant impact of personal style on the literary works' vocabulary. Secondly, it would be of interest to compile word frequency list drawing on the Russian short stories of the 21st century and then compare it to the data at hand. The above-cited paper has revealed striking differences in the upper zone of the frequency distributions that have occurred over a century (Ibid). Thirdly, an extension of the present sample is being looked forward to, to make it more representative and well-balanced, thus increasing the reliability of results. This is crucial for a broad range of research on the corpus not only within linguistics, but also in literary theory and digital humanities at large.

A special direction in the future research has to do with lexicography. Along with the comprehensive dictionaries of the Russian language that have been compiled in the Russian Academy of Sciences, there is a growing interest in the language of particular historical periods. Thus, the Russian dictionaries of the 18th and 19th centuries are currently under way. It would only be logical that the focus eventually shift to the early 20th century. This period is marked by certain distinctive features of its own, e.g. acronyms and abbreviations, coined words and phrases, novel ideological word senses, shifts in lexical use, etc. The coverage it has so far received is certainly insufficient. The corpus data will be of help to lexicographers in their future work, and frequency lemma lists have been shown to be quite useful in assessing the relative frequency of individual words [19]).

Acknowledgements

The materials of the article were presented at the IV International Conference on Engineering and Applied Linguistics "Piotrovsky Readings – 2022", dedicated to the 100th anniversary of the birth of Professor R.G. Piotrovsky at the Herzen State Pedagogical University on November 22, 2022.

REFERENCES

[1] **G. Martynenko, T. Sherstinova, T. Popova, A. Melnik, Ye. Zamirajlova,** 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]. Proc. of the XV Int. Conference on Computer and Cognitive Linguistics "TEL 2018", Kazan, 2018, pp. 180—197.

[2] **G. Martynenko, T. Sherstinova,** 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. R. Piotrowski's Readings in Language Engineering and Applied Linguistics, Proc. of the III Int. Conference on Language Engineering and Applied Linguistics (St. Petersburg, Nov., 27, 2019), CEUR Workshop Proceedings, 2552, 2020, pp. 105—120.

[3] **G. Martynenko, T. Sherstinova,** Emotional Waves of a Plot in Literary Texts: New Approaches for Investigation of the Dynamics in Digital Culture. Digital Transformation and Global Society (DTGS 2018). Communications in Computer and Information Science, 859. Springer, Cham, 2018, pp. 299—309. Available at: https://link.springer.com/chapter/10.1007/978-3-030-02846-6_24 (accessed 10.02.2023).

[4] **T. Skrebtsova,** Struktura narrativa v russkom rasskaze nachala XX veka [Narrative structure of the Russian short story in the early XX century], Proc. of the Int. Conference "Corpus Linguistics-2019", St. Petersburg University Press, St. Petersburg, 2019, pp. 426—431.

[5] **T. Sherstinova, O. Mitrofanova, T. Skrebtsova, E. Zamiraylova, M. Kirina,** Topic Modelling with NMF vs. Expert Topic Annotation: the Case Study of Russian Fiction, Advances in Computational Intelligence. 19th Mexican International Conference on Artificial Intelligence, MICAI 2020, Mexico City, Mexico, 12-17 October 2020, LNAI 12469, 2020, pp. 134—151.

[6] **M. Oakes,** Statistics for Corpus Linguistics, Edinburgh University Press, Edinburgh, 1998.

[7] **G. Leech, P. Rayson, A. Wilson,** Word Frequencies in Written and Spoken English: based on the British National Corpus, London: Longman, 2001.

[8] **A. Baron, P. Rayson, D. Archer,** Word frequency and key word statistics in historical corpus linguistics, Anglistik: International Journal of English Studies, 20 (1), 2009, pp. 41—67.

[9] **T. Sherstinova, A. Grebennikov, T. Skrebtsova, A. Guseva, M. Gukasian, I. Egoshina, M. Turygina,** Frequency Word Lists and Their Variability (the Case of Russian Fiction in 1900–1930). 27th Conference of Open Innovations Association FRUCT, University of Trento, Italy, 2020, pp. 366—373. Available at: https://fruct.org/publications/acm27/files/She.pdf (accessed 10.02.2023).

[10] **A. Grebennikov, T. Skrebtsova,** Yazykovaya kartina mira v russkom rasskaze nachala XX veka [World through the Prism of the Early XX-century Russian Short Stories], Philosophy and the Humanities in the Information Society, 3, 2019, pp. 82—92.

[11] **Zh. Anoshkina,** Podgotovka chastotnyh slovarej i konkordansov na komp'yutere [Computer-assisted Dictionary and Concordance Making], V.V. Vinogradov Russian Language Institute of the Russian Academy of Sciences Press, Moscow, 1995.

[12] **T. Sherstinova, T. Skrebtsova,** Russian Literature Around the October Revolution: A Quantitative Exploratory Study of Literary Themes and Narrative Structure in Russian Short Stories of 1900–1930, Proc. of the Int. Workshop "Computational Linguistics" (St. Petersburg, 17-20 June, 2020), CEUR Workshop Proceedings, 2813, 2021, pp. 117—128. Available at: http://ceur-ws.org/Vol-2813/rpaper09.pdf (accessed 10.02.2023).

[13] **T. Skrebtsova,** Thematic tagging of literary fiction: the case of early 20th century Russian short stories, Proc. of the Int. Workshop "Computational Linguistics" (St. Petersburg, 17-20 June, 2020), CEUR Workshop Proceedings, 2813, 2021, pp. 265—276. Available at: http://ceur-ws.org/Vol-2813/rpaper20.pdf (accessed 10.02.2023).

[14] **A. Grebennikov, G. Martynenko,** Chastotnyy slovar rasskazov A.P. Chekhova [Frequency Dictionary of Anton Chekhov's Short Stories], St. Petersburg University Press, St. Petersburg, 1999.

[15] **A. Grebennikov, G. Martynenko,** (2003), Chastotnyy slovar rasskazov L.N. Andreeva [Frequency Dictionary of Leonid Andreev's Short Stories], St. Petersburg University Press, St. Petersburg, 2003.

[16] **A. Grebennikov, G. Martynenko,** Chastotnyy slovar rasskazov A.I. Kuprina [Frequency Dictionary of Alexander Kuprin's Short Stories], St. Petersburg University Press, St. Petersburg, 2006.

[17] **A. Grebennikov, G. Martynenko,** Frequency Chastotnyy slovar rasskazov A.I. Bunina [Dictionary of Ivan Bunin's Short Stories], St. Petersburg University Press, St. Petersburg, 2011.

[18] **A. Grebennikov, N. Marusenko,** Korpus russkogo rasskaza nachala XX veka. Primer lingvostatisticheskogo analiza [The Early XX-century Russian Short Stories Corpora. An Example of Lingvo-statistical analysis], Proc. of the 23rd Int. conf. "Internet and Modern Society" (IMS-2020), St. Petersburg, 2020, pp. 21—29.

[19] **D. Lindemann, I. San Vicente,** Building corpus-based frequency lemma lists. Procedia – Social and Behavioral Sciences, 2015, pp. 266–277.

СВЕДЕНИЯ ОБ АВТОРАХ / INFORMATION ABOUT AUTHORS

Alexander O. Grebennikov
Гребенников Александр Олегович
E-mail: a.grebennikov@spbu.ru
ORCID: https://orcid.org/0000-0003-2856-5049

Natalya M. Marusenko
Марусенко Наталия Михайловна
E-mail: n.marusenko@spbu.ru
ORCID: https://orcid.org/0000-0002-3347-1373

Tatyana G. Skrebtsova
Скребцова Татьяна Георгиевна
E-mail: t.skrebtsova@spbu.ru
ORCID: https://orcid.org/0000-0002-7825-1120

Submitted: 11.02.2023; Approved: 22.03.2023; Accepted: 22.03.2023.
Поступила: 11.02.2023; Одобрена: 22.03.2023; Принята: 22.03.2023.