Frequency Word Lists and Their Variability (the Case of Russian Fiction in 1900-1930)

Tatiana Sherstinova[12], Alexander Grebennikov[1], Tatiana Skrebtsova[1], Anna Guseva[2], Mary Gukasian[2], Irina Egoshina[2], Maria Turygina[2]
[1]Saint Petersburg State University, Saint Petersburg, Russia
[2]National Research University Higher School of Economics, Saint Petersburg, Russia
{t.sherstinova, a.grebennikov, t.skrebtsova}@spbu.ru, {aaguseva_3, mngukasyan, isegoshina, miturygina}@edu.hse.ru

Abstract—Lexical system is an essential component of any natural language. Frequency word lists are a convenient representation of words functional activity in language as a whole or in some particular text. The parameters and properties of frequency word lists are in the center of attention of NLP experts, since they are used in numerous practical applications related to attribution of authorship, text automatic clustering and classification. The article explores frequency word lists of Russian fiction in the period of 1900-1930, which was marked by a series of dramatic historical events and presents unique statistical data on the most frequent words, parts of speech and keywords, and their dynamics. Special attention is paid to the issues of statistical consistency of frequency word list parameters, which becomes especially relevant when studying big text data. The study was carried out on the basis of fiction texts, which by the variety of topics, lexical and stylistic diversity reflects the variability of linguistic forms better than the other written text genres. In terms of the text corpus size and character, the research of this kind is being carried out for the first time.

I. Introduction

Lexical system is an essential component of any natural language. Frequency word lists are a convenient representation of words functional activity in language as a whole or in some particular text [Tuldava, 1986; Alekseev, 2001; Grebennikov, 2007; Popescu, 2009; Shaykevich, 2015]. The parameters and properties of frequency word lists are in the center of attention of NLP experts, since they are used in numerous practical applications related to attribution of authorship, text automatic clustering and classification. The best-known frequency word lists for Russian are the following [Josselson 1953; Steinfeldt 1963; Zasorina et al. 1977; Lönngren 1993; Lyashevskaya and Sharov 2009].

The article explores frequency word lists of Russian fiction in the period of 1900-1930 [Martynenko et al. 2018a], which was marked by a series of dramatic historical events and presents unique statistical data on the most frequent words, parts of speech and keywords, and their dynamics. The main parameters of frequency lists that determine their stability and variability for a given language genre are considered. Special attention is paid to the issues of statistical consistency of frequency word list parameters, which becomes especially relevant when studying big text data. The research was carried out on the basis of fiction texts, which by the variety of topics, lexical and stylistic diversity reflect the variability of linguistic forms better than the other written text genres. It should be emphasized, that, despite the availability of a limited number of the frequency dictionaries for the short stories of the major Russian writers, the project of this kind, i. e. dealing with the most possible number of the writers of one of the most important historical period without any limitations, is being developed for the first time.

II. Research Data: Corpus of Russian Short Stories

The research is carried out on the base of Corpus of Russian Short Stories of 1900-1930, which 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].

For the annotated corpus, 310 short stories representing literary works of 300 Russian writers were selected, among which are world-famous writers (e. g., Anton Chekhov, Leo Tolstoy, Ivan Bunin, Maxim Gorky, Alexander Kuprin, Mikhail Bulgakov, Mikhail Sholokhov, etc.), a large group of relatively famous writers (Andrei Bely, Artem Vesely, Vikenty Veresaev, Zinaida Gippius, Vladimir Korolenko, Boris Pilnyak, Andrei Platonov, Aleksey Remizov, Panteleimon Romanov, Alexander Serafimovich, Fyodor Sologub, Teffi, Aleksey Chapygin, etc.), as well as lesser-known or almost forgotten authors (e.g., Boris Verkhoustinsky, Pavel Zayakin-Uralsky, Vladimir Lensky, Ivan Kolotovkin, Sergei Kolbasiev, Eugene Opochinin, etc.) [Sherstinova, Martynenko 2020].

The corpus is divided into the three following subcorpora, referring to the main historical periods of the era in question [Martynenko et al. 2018b]. Since social backgrounds of these historical periods are very different, we can hypothesize that frequency word lists reflecting the language of these periods will also be different:

- **Period I.** Short stories of the beginning of the 20th century (1900–1913),
- **Period II.** Short stories of the war time and the acute social upheaval (1914–1922) – World War I, the February and October Revolutions and the subsequent Civil War,
- **Period III.** Short stories of the post-revolutionary era (1923–1930).

The total volume of the annotated subcorpus exceeds one million word usage (Period I has about 390,000 words, Period II – 316,000 words, and Period III – 399,000 words). In the annotated part of the corpus, consisting of 310 stories, 77 texts were written by almost forgotten or "rare" authors, so they were specially digitized to be included in the corpus.

Texts of the annotated subcorpus were written in different years and are of different size (as the selection of texts was carried out randomly, these factors were not taken into account). Since this information is important for studying frequency lists and their structure, the distribution of texts by year and by size is shown in Fig. 1-2.

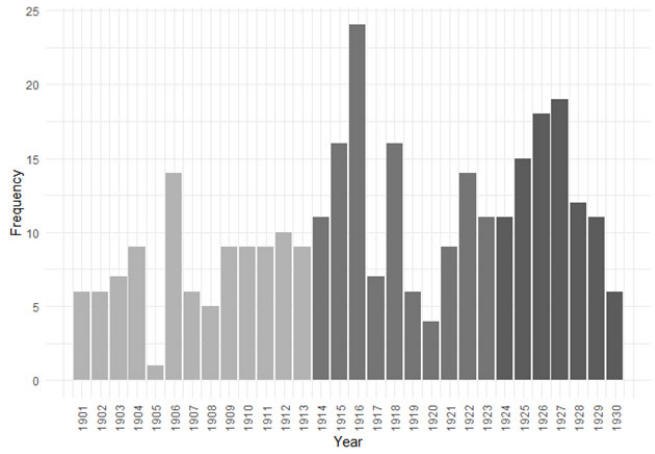

Fig. 1. Frequency distribution of texts over three periods

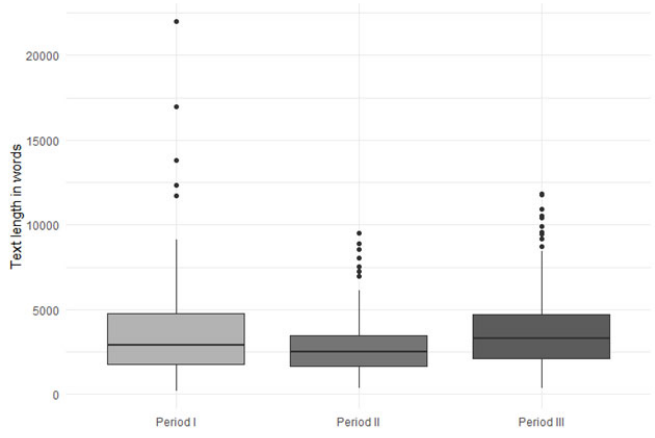

Fig. 2. The distribution of text length of the corpus by periods

The corpus was annotated at the lexical, morphological and, selectively, on syntactic and rhythmic levels. Texts are being segmented into the narrator's speech, the narrator's remarks and the characters' speech. The results of automatic text processing were edited — removal of grammatical homonymy was made as well as segmentation into the narrator's speech and the speech of the characters. The work on expert correction of the results of syntactic annotation continues. For all the stories of the annotated subcorpus, the following elements of literary annotation were introduced: 1) the type of narration (first/third person narration), 2) selected features of narrative structure, and 3) the main theme of each story [Skrebtsova 2019, 2020; Sherstinova, Skrebtsova 2020].

Lemmatization and morphological annotation (part of speech, grammatical form of words) for texts of each story, contextual automatic homonymy resolution were carried out with the use of MyStem program developed by Yandex [Mystem], which is traditionally used to lemmatize texts in Russian (including collections of the Russian National Corpus [RusCorpora]). Selective manual homonymy resolution was made as well.

Based on the results of lexical and POS annotation, frequency lists of word forms, lemmata and POS for each of 310 stories were compiled. Besides, the lists for each subcorpus and a general frequency list for the entire corpus were compiled. Statistical variables were calculated for each frequency list. Thus, statistical "images" of each story were formed in a tabular format and statistical distributions for each variable were obtained. In addition, frequency lists of syntactic constructions (n-grams) for each story, each period and the corpus as a whole were created.

The method of compiling and measuring frequency word lists is considered to be the main approach for lexical analysis in stylometric studies [Martynenko 2019].

The most typical statistical measures of frequency word lists are the following [Sherstinova, Martynenko 2020]:

- 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, for stylometric analysis an extended list of statistics was introduced. 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 (*Mo*).
- 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 ($V_σ$).
- Variation coefficient of the mean linear deviation (V_{Df}).
- Diversity coefficient (K).

Variables in the ordinal (rank) scale:

- Rank mean (R_{ave}) [Martynenko, Fomin, 1989].
- Standard Deviation (σ).
- Coefficient of variation (V_r).
- Median (Me_r).
- Golden Ratio (G_r).
- Mean Deviation (D_r).
- Coefficient of variation for D_r (V_{Dr}).
- Concentration Index (γ) [Sherstinova, Martynenko 2020].

In the following sections we use frequency lists for considering lexical features of Russian literary texts of 1900-1930 and perform statistical text parametrization.

III. Main Lexical Features of Russian Prose

As it was already noted, frequency dictionaries of lemmas and word forms were compiled for each short story, for each of the three subcorpora and for the corpus in the whole.

Table I shows the upper zone (50 most frequent lemmas) of the generalized frequency list of the era under study. The table contains the following data: *rank* of the lemma, *absolute frequency* of the correspondent lemma, *ipm* (items per million), *cumulative percentage* and *rank difference* (calculated by comparing our frequency list with that compiled for the Russian National Corpus) [Lyashevskaya, Sharov 2009].

It can be seen that four the most frequent words cover more than 10% of all texts, whereas the presented list of 50 words covers one third of all texts in concern. The rank difference makes it possible to single out the words that are used in fiction of the periods under review significantly more or less often than in Russian language in general. These words are the following: the verb *stanovit'sya* (*to become*), the nouns *glaz* (*an eye*) and *ruka* (*a hand*), and also the particle *da* (*and* or *yes*).

Time series of word frequencies allow to consider functional activity of correspondent words in diachrony. For example, we may see that the use of conjunction *i* (*and*) decreases over time (see Fig. 3). The services available online allow to compare our data with that obtained for other text collections [Zakharov, Masevich 2014].

Thus, Fig. 4 presents distribution of this conjunction according to Google Ngram Viewer [Google Ngram Viewer]. For the case of *i*, we may notice the evident similarity of trends presented in both graphs. Based on this, we can suppose that the observed fluctuations of frequencies for this conjunction are due to general language changes. This probably indicates that the language strives to grammatical simplification, since the conjunction *i* is multifunctional and can be used to express various grammatical relations.

However, it must be noted that such consent of trends is not always observed. Hypothetically, in cases where the coincidence of dynamic trends in the occurrence of words is not noticed, we can conclude that this particular lexical feature is due to the style of prose. Thus, for instance, the research revealed notable discrepancies in the frequency distribution of the pronoun *on* (*he*).

TABLE I. The Top Zone of Frequency Word List (1900-1930)

Rank	Lemmata	Abs. Frequency	ipm	Cumul. %	Rank diff.
1	*i*	46566	44655	4.47	0
2	*v*	26064	24994	6.96	0
3	*ne*	19141	18355	8.80	0
4	*on*	18136	17391	10.54	n/a
5	*na*	18033	17292	12.27	+1
6	*ya*	15883	15231	13.79	+1
7	*byt'*	12343	11836	14.98	+1
8	*s*	11937	11447	16.12	0
9	*chto*	11714	11233	17.24	0
10	*a*	10795	10351	18.28	0
11	*ona*	9461	9072	19.19	-2
12	*kak*	8224	7886	19.97	-7
13	*k*	6550	6281	20.60	-2
14	*u*	5930	5686	21.17	-7
15	*to*	5537	5309	21.70	+2
16	*po*	5473	5248	22.23	+5
17	*eto*	5221	5006	22.73	+5
18	*za*	5202	4988	23.23	-4
19	*ty*	5189	4976	23.72	-14
20	*oni*	5131	4920	24.22	+3
21	*no*	4928	4725	24.69	+5
22	*vse*	4596	4407	25.13	-13
23	*vy*	4519	4333	25.56	-15
24	*vse*	4518	4332	26.00	-3
25	*etot*	4406	4225	26.42	+11
26	*svoy*	4236	4062	26.83	+1
27	*ot*	4131	3961	27.22	-2
28	*tak*	4121	3951	27.62	-2
29	*iz*	3918	3757	27.99	+9
30	*my*	3811	3654	28.36	+12
31	*zhe*	3700	3548	28.71	-3
32	***da***	**3328**	**3191**	**29.03**	**-84**
33	*skazat'*	3222	3089	29.34	-9
34	*govorit'*	3145	3015	29.64	-24
35	***glaz***	**3077**	**2950**	**29.94**	**-75**
36	***ruka***	**3047**	**2921**	**30.23**	**-38**
37	*odin*	3043	2918	30.52	-11
38	*chelovek*	3039	2914	30.81	-1
39	*yego*	2956	2834	31.10	n/a
40	*tol'ko*	2943	2822	31.38	-3
41	*o*	2825	2709	31.65	+10
42	*yeshche*	2798	2683	31.92	-3
43	*sebya*	2717	2605	32.18	-4
44	*vot*	2599	2492	32.43	-13
45	*kotoryy*	2586	2479	32.68	+23
46	*kogda*	2502	2399	32.92	-9
47	*tot*	2432	2332	33.15	+11
48	***stanovit'sya***	**2420**	**2320**	**33.38**	**-356**
49	*moch'*	2417	2317	33.61	+12
50	*by*	2371	2273	33.84	+4

Such results allow us to trace the features of the short story as a literary genre and its lexical richness in comparison with language in general, relevance of thematic areas, etc. The

proposed methodology can be applied to any word of the corpus and, in principle, to any other corpus.

Table II summarizes statistics for each three periods in concern as well as for the corpus in the whole.

TABLE II. THE MAIN STATISTICS FOR 3 PERIODS AND TOTALLY

Statistics	Totally	Period I	Period II	Period III
Tokens	1042794	362423	284273	396098
Types	40791	21879	21611	28445
Hapax	15283	8431	8885	11611
Multiple lemmata	25508	13448	12726	16834
TTR	0.039	0.06	0.076	0.071

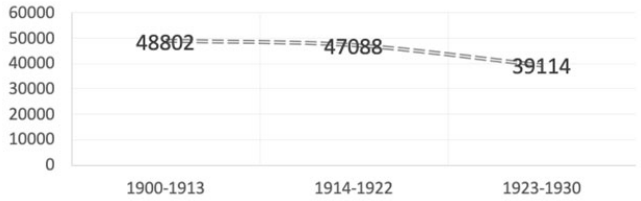

Fig. 3. Frequency distribution of the conjunction *i* (*and*) in the corpus of Russian story over three periods

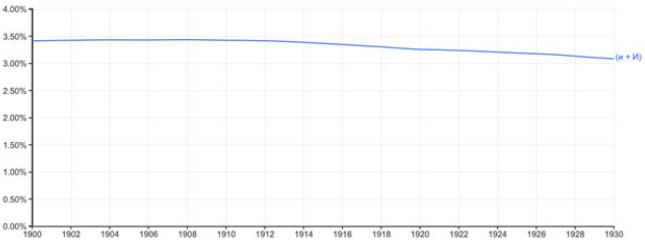

Fig. 4. Frequency distribution of the conjunction *i* (*and*) according to Google Ngram Viewer for the same period

IV. POS FREQUENCY LISTS

Part of speech annotation was made by mean of MyStem program [MyStem]. The program uses the following POS categories: S (noun), V (verb), PR (preposition), CONJ (conjunction), SPRO (pronoun/noun), A (adjective), ADV (adverb), PART (particle), APRO (pronoun/adjective), ADVPRO (pronoun), NUM (numeral), INTJ (interjection), ANUM (numeral adjective), COM (part of a composite word) [Segalovich, Titov 2011]. Frequency lists of POS categories are presented in Table III.

TABLE III. POS DISTRIBUTION

Rank	POS	Totally, %	Period I, %	Period II, %	Period III, %	Rus-corpora
1	S	26.75	24.56	26.53	28.80	30.6
2	V	19.95	19.66	19.87	20.27	15.8
3	PR	10.35	9.88	10.28	10.80	10.9
4	CONJ	8.77	9.35	9.14	7.97	7.6
5	SPRO	8.11	9.19	8.12	7.96	6.8
6	A	7.86	8.05	7.57	7.19	9.3
7	ADV	6.33	6.72	6.42	5.94	5.0
8	PART	4.70	4.79	4.87	4.48	3.8
9	APRO	4.14	4.69	4.17	3.64	4.5
10	ADVPRO	1.95	2.10	1.96	1.83	1.9
11	NUM	0.57	0.53	0.56	0.60	2.0
12	INTJ	0.24	0.22	0.25	0.26	0.1
13	ANUM	0.20	0.19	0.19	0.21	0.4
14	COM	0.0015	0.00030	0.0014	0.0025	0.9

As follows from the table, in total, Russian fiction contains mainly nouns, verbs and auxiliary parts of speech: prepositions and conjunctions. The columns Periods I-III show POS distribution obtained for each of the periods. These data clearly demonstrate that it is indeed possible to trace certain trends in the distribution of parts of speech in the analyzed periods. Table III demonstrates that the use of the "main" POS – nouns, verbs and prepositions – constantly increases. Over time, the percentage of conjunctions, pronouns and adjectives decreases. The reverse trend is observed for distribution of interjections – their use increases with each period.

Thus, it cannot be said that POS distribution remains unchanged throughout the first third of the 20th century. However, statistical tests did not allow to consider these differences as the significant ones. It is required to repeat the study on larger samples. In our case, we can assume that the shares in the general populations do not differ or differ insignificantly. This means that time factor acts rather weakly on POS distribution in the selected time intervals.

V. STATISTICAL TEXTS PARAMETRIZATION

In order to perform statistical texts parametrization, the texts have been sequentially combined by ten, and frequency word lists (in frequency descending order) have been made for those cumulative text groups. For every frequency list we fix its size (the number of lexemes) and calculate the rank mean (r) – a parameter for rank distribution – by the following formula:

$$r = \Sigma r \cdot f_r / N,$$

where r — rank, f_r — frequency for the rank, N — sample size.

The data obtained have been approximated by the technique on the base of the least square method developed by G.Ya. Martynenko [Martynenko 1988].

The Weibull function, as an analogue of Zipf's law, has been chosen as an approximation one:

$$N = N_{max} - N_{max} e^{-cx^d},$$

where N is a value of the parameter under investigation, x — sample size, N_{max} — asymptotic value of the parameter under investigation, and c, d — distribution parameters.

The empirical and theoretical values of the parameters in question for every period and all periods combined are shown in the Tables IV–VII.

TABLE IV. PERIOD I (1900-1913)

Frequency List Size		Rank Mean	
Empirical	Theoretical	Empirical	Theoretical
7 720	7 831	1001.43	1153.24
11 130	10 859	1255.76	1259.29
13 384	13 438	1330.45	1321.17
15 044	15 178	1422.45	1353.18
17 029	16 865	1487.86	1378.68
19 053	19 049	1423.13	1405.38
20 716	20 828	1396.59	1422.97
21 703	21 723	1397.31	1430.66
22 779	22 937	1389.53	1440.01
24 316	24 170	1406.92	1448.39

TABLE V. PERIOD II (1914-1922)

Frequency List Size		Rank Mean	
Empirical	Theoretical	Empirical	Theoretical
4 864	4 888	775.73	860.14
8 190	8 188	1088.16	1177.60
11 190	11 118	1328.56	1354.89
14 956	14 885	1589.55	1490.31
17 229	17 304	1621.67	1541.64
18 823	18 819	1583.75	1564.19
19 920	20 043	1550.61	1578.29
21 686	21 717	1528.57	1592.91
23 457	23 494	1519.71	1603.86

TABLE VI. PERIOD III (1923-1930)

Frequency List Size		Rank Mean	
Empirical	Theoretical	Empirical	Theoretical
6 004	6 240	970.50	1329.62
10 452	10 044	1489.53	1469.76
15 010	14 562	1875.72	1577.29
17 699	17 752	1838.33	1633.34
19 941	20 343	1725.84	1671.20
23 387	23 198	1704.95	1707.12
25 046	25 206	1682.17	1729.53
26 500	26 683	1681.56	1744.78
28 257	28 441	1694.84	1761.74
30 148	30 148	1692.77	1777.14

TABLE VII. ALL PERIODS

Sample Size	Frequency List Size		Rank Mean	
	Empirical	Theoretical	Empirical	Theoretical
45 896	7 720	6 461	1001.43	1110.22
81 437	11 130	9 761	1255.76	1221.82
119 721	13 384	12 751	1330.45	1290.02
150 017	15 044	14 837	1422.45	1326.61
183 118	17 029	16 900	1487.86	1356.60
231 871	19 053	19 616	1423.13	1389.06
276 980	20 716	21 853	1396.59	1411.21
301 641	21 703	22 983	1397.31	1421.13
337 373	22 779	24 517	1389.53	1433.43
376 513	24 316	26 073	1406.92	1444.71
397 187	25 005	26 849	1412.71	1449.92
422 397	25 392	27 754	1425.11	1455.69
451 405	26 894	28 745	1437.83	1461.65
475 475	27 767	29 529	1443.50	1466.12
498 333	28 601	30 244	1442.18	1470.02
534 794	29 773	31 328	1439.33	1475.62
586 054	30 654	32 744	1446.01	1482.44
608 478	31 299	33 328	1456.11	1485.09
641 843	32 426	34 159	1461.84	1488.72
679 414	33 283	35 045	1466.93	1492.40
704 432	34 609	35 608	1471.32	1494.65
729 568	35 063	36 152	1530.37	1496.76
756 658	36 258	36 717	1484.53	1498.88
799 506	37 713	37 566	1495.53	1501.95
837 505	38 772	38 277	1500.79	1504.41
873 798	39 798	38 921	1505.87	1506.57
920 305	41 625	39 701	1522.75	1509.08
957 765	42 514	40 294	1521.34	1510.92
988 191	43 369	40 755	1519.98	1512.31
1 027 988	44 390	41 330	1522.36	1514.00
1 070 871	45 533	41 917	1530.37	1515.67
1 077 970	45 813	42 012	1532.53	1515.94

The data having been approximated, the following theoretical values for the parameters in question have been found (Table VIII).

TABLE VIII. MAXIMUM THEORETICAL VALUES OF THE PARAMETERS

Period	Frequency List Size	Rank Mean
1900-1913	59 064	1 506.92
1914-1922	54 129	1 624.71
1923-1930	54 556	1 967.40
Totally	54 275	1 542.53

Given the asymptotic values for the parameters in question and using the inverse Weibull function we are able to find a sample size for its parameters to meet the limit ones in the most accurate way with the differences by 1% and by 1 (Table IX).

TABLE IX. SAMPLE SIZE REQUIRED TO REACH LIMIT VALUES

Period	99%	Difference by 1
Frequency List Size		
1900-1913	12 332 021	49 937 860
1914-1922	5 781 793	20 166 720
1923-1930	4 346 587	14 754 373
Totally	4 588 703	13 847 957
Rank Mean		
1900-1913	935 629	3 130 871
1914-1922	302 840	611 121
1923-1930	4 434 696	27 725 829
Totally	1 518 023	5 395 474

The data obtained allow us to make the following conclusions:

1) The frequency list size is a relatively consistent parameter for frequency vocabulary of fiction. Despite being observed as converging to a limit value the rate of the convergence is quite slow. Total and even 99% stabilization of the list sizes occurs with the sample sizes which are far beyond the real corpus size.

2) The parameter which demonstrates absolute statistical consistency for the frequency list of fiction in general and short stories in particular is the rank mean – a measure of frequency concertation in the upper zones of the list. In the vast majority of cases, 99% and total stabilization of the rank mean can be observed with the sample sizes which are practically available in the corpus.

3) Theoretical values of the parameters in question are sufficiently close to each other both for the particular periods and for the all periods combined, which confirms statistical consistency of the rank mean once more.

4) At the same time, it should be noted that the rank mean can be used as an indicator of thematic growing which is characteristic for dramatic historical periods (cf. stabilization rate for particular periods).

VI. KEYWORDS AND THEIR FEATURES

A. Keyness

Keyness is a statistical measure that reveals the most meaningful and significant words of analyzing text data [Scott and Tribble, 2006]. This indicator can help to disclose the style of a particular text, text corpus, even several corpora, as well as of some author's style. Lists of keywords were

calculated for each of the periods by means of AntConc analysis [Anthony 2019]. For compiling keyword lists, the program provides the possibility to use various statistical settings. Keywords were calculated using different statistical measures, and the obtained results were compared. Then, this data was compared with the results obtained by corpus processing using TXM platform, which employs a "specificity" measure for comparing subcorpora.

B. Lexical Specificity

Lexical specificity is a score of a word being present *f* times or more in a subcorpus of *t* words given that it appears a total of *F* times in a whole corpus of *T* words [Specificities]. This measure is being calculated by the TXM software designed for textometric analysis. TXM is an open source software designed for the preparation, processing, analysis and publication of medium-sized corpora (up to 10,000,000 words). The toolkit includes the CQP search engine, the R statistical analysis platform, and TreeTagger automatic morphological markup and lemmatization package, which requires separate installation of packages for each language. The platform also supports a large number of different formats (from TXT to TEI XML) and combines quantitative and qualitative analysis tools [Heiden 2010].

The textometric analysis of texts using the TXM software package was carried out. The obtained "keywords" are compared with the vocabulary list defined by TXM through the "measure of specificity", and a significant coincidence of the results was found.

C. Keywords and Specific Words of Russian Fiction

Results obtained by AntConc text processing show that for the Period I the most significant "keywords" which were calculated with chi-squared and log-likelihood measures are pronouns: *ona* (*she*), *on* (*he*), *ejo* (*her*), *ego* (*his*), *ej* (*her*), *ya* (*I*). In this period pronouns (in TXM also) have the highest value of keyness.

In Period II the military vocabulary is highlighted: *nemcy* (*Germans*), *oficer* (*officer*), *plenniki* (*prisoners*) — these words have the highest value of keyness in AntConc; in TXM there are: *pul'ka* (*bullet*), *nemeczkij* (*German*), *shinel'* (*trench coat*), *po-nemeczki* (*in German*), *praporshhik* (*warrant officer*) — in addition to words obtained by AntConc analysis. These results are predictive. In short stories of this period, writers wrote about the war and people, who suffered in those days.

In revolutionary period AntConc proposes the following "keywords": *kandaly* (*shackles*), *russkij* (*Russian*), *vintovka* (*rifle*), *kolonii* (*colony*), *kamera* (*ward*), *soldat* (*soldier*), *strelyat'* (*to shoot*), *tolpa* (*crowd*), *lyudi* (*people*) — which show that for this period words connected with unrest and rebels are significant. TXM highlighted the following words: *Sovet* (*Soviet*), *sudebnuyu* (*judicial*), *katorga* (*penal servitude*), and *internacional'naya* (*international*). In wartime, it was observed that writers used the words, which are specific only for war theme.

In the last Period III, when the Soviet state was in formation, the most meaningful words in AntConc turned out to be the following: *komissar* (*commissar*), *ded* (*grandfather*), *tovarishh* (*comrade*), *muzhik* (*muzhik, peasant*), *komandir* (*commanding officer*), *predsedatel'* (*chairman*), *rabotat'* (*to work*). TXM provided the following words: *sel'sovet* (*village council*), *grazhdane* (*folks*), *fabrika* (*factory*), *glava* (*leader*), *protocol* (*protocol*). All these words are also distinctive for the period under consideration.

D. Expert keywords selection

In addition to automatic keywords selection basing on frequency lists, expert selection of keywords was carried out, which was based on the analysis of upper ranks of frequency lists. It revealed a number of significant words that differ in increased frequency against the background of data from both author dictionaries and a common language dictionary (for example, *crowd, children, soul, heart, feeling, god, thought, keep silent*). Presumably, this set of words is related to the topic of stories or, more often, to the way of presenting informative material. For more details, see [Grebennikov, Skrebtsova, 2019].

VII. Word Frequency Lists and Authors Clusterization

Cluster analysis of the corpus was conducted for each period separately and for the corpus in the whole using the Stylo package for R programming language [Eder et al. 2016]. Ward's method was used in hierarchical clustering [Ward 1963]. Frequency lists were created for different levels of culling. Culling means the minimum percentage of texts where the word must be in to be included in a frequency list. Then the distance tables between texts have been computed. Clusters were built following on from the results. The interpretation of the obtained classification was based on visualization of the distance table by means of dendrograms. Whereas lexical content of frequency lists is largely correlated with the subject of the stories, it seems relevant further to perform cluster analysis of texts considering their thematic markup [Skrebtsova 2020].

Different levels of culling (from 0% to 90%) show different levels of detail, lexical preferences for authors and lexical characteristics of a specific text, which is correlated with its subject. Tables of lexical proximity between authors were made for each of the three analyzed periods, as well as for the whole corpus with a step of 10% — from 0% to 90%. The number of the most frequent words was determined for each level of culling and the lists of these words were obtained. Distance tables have been computed using different distance measures (e. g., Canberra, Euclidean, Manhattan, Delta methods, etc.). These measures have some difference in outputs which results in different clusters. Visualization of distance tables is done with dendrograms.

The next step of analysis is to be the comparison of obtained clusters with the results of expert analysis. Then it will be possible to choose the optimal model of clustering. The example of the dendrogram is presented in Fig. 5. It is built on the level of culling which equals to 10% (i. e., 2593 most frequent words were considered) and using Classic Delta Distance [Burrows 2002]. The proximity of the writers on the dendrogram (e. g., Verkhoustinsky and Chapygin) shows the similarity of the frequency dictionaries of the analyzed texts.

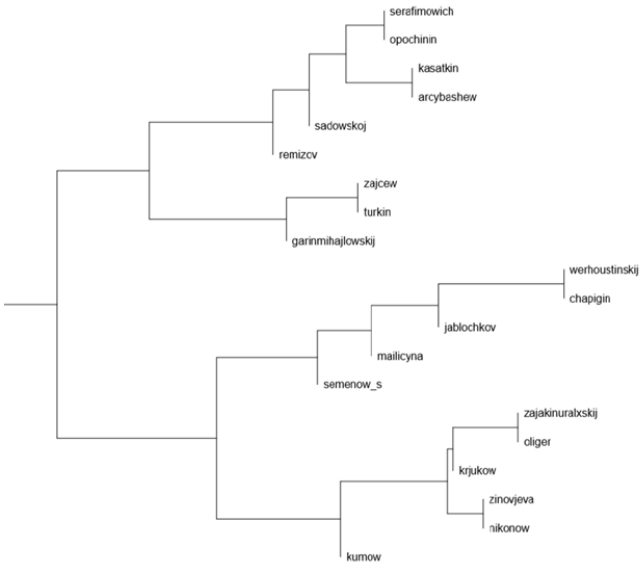

Fig. 5. One of the writers' clusters (the fragment of the dendrogram)

VIII. Conclusion

The article presents frequency dictionaries of various types (word forms, lemmas, POS) for successive historical periods and for three decades of 1900-1930 as a whole), obtained on the basis of texts of Russian short stories. Their general statistics are given.

Parametrization of the word frequency lists has been made. The dependence of the dynamic vocabulary on the sample size has been analyzed, and statistical consistency of the word list parameters has been tested. Theoretical and empirical values of parameters have been compared to each other, and the sample sizes necessary to reach the limit values of the parameters have been determined.

The obtained results are original and unique, since they are based on the application of modern linguistic technologies and machine learning methods to the unique corpus of fiction texts. The significance of these results for the development of interdisciplinary fields is determined by the fact that they test modern NLP tools on the basis of literary texts, as well as by the combination of humanitarian and "exact" research methods.

Practical relevance of the obtained results lies in the fact that they can be used to address the NLP challenges, in particular, in numerous practical applications related to attribution of authorship, text automatic clustering and classification.

Acknowledgment

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

[1] P.M. Alekseev, *Chastotnyye slovari: Uchebnoye posobiye* [*Frequency Dictionaries: Textbook*]. St. Petersburg: Publishing House of St. Petersburg University, 2001.
[2] L. Anthony, *AntConc 3.5.8*, Tokyo, Japan: Waseda University, 2019.
[3] P. Baker et al., *Glossary of Corpus Linguistics*, Edinburgh University Press, 2006.
[4] J.F. Burrows, ""Delta": a measure of stylistic difference and a guide to likely authorship", *Literary and Linguistic Computing*, 2002, 17(3), pp. 267–87.
[5] Google Ngram viewer. Web: https://books.google.com/ngrams.
[6] A.O. Grebennikov, "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, Vol. 122, 2007, pp. 256–259.
[7] A.O. Grebennikov, T.G. Skrebtsova, "Jazykovaja kartina mira v russkom rasskaze nachala XX veka" ["The linguistic picture of the world in the Russian story of the early XX century"], *Filosofija i gumanitarnye nauki v informacionnom obshhestve*, 3(25), 2019, pp. 141-143.
[8] M. Eder, J. Rybicki, M. Kestemont, "Stylometry with R: a package for computational text analysis", *R Journal*, 2016, 8(1), pp. 107-121.
[9] S. Heiden, "The TXM Platform: Building Open-Source Textual Analysis Software Compatible with the TEI Encoding Scheme", *24th Pacific Asia Conference on Language, Information and Computation - PACLIC24*, Institute for Digital Enhancement of Cognitive Development, Waseda University, Sendai, Japan, 2010, pp. 389-398.
[10] H.H. Josselson, *The Russian Word Count and Frequency Analysis of Grammatical Categories of Standard Literary Russian*, Detroit, Wayne University Press, 1953.
[11] O.N. Ljashevskaja, S.A. Sharov, *Chastotnyj slovar' sovremennogo russkogo jazyka* (*na materialah Nacional'nogo korpusa russkogo jazyka*) [*Frequency dictionary of modern Russian language* (*based on the material of the Russian National Corpus*], Moscow, Azbukovnik, 2009, Web: http://dict.ruslang.ru/.
[12] L. Lönngren, *The Frequency Dictionary of Modern Russian*. Acta Univ. Ups., Studia Slavica Upsaliensia Uppsala, 1993.
[13] G.Ya. Martynenko, *Metody matematicheskoj lingvistiki v stilisticheskih issledovaniyakh* [*Methods of mathematical linguistics in stylistic studies*]. Nestor-Istoriya, 2019.
[14] G.Ya. Martynenko, S.V. Fomin, "Ranking moments", *Nauchno-tekhnicheskaya informatsiya, Seriya 2 – Informatsionnye Protsessy i Sistemy*, Issue: 8, 1989, pp. 9–14.
[15] G.Ya. Martynenko, G.A. Martinovich, *Mnogoparametricheskij statisticheskij analiz rezul'tatov associativnogo eksperimenta* [*Multiparametric statistical analysis of the results of an associative experiment*]. St. Petersburg: Publishing House of St. Petersburg University, 2003.
[16] G.Ya. Martynenko, T.Yu. Sherstinova, T.I. Popova, A.G. Melnik, (2018a) "Metodologicheskie 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)", *Computational linguistics and computational ontologies*, Iss. 2 (Proc. of the XXI Int. United Conf. IMS-2018), ITMO University, St. Petersburg. Pp. 99–104.
[17] G.Ya. Martynenko, T.Yu. Sherstinova, T.I. Popova, A.G. Melnik, E.V. Zamirajlova, (2018b) "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'*, 2018, pp. 180–197.
[18] MyStem — Yandex Technologies, Web: https://yandex.ru/dev/mystem/.
[19] I.I. Popescu, *Quantitative Linguistics: Word Frequency Studies*. Berlin-New-York: Mouton de Gruyter, 2009.
[20] [RusCorpora] National Corpus of the Russian Language, Web: www.ruscorpora.ru.

[21] M. Scott, C. Tribble, *Textual patterns, Key words and corpus analysis in language education*, Amsterdam/Philadelphia: John Benjamins, 2006.
[22] I. Segalovich, V. Titov, *MyStem*, Russia, Moscow, 2011, Web: https://yandex.ru/dev/MyStem/
[23] A.Y. Shaykevich, "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'*], 2015, pp. 422–429.
[24] T. Sherstinova, G. Martynenko, "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. Conf. on Language Engineering and Applied Linguistics (PRLEAL-2019)*, Nov. 2019, CEUR Workshop Proceedings, Vol. 2552, 2020, pp. 105–120.
[25] 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 International Workshop «Computational Linguistics» CompLing-2020* (in print).
[26] T.G. Skrebtsova, "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, 2019, pp. 426–431.
[27] T.G. Skrebtsova, "Thematic Tagging of Literary Fiction: The Case of Early 20th Century Russian Short Stories", *Proc. of the International Workshop «Computational Linguistics» CompLing-2020* (in print).
[28] [Specificities]: *Calculate Lexical Specificity Score* Web: https://rdrr.io/cran/textometry/man/specificities.html.
[29] E.A. Steinfeldt, *Frequency dictionary of modern Russian literary language* [*Chastotnyj slovar' sovremennogo russkogo literaturnogo jazyka*]. Tallinn, 1963.
[30] Y.A. Tuldava, "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, 1986, pp. 139–162.
[31] E. Voronchak, "Metody vychisleniya pokazateley leksicheskogo bogatstva tekstov" ["Methods of calculating indicators of the lexical richness of texts"], Semiotics and artmetry. Moscow, 1972, pp. 232–250.
[32] J.H. Ward, "Hierarchical Grouping to Optimize an Objective Function", *Journal of the American Statistical Association*, no. 58, 1963, pp. 236–244.
[33] V.P. Zakharov, A.C. Masevich, "Diahronicheskie issledovaniya na osnove korpusa russkih tekstov Google Books Ngram Viewer" ["Diachronic studies based on the corpus of Russian texts Google Ngram Viewer"], *Strukturnaya i prikladnaya lingvistika*, vol. 10, St. Petersburg, 2014, pp. 303-327.
[34] L.N. Zasorina (ed.), *Frequency dictionary of the Russian language*, Moscow: Russian language, 1977.