Using TXM Platform for Research on Language Changes over Time: the Dynamics of Vocabulary and Punctuation in Russian Literary Texts

Alexei Lavrentiev, Tatiana Sherstinova, Andrey Chepovskiy, Bénédicte Pincemin

A.M. Lavrentiev, T. Yu. Sherstinova, A. M. Chepovskiy, B. Pincemin

Using TXM Platform for Research on Language Changes over Time: the Dynamics of Vocabulary and Punctuation in Russian Literary Texts[1]

The purpose of this paper is to test the methodological tools provided by TXM open-source software for research on dynamics of vocabulary and punctuation marks in diachronic corpora. TXM provides both quantitative and qualitative analysis features. It is shown that Russian revolution of 1917 did make significant changes in the core vocabulary of the corpus of Russian Short Stories (1901-1930). The same methodology may be used both for diachronic studies of literature and for various NLP tasks.

Keywords: *stylometry, textometry, TXM platform, corpus linguistics, Russian literature of the 20th century, vocabulary, punctuation, diachronic linguistics*.

Introduction

The purpose of this paper is to test the methodological tools provided by TXM platform[2] for research on dynamics of vocabulary and punctuation marks in diachronic corpora. The idea of this research was inspired by the concept of literary corpus proposed within Russian stylometric school by its leader Gregory Martynenko [1; 2]. The corpus is designed by its developers to become a testing ground for various text computation techniques, besides it should allow studying changes that occur in the language at the crucial historical moments [3] (see below section *The corpus and its sample*).

The purpose of this research is to apply TXM textometric analysis to its subcorpus referring to Russian literature of 1901–1930. It should allow us to compare and enhance the obtained results with those that are gained by other computational techniques and software. The other aim is to check if the results of this analysis can help reveal trends in literary language and Russian literature in the whole and to evaluate if deeper and more precise analysis is worth undertaking. The method consists in examining correlations between the frequency of word forms, lemmas, punctuation

[1] The reported study was funded by RFBR in the framework of research projects # 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)" and # 19-07-00806 "Research and development of methods and algorithms for complex linguistic analysis of special text corpora".

[2] TXM is the full name of the software developed by the Textometry scientific project, see http://www.textometrie.org.

marks, etc. and various external and properly literary factors. Among the external factors we consider first of all the important historical events, such as World War I (WWI), the October Revolution (OR) and the Civil War (CW) that followed it.

Our basic hypothesis is that despite the differences among individual texts, authors and styles, a general chronological evolution exists in the literary process and that some historical events cause changes in literary language that lead to formation of distinct periods. As a first step, in this paper we adopt an exploratory statistical approach [4], in order to discover from data which time divisions as well as which language features happen to be relevant, and thus provide hints for further confirmatory research.

The methodology presented in this paper may be used both for diachronic studies of literature and for various NLP tasks connected with texts processing and monitoring over time with the aim of revealing linguistic, stylistic and sentiment changes in texts influenced by some external factors such as historical events or cultural and intellectual trends.

Methodology

Information technologies for Corpus Linguistics

Automated tools for natural language text processing are applied for such tasks as text classification, authorship attribution, discovering underlying rules of natural language, building and applying models of language structure (see for instance [5, 6, 7]).

Corpus analysis software, such as TXM platform considered in this paper, is necessary for statistical analysis of the vocabulary and for retrieval and comparison of various linguistic patterns. It allows the computation of statistical characteristics of such constructions and analysis of different parts of the corpus (subcorpora). As we have shown in [8; 9], the methods of corpus analysis may be applied to examine the possibility of uncovering various differentiating features for thematic text classification, as well as to create training dataset samples for natural language text recognition.

The selection of differentiating features is a key problem for classification tasks [10] and for building correct training samples. The necessity to select a limited number of differentiating features is due to the problem the training sample increase in size as the number of features grows. In order to estimate precisely enough, it is required that each feature used would occur at least several times, which is practically impossible, for instance, in the case where all known words of the Russian language are used as a set of features. When working with relatively short texts of different authors, where considerable variation may be observed in the use of words and syntactic

constructions, the selection of differentiating features is particularly important. This is why application of corpus analysis methods for pre-processing of training datasets is so topical.

The corpus and its sample

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 [3; 11; 12]. It is intended for stylometric, linguistic and literary studies of Russian prose of 1900–1930s. The main task of corpus developers is to create a model of literary corpus, which implies the inclusion in the corpus texts of the maximum number of writers, who created their works in the corresponding era — both well-known and peripheral. The other task is to create an empirical base for studying the language and style of Russian prose in synchrony and diachrony for a given time period, as well as for conducting stylometric analysis of literary texts on phonetic (rhythmic), lexical, syntactic, semantic and structural levels [3].

The interest in this time period is determined by the fact that the first three decades of the 20th century were saturated by a series of acute social upheavals, which led to dramatic changes both in the language and in the style of fiction [ibid.]. Therefore, an important task of creating the corpus is quantitative analysis of linguistic and stylistic changes that have occurred in fiction as a result of social disasters. Particular attention is paid to the list of linguistic parameters which are planned for text data processing, such as frequency parameters (calculated for word forms, lemmas, parts of speech [13], syntactic structures [14], rhythmic patterns [15], text composition [16]) and structural parameters (word length, sentence length, paragraph length [17], text length, measures of syntactic complexity [18], the internal dynamics of the text, etc.). It should be noted that the texts originally published in old Russian spelling (prior to the reform of 1918) have been normalized according to modern spelling rules, for homogeneity considerations [19]. This is the general practice in Russian text editing, including scholarly editions, and in Russian text corpora [20: 36-37]. The proposed corpus model can be extrapolated to the creation of similar resources for other literary genres, journalism and other genres of texts, not only written, but also oral [11].

Our current research sample includes 308 texts (entire or samples). All of them belong to the "annotated" subcorpus of the project, the list of short stories for which is given in [3]. However, at this stage we did not use any pre-annotation and only relied on the NLP tools provided by TXM. For each text, the author's name and the year of publication are available as metadata. Most of the 298 authors are represented by one text (one author has three texts in the sample, and nine have two texts). Among authors one can find both major writers (such as Bunin or Gorky) and persons that are practically unknown to the general public.

A diagram of distribution of texts according to their size is shown in Fig. 1. Each text is represented by a bar, texts are ordered by descending size, and the length of the bar corresponds to the number of words. Most of them are between 1,800 and 5,000 words long, but some texts are as short as 500 words or as long as 15,000 words or more. We keep this length diversity because it is inherent to real data we want to observe. We deal with it on the one hand by grouping texts into years, which smoothes singularities, and on the other hand by choosing statistical methods that take into account size variation.

At the year level, each year is represented by several texts, and for most years the number of words ranges from 20,000 to 40,000. The corpus is therefore not perfectly balanced but it seems to be representative enough to study the evolution of short stories language and style over the selected period of time. The genre of short stories was chosen precisely because it allows collecting a relatively big amount of texts all over the period under consideration and in most cases does not require sampling.

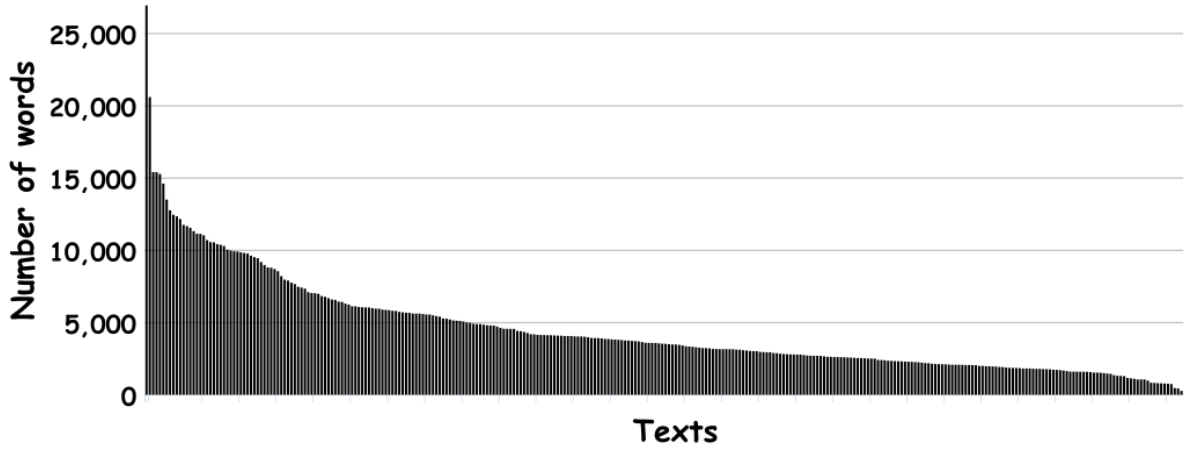

Fig. 1. Distribution of texts according to their size in words: texts are ordered by descending size, and each bar stands for a text.

The 1st third of the 20th century is a particularly interesting period in the history of Russian literature, as it starts with the “silver age” of Russian poetry and numerous innovations in the method and style of writing, goes through such dramatic socio-political events as WWI, the revolutions of 1917, and the CW, and ends in 1930s, when “socialist realism” became dominant under the pressure of communist ideology.

TXM platform

To analyze the corpus we used free and open-source TXM software platform [21; 22]. It represents the latest generation of software developed by the international community of Textual

Data Analysis, also known as Textometry. TXM is distributed in the form of a software application for all major operating systems (Linux, Windows and Mac OS) and as a GWT-based [23] application for a web server. The platform provides a wide range of tools for corpus creation, annotation, analysis and publication. It can import source texts from clipboard, simple text files (UTF-8 or other encoding), Microsoft Word and LibreOffice Writer documents, XML-tagged texts (in particular TEI XML) and some proprietary formats of corpus processing software (Alceste, Hyberbase). Metadata can be provided in a separate spreadsheet (CSV, XLSX or ODS). TreeTagger, available as an extension, may be used for automatic morphosyntactic tagging and lemmatisation [24].

Once the corpus is imported to TXM, further annotation of word properties, structural units or reference chains may be performed. TXM uses CQP [25] as a search engine for making frequency lists and advanced KWIC concordances, and R [26] for statistical analysis. Hyperlinks that exist between different TXM objects allow the user to move smoothly from quantitative tools (such as frequency lists) to qualitative interpretation (using concordances or browsing the full text). Subcorpora and partitions may be created in different modes for contrastive analysis. In this research we used two statistical tools provided by TXM — correspondence (factor) analysis and specificity.

Correspondence analysis (CA) [27; 28] is widely used in the textual data analysis community [29; 30]. Papers using CA are regularly presented at the JADT Textual Data Statistical Analysis conference [31]. Compared to other multi-dimensional analysis methods (such as principal components), it is specially designed for contingency tables, for instance those crossing texts and words [32; 33]. It may be applied to word forms, lemmas or part-of-speech tags of the corpus. In the field of textual data analysis, CA is almost entirely devoted to produce overviews through visualization of 2D factorial maps (and experimentation of 3D ones). The summary-oriented use is also based on information compression (selection of the first dimensions) and noise reduction (elimination of the last dimensions). Such a holistic approach is even required in case of Guttman effect [34] on serially-structured corpora [35].

The *specificity score* [36; 32. P. 130-136] is a non-parametric statistical test that helps evaluate the significance of the frequency of a word (or pattern) in a part of a corpus. For instance, the verb *kazat'sia* 'to seem' which counts 499 occurrences in 1901-1013 (in 942 occurrences for the whole corpus) is scored with a 25.7 specificity, which is much higher than the significance threshold of 3. The specificity score is similar to such widely used keyword tests as chi-square, log-likelihood or t-score, but it has the advantage of modeling precisely the distribution of a discrete population (in

mathematical sense) and thus allows understanding precisely the deviation it measures. The other statistical models quoted above use distribution hypotheses (approximations) that allow evaluating deviations, while specificity implements a Fisher's exact test [37]. It is generally admitted to be theoretically superior but the other tests are still used for practical reasons (approximation tests were implemented in popular software when the processing power of available computer was much lower than now) [38; 33. P. 122].

Both CA and specificity score can be applied to data that present some internal size variation: methods align with real data, rather than data align with methods. Actually, part sizes in the corpus must not be normalized, because this would distort the statistical model which takes into account raw frequencies. However, it is important to be aware of these variations to draw accurate interpretations from statistical results.

All the results provided by TXM can be exported in CSV tables and/or SVG graphic files.

In addition to the tools available through the user interface, it is possible to use the ones provided with TXM or produce new Groovy scripts and macros. TXM is efficient on medium size corpora (up to 10 million words approximately) but can be customized for working with hundreds million words.

The corpus under consideration was imported into TXM (0.7.9 version, using the XTZ+CSV module), and TreeTagger was applied to annotate every word with a morphological tag and lemma, punctuation marks were tagged as well. We used the parameter file for Russian provided on the TreeTagger website, trained on a corpus by Serge Sharoff [39].

Results and Discussion

Correspondence analysis

First, we built a partition of the corpus based on the year of publication. In this partition each year is a subcorpus that can be compared to the other subcorpora and to the corpus as a whole. As already mentioned, the corpus is relatively balanced in this respect (20,000 to 40,000 words per year). However, some years are over-represented (1927: 65,591 words, 1916: 60,630 words, 1928: 59,247 words). One year is particularly under-represented (1920: 9,749 words) and another year is totally absent (1905). We have been particularly attentive to the bias that this imbalance might create in the analysis.

A CA was performed on the dataset (lexical table) of the 200 most frequent lemmas and punctuation marks. Such an approach allows us to focus on function words and the most common

content words, and to minimize the effect of proper nouns and other words that are highly specific to a particular text or an author. Punctuation marks are analyzed in the same way as words. In our opinion, they can bear information on syntactic complexity and individual author's style. However, we also check what influence the punctuation and proper nouns have on the CA and whether the results on proximity of corpus parts hold if these categories are excluded.

The results of the CA are presented in a factorial map (Fig. 2). The axes on the map represent the first two dimensions obtained as a result of information compression. It should be noted that TXM also allows displaying the 3rd dimension instead of the 1st or the 2nd one. The first dimension is mathematically designed as the 1-D best representation of data (maximization of inertia). The second dimension is mathematically designed as to orthogonally complete the first one to provide the 2-D best representation of data, and so on. Thus, dimensions have no given *name* (they are named by their rank) but, after examination of the analysis results and indicators, they may be *interpreted* relatively to initial data. Axes cross at the center of gravity of the data cloud. The percentage provided for each axis indicates its weight in the *inertia* (variance) of the whole corpus: in Figure 2 for instance, the two first axes of the 28-dimension space[3] capture one third (20.07 % + 13.91 % = 33.98 %) of the total variation information of the corpus. The direction of axes has no particular meaning: a horizontal or vertical flip of the chart has no effect on their interpretation. Only the positions of points relative to each other count.

What is special with CA is that both categorical variables that are crossed in the data table are represented in the same geometrical space. In Figure 2, both the 29 years (parts of the corpus) and the 200 most frequent lemmas are displayed. The chart is hardly readable, as many of the 230 labels overlap. TXM users can zoom in and out the charts, and click on the labels to get more information but this is impossible in print. Therefore, one can choose, as in Figure 3, to display only one of the two sets, in order to focus on this set and to improve the readability of the chart. The coordinates of the years are the same in Figures 2 and 3. Yet the relationship with the other set can still be useful to complete the interpretation of axes: the reading of the chart must be controlled by full statistic information provided along with the graphical result.

[3] The dimensionality of the space is given by the formula: *Min (Row Number, Column Number) -1*. Here: *Min (200, 29) -1 = 28*.

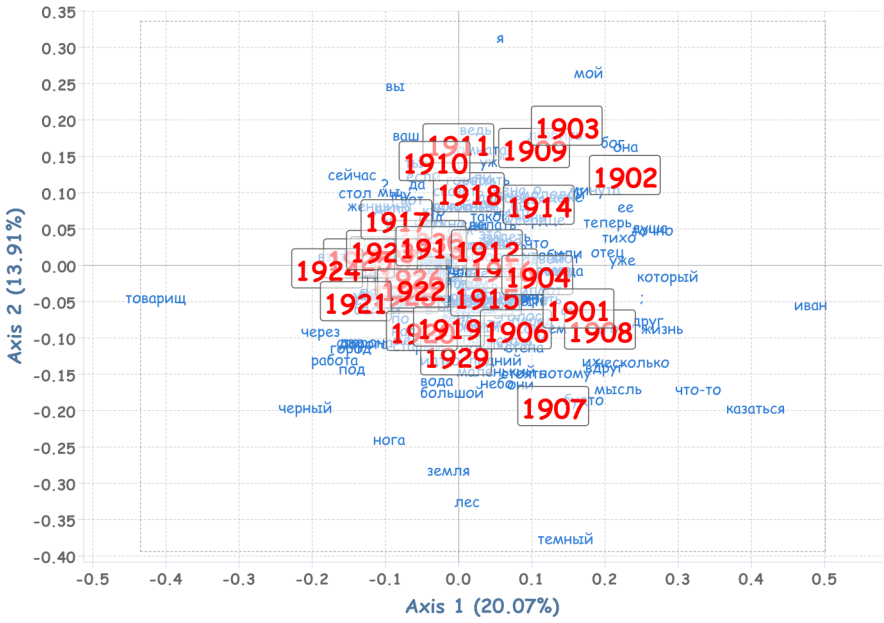

Fig. 2. CA factorial map of years crossed with the 200 most frequent lemmas

Fig. 3. CA factorial map of years crossed with the 200 most frequent lemmas (only years plotted)

Focusing our analysis on years, a chronological distribution appears very clearly on the 1st axis, although it is not linear. Most of years before 1917 are located on the right-hand side of the chart and most of years after 1918 are located on the left-hand side. Only one post-revolutionary year (1918) is located on the right-hand side (close to the center), and two pre-revolutionary years (1910 and 1913) are found in the left-hand zone. The year 1917 has a marked position in the "post-revolutionary" camp. These exceptions should be studied in more detail but they do not compromise the general trend, that is the deep impact of the 1917 revolution on the core vocabulary of the Russian literature.

Our second hypothesis, that is the influence of WWI and the CW on the core vocabulary of the corpus as opposed to peacetime, does not seem to be confirmed by this analysis. In order to test more specifically if significant differences exist between war and peaceful time periods, we built a four part partition: pre-revolutionary peace (1901-1913), WWI (1914-1916), OR and CW (1917-1922), and post-revolutionary peace time (1923-1930).

The CA of the 200 most frequent lemmas (Fig. 4) shows a very strong opposition of pre- and post-revolutionary periods on the 1st axis, which is also characterized by a high inertia (73%).

Fig. 4. CA factorial map of the table crossing 4 periods x 200 most frequent lemmas (punctuation and proper nouns included)

The position of these periods on the 1st axis is strictly chronological: the "war" periods are situated between the "peaceful" ones. With the 2nd axis (14.07% of inertia), instead of a Guttman horseshoe pattern, we observe that the strongest polarization is observed between the two "war" periods. Then the opposition between war and peace appears on the 3rd axis of the CA (9.9% of inertia) but this is nothing but an artifact created by corpus organization (4 parts of comparable size) and by the orthogonality of axes imposed by the calculation. So the only conclusion we can draw is that the frequencies of words and punctuation marks follow an overall chronological trend and that there is a major gap between the subcorpora of 1914-1916 and 1917-1922.

We can now have a deeper look into the data thanks to the table of indicators that TXM provides along with the CA charts. We can see there which rows (lemmas in our case) contribute the most to the organization of the factorial map on the 1st axis. The contribution of an item to an axis is measured by the portion that the item brings to the inertia of the axis.

The strongest contributions are that of punctuation marks (full stop, dash and semicolon), and the noun *tovarisch* ‘comrade’ associated with the post-revolutionary period, the 3rd person pronouns *on* ‘he’ and *ona* ‘she’, the conjunction *i* ‘and’ and the proper name *Ivan* associated with the pre-revolutionary period. The exact figures are given in Table 1, where the 1st column is the lemma, the second is its mass in the corpus, the 3rd is the contribution to the 1st axis (sorted descending) and the 4th is the coordinate on the axis (negative for the post-revolutionary period and positive for pre-revolutionary).

Table 1

Lemmas and the punctuation marks with the strongest contribution to the CA by periods 1st axis

Lemma	Translit.	Gloss	Mass[4]	Cont1[5]	c1[6]
.	.	.	9,86	19,75	-0,12
она	ona	she	1,12	8,16	+0,23
и	i	and	5,74	5,59	+0,09
-	-	-	5,37	4,2	-0,08
:	:	:	1,13	3,82	-0,16
он	on	he	2,21	3,66	+0,11
иван	Ivan	Ivan	0,08	3,64	+0,58
который	kotoryj	which	0,31	2,56	+0,25
товарищ	tovarisch	comrade	0,09	2,44	-0,46
на	na	on	2,19	2,4	-0,09

One might argue that punctuation marks may be altered by editing practices and that proper nouns depend on individual texts, so that they should not be taken into account to study general trends in the evolution of literature. TXM allows us to dynamically edit the lexical table on which CA is based. However, if we exclude punctuation and proper nouns, the configuration of the factorial map does not change: chronological distribution remains dominant on the 1st axis (Fig. 5). So, this test provides an additional confirmation to our basic hypothesis.

[4] The *mass* of an item is the portion of its occurrences. For instance, the first line of this table indicates that 9.86 % of occurrences of the 200 most frequent lemmas are dots.

[5] *Cont1* is contribution to the first axis, that is the portion of inertia that the item brings to the total inertia of the axis. For instance, dots represent nearly one-fifth of first axis inertia.

[6] *C1* stands for the coordinate on first axis. Opposite signs (“+” or “-”) indicate antinomical roles along the dimension.

Fig. 5. CA factorial map of the table crossing 4 periods x 200 most frequent lemmas (punctuation and proper nouns excluded)

A sharp distribution of subcorpora based on chronological criterion as a result of CA is perfectly visible in Fig. 4 and 5. This distribution indicates the possibility of chronological text classification. Such a corpus may, for instance, be used as a training dataset for identification of literary texts in Russian language. An unknown Russian literary text might then be identified as belonging to a particular time span by using various classifiers with the help of third-party computing routines.

The selection of differentiating features requires further research using CA as well other methods. The results presented in Fig. 4 and 5 show that 200 most frequent lemmas (with or without punctuation and proper nouns) may be used for such a selection. A more detailed CA including morphological types (building the datasets for analysis for different parts of speech) will allow us in the future to determine the precise differentiating features responsible for the distribution of subcorpora observed in Fig. 2 to 5.

For literary texts, individual parts of speech, as well as different sets, such as nouns and adjectives, nouns, adjectives and verbs, noun phrases, verb phrases, may be used as differentiating features in addition to the general set of lemmas. Selecting different morphological features may have a considerable impact on the thematic classification. The research results provided above show that precise recommendations for discovering the characteristic features of individual subcorpora may be formulated in the future work.

Cluster analysis

Another way to investigate and precise the organization among years is to run a cluster analysis over the 29 years. Instead of considering the full data table, we benefit from CA and extract the 5

first dimensions of our first CA (Fig. 2 and 3): thus we get a more complete representation of the data than in the factorial map, with 56% of the total variance (instead of 34 %), and still gain an efficient noise reduction. TXM implements an agglomerative hierarchical clustering with Ward criterion (minimum variance clustering). Fig. 6 displays the full hierarchical dendrogram on the 29 years. The two main branches out of three still clearly split years before 1916 and years from 1917 onwards, with a few singularities (1906, 1910, 1913 standing closer to late years, and 1918 associated with early years). As the CA was based on frequent lemmas, this global structure confirms that a major overall lexical turn happens around 1917.

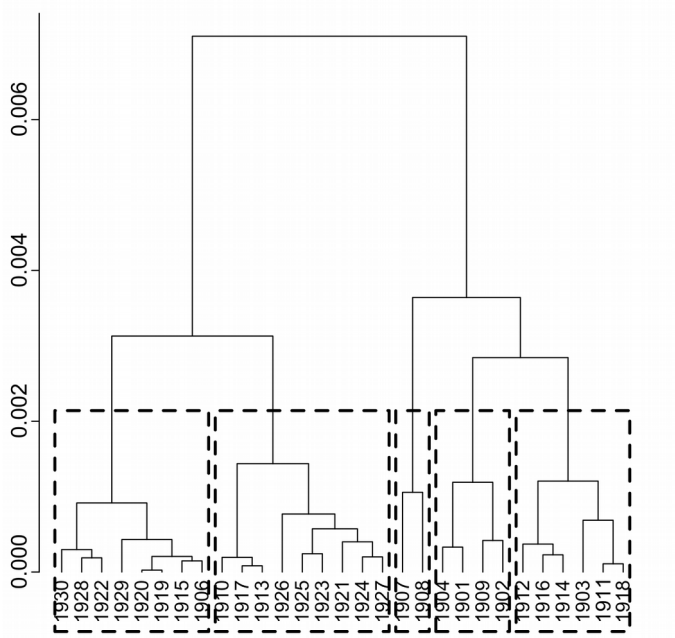

Fig. 6. Ward clustering of the 29 years represented by their 5 first coordinates in the CA space built from the 200 most frequent lemmas

Specificity

Specificity test may be used as a complement to CA in order to determine what elements (for instance, lemmas) are the most characteristic of each subcorpus and explain the variation observed in CA. As a first step, we built a general specificity table based on the 200 most frequent lemmas (including punctuation marks) in the partition by period (the same dataset as for the CA in the previous section). For the record, most of these high frequency lemmas are function words, very common content words and punctuation marks. This appears to be relevant in characterizing a deep language evolution, as with such high frequencies, we minimize the effect of proper nouns or more topical words that could be associated with a particular text or an author.

As a second step, we analyzed 10 most specific lemmas for each period obtained from the statistical test, and built a bar chart displaying the score of these lemmas in all the parts.

The specificity score indicates the probability that the number of occurrences in a given part of a corpus would be so high or so low in case of random distribution of words in the corpus. It corresponds to the absolute value of the base 10 logarithm of the probability. By convention, a positive score indicates over-representation and a negative score indicates under-representation of the element in the subcorpus. In other words, the score of 25.7 for the verb *kazat'sia* 'to seem' in the period 1901-1913 indicates that there is one chance in $10^{25.7}$ that it would occur so many times in this part of the corpus randomly. The scores between -2 and +2 (≥1% probability) are considered as "banal" (non significant). In a few cases the probability may be so low that it exceeds the possibilities of computing, so a conventional limit has to be set to the maximum score. By default the maximum score is set to 1,000 in TXM, but for the sake of readability of bar charts (with lower but significant scores), a lower limit may be set. In this research the specificity scores were capped at 30.

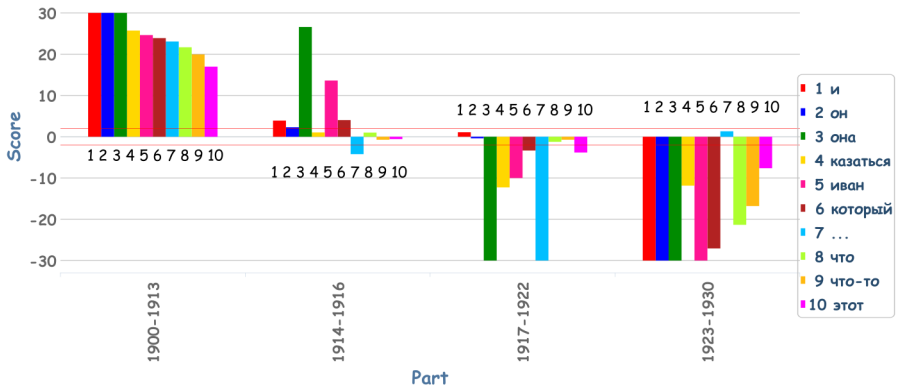

Fig. 7. Ten most specific lemmas of period 1 (1901-1913)

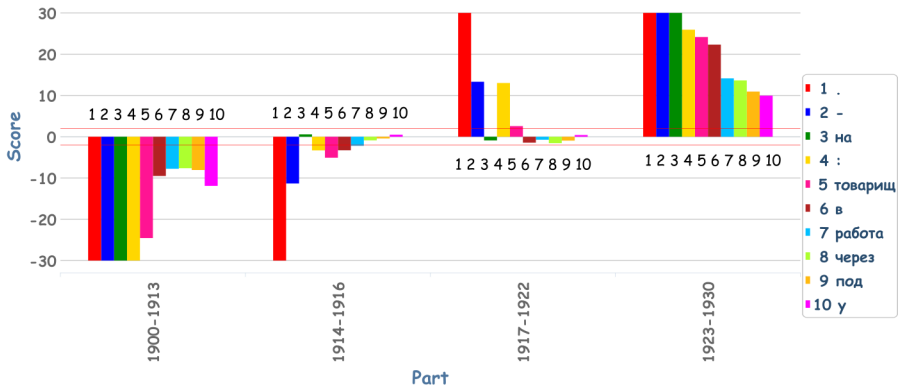

Fig. 8. Ten most specific lemmas of period 4 (1923-1930)

We first compare the most specific words in the opposite periods: 1901-1913 (Fig. 7) and 1923-1930 (Fig. 8). It appears quite vividly that the most specific words of one period are generally the most unspecific of the other, and vice versa. For instance, the score of the relative pronoun *kotoryi* is +23.9 in the first period and -27.1 in the last one.

The only exception is the ellipsis mark (…) which is highly specific in 1901-1913 and banal in 1923-1930. It is highly unspecific in 1917-1922. The intermediate periods generally display the same specificity polarity as the corresponding "extremes" with generally a lower score. This is another evidence of the chronological distribution of periods observed on the 1st axis of the CA.

The same tendency is confirmed when we look at the most specific words of the intermediate periods (Fig. 9 and 10), that is that the main distinction lies at the 1916/1917 border and that the specificity is greater in the extreme periods. This can be illustrated by two examples: the name *Ivan*, (the 3rd most specific lemma in the period 1914-1916) and the semi-colon (the 4th most specific lemma in the 1917-1922 period). Their scores are presented in Table 2.

Table 2

Specificity scores for Ivan and semi-colon

Lemma	1901-1913	1914-1916	1917-1922	1923-1930
Иван	+24.6	+13.6	–10.0	< –30.0
:	< –30.0	–3.3	+13.0	+25.9

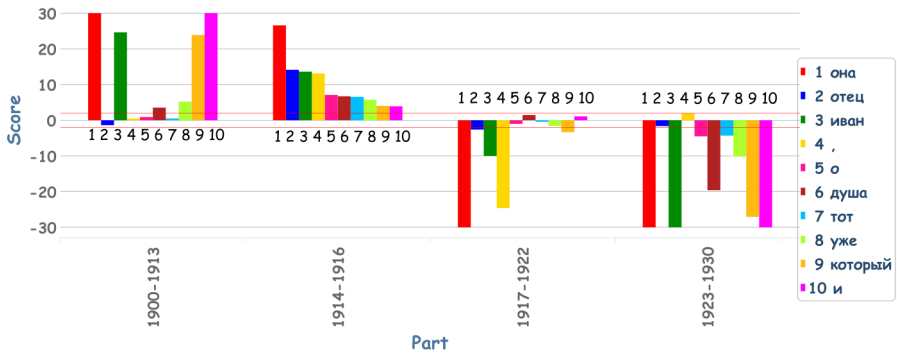

Fig. 9. Ten most specific lemmas of period 2 (1914-1916)

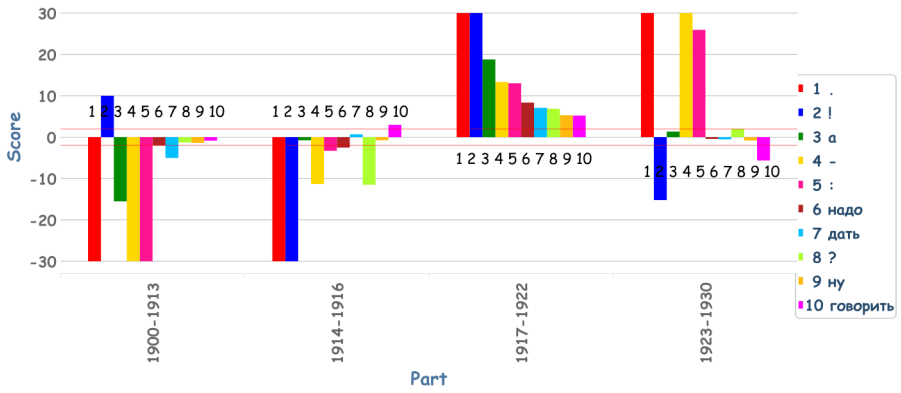

Fig. 10. Ten most specific lemmas of period 3 (1917-1922)

However, some lemmas stand out in one of the intermediate periods. For instance *otec* 'father' has the score +14.1 in 1914-1916, its score being negative or banal in all the other periods. The same kind of phenomenon occurs with the predicative adverb *nado* 'must' in 1917-1922 (score +8.4). It is banal or unspecific in the other periods.

For such words that do not seem to follow the general chronological trend it is necessary to check if their specificity is not due to some individual texts rather than to a period as a whole. For that purpose we made a partition of chronologically ordered individual texts and examined a bar chart of every word. For *otec* (Fig. 11) we see that this word is highly specific to some individual texts (score > +10 in 6 texts out of 308), and four of these texts are located in the 1914-1916 period, while in the other texts of the same period the score lies within the banality zone. In the short story *Tayna* ('Mystery') by Alexey Dementiev written in 1915, the main characters are *otec diakon* ('father deacon') and *otec Mikhail* ('father Mikhail'), where the word *otec* is used as a religious title, which explains its extremely high specificity score (over +30). It cannot therefore be considered to be characteristic of the whole period.

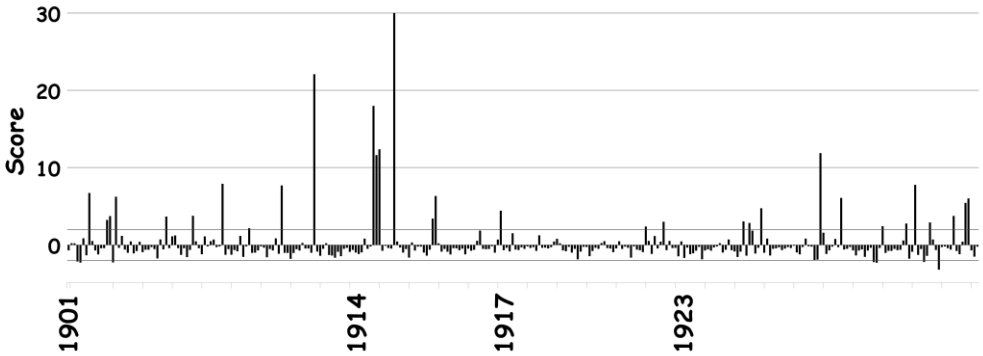

Fig. 11. Specificity bar chart for *otec* in individual texts (chronologically ordered partition)[7]

[7] The score axis notes the specificity value (a score of 10 in a text means that, if words were randomly spread, there would be less than 1 chance in 10,000,000,000 (10^{10}) that this word would

The situation is quite different for the word *nado* (Fig. 12). The score is superior to +5 in only one text (*Mladency gor* 'Babies of the mountains' by Gleb Pushkarev, 1922) but there clearly is a slight over-representation in the 1917-1922 period: 8 out of 12 texts with the score >2.1 and 6 out of 7 texts with the score >3 are located in this period.

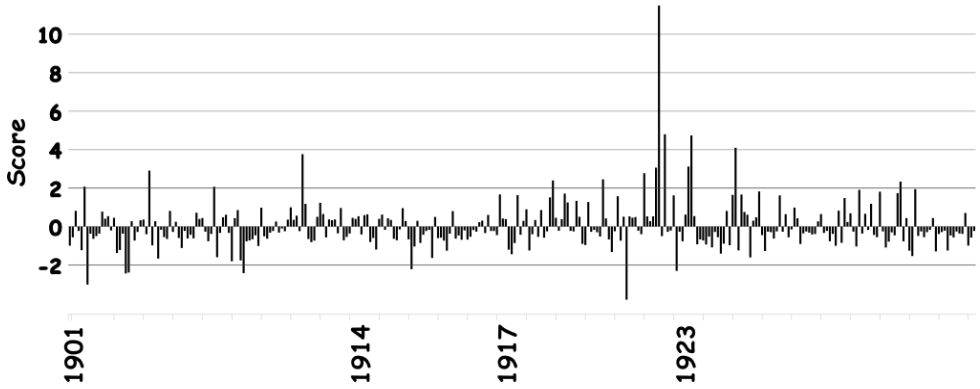

Fig 12. Specificity bar chart for *nado* in individual texts (chronologically ordered partition)

Progression

In order to locate more precisely the time point after which the frequency of a certain word increases or decreases considerably, we can use the Progression tool provided by TXM. It builds a chart where a "cumulative" curve is drawn for one or more search patterns expressed by a CQL query. The curve goes rightward with every word of the corpus and moves one step up at each occurrence of the pattern. If the number of occurrences is low, the "steps" of the curve are visible. If the pattern occurs regularly the slope of the curb indicates the *relative* frequency. Thus, no normalization transformation is needed, as the *slope* directly measures and visually shows the ratio of occurrence count to textual length, which is a normalized value. Word frequencies may differ, however word slopes can be straight compared throughout texts with various lengths.

We built a progression chart for two lemmas, *dusha* 'soul' and *tovarisch* 'comrade' that appeared among the specific words of the pre- and post-revolutionary periods respectively (Fig. 13). Vertical lines show the limits of years in the corpus. It should be noted that within a year the texts are arranged in the alphabetical order of authors, so the progression is not strictly chronological.

Both words are used all over the corpus but for each of them we observe a significant change in frequency at a certain time point. To make this change more obvious we drew straight dotted lines before and after the turning point for each word. For *tovarisch* the frequency increases in 1917 and

---

have such a relatively high frequency). So high bars flag texts in which *otec* occurs with a frequency much higher than expected. Tall negative bars (directed downwards) flag statistically significant underuse of this word.

remains stable afterwards, while *dusha* is used steadily up to 1922, and its frequency decreases afterwards.

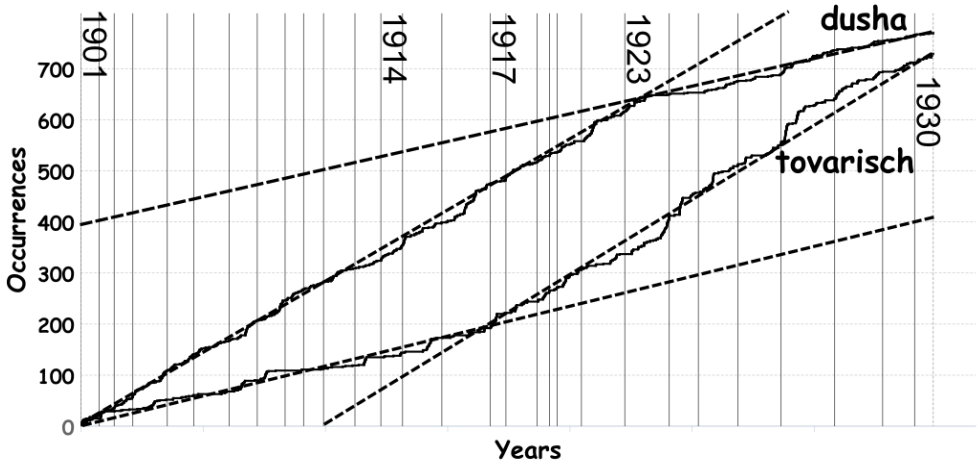

Fig. 13. Cumulative progression chart for the lemmas *dusha* 'soul' and *tovarisch* 'comrade'. The occurrences of the search terms are represented by curves. Dotted lines are drawn to visualize the average slope before and after the change in frequency.

This observation is confirmed by a simple measure of relative frequency presented in Table 3. In mathematical sense this corresponds exactly to the slope of the straight lines in Fig. 13. From a statistical point of view, we can compute the corresponding specificity scores, to assess significance: for *dusha*, the over-representation in the first period amounts to 19 (that is, less than one chance in 10^{19}); for *tovarisch*, the specificity score reaches a value of 35.

Table 3

Relative frequency (per 10,000) of the words *dusha* 'soul' and *tovarisch* 'comrade'

Lemma	**dusha**		**tovarisch**	
Period	***1901-1922***	***1923-1930***	***1901-1916***	***1917-1930***
Frequency	6.9	3.1	2.8	7.4
Specificity score	+19	-19	-35	+35

Conclusion

The results of this pilot study show the efficacy of TXM application for research on language dynamics and confirm our basic hypothesis that there is a chronological trend in the evolution of the core vocabulary of Russian short stories. The 1917 revolution represents a major event that divides the corpus into two distinct periods. The influence of WWI and the CW as opposed to peacetime is less evident, or at least the tests we used did not allow us to find evidence of lexical opposition of war and peace periods.

An interesting result is that punctuation plays an important role in the organization of the research data. The full stop is highly specific for the post-revolutionary period. This may indicate that sentences became shorter since 1917 but this observation needs to be verified by a more precise syntactic analysis. However, the chronological trend remains dominant even if the punctuation marks and proper nouns are excluded of the counts.

As we look at the specificity and relative frequency of individual words, we can observe that some of them are highly specific to a particular period (e.g. *nado* 'must' in 1917-1923) or to a small group of individual texts (e.g. *otec* 'father' for 6 texts out of 308), others demonstrate a dramatic change in relative frequency at a certain time point.

In our opinion, these preliminary observations show that textometric tools provided by TXM may be powerful and insightful in corpus analysis. Further research may be conducted in several directions. Firstly, we can apply additional NLP tools, such as stemming, detection of noun and verb phrases, and verify whether the trends observed on lemmas appear more sharply when the statistics is calculated on their output. Analysis of morphological tags may reveal completely new trends, such as the more or less "orality" features of the texts and/or the expression of nominal *vs.* verbal dominance, which appears to be a general deep trend in text statistics [40; 41 P. 147–148]. A more precise external description of the texts, such as sociolinguistic data on the authors, may also prove to be an insightful direction in the exploration of literary texts. For example, this may include the membership of the author in a particular literary group such as "futurists" or "oberiuts", certain sociological characteristics of the writers (sex, age, social background, education, profession, etc.) and their biography features [42], story topics [43] and type of the narrative [44], individual stylistic features of some authors [45], etc. Finally, other genres and text types can be added to the corpus in order to widen the part of the language it represents.

The results of our research show that the tools provided by TXM platform are efficient for analysis of Russian text corpus in the framework of computational linguistics. The proposed methodology may be used for various applied task when it is necessary to reveal language changes under the influence of some external factors or events.

References

1. Martynenko, G.Y. (1988) *Osnovy stilemetrii* [The Foundation of Stylometics]. St. Petersburg: St. Petersburg State University.

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

3. Martynenko, G.Y., Sherstinova, T.Y. (2020) 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. In: *R. Piotrowski's Readings in Language Engineering and Applied Linguistics, Proceedings of the III International Conference on Language Engineering and Applied Linguistics* (*PRLEAL-2019*), Saint Petersburg, Russia, November 27, 2019, CEUR Workshop Proceedings. Vol. 2552. pp. 105–120. URL: http://ceur-ws.org/Vol-2552/Paper10.pdf.

4. Tukey, J. W. (1980) We Need Both Exploratory and Confirmatory, *The American Statistician*, 34 (1). pp. 23-25.

5. Lüdeling, A., Kytö, M. (ed.) (2009) *Corpus linguistics: an international handbook*, Vol. 2. Berlin, New York: W. de Gruyter.

6. Osipov, G.S. (2011) *Metodi iskusstvennogo intellekta* [Methods of Artificial Intelligence]. Moscow: Fizmatlit.

7. Chepovskiy, A. M. (2015). *Informatsionnyye modeli v zadachah obrabotki tekstov na yestestvennyh yazykah* [Information Models for the Problems of Natural Text Processing]. 2nd. ed. Moscow: Natsional'nyy otkrytyy universitet "INTUIT".

8. Lavrentiev, A.M., Solovyev, F.N., Suvorova, 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 ekstremistskih 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*, vol. 16, no. 3. pp. 19-31. DOI: 10.25205/1818-7935-2018-16-3-19-31.

9. Lavrentiev, A.M., Smirnov, I.V., Solovyev, F.N., Suvorova, M.I., Fokina, A.I., Chepovskiy, A.M. (2018) Sozdaniye spetsial'nyh korpusov tekstov na osnove rasshirennoy platformy TXM [Creating text corpora for special purposes on the basis of extended TXM platform]. *Sistemy vysokoy dostupnosti*, vol. 14, no. 3. pp. 76-81. DOI: DOI: 10.18127/j20729472-201803-13.

10. Polyakov, I.V., Sokolova, T.V., Chepovskiy, A.A., Chepovskiy, A.M. (2015) Problema klassifikatsii tekstov i differentsiruyushchiye priznaki [The problem of text classification and differentiating features]. *Vestnik NSU. Series: Information Technologies*, vol. 13, no. 2. pp. 55-63.

11. Martynenko, G.Y., Sherstinova, T.Y. (2020) *Corpus of Russian Short Stories of the First Third of the 20th Century: Theoretical Issues and Linguistic Parameters*, Structural and Applied Linguistics. Vol. 14. St. Petersburg, 2020 (in print)

12. Martynenko, G., Sherstinova, T., Melnik, A., Popova, T., Zamirailova, E. (2018) On the principles of creation of the Russian short stories corpus of the first third of the 20th century. In

Proceedings of the XV International Conference on Computer and Cognitive Linguistics 'TEL 2018'. Kazan. pp. 180–197.

13. Sherstinova, T., Grebennikov, A., Skrebtsova, T., Guseva, A., Gukasian, M., Egoshina, I., Turygina, M. (2020) Frequency Word Lists and Their Variability (the Case of Russian Fiction in 1900-1930), In *Proceedings of the 27th Conference of Open Innovations Association FRUCT*. Trento: University of Trento, Italy (in print).

14. Martynenko, G., Sherstinova, T. (2019) Symmetrics of syntactic figures in fiction: the case of Russian short stories of the 20th century. *Computer Linguistics and Computing Ontologies*. 2019, 3. pp. 116-123. DOI: 10.17586/2541-9781-2019-3-116-123.

15. Kazartsev, E., Davydova, A., Sherstinova, T. (2020) Rhythmic Structures of Russian Prose and Occasional Iambs (a Diachronic Case Study). In *Proceedings of the 22nd International Conference on Speech and Computer - SpeCom 2020, St. Petersbug, LNCS (LNAI)*. Springer International Publishing (in print).

16. Sherstinova, T., Skrebtsova, T. (2020) Russian Literature around the October Revolution: A Quantitative Exploratory Study of Literary Themes and Narrative Structure in Russian Short Stories of 1900–1930. In: *Proceedings of the International Workshop "Computational Linguistics" – CompLing-2020*. St. Petersburg (in print).

17. Martynenko, G., Sherstinova, T. (2019) Analytical Distribution Model for Syntactic Variables Average Values in Russian literary Texts. In: *Proceedings of the 4th International Conference Digital Transformation and Global Society DTGS-2019, St. Petersburg, Russia, June 19–21, 2019. Revised Selected Papers. Communications in Computer and Information Science*, 1038. Springer International Publishing. pp. 719–731.

18. Sherstinova, T., Ushakova, E., Melnik, A. (2020) Measures of Syntactic Complexity and their Change over Time (the Case of Russian). In: *Proceedings of the 27th Conference of Open Innovations Association FRUCT*. Trento: University of Trento, Italy (in print).

19. Sherstinova, T., Kirina, M. (2020) Data Normalization in the Corpus of Russian Short Stories: Spelling, Literary Themes and Biographical Description of Writers (under review).

20. Savchuk, S. O. (2009) Korpus tekstov pervoy poloviny XX veka: tekuschee sostojanie i perspektivy [Text Corpus of the First Half of the 20th Century: Current State and Prospects]. In *Nacional'nyj korpus russkogo jazyka: 2006-2008. Novye rezul'taty i perspektivy* [Russian National Corpus: New Results and Prospects]. Saint-Petersburg: Nestor-Istoria. pp. 27-45.

21. Heiden, S. The TXM Platform: Building Open-Source Textual Analysis Software Compatible with the TEI Encoding Scheme. In: *Proceedings of the 24th Pacific Asia Conference on Language,*

Information and Computation. Sendai, Japan, 2010. pp. 389-398. URL: https://halshs.archives-ouvertes.fr/halshs-00549764.

22. TXM public website. URL: http://textometrie.org.

23. GWT Project website. URL: http://www.gwtproject.org.

24. Schmid, H. (1994) Probabilistic Part-of-Speech Tagging Using Decision Trees. In: *Proceedings International Conference on New Methods in Language Processing*, Manchester, UK, Sept. 1994. pp. 44-49.

25. IMS Open Corpus Workbench (CWB). URL: http://cwb.sourceforge.net.

26. The R Project for Statistical Computing. URL: https://www.r-project.org.

27. Benzécri, J.-P. *et al.* (1973) *L'analyse des données*, t. 2, *L'analyse des Correspondances*, Paris: Dunod.

28. Husson, F., Lê, S., Pagès, J. (2017) *Exploratory Multivariate Analysis by Example Using R*, 2nd ed. Boca Raton: Chapman and Hall/CRC.

29. Léon, J., Loiseau, S. (eds) (2016) *History of Quantitative Linguistics in France*. Lüdenscheid: RAM-Verlag.

30. Née, É. *et al.* (2017) *Méthodes et Outils Informatiques pour l'Analyse des Discours*. Rennes: Presses Universitaires.

31. Lexicometrica website, URL: http://jadt.org.

32. Lebart, Salem, A., Berry, L. (1998) *Exploring Textual Data*, Dordrecht: Kluwer Academic.

33. Lebart, L., Pincemin, B., Poudat, C. (2019) *Analyse des Données Textuelles*. Québec: Presses de l'université du Québec.

34. Guttman, L. (1941) The quantification of a class of attributes: A theory and method of a scale construction. In: *The predicition of personal adjustment*, New York: SSCR. pp. 251-264.

35. Salem, A. (1991) Les séries textuelles chronologiques, *Histoire et Mesure*, 6 (1). pp. 149-175.

36. Lafon, P. (1980) Sur la variabilité de la fréquence des formes dans un corpus. *Mots. Les langages du politique*. 1. pp. 127-165.

37. Fisher, R.A. (1935) *The Design of Experiments*. Edinburg: Oliver and Boyd.

38. McEnery, T., Hardie, A. (2012) *Corpus Linguistics: Method, Theory and Practice*. Cambridge University Press.

39. Sharoff, S. *Russian statistical taggers and parsers*, URL: http://corpus.leeds.ac.uk/mocky.

40. Biber, D. (2014) The ubiquitous oral versus literate dimension: A survey of multidimensional studies. In: J. Connor-Linton & L.W. Amoroso (Eds), *Measured language: Quantitative studies of acquisition, assessment, and variation*, Washington DC: Georgetown University Press. pp. 1-20.

41. Brunet, É. (2016). *Tous comptes faits, Écrits choisis*, t. 3, *Questions linguistiques*. Paris: Champion.

42. Sherstinova, T.Y. (2019) Biographical database of Russian writers (on the creation of a corpus of Russian narrative of the 20th century), In *Proceedings of the International Conference "Corpus linguistics-2019"*, St. Petersburg: Publishing House of St. Petersburg University. pp. 439–447.

43. Sherstinova, T.Y., Skrebtsova, T.G. (2020) Russian Literature Around the October Revolution: A Quantitative Exploratory Study of Literary Themes and Narrative Structure in Russian Short Stories of 1900–1930. In: *Proceedings of the International Conference DTGS-2020. Digital Transformation and Global Society*. 5th International Conference, DTGS 2020, St. Petersburg, Russia, 2020, Revised Selected Papers. Communications in Computer and Information Science (in print).

44. 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.

45. Martynenko, G.Y. (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]. In: *Proceedings of the International Conference "Corpus Linguistics – 2019"*. St. Petersburg State University. pp. 395–404.