The Dynamics of Extensive Text Variables in Russian Short Stories

Valeria Zarembo[1][0000-0002-4315-9041] and Tatiana Sherstinova[1,2][0000-0002-9085-3378]

[1]National Research University Higher School of Economics, St. Petersburg, Russia
[2] St. Petersburg State University, St. Petersburg, Russia
[1] vszarembo@edu.hse.ru; [2] tsherstinova@hse.ru

Abstract. The research presented in this paper is aimed at the analysis of dynamic organization of a literary text. Using the statistical time series method, the dynamics of the main extensive text variables — the mean paragraph length and the mean sentence length — is considered. The material for this study was the annotated subcorpus from the Corpus of the Russian Short Stories of 1900-1930, which consists of 310 stories written by 300 Russian writers. It was narrative fragments of texts (the narrator's speech) that were subjected to analysis, dialogical fragments were not taken into consideration. As a result, the most frequent dynamic profiles of paragraph length and sentence length were obtained, which reflect the most typical structures of the dynamic organization of short literary texts.

Keywords: dynamic text structure, quantitative literary studies, paragraph length, sentence length, time series, stylometrics, corpus linguistics, text composition.

1 Introduction

The research described in this paper is aimed at studying dynamic organization of literary texts expressed in the categories of paragraph length and sentence length. These quantitative measures are traditionally the focus of style and language studies [1, 2, 5, 6, 8, 14, 18, 28]. Scientific interest in investigating these extensive text variables has been intensified in recent years and may be explained by their importance for solving many urgent applied tasks related with texts classification and attribution [4, 7; 21, 25, 26]. As a rule, these variables are calculated on average over the text, their measures of variation being not taken into account. However, the question – How stable these variables are on the time scale of text composition? – remains actual [11, p. 219].

On the other hand, no less urgent is the task of comparing the dynamics of these variables with the study of text composition or its plot structure. In recent years, the interest in computer analysis of fiction texts has sharpened significantly, that may be explained by the modern technological development of society, technologies of computational linguistics and digital humanities, as well as the needs for the development of artificial intelligence systems [9, 11–13, 16]. A vivid example of a dynamic ap-

proach to the analysis of fictional texts from the point of view of the development of text tonality (emotional trajectories) is given in [20].

Developing the methodology for this study, the authors relied on the method proposed by Gregory Martynenko, who is the founder of the St. Petersburg stylometric school [10].

2 Data and Method

2.1 Material

The material for this study is the annotated subcorpus of the Corpus of Russian Short Stories of the First Third of the 20th Century [19]. This corpus is designed as a literary resource which should become the research site for various linguistic and stylistic studies, which implies the necessity to include literary texts of the maximum number of writers who wrote in a given historical period [19]. Apart from texts written by well-known and outstanding writers, the corpus contains short stories of the large number of 'second-rate' authors, which are also involved into consideration. Thereby better literary representation of different aspects of social and cultural life, as well as of language and stylistics diversity is achieved [15].

From the corpus, a subset of 300 randomly chosen short stories was created, with 100 stories for each subperiod (1900–1913, 1914–1922, 1923–1930), one per the author. Finally, 10 texts of the authors that wrote through all three subperiods were added [15].

2.2 Method

The texts were cleaned from dialogues and checked for the presence of at least 10 paragraphs or sentences. Thus, 305 and 308 texts were included in the final subset for paragraph analysis and for sentence analysis respectively.

In order to create dynamics contours, the general approach proposed in [10] was used. It was implemented in the following way:

1. It was hypothesized that there is a correlation between two extensive variables – mean sentence length and mean paragraph length – and a plot of a text. Namely, action text fragments were expected to have shorter paragraph and sentence length than descriptive or reflective episodes.
2. Each text was tokenized with R software [22]. Thereafter, two variables were measured: paragraph length (in sentences) and sentence length (in words).
3. Texts were tested for whether the changes in paragraph and sentence length are significant.
4. The sequence of paragraphs was divided into 10 groups, regardless of their size. Then, for each group mean length was measured. The same was done for sentences.
5. The results were presented as time series, with the group as the independent variable and mean paragraph or sentence length – as an independent one [3].

6. Time series were smoothed with the moving average method.
7. The final result was visualized as a line graph.

For example, the application of this method to the short story "Resort husband" ("Kurortny muzh") written by Alexander Amfiteatrov in 1911 leads to the following results (see Table 1).

Table 1. The mean absolute values for the story "Resort husband" by Alexander Amfiteatrov

Interval number	**Mean paragraph length (in sentences)**	**Mean sentence length (in words)**
1	1.73	11.2
2	1.27	9.38
3	1.55	10.9
4	3.5	11.3
5	2.45	9.61
6	3.64	9.66
7	4.1	6.88
8	4.45	12.4
9	3.91	9.59
10	3.64	12.1

Then, these numbers are smoothed with the moving average method [27], where the absolute numbers are replaced with the mean for a certain interval (three groups in this experiment). In this way, the means for each group are replaced with their smoothed values – for example, second and third ones are calculated with the following formulas:

$$y_2 = \frac{y_1 + y_2 + y_3}{3}, \quad (1.1)$$

$$y_3 = \frac{y_2 + y_3 + y_4}{3}. \quad (1.2)$$

Means for the first and the final groups are smoothed by separate formulas:

$$y_1 = \frac{2y_1 + y_2 - y_4}{2}, \quad (2.1)$$

$$y_{10} = \frac{2y_{10} + y_9 - y_7}{3}. \quad (2.2)$$

The result for the story "Resort husband" is presented in Table 2 and Figure 1.

Table 2. The mean values for the story “Resort husband” by Alexander Amfiteatrov

Interval number	Mean paragraph length, absolute value	Mean paragraph length, smoothed value	Mean sentence length, absolute value	Mean sentence length, smoothed value
1	1.73	0.614	11.2	10.2
2	1.27	1.52	9.38	10.5
3	1.55	2.11	10.9	10.5
4	3.5	2.5	11.3	10.6
5	2.45	3.2	9.61	10.2
6	3.64	3.4	9.66	8.71
7	4.1	4.06	6.88	9.65
8	4.45	4.15	12.4	9.63
9	3.91	4	9.59	10.9
10	3.64	3.54	12.1	12.1

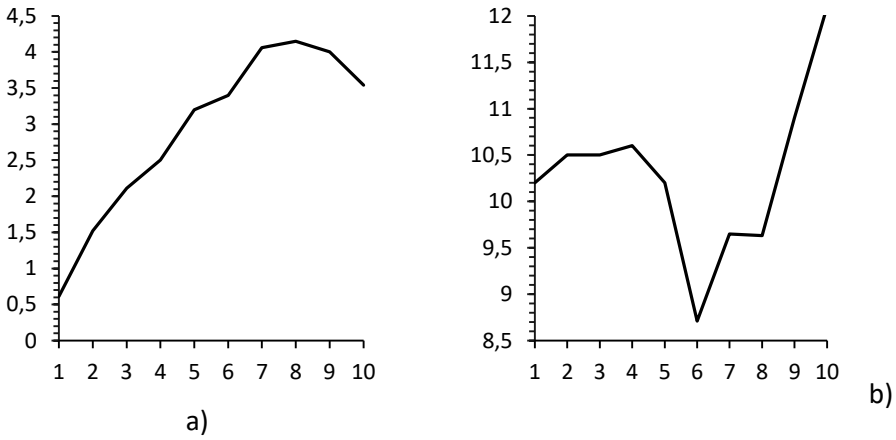


Fig. 1. The dynamic contours for the story “Resort husband” by Alexander Amfiteatrov: a) the dynamics of mean paragraph length, b) the dynamics of mean sentence length

3 The Results

3.1 Mean paragraph length

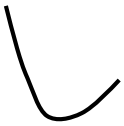
Table 3 contains some of the most frequent figures for paragraph length dynamic, which cover about 50% of the text sample.

As this table shows, the most frequent patterns are those, in which a text begins with “heavy” paragraphs that gradually shorten towards the rising action. The following changes depend on the type of a figure. For instance, types 1 and 2 have a similar dynamic – the rise of paragraph length and its subsequent fall; the key difference here lies in the nature of the rise. In type 1, a growth of paragraphs for the most part hap-

pens at the climax or near the falling action – these elements tend to be the heaviest ones. In type 2, on the other hand, the rise leans closer to the rising action and climax. Moreover, type 2 is distinguished by the lower extremity of the initial fall – mean paragraph length in the rising action rarely becomes less than the mean value across all intervals.

Another variation of the change is represented by groups 3 and 4, in which the paragraph size in the later text parts remains relatively small. Texts of both types still retain the increase of mean paragraph length, but in the first case it is closer to the falling action and is usually insignificant, while in the second case the rise is followed by the fall typical for types 1 and 2.

Table 3. The most frequent figures of paragraph length distribution

Rank	Figure	Frequency	Texts
1		29	"An autumn beam" ("Osenny luch", 1910) by Vladimir Gordin, "The man without a square" ("Chelovek bez ploshchadi", 1927) by Sergey Zayaitsky, "A wolf's dream" ("Volchya mechta", 1922) by Nina Smirnova
2		23	"Oranges" ("Apelsiny", 1907) by Alexander Grin, "The mystery" ("Tayna", 1915) by Alexey Dementyev, "Strained relations" ("Obostrenie otnosheniya", 1903) by Euhene Chirikov
3		22	"A year of their life ("God ikh zhizni", 1916) by Boris Pilnyak, "The first star" ("Pervaya zvezdochka, 1927) by Alexander Arosev, "Nothing happened" ("Nichego ne sluchilos", 1917) by Iosif Orsher
4		16	"Fire" ("Pozhar", 1926) by Sergey Belyaev, 'A pond' ("Prud", 1912) by Nikolay Oliger, "How Ivan spent his time" ("Kak Ivan provel vremya", 1912) by Semyon Podyachev
5		16	"My enemy" ("Moy vrag", 1918) by Leo Urvantsov, "In a quiet spot" ("V tikhom uglu", 1927) by Euhene Fyodorov, "Snegurochka" (1904) by Sergey Elpatyevsky
6		16	"Three portraits" ("Tri portreta", 1925) by Alexander Belyaev, "The nature of things" ("Priroda veshchey", 1916) by Ivan Danilin, "A fly" ("Mukha", 1912) by Boris Sadovskoy
7		16	"A confession" ("Ispoved", 1916) by Nikolay Kolokolov, "Propaganda train" ("Agitvagon", 1923) by Marietta Shaginyan, "Saltychikha's cove" ("Saltychikhin grot", 1926) by Olga Forsh

8		15	"A complicated case" ("Zaputanny sluchay", 1927) by Arkady Bukhov, "An accident" ("Proisshestviye", 1924) by Boris Lavrenyov, "Six days" ("Shest dney", 1925) by Nikolay Nikitin

To conclude, in the standard pattern the exposition is always composed of large paragraphs – it is probably related to the fact that this part of the text gives the reader a background of the main conflict and requires a detailed explanation. The rising action, on the other hand, should be more dynamic – short, abrupt paragraphs are more fitting. The subsequent climax has different patterns – this can happen due to two possible factors: the difference in climax nature (more "thoughtful" episodes describing the characters' feeling at the higher point of conflict (types 1 and 2) vs more dynamic, emotional ones (types 3 and 4)) or the presence of dialogues (for the cases where narration followed characters' speech and served as short remarks for it; the deletion of dialogue could lead to the drop in the mean paragraph length in such episodes). Finally, the falling action and resolution require neither a big amount of information nor a detailed description – here, smaller paragraphs can be rather used.

There are, however, several figures that differ from the "standard" pattern. For example, types 5 and 7 follow the initial "large exposition – small rising action – large climax" pattern but have a rise in paragraph length in the falling action and resolution. Texts of these groups probably required more explanations in the end, having either characters' reflection on the event of the ending or a more detailed description of the events following the climax.

Another non-standard variation is the type where only the "long climax – short resolution" pattern is retained, while the exposition is composed of shorter paragraphs. The stories of this type usually have a more dynamic beginning – perhaps, descriptions in them are replaced with the action or dialogues.

Finally, among frequent figures, there is a type opposite to types 1 and 2: a dynamic exposition, a "heavy" rising action, a subsequent dynamic climax, and a large resolution. The explanatory parts of these texts are evidently moved towards the rising action and the resolution, while the rising action and the climax are richer in action or dialogue.

This pattern is, however, the only one completely different from the standard: all the other types keep the features of the common type in one or another way. Based on this, it can be concluded that literary texts mostly lean towards the "heavy" exposition and climax and the more dynamic rising and falling action.

3.2 Mean sentence length

Table 4 contains some of the most frequent figures of the dynamic of mean sentence lengths, which cover about 44% of the sample.

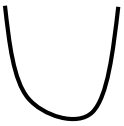
Sentences generally tend to behave similarly to paragraphs – in most cases their length diminishes along the text. The possible fluctuations are also close to the ones found in paragraph lengths, with the pattern of the rise of volume in the climax being

the most frequent. The intensity of these fluctuations also has some differences: for instance, group 1 is characterized by the milder change in mean sentence length and its earlier rise, while group 4 has a more extreme initial drop and a rise closer to the resolution. Moreover, there remains a type with a sharp drop in the rising action and only a slight rise in the climax.

At the same time, the type with longer rising action and resolution becomes more frequent. Coupled with the opposite tendency in paragraph length, it can be explained as the preference for big paragraphs filled with short sentences in the exposition and climax – and for paragraphs with few long sentences in the ending.

There is also an increase in frequency for some figures that are rare for paragraphs, such as type 6, a more radical case of "long beginning – short ending" pattern, and type 5, an U-shaped figure that might emerge either due to the need for more dynamic action episodes and more explanations in the ending or due to the presence of large dialogues, similarly to the case of paragraphs.

Table 4. The most frequent figures of sentence length distribution

Rank	Figure	Frequency	Texts
1		30	"A shipboy" ("Yunga", 1919) by Vladimir Bill-Belotserkovsky, Belyaev "Fire" ("Pozhar", 1926) by Sergey Belyaev, "The mystery" ("Tayna", 1915) by Alexey Dementyev
2		27	"A prize" ("Nagrada", 1927) by Nikolay Anov, "Tales of our days" ("Skazki nashikh dney", 1916) by Yuri Volin, "Dusk" ("Sumerki", 1909) by Sergey Semyonov
3		21	"Authority" ("Vlast", 1906) by Lydia Avilova, Z. Gippius "A madwoman" ("Sumasshedshaya", 1903) by Zinaida Guppius, "Roses" ("Rozy", 1913) by Ieronim Yasinsky
4		20	"The last will" ("Zaveshchaniye", 1901) by Boris Bentovin, "An autumn beam" ("Osenny luch", 1910) by Vladimir Gordin, "The man that visited Mars" ("Chelovek, pobyvavshij na Marse", 1927) by Graal-Arelsky
5		20	"Roma" (1919) by Alexey Kopeykin, "In a quiet spot" ("V tikhom uglu", 1927) by Euhene Fyodorov, "Strained relations" ("Obostrenie otnosheniya", 1903) by Euhene Chirikov
6		16	"Petushkov Rocket" ("Raketa Petushkova", 1924) by Gleb Alekseyev, "A glass of champagne" ("Bokal shampanskogo", 1911) by Kazimir Barancevich, "It blowed gently" ("Potyanulo", 1910) by Vasily Bashkin

To conclude, sentences by most part have little difference from the paragraphs. However, there is an increase in the frequency of the type opposite to the "standard" one.

4 Comparing the dynamics of plot and extensive variables

The study of plot dynamics on the material of the corpus of the Russian story was previously described in [24]. In this study, we will check how promising is the comparison of compositional elements and the dynamics of extensive variables.

Let us consider in what way the discovered dynamic is tied to the plot of short stories, on the example of two texts: "The mistery" ("Tayna") by Aleksey P. Dementyev written in 1915 and "Strained relationships" ("Obostrenie otnosheniya") by Euhene V. Chirikov written in 1903.

Table 5 and Figure 2 show the means and dynamic contours for "The mystery" by Aleksey P. Dementyev.

Table 5. The mean values for the story "The mystery" by Aleksey P. Dementyev

Interval number	Mean paragraph length, absolute value	Mean paragraph length, smoothed value	Mean sentence length, absolute value	Mean sentence length, smoothed value
1	4.25	4.08	25.09	25.54
2	3	3.31	13.2	16.07
3	2.67	3	9.91	13.14
4	3.33	4.44	16.3	16.8
5	7.33	4.33	24.09	18.8
6	2.33	4	15.9	17.2
7	2.33	3.22	11.6	12.44
8	5	3.22	9.82	10.04
9	2.33	2.78	8.7	9.99
10	1	1	11.45	10

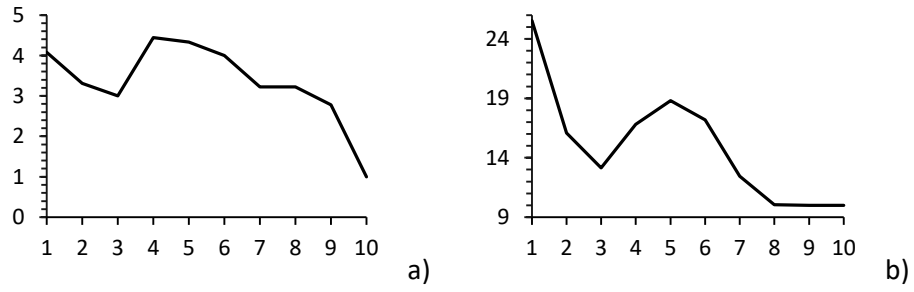


Table 6. Fig. **2.** The dynamic contours for the story "The mystery" by Aleksey P. Dementyev: a) the dynamics of mean paragraph length, b) the dynamics of mean sentence length

This short story can be considered a “standard” one: both paragraph and sentence length change according to the most frequent patterns.

The text opens with long exposition and rising action (groups 1–3) that are rich with descriptions of nature and the deacon’s thoughts. Closer to the episode of the deacon coming to pope Mikhail, paragraph and sentence lengths decrease – this part of the text is comprised mostly of dialogues and does not require long explanations.

As the plot progresses and the deacon’s “mystery” is revealed in the climax (groups 4–6), both lengths grow again. This happens, on one hand, due to the presence of long reflexive paragraphs that build the anticipation of the revelation of the “mys- tery” and, on the other hand, due to the detailed description of the deacon’s hunt and his feelings related to this experience. Here, the pacing of the narrative slows down: more importance gains not the action itself, but the sense of excitement that the main character feels from the hunt. Because of that, the climax of the story is somewhat similar to the exposition, both in its content (mostly descriptive and reflective para- graphs) and size.

In the falling action and resolution (groups 7–10), a drop of both quantitative variables takes place – these parts of the text contain a relatively short dialogue between the deacon, his wife, and pope Mikhail. The ending is given mostly through a sequence of paragraphs with lots of short sentences that describe how other characters and the deacon himself view his hunts.

A different structure can be seen in the story “Strained relationships” by Euhene V. Chirikov. Table 6 and Figure 3 show the means and dynamic contours for this text.

Like Dementyev’s story, the story by Chirikov begins with relatively long exposition and rising action (groups 1–4). In the exposition the reader is given the reason for Misha’s quarrel with parents and his refusal to eat with them – in other words, the basis of the main conflict. The rising action is built in the same way: it contains mostly reflective paragraphs flowing into the action-oriented ones.

However, in the rising action and the climax (groups 5-8), the dynamic begins to deviate from the “standard” variation: the author still uses large paragraphs but fills them with short sentences. This difference reflects the distinction in the mood of both stories – while “The mystery” had a more melancholic, slow-paced feel to it, “Strained relationships” is more energetic. The plot in it is developed first through Misha’s chaotic planning depicted by short, abrupt sentences and then through the market episode comprised mostly of action and dialogues.

The opposite occurrence can be seen in the falling action and the resolution (groups 9-10) where paragraphs become shorter and sentences – longer. This, again, is related to the fact that the consequences of Misha’s lie can be described shortly: the family got concerned about his health, and the “strained relationships” thus were resolved. Moreover, the length of the sentences is accomplished through the listing of multiple actions – this conveys the fuss and concern that overcame the mother and sister of the main character. The ending, on its end, concentrates on Misha’s feelings after the resolution of the conflict and describes them in a few large sentences.

Table 7. The mean values for the story "Strained relationships" by Euhene V. Chirikov

Interval number	Mean paragraph length, absolute value	Mean paragraph length, smoothed value	Mean sentence length, absolute value	Mean sentence length, smoothed value
1	3	3	12.43	14.43
2	2.5	3.25	11.46	10.5
3	4.25	3.08	7.62	8.85
4	2.5	3.25	7.46	8.05
5	3	4	9.07	8.13
6	6.5	4	7.85	8.05
7	2.5	4	7.23	8.31
8	3	2.83	9.85	9.05
9	3	2.6	10.08	10.5
10	1.8	2.05	11.57	12.99

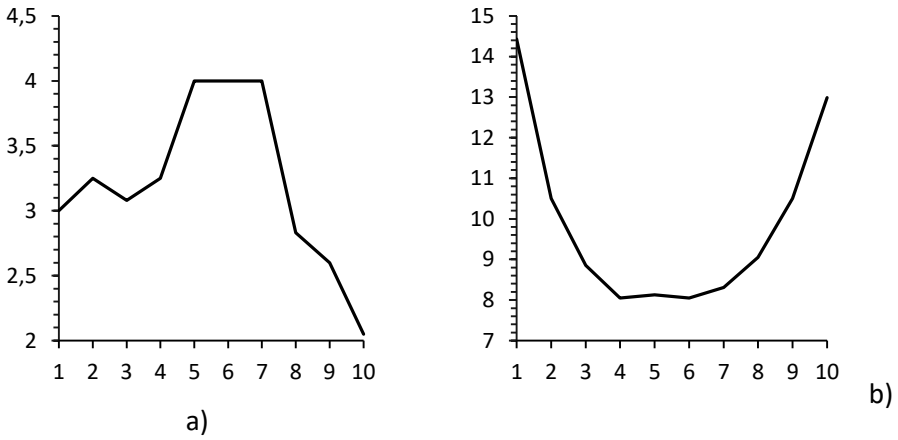


Fig. 2. The dynamic contours for the story "Strained relationships" by Euhene Chirikov: a) the dynamics of mean paragraph length, b) the dynamics of mean sentence length

5 Discussion

As it was mentioned before, the dynamic of extensive variables in question has several similarities: both paragraph and sentence length lean towards the downward trend with the large exposition and falling action; in addition, they tend to increase in volume in the climax and resolution. There are, however, other variations of dynamic contours that appear due to the nature of the plot and the structure of a specific text.

Resulting figures coincide with those obtained in [10–11] – most notably, among the most frequent were the ones similar to types 1 and 6 for paragraphs and type 3 for sentences. Moreover, the results of the experiment prove that the pattern of large sentences in the beginning, their subsequent drop in the rising action, and the growth in

climax with the final drop in the ending can be considered standard for the description of the narrative.

At the same time, in the research by Gregory Martynenko, some figures were found to be infrequent by the results of the current experiment. This difference can be explained by the characteristics of Martynenko's subset: it included only 20 texts from 4 writers. The expansion of the subset and the increase in its variety thus might specify the exact frequencies of the plot development figures.

6 Conclusion

The presented study confirmed the results obtained earlier in the analysis of small Russian prose that: 1) extensive text variables, such as the average sentence length and the average paragraph length are not constant values, but change with the development of plot narration, and 2) the change of the extensive text characteristics is not a chaotic process; general dynamic patterns (or trends) of their changes in time can be traced. Moreover, it can be argued that there are some invariant typical structures that occurs more often than others in literary texts of short prose.

It seems appropriate to continue the research by involving data on another important extensive text characteristic – the total size of literary text measured in words, as well as additional information about text internal structure (its division into chapters, sections, etc.), the topic(s) of the text [23] and other content-based characteristics related to literary annotation of data. Thus, more accurate information about the frequency and implementation features of certain typical dynamic profiles will be obtained.

7 Acknowledgements

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

References

1. Admoni, V.G.: Razmer predlozheniya i slovosochetaniya kak yavlenie sintaksicheskogo stroya [The Length of Sentences and Phrases as a Phenomenon of Syntactic Structure]. Voprosy yazykoznaniya [Topics in the Study of Language], 1966(4), 111–118 (1966).
2. Akimova, G.N.: Razmer predlozheniya kak faktor stilistiki i grammatiki [Sentence Length as a Factor of Stylistics and Grammar]. Voprosy yazykoznaniya [Topics in the Study of Language], 1973(2), 67–79 (1973).
3. Coghlan, A.A.: A Little Book of R for Time Series, Release 0.2. Wellcome Trust Sanger Institute, Cambridge (2018).

4. Grieve, J.: Quantitative Authorship Attribution: An Evaluation of Techniques. Literary and Linguistic Computing, 22(3), 251–270 (2007).
5. Huxtable, R.: Sentence length. Science, 197(4300), 208 (1977).
6. Kelih E., Grzybek P., Antić G., Stadlober E.: Quantitative Text Typology: The Impact of Sentence Length. In: Spiliopoulou, M., Kruse, R., Borgelt, C., Nürnberger, A., Gaul, W. (eds.) From Data and Information Analysis to Knowledge Engineering, Proceedings of the 29th Annual Conference of the Gesellschaft für Klassifikation e.V., University of Magdeburg, March 9–11, 2005, 382–389. Springer, Berlin (2006).
7. Lagutina, K., Lagutina, N., Boychuk, E., Vorontsova, I., Shilakhtina, E., Belyaeva, O., Paramonov, I., Demidov, P.G.: A Survey on Stylometric Text Features. In: Balandin, S., Niemi, V., Tuytina, T. (eds.). Proceedings of the 25th Conference of Open Innovations Association FRUCT, Helsinki, Finland, 184–195. Institute of Electrical and Electronic Engineers, New York (2019).
8. Lesskis, G.A.: Nekotorye statisticheskie zakonomernosti kharakteristiki prostogo i slozhnogo predlozheniya v russkoj nauchnoj i khudozhestvennoj proze XVIII–XX vv. [Some Statistical Laws of the Characteristics of Simple and Compound Sentences in Russian Scientific and Fiction Texts of 18–20th centuries]. Russkij yazyk v nacionalnoj shkole [Russian language in the national school], 1968(2), 67–80 (1968).
9. Manovich, L.: Software Takes Command. Bloomsbury Academic, New York (2013).
10. Martynenko, G.Y.: Vvedenie v chislovuyu garmoniyu teksta [The Introduction to Numeral Harmony of the Text]. St. Petersburg State University, Saint-Petersburg (2009).
11. Martynenko, G.Y.: Metody matematicheskoy lingvistiki v stilisticheskikh issledovaniyakh [Computational Linguistics Methods in the Stylistics Research]. Nestor-Istoriya, Saint-Petersburg (2019).
12. Martynenko, G.Y., Sherstinova, T.Y.: Chislovoj profil syujeta [The Numeric Profile of the Plot]. In: Proceedings of IV Congress of Russian language researchers ‘Russkij yazyk: istoricheskie sudby i sovremennost’ [Russian Language: Historical Fates and Modern Age], 524–525. Moscow State university, Moscow (2010).
13. Martynenko, G., Sherstinova, T.: Emotional waves of a plot in literary texts: new approaches for investigation of the dynamics in digital culture. In: Alexandrov, D.A., Boukhanovsky, A. V., Chugunov, A.V., Kabanov, Y., Koltsova, O. (eds.) Digital Transformation and Global Society. DTGS 2018. Communications in Computer and Information Science, 859, 299–309. Springer, Cham (2018).
14. Martynenko, G., Sherstinova, T.: Analytical Distribution Model for Syntactic Variables Average Values in Russian literary Texts. In: Alexandrov, D.A., Boukhanovsky, A.V., Chugunov, A.V., Kabanov, Y., Koltsova, O., Musabirov, I. (eds.) Digital Transformation and Global Society. DTGS 2019. Communications in Computer and Information Science, 1038, 719–731. Springer, Cham (2019).
15. Martynenko G., Sherstinova T.: 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: Ronzhin, A., Noskova, T., Karpov, A. (eds.) R. Piotrowski's Readings in Language Engineering and Applied Linguistics, Proc. of the III International Conference on Language Engineering and Applied Linguistics (PRLEAL-2019), Saint Petersburg, Russia, November 27, 2019, CEUR Workshop Proceedings, 2552, 105–120. RWTH Aachen University, Aachen (2020).
16. Martynenko, G.Y., Sherstinova, T.Y., Popova, T.I., Melnik, A.G., Zamirajlova, Y.V.: 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 XX Century]. In:

Proceedings of the 15th TEL International Conference on Computational and Cognitive Linguistics (TEL-2018), 1, 180–197. Izdatelstvo AN RT, Kazan (2018).
17. Moretti, F.: Distant Reading. Verso, London (2013).
18. Olmsted, D.: On some axioms about sentence length. Language 43(1), 303–305 (1967).
19. Martynenko, G.Y., Sherstinova, T.Y., Melnik, A.G., Popova, T.I.: Metodologicheskie problemy sozdaniya Kompyuternoy antologii russkogo rasskaza kak yazykovogo resursa dlya issledovaniya yazyka i stilya russkoy hudozhestvenny prozy v epokhy revolutsionnykh peremen (pervoy treti XX veka) [Methodological problems of creating a Computer Anthology of the Russian story as a language resource for the study of the language and style of Russian artistic prose in the era revolutionary changes (first third of the 20th century)]. In: Kompjuternaya lingvistika i vychislitelnye ontologii [Computational Linguistics and Computational Ontologies]. Issue 2 (Proceedings of the XXI International Conference "Internet i sovremennoe obshchestvo" [Internet and Modern Society], IMS-2018, St. Petersburg, 30 May 2018–2 June 2018. Collection of scientific articles), 99–104. ITMO University, St. Petersburg (2018).
20. Reagan, A.J., Mitchell, L., Kiley, D., Danforth, C.M., Dodds, P.S.: The emotional arcs of stories are dominated by six basic shapes. EPJ Data Science 5, 31 (2016).
21. Rudnicka, K.: Variation of sentence length across time and genre. Influence on syntactic usage in English. In: Whitt, R.J. (ed.) Diachronic Corpora, Genre, and Language Change (Studies in Corpus Linguistics, 85), 219–240. John Benjamins, Amsterdam (2018).
22. Silge J., Robinson, D.: Text Mining with R. O'Reilly Media, Sebastopol (2020).
23. Sherstinova, T., Mitrofanova, O., Skrebtsova, T., Zamiraylova, E., Kirina, M.: Topic Modelling with NMF vs. Expert Topic Annotation: the Case Study of Russian Fiction. In: Martínez-Villaseñor, L., Herrera-Alcántara, O., Ponce H., Castro-Espinoza, F.A. (eds) MICAI 2020, LNCS, 12469. Springer, Cham (2020).
24. Sherstinova T., Skrebtsova T.: Russian Literature Around the October Revolution: A Quantitative Exploratory Study of Literary Themes and Narrative Structure in Russian Short Stories of 1900–1930. In: Proc. of the International Workshop "Computational Linguistics" (CompLing-2020)(in print).
25. Sherstinova, T., Ushakova, E., Melnik, A.: Measures of Syntactic Complexity and their Change over Time (the Case of Russian). In: Balandin, S., Turchet, L., Tuytina, T. (eds.) Proceedings of the 27th Conference of Open Innovations Association FRUCT, Trento, Italy, 221–229. Institute of Electrical and Electronic Engineers, New York (2020).
26. Stamatatos, E.: A survey of modern authorship attribution methods. Journal of the American Society for Information Science and Technology, 60(3), 538–556 (2009).
27. Venecky, I.G., Veneckaya V.I.: Osnovniye matematiko-statisticheskie ponyatiya i formuly v ekonomicheskom analize [Basic Math and Statistics Concepts and Formulas in Economic Analysis]. Statistika, Moscow (1979).
28. Yule, G.: On sentence-length as a statistical characteristic of style in prose: with application to two cases of disputed authorship. Biometrika, 30(3/4), 363–390 (1939).