Statistical Description of Russian Texts: Parameters and Factors

Anastasia M. Amieva[1], Viktor V. Filimonov[1], Andrey A. Zhivodyorov[2,3], Anna A. Kramarenko[2]

[1]Ural Federal University, Department of printing arts and web-design, Mira, 32, Ekaterinburg 620002 Russia
[2]Ural Federal University, Department of Technical Physics, Mira, 21, Ekaterinburg 620002 Russia
[3]Central Scientific Library Ural Branch of the Russian Academy of Sciences, Sofia Kovalevskaya, 22 Ekaterinburg 620137 Russia

Abstract. This work describes the parameters for the attribution of Russian-language texts and the methods for obtaining them. The technique of machine attribution of Russian-language texts is presented. The methodology is based on applying factor analysis to study the relationships between parameters. The work was done at the department of printing arts and web-design IRIT-RtF UrFU.

Keywords: Modeling, Attribution of texts, Statistics χ^2, Law of large numbers, Diffusion coefficient, Factor analysis.

1 Introduction

This work represents a step towards the construction of a theoretical model of written text. We think that, first of all, it is necessary to understand the laws of text construction and on their basis to construct a theory for describing the text. The laws of text construction should correspond the following requirements:

- be expressed numerically;
- be summarized for different texts;
- be subjected to formal mathematical analysis and / or modeling.

With such a formulation of the problem, our research associate with the problem of interdisciplinary interactions in science (Petrov V.M.) [1]. Modern digital technologies and natural scientific concepts penetrate deeply into the humanities. Within the natural science process, areas associated with humanitarian objects are formed (cognitive studies, neuro-esthetics). In our case, it is a Russian-language texts.

Researches of a text are traditionally conducted in three directions: spatial, linguistic and statistical.

Spatial direction involves measuring the length of a line, leading, font size, etc. (Artyomov V.A. [2], Ushakova M.N. [3], Tarasov D.A. [4, 5, 6], Tyagunov A.G.

[4, 5], Sergeev A.P. [4, 5, 6], Filimonov V.V. [6], Weber A. [7], Cohn H. [8] and others).

The linguistic direction includes the study of meaning-bearing units (sentences, paragraphs) and structural features (Matezius V. [9], Skalicka V. [10], Halliday M.A.K. [11], PalmerF.R. [12], Martinet A. [13], Tenyer L. [14], Lotman Yu.M. [15, 16], Mintz Z.G. [17], and others).

Statistical studies are related to the study of quantitative characteristics of texts. Studies of quantitative parameters of texts conduct for a long time. The first theoretical result in the field of statistical studies of the text is the empirical «Zipf law». It can be formulated as follows: «The product of the frequency of occurrence of a word and its position in the frequency dictionary is approximately constant value»[1]. Quantitative parameter in the Zipf's law is the frequency of occurrence of words in the text.

In 1991, the American researcher Wentian Li proved that this law is fulfilled for any random sequence of symbols. Thus, he suggested that the law is a statistical phenomenon. And this phenomenon is not connected with the semantics of the text [18].

At this stage of the work, we develop a technique for machine attribution of texts. The technique can be used to determining authorship and evaluate the usability of the text.

We build our research on the following assumptions:

1. there are hidden structural elements in the text. They can be detected by methods of mathematical modeling and mathematical statistics;

2. The author and the reader are interpreting the text differ-ently because they have difference life experiences [19]. So we excluded «meaning» from consideration for the objectivity of the study.

Our works [19–23] was dedicated to studies of texts by the methods of frequency analysis and using the model of random walks. We considered the repeatability of individual letters and of their triples (the three vowel letters). By triples we mean three vowel letters that consistently appear in the text. These approaches were chosen by us, because they:

- are objective;
- allow to get rid of subjective and conventional effects;
- can be associated with digital data processing [19].

We base on the results of the conducted studies and we can say that the results of our work can be used to develop algorithms for machine attribution of texts without preliminary expert evaluation and without taking into account the meaning.

In the course of research, we have obtained and used a number of parameters. These parameters can be used as text attributes.

We use factor analysis to investigate the relationships between parameters in this work.

Studies were conducted on Russian-language texts (about 1,500 texts of various genres and directions) [21].

[1] https://shkolazhizni.ru/culture/articles/78456

2 Parameters of text

This section provides brief descriptions of parameters of the text. The methods for measuring the parameters are presented in the corresponding articles. References to the articles are indicated in the text.

Technique for studying texts with using the χ^2 statistics was presented in [20, 21]. We compared the calculated and observed number of triples of the vowels in the texts.

The calculated quantity was calculated using the formula:

$$n_i^{\text{theor}} = \omega_i^{\text{theor}} \cdot N \quad (1),$$

where ω_i^{theor} is the frequency of the appearance of the triples of vowel letters, *N* is the number of vowels in the text:

$$\omega_i^{\text{theor}} = \omega_{i_1} \cdot \omega_{i_2} \cdot \omega_{i_3} \quad (2),$$

where $\omega_{i_1}, \omega_{i_2}, \omega_{i_3}$ are the frequencies of occurrence of the first, second and third letters in the triples, which are calculated using the following formula:

$$\omega_{i_j} = \frac{n_j}{N} \quad (3),$$

where n_j is the number of emergence of the individual vowel in the text.

Obviously, that n_i^{theor} cannot be equal to zero.

The observed amount of triples (n_i^{emp}) was obtained by recalculating all triples in the text. Recalculation was carried out using the program «*Qlines*»[2]. In real texts, many versions of the triples are missing. Those n_i^{emp} can be equal to zero.

We used the statistics χ^2 to estimate the difference in the calculated distribution of the triples from their observed distribution:

$$\chi^2 = \sum_{i=1}^{k} \frac{\left(n_i^{\text{theor}} - n_i^{\text{emp}}\right)^2}{n_i^{\text{theor}}} \cdot \frac{50000}{N} \quad (4).$$

The texts are normalized to a length of 50 000 vowels. The normalization was applied that the values of χ^2 for different texts could be compared with each other.

The values of χ^2 were calculated for approximately 1,500 texts. The values obtained were distributed as follows:

1. from 0 to 2 000 there are not texts;
2. from 2 000 to 2 400 there are only poetic works;
3. the majority of fiction texts are located in the range from 2 000 to 6 000;
4. the majority of scientific texts are located in the range from 6 000 to 12 000;
5. the religious, socio-political and journalistic texts are located in the range from 4 000 to 12 000;
6. the administrative texts are located in the range from 10 000 to 30 000.

[2] Specially written for research by the staff of the Central Scientific Library Ural Branch of the Russian Academy of Sciences L.G. Gorbich

Thus, the texts were grouped in different groups. The groups basically coincided with the expert classification of texts. The machine did not focus on the meaning of the text and its name, but only analyzed the sequence of signs.

The next stage of the study [22] was the search for additional text attribution parameters. It was suggested that «the differences between the values of the χ^2 statistic are random and can be related to the finiteness of the length of the text» [22]. The longer the text length, the smaller the standard deviation (*SD*) of the values χ^2. In the case of «sufficiently large» texts, *SD* asymptotically tends to zero according to the law of large numbers:

$$SD = \frac{c}{\sqrt{N}} \quad (5),$$

where *c* is the coefficient of proportionality.

The coefficient *c* does not depend on the number of vowels *N* and can be related to the peculiarities of the text itself. Thus, the coefficient *c* can be an attribute of the text or be characteristic of the language as a whole.

As a result of the study it was found that:

1) $1{,}5 < c < 5$ for the main part of the texts considered;

2) religious ($0{,}3 < c < 3$) and administrative texts ($5 < c < 38$) are distinguished.

The second approach [23] is based on the use of the mathematical model of random walks.

In constructing the model, an analogy was made between the thermal motion of particles in Euclidean space and the displacement of a phase point in the state space.

Einstein's law states that the mean square of displacement of a Brownian particle in the absence of external forces is directly proportional to time. For the two-dimensional case, the law can be written as follows:

$$\bar{R}^2 = 4Dt \quad (6),$$

where *R* is the displacement, *D* is the coefficient similar to the diffusion coefficient for the physical system (hereinafter diffusion coefficient), *t* is the time (corresponds to the ordinal number of the letter from the beginning of the text).

The text is modeled as the displacement of the phase point according to certain transition rules. Each vowel is associated with a definite vector. Its length is determined by the inverse frequency of occurrence the letter. Directions of vectors corresponding to the various letters are not the same and are distributed uniformly over the circumference with intervals 40°.

The coefficients *D* were calculated for more than 100 texts. The values were distributed over two ranges: $D_1 < 124$, $D_2 > 124$. Poetic, fiction, scientific and journalistic texts are located in the first range. Administrative texts are located in the second range. Religious texts are located on the border between these ranges. Religious texts are located on the border between these ranges.

It turned out that the dependence of the mean square of displacement from time is ambiguously approximated by a straight line. Therefore, a relative correction to the Einstein law (*RC*) was analyzed. *RC* illustrates the difference between a real process and strictly random and has unique meaning for each text:

$$RC = \frac{a_2 \cdot t}{a_1}, \quad (7),$$

where a_2 the coefficient for senior term of the polynomial of the second degree, and a_1 is the coefficient for senior term of the straight line.

The values of *RC* were distributed over three ranges: $RC_1 < 12\%$, $12\% < RC_2 < 24\%$, $RC_3 > 24\%$. Fiction and journalistic texts are located in the first range, scientific and religious texts are located in the second range, and administrative texts are located in the third range. Poetry is located on the border between the first and second ranges.

3 Factor analysis

In the list of the attributes of text may be entered a large number of parameters except the parameters specified in clause 2. We used the following parameters in this work:

1. the number of vowels in the text *N*;
2. value of the statistics χ^2;
3. coefficient of proportionality of the law of large numbers *c*;
4. diffusion coefficient *D*;
5. relative correction to the law of Einstein *RC*;
6. the time when the text was created;
7. the language of the original text;
8. frequency of appearance of individual vowels in the text ω_i.

For the Russian-language text, the creation time is the year of its creation by the author, for the translated text — the year of its translation into Russian.

The language of the original text is assigned a numerical code: Russian — 1, English — 2, Japanese — 3. Texts, originals of which are written in other languages, were not considered in the present study.

The task was to determine the relationship between the listed parameters for more accurate classification of texts.

The following works were taken as a material for the study:

1. I.A. Bunin «Dark alleys»;
2. O. Wilde «Portrait of Dorian Gray»;
3. I.A. Akimov «The Legend of a Small Garrison»;
4. A.I. Kuprin «Duel»;
5. M. Twain «The Adventures of Huckleberry Finn»;
6. A. and B. Strugatsky «The Country of the Crimson Clouds»;
7. IS. Turgenev «Notes of a Hunter»;
8. A. and B. Strugatsky «The City of the Doomed»;
9. H. Murakami «Underground»;
10. S. King «Hearts in Atlantis».

We conducted an exploratory factor analysis for 16 variables (Table 1). The task of exploratory factor analysis is to ascertainment the factor structure of available data

and reduces the number of variables that are used to describe[3]. 16 parameters were grouped to five factors as a result of our study. At that the three parameters were insignificant (did not enter into in any factor), and the others were grouped into five factors.

Table 1 shows the meanings of factor loadings. The value of factor loads greater than 0,7 is standard when deciding whether to include a parameter in the factor. The values of factor loads for the parameters included in the factor are highlighted in bold type.

Table 1. The distribution of variables by factors

Variables	Factor 1	Factor 2	Factor 3	Factor 4	Factor 5
The number of vowels in the text *N*	0,364421	**0,729682**	-0,094348	0,053077	0,309197
Coefficient of proportionality of the law of large numbers *c*	0,040497	**0,816622**	-0,043738	0,184094	0,313755
Diffusion coefficient *D*	**-0,949569**	0,079862	0,070356	-0,029690	0,073218
Relative correction to the law of Einstein *RC*	-0,680325	0,005537	0,267187	-0,388376	0,388324
Value of the statistics χ^2	0,479242	-0,100362	-0,614457	0,254573	0,396370
The time when the text was created	**0,919218**	0,241454	0,026340	-0,071141	0,011079
The language of the original text	0,629812	-0,014981	-0,013261	0,068955	0,696027
Frequency of appearance the «а», ω_a	-0,443259	**0,755016**	0,255975	-0,255560	-0,099567
Frequency of appearance the «е», ω_e	0,017816	-0,305967	**0,900449**	0,022232	0,095588
Frequency of appearance the «и», ω_u	**0,746160**	-0,108156	-0,132397	0,451743	0,043413
Frequency of appearance the «о», ω_o	-0,009806	-0,299796	**-0,912499**	-0,182163	0,103690
Frequency of appearance the «у», ω_y	-0,527121	-0,030393	-0,100502	**-0,703169**	-0,303869
Frequency of appearance the «ы», $\omega_{ы}$	-0,181069	-0,294047	0,057419	0,224173	**-0,883945**
Frequency of appearance the «э», $\omega_э$	**0,880029**	0,010271	0,052785	-0,040246	0,328203
Frequency of appearance the «ю», $\omega_{ю}$	-0,085608	0,073588	0,085229	**0,891250**	-0,243617
Frequency of appearance the «я», $\omega_я$	-0,395044	0,364723	-0,090836	0,028160	**0,812817**

The diffusion coefficient *D*, the year the text was created, the frequencies of the letters «и» and «э» are included in the first factor.

The number of vowels, the coefficient of proportionality of the law of large numbers *c*, frequency of the letter «а» are included in the second factor.

The frequencies of the letters «е» and «о» are included in the third factor.

The frequencies of the letters «у» and «ю» are included in the fourth factor.

The frequencies of the letters «ы» and «я» are included in the fifth factor.

Thus, we got the opportunity to classify texts with a probability of 88%.

The distribution of parameters by factors is not interpreted by us at this stage of the study, since insufficient number of data was used.

[3]http://studme.org/79025/psihologiya/vidy_protsedury_faktornogo_analiza

4 Conclusion

The work describes methods for obtaining parameters for the attribution of Russian-language texts. Also, the use of factor analysis for these purposes is considered.

It was possible to reduce the number of variables from 16 to 5 with the help of factor analysis. The method allows you to classify texts with a probability of 88%.

Studying the interrelationships between the parameters will make it possible to obtain a more accurate classification of Russian-language texts according to their direction (poetry, fiction prose, scientific, journalistic, administrative and religious texts).

Classification can be used for machine processing of text, which will solve a number of problems:

1) construction of a mathematical model of the text;
2) research tasks (determining authorship, etc.);
3) assessment of readability (usability);
4) accounting for text specificities for editing and layout;
5) automated search for texts of a given direction in arrays of heterogeneous and poorly structured information.

References

1. Petrov, V.M. Art history and exact sciences? // Proceedings of an international scientific and practical conference dedicated to the memory of Herman Alekseevich Golitsyn (September 20–22, 2012) Quantitative methods in art history. Ekaterinburg: Artifact, 2013. — S. 6–7.
2. Artyomov, V.A. Technographic analysis of the total letters of the new alphabet [Electronic resource]. — Access mode: http://nowa.cc/printthread.php?t=282586 (circulation date is 18.05.2016).
3. Ushakova, M.N. New font for newspapers [Electronic resource]. — Access mode: http://www.liveinternet.ru/users/1013940/post267598201/ (circulation date is 22.05.2016).
4. Tarasov D.A., Akhmetova A.E., Sergeev A.P., Tyagunov A.G. Substantiation and derivation of the formula for the speed of reading based on the spatial characteristics of textual information // Proceedings of the International Scientific and Practical Conference (Ekaterinburg, 2015) Information Technologies, Telecommunications and Control Systems. Ekaterinburg: UrFU named after the first President of Russia BN. Yeltsin, 2015. — P. 140–146.
5. Tarasov, D.A. The account of spatial characteristics of a strip of a typing in the formula of reading / D.A. Tarasov, A.P. Sergeev, A.G. Tyagunov // News of Higher Educational Establishments. Problems of polygraphy and publishing. — 2014. — No. 6. — P. 3–10.
6. Tarasov, D.A. Legality of Textbooks: a literature review / D.A. Tarasov, A.P. Sergeev, V.V. Filimonov // Procedia — Social and Behavioral Sciences. — 2015. — P. 1300–1308.
7. Weber, A. Ueber die Augenuntersuchungen in den hoheren schulen zu Darmstadt [Electronic resource]. — Access mode: http://www.pearsonified.com/2011/12/golden-ratio-typography.php (circulation date is 30.05.2016).
8. Cohn, H. Die Hygiene des Auges in den Schulen. [Electronic resource]. — Access mode: http://www.pearsonified.com/2011/12/golden-ratio-typography.php (circulation date is 30.05.2016)
9. Matezius, V. On the potentiality of linguistic phenomena / V. Matezius // Selected works on linguistics. Translation from Czech and English — M., 2012. — P. 3–30.

10. Skalichka, V. Asymmetric dualism of linguistic units / V. Skalichka // History of linguistics of XIX-XX centuries in essays and extracts. — M.: Flint, 2012. — P. 54–63.
11. Halliday, MAK. Place of the functional perspective of the proposal (FPP) in the system of linguistic description [Electronic resource]. — Access mode: http://socialtranslation.ru/article.php?article_id=548 (circulation date is 2.04.2016).
12. Palmer, F.R., Mood and modality [Electronic resource]. — Access mode: www.academia.edu/3704213/Irreilis__ and_reality (circulation date is 7.04.2016)
13. Martinet, A. A functional view of language [Electronic resource]. — Access mode: https://www.questia.com/library/3057879/a-functional-view-of-language (circulation date is 19.03.2016).
14. Tenyer, L. Cours elementaire de syntaxe structurale [Electronic resource]. — Access mode: http://www.classes.ru/grammar/172.Tesniere/ source / worddocuments / _1.htm (circulation date is 19.03.2016).
15. Lotman, Yu.M. Inside the thinking worlds [Electronic resource]. — Access mode: http://lms.hse.ru/content/lessons/64192/%D1% 81% D0% B5% D0% BC% D0% B8% D0% BD% D0% B0% D1% 80% 207 -8 / lotman_yu _m_izbrannye_stati_v_treh_tomah_tom_1_stati_po_semi (circulation date is 3.05.2016).
16. Lotman, Yu.M. Articles on semiotics of culture and art [Electronic resource]. — Access mode: http://libatriam.net/read/899058/0/ (circulation date is 10.05.2016).
17. Mintz, Z.G. The structure of the sentence and the typology of artistic texts [Electronic resource]. — Access mode: http://www.studfiles.ru/view/ 3911949/ (circulation date is 16.05.2016).
18. Wentian, Li. Random Texts Exhibit Zipf's-Law-Like Word Frequency Distribution. — Santa Fe Institute, 1991. — P. 1–8.
19. Amieva, A.M. Machine attribution of Russian-language texts: a review of methods / A.M. Amieva, A.A. Kramarenko, V.V. Filimonov, A.A. Zhivodyorov // Materials of the X International Scientific and Practical Conference (Ekaterinburg, 27 February–3 March 2017) New information technologies in education and science. Ekaterinburg: RGPPU — 2017. — P. 371–375.
20. Filimonov V.V., Zhivodyorov A.A., Gorbich L.G. Expression and order in written speech // Izvestia UrFU. Series 1 The problems of education, science and culture. — 2012. — №3 (104). — P. 313–319.
21. Filimonov V.V., Amieva A.M., Sergeev A.P. Clustering of Russian-language texts using χ^2 statistics. // Proceedings of the International Scientific and Practical Conference (Ekaterinburg, January 12–13, 2016). Information: transmission, processing, perception. Ekaterinburg: UrFU named after the first President of Russia B.N. Yeltsin — 2016. — P. 164–174.
22. Filimonov V.V., Amieva A.M., Zhivodyorov A.A., Kramarenko A.A. Attribution of Russian-language texts using the law of large numbers. Proceedings of the International Scientific and Practical Conference (Ekaterinburg, January 12–13, 2017). Information: transmission, processing, perception. Ekaterinburg: UrFU named after the first President of Russia B.N. Yeltsin — 2017. — P. 10–18.
23. Kramarenko A.A., Filimonov V.V., Zhivodyorov A.A., Amieva A.M. Application of the random walk model for describing Russian-language texts. Proceedings of the International Scientific and Practical Conference (Ekaterinburg, January 12–13, 2017). Information: transmission, processing, perception. Ekaterinburg: UrFU named after the first President of Russia B.N. Yeltsin — 2017. — P. 138–164.