# LINKING WORDS IN RUSSIAN SOCIAL STUDIES COURSE BOOKS: A STUDY ON TEXT COMPLEXITY

**Marina Solnyshkina[1], Maria Kazachkova[2], Elzara Gafiyatova[3]*, Elena Varlamova[4]**

[1]Doctor in Philology, Professor, Kazan Federal University, Russia, mesoln@yandex.ru,
[2]Ph. D. in Philology, Moscow State Institute of International Relations, Russia, mbkazachkova@yandex.ru,
[3]*Ph. D. in Philology, Kazan Federal University, Russia, rg-777@yandex.ru,
[4]Ph. D. in Philology, Kazan Federal University, Russia, el-var@mail.ru

## Abstract

In this study we examine the context of transitional connectives 'firstly' (Rus. 'vo-pervyh'), secondly (Rus.'vo-vtoryh'), thirdly (Rus. 'v-tretyih'), in the fourth place (Rus. 'v-chetvertyh'), in the fifth place (Rus. 'v-paytyh') in the corpus of Russian school textbooks on Social Studies (grades 5th – 11th). The Corpus was compiled by a group of Kazan Federal University researchers and marked as RRC (Russian Readability Corpus). To ensure reproducibility of the research results and Corpus online availability, RRC developers uploaded the corpus with the shuffled order of sentences on the website (Authors' Database, 2017). The corpus is presented by texts of two sets of school textbooks written by L. N. Bogolubov (Bogolubov, 2017) and A.F. Nikitin (Nikitin, 2017), both recommended by the Ministry of Education of the Russian Federation. We view specialized corpora as preferable for discourse patterns studies and share Biber's point of view that linguistic tendencies are quite stable with ten (and to some extent even five) text samples per genre or register (Biber 2006). Based on the abovementioned assumptions, the authors view RRC with its total size of 525,748 tokens as a representative corpus of Russian academic discourse. The present study investigates two research questions: RQ1: Are the variety and frequency of numerical connectives in linear regression to text complexity? RQ2: What syntactic patterns are used after numerical connectives 'firstly' (Rus. 'vo-pervyh'), secondly (Rus.'vo-vtoryh'), thirdly (Rus. 'v-tretyih'), in the fourth place (Rus. 'v-chetvertyh'), in the fifth place (Rus. 'v-paytyh') in Russian high school academic discourse? With the help of AntConc, a concordance program which shows search results in a 'KWIC' (Key Word In Context) format, we determined the absolute frequency of transitional connectives 'firstly' (Rus. 'vo-pervyh'), secondly (Rus.'vo-vtoryh'), thirdly (Rus. 'v tretyih'), in the fourth place (Rus. 'v-chetvertyh'), in the fifth place (Rus. 'v-paytyh'). The research indicated a positive linear regression in the number of the numerical transitional connectives from 0 (in the textbooks of the 5th grade to 10 (firstly, Rus. 'vo-pervyh') in the textbooks of the 10th and 11th grades. The syntactic analysis of the context of transitional connectives indicate that the most widely used syntactic construction after all numerical connectives in the Russian academic discourse of Social science are (in descending order) the following: 1) SVO, where the subject (S) is manifested with a noun phrase (about 34%), infinitive constructions (12%), modal constructions (9%); 2) inverted word order OSV (42%) 3) VSO (3%). The results of this research provide us with insights into the general patterns of the academic discourse and cohesion.

**Keywords**: cohesion, text complexity, transitional connectives, corpus, academic discourse

# 1 INTRODUCTION

An educational text as a source of information for students and a guide for teachers to a greater degree determines success of any learning process. In recent years, the problem of qualitative school textbooks has become the subject of the World Congress of Historians, field visits of representatives and experts of the Council of Europe, international seminars, including Russia (Klokova, 2017). Modern Russian textbooks are very often described as “overloaded with conflicting and redundant information” (Shankin, 2015). Russian educators also argue that “the language of textbooks is terrible, it is abstruse and difficult to understand” (Goryainov, 2017). The poured sea of hard-to-read text in educational books hampers the educational process and creates insurmountable barriers in education (Paronjanov, 2017). Andrey Goryainov in his article “Reflections on Education and Textbooks” provides a convincing example of the above statement: “During the first week of the first school year pupils aged 6 are offered the following definition of zero: “zero is absence of elements in a set". The latter is unlikely to be simple for comprehension even for secondary school children not only for primary school. Andrey Goryainov also suggests ways of solving the problem, among which he emphasizes the need to provide different educational institutions such as lyceums, gymnasiums, special schools for children with disabilities and comprehensive schools with different textbooks. Each and every topic should be written for several age categories: in a plain language (Goryainov, 2017). The latter implies that the language used in education should guarantee that type of "communication that the public can understand and use"(Plain Language, 2017).

Unfortunately the problem of ‘plain language’ began being addressed in Russian science as late as 1970-s only. By that time American researchers had developed methods of text parameters computing and compiled representative and balanced corpora to conduct a reliable research. In late 1970-s based on the results of longitudinal studies Mikk Ya. A. concludes that reducing text complexity increases effectiveness of teaching without any additional work of teachers and schoolchildren. Mikk Ya. A. also formulated the criteria for the so-called “optimal” text. These criteria are made up of the optimal text complexity and the optimal level of comprehension. The author offers ‘optimal’ metrics of Estonian textbooks complexity in the rage from 15 to 24. According to Mikk Ya. A. optimal is the workload, in which all students can master the educational material. Based on the experimental data, the researcher proposed to reduce textbooks complexity as it “contributes to the development of cognitive activities of students and increases the amount of gained knowledge (Rakitin, 2018).

The modern paradigm of text and reading theories argues that text comprehension depends on the following three main aspects: 1) the graphic form of the text: fonts, formats, style), 2) students’ motivation, their linguistic and cognitive abilities, 3) the text itself, its vocabulary, syntax, narrativity, abstractness/ concreteness of words, referential cohesion, deep cohesion (Solnyshkina, Vishnyakova et al., 2017). The third set of factors is the one that can be automated based on the dependencies and correlations between text parameters (or metrics) and different categories of readers.

In 2017 Kazan Federal University team launched a project aimed at defining those Russian Academic texts metrics which can rank texts depending on readers’ reading profiles (Solnyshkina, Zamaletdinov et al., 2017). In the previous articles on the project we have already addressed the problem of text readability metrics and text parameters effect on its complexity (Solovyev, et al, 2018). This article is aimed at presenting results of the authors’ original research on texts parameters/metrics correlating with text complexity and thus able to profile reading texts with cognitive and linguistic abilities of readers.

# 2 RELATED WORK

Numerical connectives or transitions as linking words are defined in linguistics as phrases or words used to connect one idea to help readers progress from one idea to the next. In Russian they are represented by the following words: ‘firstly’ (Rus. ‘vo-pervyh’), secondly (Rus.’vo-vtoryh’), thirdly (Rus. ‘v tretyih’), in the fourth place (Rus. ‘v chetvertyh’), in the fifth place (Rus. ‘v-paytyh’) , etc. Numerical connectives typically show the relationship within a paragraph (or even within a sentence) between the main idea and the support the author gives for those ideas (Henry Madden Library, 2017). They are typically used to mark chronology of events or sequence of actions (The rules of Russian…, 2009).

Many researchers of language have pointed to the role of distributional frequencies in determining the relative accessibility or ease of processing associated with a particular lexical item or sentence (Ronald et al, 2007). These approaches are known by a number of names — constraint-based, competition, expectation-driven or probabilistic models — but all have in common the assumption that language processing is closely tied to a user’s experience, and that distributional frequencies of words and structures play an important (though not exclusive) role in comprehension.

## 3 METHODS

Text proficiency levels are distinguished by many different linguistic features, and numerical connectives can be one of the elements for grade level distinctions. For this purpose the team at Kazan Federal University (see Solovyev et all, 2018) compiled a corpus of textbooks on Social Studies by L. N. Bogolubov and A.F. Nikitin for the $5^{th}$ – $11^{th}$ Grades. Both sets of textbooks are from the "Federal List of Textbooks Recommended by the Ministry of Education and Science of the Russian Federation to Use in Secondary and High Schools". The course on Social Studies as a compulsory subject in all high schools of the Russian Federation is accomplished with a high stake matriculation exam after two school years: the $9^{th}$ and the $11^{th}$ grades. The choice of these particular sets of textbooks was caused by a number of reasons: (a) the fact that the texts under study were relatively free of non alphabetical symbols, graphs, figures etc., (b) the availability on the textbooks on the Internet (School textbooks and manuals, 2017). The Size of the corpus is presented in Table 1 (below). Sign "–"in Table 1 (below) marks absence of a textbook for the corresponding grade.

| Grade | Tokens | | |
|---|---|---|---|
| | BOG | NIK | TOTAL |
| $5^{-th}$ | – | 17,221 | 17,221 |
| $6^{-th}$ | 16,467 | 16,475 | 32,942 |
| $7^{-th}$ | 23,069 | 22,924 | 45,993 |
| $8^{-th}$ | 49,796 | 40,053 | 89,849 |
| $9^{-th}$ | 42,305 | 43,404 | 85,709 |
| $10^{-th}$ | 75,182 | 39,183 | 114, 365 |
| $11^{-th}$ | 100,800 | 38,869 | 139,669 |
| Total | | | 525,748 |

Table 1. *The Size of Corpus of textbooks on Social Studies by L. N. Bogolubov and A.F. Nikitin for the $5^{th}$ – $11^{th}$ Grades of Russian secondary Schools.*

The size of the collections of texts as presented in Table 1 (above) is 525,748. A token is viewed in the work as an instance of a sequence of characters in some particular document that are grouped together as a useful semantic unit for processing. Usually it refers to the total number of words in a text, corpus etc, regardless of how often they are repeated. A type is the class of all tokens containing the same character sequence (Tokenization, 2008).

To analyze the Corpus and determine the relevant frequency of the above mentioned connectives in the texts, we employed AntConc (AntConc, 2014). AntConc computes a list of words which appear with the keywords under study thus providing not only a concordance but distributional patterns of the keywords.

## 4. ANALYSIS

As it was mentioned above, the present study investigates two research questions:

RQ1: Are the variety and frequency of numerical connectives in linear regression to text complexity?

RQ2: What syntactic patterns are used after numerical connectives 'firstly' (Rus. vo-pervyh), secondly (Rus.vo-vtoryh) etc. in the Russian academic discourse?

The research demonstrated a gradual growth in the absolute amount of the linkers from 0 (for each linker) in the book for fifth-graders to10 in the book for tenth/eleventh-graders (see Table 2 and Fig. 1 below).

| Connectives | Firstly | Secondly | Thirdly | In the Fourth Place | In the Fifth Place | Total |
|---|---|---|---|---|---|---|
| Grade 5 | 0 | 0 | 0 | 0 | 0 | 0 |
| Grade 6 | 3 | 3 | 1 | 0 | 0 | 7 |
| Grade 7 | 4 | 4 | 2 | 1 | 0 | 11 |
| Grade 8 | 6 | 6 | 5 | 2 | 1 | 20 |
| Grade 9 | 7 | 7 | 4 | 1 | 1 | 20 |
| Grades 10-11 | 10 | 9 | 2 | 0 | 0 | 21 |
| Total | 30 | 29 | 14 | 4 | 2 | 93 |

Table 2. *The total amount of connectives in the textbooks of grades $5^{th}$-$11^{th}$*.

The graph below demonstrates obvious growth of the total number of numerical connectives in each of the grades under study: from 0 in the 5th Grade to 21 in the 10th -11th Grades.

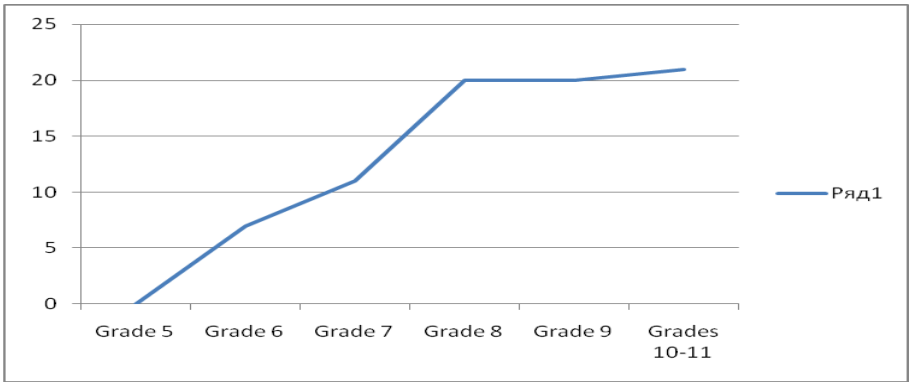


Fig.1. Growth *of raw counts of numerical connectives in Russian textbooks on Social Studies of grades 5th-11th.*

The graph clearly shows that numerical connectives in question are more frequent in sub-corpus of textbooks of 10th-11th grades, but the modern linguistic paradigm admits that since lengths of corpora/collections of texts/ separate texts are different, raw frequency counts of any linguistic feature are not indicative or reliable and as such are not to be compared to each other. Therefore, "the counts have to be normalized to a common basis and hence rendered comparable. The raw frequency counts should be divided by the total number of words of the text and then multiplied by the chosen basis" (Biber, Conrad and Reppen 1998: 263). In Table 3 below we present raw and two versions of normalized frequencies: in Column 4 we normalize the counts of connectives to the size of the smallest sub-corpus, i.e. the sub-corpus of the 5th Grade where no numerical connectives were registered. In Column 5 we offer the frequency normalized to 5000. Both ways of representing the data demonstrate stability of the metric within the range of the 6th – 9th grades: it is more than 3.6 and less than 4.0, if normalized to 17221 or about 1,0, if normalized to 5000. The latter means that in each sub-corpus of grades 6, 7, 8, 9 numerical connectives are used once in every 5000 tokens. Its normalized frequency is about 3 times less than in the sub-corpus of grades 10th and 11th.

| Grade | Tokens | Raw Frequency | Normalized Frequency (To 17221) | Normalized Frequency (To 5000) |
|---|---|---|---|---|
| 1 | 2 | 3 | 4 | 5 |
| 5th | 17221 | 0 | – | 0 |
| 6th | 32942 | 7 | 3,65937 | 1,0624 |
| 7th | 45993 | 11 | 4,11869 | 1,1958 |
| 8th | 89849 | 20 | 3,83332 | 1,1129 |
| 9th | 85709 | 20 | 4,01848 | 1,1667 |
| 10th-11th | 254034 | 21 | 1,42359 | 0,4133 |
| Total | 525748 | 93 | 3,046236 | 0,88445 |

Table 3. *Raw and normalized frequencies of numerical connectives in Russian Academic Corpus.*

| Connectives | Russian National Corpus | Raw Frequency | Normalized Frequency (To 5000) |
|---|---|---|---|
| 1 | 2 | 3 | 4 |
| Vo-pervyh (Rus. 'firstly') | 283 431 966 | 20 783 | 0,3666 |
| Vo-vtoryh (Rus. 'secondly') | 283 431 966 | 16 554 | 0, 02920 |
| V tretyih (Rus. 'thirdly') | 283 431 966 | 3 390 | 0,0059 |
| V chetvertyh (Rus. 'in the fourth place') | 283 431 966 | 687 | 0,01211 |
| V-paytyh (Rus. 'in the fifth place') | 283 431 966 | 222 | 0,00391 |
| TOTAL | | 41636 | 0,73449 |

Table 4. *Raw and normalized frequencies of numerical connectives in Russian National Corpus. (Russian National Corpus, 2018)*.

As we see in Table 4 below raw frequency of numerical connectives in the Russian National Corpus falls in the range from 20783 for 'vo-pervyh' (Rus. 'firstly') to 222 for 'v-paytyh' (Rus. in the fifth place) and normalized frequency of all the connectives under study is 0,73449.

To answer the second Research Question on distributional patterns of the contexts of numerical connectives 'firstly' (Rus. vo-pervyh), secondly (Rus. vo-vtoryh) etc. in the Russian academic discourse, we manually tagged the context of after the numerical connective.

The following tags were used:

1. Indicative pronouns: Vo-vtorykh, **eto** obshchestvenno opasnoye deyaniye, ono nanosit ser'yeznyy [Secondly, this is a socially dangerous act, it causes serious].

2. Noun phrases: Vo-vtorykh, **deyatel'nost' cheloveka v obshchestve** v tselom vo [Secondly, human activity in society as a whole].

3. Infinitive constructions: V-tret'ikh, **neobkhodimo podchinit'** svoye povedeniye moral'nym normam, privyknut' [Thirdly, it is necessary to subordinate one's behavior to moral norms, to get used to].

**4.** Adverbial modifiers of time, cause, purpose, place: V-vtorykh, **v stranakh s razvitoy rynochno yekonomikoy sushchestvuyet** [Secondly, in countries with developed market economies, there is]**.**

5. Verbs: V-chetvortykh, **ne dopuskayutsya** braki mezhdu blizkimi rodstvennikami [In the fourth place, marriages between close relatives are not allowed].

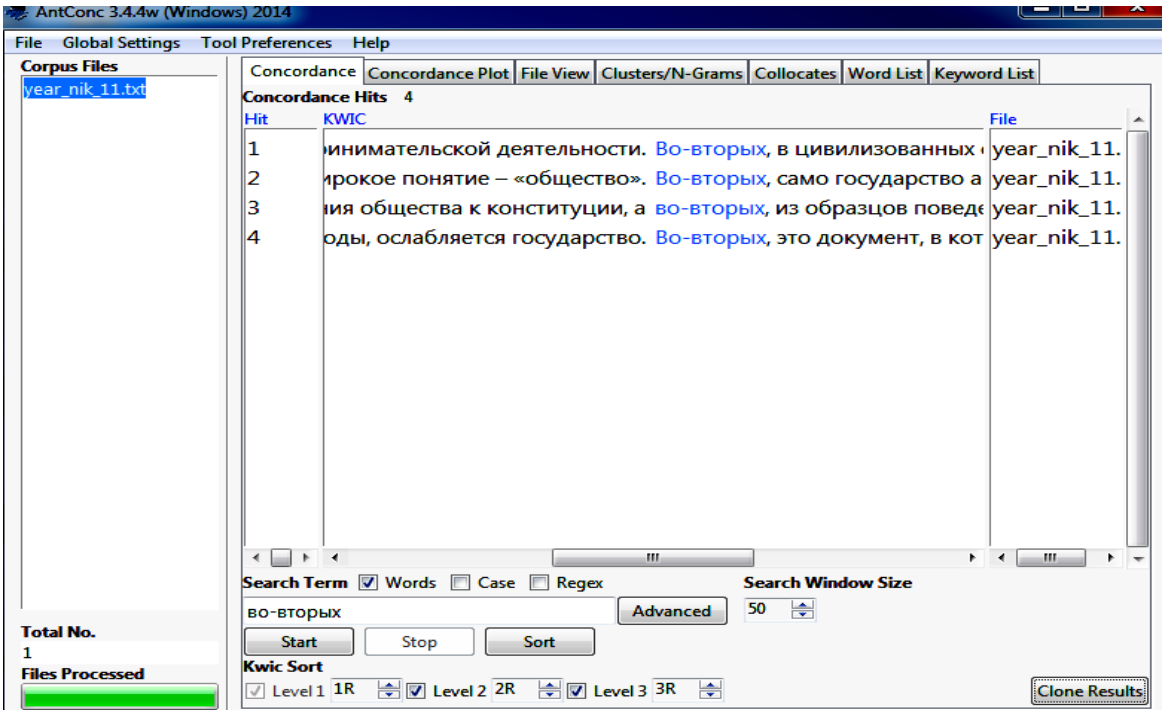


Fig.2. *Distributional patterns of the word 'Vo-vtorykh' (lit. secondly) in Grade 11 (Textbook ed. by Nikitin).*

*vo-pervykh*

GRADE 6

vo-pervykh, **v shirokom smysle vso to, chto sozdano** [in general, everything that was created]

vo-pervykh, **normy prava** zakrepleny v zakonakh i ustanavlivayutsya [First, the norms of law are fixed in laws and are established…]

vo-pervykh, **pravo grazhdan** izbirat' i byt' izbrannymi v [First, the right of citizens to be elected and elected]

GRADE 7

vo-pervykh, **nado ponyat' neobkhodimost'** i znacheniyemoral'nykh norm [Firstly, it is necessary to understand the necessity and significance of moral norms]

vo-pervykh, **fiziologicheskoye sostoyaniye cheloveka** otrozhdeniya do smerti [First, the physiological state of a person from birth to death]

vo-pervykh, **eto** zhizn' v khoroshey, zabotlivoysem'ye [First, it is life in a good, caring family]

vo-pervykh, **eto** neobkhodimost' otvechat' (derzhat' otvet, nestinakazaniye) [First, it is the need to respond (to keep an answer, to bear punishment)].

It is also informative to look at the types of differences we found between the syntactic models of the context.

| **Type of a Syntactic Construction** | % |
|---|---|
| Noun Phrase | 34 |
| Infinitive Constructions | 12 |
| Indicative Pronouns | 9 |
| Adverbial Modifiers Of Time, Cause, Purpose | 42 |
| Verbs | 3 |

Table 5. *Syntactic models of the distributional patterns used after the numerical connectives.*

The syntactic analysis of the context of numerical connectives indicates that the most widely used syntactic construction after all numerical connectives in the Russian academic discourse of Social science are (in descending order) the following: 1) SVO, where the subject (S) is manifested with a noun phrase (about 34%), infinitive constructions (12%), indicative pronouns (9%); 2) inverted word order OSV (42%) 3) VSO (3%).

## 5. DISCUSSION

The results of the study aimed at defining frequencies of numerical connectives in Russian Academic Corpus proved that numerical connectives tend to have the same normalized frequencies within the range of grades 6 – 9. The raw frequencies grow from the 5$^{th}$ grade to the 11$^{th}$ grade together with the growth of tokens in the textbooks of the same range.

We also suggest that frequencies of numerical connectives form a foundation to represent texts of different complexity. We believe that language distributional data such as that in this paper will play an important role in understanding the nature of text complexity. The perspective of the study lies in compiling a bigger corpus (of texts on other subjects) to extrapolate the results of the current research and the hypothesis to other types of texts and genre.

## 6. CONCLUSION

The article presents the results of the analysis aimed at quantitative (raw and normalized frequencies) and distributional (use in various contexts) characteristics of the following connectives: firstly “vo-pervyh” secondly “vo-vtoryh”, thirdly “v tretyih”, in the fourth place “v chetvertyh”, in the fifth place “v-ptaytyh” in the texts on Social Studies for Russian pupils of secondary school.

The positive regression of the numerical connectives in the Corpus of Russian Academic texts demonstrates the interrelation between the use of certain syntactic patterns and text complexity. The steady growth of raw counts of numerical connectives from grade to grade reflects the dependence of complexity upon different levels of linguistic awareness.

The syntactic analysis of the context of transitional connectives indicate that the most widely used syntactic construction after all numerical connectives in the Russian academic discourse of Social science are (in descending order) the following: 1) SVO, where the subject (S) is manifested with a noun phrase (about 34%), infinitive constructions (12%), modal constructions (9%); 2) inverted word order OSV (42%) 3) VSO (3%). The results of this research provide us with insights into the general patterns of the academic discourse and cohesion.

In general the study expands the range of text modeling features aimed at better identifying the notion of text complexity.

## ACKNOWLEDGEMENTS

The research presented in Part Method of the article was supported by the subsidy of the Russian Government to support the Program of Competitive Growth of Kazan Federal University.

The research presented in all other parts, including Results and Discussion, was financially supported by the Russian Science Foundation, grant № 18-18-00436.

## REFERENCE LIST

AntConc, (2018). Softpedia. http://www.softpedia.com/get/Science-CAD/AntConc.shtml

Authors' Database, (2017). http://kpfu.ru/portal/docs/F1554781210/shuffled.zip

Biber, D. (2006). A corpus-based study of spoken and written registers. University Language.

BIBER, D., CONRAD S., REPPEN R. Corpus linguistics : investigating language structure and use. Cambridge: University Press, 1998. ix, 300. ISBN~0521496225.

Bogolubov, L. N., (2017). Social Studies. 11th grade (basic level) . Textbook. - 3rd ed. Moscow. 335 p. - ISBN 978-5-09-046529-8

Goryainov, A. (2017) Reflections on education and textbooks. http://www.scienceandapologetics.org/text/488.htm

Henry Madden Library, (2017). Transition words. http://guides.library.fresnostate.edu/c.php?g=288903&p=1927133v

Klokova, G.V. (2017). New generation history textbooks: an analytical study. http://eidos.ru/journal/2004/0417-03.htm

National Corpus of Russian Language. (2018) http://www.ruscorpora.ru/search-main.html

Nikitin, AF, Gribanova, G.I, Skorobogatko, A.V, Martyanov, D.S (2017) Social Studies. Textbook. Drofa. 240 p.

Plain Writing: It's the Law! (2018). U.S. Food and Drug Administration. https://www.fda.gov/aboutfda/plainlanguage/

Paronjanov, V.D.(2017). How to write a good tutorial for good people. Textbooks dream of students and schoolchildren. Moscow. 500 p. http://drakon.su/_media/parondzhanov_.pdf.

Rakitin, P. (2018). Automated assessment of the complexity of texts. https://sites.google.com/site/moimstudentam/graduation-paper/home/kak-eto-delaut-drugie/avtomatizirovannaa-ocenka-sloznosti-tekstov.

Roland, D., Dick, F., & Elman, J. L. (2007). Frequency of Basic English Grammatical Structures: A Corpus Analysis. Journal of Memory and Language, 57(3), 348–379. http://doi.org/10.1016/j.jml.2007.03.002

Shankin, A.A. (2015) Age anatomy and physiology. Directmedia. ISBN 5447548543, 9785447548544

174 p

School textbooks and manuals (2017). Vseuchebniki. http://vseuchebniki.net/

Solnyshkina M.I, Zamaletdinov R.R, Gorodetskaya L.A. (2017) Evaluating text complexity and Flesch-Kincaid grade level//Journal of Social Studies Education Research. Vol.8, Is.3. - P.238-248

Solnyshkina M.I, Vishnyakova O.D., Gafiyatova E.V. Gabitov A. I. (2017) English textbooks for Russian students: Problems and specific features. Journal of Social Studies Education Research. Vol.8, Is.3. P.215-226

Solovyev, V. D., Solnyshkina, M.I., Ivanov, V.V., Timoshenko, S. V., (2018) Complexity of Russian Academic Texts as the Function of Syntactic Parameters, 19th International Conference on Computational Linguistics and Intelligent Text Processing, CICLing 2018, March 18 to 24, 2018, Hanoi, Vietnam

The rules of Russian spelling and punctuation. Full academic directory. (2009) Editor V.V. Lopatin. Moscow. 432 p. http://orthographia.ru/punctum_uk.php?pid=120

Tokenization. (2008) Cambridge University Press. https://nlp.stanford.edu/IR-book/html/htmledition/tokenization-1.html