Development and implementation of software application for comparative analysis of difficulty models for text data

Olga Gavenko[1, 2][0000-0003-3619-1120] and Sofia Obersht[2]

[1] Federal Research Center for Information and Computational Technologies, Novosibirsk, Russia
olga.yu.gavenko@mail.ru
[2] Novosibirsk State University, Novosibirsk, Russia
s.obersht@g.nsu.ru

Abstract. The complexity of text is a complex concept consisting of difficultness, readability and comprehensibility and describing the text structure. The determination of text complexity has applied significance in understanding and processing of information and knowledges. Subjective parameters of text include empirical data on the reader's perception of the text, physical and cognitive abilities, knowledge and education of an individual. Objective parameters are divided into quantitative such as length, frequency of usage or number of tokens, and qualitative which are related to the analysis of linguistic means of categorical language levels and their implementation. The task becomes more complicated with the usage of the large text data. Defining text as a character sequence, the estimating model of complexity can be developed, the choice of the objective parameters, as well as methods of complexity estimation can vary; most of the formulas are universal and based on the linear-regression model. The goal of this paper is the development and implementation of software application in Python and the comparative analysis of basic formulas for English and adapted for the Russian. School textbooks on Social Studies, 5-11 classes (Russian Readability Corpus), make the test sample. The experiments with the text corpus data shows incorrect results what is explained by the fact that the model development based on the texts of different genres and styles and the difference in languages; in addition, the fact, that quantitative parameters may not be sufficient to obtain reliable results, should be taken into account when expanding corpus data.

Keywords: Text complexity, readability estimates, text corpus data.

1 Introduction

One of the ways to transmit information from the carrier to the recipient is a text, which in this context is considered as a form of storing and transmitting knowledge in a certain sequence of symbols. The success of transmitting information without losing any components depends on how accurately a recipient perceives the text. The determination of text complexity has applied significance as it is allowed to be adapted to the audience for better understanding, hence knowledge transmitting. It is important when compiling educational literature, legislative and regulatory documents, various technical

documentation, etc. The task becomes more complicated when the complexity of large text data needs to be estimated

Estimation of the text complexity is one of the important tasks in the field of natural language processing. Depending on the purpose of the study classical mathematical methods or machine learning methods are used.

Despite the huge amount of research, the question of choosing a model and reliability of results by applying the model to texts of different genres and styles is opened. In case of using mathematical models of text assessment the readability formulas already developed are used: the Flesch–Kincaid Grade Level Formula, Automated Readability Index (ARI), SMOG, etc. At the same time the language differences for which a formula was developed or applied are not always taken into account. And it is necessary to consider the fact that model was obtained by analyzing data of a narrow domain and is used also for specific texts.

2 Text complexity concept

Understanding of the text in a general sense depends on various factors, which are the objects of research in different domains such as linguistics, psychology, sociology, etc. In this work, the "complexity" of the text is analyzed as an object in terms of the applicability of computational linguistics methods. The very concept of text "complexity" is complicated and must be precisely described for the formulation of the analysis target. However, due to the lack of general approach to the usage of the term, some ambiguity of the interpretation arises, what is common in linguistics. As a characteristic of "complexity" the following concepts are used: difficulty, complexity (in another meaning), readability and comprehensibility. Since there is no unified approach to defining these concepts, they are mixed and redefined from work to work by various authors. So, in I.V. Oborneva's work the difficulty of the text is meant as exclusively subjective "complexity" of the text in a wide sense of word and readability and complexity (as a characteristic of a wide concept of "complex") are considered synonyms [3]. On the other hand, in paper written by M.M. Nevdakh the concept of difficulty includes both subjective and objective criteria, at the same time, terms complexity and readability are used along with it [4]. Wherein, it should be noted that there is a tendency to divide the concepts according to what features are included in it.

Traditionally the difficulty consists of both objective and subjective parameters. Some authors include in this concept the width of knowledge about the world, social, historical, psychological, age, time and a number of other factors, the genre and style of the text. In the works of other authors, the difficulty of the text is determined on the basis of empirical data on the subjective perception of the text by the reader, his knowledge and skills [5, 6]. It is also necessary to take into account that the perception of any text is influenced by the font style, the size and color of words, the location of the text, etc.; in addition, the perception of the text depends on the physical and cognitive abilities, knowledge and education of the individual in a specific domain, and experience in general. Moreover, since text difficulty is an interdisciplinary concept, the

choice of significant characteristics may vary depending on the direction of research and the choice of scientific approach to studying the problem [7].

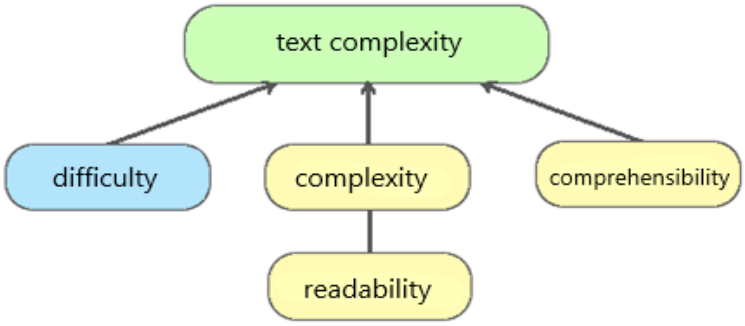

Fig. 1. Components included in concept of text "complexity". Characteristics consisting of only subjective parameters are shown in yellow, characteristics consisting of only objective parameters are shown in blue.

Despite the different interpretations of the terms, complexity, readability and comprehensibility can be distinguished into a general group, because in most cases of their usage they can be interpreted as a set of objective parameters, hence depending only on the text. In turn the objective parameters are divided into quantitative, such as length, frequency of usage or amount of language units, and the qualitative parameters related to the analysis of linguistic means of categorical language levels and their implementation in a text.

Readability in Russian scientific literature is interpreted in two ways: a measure of accessibility for understanding a written text, determined by the analysis of a number of factors, including syntactic complexity, vocabulary, expression of the topic, coherence of topics, etc. or a measurement of how readable a text is, based on the average level of readers who can read and understand it [8]. Because the development of the model estimation of the text complexity has applied significance (mostly in tasks of educational literature and legislative documents analysis), the most of readability formulas, starting from the second half of the 20th century, determine the readability index and show text complexity level depending on the age of the reader or the number of years of academic study being necessary for understanding. The readability index is traditionally calculated using qualitative parameters, such as amount of words in sentence, average word length, average sentence length, etc.

The concept of text complexity is usually used as a synonym for characteristics such as difficulty and readability. However, in the context of statistical analysis of text using only objective parameters, complexity can be interpreted as an expanded concept of readability including both quantitative and qualitative parameters of the text. In general meaning this is the objective characteristic of the text depending on internal text parameters [7]. So, E.S. Pushkina includes in concept of complexity the structural and

semantic parameters of the text: sentence size, text size, terms, length of the terminological chain, derivational structure, positional structure as a system-forming factor [9].

Comprehensibility is used much less frequently compared to complexity and readability due to its rather vague definition. In case of using this term ambiguity in parameters determination including in it appears. Ya.A. Mikk introduces the term "familiarity" along with other objective text parameters as one of the text comprehensibility parameters. It means an amount of words in the text familiar to the reader, what makes this term a subjective parameter determined experimentally [10].

3 Background

The first work on text complexity assessment appeared in the 1950s with R. Flesh's (1948) [11], E. Dale's and J. Chall's (1948) [12] and R. Gunning's (1952) [13] formulas. In the 1970s, Russian scientists analyzed the complexity of the text: M.S. Matskovsky (1976) [14], Yu.A. Tuldava (1975) [15] and others. All works in this area were dictated by applied purposes: Flesch's formula and Flesch-Kincaid test were first used by the American army for assessing the complexity of technical manuals, now they are used for assessing juristic documents and laws. Models for the Russian language were initially aimed at assessing the complexity of educational texts and students' ability to understand them.

Gradually the application area of statistical methods for determining text complexity are expanding: along with educational text analysis texts of different genres and styles, fiction and newspaper are studied.

However, due to the difference in language means, terminology and structure, which are specific for each genre and style, the specter of text features is changing and being complemented with phonological, lexical, morphological, syntactic features. Besides these parameters the American scientists include parameters related to discourse such as entity-density features, lexical-chain features, coreference, etc; part-of-speech-based grammatical features for different word classes, content and function words [16].

Russian scientists, when studying texts of various styles and genres, also identify the additional parameters: words of different part-of-speech classes, individual grammemes, general-language frequency of lemmas, multi-word expressions, cohesion assessments, word-formation patterns [17].

Modern development in the natural language processing provide the possibility to take into account a wide range of metrics. Besides traditional statistical methods of text assessment, classical machine learning [18] and deep learning approaches [19] are used. In this case readability formulas developed earlier are included in the analysis as shallow traditional features.

4 Mathematical models of text complexity

At the present time, the tendency in the development of text analysis is expanding the spectrum of model parameters, searching for connections between them and building models for various texts.

The problems appearing during the determination of text complexity are following:
– expanding of spectrum of the parameters of text complexity;
– correlation/ interconnection of the parameters;
– problem of the selection of parameters for a specific task;
– dependence of quantitative and qualitative parameters.

In this work the concept of text readability as the only component of the general concept of "complexity" for which there are evaluation formulas is considered. In the present time there are about 200 text readability formulas for various languages, most of them are developed for English, part of them are adapted for the Russian language taking into account its characteristics. There are several formulas developed for the Russian language by Mikk, Matskovsky, Nevdakh, et al. Most of the text readability formulas invented in 20 century are based on the linear-regression model, parameters of which are statistic (quantitative) text features [20]:

$$f(\boldsymbol{x}, \boldsymbol{b}) = b_0 + \sum_{i=1}^{N} b_i x_i$$

where $\boldsymbol{x}$ – independent model parameters, $\boldsymbol{b}$ – numeric coefficients.

The main features of the model are presented in Table 1.

Table 1. Linear-regression model characteristics.

Advantages	Disadvantages
simplicity of implementation	low accuracy of results
simplicity of interpretation of results	the complexity of determining parameters
availability of intermediate results	unable to apply to nonlinear processes
fast forecasting result	correlation between different methods

The main formulas of text readability estimation for English and Russian are presented in Table 2.

Table 2. Text readability formulas for English and Russian languages

Method	Formula	Parameters	Language
Flesch [11]	206.835 – 1.015*ASL – 84.6*AWL	ASL – average sentence length, AWL – average word length	eng
Flesch (modified by Oborneva) [3]	206.836 – 1.52*ASL –65.14*AWL	ASL – average sentence length, AWL– average word length	rus

Mikk [21]	$0.01x_1 + 0.27x_2 + 0.54x_3$	x_1 – sentence length in symbols; x_2 – percentage of different unknown words, x_3 – abstractness of repeated concepts expressed by nouns	rus
Matskovsky [14]	$x_1 = 0.62x_2 + 0.123x_3 + 0.051$	x_1 – assessing difficulty obtained by applying the successive interval method, x_2 – ASL– average sentence length, x_3 – percentage of words with >3 syllables	rus
Spache [22]	0.121*ASL + 0.082*PDW + 0.659	ASL – average sentence length, PDW – percentage of difficult words	eng
Gunning [23]	0.4*(ASL + 100*PDW)	ASL – average sentence length, PDW – percentage of difficult words	eng
Dale-Chall Formula [24]	$0.1579x_1 + 0.0496x_2$	x_1 – percentage of difficult words, x_2 – average sentence length in words	eng
Flesch-Kincaid [25]	0.39*ASL+ 11.8*AWL – 15.59	ASL – average sentence length, AWL – average word length	eng
Tuldava [15]	R(i,j)= i*lg(j)	i – average word length in syllables, j – average sentence length in words	rus
Shpakovsky [26]	$20.24 + 0.48x_1 + 0.58x_2 + 0.41x_3$	x_1 – percentage of words that are 9 or more letters long, x_2 – percentage of terms, x_3 – percentage of the number of symbols in chemical reactions	rus
Nevdakh [4]	$Y = -16.7873 + 0.7602x_1 - 0.1002x_2 + 1.4484x_3 + 0.0283x_4$ $Y = -20.3376 + 0.4448x_1 - 0.0419x_2 + 1.052x_3 + 0.679x_4$	x_1 – average paragraph length in words, x_2 – average paragraph length in letters, x_3 – percentage of words 11 or more letters long, x_4 – percentage of words 13 or more letters long	rus

Automated readability index (ARI) [27]	4.71*(number of letters/number of words) + 0.5*(number of words/number of sentences) – 21.43	–	eng
Coleman-Liau [28]	0.0588L – 0.296S – 15.8	L – average number of letters per 100 words, S – average number of sentences per 100 words	eng
Mc Laughlin's SMOG test [29]	1.043*sqrt (x_1*30/x_2) + 3.1291	x_1 – number of polysyllabic words, x_2 – number of sentences	eng
Sticht (FORCAST formula) [30]	20 – k*0.667*b	k – number of words in text, b – number of monosyllabic words	eng

5 Formulas' adaptation to the Russian language

Talking about the adaptation of English indices to Russian language it is necessary to take into account the categorical characteristics of each language. Russian is a synthetic type of language in which the grammatical relationships are expressed through changes in the morphemes of one word. Wherein the word order in a sentence is not fixed and there is inversion, which does not affect the perception of meaning stated in the sentence. Inversion of the words in the sentence, ellipsis and different types of sentences inherent in Russian speech do not affect the perception of the semantic too. English is an analytic type of language, in which the grammatical and lexical relationships are beyond one word. They are expressed through individual function words such as prepositions and modal verbs, lexically significant units, which almost do not change. In addition, the words are grouped around the predicate, which cannot exist without the subject. As a result, English tends to shorten word length and increase the number of words to convey the same information in different ways.

A major contribution to the adaptation of formulas for the Russian was made by I.V. Oborneva who recalculated coefficients of Flesch index taking into account the fact that average word length in Russian is longer compared to English and average sentence length is shorter simultaneously. To clarify the coefficients in Flesch's formula a study on average word length in Russian and English was conducted. Russian dictionary edited by S.I. Ozhegov (39174 words) and English-Russian dictionary edited by V.K. Muller (41977 words) was taken as a sample. According to the research the average word length in English dictionary is 2.97 syllables, average word length in Russian is 3.29 syllables. In addition, the number of polysyllabic words (with the number of syllables of 3 and more) in the Russian dictionary are 7% more. Following results were obtained by studying literary texts in English and their translations into Russian with a total volume of 6 million words. As a result, number of syllables in English is 0.71 times less compared to Russian, sentence length is 1.25 times greater [3].

Similar recalculations were made by I.V. Begtin for the basic English models: Flesch index, Automated readability index (ARI), Coleman-Liau index, Dale-Chall formula, McLaughlin's SMOG test. Results are presented as an Internet source [31].

6 Models and dataset

The most used formulas for the estimation of quantitative parameters related to text complexity are the following: Flesch index, Flesch-Kincaid test, Automated readability index (ARI), Coleman-Liau index, Coleman-Liau index, McLaughlin's SMOG test. They are developed for English texts but are considered to be universal. However, the fundamental language differences are not taken into account in full during adaptation to Russian. Flesch and ARI indices are adapted to Russian.

Listed indices are included in the experiment, the main goal of which is the development and implementation of software application in Python able to do the comparative analysis of existing readability assessments for text data and the identifying the formula to estimate quantitative parameters of text complexity as accurately as possible. The selected indices are applied to the test sample for assessment of the quality of the results taking into account the language for which the index is developed or adapted. Despite the rise of research on text assessment using machine learning approaches, there is a lack of data in Russian language in open access. Two collections of text were assembled for the initial research as a test sample. Both are included in Russian Readability Corpus and available for use [32]. There are textbooks on Social Studies by L.N. Bogolyubov with 6-11 grade levels and by A.F. Nikitin with 5-11 grade level. The main information about text data is described in Table 3.

Table 3. Text data for the experiment.

	L.N. Bogolubov			A.F. Nikitin		
Grade level	Total number of words	ASL	AWL	Total number of words	ASL	AWL
5	–	–	–	16804	11.4468	2.100
6	16034	13.099	2.309	15850	14.1897	2.3784
7	22226	13.934	2.555	22122	14.281	2.386
8	47839	16.080	2.643	38410	14.549	2.559
9	40619	17.240	2.697	41594	15.684	2.657
10	72369	17.488	2.730	37830	17.281	2.733
10*	95449	17.414	2.723	–	–	–
11	–	–	–	37134	18.061	2.7365
11*	96541	16.607	2.865	–	–	–

ASL – average sentence length, AWL – average word length. Star sign (*) denotes the advanced version of book for the corresponding grade, dash sign (–) denotes the absence of book.

7 Results

The Flesch reading-ease test is a score between 1 and 100, scaled by 10, where 1 represents 5-th U.S. school level and 100 indicates text best understood by students and academicians. The result of the Flesch–Kincaid Grade Level Formula and Automated Readability Index (ARI) is the age needed for text understanding, which can be interpreted as the number of years of schooling in the American educational grading system or Russian educational grading system for adapted indices.

Results for Bogolubov and Nikitin books are presented in Tables 4 and 5.

Table 4. Results for Bogolubov textbooks

Grade	Flasch		Flasch adapted		Flasch-Kincaid		Flasch-Kincaid adapted		ARI		ARI adapted	
	R	I	R	I	R	I	R	I	R	I	R	I
6	-1.8	out of range	36.5	student	16.8	11	10.4	5	11.7	6	7.9	3
7	-23.5	out of range	19.2	student	20	student	12.8	7	14.9	9	11.9	7
8	-33.1	out of range	10.2	student	21.9	student	14.7	7-8	17	11	13.9	8
9	-38.9	out of range	4.9	student	22.9	student	15.7	8-9	18	12	14.8	9
10	-41.9	out of range	2.4	student	23.5	student	16.1	10	18.4	student	15.2	10
10*	-41.3	out of range	2.9	student	23.3	student	16	10	18.2	student	14.9	9
11*	-52.4	out of range	-5.1	out of range	24.7	student	16.8	11	19.5	student	16.9	11

"R" denotes results: "out of range" – result is out of index bounds, "I" denotes interpretation of the index: "student" denotes texts being understood by university graduates.

Table 5. Results for Nikitin textbooks

Grade	Flasch		Flasch adapted		Flasch-Kincaid		Flasch-Kincaid adapted		ARI		ARI adapted	
	R	I	R	I	R	I	R	I	R	I	R	I
5	17.5	student	52.6	10	13.7	8	7.8	2	8.5	3	4.4	preschool
6	-8.8	out of range	30.3	student	18	12	11.5	6	13.1	7	9.4	4
7	-10	out of range	29.7	student	18.1	12	11.6	6	13.3	8	9.6	4
8	-24	out of range	18	student	20.3	student	13.2	7	15.3	9	12.2	7
9	-40	out of range	9.9	student	21.9	student	14.6	9	16.9	11	13.9	8
10	-42	out of range	2.5	student	23.4	student	16	10	18.4	student	15.3	10
11	-43	out of range	1.1	student	23.7	student	16.4	11	18.9	student	15.6	10

"R" denotes results: "out of range" – result is out of index bounds, "I" denotes interpretation of the index: "student" denotes texts being understood by university graduates, "preschool" denotes extremely easy text.

As a result of applying formulas to the test sample a number of incorrect results are observed, in particular, Flesch index interprets 6th grade level text as difficult as students' books are. In general, the indices developed for the English language give less accurate results compared to adapted ones, what is due to the difference in languages. Models are sensitive to initial parameter changes, what can be explained by model development based on the texts of different genres and styles, which have various linguistic means, terminology and structure.

Thus, despite the fact that the indices taken for analysis are considered to be universal and potentially applicable to texts of any genre and styles, the experiments carried out using the developed software application show that it is necessary to take into account the language, for which formulas are developed, and the fact, that quantitative parameters may not be sufficient to obtain reliable results.

Future research assume to expand the text data in the purpose of more accurately determination of the text parameters used in formulas that affect the objectivity of the results; to modify the formula given more accuracy result to Russian language; to develop a methodology allowed to take into account a wide range of quantitative characteristics of complexity including both quantitative and qualitative parameters.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Great Dictionary of Russian language. Shangwu Yinshuguan, 1481 p. (2020).
2. Dictionary of modern Russian language. Publishing House of the USSR Academy of Sciences, Leningrad-Moscow, in 16 v. (1957).
3. Oborneva I.V.: Automated estimation of complexity of educational texts on the basis of statistical parameters. Pedagogy Cand. Diss, Moscow, 165 p. (2006).
4. Nevdakh M.M., Zil'bergleyt M.A.: The development of the methodology for text complexity estimation and its software implementation. Proceedings of BSTU, Series 4: Print and media technology. 9, pp. 84-88 (2010).
5. Solnyshkina M.I., Kisel'nikov A.S.: Text complexity: stages of study in domestic applied linguistics. Vestnik of Tomsk State University. Philology, 6 (38), pp.86-99 (2015).
6. Shpakovsky Yu.F.: Assessment of text comprehension difficulty. Proceedings of BSTU, Series 4: Print and media technology, 9, pp. 72-75 (2012).
7. Tomina Yu.A.: Objective assessment of the language difficulty of texts (description, narration, reasoning, evidence). 225 p. (1985).
8. Reber A.S.: Oxford Dictionary of Psychology. 864 p. (2002).
9. Pushkina E.S.: Theoretical and experimental study of structural and semantic parameters of the text. Philology Cand. Diss., Kemerovo, Kemerovo State University, 155p. (2004).
10. Kisel'nikov A.S.: To the problem of the text characteristics: readability, comprehensibility, complexity, difficulty. Philology science. Theory and practice questions, 11 (2), pp. 79-84 (2015).
11. Flesch R.: A new readability yardstick. Journal of applied psychology, 32 (3), 221 p. (1948).
12. Dale E., Chall J.S.: A formula for predicting readability. Bureau of Educational Research, Ohio State University (1948).
13. Gunning R.: The Technique of Clear Writing. (1952).
14. Matskovskiy, M.S.: Problems of printed material readability. Semantic perception of verbal communication in the context of mass communication, pp. 126-142 (1976).
15. Tuldava, Yu.A.: On measuring the complexity of the text. Scientific notes by Tartu State University. Works on methods of teaching foreign languages. 345, pp. 102–120 (1975).
16. Feng L. et al.: A comparison of features for automatic readability assessment. Coling 2010: Posters, pp. 276-284 (2010).
17. Blinova O., Tarasov N.: A hybrid model of complexity estimation: Evidence from Russian legal texts. Front. Artif. Intell., 5:1008530, (2022).
18. Lyashevskaya O., Panteleeva I., Vinogradova O.: Automated assessment of learner text complexity, Assessing writing, s. 100529, v. 49 (2021).
19. Lee B. W., Jang Y. S., Lee J.: Pushing on Text Readability Assessment: A Transformer Meets Handcrafted Linguistic Features. Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, s. 10669-10686, (2021).
20. Mizernov I.Yu., Grashchenko L.A.: The analysis of methods of text complexity assessment. New information technologies in automated systems, 18, pp. 572-581 (2015).
21. Mikk, Ya.A.: Factors of educational text clarity. Abstract of Pedagogy Cand. Diss., Tartu, (1970).
22. Spache G.: A new readability formula for primary-grade reading materials. The Elementary School Journal, 53 (7), pp. 410-413 (1953).
23. Gunning R.: The technique of clear writing. McGraw-Hill, New York, (1952).
24. Dale E., Chall J. S.: A formula for predicting readability: Instructions. Educational research bulletin, pp. 37-54 (1948).

25. Kincaid J.P., Fishburne R.P., Rogers R.L., Chissom B.S.: Derivation of new readability formulas (Automated Readability Index, Fog Count and Flesch Reading Ease Formula) for Navy enlisted personnel (Research Branch Report 8–75). Memphis, TN: Naval Air Station, p. 40 (1975).
26. Spakovsky Yu.F.: Evaluation of perception challenges and complexity optimization of the educational text (on a material of texts on Chemistry). Abstract of Philology Cand. Diss., Minsk, (2007).
27. Smith E. A.: Automated readability index. AMRL-TR, (1967).
28. Coleman M., Liau T. L.: A computer readability formula designed for machine scoring. Journal of Applied Psychology, 60 (2), 283 p. (1975).
29. Mc Laughlin G. H.: SMOG grading-a new readability formula. Journal of reading, 12 (8), pp. 639-646 (1969).
30. Sticht T. G.: Research Toward the Design, Development and Evaluation of a Job-Functional Literacy Training Program for the United States Army. Literacy Discussion, 4 (3), pp. 339-369 (1973).
31. Begtin I.V.: What is “understandable Russian language” in terms of technology. Let’s look at the metrics of text readability. http://habrahabr.ru/company/infoculture/blog/238875/, (2014), last accessed 2024/02/16.
32. Solovyev V., Ivanov V. and Solnyshkina M.: Assessment of reading difficulty levels in Russian academic texts. Approaches and metrics, Journal of Intelligent & Fuzzy Systems, 34(5), pp. 3049-3058 (2018).