Computational Linguistics and Intellectual Technologies:
Proceedings of the International Conference "Dialogue 2023"

June 14–16, 2023

Estimating cognitive text complexity with aggregation of quantile-based models

Arseniy Veselov
Lomonosov Moscow State University
Moscow, Russia
arseniy.veselov@yandex.ru

Maksim Eremeev
New York University
New York, USA
eremeev@nyu.edu

Konstantin Vorontsov
Moscow Institute of Physics and Technology
Moscow, Russia
vokov@forecsys.ru

Abstract

In this paper, we introduce a novel approach to estimating the cognitive complexity of a text at different levels of language: phonetic, morphemic, lexical, and syntactic. The proposed method detects tokens with an abnormal frequency of complexity scores. The frequencies are taken from the empirical distributions calculated over the reference corpus of texts. We use the Russian Wikipedia for this purpose. Ensemble models are combined from individual models from different language levels. We created datasets of pairs of text fragments taken from social studies textbooks of different grades to train the ensembles. Empirical evidence shows that the proposed approach outperforms existing methods, such as readability indices, in estimating text complexity in terms of accuracy. The purpose of this study is to create one of the important components of the system of recommendation of scientific and educational content.

Keywords: cognitive complexity of texts, language levels, ensemble learning
DOI: 10.28995/2075-7182-2023-22-525-538

Оценивание когнитивной сложности текста с помощью агрегирования моделей, основанных на квантилях

Веселов А.С.
Московский государственный
университет им. М. В. Ломоносова
Москва, Россия
arseniy.veselov@yandex.ru

Еремеев М.А.
Нью-Йоркский
университет
Нью-Йорк, США
eremeev@nyu.edu

Воронцов К.В.
Московский физико-технический
институт
Москва, Россия
vokov@forecsys.ru

Аннотация

В данной работе описывается подход к оцениванию когнитивной сложности текста на разных уровнях языка: на фонетическом, морфемном, лексическом и синтаксическом. В его основе лежит определение токенов с аномальной частотой их сложностей. Частоты определяются по эмпирическим распределениям, построенным на основе референтного корпуса текстов, в качестве которого используется русскоязычная Википедия. Из отдельных моделей с разных уровней языка создаются агрегированные модели. Для их обучения мы создали выборки пар фрагментов текстов, взятых из учебников по обществознанию разных учебных классов. Проведённые в работе эксперименты показывают у предлагаемого подхода более высокую точность ранжирования текстов по сложности в сравнении с индексами удобочитаемости. Целью проведения данного исследования является создание одного из важных компонентов системы рекомендации научно-образовательного контента.

Ключевые слова: когнитивная сложность текстов, уровни языка, ансамблевое обучение

1 Introduction

Many readability indices have been developed for the task of estimating the complexity of the text. Most of them are a linear combination of some trivial statistical parameters of the text based on the number of letters, syllables, words, and sentences. In this paper, we continue the research and improvement of the generalised quantile-based approach to the estimation of the cognitive complexity of the text at different levels of the language (phonetic, morphemic, lexical, and syntactic). The idea of such an approach was first presented by Eremeev M.A. and Vorontsov K.V. in (Eremeev and Vorontsov, 2019). It is based on the detection of tokens with an abnormal frequency of their complexity scores. We use the reference corpus of texts, which is the Russian-language Wikipedia, to construct the empirical distributions for this purpose. This paper is devoted to the study of the aggregation of individual quantile-based models in order to take information from different levels of the language into account, and this is its novelty. We train aggregated models on datasets of pairs of text fragments, which we created on the basis of social studies textbooks of different educational grades. In this paper, we conduct experiments to compare the accuracy of our models with adapted readability indices, including the comparison of accuracy over each pair of educational grades. The analysis of the contribution of individual components to the aggregated model (ablation study) and the analysis of the dependence of the ranking accuracy on the average length of a text fragment in a dataset are also carried out. The experiments conducted in the paper demonstrate that the proposed approach has a higher accuracy of ranking texts in terms of cognitive text complexity compared to readability indices. The purpose of this study is to create one of the important components of the system of recommendation of scientific and educational content.

2 Readability indices review

Historically linguists use readability indices for estimating text complexity of the educational literature. Many of them were initially developed for the US education system and were therefore adapted for the English language.

The automated readability index (ARI) was developed by R.J. Senter and E.A. Smith in 1967 (Senter and Smith, 1967). It approximates a representation of the US grade level required to understand the analysed text. For a document d written in English ARI has the following calculation formula:

$$\mathrm{ARI}(d) = 4.71 \times \frac{C}{W} + 0.5 \times \frac{W}{S} - 21.43,$$

where C is the number of letters and digits, W is the number of words, and S is the number of sentences in the text of the document d.

Läsbarhetsindex (LIX) was developed by Swedish scientist Carl-Hugo Björnsson in 1968 (Björnsson, 1968). Index value monotonically increases with respect to text complexity. LIX does not take into account the language in which the text is written and is calculated as follows:

$$\mathrm{LIX}(d) = \frac{A}{B} + 100 \times \frac{C}{A},$$

where A is the number of letters, B is the number of sentences, and C is the number of words longer than 6 letters in the text of the document d.

In 1969 G. Harry McLaughlin developed the Simple Measure of Gobbledygook (SMOG) (McLaughlin, 1969). This readability index produces an approximate number of years of study needed to comprehend the text. SMOG is calculated for the document d written in English with the following formula:

$$\mathrm{SMOG}(d) = 1.0430\sqrt{A \times \frac{30}{B}} + 3.1291,$$

where A denotes the number of polysyllabic words (3 and more syllables in English), and B is the number of sentences.

Coleman–Liau index (CLI), developed in 1975 by Meri Coleman and T.L. Liau (Coleman and Liau, 1975), approximates a representation of the US grade level necessary to understand the given text. For the document d written in English CLI has the following calculation formula:

$$\mathrm{CLI}(d) = 0.0588 \times L - 0.296 \times S - 15.8,$$

where L denotes the average number of letters per 100 words, and S refers to the average number of sentences per 100 words.

In 1948 Rudolf Flesch developed the most popular measure of text complexity — the Flesch reading-ease score (FRES) (Flesch, 1948). The index value monotonically declines with respect to text complexity. FRES is calculated for the document d written in English as follows:

$$\mathrm{FRES}(d) = 206.835 - 1.015 \times \mathrm{ASL} - 84.6 \times \mathrm{ASW},$$

where ASL is the average sentence length in words, and ASW is the average number of syllables per word.

Flesch–Kincaid grade level (FKGL) was developed by J. Peter Kincaid in 1975 (Kincaid et al., 1975). This readability index approximates a representation of the US grade level. FKGL has the following formula for calculation for the document d written in English:

$$\mathrm{FKGL}(d) = 0.39 \times \mathrm{ASL} + 11.8 \times \mathrm{ASW} - 15.59.$$

The Estonian linguist Juhan Tuldava proposed in 1975 his own readability index (Tuldava, 1975), which we refer to in our article as the Tuldava index (TI). TI does not take the language of the text into account and is calculated as follows:

$$\mathrm{TI}(d) = \mathrm{ASW} \times lg(\mathrm{ASL}).$$

In this paper, we estimate the complexity of Russian texts. Therefore, we use adapted versions of indices for comparison with the proposed quantile-based approach.

Irina Oborneva made a significant contribution to the development of the readability formulae for texts in Russian by adapting the FRES and FKGL indices in 2005 (Oborneva, 2005):

$$\mathrm{FRES}_{\mathrm{ru}}(d) = 206.835 - 1.3 \times \mathrm{ASL} - 60.1 \times \mathrm{ASW},$$

$$\mathrm{FKGL}_{\mathrm{ru}}(d) = 0.5 \times \mathrm{ASL} + 8.4 \times \mathrm{ASW} - 15.59.$$

Later, the results of the adaptation of the readability formulae for automated analysis of texts in Russian were presented by Ivan Begtin in 2014 (Begtin, 2014). These implementations were collected in the Python library ruTS by Sergey Shkarin in 2021 (Shkarin, 2021). We utilise this library to reproduce baseline results for this paper. We have extended it by adding the Tuldava index and correcting the wrong coefficients in the Coleman–Liau index. See the formulae for ARI_{ru}, $SMOG_{ru}$, and CLI_{ru} readability indices adapted for the Russian language below (the variables that are not explained below are the same as for the formulae for English):

$$\mathrm{ARI}_{\mathrm{ru}}(d) = 6.26 \times \frac{C}{W} + 0.2805 \times \frac{W}{S} - 31.04.$$

$$\mathrm{SMOG}_{\mathrm{ru}}(d) = 1.1\sqrt{A \times \frac{64.6}{B}} + 0.05,$$

where A denotes the number of polysyllabic words (4 and more syllables in Russian).

$$\mathrm{CLI}_{\mathrm{ru}}(d) = 0.055 \times L - 0.35 \times S - 20.33.$$

Text complexity estimates have many applications. For example, Arina Dmitrieva describes the methods of analysing legal documents in Russian based on readability indices (Dmitrieva, 2017). The FKGL readability index was developed in order to compile the texts of instructions for the use of weapons or technical means, and the SMOG index was used to study the text complexity of instructions for medicines and preparations. Many indices are used to estimate the comprehensibility of textbooks offered to students of different ages. The use of text complexity estimation can be helpful for predicting the time spent processing regulations, documents, and educational literature.

3 Generalised text complexity model

Let d be an arbitrary document of length n consisting of tokens $x_1, \dots, x_n$ from a fixed finite alphabet A_h, where h denotes the level of the language: phonetic, morphemic, lexical or syntactic. In this paper, we consider letters, syllables, words, or sentences (or structures describing a part of speech and the syntactic function of words) as tokens, depending on the level of the language. Suppose that every token x_i of the document d has its own processing complexity c_i caused by its context or by its internal structure. Also assume that each token $a \in A_h$ has its usual processing complexity, which is a result of the language evolution within a historical and cultural environment. If the current processing complexity of the token $x_i = a$ in the analysed text turns out to be abnormally high compared to the usual processing complexity of token a, then we will assume that the token x_i carries an excessive difficulty of perception. The information about usual complexity of tokens can be retrieved from a *reference collection* denoted by K, which is a large union of texts of medium complexity. In order to determine if the token $x_i \in d,\ x_i = a$ is abnormally complex we need to construct an empirical distribution of complexity scores $\hat{c}_j$ of every token $\hat{x}_j \in K$ such that $\hat{x}_j = a$. The token x_i is considered as abnormally complex if its complexity score is greater than the γ-quantile $C_\gamma(x_i)$ of the constructed distribution for token (see Figure 1).

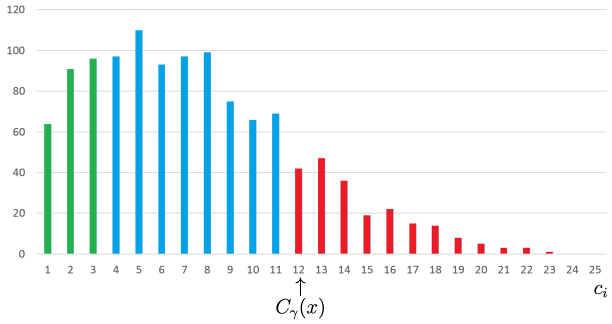

Figure 1: Histogram for empirical distribution of complexity scores and its γ-quantile

In Figure 1 the red zone corresponds to an abnormally high complexity. The green zone corresponds to a low complexity. The blue zone indicates the usual complexity of the token.

We shall call the nonlinear sum of weights w_i of tokens with abnormal complexities the *document complexity score* and denote it by $S(d)$.

$$S(d) = \sum_{i=1}^{n} w_i^p \left[c_i > C_\gamma\left(x_i\right)\right], \tag{1}$$

where [ ] is the Iverson bracket (i.e. [true] = 1, [false] = 0), p is a positive integer.

The weight w_i is a non-negative value that does not decrease with increasing complexity c_i. Complexity c_i is defined up to an arbitrary increasing function.

Table 1 shows several examples of possible weights.

w_i	Meaning of w_i
1	number of complex tokens
$1/n \times 100\%$	percentage of complex tokens
c_i	total complexity
c_i/n	mean complexity
$c_i - C_\gamma\left(x_i\right)$	excessive complexity
$\left(c_i - C_\gamma\left(x_i\right)\right)/n$	mean excessive complexity

Table 1: Examples of weights w_i

4 Token complexity functions

4.1 Distance-based complexity function

Let r_i be a distance from the previous occurrence of the token x_i to its current occurrence in the text:

$$\ldots \boxed{x_{i-r_i} = a} \quad \underbrace{x_{i-r_i+1} \quad x_{i-r_i+2} \quad \ldots \quad x_{i-2} \quad x_{i-1} \quad \boxed{x_i = a}}_{r_i} \ldots$$

$$r_i = \min_{1 \leqslant j < i} \{i - j \mid x_i = x_j\}.$$

In the first occurrence of the token a in the text, at the position i, the distance r_i is undefined. In that case r_i is redefined such that the sum of all distances r_j for this token $x_j = a$ equals to the document length n.

To obtain a frequency model of complexity as a special case of the generalised model, the parameters c_i are defined as some decreasing function of r_i, for example:

$$c_i = -r_i \tag{2}$$

4.2 Counter-based complexity function

In the counter-based approach, as in the special case of the generalised approach, it is assumed that the alphabet A_h consists of a single token $A_h = \{a\}$, i.e. we distinguish not the tokens themselves, but only their complexity. The complexity of tokens is determined by their linguistic properties, and each token has exactly one possible complexity value. Thereby, just one empirical distribution of token complexities is constructed over all tokens from the reference collection. In that case, $C_\gamma(x_i) = C_\gamma$.

5 Considered models

In this section, we describe individual models at different levels of language in terms introduced when considering a generalised model above, i.e. by specifying the alphabets of tokens and complexity functions. The available means of morphological, lexical, and syntactic analysis can be used to form alphabets of tokens and characteristics of their complexity.

5.1 Phonetic level

We consider individual letters as tokens here. For this type of the models we use the distance-based approach. The name of the model implemented in that way is $letter\_dist\_model$.

5.2 Morphemic level

There are two possible ways to form tokens: either take the original syllables, or rearrange the letters in them in alphabetical order so that the order of the letters is not taken into account. Therefore, we consider two distance-based models, which we refer to as $syllab\_dist\_model$ and $syllabsort\_dist\_model$.

5.3 Lexical level

The tokens here are individual words. For models at this level (except the $lexical\_len\_model$) we consider different forms of one word to be equal and use the lemmatization of words as a preprocessing.

Distance-based model at this level is called $lexical\_dist\_model$.

Word length counter-based model considers the length of the word as its complexity score. To implement such a model, we construct an empirical distribution of lengths of all words in the reference collection. We refer to this model as $lexical\_len\_model$.

Counter-based model at lexical level is based on the assumption that the rarer a word is encountered in the reference corpus, the more specific and difficult it is. In the experiments, the following complexity function is used:

$$c_i = -\operatorname{count}(x_i), \tag{3}$$

where $\operatorname{count}(x_i)$ is the number of token x_i occurences in the reference collection. We refer to this model as $lexical\_cnt\_model$.

5.4 Syntactic level

The tokens here are sentences or structures describing the part of speech and the syntactic function of words in the sentence. In this paper, we use the UDPipe library to divide the text into sentences and extract the syntactic dependencies and parts of speech (Straka and Straková, 2017).

Counter-based model at this level uses the maximum length of the syntactic dependency in the sentence as a complexity score of this sentence. We refer to this model as $syntax\_len\_model$.

Distance-based model considers a sentence as a set structures describing a part of speech and the syntactic function of words in the sentence. Each such structure corresponds to one word. The word itself is ignored, but information about its part of speech and syntactic role in the sentence is considered. We refer to this model as $syntaxpos\_dist\_model$.

6 Experiments

6.1 Reference collection and datasets

We use the Russian Wikipedia (1.5 million articles) as a reference collection for our experiments. The ruwiki-latest-pages-articles.xml.bz2 archive was processed by the WikiExtractor parser. After the additional preprocessing it was translated into a format where each article corresponds to its own TXT document.

As a dataset we use the sets of social studies textbooks, prepared in (Solovyev et al., 2018): textbooks by L.N. Bogolyubov for 6, 7, 8, 9, 10, 10+, 11+ grades («+» denotes a version with in-depth study) and textbooks by A.F. Nikitin for 5, 6, 7, 8, 9, 10, 11 grades. In this dataset, each document contains randomly shuffled sentences from the textbook. In order to create a dataset for the training and validation of models, we first combined the texts of the textbooks intended for the same grade and then cut them into pieces of similar length consisting of whole sentences. Afterwards, the fragments of texts of different grades were combined into pairs, where a piece of text from a textbook of a higher grade comes second: $D = \{(d,\ d') \mid d'$ more complex than $d)\}$. The complexity of the textbook is determined by its grade, which should be a fairly reliable characteristic to estimate the cognitive complexity of the text, since textbooks are created in accordance with educational standards.

Eight datasets with different number of pairs were prepared. They are available at this link. For this purpose, the length of one text fragment varied (see Table 2). Each dataset consists of all possible pairs in such a way that each text piece of one grade is compared with each text piece of each other grade.

Dataset name	Number of pairs of text fragments	Average number of symbols in one text fragment
D1	1027	94 100
D2	2532	59 850
D3	5001	42 650
D4	10 041	30 100
D5	45 058	14 200
D6	250 152	6000
D7	1 008 881	2950
D8	5 400 136	1250

Table 2: Datasets

We create such a number of datasets in order to investigate the dependence of ranking accuracy on the average length of a text fragment in the last experiment. In other experiments, only datasets D1, D2, D3, D4 are used, because their average lengths of a text fragment are large enough to provide as much information as possible to models and readability indices to estimate the complexity of text fragments. Moreover, we will focus more on D4 in further experiments since there are quite a lot of pairs of text fragments in this dataset, so that we can get more different possible values of the quality criterion as well as train aggregated models on a larger number of pairs.

6.2 Quality criterion

As a quality criterion we consider accuracy, i.e. the ratio of the number of correctly estimated pairs of text pieces to the total number of pairs:

$$\text{accuracy(S)} = \frac{\sum_{(d,d')\in D}\left[S(d') > S\left(d\right)\right]}{|D|}, \tag{4}$$

where S denotes a model (or readability index), which produces document complexity score.

6.3 Separate models

In experiments (see Table 3), the best parameters $(p,\ w_i,\ \gamma)$ (see the formula 1) of the models that maximized the quality criterion are selected on D3 and D4 datasets (since they have more pairs, and this means that it is potentially possible to get more different values of accuracy) or on similar-sized datasets based on a series of textbooks by only one of the authors.

The weights w_i are searched over the grid $\{1,\ c_i,\ c_i/n,\ c_i - C_\gamma\left(x_i\right),\ \left(c_i - C_\gamma\left(x_i\right)\right)/n\}$. The parameter γ is searched over the following grid with a step 0.05: $\{0.01\} \cup [0.05,\ 0.1,\ 0.15,\ \ldots,\ 0.9,\ 0.95] \cup \{0.99\}$. The parameter p is searched over the grid $[1,\ 2,\ 3,\ 4]$. In addition to the distance-based models with the complexity function (2), the experiments also estimated the quality of the models, which are based on the opposite hypothesis that the rarer the same tokens are found in the analysed text, the more difficult they are to comprehend, with a complexity function $C_i = r_i$. But the quality of such models was in the range of 30-60%, so they are not presented further.

№	**Model name**	**Hyperparameters**			**Accuracy on dataset, %**			
		$\boldsymbol{w_i}$	$\boldsymbol{p}$	$\boldsymbol{\gamma}$	**D1**	**D2**	**D3**	**D4**
1	***letter_dist_0***	c_i/n	1	0.10	79.45	77.49	77.70	76.36
2	***letter_dist_1***	c_i/n	1	0.85	81.60	77.13	76.42	75.74
3	*letter_dist_2*	c_i	1	0.05	80.92	72.08	80.66	67.29
4	*syllab_dist_0*	c_i/n	1	0.01	63.49	76.46	78.08	78.23
5	*syllab_dist_1*	c_i	1	0.65	73.61	63.19	72.27	59.57
6	*syllabsort_dist_0*	c_i	1	0.05	79.07	67.65	82.04	76.25
7	*lexical_dist_0*	$\left(c_i - C_\gamma\left(x_i\right)\right)/n$	1	0.01	75.17	76.11	84.88	82.01
8	*lexical_dist_1*	c_i	1	0.99	82.38	76.94	85.64	76.17
9	***lexical_len_0***	$\left(c_i - C_\gamma\left(x_i\right)\right)/n$	1	0.55	92.02	90.72	89.58	89.57
10	***lexical_len_1***	c_i/n	1	0.85	88.41	87.52	87.00	86.86
11	***lexical_len_2***	c_i/n	1	0.45	92.11	90.84	91.10	91.08
12	***lexical_len_3***	$\left(c_i - C_\gamma\left(x_i\right)\right)/n$	1	0.30	**93.48**	**92.06**	91.70	**91.28**
13	***lexical_len_4***	c_i/n	2	0.65	90.36	89.38	**92.76**	87.72
14	*lexical_cnt_0*	c_i/n	2	0.35	70.79	61.77	72.02	63.10
15	*lexical_cnt_1*	c_i/n	2	0.15	84.32	83.10	87.16	80.78
16	*lexical_cnt_2*	c_i	1	0.85	63.68	57.70	63.25	57.50
17	*lexical_cnt_3*	c_i	1	0.45	73.81	67.69	71.25	60.92
18	***syntax_len_0***	c_i/n	2	0.01	88.61	83.77	86.14	83.95
19	***syntax_len_1***	c_i/n	2	0.35	88.51	83.81	85.80	83.89
20	*syntaxpos_dist_0*	c_i	1	0.45	81.60	81.67	85.58	78.99
21	*syntaxpos_dist_1*	c_i	1	0.35	83.93	82.39	86.30	80.84

Table 3: Selected parameters and accuracy of individual models on D3 and D4. Bold lines separate different types of models. Models highlighted in bold show the greatest contribution to the ensemble in ablation studies

As a result, 21 models were selected to be used in aggregation experiments. The word length counter-based lexical models show the best quality amongst the individual models, surpassing all the readability indices, whose accuracy on the same datasets is shown in the table 4. $\mathrm{FKGL_{ru}}$ and $\mathrm{FRES_{ru}}$ demonstrate the best accuracy amongst the readability indices. If we focus only on the D4 dataset, then the best index is $\mathrm{FRES_{ru}}$.

Index	**Accuracy on dataset, %**			
	D1	**D2**	**D3**	**D4**
$\mathrm{FKGL_{ru}}$	**91.04**	**90.00**	89.94	89.49
$\mathrm{FRES_{ru}}$	90.75	**90.00**	**90.30**	**90.50**
$\mathrm{CLI_{ru}}$	89.97	89.26	89.76	89.09
$\mathrm{SMOG_{ru}}$	90.26	88.63	88.24	87.80
$\mathrm{ARI_{ru}}$	90.36	89.69	90.14	89.64
LIX	90.65	89.22	89.44	88.79
TI	90.94	89.97	89.92	89.55

Table 4: Accuracy of readability indices on D1, D2, D3, and D4

6.4 Ensemble models

From the selected separate models (Table 3) the ensemble models are constructed. Due to the small size of the datasets, linear regression with non-negative weights is used for the ensembling.

$$S(d, \boldsymbol{\alpha}) = \sum_{k=1}^{K} \alpha_k S_k(d), \ \alpha_k \geqslant 0, \tag{5}$$

where vector $\boldsymbol{\alpha}$ is a solution of the following optimization problem:

$$\sum_{(d,\ d')\in D} \mathcal{L}(S(d', \boldsymbol{\alpha}) - S(d, \boldsymbol{\alpha})) + \lambda \operatorname{Reg}(\boldsymbol{\alpha}) \to \min_{\boldsymbol{\alpha}}, \tag{6}$$

where $\mathcal{L}(M)$ is a non-increasing function of margin M, and Reg is a regularizer. The ensemble models are trained on 80% of the dataset and validated on the remaining 20%.

We compare the ensemble models with and without regularization in experiments. For this purpose, L1, L2, or elastic net regularization with a mixing hyperparameter equal to 0.5 are used. The hyperparameter λ is optimized over the grid $[10^{-4}, 10^{-3}, 10^{-2}, 0.1, 1]$. The following loss functions $\mathcal{L}$ of margin M are used:

$\mathcal{L}_1(M) = (1 - M).clip(min = 0), \quad \mathcal{L}_2(M) = |1 - M|, \quad \mathcal{L}_3(M) = (1 - M^2),$

$\mathcal{L}_4(M) = \log(1 + e^{-M}), \quad \mathcal{L}_5(M) = \frac{1}{1+e^M}, \quad \mathcal{L}_6(M) = e^{-M}.$

Tables 5, 6 show for each loss function the ensembles of 21 separate models from Table 3) of the best validation accuracy on the datasets D4, D3, respectively.

The loss function $\mathcal{L}_6(M)$ had an overflow problem, so its results are not shown in Tables 5, 6. The following functions proved to be bad for our problem, thus their results are not presented in this paper: $\mathcal{L}_7(M) = -|M|,\ \mathcal{L}_8(M) = -M^2,\ \mathcal{L}_9(M) = 1 - M,\ \mathcal{L}_{10}(M) = (-M)^3.$

№	**Loss function**	**Reg**	**λ**	**Acc. on D4 [val.], %**
1	$\mathcal{L}_1$	L2	10^{-4}	**92.78**
2	$\mathcal{L}_2$	L1	10^{-2}	91.24
3	$\mathcal{L}_3$	L1	10^{-3}	92.14
4	$\mathcal{L}_4$	L1	10^{-3}	88.00
5	$\mathcal{L}_5$	L1	1	82.73

Table 5: Validation accuracy of ensembles of 21 separate models for each loss function on D4

№	Loss function	Reg	λ	Acc. on D3 [val.], %
1	$\mathcal{L}_1$	L2	10^{-3}	93.61
2	$\mathcal{L}_2$	L1	10^{-2}	93.31
3	$\mathcal{L}_3$	L2	10^{-3}	**94.51**
4	$\mathcal{L}_4$	L1	0.1	91.11
5	$\mathcal{L}_5$	L1	1	90.81

Table 6: Validation accuracy of ensembles of 21 separate models for each loss function on D3

The experiments have shown the loss functions $\mathcal{L}_1(M)$ and $\mathcal{L}_2(M)$ to consistently be of the highest quality, i.e. they are less sensitive to the selection of hyperparameters.

The accuracy of the readability indices on the same validation parts of the datasets D3, D4 is presented in Table 7.

Index	Acc. on D3, %	Acc. on D4, %
$\mathrm{FKGL}_{\mathrm{ru}}$	89.71	88.40
$\mathrm{FRES}_{\mathrm{ru}}$	**89.81**	**89.90**
$\mathrm{CLI}_{\mathrm{ru}}$	89.11	87.90
$\mathrm{SMOG}_{\mathrm{ru}}$	87.31	86.71
$\mathrm{ARI}_{\mathrm{ru}}$	89.31	88.75
LIX	89.01	87.76
TI	**89.81**	88.60

Table 7: Accuracy of readability indices on validation part of D3 and D4

6.5 Accuracy over grade pairs

The accuracy of the best ensemble of 21 separate models (first in Table 5) is examined in more detail in the following section. Table 8 shows the values of the quality criterion (4) on every pair of grades separately.

Acc.	6	7	8	9	10	10+	11	11+
5	1	1	1	1	1	1	1	1
6	—	0.95	1	1	1	1	1	1
7	—	—	0.975	1	1	1	1	1
8	—	—	—	0.955	0.97	1	1	1
9	—	—	—	—	0.636	0.953	0.935	1
10	—	—	—	—	—	0.705	0.736	0.98
10+	—	—	—	—	—	—	0.591	0.984
11	—	—	—	—	—	—	—	0.98

Table 8: Validation accuracies of ensemble of 21 separate models on D4 over grade pairs

Table 8 demonstrates that the ensemble model accurately ranks by complexity the text pieces from grades that are more than one—two years apart. It is also noticeable that the lower the grades of both text pieces in a pair, the easier it is for the model to arrange them correctly. That looks logical, since the increase in the complexity of texts of middle school textbooks should be more dramatic than that of high school textbooks.

Table 9 shows the results for the $FRES_{ru}$ readability index, which demonstrated the best accuracy among other indices.

Acc.	**6**	**7**	**8**	**9**	**10**	**10+**	**11**	**11+**
5	1	1	1	1	1	1	1	1
6	—	0.8	1	1	1	1	1	1
7	—	—	0.975	1	1	1	1	1
8	—	—	—	0.736	0.993	1	1	1
9	—	—	—	—	0.882	0.915	0.871	0.991
10	—	—	—	—	—	0.524	0.491	0.967
10+	—	—	—	—	—	—	0.341	0.992
11	—	—	—	—	—	—	—	1

Table 9: Accuracy of $FRES_{ru}$ on validation part of D4 over grade pairs

6.6 Ablation study

In this experiment, we reduce the number of individual models in the ensemble model so as not to degrade, but even to improve the quality.

For this purpose, we examine the vector $\boldsymbol{\alpha}$, computed as a result of training ensemble of 21 separate models (first in Table 5). We sort its components in descending order: the weights corresponding to the individual models that make the greatest contribution to estimating the text complexity are the first (see Figure 2).

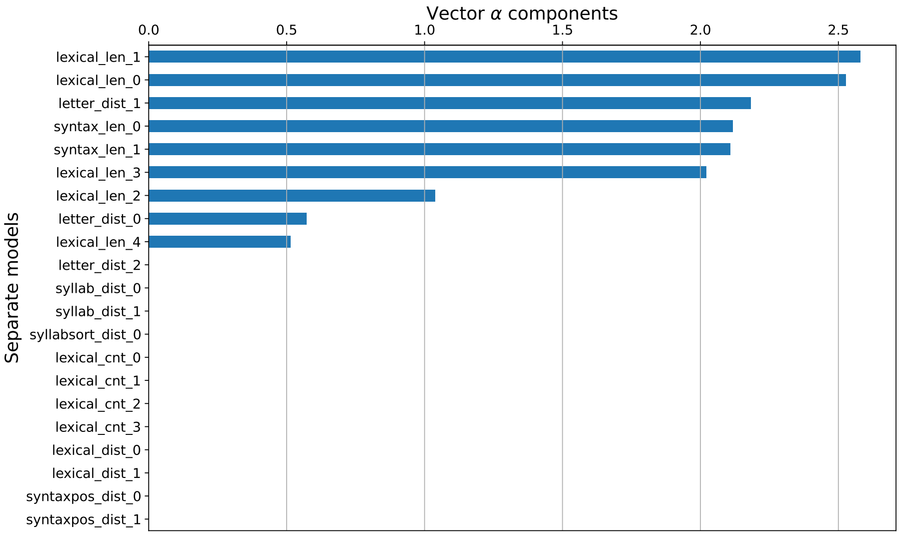

Figure 2: Importance of separate models

Further, a comparison of the validation accuracy on the dataset D4 of different ensembles with one removed block of separate models of one type has shown that deleting the block with word length counter-based lexical models or counter-based syntactic models leads to a significant loss of quality in all ensembles with the loss function $\mathcal{L}_1$ and regularization. We examine ensembles with the loss function $\mathcal{L}_1$ and regularization because this combination proved to be the best. Deleting the distance-based phonetic models block leads to a drop in accuracy on most of these ensembles. Deleting the distance-based syntactic models, counter-based lexical models, distance-based lexical models or distance-based morphemic models block almost does not lead to significant quality losses, and in some cases even increases it. Figure 3 shows how the accuracy of the best ensemble changes when one of the blocks is removed.

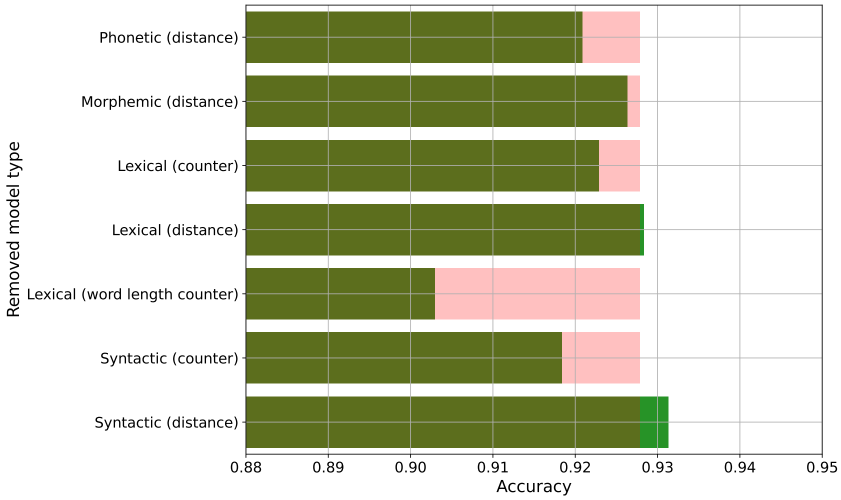

Figure 3: The change in validation accuracy on D4 when removing the block of models of one type: pink shows a decline in accuracy with respect to an ensemble of 21 models; bright green, respectively, shows an improvement

As a next step, the comparison of ensembles of different sets of blocks without separate models of the least importance (Figure 2) is carried out. As a result, the following ensemble model of nine separate models proved to be the best:

- Distance-based phonetic models: $letter\_dist\_0$, $letter\_dist\_1$;
- Word length counter-based lexical models: $lexical\_len\_0$, $lexical\_len\_1$, $lexical\_len\_2$, $lexical\_len\_3$, $lexical\_len\_4$;
- Counter-based syntactic models: $syntax\_len\_0$, $syntax\_len\_1$.

Tables 10, 11 show for each loss function ensembles of nine separate models of the best validation accuracy on the datasets D4, D3, respectively.

№	**Loss function**	**Reg**	λ	**Acc. on D4 [val.], %**
1	$\mathcal{L}_1$	elastic net	10^{-4}	**93.48**
2	$\mathcal{L}_2$	L2	10^{-2}	92.33
3	$\mathcal{L}_3$	L2	0.1	92.33
4	$\mathcal{L}_4$	—	0	93.23
5	$\mathcal{L}_5$	—	0	93.33
6	$\mathcal{L}_6$	L1	10^{-3}	93.23

Table 10: Validation accuracy of ensembles of 9 separate models for each loss function on D4

№	**Loss function**	**Reg**	λ	**Acc. on D3 [val.], %**
1	$\mathcal{L}_1$	—	0	94.91
2	$\mathcal{L}_2$	elastic net	10^{-2}	94.51
3	$\mathcal{L}_3$	—	0	**95.60**
4	$\mathcal{L}_4$	—	0	93.51
5	$\mathcal{L}_5$	L1	10^{-4}	94.61
6	$\mathcal{L}_6$	L2	10^{-4}	94.91

Table 11: Validation accuracy of ensembles of 9 separate models for each loss function on D3

The experiments have shown that of all the loss functions $\mathcal{L}_1(M)$, $\mathcal{L}_2(M)$, $\mathcal{L}_3(M)$ and $\mathcal{L}_6(M)$ consistently demonstrate a high accuracy with different values of hyperparameters. As for $\mathcal{L}_4(M)$, $\mathcal{L}_5(M)$, it is better not to use regularization at all, since with it the quality drops quickly. It is also noticeable that with a good set of separate models for ensembling, one can get an acceptable quality with almost any loss function.

Thus, the experiments show that using the loss function $\mathcal{L}_1(M)$ and any weak regularization with hyperparameter $\lambda = 10^{-4} \dots 10^{-3}$ (or without regularization at all) is the best option.

6.7 Dependence of accuracy on the text fragment average length

In this experiment, the analysis of the dependence of the ranking accuracy on the average length of a text fragment in a dataset is carried out. For this purpose, all built datasets based on textbooks are used (see Table 2).

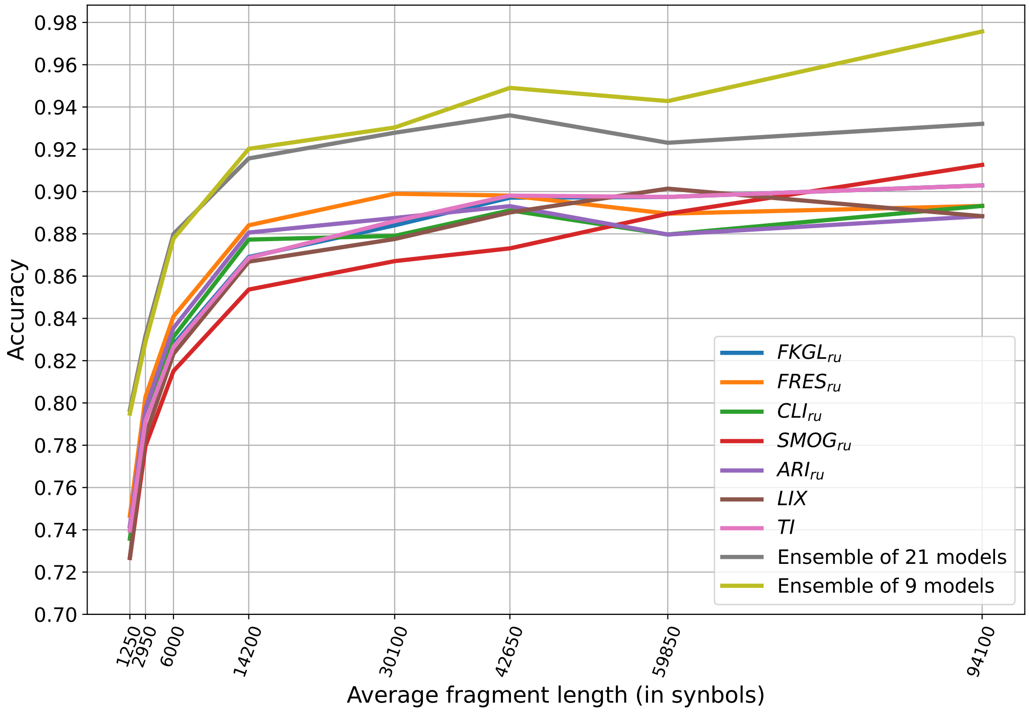

Figure 4:

Figure 4 demonstrates that the accuracy begins to decrease as the length of the text fragment decreases, both for models and for readability indices. A particularly sharp drop is noticeable, starting with a length of 14200 characters or less. While with lengths of more than 14200 symbols, many indices and models have a plateau in ranking accuracy. It is also clear from the figure that the aggregated models demonstrate a higher quality than the readability indices for all the lengths of text fragments, and the ensemble of 9 models shows higher accuracy than the ensemble of 21 models. For this experiment, an ensemble of 21 models with a loss function $\mathcal{L}_1(M)$ and with L2 regularization with hyperparameter $\lambda = 10^{-4}$, and an ensemble of 9 models with loss function $\mathcal{L}_1(M)$ without regularization were selected.

7 Conclusion

In this paper, a method of estimating the cognitive complexity of a text based on quantile-based models is investigated. In particular, models are implemented at the phonetic, morphemic, lexical, and syntactic levels of the language, as well as their ensembling. For the individual models the empirical distributions of tokens over the reference collection of Russian Wikipedia articles are calculated. Ensemble models are trained on the datasets formed from social studies textbooks for different grades. All the models considered are compared in accuracy with the readability indices adapted for the Russian language. Among the individual models, the word length counter-based lexical models have shown the best accuracy, surpassing all the readability indices. The ensemble of 21 best separate models of all types has even more significantly surpassed all the readability indices in terms of the accuracy of ranking pairs of text fragments. The results of analysis of its accuracy for each pair of grades separately are consistent with our ideas about the complexity of school textbooks. It is observed that the ensemble model accurately ranks the text pieces by complexity from grades that are more than one to two years apart. It is also noticeable that the lower the grades of both text pieces in a pair, the easier it is for the model to arrange them correctly. The selection of the best ensemble (ablation study) is carried out, as a result of which the ensemble of nine separate models shows further significant improvement in quality. It consists of models of the following types: distance-based phonetic model, word length counter-based lexical model, and counter-based syntactic model. The paper also analyzes the dependence of the ranking accuracy on the average length of a text fragment in a dataset. As a result, it turned out that the accuracy decreases as the average length of the fragment decreases. A particularly sharp drop begins when the number of symbols is less than 14200.

References

Ivan Viktorovich Begtin. 2014. Plain russian language. `https://github.com/infoculture/plainrussian`.

Carl-Hugo Björnsson. 1968. *Läsbarhet: Lesbarkeit durch Lix*. Liber, Stockholm, Sweden.

Meri Coleman and Ta Lin Liau. 1975. A computer readability formula designed for machine scoring. *Journal of Applied Psychology*, 60(2):283–284.

Arina Viktorovna Dmitrieva. 2017. «iskusstvo yuridicheskogo pis'ma»: kolichestvenniy analiz resheniy konstitucionnogo suda rossiyskoy federacii. *Sravnitel'noe konstitucionnoe obozrenie*, 3(118):125–133. Online available at: `https://academia.ilpp.ru/wp-content/uploads/2021/10/SKO-3-118-2017-125-133-Dmitrieva.pdf`.

Maksim A. Eremeev and Konstantin V. Vorontsov. 2019. Lexical quantile-based text complexity measure. // *Proceedings of the International Conference on Recent Advances in Natural Language Processing*, P 270–275. Online available at: `https://aclanthology.org/R19-1031.pdf`.

Rudolf Flesch. 1948. A new readability yardstick. *Journal of Applied Psychology*, 32:221–233.

J. Peter Kincaid, Robert P. Fishburne Jr, Richard L. Rogers, and Brad S. Chissom. 1975. Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel. *Naval Technical Training Command Millington TN Research Branch*, P 40. Online available at: `https://web.archive.org/web/20220409163459/https://apps.dtic.mil/sti/pdfs/ADA006655.pdf`.

G. Harry McLaughlin. 1969. Smog grading — a new readability formula. *Journal of reading*, 12(8):639–646. Online available at: `https://web.archive.org/web/20220119124738/https://ogg.osu.edu/media/documents/health_lit/WRRSMOG_Readability_Formula_G._Harry_McLaughlin__1969_.pdf`.

Irina Vladimirovna Oborneva. 2005. Avtomatizaciya ocenki kachestva vospriyatiya teksta. *Vestnik Moskovskogo gorodskogo pedagogicheskogo universiteta. Seriya: Informatika i informatizaciya obrazovaniya*, 5:86–91.

R. J. Senter and Edgar A. Smith. 1967. Automated readability index. *AMRL-TR*, 66(220). Online available at: `https://web.archive.org/web/20160305161235/http://www.dtic.mil/get-tr-doc/pdf?AD=AD0667273`.

Sergey Shkarin. 2021. ruts, a library for statistics extraction from texts in russian. `https://github.com/SergeyShk/ruTS`.

Valery Solovyev, Vladimir Ivanov, and Marina Solnyshkina. 2018. Assessment of reading difficulty levels in russian academic texts: Approaches and metrics. *Journal of intelligent & fuzzy systems*, 34(5):3049–3058. Online available at: `https://www.researchgate.net/publication/324583915_Assessment_of_reading_difficulty_levels_in_Russian_academic_texts_Approaches_and_metrics`.

Milan Straka and Jana Straková. 2017. Tokenizing, pos tagging, lemmatizing and parsing ud 2.0 with udpipe. // *Proceedings of the CoNLL 2017 shared task: Multilingual parsing from raw text to universal dependencies*, P 88–99.

J.A. Tuldava. 1975. On measuring text difficulty. // *Proceedings of Tartu State University*, volume 345, P 102–119.