Preprints.org

Article

Not peer-reviewed version

Towards Measuring Text Readability for Bulgarian

Ivelina Stoyanova *

Posted Date: 19 November 2026

Keywords: readability; readability indices; readability index calibration; Bulgarian

Preprints.org is a free multidisciplinary platform providing preprint service that is dedicated to making early versions of research outputs permanently available and citable. Preprints posted at Preprints.org appear in Web of Science, Crossref, Google Scholar, Scilit, Europe PMC, OpenAlex.

Copyright: This open access article is published under a Creative Commons CC BY 4.0 license, which permit the free download, distribution, and reuse, provided that the author and preprint are cited in any reuse.

Preprints.org | NOT PEER-REVIEWED | Posted: 19 November 2026 https://doi.org/10.20944/preprints202611.1991.v1

Disclaimer/Publisher's Note: The statements, opinions, and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions, or products referred to in the content.

Article

Towards Measuring Text Readability for Bulgarian

Ivelina Stoyanova

Department of Computational Linguistics, Institute for Bulgarian Language, Bulgarian Academy of Sciences; iva@dcl.bas.bg

Abstract

Readability indices are mathematical formulae estimating the difficulty of a text based on its features, most commonly simple counts of words, sentences, letters and syllables, and are widely used in education, public communication and natural language processing. The constants in the readability formulae are fitted to the language for which they were developed. So far, there are no readability indices developed specifically for Bulgarian and no indices have been tested systematically for it. We examine six widely used indices – Flesch Reading Ease, LIX, the Coleman–Liau Index, the Gunning Fog index, SMOG and Pisarek's formulae, as well as a neural measure based on a language model, on the texts of the Bulgarian national examinations at the end of grades 4, 7, 10 and 12. The indices are adapted to Bulgarian by comparing the values of their variables on parallel texts in Bulgarian and in the language of the original formula, and the adaptations are checked by fitting the formulae to the examination texts. Applied with their original constants, all indices give strongly shifted values for Bulgarian. The adapted indices give plausible values ordered by grade and separate the three levels of the examinations – elementary school (ending at grade 4), primary school (ending at grade 7) and high school (ending at grade 12).

Keywords: readability; readability indices; readability index calibration; Bulgarian

1. Introduction and Motivation

Readability is a property of the written text that determines how easily it can be read and understood by a given group of readers. It depends jointly on the characteristics of the text – the length and familiarity of its words, the complexity of its grammatical forms and syntactic structures, the way its parts are connected, as well as on the reading skills of the reader. Matching texts to the reading abilities of their intended readers is a long-standing practical need. Texts that are too difficult for their readers impede both the development of reading skills and the acquisition of the knowledge the texts convey, while texts that are too simple do not support further progress. For Bulgarian, this need is particularly pressing, as international assessments have consistently documented low reading literacy among Bulgarian students [1].

Readability indices are the most widely used tools for estimating the difficulty of a text. Typically, they estimate readability from a small number of easily computable surface variables, mainly the average length of the sentences and the average length of the words, and return a single numerical value interpreted on a scale of reading ease or school grades. Their simplicity, transparency and ease of automatic computation make them applicable in a wide range of settings [2,3]: in education, for selecting and grading reading materials, designing textbooks and tests, and matching texts to the reading level of students and target groups (e.g., L2 learners or groups with special needs), in health and public communication, for checking the accessibility of patient information, official documents and administrative texts, in publishing and in writing tools that give authors immediate feedback on the difficulty of their texts, in natural language processing as features, baselines and evaluation measures for automatic readability assessment and text simplification [4,5]. Decades after their introduction, and despite well-documented criticism of their linguistic assumptions [3], the classical indices remain the default choice in many of these applications.

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org | NOT PEER-REVIEWED | Posted: 19 November 2026 https://doi.org/10.20944/preprints202611.1991.v1

The constants of the readability formulae are known to be language-specific. They encode the typical length of words and sentences in the language for which the formula was developed, and the relation between these values and the reading skills of its readers. Most of the widely used indices were developed for English, and a formula developed for one language yields systematically shifted and hardly interpretable values when applied directly to another. Adaptations have been proposed for a number of languages, but for the Slavic languages the research remains limited. For Russian, the Flesch index was recalibrated by Oborneva [6], and several formulae have been developed for school texts. For Polish, the formulae of Pisarek [7] have been tested on larger data and implemented in automatic tools [8]. A recent study on Ukrainian concludes that neither the English formulae nor those developed for Russian and Polish can be applied to Ukrainian school texts without recalibration [9].

For Bulgarian, readability research is at an early stage. Kazakov et al. [10] make the first steps towards a Bulgarian readability index. Empirical studies of reading in Bulgarian primary school, based on finger-tracking data [11–14], provide reading-speed data that relate reading difficulty to the properties of words and sentences, and Stoyanova [15] uses these data to identify the lexical and grammatical features that predict reading difficulty for Bulgarian children in grades 2–5. To our knowledge, there is no systematic study of the applicability of the established readability indices to Bulgarian, of the recalibration of their constants.

The direct application of the established indices to Bulgarian does not take into account the specific linguistic features of the language. Like the other Slavic languages, Bulgarian has a rich inflectional and derivational morphology, which results in longer word forms. Unlike most Slavic languages, however, it has lost the nominal case system and expresses syntactic relations analytically, by prepositions and word order. Its verb system is complex, with a large number of tense forms, including analytic ones, and grammatical aspect expressed lexically through aspectual pairs of verbs [16]. Bulgarian is also a pro-drop language with relatively free word order. Finally, the definite article is postpositive and is attached to the end of the first word of the noun phrase as an enclitic, adding one to three letters, and often a syllable, to a large proportion of nouns and adjectives. Since the readability indices measure word length in letters or syllables, the article lengthens words without making them more difficult: the reading data show that it has practically no effect on reading speed even in grade 2 [15]. Such features suggest that, for Bulgarian, not only the constants of the formulae but also the thresholds of some of the features that define a ‘long’ or ‘complex’ word, may need to be redefined.

The present paper focuses on readability for readers of Bulgarian as a first language (L1). Readability for second-language (L2) learners and for specific target groups of readers relies on different reference data and different notions of difficulty, and is outside the scope of the present study. The paper addresses the following research questions:

RQ1. Are widely adopted readability indices applicable to Bulgarian? Can they be recalibrated, in terms of the constants in the formulae, and does recalibration improve their fit to the expected reading level of the texts?

RQ2. Can the variables used in the indices be redefined so that the indices become more sensitive to language-specific features of Bulgarian without losing their simplicity? Can accuracy be improved at the cost of simplicity, and to what extent?

RQ3. How much data is needed to calibrate a readability formula? Is there enough data for Bulgarian to establish a good fit for readability indices, and what kind of data is needed?

Answering these questions systematically, for Bulgarian and for other languages, can help to explain how readability indices work: how the simple text properties they use, and the way they combine them, manage to reflect the readability of a text in practice, with direct application in education.

The remainder of the paper is organised as follows. Section 2 presents a brief typology of readability measures and the indices considered in the study. Section 3 describes the Bulgarian reference data and the parallel texts used for the adaptation of the indices. Section 4 examines the

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org | NOT PEER-REVIEWED | Posted: 19 November 2026 https://doi.org/10.20944/preprints202611.1991.v1

applicability of the indices to Bulgarian and their recalibration. Section 5 discusses the results with respect to the research questions, Section **??** concludes the paper, and Section 6 outlines its limitations.

2. Related Works

2.1. Types of Readability Measures

Readability measures relate the characteristics of a text to the expected difficulty of reading and understanding it. With respect to how they are developed, two groups of measures can be distinguished. Theoretically (heuristically) motivated measures combine text characteristics that are assumed to make reading more difficult, without deriving their weights from reader data. Empirical measures are calibrated against an external criterion: results from comprehension tests, the school grade for which a text is intended, reader judgements, or reading speed. Most of the classical formulae are linear regression models fitted on such data.

With respect to the information they use, readability measures can be grouped into three broad types:

- formulae based on surface characteristics, which combine the average length of sentences with the average length of words, measured in letters or syllables, or with the share of long or polysyllabic words;
- multi-feature models, which combine a larger number of language-specific lexical, morphological, syntactic and discourse features, and, more recently, neural models and large language models.

The present study focuses on the first type. These formulae use only the counts of letters, syllables, words and sentences, which can be obtained for any text without language-specific resources, and are therefore the natural starting point for a new language. Moreover, despite their simplicity, they prove to be able to predict the text difficulty with higher level of certainty, and their results are comparable to those of much more complex formulae. They main deficiency is that they only provide a general quantitative measure and do not justify or identify any particular area or features of difficulty.

2.2. Readability Indices Based on Surface Characteristics

Readability indices based on surface characteristics return a single numerical value, which in general is interpreted on one of two types of scales. Reading-ease scales assign difficulty level (higher values correspond to easier texts or vise versa), while grade-level scales give the number of years of schooling, usually in the US or other English-speaking school system, that a reader needs in order to understand the text.

The list of abbreviations of the features used in the formulae below is provided at the end of the paper.

The Flesch Reading Ease index (FRE) [17] was developed for English and gives values from 0 (very difficult) to 100 (very easy):[1]

$$\mathrm{FRE} = 206.835 - 1.015 \times \mathrm{ASL} - 84.6 \times \mathrm{ASW} \tag{1}$$

Flesch [18] relates the scale to the school grade completed by the reader: texts with values from 90 to 100 are considered very easy and accessible after grade 4, texts from 60 to 70 standard, and texts below 30 very difficult, suitable for readers with university education. For Russian, Oborneva [6] recalibrated the coefficients by comparing the average length of words and sentences in English texts and their Russian translations, with the aim that equivalent texts receive equal values:[2]

$$\mathrm{FRE}_{\mathrm{ru}} = 206.836 - 1.3 \times \mathrm{ASL} - 60.1 \times \mathrm{ASW} \tag{2}$$

[1] https://quanteda.io/reference/textstat_readability.html

[2] https://nauka-pedagogika.com/pedagogika-13-00-02/dissertaciya-avtomatizirovannaya-otsenka-slozhnosti-uchebnyh-tekstov-na-osnove-statisticheskih-parametrov

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org | NOT PEER-REVIEWED | Posted: 19 November 2026 https://doi.org/10.20944/preprints202611.1991.v1

The LIX index (Swedish *läsbarhetsindex*) [19] was developed for Swedish and combines the average sentence length with the percentage of long words, defined as words with more than six letters. The formula is given in English by Anderson [20]:

$$\mathrm{LIX} = \mathrm{ASL} + 100 \times \frac{\text{long words}}{\text{words}} \tag{3}$$

LIX does not require syllable counting and has been widely used for other European languages. Its values range in practice from about 20 (very easy texts, such as children's books) to about 60 (very difficult texts, such as specialised and official texts) [15].

The Coleman–Liau Index (CLI) [21] was developed for English and uses letters instead of syllables:

$$\mathrm{CLI} = 0.0588 \times L - 0.296 \times S - 15.8 \tag{4}$$

where L is the number of letters per 100 words and S the number of sentences per 100 words. The result is interpreted as a US school grade.

Russian coefficients for the formula, fitted on texts labelled with school grades, are used in the Plain Russian Language project [22] and in the ruTS library [23].[3]

$$\mathrm{CLI}_{\mathrm{ru}} = 0.055 \times L - 0.35 \times S - 20.33 \tag{5}$$

The Gunning Fog index [24] was developed for English and combines the average sentence length with the percentage of complex words, defined as words with three or more syllables [2]:

$$\mathrm{Fog} = 0.4 \times \left(\mathrm{ASL} + 100 \times \frac{\text{complex words}}{\text{words}} \right) \tag{6}$$

The result is interpreted as the number of years of formal education needed to understand the text. In the Russian adaptation, complex words are defined as words with five or more syllables [22,23].

The SMOG grade [25] was developed for English on the basis of the number of polysyllabic words (with three or more syllables) in a sample of 30 sentences:[4]

$$\mathrm{SMOG} = 1.0430 \times \sqrt{\text{polysyllabic words} \times \frac{30}{\text{sentences}}} + 3.1291 \tag{7}$$

The result is interpreted as the US school grade needed to fully understand the text.

The Russian adaptation counts words with five or more syllables and uses different constants [23].

$$\mathrm{SMOG}_{\mathrm{ru}} = 1.1 \times \sqrt{64.6 \times \text{polysyllabic words}/\text{sentences}} + 0.05 \tag{8}$$

The formulae of Pisarek [7] were developed for Polish and combine the average sentence length with the percentage N of words with four or more syllables, in a linear and a non-linear form [8]:[5]

$$L_{\mathrm{lin}} = \frac{\mathrm{ASL}}{3} + \frac{N}{3} + 1 \tag{9}$$

$$L_{\mathrm{nlin}} = \frac{1}{2}\sqrt{\mathrm{ASL}^2 + N^2} \tag{10}$$

Higher values indicate more difficult texts. Pisarek's index is used in the Jasnopis system for Polish [8][6].

[3] https://sergeyshk.github.io/ruTS/en/stats/readability_stats/
[4] https://ogg.osu.edu/media/documents/health_lit/WRRSMOG_Readability_Formula_G._Harry_McLaughlin__1969_.pdf
[5] https://aclanthology.org/L14-1366/
[6] https://www.jasnopis.pl

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org | NOT PEER-REVIEWED | Posted: 19 November 2026 https://doi.org/10.20944/preprints202611.1991.v1

Its threshold of four syllables, higher than the three syllables used in English, reflects the longer words of an inflectional Slavic language.

A similar formula (to the linear variant of Pisarek's index) for Russian was proposed by Matskovskii [26]:

$$M_{\text{ru}} = 0.62 \times \text{ASL} + 0.123 \times N + 0.051 \tag{11}$$

Solnyshkina et al. [27] propose a readability formula for Russian school texts, which combines the average sentence length (ASL) and the average number of syllables per word (ASW) with two measures of the relation between nouns, adjectives and verbs: UNAV, the ratio between the number of unique adjectives and nouns and the number of unique verbs, and NAV, the ratio between the type–token ratios of adjectives and nouns and the type–token ratio of verbs. The formula includes the squares of the variables and their products:

$$\begin{aligned} S = {} & -0.124 \times \text{ASL} + 0.018 \times \text{ASW} - 0.007 \times \text{UNAV} + 0.007 \times \text{NAV} \\ & - 0.003 \times \text{ASL}^2 + 0.184 \times \text{ASL} \times \text{ASW} + 0.097 \times \text{ASL} \times \text{UNAV} \\ & - 0.158 \times \text{ASL} \times \text{NAV} + 0.09 \times \text{ASW}^2 + 0.091 \times \text{ASW} \times \text{UNAV} \\ & + 0.023 \times \text{ASW} \times \text{NAV} - 0.157 \times \text{UNAV}^2 - 0.079 \times \text{UNAV} \times \text{NAV} \\ & + 0.058 \times \text{NAV}^2 \end{aligned} \tag{12}$$

Unlike the classical indices, the formula uses information about parts of speech and requires morphological annotation. When applied to Ukrainian school texts without recalibration, it did not show a consistent increase with the level of the texts [9].

Prykhodchenko and Prykhodchenko [9] applied the original formulae of Flesch and Flesch–Kincaid for English, Pisarek for Polish, and Matskovskii and Solnyshkina et al. [27] for Russian to Ukrainian school texts without recalibration. None of the formulae could reliably determine the level of the texts: Pisarek's and Matskovskii's formulae increased with the level, but with a large spread within each level, and the formula of Solnyshkina et al. [27] showed a partly reversed trend. The results show that formulae cannot be transferred directly not only from English or Polish, but even between closely related Slavic languages such as Russian and Ukrainian, and the authors conclude that Ukrainian needs a formula with parameters adapted to the language.

The application of these indices to Slavic languages shows two main problems. First, the constants of the English formulae do not fit the longer words of the Slavic languages, and the values obtained by direct application are shifted towards greater difficulty. Second, even after recalibration, the indices separate well only distant levels of difficulty [9]. Similar results are reported for primary-school texts in Mexican Spanish, where the direct application of the Flesch formula gives very poor results and most formulae cannot distinguish between grades 4 and 6 [28].

2.3. Online Calculators and Software Tools

Readability indices are implemented in a number of online calculators[7] and software libraries, such as `textstat` for Python [29] and `quanteda` for R [30]. Language-specific tools have been developed for some languages, for example Jasnopis for Polish [8] and ruTS for Russian [23]. Most tools are designed for English and count syllables by English rules, and the implementations differ in details such as sentence segmentation and the definition of long words, so that the same index can give different values for the same text. For Bulgarian, the syllable-based indices require counting syllables by Bulgarian rules, which, in a language with transparent orthography, is generally equal to the number of vowel letters in the word.

[7] E.g., https://readabilityformulae.com/readability-scoring-system.php.

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org | NOT PEER-REVIEWED | Posted: 19 November 2026 https://doi.org/10.20944/preprints202611.1991.v1

2.4. Multi-Feature Models and Large Language Models

Since the early 2000s, readability assessment has been treated as a classification or regression task with a much larger set of features, as in Coh-Metrix [31], and, more recently, with neural models and large language models [5]. Language models can also be used to estimate readability without labelled data. Martinc et al. [32] propose the unsupervised Ranked Sentence Readability Score (RSRS), computed from the probabilities that a masked language model assigns to the words of a sentence:[8]

$$\mathrm{RSRS} = \frac{1}{S}\sum_{i=1}^{S}\left(\sqrt{i}\right)^{\alpha} \times \mathrm{WNLL}(i) \tag{13}$$

where S is the number of words in the sentence, $\mathrm{WNLL}(i)$ is the negative log-likelihood of the word with rank i (the words being ranked by increasing WNLL), and α is 2 for words that are not in the vocabulary of the model and 1 otherwise. Higher values indicate more difficult texts. RSRS does not require training data, which makes it a natural point of comparison with the surface indices for a language without annotated data.

These approaches reportedly achieve higher accuracy on the data on which they are developed, but they require large annotated data sets and language-specific models, and their results are more difficult to interpret. For Bulgarian, their application requires a better understanding of which characteristics of the text are important and how they should be measured. The simple surface indices are a good basis for such an understanding, because the characteristics they measure are among the most robust predictors of reading difficulty. Word length is the strongest single predictor of word-level reading difficulty for Bulgarian children [15], sentence length shows the highest correlation with the grade level of Mexican Spanish school texts [28], and Ehara [33] shows that the Flesch–Kincaid formulae can be interpreted as the average number of syllables per sentence, a measure that remains stable across ages.

3. Data

3.1. Supplementary Parallel Texts

To adapt the indices to Bulgarian, we compare the values of their variables in Bulgarian and in the languages for which the indices were originally developed. Following Oborneva [6], the comparison is made on parallel texts, so that the differences reflect the languages and not the content. We use the Bible translations from the parallel corpus of Christodoulopoulos and Steedman [34], which is available for all the languages we need: English, Swedish, Russian, Polish and Bulgarian. The comparison is made on the 30,367 verses available in all five translations. Sentences are counted within each verse at the sentence-final punctuation marks. Syllables are counted as vowel letters for Bulgarian, Russian, Polish and Swedish, and with the syllable counter of the textstat library [29] for English.

Table 1 shows the main values derived from the parallel texts. Bulgarian sentences are shorter than English and Swedish sentences, but longer than Russian and Polish ones. Bulgarian words have more syllables than English, Swedish and Polish words, but fewer than Russian words, which have case endings. The share of long words in Bulgarian is much higher than in English and Swedish: 21% of the Bulgarian words have more than six letters, compared with 12–14%, and a third of the Bulgarian words have three or more syllables, compared with 5% in English. The English Bible text is archaic and has an unusually small number of polysyllabic words, so the ratios for English should be treated with caution. An independent comparison of modern translations of the Gospel of Matthew gives a similar ratio between the sentence lengths in English and Bulgarian (1.24) [35].

[8] https://aclanthology.org/2021.cl-1.6/

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org | NOT PEER-REVIEWED | Posted: 19 November 2026 https://doi.org/10.20944/preprints202611.1991.v1

Table 1. Values of the variables derived from the parallel texts.

Language	ASL	ASW	Letters per word	>6 letters (%)	≥3 syll. (%)	≥4 syll. (%)	≥5 syll. (%)
English	22.5	1.28	4.08	12.1	4.9	1.0	0.1
Swedish	18.7	1.62	4.37	14.0	12.6	3.4	0.7
Russian	15.1	2.13	4.75	25.5	31.6	13.3	4.4
Polish	16.3	1.96	5.00	28.0	25.1	8.3	1.9
Bulgarian	17.4	2.07	4.40	21.0	33.1	11.3	3.5

3.2. Texts for Determining the Levels in Calibration

Calibrating a readability index requires texts for which the intended reading level is known. For Bulgarian, several types of texts could be considered: textbooks and reading books for each grade, books recommended for extracurricular reading, and the texts of the national examinations. We use only the examination texts, for the following reasons.

First, the examination texts are designed to be read and understood by the majority of students at the end of a given grade. They therefore represent the expected reading level of the grade. Second, they represent a fixed point in the development of reading skills – the level expected at the end of the school year, rather than the progression from one grade to the next during the year. Third, they are prepared by experts under the same procedure every year, which makes the texts of different years comparable. Finally, they are officially published and publicly available.

Teaching materials, in contrast, are less suitable for this purpose. Textbooks cover the whole school year, and in order to support the progress of the students, they often include texts and features that are deliberately more challenging than the current reading level. Including them would blur the relation between a text and the reading level it is intended for.

The examination texts also have limitations. They are not selected on the basis of explicit readability criteria, but for the tasks they serve: dictation, retelling, comprehension questions or literary analysis. As a result, their length depends on the task rather than on the grade, and the proportion of literary and informational texts differs between the grades. The number of texts is small, as it is limited to the examinations held since their introduction (in 2016).

The texts were collected from the published materials of the Bulgarian Ministry of Education and Science for four examinations in Bulgarian language and literature:

- the National External Assessment at the end of grade 4 (end of primary education, age 10–11), 2007–2026;
- the National External Assessment at the end of grade 7 (end of lower secondary education, age 13–14), 2010–2026;
- the National External Assessment at the end of grade 10 (end of the first stage of upper secondary education and of compulsory education, age 16–17), 2021–2026;
- the State Matriculation Examination at the end of grade 12 (end of secondary education, age 18–19), 2022–2026.

From the examination materials, we extracted all texts that the students are expected to read, excluding instructions, tasks and assessment criteria. In total, 99 texts were extracted. Five of them were excluded: three texts with deliberate spelling errors, used in editing tasks, and two texts from model examinations that are repeated in the examination of the following year. The texts were classified into three types: informational texts (popular science and journalistic texts, interviews, essays), literary texts (tales, fables, stories, excerpts from novels, descriptions) and poems. The nine poems are excluded from the analysis, because their line structure makes the sentence-based variables of the indices unreliable. Table 2 shows the composition of the set.

In the analysis, the texts for grades 10 and 12 are treated together as one level, denoted as level 10. There are three reasons for this. First, grade 10 marks the end of compulsory education, and the texts for grades 10 and 12 are intended for readers at the level of educated adults. At this level,

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org | NOT PEER-REVIEWED | Posted: 19 November 2026 https://doi.org/10.20944/preprints202611.1991.v1

differences in difficulty depend more on the subject and the domain of the text than on the grade, and they are difficult to establish. Second, the texts for the two grades do not differ in any of the measured characteristics (Section 3.4), and the grade 12 texts even have slightly shorter sentences and words. Third, the number of texts for each of the two grades is small (14 and 13), and merging them gives a group of a similar size to the other two (27 texts).

Table 2. Composition of the set of examination texts by grade and type of text.

Grade	Years	Informational	Literary	Poems	Total
4	2007–2026	3	21	–	24
7	2010–2026	17	17	–	34
10	2021–2026	12	2	4	18
12	2022–2026	10	3	5	18
Total		42	43	9	94

Moreover, this decision was justified by the data, which showed very close values for the two grades on all measured characteristics: the average sentence length (16.4 and 16.1 words), the average number of syllables per word (2.39 and 2.34), the average number of letters per word (5.27 and 5.18) and the share of long words (Table 3). None of the readability indices distinguished the two grades either, and the grade 12 texts even had slightly shorter sentences and words than the grade 10 texts.

The level of grade 12, and especially the levels above it, such as university texts, need other reference data. A proper reference for grade 12 would require a collection of common informational texts for adults from various domains, and a reference for the level above grade 12 would require university-level texts. In the present study, we therefore use the three levels: 4, 7 and 10.

3.3. Linguistic Processing

The examination texts were processed with UDPipe [36], which performs sentence segmentation, tokenisation, part-of-speech tagging, morphological analysis, lemmatisation and dependency parsing according to the Universal Dependencies framework. The counts used by the indices were obtained from the annotated texts:

- words are the tokens that contain at least one letter, so that punctuation marks and numbers written in digits are not counted;
- sentences are the sentences identified by UDPipe that contain at least one word;
- letters are the Cyrillic letters in the words;
- syllables are counted as the number of vowel letters in the word, with a minimum of one syllable per word.

Because Bulgarian orthography is transparent, the number of vowel letters gives the number of syllables of a word, and no hyphenation dictionary is needed. Words without a vowel, such as prepositions, are counted as one syllable.

3.4. Characteristics of the Texts

Table 3 presents the main characteristics of the 85 texts, which contain 27,929 words in 2,353 sentences, and Figure 1 shows the distribution of three variables by level and type of text.

Average sentence length increases clearly from grade 4 (9.3 words) to level 10 (16.3 words), while word length changes much less: the average number of syllables per word increases only from 2.18 to 2.37. The type of text has a stronger effect on the variables than the grade (Figure 1): the informational texts have much longer sentences and words than the literary texts, and the literary texts change little from grade to grade. Since the share of informational texts increases from 3 of 24 texts in grade 4 to 22 of 27 texts in level 10, part of the differences between the levels is due to the composition of the set.

In general, Bulgarian words are long compared to the languages for which most of the indices were developed. In the table relating the Flesch scale to the school grade of the reader, the average

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org | NOT PEER-REVIEWED | Posted: 19 November 2026 https://doi.org/10.20944/preprints202611.1991.v1

number of syllables per word ranges from 1.23 for very easy texts to 1.92 for very difficult texts [18], while even the texts for grade 4 have 2.18 syllables per word on average. This is the main reason why the direct application of the indices gives shifted values for Bulgarian.

Table 3. Characteristics of the texts by grade: number of texts, number of words, range of text length in words, average sentence length (ASL), average number of syllables per word (ASW), average number of letters per word, and percentage of long words by different thresholds (mean values over the texts).

Grade	Texts	Words	Length (min–max)	ASL	ASW	Letters per word	>6 letters (%)	≥8 letters (%)	≥4 syll. (%)	≥5 syll. (%)
4	24	3,493	47–459	9.3	2.18	4.80	28.2	18.0	13.7	2.7
7	34	14,589	127–716	13.2	2.26	5.00	30.7	21.2	16.4	6.0
10	14	4,528	239–495	16.4	2.39	5.27	33.8	25.3	21.1	8.9
12	13	5,319	235–813	16.1	2.34	5.18	32.2	23.5	18.7	8.0
10 (10+12)	27	9,847	235–813	16.3	2.37	5.23	33.0	24.4	20.0	8.5

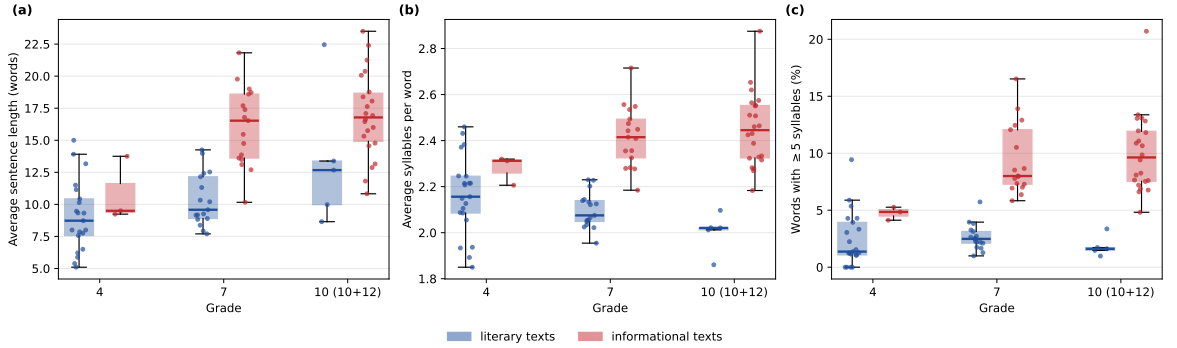


Figure 1. Distribution of (**a**) the average sentence length, (**b**) the average number of syllables per word and (**c**) the percentage of words with five or more syllables by level and type of text.

4. Evaluating Applicability of Readability Indices for Bulgarian

Each index is tested in three steps. First, the index is applied with its original constants, and, where it exists, with its Russian adaptation. Second, the index is adapted to Bulgarian with the help of the parallel texts (Table 1). The adaptation follows the idea of Oborneva [6]: a text and its translation should receive the same value of the index. If a variable has a different average value in the source language and in Bulgarian, its coefficient is multiplied by the ratio between the two values, and the constant of the formula is kept. For indices that count long words, the threshold for a long word can be changed instead, so that the share of long words in Bulgarian is the same as in the source language. Where a Russian formula exists, we derive a Bulgarian formula both from the original formula and from the Russian one, and compare the two. Third, the Bulgarian formula is checked by fitting its coefficients directly to the examination texts, with the structure and the constant of the formula kept.

The results are evaluated by four measures. The Spearman correlation (ρ) between the index and the level shows whether the index orders the texts correctly. The root mean squared error (RMSE) and the mean difference (bias) show how far the values are from the expected values: the grade itself for the indices that give a grade, and the values of Flesch's scale for FRE (90 for grade 4, 65 for grade 7 and 55 for level 10). The Δ_p index of Fajardo-Delgado et al. [28], computed from the p-values of Tukey's test for all pairs of levels, shows how well the levels are separated; for three levels, its maximum value is 3. Finally, we test whether the values can be interpolated between the levels: a logarithmic curve, $y = a + b \times \ln g$, where g is the grade, is fitted to the means for grades 4 and 10 and used to predict the mean for grade 7.

Table 4 summarises the results. The Bulgarian formulae recommended for each index are in bold.

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org | NOT PEER-REVIEWED | Posted: 19 November 2026 https://doi.org/10.20944/preprints202611.1991.v1

Table 4. Results for the examination texts (85 texts): mean values by level, Spearman correlation with the level ($|\rho|$), RMSE and bias with respect to the expected values, and separation of the levels (Δ_p, maximum 3). EN, SV, PL and RU indicate the language from which the Bulgarian formula is derived.

Index	Formula	4	7	10	$\lvert\rho\rvert$	RMSE	Bias	Δ_p
FRE	original (Equation (1))	13.4	2.4	−9.9	0.43	69.9	−67.3	2.80
	Russian (Equation (2))	64.0	54.0	43.4	0.48	21.8	−15.4	2.92
	Bulgarian, EN (Equation (14))	**80.6**	**71.1**	**61.4**	**0.48**	**14.5**	**1.8**	**2.94**
	Bulgarian, RU (Equation (15))	61.8	52.3	42.1	0.47	23.1	−17.1	2.89
LIX	original (Equation (3))	37.5	43.9	49.3	0.46	–	–	2.87
	Bulgarian, SV (Equation (16))	**27.3**	**34.5**	**40.7**	**0.48**	–	–	**2.94**
CLI	original (Equation (4))	9.0	11.1	13.0	0.44	5.0	4.0	2.87
	Russian (Equation (5))	2.0	4.3	6.1	0.48	4.1	−2.9	2.91
	Bulgarian, EN (Equation (17))	7.7	9.6	11.2	0.43	3.8	2.5	2.84
	Bulgarian, RU (Equation (18))	**3.5**	**6.0**	**8.1**	**0.48**	**3.3**	**−1.2**	**2.92**
Fog	original (Equation (6))	18.8	21.1	22.9	0.43	14.3	13.9	2.72
	Bulgarian, EN (Equation (19))	**4.8**	**7.7**	**9.9**	**0.61**	**2.5**	**0.5**	**2.99**
	Bulgarian, RU (Equation (21))	4.6	7.6	9.9	0.60	2.7	0.4	2.99
SMOG	original (Equation (7))	13.6	16.1	17.8	0.56	9.1	8.8	2.97
	Russian (Equation (8))	3.8	7.6	10.1	0.58	3.2	0.2	2.98
	Bulgarian, EN (Equation (20))	**5.6**	**8.0**	**9.6**	**0.58**	**2.3**	**0.7**	**2.98**
	Bulgarian, RU (Equation (22))	4.0	8.0	10.5	0.58	3.4	0.6	2.98
Pisarek	original (Equation (9))	8.7	10.9	13.1	0.52	4.7	3.8	2.97
	Bulgarian, PL (Equation (23))	**7.2**	**9.1**	**11.0**	**0.53**	**3.1**	**2.1**	**2.98**
Matskovskii	Russian (Equation (11))	7.5	10.3	12.6	0.60	4.0	3.1	2.99
	Bulgarian, RU (Equation (24))	7.0	9.5	11.7	0.59	3.4	2.4	2.99
Solnyshkina	Russian (Equation (12))	1.3	2.8	3.5	0.44	5.0	−4.5	2.52
	Bulgarian, RU	1.3	2.6	3.3	0.46	5.1	−4.6	2.69
	Bulgarian, fitted (Equation (25))[a]	5.9	7.0	8.3	0.61	1.8	0.0	2.99
RSRS	XLM-RoBERTa (Equation (13))	47.9	55.0	61.1	0.33	–	–	2.25
ASL	sentence length alone	9.3	13.2	16.3	0.60	–	–	2.99

[a] Values from leave-one-out cross-validation.

4.1. Flesch Reading Ease

Oborneva [6] found that English sentences are 1.25 times longer than Russian ones and that English words have 0.71 times as many syllables, and obtained the Russian coefficients $1.015 \times 1.25 \approx 1.3$ and $84.6 \times 0.71 \approx 60.1$. On the parallel texts, English sentences are 1.29 times longer than Bulgarian ones, and English words have 0.62 times as many syllables. Derived from the original formula, the Bulgarian formula is:

$$\mathrm{FRE}_{bg}^{en} = 206.835 - 1.3 \times \mathrm{ASL} - 52.5 \times \mathrm{ASW} \tag{14}$$

Derived from the Russian formula, with Russian sentences 0.87 times as long as Bulgarian ones and Russian words with 1.03 times as many syllables, it is:

$$\mathrm{FRE}_{bg}^{ru} = 206.836 - 1.13 \times \mathrm{ASL} - 61.8 \times \mathrm{ASW} \tag{15}$$

Both Bulgarian formulae have a lower coefficient for word length than the original, because Bulgarian words have more syllables than English ones, but the two formulae give different levels. The formula derived from English places grade 4 at 80.6 (easy texts), grade 7 at 71.1 and level 10 at 61.4 (standard texts), close to Flesch's scale. The formula derived from Russian, like the Russian formula itself, places all levels about 19 points lower. The difference comes from the parallel texts: on the Bible translations, the ratios between English and Russian differ from those of Oborneva [6], so the Russian formula does not fit these texts. The formula derived from English is also confirmed by the

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org | NOT PEER-REVIEWED | Posted: 19 November 2026 https://doi.org/10.20944/preprints202611.1991.v1

examination texts: fitting its coefficients to the texts with the original constant gives 1.28 × ASL and 53.1 × ASW, practically the same values.

In all variants, the literary texts have similar values at all levels (82–84 for FRE^{en}_{bg}), while the values of the informational texts decrease from 73 to 56. The logarithmic curve through grades 4 and 10 predicts the mean for grade 7 within one standard error and gives expected values for the grades between the levels (Figure 2), so that the value of a text can be expressed as a grade equivalent.

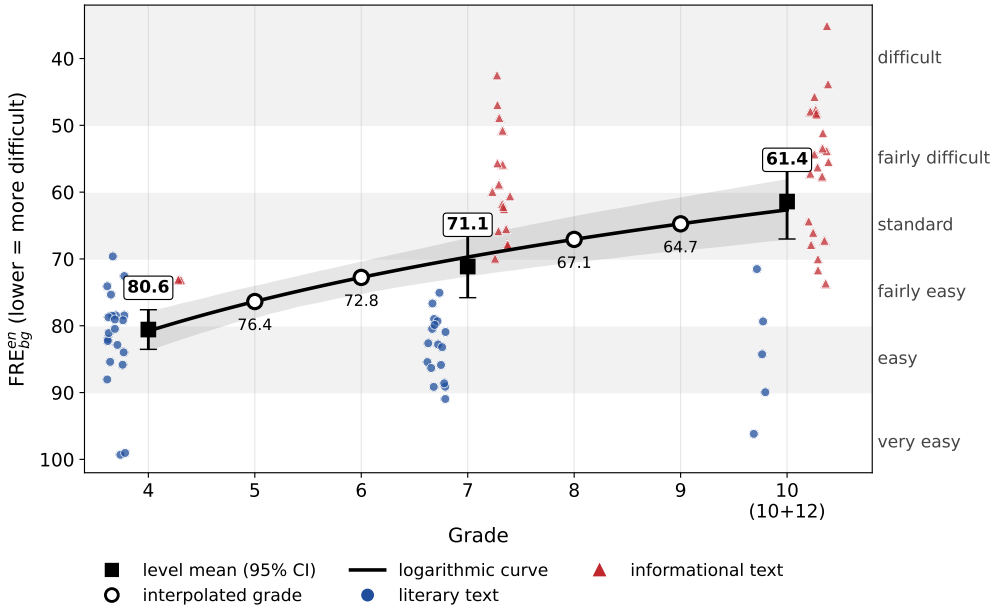


Figure 2. FRE^{en}_{bg} (Equation (14)) for the examination texts by level, with the logarithmic curve fitted to the means of the levels (shaded: 95% bootstrap band) and the interpolated values for grades 5, 6, 8 and 9. The levels of Flesch's scale are shown on the right.

4.2. LIX

LIX has no Russian adaptation, so the Bulgarian formula is derived from Swedish only. The sentences in the two languages have almost the same length (ratio 1.07), but Bulgarian has many more long words: 21% of the Bulgarian words have more than six letters, compared with 14% of the Swedish words. The share of Swedish words with more than six letters corresponds to the share of Bulgarian words with eight or more letters (13.4%), so we keep the formula and change only the definition of a long word:

$$\mathrm{LIX}^{sv}_{bg} = \mathrm{ASL} + \mathrm{LW}_{\geq 8} \tag{16}$$

where $\mathrm{LW}_{\geq 8}$ is the percentage of words with eight or more letters. Rescaling the coefficients instead ($1.07 \times \mathrm{ASL} + 0.67 \times \mathrm{LW}_{>6}$) gives almost the same results.

The Bulgarian formula lowers the values by about ten points, so that grade 4 is at 27 (easy texts) and level 10 at 41. LIX has no generally accepted scale related to school grades, so it cannot be checked against expected values. The threshold, however, is important: when it is raised from more than six to more than ten letters, the correlation of LIX with the level increases from 0.46 to 0.61. In Bulgarian, many common inflected forms and forms with the definite article have seven or eight letters, so only longer words are a reliable sign of a more complex vocabulary.

4.3. Coleman–Liau Index

The Coleman–Liau Index gives a grade directly, so its constant determines the level of the values. Derived from the original formula, with English words having 0.93 times as many letters as Bulgarian ones and English texts having 0.78 times as many sentences per 100 words, the Bulgarian formula is:

$$\mathrm{CLI}^{en}_{bg} = 0.0545 \times L - 0.2294 \times S - 15.8 \tag{17}$$

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org | NOT PEER-REVIEWED | Posted: 19 November 2026 https://doi.org/10.20944/preprints202611.1991.v1

Derived from the Russian formula (Equation (5)), with Russian words having 8% more letters than Bulgarian ones and Russian sentences being 13% shorter, it is:

$$\mathrm{CLI}^{ru}_{bg} = 0.0594 \times L - 0.4035 \times S - 20.33 \tag{18}$$

The original formula overestimates the grade of the Bulgarian texts by four grades, and the Russian formula underestimates it by three: Bulgarian is between the two languages. The formula derived from English reduces the overestimation, but cannot remove it, because the English constant is kept. The formula derived from Russian gives the best level, with values close to the grade for grades 4 and 7 (3.5 and 6.0), but it underestimates level 10 (8.1). It is, however, not confirmed by the examination texts: when the coefficients are fitted with the Russian constant, the weight of the sentences almost disappears. None of the variants orders the texts better than the original formula.

4.4. Gunning Fog Index and SMOG

For the Gunning Fog index and SMOG, the main parameter is the threshold for a polysyllabic (complex) word. With the English threshold of three syllables, a third of the Bulgarian words count as complex, compared with 5% in English. The share of Bulgarian words with five or more syllables (3.5%) is the closest to the English share. Derived from the original formulae, the Bulgarian formulae keep the constants and change only the threshold:

$$\mathrm{Fog}^{en}_{bg} = 0.4 \times (\mathrm{ASL} + \mathrm{PW}_{\geq 5}) \tag{19}$$

$$\mathrm{SMOG}^{en}_{bg} = 1.0430 \times \sqrt{\mathrm{NPW}_{\geq 5} \times \frac{30}{\text{sentences}}} + 3.1291 \tag{20}$$

where $\mathrm{PW}_{\geq 5}$ is the percentage and $\mathrm{NPW}_{\geq 5}$ the number of words with five or more syllables. Fog^{en}_{bg} coincides with the Russian adaptation of Fog, which uses the same threshold. Derived from the Russian formulae, with Russian sentences 0.87 times as long as Bulgarian ones and Russian texts having 1.26 times as many words with five or more syllables, the formulae are:

$$\mathrm{Fog}^{ru}_{bg} = 0.4 \times (0.87 \times \mathrm{ASL} + 1.26 \times \mathrm{PW}_{\geq 5}) \tag{21}$$

$$\mathrm{SMOG}^{ru}_{bg} = 1.1 \times \sqrt{70.4 \times \mathrm{NPW}_{\geq 5}/\text{sentences}} + 0.05 \tag{22}$$

The original formulae overestimate the grade by 14 (Fog) and 9 (SMOG) grades. With the threshold of five syllables, the overestimation almost disappears. Fog^{en}_{bg} gives 4.8 for grade 4, 7.7 for grade 7 and 9.9 for level 10, with a mean difference from the grade of only 0.5, and orders the texts best of all indices ($|\rho| = 0.61$). SMOG^{en}_{bg} also gives values close to the grade, with the smallest RMSE (2.3). The formulae derived from Russian give almost the same results, which confirms that the threshold, and not the coefficients, is the important parameter for Bulgarian. The fit to the examination texts gives coefficients of 0.35 instead of 0.4 for Fog and 0.81 instead of 1.04 for SMOG, so the structure of both formulae is confirmed. With the threshold of five syllables, the inflected forms and the forms with the definite article of common words are not counted as complex, while long derived and compound words, typical of informational texts, are.

4.5. Pisarek's and Matskovskii's formulae

Pisarek's formulae were developed for Polish and already use a higher threshold. The share of words with four or more syllables is 8.3% in Polish and 11.3% in Bulgarian, so the threshold is kept, and only the coefficients are rescaled by the ratios between Polish and Bulgarian (0.94 for the sentence length and 0.73 for the share of long words):

$$L^{pl}_{\mathrm{lin},bg} = \frac{0.94 \times \mathrm{ASL}}{3} + \frac{0.73 \times \mathrm{PW}_{\geq 4}}{3} + 1 \tag{23}$$

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org | NOT PEER-REVIEWED | Posted: 19 November 2026 https://doi.org/10.20944/preprints202611.1991.v1

Matskovskii's formula for Russian (Equation (11)) uses the same variables. Derived from it, with Russian texts having 1.18 times as many words with four or more syllables as Bulgarian texts, the Bulgarian formula is:

$$M_{bg}^{ru} = 0.54 \times \mathrm{ASL} + 0.145 \times \mathrm{PW}_{\geq 4} + 0.051 \tag{24}$$

Applied directly, Pisarek's and Matskovskii's formulae order the texts better than the original FRE, LIX, CLI and Fog, which suggests that formulae developed for Slavic languages transfer better to Bulgarian. The adaptation lowers the values of both formulae by 0.5–2 grades, but they remain about two grades above the actual grade. Interestingly, Polish words have more letters than Bulgarian words, but fewer syllables (Table 1), so the choice between letters and syllables changes the comparison between the two languages. When the coefficients of Pisarek's formula are fitted to the examination texts, the weight of the long words drops (0.22 instead of 0.73), so the adaptation is only partly confirmed.

4.6. Solnyshkina's Formula

The formula of Solnyshkina et al. [27] (Equation (12)) combines the surface variables with two variables that describe the relation between nouns, adjectives and verbs, and may therefore capture properties of the Slavic languages that the surface formulae do not. We computed UNAV and NAV from the part-of-speech annotation: UNAV from the lemmas of adjectives, nouns and verbs, and NAV from the type–token ratios of their word forms. Applied with the Russian coefficients, the formula gives values that increase with the level but are far below the grades (1.3 for grade 4 and 3.5 for level 10), and the difference between grade 7 and level 10 is not significant. Rescaling ASL and ASW by the ratios between Russian and Bulgarian changes the results only slightly.

As a small-scale test of whether the structure of the formula is suitable for Bulgarian, we calibrated all its coefficients on the examination texts, with the level as the target. Since the formula has 14 terms and the data contain only 85 texts, the coefficients were estimated with ridge regression, which reduces overfitting by keeping the coefficients small, and the result was evaluated by leave-one-out cross-validation. The calibrated formula is:

$$\begin{aligned} S_{bg}^{ru} = {} & -1.945 + 0.106 \times \mathrm{ASL} - 0.344 \times \mathrm{ASW} - 0.111 \times \mathrm{UNAV} + 1.556 \times \mathrm{NAV} \\ & - 0.005 \times \mathrm{ASL}^2 + 0.079 \times \mathrm{ASL} \times \mathrm{ASW} + 0.002 \times \mathrm{ASL} \times \mathrm{UNAV} \\ & + 0.084 \times \mathrm{ASL} \times \mathrm{NAV} + 0.300 \times \mathrm{ASW}^2 + 0.051 \times \mathrm{ASW} \times \mathrm{UNAV} \\ & + 0.164 \times \mathrm{ASW} \times \mathrm{NAV} - 0.033 \times \mathrm{UNAV}^2 - 0.116 \times \mathrm{UNAV} \times \mathrm{NAV} \\ & + 0.124 \times \mathrm{NAV}^2 \end{aligned} \tag{25}$$

On texts not used for the fit, the calibrated formula orders the texts as well as the best surface indices ($|\rho| = 0.61$, RMSE 1.8 grades), and separates the three levels. Its values, however, are compressed towards the middle of the scale (5.9 for grade 4 and 8.3 for level 10), and it improves only slightly on the average sentence length alone (RMSE 1.9). The terms with the largest contribution all include the sentence length (ASL × ASW, ASL × NAV, ASL^2 and ASL), while the terms with UNAV and NAV alone contribute little: on their own, UNAV and NAV correlate only weakly with the level ($\rho = 0.33$ and 0.17). Eleven of the 14 coefficients have the same sign as in the Russian formula, but the signs of ASL, ASW and ASL × NAV are reversed.

The results are therefore mixed. The structure of the formula can be calibrated for Bulgarian and gives results comparable to the best surface indices, but the part-of-speech variables add little to the sentence length, and some of the relations between the variables differ from Russian. This suggests that not all features that are relevant for one Slavic language are equally relevant for another, even a closely related one. A reliable calibration of a formula with so many coefficients requires much larger data, and the present results should be seen as a first test.

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org | NOT PEER-REVIEWED | Posted: 19 November 2026 https://doi.org/10.20944/preprints202611.1991.v1

4.7. Ranked Sentence Readability Score

RSRS was computed with the multilingual masked language model XLM-RoBERTa (base) [37], which includes Bulgarian. Its values have no fixed scale, so it is evaluated only by its correlation with the level and the separation of the levels. The mean values increase with the level, but RSRS orders the texts worse than any of the surface indices ($|\rho| = 0.33$), and the differences between grades 4 and 7 and between grade 7 and level 10 are not significant. RSRS is strongly correlated with the average sentence length ($\rho = 0.64$), because the score of a sentence is a sum over its words, and it does not improve the prediction when combined with the sentence length.

5. Discussion

None of the indices can be applied to Bulgarian with its original constants (RQ1). The English formulae place even the texts for grade 4 at the level of difficult texts, while the Russian formulae are shifted in the opposite direction. Adapted using parallel data, all surface indices give plausible values ordered by grade. The comparison of the two routes of adaptation shows that a formula from a closely related language is not necessarily a better starting point: for FRE, the formula derived from English fits the Bulgarian texts better than the one derived from Russian, while for CLI the Russian constant gives a more plausible level.

The adaptation improves the order of the texts only when the definition of a long word changes or when the formula comes from a Slavic language (Figure 3a). For Bulgarian, the definition of the variables is more important than the constants (RQ2). The best results are obtained with higher thresholds for long words, which compensate for the longer word forms of Bulgarian, due to inflection, derivation and the definite article rather than to more difficult vocabulary. With the threshold of five syllables, Fog^{en}_{bg} gives values close to the grade and orders the texts as well as the formula calibrated on the examination texts.

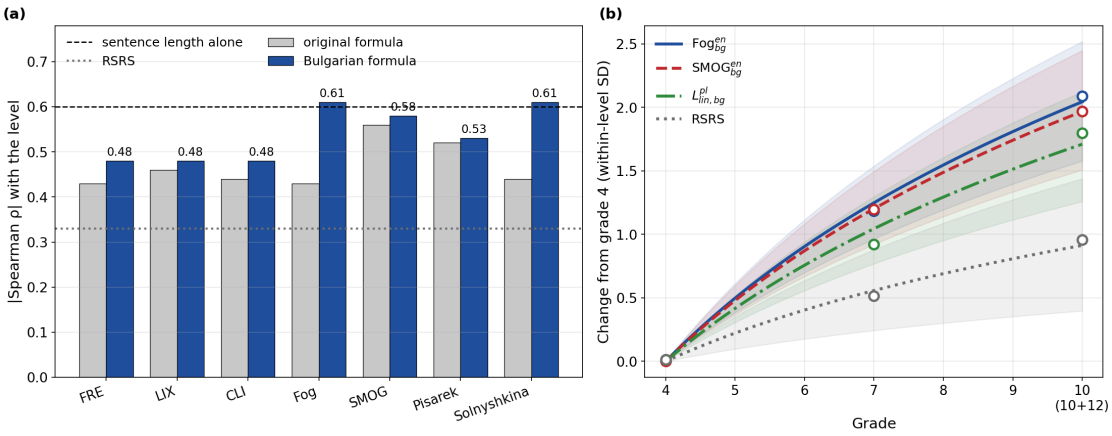


Figure 3. Comparison of the indices. (**a**) Spearman correlation with the level for the original formulae and the Bulgarian formulae, compared with the average sentence length alone and with RSRS. (**b**) Logarithmic curves fitted to the means of the levels for the three best surface indices and for RSRS, with 95% bootstrap bands.

The morphology of Bulgarian can also be taken into account by measuring the length of words by removing the postpositive definite article, which accounts for 13–15% of the words in the examination texts. In a preliminary test, removing the article before counting letters and syllables increased the correlation with the level from 0.48 to 0.51 for FRE^{en}_{bg} and from 0.48 to 0.54 for LIX^{sv}_{bg}, and raised the value of FRE^{en}_{bg} for grade 4 to 87.6, close to Flesch's scale. This agrees with the reading data, which show that the article does not affect reading speed [15]. Other features could be included in a similar way while keeping the formulae simple. In a preliminary experiment, adding three such features to Fog^{en}_{bg} increased the cross-validated correlation with the level from 0.59 to 0.65, so accuracy can be improved at the cost of simplicity.

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org | NOT PEER-REVIEWED | Posted: 19 November 2026 https://doi.org/10.20944/preprints202611.1991.v1

The calibration of Solnyshkina's formula shows the limits of this approach on small data. Calibrated on the Bulgarian texts, the formula orders the texts as well as the best surface indices, but its part-of-speech variables add little to the sentence length, and some coefficients change their sign compared with Russian. Not all features that are relevant for one Slavic language are equally relevant for another, and formulae with many coefficients need much larger data to be calibrated reliably.

No index orders the texts much better than the average sentence length alone, which carries most of the information about the grade in all formulae, and the neural measure RSRS orders them worse than any surface index. The values of all surface indices, however, can be interpolated between the levels with a logarithmic curve (Figure 3b), so that the indices can be used on a continuous grade scale. The type of text has a stronger effect than the grade: the literary texts have similar values at all levels, so a practical index may need separate reference values for literary and informational texts.

The examination texts are sufficient to separate the three levels: about 15–30 texts per level are needed for this, which the data meet (RQ3). The texts for grades 10 and 12, however, cannot be separated by any index, and the levels of grade 12 and above require different reference data: common informational texts for adults from various domains for grade 12, and university-level texts above it. A more precise calibration would also need reference texts for more grades, a balanced proportion of literary and informational texts, and data on actual readers, such as comprehension results and reading times.

6. Limitations

The main limitation of the study is the limited volume of the reference data and its internal diversity and inconsistency of features due to the different proportions of literary and informational texts at each level. This weakness should be overcome by collecting more data and validating indices against text types as well.

The grade of a text is used as the measure of its difficulty, as no comprehension data were available. However, the grade is a coarse measure and it reflects the judgement of the authors of the examinations rather than the performance of the readers. A more precise validation would require data on actual readers, such as comprehension results and reading times. Further, the interpolation between the levels assumes that the expected values change smoothly from grade to grade.

The calculation of some features is preliminary, as more detailed annotation is required. The calibration of Solnyshkina's formula, with 14 coefficients estimated on 85 texts, should be seen as a preliminary test only.

Finally, the parallel texts used for the adaptation of the indices are Bible translations, whose archaic language affects the ratios between the languages, especially for English.

Data Availability Statement: The examination texts, their annotation and the values of all variables are available at https://github.com/DCL-IBL/READABILITY-BG.

Acknowledgments: During the preparation of this manuscript, the author used AI for data processing, generation of diagrams, formatting and proofreading the manuscript. The author has reviewed the manuscript and takes full responsibility for the content of this publication.

Conflicts of Interest: The author declares no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

ASL	Average sentence length (number of words per sentence)
ASW	Average number of syllables per word
CLI	Coleman–Liau Index
FRE	Flesch Reading Ease
Fog	Gunning Fog index
L	Number of letters per 100 words

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org | NOT PEER-REVIEWED | Posted: 19 November 2026 https://doi.org/10.20944/preprints202611.1991.v1

L_{lin}, L_{nlin}	Pisarek's linear and non-linear formulae
LIX	Läsbarhetsindex (readability index)
$\text{LW}_{>6}$, $\text{LW}_{\geq 8}$	Percentage of long words (more than six letters; eight or more letters)
M	Matskovskii's formula
NAV	Ratio of the type–token ratios of adjectives and nouns to the type–token ratio of verbs
$\text{NPW}_{\geq 5}$	Number of words with five or more syllables
$\text{PW}_{\geq 4}$, $\text{PW}_{\geq 5}$	Percentage of words with four or more (five or more) syllables
RMSE	Root mean squared error
RSRS	Ranked Sentence Readability Score
S	Number of sentences per 100 words (in CLI); Solnyshkina's formula
SMOG	Simple Measure of Gobbledygook
UNAV	Ratio of the number of unique adjectives and nouns to the number of unique verbs
WNLL	Word negative log-likelihood
XLM-RoBERTa	Cross-lingual Language Model – Robustly Optimised BERT Approach
Δ_p	Separability index of the levels, based on Tukey's test

References

1. OECD. *PISA 2022 Results (Volume I): The State of Learning and Equity in Education*; OECD Publishing: Paris, 2023. https://doi.org/10.1787/53f23881-en.
2. DuBay, W.H. The Principles of Readability. Impact Information; ERIC ED490073, 2004.
3. Benjamin, R.G. Reconstructing readability: Recent developments and recommendations in the analysis of text difficulty. *Educational Psychology Review* **2012**, *24*, 63–88.
4. Collins-Thompson, K. Computational assessment of text readability: A survey of current and future research. *ITL – International Journal of Applied Linguistics* **2014**, *165*, 97–135. https://doi.org/10.1075/itl.165.2.01col.
5. Vajjala, S. Trends, Limitations and Open Challenges in Automatic Readability Assessment Research. In Proceedings of the Proceedings of the Thirteenth Language Resources and Evaluation Conference, Marseille, France, 2022; pp. 5366–5377.
6. Oborneva, I.V. Avtomatizirovannaya otsenka slozhnosti uchebnyh tekstov na osnove statisticheskih parametrov. PhD thesis, Institut soderzhaniya i metodov obucheniya RAO, Moskva, 2006.
7. Pisarek, W. Jak mierzyć zrozumiałość tekstu? *Zeszyty Prasoznawcze* **1969**, pp. 35–48.
8. Broda, B.; Nitoń, B.; Gruszczyński, W.; Ogrodniczuk, M. Measuring Readability of Polish Texts: Baseline Experiments. In Proceedings of the Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14), Reykjavik, 2014.
9. Prykhodchenko, S.D.; Prykhodchenko, O.Y. Application of Existing Readability Methods to the Ukrainian Language: A Comprehensive Study. In Proceedings of the Proceedings of the Third Workshop on Quantitative Syntax (QUASY, SyntaxFest 2025), Ljubljana, Slovenia, 2025; pp. 17–25.
10. Kazakov, D.; Temnikova, I.; Minkov, S.; Margova, R.; Emanuilov, I. Towards Creating a Bulgarian Readability Index. In Proceedings of the Proceedings of the 1st Workshop on Advancing NLP for Low-Resource Languages associated with RANLP 2025, Varna, Bulgaria, 2025; pp. 192–200. https://doi.org/10.26615/978-954-452-100-4-018.
11. Marzi, C.; Ferro, M.; Nadalini, A.; Pirrelli, V.; Todorova, M.; Dimitrova, T.; Stefanova, V.; Kukova, H.; Koeva, S. Comparable Reading Development in Bulgarian and Italian: Cross-Linguistic Insights from a Finger-Tracking Study. *Languages* **2026**, *11*, 70. https://doi.org/10.3390/languages11040070.
12. Lento, A.; Nadalini, A.; Ferro, M.; Marzi, C.; Pirrelli, V.; Dimitrova, T.; Kukova, H.; Stefanova, V.; Todorova, M.; Koeva, S. Assessing reading literacy of Bulgarian pupils with finger-tracking. In Proceedings of the Proceedings of the Sixth International Conference on Computational Linguistics in Bulgaria (CLIB 2024), Sofia, Bulgaria, 2024; pp. 140–149. https://doi.org/10.47810/CLIB.24.14.
13. Pirrelli, V.; Koeva, S. Developing materials for assessing reading literacy and comprehension of early graders in Bulgaria and Italy. *Foreign Language Teaching (Chuzhdoezikovo Obuchenie)* **2024**, *51*, 29–37. https://doi.org/10.53656/for2024-01-03.
14. Koeva, S.; Stefanova, V.; Stoyanova, I.; Todorova, M. Otsenka na umeniyata za chetene v nachalna uchilishtna vazrast. *Balgarski ezik* **2023**, *70*, 92–102. In Bulgarian, https://doi.org/10.47810/BL.70.23.04.08.
15. Stoyanova, I. Readability Criteria for Bulgarian: Lexical and Grammatical Dimensions. *Computational Linguistics in Bulgaria* **2026**, *2*, 98–124. https://doi.org/10.47810/JCLIB.2.2026.05.
16. Nicolova, R. *Balgarska gramatika. Morfologoya*; Sofia University Press: Sofia, 2008.

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org | NOT PEER-REVIEWED | Posted: 19 November 2026 https://doi.org/10.20944/preprints202611.1991.v1

17. Flesch, R. A new readability yardstick. *Journal of Applied Psychology* **1948**, *32*, 221–233. https://doi.org/10.1037/h0057532.
18. Flesch, R. *The Art of Readable Writing*; Harper: New York, 1949.
19. Björnsson, C.H. *Läsbarhet*; Liber: Stockholm, 1968.
20. Anderson, J. Lix and Rix: Variations on a little-known readability index. *Journal of Reading* **1983**, *26*, 490–496.
21. Coleman, M.; Liau, T.L. A computer readability formula designed for machine scoring. *Journal of Applied Psychology* **1975**, *60*, 283–284. https://doi.org/10.1037/h0076540.
22. Begtin, I. Plain Russian Language. GitHub repository, n.d.
23. Shkarin, S. ruTS: A library for extracting statistics from Russian texts. Readability metrics. Software documentation, 2026. Accessed 2026-11-06.
24. Gunning, R. *The Technique of Clear Writing*; McGraw-Hill: New York, 1952.
25. McLaughlin, G.H. SMOG grading: A new readability formula. *Journal of Reading* **1969**, *12*, 639–646.
26. Matskovskii, M.S. Problems of Readability of Printed Material. In *Semantic Perception of a Speech Message in Mass Communication*; Nauka: Moscow, 1976; pp. 126–142.
27. Solnyshkina, M.; Ivanov, V.; Solovyev, V. Readability Formula for Russian Texts: A Modified Version. In Proceedings of the Advances in Computational Intelligence. MICAI 2018, Cham, 2018; Vol. 11289, *Lecture Notes in Computer Science*, pp. 132–145. https://doi.org/10.1007/978-3-030-04497-8_11.
28. Fajardo-Delgado, D.; Rodriguez-Coayahuitl, L.; Sánchez-Cervantes, M.G.; Álvarez-Carmona, M.Á.; Rodríguez-González, A.Y. Readability Formulas for Elementary School Texts in Mexican Spanish. *Applied Sciences* **2025**, *15*, 7259. https://doi.org/10.3390/app15137259.
29. textstat contributors. textstat: Python package to calculate readability statistics of a text object. GitHub, n.d.
30. Benoit, K.; Watanabe, K.; Wang, H.; Nulty, P.; Obeng, A.; Müller, S.; Matsuo, A. quanteda: An R package for the quantitative analysis of textual data. *Journal of Open Source Software* **2018**, *3*, 774. https://doi.org/10.21105/joss.00774.
31. Graesser, A.C.; McNamara, D.S.; Kulikowich, J.M. Coh-Metrix: Providing multilevel analyses of text characteristics. *Educational Researcher* **2011**, *40*, 223–234. https://doi.org/10.3102/0013189X11413260.
32. Martinc, M.; Pollak, S.; Robnik-Šikonja, M. Supervised and Unsupervised Neural Approaches to Text Readability. *Computational Linguistics* **2021**, *47*, 141–179. https://doi.org/10.1162/coli_a_00398.
33. Ehara, Y. An Analytical Study of the Flesch-Kincaid Readability Formulae to Explain Their Robustness over Time. In Proceedings of the Proceedings of the 38th Pacific Asia Conference on Language, Information and Computation, Tokyo, Japan, 2024; pp. 989–997.
34. Christodoulopoulos, C.; Steedman, M. A massively parallel corpus: the Bible in 100 languages. *Language Resources and Evaluation* **2015**, *49*, 375–395. https://doi.org/10.1007/s10579-014-9287-y.
35. Matricciani, E. Readability across Time and Languages: The Case of Matthew's Gospel Translations. *AppliedMath* **2023**, *3*, 497–509. https://doi.org/10.3390/appliedmath3020026.
36. Straka, M.; Straková, J. Tokenizing, POS Tagging, Lemmatizing and Parsing UD 2.0 with UDPipe. In Proceedings of the Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies, Vancouver, Canada, 2017; pp. 88–99.
37. Conneau, A.; Khandelwal, K.; Goyal, N.; Chaudhary, V.; Wenzek, G.; Guzmán, F.; Grave, E.; Ott, M.; Zettlemoyer, L.; Stoyanov, V. Unsupervised Cross-lingual Representation Learning at Scale. In Proceedings of the Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 2020, pp. 8440–8451. https://doi.org/10.18653/v1/2020.acl-main.747.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.