Application of Existing Readability Methods to the Ukrainian Language: A Comprehensive Study

Serhii D. Prykhodchenko*, **Oksana Yu. Prykhodchenko**[x]
∗ Computer Systems Software Department, Dnipro University of Technology, Dnipro, Ukraine
prykhodchenko.s.d@nmu.one
x Computer Systems Software Department, Dnipro University of Technology, Dnipro, Ukraine
prykhodchenko.o.yu@nmu.one

Abstract

The Ukrainian language currently lacks a well-developed framework for assessing text readability. This study addresses this gap by focusing on three key contributions. First, we present the creation of UkrTB, a Ukrainian-language corpus of texts categorized by reader age. Second, we conduct a statistical analysis of the corpus, evaluating key linguistic features such as sentence length, word complexity, and part-of-speech distribution. Third, we systematically assess the applicability of existing readability formulas, including Flesch, Flesch-Kincaid, Matskovskii, Pisarek, and Solnyshkina et al., to Ukrainian texts. Our findings indicate that readability models developed for English and other Slavic languages exhibit significant limitations when applied to Ukrainian. While some methods demonstrate partial correlation with expected readability levels, others produce inconsistent results, underscoring the need for a specialized readability metric tailored to Ukrainian. This work lays the foundation for further research in Ukrainian readability assessment and the development of language-specific models.

Keywords: Readability, Ukrainian language, Natural Language Processing, Corpus Linguistics, Text Complexity

1 Introduction

Research on quantitative readability assessment began in the 1940s with Flesch's early work (Flesch, 1948), leading to the development of various readability metrics. However, these methods were primarily designed for English and did not account for structural differences in other languages. English, as an analytical language, relies on a fixed word order and auxiliary words to convey meaning, whereas Slavic languages, including Ukrainian, use a more synthetic structure characterized by extensive inflection, case systems, and grammatical gender. These differences complicate the direct application of English-based readability models to Ukrainian, as they fail to capture the complexity introduced by its flexible word order and morphological variation. Ukrainian is characterised by rich morphology (cases, long word forms), syllabic structure (complex syllable divisions, less vowel reduction) and syntax (long sentences, free word order). Polish has consonant clusters (ex. 'szczegółowy'). At the same time, Russian is characterised by vowel reduction - this does not coincide with Ukrainian, where vowels are fuller. Formula coefficients must be calibrated on the Ukrainian corpus of texts and tested on native speakers, otherwise they give incorrect results. Since the 1960s, readability studies have been conducted for some Slavic languages, particularly Polish and Russian, leading to the development of language-specific models. However, Ukrainian remains underexplored in this context, despite having a substantial number of native speakers and a distinct grammatical system. The absence of a dedicated readability assessment framework for Ukrainian presents a challenge for text classification, educational content adaptation, and NLP applications. Existing readability formulas, whether developed for English or other Slavic languages, may not be directly transferable to Ukrainian due to its unique linguistic features.

To address this gap, this study makes three key contributions. First, we construct a Ukrainian-language text corpus categorized by readability levels to provide a foundation for further research. Second, we conduct a statistical analysis of the corpus, evaluating linguistic features such as sentence structure, word complexity, and part-of-speech distribution. Third, we systematically

Proceedings of the Third Workshop on Quantitative Syntax (QUASY, SyntaxFest 2025), pages 17–25

assess the applicability of existing readability formulas, including those developed for English, Polish, and Russian, to determine their effectiveness for Ukrainian. Our findings will help establish whether current models can be adapted or if a new methodology is required to develop an accurate readability assessment framework for the Ukrainian language.

2 Related work

To consider the task of determining the applicability of existing methods for calculating readability to Ukrainian-language texts, it is necessary to solve several successive tasks:

1. Create a text corpus suitable for testing the research hypotheses and, in the long term, for further research.
2. Determine existing methods for determining readability that use the desired parameters for determining readability created for Ukrainian and other languages.
3. Apply previously determined methods, evaluate their accuracy for the Ukrainian language, and draw conclusions for further research.

Prior research has explored readability assessment across multiple domains. Studies on text complexity for second-language learners have focused (Xia et al., 2016) on CEFR-graded English datasets, addressing the challenge of limited annotated data. These works have adapted models trained on native corpora, leveraging domain adaptation and self-training techniques to improve performance. High accuracy and strong correlation coefficients suggest that similar approaches could be explored for the Ukrainian language.Lexical richness has also been shown to correlate with perceived quality in ESL learners' oral narratives (Lu, 2012), suggesting that vocabulary diversity is an important dimension of readability and language proficiency assessment.

Efforts to standardize readability assessment in educational materials have also been undertaken, particularly in the context of Ukrainian textbooks. Research from Lviv Polytechnic (Krychkovska et al., 2014) investigated text complexity using the "Chitanka" program, analyzing the accessibility of scientific content. The study emphasized the importance of systematically increasing textual complexity in educational materials, advocating for a more structured approach to readability in Ukrainian academia.

Recent advancements in multilingual readability estimation highlight the potential of cross-lingual transfer learning. The paper "An Open Multilingual System for Scoring Readability of Wikipedia" (Trokhymovych et al., 2024) presents a novel approach to assessing the readability of Wikipedia articles across multiple languages. It introduced a model trained on Wikipedia articles in 14 languages, demonstrating significant improvements over previous benchmarks. By aligning document pairs across different languages, the model effectively assessed text complexity even in languages with limited resources, providing a strong foundation for future Ukrainian readability models.

The UKRMED corpus (Cherednichenko et al., 2020), focuses on Ukrainian medical texts, including clinical protocols and forums, categorized by complexity. Its creation involved preprocessing, tokenization, and statistical analysis, with crowdsourcing to improve quality. Studies using Pymorphy2 showed that frequency-based methods are insufficient, highlighting the need for advanced linguistic features and lexical resources—relevant for broader readability assessment.

Next studies (Cherednichenko et al., 2018) highlight the diverse methodologies for readability evaluation and corpus development. While significant progress has been made in multilingual and domain-specific readability assessment, Ukrainian remains underexplored. By leveraging insights from existing research, this study aims to construct a robust Ukrainian readability framework, integrating statistical, linguistic, and computational approaches. They used the Pymorphy2 morphological analyzer and stopword lists for preprocessing. Their results underscored the need for specialized lexical resources and ontologies to simplify medical texts for better comprehension by non-specialist readers effectively.

The article (Cherednichenko and Kanishcheva. 2021) examined the application of readability formulas to Ukrainian medical texts using the UKRMED corpus, which categorizes texts into three levels of complexity: simple, moderate, and complex. The results indicated that while existing formulas produce similar rankings, medical texts remain inherently difficult to understand. The study suggests that further refinement of readability metrics and detailed text markup could improve accessibility for both non-native speakers and individuals with varying educational backgrounds.

The paper (Vajjala and Meurers, 2012) investigated the impact of second-language acquisition (SLA) research on readability classification. By integrating SLA-based complexity measures with traditional readability metrics, the researchers improved classification accuracy. Their model, which combined lexical and syntactic features, outperformed conventional methods, achieving over 93% accuracy in predicting text difficulty across different grade levels. This work demonstrates the potential of interdisciplinary approaches in enhancing automated readability assessment, particularly in educational and language-learning contexts.

For Russian texts, (Solnyshkina, 2018) proposed a modified readability formula incorporating syntactic, lexical, and frequency features. Tested on Russian school textbooks, it showed improved accuracy over previous models and highlighted the potential of regression-based approaches for genre-specific readability assessment.

Similarly, research (Broda et al., 2014) on Polish texts evaluated multiple readability assessment methods, including traditional formulas such as Gunning-Fog and Pisarek's method. In addition, the study introduced novel approaches, such as distributional lexical similarity and an automated Taylor test based on statistical language modelling. The authors developed an online tool for readability evaluation, showing a strong correlation between various readability indices and effective classification of text complexity levels.

Corpus parameterization has also been a focus in readability studies (Starko and Cheylytko, 2013), as highlighted in research on optimizing corpus balance and representativity. A hybrid methodology combining statistical models, expert evaluations, and adaptive monitoring was proposed to address the proportionality of text types in linguistic corpora. This study emphasizes the importance of dynamic, data-driven strategies in corpus design, ensuring that robust and well-structured datasets support readability research.

Adapting readability assessment methods for the German language has seen significant advancements through various innovative approaches. Recent research has focused on integrating machine learning and natural language processing techniques to enhance the accuracy and applicability of these methods. An online service has been developed that utilizes five statistical methods and two machine learning models, including BERT, to evaluate German text readability at the sentence level (Pickelmann, 2023). This tool is particularly beneficial in educational contexts, helping to assess the suitability of teaching materials for different grade levels.

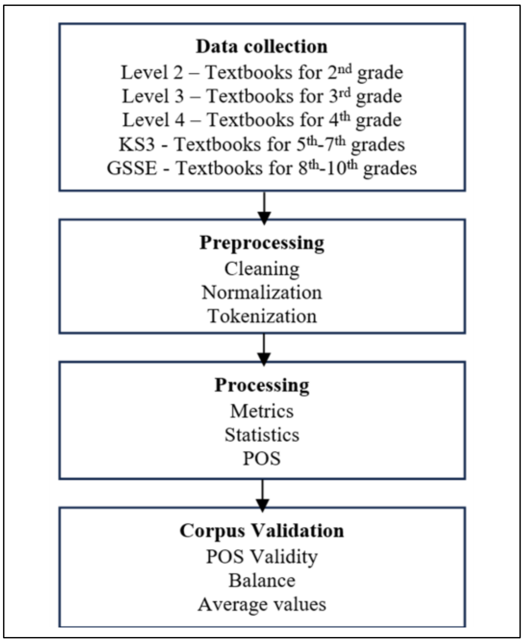

Figure 1: Algorithm of Text corpus creation

Recent research in German has demonstrated the use of neural models (e.g., BERT-based) for sentence-level readability prediction, showing potential for future adaptation to low-resource languages.(Blaneck et al., 2022), (Mohtaj et al., 2022).

3 Methods

This study follows a structured approach to evaluate the applicability of existing readability metrics to Ukrainian texts. The first step involved constructing UkrTB, a corpus of 750 texts extracted from Ukrainian educational materials, categorized into five readability levels corresponding to different school grades. The texts were preprocessed through cleaning, normalization, and tokenization. Part-of-speech tagging was performed using pymorphy3, leveraging OpenCorpora dictionaries to ensure consistency in morphological analysis. We computed statistical features — including sentence length, word length, and syllable count — to quantitatively evaluate text readability.

To assess readability, we applied six established formulas: Flesch Readability Ease (Flesch, 1948), Flesch-Kincaid Grade Level (Kincaid et al., 1975), Pisarek's (Pisarek, 1969) linear and nonlinear models, Matskovskii's readability index (Matskovskii, 1976), and Solnyshkina et al.'s model (Solnyshkina, 2018). These metrics were chosen based on their relevance to English and Slavic languages. Each formula was systematically applied to the UkrTB corpus, producing readability scores for each text. The evaluation of these methods involved a correlation analysis using Spearman and Pearson coefficients to examine the relationship between formula scores and the intended educational levels. The correlation between computed readability scores and the predefined grade levels was quantified using both Pearson's and Spearman's coefficients. These statistical measures allow for evaluating the degree of linear and monotonic relationships, respectively. A distribution analysis was also conducted to assess intra-class variability, determining whether readability scores were consistent within the same grade level.

4 Collection of the dataset

Creating a text corpus for a source dataset is challenging, as the source data strongly influences the processing results. Currently, there are no standard techniques for creating text corpora in linguistics. Our method for collecting a corpus of texts has similarities with the WeeBit corpus, which involved Weekly Reader texts ranked by reader age (Vajjala and Meurers, 2012). We propose our own method for collecting texts to create a training and general corpus based on the work of (Cherednichenko and Kanishcheva. 2021)(Vajjala and Meurers, 2012). The main requirement for the corpus is to provide Ukrainian text data ranked by readers' ages to study language problems.

The corpus of Ukrainian-language texts we are creating, which we call UkrTB - Ukrainian Textbooks[1], was originally conceived as a corpus without saved tags and attributes. The first reason is the availability of modern tagging and attribution techniques, including the rapid creation of keyword lists using artificial intelligence elements. The second reason for not tagging is the perceived versatility of the corpus, which, as research progresses, is planned to be expanded to the size of WeeBit, i.e. about 6000 texts, divided into appropriate ages of potential readers, where texts range from folk tales to excerpts from history or physics textbooks.

This study proposes the following process (Fig. 1) based on the work (Cherednichenko et al., 2020) to create a text corpus ranging by readers' ages in Ukrainian. As data sources, texts from the electronic library formed by the Ministry of Education of Ukraine, which contains all textbooks and auxiliary literature recommended by the Ministry of Education from the first to the 11th grade, are considered. These texts include both folk tales and samples of literature from the 11th to 21st centuries, as well as scientific explanations of, for example, physics or chemistry,

It should be explained that there is no single textbook for the discipline studied in Ukraine; an author may submit their textbook, which meets the specified criteria, to the Ministry of Education, get approval, and then their textbook will be included in the list of recommended textbooks. Thus, on the recommended textbooks website, we can find 5 history textbooks for 7th grade and 6 Ukrainian literature textbooks for 8th grade, which may differ in the selection of materials and writing style. Nevertheless, the Ministry's recommendation confirms that these texts fulfil the criteria set out for textbooks, which means that a pupil of the relevant grade will be able to read the material.

The corpus design and validation processes were inspired by the works of Vasyl Starko (Starko and Cheylytko, 2013), who is one of the creators of the Brown Corpus of the Ukrainian language. Obviously, his works are based on the ideas of creating the Brown Corpus of English (Standard Corpus of Present-Day Edited American English, or shortened Brown Corpus). At the same time, we will note that the exact date of the creation of the text in this work is of little importance to us in comparison with the indication of the age audience of the reader. Thus, according to our criteria, a folk tale recorded more than a hundred years ago but present in a modern textbook for the 3rd grade is suitable for inclusion in the corpus. The average number of words per page of a textbook depends

[1] github.com/prykhodchenkosd/ukrtb

on the format, font and text density. Still, it averages around 250-300 words for a standard A4 page or a textbook without many illustrations and formulas. Textbooks with a dense layout (e.g., high school textbooks) may have up to 400 words, and those with many graphics or examples (e.g., elementary school textbooks) may have about 200 words. To create this corpus, we used an average article length of 2 pages, i.e., an average of about 500 words on the topic of one lesson.

The data are collected by dividing the texts by the reader's age group. We identify five levels of text complexity: Level2, Level3, Level4, KS3, and GCSE, similar to the age categories of the WeeBit corpus (Vajjala and Meurers, 2012) KS3 (Key Stage 3, UK educational stage for ages 11–14) and GCSE (General Certificate of Secondary Education, typically for ages 14–16). In the first stage, we collected equal amounts of data for each group using semantically similar sources, prioritizing textbooks from the same authors or editors. Texts were then cleaned, encoded uniformly, and processed using tokenization, normalization, and statistical analysisAt the processing stage, the program calculates statistical indicators of the text and metrics, such as the number of letters in a sentence, the number of words in a sentence, the average number of letters in a word, the average number of words in a sentence, the average number of syllables in a word, etc. The primary metrics and statistical indicators are given in Table 1. In further studies of readability formulas, each level will appear under a specific number: 1 – Level2, 2 – Level3, 3 – Level4, 4 – KS3, and 5 – GCSE. This coding is present in Fig. 2-6 in the form of marks on the abscissa axis.

To evaluate statistical information about parts of speech, the work used automatic POS tagging, carried out using the pymorphy3[2] project. The original versions 1 and 2 of the pymorphy project can currently be considered abandonware; the development and support of the latest versions of Python is provided by its fork, pymorphy3. This project is based on the dictionaries of the OpenCorpora project, which are the basis for a significant number of scientific works in the field of text corpora processing (Korobov, 2015), (D. Kalugin-Balashov. 2023), (Tmienova and Sus, 2019), which allows us to talk about the quality and reliability of the assessments of this solution. The program using pymorphy3 produces a significant array of data on each of the words of the text, including the POS decision, which is issued in an abbreviated form[3].

Grade **Level**	Age in Years	Num. of articles	Avg. number of sentences per article	Avg. syllables in a word	Avg. words in a sentence
Level 2	7-8	150	24,17	2,27	10,6542
Level 3	8-9	150	31,16	2,34236	10,28945
Level 4	9-10	150	41,35	2,15672	13,10985
KS3	11-14	150	37,82	2,34837	12,9896
GCSE	14-16	150	27,91	2,51437	15,7587

Table 1: Primary metrics and statistical indicators

As a result of processing the corpus texts, categorizations of words by parts of speech were obtained for each text passage that makes up the corpus. The results are presented in Table 2.

5 Applicability of existing readability methods to Ukrainian texts

Classic works on calculating the readability coefficient are the works of Flesch, who defined a general readability formula and, later, Kincaid derived the grade-level readability formula, which determined the readability index depending on the level of education of the reader. Flesch's works were aimed at studying the English-language texts; they also served as a starting point for several similar studies of later times, which resulted in the methods for determining readability ARI, SMOG, and several others, which were also focused on English-language texts. Attempts to use these methods for other languages, even of the related Germanic language group, usually concluded with the need to create a method entirely adapted to the corresponding language (Pickelmann et al., 2023), (Blaneck et al., 2022), (Mohtaj et al., 2022).

For the languages of the Slavic group, such studies began to be conducted in the 1960s-70s, gaining recognition for such languages as Russian and Polish, for which studies of readability

[2] https://pypi.org/project/pymorphy3/

[3] https://github.com/no-plagiarism/pymorphy3/blob/master/docs/user/grammemes.rst

	Level 2	Level 3	Level 4	KS3	GCSE
Total	26276	36132	53548	59498	55109
NOUN	9957	13622	18357	23409	23623
ADJF	2115	3516	3915	6444	7880
VERB	4721	6274	9599	8996	6700
NPRO	2389	3285	5682	5219	3651
ADVB	1253	1640	2588	2688	2153
PRCL	1606	2064	4030	3329	2525
CONJ	2514	3278	5775	5016	4355
PREP	1479	2081	3005	3684	3617
GRND	99	171	252	355	270
NUMR	143	201	345	358	335

Table 2: Number of part-of-speech in the text corpus

assessment methods and software products based on these methods are still relevant. At the same time, evaluations of other reading methods for other Slavic languages, including Ukrainian, have either not been conducted at all or have not received sufficient publicity and recognition.

The Ukrainian language has the most remarkable lexical similarity with the Belarusian language and, to a lesser extent, with Russian and Polish. Studies of readability assessment methods for the Belarusian language have also not been conducted, so in this paper, we will try to apply methods applicable to Russian and Polish, as well as the Flesch and Flesch-Kincaid methods, which have become classics, for assessing the readability of Ukrainian-language texts collected in the UkrTB corpus. On the way to this part of the study, we manually checked the automatically performed morphological partitioning of the corpus presented in Table 2.

Flesch's work assumed the calculation of the readability coefficient based on such parameters as total words, sequences, and syllables. Its consideration is still encountered today when considering readability coefficients for languages different from English. Formula (1), proposed by Flesch, was calculated for the texts of the UkrTB corpus, resulting in the graph presented in Fig. 2. This graph shows each class's maximum, average, and minimum values.

$$F = 206.835 - 1.015\left(\frac{total\ words}{total\ sentences}\right) - \\ -84.6\left(\frac{total\ syllables}{total\ words}\right) \quad (1)$$

In 1975, Kincaid at al. improved Flesch's readability method (Kincaid et al., 1975) by including a proposed division by educational level, thus calibrating the readability coefficient by the level of the proposed reader. This method relies on the same parameters as in the original work - total words, sequences, and syllables, and Formula 2 is a mathematical description of this method. In this study, all corpus texts were also processed using this method, resulting in the graphs presented in Fig. 3, which also determined the maximum, minimum, and average coefficients for each class of readers. Although Figure 3 shows an upward trend in Flesch-Kincaid scores with increasing grade level, the within-group variance is substantial. Furthermore, the score ranges overlap significantly between neighboring levels, making it difficult to reliably distinguish readability tiers based on these scores alone. This explains the assessment of low correspondence, despite the overall trend.

$$FK = 0.39\left(\frac{total\ words}{total\ sentences}\right) + 11.8\left(\frac{total\ syllables}{total\ words}\right) - \\ -15.59 \quad (2)$$

Later, linguists developed readability methods for Slavic languages, as English-based formulas proved unsuitable. Matskovskii and Pisarek's formulas were chosen due to their relevance to Slavic structures. Pisarek's Polish method, adapted for inflected languages with complex morphology, is a strong candidate for Ukrainian. Matskovskii's formula, designed for Russian, incorporates syntactic and lexical complexity, offering insights into how readability metrics perform in another East Slavic language. Given Ukrainian's linguistic position between Polish and Russian, testing these models provides a comparative framework for assessing their applicability and identifying necessary modifications for a more accurate Ukrainian readability metric.

The author of the readability coefficient for the Polish language is prof. Pisarek (Pisarek, 1969) proposed calculating the linear (3) and nonlinear

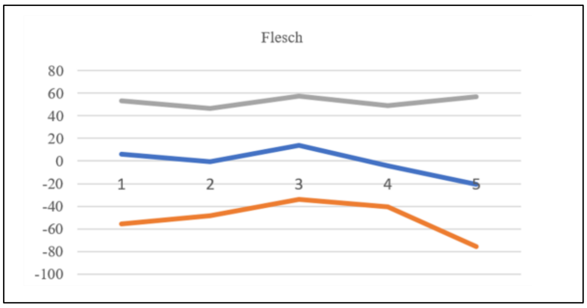

Figure 2: Classes of UkrTB by Flesch readability ease

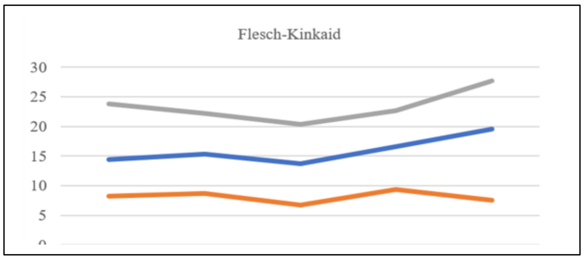

Figure 3: Classes of UkrTB by Flesch-Kinkaid readability ease

(4) dependencies for the readability coefficient of the Polish language based on ASL (average number of words per sentence) and PCW (percentage of complex words). His work in this field is still used to analyze Polish-language texts (Broda et al., 2014). In our implementation, complex words (PCW) were defined as those exceeding three syllables and not belonging to a predefined list of stopwords or functional parts of speech. This operationalization was chosen to approximate the lexical difficulty as perceived by native readers, following practices in Polish readability research.

$$P_l = \frac{1}{3} ASL \cdot \frac{1}{3} PCW + 1 \qquad (3)$$

$$P_{nl} = \frac{1}{2}\sqrt{ASL^2 + PCW^2} \qquad (4)$$

This study applied these techniques to UkrTB corpus. The results can be seen in Figure 4.

Some methods have been proposed for the Russian language, which differs significantly from each other in the time of creation and the number and composition of the studied text features. In this study, we considered the method (5) (Matskovskii, 1976), and based on such parameters as average sentence length and X_3 is the percentage of words of more than 3 syllables in the text.

$$M = 0.62ASL + 0.123X_3 + 0.051 \qquad (5)$$

We also considered a relatively recent work by M. Solnyshkina and co-authors (Solnyshkina, 2018), based on the analysis of ASW (average number of syllables per word), ASL (average number of words per sentence), UNAV (relation between the number of unique words in text: (number of unique Adjectives + number of unique Nouns)/(number of unique Verbs)) and NAV (relation between the number of words in text: (type-token ratio of Adjectives + type-token ratio of Nouns)/(type-token ratio of Verbs)).

$$\begin{gathered} S = -\ 0.124ASL\ +\ 0.018\ ASW\ -\ 0.007\ UNAV\ + \\ +\ 0.007\ NAV\ - 0.003\ ASL^2 +\ 0.184\ ASL\ ASW\ + \\ +\ 0.097\ ASL\ UNAV\ -\ 0.158\ ASL\ NAV\ + \\ +0.09\ ASW^2\ +\ 0.091\ ASW\ UNAV\ + \\ +\ 0.023\ ASW\ NAV\ -\ 0.157\ UNAV^2\ - \\ -0.079\ UNAV\ NAV\ +\ 0.058\ NAV^2 \qquad (6) \end{gathered}$$

The results of applying formulas 5 and 6 to the corpus texts can be seen in Fig. 5.

The overall figure comparing the applications of the above-described methods can be seen in Fig. 6. To facilitate the visual comparison of readability scores across methods with different scales and directionalities, we applied min-max normalization to each set of scores before plotting. The normalized scores are shown on a unified vertical axis, where higher values indicate higher perceived text complexity. This allows approximate comparison of method behavior across grade levels.

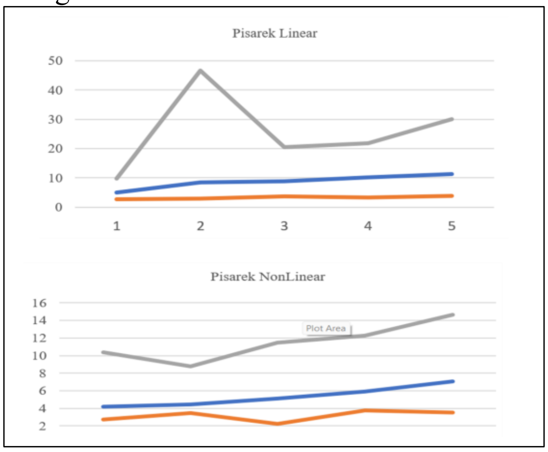

Figure 4: Classes of UkrTB by Pisarek linear and non-linear readability ease

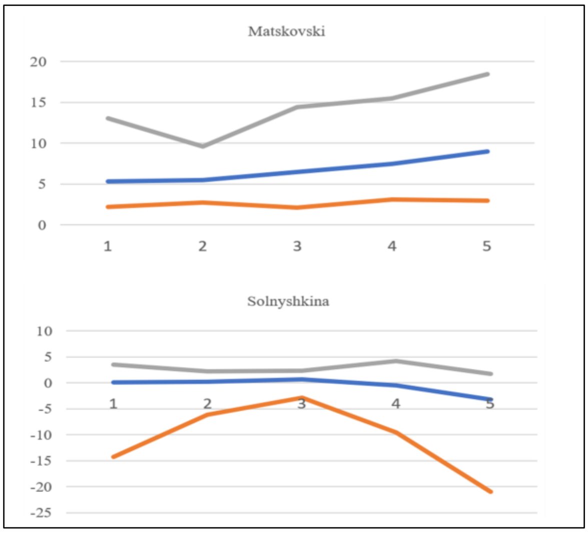

Figure 5: Classes of UkrTB by Matskovskii and Solnyshkina readability ease

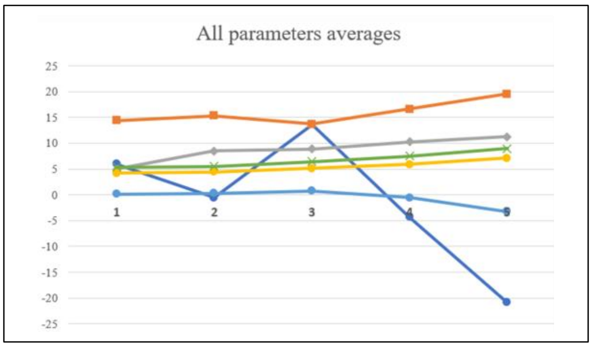

Figure 6: Classes of UkrTB by methods 1-6 of readability ease

The examination of the above-described methods applied to the texts of the UkrTB corpus showed the following results:

1. The Flesch and Flesch-Kincaid methods showed a low correspondence of the calculated readability coefficients to the expected change in the level of text complexity from low to high, which was initially assumed due to the initial calculation of the applicability of these methods for the English language.
2. Pisarek's methods, on average, showed a proportional increase in the expected complexity of texts. Still, the difference is not obvious when considering the range of values within classes, which is especially important for the linear model.
3. Among the considered models of the Russian language applied to the texts of the UkrTB corpus, Solnyshkina's model showed an average inversely proportional nonlinear dependence, with an initial directly proportional one, and Matskovskii's model - a directly proportional one. However, as in the case of Pisarek's models, the results of both models within a class show a wide range of values, which does not allow us to talk about sufficient accuracy in determining the level of readability for the Ukrainian language.

6 Conclusions

As a result of the research work, two main results were obtained, the first of which is the creation of a Ukrainian-language corpus of texts divided by the assumed levels of education into 5 classes.

The second result of this work is a series of tests of the applicability of existing methods for determining readability coefficients to corpus texts written in Ukrainian.

To summarize, some methods for determining readability in average results show dependencies between the level of readability and the assumed level of education. In some cases, this dependence is the opposite of what was declared by the creators of the method. Nevertheless, the spread of values obtained as a result of processing is such that in the presence of two texts of different classes, false identifications of the readability level are possible not only with the nearest class but - through a class, two, and in some cases - through three, i.e., at the other end of the readability level and the assumed level of education. Consequently, the methods proposed earlier for languages other than Ukrainian can be applied to Ukrainian texts only in limited cases, using additional coefficients, and for large average samples, which makes their real application difficult. For instance, one text designed for Level 2 (age 7–8), consisting of short sentences and simple vocabulary, was assessed by the Solnyshkina model as equivalent in complexity to GCSE-level texts. This highlights how non-adapted formula parameters may misinterpret language features that are not penalized in Ukrainian, such as long noun phrases or complex morphologies.

In general, this study shows that readability formulas such as Flesch, Flesch-Kincaid (for English), Pisarek (for Polish) and Matskowski and Solnyskina (for Russian) are not suitable for Ukrainian because of linguistic and statistical differences. Formulas like Flesch count the length of sentences and words, but their coefficients are customised for English. Readers' cultural expectations are also important: complex constructions in Ukrainian are perceived naturally, but existing formulas “penalise” them.

The Ukrainian language needs a new formula with empirical data and adapted parameters that consider specificity. Thus, as a result of these studies, it can be concluded that additional research is necessary to determine the dependence of the parameters of Ukrainian-language texts on the expected level of education and thus determine a specialized methodology for determining the readability coefficient for the Ukrainian language.

References

P. G. Blaneck, T. Bornheim, N. Grieger, and S. Bialonski. 2022. Automatic Readability Assessment of German Sentences with Transformer Ensembles. https://doi.org/10.48550/arXiv.2209.04299.

B. Broda, B. Nitoń, W. Gruszczyński, and M. Ogrodniczuk. 2014. Measuring Readability of Polish Texts: Baseline Experiments. In Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14), pages 573–580, Reykjavik, Iceland. European Language Resources Association (ELRA).

O. Cherednichenko and O. Kanishcheva. 2021. Readability Evaluation for Ukrainian Medicine Corpus (UKRMED). COLINS, pages 402–412.

O. Cherednichenko, O. Kanishcheva, and N. Babkova. 2018. Complex term identification for Ukrainian medical texts. CEUR Workshop Proceedings, 2255:146–154.

O. Cherednichenko, O. Kanishcheva, and O. Yakovleva. 2020. Collection and Processing of a Medical Corpus in Ukrainian. Computational Linguistics and Intelligent Systems, COLINS, CEUR Workshop Proceedings, 2604:272–282.

Rudolf Flesch. 1948. A new readability yardstick. Journal of Applied Psychology, 32(3):221–233. https://doi.org/10.1037/h0057532. PMID 18867058.

D. Kalugin-Balashov. 2023. Advancing Full-Text Search Lemmatization Techniques with Paradigm Retrieval from OpenCorpora. https://doi.org/10.48550/arXiv.2305.10848.

J. P. Kincaid, R. P. Fishburne, R. L. Rogers, and B. S. Chissom. 1975. Derivation of new readability formulas (automated readability index, fog count, and flesch reading ease formula) for Navy enlisted personnel. Research Branch Report 8–75. Chief of Naval Technical Training: Naval Air Station Memphis.

M. Korobov. 2015. Morphological Analyzer and Generator for Russian and Ukrainian Languages. In: M. Khachay, N. Konstantinova, A. Panchenko, D. Ignatov, and V. Labunets (Eds.), Analysis of Images, Social Networks and Texts. AIST 2015, Communications in Computer and Information Science, vol 542. Springer, Cham. https://doi.org/10.1007/978-3-319-26123-2_31.

Lu, Xiaofei. (2012). The Relationship of Lexical Richness to the Quality of ESL Learners' Oral Narratives. The Modern Language Journal. 96. 190-208. 10.2307/41684069.

M. S. Matskovskii. 1976. Problems of Readability of Printed Material In Semantic Perception of a Speech Message in Mass Communication, pages 126–142. Nauka, Moscow.

S. Mohtaj, B. Naderi, S. Möller, F. Maschhur, C. Wu, and M. Reinhard. 2022. A Transfer Learning Based Model for Text Readability Assessment in German. https://doi.org/10.48550/arXiv.2207.06265.

A. M. Krychkovska, Zh. P. Parashchyn, O. V. Shved, I. I. Gubytska, L. D. Bolibrukh, V. P. Novikov . Application of information technologies for standardization of the methodology of creating educational literature // Bulletin of the Lviv Polytechnic National University. Informatization of a higher educational institution. -2014. - No. 803. - S. 81-85.

F. Pickelmann, M. Färber, and A. Jatowt. 2023. Ablesbarkeitsmesser: A System for Assessing the Readability of German Text. https://doi.org/10.1007/978-3-031-28241-6_28.

W. Pisarek. 1969. Jak mierzyć zrozumiałość tekstu. Zeszyty Prasoznawcze, (4):35–48.

M. Solnyshkina, V. Ivanov, and V. Solovyev. 2018. Readability Formula for Russian Texts: A Modified Version. In 17th Mexican International Conference on Artificial Intelligence, MICAI 2018, Guadalajara, Mexico, October 22–27, 2018, Proceedings, Part II. https://doi.org/10.1007/978-3-030-04497-8_11.

V. Starko and N. Cheylytko. 2013. Parameterization of the corpus as a way to increase its representativeness and balance. Ukrainian Linguistics, 43:87–94.

N. Tmienova and B. Sus. 2019. System of Intellectual Ukrainian Language Processing. CEUR Workshop Proceedings, 2577:199–209.

M. Trokhymovych, I. Sen, and M. Gerlach. 2024. An Open Multilingual System for Scoring Readability of Wikipedia. https://doi.org/10.48550/arXiv.2406.01835.

Sowmya Vajjala and Detmar Meurers. 2012. On improving the accuracy of readability classification using insights from second language acquisition. In Proceedings of the Seventh Workshop on Building Educational Applications Using NLP, pages 163–173. ACL.

Mingqiang Xia, Ekaterina Kochmar, and Ted Briscoe. 2016. Text Readability Assessment for Second Language Learners. Pages 12–22. https://doi.org/10.18653/v1/W16-0502.