Computer Assisted Methods in Engineering and Science
32(3): 271–292, 2025, https://doi.org/10.24423/cames.2025.1657

Matrix Similarity Analysis of Texts Written in Belarusian and Ukrainian

Artur NIEWIAROWSKI*, Anna PLICHTA

Department of Computer Science, Faculty of Computer Science and Telecommunications, Cracow University of Technology, Kraków, Poland

* *Corresponding Author e-mail: artur.niewiarowski@pk.edu.pl*

This publication presents the results of a study on text similarity between Belarusian and Ukrainian, utilizing a matrix-based analysis method grounded in edit distance. A distinctive feature of this approach is the absence of language-specific vocabulary rules, highlighting the algorithm's linguistic universality in similarity analysis. The analyzed texts were sourced from excerpts of online encyclopedias, translated using AI-powered online translation services provided by well-known companies. The primary objective of this study is to determine whether it is possible to compare texts written in these languages without prior translation into a common language. Additionally, it aims to assess whether a method that does not belong to the large language model (LLM) family or the broader category of AI-based approaches can effectively compare languages within the same linguistic group. Furthermore, the study provides insights into the degree of similarity between Belarusian and Ukrainian, investigating the extent to which speakers of one language might partially understand the other.

Keywords: text-mining, anti-plagiarism, text similarity analysis, Levenshtein edit distance, matrix-based text analysis, Belarusian language, Ukrainian language, East Slavic language group, Old Russian language, Indo-European language, *беларуская мова*, *Bielaruskaja mova*, *украї́нська мо́ва*, *Ukrainska mova.*

1. Introduction

The following languages, written in the variants of Cyrillic script, are used in European countries: Russian in the Russian Federation (European Russia), Ukrainian in Ukraine, Belarusian in Belarus, Bulgarian in Bulgaria, Macedonian in North Macedonia, and Bosnian-Croatian-Montenegrin-Serbian (BCMS), which in Bosnia-Herzegovina, Serbia, and Montenegro is officially written in both the Latin and a modified Cyrillic scripts while Croatia uses exclusively the Latin script.

This article focuses on investigating whether texts written in two closely related languages – Belarusian and Ukrainian – can be compared in terms of text reuse or text overlap without prior translation into a common language. The goal is to verify whether this is possible using a method based on edit distance, combined with a matrix constructed from the alignment of two texts. This matrix reflects the relationships between individual characters in the compared texts. Belarusian and Ukrainian were selected for this study due to their linguistic proximity, lexical similarities, and shared influences from the Russian language.

The proposed method does not rely on models such as BERT or other transformer-based architectures; it requires neither tokenization nor the training of complex models. This makes it possible to apply the approach on standard computer hardware without significant computational requirements.

Belarusian (also known as *беларуская мова*, *Bielaruskaja mova*) and Ukrainian (*украї́нська мо́ва*, *Ukrainska mova*), the official language of Ukraine [5, 23], are Indo-European languages belonging to the East Slavic group, which also includes Russian [5, 21]. They are primarily spoken in the territories of present-day Belarus, Russia, Ukraine, Latvia, Lithuania and Poland, and both evolved gradually from Proto-Slavic, a language spoken by Slavic peoples approximately 1500 years ago. Although many features common to East Slavic languages were established around the 9th to 10th centuries, the process of forming distinct linguistic features in Belarusian and Ukrainian as separate languages continued until at least the 14th to 15th centuries. From the 14th century onward, the broader region where these languages were evolving became part of Poland and Lithuania, which were united in 1569 as the Polish-Lithuanian Commonwealth. Only four hundred years later, following the economic and political collapse of that Commonwealth and its eventual partition by Russia, Prussia and Austria in the 18th century, most speakers of Belarusian and Ukrainian became citizens of the Russian Empire (and later Soviet Russia). As a result, both languages were first significantly influenced by Polish, while the impact of Russian, though substantial, is more recent and dates back mainly to the period after ca. 1700/1800 (depending on the region).

The Belarusian and Ukrainian languages share many common features, but there are several key differences between them in terms of phonetics, grammar, and lexis. Here are some of them:

- Phonetics – apart from having different phonetic inventories, one of the key distinctions between Belarusian and Ukrainian lies in phonotactics, as Belarusian very clearly differentiates between the pronunciation of stressed and unstressed vowels, merging all unstressed ‘a’ and ‘o’ sounds into characteristic ‘a’, which is consistently marked in spelling. A similar phenomenon occurs in Russian, where unstressed vowels are turned into weak 'e' or 'i' sounds, however, in Russian, this change is not marked in spelling.

- Grammar and morphology – there are some notable differences in the applied inflectional endings, as well as in the formation and use of tenses, etc. When it comes to morphology, Ukrainian regularly employs consonant and vowel mutations of the stem in the noun declension (with the most characteristic change of 'i' to 'o' or 'e', depending on the syllable structure). This feature is much more limited in Belarusian.
- Lexis – many words are different in the two languages, although they are usually mutually intelligible to speakers of both.
- Alphabet – both languages use modified Cyrillic script. Ukrainian has several unique letters not found in Belarusian, such as 'г', 'ї', 'є', and 'й' (the letter 'й' functions as a vowel in Belarusian, whereas in Ukrainian, it is treated as a consonant). In contrast, Belarusian has a letter: 'Ў' ('u' with a breve).
- Dialects – Ukrainian has a number of dialects, which vary considerably between different regions of Ukraine. Belarusian, on the other hand, is more uniform, although there are some regional differences within Belarus.

A substantial part of this article overlaps with Chapter 4.3 of the PhD dissertation of one of the authors [1], as the doctoral research was conducted in parallel with the preparation of this paper due to the extended duration of the review process.

2. Description of the problem and objectives

The two languages addressed in this publication are closely related and mutually intelligible to some extent, yet they each possess unique features and differences. Both languages belong to the broader Slavic language group [28]. The research presented in this publication attempts to determine whether an effective comparative analysis can be conducted between texts written in Belarusian and Ukrainian using a computer algorithm that does not incorporate grammatical rules specific to East Slavic languages, This includes the absence of implemented stemming and lemmatization methods [10–12, 26, 29].

In addition to stemming and lemmatization, there are more advanced approaches to text comparison, including word embeddings [23] such as Word2Vec, GloVe, and BERT, as well as LLMs [24] such as GPT-3/4 and mBERT. Furthermore, techniques based on embedding distance calculations, such as word mover's distance (WMD) [25], allow for measuring similarity between texts at deep semantic level. However, these methods require significant computational resources and produce large trained datasets, often reaching hundreds of gigabytes in storage size.

In contrast, the method presented in this study is relatively simple and computationally efficient, as it does not rely on resource-intensive machine learning models or pre-trained embeddings. This approach makes it more accessible for practical applications, as it does not require high-performance computing infrastructure. Despite its simplicity, the proposed method demonstrates that it is possible to carry out an effective comparative analysis between texts written in Belarusian and Ukrainian.

In this study, the term "effective comparative analysis between texts" refers to a quantitative approach based on word similarity, measured using the Levenshtein distance algorithm. The analysis does not focus on semantic or stylistic aspects but primarily on syntactic similarity, as it compares texts by assessing how similar individual words are at the character level. The method does not incorporate language-specific grammar rules, but instead relies on a binary similarity matrix, where a match between words is marked as 1 and a non-match as 0. This approach is particularly useful for detecting text reuse or potential plagiarism, as it highlights patterns of lexical similarity between documents. However, it does not aim to analyze language typology, as it does not account for structural differences between languages beyond the direct comparison of word forms.

3. Presentation of the applied method

The concept of matrix analysis of text data based on Levenshtein's edit distance [8] is described in detail in [2]. Levenshtein's distance has a variety of applications, including DNS analysis [14] and spelling error correction [13]. This section introduces the algorithm's general concept, with texts written in Spanish and Romanian used to illustrate it in more detail [7, 9]. Detailed comparative analyses of these two languages, which belong to the Romance language group [6] are included in [20] and one such analysis was also posted as a video [A20].

3.1. Presentation of the algorithm for analyzing text data

The algorithm is based on the idea of constructing a matrix $\mathbf{M}$ from two analyzed documents whose matrix's dimensions are equal to the number of words in each document. The matrix is populated with binary values (1 or 0) depending on whether the similarity value calculated between the words at each iterative step, based on the Levenshtein distance [8], falls within a given range set as one of the parameters by the user, see formula (1):

$$\begin{gathered}\mathop{I}\limits_{ti=1}^{tn}\ \mathop{I}\limits_{tj=1}^{tm}\ \mathbf{M}\,[ti,\, tj] = \beta,\\ \begin{cases} \beta = 1 : fp(\mathbf{Doc1}\,[ti],\, \mathbf{Doc2}\,[tj]) \geq bp,\\ \beta = 0 : fp(\mathbf{Doc1}\,[ti],\, \mathbf{Doc2}\,[tj]) < bp,\end{cases}\end{gathered} \tag{1}$$

where **M** – a document matrix of size *tn* by *tm*, formed based on two documents: **Doc1** and **Doc2**, *tn* – the number of terms in **Doc1**, *ti* – the index number of a given term in **Doc1**, Doc1 [*ti*] – the term (element) of **Doc1**, separated by a space from a next element in the document, *fp* – a function that returns the similarity measure *p*, as defined in formula (2), *bp* – the acceptable boundary value of similarity measure parameterized by the user, e.g., corresponds to document type (e.g., scientific vocabulary) or document language – setting the appropriate value of this parameter must be preceded by prior analyses of the given languages, language groups, and document types, which is also addressed among others, in this article, and β – a value of either 0 or 1.

The similarity measure *p* is calculated by the following formula:

$$p = 1 - \left(\frac{k}{k_{\max}}\right); \quad k_{\max} = \max\,(n,\, m), \quad \begin{gathered} k \geq 0,\, m > 0,\, n > 0,\\ p \in \langle 0,\, 1\rangle,\end{gathered} \tag{2}$$

where *m*, *n* – the lengths of the two terms/text strings (i.e., the number of characters), *k* – the Levenshtein distance between the two terms/text strings, and $k_{\max}$– the length of the longer of the two analyzed terms/text strings.

Each language has its own specific characteristics, including average word length, grammatical inflection, plural formation, and word distribution within sentences. These differences influence how text similarity can be analyzed across languages and which parameters should be taken into account. The parameterization of the method for a given language or language group involves identifying a universal set of values that are well-suited to the linguistic characteristics of that language, enabling effective text similarity analysis. This means that these values should provide reliable results for a broad range of texts in the given language, though they may not always be optimal for every possible case.

Therefore, while adjusting parameters for a specific language improves the method's precision, there will always be a margin of inaccuracy due to the diversity of linguistic structures and text variations. Consequently, parameter selection should be based on the analysis of representative data and experiments to determine the optimal values for a given language or group of languages.

The example to illustrate the method, is based on two different languages from the Romance group. A detailed description of the method is presented in [2]. A graphical visualization (matrix **M** described in formula (1)) comparing two short texts written in Spanish (Doc1) and Romanian (Doc2) [5, 6] is

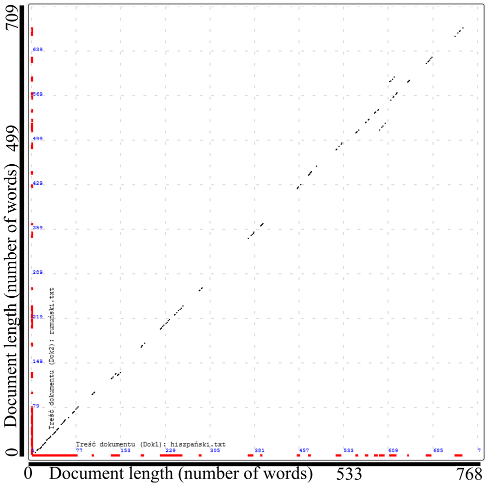

Fig. 1. Example of the result in the form of a graphical matrix **M** analysis of two texts written in languages belonging to the Romance language group.

shown in Fig. 1. The texts, analyzed below, were sourced from a Wikipedia article about Spain written in Polish. The text was then translated into both Spanish and Romanian using Google Translate. In the matrix, individual points represent locations with logical values of '1', indicating, fragments that are similar between the text strings. Excerpt from the Spanish text: *España, también denominado Reino de España, es un país soberano transcontinental, miembro de la Unión Europea, constituido en Estado social y democrático de derecho, cuya forma de gobierno es la monarquía parlamentaria. Su territorio, con capital en Madrid, está organizado en diecisiete comunidades autónomas, formadas a su vez por cincuenta provincias; y dos ciudades autónomas [...]* [A1]. Excerpt from the text in Romanian: *Spania, cunoscută și sub denumirea de Regatul Spaniei, este o țară suverană transcontinentală, membră a Uniunii Europene, constituită ca stat social și democratic de drept, a cărui formă de guvernare este monarhia parlamentară. Teritoriul său, cu capitala la Madrid, este organizat în șaptesprezece comunități autonome, formate la rândul lor din cincizeci de provincii; și două orașe autonome [...]* [A2].

Particularly noteworthy is the curve that forms in Fig. 1. It reflects the sequence of similar words occurring in succession in the analyzed texts. The fewer the gaps between the words, the greater the similarity. This curve closely corresponds to the data displayed in Table 1. If the points deviate from the main diagonal, it suggests that, for example, a sequence of words appearing consecutively in one text document is shifted relative to the other – appearing earlier or later depending on whether the point is above or below the diagonal. Another reason for point deviation could be the repetition of a specific sequence of text that appears in a similar form in the other document.

Table 1. Examples of text excerpts considered similar.

ID	Spanish language	Romanian language
1	[...] En Europa, ocupa la mayor parte de la península ibérica, conocida como España [...].	[...] În Europa, ocupă cea mai mare parte a Peninsulei Iberice, cunoscută drept Spania [...].
2	[...] de facto del G. La primera presencia constatada de homínidos del género Homo se remonta a millones de años antes del presente, como atestigua el descubrimiento [...].	[...] de facto membră a G. Prima prezență confirmată a hominidelor din genul Homo datează cu milioane de ani înainte de prezent, fapt dovedit de descoperirea [...].
3	[...] monarcas españoles dominaron el primer imperio de ultramar global, que abarcaba territorios en los cinco continentes, nota dejando un vasto acervo cultural y lingüístico por el globo. [...]	[...] monarhii spanioli dominau primul imperiu global de peste mări, care cuprindea teritorii de pe cinci continente, nota lăsând o vastă moștenire culturală și lingvistică pe tot globul. [...]

The points on the chart, which represent the positions of identical words between documents, have been filtered to enhance readability. This filtering highlights sequences of matching words against a background of seemingly random points. The red color along the axes indicates sections where the documents share similar or identical text.

The additional parameters introduced in this study for calculating document overlap are as follows:

- Maximum allowable gap between words to maintain text continuity (denoted as gw in the following sections).
- Minimum required number of words to construct a text continuity vector (denoted as wv in the following sections).

As part of this parameterization, the concept of text continuity has been introduced. Text continuity can be defined using the following formula (3):

$$\mathbf{T_j} = \left(t_{j,i},\, t_{j,i+1},\, ...,\, t_{j,n-1},\, t_{j,n}\right), \qquad i \in \langle 1,\, n_j \rangle\,, \tag{3}$$

where $\mathbf{T}$ – the word continuity vector (a fragment of text in one document that overlaps with a corresponding fragment in the other document), j – the vector number in the results of the text comparison, i – a natural number indicating the position of a word within the vector, n_j – a natural number representing the total number of similar and dissimilar words in the j-th vector and, $t_{j,i}$ – a Boolean value indicating the similarity of specific terms between documents.

The word continuity vector was introduced in order to filter noise from the graph, i.e., irrelevant fragments that may have been mistakenly identified as similar. The maximum length of the $\mathbf{T}$ vector is not user-parameterized. Its upper limit could, for example, be equal to the size of the document, i.e., the total number of words in it. This occurs when the analyzed documents are identical. The $\mathbf{T}$ vector consists of Boolean values (1 – similar word, 0 – dissimilar word, i.e., true or false). The minimum size of the vector is determined by the wv parameter and is adjusted by the user based on the characteristics of the analyzed text (e.g., the language in which it is written). The wv parameter includes words that are similar and appear consecutively or with a maximum allowable gap *gw* between them (i.e., dissimilar words). The smaller the gap defined by *gw* and the larger the *wv* parameter, the stricter the constraints imposed on the text comparison analysis. In other words, text sequences must be more precise to be considered identical. The following example (Figs. 2 and 3) serves as a visual complement to the explanation above, illustrating the structure and functioning of the continuity vector **T**. The parameter *bp* is not relevant for the following explanation of the introduction of the continuity vector **T**.

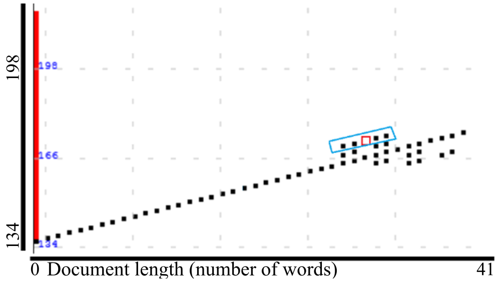

Fig. 2. Analysis parameters – *wv*: 4, *gw*: 1.

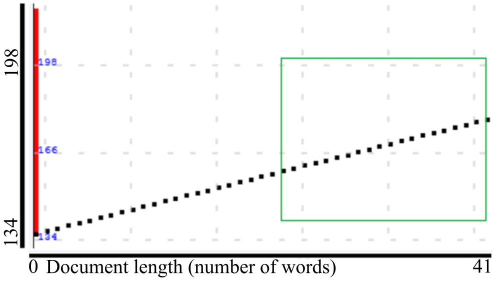

Fig. 3. Analysis parameters – *wv*: 5, *gw*: 1.

The points outlined in blue in Fig. 2 indicate the words that form one of the word continuity vectors **T**. According to the configured analysis parameters, these points remain in the similarity result because there are at least four of them (as defined by the *wv* parameter), and the maximum allowable gap between them (*gw* parameter) is set to one word. If the *wv* parameter were increased by just one additional required word, or, if the allowable gap *gw* were reduced to zero, the result would be more denoised, as shown in Fig. 3. In that case, the words represented by the removed points in Fig. 2 would no longer be included in the similarity outcome.

3.2. Tool for data analysis

The tool used in the research is a program called N-DMS Antyplagius [27], which operates based on the matrix analysis of text data concept presented earlier. The computer application was developed by one of the authors of this paper and it is one of the results of scientific research on algorithms from the text-mining field [2–4]. It performs similarity analysis between text data and does not contain implemented language-dependent vocabulary rules. The program analyzes documents written in languages of European origin, using both Latin and Cyrillic alphabets (with the option of automatic transcription), as well as Chinese characters. It allows users to customize analysis parameters for documents under study, including, among others, the degree of word similarity and the size of breaks in sentence continuity. In addition, it has a predefined set of parameters tailored to specific document types of documents studied, i.e., papers, homework, theses, journal articles and, books, including mixed content, and languages (Belarusian, Bulgarian, Chinese, Czech, Danish, Finnish, French, German, Italian, Dutch, Norwegian, Polish, Portuguese, Russian, Romanian, Slovak, Swedish, Taiwanese, Ukrainian). The application also features a built-in OCR module (based on the world-famous Tesseract OCR Engine), which enables text recognition from images [15] and perfectly complements the text similarity method based on edit distance. Addressing and correcting shortcomings of the OCR mechanism. The program is resistant to misrepresentation in the form of character substitutions, spelling, and grammatical errors, as well as occasional word substitutions. The analysis results include text fragments considered similar and a diagram of the relationship between documents.

3.3. Data analysis and results

The analyzed text strings were sourced from online encyclopedia articles from two well-known encyclopedias. They were translated using AI-powered machine translation systems, which are now recognized for their high effectiveness (translations were carried out using Google [30] and Microsoft [31] translators). Each system relies on distinct models and language processing methods, enabling more diverse results. This approach ensures that the analysis is not limited to a single translation algorithm but instead incorporates different methodologies applied in practice. As a result, it provides a more objective assessment of text similarity and reduces the risk that findings are merely artifacts of a specific translation system. Three tests are described below. The first test (3.3.1) involves adapting one language to another by translating the former and then analyzing their similarity. The second test (3.3.2) begins by translating an article written in English into the two languages under study. These two approaches are a reference to the research that is taking place on issues related to cross-language plagiarism, a growing issue in schools and universities around the world [18, 19]. The third test (3.3.3) involves analyzing the similarity between texts written in the two languages examined in this study (translated from English) and several selected Cyrillic-language texts.

A visualization of this type of analysis is available as a video on a YouTube channel [A3], demonstrating the steps performed in the program to obtain a text similarity result. The analysis focuses on two encyclopedic articles about the Polish Academy of Sciences, written in the studied languages and not translated using machine translators.

3.3.1. Encyclopedia article on Belarus. This analysis uses an encyclopedic article about Belarus [A4] sourced from an online encyclopedia (the website address of the encyclopedia site is available in reference [A5]). The article, originally written in Belarusian, was translated into Ukrainian by a machine translator [A6, A7]. The two texts were then compared. A graphical interpretation of the compared texts is given below. The *gw* constant remains consistent across all tests in this section, as the specifics of the problem do not require constant adjustment of this parameter. The *gw* constant is applied in analyses of texts where there is a significant likelihood of content misrepresentation by deliberately changing the structure of sentences, including word reordering, word deletion, and substitution with synonyms. The value of *gw* was selected from previous studies of texts written in different languages within the same language family.

Result (1) represents the percentage of text similarity with respect to Document (2), while Result (2) represents the percentage of text similarity with respect to Document (1). The reported similarity is calculated as the ratio of words in the **T** vector identified as similar to the total number of words in the given document.

The above results should be considered primarily in relation to the thesis posed in Sec. 2, addressing the tackled issue. From the graphs above, it can be seen that regardless of the comparison parameters applied (chosen sensibly, of course, and within the limits of analytical correctness), the texts show considerable similarity and, in some cases, near identity – this is evidenced by the line running diagonally across the matrix. It can be seen in the figures and the table that the best result showing similarity between the texts in Belarusian and Ukrainian is obtained by setting the word similarity *bp* at a level higher than 70%, while the minimum number of words in the sequence vector *wv* and the maximum permissible gap between words *gw* remain unchanged at 5 and 8,

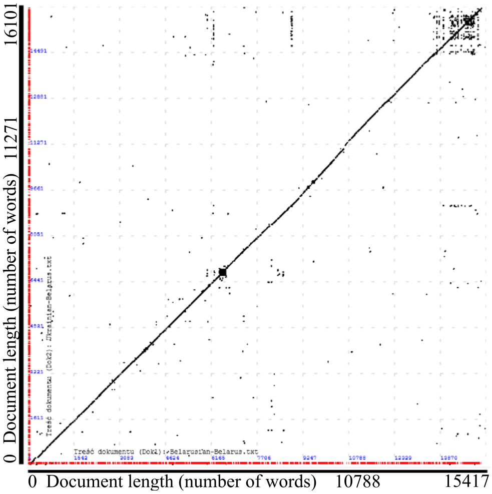

FIG. 4. Analysis parameters – *bp*: 50%, *wv*: 5, *gw*: 8.

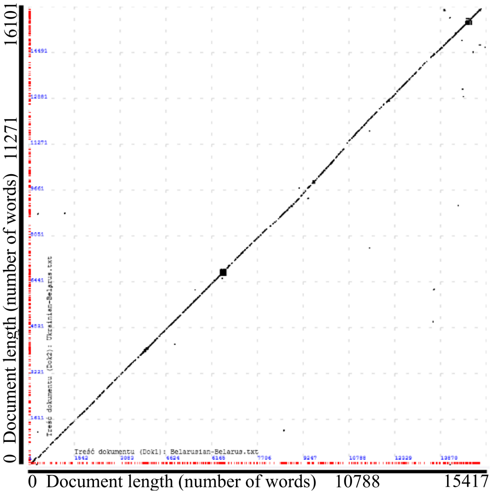

FIG. 5. Analysis parameters – *bp*: 70%, *wv*: 5, *gw*: 8.

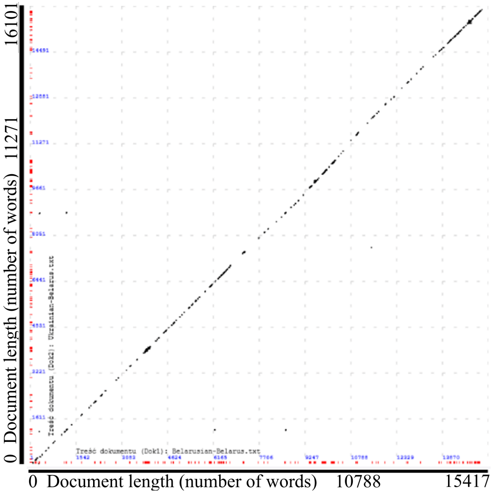

FIG. 6. Analysis parameters – *bp*: 90%, *wv*: 5, *gw*: 8.

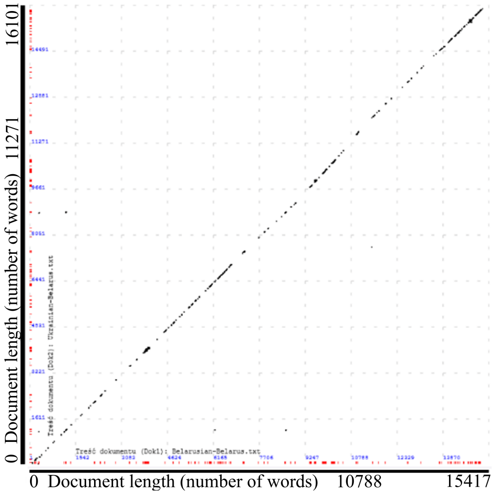

FIG. 7. Analysis parameters – *bp*: 100%, *wv*: 5, *gw*: 8.

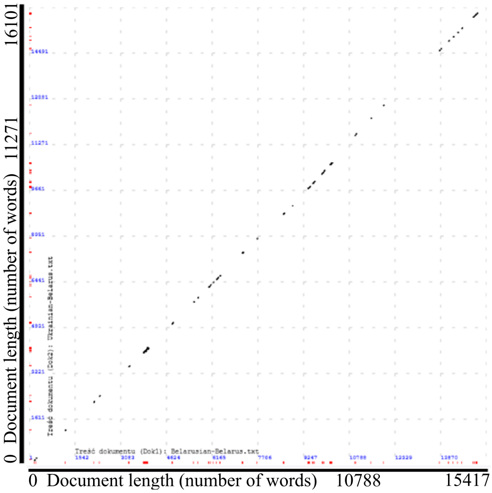

FIG. 8. Analysis parameters – *bp*: 100%, *wv*: 8, *gw*: 8.

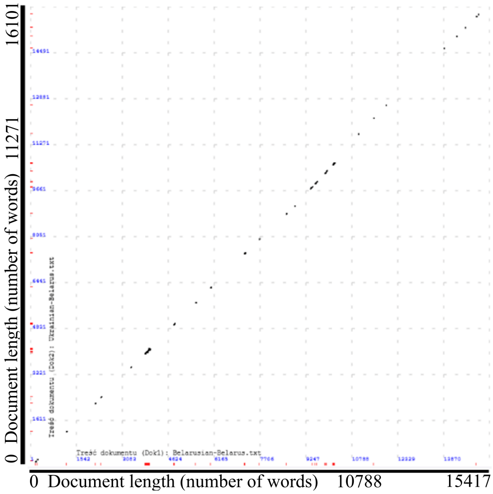

FIG. 9. Analysis parameters – *bp*: 100%, *wv*: 8, *gw*: 2.

respectively (Figs. 5 and 7). Setting the word similarity value *bp* below this threshold makes the graphical result of the comparison less clear, noise appears, and the diagonal line, which is responsible for visual confirmation of text similarity, becomes less visible (Fig. 4). Reducing the constant *bp* to 0% generates similarity score of 100% between the documents – which is an obvious error and is due to the principle of the algorithm – i.e., each word under study is considered identical to every other. Increasing the *wv* variable and decreasing the *gw* value decreases the similarity score; however, the result here is definitely not distorted (Figs. 8 and 9). In fact, the result is more accurate, as gaps in

Table 2. Results of multiple text comparison analyses, conducted with different analysis parameters.

ID	Document language (1)	Document language (2)	*bp* [%]	*wv*	*gw*	Number of words (number of characters) (1) × (2)	Result (1) [%]	Result (2) [%]
1	Belarusian	Ukrainian	42	5	8	15 417 × 16 101 (117 333 × 115 464)	59.71	56.49
2	Belarusian	Ukrainian	45	5	8	15 417 × 16 101 (117 333 × 115 464)	57.92	54.51
3	Belarusian	Ukrainian	50	5	8	15 417 × 16 101 (117 333 × 115 464)	55.82	52.62
4	Belarusian	Ukrainian	70	5	8	15 417 × 16 101 (117 333 × 115 464)	29.89	26.87
5	Belarusian	Ukrainian	90	5	8	15 417 × 16 101 (117 333 × 115 464)	12.59	9.9
6	Belarusian	Ukrainian	100	5	8	15 417 × 16 101 (117 333 × 115 464)	11.79	9.03
7	Belarusian	Ukrainian	100	8	8	15 417 × 16 101 (117 333 × 115 464)	6.08	4.51
8	Belarusian	Ukrainian	100	8	5	15 417 × 16 101 (117 333 × 115 464)	5.38	3.94
9	Belarusian	Ukrainian	100	8	2	15 417 × 16 101 (117 333 × 115 464)	4.4	3.11
10	Belarusian	Ukrainian	100	8	1	15 417 × 16 101 (117 333 × 115 464)	4.03	2.73

the occurrence of consecutive terms, forming a continuity vector that indicates a similar passage between documents, are either not taken into account or are taken into account to a lesser extent. These gaps are to be understood as other words placed between terms or a change in word order, resulting primarily from

Table 3. Examples of text passages considered similar in the ID 5 analysis from Table 2.

ID	Belarusian language	Ukrainian language
1	[...] 19 верасня 1991 гады краіна стала называца Рэспубліка Беларусь, у гэты ж час былі прынятыя новыя герб і сцяг заменены на сучасныя герб і сцяг 7 чэрвеня 1995 [...].	[...] On 19 вересня 1991 року країна стала називатися Республіка Білорусь, тоді ж були прийняті нові герб і прапор замінений на сучасні герб і прапор 7 червня 1995 [...].
2	[...] XVI ст. З 1620х гадоў назва замацавалася за ўсходнімі землямі Вялікага Княства Літоўскага - падзвінскімі і падняпроўскімі паветамі. На думку [...].	[...] XVI ст. З 1620х років назва закріпилася за східними землями Великого князівства Литовського - Подвинським і Наддніпрянським і Наддніпрянським повітами. На думку [...].
3	[...] і мае тытул "Масква сталіца ўсёй Белай Русііі" Moscovia urbs metropolis tutius Russiæ Albæ. План горада павёрнуты на 90 градусаў поўнач - справа, зверху - захад Карта "Вялікае Княства Маскоўскае ці Царства Белай Русі паводле апошніх паведамленняў" Estats du Grandduc de Moscovie ou de l'Empereur de la Russie Blanche suivant les derniers relations, каля 1749 г. Картограф Гендрык дэ Лет Нідэрланды Карта [...].	[...] і має назву "Москва, столиця всієї Білої Русі" Moscovia urbs metropolis tutius Russiæ Albæ . План міста повернуто на 90 градусів справа - північ, зверху - захід Карта "Вялікае Княства Маскоўскае ці Царства Белай Русі паводле апошніх паведамленняў" Estats du Grandduc de Moscovie ou de l'Empereur de la Russie Blanche suivant les derniers relations, каля 1749 г. Картограф Гендрык дэ Лет Нідэрланды Карта [...].

differences between languages. Therefore, due to the fact that one compares two different languages, the *wv* and *gw* variables, in most analyses, were set according to the given values and were not changed.

Table 3 contains selected passages considered similar, resulting from the analysis of the text comparison. Each of the above words considered similar to its counterpart in the other text is visually represented as a point on the matrix.

3.3.2. Encyclopedia article on Ukraine. The above analysis juxtaposes two excerpts from articles about Ukraine [A8], obtained from a different online encyclopedia [A9] (the address of the encyclopedia's website is available at [A10]) than those used in previous analyses. The English text was translated by machine translation systems from two different providers into Belarusian [30] and Ukrainian [31]. The parameters used in this comparison analysis are analogous to those in the previous section. A graphical interpretation of the comparison is presented in Figs. 10–13.

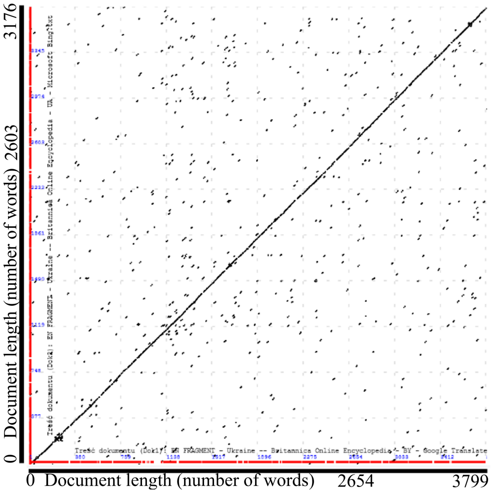

FIG. 10. Analysis parameters – *bp*: 30%, *wv*: 5, *gw*: 8.

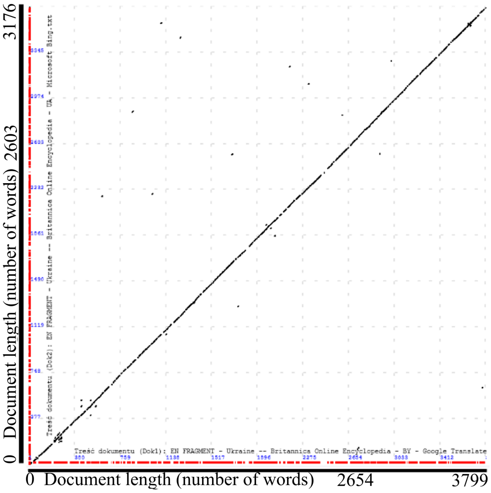

FIG. 11. Analysis parameters – *bp*: 42%, *wv*: 5, *gw*: 8.

As in the previous test, the best-fitting parameters for analyzing these two languages are a *bp* word similarity above 50%, but no more than 70% (Fig. 12). Within this range, noise is reduced and the curve is more clearly defined.

Raising the word similarity *bp* to 100%, that is, making words identical by force, results in a 0% similarity; in such a case, the texts are considered dissimilar (Fig. 13).

Table 5 contains selected text passages identified as similar by the algorithm. The data is the result of the analysis based on the parameters of ID 3 from Table 4. It can be seen that there is a clear similarity between texts, even for very

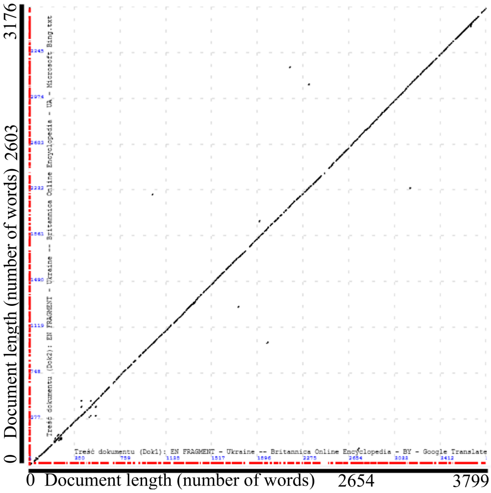

Fig. 12. Analysis parameters – *bp*: 50%, *wv*: 5, *gw*: 8.

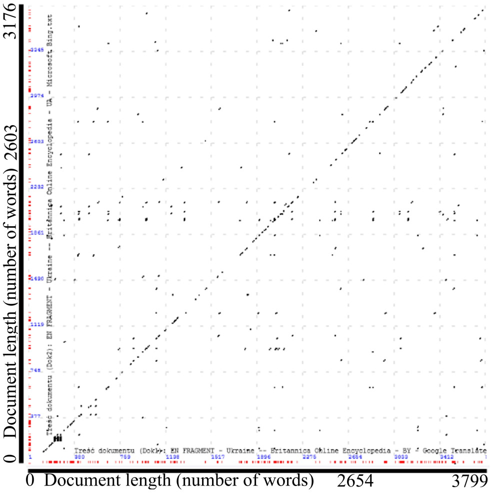

Fig. 13. Analysis parameters – *bp*: 100%, *wv*: 3, *gw*: 8.

Table 4. Results of multiple text comparison analyses, conducted with different analysis parameters.

ID	Document language (1)	Document language (2)	*bp* [%]	*wv*	*gw*	Number of words (number of characters) (1) × (2)	Result (1) [%]	Result (2) [%]
1	Belarusian	Ukrainian	42	5	8	3799 × 3716 (28 616 × 28 633)	55.41	56.38
2	Belarusian	Ukrainian	45	5	8	3799 × 3716 (28 616 × 28 633)	52.88	53.90
3	Belarusian	Ukrainian	50	5	8	3799 × 3716 (28 616 × 28 633)	51.22	52.10
4	Belarusian	Ukrainian	70	5	8	3799 × 3716 (28 616 × 28 633)	21.32	21.61
5	Belarusian	Ukrainian	90	5	8	3799 × 3716 (28 616 × 28 633)	3.05	3.12
6	Belarusian	Ukrainian	100	5	8	3799 × 3716 (28 616 × 28 633)	2.45	2.50
7	Belarusian	Ukrainian	100	8	8	3799 × 3716 (28 616 × 28 633)	0.47	0.48
8	Belarusian	Ukrainian	100	8	5	3799 × 3716 (28 616 × 28 633)	0	0
9	Belarusian	Ukrainian	100	8	2	3799 × 3716 (28 616 × 28 633)	0	0
10	Belarusian	Ukrainian	100	8	1	3799 × 3716 (28 616 × 28 633)	0	0

long words. The texts consist of both similar words (e.g., Table 5, ID 4: індустрыялізаваныя vs. індустріалізовані), as well as words that are completely different in form (e.g., Table 5, ID 4: гарады vs. міста), but meaning exactly the same in both languages under study. Appropriate parameter settings in the

TABLE 5. Examples of text passages considered similar in the ID 3 analysis from Table 4.

ID	Belarusian language	Ukrainian language
1	[...] Украіна, краіна, размешчаная ва ўсходняй Еўропе, другая па велічыні на кантыненце пасля Расііі. Сталіца - Кіеў, размешчаны на рацэ Днепр у паўночнацэнтральнайй [...].	[...] Україна, країна розташована на сході Європи, друга за величиною на континенті після Росії. Столицею є Київ, розташований на річці Дніпро в північноцентральній [...].
2	[...] Саюза як Украінская Савецкая Сацыялістычная Рэспубліка С. С. Р. Калі Савецкі Саюз пачаў распадаца ў 1990-91 гадах, заканадаўчая ўлада Украінскай ССР. абвясцілі суверэнітэт 16 ліпеня 1990 г., а затым поўную незалежнасць 24 жніўня 1991 г., крок, які быў пацверджаны народным адабрэннем на плебісцыце 1 снежня 1991 г. Пасля распаду СССР у снежні 1991 года Украіна атрымала поўную незалежнасць. Краіна змяніла сваю афіцыйную назву на Украіна, і гэта дапамагло заснаваць Садружнась Незалежных Дзяржаў СНД, абяднанне краін, якія раней былі рэспублікамі Савецкага Саюза. Клімат Украіна знаходзіца ва ўмераным кліматычным поясе, на які паступае ўмерана цёплае вільготнае паветра з Атлантычнага акіяна. Зімы на захадзе значна мякчэй, чым на ўсходзе. [...]	[...] Союзу як Українська Радянська Соціалістична Республіка СРСР. Коли Радянський Союз почав розпадатися в 1990-91 роках, законодавчий орган УРСР проголосив суверенітет 16 липня 1990, а потім повну незалежність 24 серпня 1991, крок, який був підтверджений народним схваленням на плебісциті 1 грудня 1991. З розпадом СРСР у грудні 1991 року Україна отримала повну незалежність. Країна змінила свою офіційну назву на Україна, і це допомогло заснувати Співдружність Незалежних Держав СНД, обєднання країня, які раніше були республіками Радянського Союзу. Клімат Україна лежить у помірному кліматичному поясі, на який впливає помірно тепле, вологе повітря з Атлантичного океану. Зими на заході значно мякші, ніж на сході. [...]
3	[...] сельскагаспадарчага рэгіёна займаюць ворныя землі лясы займаюць толькі каля 1/8 тэрыторыі. Далей на поўдзень, каля Чорнага, Азоўскага мораў і Крымскіх гор, лесастэп злучаецца са стэпавай зонай, плошча якой складае каля 89 000 квадратных міль 231 000 квадратных кіламетраў. Многія з плоскіх бязлесных раўнін у гэтым рэгіёне апрацоўваюца, хаця малая гадавая колькасць ападкаў і гарачае лета робяць неабходным дадатковае абрашэнне. [...]	[...] сільськогосподарського регіону займають орні землі Ліси займають лише близько восьмої частини площі. Далі на південь, біля Чорного моря, Азовського моря і Кримських гір, лісостеп приєднується до степової зони, площа якої становить близькоко 89 000 квадратних миль 231 000 квадратних км. Багато плоских, безлісих рівнин в цьому регіоні обробляються, хоча низька річна кількість опадів і спекотне літо роблять необхідним додаткове зрошення. [...]
4	[...] Запарожжа. Слаба індустрыялізаваныя гарады на захадзе, такія як Ужгарад і Хмяльніцкі, сутыкаюцца з забруджваннем паветра, выкліканым перавагай неэфектыўных аўтамабіляў. Асноўныя рэкі, у тым ліку Днепр, Днестр, Інгул і Данец, сурёзна забруджаныя хімічнымі ўгнаеннямі і пестыцыдамі [...].	[Запоріжжя. Слабо індустріалізовані міста на заході, такі як Ужгород та Хмельницький, стикаються із забрудненням повітря, спричиненим переважанням неефективних автомобілів. Великі річки, включаючи Дніпро, Дністер, Інгул і Донець, серйозно забруднені хімічними добривами і пестицидами [...].

analysis allow such words to be ignored in order to ensure continuity in the vector of similar words.

3.3.3. Encyclopedia article on Ukraine in other Cyrillic languages. For the sake of completeness and to aid in the interpretation of the results, this section presents comparative analyses of an excerpt [A11] taken from the above texts, written in Belarusian and Ukrainian (the translation performed using the

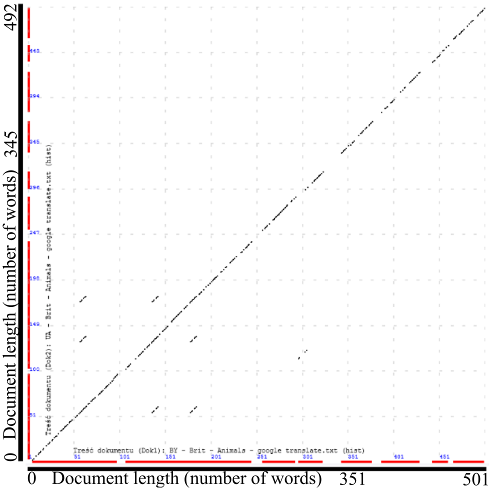

FIG. 14. Languages compared: Belarusian – Ukrainian.

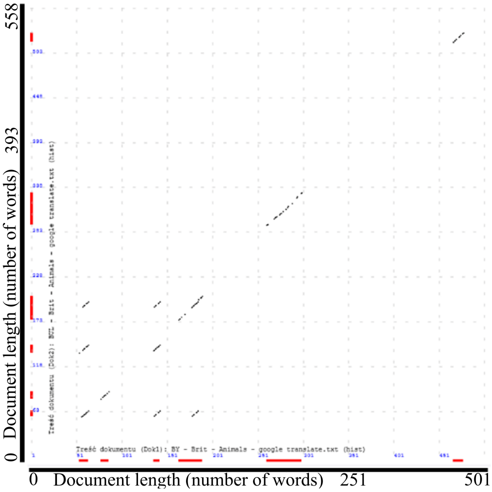

FIG. 15. Languages compared: Belarusian – Bulgarian.

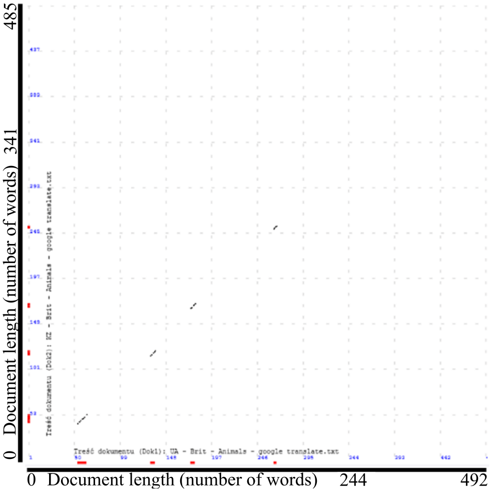

FIG. 16. Languages compared: Belarusian – Kazakh.

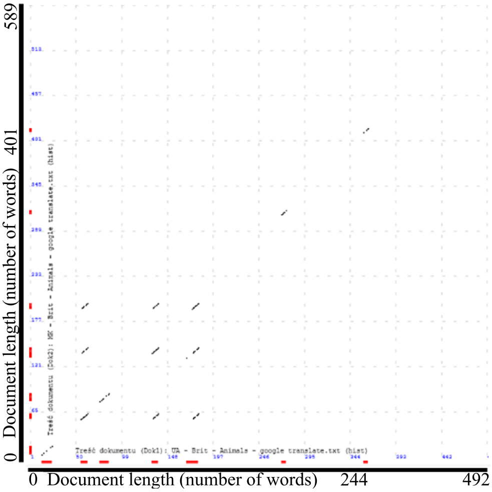

FIG. 17. Languages compared: Ukrainian – Macedonian.

Google tool [30]), in comparison with other Cyrillic languages (Figs. 14–17). (The Cyrillic script is an alphabet used by many languages, mainly in Eastern Europe and Central Asia. Here are some of the languages that use it: Russian, Ukrainian, Belarusian, Bulgarian, Serbian, Montenegrin, Macedonian, Kazakh, Kyrgyz, Tajik, Uzbek (sometimes interchangeable with the Latin script), Mongolian, Ossetian, Abkhazian, Bashkir, Chuvash, Komi, Mordvin, Tatar, Tuvan, and Yakut.) The languages considered are from both European and Asian countries. The analysis parameters are similar to those used in previous tests and are set as follows: *bp*: 50%, *wv*: 5, *gw*: 4–8.

Table 7 contains the results of the full similarity analysis for ID 5 from Table 6.

Table 6. Results of multiple text comparison analyses, between selected Cyrillic languages, using the same analysis parameters.

ID	Document language (1)	Document language (2)	*bp* [%]	*wv*	*gw*	Number of words (number of characters) (1) × (2)	Result (1) [%]	Result (2) [%]
1	Belarusian [A12]	Ukrainian [A13]	50	5	8	501 × 492 (3687 × 3611)	57.49	57.93
2	Belarusian	Bulgarian [A14]	50	5	8	501 × 558 (3687 × 3715)	13.37	12.01
3	Belarusian	Serbian [A15]	50	5	8	501 × 512 (3687 × 3444)	10.78	10.55
4	Belarusian	Macedonian [A16]	50	5	4	501 × 569 (3687 × 3815)	6.39	5.62
5	Belarusian	Kazakh [A17]	50	5	4	501 × 485 (3687 × 3729)	4.19	4.33
6	Ukrainian	Bulgarian	50	5	8	492 × 558 (3611 × 3715)	17.48	15.77
7	Ukrainian	Serbian	50	5	8	492 × 512 (3611 × 3444)	14.43	14.06
8	Ukrainian	Macedonian	50	5	4	492 × 569 (3611 × 3815)	9.35	8.08
9	Ukrainian	Kazakh	50	5	4	492 × 485 (3611 × 3729)	4.33	4.27

Table 7. Complete result of the comparative analysis of texts written in Belarusian and Kazakh.

ID	Belarusian language (1)	Kazakh language (2)	Text excerpt (1)	Text excerpt (2)
1	[...] - каля 44 000 квадратных міль 114 000 квадратных кіламетраў - [...].	[...] - шамамен 44 000 шаршы миль 114 000 шаршы км - [...].	55–65	45–55
2	[...] 78,000 квадратных міль 202,000 [...].	[...] 78,000 шаршы мильді 202,000 [...].	136–141	117–122
3	[...] 89,000 квадратных міль 231,000 [...].	[...] 89,000 шаршы мильді 231,000 [...].	178–183	167–172
4	[...] 6 міль 10 км [...].	[...] 6 миль 10 км [...].	269–272	251–254

As can be seen from the above analyses, the same texts written in languages that use Cyrillic do not automatically ensure a high degree of similarity between them. A large role is played here by the languages' belonging to particular language groups and their common history of linguistic evolution. In Table 7, one can see how little similarity some texts show, and the main factors that increase similarity values are numbers and measurement units. The occurrences of points on both sides of the diagonal result from the repetition of text sequences in different sections of the analyzed texts – primarily the numbers and measurement units mentioned earlier.

4. Applications

To sum up, it turns out that the accuracy of the analysis results depends primarily on the word similarity parameter *bp*. This parameter is responsible for determining, within the matrix concept of text analysis, whether the words analyzed in a given iteration step should be considered identical and, if so, filling the corresponding matrix cell with a positive value. Based on the above results, it can be seen that an algorithm, even without implementing grammar rules for a particular language, is able to correctly estimate the existing similarity between texts, despite the additional differences arising from the different languages involved.

This method, like any other, is not perfect. The calculation parameters should be adjusted according to the specific scenario, particularly the languages being analyzed. An additional aspect, in this case, could be the development of a supplementary method for selecting analysis parameters based on the analyzed text characteristics. However, even when the optimal parameters are not selected for the analysis of a specific set of textual data, the results may still strongly suggest misconduct in the form of plagiarism, or simply indicate a high degree of text similarity.

5. Summary and future work

The matrix text analysis algorithm based on Levenshtein's edit distance [8], confirmed the similarity of languages within the same language group, as described by linguists [5]. The algorithm does not use a thesaurus, so words with similar meanings but a large edit distance are not considered identical. However, this should not significantly affect the result of the text similarity analysis, since it is impossible to swap most words in a text document so that it still carries the same message while consisting of entirely different terms with similar meanings. And even if this were possible, it would be difficult in such a case to classify it as simple plagiarism of the text. However, a dictionary of closely related words could be an interesting factor to strengthen the algorithm, so this will be the subject of future research, especially in terms of optimizing the overall calculation process. In addition, the approach presented in this study will be used to analyze the similarity of essays created by the ChatGPT program (GPT – generative pre-trained transformer, https://chat.openai.com), which is currently being studied by researchers around the world and which is becoming an increasing ethical issue in academia [16, 17]. First steps in, this regard, have already been taken, and the results can be viewed at the following on-line resources: [A18] and [A19].

Conflict of interest

The author declares that they are the developer of the N-DMS Antyplagius software, which is a commercial solution. This fact did not influence the content or conclusions of this publication. The author's intent is not to promote the software but to present the method described in this study. Any reader interested in the software may contact the author by email to obtain the program and a license for its use.

References

1. A. Niewiarowski, *Zastosowanie algorytmu odległości edycyjnej do ilościowej analizy danych tekstowych* [in Polish], PhD dissertation, IPPT PAN, Warsaw, 2024.
2. A. Niewiarowski, Similarity detection based on document matrix model and edit distance algorithm, *Computer Assisted Methods in Engineering and Science*, **26**(3–4): 163–175, 2019, https://doi.org/10.24423/cames.277.
3. A. Niewiarowski, Short text similarity algorithm based on the edit distance and thesaurus, *Technical Transactions*, **113**(1-NP): 159–173, 2016, https://doi.org/10.4467/2353737XCT.16.149.5760.
4. A. Niewiarowski, M. Stanuszek, Parallelization of the Levenshtein distance algorithm, *Technical Transactions*, **111**(3-NP): 109–122, 2014, https://doi.org/10.4467/2353737XCT.14.319.3407.
5. K. Katzner, *The Languages of the World*, Taylor & Francis, London, 2002.
6. R. Posner, *The Romance Languages*, Cambridge Language Surveys, Cambridge University Press, Cambridge, 1996.
7. R. Penny, *A History of the Spanish Language*, Cambridge University Press, Cambridge, 2002.
8. V.I. Levenshtein, Binary codes for correcting dropouts, inserts, and symbol substitutions [in Russian], *Reports of the Academy of Sciences of the USSR*, **163**(4): 845–848, 1965.
9. P.R. Petrucci, *Slavic Features in the History of Rumanian*, LINCOM Europa, München, 1999.
10. A. Dziob, M. Piasecki, Implementation of the verb model in plWordNet 4.0, [in:] *Proceedings of the 9th Global Wordnet Conference, Singapore*, January 8–12, pp. 113–122, Nanyang Technological University, 2018.
11. W.B.A. Karaa, A new stemmer to improve information retrieval, *International Journal of Network Security & Its Applications*, **5**(4): 143–154, 2013, https://doi.org/ 10.5121/ijnsa.2013.5411.
12. D. Khyani *et al.*, An interpretation of lemmatization and stemming in natural language processing, *Journal of University of Shanghai for Science and Technology*, **22**(10): 350–357, 2021.
13. M.M. Maulana, R. Arifudin, A. Alamsyah, Autocomplete and spell checking Levenshtein distance algorithm for text suggestion error data searching in library, *Scientific Journal of Informatics*, **5**(1): 75, 2018, https://doi.org/10.15294/sji.v5i1.14148.

14. R. Gabrys, E. Yaakobi, O. Milenkovic, Codes in the Damerau distance for DNA storage, [in:] *2016 IEEE International Symposium on Information Theory (ISIT)*, Barcelona, Spain, pp. 2644–2648, 2016, https://doi.org/10.1109/ISIT.2016.7541778.

15. R. Smith, An overview of the Tesseract OCR engine, [in:] *Ninth International Conference on Document Analysis and Recognition (ICDAR 2007)*, Curitiba, Brazil, Vol. 2, pp. 629–633, 2007, https://doi.org/10.1109/ICDAR.2007.4376991.

16. B.D. Lund, T. Wang, Chatting about ChatGPT: How may AI and GPT impact academia and libraries?, *Library Hi Tech News*, **40**(3): 26–29, 2023, https://doi.org/10.1108/LHTN-01-2023-0009.

17. A.J. Adetayo, Artificial intelligence chatbots in academic libraries: The rise of ChatGPT, *Library Hi Tech News*, **40**(3): 18–21, 2023, https://doi.org/10.1108/LHTN-01-2023-0007.

18. O. Bakhteev *et al.*, Cross-language plagiarism detection: A case study of European languages academic works, [in:] S. Bjelobaba, T. Foltýnek, I. Glendinning, V. Krásničan, D.H. Dlabolová [Eds.], *Academic Integrity: Broadening Practices, Technologies, and the Role of Students*, Ethics and Integrity in Educational Contexts, Vol. 4, Springer, Cham, pp. 143–161, 2022, https://doi.org/10.1007/978-3-031-16976-2_9.

19. B. Agarwal, Cross-lingual plagiarism detection techniques for English-Hindi language pairs, *Journal of Discrete Mathematical Sciences and Cryptography*, **22**(4): 679–686, 2019, https://doi.org/10.1080/09720529.2019.1642626.

20. A. Niewiarowski, A. Plichta, Matrix similarity analysis of texts written in Romanian and Spanish, [in:] *ECMS 2023: Proceedings of the 37th ECMS International Conference on Modelling and Simulation*, Florence, Italy, June 20–23, **37**(1): 507–512, 2023.

21. V. Komorovskaya, The future of the Belarusian language: Is it doomed to extinction? Controversies and challenges in language maintenance and revitalization, *Acta Philologica*, **48**: 15–28, 2016.

22. M.S. Flier, A. Graziosi, The battle for Ukrainian: An introduction, *Harvard Ukrainian Studies: The Journal of the Ukrainian Research Institute at Harvard University*, **35**(1–4): 11–30, 2017–2018.

23. E. Agirre, Cross-lingual word embeddings, *Computational Linguistics*, **46**(1): 245–248, 2020, https://doi.org/10.1162/coli_r_00372.

24. N.R. Schneider, A. Das, K. O'Sullivan, H. Samet, Cross-lingual clustering using large language models, [in:] *Proceedings of the 7th ACM SIGSPATIAL International Workshop on AI for Geographic Knowledge Discovery (GeoAI '24)*, Association for Computing Machinery, New York, USA, pp. 1–10, 2024, https://doi.org/10.1145/3687123.3698280.

25. S. Dutta, "Alignment is all you need": Analyzing cross-lingual text similarity for domain-specific applications, [in:] *Proceedings of the International Workshop on Cross-lingual Event-centric Open Analytics*, *CEUR Workshop Proceedings*, Vol. 2829, pp. 13–22, 2021.

26. C.D. Manning, P. Raghavan, H. Schütze, *Introduction to Information Retrieval*, Cambridge University Press, Cambridge, 2008.

27. Website: https://antyplagius.n-dms.com, New Data Mining Systems sp. z o.o., YouTube channel of the project: https://youtube.com/@n-dms.

28. Slavic languages, *Britannica*, https://www.britannica.com/topic/Slavic-languages.

29. C.D. Manning, P. Raghavan, H. Schütze, Stemming and lemmatization, [in:] *Introduction to Information Retrieval*, Cambridge University Press, 2008, https://nlp.stanford.edu/IR-book/html/htmledition/stemming-and-lemmatization-1.html.

30. Google Translate, https://translate.google.com.

31. Microsoft Translator, https://www.bing.com/translator.

Additional Online Resources

A1. A full excerpt from the text is available at https://antyplagius.n-dms.com/tests/Spanish-Romanian/Espania-Spanish-wikipedia-google-translate.txt.

A2. A full excerpt from the text is available at https://antyplagius.n-dms.com/tests/Spanish-Romanian/Espania-Romanian-wikipedia-google-translate.txt.

A3. N-DMS, Belarusian and Ukrainian – an analysis of similarities. Antyplagiat N-DMS Antyplagius [in Polish], YouTube, 06.01.2022, https://youtu.be/d6o3QAQDWPk.

A4. N-DMS ANTYPLAGIUS, Belarus – Wikipedia, https://antyplagius.n-dms.com/tests/Belarusian-Ukrainian/Belarus-Wikipedia.pdf.

A5. Wikipedia, Беларусь, https://be.wikipedia.org/wiki/%D0%91%D0%B5%D0%BB%D0%B0%D1%80%D1%83%D1%81%D1%8C.

A6. N-DMS ANTYPLAGIUS, Belarusian-Belarus, https://antyplagius.n-dms.com/tests/Belarusian-Ukrainian/Belarusian-Belarus.txt.

A7. N-DMS ANTYPLAGIUS, Ukrainian-Belarus, https://antyplagius.n-dms.com/tests/Belarusian-Ukrainian/Ukrainian-Belarus.txt.

A8. N-DMS ANTYPLAGIUS, EN Fragment – Ukraine – Britannica Online Encyclopedia, https://antyplagius.n-dms.com/tests/Belarusian-Ukrainian/EN%20FRAGMENT%20-%20Ukraine%20–%20Britannica%20Online%20Encyclopedia.txt.

A9. N-DMS ANTYPLAGIUS, Ukraine – Britannica Online Encyclopedia, https://antyplagius.n-dms.com/tests/Belarusian-Ukrainian/Ukraine%20–%20Britannica%20Online%20Encyclopedia.pdf.

A10. Ukraine, *Britannica*, https://www.britannica.com/place/Ukraine.

A11. N-DMS ANTYPLAGIUS, Chapter "Plant and animal life" in English from encyclopaedia: https://antyplagius.n-dms.com/tests/Belarusian-Ukrainian/CYR/ENG%20-%20Brit%20-%20Animals%20-%20google%20translate.txt.

A12. N-DMS ANTYPLAGIUS, Belarusian, https://antyplagius.n-dms.com/tests/Belarusian-Ukrainian/CYR/BY%20-%20Brit%20-%20Animals%20-%20google%20translate.txt.

A13. N-DMS ANTYPLAGIUS, Ukrainian, https://antyplagius.n-dms.com/tests/Belarusian-Ukrainian/CYR/UA%20-%20Brit%20-%20Animals%20-%20google%20translate.txt.

A14. N-DMS ANTYPLAGIUS, Bulgarian, https://antyplagius.n-dms.com/tests/Belarusian-Ukrainian/CYR/BUL%20-%20Brit%20-%20Animals%20-%20google%20translate.txt.

A15. N-DMS ANTYPLAGIUS, Serbian, https://antyplagius.n-dms.com/tests/Belarusian-Ukrainian/CYR/SR%20-%20Brit%20-%20Animals%20-%20google%20translate.txt.

A16. N-DMS ANTYPLAGIUS, Macedonian, https://antyplagius.n-dms.com/tests/Belarusian-Ukrainian/CYR/MK%20-%20Brit%20-%20Animals%20-%20google%20translate.txt.

A17. N-DMS ANTYPLAGIUS, Kazakh, https://antyplagius.n-dms.com/tests/Belarusian-Ukrainian/CYR/KZ%20-%20Brit%20-%20Animals%20-%20google%20translate.txt.

A18. N-DMS, Antyplagius vs chatGPT-4 – Review of the film Troy [in Polish], YouTube, 10.04.2023, https://youtu.be/_ejk1xTPDDQ.

A19. N-DMS, Antyplagius vs chatGPT (part 1), YouTube, 27.04.2023, https://youtu.be/PxrVB9AwcR0.

A20. N-DMS, Are the Spanish and Romanian languages similar to each other? Test using the Antyplagiat N-DMS Antyplagius [in Polish], YouTube, 22.01.2022, https://youtu.be/JhfdwbyIsFc.

Received March 24, 2024; revised version July 11, 2025;
accepted July 29, 2025; published online September 2, 2025.