Matrix similarity analysis of texts written in Romanian and Spanish

Artur Niewiarowski

Department
of Computer Science
Cracow University
of Technology
Cracow, Poland
New Data Mining
Systems sp. z o.o.
58 Jana Pawła II Street
30-444 Libertów, Poland

Anna Plichta

Department
of Computer Science
Cracow University
of Technology
Cracow, Poland

KEYWORDS

text-mining; anti-plagiarism; text similarity analysis; Levenshtein's edit distance; matrix analysis of texts; Romanian language; Spanish language; Romance language group

ABSTRACT

This publication presents the results of a study of similarity between texts written in Romanian and Spanish, using a matrix analysis method based on Levenshtein's edit distance. The method used in the study does not contain implemented language-dependent vocabulary rules and exhibits the feature of linguistic universality in terms of similarity analysis. The study was carried out on the basis of the commercial computer program Antyplagius, created by the New Data Mining Systems company, which performs similarity analysis exclusively using the aforementioned method. The texts being compared were taken from excerpts from Wikipedia translated by online translators of popular companies which are based on artificial intelligence solutions.

I. INTRODUCTION

Romanian is spoken by some 24 million people, primarily in Romania and Moldova but also in Bulgaria, Serbia, Ukraine, Hungary, and among the members of the Romanian diaspora in the US, Canada and Germany [12], [22], [5]. Spanish, on the other hand, belongs to the most widely spoken languages in the world, ranking second in terms of usage with 360 million people worldwide for whom it is the first language [20]. Compared to Romanian, Spanish is a global language spoken in Europe, parts of Africa and in most countries of the New World where it is an official language in such countries as Argentina, Cuba, Mexico, Peru, and Venezuela [6]. Due to its strong global position, it is also one of the official languages of the UN, the EU, as well as UNESCO. Texts written in Spanish and Romanian can be analyzed for similarity, since they belong to the same Romance language group [22], [4]. However, one of the factors taken into consideration in the above publication, was that Romanian and Spanish are probably the least similar languages in the mentioned group. Historically speaking, both are descendants of vernacular Latin which were evolving in relative geographical isolation and under heavy influence of different non-Romance languages. While Spanish lexicon and phonology were impacted by Arabic during the seven centuries of Reconquista, Romanian was completely cut off from the other Romance languages after the fall of the Roman empire and evolved as a Romance linguistic island surrounded by Slavic languages; later, it came into long contact with Hungarian and Ottoman Turkish. Therefore, it is no longer mutually intelligible with Spanish.

II. RESEARCH PROBLEM

A. *Description of the problem*

This study is intended to check whether it is possible to carry out an effective comparative analysis between texts written in Spanish and Romanian using a computer algorithm that does not contain grammar rules dedicated to Romance languages as such, including the implemented stemming and lemmatization methods [16], [3]. The task is hindered by the fact that Spanish and Romanian are the least similar to each other in the same language group; moreover, in Romanian the articles do not precede the noun as in Spanish but are attached to it as an ending (suffix), which significantly alters the structure of the words and can make similarity analysis difficult.

B. *Text analysis*

The concept of matrix analysis of text data based on Levenshtein's edit distance [14], [25], [11], is described in detail in [17].

At present, there are no competing algorithms for

Communications of the ECMS, Volume 37, Issue 1,
Proceedings, ©ECMS Enrico Vicario, Romeo Bandinelli,
Virginia Fani, Michele Mastroianni (Editors) 2023
ISBN: 978-3-937436-80-7/978-3-937436-79-1 (CD) ISSN 2522-2414

performing such comparisons (i.e., between different languages), so alternatives are not mentioned in this publication. Currently known methods for analyzing text comparisons are based on stemming and lemmatization algorithms tailored to a specific language rather than a language group, such as those described in publications [24], [13]. This chapter presents the concept of the algorithm in general, based on Spanish and Portuguese texts.

C. *Presentation of the algorithm for analyzing text data*

Generally, the idea is to build a matrix from two analyzed documents of size equal to the number of words of one and the other document, respectively. The matrix is completed with logical values (1 or 0) depending on whether the similarity value calculated between the words in a given iterative step based on Levenshtein distance [14], falls within a given range set as one of the parameters by the user (formula 1).

$$\mathrm{I}_{ti=1}^{tm}\,\mathrm{I}_{tj=1}^{tn}\,\mathbf{M}[ti,tj]=\beta \qquad (1)$$
$$\begin{cases}\beta=\mathrm{true}:\mathrm{fp}(\mathbf{Doc1}[ti],\mathbf{Doc2}[tj])\geq bp\\ \beta=\mathrm{false}:\mathrm{fp}(\mathbf{Doc1}[ti],\mathbf{Doc2}[tj])<bp\end{cases}$$

where:
fp - function returns similarity measure p (formula 2),
bp - acceptable boundary value of similarity measure parameterized by the user e.g. corresponds to the type of the documents (e.g., scientific vocabulary) or the language of documents - the determination of the appropriate value of this parameter must be preceded by prior analyses against the given languages, language groups and document types, which is done, among others, in this publication,
ß - values either false or true,
Doc1[ti] - element (term) of document 1 separated by a space from a next element of the document.

The similarity measure p is calculated by the formula:

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

where:
m, n – lengths of two terms/text strings (i.e. number of characters),
kmax - length of the longest of analyzed two terms/text strings (i.e. pessimistic case where k is equal to the length of the longest term).

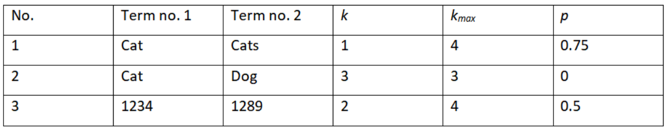
The Levenshtein distance k is equal to the D[m,n] element of the so-called Levenshtein matrix D:

$$k=\mathbf{D}[m,n]=\mathrm{LevenshteinDistance}(\mathrm{Term1},\mathrm{Term2}) \qquad (3)$$

where:

D[m,n] – result of the Levenshtein distance algorithm, the last element of matrix D, i.e. minimum number of operations: insertion, deletion and substitution required to convert one term (text string) into the other.

Examples of the similarity measure p, based on formula 2 are presented below.



No.	Term no. 1	Term no. 2	k	k_{max}	p
1	Cat	Cats	1	4	0.75
2	Cat	Dog	3	3	0
3	1234	1289	2	4	0.5

Fig. 1: Examples of the similarity measure p, based on formula 2

The figure 2 shows a graphical visualization of the comparison of two very short texts written in Spanish and Portuguese. The individual points represent locations on the rectangular matrix with logical "true" values, i.e., fragments similar between the text strings. The content of the document written in Spanish is as follows: Pero, a pesar de esta variedad de posibilidades que la voz posee, sería un muy pobre instrumento de comunicación si no contara más que con ella. La capacidad de expresión del hombre no dispondría de más medios que la de los animales. La voz, sola, es para el hombre escasamente una materia informe, que para convertirse en un instrumento perfecto de comunicación debe ser sometida a un cierto tratamiento. Esa manipulación que recibe la voz son las "articulaciones." The content of the document written in Portuguese reads: Mas, apesar da variedade de possibilidades que a voz possui, seria um instrumento de comunicação muito pobre se não se contasse com mais do que ela. A capacidade de expressão do homem não disporia de mais meios que a dos animais. A voz, sozinha, é para o homem apenas uma matéria informe, que para se converter num instrumento perfeito de comunicação deve ser submetida a um certo tratamento. Essa manipulação que a voz recebe são as "articulações."

D. *A tool to perform data analysis*

In the research, the N-DMS Antiplagius program was used whose operation principle is based on the

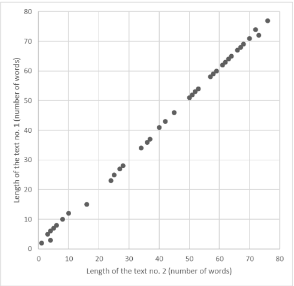


Fig. 2: Example of the result in the form of a graphical matrix analysis of two short texts written in two languages both belonging to the Romance language group

presented concept of matrix analysis of text data [1]; YouTube channel of the project: `https://youtube.com/@n-dms`. The application is one of the results of scientific research on algorithms from the text-mining family [19], [18], [17]. It performs similarity analysis between text data.

It does not contain implemented language-dependent vocabulary rules and is based on the author's linguistically universal solutions[17], including additionally the Levenshtein edit distance. The Levenshtein distance is a generalization of the Hamming path and has numerous applications in text string analysis, such as in text processors [25], or DNA sequence analysis (Levenshtein-Damerau distance) [11].

Consequently, it can analyze documents in languages of European roots using the Latin alphabet, Cyrillic script (additional possibility of automatic transcription of documents) and the Chinese characters. It allows the user to customize the analysis parameters for the documents under study, including, among others: the degree of word similarity and the size of breaks in sentence continuity. In addition, it has a defined set of parameters for the types of documents studied [i.e., paper, homework, thesis, journal, article, book, mixed] and the language. The program has defined analysis parameters for the following languages: Belarusian, Bulgarian, Chinese, Czech, Danish, Finnish, French, German, Italian, Dutch, Norwegian, Polish, Portuguese, Russian, Romanian, Slovak, Swedish, Taiwanese, and Ukrainian. The application additionally has a built-in OCR module (based on the world-famous Tesseract OCR Engine) that recognizes text in images [23] and perfectly complements the text similarity method based on editing distance correcting the 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 results of the analysis are: text fragments considered similar and a diagram of the relationship between documents.

E. ***Data analysis and results***

The text strings analyzed came from online sources in the form of entries from two different well-known online encyclopedias. They were subjected to machine translation based on artificial intelligence solutions[translators used in translation: `https://translate.google.com` and `https://www.bing.com/translator`] which are now considered extremely effective. The two tests represent separate approaches related to translation. The first approach is to adapt one language to the other by translating the former. The second test begins by translating English into the two languages being tested. In a way, the approaches presented are a reference to the research that is taking place on issues related to cross-language plagiarism, which is being committed with increasing frequency around the world in schools and universities [8], [9].

F. ***Encyclopedia article on Spain***

This analysis uses an encyclopedic article about Spain `https://es.wikipedia.org/wiki/EspaC3B1a`, written in Spanish and machine-translated into Romanian as described here [2]. The texts were compared with each other, in addition, one of the analyses was posted as a video on the YouTube channel `https://www.youtube.com/watch?v=JhfdwbyIsFc`. Graphical interpretation of the compared texts is below, where: **bp** - acceptable boundary value of similarity measure parameterized by the user; **wv** - minimum number of words in sequence vector, and **gw** - maximum acceptable gap between words. The constant gw will be the same for all the tests in the chapter, since the specifics of the problem do not force its constant adjustment to the text. The gw constant is provided for the analysis of texts where there is a significant likeliness of attempting to misrepresent the content by deliberately changing the structure of sentences, including reordering terms, deleting words and inserting equivalents in the form of synonyms. The value has been selected through previous research on texts written in different languages.

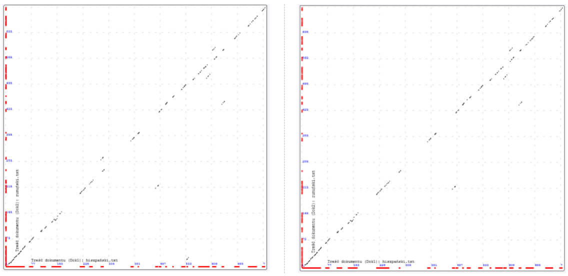


Fig. 3: Analysis parameters - bp: 42%, wv: 5, gw: 8 and Analysis parameters - bp: 45%, wv: 5, gw: 8

In the figures and in the table, it can be seen that the best result showing significant similarity between the texts is obtained by setting the word similarity to between 42% and 50% Setting the word similarity (bw) value below 42% for the above example makes the graphical result of the comparison less clear, noise appears, and the diagonal line (possibly smaller diagonal

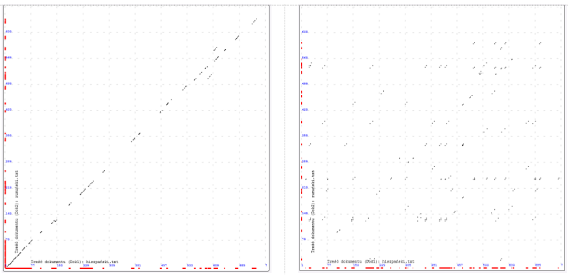


Fig. 4: Analysis parameters - bp: 50%, wv: 5, gw: 8 and Analysis parameters - bp: 100%, wv: 3,gw: 8

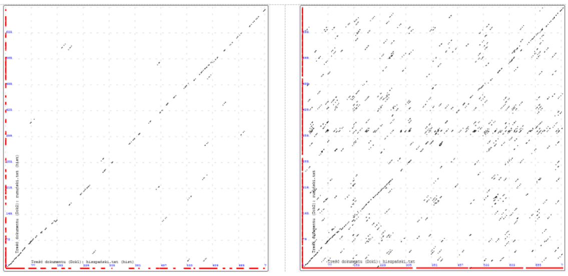
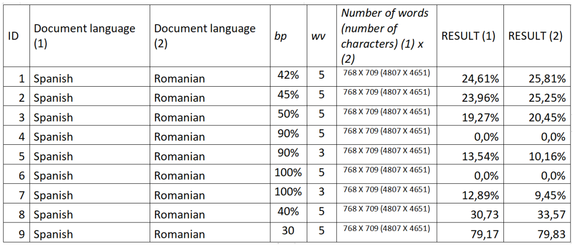


Fig. 5: Analysis parameters - bp: 40%, wv: 5,gw: 8 and Analysis parameters - bp: 30%, wv: 5,gw: 8

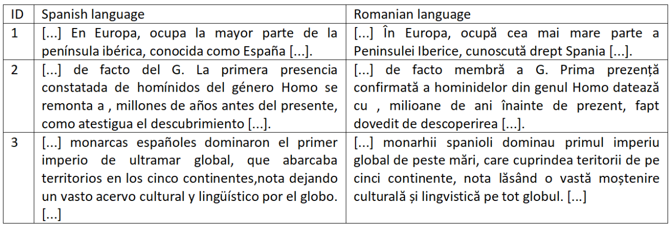


ID	Document language (1)	Document language (2)	*bp*	*wv*	*Number of words (number of characters) (1) x (2)*	RESULT (1)	RESULT (2)
1	Spanish	Romanian	42%	5	768 X 709 (4807 X 4651)	24,61%	25,81%
2	Spanish	Romanian	45%	5	768 X 709 (4807 X 4651)	23,96%	25,25%
3	Spanish	Romanian	50%	5	768 X 709 (4807 X 4651)	19,27%	20,45%
4	Spanish	Romanian	90%	5	768 X 709 (4807 X 4651)	0,0%	0,0%
5	Spanish	Romanian	90%	3	768 X 709 (4807 X 4651)	13,54%	10,16%
6	Spanish	Romanian	100%	5	768 X 709 (4807 X 4651)	0,0%	0,0%
7	Spanish	Romanian	100%	3	768 X 709 (4807 X 4651)	12,89%	9,45%
8	Spanish	Romanian	40%	5	768 X 709 (4807 X 4651)	30,73	33,57
9	Spanish	Romanian	30	5	768 X 709 (4807 X 4651)	79,17	79,83

Fig. 6: Table showing the results of several text comparison analyses, with different analysis parameters

lines), which can be said to be responsible for the visual confirmation of the similarity of the texts, is less visible and blurs in the noise. Reducing the bw constant to 0% will generate a result of 100% similarity between documents - which will be an obvious error. Increasing the degree of word similarity closer to 100 will result in the disappearance of points on the matrix and no visible similarity between text strings. Examples of text passages considered similar for ID 1 analysis from Table - Fig.6 are presented below.

The above table contains selected passages considered similar that are the result of an analysis of the comparison of the texts in question. Each of the above words considered similar to its counterpart in the other text has its graphical interpretation in the form of a point on the matrix.



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. [...]

Fig. 7: Examples of text passages considered similar for ID 1 analysis from Table - Fig.6.

G. *Encyclopedia article on Romania*

The above analysis juxtaposes two texts about Romania from the online encyclopedia `https://www.britannica.com/place/Romania`. Previously, the English text was translated by a different translator into Romanian and Spanish `https://www.bing.com/translator`. Below is a graphic interpretation of the comparison.

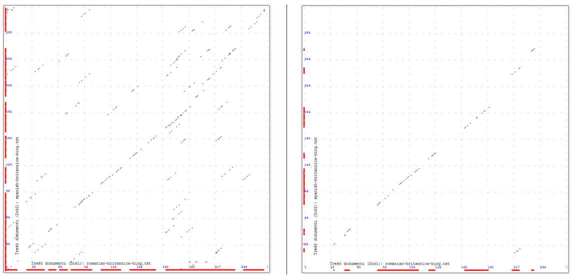


Fig. 8: Analysis parameters - bp: 30%, wv: 5, gw: 8 and Analysis parameters - bp: 42%, wv: 5, gw: 8

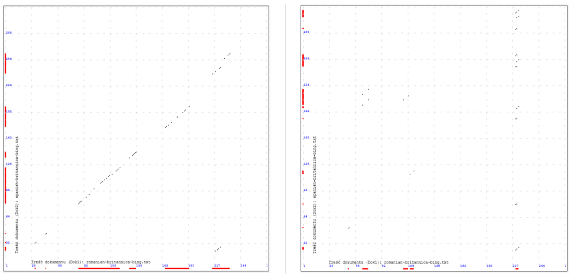
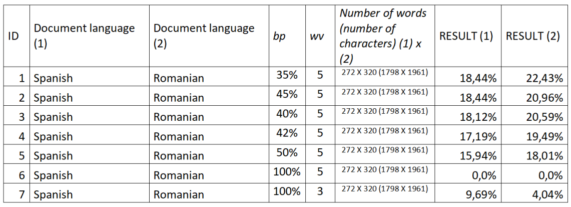


Fig. 9: Analysis parameters - bp: 50%, wv: 5, gw: 8 and Analysis parameters-bp: 100%, wv:3,gw: 8

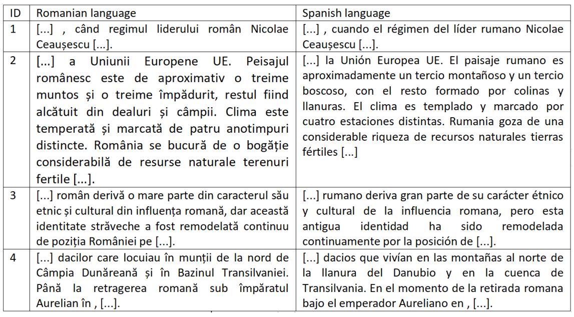
As in the previous test, the best-fitting parameters for the analysis of the two languages are similarities between 42% and 50%.

The above table contains selected text passages considered similar by the algorithm. The data are the result of the analysis based on parameters No. 2 from the Fig.10.



ID	Document language (1)	Document language (2)	*bp*	*wv*	*Number of words (number of characters) (1) x (2)*	RESULT (1)	RESULT (2)
1	Spanish	Romanian	35%	5	272 X 320 (1798 X 1961)	18,44%	22,43%
2	Spanish	Romanian	45%	5	272 X 320 (1798 X 1961)	18,44%	20,96%
3	Spanish	Romanian	40%	5	272 X 320 (1798 X 1961)	18,12%	20,59%
4	Spanish	Romanian	42%	5	272 X 320 (1798 X 1961)	17,19%	19,49%
5	Spanish	Romanian	50%	5	272 X 320 (1798 X 1961)	15,94%	18,01%
6	Spanish	Romanian	100%	5	272 X 320 (1798 X 1961)	0,0%	0,0%
7	Spanish	Romanian	100%	3	272 X 320 (1798 X 1961)	9,69%	4,04%

Fig. 10: Table showing the results of several text comparison analyses, with different analysis parameters



ID	Romanian language	Spanish language
1	[...] , când regimul liderului român Nicolae Ceaușescu [...].	[...] , cuando el régimen del líder rumano Nicolae Ceaușescu [...].
2	[...] a Uniunii Europene UE. Peisajul românesc este de aproximativ o treime muntos și o treime împădurit, restul fiind alcătuit din dealuri și câmpii. Clima este temperată și marcată de patru anotimpuri distincte. România se bucură de o bogăție considerabilă de resurse naturale terenuri fertile [...].	[...] la Unión Europea UE. El paisaje rumano es aproximadamente un tercio montañoso y un tercio boscoso, con el resto formado por colinas y llanuras. El clima es templado y marcado por cuatro estaciones distintas. Rumania goza de una considerable riqueza de recursos naturales tierras fértiles [...]
3	[...] român derivă o mare parte din caracterul său etnic și cultural din influența romană, dar această identitate străveche a fost remodelată continuu de poziția României pe [...].	[...] rumano deriva gran parte de su carácter étnico y cultural de la influencia romana, pero esta antigua identidad ha sido remodelada continuamente por la posición de [...].
4	[...] dacilor care locuiau în munții de la nord de Câmpia Dunăreană și în Bazinul Transilvaniei. Până la retragerea romană sub împăratul Aurelian în , [...].	[...] dacios que vivían en las montañas al norte de la llanura del Danubio y en la cuenca de Transilvania. En el momento de la retirada romana bajo el emperador Aureliano en , [...].

Fig. 11: Examples of text passages considered similar for ID 2 analysis from Table - Fig.10.

H. Applications

Summarizing the above results, it turns out that the accuracy of the analysis results depends primarily on the word similarity parameter bp. It is responsible for recognizing in the matrix concept of text analysis whether the words analyzed in a given iteration step are to be considered identical and fill the matrix cell with a positive value. Based on the above results, it can be seen that the algorithm, in which the grammatical rules for a particular language are not implemented, is able to correctly estimate the existing similarity between texts despite additional differences due to different languages. In addition, the matrix analysis algorithm based on Levenshtein's edit distance [14] confirmed the similarity of languages from the common Romance language group described by linguists [22], [12].

III. Summary and Future Work

The matrix text analysis algorithm based on Levenshtein's edit distance[14], confirmed the similarity of languages from the common Romance language group described by linguists[12], [22]. The algorithm does not use Thesaurus, so words with similar meaning whose edit distance is large are not considered identical. However, this should not significantly affect the result of the text similarity analysis, because it is impossible to swap most of the words of a text document so that it still carries the same message and at the same time consists of other terms with similar meaning. And even if this were possible, it is difficult in such a case to talk about simple plagiarism of the text. However, a dictionary of similar words would be an interesting component to strengthen the algorithm, so this will be the subject of further research, especially in terms of optimizing the overall calculation.

In addition, the approach presented in the publication will be used to analyze the similarity of the texts 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 a growing moral issue in academia[7], [15]. The first steps in this regard have already been made, and the results can be viewed at the following links: `https://youtu.be/_ejk1xTPDDQ` and `https://youtu.be/PxrVB9AwcR0`.

References

[1] https://antyplagius.n-dms.com,.

[2] https://antyplagius.n-dms.com/tests/spanish-romanian/espania-spanish-wikipedia-google-translate.txt.

[3] https://nlp.stanford.edu/ir-book/html/htmledition/stemming-and-lemmatization-1.html.

[4] https://www.britannica.com/topic/romance-languages.

[5] https://www.britannica.com/topic/romanian-language.

[6] https://www.britannica.com/topic/spanish-language.

[7] A. J. Adetayo. Artificial intelligence chatbots in academic libraries: the rise of chatgpt. *Library Hi Tech News*, 2023.

[8] B. Agarwal. Cross-lingual plagiarism detection techniques for english-hindi language pairs. *Journal of Discrete Mathematical Sciences and Cryptography*, 22(4):679–686, 2019.

[9] O. Bakhteev, Y. Chekhovich, A. Grabovoy, G. Gorbachev, T. Gorlenko, K. Grashchenkov, A. Ivakhnenko, A. Kildyakov, A. Khazov, V. Komarnitsky, et al. Cross-language plagiarism detection: A case study of european languages academic works. In *Academic Integrity: Broadening Practices, Technologies, and the Role of Students: Proceedings from the European Conference on Academic Integrity and Plagiarism 2021*, pages 143–161. Springer, 2023.

[10] A. Dziob and M. Piasecki. Implementation of the verb model in plwordnet 4.0. In *Proceedings of the 9th Global Wordnet Conference*, pages 113–122, 2018.

[11] R. Gabrys, E. Yaakobi, and O. Milenkovic. Codes in the damerau distance for dna storage. In *2016 IEEE International Symposium on Information Theory (ISIT)*, pages 2644–2648. IEEE, 2016.

[12] K. Katzner and K. Miller. *The languages of the world.* Routledge, 2002.

[13] D. Khyani, B. Siddhartha, N. Niveditha, and B. Divya. An interpretation of lemmatization and stemming in natural language processing. *Journal of University of Shanghai for Science and Technology*, 22(10):350–357, 2021.

[14] V. Levenshtein. Binary codes for correcting dropouts, inserts, and symbol substitutions. *Reports of the Academy of Sciences of the USSR*, 163(4):845–848, 1965.

[15] B. D. Lund and T. Wang. Chatting about chatgpt: how may ai and gpt impact academia and libraries? *Library Hi Tech News*, 2023.

[16] C. Manning. *Prabhakar Raghavan and Hinrich Schütze, Introduction to Information Retrieval.* Cambridge University Press, 2008.

[17] 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.

[18] A. Niewiarowski et al. Short text similarity algorithm based on the edit distance and thesaurus. *Czasopismo Techniczne*, 2016(Nauki Podstawowe Zeszyt 1-NP 2016):159–173, 2016.

[19] A. Niewiarowski, M. Stanuszek, et al. Parallelization of the levenshtein distance algorithm. *Czasopismo Techniczne*, 2014(Nauki Podstawowe Zeszyt 3 NP (17) 2014):109–112, 2014.

[20] R. Penny and R. J. Penny. *A history of the Spanish language.* Cambridge University Press, 2002.

[21] P. R. Petrucci. Slavic features in the history of rumanian. 1994.

[22] R. Posner. *The Romance Languages.* Cambridge Language Surveys, 1996.

[23] R. Smith. An overview of the tesseract ocr engine. In *Ninth international conference on document analysis and recognition (ICDAR 2007)*, volume 2, pages 629–633. IEEE, 2007.

[24] B. Wahiba. A new stemmer to improve information. *International Journal of Network Security & Its Applications (IJNSA)*, pages 143–154, 2013.

[25] M. M. Yulianto, R. Arifudin, and A. Alamsyah. Autocomplete and spell checking levenshtein distance algorithm to getting text suggest error data searching in library. *Scientific Journal of Informatics*, 5(1):75, 2018.

AUTHOR BIOGRAPHIES

ARTUR NIEWIAROWSKI academic teacher, assistant at the Department of Computer Science and Telecommunications, Cracow University of Technology. Academic achievements cover the area of technical sciences in the discipline of computer science. Author of usable, scientific and financial computer programs and data management systems for universities. Co-owner and president of the board of directors of the IT company New Data Mining Systems sp. z o.o. Member of the Polish Informatics Society.

ANNA PLICHTA She graduated computer science at Cracow University of Technology in 2010. In 2019 she obtained a Ph.D. in computer science at Wroclaw University of Science and Technology. Currently, she is an Assistant Professor at Cracow University of Technology. The main topics of her research are pattern recognition, databases, artificial intelligent systems and e-learning technologies.