***Even though the plagiarism identification issue remains relevant, modern detection methods are still resource-intensive. This paper reports a more efficient alternative to existing solutions.***

***The devised system for identifying patterns in multilingual texts compares two texts and determines, by using different approaches, whether the second text is a translation of the first or not. This study's approach is based on Renyi entropy.***

***The original text from an English writer's work and five texts in the Russian language were selected for this research. The real and "fake" translations that were chosen included translations by Google Translator and Yandex Translator, an author's book translation, a text from another work by an English writer, and a fake text. The fake text represents a text compiled with the same frequency of keywords as in the authentic text.***

***Upon forming a key series of high-frequency words for the original text, the relevant key series for other texts were identified. Then the entropies for the texts were calculated when they were divided into "sentences" and "paragraphs".***

***A Minkowski metric was used to calculate the proximity of the texts. It underlies the calculations of a Hamming distance, the Cartesian distance, the distance between the centers of masses, the distance between the geometric centers, and the distance between the centers of parametric means.***

***It was found that the proximity of texts is best determined by calculating the relative distances between the centers of parametric means (for "fake" texts – exceeding 3, for translations – less than 1).***

***Calculating the proximity of texts by using the algorithm based on Renyi entropy, reported in this work, makes it possible to save resources and time compared to methods based on neural networks. All the raw data and an example of the entropy calculation on php are publicly available***

***Keywords: Google Translator, Yandex.Translator, Renyi entropy, Minkowski metric, Hamming distance***

UDC 00.004.9

DOI: 10.15587/1729-4061.2021.228695

# DEVISING AN ENTROPY-BASED APPROACH FOR IDENTIFYING PATTERNS IN MULTILINGUAL TEXTS

**Gulnur Yerkebulan**
Master of Engineering Science, Doctoral Candidate*
E-mail: erkgulnur@mail.ru
**Valentina Kulikova**
PhD, Associate Professor*
E-mail: valentina@nkzu.kz
**Vladimir Kulikov**
PhD, Associate Professor*
E-mail: qwertyrant@nkzu.kz
**Zaru Kulsharipova**
PhD, Associate Professor
Higher School of Pedagogy
Pavlodar Pedagogical University
Mira str., 60, Pavlodar,
Republic of Kazakhstan, 140000
E-mail: kulsharipovazk@mail.ru
*Department of Information
and Communication Technologies
Manash Kozybayev North Kazakhstan University
Pushkin str., 86, Petropavlovsk,
Republic of Kazakhstan, 150000

*Received date 01.03.2021*
*Accepted date 01.04.2021*
*Published date 30.04.2021*

***How to Cite:*** *Yerkebulan, G., Kulikova, V., Kulikov, V., Kulsharipova, Z. (2021). Devising an entropy-based approach for identifying patterns in multilingual texts. Eastern-European Journal of Enterprise Technologies, 2 (2 (110)), 16–22. doi: https://doi.org/10.15587/1729-4061.2021.228695*

## 1. Introduction

Plagiarism detection is still a pressing issue, especially with the advent of websites that automatically generate texts, as well as such translator websites that enable translating from one language to another while making changes to the original text.

Various methods are used to detect plagiarism, among which neural network-based techniques are rapidly evolving. Neural networks are used wherever one wants to solve prediction, classification, or management tasks. The benefits of neural networks include problem-solving under unknown patterns, the resistance to noisy input data, potential ultra-high performance, as well as failure-free operation in the hardware implementation of a neural network [1]. At the same time, their training and operation require enormous computational resources [2].

Finding less expensive solutions to this problem could save time, bring down hardware and software costs. The search for borrowings would become more accessible for the user who does not have special knowledge. Technological advancements are expected to expand the list of tasks to be solved. For example, in addition to determining a borrowing, it could be possible to find the original sources of news and articles, regardless of language.

## 2. Literature review and problem statement

The entropy approach, which is much more resource-efficient than neural networks, has previously been used in certain "niches" when working with text documents. We shall consider some of them.

Paper [3] reports the results of a study into extracting concepts for a structured text using the entropy weight method. Classical methods of extracting concepts are based on frequency (word frequency, frequency of documents, and TF-IDF). The authors of the cited pa-

per decided additionally to take into consideration the weight of each module by using information entropy. To this end, they evaluated the contribution of each module to the weight assessment of the concept. If the weight of entropy is zero, that module is too weak to contribute useful information to the calculations. At the same time, the authors performed calculations related to academic texts; however, they did not fully examine the application of advancements to other structured texts and documents.

Work [4] reports the results of a study into the entropy relationship between text length and lexical wealth. It is shown that lexical wealth by Shannon increases rapidly with shorter texts, and then it reaches a certain point from which it stabilizes, despite the continuous increase in the length of the text. This stability can be explained by the stabilization of the probability of a word appearing in the texts. However, the cited study is limited to data from only one language (English); another study to address the issues of natural entropy in different languages is planned by researchers in the future.

Paper [5] explores entropy analysis of questionable text sources using the example of Voynich manuscript. Some scholars believe that the Voynich manuscript is genuine, others believe it is a hoax. Three methods, including the entropy calculation by Shannon and Renyi, show that the mysterious manuscript is a significant human art, rather than a hoax. The methods developed are applicable to any text source. The paper notes that the method based on Shannon's entropy has a disadvantage, which is that one value does not make it possible to draw unambiguous conclusions. However, this method generally demonstrates that the Voynich manuscript is not an encrypted text. It is also noted that the calculations of Renyi entropy depend on the scale when applied to continuous distributions and, therefore, their absolute values are meaningless; it is necessary to consider the whole chart. Owing to the method based on Renyi entropy, the cited paper's authors concluded that the manuscript was mostly similar to a natural language.

In the post-Soviet space, there is also an example of investigating authorship using Shannon's entropy. There are widely known attempts to "tie the authorship" of the document "Silent Don" not to Sholokhov but to other writers. The authorship was mathematically proven by the Burroughs Delta method [6]. Among the approaches to the controversial issue was the use of computer processing by an archiver (that is, the most unified entropy assessment) and a direct assessment by Shannon's entropy of the words used by authors. Thus, the authors of work [7] calculated Shannon's entropy based on the probabilities of words and the probabilities of letters in a text, then analyzed the difference in entropies received. As a result, four volumes of Sholokhov's text had a difference ranging from 0.214968 to 0.233365, and in four of Kryukov's stories, it ranged from 0.181743 to 0.253215. It was concluded that the texts differed according to the formal criteria set. At the same time, one can note the drawback in those calculations such as overlapping intervals, so somewhere the texts by different authors are similarly based on the data from entropy calculations.

Work [8] links entropy assessment to machine translation and suggests methods for dealing with insufficient translation by two-phase process splitting. The first step introduces a simple strategy to reduce the entropy of highly entropy words by building pseudo-representations. The detail phase offers a pre-learning method, a multitasking method, and a two-run method to stimulate the neural model to correctly translate high-entropic words. However, the resource-cost aspect of that solution renders additional relevance to the search for alternative approaches to entropy-based interaction with translations.

The idea of establishing a pattern of "being translated" based on the entropy evaluation of texts is original. Renyi entropy, which has been found in a series of "geometric" applications, is more acceptable than Shannon's classic approach because of the additional parameter and generalizing nature for the Kullback-Leibler distances [9]. Although Tsalis entropy may also serve the parametrized entropy to generalize the Shannon entropy, it is not additive, which does not allow its application in our approach to the analysis of texts.

## 3. The aim and objectives of the study

The aim of this work is to determine whether it is possible to apply an entropy approach to calculate whether a particular text is a translation of the original text in the system of identification of patterns in polylingual texts.

To accomplish the aim, the following tasks have been set:

– to calculate the sets of entropies for patterns "sentence" and "paragraph";

– to calculate distances between the sets of texts' entropies based on Minkowski metric and define a pattern when comparing the original text with translations and "fake" texts, or the lack of it.

## 4. Materials and methods to study patterns in polylingual texts

### 4. 1. Entropy calculation

Our hypothesis assumes building numerical series by calculating Renyi entropy in a text by dividing it into "sentences" and "paragraphs". The proximity of the series would indicate that one text is a translation of the second text, while a distance in the series would mean that the texts are different.

During the development, we used a "sentence" pattern and a "paragraph" pattern, where the two texts are compared based on calculations for each sentence and paragraph, respectively. In the future, it is possible to involve other patterns. For example, under a "page" pattern, all calculations would be performed for each page separately; under a "figure/table" pattern, all calculations would be carried out for a text divided into parts at the beginning of a figure or table. For this study's experiments, "text" and "paragraph" patterns were chosen as they are simple enough to validate our hypothesis and are present in all the texts under consideration.

It is also important that the statistically "similar" texts should not, if possible, be confused with the translation in order not to detect the "authorship" inadvertently, including that by a translator.

The first chapter of The Adventures of Sherlock Holmes by Conan Doyle in English (En), as well as the following texts, have been chosen for the comparison:

– the authorized translation of this chapter into the Russian language (RuAuth);

– the translation of En into Russian by Yandex.Translator (RuYandexTr);

– the translation of En into Russian by Google Translate (RuGoogleTr);

– a fake text created with Google search engine, with a frequency series of the original text (RuImitation);

– another work by Conan Doyle (RuOtherEnWork).

The data considered here can be accessed at website [10], in the "Texts Used" section.

To form a list of high-frequency words of the text, we used the calculation of word statistics at a specialized Internet resource [11].

We selected the first 20 English words from the original text En, the translation of which was found in the frequency list of words from the authorized translation. At the same time, stop-words were excluded, among which are conjunctions, common nouns, words that have too many synonyms in translation dictionaries, or have little meaning. In addition, the number of occurrences of these words was taken into consideration. The number of occurrences of a Russian word should not exceed the number of occurrences of the English word.

Thus, a key series from the original text in English were compiled ('said', 'know', 'eyes', 'majesty', 'little', 'matter', 'case', 'indeed', 'note', 'paper', 'photograph', 'address', 'face', 'German', 'good', 'himself', 'just', 'king', 'looked', 'mask'), as well as a key series from the authorized translation in the Russian language ('сказал', 'знаете', 'глаза', 'величество', 'маленьким', 'случае', 'дело', 'действительно', 'заметил', 'бумага', 'фотографию', 'адрес', 'лицо', 'по-немецки', 'хорошо', 'себя', 'только', ' король', 'посмотрел', 'маску'). The sequence of words in the key series of the authorized translation is maintained in accordance with the sequence of words in the key series from the original text.

Next, formula (1) was used to count entropies for the first word from the key series of the text under consideration, for the first-second word, for the first-second-third word, etc. for each sentence and each paragraph. Thus, we compiled two series of numbers (the "sentence" pattern and the "paragraph" pattern) for each text.

$$S_R = \frac{1}{1-q}\log\sum\nolimits_{ij}^{N} p_{ij}^{q}, \tag{1}$$

where $p_{ij} = \frac{n_{ij}}{N}$, $\sum_{ij}^{N} p_{ij} = 1$, $q$=2, $n_{ij}$ is the number of words in a key series $i$ (the actual rank number to which, from the frequency series we have chosen, the words are counted) in a sentence (paragraph) $j$, $N$ is the number of sentences (paragraphs) in the text used for calculation.

Two more texts were created for the research, obtained by translating the original text with the help of the online translators Google Translate and Yandex Metric. They also underwent a procedure of building the key series of high-frequency words and the series of numbers for two patterns.

For further research, a fake text was created in the Russian language using a Google search engine. First, we searched for words in the Russian language that corresponded to the frequency series of the original text, then copied the pieces of text from the search results to achieve the same number of keyword occurrences as that in the original text. This took into consideration the coincidence of the number of sentences and paragraphs in the original text and the fake text.

For the fake text, and another work by Conan Doyle, the procedure of compiling the key series of high-frequency words and the series of entropy for two patterns was repeated.

Calculating the entropy values for "sentence" and "paragraph" patterns was based on the php script we created, which included the text under consideration and the corresponding key series of 20 words. Porter's stemmer was used for Russian [12]. Porter's stemming algorithm, based on certain rules, cuts off the suffixes and endings according to the specificity of a language [13].

The resulting calculations can be accessed at website [10] by following the "Entropy Series for Sentences and Paragraphs" link, which also includes an example of entropy calculations for "paragraph" and "sentence" patterns for RuAuth and 20 words.

### 4. 2. Calculating the distances between texts' entropies

In the terminology of image recognition (classification without a trainer), the concept of best parsing is refined in each specific task by selecting the optimal break-up criterion, reflecting the "similarity" among the elements related to a given cluster. The measure of similarity and difference of distance type (a distance function) was considered. In this type, objects are considered to be the more similar the smaller the distance between them. The distance between the coordinates of the entropy of the studied texts was calculated.

Minkowski metric was considered: formula (2)

$$\rho_{ij} = \sqrt[r]{\sum_{k=1}^{J}\left|t_{ik} - t_{jk}\right|^{r}}, \tag{2}$$

where, at $r$=1, we obtain the Manhattan distance (Hamming distance); at $r$=2, we obtain the Euclidean metric (Cartesian distance); at $r$®$\mu$, we obtain a metric of dominance (Chebyshev distance).

The mathematical expectation is the center of the distribution of its probabilities. For discrete and continuous random values, the mathematical expectation is calculated from formulae (3):

$$M(X) = \sum_i x_i p_i, \quad \sum_i p_i = 1$$

and

$$M(X) = \int_{-\infty}^{\infty} x \cdot p(x)\mathrm{d}x, \tag{3}$$

where $x_i$ is the values taken by a random value at probability $p_i$, $p(x)$ is the density of the distribution of the probability of a random value $X$.

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i n_i = \sum_{i=1}^{n} x_i \varpi_i, \tag{4}$$

$$s_{\bar{x}} = \frac{\sigma}{\sqrt{n}}, \tag{5}$$

where $n_i$ and $v_i$ are, respectively, the absolute and relative frequency of the $i$-th value of a random value.

A general form of the parametric mean of a variation series of the $k$ order is formula (6):

$$\bar{x}_k = \left( \frac{\sum_{i=1}^{m} n_i x_i^k}{\sum_{i=1}^{m} n_i} \right)^{\frac{1}{k}} = \left( \frac{\sum_{i=1}^{m} n_i x_i^k}{n} \right)^{\frac{1}{k}}, \quad (6)$$

where $\bar{x}_k$ is the mean; $x_i$ is the variant of the $i$-th class of the examined totality; $n_i$ is the weights (class size); $m$ is the number of classes; $n$ is the population size. This formula lowers or increases the contribution of variants. The first ten is considered to be the first variant occurring 9 times less often, the second – 8 times less often, and so on until variant 9. Starting from variant 10, the frequency of occurrence is leveled to be calculated as similar. The weights were taken equal to the ordinate number of the coordinate point since the likelihood of their "usable contribution" is decreasing in proportion to a decrease in the coordinate. The $k$ indicator was accepted to equal unity.

The following distances between texts were calculated using a Minkowski metric:

– Hamming distance;
– Cartesian distance;
– the distance between the centers of masses;
– the distance between geometric centers;
– the distance between the centers of parametric means.

A cross-platform build of the XAMPP webserver was used [14], which includes a PHP script interpreter.

## 5. The results of calculating Renyi entropy and distances between entropies in the system of polylingual text pattern identification

### 5. 1. Calculating the set of entropies for "sentence" and "paragraph" patterns

We examined 6 texts in our experiments; the statistics on paragraphs and sentences are given in Table 1.

Table 1

Statistics on sentences and paragraphs

| Parameter | En | RuAuth | RuYandexTr | RuGoogleTr | RuImitation | RuOtherEnWork |
|---|---|---|---|---|---|---|
| Paragraph quantity | 116 | 116 | 116 | 116 | 116 | 116 |
| Sentence quantity | 292 | 293 | 281 | 302 | 292 | 292 |

Below are the charts comparing entropy calculations for the original text and other texts for "sentence" and "paragraph" patterns (Fig. 1–5).

For greater visibility, a chart was used on which entropies by paragraphs are marked on the abscissa, and entropies by sentences (Fig. 6) are marked on the ordinate.

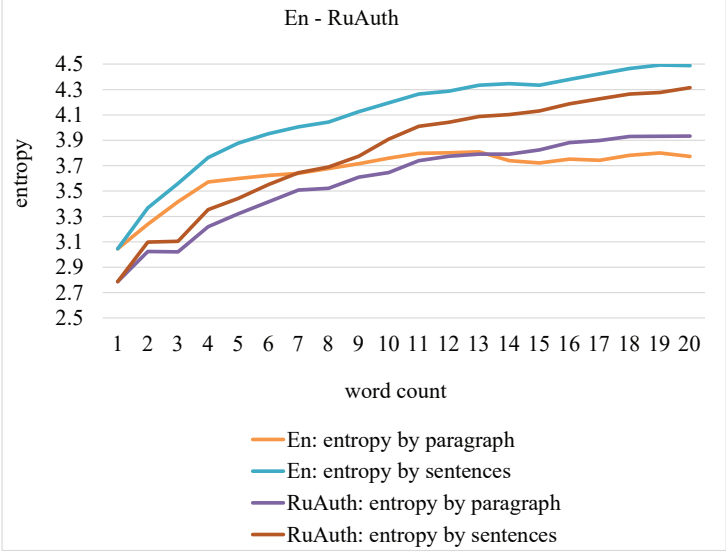


Fig. 1. Comparison of entropy calculations for En and RuAuth

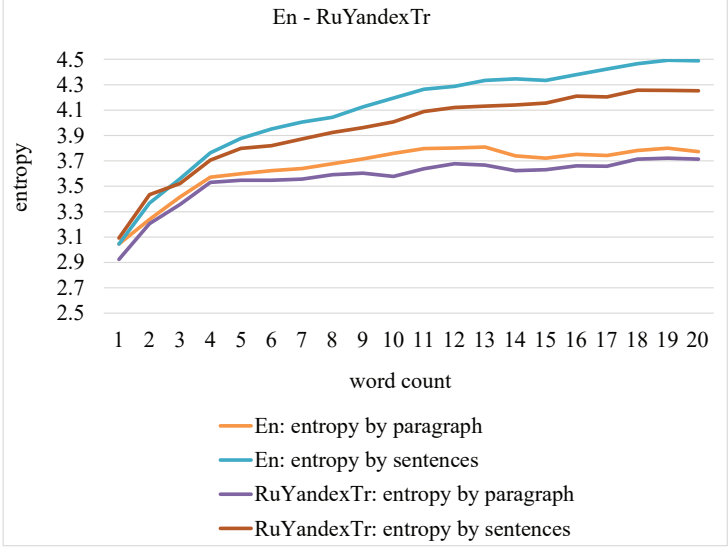


Fig. 2. Comparison of entropy calculations for En and RuYandexTr

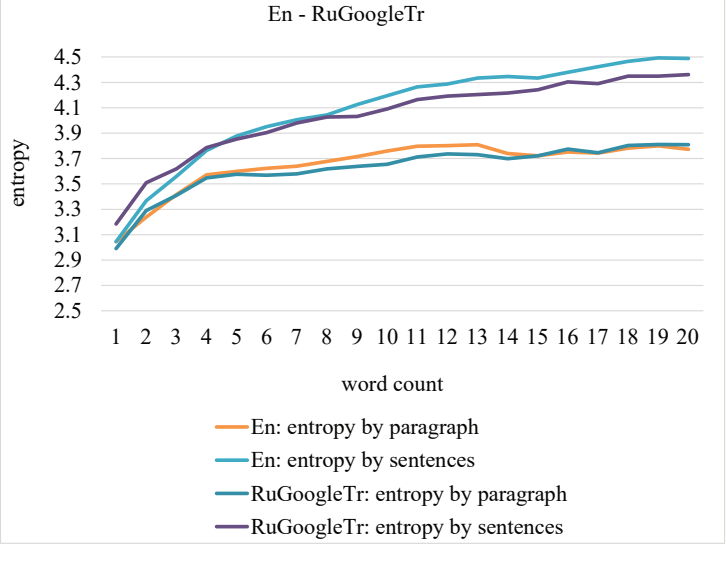


Fig. 3. Comparison of entropy calculations for En and RuG

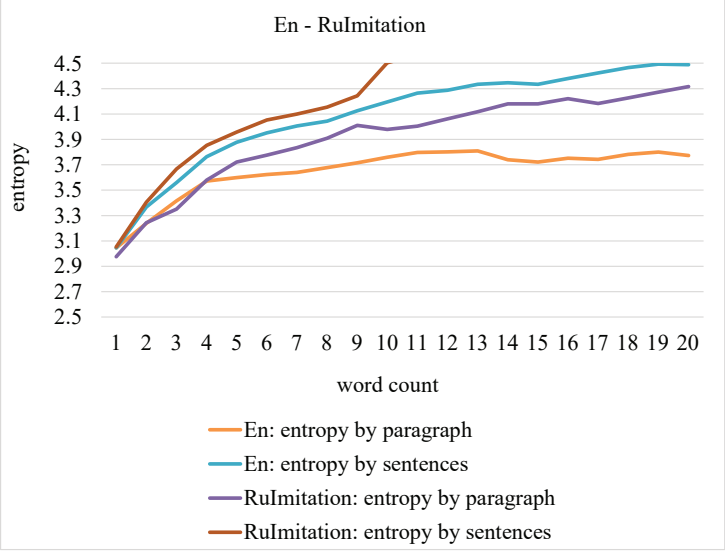


Fig. 4. Comparison of entropy calculations for En and RuImitation

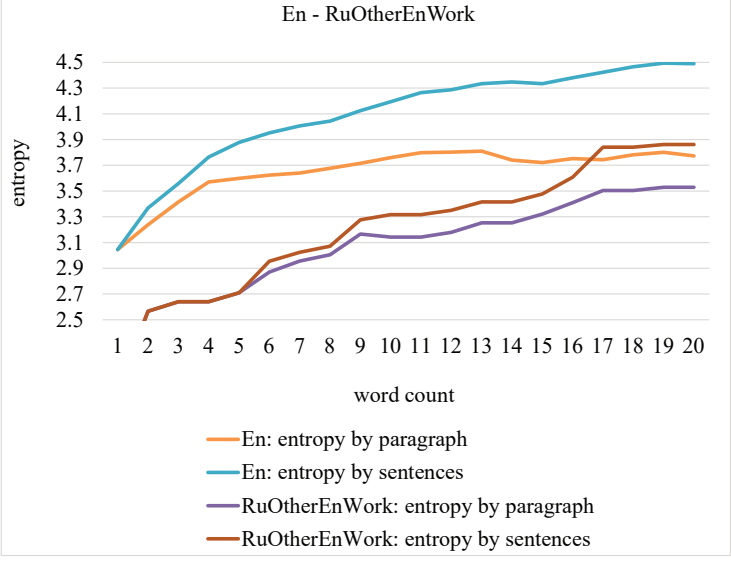


Fig. 5. Comparison of entropy calculations for En and RuOtherEnWork

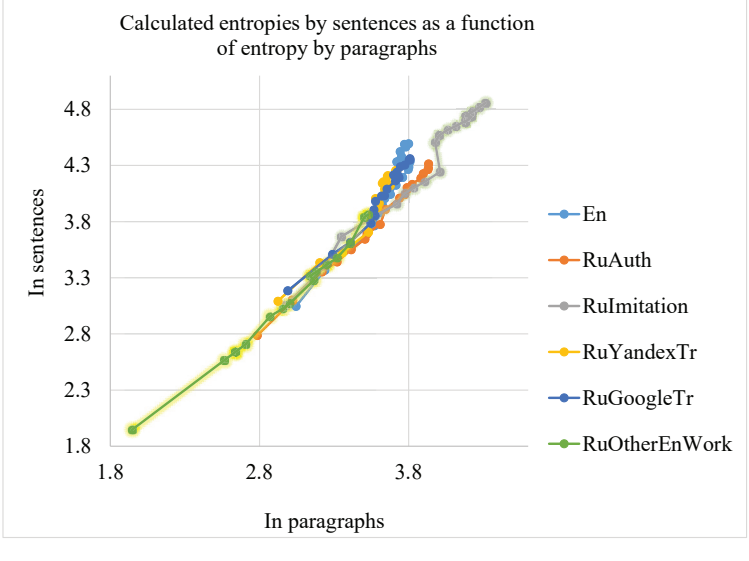


Fig. 6. Calculated entropies by sentences as a function of entropy by paragraphs

Fig. 6 shows that the accumulated differences between the pairs of chart points are significantly greater for another work (RuOtherEnWork) and the imitation of statistics (RuImitation) than those for the translation options (RuAuth, RuYandexTr, RuGoogleTr) and the original text (En).

### 5. 2. Calculating distances between the sets of texts' entropies

*Hamming distance.* By taking the entropy by paragraphs on the abscissa, and the entropy by sentences on the ordinate, we calculated the Hamming distance from the original text (En) to other texts for all sequences of keywords. It is noted that the authorized translation (RuAuth), unlike the original text (En), uses frequency words in a different way in the interval of calculations for sequences of 1–10 words, which is shown in Fig. 1, so we additionally calculated the Hamming distance for sequences of 11–20 words (Table 2).

Table 2

Hamming distance

| *i\j* | Ru-Auth | RuImi-tation | RuOther-EnWork | RuYan-dexTr | RuGoog-leTr |
|---|---|---|---|---|---|
| En (1–20 coor-dinates) | 8.953 | 9.786 | 29.362 | 4.875 | 2.698 |
| En (11–20 coor-dinates) | 3.155 | 7.369 | 11.924 | 3.012 | 1.511 |

*Cartesian distance.* When calculating the Cartesian distance between the original text and other texts, there is no pronounced difference between the fake texts and translations, so we calculated the comparison of the authorized translation (RuAuth) with other texts.

Table 3

Cartesian distance for texts' entropies

| *i\j* | En | Ru-Auth | RuIm-itation | RuOth-erEnWork | RuYan-dexTr | RuGoo-gleTr |
|---|---|---|---|---|---|---|
| En | 0 | 1.584 | 1.814 | 4.616 | 1.903 | 1.460 |
| Ru-Auth | 1.584 | 0 | **2.738** | **3.584** | 1.252 | 1.449 |

*Distance between the centers of masses.* The coordinates of the mass centers of the sets of entropies in the axes of paragraphs/sentences were calculated as the mean arithmetic along each axis. Then we calculated the distances between the centers of mass of the sets of texts' entropies according to formula (1) at $r$=2 (for greater radicality).

The distances between the original text and other texts, the authorized translation and other texts, the Yandex translation and other texts were calculated; the relative distance was computed by dividing the resulting distance by the distance to the original text (Table 4).

Table 4

Distance between the centers of mass of the sets of texts' entropies

| *i**j* | En | RuAuth | RuImitation | RuOtherEnWork | RuYandexTr | RuGoogleTr |
|---|---|---|---|---|---|---|
| En | 0 | 0.296 | 0.338 | 1.059 | 0.168 | 0.062 |
| RuAuth | 0.296 | 0 | 0.604 | 0.786 | 0.149 | 0.237 |
| RuAuth (relative) | 1 | 0 | **2.037** | **2.651** | 0.504 | 0.799 |
| RuYandexTr | 0.168 | 0.149 | 0,501 | 0.891 | 0 | 0.106 |
| RuYandexTr (relative) | 1 | 0.891 | **2.987** | **5.314** | 0 | 0.632 |

*Distance between geometric centers.* The distance between the texts is calculated using the mean geometric, that is, by taking the root of power $n$ from the product of $n$-elements (in this case, $n$=20). The calculation results are given in Table 5.

Table 5

Calculating distances between the geometric centers of the sets of texts' entropies

| *i**j* | En | RuAuth | RuImitation | RuOtherEnWork | RuYandexTr | RuGoogleTr |
|---|---|---|---|---|---|---|
| En | 0 | 0.306 | 0.320 | 1.091 | 0.161 | 0.055 |
| RuAuth | 0.306 | 0 | 0.600 | 0.806 | 0.163 | 0.254 |
| RuAuth (relative) | 1 | 0 | **1.956** | **2.630** | 0.531 | 0.828 |
| RuYandexTr | 0.225 | 0.172 | 0.676 | 1.290 | 0 | 0.148 |
| RuYandexTr (relative) | 1 | 0.765 | **3.009** | **5.737** | 0 | 0.660 |

*Distance between the centers of parametric means.* To smooth out the unfavorable start of the series, the weights were calculated by deriving the parametric mean.

We calculated the centers of parametric means for all series of coordinates along the paragraph/sentence axes. The data that are given in Table 6 were acquired from formula (1) at $r$=1.

Table 6

Calculating distances between the centers of parametric means

| *i**j* | En | RuAuth | RuImitation | RuOtherEnWork | RuYandexTr | RuGoogleTr |
|---|---|---|---|---|---|---|
| En | 0 | 0.182 | 0.435 | 0.903 | 0.197 | 0.088 |
| RuAuth | 0.182 | 0 | 0.602 | 0.736 | 0.129 | 0.137 |
| RuAuth (relative) | 1 | 0 | **3.303** | **4.044** | 0.708 | 0.753 |
| RuYandexTr | 0.197 | 0.129 | 0.632 | 0.706 | 0 | 0.109 |
| RuYandexTr (relative) | 1 | 0.654 | **3.208** | **3.582** | 0 | 0.553 |

Thus, in some experiments, we obtained the clear-cut boundaries of intervals between the actual and "fake" translations while in other experiments these intervals overlap.

Of all the calculations of distances between the texts, the calculation of distances between the centers of the parametric means (Table 6) stands out. A feature was revealed: when calculating the relative distance between the centers of the parametric means of the authorized translation and other texts, the Yandex translation and other texts, the value for fake texts exceeded 3, for translations – less than 1.

## 6. Discussion of results of studying the possibility of calculating the proximity of texts using Renyi entropy

The poor difference between the fake translations and actual translations in some distances' calculations can be explained by the difference in the use of frequency words in each text. For example, consider that the authorized translation (RuAuth), as opposed to the original text (En), uses frequency words in a different way in the interval of calculations for sequences of 1–10 words. Thus, when calculating the distances between the centers of the masses of the sets of texts' entropies (Table 4), it is preferable to separate translations from fakes by computing relative distances from the center of the mass of the Yandex translation to the centers of the mass of other series. The same conclusion can be drawn from the calculations of distances between the geometric centers of the sets of texts' entropies (Table 5). At the same time, good results were demonstrated by the calculations of relative distances between the centers of parametric means (Table 6). When searching for relative distances between the authorized translation and other texts, the Yandex translation and other texts, we received similar results with the values for the "fake" texts exceeding 3, for translations – less than 1. We believe that the favorable component in these calculations was the introduction of weights that correct the discrepancy in the frequency of keyword occurrence between the original text and the authorized translation.

The main advantage of using Renyi entropy in the system for identifying patterns in polylingual texts, when calculating the relative distances between the centers of parametric means, is performance speed. Unlike neural networks, which take a lot of time to be trained, and are characterized by the high cost of hardware while requiring high-skilled specialists, our solution can be deployed on an average computer provided there is a free php-handler.

The advantages also include the scalability of the system when a similar algorithm can be used to adjust the execution of calculations for other language pairs with the connection of the appropriate Porter stemmers.

The use of Renyi entropy, which is employed in a number of "geometric" applications, seems more acceptable than using Shannon's classic approach, due to the additional parameter and generalization for the Kullback-Leibler distances. The results reported here are easily reproduced for other languages; the experiment does not require large hardware and software costs.

At present, the study limitation is that the system is applicable for the English-Russian language pair, however, other language pairs are planned to be added soon.

The caveat of our research is the lack of automatic generation of key series of high-frequency words. That is, although we use the high-frequency word counting service, the exclusion of stop words that have too many synonyms in translation dictionaries or have little meaning is done manually. In the future, this task can be solved by analyzing dictionaries and identifying certain patterns.

The current study may be advanced by connecting to the system of identification of patterns in polylingual texts of a web crawler, which would automatically select similar texts from the Internet and pass them on for entropy calculation.

We note that one of the difficulties that an algorithm developer may face in the future relates to the stemming for each particular language.

## 7. Conclusions

1. Counting entropies for "sentence" and "paragraph" patterns is informative enough to distinguish fake texts from real ones.

2. It has been established that the proximity of texts is best determined by calculating the relative distances between the centers of parametric means. When searching for the relative distances between the authorized translation and other texts, the Yandex translation and other texts, we received similar results with values for the "fake" texts exceeding 3, for translations – less than 1. A favorable component in our calculation was the introduction of weights that correct the discrepancy in the frequency of keyword occurrence between the original text and the authorized translation.

## References


1. Imran, M. (2020). Advantages of Neural Networks - Benefits of AI and Deep Learning. Folio3. Available at: https://www.folio3.ai/blog/advantages-of-neural-networks/
2. Hanlon, J. (2017). Why is so much memory needed for deep neural networks? Graphcore. Available at: https://www.graphcore.ai/posts/why-is-so-much-memory-needed-for-deep-neural-networks
3. Yu, J., Chen, R., Xu, L., Wang, D. (2019). Concept extraction for structured text using entropy weight method. 2019 IEEE Symposium on Computers and Communications (ISCC). doi: https://doi.org/10.1109/iscc47284.2019.8969759
4. Shi, Y., Lei, L. (2020). Lexical Richness and Text Length: An Entropy-based Perspective. Journal of Quantitative Linguistics, 1–18. doi: https://doi.org/10.1080/09296174.2020.1766346
5. Kouyama, N., Köppen, M. (2019). Entropy Analysis of Questionable Text Sources by Example of the Voynich Manuscript. Soft Computing in Data Science, 3–13. doi: https://doi.org/10.1007/978-981-15-0399-3_1
6. Authorship Proven by Mathematics Burrow's Delta helps determine the real author of And Quiet Flows the Don. IQ: Research and Education Website. Available at: https://iq.hse.ru/news/367813734.html
7. Bubnov, V. A., Survilo, A. V. (2016). Comparative Computer Analysis of the Text the Novel «The Quiet Don» with Texts of Four Fyodor Kryukov's Stories. Vestnik Rossiyskogo universiteta druzhby narodov. Seriya: Informatizatsiya obrazovaniya, 1, 60–69. Available at: https://cyberleninka.ru/article/n/sravnitelnyy-kompyuternyy-analiz-teksta-romana-tihiy-don-s-tekstami-chetyreh-rasskazov-fyodora-kryukova/viewer
8. Zhao, Y., Zhang, J., Zong, C., He, Z., Wu, H. (2019). Addressing the Under-Translation Problem from the Entropy Perspective. Proceedings of the AAAI Conference on Artificial Intelligence, 33, 451–458. doi: https://doi.org/10.1609/aaai.v33i01.3301451
9. Bromiley, P., Thacker, N., Bouhova-Thacker, E. (2010). Shannon Entropy, Renyi Entropy, and Information. TINA. Available at: https://www.academia.edu/32317926/Shannon_Entropy_Renyi_Entropy_and_Information
10. Investigation of distances between sets of entropies. Available at: http://102030.kz/entropyR2.php
11. Word and Character Counter. Available at: https://countwordsfree.com/
12. Russian stemming algorithm. Available at: http://snowball.tartarus.org/algorithms/russian/stemmer.html
13. The Porter Stemming Algorithm. Available at: https://tartarus.org/martin/PorterStemmer/
14. XAMPP Installers and Downloads for Apache Friends. Available at: https://www.apachefriends.org/index.html