

Nonparametric Methods of Authorship Attribution in Ukrainian Literature

Dmitriy Klyushin¹, Yulia Nykyporets¹

¹ Taras Shevchenko National University of Kyiv, Ukraine, 03680, Kyiv, Akademika Glushkova Avenue 4D

Abstract

The paper presents the results of the comparison of two nonparametric methods of authorship identification of the Ukrainian literature texts. The paper describes the implementation of the corresponding methods based on the Klyushin–Petunin tests and its simplified version. The method of n-gram selection is applied. For testing a collection of texts up to 200,000 characters from 10 authors was used. As a result of carrying out the test, it was found out that the simplified test appears to be more sensitive and specific, and monograms and bigrams in opposite to trigrams provide clear detection of authorship.

Keywords ¹

authorship identification, nonparametric methods, p-statistics, confidence intervals

1. Introduction

It is generally accepted that the problem of authorship attribution traditionally is solved by methods that use two paradigms: classification and similarity [1, 2]. The former approach entails the using of training sets of a priori known authors, and the latter approach uses a metrics of similarity of two tests and selects the authors by the nearest neighbor method. In our opinion, this division is quite artificial because similarity metrics are wide used for classification [3, 4]. In this paper, we propose new method of authorship attribution of Ukrainian writers using an original statistical similarity measure. As a basis for statistical analysis, we selected n-grams that are widely used as a characteristic of author style.

The use of letter combinations (n-grams) as a stylistic feature for recognition by the author was first suggested by B. Kjell. This approach was developed by E. Stamos, P. Juola, Y. Orlov, K. Osminin, L. Borisov, D. Shalimov, J. Peng, V. Keselj, C. Boughaci, D. Klyushin etc. In these works, different approaches to automatic identification of authorship were considered and experiments were carried out to assess the accuracy and effectiveness of the proposed methods for attribution of literary texts written in different languages.

The Kjells idea was to compute the relative frequencies of letter pairs within texts samples and compare them using neural network classifier [5, 6]. Stamos et al. considered computational issues of the authorship identification and discussed methodologies and criteria for authorship attribution [7, 8]. Juola [9] surveyed the history and state-of-the-art of the discipline authorship identification, presented some comparative results and concluded that current methods are difficult to apply, their rate of errors are quite unknown and there are very little widely approved practices in this area.

Orlov and his co-authors proposed a statistical test for classification of literature texts using information about distribution functions [10, 11]. Orlov et al. [11] investigated distributions of distances between distributions of trigrams, estimated the accuracy of the classification using the frequencies of combinations of letters depending on the length of the text, and estimated of the specificity and sensitivity of the proposed method. This method may be considered as a first attempt to solve the problem of authorship identification using rigorous statistical test. This approach was further developed in papers [12–14]. These issues are at the focus of investigation during last years.


¹ICTERI-2021, Vol II: Workshops, September 28 – October 2, 2021, Kherson, Ukraine

EMAIL: dokmed5@gmail.com (Dmitriy Klyushin); Nykyporets.y@gmail.com (Yulia Nykyporets)

ORCID: 0000-0003-4554-1049 (Dmitriy Klyushin)



© 2020 Copyright for this paper by its authors.

Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).  CEUR Workshop Proceedings (CEUR-WS.org)

Wright [15] investigated Enron Email Corpus containing texts of 176-author consisting of 2.5 million-word and successfully tested the accuracy of n-grams for identifying the anonymous authors of email.

Grieve et al. [16] used n-gram tracing to identify the authorship of the Bixby Letter that hypothetically is written by President Abraham Lincoln. The authors claimed that the true author was John Hay, Lincoln's personal secretary. It is interesting that the method of n-gram tracing shown good results despite that the letter consists of only 139 words and the standard methods were ineffective in this case.

Singh and Murthy [17] used a filter removing noun and verb groups and compared the filtered n-grams with the traditional or unfiltered n-grams for authorship attribution. Due to this filtering the authors constructed new n-grams and found that this improved the performance.

Georgieva-Trifonova and Duraku [18] investigated feature selection methods in terms of the accuracy and F-measure of text classification using N-grams of words, different classifiers and different datasets. They proved that to obtain high performance of classification it is necessary implement pre-processing steps.

Ramezani [19] proposed a solution of text classification problem introducing a new measure for identifying important words and using the Term Frequency-Inverse Document Frequency (TF_IDF) scheme. He achieved accuracy that was 0.902 for an English dataset and 0.931 for a Persian dataset. This fact allowed him to claim that his approach is a language-independent one.

Romanov et al. [20] described the identification of authors of Russian-language texts using support vector machine (SVM) and deep neural networks with long short-term memory (LSTM) and convolutional neural networks (CNN). The authors claimed that due to successful feature selection the SVM provided the best results: average accuracy of SVM was 96% in opposite to 93% accuracy of deep neural networks.

Kosmajac and Kešelj [21] extended experiments on authorship attribution conducted by Gamallo et al. [22] and compared texts in 41 European languages using a distance measure that was proven to work well in authorship attribution tasks.

All these papers stressed that one of the most desirable properties of the method of the authorship identification is its independence on a language of an identified text. The variety of considered languages proves that this approach is a language-independent one.

In the frame of this approach, Klyushin et al. [23, 24] proposed an original rigorous statistical test for authorship detection using estimations of homogeneity of distributions of the n-grams (monograms, bigrams and trigrams) in text of literature texts in English and Russian. This method is based on the Klyushin–Petunin test [25]. In [23] 100 texts of 11 Russian authors were generated and the following accuracy was obtained: 62.5% for monograms, 87.5 for bigrams, and 90.6% for trigrams. In [24] more than 800 texts of 16 English authors were used for testing. The accuracy was 81% for monograms, 85% for bigrams and 81% for trigrams.

In this work, the attribution of authorship of Ukrainian literary texts is carried out. The author's identification method consists in testing the statistical hypothesis that the text belongs to a certain text corpus using the measure of the homogeneity of the distribution of n-grams in training and test samples.

2. Theoretical background

According to the Hill's assumption $A_{(n)}$ [26], if a sample of exchangeable random values x_1, x_2, \dots, x_n has no ties than

$$P\left(x_{n+1} \in \left(x_{(i)}, x_{(j)}\right)\right) = p_{ij} = \frac{j-i}{n+1}, \quad j > i,$$

where x_{n+1} follows the same distribution function and $x_{(i)}$ is the i -th order statistics.

Let samples $x = (x_1, x_2, \dots, x_n)$ and $y = (y_1, y_2, \dots, y_m)$ follow distributions F_1 and F_2 . Let $A_{ij}^{(k)} = \{x_{(i)} < y_k < x_{(j)}\}$ and $h_{ij} = \# A_{ij} / m$. Consider the Wilson confidence interval for the probability of $A_{ij}^{(k)}$:

$$p_{ij}^{(1)} = \frac{h_{ij}m + z^2/2 - z\sqrt{h_{ij}(1-h_{ij})m + z^2/4}}{m + z^2}, p_{ij}^{(2)} = \frac{h_{ij}m + z^2/2 + z\sqrt{h_{ij}(1-h_{ij})m + z^2/4}}{m + z^2}.$$

The lower and upper bounds of the Wilson confidence interval $I_{ij}^{(n,m)} = (p_{ij}^{(1)}, p_{ij}^{(2)})$ depend on the parameter z . If $z = 3$ than the significance level of $I_{ij}^{(n,m)}$ is less than 0.05 [25]. Put $N = (n-1)n/2$ and $L = \#\left\{p_{ij} = \frac{j-i}{n+1} \in I_{ij}^{(n,m)}\right\}$. Then, $\rho(x, y) = L/N$ is the probability that x and y follow the same distribution function. We shall refer it as p -statistics. Since the p -statistics is a binomial proportion, then constructing the Wilson confidence interval $I = (p_1, p_2)$ for the p -statistics we can formulate the following decision rule: if $p_2 \leq 0.95$ then the null hypothesis is accepted, else it is rejected.

As far the samples x and y in the test play different roles (the sample x is ordered and used for construction of a variational series and the sample y is sieved through intervals formed by ordered statistics $x_{(i)}$), the p -statistics is nonsymmetrical, i.e. $\rho(x, y) \neq \rho(y, x)$. It is easy to see, that we can construct a symmetrical p -statistics by averaging $\rho(x, y)$ and $\rho(y, x)$: $\rho^*(x, y) = \frac{1}{2}(\rho(x, y) + \rho(y, x))$.

In the original version of the Klyushin–Petunin all the pairs $x_{(i)}$ and $x_{(j)}$ are exhausted. Therefore, the algorithmic complexity of this test is $O(n^2)$. But, if we use only M random intervals $(x_{(i)}, x_{(j)})$ where M is more than 100 but much less than $n(n-1)/2$, then the algorithmic complexity may be reduced to $O(n)$ because of the well-known fact that the relative frequency in the Bernoulli schemes is stabilized rather quickly.

Let us select M times random numbers i and j such that $i < j \leq n$. Find the relative frequency of the event $A_{ij}^{(k)} = \{x_{(i)} < y_k < x_{(j)}\}$. Then, construct the Wilson confidence interval $I_{ij}^{(n,m)} = (p_{ij}^{(1)}, p_{ij}^{(2)})$ and compute $L = \#\left\{\frac{j-i}{n+1} \in I_{ij}^{(n,m)}\right\}$. Find $\rho(x, y) = L/M$. Constructing the Wilson confidence interval $I = (p_1, p_2)$ for the p -statistics, we can formulate the following decision rule: if $p_2 \leq 0.95$ then the null hypothesis is accepted, else it is rejected. According to practical recommendations, we used 100 trials.

In this work, attribution of authorship of literary works written in the Ukrainian language is carried out. The author's identification method consists in testing the statistical hypothesis that a text belongs to a certain corpus using a homogeneity measure (p -statistics) of the distributions of n -grams.

3. Experiments and results

Testing was carried out on a set of texts of 10 Ukrainian writers: 1) Yuriy Andrukhovych, 2) Ivan Bagryaniy, 3) Lyubko Deresh, 4) Oleksandr Dovzhenko, 5) Serhiy Zhadan, 6) Irena Karpa, 7) Mikhaïlo Kotsyubinsky, 8) Lyuko Dashvar, 9) Ivan Nechui-Levytskiy, 10) Osip Turyansky. We used 200,000 first characters of the texts, each of which was divided into K parts, from which samples of the corresponding size were selected to determine the frequencies of n -gram.

The accuracy of the author's identification depends on the number of fragments into which the text is divided. With a decrease in their number, it drops sharply because the relational frequency of n -grams worse approximate a corresponding probability. Fig. 1 shows the results for monograms at $K = 30$. In this case, both the original and the simplified p -statistics also give equally accurate results. Fig. 2 shows the results for bigrams. The best results were obtained by splitting the text into 15 parts.

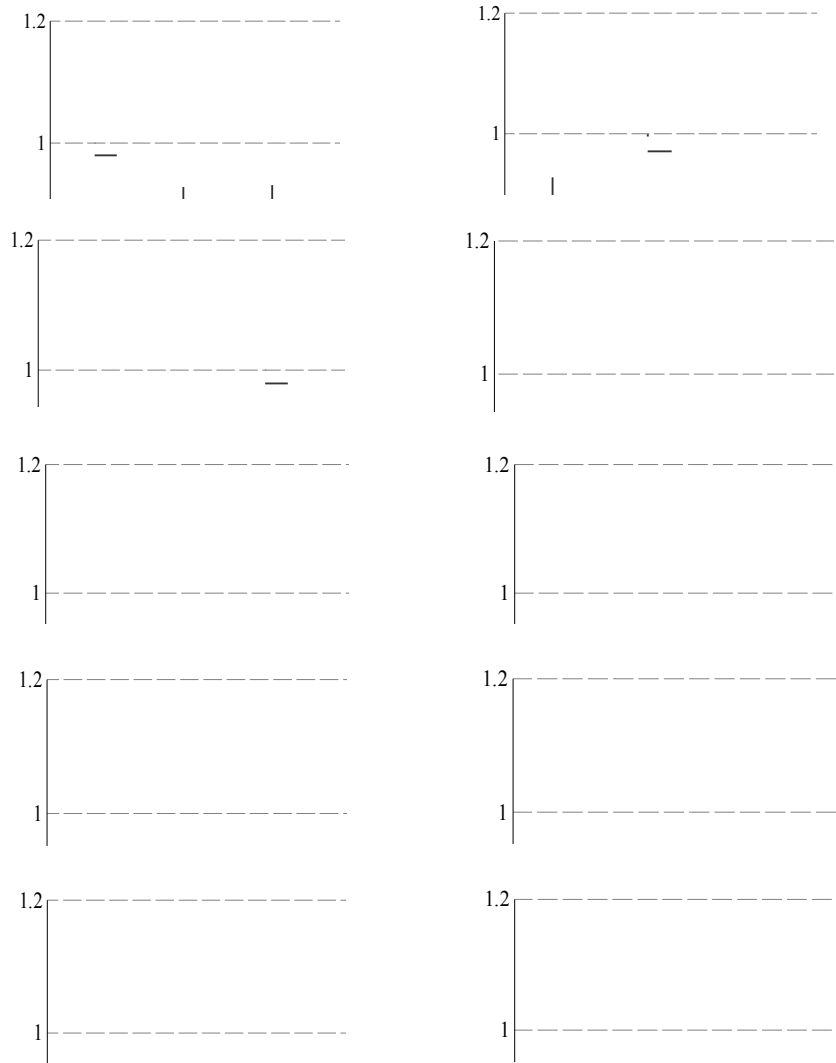


Figure 1: Simplified p -statistics for monograms and their confidence limits

Visual analysis of the graphs shows that both versions of p -statistics almost equally accurately identify the author, but the simplified p -statistics showed slightly better results when comparing the texts of the same author, i.e. it may have a higher specificity. Fig. 3 shows the results for trigrams with $K = 8$. Both tests equally well identify different authors, but for trigrams it is better to use longer text, since the number of trigrams is not enough for a reliable result because for stabilization of the relative frequency we must have many observations.

As we see, the homogeneity measure of monograms, bigrams and trigrams in texts of 10 selected Ukrainian authors is quite stable and demonstrates expected properties. First, it attains a maximum when we compare texts of the same author. Second, it allows clear distinguishing the texts of the different authors. However, it has several deficiencies: the homogeneity measure has very narrow confidence interval and the maximum of p -statistics for bigrams is more slightly expressed than for monogram and in case of trigram the p -statistics becomes a constant.



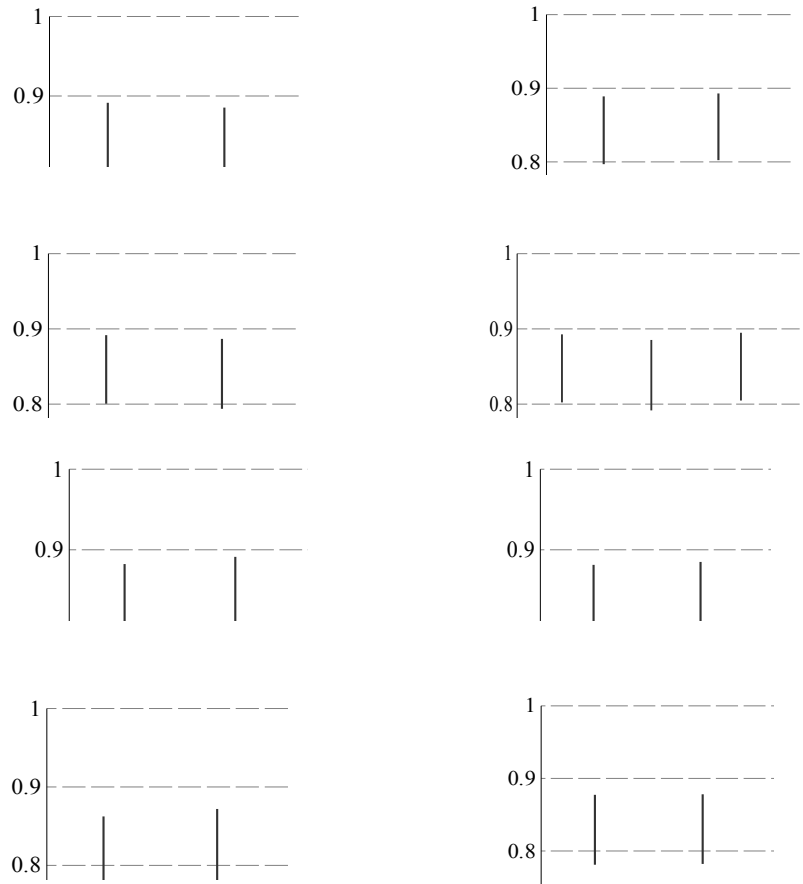


Figure 2: Simplified p -statistics for bigrams and their confidence limits

As we see in Figures 1-3 the pairwise homogeneity measure varies very slightly for monograms, bigrams, and trigrams. For all pairs consisting of different authors is bounded by 0.75 below and 0.83 upper (see Tables 1 and 2). Note, that the homogeneity measure between texts of the same author is greater than between texts of different authors. It is interesting, that the maximum of the graphs of the homogeneity measure corresponding to the comparison between the texts of the same author is the most clear for monograms, less clear for bigrams and almost invisible for trigrams.



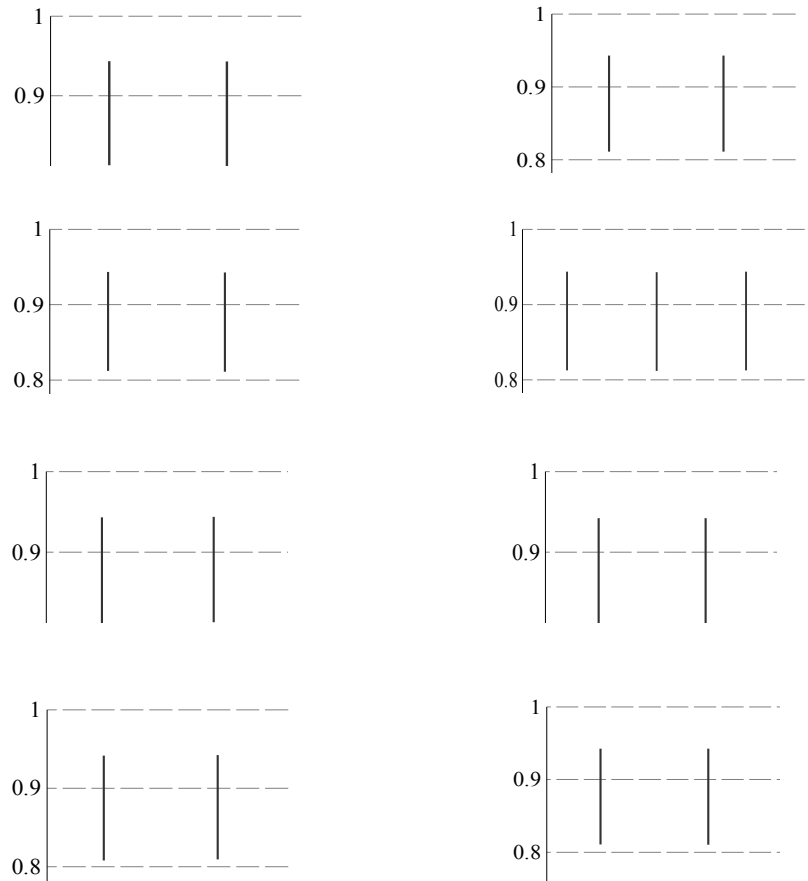


Figure 3: Simplified p statistics for trigrams and their confidence limits

The nature of the graphs corresponding to monograms, bigrams, and trigrams is dual. From one side, we see very flat graphs in all cases. From the other side, the graphs demonstrate different behaviors. The graph at the Figure 1, corresponding to monograms, is quite unstable in comparing with other graphs. The graph at the Figure 2, corresponding to bigrams, has moderate variance. And the graph at the Figure 3 is almost constant.

Therefore, to identify the authorship with statistical significance we cannot rely only on the graphs. We must analyze the confidence intervals for the homogeneity measures and use the following decision rule: if the upper confidence bound for the homogeneity measure is greater than 0.95 the texts are considered as written by the same author; if the upper confidence bound for the homogeneity measure is less than 0.95 the texts are considered as written by different authors.

Table 1
P-statistics for bigrams

	1	2	3	4	5	6	7	8	9	10
1	0.83	0.80	0.80	0.79	0.79	0.80	0.78	0.78	0.76	0.78
2	0.80	0.83	0.80	0.80	0.78	0.80	0.80	0.78	0.78	0.78
3	0.80	0.79	0.82	0.78	0.78	0.80	0.78	0.77	0.76	0.77
4	0.79	0.80	0.79	0.83	0.79	0.78	0.79	0.78	0.78	0.78
5	0.80	0.79	0.79	0.79	0.82	0.79	0.78	0.79	0.76	0.77
6	0.80	0.79	0.80	0.78	0.78	0.82	0.77	0.77	0.75	0.77
7	0.78	0.79	0.78	0.79	0.77	0.78	0.82	0.77	0.78	0.78
8	0.78	0.79	0.78	0.78	0.78	0.78	0.78	0.82	0.77	0.76
9	0.76	0.77	0.76	0.77	0.75	0.75	0.77	0.76	0.81	0.75
10	0.78	0.78	0.77	0.78	0.76	0.77	0.78	0.76	0.76	0.82

Table 2
Simplified p-statistics for bigrams

	1	2	3	4	5	6	7	8	9	10
1	0.82	0.79	0.79	0.78	0.79	0.79	0.78	0.78	0.75	0.78
2	0.79	0.82	0.79	0.79	0.78	0.79	0.79	0.77	0.78	0.78
3	0.79	0.79	0.82	0.77	0.78	0.80	0.77	0.78	0.76	0.76
4	0.79	0.80	0.78	0.83	0.79	0.78	0.79	0.78	0.77	0.78
5	0.79	0.79	0.79	0.79	0.81	0.79	0.77	0.78	0.75	0.76
6	0.79	0.78	0.80	0.77	0.77	0.82	0.77	0.77	0.74	0.76
7	0.78	0.79	0.78	0.78	0.76	0.77	0.82	0.77	0.77	0.77
8	0.78	0.78	0.78	0.78	0.77	0.77	0.78	0.82	0.77	0.76
9	0.75	0.76	0.76	0.76	0.74	0.75	0.77	0.76	0.80	0.75
10	0.77	0.78	0.77	0.77	0.75	0.77	0.78	0.76	0.75	0.82

Table 3
Confidence intervals for p-statistics for monograms (K=30)

	1	2	3	4	5	6	7	8	9	10
1	(0.68, 0.99)	(0.55, 0.97)	(0.55, 0.97)	(0.48, 0.95)	(0.56, 0.98)	(0.53, 0.97)	(0.52, 0.97)	(0.48, 0.95)	(0.47, 0.95)	(0.52, 0.97)
2	(0.56, 0.98)	(0.68, 0.99)	(0.51, 0.96)	(0.57, 0.98)	(0.51, 0.96)	(0.52, 0.97)	(0.56, 0.98)	(0.47, 0.95)	(0.53, 0.97)	(0.51, 0.96)
3	(0.56, 0.98)	(0.50, 0.96)	(0.68, 1.00)	(0.45, 0.94)	(0.46, 0.94)	(0.57, 0.98)	(0.50, 0.96)	(0.48, 0.95)	(0.49, 0.96)	(0.45, 0.94)
4	(0.50, 0.96)	(0.58, 0.98)	(0.49, 0.95)	(0.68, 0.99)	(0.50, 0.96)	(0.50, 0.96)	(0.47, 0.95)	(0.55, 0.98)	(0.48, 0.95)	(0.55, 0.98)
5	(0.56, 0.98)	(0.48, 0.95)	(0.51, 0.96)	(0.51, 0.96)	(0.65, 0.99)	(0.47, 0.95)	(0.48, 0.95)	(0.45, 0.94)	(0.45, 0.94)	(0.50, 0.96)
6	(0.53, 0.97)	(0.51, 0.96)	(0.52, 0.97)	(0.49, 0.96)	(0.48, 0.95)	(0.68, 0.99)	(0.49, 0.96)	(0.50, 0.96)	(0.47, 0.95)	(0.51, 0.96)
7	(0.50, 0.96)	(0.56, 0.98)	(0.49, 0.96)	(0.51, 0.96)	(0.47, 0.95)	(0.50, 0.96)	(0.68, 1.00)	(0.48, 0.95)	(0.54, 0.97)	(0.52, 0.97)
8	(0.46, 0.94)	(0.49, 0.95)	(0.48, 0.95)	(0.48, 0.95)	(0.45, 0.94)	(0.51, 0.96)	(0.47, 0.95)	(0.66, 0.99)	(0.49, 0.96)	(0.43, 0.93)
9	(0.48, 0.95)	(0.49, 0.96)	(0.44, 0.93)	(0.46, 0.94)	(0.43, 0.93)	(0.49, 0.96)	(0.51, 0.96)	(0.47, 0.95)	(0.65, 0.99)	(0.46, 0.94)
10	(0.50, 0.96)	(0.53, 0.97)	(0.46, 0.94)	(0.53, 0.97)	(0.49, 0.96)	(0.49, 0.96)	(0.51, 0.96)	(0.43, 0.93)	(0.43, 0.93)	(0.68, 0.99)

Table 4Confidence intervals for simplified p-statistics for monograms ($K=30$)

	1	2	3	4	5	6	7	8	9	10
1	(0.97, 1.00)	(0.83, 0.92)	(0.84, 0.93)	(0.79, 0.89)	(0.83, 0.92)	(0.82, 0.91)	(0.79, 0.89)	(0.74, 0.86)	(0.75, 0.86)	(0.79, 0.89)
2	(0.83, 0.92)	(0.97, 0.99)	(0.77, 0.88)	(0.85, 0.93)	(0.78, 0.88)	(0.80, 0.90)	(0.84, 0.93)	(0.76, 0.87)	(0.80, 0.90)	(0.81, 0.91)
3	(0.83, 0.9)	(0.78, 0.88)	(0.97, 1.00)	(0.75, 0.86)	(0.77, 0.87)	(0.85, 0.93)	(0.77, 0.88)	(0.76, 0.87)	(0.73, 0.85)	(0.72, 0.84)
4	(0.77, 0.88)	(0.85, 0.94)	(0.75, 0.86)	(0.97, 0.99)	(0.78, 0.88)	(0.77, 0.88)	(0.82, 0.91)	(0.76, 0.87)	(0.75, 0.86)	(0.83, 0.92)
5	(0.82, 0.91)	(0.77, 0.87)	(0.76, 0.87)	(0.78, 0.89)	(0.93, 0.98)	(0.75, 0.86)	(0.75, 0.86)	(0.72, 0.84)	(0.70, 0.82)	(0.77, 0.88)
6	(0.82, 0.91)	(0.81, 0.90)	(0.84, 0.93)	(0.78, 0.88)	(0.75, 0.86)	(0.97, 1.00)	(0.77, 0.88)	(0.77, 0.88)	(0.75, 0.86)	(0.78, 0.88)
7	(0.78, 0.88)	(0.83, 0.92)	(0.76, 0.87)	(0.81, 0.91)	(0.75, 0.86)	(0.76, 0.87)	(0.97, 1.00)	(0.75, 0.86)	(0.82, 0.91)	(0.81, 0.91)
8	(0.74, 0.85)	(0.76, 0.87)	(0.76, 0.87)	(0.76, 0.87)	(0.70, 0.82)	(0.77, 0.88)	(0.75, 0.86)	0.95, 0.99	(0.74, 0.86)	(0.69, 0.81)
9	(0.74, 0.85)	(0.79, 0.89)	(0.72, 0.84)	(0.74, 0.86)	(0.70, 0.82)	(0.74, 0.85)	(0.81, 0.91)	(0.74, 0.85)	(0.95, 0.99)	(0.70, 0.82)
10	(0.79, 0.89)	(0.80, 0.90)	(0.72, 0.83)	(0.82, 0.91)	(0.77, 0.88)	(0.77, 0.88)	(0.80, 0.90)	(0.69, 0.81)	(0.71, 0.83)	(0.97, 0.99)

Table 5

Confidence intervals for p-statistics for bigrams

	1	2	3	4	5	6	7	8	9	10
1	(0.69, 0.91)	(0.66, 0.89)	(0.66, 0.89)	(0.65, 0.88)	(0.65, 0.88)	(0.66, 0.89)	(0.64, 0.88)	(0.64, 0.87)	(0.61, 0.86)	(0.64, 0.87)
2	(0.66, 0.89)	(0.69, 0.91)	(0.66, 0.89)	(0.66, 0.89)	(0.64, 0.88)	(0.66, 0.89)	(0.66, 0.89)	(0.64, 0.88)	(0.64, 0.87)	(0.64, 0.88)
3	(0.66, 0.89)	(0.65, 0.88)	(0.69, 0.91)	(0.64, 0.87)	(0.64, 0.87)	(0.66, 0.89)	(0.64, 0.87)	(0.63, 0.87)	(0.62, 0.86)	(0.63, 0.87)
4	(0.65, 0.88)	(0.66, 0.89)	(0.65, 0.88)	(0.69, 0.91)	(0.65, 0.88)	(0.64, 0.88)	(0.65, 0.89)	(0.64, 0.88)	(0.64, 0.87)	(0.64, 0.88)
5	(0.66, 0.89)	(0.65, 0.88)	(0.65, 0.88)	(0.66, 0.89)	(0.69, 0.91)	(0.65, 0.88)	(0.64, 0.87)	(0.65, 0.88)	(0.61, 0.86)	(0.63, 0.86)
6	(0.66, 0.89)	(0.65, 0.88)	(0.66, 0.89)	(0.64, 0.87)	(0.64, 0.87)	(0.69, 0.91)	(0.63, 0.87)	(0.63, 0.87)	(0.61, 0.85)	(0.63, 0.87)
7	(0.64, 0.88)	(0.66, 0.89)	(0.64, 0.88)	(0.65, 0.88)	(0.63, 0.87)	(0.64, 0.87)	(0.68, 0.90)	(0.63, 0.87)	(0.64, 0.87)	(0.64, 0.87)
8	(0.64, 0.88)	(0.65, 0.88)	(0.64, 0.88)	(0.64, 0.88)	(0.64, 0.87)	(0.64, 0.87)	(0.64, 0.88)	(0.69, 0.912)	(0.63, 0.87)	(0.62, 0.86)
9	(0.61, 0.86)	(0.63, 0.87)	(0.62, 0.86)	(0.63, 0.87)	(0.60, 0.85)	(0.61, 0.86)	(0.63, 0.87)	(0.62, 0.86)	(0.67, 0.90)	(0.61, 0.86)
10	(0.64, 0.87)	(0.64, 0.87)	(0.63, 0.87)	(0.64, 0.87)	(0.62, 0.86)	(0.63, 0.87)	(0.64, 0.87)	(0.61, 0.86)	(0.62, 0.86)	(0.68, 0.90)

Table 8Confidence intervals for simplified p-statistics for trigrams ($K=8$)

	1	2	3	4	5	6	7	8	9	10
1	(0.47, 0.95)	(0.46, 0.95)	(0.47, 0.95)	(0.46, 0.95)	(0.46, 0.95)	(0.47, 0.95)	(0.46, 0.95)	(0.46, 0.95)	(0.46, 0.95)	(0.46, 0.95)
2	(0.46, 0.95)	(0.47, 0.95)	(0.47, 0.95)	(0.47, 0.95)	(0.47, 0.95)	(0.47, 0.95)	(0.47, 0.95)	(0.46, 0.95)	(0.46, 0.95)	(0.47, 0.95)
3	(0.46, 0.95)	(0.46, 0.95)	(0.47, 0.95)	(0.47, 0.95)	(0.46, 0.95)	(0.47, 0.95)	(0.46, 0.95)	(0.46, 0.95)	(0.46, 0.95)	(0.46, 0.95)
4	(0.46, 0.95)	(0.46, 0.95)	(0.46, 0.95)	(0.44, 0.95)	(0.46, 0.95)	(0.46, 0.95)	(0.46, 0.95)	(0.46, 0.95)	(0.46, 0.95)	(0.46, 0.95)
5	(0.46, 0.95)	(0.46, 0.95)	(0.46, 0.95)	(0.46, 0.95)	(0.47, 0.95)	(0.46, 0.95)	(0.46, 0.95)	(0.46, 0.95)	(0.46, 0.95)	(0.46, 0.95)
6	(0.46, 0.95)	(0.46, 0.95)	(0.47, 0.95)	(0.46, 0.95)	(0.46, 0.95)	(0.47, 0.95)	(0.47, 0.95)	(0.46, 0.95)	(0.46, 0.95)	(0.46, 0.95)
7	(0.46, 0.95)	(0.46, 0.95)	(0.47, 0.95)	(0.46, 0.95)	(0.46, 0.95)	(0.46, 0.95)	(0.46, 0.95)	(0.46, 0.95)	(0.46, 0.95)	(0.46, 0.95)
8	(0.46, 0.95)	(0.46, 0.95)	(0.46, 0.95)	(0.46, 0.95)	(0.46, 0.95)	(0.46, 0.95)	(0.46, 0.95)	(0.47, 0.95)	(0.47, 0.95)	(0.46, 0.95)
9	(0.46, 0.95)	(0.46, 0.95)	(0.45, 0.95)	(0.46, 0.95)	(0.45, 0.95)	(0.46, 0.95)	(0.46, 0.95)	(0.46, 0.95)	(0.46, 0.95)	(0.46, 0.95)
10	(0.46, 0.95)	(0.46, 0.95)	(0.46, 0.95)	(0.46, 0.95)	(0.46, 0.95)	(0.46, 0.95)	(0.46, 0.95)	(0.46, 0.95)	(0.46, 0.95)	(0.47, 0.95)

Tables 3-8 contain confidence limits for the p-statistics for bigrams. It is easy to see that 1) the p-statistics and the simplified p-statistics are varying in very narrow range, 2) all p-statistics are almost constant for all writers; 3) the variations are decreasing in the sequence monograms, bigrams, and trigrams. To estimate the specificity and sensitivity of the proposed test we applied the decision rule formulated above. As far as the test is statistical we selected the significance level. It was found that this parameter has great influence on the accuracy.

Table 3 contains confidence limits for the p-statistics for monograms were K was equal 30 at the significance level 0.05. The sensitivity of the test for monograms is 100% but the specificity is equal only 17%. However, the corresponding simplified p-statistics was appeared more informative (Table 4). Its sensitivity and specificity are equal to 100%, i.e. we correctly identified the texts of the same author and distinguish them from the texts of other authors.

Table 4 contains confidence limits for the p-statistics for bigrams at the significance level 0.05. These confidence intervals are narrower comparing with the confidence intervals for monograms and do not contain 0.95 in all the cases. This means that the sensitivity of the test for bigrams is 100% but the specificity is equal 0% (all the texts are considered as different). However, it is remarkable that increasing of the significance level up to 0.1 instead of 0.05 (substituting 0.95 with 0.90) makes this test an ideal one since its sensitivity and specificity becomes equal to 100%. As in the previous experiment, the corresponding simplified p-statistics was appeared much more informative (Table 5). Its sensitivity and specificity are equal to 100%, i.e. we correctly identified the texts of the same author and distinguish them from the texts of other authors.

Table 6 contains confidence limits for the p-statistics for trigrams with $K = 8$ at the significance level 0.05. These confidence intervals in the case are narrowest comparing with the confidence intervals for monograms and bigrams and do not contain 0.95 in all the cases. This means that the sensitivity of the test for binograms is 100% but the specificity is equal 0% (all the texts are different). Note that, as in the previous case, increasing of the significance level up to 0.1 instead of 0.05 (substituting 0.95 with 0.90) makes this test an ideal one since its sensitivity and specificity becomes equal to 100%. In opposite to the previous experiments, the corresponding simplified p-statistics was

not much more informative (Table 8). Its sensitivity is equal to 0% and specificity is equal to 100% (all the texts are considered as texts of the same author).

4. Conclusions

Both variants of p -statistics used to identify authorship of a text provide high accuracy when the volume of the text is not less than 200,000 characters and the significance level is 0.1. The simplified p -statistic gives the best results for monograms. For bigrams and trigrams original and simplified p -statistics give same results. The larger the size of n -gram, the text is broken into a smaller number of fragments. That is why the clearest results were obtained for monograms and less clear for bigrams. In the case of trigrams the p -statistics is a constant and does not allow detecting the authorship in this investigation. Thus, the use of the simplified p -statistic provides high accuracy and, at the same time, significantly reduces the time spent.

It was demonstrated that varying the significance level of the tests we can control their specificity and sensitivity. In some cases the significance level of 0.05 is a very strict demand and increasing the significance level up to 0.01 may provide much better results.

Our investigation of Ukrainian literature texts is the next in the series of investigations of Russian and English literature texts using the Klyushin-Petunin test. The results of these investigations allow claim that the proposed method is a language-independent. Its future evolution is connected with decreasing of required size of samples possibly due to some preprocessing of data (filtering n -grams and so on).

5. References

- [1] E. Stamatatos, A survey of modern authorship attribution methods. *Journal of the American Society for Information Science and Technology* 60 (2009) 538–556.
- [2] C. Saedi, M. Dras, Siamese networks for large-scale author identification, *Computer Speech & Language*, 70 (2021) 101241.
- [3] A. Bernal, K. Hospevian, T. Karadeniz., JL Lassez, Similarity Based Classification. In: R. Berthold M., Lenz HJ., Bradley E., Kruse R., Borgelt C. (eds) *Advances in Intelligent Data Analysis V. IDA 2003. Lecture Notes in Computer Science*, vol. 2810, Springer, Berlin, Heidelberg, 2007.
- [4] Y. Chen, E. Garcia, M. Gupta, A. Rahimi, L. Cazzanti Similarity-based Classification: Concepts and Algorithms, *Journal of Machine Learning Research* 10 (2009) 747–776.
- [5] B. Kjell, Authorship determination using letter pair frequency features with neural network classifiers. *Literary and Linguistic Computing* 9 (1994) 119–124.
- [6] B. Kjell, W. Woods, O. Frieder, Discrimination of authorship using visualization. *Information Processing and Management* 30 (1994) 141–150.
- [7] J. Houvardas, E. Stamatatos N-gram feature selection for authorship identification. In: Euzenat J., Domingue J. (eds) *Artificial Intelligence: Methodology, Systems, and Applications. AIMSA 2006. Lecture Notes in Computer Science* 4183:77–86.
- [8] Stamatatos E (2009) Intrinsic Plagiarism Detection Using character n -gram profiles. In: Stein B., Rosso P., Stamatatos E., Koppel M., and Agirre E. (eds.), *SEPLN 2009 Workshop on Uncovering Plagiarism, Authorship, and Social Software Misuse (PAN 09)* Universidad Politecnica de Valencia and CEUR-WS.org, September 2009. P. 38-46.
- [9] P. Juola, Authorship attribution. *Found Trends in Information Retrieval* 1 (2008) 233–334.
- [10] Y. Orlov, K. Osminin, Determination of the genre and author of a literary work by statistical methods. *Applied Informatics* 26 (2010) 95–108.
- [11] L. Borisov, Y. Orlov, K. Osminin, Identification of a text author by the letter frequency empirical distribution. *Keldysh Institute preprints* 027 (2013) (In Russian).
- [12] P. Diurdeva, E. Mikhailova, D. Shalymov, Writer identification based on letter frequency distribution. In: Tyutina T., Balandin S. (ed.), *19th Conference of Open Innovations Association (FRUCT 2016)*, pp. 24–33.

- [13] J. Peng, K. Choo, H. Ashman, Bit-level n-gram based forensic authorship analysis on social media: Identifying individuals from linguistic profiles. *Journal of Networked and Computer Applications* 70 (2016) 171–182.
- [14] D. Boughaci, M. Benmesbah, A. Zebiri. An improved n-grams based model for authorship attribution. In: *Proceedings of the 2019 International Conference on Computer and Information Sciences (ICCIS)*, Sakaka, Saudi Arabia, pp. 1–6.
- [15] D. Wright, Using word n-grams to identify authors and idiolects: A corpus approach to a forensic linguistic problem. *International Journal of Corpus Linguistics* 22 (2017) 212–241.
- [16] J. Grieve, I. Clarke, E. Chiang, H. Giddeon, A. Heini, A. Nini, E. Waibel, Attributing the Bixby Letter using n-gram tracing, *Digital Scholarship in the Humanities* 34 (2019) 493–512.
- [17] M. Singh, K. Murthy. Authorship attribution using filtered n-grams as features. In: Reddy K.A., Devi B.R., George B., Raju K.S. (eds) *Data Engineering and Communication Technology. Lecture Notes on Data Engineering and Communications Technologies* 63 (2021) 379-390.
- [18] T. Georgieva-Trifonova, M. Duraku, Research on n-grams feature selection methods for text classification. *IOP Conference Series: Materials Science and Engineering* 1031 (2020) 012048.
- [19] R. Ramezani, A language-independent authorship attribution approach for author identification of text documents. *Expert Systems with Applications* 180 (2021) 115139.
- [20] A. Romanov, A. Kurtukova, A. Shelupanov, A. Fedotova, V. Goncharov, Authorship Identification of a Russian-Language Text Using Support Vector Machine and Deep Neural Networks. *Future Internet* 13 (2021) 3.
- [21] D. Kosmajac, V. Kešelj, Language distance using common n-grams approach. In: *Proceedings of the 19th International Symposium INFOTEH-JAHORINA, 2020*, pp. 1–5.
- [22] P. Gamallo, J. R. Pichel, I. Alegria, From language identification to language distance, *Physica A: Statistical Mechanics and its Applications* 484 (2017) 152-162.
- [23] A. Yaroshevskiy, D. Klyushin, Nonparametric Methods of Authorship Attribution in Classic and Modern Literature. In: *Proceedings of 2019 IEEE International Conference on Advanced Trends in Information Theory (ATIT)*, 2019, pp. 465–469.
- [24] D. Klyushin, V. Michayliuk, Nonparametric methods of authorship attribution in English literature. *Journal of Numerical and Applied Mathematics* 133 (2020) 50–58.
- [25] D. Klyushin, Y. Petunin, A nonparametric test for the equivalence of population a measure of proximity of samples. *Ukrainian Mathematical Journal* 55 (2003)181–198.
- [26] B. M. Hill. Posterior distribution of percentiles: Bayes’ theorem for sampling from a population. *Journal of American Statistical Association* 63 (1968) 677–691.