The Multifactor Method Applied for Authorship Attribution on the Phonological Level

Iryna Khomytska [0000-0003-3470-7191] and Vasyl Teslyuk [0000-0002-5974-9310]

Lviv Polytechnic National University, Lviv 79013, Ukraine
Iryna.khomytska@ukr.net,vasyl.m.teslyuk@lpnu.ua

Abstract. The multifactor method has been developed to enhance test validity of authorship attribution. The method is style based. An author of a text can be identified by three major factors: a style based factor, a topic based factor and an authorial style based factor. For each factor certain statistical parameters are determined. The statistical parameters are actual distributions of frequencies of occurrence of the researched language units. As the research is done on the phonological level, the language units are phonemes. To differentiate texts by different authors, the powerful statistical tests have been applied (the Kolmogorov-Smirnov's test, the chi-square test, the Student's t-test).

Keywords: Phoneme Group, Multifactor Method, Style Based Factor, Topic Based Factor, Authorial Style Based Factor.

CEUR-WS.org/Vol-2604/paper14.pdf

1 Introduction

Anonymous information in the Internet has always been an important problem for researchers. The necessity to solve this problem is growing as the number of people using this network is increasing. The anonymous texts occur in different areas of communication. In certain cases the information doesn't bother anyone and can be negligible. But it does disturb when it threatens, harasses and is inappropriate. If the reader considers the information personal and sensitive, the anonymous or given under a pseudonym information must be studied. It should be noted that plagiarism detection is closely connected with disputed authorship. It is particularly important to determine the real author of a text of the same or similar content, but having several authors. In certain cases it is necessary to establish the author of the text written a long time ago, not clear when. The task of determining an author involves some text categorization and classification. The choice of most optimal classifiers and feature sets depends on various factors. The size of the text under study is of importance. The method efficient for the short texts may be inefficient for long texts. Each language level and feature set on this level have their specificity. The methods selected for a certain language level and feature sets are sure to behave differently. The number of features in a set may be increased or decreased, depending on the applied method. The problem of authorship attribution implies inferring a certain style, a certain topic and an authorial style. A text by an author can be of different complexity. The simplest case is when two studied texts are

Copyright © 2020 for this paper by its authors.
Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

of the same style, genre and topic. The most complex case is when the texts are of different style, genre and topic. There also intermediate cases with greater or less similarity or difference. The three mentioned factors lie in the basis of the developed multifactor method. Each factor effect is determined by three efficient statistical methods – the Kolmogorov-Smirnov's test, the chi-square test and the Student's t-test. The purpose of the research is to enhance test validity of authorship attribution with the help of the multifactor method. According to the results of recent research, different authorship attribution approaches have been used. Thus, in the field of digital text forensics, informal chat conversations have been researched. The algorithmic solutions have been obtained with 72,7%, 75% accuracy [1, 2]. The problem of author identification in short texts of Internet communication has been studied. In this research the temporal changes of word usage are relevant [3]. The analysis of principle components for authorship identification has been conducted in business systems research [4]. The multi sequence word selection method has been chosen to determine the author of a text [5]. Method of similar textual content selection based on thematic information retrieval has been applied for an analysis of the text under study [6]. The quantitative methods have been used to study lexical and stylistic peculiarities of a text [7 – 10]. The recurrent neural networks have been used to model the flow of the text for authorship attribution. For this study a large corpus has been recommended [11]. The unmasking approach has been used in the forensic field for short texts – four pages. The accuracy is 75%, 80% [12]. Large candidate sets have been researched by machine learning techniques. This is a novel approach, as the previous one studied a limited number of candidates [13]. The Twitter site has been analyzed for stylometric features. The author for an illegitimate text has been inferred [14]. Similarity-based methods have been used to consider authorship attribution in the wild. Anonymous texts have been analyzed on the lexical level [15]. In the investigation conducted by the support vector machine classifier, good results have been obtained – around 95% on the feature set of bag of words [16]. In comparison with the mentioned research, the novel approach in this study consists of applying the proposed combination of the three methods: the Kolmogorov-Smirnov's test, the chi-square test and the Student's t-test which have proved efficient in authorship attribution. To maximize the accuracy, the language level with an unchangeable number of elements has been chosen – the phonological level. The success rate is 95%, 97% and 98% [17].

2 Mathematical Support of Software System

2.1 The Method Developed

The problem of authorship attribution is aimed at determining if two compared pieces of text were written by a single author. Texts from poetry (G. Byron, T. Moore) and the publicist style (B. Obama, D. Trump, D. Webster, S. Logan) have been selected for experiments. An author of a text can be revealed by three major factors: a style based factor, a topic based factor and an authorial style based factor. The style based factor consists of showing the difference between the two styles, the topic based factor – the difference between the two texts on different topic, the authorial style based factor – the difference between the two texts by different authors. The scheme is style – topic –

author. The average value of the three factor based values is considered general style markedness of a text. The steps of the multifactor method algorithm are given below. More detailed information regarding the steps of the Student's t-test, the Kolmogorov-Smirnov's test and the chi-square test was presented in the previous research [18, 20].

1) the Student's t-test is performed for the texts from different styles, on different topics and by different authors [18, 19, 20]:

$$t=\frac{\overset{-a}{x_1}-\overset{-a}{x_2}}{S}\sqrt{\frac{n_1\cdot n_2}{n_1+n_2}}\,, \tag{1}$$

where $\overset{-a}{x}_1-\overset{-a}{x}_2$ is a difference of average frequencies for two samples for the fixed group of consonants *a*, S is a variance, n is a sample size.

2) the Kolmogorov-Smirnov's test is performed for the texts from different styles, on different topics and by different authors [18, 21, 22, 23]:

$$D_{n,m}=\sup_{-\infty<z<\infty}\left|F_n(z)-F_m(z)\right|, \tag{2}$$

where $F_n(z)$ and $F_m(z)$ are two empirical distribution functions for n and m. Samples. If $\lambda_{n,m}\geq\lambda_\alpha$, the texts can be differentiated:

$$\lambda_{n,m}=\sqrt{\frac{nm}{n+m}}D_{n,m}=\sqrt{\frac{nm}{n+m}}\sup_{-\infty<z<\infty}\left|F_n(z)-F_m(z)\right|\cdot \tag{3}$$

3) the chi-square test is performed for the texts from different styles, on different topics and by different authors [24, 25, 26, 27]:

$$\hat{\chi}_n^2=\sum_{i=1}^{s}\sum_{j=1}^{k}\frac{\left(\upsilon_{i,j}-\frac{n_j\upsilon_j}{n}\right)^2}{\frac{n_j\upsilon_j}{n}},\quad \upsilon_j=\sum_{j=1}^{k}\upsilon_{ij}\,, \tag{4}$$

where $\vartheta_{i,j}$ is a realization number in *j*-th series, s is a number of consonant groups, k is a number of samples, n_j is a number of sample portions, n is a number of portions for two samples. The texts can be differentiated if $\hat{\chi}_n^2\geq\chi^2_{1-\alpha,(s-1)(k-1)}$.

4) determining the style based factor value t_{f_1} for a phoneme group *a*: $\overset{-a}{x}_{1s}-\overset{-a}{x}_{2s}$ (s is a style);

5) determining the topic based factor value t_{f_2} : $\overset{-a}{x}_{1t}-\overset{-a}{x}_{2t}$ (t is a topic);

6) determining the authorial style based factor value t_{f_3} : $\overset{-a}{x}_{1a}-\overset{-a}{x}_{2a}$ (*a* is an author);

7) determining the general style markedness:

$$sm = \frac{t_{f_1} + t_{f_2} + t_{f_3}}{3}. \qquad (5)$$

8) the authorship attribution is calculated by the difference of values of the general style markedness for two authors.

2.2 The Developed Software

To develop the software for authorial differentiation, the Java programming language has been used. The programming language is cross-platform and this is an advantage of the chosen programming language. The developed program system realizes the following algorithm (Fig. 1):

The structure of the developed software has the following tabs: “Text”, “Transcription Symbols”, “Consonant Phoneme Sample”, “Portion Division”, “Group Division”, “Calculating Phonemes in Portions”, “Calculating Phonemes in Groups”, “Statistical Test”, “Style Based Factor Value”, “Topic Based Factor Value”, “Authorial Style Based Factor Value”, “General Style Markedness Values”, “Difference by General Style Markedness Values”.

The software classes are shown in the diagram s in Fig. 2.

One of the advantages of this program system is its relative independence of the transcription site on which an English text is transcribed. The transcription site is used when the first experiments are made. The bag of words gets larger every time new texts are processed. Therefore, it is advisable to process large samples. These are some short documents in the forensic field. However, large samples are of interest when it is necessary to characterize literary legacy of some author. Such problems are usually researched in corpus linguistics. In this investigation, both short and long texts are analyzed. The sample size is 50 000 phonemes and more.

3 Results of the Study

For the first experiment of author identification, two pieces of poetry have been selected: one by G. Byron and another by T. Moore. According to the proposed scheme, the two pieces of poetry must be analyzed in a comparison with some style having most common bag of words. Evidently, this may be the conversational style in its literary version which has few colloquial elements. This is particularly relevant in the comparison with poems by romanticists who tried to use conversational elements. On the other hand, it is necessary to compare the poems with another genre of the style of fiction. It may be Byron’s emotive prose. These two samples are sure to have common language units both being of fiction style. In the third stage of the study, the poems by the two poets are compared. The multifactor method makes it possible to calculate the average value of the values calculated for each mentioned above comparisons. The average value is general style markedness (Table 1).

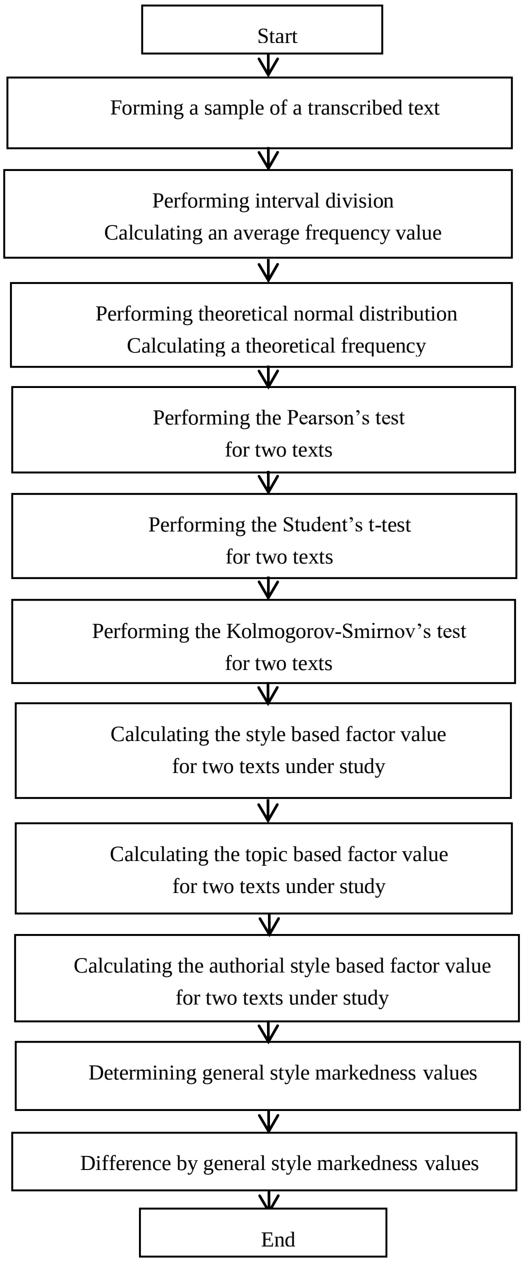
Start
Forming a sample of a transcribed text
Performing interval division
Calculating an average frequency value
Performing theoretical normal distribution
Calculating a theoretical frequency
Performing the Pearson's test
for two texts
Performing the Student's t-test
for two texts
Performing the Kolmogorov-Smirnov's test
for two texts
Calculating the style based factor value
for two texts under study
Calculating the topic based factor value
for two texts under study
Calculating the authorial style based factor value
for two texts under study
Determining general style markedness values
Difference by general style markedness values
End

Fig. 1. An algorithm of the developed program system.

Having calculated the value of general style markedness for each poet, the author identification test can be performed. The difference of the two authorial styles is calculated by the difference of values general style markedness (Table 2). It is equal to 1.8.

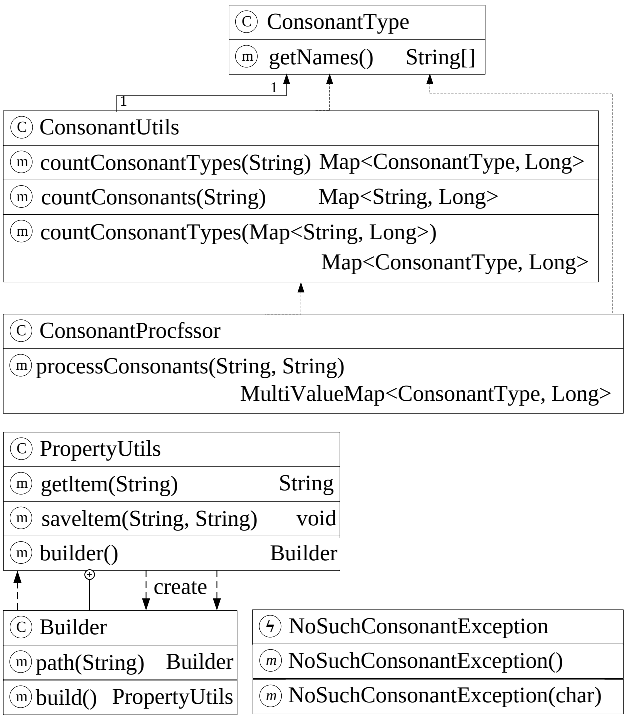
C ConsonantType
m getNames() String[]
1
1
C ConsonantUtils
m countConsonantTypes(String) Map<ConsonantType, Long>
m countConsonants(String) Map<String, Long>
m countConsonantTypes(Map<String, Long>) Map<ConsonantType, Long>
C ConsonantProcfssor
m processConsonants(String, String) MultiValueMap<ConsonantType, Long>
C PropertyUtils
m getltem(String) String
m saveltem(String, String) void
m builder() Builder
create
C Builder
m path(String) Builder
m build() PropertyUtils
NoSuchConsonantException
m NoSuchConsonantException()
m NoSuchConsonantException(char)

Fig. 2. A diagram of the system classes for authorship attribution

Table 1. Results of determining general style markedness

Comparison with the 1-st style	Comparison with the 2-nd style	Comparison with another author	Value of general style markedness
GB-CLS – 16	GB-GBP – 17	GB-TM – 7	GB – 13.3
TM-CLS –15	TM-GBP –13	TM-GB –7	TM –11.5

In Table 1 the following designations are used: GB is Byron's poetry, TM is Moore's poetry, GBP is Byron's emotive prose, CLS is the conversational style.

Table 2. Results of the comparison of Byron's and Moore's poetry

Compared texts by Byron and Moore in dorsal phoneme group	General style markedness of Byron's poetry	General style markedness of Moore's poetry	Essential difference value
Byron-Moore's poetry	13.3	11.5	1.8

For the second experiment four authors have been selected. They are: B. Obama, D. Trump, D. Webster, S. Logan. The pieces of writing represent the publicist style. In this case the multifactor method involves comparison with the 1-st author, comparison with the 2-nd author and comparison with the 3-rd author. The average value calculated from the three values got in each comparison is the general style markedness by which the author can be identified. Though the samples are of the same style, the topic varies from sample to sample having common bag of words. The topic reflects international political events all over the world. Therefore, the content is relatively homogeneous. This relative homogeneity creates a problem of its own. The topic based factor affects the final result of author identification. Sometimes it is rather difficult to draw a distinct demarcation line between effect of the topic based factor and the authorial style based factor. The case is easier with documents following strict standards of conveying information. The same situation can hardly be observed in the publicist style. On the other hand, the publicist style is the style in which the individual peculiarities of an author's manner of writing can be vividly revealed. The effect of the three factors mentioned is expressed in the value of the general style markedness given in Table 3.

Table 3. Results of determining general style markedness

Comparison with the 1-st author	Comparison with the 2-nd author	Comparison with the 3-rd author
Obama-Trump–14	Obama-Webster –14	Obama-Logan –14
Trump-Obama –14	Trump-Webster –17	Trump-Logan –16
Logan-Webster –17	Logan-Obama –14	Logan-Trump –16
Webster-Logan –17	Webster-Obama –14	Webster-Trump –17

The value of general style markedness is the highest for Webster's authorial style. It equals to 16. The lowest value is for Obama's writing style (14). But, as a significant role is played here by the topic based factor, the authorial writing characteristics can be different in another verbal content. Thus the essential information about author identification can be got with the help of the multifactor method. The results of comparisons of texts by different authors are shown in Table 4.

Table 4. Results of comparisons of texts by B. Obama, D. Trump, D. Webster, S. Logan

Compared texts by different authors in 8 phoneme groups	General style markedness	General style markedness	Value of essential difference
Obama-Trump	Obama – 14	Trump – 15.6	1.6
Obama-Webster	Obama – 14	Webster – 16	2
Obama-Logan	Obama – 14	Logan – 15.6	1.6
Trump-Webster	Trump – 15.6	Webster – 16	0.4
Trump-Logan	Trump – 15.6	Logan – 15.6	0
Webster-Logan	Webster – 16	Logan – 15.6	0.4

Table 4 shows that different effect of the topic and author based factors causes great difference in a comparison Obama-Webster, less difference – Obama-Trump, Obama-Logan, still less difference – Trump-Webster, Webster-Logan and practically no difference – Trump-Logan. The last pair of texts shows similarity of bag of words.

Among the used statistical tests two are the most powerful. These are the Kolmogorov-Smirnov's test and the chi-square test. With the help of the former, the authorial styles differ essentially in all eight phoneme groups. The latter is a little less powerful – the difference in six of eight groups.

The efficiency of the multifactor method may be analyzed for each of eight groups of phonemes. Reduction of the number of phoneme groups makes the whole procedure more economical. Consequently, it is necessary to analyze author-differentiating capability for every group.

The degree of author-differentiating capability of phoneme group depends on the number of times the essential differences have been established. If the essential difference has been revealed by three statistical tests, the phoneme group takes number 3, by two statistical tests – number 2, by one statistical test – number 1. In Table 5, for labial group, number 3 (Obama-Logan) has been got once, number 1 – thrice. The group takes the second degree of differentiation power.

Table 5. The author-differentiating capability of labial phoneme group

Compared texts by different authors	Author-differentiating capability
Obama-Trump	1
Obama-Webster	1
Obama-Logan	3
Trump-Webster	2
Trump-Logan	2
Webster-Logan	1
Byron-Moore	2

Compared with labial group, dorsal group has higher degree of differentiating capability (Table 6). Only in two pairs of texts one statistical test has proved efficient. More differences have been obtained by two tests. The group takes the first degree of differentiation power.

The results of the conducted research have shown that the multifactor method is efficient in authorship attribution. The established general style markedness of a text has made it possible to classify each sample under study in accordance with three basic factors – style, topic and author's manner of writing. Taking into account the three mentioned factors is particularly efficient when the compared samples represent different style and topic. In this case it is impossible to characterize the authorial specificity of writing because of the influence of style and topic factors. Having determined the style based and topic based features, the authorial features can be identified.

4 Conclusions

In order to single out particular features of an individual writing style, the style based features and topic based features must be separated. To solve this task, the multifactor method must be applied. In accordance with this method, the average value of the three factor based values is calculated. The three factor based values involve: comparison with the text least marked by the style elements, comparison with the text of the same topic, but different author and comparison with the text of the same style and topic, but another author. The average value of these three values is general style markedness. The author identification is calculated by the difference of general style markedness values. The results show that the greatest difference is in the pair Obama – Webster (2), less difference – in the pairs Obama – Trump, Obama – Logan (1.6), still less – in the pairs Trump – Webster, Webster – Logan (0.4), the least in the pair Trump – Logan. The test validity has been enhanced up to 95%, 97%.

The developed software on the Java programming language has performed the author identification procedure in a fewer number of consonant groups making it more automated. The next step in our research will be concentrated on the other statistical methods.

Table 6. The author-differentiating capability of dorsal phoneme group

Compared texts by different authors	Author-differentiating capability
Obama-Trump	1
Obama-Webster	1
Obama-Logan	3
Trump-Webster	2
Trump-Logan	2
Webster-Logan	2
Byron-Moore	2

References

1. Halvani, O., Winter, Ch., Graner, L.: Assessing the Applicability of Authorship Verification Methods. In: Proceedings of the 14th International Conference on Availability, Reliability and Security, No.: 38. pp. 1–10. (2019). https://doi.org/10.1145/3339252.3340508.
2. Koppel, M., Schler, J., Argamon, Sh.: Authorship Attribution: What's Easy and What's Hard? In: Computer Science, (2013). DOI: 10.2139/ssrn.2274891.
3. Azarbonyad, H., Dehghani, M., Marx, M., Kamps, J.: Time-Aware Authorship Attribution for Short Text Streams. In: Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval, New York, USA, pp. 727 –730. (2015).
4. Jamak, A., Alen, S., Can, M.: Principal Component Analysis for Authorship Attribution. In: Business Systems Research, 3(2), pp. 49–56. (2012).
5. Mubin, S. T., Rajesh S. P.: Authorship Identification with Multi Sequence Word Selection Method. In: Thermal Stresses—Advanced Theory and Applications, pp. 653–661, (2019). 10.1007/978-3-030-16657-1_61.
6. Vysotska, V., Lytvyn, V., Kovalchuk, V., Kubinska, S., Dilai, M., Rusyn, B., Pohreliuk, L., Chyrun, L., Chyrun, S., Brodyak, O.: Method of similar textual content selection based on thematic information retrieval. In: CSIT, Proceedings of the XIVth Scientific and Technical Conference, Lviv, pp. 1–6. (2019).
7. Kulchytskyi, I., Shandruk, U.: The quantitative research of scientific texts at the symbolic level. In: Computational linguistics and intelligent systems. Lviv: Lviv Polytechnic National University, 25 – 27 June, vol 2, pp. 71–80. (2018).

8. Karamysheva, I., Nazarchuk, R., Fedoruk, M.: Synonymic connections of cognitive verbs in English and Ukrainian languages: applied aspect. In: CSIT, Proceedings of the XIIIth Scientific and Technical Conference. Lviv, pp. 1–4. (2018).
9. Romanyshyn, N.: Application of computer technologies in conceptual analysis. In: CSIT, Proceedings of the XIIIth Scientific and Technical Conference. Lviv, pp. 55–57. (2018).
10. Peleshchyshyn, A., Markovets, O., Vus, V., Albota, S.: Identifying specific roles of users of social networks and their influnce methods. In: CSIT, Proceedings of the XIIIth Scientific and Technical Conference. Lviv, pp. 39–42. (2018).
11. Bagnall, D.: Author Identification Using Multi-headed Recurrent Neural Networks. Conference and Labs of the Evaluation forum, Toulouse, France, pp. 1 – 8. (2015).
12. Bevendorff, J., Stein, B., Hagen, M., Potthas, M.: Generalizing Unmasking for Short Texts. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Minneapolis, Minnesota, vol. 1, pp. 654 – 659. (2019).
13. Koppel, M., Schler, J., Argamon, Sh., Winter, Ya.: The "Fundamental Problem" of Authorship Attribution, vol. 93, issue 3, pp. 284 – 291. (2012) DOI:10.1080/0013838X.2012.668794
14. Bagnall, D.: Author Identification Using Multi-headed Recurrent Neural Networks. In: Conference and Labs of the Evaluation forum, Toulouse, France, pp. 1 – 8. (2015).
15. Bhargava, M., Mehndiratta, P., Asawa, K.: Stylometric Analysis for Authorship Attribution on Twitter. In: Proceedings of the Second International Conference on Big Data Analytics, vol. 8302,. pp. 37–47. (2013). https://doi.org/10.1007/978-3-319-03689-2_3
16. Koppel, M., Schler, J., Argamon, Sh.: Authorship attribution in the wild. In: Language Resources and Evaluation, vol. 45, No. 1, (2011). URL: https://doi.org/10.1007/s10579-009-9111-2
17. Bozkurt, I., N., Baghoglu, O., Uyar, E.: Authorship attribution. In: 22nd International Symposium on Computer and Information Sciences (ISCIS), pp. 158 – 162. (2007). DOI: 10.1109/ISCIS. 2007. 4456854.
18. Khomytska, I., Teslyuk, V.: Statistical Models for Authorship Attribution. In: Advances in Intelligent Systems and Computing III / Natalia Shakhovska editor, Lviv,. vol. 1080. pp. 579–592. (2019).
19. Gomez, P., C.: Statistical Methods in Language and Linguistic Research. Spain: Unibersity of Murcia, (2013).
20. Khomytska, I., Teslyuk, V., Kryvinska, N., Beregovskyi, V.: The Nonparametric Method for Differentiation of Phonostatistical Structures of Authorial Style. In: Procedia Computer Science: Proceedings of the 10th International Conference on Emerging Ubiquitous Systems and Pervasive Networks, Coimbra, Portugal, vol. 160, pp. 38-45. (2019).
21. Kolmogorov, A. N.: Mathematics and its Historical Development Edited by V. A. Uspensky, Published by Nauka, Moscow (1991).
22. Gnedenko, B. V., Kolmogorov, A. N.: Limit Distributions for Sums of Independent Variables Published by Addison-Wesley (1968).
23. Watanabe, S.: Probability Theory and Mathematical Statistics. Springer (1988).
24. Gries, Th. S.: Statistics for Linguistics with R: A Practical Introduction (Trends in Linguistics: Studies & Monographs), p. 348. (2009).
25. Rozanov, Iu. A., Silverman, R. A.: Probability Theory: A Concise Course Dover Publications Inc. (2007).
26. Jorgensen, P.E.T.: Analysis and Probability. Springer (2006).
27. Bhattacharya, R., Waymire, E. C.: A Basic Course in Probability Theory Springer; 2nd ed. 2016 edition, February 16, (2017).