Розглянуто особливості застосування технологій лінгвостатистики для ідентифікації стилістики автора текстового контенту науково-технічного профілю. Квантитативний лінгвістичний аналіз тексту використовує переваги контент-моніторінгу на основі методів NLP для визначення та аналізу множини стопових слів, ключових слів, стійких словосполучень та дослідження N-грам. Останні використовують в методах лінгвометрії для визначення приналежності аналізованого тексту конкретному авторові у відсотках. Розроблено квантитативний метод автоматичного визначення авторства текстового контенту на основі статистичного аналізу розподілу 3-грам. Запропоновано підхід реалізації визначення автора україномовного тексту науково-технічного профілю. Отримано експериментальні результати запропонованого методу для визначення приналежності аналізованого тексту конкретному автору за наявності еталонного авторського тексту. Застосування лінгвостатистичного аналізу 3-грам до множини статей дозволить сформувати підмножину подібних за лінгвістичними характеристиками публікацій. Накладання на підмножину додаткових умов у вигляді проведення статистичних та квантитативних аналізів (множини ключових слів, стійких словосполучень, стилеметричного, лігвометричного тощо) дозволить значно скоротити цю підмножину, уточнивши список ймовірніших авторських робіт. Для якісного та ефективного аналізу контенту при визначенні ступеня авторства конкретному автору пропонуємо аналізувати еталонного тексту та досліджуваного в декілька етапів: лінгвометричний аналіз коефіцієнтів різноманіття авторського мовлення, стилометричний аналіз, аналіз стійких словосполучень, лінгвостатистичний аналіз 3-грам. Для автоматизованого опрацювання тексту має велике значення не тільки частота появи тієї чи іншої категорії, а взагалі присутність в досліджуваному тексті. Кількісний підрахунок дозволяє зробити об'єктивні висновки щодо спрямованості матеріалів за кількістю уживань одиниць аналізу в досліджуваних текстах. Якісний аналіз робить те саме, але внаслідок дослідження того, чи зустрічається (і в якому контексті) певна важлива оригінальна категорія взагалі

Ключові слова: NLP, контент, контент-моніторінг, стоп-слова, контент-аналіз, статистийний лінгвістичний аналіз, квантитативна лінгвістика, статистична лінгвістика, лінгвометрія

UDC 004.89
DOI: 10.15587/1729-4061.2019.186834

DEVELOPMENT OF THE QUANTITATIVE METHOD FOR AUTOMATED TEXT CONTENT AUTHORSHIP ATTRIBUTION BASED ON THE STATISTICAL ANALYSIS OF N-GRAMS DISTRIBUTION

V. Lytvyn
Doctor of Technical Sciences, Professor*
E-mail: vasyl.v.lytvyn@lpnu.ua
V. Vysotska
PhD, Associate Professor*
I. Budz
PhD, Associate Professor
Department of Computational Mathematics and Programming**
Y. Pelekh
PhD, Associate Professor
Department of Computational Mathematics and Programming**
N. Sokulska
PhD***
R. Kovalchuk
PhD, Associate Professor***
L. Dzyubyk
PhD***
O. Tereshchuk
PhD***
M. Komar
PhD
Department of Information and Computing Systems and Control
Ternopil National Economic University
Lvivska str., 11, Ternopil, Ukraine, 46009
*Department of Information Systems and Networks**
**Lviv Polytechnic National University
S. Bandery str., 12, Lviv, Ukraine, 79013
***Department of Engineering Mechanics (Weapons and Equipment of Military Engineering Forces)
Hetman Petro Sahaidachnyi National Army Academy
Heroiv Maidanu str., 32, Lviv, Ukraine, 79026

Received date 06.10.2019
Accepted date 04.12.2019
Published date 25.12.2019

1. Introduction

Due to the increasing availability and distribution of the text content in the Internet, the degree of importance of using automatic methods of text content analysis is increasing [1]. The tasks of content analysis include the problems of classification and clustering of text-based publications according to various criteria, for example, genre, epoch of

writing, format (a novel, an essay, and a scientific article), emotional coloration, style of speech, as well as the problem of text authorship attribution [2]. With the simplification of access to various data, expanding the ability of finding, copying and distributing information in the Internet, the problem of the authorship attribution is becoming relevant [3]. The problems related to authorship attribution are important in linguistic, historical, and forensic research [4]. The availability of electronic devices makes it possible to push the author's recognition with the involvement of a large number of experts to the background, accelerate and simplify this process through its automation [5]. The concept of the authorship attribution is defined as the process of recognizing the author by a set of general and individual features of a text constituting the author's style [6].

The statistical methods for authorship attribution, based on the search for the author's invariant enjoy great popularity today [7]. This is a characteristic of the linguistic peculiarity of a text (lexical, grammatical, phraseological, etc.) [8]. In particular, an invariant is the percentage of using vowels/consonants, frequency of using of a certain part of speech, probability of transitions from one part of speech to another, parasite words, information entropy, etc. [9]. The statistical method of identification of the author and the genre of a text, based on the distribution of frequencies of letter combinations (*N*-grams), is also effective. However, the accuracy of the statistical methods for authorship attribution relies heavily on the specifics of the data used: the speech style and text length [11]. Due to this, it is difficult to make conclusion about the accuracy of this approach on scientific and technical articles [12]. For this reason, it is necessary to analyze the applicability of such mathematical apparatus as the distribution of frequencies of different languages simultaneously with other techniques in solving the problem of attribution of authorship of texts with different lengths and written in different speech styles [13]. The methods for authorship attribution for the Ukrainian-language text content of the scientific and technical area are proposed and studied in papers [1–7]. Various algorithms can be used to implement these methods [14], including quantitative [15]. Therefore, there arises the problem of analyzing such algorithms to find the most effective one [16]. The authorship attribution is the technique for determining the author of a text, when it is ambiguous, who wrote it [17]. This is useful when several people claim for the authorship of one publication [18] or in cases when no one claims to be the author of the textual content [19], for instance, so-called trolls in social networks during information warfare. The complexity of the problem of the author's test is evidently exponentially higher, and the number of likely authors is larger [21]. The existence of the author's text samples is also significant in advancing this problem [22]. The author's text attribution includes the following three issues [23]:

– identification of the author of a text from the group of probable or expected authors, where the author is always in the group of suspects [24];

– non-identification of the text author from the group of probable or expected authors, where the author may not be in the group of suspects [25];

– evaluation of the possibility that this text is written by this author or not.

Therefore, the problem of automatic authorship attribution for the text content of scientific and technical direction is relevant and requires new (more advanced) approaches to its solution [27].

2. Literature review and problem statement

Papers [1–3, 28] show the results of the studies of language and speech material on a representative array of texts. This should be a homogeneous array (body) of certain units, that is, the general totality (GT) [3]. It was shown that the volume and the character of the GT depend on the tasks of the study. For example, if the peculiarities of Ivan Franko's style are explored, the GT is all his works. If we explore the Ukrainian language of the XX century, the GT is all the texts (spoken and written) of the XX century [3]. The boundaries of the latter are difficult to identify and it is simply impossible to explore entire oral speech [29], especially when analyzing the author's speech. The issues related to attribution of the authorship of the text in the collective works of scientific and technical direction based on the analysis of reference sample also remain unresolved. The reason is the lack of experiments in this direction. Another reason is the existence of insufficient statistics for the formation of conclusions due to the fact that the authors in this direction rarely write papers individually, and in some areas, works are generally rarely written cooperatively. An option to overcome the corresponding difficulties, when the overall study of the GT is impossible, is sampling and the formation of a set of parameters for the corresponding analysis [1–3, 30]. All this gives grounds to argue that it is appropriate to conduct research dedicated to the textual content authorship attribution based on statistical analysis of the distribution of characteristics of the author's speech with a sufficient number of data samples.

Samples are a certain amount of material, by studying which it is possible to make correct conclusions about the entire GT [31]. The basic requirements for the samples include representativity and homogeneity [32]. To be representative, the sampling should [33]:

1) be evenly distributed around the GT [34];

2) have a sufficiently large volume, which is enough for the correct conclusions about the GT [35].

There are two types of sampling uniformity: linguistic and statistical.

Within the linguistic homogeneity of the sampling, they distinguish [3]:

– chronological (sample texts must have chronological boundaries) [36];

– genre (sample texts should be genre-limited) [37];

– thematic (texts should be thematically limited) [38].

Statistically homogeneous samples are the samples, in which the studied units have statistical behavior that does not differ substantially among them [39]. If the average frequency of a phenomenon (letters, morphemes, words, word length, sentence length, etc.) in a single sampling does not significantly differ from its frequency in other samples, then these samples are statistically homogeneous in relation to this phenomenon [40].

According to the way of organization, the following kinds of samples are distinguished [3]:

– mechanical – organized taking into consideration the uniformity of distribution of a studied unit in the general totality [41]. All texts of the general totality are numbered, and then, for example, a segment of the necessary length is chosen from each fifth and tenth or the twentieth text [42];

– random – arranged by random choice of the texts from the GT [43]. At the core of this method of organization of samples, there is the hypothesis that quite a large number of

randomly chosen units from the GT must adequately represent it [44]. Thus, each page, section, or other unit of a text of the GT should have the same chance of getting into the sampling. Therefore, as a rule, random sampling is based on the table of random numbers [45];

– zonal (typical) – organized based on the linguistically homogeneous totality of texts, that is, zones [46]. Depending on the purpose of the study, a zone can be considered to be prose, poetry and drama in fiction; works by one author or a specific work; totality of words of a certain morpheme structure (for example, prefixed or monomorphemic), etc. [47].

Samples may be structural, that is, composed of smaller parts (sub-samples) and non-structural, that is continuous [48].

The ratio between the frequency of language and speech units will be shown by the example: "if one takes 33 bingo barrels, glue down the Ukrainian alphabet on them and mix, then the probability that the first taken barrel is purely vowel will be 6:33 (6 pure vowels letters (а, о, у, є, и, і) to all 33 letters of the Ukrainian alphabet), that is, approximately 16 %" [3]. If one takes a random Ukrainian text and chooses randomly one letter from it, the probability that it will turn out to be pure vowel will be approximately 30 % [3]. In the first case, it goes about the probability of a group of six letters at the paradigmatic level (language), in the second case – at the syntagmatic level (speech) [49]. To assume that all vowels or all the case forms, or all the members of the sentence are equally probable, would mean to replace the natural speech with its scheme [50]. Thus, speech prefers a small number of units (the prevalence law), which constitutes the core of the speech subsystem, whereas in the language, all units are equally probable [51]. In different languages, the frequency of the same letter or a sequence of letters is different, so knowing the order of the most frequent letters, bigrams, trigrams, four-grams of a particular language, it is possible to identify it automatically [52]. The frequency of these units in a language is determined using representative sampling, since frequency in the works by specific authors, styles or themes is also different [53]. For example, it is found for Ukrainian texts that frequency of vowels, consonants, spaces between words, as well as the groups of consonants: soft, sonoric can be considered statistical style parameters [1–7]. The frequencies of letters in the texts were studied for the needs of cryptography (science about ciphering and deciphering messages), in particular, Morse code (the more frequent a letter or a letter combination, the shorter the slashes for their designation), for shorthand, for automatic determining of a language, confirmation or denial of authorship of a work, etc. [54]. Morphemes and grammatical categories also have their own quantitative characteristics:

– non-uniform use of morpheme of a foreign language and specific language morphemes [55];

– verbs of present, past and future tense, indicative, conditional, imperative moods [56];

– forms of the verb (infinitive, personal forms, participles, impersonal forms ending in -но, -то) [57];

– different parts of speech depending on the style [58].

The regularity was found that in different functional styles, the quantitative ratio of functioning of different cases is not the same [59]. For example, scientific prose prefers genitive case and neglects nominative case, but the opposite is true for the spoken language, etc. [3].

The quantitative characteristics of words are best seen in the WF [59]. The functional dependence of the relationship between the word frequency and polysemy, as well as between the word frequency and its rank in a dictionary by descending frequencies is expressed by the Zipf-Mandelbrot law. The most frequent are functional parts of speech or general abstract concepts [60]. Instead, words with a specific meaning (required for a conversation in a usual situation) are low-frequency words [61]. Although they are rarely used, they are always in the speaker's mind [62]. In other words, the frequency criterion is supplemented by the subject-matter criterion [3]. The formula to establish the synonymy degree (semantic proximity) of words [63]: $C=2c/(n_1+n_2)$, where n_1 is the number of meanings of the first word, n_2 is the number of meanings of the second word, c is the number of common meanings in the given pair of words. Quantitative characteristics of the syntactic constructions also depend on a functional style: simple uncomplicated, even incomplete and broken sentences prevail in the colloquial everyday style, while composite sentences, complicated by constructions, parentheses and inserted structures dominate in the official style [64].

The tempo of speech-thought can be represented in a simple way as a ratio of the number of independent words to the number of simple sentences, since the fewer words are included in one sentence, the more frequent the sentences are (subsequently, the thoughts) [65]. It was revealed that the tempo of speech-thought in a fairy tale is 2.39, and in a scientific text – only 0.42 [3]. This means that speech and action in a fairy tale unfold almost 6 times as fast. And this is understandable: in a fairy tale, thoughts and statements expressed are simple in structure, and therefore faster, easier to construct in a dynamic sequence; in a scientific article, the structure of a speech-thought is much more complicated, so consciousness channels pass the units of such speech-thought slower [66].

It is logical to measure the speech coherence coefficient based on the ratio of the number of prepositions and conjunctions to the number of separate sentences [67]. Let this coefficient be equal to unity when one sentence has three connecting elements (prepositions and conjunctions) [3]: $K_z=(P+C)/(3N)$, where P is the number of prepositions, C is the number of conjunctions, N is the number of separate sentences. It was found that the text of a fairy tale has coherence coefficient 0.77, and the text of a scientific article – 3.0, that is, coherence in the second text is 3.9 times stronger than in the first one [3].

The concept of language synthetics is determined as M/W, where M is the number of morphs in a certain segment of the text, W is the number of words in this text [68]. Languages with index from 1 to 2 are considered analytical, from 2 to 3 – synthetic, and 3 or more – polysynthetic [69]. The Vietnamese language has the lowest magnitude – 1.06, that is, there are 196 morphs per 100 words, the Eskimos language has the highest magnitude – 3.72, that is, there are 372 morphs per 100 words [3]. English has an indicator of 1.68, Russian – 2.33 [3]. Based on the synthetics index, analytical languages include Vietnamese, Chinese, Persian, Italian, German, and Danish; synthetic languages include Ukrainian, Russian, Sanskrit, Lithuanian, Czech, Polish, Yakut: polysynthetic languages include Eskimos, Tubal-American, Iberic-Caucasian [69]. Due to the increasing availability and spreading of text documents in the electronic form, the importance of using automatic methods for analysis of the documents' content increased [70]. It is possible to consider that text analysis includes the tasks of documents' classification and clustering by various criteria,

for example, genre, epoch of writing, format (novel, essay, sketch), emotional coloration, style of speech, as well as task of determining the author of the text [71]. With the simplification of access to different data, expanded data search, copy, and distribution capabilities in networks, the task of authorship attribution is becoming even more urgent [72]. Similarly, the issues related to authorship attribution are important in linguistic, historical and forensic studies [73]. The availability of electronic devices makes it possible to push back the authorship attribution with the involvement of a large number of experts, to accelerate and simplify this process through its automation [74]. The notion of the authorship attribution is defined as the process of finding the author by many general and private features of a text that constitute the author's style [75].

In the existing systems of text authorship attribution, statistical methods based on the search for the "author's invariant" enjoy popularity [76]. The "author's invariant" characterizes the language peculiarity (lexical, grammatical, phraseological and other) of a text [77]. An invariant can include: proportion of vowels or consonants, frequency of using a certain part of speech, probability of transitions from one part of speech to another, "favorite" words, information entropy and so on [78]. In paper [3], the statistical method for determining the author and the genre of text based on the distribution of frequencies of letter combinations (*n*-grams) was proposed. The method showed decent results for the works of Russian literature [79]. However, the accuracy of the statistical methods for authorship attribution relies heavily on the specifics of the used data: on the language, on which the texts are written [80], on the style of speech of the text [82], and, above all, on the lengths of texts under research [83]. Because of this, it is difficult to conclude about the accuracy of this approach for the data of another nature. For this reason, the purpose of this work was to analyze the applicability of such mathematical apparatus as the distribution frequencies of letter combination for different languages in solving the problem of identification of the authorship of texts of different lengths and written in different styles of speech [84].

Calculation of the distance between the corresponding vectors is used as the criterion of proximity of two texts [85]. The sets of parameters and speech coefficients are represented as normal vectors in the *n*-dimensional Cartesian space from coordinate origin [86]. Then the distance between the texts is the usual Cartesian distance between the ends of the corresponding vectors. This norm distance is an integral characteristic of text differences [87]. And the texts with a large distance are highly likely to belong to different authors. Thus, in order to compare the authorship of two texts, it is enough to compute the parameters and determine the distance for them [88]. To associate a text with the author, the vectors the author's parameters and of the text are compared, that is, actually, two texts are compared again – a text, the author of which is known (the reference text), and the text, the authorship of which is necessary to identify, confirm or refute (analyzed/ researched text) [89]. The vectors of formal parameters that recognize not specific authors (or groups), but rather distinguish certain characteristics of authors (for example, educational level) are constructed [90]. In most cases, according to [91], statistic characteristics are selected as characteristic parameters of a text:

– the number of applications of certain parts of speech, some specific words, punctuation marks, phraseological units, archaisms, rare and foreign words [92];

– the number and the length of sentences (measured in words, syllables, and signs), average sentence length [93];

– the number of notional and functional words [94];

– vocabulary volume, ratio of the number of verbs to the total number of words used in a text, etc. [95].

The main problem of the formal methods for authorship analysis is to select the parameters and speech coefficients [96]. There are a number of formal statistical characteristics of the texts unsuitable for the identification of authorship because of one of the two shortcomings [1–7, 97]:

– Lack of stability. The scatter of parameter values for the texts of the same author is so large that the ranges of possible values for different authors intersect [99]. Obviously, this parameter will not help distinguish between the authors, and when used as a part of the parameter group, will only play the role of additional information noise [100].

– Lack of the distinguishing ability. The parameter can accept close values for all or most authors, because its values are determined by the properties of the language, in which the texts are written, rather than by the individual features of the author of a text [101]. That is why parameters must be previously studied in terms of stability and distinguishing ability, preferably in the texts of a large number of different authors [102].

In articles [1–7], the following conditions of applicability of formal speech coefficient of the author's style were separated:

– Mass character (the use of those characteristics of the text, which are poorly controlled by an author at the conscious level to eliminate the possibility of conscious distortion by the author of his characteristic style or imitation of the style of another author) [103].

– Stability (retaining a constant value for one participant, but some deviation of values from the mean should be rather small) [104].

– Distinguishing ability (takes substantially different values for different authors, that is, exceeds the fluctuations that are possible for one participant) [105].

It is very difficult to choose speech coefficients and parameters that are sure to distinguish between any two authors [106]. Whatever the parameters are, there is always the possibility that two or more participants are close by these parameters due to random coincidence [107]. That is why in practice it is sufficient for a parameter to make it possible to distinguish between different subsets of authors, that is, there should be quite a large number of subsets of the authors, for who mean values of the parameter are significantly different [108]. The parameter, obviously, will not help distinguish between the authors' texts from one subset, but will enable us to distinguish confidently between the texts of the authors that got in different subsets [1–7]. It is possible to distinguish between the texts of the authors of one subset by using simultaneously a rather large vector of parameters that are different by nature – in this case, the probability of random coincidence will be noticeably smaller. For confident output of the texts, for which the formally calculated distance is small, it is necessary to conduct additional research by expert methods, for example, analysis of key and/or stop (auxiliary) words [1–7].

Thus, there is a need to conduct the research in this direction due to the lack of practical experiments for identification of the author's style for Ukrainian scientific- technical texts. Recently, many systems have been developed to solve the problem of plagiarism as copyright. When it comes

to re-writing, it is quite difficult to solve such a problem for the Slavic languages due to the existence of a large set of synonyms and the possibility of restructuring sentences using other endings. This issue does not apply to the use of auxiliary words, as most people do not even pay attention to them in case of plagiarism. That is why this encourages the exploration of the problem of the author's style identification to determine the degree of belonging of a particular text to a particular author.

3. The aim and objectives of the study

The aim of this study is to develop the method for automatic text content authorship attribution based on statistical analysis of *N*-gram distribution.

To achieve the set aim, the following tasks were to be solved:

– to develop the quantitative method for identifying the potential author of a text from a set of possible ones comparing the results of analysis of the reference text with the studied text;

– to develop content-monitoring software to determine the author in the texts in the Ukrainian language based on the linguo-statistical analysis of the reference text content;

– to obtain and analyze the results of the experimental approbation of the proposed content monitoring method to determine the author of scientific texts of technical profile in the Ukrainian language.

4. Quantitative method

The quantitative method for identification of the potential author of a text from a set of possible ones is based on comparison of the results of analysis of the reference text with the studied text.

Linguometry is the field of applied linguistics that detects, measures, and analyzes quantitative characteristics of the units of different language or speech levels [3]. Using the apparatus of mathematical statistics, linguometry participates in solving the following problems of linguistics, as creation of:

– dictionaries (including frequency and statistical) and comparison;

– automatic dictionaries, thesauruses;

– systems of transcripts;

– methods and means of automatic language determining;

– methods and means of information search, etc.

Each language has its own statistical parameters, and knowledge of the frequency of occurrence of letters and their combinations (2-gram, 3-gram, 4-gram) of a particular language makes it possible to identify it automatically. For example, for Ukrainian texts, it was found that statistical parameters of styles can be frequencies of vowels, consonants, spaces between words, as well as soft and sonoric groups of consonants [3]. We will show how to evaluate the speech of a certain author taking a particular passage of his work [77] with the help of a certain reference – for example, the data on the frequency of the letters of the Ukrainian language. Consider two passages of a technical text in Ukrainian, presented in the format, where the letters are arranged in the order of descending frequencies of their occurrence in the passage (frequencies are presented in Table 1), and small and capital letters are not distinguished. We will find the type of correlation of the letters of the passages [76] and the standard [77], and the results proving these conclusions will be presented, in particular, in a graphical form.

Table 1

Frequencies of letter occurrence in reference passage and in studied passage

Letter	Frequency of using the letters of the Ukrainian language (reference sample)	Absolute frequency of letters in Passage 1	Relative frequency of letters in Passage 1	Absolute frequency of letters in Passage 2	frequency of letters in Passage 2
« »	0.133	80	0.14	82	0.15
О	0.082	37	0.07	41	0.08
а	0.074	43	0.08	31	0.06
н	0.068	33	0.06	30	0.06
и	0.054	27	0.05	27	0.05
в	0.047	29	0.05	19	0.04
т	0.046	25	0.04	20	0.04
е	0.038	26	0.05	45	0.08
р	0.036	15	0.03	16	0.03
с	0.033	22	0.04	27	0.05
м	0.031	10	0.02	13	0.02
к	0.031	22	0.04	20	0.04
л	0.028	17	0.03	30	0.06
д	0.028	16	0.03	4	0.01
у	0.025	19	0.03	14	0.03
п	0.025	11	0.02	21	0.04
я	0.024	15	0.03	6	0.01
з	0.018	9	0.02	8	0.01
б	0.016	7	0.01	5	0.01
ч	0.015	5	0.01	11	0.02
г	0.012	4	0.01	6	0.01
ю	0.012	2	0.00	2	0.00
б	0.011	7	0.01	5	0.01
х	0.01	4	0.01	7	0.01
ц	0.009	7	0.01	1	0.00
ж	0.007	3	0.01	7	0.01
й	0.007	4	0.01	6	0.01
ш	0.005	3	0.01	2	0.00
щ	0.004	3	0.01	1	0.00
ф	0.003	1	0.00	0	0.00
others	0.0605	51	0.09	34	0.06

The following data were entered in Table 1 for convenience: frequency of using the letters of the Ukrainian language, absolute and relative frequencies of using the letters in studied Passage 1 by Author 1 [76] and Passage 2 by Author 2 [77]. Note that Passage 1 contains 556 characters, Passage 2 contains 541 characters. Note that the term "others" in the column of letters contains authentic letters for the Ukrainian language (ї, є, ґ, і), which are less used in most technical texts. This makes it possible to achieve certain independence in analysis. Show the obtained results graphically in Fig. 1.

The graphical representation of relative frequencies of letters' occurrence in the passages gives a convincing answer to the question which of the passages is written by which author.

1-gram distribution in the works is different. 3-gram analysis gives the optimal indicators of text research [3]. This will be checked at the following stages of research. There is an abrupt jump in relative frequency of occurrence of letter

"e" for Passage 2 relative to the reference values of Reference passage 1 [77] (Fig. 2), so we will assume that it is more likely that Reference passage 1 was written by the author of Passage 1 [76]. We will also give numerical values for the correlation of letters' frequency in the Passages and in the Reference passage. Find two correlation factors: for the reference passage and for Passage 1 [76] and for the reference passage and Passage 2 [78]; the factor that is closest to unity will indicate the greater probability of belonging to the corresponding passage to the reference passage. Calculations of the correlation factor for the reference passage and Passage 1 give $R_{e\text{-}P1}$=0.962716, and correlation factor for the reference passage and Passage 2 – $R_{e,P2}$=0.909958. Similarly, the values of relative frequencies in Reference passage 2 and Passages 1, 2 in Fig. 3 are substantially different, so it is likely that the author of Reference passage 2 [75] is not the author of Passages 1, 2.

The obtained values of the factors, as well as analysis of graphic results, suggest that the probability of belonging of Passage 1 [76] to Reference passage 1 [77] is higher than for Passage 2 [75].

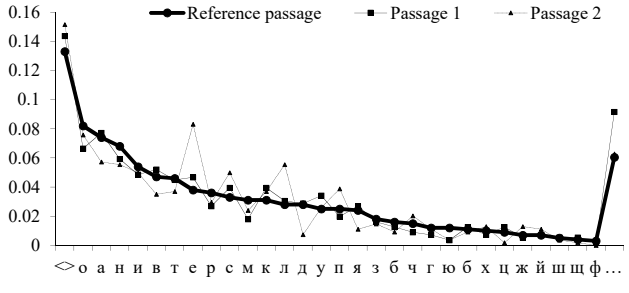


Fig. 1. Graphical representation of relative frequencies of letters' occurrence in the reference passage and in the studied passage

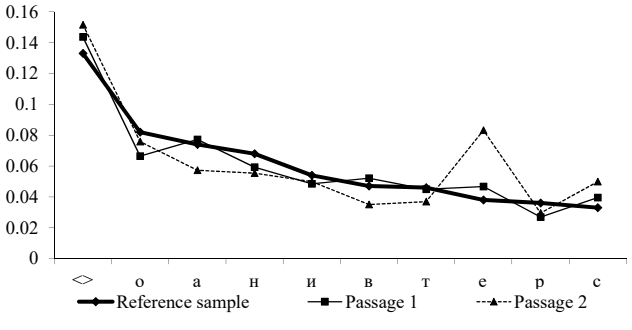


Fig. 2. Graphic representation of relative frequencies of occurrence of ten most frequent characters in Reference passage 1 and in the studied Passages 1, 2, including spaces

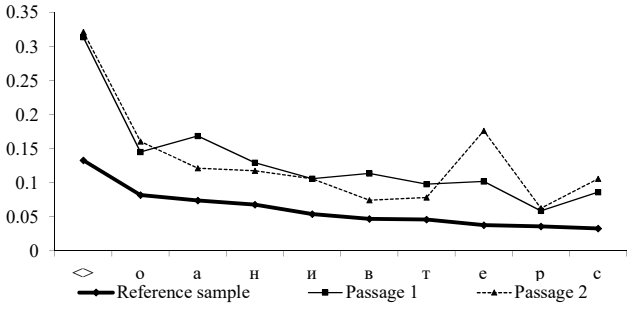


Fig. 3. Graphic representation of relative frequencies of occurrence of ten most frequent characters in Reference passage 2 and studied Passages 1, 2, including spaces

5. Content-monitoring software to identify the author in Ukrainian language texts

To achieve the aim of the research, the system with the possibility to choose the language/languages of the analyzed content, which is realized at the Web-resource Victana, was developed. For qualitative and effective content analysis, when determining the degree of authorship of a particular person, we propose to analyze the reference text and the studied one at several stages:

– linguometric analysis of coefficients of diversity of the author's speech (alg. 1);

– stylometric analysis (alg. 2);

– analysis of set expressions (alg. 3);

– linguistic analysis via *N*-gram (alg. 4).

At the Web-resource, there are the following fields for linguometric analysis (Fig. 4):

– Characters. (The input text must contain not less than 100 and not more than 10,000 characters.) – the maximum size of content is set.

– Content – the field where the studied text is copied from the buffer.

– Calculate – calculation run.

– Clear – clear the input data.

Algorithm 1. Linguometric analysis of the text for authorship attribution

Step 1. Checking the text length – the excess is removed.

Step 2. Determining the number of sentences.

Step 3. Clearing the studied text (figures, special characters).

Step 4. Determining the total number of the words in text *N*.

Step 5. Determining the number of words *W*.

Step 6. Determining the number of prepositions *Z*.

Step 7. Determining the number of conjunctions *S*.

Step 8. Calculation of author's speech coefficients

Step 9. Results output to the end user (Table 2, Fig. 4).

Table 2

Example of calculations of author's speech coefficients

No.	Coefficient	Input data	Calculation
1.	Coefficient of lexical diversity: $K_l=W/N$	W=184 N=295	K_l=0.62372881355932
2.	Coefficient of syntactic complexity: $K_s=1-P/W$	P=18 W=184	K_s=0.90217391304348
3.	Coefficient of speech coherence: $K_z=(Z+S)/(3*P)$	Z=20 S=28 P=18	K_z=0.88888888888889
4.	Exclusiveness index: $I_{wt}=W_1/W$	W_1=141 W=184	I_{wt}=0.76630434782609
5.	Concentration index: $I_{kt}=W_{10}/W$	W_{10}=2 W=184	I_{kt}=0.010869565217391

There are the following fields for stylometric analysis at the Web-resource (Fig. 5):

– Reference text is the field where the Reference text is copied from the buffer.

– Choose Passage 1 (2, 3) – open the access to the passages. The access to the following passage is only after activation of the access to the previous one. The access is opened sequentially from the smaller to the larger number.

– Passage 1 (2, 3) is the field where the text of the corresponding passage is copied.

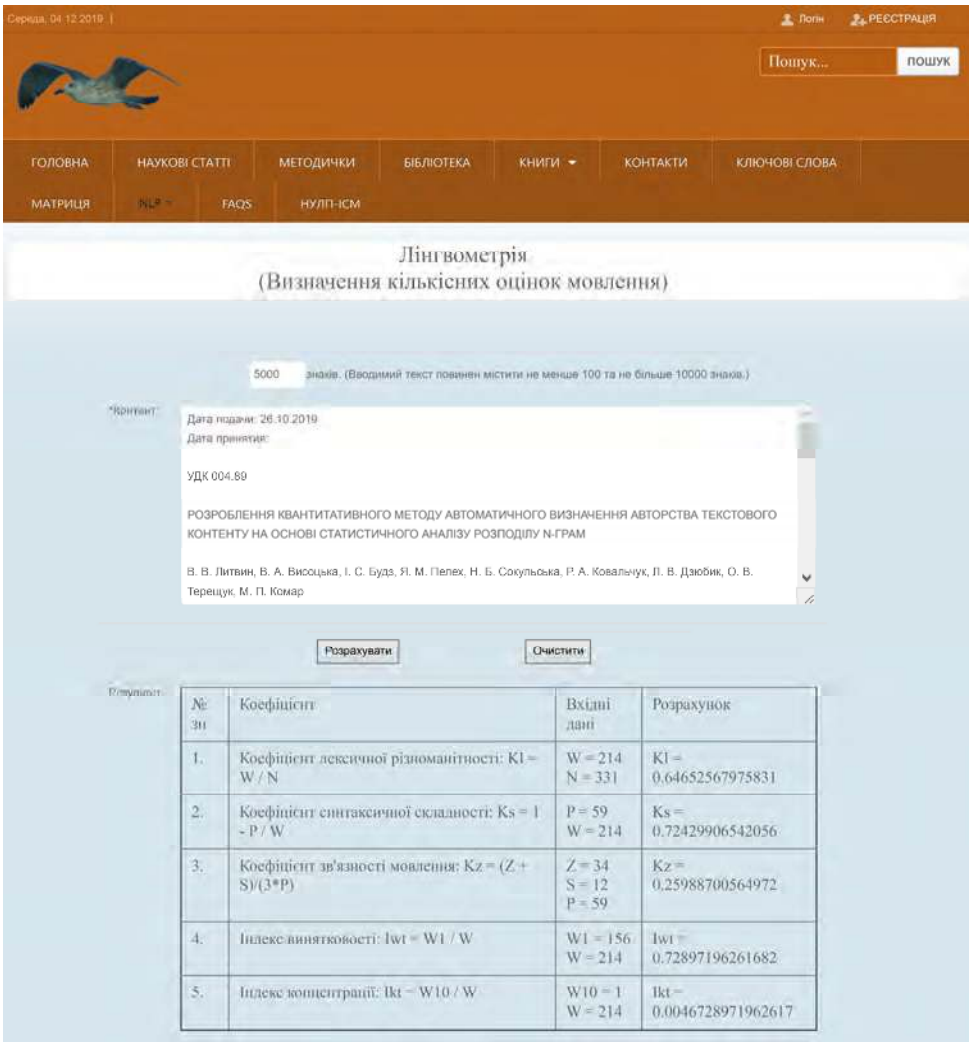


Fig. 4. Example of the result of using linguometric analysis

The input text should have not less than 100 characters. (Now 0) – After the calculation run, the actual number of characters of each passage separately will be computed and shown.

– Calculate – calculation run.

– Clear – clearing the input data.

Algorithm 2. Stylometric analysis of the text for authorship attribution.

Step 1. Checking the lengths of the reference text and of selected passages and bringing the length of the reference text to the minimum of the checked texts.

Step 2. Clearing the reference text from special characters and others.

Step 3. Determining the number of words in the reference text.

Step 4. Determining the number of stop-words (preposition+conjunctions+particles) in the text of the reference sample (Fig. 6, 7).

Step 5. The length of Passage 1 is not more than the minimal text.

Step 6. Clearing Passage 1 from special characters and others.

Step 7. Determining the number of words W1 for Passage 1.

Step 8. Determining the number of stop-words (preposition+conjunctions+particles) in the text.

Step 9. Preparation of separate arrays (passage and reference text) for the calculation of correlation factor (Fig. 7).

Step 10. Click the function for calculation of correlation factor.

Step 1. Generation of the array for formation of the graphic image of the relative frequency of occurrence of stop-words in Passage 1 and in the reference sample.

Step 12. Click the function for calculation of the diagram of RF (Fig. 8).

Step 13. Click the function for calculation of correlation factor for Passages 2 (3) for each of auxiliary words.

Step 14. Form the words of the Swadesh list from the directory, determining the number of words from the Swadesh list in the text of the passage (for the reference text and selected passages – Table 3).

Step 15. From the lists common for the Reference sample, Passages 1–3 and the Swadesh list.

Step 16. Research results are displayed on the screen (Table 4).

For automated text processing, not only the frequency of occurrence of a particular category in the text is important, but also its presence in the studied text. Quantitative counting makes it possible to draw objective conclusions about the orientation of the materials by the number of using the units of the analysis (key quotations) in the studied texts. Qualitative analysis does the same, but as a result of studying whether (and in what context) some important, original category is found in general. Summing up, it should be noted that the use

of content-analysis for the creation of information systems makes it possible to catch the prevalence of a particular feature of the studied totality of texts. In this case, not absolute, but rather relative value of the feature, that is, the characteristic of its place (fraction) among other features is important. Measurement of the ratio between the features in the texts gives empirical material to understand the functional relations between the elements of reality displayed in texts. In the presence of texts that have a chronological sequence, it is possible to have a number of time-fixed portraits of the studied reality, which makes it possible to put forward the hypotheses of predictive nature about functioning of the elements of the system. For example, frequency characteristics of a text (the average sentence length) may indicate certain specificity of intellectual abilities of a person in terms of verbal representation of thoughts. Determining the average sentence length, it is possible to characterize a change in the emotional state of an individual.

The choice of analysis of the vocabulary variant in context dependence is one of the most significant and powerful in psycholinguistic diagnostics. Due to the established coefficient of vocabulary diversity (Table 5) in the speech of a person, it is possible to identify psychopathology, for example, schizophrenia, as well as a tendency to it.

Another criterion of language competence is a verb coefficient (aggressiveness). The essence of this coefficient is the ratio of the number of verbs and verb forms (participles) to the total number of all words. Like in psychology, a high coefficient of aggressiveness indicates a possible high emotional tension of an author, which is reflected in the text by manifestations of a change in the dynamics of events and other characteristic features. Coefficient of logical coherence is also calculated from the formula of the ratio of the total number of functional words (conjunctions, prepositions and particles) to the total number of sentences. At magnitudes within unity, rather harmonious relations between functional words and syntactic constructions are ensured. Embolus coefficient (med. embolus – a blockage of a blood vessel), or speech "contamination" is the ratio of the total number of emboli (words that do not carry semantic load) to the total number of words in a sentence. The structure of the emboli includes exclamations (nu-nu, ha-ha, ehe, zh, oi, etc.), vulgarisms (rough vocabulary), and unnecessary repetitions. The embolic coefficient demonstrates the peculiarities of verbal intelligence and the emotional state of a speaker/author of the text. It can also give an idea of the general culture of speech. Even taking into consideration the fact that a belle-letter text is generally considered to be androgenic and is a weave of subordination functions – the qualities of the author's "I", which are in some way graded depending on the characterological profile of a particular author. In other words, the text of the original and the text of the translation depend on their authors.

Fig. 5. Example of input data for stylometric analysis

There are the following fields for analysis of set expressions in the Web-resource (Fig. 9):

– enter the number of phrases to be displayed on the screen (10; 100) – so many word combinations will be displayed on the screen after calculation;

– select Passage 1 (2, 3) – open the access to passages.

Access to the next passage is only after activating the access to the previous one. Access opens sequentially from a smaller number to a larger one. (Not implemented – only one passage is analyzed);

– passage 1 – the field where the text of the corresponding passage is copied;

– used: 57 % The input test must contain not less than 100 characters. (Limit: 4000) – analysis of the text size.

– calculate – run calculation.

– clear – input data clearing.

Algorithm 3. Quantitative analysis of set expressions.

Step 1. Clearing the obtained content from special characters and others.

Step 2. Form the list of blocked words from the database depending on the chosen language of the context.

Step 3. Preparation to the formation of the arrays of double word combinations and all words. The array at the input: clue – figures, meanings – text, split into sentences (divider dot). The words are compared with the database of the given keywords and by the rule, described in the database, it brings the given word to the word base if it itself is not the word base.

Step 4. Determining set expressions by the FREG method: to obtain the absolute frequency of word combinations (Fig. 10).

Step 5. Determining set expressions by the *t*-test method: $P(W1)*P(W2)$ accounting not only the pairs, but also the frequency of using separate words (those that make up the pair).

Step 6. Determining set expressions by the LR method.

Step 7. Determining set expressions by the X2 method (Table. 6).

Step 8. Research results are displayed on the screen.

If a word is missing in the database, it is added automatically. For this word, a moderator needs to describe the rule of bringing the word to the word base.

When identifying the author of a text, it is assumed that the text reflects the individual manner of author's writing, which makes it possible to distinguish it from others. To compare the texts with each other, it is necessary to compare a text with some numerical characteristic that would be close to the texts of the same author, and would be substantially different for the works by other authors. Such a characteristic can be the density of distribution function (DDF) of letter combinations from three consecutive characters (3-grams). The DDF is defined as a set of empirical frequencies of using the letters or their combinations. When analyzing the text using the DDF, the inclusion of punctuation marks, spaces and figures is not taken into consideration. The task of identification of the author of an unknown text in terms of the DDF is stated as follows. There is a set of texts, which contains works A by well-known authors. Let us assume that K_a is the amount of content by the a-th author, $N_{i,a}$ is the number of characters in the i-th content of the a-th participant, i=1, ..., K_a All the texts in this set are given in the DDF form.

Table 3

Passage 1 words: 153. Reference text words: 153

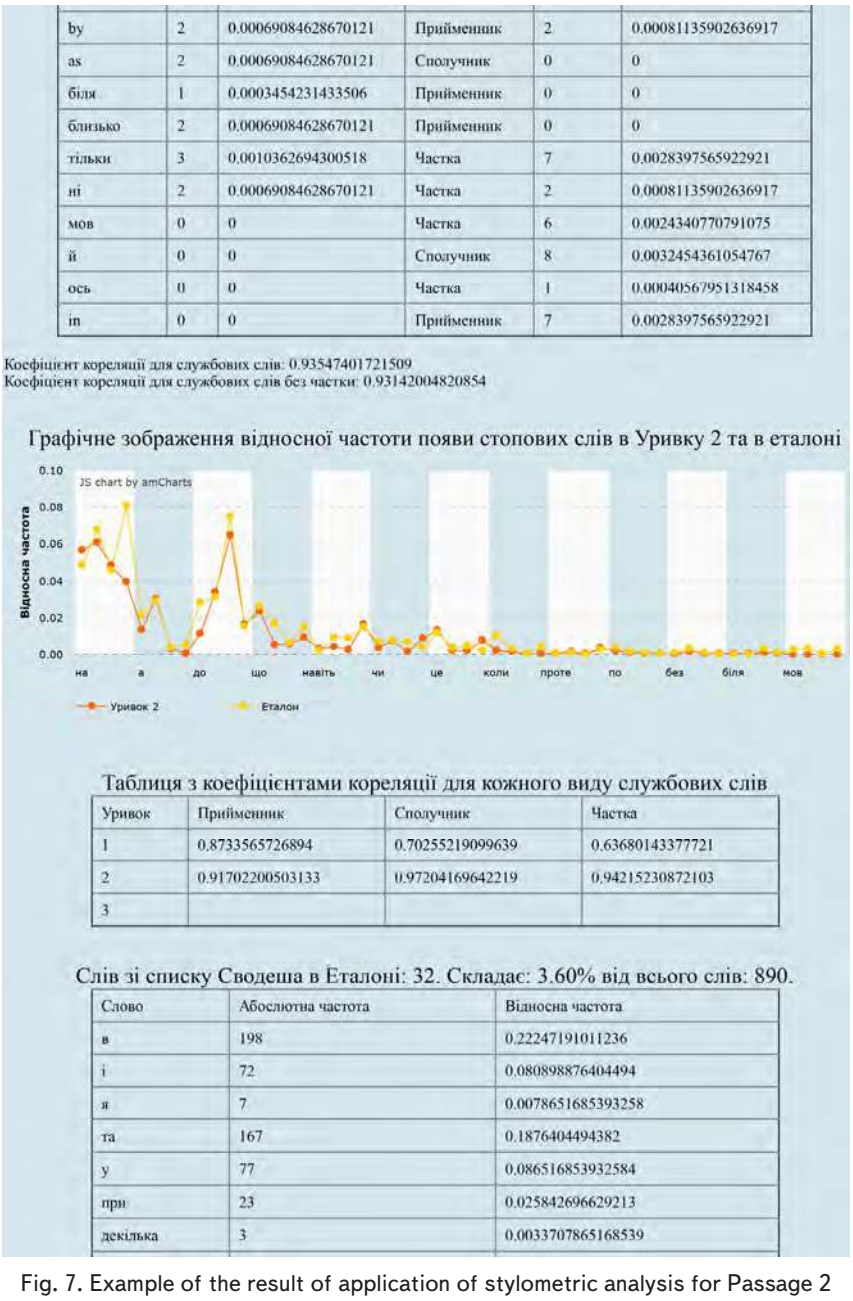
Word	AF	RF	Part of speech	AF reference	RF in reference text
в	5	0.032679738562	Preposition	5	0.032679738562
а	2	0.0130718954248	Conjunction	2	0.0130718954248
це	1	0.0065359477124	Particle	1	0.0065359477124
та	16	0.1045751633987	Conjunction	16	0.1045751633987
для	7	0.0457516339869	Preposition	7	0.0457516339869
з	2	0.0130718954248	Preposition	2	0.0130718954248
ж	1	0.0065359477124	Particle	1	0.0065359477124
і	3	0.019607843137	Conjunction	3	0.019607843137
також	2	0.0130718954248	Conjunction	2	0.0130718954248
мов	2	0.0130718954248	Particle	2	0.0130718954248
у	1	0.0065359477124	Preposition	1	0.0065359477124
що	1	0.0065359477124	Conjunction	1	0.0065359477124
за	1	0.0065359477124	Preposition	1	0.0065359477124

Розрахувати Очистити

Уривок 1 слів: 3046. Еталонний текст слів: 2465.

Стоп-слово	АЧ	ВЧ	Частина мови	АЧ етал.	ВЧ в еталоні
та	158	0.051871306631648	Сполучник	167	0.067748478701826
з	149	0.048916611950098	Прийменник	113	0.045841784989858
в	129	0.042350623768877	Прийменник	198	0.080324543610548
а	44	0.014445173998687	Сполучник	53	0.021501014198783
і	99	0.032501641497045	Сполучник	72	0.02920892494929
for	33	0.010833880499015	Прийменник	8	0.0032454361054767
and	136	0.044648719632305	Сполучник	13	0.00527383367139 96
для	166	0.054497701904137	Прийменник	183	0.074239350912779
по	33	0.010833880499015	Прийменник	9	0.0036511156186613
це	10	0.0032829940906106	Частка	29	0.011764705882353
від	14	0.0045961917268549	Прийменник	42	0.017038539553753
до	31	0.010177281680893	Прийменник	70	0.028397565922921
через	22	0.0072225869993434	Прийменник	2	0.00081135902636917
без	6	0.0019697964543664	Прийменник	2	0.00081135902636917
або	2	0.00065659881812213	Частка	38	0.015415821501014
за	48	0.015758371634931	Прийменник	37	0.01501014198783
чи	9	0.0029546946815496	Частка	16	0.0064908722109533
на	128	0.042022324359816	Прийменник	120	0.04868154158215
якщо	1	0.00032829940906106	Сполучник	10	0.0040567951318458
не	33	0.010833880499015	Частка	37	0.01501014198783
то	1	0.00032829940906106	Частка	6	0.0024340770791075
так	13	0.0042678923177938	Частка	9	0.0036511156186613
що	16	0.005252790544977	Сполучник	64	0.025963488843813
при	7	0.0022980958634274	Прийменник	23	0.0093306288032454
щоб	16	0.005252790544977	Сполучник	5	0.0020283975659229
коли	4	0.001313197636244 3	Сполучник	25	0.010141987829615
лише	1	0.00032829940906106	Частка	11	0.0044624746450304

Fig. 6. Example of the result of application of stylometric analysis for Passage 1



by	2	0.00069084628670121	Прийменник	2	0.00081135902636917
as	2	0.00069084628670121	Сполучник	0	0
біля	1	0.0003454231433506	Прийменник	0	0
близько	2	0.00069084628670121	Прийменник	0	0
тільки	3	0.0010362694300518	Частка	7	0.0028397565922921
ні	2	0.00069084628670121	Частка	2	0.00081135902636917
мов	0	0	Частка	6	0.0024340770791075
й	0	0	Сполучник	8	0.0032454361054767
ось	0	0	Частка	1	0.00040567951318458
in	0	0	Прийменник	7	0.0028397565922921

Уривок	Прийменник	Сполучник	Частка
1	0.8733565726894	0.70255219099639	0.63680143377721
2	0.91702200503133	0.97204169642219	0.94215230872103
3			

Слово	Абослютна частота	Відносна частота
в	198	0.22247191011236
і	72	0.080898876404494
я	7	0.0078651685393258
та	167	0.1876404494382
у	77	0.086516853932584
при	23	0.025842696629213
декілька	3	0.0033707865168539

Fig. 7. Example of the result of application of stylometric analysis for Passage 2

Table 4

Common words for Reference sample, Passages 1–3 and Swadesh list: 8. It is 26.67 % of the total number of words: 30

No	Common	AF	Reference sample	Passage 1	Passage 2	Passage 3
1	в	5	0.167	0.167	0.167	0.167
2	це	1	0.033	0.033	0.033	0.033
3	та	16	0.533	0.533	0.533	0.533
4	з	2	0.167	0.167	0.167	0.167
5	коло	1	0.033	0.033	0.033	0.033
6	і	3	0.1	0.1	0.1	0.1
7	у	1	0.033	0.033	0.033	0.033
8	що	1	0.033	0.033	0.033	0.033

The DDF of the content, the volume of which is equal to $N_{i,a}$, is assigned as the set of values $f_{i,a}(j)=k_j/N_{i,a}$, k_j is the number of using N-grams by number j. Argument $j=1, ..., a(n,M)$, corresponds to the number of the letter combination (N-gram) at the alphabetical order, where M is the capacity of the alphabet of the language of the written text, n is the number of N-gram, that is the number of characters in the letter combination. $a(n, M)=M^n$ is the number of N-grams in the given alphabet. Each author is identified with his averaged weighed DDF according to formula

$$F_a(j)=\frac{1}{N_a}\sum_{i=1}^{K_a} f_{i,a}N_{i,a}.$$

These DDF are the author's references. To compare two texts, or a text and the author's reference text, one must assign the distance between the corresponding distribution functions. The norm in the space of functions as summands is used as distance metrics. Thus, for example, distance $p_{0,a}$ between the DDF of an unknown text f_0 and any author's DDF F_a is calculated as follows:

$$p_{0,a} = \left\| f_0 - F_a \right\| = \sum_{j=1}^{a(n,M)} \left| f_0(j) - F_a(j) \right|.$$

Table 5

Coefficients of frequency characteristics of a text

Coefficients	Formula
Vocabulary diversity	$K_{\text{vocab. diversity}}$=different words/$2N_{\text{all words}}$
Verbs (aggressiveness)	K_{verb}=verbs/$N_{\text{all words}}$·100 %
Emotiveness of a text	$K_{\text{adject.}}$=adjectives/$2N_{\text{all words}}$
Logical coherence	$K_{\text{log.coher.}}$=auxiliary words/$3N_{\text{реч}}$
Embolus (contamination)	$K_{\text{emb.}}$=embol/$N_{\text{all words}}$·100 %

Accordingly, text "0" will belong to the author, the distance of which to the DDF will be the shortest. In solving the classification problem, the dataset was not split explicitly into the test and training sets. Weighted average DDF were constructed throughout the set of the content of one author.

The distance from content i to specific author a was calculated as:

$$p_{i,a} = \frac{\left\| f_{i,a} - F_a \right\|}{1 - N_{i,a} / N_a}.$$

The formula makes it possible to exclude participation of the DDF content i in the average DDF of a specific author. In the Web-resource, there are the following fields for N-gram analysis (Fig. 11):

– Choose the language of the text – language of the text for analysis (research). By default, "Ukrainian".

– The number of grams – the number of characters in the gram. By default, 3. It can be changed for 1, 2, 3, 4.

– Text limitation in characters.

– Text is the field, where the studied text is copied from the buffer.

– Generate – to run N-gram generation.

– Clear – clearing the entered data.

Слова, спільні для Еталону, Уривків 1–3 та списку Сводеша

№ зп	Спільні слова	ВЧ в Еталоні	ВЧ в Уривку 1	ВЧ в Уривку 2	ВЧ в Уривку 3
1	в	0.22247191011236	0.17503392130258	0.12166488794023	
2	і	0.080898876404494	0.13432835820896	0.093916755602988	
3	та	0.1876404494382	0.21438263229308	0.18676627534685	
4	у	0.086516853932584	0.084124830393487	0.10458911419424	
5	при	0.025842696629213	0.0094979647218453	0.012806830309498	
6	те	0.0044943820224719	0.005427408412483	0.0085378868729989	
7	той	0.002247191011236	0.001356852103120	0.0021344717182497	
8	зі	0.0044943820224719	0.0094979647218453	0.0032017075773746	
9	з	0.12696629213483	0.20217096336499	0.14834578441836	
10	цей	0.0044943820224719	0.0040705563093623	0.0064034151547492	
11	як	0.025842696629213	0.025780189959294	0.071504802561366	
12	що	0.071910112359551	0.021709633649932	0.073639274279616	
13	це	0.032584269662921	0.013568521031208	0.040554962646745	
14	коли	0.028089887640449	0.005427408412483	0.0064034151547492	
15	три	0.0044943820224719	0.001356852103120	0.0021344717182497	
16	де	0.0078651685393258	0.013568521031208	0.0021344717182497	
17	ні	0.002247191011236	0.0027137042062415	0.0021344717182497	
18	якщо	0.01123595505618	0.001356852103120	0.026680896478122	
19	мати	0.0056179775280899	0.0027137042062415	0.0096051227321238	
20	великий	0.002247191011236	0.001356852103120	0.0010672358591249	
21	два	0.0056179775280899	0.001356852103120	0.0021344717182497	
22	багато	0.001123595505618	0.0067842605156038	0.0032017075773746	

Fig. 8. Example of the result of stylometric analysis for Passages1–3

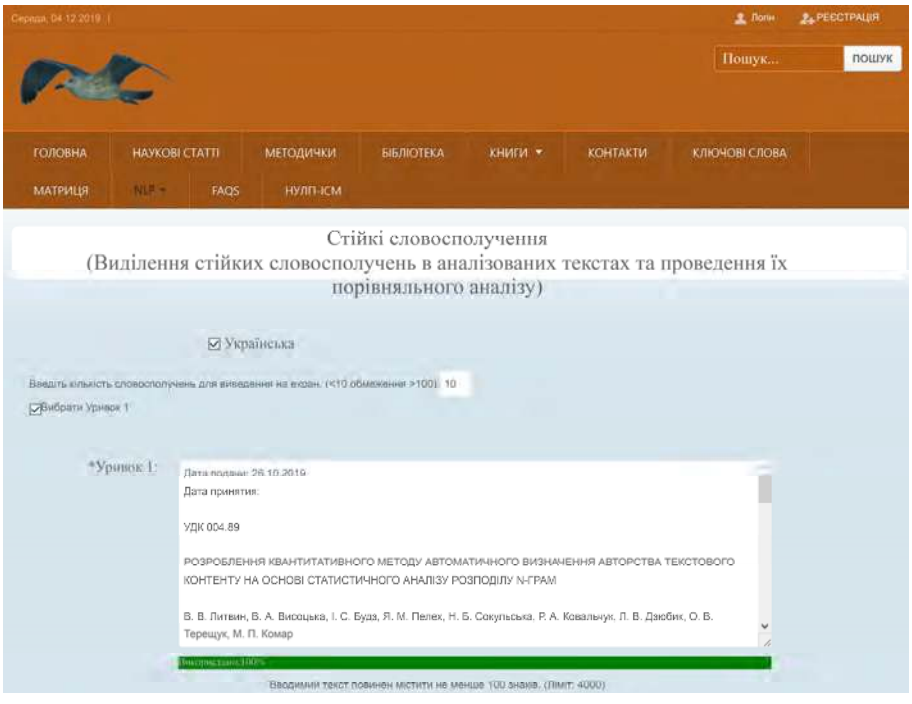


Fig. 9. Example of using analysis of set expressions

Table 6

The list based on rating frequency of set expressions occurrence for article 1, word combinations: 45. Total number of words: 108

No	FREG			t-test		LR		X2	
	Word combination	AF	RF	Word combination	*t*	Word combination	logL	Word combination *e*	2
1	system electronic	4	0.088889	system electronic	1.822222	information technology	5.03e–1	decision making	45.000000
2	information system	4	0.088889	electronic content-commerce	1.578091	intelligent system	2.13e–1	system electronic	45.000000
3	electronic content-commerce	3	0.066667	section scientific	1.319933	information system	8.36e–2	electronic content-commerce	32.946429
4	section scientific	2	0.044444	information system	1.222222	portal scientific	5.58e–2	section scientific	29.302326
5	portal scientific	1	0.022222	decision making	0.977778	course technology	3.31e–2	course technology	21.988636
6	intelligent system	1	0.022222	course technology	0.955556	storage data	3.31e–2	storage data	21.988636
7	decision making	1	0.022222	storage data	0.955556	decision making	8.27e–3	portal scientific	14.318182
8	course technology	1	0.022222	portal scientific	0.933333	section scientific	1.89e–3	information system	5.848550
9	storage data	1	0.022222	intelligent system	0.777778	electronic content-commerce	1.55e–4	intelligent system	3.579545
10	information technology	1	0.022222	information technology	0.688889	system electronic	1.37e–6	information technology	1.890409

Algorithm 4. Linguo-statistical analysis of *N*-grams of the text.

Step 1. Clearing the studied text (figures, special characters).

Step 2. Calculate the number of the words in the text

Step 3. All three words in the text are put into lower case.

Step 4. Delete spaces.

Step 5. Depending on the chosen language, substitute the appropriate alphabet.

Step 6. Depending on the established number of grams, run the appropriate function, which calculate all possible variants of grams and stores them in the array.

Step 7. Next, run the function of calculation of the number of occurrence of words.

Here, we calculate relative occurrence frequency and store in the array: the sequence number of the gram, the gram itself, the number of occurrences of this gram, relative frequency of occurrence of this gram.

Step 8. The next function forms the array for exporting to the CSV file, obtained in the previous function. This file is stored on the server. It can be downloaded to the user's (researcher's) computer by the link, which will be accessed after generating the form with the research results.

Step 9. Research results are displayed on the screen (only the grams found in the text).

Step 10. Access to the export file is opened.

Step 11. The summarizing results are displayed:

– alphabet size;

– number of words in the text;

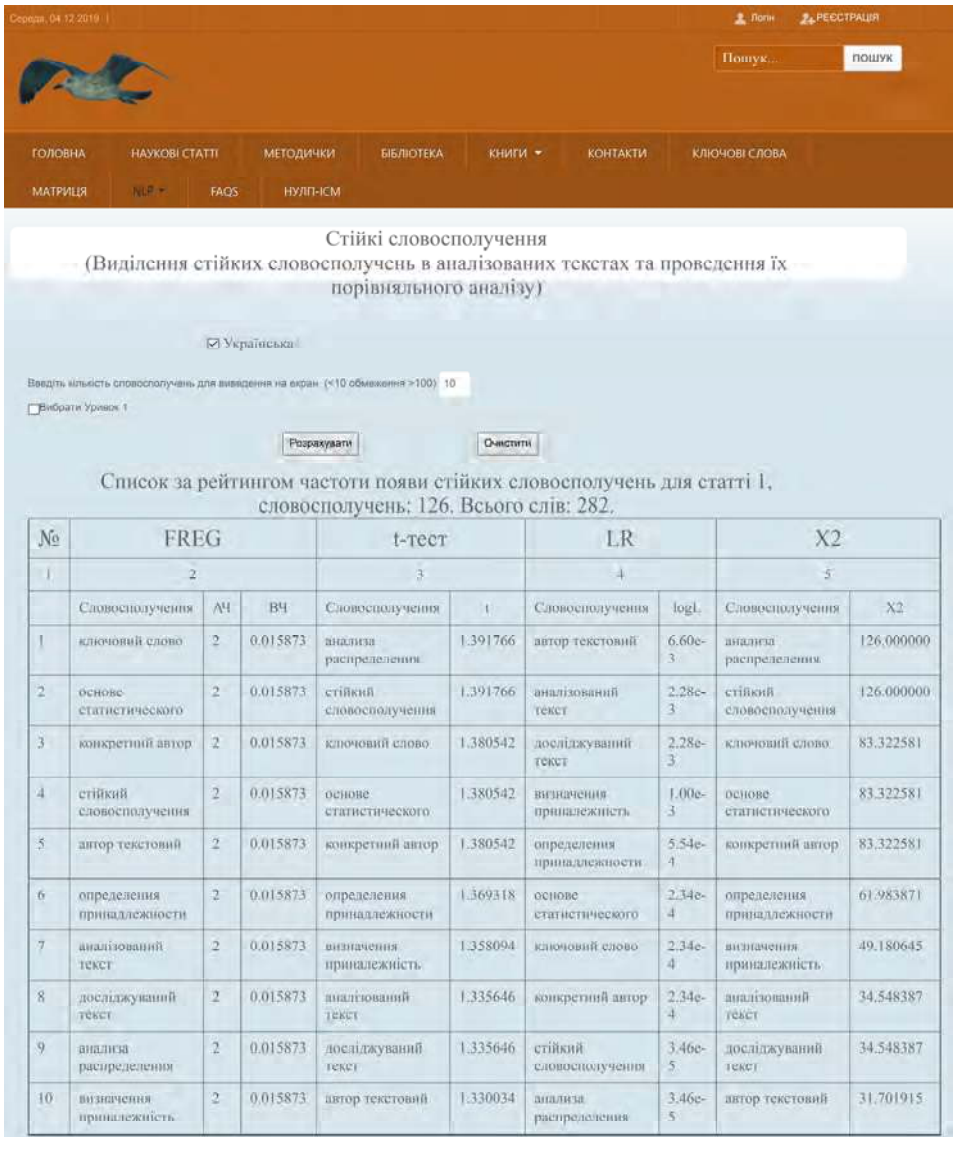
– number of characters in the text with spaces;

– number of characters in the text that was completely cleared;

– total number of *N*-grams;

– total number of found *N*-grams without repetitions;

– total number of found *N*-grams with repetitions.



№	FREG			t-тест		LR		X2	
1	2			3		4		5	
	Словосполучення	АЧ	ВЧ	Словосполучення	t	Словосполучення	logL	Словосполучення	X2
1	ключовий слово	2	0.015873	анализа распределения	1.391766	автор текстовий	6.60e-3	анализа распределения	126.000000
2	основе статистического	2	0.015873	стійкий словосполучення	1.391766	аналізований текст	2.28e-3	стійкий словосполучення	126.000000
3	конкретний автор	2	0.015873	ключовий слово	1.380542	досліджуваний текст	2.28e-3	ключовий слово	83.322581
4	стійкий словосполучення	2	0.015873	основе статистического	1.380542	визначення приналежність	1.00e-3	основе статистического	83.322581
5	автор текстовий	2	0.015873	конкретний автор	1.380542	определения принадлежности	5.54e-4	конкретний автор	83.322581
6	определения принадлежности	2	0.015873	определения принадлежности	1.369318	основе статистического	2.34e-4	определения принадлежности	61.983871
7	аналізований текст	2	0.015873	визначення приналежність	1.358094	ключовий слово	2.34e-4	визначення приналежність	49.180645
8	досліджуваний текст	2	0.015873	аналізований текст	1.335646	конкретний автор	2.34e-4	аналізований текст	34.548387
9	анализа распределения	2	0.015873	досліджуваний текст	1.335646	стійкий словосполучення	3.46e-5	досліджуваний текст	34.548387
10	визначення приналежність	2	0.015873	автор текстовий	1.330034	анализа распределения	3.46e-5	автор текстовий	31.701915

Fig. 10. Example of the result of using analysis of set expressions

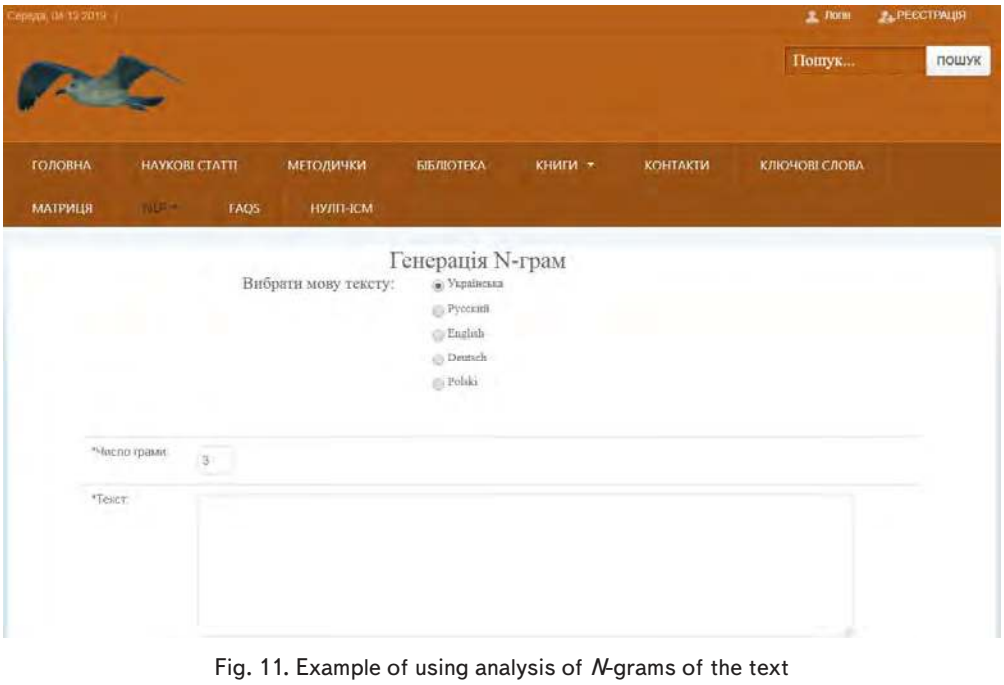


Fig. 11. Example of using analysis of *N*-grams of the text

6. Results of experimental testing of the proposed content-monitoring method for text authorship attribution

Compare three publications [1, 74, 77] in scientific technical area based on the linguo-statistical analysis of 3-grams. Article 1, 2 were written by one team of authors [1, 74], Article 3 was written by another author [77] (Table 7). The language of the text is Ukrainian (there are 33 letters in the alphabet, then the number of all possible *N*-grams is 35,937).

Table 7

Values of parameters for the analyzed articles 1–3

Parameters	Article 1	Article 2	Article 3
Total number of *N*-grams	35,937	35,937	35,937
Total number of found *N*-grams without repetitions)	4,354	4,377	3,890
Total number of found *N*-grams with repetitions	29,494	29,862	36,383
Total number of words	5,475	5,358	6,060
Total number of characters in uncleared text	39,792	39,663	47,084
Total number of characters in cleared text	29,967	32,570	37,062

When comparing the articles, we will take into account only the 3-grams, which were found in the text in three articles simultaneously at least one time. That is why for this specific example, there are in total 2147 3-grams. That is, for Article 1, we analyze 78.4814 % of 3-grams, for Article 2 – 72.6332 % and for Article 3 – 84.1271 %. Accordingly, the difference in using the corresponding 3-grams between Articles 1 and 2 is R_{12}=56.5254 %, between Articles 2 and 3 – R_{23}=69.4271 %, between Articles 1 and 3 – R_{13}=62.9839 %. These indicators show that the characteristics of Articles 1 and 2 are more similar ($R_{23}>R_{12}$ by 12.9017 %, $R_{23}>R_{13}$ by 6.4432 %, $R_{13}>R_{12}$ by 6.4585 %, that is, $R_{23}>R_{13}>R_{12}$), than the characteristics of Articles 1–3 and 2–3, respectively. The lower R_{ij}, the higher the degree of confidence that the articles were written by the same author. Then Article 1 and 2 are more likely to be written by one author/ team of authors, than Articles 2–3 and Articles 1–3, respectively. Analyze the use of separate clusters of 3-grams in corresponding articles and compare the results obtained. Fig. 12, 13 show the results of using in Articles 1–3 of 3-grams, which begin with letter a (occurrence in Articles 1–3 in the range of 6.1125–6.7087 %). Most often, the curves for Articles 1–2 (4.2322 %) and Articles 1–3 (4.197 %) coincide or approach each other (average divergence 0.02713 % and 0.0269 %, respectively). But not always – there is a coincidence with Articles 2–3 (4.6322 %) and there are substantial divergences (average divergence is 0.02969 %). If we analyze only such 3-grams, it follows that all 3 articles are most likely to have been written by one author. This is explained by the fact that this letter is one of the most often used for the formation of Ukrainian words.

Fig. 14, 15 show the results of analysis of using in Articles 1–3 of 3-grams beginning with letter б (occurrence in articles 1–3 in the range of 0.48884–0.77738 %). Most often the curves for Articles 1–2 (0.594 %), unlike for Articles 1–3 (0.7072 %) and Articles 2–3 (1.1208 %), coincide or approach each other. But the trajectories of the curve of Article 1 and Article 3 coincide most often (most likely the articles were written by one author – average divergence is 0.01809 %, while for articles 1–2 – 0.0261 % and for Articles 2–3 – 0.02866 %). If we analyze only the 3-grams (which are less common), it turns out that all Articles 1–2 are more likely to have been written by one author, and Article 3 – by another author. This is explained by the fact that this letter б is rare in the formation of Ukrainian words. And some authors use such words more often because of the habit and/or because of the subject of their publications (this needs additional research).

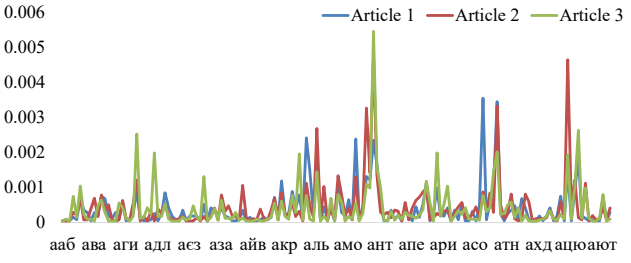


Fig. 12. The graph of using 3-grams beginning with letter *а*

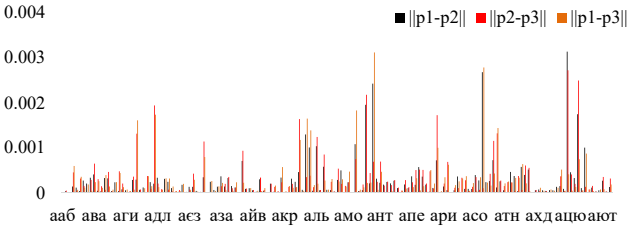


Fig. 13. The graph of the difference of using 3-grams beginning with letter *а*

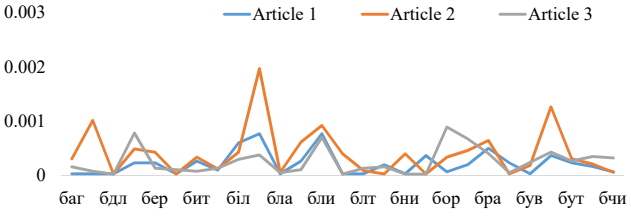


Fig. 14. Graph of using 3-grams beginning with letter б

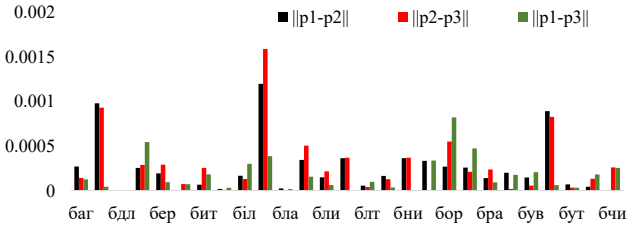


Fig. 15. The graph of the difference of using 3-grams beginning with letter б

Fig. 16 shows the results of analysis of using in Articles 1–3 of 3-graphs beginning with letter в (occurrence in articles 1–3 in the range of 4.2622–4.5219 %). Most often the lines of curves for Articles 1–2 (3.55581 %), Articles 1–3 (3.6523 %) and Articles 2–3 (4.1064 %) coincide or approach each other (average divergence is 0.03067 %, 0.03149 % and 0.0354 % respectively). According to these data, all three articles are most likely to have been written by one author.

Fig. 17, 18 show the results of analysis of using 3-grams beginning with the letter г (occurrence in articles 1–3 in the range of 0.7493–1.4544 %) in Articles 1–3.

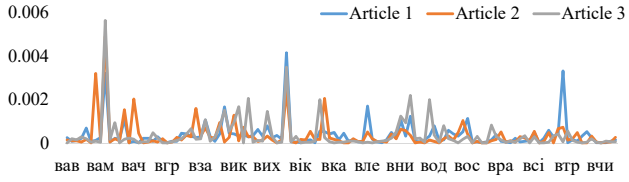


Fig. 16. The graph of using 3-grams beginning with letter в

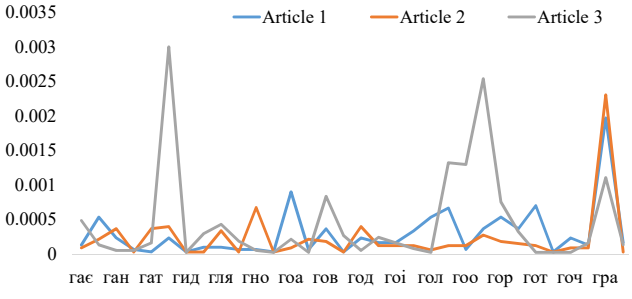


Fig. 17. The graph of using 3-grams beginning with letter г

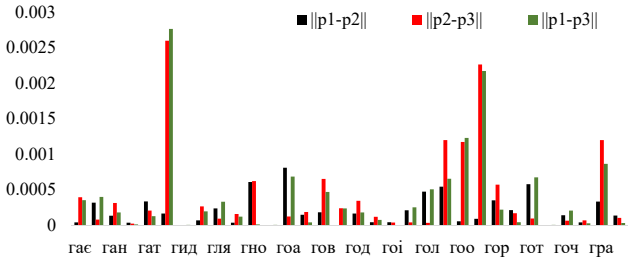


Fig. 18. The graph of the difference of using 3-grams beginning with letter г

Most often the lines of curves for Articles 1–2 (0.6551 %), unlike Articles 1–3 (1.309 %) and Articles 2–3 (1.3451 %), coincide or approach each other. But the trajectories of the curves of Article 1 and Article 2 coincide most often (most often they are written by one author, average divergence is 0.02047 %, while for articles 2–3 it is 0.04203 %, for articles 1–3 – 0.04091 %). If we analyze only such 3-grams (which are less common), it turns out that Articles 1–2 were written more likely by one author, Articles 2–3 and Articles 1–3 were definitely written by different authors.

7. Discussion of results of research into the authorship attribution in Ukrainian-language texts based on the technology of statistical linguistics

According to the data in Table 8 and Fig. 19, a part of the letters in the Ukrainian language are most often used, the others are used much more rarely. For most used letters, the frequency of occurrence of 3-grams with such initial letters will have almost the same distribution (peak values in the graph in Fig. 19), while for other letters, the distribution will not be the same.

Therefore, it is advisable to investigate only trigrams for the beginning letters, which are less common in texts of a specific language in order to determine the degree of belonging of a text to a particular author (for example, Fig. 20, 21). Thus, for the 3-grams beginning with letter є (occurrence in articles 1–3 in the range of 0.2517 –0.707 %), most frequently the curves for Articles 1–2 (0.2508 %), unlike for Articles 1–3 (0.6077 %) and Articles 2–3 (0.5443 %) coincide or approach each other. However, the trajectories of the curve of Article 1 and Article 2 most often coincide (the articles are most likely to have been written by one author – average divergence is 0.0114 %, while for Articles 2 – 3 – 0.02478 % and for articles 1–3 – 0.02762 %, this value is 2 times as high).

Table 8

Distribution of frequencies of appearance of 1-grams in Articles 1–3

No	1-gram	Article 1		Article 2		Article 3	
		Number	RF	Number	RF	Number	RF
1	А	2,255	0.075252	2,698	0.082837	2,491	0.066685
2	Б	284	0.009477	569	0.017470	428	0.011458
3	В	1,654	0.055196	1,590	0.048818	1,915	0.051265
4	Г	408	0.013615	373	0.011452	651	0.017427
5	Д	859	0.028666	939	0.028830	1,319	0.035310
6	Е	1,404	0.046853	1,453	0.044612	2,090	0.055950
7	Є	188	0.006274	165	0.005066	347	0.009289
8	Ж	246	0.008209	210	0.006448	176	0.004712
9	З	623	0.020790	644	0.019773	946	0.025325
10	И	1,732	0.057799	1,852	0.056862	2,036	0.054504
11	І	1,789	0.059701	1,967	0.060393	2,250	0.060233
12	Ї	174	0.005807	217	0.006663	270	0.007228
13	Й	239	0.007976	260	0.007983	265	0.007094
14	К	1,279	0.042682	1,110	0.034080	1,453	0.038897
15	Л	1,116	0.037242	927	0.028462	906	0.024254
16	М	808	0.026964	976	0.029966	1,399	0.037451
17	Н	2,471	0.082460	2,370	0.072766	2,888	0.077312
18	О	2,824	0.094240	2,472	0.075898	3,870	0.103601
19	П	647	0.021591	825	0.025330	1,138	0.030464
20	Р	1,335	0.044550	1,722	0.052871	1,893	0.050676
21	С	1,549	0.051692	1,327	0.040743	1,384	0.037050
22	Т	2,102	0.070146	1,956	0.060055	2,141	0.057315
23	У	987	0.032937	960	0.029475	1,195	0.031990
24	Ф	179	0.005973	209	0.006417	137	0.003668
25	Х	355	0.011847	384	0.011790	482	0.012903
26	Ц	224	0.007475	334	0.010255	299	0.008004
27	Ч	459	0.015317	289	0.008873	574	0.015366
28	Ш	117	0.003904	169	0.005189	281	0.007522
29	Щ	95	0.003170	52	0.001597	128	0.003427
30	Ь	498	0.016619	418	0.012834	613	0.016410
31	Ю	156	0.005206	277	0.008505	289	0.007737
32	Я	647	0.021591	681	0.020909	864	0.023129

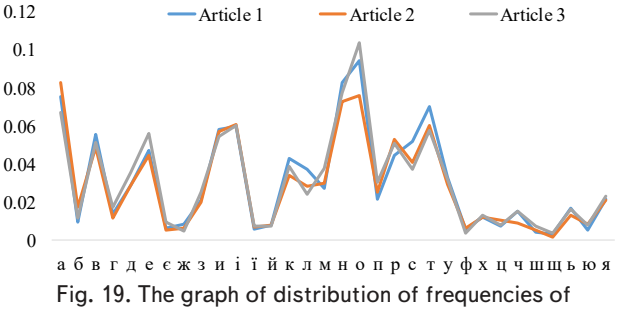


Fig. 19. The graph of distribution of frequencies of occurrence of 1-grams in Articles 1–3

However, often it does not work. Thus, for 3-grams beginning with letter ж (occurrence in Articles 1–3 in the range of 0.3408–0.4738 %), all the curves for Articles 1–2 (0.25 %), Articles 1–3 (0.2126 %) and Articles 2–3 (0.2302 %) coincide or approach each other. Average divergence for Ar-

ticles 1–2 is 0.01786 %, while for Articles 2–3 – 0.01644 % and for Articles 1–3 – 0.01519 %. It looks like all the articles were written by one author. Although the trajectories of the curves in Fig. 22 and the columns of the graph in Fig. 23 show that Articles 1–2 are likely to have been written by one author and Article 3 – by another.

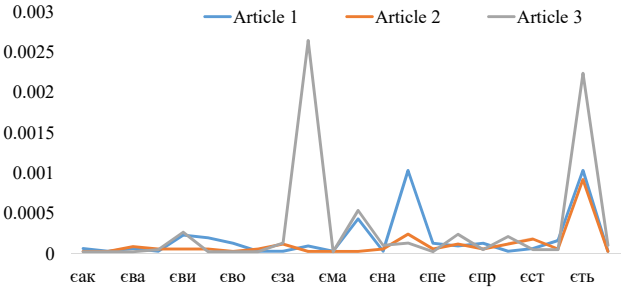


Fig. 20. The graph of using 3-grams beginning with letter є

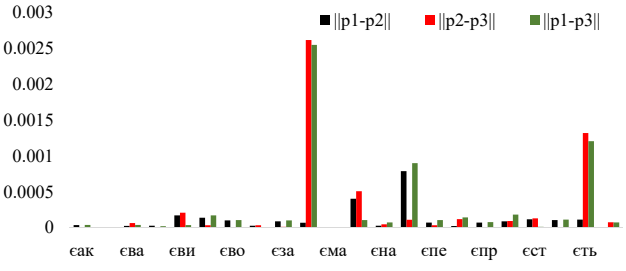


Fig. 21. The graph of the difference of using 3-grams beginning with letter є

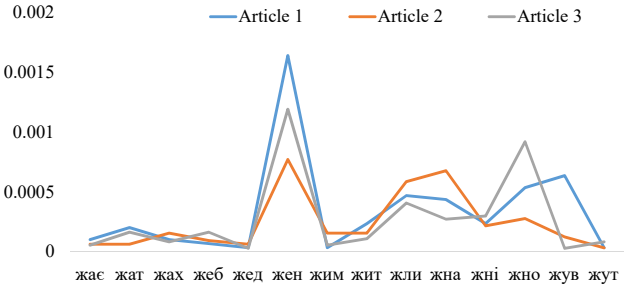


Fig. 22. The graph of using 3-grams beginning with letter ж

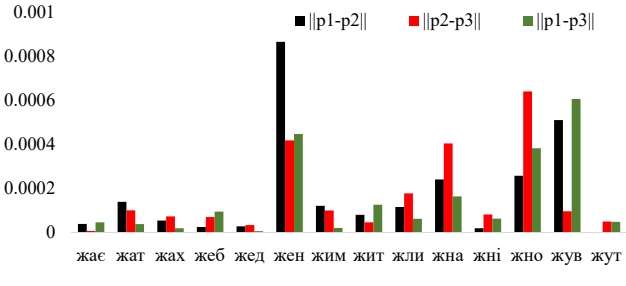


Fig. 23. The graph of the difference of using 3-grams beginning with letter ж

Check again. For 3-grams beginning with letter з (occurrence in articles 1–3 in the range of 1.3108–1.973 %), the curves for Articles 1–2 (1.1879 %), Articles 1–3 (1.3259 %) and Articles 2–3 (1.25 %) coincide or approach each other. Average divergence for Articles 1–2 is 0.02121 %, while for Articles 2–3 – 0.02232 % and for Articles 1–3 – 0.02368 %. It looks like all the articles were written by one author. Although the trajectory of the curves in Fig. 24 shows that Articles 1 are 2 are likely to have been written by one author, and Article 3 – by another one.

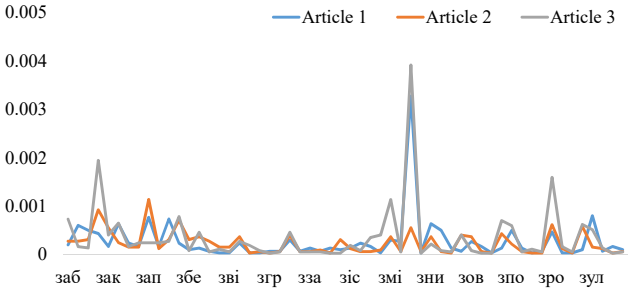


Fig. 24. The graph of using 3-grams beginning with letter з

For 3-grams beginning with letter й (occurrence in articles 1–3 in the range of 0.301–0.4319 %), the curves for Articles 1–2 (0.3352 %), Articles 1–3 (0.3483 %) and Articles 2–3 (0.3469 %) coincide or approach each other. Average divergence for Articles 1–2 is 0.01457 %, while for Articles 2–3 – 0.01508 % and for Articles 1–3 – 0.01514 %. It looks like all the articles were written by one author. Though the trajectory of the curves in Fig. 25 shows that Articles 1–2 are more likely to have been written by one author, and Article 3 – by another.

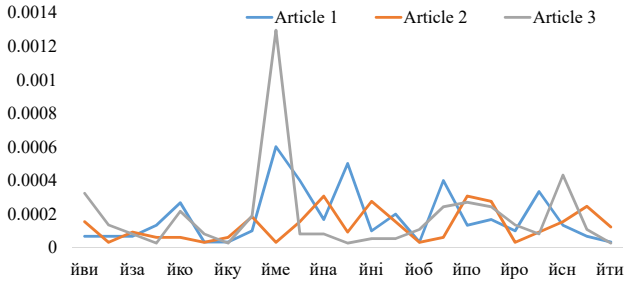


Fig. 25. The graph of using 3-grams beginning with letter й

For 3-gram beginning with letter м (occurrence in articles 1–3 in the range of 2.1681–3.1225 %), the curves for Articles 1–2 (1.7619 %) and Articles 1–3 (1.8193 %), unlike Articles 2–3 (2.6606 %), coincide or approach each other. Average divergence for Articles 1–2 is 0.01936 %, while for Articles 2–3 – 0.02936 % and for Articles 1–3 – 0.02 %. It looks like all the articles were written by one author. Though the trajectory of the curves in Fig. 26 shows that Articles 1 and 2 are more likely to have been written by one author, and Article 3 – by another. Thus, not only the number of occurrence of trigrams with a certain initial letter influences the correctness of the result, but also the frequency of occurrence of such 3-grams.

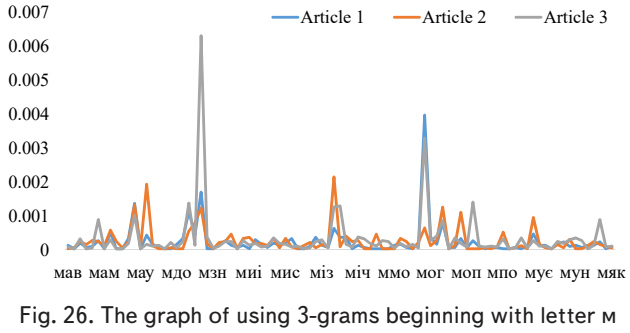


Fig. 26. The graph of using 3-grams beginning with letter м

For 3-grams beginning with letter п (occurrence in articles 1–3 in the range of 1.8583–2.8092 %), all curves for Articles 1–2 (1.6619 %), unlike Articles 1–3 (2.1261 %) and Articles 2–3 (2.5456 %) coincide or approach each other (Fig. 27). Average divergence for Articles 1–2 is 0.04261 %, while for Articles 2–3 – 0.06527 % and Articles 1–3 – 0.05452 %. Articles 1–2 were written by one author, and Article 3 – by another.

Fig. 27. The graph of using 3-grams beginning with letter п

For 3-grams beginning with letter р (occurrence in articles 1–3 in the range of 3.69–4.3802 %), the curves for Articles 1–2 (3.1902 %), Articles 1–3 (3.4834 %) and Articles 2–3 (4.3566 %) coincide or approach each other (Fig. 28). Average divergence for Articles 1–2 is 0.03323 %, while for Articles 2–3 – 0.04538 % and for Articles 1–3 – 0.03629 %. It looks like all the articles were written by one author.

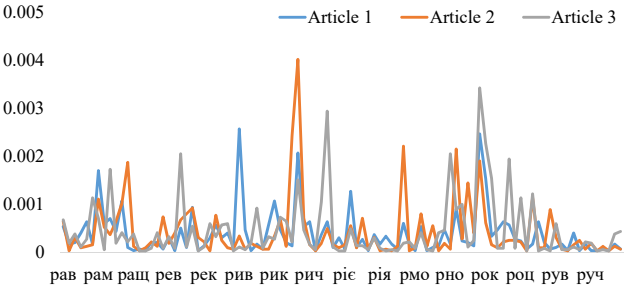


Fig. 28. The graph of using 3-grams beginning with letter р

For 3-grams beginning with letter у (occurrence in articles 1–3 in the range of 2.1927–2.7261 %), the curves for Articles 1–2 (1.7905 %), unlike for Articles 1–3 (1.9443 %) and Articles 2–3 (1.9852 %), coincide or approach each other (Fig. 29). Average divergence for Articles 1–2 is 0.02184 %, while for Articles 2–3 – 0.02421 % and Articles 1–3 – 0.02371 %. It looks like all the articles were written by one author. If the trigram beginning with the certain letter is used in the text more than 1 %, average divergence during comparison some articles, irrespective of the authorship, will be almost the same. Then it is necessary to take into account only the trigrams that have the percentage of occurrence less than 1.

For 3-grams beginning with letter ф (occurrence in articles 1–3 in the range of 0.3069–0.4759 %), all curves for Articles 1–2 (0.2762 %) and Articles 1–3 (0.299 %), unlike Articles 2–3 (0.495 %) coincide or approach each other (Fig. 30). Average divergence for Articles 1–2 is 0.03453 %, while for Articles 2–3 – 0.06188 % and for Articles 1–3 – 0.03738 %. It looks like articles 1–2 were written by one author, similarly, articles 1–3 were written by one author, and articles 1–3 were written by different authors. But the trajectory of the curves coincides for Articles 1–2. Though the frequency of appearance of trigram with the initial letter ф is less than 1 % in each article (this allowed splitting the set of potential authors into two subsets), the frequency of appearance of each trigram beginning with letter ф is very small (8 trigrams out of 333 possible). In comparison with letter а – 156 trigrams out of 333 possible ones. The best results are shown by trigrams with a specific initial letter when their number in within (30.90). These numbers are approximate. To specify their values, when studying scientific and technical texts in the Ukrainian language, it is necessary to carry out additional research using a sufficiently large volume of texts (over 1000) among a large number of authors (over 100) and to have accurate reference sole author's texts (with confirmation of their authorship). This is next to impossible to do, due to the fact that most scientific and technical literature is co-authored. This imposes subjective characteristics on the analyzed text.

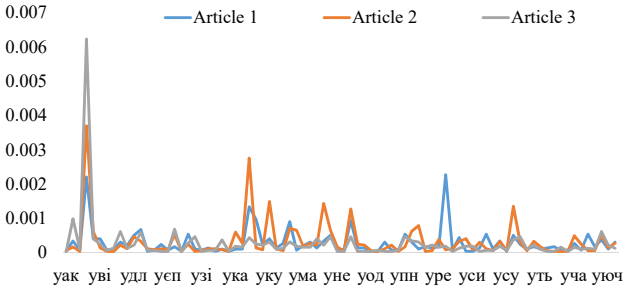


Fig. 29. The graph of using 3-grams beginning with letter у

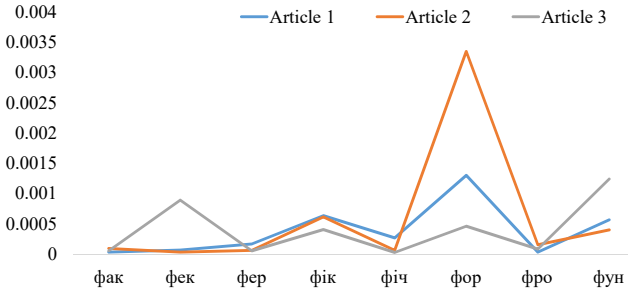


Fig. 30. The graph of using 3-grams beginning with letter ф

Check in practice. For 3-grams beginning with letter х (occurrence in articles 1–3 in the range of 0.5732–0.9339 %, the total of 37 trigrams), the curves for Articles 1–2 (0.5083 %), unlike Articles 1–3 (0.7957 %) and Articles 2–3 (0.7426 %), coincide or approach each other (Fig. 31). Average divergence for Articles 1–2 is 0.01374 %, while for articles 2–3 – 0.02007 % and for articles 1–3 – 0.02151 %. Articles 1–2 were written by one author, Article 3 – by another.

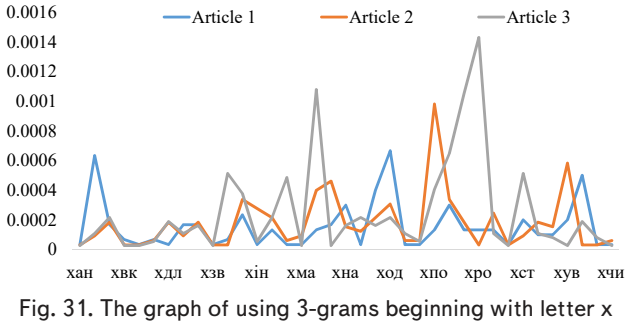


Fig. 31. The graph of using 3-grams beginning with letter х

For 3-grams beginning with letter ц (occurrence in Articles 1–3 in the range of 0.5906–0.829 %, in total

24 trigrams), all curves for Articles 1–2 (0.568 %), Articles 1–3 (0.4748 %) and Articles 2–3 (0.4416 %), coincide or approach each other (Fig. 32). Average divergence for Articles 1–2 is 0.02367 %, while for Articles 2–3 – 0.0184 % and for Articles 1–3 – 0.01978 %. It looks like all the articles were written by one author.

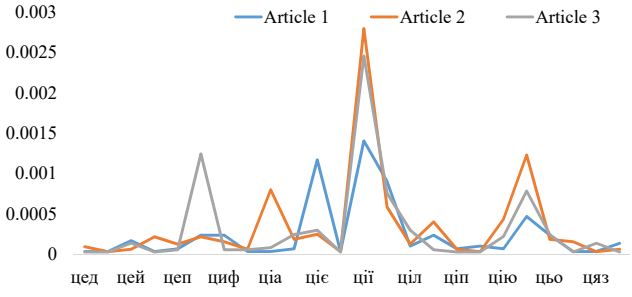


Fig. 32. The graph of using 3-grams beginning with letter ц

For 3-grams beginning with letter ч (occurrence in articles 1–3 in the range of 0.5128–1.3244 %), the curves for Articles 1–2 (1.0044 %), Articles 1–3 (0.6924 %) and Articles 2–3 (0.9368 %), coincide or approach each other (Fig 33). Average divergence for Articles 1–2 is 0.04367 %, while for Articles 2–3 – 0.04073 % and for Articles 1–3 – 0.0301 %. It looks like all the articles were written by one author.

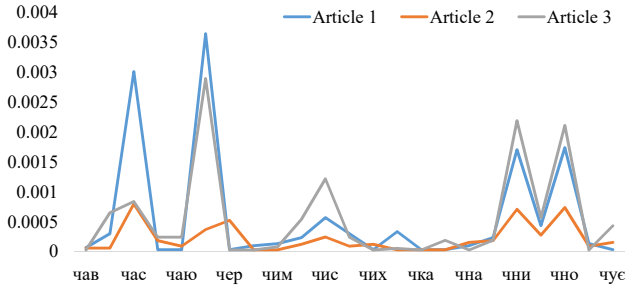


Fig. 33. The graph of using 3-grams beginning with letter ч

For 3-grams beginning with letter ь (occurrence in articles 1–3 in the range of 0,9981–1,2848 %, the total of 30 trigrams), the curves for Articles 1–2 (0.6593 %), Articles 1–3 (0.7326 %) and Articles 2–3 (0.7983 %), coincide or approach each other (Fig. 34). Average divergence for Articles 1–2 is 0.01691 %, while for Articles 2–3 – 0.02047 % and for Articles 1–3 – 0.01878 %. Articles 1–2 were written by one author, and Article 3 – by another. But the values themselves are boundary because the occurrence of the trigram beginning with letter ь in the range of (0.9;1).

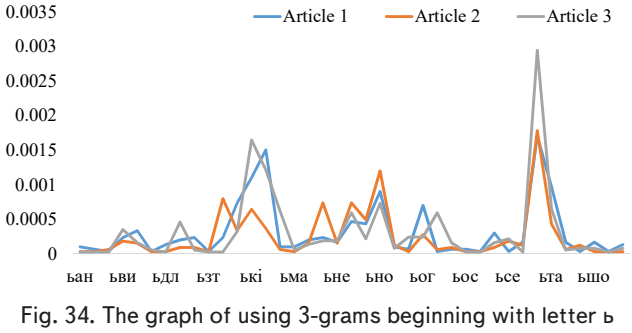


Fig. 34. The graph of using 3-grams beginning with letter ь

For 3-grams beginning with letter ш-щ (occurrence in articles 1–3 in the range of 0.357–0.8258 %, in total 22 trigrams), the curves for Articles 1–2 (0.2625 %), unlike Articles 1–3 (0.667 %) and Articles 2–3 (0.7209 %), coincide or approach each other (Fig. 35). Average divergence for Articles 1–2 is 0.01193 %, while for Articles 2–3 – 0.03277 % and Articles 1–3 – 0.03032 %. For sure, Articles 1–2 were written by one author and Article 3 – by another. It is explained by the fact that we compare two different set of trigrams with two different initial letters.

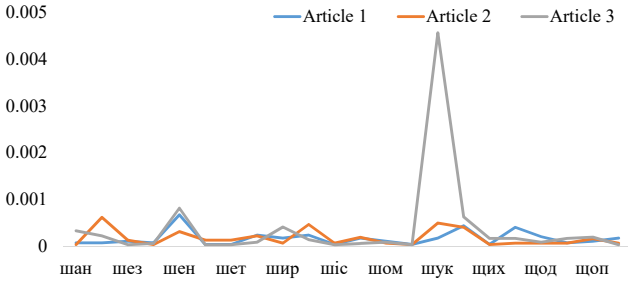


Fig. 35. The graph of using 3-grams beginning with letters ш and щ

For 3-grams beginning with letter ю (occurrence in articles 1–3 in the range of 0.2768–0.4939 %), all curves for Articles 1–2 (0.1558 %), unlike Articles 1–3 (0.2673 %) and Articles 2–3 (0.2005 %), coincide or approach each other (Fig. 36). Average divergence for Articles 1–2 is 0.0097375 %, while for Articles 2–3 – 0.01878 % and for Articles 1–3 – 0.01671 %. Definitely, Articles 1–2 were written by one author, and Article 3 – by another. This is explained by the fact that the frequency of occurrence of such trigrams is much less than 1 %, to be more exact, even less than 0.5 %.

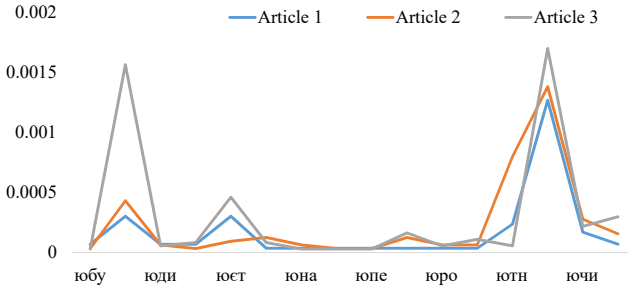


Fig. 36. The graph of using 3-grams beginning with letter ю

For 3-grams beginning with letter я (occurrence in articles 1–3 in the range of 1.4442–1.5541 %, in total 72 trigrams), all curves for Articles 1–2 (0.9522 %), Articles 1–3 (0.9361 %) and Articles 2–3 (1.0555 %) coincide or approach each other (Fig. 37). Average divergence for Articles 1–2 is 0.013225 %, while for Articles 2–3 – 0.01466 % and for Articles 1–3 – 0.013 %. It looks like all the articles were written by one author. This is explained by the fact that the frequency of occurrence of such trigrams is much more than 1 %.

Therefore, we compare the frequencies of appearance of all the trigrams that begin with a specific letter (Fig. 38, 39).

According to these graphs, Article 1 and Article 2 are most likely to have been written by one author, although Article 1 and Article may also have been written by one author (but this is not the truth). However, articles 2–3 defini/tely have been written by different authors. Application of linguo-statistical analysis of 3-grams to a set of articles will make it possible to form a subset of publications that are similar by linguistic characteristics. The imposition on

this subset of additional conditions in the form of statistical and quantitative analyses (sets of keywords, set expressions, stylometric, linguometric analyses, etc.) will significantly reduce this subset by specifying the list of the most probable author's papers. Thus, the analysis of the content and frequency of occurrence of only functional words will separate article 1 and 3 in different subsets, articles 1 and 2 will remain in the same subset.

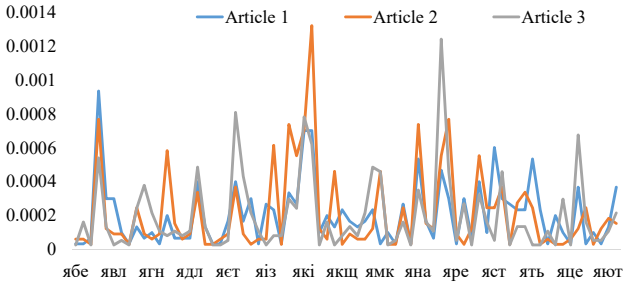


Fig. 37. The graph of using 3-grams beginning with letter я

Fig. 38. The graph of using 3-grams beginning with a specific letter

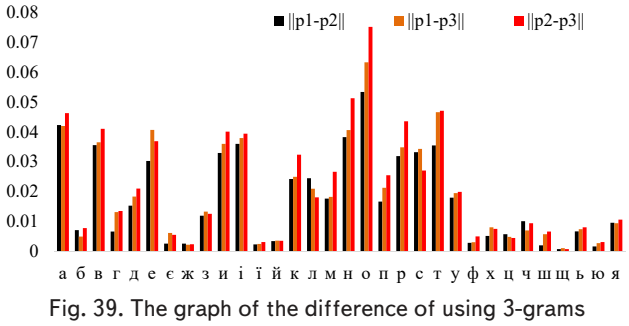


Fig. 39. The graph of the difference of using 3-grams beginning with a specific letter

This study does not imply solving the task of identifying the author to the full because the distinction of the author's traits is subjective in nature and depends on the limitations imposed on the author's creative process. However, as a result, the system, which implements such methods, can provide recommendations on the degree of belonging of a text to a specific author. The testing of the proposed method for identification of the author's style for other categories of texts – scientific humanities, fiction, journalism, etc. requires further experimental research

8. Conclusions

1. The quantitative method for identifying a potential author of a text from set of the possible ones was developed based on the comparison of the results of analysis of the reference text with the studied text. The algorithm of determining the stop words of the text content based on linguistic analysis of text content was developed. The algorithm of lexical analysis of texts in the Ukrainian language and the algorithm of the syntax analyzer of text content were developed. Its specific features are the adaptation of morphological and syntactic analysis of lexical units to the peculiarities of Ukrainian-language words/ texts. The theoretical and experimental substantiation of the method of content-monitoring and determining stop words of a text in the Ukrainian language was presented. The method is aimed at automatic detection of significant stop words of a text in the Ukrainian language due to the proposed formal approach to the implementation of content parsing.

2. The approach to the development of content monitoring software was proposed to identify the style of the author in text written in Ukrainian based on Web Mining. The problem of realization of identification of the author of a text in the Ukrainian language using reference characteristics of the author's speech based on the methods of NLP and stylometry was considered. This is important because the introduction of information technologies of stylometry for textual content authorship attribution leads to a higher coefficient of reliability of authorship identification for the studied text. However, there are objective difficulties, related to the accuracy of authorship attribution of a particular person, because sampling of individual scientific and technical publications is small (most articles in this field are written by co-authors). Only taking into consideration their personal characteristics through system learning can significantly reduce the range of potential authors of a particular technical text. As a part of the research described in this article, the quantitative method for automatic textual content authorship attribution based on statistical analysis of the *N*-gram distribution was developed. The system, based on the modern methods of NLP and stylometry with consideration of the metrics of evaluation of the analyzed text compared to the reference text, was developed. In addition, based on the modern methods of Machine Learning, the developed system learns to specify the results of text analysis for the authorship degree compared to the reference sample. This makes it possible to approach reasonably the determining of the quality of automatic identification of the author of a scientific technical text and obtain certain effects from its implementation in production. In particular, the coefficients of the author's speech can be clarified. In short, the algorithms of authorship attribution based on modern approaches of the NLP and stylometry taken together enable us to decrease the set of potential authors of the studied text. Further analysis of keywords, use of functional words and set expressions makes it possible to determine more accurately the degree of belonging of a paper to a particular author.

3. The results of experimental testing of the proposed content-monitoring method for determining the style of an author of scientific technical texts written in the Ukrainian language were explored. Typically, the author attribution systems use plagiarism detection algorithms on copyright and rewrite metrics. It is necessary only in order to determine whether a paper was not borrowed in whole or in part. However, they do not take into consideration the situation when a paper has not yet been published. The quantitative content-analysis of the textual content of scientific technical direction uses the benefits of content-monitoring and content-analysis of a text based on the methods of NLP, Web-Mining and stylometry for determining of a set of authors, whose writing styles are similar to the studied text

passage. This narrows the range of the search at further use of stylometry methods to determine the degree of belonging of the analyzed text to a particular author. We performed the decomposition of the authorship attribution method based on analysis of such speech coefficients as lexical diversity, degree (measure) of syntax complexity, speech coherence, exclusivity index and text concentration index. In parallel, such parameters of the author's style were analyzed, as the number of words in a certain text, total number of words in this text, the number of sentences, the number of prepositions, the number of conjunctions, the number of words with frequency 1, the number of words with frequency 10 and more. 3-grams from 3 articles were analyzed as an example. For Article 1, 78.4814 % of 3-grams were analyzed, for Article 2 – 72.6332 % and for Article 3 – 84.1271 %. Accordingly, the difference in using the corresponding 3-grams between Articles 1 and 2 is R_{12}=56.5254 %, between Articles 2 and 3 – R_{23}=69.4271 %, between Articles 1 and 3 – R_{13}=62.9839 %. These indicators show that the characteristics of Articles 1 and 2 are more similar ($R_{23}>R_{12}$ by 12.9017 %, $R_{23}>R_{13}$ by 6.4432 %, $R_{13}> R_{12}$ by 6.4585 %, that is, $R_{23}>R_{13}>R_{12}$), than the characteristics of Articles 1–3 and 2–3, respectively. The lower R_{ij}, the greater degree that the articles were written by the same author. In this case, Articles 1 and 2 are more likely to have been written by one author/team than Articles 2–3 and Articles 1–3, respectively. This paper contains the materials from the completed scientific research in the field of information technology, related to computer linguistics, artificial intelligence, and Machine Learning. The obtained results, presented in the article, give grounds to argue about the possibility of their implementation in actual industrial production.

References

1. Lytvyn, V., Vysotska, V., Pukach, P., Nytrebych, Z., Demkiv, I., Kovalchuk, R., Huzyk, N. (2018). Development of the linguometric method for automatic identification of the author of text content based on statistical analysis of language diversity coefficients. Eastern-European Journal of Enterprise Technologies, 5 (2 (95)), 16–28. doi: https://doi.org/10.15587/1729-4061.2018.142451
2. Lytvyn, V., Vysotska, V., Uhryn, D., Hrendus, M., Naum, O. (2018). Analysis of statistical methods for stable combinations determination of keywords identification. Eastern-European Journal of Enterprise Technologies, 2 (2 (92)), 23–37. doi: https://doi.org/10.15587/1729-4061.2018.126009
3. Buk, S. (2008). Osnovy statystychnoi linhvistyky. Lviv, 124.
4. Lytvyn, V., Vysotska, V., Pukach, P., Brodyak, O., Ugryn, D. (2017). Development of a method for determining the keywords in the slavic language texts based on the technology of web mining. Eastern-European Journal of Enterprise Technologies, 2 (2 (86)), 14–23. doi: https://doi.org/10.15587/1729-4061.2017.98750
5. Lytvyn, V., Vysotska, V., Pukach, P., Bobyk, I., Uhryn, D. (2017). Development of a method for the recognition of author's style in the Ukrainian language texts based on linguometry, stylemetry and glottochronology. Eastern-European Journal of Enterprise Technologies, 4 (2 (88)), 10–19. doi: https://doi.org/10.15587/1729-4061.2017.107512
6. Lytvyn, V., Pukach, P., Bobyk, I., Vysotska, V. (2016). The method of formation of the status of personality understanding based on the content analysis. Eastern-European Journal of Enterprise Technologies, 5 (2 (83)), 4–12. doi: https://doi.org/10.15587/1729-4061.2016.77174
7. Lytvyn, V., Vysotska, V., Pukach, P., Vovk, M., Ugryn, D. (2017). Method of functioning of intelligent agents, designed to solve action planning problems based on ontological approach. Eastern-European Journal of Enterprise Technologies, 3 (2 (87)), 11–17. doi: https://doi.org/10.15587/1729-4061.2017.103630
8. Khomytska, I., Teslyuk, V. (2016). The Method of Statistical Analysis of the Scientific, Colloquial, Belles-Lettres and Newspaper Styles on the Phonological Level. Advances in Intelligent Systems and Computing, 149–163. doi: https://doi.org/10.1007/978-3-319-45991-2_10
9. Khomytska, I., Teslyuk, V., Holovatyy, A., Morushko, O. (2018). Development of methods, models, and means for the author attribution of a text. Eastern-European Journal of Enterprise Technologies, 3 (2 (93)), 41–46. doi: https://doi.org/10.15587/1729-4061.2018.132052
10. Khomytska, I., Teslyuk, V. (2018). Authorship and Style Attribution by Statistical Methods of Style Differentiation on the Phonological Level. Advances in Intelligent Systems and Computing III, 105–118. doi: https://doi.org/10.1007/978-3-030-01069-0_8
11. Khomytska, I., Teslyuk, V. (2016). Specifics of phonostatistical structure of the scientific style in English style system. 2016 XIth International Scientific and Technical Conference Computer Sciences and Information Technologies (CSIT). doi: https://doi.org/10.1109/stc-csit.2016.7589887
12. Khomytska, I., Teslyuk, V. (2017). Modelling of phonostatistical structures of English backlingual phoneme group in style system. 2017 14th International Conference The Experience of Designing and Application of CAD Systems in Microelectronics (CADSM). doi: https://doi.org/10.1109/cadsm.2017.7916144
13. Khomytska, I., Teslyuk, V. (2017). Modelling of phonostatistical structures of the colloquial and newspaper styles in english sonorant phoneme group. 2017 12th International Scientific and Technical Conference on Computer Sciences and Information Technologies (CSIT). doi: https://doi.org/10.1109/stc-csit.2017.8098738
14. Khomytska, I., Teslyuk, V. (2018). Authorship Attribution by Differentiation of Phonostatistical Structures of Styles. 2018 IEEE 13th International Scientific and Technical Conference on Computer Sciences and Information Technologies (CSIT). doi: https://doi.org/10.1109/stc-csit.2018.8526739
15. Khomytska, I., Teslyuk, V. (2019). The Software for Authorship and Style Attribution. 2019 IEEE 15th International Conference on the Experience of Designing and Application of CAD Systems (CADSM). doi: https://doi.org/10.1109/cadsm.2019.8779346

16. Khomytska, I., Teslyuk, V. (2019). Mathematical Methods Applied for Authorship Attribution on the Phonological Level. CSIT: Proceedings of the XIVth Scientific and Technical Conference, 7–11.
17. Bol'shakova, E., Klyshinskiy, E., Lande, D., Noskov, A., Peskova, O., Yagunova, E. (2011). Avtomaticheskaya obrabotka tekstov na estestvennom yazyke i komp'yuternaya lingvistika. Moscow: MIEM, 272.
18. Anisimov, A., Marchenko, A. (2002). Sistema obrabotki tekstov na estestvennom yazyke. Iskusstvennyy intellekt, 4, 157–163.
19. Perebyinis, V. (2000). Matematychna linhvistyka. Ukrainska mova. Kyiv, 287–302.
20. Perebyinis, V. (2013). Statystychni metody dlia linhvistiv. Vinnytsia, 176.
21. Braslavskiy, P. I. (2006). Intellektual'nye informatsionnye sistemy. Tema 7 Tematicheskaya kategorizatsiya. Available at: https://docplayer.ru/36866470-Intellektualnye-informacionnye-sistemy-tema-7-tematicheskaya-kategorizaciya-pavel-isaakovich-braslavskiy-vesenniy-semestr-2006.html
22. Lande, D., Zhyhalo, V. (2008). Pidkhid do rishennia problem poshuku dvomovnoho plahiatu. Problemy informatyzatsiyi ta upravlinnia, 2 (24), 125–129.
23. Varfolomeev, A. (2000). Psihosemantika slova i lingvostatistika teksta. Kaliningrad, 37.
24. Marchenko, O. (2006). Modeliuvannia semantychnoho kontekstu pry analizi tekstiv na pryrodniy movi. Visnyk Kyivskoho universytetu, 3, 230–235.
25. Jivani, A. G. (2011). A Comparative Study of Stemming Algorithms. Int. J. Comp. Tech. Appl., 2 (6), 1930–1938.
26. Linhvometriya. Victana. Available at: http://victana.lviv.ua/nlp/linhvometriia
27. Sushko, S. O., Fomychova, L. Ya., Barsukov, Ye. S. (2010). Chastoty povtoriuvanosti bukv i bihram u vidkrytykh tekstakh ukrainskoiu movoiu. Ukrainian Information Security Research Journal, 12 (3 (48)). doi: https://doi.org/10.18372/2410-7840.12.1968
28. Kognitivnaya stilometriya: k postanovke problemy. Available at: http://www.manekin.narod.ru/hist/styl.htm
29. Kocherhan, M. (2005). Vstup do movoznavstva. Kyiv, 368.
30. Rodionova, E. (2008). Metody atributsii hudozhestvennyh tekstov. Strukturnaya i prikladnaya lingvistika, 7, 118–127.
31. Meshcheryakov, R. V., Vasyukov, N. S. (2005). Modeli opredeleniya avtorstva teksta. Izmereniya, avtomatizatsiya i modelirovanie v promyshlennosti i nauchnyh issledovaniyah, 25–29.
32. Morozov, N. A. Lingvisticheskie spektry. Available at: http://www.textology.ru/library/book.aspx?bookId=1&textId=3
33. Mobasher, B. (2007). Data Mining for Web Personalization. The Adaptive Web. Lecture Notes in Computer Science, 90–135. doi: https://doi.org/10.1007/978-3-540-72079-9_3
34. Dinucă, C. E., Ciobanu, D. (2012). Web Content Mining. Annals of the University of Petroşani. Economics, 12 (1), 85–92.
35. Xu, G., Zhang, Y., Li, L. (2010). Web Content Mining. Web Mining and Social Networking, 71–87. doi: https://doi.org/10.1007/978-1-4419-7735-9_4
36. Mishler, A., Crabb, E. S., Paletz, S., Hefright, B., Golonka, E. (2015). Using Structural Topic Modeling to Detect Events and Cluster Twitter Users in the Ukrainian Crisis. HCI International 2015 - Posters' Extended Abstracts, 639–644. doi: https://doi.org/10.1007/978-3-319-21380-4_108
37. Bubleinyk, L. (2000). Osoblyvosti khudozhnoho movlennia. Lutsk, 179.
38. Kowalska, K., Cai, D., Wade, S. (2012). Sentiment Analysis of Polish Texts. International Journal of Computer and Communication Engineering, 1 (1), 39–42. doi: https://doi.org/10.7763/ijcce.2012.v1.12
39. Kotsyba, N. (2009). The current state of work on the Polish-Ukrainian Parallel Corpus (PolUKR). Organization and Development of Digital Lexical Resources, 55–60.
40. Lytvyn, V., Vysotska, V., Rzheuskyi, A. (2019). Technology for the Psychological Portraits Formation of Social Networks Users for the IT Specialists Recruitment Based on Big Five, NLP and Big Data Analysis. Proceedings of the 1st International Workshop on Control, Optimisation and Analytical Processing of Social Networks (COAPSN-2019), 2392, 147–171.
41. Lytvyn, V., Vysotska, V., Rusyn, B., Pohreliuk, L., Berezin, P., Naum, O. (2019). Textual Content Categorizing Technology Development Based on Ontology. Workshop Proceedings of the 8th International Conference on "Mathematics. Information Technologies. Education", 2386, 234–254.
42. Lytvyn, V., Kuchkovskiy, V., Vysotska, V., Markiv, O., Pabyrivskyy, V. (2018). Architecture of System for Content Integration and Formation Based on Cryptographic Consumer Needs. 2018 IEEE 13th International Scientific and Technical Conference on Computer Sciences and Information Technologies (CSIT). doi: https://doi.org/10.1109/stc-csit.2018.8526669
43. Berko, A., Alieksieiev, V. (2018). A Method to Solve Uncertainty Problem for Big Data Sources. 2018 IEEE Second International Conference on Data Stream Mining & Processing (DSMP). doi: https://doi.org/10.1109/dsmp.2018.8478460
44. Gozhyj, A., Kalinina, I., Vysotska, V., Gozhyj, V. (2018). The Method of Web-Resources Management Under Conditions of Uncertainty Based on Fuzzy Logic. 2018 IEEE 13th International Scientific and Technical Conference on Computer Sciences and Information Technologies (CSIT). doi: https://doi.org/10.1109/stc-csit.2018.8526761
45. Lytvyn, V., Vysotska, V., Dosyn, D., Burov, Y. (2018). Method for ontology content and structure optimization, provided by a weighted conceptual graph. Webology, 15 (2), 66–85.
46. Nytrebych, Z. M., Malanchuk, O. M., Il'kiv, V. S., Pukach, P. Ya. (2017). Homogeneous problem with two-point conditions in time for some equations of mathematical physics. Azerbaijan Journal of Mathematics, 7 (2), 180–196.
47. Nytrebych, Z., Il'kiv, V., Pukach, P., Malanchuk, O. (2018). On nontrivial solutions of homogeneous Dirichlet problem for partial differential equations in a layer. Kragujevac Journal of Mathematics, 42 (2), 193–207. doi: https://doi.org/10.5937/kgjmath1802193n

48. Nytrebych, Z., Malanchuk, O., Il'kiv, V., Pukach, P. (2017). On the solvability of two-point in time problem for PDE. Italian Journal of Pure and Applied Mathematics, 38, 715–726.
49. Pukach, P. Ya., Kuzio, I. V., Nytrebych, Z. M., Ilkiv, V. S. (2017). Analytical methods for determining the effect of the dynamic process on the nonlinear flexural vibrations and the strength of compressed shaft. Naukovyi Visnyk Natsionalnoho Hirnychoho Universytetu, 5, 69–76.
50. Pukach, P. Y., Kuzio, I. V., Nytrebych, Z. M., Il'kiv, V. S. (2018). Asymptotic method for investigating resonant regimes of nonlinear bending vibrations of elastic shaft. Scientific Bulletin of National Mining University, 1, 68–73. doi: https://doi.org/10.29202/nvngu/2018-1/9
51. Nytrebych, Z., Ilkiv, V., Pukach, P., Malanchuk, O., Kohut, I., Senyk, A. (2019). Analytical method to study a mathematical model of wave processes under twopoint time conditions. Eastern-European Journal of Enterprise Technologies, 1 (7 (97)), 74–83. doi: https://doi.org/10.15587/1729-4061.2019.155148
52. Pukach, P., Il'kiv, V., Nytrebych, Z., Vovk, M., Pukach, P. (2017). On the Asymptotic Methods of the Mathematical Models of Strongly Nonlinear Physical Systems. Advances in Intelligent Systems and Computing, 421–433. doi: https://doi.org/10.1007/978-3-319-70581-1_30
53. Lavrenyuk, S. P., Pukach, P. Y. (2007). Mixed problem for a nonlinear hyperbolic equation in a domain unbounded with respect to space variables. Ukrainian Mathematical Journal, 59 (11), 1708–1718. doi: https://doi.org/10.1007/s11253-008-0020-0
54. Pukach, P. Y. (2016). Investigation of Bending Vibrations in Voigt–Kelvin Bars with Regard for Nonlinear Resistance Forces. Journal of Mathematical Sciences, 215 (1), 71–78. doi: https://doi.org/10.1007/s10958-016-2823-0
55. Pukach, P., Il'kiv, V., Nytrebych, Z., Vovk, M. (2017). On nonexistence of global in time solution for a mixed problem for a nonlinear evolution equation with memory generalizing the Voigt-Kelvin rheological model. Opuscula Mathematica, 37 (45), 735. doi: https://doi.org/10.7494/opmath.2017.37.5.735
56. Pukach, P. Y. (2012). On the unboundedness of a solution of the mixed problem for a nonlinear evolution equation at a finite time. Nonlinear Oscillations, 14 (3), 369–378. doi: https://doi.org/10.1007/s11072-012-0164-6
57. Pukach, P. Y. (2014). Qualitative Methods for the Investigation of a Mathematical Model of Nonlinear Vibrations of a Conveyer Belt. Journal of Mathematical Sciences, 198 (1), 31–38. doi: https://doi.org/10.1007/s10958-014-1770-x
58. Bezobrazov, S., Sachenko, A., Komar, M., Rubanau, V. (2016). The Methods of Artificial Intelligence for Malicious Applications Detection in Android OS. International Journal of Computing, 15 (3), 184–190.
59. Dunets, O., Wolff, C., Sachenko, A., Hladiy, G., Dobrotvor, I. (2017). Multi-agent system of IT project planning. 2017 9th IEEE International Conference on Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications (IDAACS). doi: https://doi.org/10.1109/idaacs.2017.8095141
60. Vysotska, V., Lytvyn, V., Burov, Y., Berezin, P., Emmerich, M., Basto Fernandes, V. (2019). Development of Information System for Textual Content Categorizing Based on Ontology. CEUR Workshop Proceedings, 53–70.
61. Vysotska, V., Lytvyn, V., Burov, Y., Gozhyj, A., Makara, S. (2018). The consolidated information web-resource about pharmacy networks in city. Proceedings of the 1st International Workshop on Informatics & Data-Driven Medicine (IDDM 2018), 2255, 239–255. Available at: http://ceur-ws.org/Vol-2255/paper22.pdf
62. Rusyn, B., Vysotska, V., Pohreliuk, L. (2018). Model and Architecture for Virtual Library Information System. 2018 IEEE 13th International Scientific and Technical Conference on Computer Sciences and Information Technologies (CSIT). doi: https://doi.org/10.1109/stc-csit.2018.8526679
63. Lytvyn, V., Vysotska, V., Dosyn, D., Lozynska, O., Oborska, O. (2018). Methods of Building Intelligent Decision Support Systems Based on Adaptive Ontology. 2018 IEEE Second International Conference on Data Stream Mining & Processing (DSMP). doi: https://doi.org/10.1109/dsmp.2018.8478500
64. Lytvyn, V., Vysotska, V., Burov, Y., Bobyk, I., Ohirko, O. (2018). The Linguometric Approach for Co-authoring Author's Style Definition. 2018 IEEE 4th International Symposium on Wireless Systems within the International Conferences on Intelligent Data Acquisition and Advanced Computing Systems (IDAACS-SWS). doi: https://doi.org/10.1109/idaacs-sws.2018.8525741
65. Zdebskyi, P., Vysotska, V., Peleshchak, R., Peleshchak, I., Demchuk, A., Krylyshyn, M. (2019). An Application Development for Recognizing of View in Order to Control the Mouse Pointer. Workshop Proceedings of the 8th International Conference on "Mathematics. Information Technologies. Education", 55–74.
66. Veres, O., Rusyn, B., Sachenko, A., Rishnyak, I. (2018). Choosing the method of finding similar images in the reverse search system. Proceedings of the 2nd International Conference on Computational Linguistics and Intelligent Systems. Volume I: Main Conference (COLINS 2018), 2136, 99–107.
67. Vysotska, V., Lytvyn, V., Hrendus, M., Kubinska, S., Brodyak, O. (2018). Method of Textual Information Authorship Analysis Based on Stylometry. 2018 IEEE 13th International Scientific and Technical Conference on Computer Sciences and Information Technologies (CSIT). doi: https://doi.org/10.1109/stc-csit.2018.8526608
68. Gozhyj, A., Chyrun, L., Kowalska-Styczen, A., Lozynska, O. (2018). Uniform Method of Operative Content Management in Web Systems. Proceedings of the 2nd International Conference on Computational Linguistics and Intelligent Systems. Volume I: Main Conference (COLINS 2018), 2136, 62–77. Available at: http://ceur-ws.org/Vol-2136/10000062.pdf
69. Vysotska, V., Burov, Y., Lytvyn, V., Demchuk, A. (2018). Defining Author's Style for Plagiarism Detection in Academic Environment. 2018 IEEE Second International Conference on Data Stream Mining & Processing (DSMP), 128–133. doi: https://doi.org/10.1109/dsmp.2018.8478574

70. Chyrun, L., Vysotska, V., Kis, I., Chyrun, L. (2018). Content Analysis Method for Cut Formation of Human Psychological State. 2018 IEEE Second International Conference on Data Stream Mining & Processing (DSMP). doi: https://doi.org/10.1109/dsmp.2018.8478619
71. Gozhyj, A., Vysotska, V., Yevseyeva, I., Kalinina, I., Gozhyj, V. (2018). Web Resources Management Method Based on Intelligent Technologies. Advances in Intelligent Systems and Computing III, 206–221. doi: https://doi.org/10.1007/978-3-030-01069-0_15
72. Chyrun, L., Kis, I., Vysotska, V., Chyrun, L. (2018). Content Monitoring Method for Cut Formation of Person Psychological State in Social Scoring. 2018 IEEE 13th International Scientific and Technical Conference on Computer Sciences and Information Technologies (CSIT). doi: https://doi.org/10.1109/stc-csit.2018.8526624
73. Demchuk, A., Lytvyn, V., Vysotska, V., Dilai, M. (2019). Methods and Means of Web Content Personalization for Commercial Information Products Distribution. Lecture Notes in Computational Intelligence and Decision Making, 332–347. doi: https://doi.org/10.1007/978-3-030-26474-1_24
74. Lytvyn, V., Vysotska, V., Kuchkovskiy, V., Bobyk, I., Malanchuk, O., Ryshkovets, Y. et. al. (2019). Development of the system to integrate and generate content considering the cryptocurrent needs of users. Eastern-European Journal of Enterprise Technologies, 1 (2 (97)), 18–39. doi: https://doi.org/10.15587/1729-4061.2019.154709
75. Vysotska, V., Fernandes, V. B., Lytvyn, V., Emmerich, M., Hrendus, M. (2018). Method for Determining Linguometric Coefficient Dynamics of Ukrainian Text Content Authorship. Advances in Intelligent Systems and Computing III, 132–151. doi: https://doi.org/10.1007/978-3-030-01069-0_10
76. Kravets, P. (2010). The control agent with fuzzy logic. Perspective Technologies and Methods in MEMS Design, 40–41.
77. Kravets, P. (2007). The Game Method for Orthonormal Systems Construction. 2007 9th International Conference - The Experience of Designing and Applications of CAD Systems in Microelectronics. doi: https://doi.org/10.1109/cadsm.2007.4297555
78. Kravets, P. (2016). Game Model of Dragonfly Animat Self-Learning. Perspective Technologies and Methods in MEMS Design, 195–201.
79. Basyuk, T. (2015). The main reasons of attendance falling of internet resource. 2015 Xth International Scientific and Technical Conference "Computer Sciences and Information Technologies" (CSIT). doi: https://doi.org/10.1109/stc-csit.2015.7325440
80. Chyrun, L., Kowalska-Styczen, A., Burov, Y., Berko, A., Vasevych, A., Pelekh, I., Ryshkovets, Y. (2019). Heterogeneous Data with Agreed Content Aggregation System Development. Workshop Proceedings of the 8th International Conference on "Mathematics. Information Technologies. Education", 2386, 35–54.
81. Chyrun, L., Burov, Y., Rusyn, B., Pohreliuk, L., Oleshek, O. et. al. (2019). Web Resource Changes Monitoring System Development. Workshop Proceedings of the 8th International Conference on "Mathematics. Information Technologies. Education", 2386, 255–273.
82. Vysotska, V., Burov, Y., Lytvyn, V., Oleshek, O. (2019). Automated Monitoring of Changes in Web Resources. Lecture Notes in Computational Intelligence and Decision Making, 348–363. doi: https://doi.org/10.1007/978-3-030-26474-1_25
83. Chyrun, L., Gozhyj, A., Yevseyeva, I., Dosyn, D., Tyhonov, V., Zakharchuk, M. (2019). Web Content Monitoring System Development. Proceedings of the 3rd International Conference on Computational Linguistics and Intelligent Systems (COLINS-2019). Volume I: Main Conference, 2362, 126–142.
84. Rzheuskyi, A., Gozhyj, A., Stefanchuk, A., Oborska, O., Chyrun, L., Lozynska, O. et. al. (2019). Development of Mobile Application for Choreographic Productions Creation and Visualization. Workshop Proceedings of the 8th International Conference on "Mathematics. Information Technologies. Education", 2386, 340–358.
85. Lytvynenko, V., Savina, N., Krejci, J., Voronenko, M., Yakobchuk, M., Kryvoruchko, O. (2019). Bayesian Networks' Development Based on Noisy-MAX Nodes for Modeling Investment Processes in Transport. Workshop Proceedings of the 8th International Conference on "Mathematics. Information Technologies. Education", 2386, 1–10.
86. Lytvynenko, V., Lurie, I., Krejci, J., Voronenko, M., Savina, N., Taif, M. A. (2019). Two Step Density-Based Object-Inductive Clustering Algorithm. Workshop Proceedings of the 8th International Conference on "Mathematics. Information Technologies. Education", 2386, 117–135.
87. Antonyuk, N., Medykovskyy, M., Chyrun, L., Dverii, M., Oborska, O., Krylyshyn, M. et. al. (2019). Online Tourism System Development for Searching and Planning Trips with User's Requirements. Advances in Intelligent Systems and Computing, 831–863. doi: https://doi.org/10.1007/978-3-030-33695-0_55
88. Rzheuskyi, A., Kutyuk, O., Voloshyn, O., Kowalska-Styczen, A., Voloshyn, V., Chyrun, L. et. al. (2019). The Intellectual System Development of Distant Competencies Analyzing for IT Recruitment. Advances in Intelligent Systems and Computing, 696–720. doi: https://doi.org/10.1007/978-3-030-33695-0_47
89. Rusyn, B., Pohreliuk, L., Rzheuskyi, A., Kubik, R., Ryshkovets, Y., Chyrun, L. et. al. (2019). The Mobile Application Development Based on Online Music Library for Socializing in the World of Bard Songs and Scouts' Bonfires. Advances in Intelligent Systems and Computing, 734–756. doi: https://doi.org/10.1007/978-3-030-33695-0_49
90. Chyrun, L., Leshchynskyy, E., Lytvyn, V., Rzheuskyi, A.,Vysotska, V., Borzov, Y. (2019). Intellectual Analysis of Making Decisions Tree in Information Systems of Screening Observation for Immunological Patients. CEUR Workshop Proceedings, of the 2nd International Workshop on Informatics & Data-Driven Medicine (IDDM 2019), 281–296. Available at: http://ceur-ws.org/Vol-2488/paper25.pdf
91. Lytvyn, V., Burov, Y., Kravets, P., Vysotska, V., Demchuk, A., Berko, A. et. al. (2019). Methods and Models of Intellectual Processing of Texts for Building Ontologies of Software for Medical Terms Identification in Content Classification. CEUR Workshop

Proceedings, of the 2nd International Workshop on Informatics & Data-Driven Medicine (IDDM 2019), 354–368. Available at: http://ceur-ws.org/Vol-2488/paper31.pdf

92. Antonyuk, N., Chyrun, L., Andrunyk, V., Vasevych, A., Chyrun, S., Gozhyj, A. et. al. (2019). Medical News Aggregation and Ranking of Taking into Account the User Needs. CEUR Workshop Proceedings, of the 2nd International Workshop on Informatics&Data-Driven Medicine (IDDM 2019), 369–382. Available at: http://ceur-ws.org/Vol-2488/paper32.pdf
93. Babichev, S., Taif, M. A., Lytvynenko, V., Osypenko, V. (2017). Criterial analysis of gene expression sequences to create the objective clustering inductive technology. 2017 IEEE 37th International Conference on Electronics and Nanotechnology (ELNANO). doi: https://doi.org/10.1109/elnano.2017.7939756
94. Babichev, S., Korobchynskyi, M., Lahodynskyi, O., Korchomnyi, O., Basanets, V., Borynskyi, V. (2018). Development of a technique for the reconstruction and validation of gene network models based on gene expression profiles. Eastern-European Journal of Enterprise Technologies, 1 (4 (91)), 19–32. doi: https://doi.org/10.15587/1729-4061.2018.123634
95. Babichev, S., Lytvynenko, V., Osypenko, V. (2017). Implementation of the objective clustering inductive technology based on DBSCAN clustering algorithm. 2017 12th International Scientific and Technical Conference on Computer Sciences and Information Technologies (CSIT). doi: https://doi.org/10.1109/stc-csit.2017.8098832
96. Babichev, S. A., Gozhyj, A., Kornelyuk, A. I., Lytvynenko, V. I. (2017). Objective clustering inductive technology of gene expression profiles based on sota clustering algorithm. Biopolymers and Cell, 33 (5), 379–392. doi: https://doi.org/10.7124/bc.000961
97. Pasichnyk, V., Shestakevych, T. (2016). The Model of Data Analysis of the Psychophysiological Survey Results. Advances in Intelligent Systems and Computing, 271–281. doi: https://doi.org/10.1007/978-3-319-45991-2_18
98. Zhezhnych, P., Markiv, O. (2017). Linguistic Comparison Quality Evaluation of Web-Site Content with Tourism Documentation Objects. Advances in Intelligent Systems and Computing, 656–667. doi: https://doi.org/10.1007/978-3-319-70581-1_45
99. Davydov, M., Lozynska, O. (2017). Information system for translation into ukrainian sign language on mobile devices. 2017 12th International Scientific and Technical Conference on Computer Sciences and Information Technologies (CSIT). doi: https://doi.org/10.1109/stc-csit.2017.8098734
100. Davydov, M., Lozynska, O. (2017). Mathematical Method of Translation into Ukrainian Sign Language Based on Ontologies. Advances in Intelligent Systems and Computing, 89–100. doi: https://doi.org/10.1007/978-3-319-70581-1_7
101. Davydov, M., Lozynska, O. (2016). Linguistic models of assistive computer technologies for cognition and communication. 2016 XIth International Scientific and Technical Conference Computer Sciences and Information Technologies (CSIT). doi: https://doi.org/10.1109/stc-csit.2016.7589898
102. Vysotska, V., Chyrun, L. (2015). Methods of information resources processing in electronic content commerce systems. Proceedings of 13th International Conference: The Experience of Designing and Application of CAD Systems in Microelectronics, CADSM 2015-February.
103. Andrunyk, V., Chyrun, L., Vysotska, V. (2015). Electronic content commerce system development. Proceedings of 13th International Conference: The Experience of Designing and Application of CAD Systems in Microelectronics, CADSM 2015-February.
104. Alieksieieva, K., Berko, A., Vysotska, V. (2015). Technology of commercial web-resource processing. Proceedings of 13th International Conference: The Experience of Designing and Application of CAD Systems in Microelectronics, CADSM 2015-February.
105. Mykich, K., Burov, Y. (2016). Uncertainty in situational awareness systems. 2016 13th International Conference on Modern Problems of Radio Engineering, Telecommunications and Computer Science (TCSET). doi: https://doi.org/10.1109/tcset.2016.7452165
106. Mykich, K., Burov, Y. (2016). Algebraic Framework for Knowledge Processing in Systems with Situational Awareness. Advances in Intelligent Systems and Computing, 217–227. doi: https://doi.org/10.1007/978-3-319-45991-2_14
107. Mykich, K., Burov, Y. (2016). Research of uncertainties in situational awareness systems and methods of their processing. Eastern-European Journal of Enterprise Technologies, 1 (4 (79)), 19–27. doi: https://doi.org/10.15587/1729-4061.2016.60828
108. Mykich, K., Burov, Y. (2016). Algebraic model for knowledge representation in situational awareness systems. 2016 XIth International Scientific and Technical Conference Computer Sciences and Information Technologies (CSIT). doi: https://doi.org/10.1109/stc-csit.2016.7589896