UDC 572. 511:61+004.94
DOI: 10.15587/1729-4061.2018.141451

DEVELOPMENT OF THE LINGUOMETRIC METHOD FOR AUTOMATIC IDENTIFICATION OF THE AUTHOR OF TEXT CONTENT BASED ON STATISTICAL ANALYSIS OF LANGUAGE DIVERSITY COEFFICIENTS

V. Lytvyn
Doctor of Technical Sciences, Professor
Department of Information Systems and Networks**
E-mail: yevhen.v.burov@lpnu.ua
V. Vysotska
PhD, Associate Professor
Department of Information Systems and Networks**
E-mail: victoria.a.vysotska@lpnu.ua
P. Pukach
Doctor of Technical Sciences, Professor *
E-mail: ppukach@gmail.com
Z. Nytrebych
Doctor of Physical and Mathematical Sciences, Professor
Department of Mathematics**
E-mail: znytrebych@gmail.com
I. Demkiv
Doctor of Physical and Mathematical Sciences,
Associate Professor
Department of
Computational Mathematics and Programming**
E-mail: ihor.i.demkiv@lpnu.ua
R. Kovalchuk
PhD, Associate Professor*
E-mail: roma_kov@meta.ua
N. Huzyk
PhD*
E-mail: hryntsiv@ukr.net
* Department of Engineering Mechanics
(Weapons and Equipment of Military Engineering Forces)
Hetman Petro Sahaidachnyi National Army Academy
Heroiv Maidanu str., 32, Lviv, Ukraine, 79026
**Lviv Polytechnic National University
S. Bandery str., 12, Lviv, Ukraine, 79013

Розробленно лінгвометричний метод алгоритмічного забезпечення процесів контент-моніторінгу для розв'язання задачі автоматичного визначення автора україномовного текстового контенту на основі технології статистичного аналізу коефіцієнтів мовної різноманітності. Проведено декомпозицію методу визначення автора на основі аналізу таких коефіцієнтів мовлення як лексична різноманітність, ступінь (міра) синтаксичної складності, зв'язність мовлення, індекси винятковості та концентрації тексту. Проаналізовані також параметри авторського стилю як кількість слів у певному тексті, загальна кількість слів цього тексту, кількість речень, кількість прийменників, кількість сполучників, кількість слів із частотою 1, та кількість слів із частотою 10 та більше.

Особливостями розробленого є адаптація морфологічного та синтаксичного аналізу лексичних одиниць до особливостей конструкцій україномовних слів/текстів. Тобто при аналізі лінгвістичних одиниць типу слів враховувалась належність до частини мови та відмінювання в межах цієї частини мови. Для цього проводився аналіз флексій цих слів для класифікації, виділення основи для формування відповідних алфавітно-частотних словників. Наповнення цих словників в подальшому враховувалися на наступних кроках визначення авторства тексту як розрахунок параметрів та коефіцієнтів авторського мовлення. Для індивідуального стилю письменника показовими є саме службові (стопові або опорні) слова, оскільки вони ніяк не пов'язані з темою і змістом публікації.

Проведено порівняння результатів на множині 200 одноосібних робіт технічного спрямування біля 100 різних авторів за період 2001–2017 рр. для визначення, чи змінюються і як коефіцієнти різноманітності тексту цих авторів в різні проміжки часу. Виявлено, що для обраної експериментальної бази з понад 200 робіт найкращих результатів за критерієм щільності досягає метод аналізу статті без початкової обов'язкової інформації як анотації та ключові слова різними мовами, а також списку літератури

Ключові слова: NLP, контент-моніторінг, стоп-слова, контент-аналіз, статистийний лінгвістичний аналіз, квантитативна лінгві стика

1. Introduction

Important tasks of linguistics-based linguometry is creation and comparison of dictionaries (including frequency and statistics dictionaries), automatic dictionaries, thesauruses, shorthand systems, automatic language identification, information search, etc. [1]. For example, statistical and transition probabilities of morphemes of a text are found in order to model information search processes [2]. Based on the constructed tables, proof-reading

of a studied word is modeled and some most probable variants are offered [3]. In turn, stylemetry as a subdivision of applied linguistics reveals and analyzes the quantitative characteristics of a certain functional style of the language or speech of the authors of text content, that is, the author's attributions [4]. Attribution implies determining with the use of methods of quantitative linguistics the validity, authenticity of the author's work, its author, the place and time of its creation based of the analysis of the technological and stylistic patterns and characteristics of coefficients of the language diversity of a particular author and/or of a particular text of the work [5]. For example, one of the known linguistic problems is the process of determining the author's attribution of passages of a particular text content [6]. To do this, the frequencies of word usage of the proposed passages are calculated [7]. Using frequency dictionaries of the author's creative work in general or of his separate works, the author of a work (or a work – if a dictionary makes it possible) are identified [8]. The disadvantage is saving or auto-generation of large data arrays in the form of frequency dictionaries of author's works [9]. Processing of such dictionaries requires a lot of time, while saving them demands a lot of resources [10]. In their turn, there are the authors who have not created a large number of works, which makes it impossible to reproduce exactly the results of the analysis of the author's attribution [11]. A well-known method of dating in order to determine the duration of the separate existence of two closely related languages is based on the assumption that the bulk of the lexical structure of any language (nuclear vocabulary) changes at the same rate and requires counting a percentage of common elements in the basic vocabulary [12]. The modified methods of glottochronology are used to determine the dynamics of a change of the author's speech in his text content for a long time in order to date the approximate period, within which a particular text was created by this author [13]. That is why the problem of automatic identification of the author of the text content is relevant and requires new (improved) approaches to its solution, for example, based on statistical analysis of language diversity coefficients.

2. Literature review and problem statement

The separation (distribution) of a linguistic unit in a text – the presence of a linguistic unit in various (usually equal) subsamples (passages) – is of great importance in quantitative linguistics [15]. If a studied linguistic unit operates only in one subsample, although with a high frequency, such sample is non-representative in respect to this linguistic unit [16]. It is important, when a studied linguistic unit is evenly distributed in the general totality [17]. To do this, the distribution factor is analyzed [18]: $K_r=N_p/N_z$, where N_p is the ratio of the number of subsamples with a certain linguistic unit, N_z is the general number of subsamples. However, the characteristics, obtained from the material of a sample, usually differ from the real characteristics of the general totality, as there is a relative inaccuracy of research in quantitative linguistics [19]. Distribution of frequency of linguistic units in text content has a certain regularity and forms its statistical (frequency, probabilistic) structure [20]. This distribution is different for each of the language elements – lexemes, morphemes, phonemes, etc. [21]. That is why the linguo-statistic parameters of the authors' styles, established at different levels (phonemic, morphemic, N-gram, lexemic, etc.), have a different style identifiable power of the authors' speech for different pairs of styles [22]. For example, related styles are more clearly distinguished at the semantic level, while less related – at the lexical level [23]. To do this, frequency dictionaries of certain linguistic units are created and with their use, the average frequency of words in a text, the hapax legomena coefficient (words that have frequency 1 in the studied sample), exclusivity index, concentration index, etc. are analyzed [1–5, 14, 24].

According the WF, one calculates such characteristics as vocabulary richness, *diversity index* (K_l) – a ratio of the volume of lexeme vocabulary (W) to text volume (N), that is $K_l=W/N$. In accordance to Table 1, the most diverse, the richest lexis is found in poetry, then in descending order, in prose, everyday colloquial style, journalism, scientific and formal business style [14, 25].

Table 1

Results of speech coefficients according to the styles of the Ukrainian language [14]

Style	W/N	W_1/N	W_1/W	W_{10}/W	W_{10t}/N
scientific	0.059	0.427	0.025	0.189	0.890
journalistic	0.070	0.450	0.031	0.121	0.804
formal business	0.030	0.280	0.0085	0.303	0.935
poetry	0.103	0.495	0.052	0.098	0.789
fiction prose	0.067	0.430	0.029	0.149	0.821
colloquial	0.073	0.465	0.034	0.161	0.789

The average frequency of a word in text A is the ratio of text volume N to the volume of lexeme vocabulary W (inverse to diversity index), that is, $A=N/W$ [26]. According to the WF data, each word in the colloquial everyday style on the average was used 14 times, and in the scientific style – 17 times [27].

Exclusivity index characterizes lexis variability, that is, percentage of a text (vocabulary), occupied by the words that were found 1 time (Table 1) [28]:

– *of vocabulary* I_{wt} – the ratio of the number of lexemes with frequency 1 W_1 to the total number of lexemes: $I_{wt}=W_1/W$ [14];

– *of text* I_t – the ratio of the number of lexemes with frequency 1 W_1 to text volume N: $I_t=W_1/N$ [14].

Concentration index indicates percentage of a text (vocabulary), occupied by the words that were used 10 or more times (Table 1) [29]:

– *of vocabulary* I_{kt} – the ratio of the number of words in vocabulary with absolute frequency 10 and more (W_{10}) to the total number of words in vocabulary (W): $I_{kt}=W_{10}/W$ [14];

– *of text* I_{tn} – the ratio of the sum of absolute frequencies of words with absolute frequency 10 and more W_{10t} to text volume N: $I_{tn}=W_{10t}/N$ [14].

As indicated by WF, speech gives preference to a small number of units, which are often used [30]. They form the core of any speech subsystem, while most units are of low frequency [31]. This regularity was noticed by Dewey at the beginning of the XX century, calling it the *outweigh law* [32]. This regularity was more researched by the German linguist J. Zipf, who formulated the *Zipf's law*, which sets the dependences [33]:

– of the word frequency and its rank in the vocabulary: the more frequent a word, the higher its rank at $F\times i=const$,

where F is the word frequency in the frequency vocabulary, i is the rank of its word [34];

– of the word frequency and its length: the more frequent a word, the shorter it is at $k=C \lg r$, where k is the length of a word in phonemes, C is the constant, r is the rank [35];

– of the word frequency and the number of its meanings: the more frequent a word, the more meanings it has at $m=C\sqrt{f}$, where m is the number of meanings of a word, C is the constant, f is the word frequency [36];

– of the word frequency and its origin: the longer a word, the more frequent it is [37].

According to the law of the German linguist P. Mencerat, the length of a language structure (word, word combination, super phrase unity, sentence) is inversely proportional to the length of its components (syllables, words, word combinations, etc.), that is, the longer the language structure, the shorter its components [14]. According to the research of G. Altmann, $y=ax^b$, where y is the average length of the constituents, x is the length of a language structure, b is the indicator that characterizes the dynamics of a change in the length of the components (the law is valid, if b<0) [38].

The *Krilov law* establishes the relationship between the number of polysemic words and frequency:

$$p_x = 1/2^x, \quad px = (\omega-1)^{x-1}/\omega^x,$$

where p_x is the probability of using a word, which has x meanings, ω is the average number of meanings of a word in the dictionary [14].

Some major quantitative characteristics of a language are very simple. For example, the difference between the number of words (10^4–10^5), the number of morphemes (several thousand), the number of syllables (from several hundred to several thousand) and the number of phonemes (from 10 to 80) [31–49]. There is an assumption that such ratios are associated with the property of human memory [39]. We will also note that the more frequent a word, the faster a person can recollect it [40]. However, there is no research in the field of dependence of changes in the coefficients of lexical author's speech during the period of his creative work [41].

3. The aim and objectives of the study

The aim of this work is to develop a method for identifying the author in texts in the Ukrainian language based on the linguometry technology.

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

– based on the analysis of coefficients of lexical author's speech in the reference text, to develop the algorithms for identification of the author of a text;

– to develop software of content monitoring to identify the author of the texts in the Ukrainian language based on the linguometric analysis of the identified stop words in text content;

– to carry out analysis of the results of experimental testing of the proposed method of content monitoring to identify the author in Ukrainian scientific texts in the technical area.

4. The method for identifying the style of the author of text content

Linguometry is a branch of applied linguistics that detects, measures, and analyzes the quantitative characteristics of the units of different levels of a language or speech [42]. One of the ways to characterize the literary richness of a text is the evaluation of the character of using language units at all language levels [43]. This makes it possible to equate the concept of *richness* and *diversity* of speech [44]. The calculation of linguistic diversity coefficients should assume the relationship of such coefficients as *lexical diversity, degree (measure) of syntactic complexity* [14], *speech coherence, indices of exclusivity and concentration of a text* [45]. Because a coefficient is an absolute value, it is possible to neglect in certain limits the length of the compared texts [46]. It is of theoretical interest to study the internal "dynamics" of a text in the part of matching coefficients from its various sections and the coefficient that is general for the entire text (Table 2) [47]:

– for lexical diversity, the larger the resulting decimal fraction, the higher the lexical diversity of a text [48];

– for syntactic complexity, the larger the fraction (within [0; 1]), the more words in general there are is such a text, and therefore, the higher the possibility of diversity of syntactic relationships between the words in a single sentence [49];

– for speech coherence, it is equal to unity when there are three connective elements in one sentence (prepositions and conjunctions) [50].

Table 2

Diversity coefficients of a text [1–5, 14, 41, 47]

Coefficient	Definition	Formula	Explanation
Lexical diversity	The ratio of the number of words to the total number of word forms. The value of the coefficient lies within [0; 1]	$K_l=W/N$	K_l is the lexical diversity coefficient, W is the number of words in a certain text, N is the total number of words of this text
Syntactic complexity	The ratio of the number of sentences to the number of words of a certain text	$K_s=1-P/W$	K_s is the coefficient of syntactical complexity, P is the number of sentences, W is the number of words in the entire text
Speech coherence coefficient	The ratio of the number of prepositions and conjunctions to the number of separate sentences	$K_z=(Z+S)/(3P)$	Z is the number of prepositions, S is the number of conjunctions, P is the number of separate sentences
Exclusiveness index	Variability of lexis, that is, the proportion of the text, which is occupied by the words found 1 time	$I_{wt}=W_1/W$	I_{wt} is the exclusiveness index, W_1 is the number of words with frequency 1, W is the number of words in the entire text
Concentration index	The proportion of the text, occupied by the words found 10 times or more	$I_{kt}=W_{10}/W$	I_{kt} is the concentration index, W_{10} is the number of words with frequency of 10 and more, W is the number of words in the whole text

It was found [14], that the text of a Ukrainian fairy tale has K_z=0,77, and the text of a scientific article in Ukrainian – 3.0, that is, the coherence in the second text is 3.9 times stronger than in the first one. There are no official standards for speech diversity coefficients for K_l and K_s [51], but the reference point for comparison and evaluation of a text in a homogeneous group of texts is the average statistical rate of the value of the coefficient for passages of the equal length [52]. 100 words will be accepted as the minimum size (length) of a passage, we will consider that the coefficients have already stabilized, reflecting the real features of the author's language [53]. The proximity or remoteness of a separate individual coefficient from the average one serves as a basis for evaluating the speech diversity in the respective text [54]. The texts, the diversity coefficients of which fall within the area of the root mean square deviations D from a certain mean, are considered satisfactory.

$$D = \sqrt{\frac{\sum_{i=1}^{n} x_i^2}{n} - (\bar{x})^2} \quad [14, 55].$$

Analysis and interpretation at the linguistic level of the stylistic features and peculiarities of the writing style of a certain author (or a certain literary epoch) includes the most basic stages, presented in algorithm 1 [14, 56].

Algorithm 1. Analysis and interpretation at the linguistic level of the stylistic features and peculiarities of the writing style of a certain author

Stage 1. Selection and primary processing of text content. For selection, the text filters are built by the parameters (the main language of the text, the text sample volume, time of the publication, publication source, format etc.) [41, 57]. The basic steps of primary text processing are:

– bringing it to a unified format (such as removing tags, if a previous publication is in the Internet-resource in the form of a static page);

– eliminating information noise (pictures, formulas, references, abstracts in other languages, etc.), which does not affect the outcome, but increases the time of processing;

– bringing to a unified volume (shortening if necessary, removing non-informative sections of the beginning and ending of a text).

Stage 2. Lemmatization of text linguistic units. Uniting word forms under the language lemma [14, 58].

Stage 3. Removal of non-homogeneity of text linguistic units. Solving the problem of non-homogeneity of text linguistic units, for example, from the position of belonging to different kinds of a language (author's, non-author's, etc.)

Stage 4. Construction of a system of frequency dictionaries based on statistical distributions in necessary frequency scales. A frequency dictionary is the type of a dictionary, which contains the number of usages (frequency) of a certain linguistic unit of a language (composition, words, word forms, word combinations, idioms, idioms) in various texts of a certain volume [59]. Usually, absolute and relative frequencies of the usage of language units are presented, dictionary article are placed in the descending order of frequencies [60].

Stage 5. Search for parameters that adequately reflect the structure of the frequency dictionary. The following parameters make it possible to formulate some basic linguostatistic methods for researching a text [61]:

– the method of anchor words (calculation of the total frequency of usage and finding percentage of syntactic words [62]: prepositions, conjunctions, particles);

– the punctuation signs method (calculation only of the number of internal and external punctuation signs) [63];

– the method of words (calculation only of the words of a certain length) [64];

– the method of sentences (calculation of only the sentences of a certain length) [65];

– the syntactical method (calculation of punctuation signs, words, and sentences of a certain length) [66];

– the combined method (a combination of the syntactical method and the method of anchor words) [67].

Stage 6. Checking effectiveness of parameters. Analysis and comparison of obtained results from the well-known author's works to identify the patterns of influence of the author's stylistics on formation of the author's structure of the frequency dictionary by these parameters [68].

Stage 7. Mathematic modeling of lexicostatistic distributions [69].

Stage 8. Construction of statistic classifications, i.e., author's reference, which reflect stylistic patterns within the works by a certain author or a certain literary style taking into consideration a literary epoch or specifics of the language, in which the analyzed works are written [70].

Stage 9. Interpretation of results from the positions of stylistic ideas over a specific time, the general and the author's style, taking into consideration time parameters [71]. Thus, we will also solve the problem of the author's attribution, which we will form as follows. Let us assume that there is a statistically processed work by an author (reference work). It is necessary to estimate belonging of certain passages to the reference work with the use of appropriate methods. A graphic representation of the relative frequency of occurrence of syntactic words in Passage 4 and in the reference work is shown in Fig. 1. The correlation coefficient for the syntactic words in this case makes up R_{e-U4}=0,7326. We will also present the correlation factors for each of the syntactic words for passages 1–4 (Table 4). Analyzing the correlation factors for syntactic words, we conclude that the probability of belonging of passages to the studied reference is the highest for Passage 4, followed by Passage 2, Passage 1, and Passage 3. We will note that for all the four passages, consistently high correlations are observed for particles, which can by understood as the lack of influence of particles on the author's style. In addition, we will analyze the frequencies of occurrences only of prepositions and conjunctions for passages, find the appropriate correlation factors and compare results (Table 3).

Table 3

Correlation factors for a syntactic part of speech and each of the passages

Passage	Preposition	Conjunction	Particle	R_{e-U}	R'_{e-U}
1	0.72	0.79	1	0.6076	0.6900
2	0.4928	0.5714	0.9580	0.7066	0.4913
3	0.1517	0.1624	0.8800	0.2810	0.2254
4	0.5639	0.9544	0.9594	0.7326	0.6905

Passage 4 remained a likely candidate to belong to the reference sample, followed with a slight margin by Passage 1, then Passage 2. Passage 3, like in the previous study, is the least likely to belong to the reference sample. To prove the results, we will turn to [1–4], from which the three passages were taken to be studied.

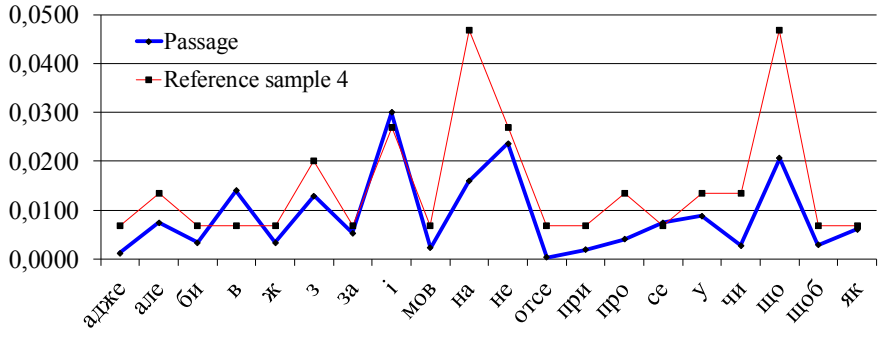

Fig. 1. Relative frequency of occurrence of syntactic words in Passage 4 and in the reference sample

5. Results of research into identifying the author in the Ukrainian scientific and technical texts

In the course of the research, we developed the system with the probability of selecting the language/languages of the analyzed content, which is implemented at the Internet site Victana [25] (Fig. 2). Analyzing the components of the formulas for estimating the richness of the work, we conclude that it is necessary to find such magnitudes as the number of words and word forms, sentences, conjunctions and prepositions, the words with the frequency of 1 and not less than 10. The algorithm for the analysis of the whole text is enabled on the server after starting the process of calculating the coefficients of text diversity (Algorithm 2).

Algorithm 2. Analysis of the style of the author's speech

Stage 1. Checking the text length – extra things are removed.

Stage 2. Clearing the studied text (figures, special characters, formulas, drawings).

Stage 3. Determining the number of sentences P.

Stage 4. Determining the total number of words in text N.

Stage 5. Determining the number of words W (by the word base frequency dictionary).

Stage 6. Calculation of language diversity coefficient: $K_l=W/N$.

Stage 7. Calculation of coefficient of syntactic complexity: $K_s=1-P/W$.

Stage 8. Determining the number of words that occurred exactly one time, that is, W_1.

Stage 9. Calculation of text exclusivity index: $I_{wt}=W_1/W$.

Stage 10. Determining the number of words that occurred more than 9 times, that is, W_{10}.

Stage 11. Calculation of text concentration index: $I_{kt}=W_{10}/W$.

Stage 12. Determining the number of prepositions Z.

Stage 13. Determining the number of conjunctions S.

Stage 14. Calculation of speech coherence coefficient:

$$K_z=(Z+S)/(3*P).$$

Stage 15. Displaying results of the Internet page at the web site Victana [25].

Analyzing the components of formulas for evaluating the richness of the work, we see that it is necessary to find the number of sentences, words and word forms, prepositions, and conjunctions, words with the frequency of 1 and with the frequency of not less than 10. For convenience, we will enter the results into the table. The generated table (Table 4) and the obtained results of the study are displayed on the screen on the information resource.

Based on the above, we will estimate the richness of passages in the scientific articles in the technical field from Visnyk of the National University "Lviv Polytechnic" in the series "Information systems and networks", written by one author over the period of 2001–2017 [25], using coefficients of diversity and speech coherence, exclusivity, and concentration indices. For the analysis, we will select the first part (10,000 characters) of each article (Alg. 3).

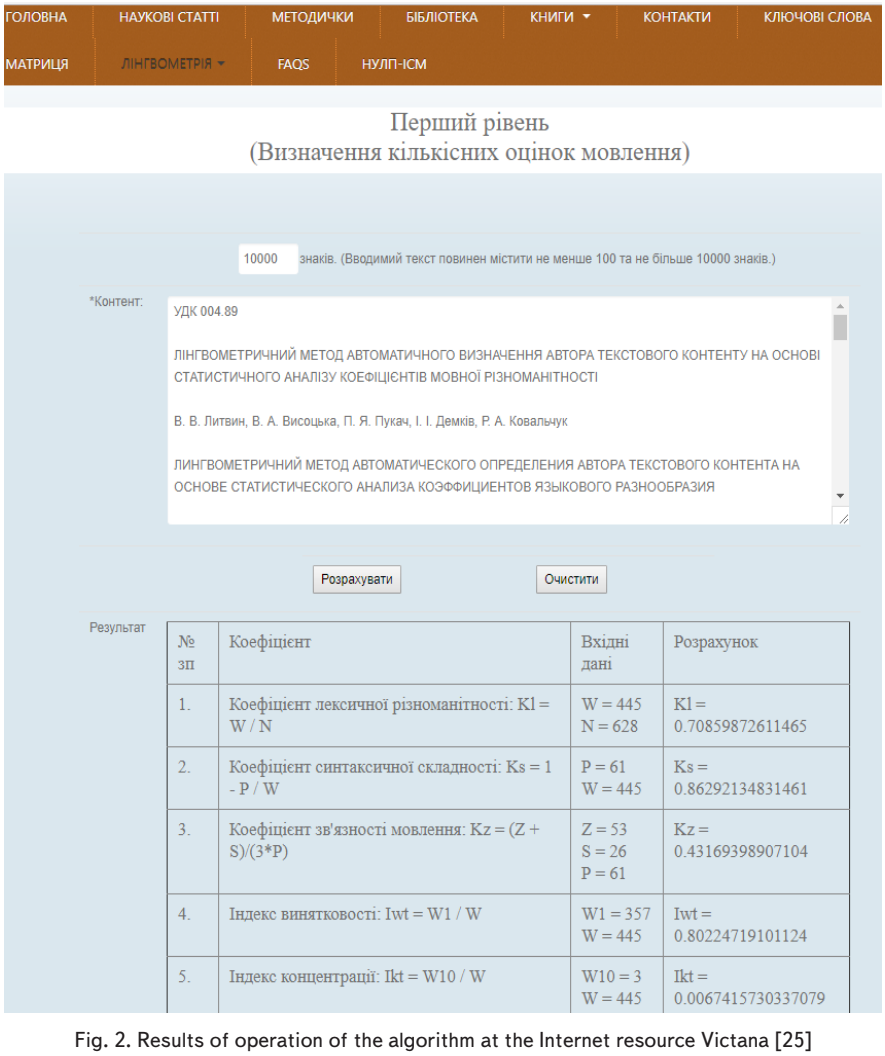

Fig. 2. Results of operation of the algorithm at the Internet resource Victana [25]

Table 4

Example of a generated table as a result of operation of the algorithm for analysis of the style of the author of a publication at the Internet site Victana [25]

Coefficient	Data	Calculation
lexical diversity: $K_l=W/N$	W=184; N=295	K_l=0.6237
syntactic complexity: $K_s=1-P/W$	P=18; W=184	K_s=0.902
speech coherence: $K_z=(Z+S)/(3*P)$	Z=20; S=28; P=18	K_z=0.889
exclusivity: $I_{wt}=W_1/W$	W_1=141; W=184	I_{wt}=0.7663
concentration: $I_{kt}=W_{10}/W$	W_{10}=2; W=184	I_{kt}=0.01

Algorithm 3. Analysis of statistics of functioning of the system of stop words identification from 215 scientific articles of the technical area

Stage 1. Analysis of 100 scientific articles to determine the range of the optimal size of the text. First, the full volume of the texts was analyzed, then these texts were analyzed to identify different numbers of characters. The results showed that the optimal study of the texts is in the range of [100; 10,000] characters. If there are less than 100 characters, the obtained information is non-informative, very often the values of the coefficients of different authors are similar and are significantly different for one author on various tests. If there are more than 10,000 characters, the coefficients do not change significantly, but the analogs for studying have a different length due to the lack of diversity of the analogues of a large length, so the maximal number of 10,000 was selected for analysis.

Stage 2. Analysis of over 200 one-author papers in technical area of over 50 different authors over the period of 2001–2017 to determine if and how the text diversity coefficients of these authors change within different periods of time.

Stage 3. Analysis of over 200 one-author papers in technical area by over 100 different authors over the period of 2001–2017 to determine if and how the text diversity coefficients of these authors change within different periods of time.

Stage 4. Analysis of over 200 one-author papers in technical area by over 100 different authors over the period of 2001–2017 to determine the speech style of these authors.

Stage 5. Analysis of the obtained coefficient of speech of more than 100 different authors over the period of 2001–2017 to determine the subset of authors with the style that is similar to 4 reference papers (collective papers, the authors of which are among the studied one-author papers).

Stage 6. Analysis of the results, obtained at stage 5. Checking if there are actual authors of these reference texts in the obtained texts. To select the best algorithm for studying the author's style in the Ukrainian scientific and technical texts based on the technology of quantitative linguistics.

For the accuracy of research, it is necessary to analyze, if the time of the publication of papers influences the text diversity coefficients, that is, if these coefficients change over time based on the sample of the same authors and texts. First, we will analyze how the total volume of words with the passages of the same size change in the range 2001–2017. As one can see, over time the same authors use shorter words more often (Fig. 3*a*).

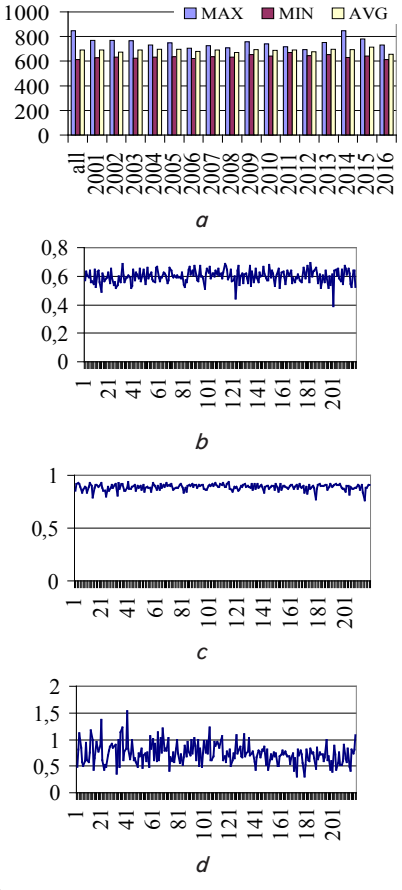

Fig. 3. Distribution: *a* – of the words and speech coefficients for passages of equal size in the range of 2001–2017: *b* – K_l; *c* – K_s; *d* – K_z

Over time, lexical diversity coefficient K_l does not change substantially (Fig. 3, b–d). Similarly, over time syntactic complexity coefficient K_s does not substantially change either. But speech coherence coefficient K_z changes insignificantly over 16 years. In the beginning (2001), it varies in the range of [0.5; 1.2], and at the end of the period – in the range of [0.4; 0.9] (Fig. 4).

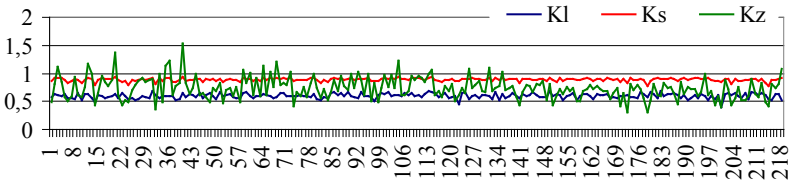

Fig. 4. Comparison of distribution of speech coefficients K_l, K_s and K_z

Similarly, we will compare distributions of indices of exclusivity and concentration (Fig. 5). While the scope of distribution does not change significantly over time for I_{wt}, significant changes were recorded for I_{kt}. Over time, the authors of these papers more often repeat some terms in their papers more than 10 times, narrowing down the circle of their research. Fig. 5, *d* shows the result of analysis of speech coefficients for the passages of the equal size in the range 2001–2017 as minimum, maximum, and mean values for this period (determining the fluctuations of values in this period). More substantial fluctuations are observed for K_z (Fig. 6).

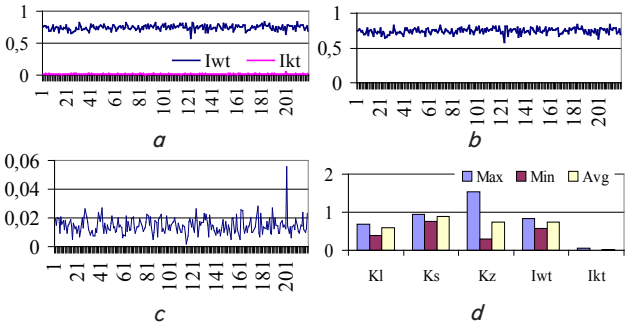

Fig. 5. Distribution of speech indices for: *a* — both indices; *b* — I_{wt}; *c* — I_{kt}; *d* — minimal, maximal, and mean values for all coefficients

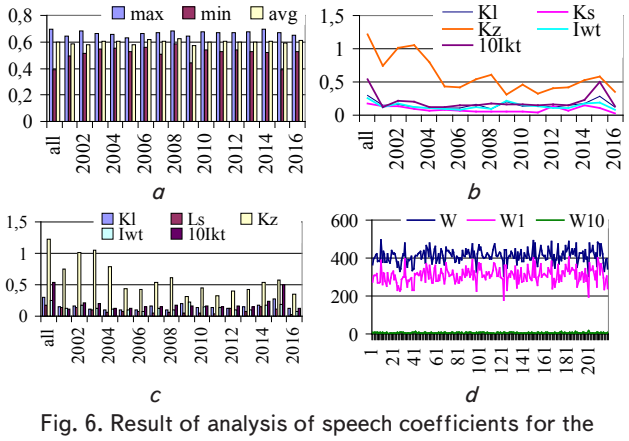

Fig. 6. Result of analysis of speech coefficients for the passages of equal size within 2001–2017: *a* — minimal, maximal, mean values for this period for K_i; *b* — diagram of dynamics of a change in coefficients within the specified period; *c* — histogram of dynamics of a change of all coefficients within the specified period; *d* — usage of word forms (all, one time, more than 10 times)

We will analyze separately the distribution of the usage of all word forms (Fig. 6, *d*), the words, used once, the words used 10 times in the studied texts for the passages of equal size in the range 2001–2017 (Fig. 7). Fig. 7, *b* shows the analysis of the usage of prepositions, conjunctions, and separate sentences in the studied texts in the passages of equal size in the range of 2001–2017, where *Z* is the number of prepositions, *S* is the number of conjunctions, *P* is the number of separate sentences. According to Fig. 7, *c*, over time, the authors use shorter sentences to describe the subject area than at the beginning of the studied period. While the number of prepositions decreases, the distribution of the use of conjunctions does not change essentially (Fig. 7, *e*). Fig. 8, *a*–*b* shows the analysis of a change in the dynamics of the use of words in the studied texts within a specified period. Fig. 8, *c*, 9, *d* show the result of the analysis of a change in the dynamics of the use of prepositions, conjunctions, and sentences in the studied texts for the specified period.

It was proved that there exists the dynamics of a change of not only in the speech coefficients of the author's text within the specified period of his work, but also the dynamics of a change in the separate components, such as the number of the use of word forms per total number of words, conjunctions, and prepositions, sentences in the determined volume of the passage, word forms that are used only once and those used more than 10 times.

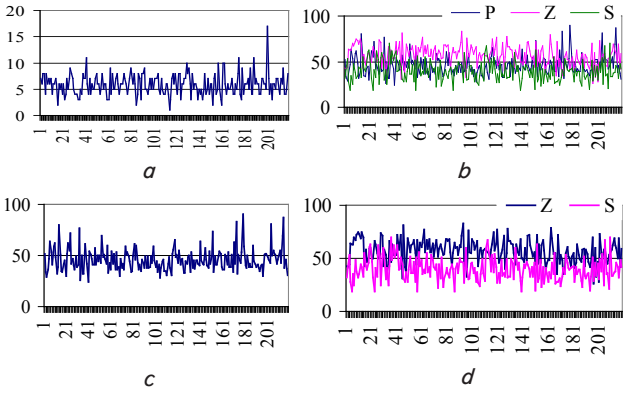

Fig. 7. Analysis of the frequency of the use of words: *a* — more than 9 times (W_{10}); *b* — parameters of speech coherence; *c* — sentences; *d* — prepositions and conjunctions

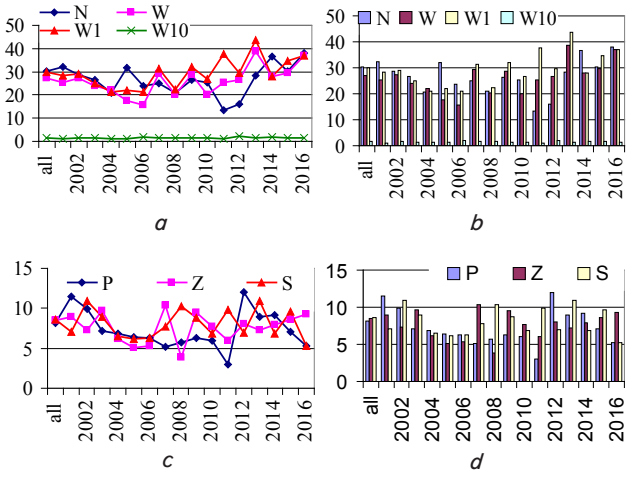

Fig. 8. Result of analysis of a change in the dynamics of the use of words in the studied texts within a certain period of time: *a* — dynamics of a change in speech parameters; *b* — distribution of values of speech parameters within the specified period of research; *c* — dynamics of a change in the use of word combinations, prepositions and sentences in the studied texts; *d* — distribution of values of the use of word combinations, prepositions and sentences for the specified period of research of authors' styles

7. Discussion of results of research into identifying the author of the Ukrainian scientific and technical texts

For more accurate identification of the magnitude of an increase in each studied parameter, it is necessary to do more substantial research on a large sample of papers, written by one author and to increase the range of research into creative work of different authors by a longer period of their creative work.

Then, we will analyze the sample for the author's style and select the best algorithm to determine the style of the author. In Fig. 9, *a*, the diagram displays the identification of the author's style by speech coefficients. In Fig. 9, *b*, the diagram with accumulation displays changes in the total sum by the speech coefficients. In Fig. 9, *c*, the normalized diagram reflects a change of contribution of each value by speech coefficients.

As we can see, coefficients of the author's speech, except for *Kz*, do not change much depending on the style of a spe-

cific author for Ukrainian scientific and technical texts. Or it changes to little extent, which complicates the process of identification of the features of the style of speech of a particular author in the totality of the analyzed authors' styles. And the larger such set, the more difficult the process of identification of a specific style of the author without any additional parameters. Then, we will analyze the sample for the author's style by such additional parameters as the total number of sentences in the passages that are equal in volume, the number of words in the sample, the frequency and occurrence of prepositions and conjunctions. In Fig. 10, the diagram displays the identification of the author's style by the additional parameters of the author's speech.

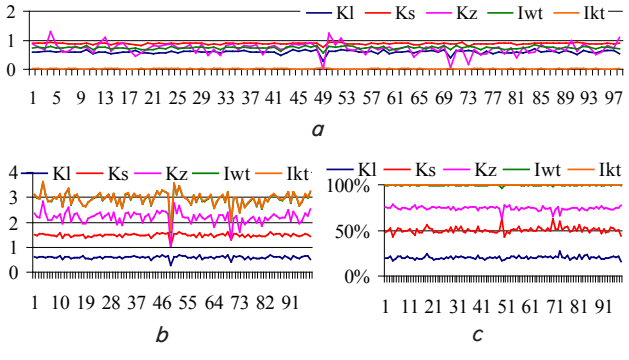

Fig. 9. Detailed analysis: *a* — of the process of identification of the author's style by speech coefficients over time; *b* — of a change of the total sum by speech coefficients; *c* — of a change of contribution of each value by speech coefficients

In Fig. 10, *b*, the diagram with accumulation displays changes in the total sum according to the parameters. In Fig. 10, *c*, the normalized diagram displays a change in contribution of each value by the parameters. As we see, the introduction of the additional parameters will decrease the set of authors, whose speech styles are similar to the Ukrainian scientific and technical style of publications. We will introduce the additional parameters, such as the number of sentences, conjunctions, and prepositions (Fig. 11) and will analyze the dynamics (Table 5).

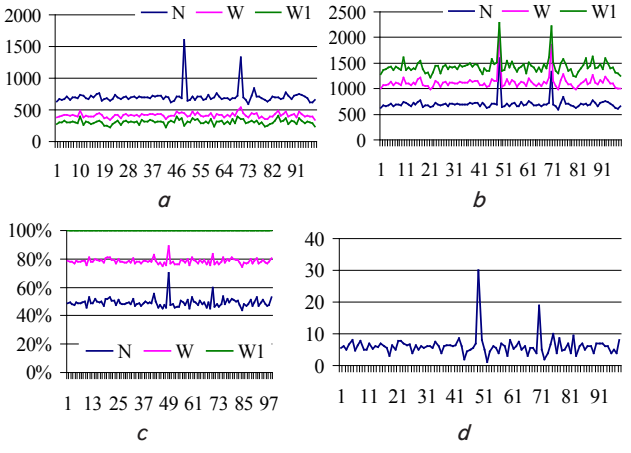

Fig. 10. Detailed analysis: *a* — of the process of identification of the author's style by parameters of speech; *b* — of a change of the total sum by speech coefficients; c — of a change of contribution of each value by speech coefficients; *d* — of a change in the parameter such as occurrence of a word over 10 times (W10)

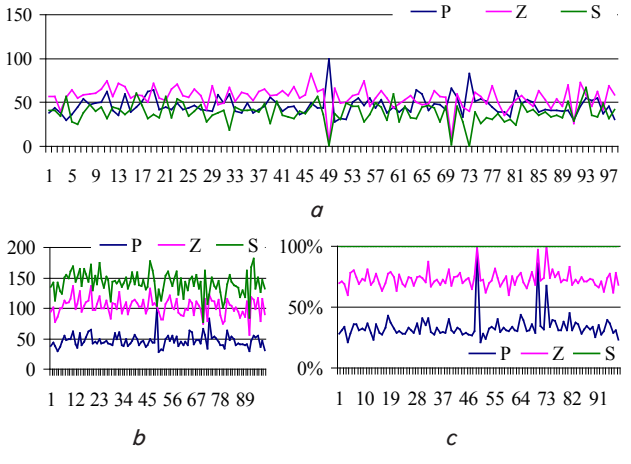

Fig. 11. Detailed analysis: *a* — of the process of determining the author's style by speech parameters; *b* — of a change in the total sum by speech coefficients; *c* — of a change in the introduction from each value by coefficients

Table 5

Result of operation of the algorithm for analysis of the style of the author of the publication at the Internet site Victana [25]

No.	N	W	W_1	W_{10}	P	Z	S	K_l	K_s	K_z	I_{wt}	I_{kt}
1	2	3	4	5	6	7	8	9	10	11	12	13
1	671.3	395.6	299	6	44.2	57.1	41.1	0.59	0.89	0.76	0.76	0.015
2	662.5	410.3	303	5	37.8	39.8	34.8	0.61	0.9	0.67	0.74	0.012
3	668.8	418.3	325.8	6.8	29.8	56	57	0.63	0.93	1.28	0.78	0.016
4	708	419	309	8	36	64	28	0.59	0.91	0.85	0.74	0.019
5	661.1	402.7	299.7	4.7	44.7	54.7	24.8	0.61	0.89	0.6	0.74	0.012
6	694.5	417.4	313.1	6.4	54.3	58.5	38.1	0.6	0.87	0.62	0.75	0.015
7	691.8	403.4	301.6	7.8	47.8	60	47.8	0.58	0.88	0.79	0.75	0.019
8	682.5	394.2	291	5	49	61	39.7	0.58	0.88	0.74	0.74	0.013
9	733.5	486.5	392	5	50	65	45	0.66	0.9	0.76	0.8	0.01
10	729	380	261	7	62	75	32	0.52	0.84	0.58	0.69	0.018
11	686.5	414.5	312.6	5.9	41.1	56.9	45	0.6	0.9	0.86	0.75	0.012
12	665.5	399	299	6	35.5	72	43	0.6	0.91	1.09	0.75	0.015
13	724.2	394.2	278.8	5.8	59.6	68.4	36.8	0.55	0.85	0.61	0.71	0.015
14	691	396.7	289	7	39	55.3	42.3	0.57	0.9	0.85	0.73	0.018
15	745	439	319	6	45	59	61	0.59	0.9	0.89	0.73	0.014

Continuation of Table 5

1	2	3	4	5	6	7	8	9	10	11	12	13
16	768	452.5	323	5.5	51.5	58	47	0.59	0.89	0.68	0.71	0.012
17	647	422	308	3	62	50	32	0.65	0.85	0.44	0.73	0.007
18	677.5	373.5	255	6.5	64.5	72	36	0.55	0.86	0.57	0.68	0.018
19	680	379	251	5	42	55	33	0.56	0.89	0.7	0.66	0.013
20	642	337.5	230.3	7.8	44.8	52.3	56.8	0.52	0.87	0.81	0.68	0.023
21	665	376	275.7	7.7	41.7	65	32.3	0.57	0.89	0.79	0.73	0.02
22	731	420	301	7	49	71	54	0.57	0.88	0.85	0.72	0.017
23	691.7	425.7	331.3	6.5	41.8	58.2	50	0.62	0.9	0.88	0.78	0.015
24	668.8	368.3	262.5	6.8	44	55.8	34.5	0.55	0.88	0.73	0.71	0.018
25	691	421	311	4	47	65	40	0.6	0.89	0.74	0.74	0.01
26	708.5	434	323.5	6.5	42	57.5	47.5	0.61	0.9	0.84	0.75	0.015
27	665	406	309	5	41	42	28	0.61	0.9	0.57	0.76	0.012
28	700	418.5	320.5	6	40	68.5	35	0.6	0.9	0.88	0.77	0.014
29	704.5	412	303.5	5.5	59	47.5	38	0.58	0.86	0.49	0.74	0.013
30	688.8	416.8	321.9	6	49.7	49.3	41.3	0.6	0.88	0.67	0.77	0.016
31	711	396	268	6	60	67	19	0.56	0.85	0.48	0.68	0.015
32	691	436.7	336.7	5.7	40	51	44.7	0.63	0.91	0.82	0.77	0.013
33	695	422.5	318.3	7.5	38.5	61.3	41	0.6	0.91	0.89	0.75	0.018
34	699	427	314	6	49.5	60	41	0.61	0.88	0.69	0.74	0.014
35	683	438	339	4	38	52	42	0.64	0.91	0.82	0.77	0.009
36	730	440	323	6	42	62	39	0.6	0.9	0.8	0.73	0.014
37	714.5	418.5	304.5	6.5	46	65	48.5	0.59	0.89	0.86	0.73	0.016
38	717.5	433.5	321.5	6.5	56	57.5	26.5	0.6	0.87	0.5	0.74	0.015
39	728	430	313	6	49	59	51	0.59	0.89	0.75	0.73	0.014
40	666	401.5	305	6.5	40	63	35.5	0.6	0.9	0.82	0.76	0.016
41	715.5	352	223.5	8.5	45	58	34	0.49	0.87	0.68	0.63	0.024
42	699	401	302	6	46	68	32	0.57	0.89	0.72	0.75	0.015
43	620	411	323	2	36	55	40	0.66	0.91	0.88	0.79	0.005
44	645	403	302.3	4.3	39.3	58.7	37.7	0.62	0.9	0.84	0.74	0.011
45	708	475	392	5	49	83	46	0.67	0.9	0.88	0.83	0.011
46	708	442.5	336.5	5.5	43.5	62	56.5	0.63	0.9	0.91	0.76	0.012
47	689	458	369	7	44	65	36	0.66	0.9	0.77	0.81	0.015
48	1602	442	245	30	100	3	1	0.28	0.77	0.01	0.55	0.068
49	644	400	310	8	28	66	37	0.62	0.93	1.23	0.78	0.02
50	661.5	402.5	302	5	32	49.5	31	0.6	0.92	0.84	0.75	0.012
51	705	474	369	1	31	50	49	0.67	0.93	1.06	0.78	0.002
52	656	422.5	341.5	4.5	50	57.5	46	0.64	0.88	0.69	0.81	0.011
53	704.8	458.8	360	6	54.8	60	45.8	0.65	0.88	0.66	0.78	0.013
54	716	413.5	293	5.5	47	74.5	27.5	0.58	0.89	0.73	0.71	0.013
55	652	389	287	4	55	46	36	0.6	0.86	0.5	0.74	0.01
56	666	412	318	7	44	55	49	0.62	0.89	0.79	0.77	0.017
57	732	402	290	6	53	63	45	0.55	0.87	0.68	0.72	0.015
58	670	449	356	3	38	55	30	0.67	0.92	0.75	0.79	0.007
59	693	366	242	8	45	44	60	0.53	0.88	0.77	0.66	0.022
60	761	440	315.8	5.3	39.3	48.5	28.3	0.58	0.91	0.65	0.71	0.012
61	717	422	310	6	45	53	46	0.59	0.89	0.73	0.73	0.014
62	673.5	419	329	7.5	39	58	33	0.62	0.91	0.78	0.79	0.018
63	679	381	280	5	64	50	32	0.56	0.83	0.43	0.73	0.013
64	682.6	416.2	318	6.2	60	47.8	45	0.6	0.86	0.59	0.76	0.015
65	658	399	277	3	41	48	47	0.6	0.9	0.78	0.69	0.008
66	683	446	357	5.5	48.5	63	43.5	0.65	0.89	0.74	0.8	0.012
67	689.5	407.5	296	5.5	47.5	57	28	0.59	0.88	0.61	0.73	0.014
68	726	493	399	4	42	56	46	0.68	0.91	0.81	0.81	0.008
69	1325	538	360	19	66	9	2	0.4	0.88	0.06	0.67	0.035
70	697	450	361.5	5	56	59.5	46	0.65	0.88	0.63	0.8	0.011
71	652	405	296	2	34	45	28	0.62	0.92	0.72	0.73	0.005
72	598	386	309	4	83	40	0	0.65	0.78	0.16	0.8	0.01
73	726.3	441.3	332.3	6.7	51	61.3	39	0.6	0.88	0.68	0.75	0.015
74	846	440	299	10	54	57	26	0.52	0.88	0.51	0.68	0.023
75	712.5	442.5	331.5	4	51	48	33	0.62	0.88	0.53	0.75	0.009
76	706	374	275	8.5	45	68.5	31	0.53	0.88	0.74	0.73	0.023
77	682.3	398.7	296.3	4.7	39	50.3	37.7	0.58	0.9	0.75	0.74	0.012

Continuation of Table 5

1	2	3	4	5	6	7	8	9	10	11	12	13
78	654	361	240	5	39	35	28	0.55	0.89	0.54	0.66	0.014
79	631	350	249	7	34	45	31	0.55	0.9	0.75	0.71	0.02
80	661	391	275	4	63	53	24	0.59	0.84	0.41	0.7	0.01
81	709.5	399	292.5	9.5	48	58	49.5	0.56	0.88	0.75	0.73	0.024
82	695	436	332	3	53	51	39	0.63	0.88	0.57	0.76	0.007
83	700	485	406	6	50	46	42	0.69	0.9	0.59	0.84	0.012
84	674	404	316	7	39	63	35	0.9	0.9	0.84	0.78	0.017
85	685	432	333	5	42	53	39	0.63	0.9	0.73	0.77	0.012
86	780	479	366	6	41	43	34	0.61	0.91	0.63	0.76	0.013
87	723	401	280	6	41	54	35	0.55	0.9	0.72	0.7	0.015
88	665	425	324	4	40	46	33	0.64	0.91	0.66	0.76	0.009
89	730	433	317	7	41	70	51	0.59	0.91	0.98	0.73	0.016
90	734	381	273	7	30	26	29	0.52	0.92	0.61	0.72	0.018
91	749	478	375	7	46	73	49	0.64	0.9	0.88	0.78	0.015
92	732	429	329	6	55	59	67	0.59	0.87	0.76	0.77	0.014
93	709	398	285	6	52	46	35	0.56	0.87	0.52	0.72	0.015
94	680	414	314	4	55	62	34	0.6	0.87	0.58	0.76	0.01
95	622	397	305	5	37	42	48	0.64	0.91	0.81	0.77	0.013
96	614	391	287	4	46	69	32	0.64	0.88	0.73	0.73	0.01
97	658	345	241	8	31	59	42	0.52	0.91	1.07	0.7	0.023
98	631.3	377.7	277.7	5.7	38	56.7	40.7	0.6	0.9	0.88	0.73	0.015

Table 5 shows the results of analysis of the style of 94 authors in papers written by one author (over 200 papers) in technical field over the period of 2001–2017. For each author, we will derive arithmetic mean value of each coefficient and parameter of speech based on the analysis of several of his work within the specified period. The styles of 4 articles of one team of authors at numbers 95–98 (in the Table they are highlighted in yellow), a part of the authors of which are in Table 5 at number 6 and 30 (in the Table, they are highlighted in blue).

However, too small sample of texts for analysis (more than 200) and the number of authors (94) does not guarantee exact results. The study should be extended to a greater number of texts, which are not always easily accessible. In the future, it is necessary to improve the method due to the analysis of texts by the methods of stylemetry and glottochronology.

6. Conclusions

1. The method for identifying the author of the text based on the analysis of coefficients of the lexical author's speech in the reference sample passage of the author's text was developed. The algorithm of lexical analysis of Ukrainian texts and the algorithm of the parser of text content based on analysis of each word taking into consideration its part of speech and declension was designed. That is, when analyzing the linguistic units of the type of words, belonging to a part of speech and declension within this part of speech were taken into consideration. For this, the analysis of flexions of these words for classification, separation of the base for the formation of the corresponding alphabetical-frequency dictionaries was performed. Filling these dictionaries was subsequently considered at the following stages of determining the authorship of a text as calculation of parameters and coefficients of the author's speech. Syntactic words (stop or anchor) words are most essential for an individual style of an author, as they are not related to the subject and content of the publication. The algorithm for determining stop words in the text content based on linguistic analysis of text content was developed. Its features are the adaptation of morphological and syntactic analysis of the lexical units to the features of the structure of Ukrainian words/texts. The theoretical and experimental substantiation of the method for content monitoring and determining stop words of the Ukrainian text were presented. The method is aimed at automatic detection of significant stop words of the Ukrainian text at the expense of the proposed formal approach to implementation of parsing of text content in the scientific and technical area.

2. The approach to the development of software of content monitoring to identify the author in Ukrainian scientific and technical texts based on NLP, stylemetry, and Web Mining was proposed. More than 200 scientific publications written by one author from all issues in Visnyk of the National University "Lviv Polytechnic" from the series "Information systems and networks" (Ukraine) over the period of 2001–2017 were analyzed by the developed system. The internal "dynamics" of these texts of randomly selected authors was studied through the analysis of coefficients of speech coherence, lexical diversity, and syntactic complexity, as well as indices of concentration and exclusivity for the first k, n and m (without a title) words of the author's passage and the one that was analyzed.

3. The results of experimental testing of the proposed method of content monitoring for identifying the author in Ukrainian scientific texts of the technical profile were studied. We compared the results in a set of 200 one-author papers in the technical area of more than 100 different authors over the period of 2001–2017 to determine if and how the coefficients of diversity of a text of these authors change within different periods of time. Based on the developed software, we obtained the results of experimental testing of the proposed method of content monitoring to identify and analyze stop words in Ukrainian scientific texts of the technical area based on Web Mining technology. It was found that for the selected experimental base of more than

200 papers, the best results according to the density criterion are reached by the method for analysis of an article without the initial compulsory information, such as abstracts and keywords in different languages, as well as the list of literature. Testing the proposed method for identification of the author's style from other categories of texts – scientific, humanitarian, artistic, journalistic, etc. – requires further experimental research.

References

1. Development of a method for the recognition of author's style in the Ukrainian language texts based on linguometry, stylemetry and glottochronology / Lytvyn V., Vysotska V., Pukach P., Bobyk I., Uhryn D. // Eastern-European Journal of Enterprise Technologies. 2017. Vol. 4, Issue 2 (88). P. 10–19. doi: https://doi.org/10.15587/1729-4061.2017.107512
2. Development of a method for determining the keywords in the slavic language texts based on the technology of web mining / Lytvyn V., Vysotska V., Pukach P., Brodyak O., Ugryn D. // Eastern-European Journal of Enterprise Technologies. 2017. Vol. 2, Issue 2 (86). P. 14–23. doi: https://doi.org/10.15587/1729-4061.2017.98750
3. The method of formation of the status of personality understanding based on the content analysis / Lytvyn V., Pukach P., Bobyk I., Vysotska V. // Eastern-European Journal of Enterprise Technologies. 2016. Vol. 5, Issue 2 (83). P. 4–12. doi: https://doi.org/10.15587/1729-4061.2016.77174
4. Method of functioning of intelligent agents, designed to solve action planning problems based on ontological approach / Lytvyn V., Vysotska V., Pukach P., Vovk M., Ugryn D. // Eastern-European Journal of Enterprise Technologies. 2017. Vol. 3, Issue 2 (87). P. 11–17. doi: https://doi.org/10.15587/1729-4061.2017.103630
5. Analysis of statistical methods for stable combinations determination of keywords identification / Lytvyn V., Vysotska V., Uhryn D., Hrendus M., Naum O. // Eastern-European Journal of Enterprise Technologies. 2018. Vol. 2, Issue 2 (92). P. 23–37. doi: https://doi.org/10.15587/1729-4061.2018.126009
6. Khomytska I., Teslyuk V. 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). 2016. doi: https://doi.org/10.1109/stc-csit.2016.7589887
7. Khomytska I., Teslyuk V. The Method of Statistical Analysis of the Scientific, Colloquial, Belles-Lettres and Newspaper Styles on the Phonological Level // Advances in Intelligent Systems and Computing. 2016. P. 149–163. doi: https://doi.org/10.1007/978-3-319-45991-2_10
8. Mobasher B. Data Mining for Web Personalization // Lecture Notes in Computer Science. 2007. P. 90–135. doi: https://doi.org/10.1007/978-3-540-72079-9_3
9. Dinucă C. E., Ciobanu D. Web Content Mining // Annals of the University of Petroşani. Economics. 2012. Vol. 12, Issue 1. P. 85–92.
10. Xu G., Zhang Y., Li L. Web Content Mining // Web Mining and Social Networking. 2010. P. 71–87. doi: https://doi.org/10.1007/978-1-4419-7735-9_4
11. Avtomaticheskaya obrabotka tekstov na estestvennom yazyke i komp'yuternaya lingvistika / Bol'shakova E., Klyshinskiy E., Lande D., Noskov A., Peskova O., Yagunova E. Moscow: MIEM, 2011. 272 p.
12. Anisimov A., Marchenko A. Sistema obrabotki tekstov na estestvennom yazyke // Iskusstvenniy intellekt. 2002. Issue 4. P. 157–163.
13. Perebyinis V. Matematychna linhvistyka. Ukrainska mova. Kyiv, 2000. P. 287–302.
14. Buk S. Osnovy statystychnoi lingvistyky. Lviv, 2008. 124 p.
15. Perebyinis V. Statystychni metody dlia linhvistiv. Vinnytsia, 2013. 176 p.
16. Braslavskiy P. I. Intellektual'nye informacionnye sistemy. URL: http://www.kansas.ru/ai2006/
17. Lande D., Zhyhalo V. Pidkhid do rishennia problem poshuku dvomovnoho plahiatu // Problemy informatyzatsiyi ta upravlinnia. 2008. Issue 2 (24). P. 125–129.
18. Varfolomeev A. Psihosemantika slova i lingvostatistika teksta. Kaliningrad, 2000. 37 p.
19. Sushko S., Fomychova L., Barsukov Ye. Chastoty povtoriuvanosti bukv i bihram u vidkrytykh tekstakh ukrainskoiu movoiu // Ukrainian Information Security Research Journal. 2010. Vol. 12, Issue 3 (48). doi: https://doi.org/10.18372/2410-7840.12.1968
20. Kognitivnaya stilometriya: k postanovke problemy. URL: http://www.manekin.narod.ru/hist/styl.htm
21. Kocherhan M. Vstup do movoznavstva. Kyiv, 2005. 368 p.
22. Rodionova E. Metody atribucii hudozhestvennyh tekstov // Strukturnaya i prikladnaya lingvistika. 2008. Issue 7. P. 118–127.
23. Meshcheryakov R. V., Vasyukov N. S. Modeli opredeleniya avtorstva teksta. URL: http://db.biysk.secna.ru/conference/conference.conference.doc_download?id_thesis_dl=427
24. Morozov N. A. Lingvisticheskie spektry. URL: http://www.textology.ru/library/book.aspx?bookId=1&textId=3
25. Victana. URL: http://victana.lviv.ua/nlp/linhvometriia
26. Method of Integration and Content Management of the Information Resources Network / Kanishcheva O., Vysotska V., Chyrun L., Gozhyj A. // Advances in Intelligent Systems and Computing. 2017. P. 204–216. doi: https://doi.org/10.1007/978-3-319-70581-1_14
27. Information resources processing using linguistic analysis of textual content / Su J., Vysotska V., Sachenko A., Lytvyn V., Burov Y. // 2017 9th IEEE International Conference on Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications (IDAACS). 2017. doi: https://doi.org/10.1109/idaacs.2017.8095038
28. The risk management modelling in multi project environment / Lytvyn V., Vysotska V., Veres O., Rishnyak I., Rishnyak H. // 2017 12th International Scientific and Technical Conference on Computer Sciences and Information Technologies (CSIT). 2017. doi: https://doi.org/10.1109/stc-csit.2017.8098730

29. Peculiarities of content forming and analysis in internet newspaper covering music news / Korobchinsky M., Chyrun L., Chyrun L., Vysotska V. // 2017 12th International Scientific and Technical Conference on Computer Sciences and Information Technologies (CSIT). 2017. doi: https://doi.org/10.1109/stc-csit.2017.8098735
30. Intellectual system design for content formation / Naum O., Chyrun L., Vysotska V., Kanishcheva O. // 2017 12th International Scientific and Technical Conference on Computer Sciences and Information Technologies (CSIT). 2017. doi: https://doi.org/10.1109/stc-csit.2017.8098753
31. The Contextual Search Method Based on Domain Thesaurus / Lytvyn V., Vysotska V., Burov Y., Veres O., Rishnyak I. // Advances in Intelligent Systems and Computing. 2017. P. 310–319. doi: https://doi.org/10.1007/978-3-319-70581-1_22
32. Marchenko O. Modeliuvannia semantychnoho kontekstu pry analizi tekstiv na pryrodniy movi // Visnyk Kyivskoho universytetu. 2006. Issue 3. P. 230–235.
33. Jivani A. G. A Comparative Study of Stemming Algorithms // Int. J. Comp. Tech. Appl. 2011. Vol. 2, Issue 6. P. 1930–1938.
34. Using Structural Topic Modeling to Detect Events and Cluster Twitter Users in the Ukrainian Crisis / Mishler A., Crabb E. S., Paletz S., Hefright B., Golonka E. // Communications in Computer and Information Science. 2015. P. 639–644. doi: https://doi.org/10.1007/978-3-319-21380-4_108
35. Rodionova E. Metody atribucii hudozhestvennyh tekstov // Strukturnaya i prikladnaya lingvistika. 2008. Issue 7. P. 118–127.
36. Bubleinyk L. Osoblyvosti khudozhnoho movlennia. Lutsk, 2000. 179 p.
37. Kowalska K., Cai D., Wade S. Sentiment Analysis of Polish Texts // International Journal of Computer and Communication Engineering. 2012. P. 39–42. doi: https://doi.org/10.7763/ijcce.2012.v1.12
38. Kotsyba N. The current state of work on the Polish–Ukrainian Parallel Corpus (PolUKR) // Organization and Development of Digital Lexical Resources. 2009. P. 55–60.
39. Single-frame image super-resolution based on singular square matrix operator / Rashkevych Y., Peleshko D., Vynokurova O., Izonin I., Lotoshynska N. // 2017 IEEE First Ukraine Conference on Electrical and Computer Engineering (UKRCON). 2017. doi: https://doi.org/10.1109/ukrcon.2017.8100390
40. Learning-Based Image Scaling Using Neural-Like Structure of Geometric Transformation Paradigm / Tkachenko R., Tkachenko P., Izonin I., Tsymbal Y. // Studies in Computational Intelligence. 2017. P. 537–565. doi: https://doi.org/10.1007/978-3-319-63754-9_25
41. Vysotska V. Linguistic analysis of textual commercial content for information resources processing // 2016 13th International Conference on Modern Problems of Radio Engineering, Telecommunications and Computer Science (TCSET). 2016. doi: https://doi.org/10.1109/tcset.2016.7452160
42. Detection of near dublicates in tables based on the locality-sensitive hashing method and the nearest neighbor method / Lizunov P., Biloshchytskyi A., Kuchansky A., Biloshchytska S., Chala L. // Eastern-European Journal of Enterprise Technologies. 2016. Vol. 6, Issue 4 (84). P. 4–10. doi: https://doi.org/10.15587/1729-4061.2016.86243
43. Conceptual model of automatic system of near duplicates detection in electronic documents / Biloshchytskyi A., Kuchansky A., Biloshchytska S., Dubnytska A. // 2017 14th International Conference The Experience of Designing and Application of CAD Systems in Microelectronics (CADSM). 2017. doi: https://doi.org/10.1109/cadsm.2017.7916155
44. Vysotska V., Rishnyak I., Chyryn L. Analysis and Evaluation of Risks in Electronic Commerce // 2007 9th International Conference – The Experience of Designing and Applications of CAD Systems in Microelectronics. 2007. doi: https://doi.org/10.1109/cadsm.2007.4297570
45. Vysotska V., Chyrun L., Chyrun L. Information technology of processing information resources in electronic content commerce systems // 2016 XIth International Scientific and Technical Conference Computer Sciences and Information Technologies (CSIT). 2016. doi: https://doi.org/10.1109/stc-csit.2016.7589909
46. Vysotska V., Chyrun L., Chyrun L. The commercial content digest formation and distributional process // 2016 XIth International Scientific and Technical Conference Computer Sciences and Information Technologies (CSIT). 2016. doi: https://doi.org/10.1109/stc-csit.2016.7589902
47. Content linguistic analysis methods for textual documents classification / Lytvyn V., Vysotska V., Veres O., Rishnyak I., Rishnyak H. // 2016 XIth International Scientific and Technical Conference Computer Sciences and Information Technologies (CSIT). 2016. doi: https://doi.org/10.1109/stc-csit.2016.7589903
48. Lytvyn V., Vysotska V. Designing architecture of electronic content commerce system // 2015 Xth International Scientific and Technical Conference "Computer Sciences and Information Technologies" (CSIT). 2015. doi: https://doi.org/10.1109/stc-csit.2015.7325446
49. Vysotska V., Chyrun L. Analysis features of information resources processing // 2015 Xth International Scientific and Technical Conference "Computer Sciences and Information Technologies" (CSIT). 2015. doi: https://doi.org/10.1109/stc-csit.2015.7325448
50. Application of sentence parsing for determining keywords in Ukrainian texts / Vasyl L., Victoria V., Dmytro D., Roman H., Zoriana R. // 2017 12th International Scientific and Technical Conference on Computer Sciences and Information Technologies (CSIT). 2017. doi: https://doi.org/10.1109/stc-csit.2017.8098797
51. Maksymiv O., Rak T., Peleshko D. Video-based Flame Detection using LBP-based Descriptor: Influences of Classifiers Variety on Detection Efficiency // International Journal of Intelligent Systems and Applications. 2017. Vol. 9, Issue 2. P. 42–48. doi: https://doi.org/10.5815/ijisa.2017.02.06
52. Peleshko D., Rak T., Izonin I. Image Superresolution via Divergence Matrix and Automatic Detection of Crossover // International Journal of Intelligent Systems and Applications. 2016. Vol. 8, Issue 12. P. 1–8. doi: https://doi.org/10.5815/ijisa.2016.12.01

53. The results of software complex OPTAN use for modeling and optimization of standard engineering processes of printed circuit boards manufacturing / Bazylyk O., Taradaha P., Nadobko O., Chyrun L., Shestakevych T. // 2012 11th International Conference on «Modern Problems of Radio Engineering, Telecommunications and Computer Science» (TCSET). 2012. P. 107–108.
54. The software complex development for modeling and optimizing of processes of radio-engineering equipment quality providing at the stage of manufacture / Bondariev A., Kiselychnyk M., Nadobko O., Nedostup L., Chyrun L., Shestakevych T. // TCSET'2012. 2012. P. 159.
55. Riznyk V. Multi-modular Optimum Coding Systems Based on Remarkable Geometric Properties of Space // Advances in Intelligent Systems and Computing. 2017. Vol. 512. P. 129–148. doi: https://doi.org/10.1007/978-3-319-45991-2_9
56. Development and Implementation of the Technical Accident Prevention Subsystem for the Smart Home System / Teslyuk V., Beregovskyi V., Denysyuk P., Teslyuk T., Lozynskyi A. // International Journal of Intelligent Systems and Applications. 2018. Vol. 10, Issue 1. P. 1–8. doi: https://doi.org/10.5815/ijisa.2018.01.01
57. Basyuk T. The main reasons of attendance falling of internet resource // 2015 Xth International Scientific and Technical Conference "Computer Sciences and Information Technologies" (CSIT). 2015. doi: https://doi.org/10.1109/stc-csit.2015.7325440
58. Pasichnyk V., Shestakevych T. The model of data analysis of the psychophysiological survey results // Advances in Intelligent Systems and Computing. 2017. Vol. 512. P. 271–281. doi: https://doi.org/10.1007/978-3-319-45991-2_18
59. Zhezhnych P., Markiv O. Linguistic Comparison Quality Evaluation of Web-Site Content with Tourism Documentation Objects // Advances in Intelligent Systems and Computing. 2018. Vol. 689. P. 656–667. doi: https://doi.org/10.1007/978-3-319-70581-1_45
60. Chernukha O., Bilushchak Y. Mathematical modeling of random concentration field and its second moments in a semispace with erlangian disrtibution of layered inclusions // Task Quarterly. 2016. Vol. 20, Issue 3. P. 295–334.
61. Davydov M., Lozynska O. 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). 2017. doi: https://doi.org/10.1109/stc-csit.2017.8098734
62. Davydov M., Lozynska O. Mathematical Method of Translation into Ukrainian Sign Language Based on Ontologies // Advances in Intelligent Systems and Computing. 2018. Vol. 689. P. 89–100. doi: https://doi.org/10.1007/978-3-319-70581-1_7
63. Davydov M., Lozynska O. Linguistic models of assistive computer technologies for cognition and communication // 2016 XIth International Scientific and Technical Conference Computer Sciences and Information Technologies (CSIT). 2016. doi: https://doi.org/10.1109/stc-csit.2016.7589898
64. Mykich K., Burov Y. Uncertainty in situational awareness systems // 2016 13th International Conference on Modern Problems of Radio Engineering, Telecommunications and Computer Science (TCSET). 2016. doi: https://doi.org/10.1109/tcset.2016.7452165
65. Mykich K., Burov Y. Algebraic Framework for Knowledge Processing in Systems with Situational Awareness // Advances in Intelligent Systems and Computing. 2016. P. 217–227. doi: https://doi.org/10.1007/978-3-319-45991-2_14
66. Mykich K., Burov Y. Research of uncertainties in situational awareness systems and methods of their processing // Eastern-European Journal of Enterprise Technologies. 2016. Vol. 1, Issue 4 (79). P. 19–27. doi: https://doi.org/10.15587/1729-4061.2016.60828
67. Mykich K., Burov Y. 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
68. Kravets P. The control agent with fuzzy logic // Perspective Technologies and Methods in MEMS Design, MEMSTECH'2010 – Proceedings of the 6th International Conference. Lviv, 2010. P. 40–41.
69. On the Asymptotic Methods of the Mathematical Models of Strongly Nonlinear Physical Systems / Pukach P., Il'kiv V., Nytrebych Z., Vovk M., Pukach P. // Advances in Intelligent Systems and Computing. 2018. Vol. 689. P. 421–433. doi: https://doi.org/10.1007/978-3-319-70581-1_30
70. Kravets P. The Game Method for Orthonormal Systems Construction // 2007 9th International Conference – The Experience of Designing and Applications of CAD Systems in Microelectronics. 2007. doi: https://doi.org/10.1109/cadsm.2007.4297555
71. Kravets P. Game Model of Dragonfly Animat Self-Learning // Perspective Technologies and Methods in MEMS Design. 2016. P. 195–201.