Запропоновано формальний підхід реалізації визначення автора україномовного тексту. Дослідження проводилось в україномовних наукових текстах технічного профілю. Проаналізовані результати застосування розроблених алгоритмів автоматичного визначення автора текстового контенту на основі методів NLP та стилеметрії. Розглянуто перспективи та особливості застосування інформаційних технологій стилеметрії для визначення автора текстового контенту. Квантитативний контент-аналіз текстового контенту науково-технічного спрямування використовує переваги контент-моніторингу та контент-аналізу тексту на основі методів NLP, Web-Mining та стилеметрії для визначення множини авторів, стилі мовлення яких подібні з досліджуваним уривком тексту. Це звужує коло пошуку при подальшому використанні в методах стилеметрії для визначення ступеня приналежності аналізованого тексту конкретному авторові.

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. Паралельно проаналізовані такі 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. Подальшого експериментального дослідження потребує апробація запропонованого методу для визначення ключових слів з інших категорій текстів – наукових гуманітарного профілю, художніх, публіцистичних тощо

Ключові слова: NLP, контент-моніторинг, стоп-слова, контент-аналіз, статистичний лінгвістичний аналіз, квантитативна лінгвістика

UDC 004.89
DOI: 10.15587/1729-4061.2018.149596

ANALYSIS OF THE DEVELOPED QUANTITATIVE METHOD FOR AUTOMATIC ATTRIBUTION OF SCIENTIFIC AND TECHNICAL TEXT CONTENT WRITTEN IN UKRAINIAN

V. Lytvyn
Doctor of Technical Sciences, Professor*
V. Vysotska
PhD, Associate Professor*
E-mail: victoria.a.vysotska@lpnu.ua
P. Pukach
Doctor of Technical Sciences, Professor***
Z. Nytrebych
Doctor of Physical and Mathematical Sciences, Professor
Department of Mathematics**
I. Demkiv
Doctor of Physical and Mathematical Sciences, Associate Professor
Department of Computational Mathematics and Programming**
A. Senyk
PhD, Associate Professor***
O. Malanchuk
PhD
Department of Biophysics
Danylo Halytsky Lviv National Medical University
Pekarska str., 69, Lviv, Ukraine, 79010
S. Sachenko
PhD
Department of Economic Assessment and Business Audit
Ternopil National Economic University
Lvivska str., 11, Ternopil, Ukraine, 46009
R. Kovalchuk
PhD, Associate Professor***
N. Huzyk
PhD***
*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, 79012

1. Introduction

The scheme of combining methods of attribution of Ukrainian scientific and technical text content consists of lexical and syntactic levels [1]. Use of the syntactic level involves calculation of linguistic correlations in word combinations [2]. A model for constructing author's style profile is proposed in [3]. It consists of a characteristic author's vocabulary and author's syntax [4]. To describe syntax, it is necessary to use formalized description of linguistic relations

between lexical units of the phrase in a theoretical-plural language [5]. Formal description of any text is put forward in [6], however, formal description of linguistic relations between lexical units is not updated. Formal text description is also contained in [7]. Formal text presentation is given in the reference-book [8] to automate the procedure of analysis of scientific and educational texts in order to identify semantically significant fragments [9]. The study [10] describes theoretical-plural description of linguistic relations in phrases. Such models can be used to describe images of author's vocabulary and author's syntax. However, they do not take into account statistical information about frequency of vocabulary and syntax [11]. Formalized description which was used to analyze text of a term vocabulary in order to construct a semantic network of its terms is outlined in the reference-book [12]. However, the proposed model does not provide for accounting of statistical information on frequency of vocabulary and syntax occurrence [13] alike.

Methods of attribution of Ukrainian scientific and technical text content were proposed and studied in [1–5]. To implement these methods, various algorithms [14], in particular quantitative one [15], can be used. Therefore, a problem arises in analysis of such algorithms in order to find the most effective of them [16].

Authorship identification is a technique for text attribution when it is questionable who wrote it [17]. It is useful when several people claim to be the authors of the same publication [18] or in cases where nobody claims to be the author of text content [19], for example, so-called trolls in social networks during information warfare [20]. Complexity of the problem of the author's text, obviously, is exponentially higher the greater the number of probable authors [21]. Availability of author's text samples is also significant in advancing this problem [22]. Text attribution includes the following three problems [23]:

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

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

– assessment of the possibility whether or not this text could be written by the author under consideration [26].

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

2. Literature review and problem statement

Text attribution is the text study in order to establish its author or obtain any information about the author and conditions of the textual document creation [17]. Attribution is divided into identification and diagnostic tasks [18]. Identification tasks make it possible to verify authorship [19]:

– confirm/exclude authorship of a certain person [20];

– check the fact that the author of the whole text is one person [21];

– check whether the author of the text simultaneously is its actual author [22].

Identification tasks are solved with the assumption that author of the text is known [23]. Diagnostic tasks make it possible to determine personal characteristics of the author (educational level, mother tongue, origin, knowledge of foreign languages, place of permanent residence, etc.) and/or the fact of conscious distortion of the written language [24]. Diagnostic tasks are solved with the assumption that the text author is unknown [25]. In these cases, it is usually impossible to compare the text under study with texts of another author [26]. Attribution methods enable study of the text at five levels [27]:

– punctuation (feature of the use of punctuation marks and characteristic errors) [28];

– spelling (characteristic errors in word spelling) [29];

– syntactic (features of constructing sentences, giving preference to one or another language structures, use of times, active or passive voice, word order, characteristic syntactic errors) [30];

– lexical-phraseological (author's vocabulary [31], peculiarities of use of words and expressions [32], tendency to use rare and foreign words, dialecticisms, archaisms, neologisms, professionalisms, argotisms [33], skill of using phraseologisms, proverbs, sayings, winged words, etc.) [34];

– stylistic (genre [35], general structure of the text [36], plot for literary works [37], characteristic pictorial tools (metaphor, irony, allegory, hyperbole, comparison) [38], stylistic figures (gradation, antithesis, rhetorical questions, etc.) [39], other linguistic techniques.

There are quite a lot of methods for style analysis [42]. In general, there are two large groups: expert and formal methods [43]. Expert methods involve the text study by a professional linguist [44]. Formal approaches include techniques from probability theory and mathematical statistics, algorithms of cluster analysis and neural networks [46]. The most complete classification of the main formal methods of text attribution is given, e.g. in studies [1 8, 47]. As can be seen, formal methods are most often based on comparison of computational characteristics of texts, as in the theory of image recognition [49]. Applying the theory of image recognition to the task of text attribution can be found, e. g. in [50]. In general, the text is displayed in a vector of parameters calculated for it, each of them objectively characterizing a certain set of text features [51]. Thus, text is graphically displayed to some point in the n-dimensional space [52]. With such formalization, the author is presented in the form of a similar vector of parameters. This vector is the vector of texts written by the author [53].

The distance between corresponding vectors is calculated as a criterion for proximity of two texts [54]. The sets of parameters and dispersion factors are presented as ordinary vectors in the n-dimensional Cartesian space from the origin of coordinates. Then the distance between the texts is the usual Cartesian distance between the ends of corresponding vectors. Such normal distance is an integral characteristic of difference between texts and the texts with a large distance between them belong with high probability to different authors. So, in order to compare authorship of two texts, it is enough to calculate parameters for them and determine distance [55]. To juxtapose the text with the author, vectors of the author parameters and the given text are compared, that is, actually two texts are compared again: the text with the known author (reference text) and the text whose authorship has to be established, confirm or refute (the analyzed/investigated text) [56]. Also, vectors of formal parameters that distinguish not concrete authors (or groups) but establish certain characteristics of authors (e. g. educational level) are constructed [57]. In most cases, statistical features are chosen as characteristic parameters of the text:

– amount of use of certain parts of speech, certain specific words, punctuation marks, phraseologisms, archaisms, rare and foreign words;

– number and length of sentences (measured in words, syllables, signs), mean sentence length;

– number of meaningful and auxiliary words;

– volume of vocabulary, ratio of the number of verbs to the total number of words in the text, etc. [58].

The main problem of formal methods of authorship analysis is precisely the choice of parameters and coefficients of talking [59]. There are a number of formal statistical characteristics of texts that are not suitable for attribution because of one of two shortcomings [1–5, 60]:

– lack of stability. The spread of parameter values in texts of the same author is so great that the ranges of possible values for different authors intersect. Obviously, this parameter will not help distinguish authors and when used as a part of the group of parameters, they only play the role of additional informational noise [61];

– lack of distinguishing ability. The parameter may accept close values for all or most authors since its value is determined by properties of the language in which the texts are written and not by individual features of the text author. Therefore, parameters must be investigated beforehand for stability and ability to distinguish, preferably in the texts of many different authors [62].

The following conditions of applicability of the formal talking coefficient of the author's style are determined in [3, 63–65]:

– mass character (the use of those characteristics of the text that are poorly controlled by the author at the conscious level in order to eliminate possibility of conscious distortion by the author of the style characteristic for him or imitation of the style of another author) [3, 63];

– stability (constant value for one participant is maintained but some deviation of values from the mean value should be rather small) [3, 64];

– ability to distinguish (takes substantially different values for different authors, that is, exceeds variations that are possible for one participant) [3, 65].

It is very hard to choose coefficients and talking parameters which assuredly distinguish any two authors [66]. Whatever the parameters, there is always a probability that two or more participants are close by virtue of accidental coincidence [67]. Therefore, it is sufficient in practice that the parameter allows us to confidently distinguish between different subsets of authors, that is, there would exist a sufficiently large number of subsets of authors for which mean values of the parameter differ significantly [68]. The parameter obviously will not help distinguish the authors' texts from one subset but will allow us to confidently distinguish between texts of authors who fall into different subsets [69]. Texts of authors of one subset can be distinguished by simultaneous use of a sufficiently large vector of parameters with different characters. In this case, probability of accidental coincidence will be noticeably less [70]. For a confident elucidation of texts for which formally calculated parametric distance is small, an additional study by expert methods, e.g. analysis of key and/or stop (auxiliary) words is necessary [70].

Consequently, it is necessary to conduct a study in this direction because of the lack of practical experiments to know the author's style for Ukrainian scientific and technical texts. To solve the task of plagiarism or copyright, many systems have already been developed recently. As for rewrite, it is quite difficult in Slavic languages to solve such a task in the presence of a large number of synonyms and the possibility of restructuring sentences with the use of other endings. This question does not apply to the use of auxiliary words as most people do not even pay attention to them when referring to plagiarism. Therefore, this induces to explore the problem of identifying the author's style to determine the degree of belonging of a particular text to a particular author.

3. The aim and objectives of the study

This study objective was to analyze quantitative algorithms for automatic attribution of the Ukrainian scientific and technical text content on the basis of stylistics and NLP methods.

To achieve the objective, the following tasks were formulated:

– develop a method for attribution of the text based on analysis of algorithms and coefficients of author's lexical talking in a reference text;

– develop a content analysis software for attribution of Ukrainian texts based on stylistic analysis of coefficients of talking of the text content;

– analyze results of experimental testing of the proposed method based on the content analysis for comparison of algorithms of automatic attribution of Ukrainian scientific and technical texts.

4. Method for determining a style of the text content

Several algorithms were taken as the basis of the developed method.

Algorithm I. Pre-processing of data based on content analysis (parsing, segmentation and tokenization of the text and linguistic analysis of the text).

Algorithm II. Calculation and analysis of talking parameters for each author (frequency of word use, number of punctuation marks, symbols, sentences, words and the ratio of the number of signs to the number of sentences).

Algorithm III. Calculation and analysis of talking coefficients for each author (lexical diversity, degree of syntactic complexity, talking coherency, uniqueness indexes and text concentration).

Algorithm IV. According to these factors, classify project participants (use of three classifiers as fuzzy, SVM and a combination of these two).

Algorithm V. Performance analysis to determine exactness of each classifier.

Algorithm VI. Identification of a subset of probable authors from the set of all investigated participants (algorithms VIII–XI) by superimposing the filters.

To achieve the study objective, a lexer-type system with the ability to select language/languages of the analyzed content implemented at the Victana web-site [16] has been developed. Lexer (tokenizer, segmentator) is the section of the text analyzer in a natural language. The lexer task was to define basic structural units in the text, lexems, and recognize by comparing with dictionary forms or other morphological samples. As a result of the lexer functioning, a complex data structure, namely the tokenization graph is

obtained. The tokenization graph is the source material for operation of the syntactic parser (Fig. 1). There are markers in the graph nodes. Each token stores information about location of the extracted word in the original text (index of the first character and the number of characters in the word), the word itself and the results of its identification. At the left, there is always a special token of identification of the sentence beginning. Each letter in the graph is a special token of sentence ending. Each path in the graph ends with a special token. For most cases, this token denotes the right boundary of the sentence. Thus, the parser has the ability to take into account proximity of words to the limits of expression which is useful for optimizing some rules of token filtering.

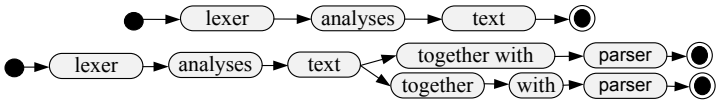


Fig. 1. Examples of tokenization graphs for a Ukrainian sentence

The lexer works in very close collaboration with the text parser. The words recognized by the lexer confirm/refute the parser's hypothesis on syntactic structure of the text. On the basis of the current context, parser puts forward new hypotheses that interrupt/continue certain paths of tokenization in the graph. Thus, tokens elongate during functioning of the parser rules and are immediately checked for compliance of conditions in these rules. For example, without taking in account syntactic rules, the chain of "*ямаладонькататамого*" allows a plurality of variants of breakdown into words (Fig. 2).

Some paths are interrupted because of inability to find an appropriate word in the vocabulary. Simple greedy algorithm does not work by the rule "identify from the input buffer the maximum long word found in the lexicon". The lexer appears in the text processing system [16] as a result of decomposition of the parsing problem. It simplifies implementation of morphological and syntactic analyzers (Fig. 3) since it allows one to work with larger units, lexemes. The simplicity introduced in this way implicitly limits commonality of the entire system since the idea of splitting the text into independent lexemes in itself does not work out with all languages. Moreover, even for languages with natural highlighting of words in writing, complex effects of merging words in larger units appear in audio representation. In Germanic languages, this is reflected in writing as articles and prepositions merged with other words.

Lexer and tokenizer work without explicit rules at all, they only use information in the lexicon. Only the task of specifying the type of word boundaries in the language and the list of separator symbols is more or less binding.

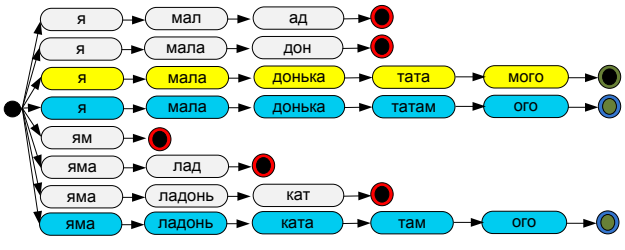


Fig. 2. An example of a tokenization graph with no taking into account syntactic rules

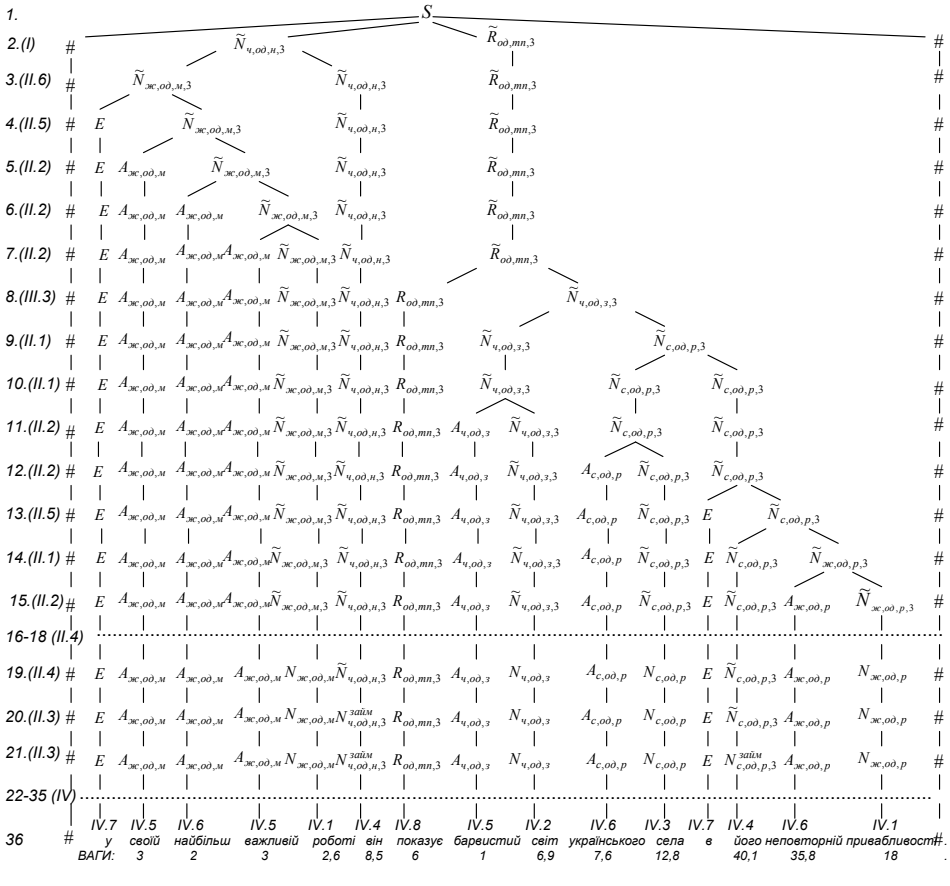


Fig. 3. The result of syntactic analysis of a Ukrainian sentence

Additional rules help solve some practical tasks increasing efficiency of the grammar slider. In particular, the rules make it possible to reduce ambiguity of word identification due to the partial removal of homonymy. Algorithms that are designed to solve the above tasks, allow different settings for the object language and the features of the processed texts and messages. Rules for each type of setting were created. Rules are written in text source files of the vocabulary according to the given specifications. The specifications are designed so that the rules can be easily edited in any plain text editor or generated by software, e.g. as a result of statistical processing of the language bodies. When translating these specifications, the vocabulary compiler formally checks correctness of the rules, optimizes and saves them in a special internal representation. Then, the slider loads compiled rules during the text parsing, usually without wasting time for syntax (algorithm VII), a compromise between convenience of writing rules and effectiveness of using the slider is achieved in this way.

Algorithm VII. Segmentation of the text content.

Step 1. Word recognition.

Step 2. Definition of the lexeme boundaries.

Step 3. Definition of complete word-forms.

Step 4. Identification of indivisible tokens having dots, spaces, etc.

Step 5. Splitting the text into sentences.

The characters that are delimiters of sentences (point, question, and exclamation marks) are determined by the appropriate parameter in the language description. Another parameter in the language description specifies maximum length of the sentence. It is used to prevent overflow of internal buffers and looping when parsing complex formatted texts when the algorithm cannot find the sentence end marker. If a point is used as a separator, then it is processed in a special way, unlike '?' and '!' signs. The fact is that some words may contain a point, and this should not cause the sentence break. A typical example is shortening of the "etc." type or abbreviations such as "N.Y.". Analogously, numbers with a decimal point like "9.3" are specially processed. Processing of such exceptions (tokens with a dot inside) relies on the tokenizer ability to recognize special chains with separators inside in a thread of characters.

A point after a complete word-form is considered an absolute divisor of the sentence. A list of special tokens is used to do this. If a complete word-form follows such a token, it starts with a capital letter and it is noted in the vocabulary lexicon that the entry does not begin with a capital letter, then a special token is the boundary of the sentence. For example, the first sentence in the text "Text, video, etc. Message, article, etc. "will be cut after "etc." since such word (Message) starts with a capital letter.

The values of minimum length of the full word-form are used in the case when point stands after the full word. Since the segmenter usually sees that characters stand in front and considers the event when the next word begins with a capital letter is the sentence boundary, the text "sun. sea. sand" will be considered one sentence. A corresponding rule forces the segmentator to check the word before the point according to the lexicon and in case of success, consider the point the sentence boundary regardless of the shift of the next word characters. The corresponding parameter enables avoiding of unnecessary checks for cases like "etc. characters". It sets minimum length of the analyzed complete word.

In addition to defining boundaries of lexemes, the lexer also preliminarily recognizes morphological attributes of words by transforming lexemes into tokens. To this end, the lexer uses information in the lexicon and rules of recognizing non-vocabulary tokens as well as a number of auxiliary algorithms, including fuzzy recognition. When recognizing a word, characteristics such as belonging to a certain part of speech and a set of grammatical attributes are defined. Noun phrases, $\tilde{N}$, and verbal phrases, $\check{R}$, are distinguished in the structure of Ukrainian sentences with direct word order (Fig. 4, 5) [1–5].

A user can only observe how the tree of constituents or syntactic structure of the analyzed sentence (Fig. 6) is obtained.

A vocabulary entry of a lexeme form is also defined for vocabulary lexemes. In alpha-frequency dictionaries, word characteristics are given after a slash (Fig. 7) where A is verb, the uppercase English letters are additional characteristics of the verb, V is adjective, the lowercase English letters indicate the noun characteristics.

The rules for reducing words to their stems are kept in the database (Fig. 9, *a*) where the 'flag' is the rule for word type identification (e.g. noun, singular), 'mask' is the word flexion (exclusion is given in square brackets), 'find' is the word flexions, subjective case, 'repl' is the word flexions in conjugation (Fig. 10).

In addition, the database (Fig. 9, *b*) contains a vocabulary of auxiliary words, that is, the words that are additional parameters for analysis of the author's talking style and their taking into consideration in analysis of texts has a significant effect on the final result.

I) $S \to \# \tilde{N}_{GD,NB,nm,PS} \tilde{R}_{NB,pr,PS} \# .$

II) $\tilde{N} = \{AN\}$ or $\tilde{N} = N^p$

1) $\tilde{N}_{GD,NB,CS,3} \to \tilde{N}_{GD,NB,CS,3} \tilde{N}_{GD',NB',gn,PS'};$

2) $\tilde{N}_{GD,NB,CS,3} \to A_{GD,NB,CS} \tilde{N}_{GD,NB,CS,3};$

3) $K_1 \tilde{N}_{GD,NB,CS,PS} K_2 \to K_1 \tilde{N}^{pron}_{GD,NB,CS,PS} K_2;$

4) $\tilde{N}_{GD,NB,CS,3} \to N_{GD,NB,CS};$

5) $\tilde{N}_{GD,NB,CS,3} \to E\tilde{N}_{GD,NB,CS,3};$

6) $\tilde{N}_{GD,NB,CS,3} \to \tilde{N}_{GD,NB,CS,3} \tilde{N}_{GD,NB,lc,3}.$

III) $\tilde{R} = R\tilde{N}$ or $\tilde{R} = \tilde{N}R$

1) $\tilde{R}_{NB,pr,PS} \to R_{NB,pr,PS} \tilde{N}_{GD',NB',ac,PS'} \tilde{N}_{GD*,NB*,ab,PS*};$

2) $\tilde{R}_{NB,pr,PS} \to R_{NB,pr,PS} \tilde{N}_{GD',NB',ab,PS'} \tilde{N}_{GD*,NB*,ac,PS*};$

3) $\tilde{R}_{NB,pr,PS} \to R_{NB,pr,PS} \tilde{N}_{GD',NB',ac,PS'};$

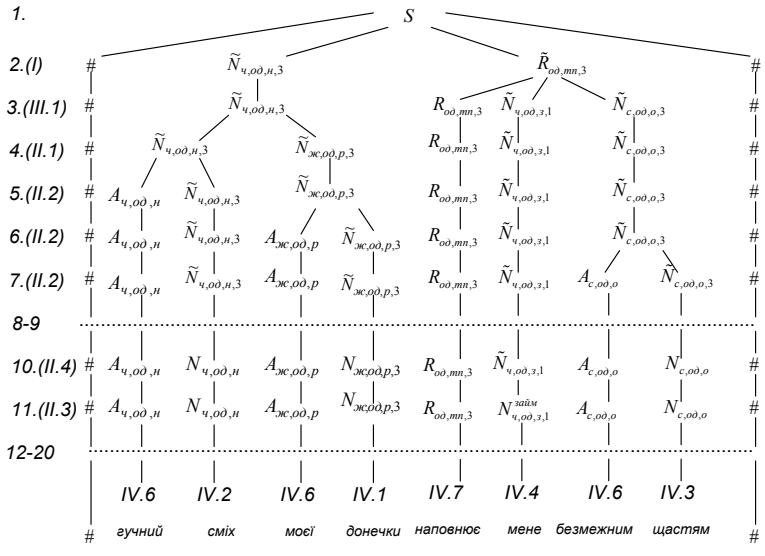
4) $\tilde{R}_{NB,pr,PS} \to R_{NB,pr,PS} \tilde{N}_{GD',NB',ab,PS'};$

5) $\tilde{R}_{NB,pr,PS} \to R_{NB,pr,PS} E\tilde{N}_{GD,NB,lc,3};$

6) $\tilde{R}_{NB,pr,PS} \to E\tilde{N}_{GD,NB,lc,3} R_{NB,pr,PS}.$

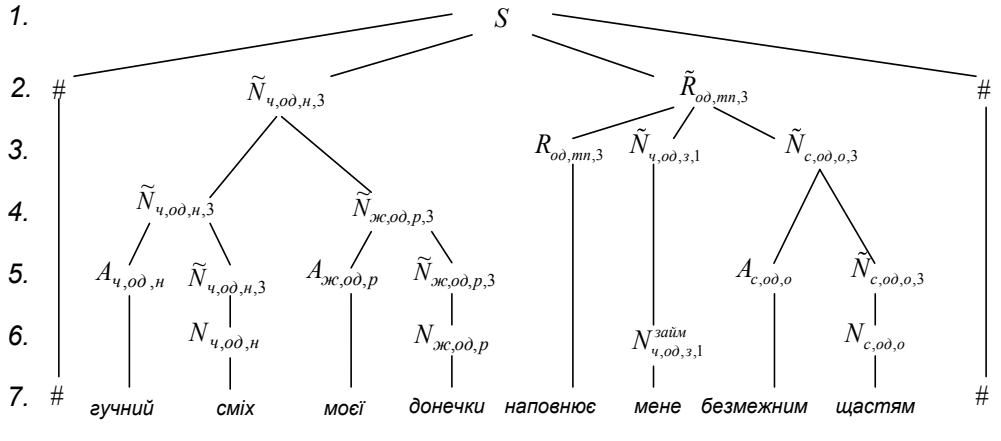
IV) $Words = \{x_1, x_2, x_3, \ldots, x_n\}$

Fig. 4. Rules for analysis of a Ukrainian sentence where *A* is adjective, *N* is noun, N^{pron} is pronoun; *NB* is number (*sn*, *pl*); *CS* is case (*nm*, *gn*, *dt*, *ac*, *ab*, *lc*, *vc*); *GD* is gender (*m, f, n*), *PS* is person (1, 2, 3); *TN* is time (*pr*, *ps*, *ft*)



1. S
2. (I) $\# \tilde{N}_{ч,од,н,3} \; \tilde{R}_{од,тп,3} \#$
3. (III.1) $\# \tilde{N}_{ч,од,н,3} \; R_{од,тп} \; \tilde{N}_{ч,од,з,1} \; \tilde{N}_{с,од,о,3} \#$
4. (II.1) $\# \tilde{N}_{ч,од,н,3} \; \tilde{N}_{ж,од,р,3} \; R_{од,тп} \; \tilde{N}_{ч,од,з,1} \; \tilde{N}_{с,од,о,3} \#$
5. (II.2) $\# A_{ч,од,н} \; \tilde{N}_{ч,од,н,3} \; \tilde{N}_{ж,од,р,3} \; R_{од,тп} \; \tilde{N}_{ч,од,з,1} \; \tilde{N}_{с,од,о,3} \#$
6. (II.2) $\# A_{ч,од,н} \; \tilde{N}_{ч,од,н,3} \; A_{ж,од,р} \; \tilde{N}_{ж,од,р,3} \; R_{од,тп} \; \tilde{N}_{ч,од,з,1} \; \tilde{N}_{с,од,о,3} \#$
7. (II.2) $\# A_{ч,од,н} \; \tilde{N}_{ч,од,н,3} \; A_{ж,од,р} \; \tilde{N}_{ж,од,р,3} \; R_{од,тп} \; \tilde{N}_{ч,од,з,1} \; A_{с,од,о} \; \tilde{N}_{с,од,о,3} \#$
8. (II.4) $\# A_{ч,од,н} \; N_{ч,од,н} \; A_{ж,од,р} \; \tilde{N}_{ж,од,р,3} \; R_{од,тп} \; \tilde{N}_{ч,од,з,1} \; A_{с,од,о} \; \tilde{N}_{с,од,о,3} \#$
9. (II.4) $\# A_{ч,од,н} \; N_{ч,од,н} \; A_{ж,од,р} \; N_{ж,од,р} \; R_{од,тп} \; \tilde{N}_{ч,од,з,1} \; A_{с,од,о} \; \tilde{N}_{с,од,о,3} \#$
10. (II.4) $\# A_{ч,од,н} \; N_{ч,од,н} \; A_{ж,од,р} \; N_{ж,од,р} \; R_{од,тп} \; \tilde{N}_{ч,од,з,1} \; A_{с,од,о} \; N_{с,од,о} \#$
11. (II.3) $\# A_{ч,од,н} \; N_{ч,од,н} \; A_{ж,од,р} \; N_{ж,од,р} \; R_{од,тп} \; N^{займ}_{ч,од,з,1} \; A_{с,од,о} \; N_{с,од,о} \#$
12. (IV.6) *#гучний* $N_{ч,од,н} \; A_{ж,од,р} \; N_{ж,од,р} \; R_{од,тп} \; N^{займ}_{ч,од,з,1} \; A_{с,од,о} \; N_{с,од,о} \#$
13. (IV.2) *#гучний сміх* $A_{ж,од,р} \; N_{ж,од,р} \; R_{од,тп} \; N^{займ}_{ч,од,з,1} \; A_{с,од,о} \; N_{с,од,о} \#$
14. (IV.6) *#гучний сміх моєї* $N_{ж,од,р} \; R_{од,тп} \; N^{займ}_{ч,од,з,1} \; A_{с,од,о} \; N_{с,од,о} \#$
15. (IV.1) *#гучний сміх моєї* $N_{ж,од,р} \; R_{од,тп} \; N^{займ}_{ч,од,з,1} \; A_{с,од,о} \; N_{с,од,о} \#$
16. (IV.7) *#гучний сміх моєї донечки* $R_{од,тп} \; N^{займ}_{ч,од,з,1} \; A_{с,од,о} \; N_{с,од,о} \#$
17. (IV.7) *#гучний сміх моєї донечки наповнює* $N^{займ}_{ч,од,з,1} \; A_{с,од,о} \; N_{с,од,о} \#$
18. (IV.4) *#гучний сміх моєї донечки наповнює мене* $A_{с,од,о} \; N_{с,од,о} \#$
19. (IV.6) *#гучний сміх моєї донечки наповнює мене безмежним* $N_{с,од,о} \#$
20. (IV.3) *#гучний сміх моєї донечки наповнює мене безмежним щастям#*

Fig. 5. An example of analysis of a Ukrainian sentence



S
(1) $N_{ж,од,н}$
(2) *гучний* $N_{ж,од,н}$
(4) *гучний сміх* $R_{од,з}$
(8) *гучний сміх наповнює* $N_{с,од,з}$
(3) *гучний сміх наповнює безмежним* $N_{с,од,з}$
(7) *гучний сміх наповнює безмежним щастям*.

Fig. 6. An example of syntactic structure of the analyzed sentence

Fragment 1	Fragment 2	Fragment 3
буферизувати/**ABGH**	клавіатурний/**V**	консоль/**ij**
відформатувати/**AB**	Кобол/**e**	конфігуратор/**efg**
декодувати/**ABGH**	кодек/**efg**	копілефт/**e**
кешувати/**ABGH**	кодер/**efg**	копірайт/**e**
кириличний/**V**	кодогенератор/**efg**	криптографічний/**V**
кілобайтовий/**V**	кодосумісний/**V**	криптозахищений/**V**
кілобайт/**efg**	комбосписок/**ab**	крос-асемблер/**efg**
кілобітовий/**V**	комутований/**V**	крос-компілятор/**efg**
кілобіт/**efg**	конкатенація/**ab**	кука/**ab**
кілобод/**efg**	консольний/**V**	курсорний/**V**

Fig. 7. An example of an alpha-frequency vocabulary

```
Файл  Правка  Вид  Справка
##################################################################
# Групи a b c d o
#
# -- Перша відміна: іменники жіночого та чоловічого та середнього роду
#
# -- Друга відміна: іменники чоловічого роду із закінченням на -ар -ир
#                   наголошені (Мішана група на -ар -ир)
#
# -- Друга відміна: іменники чоловічого роду з чергуванням -і -о
#
# -- Числівники -ять, -сят, -сто
#
#

SFX a Y 235

#
# ОДНИНА (множина перенесена в гр. b)
#
# Спочатку перша відміна
#
# тверда група в Називному відмінку однини з закінченням на -а
# однина
SFX a   а и [^жчшщ]а  # хата  хати  (Р.)
SFX a   а і [^гґкх]а  # хата  хаті  (Д.М.)
SFX a   а у а    # хата  хату  (З.)
SFX a   а ою [^жчшщ]а  # хата  хатою  (О.)
```

Fig. 8. Vocabulary of rules for morphological analysis of nouns

id	ordering	state	flag	type	lang	mask	find	repl
26	26	1	a	SFX	uk	ін	ін	оном
27	27	1	a	SFX	uk	ін	ін	оні
28	28	1	a	SFX	uk	ір	ір	огу
29	29	1	a	SFX	uk	ір	ір	огові
30	30	1	a	SFX	uk	ір	ір	огом
31	31	1	a	SFX	uk	ір	ір	озі
32	32	1	a	SFX	uk	[^л]ід	ід	оду
33	33	1	a	SFX	uk	[^л]ід	ід	одові
34	34	1	a	SFX	uk	[^л]ід	ід	одом
35	35	1	a	SFX	uk	[^л]ід	ід	оді
36	36	1	a	SFX	uk	[^пг]лід	ід	ьоду
37	37	1	a	SFX	uk	[^пг]лід	ід	ьодові
38	38	1	a	SFX	uk	[^пг]лід	ід	ьодом
39	39	1	a	SFX	uk	[^пг]лід	ід	ьоді
40	40	1	a	SFX	uk	[пг]лід	ід	оду
41	41	1	a	SFX	uk	[пг]лід	ід	одові
42	42	1	a	SFX	uk	[пг]лід	ід	одом
43	43	1	a	SFX	uk	[пг]лід	ід	оді
44	44	1	a	SFX	uk	іб	іб	обу
45	45	1	a	SFX	uk	іб	іб	обові
46	46	1	a	SFX	uk	іб	іб	обом
47	47	1	a	SFX	uk	іб	іб	обі
48	48	1	a	SFX	uk	in	in	onу

a

id	ordering	state	word	lang
1	1	1	після	uk
2	2	1	між	uk
3	3	1	are	en
4	4	1	and	en
5	5	7	між	uk
6	6	1	been	en
7	7	1	has	en
8	8	1	their	en
9	9	1	any	en
10	10	1	the	en
11	11	1	with	en
12	12	1	таких	uk
13	13	1	їхніми	uk
14	14	1	как	ru
15	15	1	такої	uk
16	16	1	на	uk
17	17	1	на	ru
18	18	1	ними	uk
20	19	1	для	uk
21	20	1	что	ru
22	21	1	или	ru
23	22	1	это	ru
24	23	1	этих	ru

b

Fig. 9. Rules of identifying word stems (*a*) and auxiliary words (*b*)

```
# The 1st line describes nouns ending in -in with vowel interchange -i -o
SFX a ін    ону   ін     # загін загону      (Д.Р.)
SFX a ін    онові ін     # загін загонові   (Д.)
SFX a ін    оном  ін     # загін загоном    (О.)
SFX a ін    оні   ін     # загін загоні     (М.)
# The 3rd line describes nouns ending in -iz with vowel interchange -i -o
SFX a ir    огу   ir     # батіг батогу      (Д.Р.)
SFX a ir    огові ir     # батіг батогові    (Д.М.)
SFX a ir    огом  ir     # батіг батогом     (О.)
SFX a ir    озі   ir     # батіг батозі      (М.)
# The 9th line describes nouns ending in -id with vowel interchange -i -o
SFX a ід    оду   [^л]ід # провід проводу (Д.Р.)
SFX a ід    одові [^л]ід # провід проводові  (Д.)
SFX a ід    одом  [^л]ід # провід проводом   (О.)
SFX a ід    оді   [^л]ід # провід проводі  (М.)
```

Fig. 10. An example of rules for identifying the word stems by analysis of flexions

5. Results obtained in studies of author's style attribution in Ukrainian scientific and technical texts

Let us analyze four algorithms designed to assess the optimal method we have developed for identifying author's style of a publication based on analysis of his composite papers.

Algorithm VIII. Filtering the set of author's styles analyzed

```
int i=0, j=0;
while (i<4){
int c1=0, c2=0, cc2=0;
while (j<94){
        int s=0;
        while (l<12){
        if ((K[i][l]+abs(F[l]-K[i][l]))>A[j][l]) &&
                ((K[i][l]-abs(F[l]-K[i][l]))< A[j][l])
        s+=1;
        if (l>6) &&
        ((K[i][l]+abs(F[l]-K[i][l]))>A[j][l]) &&
                ((K[i][l]-abs(F[l]-K[i][l]))< A[j][l])
        cc2+=s;
        l+=1;
        }
        A2[j]=s;
        A3[j]=cc2;
        c1+=s;
        c2+=s;
        j+=j;
        }
float t1=c1/94, t2=c2/94;
int filtr1=0, filtr2=0, filtr3=0
while (j<94){
        if(A2[j]>=t1) filtr1+=1;
        if(A3[j]>=t2) filtr2+=1;
        if (A2[j]>=t1)&&(A3[j]>=t2)filtr3+=1;
        j+=1;
i+=1;
}
```

Array K[i][l]: parameters and coefficients of style for 4 composite papers (lines 1–4 in Table 1 highlighted in yellow). Array A[j][l]: parameters and coefficients of style for all 94 authors participating in the project. Array F[l]: mean values of parameters and coefficients of style for all 94 authors. The algorithm determines whether values of parameters and coefficients of style talking of the *j*-th author fall within the limits $[x_i+x_{cep}; x_i-x_{cep}]$ of deviation of values of parameters and talking ratios of the certain composite paper style. Two are filled through filters under the arrays A2 (authors, the values of majority of parameters and coefficients are similar to the team style, *i*) and A3 (authors, the values of majority of coefficients only similar to the team style, *i*). Subsequently, a new subset of authors (whose styles are more similar to the collective style, i. e. the *i*-th work style) was created from the previously obtained subarrays by overlaying of a new filter.

The results obtained in style analysis of more than 200 individual scientific and technical papers written by 94 authors for the period of 2001–2017 are given in Table 1. For each author, mean arithmetic value of each coefficient and talking parameter was derived based on analysis of several his papers written during the above period. In addition, styles of 4 papers (Nos. 1–4 in Table 1 highlighted in yellow) of one team of authors were analyzed. Some of these authors are in Table 3 (Nos. 6 and 30). They are highlighted in blue in Table 1,

As a result, we obtained values given in Table 2 (algorithm VIII). Columns A contain the result of analysis of all values of vectors of coefficients and talking parameters for the authors from Table 1. Columns B contain the result of analysis of only the last 5 columns in Table 1. Unfortunately, this algorithm has provided such results: it is unlikely that the mentioned authors have written the papers by themselves (the best results are highlighted in red) and not enough to assert that they are actual authors of more than 50 % of these composite papers. On the other hand, although this algorithm yields good results: reduced number of authors at the first stage of attribution (up to 34.04 % of the total number of project participants). This is necessary for further filtering by means of analysis of stop words (prepositions and conjunctions) and keywords, features of semantics and vocabulary in construction of sentences, etc.

Table 1

Result of work of the algorithm for analyzing the author's publication style at Victana's information resource [16]

No.	N	W	W_1	W_{10}	P	Z	S	K_l	K_s	K_z	I_{wt}	I_{kt}
1	622	397	305	5	37	42	48	0.64	0.91	0.81	0.77	0.013
2	614	391	287	4	46	69	32	0.64	0.88	0.73	0.73	0.01
3	658	345	241	8	31	59	42	0.52	0.91	1.07	0.7	0.023
4	631.3	377.7	277.7	5.7	38	56.7	40.7	0.6	0.9	0.88	0.73	0.015
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
.....												
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
.....												
94	680	414	314	4	55	62	34	0.6	0.87	0.58	0.76	0.01

Table 2

Result of work of algorithms I–IV at Victana information resource [16]

Algorithm	Team	Mean value		Author				Filter			%
				6		30		1	2	3	
		A	B	A	B	A	B	A	B		
VIII	1	5.55319	2.3617	3	2	**6**	2	48	39	35	37.2
	2	7.361702	3.21277	6	3	6	3	40	37	25	26.6
	3	7.521277	3.925532	**8**	**5**	5	**5**	58	35	35	37.2
	4	4.148936	1.457447	3	2	3	0	41	43	33	35.1
	$\overline{x}_i$	6.15	2.4	5.0	3.0	5.0	2.5	46.8	38.5	32.0	34.0
IX	1	5.85106	2.75532	5	2	**8**	**3**	53	53	46	48.9
	2	5.6383	2.7234	**6**	**4**	4	**3**	53	56	43	45.7
	3	3.45745	1.04255	3	0	2	0	40	21	15	15.9
	4	6.2766	2.90426	6	**3**	5	2	44	54	41	43,6
	$\overline{x}_i$	5.31	2.36	5.0	2.3	4.8	2.0	47.5	46.0	36.3	38.6
X	1	6.44681	2.6383	**9**	**3**	**6**	**3**	46	55	42	44.7
	2	7.23404	3.39362	**8**	**4**	**8**	3	45	46	34	36.2
	3	6.46809	2.55319	**8**	**4**	**9**	**4**	48	46	39	41.5
	4	7.8516	3.54255	**9**	3	**9**	**5**	53	51	43	45.7
	$\overline{x}_i$	7.00	3.03	8.5	3.5	8.0	3.8	48.0	49.5	39.5	42.0
XI	1	6.31915	2.11702	**3**	2	**8**	**3**	45	35	29	30.9
	2	4.82979	2.14894	**6**	**3**	**6**	2	51	36	30	31.9
	3	5.89362	2.5	**8**	**4**	**9**	**4**	56	42	41	43.6
	4	5.53191	2.58511	**8**	**3**	**7**	2	49	53	43	45.7
	$\overline{x}_i$	5.64	2.34	6.3	3.0	7.5	2.8	50.3	41.5	35.8	38.0

Next, let us analyze the second algorithm. It does not differ significantly from the previous, just by condition in the third cycle:

if ((K[i][l]+V[l])>A[j][l]) && ((K[i][l]-V[l])<A[j][l]) s+=1

where V[l] is the array of mean absolute values of deviations of data points from the mean value. As a result, the values given in Table 2 (algorithm IX) were obtained. The results have improved a bit but not so much as to assert that the authors Nos. 6 and 30 are the actual authors of the composite papers 1–4 although they actually wrote them. On the other hand, the number of authors increased slightly (up to 38.56 % of the total number of project participants) with similarity in the style of talking. Now, let us analyze the algorithm X. Also, replace condition in the third cycle of algorithm 1 with the following:

if (abs(A[j][l]-K[i][l])>abs(K[i][l]-F[l])) s+=1

As a result, the values given in Table 2 (algorithm X) are obtained. As can be seen, the obtained values give firm grounds to assert that style of authors Nos. 6 and 30 is rather close (over 75–100 %) to the style of composite papers 1–4 accordingly (positive results are highlighted in red). Although the number of authors with similarity in the talking style increased significantly (up to 42.02 % of the total number of project participants). On the other hand, many authors who did not fall in the list at the previous stages of the study were found there at present and on the contrary, those who fell in the list at the previous two stages fell out of the present list. Next, let us try to reduce that total number by applying algorithm XI to the obtained initial data, namely parameters and talking coefficients of 94 participants of the project. Improve condition in the third cycle (by filtering) in algorithm X as follows:

if ((abs(A[j][l]- K[i][l])>abs(K[i][l]-F[l])) && (abs(A[j][l]-F[l])>abs(K[i][l]-F[l])))
‖ ((abs(A[j][l]-K[i][l])<abs(K[i][l]-F[l])) && (abs(A[j][l]-F[l])<abs(K[i][l]-F[l])))
s+=1

As a result, the values given in Table 2 (algorithm XI) are obtained. The obtained values also confirm that the style of authors Nos. 6 and 30 is sufficiently close (over 75–100 %) to the style of composite papers 1–4 accordingly (positive results are highlighted in red). Also, the number of authors (up to 38.03 % of the total number of the project participants) with similarity in the talking style has significantly reduced.

6. Discussion of results obtained in the study of author's style in Ukrainian scientific and technical texts

Detailed graphs of the results obtained in using algorithms VIII–XI (Nos. 1–4, respectively) for analysis of our method of style attribution are given in Fig. 11. At the next stage, analysis of stop words (prepositions and conjunctions) and key words in papers of the authors who fell to those 38.03 % was used to attribute author's style (Fig. 12). Each individual uses its own special vocabulary to convey its thoughts, including the so-called filler words (e.g. "that is", "therefore", "though", etc.) and auxiliary words ("and", "but", "at least").

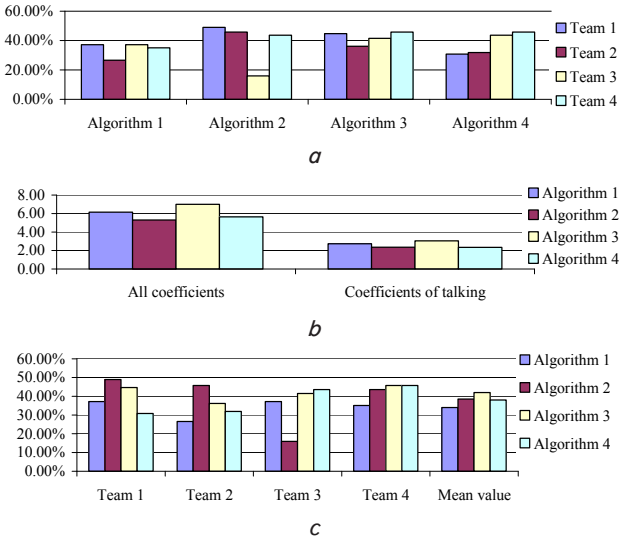


Fig. 11. Detailed analysis of the process of attribution of the author's style: according to the developed algorithms (*a*); with taking into account all parameters and only coefficients of talking (*b*); for the analyzed composite papers (*c*)

Fig. 12. Detailed analysis of the process of attribution of the author's style at the second stage for a full text with compilation of a frequency vocabulary containing 100 words (*a*); a main text with compilation of a frequency vocabulary containing 100 words (*b*); a complete text with compilation of a frequency vocabulary containing 200 words (*c*); a main text with compilation of a frequency vocabulary containing 200 words (*d*); a complete text with compilation of a frequency vocabulary containing 50 words (*e*); a main text with compilation of a frequency vocabulary containing 50 words (*f*)

An example of analysis of author's style at the second stage through analysis of frequency of appearance of auxil-

iary and key words with taking into account various filters, analysis of full texts with references and abstracts in various languages and analysis of only informative part of the publication, i. e. the main text, with compilation of frequency vocabularies respectively containing 200, 10 and 50 words is given in Fig. 12.

However, it should be noted that the number of texts sampled for analysis (over 200) and the number of authors (94) are small to guarantee exact results. The study should be continued with more texts (it should be noticed that they are not always available). In the future, it is also necessary to improve the method by analyzing the texts using methods of stylemetry and glotochronology.

6. Conclusions

1. A method for attribution of texts based on analysis of coefficients of lexical author's talking in a reference excerpt of the author's text was developed. Establishment of the author's style is based on a comparative analysis of coefficients of lexical author's talking: speech connectivity, lexical diversity, syntactic complexity, concentration and exclusiveness indexes for the author's excerpt and other analyzed excerpt for further comparison and determination of the degree of belonging of the analyzed text to a particular author. The main stylistic coefficients for the author's excerpt and other analyzed excerpt include speech connectivity, lexical diversity, syntactic complexity, as well as concentration and exclusivity indexes. Further analysis is needed to compare values of the coefficients and determine the degree of attribution of the analyzed text to a particular author. The developed method features adaptation of the morphological and syntactic analysis of lexical units to peculiarities of Ukrainian word/text structure. That is, analysis of linguistic units of the word type took into account their belonging to a part of speech and conjugation within this part of speech. To this end, we analyzed flexions of these words for classification and extraction of word stems for compilation of corresponding alphabetic and frequency dictionaries. Supplementation of these dictionaries was further taken into account in subsequent steps of text attribution as calculation of parameters and coefficients of the author's talking. Namely auxiliary (stop or reference) words are indicative for the individual writer style because they are in no way related to the topic and content of the publication. An algorithm of definition of stop words of the text content on the basis of linguistic analysis of the text content was designed. It features adaptation of the morphological and syntactic analysis of lexical units to peculiarities of structure of Ukrainian words/texts. Theoretical and experimental substantiation of the method of content monitoring and definition of stop words of Ukrainian texts was made. The method is aimed at automatic detection of significant stop words in a Ukrainian text by means of the proposed formal approach to implementation of parsing of scientific and technical text content.

2. A formal approach to attribution of Ukrainian texts was proposed. The study was conducted with Ukrainian scientific and technical texts. Decomposition of the author's method of attribution based on analysis of such talking coefficients as lexical diversity, degree of syntactic complexity, talking connectivity, indexes of text exclusiveness and concentration was made. Parallel analysis of author's style parameters such as the number of words in a particular text, the total number of words in this text, the number of sentences, the number of prepositions and conjunctions, the number of words with one occurrence and the number of words with 10 or more occurrences. The developed system has analyzed over 200 individual scientific publications from all issues of Lviv Politechnic National University Bulletin, Information Systems and Networks series, for the period from 2001 to 2017.

3. The results of application of the designed algorithms of automatic attribution of the text content on the basis of NLP and stylemetry methods were analyzed. Prospects and peculiarities of application of information stylemetry methods for attribution of text contents were considered. Quantitative analysis of scientific and technical text contents uses benefits of content monitoring and content analysis of texts based on NLP, Web-Mining and stylemetry methods to determine the number of authors whose talking styles are similar to that of the text fragment being studied. This has narrowed the search circle for later use in stylemetry methods to determine the degree of belonging of the analyzed text to a particular author. Comparison of the results obtained with 200 individual technical papers written by about 100 different authors during the period from 2001 to 2017 has been made to determine whether the coefficients of diversity of the text of these authors varied at different time intervals. Experimental results of the proposed approach were obtained to determine belonging of the analyzed text to a particular author in the presence of a reference information stream of author's text content. Absence of analysis of introduction and conclusion sections somewhat improved results as the main section usually discloses its style when describing the study essence. This is achieved through training the system and checking the clarified blocked words and due to the refined idioglossary.

References

1. 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
2. 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
3. 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
4. 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: ttps://doi.org/10.15587/1729-4061.2016.77174

5. 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
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. Vol. 512. Springer, 2016. P. 149–163. doi: https://doi.org/10.1007/978-3-319-45991-2_10
8. Buk S. Osnovy statystychnoi linhvistyky. Lviv, 2008. 124 p.
9. 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.
10. Anisimov A., Marchenko A. Sistema obrabotki tekstov na estestvennom yazyke // Iskusstvenniy intellekt. 2002. Issue 4. P. 157–163.
11. Perebyinis V. Matematychna linhvistyka. Ukrainska mova. Kyiv, 2000. P. 287–302.
12. Perebyinis V. Statystychni metody dlia linhvistiv. Vinnytsia, 2013. 176 p.
13. Braslavskiy P. I. Intellektual'nye informacionnye sistemy. URL: http://www.kansas.ru/ai2006/
14. Lande D., Zhyhalo V. Pidkhid do rishennia problem poshuku dvomovnoho plahiatu // Problemy informatyzatsii ta upravlinnia. 2008. Issue 2 (24). P. 125–129.
15. Varfolomeev A. Psihosemantika slova i lingvostatistika teksta. Kaliningrad, 2000. 37 p.
16. Victana. URL: http://victana.lviv.ua/nlp/linhvometriia
17. 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
18. Kognitivnaya stilometriya: k postanovke problemy. URL: http://www.manekin.narod.ru/hist/styl.htm
19. Kocherhan M. Vstup do movoznavstva. Kyiv, 2005. 368 p.
20. 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
21. Rodionova E. Metody atribucii hudozhestvennyh tekstov // Strukturnaya i prikladnaya lingvistika. 2008. Issue 7. P. 118–127.
22. Meshcheryakov R. V., Vasyukov N. S. Modeli opredeleniya avtorstva teksta // Izmereniya, avtomatizaciya i modelirovanie v promyshlennosti i nauchnyh issledovaniyah. 2005. P. 25–29.
23. Morozov N. A. Lingvisticheskie spektry. URL: http://www.textology.ru/library/book.aspx?bookId=1&textId=3
24. Mobasher B. Data mining for web personalization // The adaptive web. 2007. P. 90–135. doi: https://doi.org/10.1007/978-3-540-72079-9_3
25. Dinucă C. E., Ciobanu D. Web Content Mining // Annals of the University of Petroşani. Economics. 2012. Vol. 12, Issue 1. P. 85–92.
26. Xu G., Zhang Y., Li L. Web content mining // Web Mining and Social Networking. 2011. P. 71–87. doi: https://doi.org/10.1007/978-1-4419-7735-9_4
27. 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. Vol. 689. Springer, 2017. P. 204–216. doi: https://doi.org/10.1007/978-3-319-70581-1_14
28. 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
29. 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
30. 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
31. 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
32. The Contextual Search Method Based on Domain Thesaurus / Lytvyn V., Vysotska V., Burov Y., Veres O., Rishnyak I. // Advances in Intelligent Systems and Computing. Vol. 689. Springer, 2017. P. 310–319. doi: https://doi.org/10.1007/978-3-319-70581-1_22
33. Marchenko O. Modeliuvannia semantychnoho kontekstu pry analizi tekstiv na pryrodniy movi // Visnyk Kyivskoho universytetu. 2006. Issue 3. P. 230–235.
34. Jivani A. G. A Comparative Study of Stemming Algorithms // Int. J. Comp. Tech. Appl. 2011. Vol. 2, Issue 6. P. 1930–1938.

35. 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. Vol. 528. Springer, 2015. P. 639–644. doi: https://doi.org/10.1007/978-3-319-21380-4_108
36. Rodionova E. Metody atribucii hudozhestvennyh tekstov // Strukturnaya i prikladnaya lingvistika. 2008. Issue 7. P. 118–127.
37. Bubleinyk L. Osoblyvosti khudozhnoho movlennia. Lutsk, 2000. 179 p.
38. Kowalska K., Cai D., Wade S. Sentiment Analysis of Polish Texts // International Journal of Computer and Communication Engineering. 2012. Vol. 1, Issue 1. P. 39–42. doi: https://doi.org/10.7763/ijcce.2012.v1.12
39. 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.
40. 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
41. Learning-based image scaling using neural-like structure of geometric transformation paradigm / Tkachenko R., Tkachenko P., Izonin I., Tsymbal Y. // Studies in Computational Intelligence. Vol. 730. Springer, 2018. P. 537–565. doi: https://doi.org/10.1007/978-3-319-63754-9_25
42. 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
43. 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
44. 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
45. 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
46. 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
47. 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
48. 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
49. 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. Vol. 9, Issue 2. P. 42–48. doi: https://doi.org/10.5815/ijisa.2017.02.06
50. 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
51. 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.
52. 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.
53. Riznyk V. Multi-modular Optimum Coding Systems Based on Remarkable Geometric Properties of Space // Advances in Intelligent Systems and Computing. Vol. 512. Springer, 2017. P. 129–148. doi: https://doi.org/10.1007/978-3-319-45991-2_9
54. 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
55. 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
56. Pasichnyk V., Shestakevych T. The model of data analysis of the psychophysiological survey results // Advances in Intelligent Systems and Computing. Vol. 512. Springer, 2017. P. 271–281. doi: https://doi.org/10.1007/978-3-319-45991-2_18

57. Zhezhnych P., Markiv O. Linguistic Comparison Quality Evaluation of Web-Site Content with Tourism Documentation Objects // Advances in Intelligent Systems and Computing. Vol. 689. Springer, 2018. P. 656–667. doi: https://doi.org/10.1007/978-3-319-70581-1_45
58. 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.
59. 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
60. Davydov M., Lozynska O. Mathematical Method of Translation into Ukrainian Sign Language Based on Ontologies // Advances in Intelligent Systems and Computing. Vol. 689. Springer, 2018. P. 89–100. doi: https://doi.org/10.1007/978-3-319-70581-1_7
61. 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
62. 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
63. Mykich K., Burov Y. Algebraic Framework for Knowledge Processing in Systems with Situational Awareness // Advances in Intelligent Systems and Computing. Vol. 512. Springer, 2017. P. 217–227. doi: https://doi.org/10.1007/978-3-319-45991-2_14
64. 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
65. 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). 2016. doi: https://doi.org/10.1109/stc-csit.2016.7589896
66. 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.
67. 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. Vol. 689. Springer, 2018. P. 421–433. doi: https://doi.org/10.1007/978-3-319-70581-1_30
68. 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
69. Kravets P. Game Model of Dragonfly Animat Self-Learning // Perspective Technologies and Methods in MEMS Design (MEMSTECH 2016): Proc. of XII-th Int. Conf. Lviv: Lviv Politechnic Publishing House, 2016. P. 195–201.
70. Vysotska V., Fernandes V. B., Emmerich M. Web content support method in electronic business systems // Proceedings of the 2nd International Conference on Computational Linguistics and Intelligent Systems. Vol. I: Main Conference. Lviv, 2018. P. 20–41. URL: http://ceur-ws.org/Vol-2136/10000020.pdf