Розглянуто особливості застосування технологій лінгвометрії, стилеметрії та глоттохронології для визначення стилю автора публікацій. Лінгвостатистичний аналіз авторського тексту використовує переваги контент-моніторінгу на основі методів NLP для визначення стопових слів. Квантитативний аналіз стопових слів використано для визначення степеня приналежності аналізованого тексту конкретному авторові. Запропоновано метод визначення стилю автора україномовного тексту

Ключові слова: стиль автора, статистичний лінгвістичний аналіз, квантитативна лінгвістика, авторська атрибуція

Рассмотрены особенности применения технологий лингвометрии, стилеметрии и глоттохронологии для определения стиля автора публикаций. Лингвостатистический анализ авторского текста использует преимущества контент-мониторинга на основе методов NLP для определения стоповых слов. Квантитативный анализ стоповых слов использован для определения степени принадлежности анализируемого текста конкретному автору. Предложен метод определения стиля автора украиноязычного текста

Ключевые слова: стиль автора, статистический лингвистический анализ, квантытативная лингвистика, авторская атрибуция

UDC 004.89
DOI: 10.15587/1729-4061.2017.107512

DEVELOPMENT OF A METHOD FOR THE RECOGNITION OF AUTHOR'S STYLE IN THE UKRAINIAN LANGUAGE TEXTS BASED ON LINGUOMETRY, STYLEMETRY AND GLOTTOCHRONOLOGY

V. Lytvyn
Doctor of Technical Sciences, Professor*
E-mail: yevhen.v.burov@lpnu.ua
V. Vysotska
PhD, Associate Professor*
E-mail: victoria.a.vysotska@lpnu.ua
P. Pukach
Doctor of Technical Sciences, Associate Professor**
E-mail: petro.y.pukach@lpnu.ua
I. Bobyk
PhD, Associate Professor**
E-mail: igor.bobyk@gmail.com
D. Uhryn
PhD, Associate Professor
Department of Information Systems
Chernivtsi Faculty of National Technical University
«Kharkiv Polytechnic Institute»
Holovna str., 203-a, Chernivtsi, Ukraine, 58000
E-mail: ugrund38@gmail.com
*Department of Information Systems and Networks***
Department of Mathematics*
***Lviv Polytechnic National University
S. Bandery str., 12, Lviv, Ukraine, 79013

1. Introduction

The impetus of research into statistical linguistic (quantitative linguistics) was the emergence and active development of information technologies (IT) in the area of NLP and Web Mining [1]. In the early 1960s, at the Institute of Linguistics named after O. Potebnya of the Academy of Science of the USSR, a group of structural and mathematical linguistics was organized [2]. It began a straightforward statistical research into Ukrainian texts of belles-lettres, scientific-technical and socio-political functional styles. This made it possible to reveal their statistical parameters. It was at that time that the project on compiling a series of frequency dictionaries started: belles-lettres prose, drama, poetry, journalism, scientific prose, in which the laboratory of computer linguistics of Taras Shevchenko National University of Kyiv (Ukraine) was also involved [3]. The major trend of applied statistical linguistics and sciences, related to it, is development of methods and technologies for determining the statistical structure of a text for solving problems, in particular, of linguometry [4], stylemetry [5], and glottochronology [6]. These problems include, for example, automation of lexicographic processes, comparison of dictionaries, creation of shorthand systems, and automatic recognition of a language [7]. To recognize the author's style, the linguistically statistical problems are used:

– automatic language recognition;

– calculation and analysis of coefficients of lexical author's language;

– determining of a degree of plagiarism;

– identification of the author of a text or a text itself;

– analysis of authorship phenomenon and dynamics of changes in the author's style;

– determining and analysis of a degree of the author's attribution [8].

Essential tasks of linguistics include creation and comparison of dictionaries with the use of linguometry (including frequency and statistic dictionaries), creation of automatic dictionaries, thesauruses, creation of shorthand systems, automatic language recognition, information search, etc. For modeling of some processes of content monitoring and content analysis, statistical and transition probabilities of the morphemes of a text are found. Based on the constructed tables, the proofreading of an explored word is modeled and some of the most probable options are proposed.

The purpose of stylemetry is typology, attribution (author's, temporal, style for using, for example, in judicial and criminal linguistics), diagnostics, reconstruction, of texts and their parts, etc. An example of solving a linguistic problem is the process of the author's attribution of text fragments. For this purpose, word usage frequencies in the analyzed text are calculated. With the use of frequency dictionaries of literary activity of writers in general or of their separate works, it is possible to recognize the author of a piece of literature (or a piece of literature – if a dictionary allows it).

Glottochronology explores the rate of language changes and on this basis determines the time of separation of related languages and a degree of closeness between them. The dating method, which is used to determine duration of the period when two closely related languages existed separately, is based on the assumption that the bulk of the lexical structure of any language (nuclear lexis) changes at the same rate and requires calculation of percentage of shared elements in their basic vocabulary.

Each language has its own statistical parameters, and knowledge of the frequency of occurrences of letters and their combinations (bigrams, trigrams, and four-grams) of a certain language enables us to identify it automatically. For example, for Ukrainian texts, it was found that statistical parameters of styles include frequencies of vowels, consonants, spaces between words, as well as palatelized and resonant groups of consonants.

2. Literature review and problem statement

For automatic recognition of a language, formatted fragments of a text are analyzed: the letters, arranged by decreasing of frequency of their occurrence in the fragment (frequencies are given); small and capital letters are not distinguished. It is possible to analyze the data and recognize the author's language of formatted fragments with the use of three methods through research of [9]:

1) frequencies of vowels and consonants in a text;

2) resonant, voiced and voiceless consonants and their assessments;

3) frequency of usage of the letters of a language.

To explore the special features of the author's style, coefficients of the lexical author's language are determined and analyzed. They include coherence of speech, lexical diversity, syntactical complexity, indices of concentration and exclusivity for the author's fragment and another analyzed fragment. Subsequently, the internal "dynamics" of a text through analysis of these coefficients is explored and a degree of belonging of this text to a particular author is determined [10].

To determine a degree of plagiarism, a summary group table is constructed. There, we enter calculated group mean values of speech coherence, lexical diversity and syntactic complexity, as well as indices of concentration and exclusivity for sets of texts, similar by content [11]. The area of standard deviations is calculated and thus, lexical similarity of each analyzed text in comparison with a reference fragment is assessed [12].

Recognition of the author of a text or identification of a text is conducted according to results of analysis of its formatted fragment [13]. Word usages are arranged in descending order of frequency of their occurrence in the fragment. The type of the language, to which the word usage belongs (author's or not author's language), is specified. Proper names are deleted from the text of a fragment. Based on frequency dictionaries, if possible, the author of the passage or the passage itself is recognized [14]. Analysis of the authorship phenomenon lies in determining of differences between the styles of writers [15]. This makes the author's language dynamic, exciting, easy to understand, determines, which characteristics are individual, and which may be regarded as shared [16]. A degree of the author's attribution is analyzed: reliability, authenticity of the literary piece, its author, the place and the time of its creation based on stylistic and technological features [17].

Dynamics of change of the author's style is also analyzed. From the literary heritage of the authors of the works, written in one language and belonging to the same period of time, the couples of theme works are chosen, each following couple is chosen with the step of h years [6]. For each set of works, it is necessary to process 1000 word usages from every set and find out how many of these words belong to the 100-word Swadesh list. It is a tool for assessment of a degree of closeness between different languages/speeches by such quality as similarity of the most set basic dictionary; it is enumeration of basic lexemes of a specific language/speech that is sorted by order of decreasing of their "being basic". Minimum set of the most essential ("nuclear") lexis is contained in the Swadesh 100-word list. 200- and 207-word lists are used as well. Comparison of results, obtained within a group, allows us to reveal a tendency to an increase (a decrease) in the number of shared words from the Swadesh list in the works of these authors. It also determines their divergence in order to determine authorship in joint journalistic scientific works [6].

The problem of establishing of authorship of anonymous and pseudo-anonymous texts is associated with both historical-philological and natural-technical sciences, among which statistics and theory of probability are becoming increasingly essential for solving this problem. Moreover, the problem setting and use of results are related to literary studies, and the apparatus and methods for obtaining a result – to the mathematical field that requires the use of modern scientific theories and computational tools [18].

For description of an individual style, linguo-mathematical methods are used, which contributes to accumulation of data about properties of the language units and formation of a special scientific apparatus of texts attribution. With its help, stylemetry takes part in solution of the main practical problems of four groups [3, 19].

1. *Research into publications or historical facts.* It is just worth recalling "Shakespeare issue", which is still the point of argument for scientists throughout the world, beginning with 1785, since Rev. James Wilmot expressed the assumption that the real author of Shakespeare's plays was Francis Bacon. Researchers also claim that not all the works, attributed to Moliere, belong to him; the authorship of "And Quiet Flows the Don" is disputable, besides, there are a number of anonymous works with unknown or disputable authorship – the procedures of the author's attribution can help solve these issues as well. From a historical standpoint, it is necessary to link various archival documents with the author and the period when they were written. Only in this case, can you make conclusions based on the content of historical texts.

2. Area of education, science and psychology. With development of the Internet, researchers less and less work independently, using finished works or fragments from them. It is not seldom that quoted passages of a text exceed contribution of an author and often do not contain reference to the original source. Using the methods of authorship determining, it is possible to reveal a similar plagiarism, thus taking control and assessing a paper properly [11, 20]. Similarly, it applies to scientific papers, not only in determining of text uniqueness coefficients (copyright and rewrite), but also percentage of the author's contribution to joint papers of a team of authors.

3. *Judicial practice.* The objects of research are the issues of copyright and plagiarism, written evidence of witnesses or evidence, made under pressure, as well as agreements, wills, anonymous letters, etc. One of the most modern directions of the author's attribution is identification of creators of computer viruses. Relevant in the search for the author's attribution of texts is, for example, a study of preservation of the author's style in the translations of texts [3, 21].

4. *Cybersecurity.* Recognition of the author's style with the rapid development of IT and activeness of the Internet users is quite important for identification of fraudsters through their history in social networks. This not only helps find offenders, but also may contribute to prevention of crimes (for example, activity of ill-intended organization "Blue cat" or activity in the nets of so-called trolls in the notorious information war between Slavic states).

3. The goal and objectives of research

The goal of the present research is to develop a formal approach to recognition of the author's style in the Ukrainian texts based on technology of statistical linguistics.

To accomplish the set goal, the following tasks were formulated:

– to develop the method for recognition of the style of the author of a text based on analysis of coefficients of lexical author's language in the reference fragment of the text by this author;

– to develop a formal approach to designing software for content monitoring for determining of the style of the author in Ukrainian texts based on Web Mining and lexical analysis of determined stop words in the text content;

– to obtain and analyze results of experimental testing of the proposed method of content monitoring for recognition of the style of an author in scientific texts of technical profile in Ukrainian.

4. Method of determining style of the text content's author

Linguo-statistic fundamentals for the implementation of study for the purpose of text attribution include [3, 18–24]:

1) preliminary processing of linguistic data (construction of distribution series, calculation of statistics, statistical evaluations and other parameters of linguometry);

2) lexicographical processing of text data (creation of frequency and alphabetical-frequency dictionaries, dictionaries-concordances, word indices, reverse dictionaries, glossaries of keywords of a writer's style, etc.

Application of procedures of linguometry for statistical description of a text allows us to perform research relating to the authorship phenomenon [25]. The method of analysis and interpretation of stylistic peculiarities and patterns of writing style of a certain author (or of a specific literary epoch) at the linguistic level uses algorithm 1.

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

Stage 1. *Selection of texts.* The way of organization of selection and the volume of the text sample are important: in order to determine characteristics, it should include at least 18 thousand words [23–25].

Stage 2. *Lemmatization of text units.* Incorporation of word forms in a language lemma [5].

Stage 3. *Elimination of inhomogeneity of text units.* Solution of the problem of inhomogeneity of text units, for example, from the standpoint of their relation to the various types of a language (author's or not author's, etc.).

Stage 4. *Construction of a system, organization on this basis of statistic spread in the required frequency dictionary scales.* A frequency dictionary is the type of dictionary, which gives the number of usages (frequency) of a particular language unit (combination, word, word form, idiom, phraseological unit) in various texts of a certain volume. Usually, absolute and reference frequency of usage of language units is given, dictionary articles are arranged in order of decreasing of frequency [3].

Stage 5. *Search for parameters that adequately reflect the structure of the frequency dictionary.* The number of parameters is varied, for example, to describe the French texts of the XVII century, 51 parameters were proposed [25]. The parameters, found in papers [26-31], allow us to formulate some basic linguo-statistical methods of text research:

– anchor words method (calculation of total frequency of usage and finding percentage composition of syntactic words [18–22]: prepositions, conjunctions, particles);

– punctuation marks method (calculation only of the number of internal and external punctuation marks);

– words method (calculation only of words of a certain length);

– sentences method (calculation only of sentences of a certain length);

– syntactical method (calculation of punctuation marks, words and sentences of a certain length);

– combined (combination of anchor words method and syntactical method).

Stage 6. *Checking effectiveness of parameters.* The use of general methods of checking effectiveness of selected parameters

Stage 7. *Mathematical modeling of lexical-statistical distributions.* The use of general methods of mathematic apparatus of modeling of lexical-statistical distributions.

Stage 8. *Construction of statistical classifications* (author's reference fragment), which reflect stylistic patterns within the works of a certain author or a certain epoch (or a sequence of literary epochs).

Stage 9. *Interpretation of obtained results* from the standpoints of historical and literary ideas, general and historical stylistics.

Using algorithm 1, we can solve the problem of the author's attribution, which can be formulated, for example, in the following way. Let assume there is a statistically processed works, created by an author (reference fragment). It is necessary to estimate belonging of certain fragments to the reference fragment with the use of appropriate methods. To illustrate this, consider creative work of Author I and his publications from [24]. In this case, we will assume that the author's reference fragment has already been built – problems of texts selection, lemmatization and problems of inhomogeneity have been solved, the processed material has been formed as a frequency dictionary [3]. For attribution, we will use the method of anchor words, results will be shown in the form of correlation coefficients and graphically. Separately, we will mention the evolution of significance of one of the text parameters – syntactic words – in the author's attribution of texts (Table 1).

Table 1

Syntactic parts of the Ukrainian language (stop words)

Part of speech	List of stop words
Prepositions	в, на, з, за, до, по, у, біля, від, для, без, про, через, при, над, з-за, з-під, під, близько, вглиб, крізь, поза, проміж
Conjunctions	і, й, що, так, хоча, коли, або, щоб, якщо, також, чи, тобто, проте, немов, а, але, та, через те що, одначе, та й
Particles	не, так, ж, же, навіть, би, або, лише, то, ні, адже, он, тобто, уже, чи, аякже, це, тільки, ось, ледве чи, мов, немов

For the individual style of an author, it is syntactic words that are significant, as they are not related to the theme and the content of a book [3]. We will consider the specified parameter of text research to be effective and accept a list of stop words (syntactic words) [25], presented in Table 1 (71 words in total).

5. Results of research into the author's style in the Ukrainian texts based on technology of statistical linguistics

100 scientific publications from two issues (783 and 805) of the Visnyk of the National University "Lviv Polytechnika" from a series "Information systems and networks" were analized. Consider four arbitrary fragments from analyzed texts, formatted with respect to the choice of the method for attribution: from each fragment, we selected only prepositions, conjunctions and particles. The total number of word usage in the passage is given, proper names are not included. Table 2 for each of the fragments specifies absolute frequency (AF) and relative frequency (RF) of occurrence of a syntactic word for each fragment, as well as relative frequency of occurrence of a specified word in the reference fragment.

Fig. 1 shows graphic representation of relative frequency of occurrence of stop words in Fragment 1 and in the reference fragment. Correlation coefficient for syntactic words in this case makes up $R_{r\text{-}F1}$=0.6076. Graphic representation of relative frequency of occurrence of syntactic words in Fragment 2 and in the reference fragment is shown in Fig. 2. Correlation coefficient for syntactic words in this case is $R_{r\text{-}F2}$=0.7066.

Graphic representation of relative frequency of occurrence of syntactic words in Fragment 3 and in the reference fragment is shown in Fig. 3. Correlation coefficient for syntactic words in this case is $R_{r\text{-}F3}$=0.2810.

Table 2

Absolute and relative frequencies of occurrence of stop words in Fragment and in reference fragment

Fragment	Stop word	AF	RF	Part of speech	RF in reference fragment
1	2	3	4	5	6
1 (107 words)	але	1	0.0093	Conjunction	0.0074
	в	2	0.0187	Preposition	0.0140
	для	3	0.0280	Preposition	0.0024
	до	1	0.0093	Preposition	0.0113
	з	1	0.0093	Preposition	0.0129
	і	14	0.1308	Conjunction	0.0300
	й	1	0.0093	Conjunction	0.0038
	мов	1	0.0093	Particle	0.0022
	не	2	0.0187	Particle	0.0237
	про	2	0.0187	Preposition	0.0040
	та	2	0.0187	Conjunction	0.0047
	що	1	0.0093	Conjunction	0.0206

Continuation of Table 2

1	2	3	4	5	6
2 (117 words)	а	2	0.0171	Conjunction	0.0116
	в	3	0.0256	Preposition	0.0140
	від	1	0.0085	Preposition	0.0034
	до	1	0.0085	Preposition	0.0113
	ж	1	0.0085	Conjunction	0.0033
	з	2	0.0171	Preposition	0.0129
	за	1	0.0085	Preposition	0.0053
	і	2	0.0171	Conjunction	0.0300
	й	2	0.0171	Conjunction	0.0038
	на	1	0.0085	Preposition	0.0159
	над	1	0.0085	Preposition	0.0005
	не	2	0.0171	Particle	0.0237
	ні	1	0.0085	Particle	0.0024
	ось	1	0.0085	Particle	0.0012
	от	1	0.0085	Particle	0.0005
	се	1	0.0085	Particle	0.0074
	хіба	1	0.0085	Particle	0.0006
	хоч	1	0.0085	Particle	0.0010
	що	2	0.0171	Conjunction	0.0206
	як	1	0.0085	Conjunction	0.0060
3 (162 words)	а	4	0.0247	Conjunction	0.0116
	але	2	0.0123	Conjunction	0.0074
	без	1	0.0062	Preposition	0.0008
	бо	1	0.0062	Conjunction	0.0012
	в	1	0.0062	Preposition	0.0140
	від	1	0.0062	Preposition	0.0034
	ж	1	0.0062	Conjunction	0.0033
	з	4	0.0247	Preposition	0.0129
	за	2	0.0123	Preposition	0.0053
	і	1	0.0062	Conjunction	0.0300
	й	4	0.0247	Conjunction	0.0038
	на	6	0.0370	Conjunction	0.0159
	навіть	2	0.0123	Particle	0.0011
	не	3	0.0185	Particle	0.0237
	під	4	0.0247	Preposition	0.0011
	таки	1	0.0062	Particle	0.0004
	тож	1	0.0062	Conjunction	0.0001
	у	4	0.0247	Preposition	0.0088
	що	3	0.0185	Conjunction	0.0206
	щоб	1	0.0062	Conjunction	0.0028
	як	1	0.0062	Conjunction	0.0060

Continuation of Table 2

1	2	3	4	5	6
4 (149 words)	адже	1	0.00671	Particle	0.0011
	але	2	0.01342	Conjunction	0.0074
	би	1	0.00671	Particle	0.0033
	в	1	0.00671	Preposition	0.0140
	ж	1	0.00671	Conjunction	0.0033
	з	3	0.02013	Preposition	0.0129
	за	1	0.00671	Preposition	0.0053
	і	4	0.02685	Preposition	0.0300
	мов	1	0.00671	Particle	0.0022
	на	7	0.04698	Preposition	0.0159
	не	4	0.02685	Particle	0.0237
	отсе	1	0.00671	Particle	0.0003
	при	1	0.00671	Preposition	0.0018
	про	2	0.01342	Preposition	0.0040
	се	1	0.00671	Particle	0.0074
	у	2	0.01342	Preposition	0.0088
	чи	2	0.01342	Conjunction	0.0027
	що	7	0.04698	Conjunction	0.0206
	щоб	1	0.00671	Conjunction	0.0028
	як	1	0.00671	Conjunction	0.0060

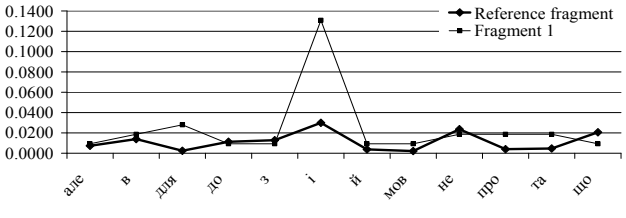

Fig. 1. Relative frequency of occurrence of syntactic words in Fragment 1 and in reference fragment

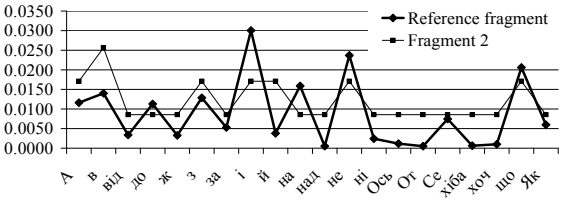

Fig. 2. Relative frequency of occurrence of syntactic words in Fragment 2 and in reference fragment

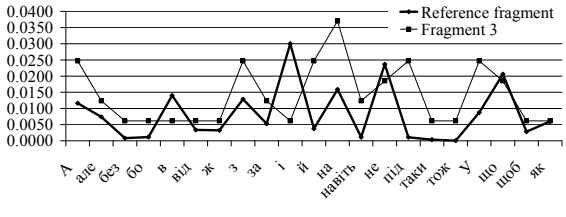

Fig. 3. Relative frequency of occurrence of syntactic words in Fragment 3 and in reference fragment

Fig. 4 shows graphic representation of relative frequency of occurrence of syntactic words in Fragment 4 and in the reference fragment. Correlation coefficient for syntactic words in this case is $R_{r\text{-}F4}$=0.7326.

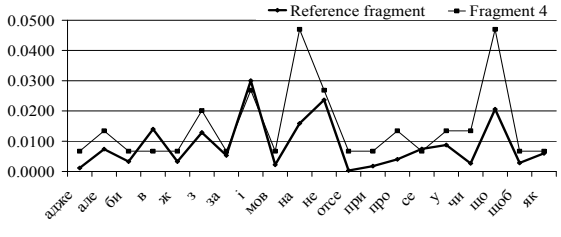

Fig. 4. Relative frequency of the occurrence of syntactic words in Fragment 4 and in reference fragment

Here are the correlation coefficients for each syntactic word for fragments 1–4 (Table 3).

Table 3

Correlation coefficients for syntactic parts of speech

Fragment	Preposition	Conjunction	Particle
1	$R_{r\text{-}F1Z}$=0.72	$R_{r\text{-}F1S}$=0.79	$R_{r\text{-}F1C}$=1
2	$R_{r\text{-}F2Z}$=0.4928	$R_{r\text{-}F2S}$=0.5714	$R_{r\text{-}F2C}$=0.9580
3	$R_{r\text{-}F3Z}$=0.1517	$R_{r\text{-}F3S}$=0.1624	$R_{r\text{-}F3C}$=0.8800
4	$R_{r\text{-}F4Z}$=0.5639	$R_{r\text{-}F4S}$=0.9544	$R_{r\text{-}F4C}$=0.9594

Analyzing correlation coefficients for syntactic words, we come to conclusion that probability of belonging of fragments to the studied reference fragment is the highest for Fragment 4, which is followed by Fragment 2, Fragment 1 and Fragment 3.

We will note that for all four fragments, we can trace consistently high correlation coefficients for particles, which may be understood as the lack of influence of particles on the author's style. In addition, we will analyze frequency of occurrences only of prepositions and conjunctions for fragments, find appropriate correlation coefficients and compare results (Table 4).

Table 4

Correlation coefficients for each fragment

Fragment	Fragment 1	Fragment 2	Fragment 3	Fragment 4
Coefficient $R_{r\text{-}F}$	$R_{r\text{-}F1}$=0.6076	$R_{r\text{-}F2}$=0.7066	$R_{r\text{-}F3}$=0.2810	$R_{r\text{-}F4}$=0.7326
Coefficient $R'_{r\text{-}F}$	$R'_{r\text{-}F1}$=0.6900	$R'_{r\text{-}F2}$=0.4913	$R'_{r\text{-}F3}$=0.2254	$R'_{r\text{-}F4}$=0.6905

Fragment 4 still remained the most likely candidate to belong to the reference fragment, followed with a slight gap by Fragment 1, then by Fragment 2. Fragment 3, like in the previous study, has the least probability of belonging to the reference fragment. To prove the results, we will turn to analyzed texts, from which the three fragments for research were taken.

Thus, application of the method of anchor words produced the following results: among the studied fragments, the fragment, which belongs to analyzed texts, has the highest probability of belonging to the reference fragment. Other results also prove effectiveness of the method of anchor words in the author's attribution of texts. Thus, in the first study, the fragment from another work by the same author has the second highest probability of belonging to the reference fragment. Fragment 1, which also belongs to the reference fragment, "lost" only one-tenth of correlation coefficient to Fragment 4. The result for Fragment 3, separated from the reference fragment by a 100-year period, is also relevant. The assumption about insignificant influence of a particle as a parameter of the method, put forward in [25], led to a decrease in correlation coefficients, but arranged the probability of the fragments to belong to the reference fragment in the right order. Above all, the difference between correlation coefficients for Fragment 1 and Fragment 4 significantly decreased and amounted to 0.0005. However, to prove or to deny the fact that particles are not a determining factor, it is necessary to carry out more fundamental research.

6. Consideration of results of research into the analyzed Ukrainian language content for the recognition of the author's style

To accomplish the goal of the research, we developed a system with possibility of selecting a language/languages of the analyzed content, which was implemented on Web-resource Victana [24]. Analysis of statistics of functioning of the system for recognition of a set of stop words from 100 scientific articles in the technical field included 3 stages (algorithm 2).

Algorithm 2. Analysis and interpretation of linguo-statistic research into recognition and analysis of the author's style.

Stage I. Lexical analysis of a text for determining of stop words and calculation of coefficients of lexical author's language (text diversity).

Stage II. Recognition of the author's style by methods of stylemetry.

Stage III. Analysis of fragments of a text by methods of glottochronology, using the Swadesh lists.

Stages I, II were considered in the previous section of the article. So, let us consider stage III.

The main objective is to determine the number of words from the 200-word Swadesh list, which are present in the works of different time sample, and to determine percentage of such words in fragments. We will also explore the number of common words from the Swadesh list for the selected passages. For consideration, we will find the fragments, written with a gap of several years. Let the fragments contain, for example, 250 words not including the title and proper names. Comparison of the 200-word Swadesh list and Fragment 1 from analyzed texts is given in Table 5.

Table 5

Words from the Swadesh list in Fragment 1

No.	Word	Absolute frequency	Relative frequency
1	все	4	0.0526
2	і	19	0.2500
3	на	3	0.0395
4	он	5	0.0658
5	слухати	1	0.0132
6	як	2	0.0263
7	я	6	0.0789
8	в	4	0.0526
9	знати	2	0.0263
10	довго	2	0.0263
11	чоловік	1	0.0132
12	багато	1	0.0132
13	ім'я	1	0.0132
14	ні	3	0.0395
15	старий	2	0.0263
16	сонце	1	0.0132
17	що	6	0.0789
18	там	3	0.0395
19	what	1	0.0132
20	який	2	0.0263
21	з	5	0.0658
22	рік	1	0.0132
23	ви	1	0.0132
Total		76	

In Fragment 1, containing 253 words, there are 23 words from the 200-word Swadesh list. These words make up 30.04 % of the entire fragment. Fragment 2 is a fragment of analyzed texts. Comparison of the 200-word Swadesh list and Fragment 2 is shown in Table 6.

Table 6

Words from the Swadesh list in Fragment 2

No.	Word	Absolute frequency	Relative frequency
1	все	4	0.0816
2	і	6	0.1224
3	на	1	0.0204
4	назад	1	0.0204
5	далеко	1	0.0204
6	товстий	1	0.0204
7	потік	1	0.0204
8	тут	2	0.0408
9	якщо	1	0.0204
10	в	7	0.1429
11	знати	2	0.0408
12	ні	1	0.0204
13	один	2	0.0408
14	інший	1	0.0204
15	дещо	1	0.0204
16	що	3	0.0612
17	там	2	0.0408
18	це	2	0.0408
19	кидати	1	0.0204
20	який	4	0.0816
21	білий	1	0.0204
22	хто	1	0.0204
23	з	2	0.0408
24	ви	1	0.0204
Total		49	

In Fragment 2, containing 262 words, there are 24 words from the 200-word Swadesh list. These words make up 18.7 % of the entire fragment. Fragment 3 is a fragment of analyzed texts. Comparison of 200-word Swadesh list and Fragment 3 is shown in Table 7.

Table 7

Words from the Swadesh list in Fragment 3

No.	Word	Absolute frequency	Relative frequency
1	все	3	0.0652
2	і	10	0.2174
3	на	1	0.0217
4	приходити	1	0.0217
5	тут	1	0.0217
6	якщо	1	0,0217
7	в	4	0.087
8	знати	2	0.0435
9	довго	1	0.0217
10	ні	7	0.1522
11	інший	1	0.0217
12	казати	1	0.0217
13	що	4	0.087
14	там	1	0.0217
15	вони	2	0.0435
16	це	1	0.0217
17	який	1	0.0217
18	хто	2	0.0435
19	з	2	0.0435
Total		46	

In Fragment 3, containing 246 words, there are 19 words from the 200-word Swadesh list. These words make up 18.7 % of the entire fragment. Analyzing the obtained data, we notice that the words from the Swadesh list in Fragment 1 make up 30 % of the fragment, which is significantly more than 18.7 %, similar to Fragments 2 and 3 (Fig. 5). Such results are objective and transparent: over time, the vocabulary of a person is enriched. For these fragments, Fig. 6 graphically shows the following results:

– in the nodes, the fragment, and the number of words from the Swadesh list in it, are specified;

– on the arcs, the number of common words from the Swadesh list for these passages and correlation coefficient for these passages are specified;

– in the center, the total number of words, common for the fragments and the Swadesh list is specified (Table 8).

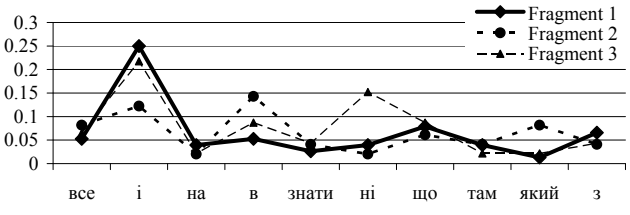

Fig. 5. Numerical results of examination of fragments

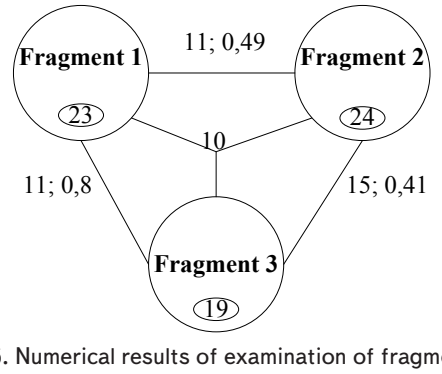

Fig. 6. Numerical results of examination of fragments

Table 8

Words that are common for Fragments 1–3 and Swadesh list

No.	Common words	Relative frequency in Fragment 1	Relative frequency in Fragment 2	Relative frequency in Fragment 3
1	все	0.0526	0.0816	0.0652
2	і	0.25	0.1224	0.2174
3	на	0.0395	0.0204	0.0217
4	в	0.0526	0.1429	0.087
5	знати	0.0263	0.0408	0.0435
6	ні	0.0395	0.0204	0.1522
7	що	0.0789	0.0612	0.087
8	там	0.0395	0.0408	0.0217
9	який	0.0132	0.0816	0.0217
10	з	0.0658	0.0408	0.0435

The scope of the conducted studies does not allow us to state that such a high correlation coefficient as we have between Fragment 1 and Fragment 3, is objective. At present, the coefficient allows us to put forward the hypothesis that, in general, Fragment 1 is either written in a different period of time than Fragments 2, 3, or written by another person.

The fact that such dependence actually exists, or that it is a random coincidence because of a poorly chosen fragment, requires much wider research.

7. Conclusions

1. We developed the method of recognition of the style of the text's author based on coefficients of lexical author's language in the reference fragment of the author's text. The method lies in comparative analysis of the author's attribution in the statistically processed works of literature, created by the author (reference fragment), with an arbitrary analyzed fragment. The method estimates belonging of certain fragments to the reference fragment with analysis of relevant coefficients of lexical author's language. Moreover, the method works provided that the author's reference fragment has already been built and analyzed – problems of texts selection, lemmatization and problems of inhomogeneity have been solved, the processed material is formed as a frequency dictionary of syntactic words (stop words). For attribution, we used the method of anchor words, the results are presented in the form of correlation coefficients. We will separately mention evolution of significance of one of the text parameters – in the author's attribution of texts.

The algorithm of determining the stop words in text content based on an linguistic analysis of the text content was developed. For the individual style of a writer, it is syntactic words that are the most significant, because they are in no way related to the theme and the content of a book. The analyzed fragments are formatted with respect to selection of the method of attribution: for any fragment, only prepositions, conjunctions and particles were automatically chosen. The total number of word usages in the fragment was calculated, proper names were not taken into account. For each fragment, absolute and relative frequencies of occurrence of stop words were analyzed and compared with reference values. Therefore, application of the method of anchor words gives the following results: among the studied fragments, finding the one that most likely belongs to the reference fragment. Other results also prove effectiveness of the method of anchor words in the author's attribution of texts. The assumption that was made about an insignificant impact of a particle as a parameter of the method on the results led to a decrease in correlation coefficients, but arranged the probability of belonging of fragments to the reference fragment in the right order. However, to prove or refute the fact that particles are not a determining factor in the author's style, it is necessary to carry out a deeper fundamental research.

The algorithm of lexical analysis of texts in Ukrainian and the algorithm of a syntactic parser of text content was developed. Special features of the algorithm include adaptation of morphological and syntactic analysis of lexical units to structural features of words/texts in Ukrainian. Theoretical and experimental substantiation of the method of content monitoring and determining of stop words of a text in Ukrainian was presented. The method is aimed at automatic detection of notional stop words in a Ukrainian text with the use of the proposed formal approach to implementation of content parsing.

2. We proposed an approach to development of software of content monitoring for recognition of the style of an author in Ukrainian texts based on Web Mining. The peculiarity of the approach is in the adaptation of linguo-statistical analysis of lexical units to structural features of words/texts in Ukrainian.

3. We studied results of experimental testing of the proposed method for content-monitoring for recognition of the style of an author in Ukrainian scientific texts in the technical area. 100 scientific publications from two issues (783 and 805) of the Visnyk of the National University "Lviv Polytechnika" from a series "Information systems and networks" were examined. Testing of the proposed method for the recognition of the author's style for other categories of texts – scientific humanitarian, belles-lettres, journalistic, etc. – requires subsequent experimental research.

References

1. Anisimov, A. Sistema obrabotki tekstov na estestvennom yazyke [Text] / A. Anisimov, A. Marchenko // Iskusstvennyy intellekt. – 2002. – Issue 4. – P. 157–163.
2. Perebyinis, V. Matematychna linhvistyka. Ukrainska mova [Text] / V. Perebyinis. – Kyiv: Ukrainska entsyklopediya, 2000. – P. 287–302.
3. Buk, S. N. Osnovy statystychnoi lingvistyky [Text] / S. N. Buk; F. S. Batsevych (Ed.). – Lviv: Vydavnychyi tsentr LNU im. I. Franka, 2008. – 124 p.
4. Varfolomeev, A. P. Psihosemantika slova i lingvostatistika teksta [Text] / A. P. Varfolomeev. – Kaliningrad: KGU, 2000. – 37 p.
5. Kognitivnaya stilometriya: k postanovke problemy [Electronic resource]. – Available at: http://www.manekin.narod.ru/hist/styl.htm
6. D'yachok, M. T. Glottohronologiya: pyat'desyat let spustya [Text] / M. T. D'yachok // Sibirskiy lingvisticheskiy seminar. – 2002. – Issue 1. – P. 44–69.
7. Perebyinis, V. I. Statystychni metody dlia lingvistiv [Text] / V. I. Perebyinis. – Vinnytsia: Nova knyha, 2013. – 176 p.
8. Kochergan, M. P. Vstup do movoznavstva [Text] / M. P. Kochergan. – Kyiv: Akademiya, 2005. – 329 p.
9. Sushko, S. Chastoty povtoriuvanosti bukv i bihram u vidkrytykh tekstakh ukrainskoiu movoiu [Text] / S. Sushko, L. Fomychova, Ye. Barsukov // Ukrainian Information Security Research Journal. – 2010. – Vol. 12, Issue 3. doi: 10.18372/2410-7840.12.1968
10. Hmelev, D. Kak opredelit' pisatelya? [Electronic resource] / D. Hmelev // Komp'yuterra-Onlayn. – 2000. – Available at: http://old.computerra.ru/2000/338/195699/
11. Lande, D. V. Pidkhid do rishennia problem poshuku dvomovnoho plahiatu [Text] / D. V. Lande, V. V. Zhyhalo // Problemy informatyzatsii ta upravlinnia. – 2008. – Issue 2 (24). – P. 125–129.
12. Morozov, N. A. Lingvisticheskie spektry: sredstvo dlya otlicheniya plagiatov ot istinnyh proizvedeniy togo ili inogo neizvestnogo avtora. Stilemetricheskiy ehtyud [Electronic resource] / N. A. Morozov // Izvestiya otd. russkogo yazyka i slovesnosti Imp. Akad. nauk. – 1915. – Vol. XX. – Available at: http://www.textology.ru/library/book.aspx?bookId=1&textId=3

13. Bubleinyk, L. V. Osoblyvosti khudozhnoho movlennia [Text] / L. V. Bubleinyk. – Lutsk: Vezha, 2000. – 179 p.
14. Rodionova, E. S. Metody atribucii hudozhestvennyh tekstov [Text] / E. S. Rodionova // Strukturnaya i prikladnaya lingvistika. – 2008. – Issue 7. – P. 118–127.
15. Meshcheryakov, R. V. Modeli opredeleniya avtorstva teksta [Text] / R. V. Meshcheryakov, N. S. Vasyukov // Izmereniya, avtomatizaciya i modelirovanie v promyshlennosti i nauchnyh issledovaniyah. – 2005. – P. 25–29. – Available at: http://db.biysk.secna.ru/conference/conference.conference.doc_download?id_thesis_dl=427
16. Khomytska, I. The Method of Statistical Analysis of the Scientific, Colloquial, Belles-Lettres and Newspaper Styles on the Phonological Level [Text] / I. Khomytska, V. Teslyuk // Advances in Intelligent Systems and Computing. – 2016. – P. 149–163. doi: 10.1007/978-3-319-45991-2_10
17. Khomytska, I. Specifics of phonostatistical structure of the scientific style in English style system [Text] / I. Khomytska, V. Teslyuk // 2016 XIth International Scientific and Technical Conference Computer Sciences and Information Technologies (CSIT). – 2016. doi: 10.1109/stc-csit.2016.7589887
18. Lytvyn, V. Classification Methods of Text Documents Using Ontology Based Approach [Text] / V. Lytvyn, V. Vysotska, O. Veres, I. Rishnyak, H. Rishnyak // Advances in Intelligent Systems and Computing. – 2016. – P. 229–240. doi: 10.1007/978-3-319-45991-2_15
19. Lytvyn, V. The method of formation of the status of personality understanding based on the content analysis [Text] / V. Lytvyn, P. Pukach, I. Bobyk, V. Vysotska // Eastern-European Journal of Enterprise Technologies. – 2016. – Vol. 5, Issue 2 (83). – P. 4–12. doi: 10.15587/1729-4061.2016.77174
20. Vysotska, V. Linguistic analysis of textual commercial content for information resources processing [Text] / V. Vysotska // 2016 13th International Conference on Modern Problems of Radio Engineering, Telecommunications and Computer Science (TCSET). – 2016. doi: 10.1109/tcset.2016.7452160
21. Vysotska, V. Information technology of processing information resources in electronic content commerce systems [Text] / V. Vysotska, L. Chyrun, L. Chyrun // 2016 XIth International Scientific and Technical Conference Computer Sciences and Information Technologies (CSIT). – 2016. doi: 10.1109/stc-csit.2016.7589909
22. Vysotska, V. The commercial content digest formation and distributional process [Text] / V. Vysotska, L. Chyrun, L. Chyrun // 2016 XIth International Scientific and Technical Conference Computer Sciences and Information Technologies (CSIT). – 2016. doi: 10.1109/stc-csit.2016.7589902
23. Marchenko, O. O. Modeliuvannia semantychnoho kontekstu pry analizi tekstiv na pryrodnii movi [Text] / O. O. Marchenko // Visnyk Kyivskoho universytetu. – 2006. – Issue 3. – P. 230–235.
24. Bloh Viktorii Anatoliivny [Electronic resource]. – Available at: http://victana.lviv.ua/index.php/kliuchovi-slova
25. Rodionova, E. S. Metody atribucii hudozhestvennyh tekstov [Text] / E. S. Rodionova // Strukturnaya i prikladnaya lingvistika. – 2008. – Issue 7. – P. 118–127. – Available at: http://epir.ru/pragmat!/projects/corneille/files/Metody_atributsii.pdf
26. Lytvyn, V. Development of a method for determining the keywords in the slavic language texts based on the technology of web mining [Text] / V. Lytvyn, V. Vysotska, P. Pukach, O. Brodyak, D. Ugryn // Eastern-European Journal of Enterprise Technologies. – 2017. – Vol. 2, Issue 2 (86). – P. 14–23. doi: 10.15587/1729-4061.2017.98750
27. Lytvyn, V. Content linguistic analysis methods for textual documents classification [Text] / V. Lytvyn, V. Vysotska, O. Veres, I. Rishnyak, H. Rishnyak // 2016 XIth International Scientific and Technical Conference Computer Sciences and Information Technologies (CSIT). – 2016. doi: 10.1109/stc-csit.2016.7589903
28. Lytvyn, V. Designing architecture of electronic content commerce system [Text] / V. Lytvyn, V. Vysotska // 2015 Xth International Scientific and Technical Conference "Computer Sciences and Information Technologies" (CSIT). – 2015. doi: 10.1109/stc-csit.2015.7325446
29. Vysotska, V. Analysis features of information resources processing [Text] / V. Vysotska, L. Chyrun // 2015 Xth International Scientific and Technical Conference "Computer Sciences and Information Technologies" (CSIT). – 2015. doi: 10.1109/stc-csit.2015.7325448
30. Chen, J. Smart Data Integration by Goal Driven Ontology Learning [Text] / J. Chen, D. Dosyn, V. Lytvyn, A. Sachenko // Advances in Big Data. Proceedings of the 2nd INNS Conference on Big Data. – October 23-25, 2016. – Thessaloniki, Greece. – P. 283-292.
31. Mykhailiuk, A. A Creation of the Linguistic Ontology Based on a structured Electronic Encyclopedic Resource [Text] / A. Mykhailiuk, O. Mykhailiuk, O. Pylypchuk, V. Tarasenko // International Journal of Computing. – 2012. – Vol. 11, Issue 3. – P. 191-202.