Linguistic intellectual analysis methods for Ukrainian textual content processing

Victoria Vysotska

Lviv Polytechnic National University, Stepan Bandera 12, 79013 Lviv, Ukraine

Abstract

The peculiarities of the method of syntactic analysis of Ukrainian-language text content aimed at automatic detection of significant keywords of input texts are considered. The role and formal features of the parser in the process of identifying keywords of the content topic are defined, and the procedures of the proposed method are decomposed into 4 stages. Compared to well-known parsers, the proposed method provides self-improvement and self-learning of the automated keyword identification system due to the mechanism of identification of significant statistical parameters within the limits defined by the moderator. The experimental study confirmed the reliability of the method - for various methods of processing the primary text, the average coincidence of the lists of identified keywords with the authors varies in the range of 52.6-68.5%. The accuracy of matching keywords with the author's keywords ranges from 43.6 to 62.9%. The average match of meaningful keywords compared to all found by the system ranges from 38.9 to 75.8% according to the stages of article text analysis. The accuracy of matching keywords compared to all found by the system varies between 34.3-71.9% according to the stages of analysis of the texts of the articles.

The reliability of scientific and practical results is confirmed by relevant materials on the implementation of dissertation research, as well as by comparing the obtained practical results on different samples of reliable input data. CLS was developed on the information resource http://victana.lviv.ua using CMS Joomla! (for developing the e-framework of articles), PHP (for implementing text content processing methods), HTML (for implementing page markup), CSS (for describing page styles), and MySQL (for storing data and dictionaries). The experimental study confirmed the reliability of the method of determining keywords - for different algorithms for processing the primary text, the average coincidence of the lists of identified keywords with the authors varies in the range of 52.6-68.5%. The accuracy of matching keywords with the author's keywords ranges from 43.6 to 62.9%. The average match of meaningful keywords compared to all found by the system ranges from 38.9-75.8%, depending on the stages of analysis of article texts. The accuracy of matching keywords compared to all found by the system varies between 34.3-71.9%, depending on the stages of analysis of the text of the articles.

Keywords

computer linguistics, system, NLP, Ukrainian language, information resource, system modelling

CLW-2024: Computational Linguistics Workshop at 8th International Conference on Computational Linguistics and Intelligent Systems (CoLInS-2024), April 12–13, 2024, Lviv, Ukraine

* Corresponding author.

† These authors contributed equally.

victoria.a.vysotska@lpnu.ua (V. Vysotska)

0000-0001-6417-3689 (V. Vysotska)

1. Introduction

The identification of keywords of the text content $\varsigma\,(C, U, R, D, T) \to C'$ is a mapping of the input text content C into the new state C', which, unlike the previous one, is supplemented with a set of keywords as the main markers of the text content. For this purpose, the multi-level linear (sequences) [1-3]. And, if necessary, hierarchical/network (interconnections) structure of the text is linguistically investigated as symbols, N-grams, morphological features, weights of words and phrases, features of sentences and interconnected units (Fig. 1) [4-9].

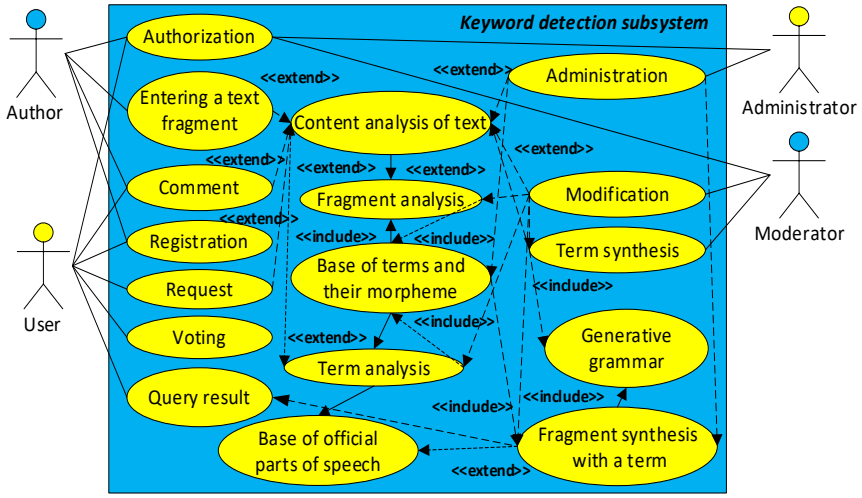


Figure 1: Keyword identification use case diagram

2. Models and methods

2.1. Peculiarities of defining keywords of the Ukrainian-language text

Web Mining technology is based on the use of methods of intellectual analysis of the flow of information content to identify patterns in the Internet or Web-site [10-12]. The main technology of Web Mining is Text Mining, which is used to extract structured/unstructured data from Web-pages, Web-sites, link structures, etc. [13-15].

Algorithm 1. Content keyword identification based on Web Mining

Stage 1. Integration/downloading of textual content for further analysis.

Stage 2. Grapheme analysis of textual content C.

Step 1. Formatting of incoming text content, for example, the same apostrophes for Ukrainian text.

Step 2 Removal of the service part of the C content, such as tags.

Step 3. Removal of the non-character part of C content, such as dates, numbers, financial symbols, mathematical formulas, images, etc. Removal of special characters that are not included in the alphabet, except for service ones such as space, and apostrophe.

Step 4. Analysis of abbreviations and abbreviations of content C. If $\leq n$ used in the text and not in the dictionary D, then step 5, otherwise step 6.

Step 5. If necessary, edit the thematic dictionary D, for example, add new abbreviations or abbreviations.

Step 6. Segmentation of the input array of text C into sentences and paragraphs with appropriate marking of the corresponding boundaries.

Step 7. Segmentation of the sequence of symbols of sentences of content C into tokens.

Stage 3. Morphological analysis of the Ukrainian-language text C.

Step 1. Selection of bases (word forms without inflexions).

Step 2. Analysis of the resulting inflexion to determine the part of speech.

Step 3. Marking the word with the appropriate part of speech.

Step 4. Word forms are marked by a collection of morphological features: case, gender, declension, singular/plural, person, etc.).

Step 5. If the part of speech word is a noun, mark it as a potential keyword. If the part of speech of the word is an adjective, mark it and the next word (if it is a noun) as a phrase that could potentially be a keyword.

Step 6. Formation of a linear chain of labelled structures.

Stage 4. Lexical analysis of the Ukrainian text C.

Step 1. Search for the base in the base dictionary for further normalization taking into account the part of the language used in a specific place of the text C.

Step 2. Normalization of marked morphological structures.

Step 3. Segmentation and analysis of a chain of normalized tokens of content C into tokens and word types taking into account marked sentence boundaries.

Step 4. Formation of collections of tokens (sequences of symbols according to appropriate templates) as lexemes with further identification of their types, taking into account their interrelationships in the textual content C.

Step 5. If the dimensionality of the text content is $\leq N_1$, then step 9, otherwise step 5.

Stage 5. Syntactic analysis of textual content C.

Step 1. Selection of tokens $U_1 \in U$ for text content C.

Step 2. Identification of a sequence of tokens as an expression or sentence.

Step 3. Identification of the nominal group of the expression based on the dictionary of word bases D.

Step 4. Definition of the verb group of the sentence based on the dictionary of word bases D.

Step 5. Formation of a left-to-right parsing tree of linguistic variables.

Step 6. Analysis of noun phrase group for textual content C.

Step 7. Analysis of the verb group of the sentence for textual content C.

Step 8. Study of syntactic categories by word forms.

Step 9. If not the end of content C, then go to step 2, otherwise go to step 9.

Stage 6. Semantic analysis of the Ukrainian text C.

Step 1. Expression tokens are compared with the semantic classes of the dictionary D.

Step 2. Definition of morpho-semantic analogues for a specific sentence.

Step 3. Combining tokens into a common structure.

Step 4. Generating a tuple of superpositions of lexical functions and semantic classes.

Stage 7. Referential analysis for determining interphase unities of the text C.

Step 1. Contextual analysis of C content for identification of local references (which, this, his) and selection of utterances - kernels of unity.

Step 2. Thematic analysis to highlight the thematic structure.

Step 3. Identification of the identity of references; synonymizing, duplication and re-nomination of tokens; implications based on situational connections.

Stage 8. Structural analysis of textual content C.

Step 1. Identification of the basic tuple of rhetorical connections between entities.

Step 2. Construction of a nonlinear network of units.

Stage 9. Identifying a set of content keywords $\varsigma\ (C, U, R, D, T) \to C'$.

Step 1. Formation of an alphabetic-frequency dictionary $Vocab = \upsilon(C, D, R)$.

Step 2. Identification of terms $(Noun \in U_1) \cap (Noun \in Vocab)$ as nouns, noun phrases, an adjective with a noun, or abbreviations.

Step 3. Formation of a shortened list of words whose frequencies correspond to the conditions of formation of potential keywords – $Filter \subseteq Vocab$.

Step 4. Determination of the level of uniqueness $\forall Noun\ Unicity(Noun), Noun \in Filter$.

Step 5. Nmb_{Smb} calculation (number of characters without spaces) for $Noun \in Filter$ at $Unicity \geq 80$.

Step 6. Calculation of Us_{Fr} (keyword usage frequency). For terms with $Nmb_{Smb} \leq 2000$ frequency $Us_{Fr} \in (6; 8]$%, з $2000 > Nmb_{Smb} < 3000$ frequency $Us_{Fr} \in [4; 6]$%, with $Nmb_{Smb} \geq 3000$ frequency $Us_{Fr} \in [2; 4)$%.

Step 7. Calculation of the probability of using the keywords Bs_{Fr} (at the beginning of the text), Is_{Fr} (in the middle of the text content) and Es_{Fr} (at the end of the text content).

Step 8. Comparison of Bs_{Fr}, Is_{Fr} та Es_{Fr} values for keyword prioritization under the condition $Bs_{Fr} \gg Is_{Fr} \gg Es_{Fr}$.

Step 9. Sorting keywords according to defined priorities.

Step 10. Comparison of $Filter \subseteq Vocab$ content with the $Thematic \in D$ list.

Step 11. Formation of a new list of $Resvoc = Filter \cap Thematic$ tokens.

Step 12. Formation of the collection of keywords C' with $KeyWords \in Resvoc$, $KeyWords = \{Noun, Unicity \geq 80, Nmb_{Smb}, Us_{Fr}, Bs_{Fr}, Is_{Fr}, Es_{Fr}\}$.

2.2. Method of identifying keywords of Ukrainian-language content

The analysis of the text flow of C content for the identification of keywords is usually implemented on Zipf's law and reduced to the selection of words with an average frequency of occurrence [16-18]. This is easy to implement for English-language texts. It will not work for Ukrainian-language texts. It is necessary to adapt the parser and stemming algorithms to the Ukrainian language based on thematic frequency dictionaries of the basics [19-27].

Algorithm 2. Adaptation of parser/stemming algorithms of Ukrainian texts.

Stage 1. Based on the parser, a set of words with a frequency of occurrence within a certain limit is identified, for example, 4-6% with ≤ 2000 characters without spaces;

Stage 2. Based on the parser and stemming, a subset of frequently used semantically loaded words is generated by extracting/marking words from the blocked dictionary, for example, such as prepositions, conjunctions, pronouns, verbs, particles, etc.;

Stage 3. If the keyword is an adjective (inflexion of the normalized word **ий** [yy]), then all bases to the right of it are found in the text and a frequency dictionary is built for them. Those phrases that are used more than the corresponding threshold value (but less than this adjective) are keywords. The threshold value is determined by the moderator. Repeat multiple keywords

Stage 4. If the keyword is a noun (the inflexion of the word is not **ий** [yy]), then all bases and their inflexions on both sides of it are examined.

Step 1. All words to the left of the noun are analysed for the presence of inflexions **ий** [yy] and compared with the frequency dictionary. A set of words that are used most often above the threshold value is identified - these are new keywords.

Step 2. All bases and their inflexions on the right are analysed - without inflexion **ий** [yy] and inflexions of other parts of speech, except nouns, are compared with the frequency dictionary, which determines the set of keywords.

Stage 5. The new subset is compared with the thematic dictionary of the basics of Ukrainian words to form a set of keywords;

Stage 6. If there is no analogue of the word, add it to the thematic dictionary of word bases through the buffer dictionary (edited by the moderator) to accumulate statistics for various stylistic text content.

3. Experiments, results and discussion

3.1. Content keyword identification based on Web Mining technology

100 scientific articles of the "Lviv Polytechnic" NU Bulletin of the "Information Systems and Networks" series (http://science.lp.edu.ua/sisn), two numbers 783 (http://science.lp.edu.ua/SISN/SISN-2014) and 805 (http://science.lp.edu.ua/sisn/vol-cur-805-2014-2) were chosen as the experimental base for the relevant research. To achieve the goal of the research, IS was developed (Fig. 2), placed on the Victana resource (http://victana.lviv.ua/index.php/kliuchovi-slova) using the following tools: CMS Joomla! for IS e-framework, PHP for algorithm implementation, MySQL for data storage and dictionaries, HTML for implementation of Web-pages markup and CSS for description of Web-page styles.

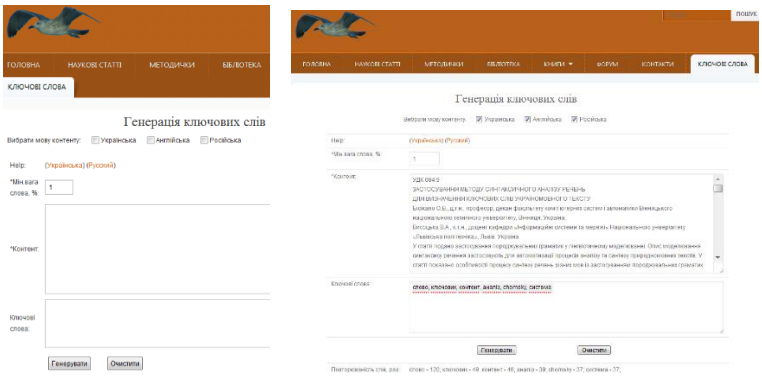


Figure 2: IS dialogue box for identifying keywords in text content

The developed IS has the following main components.

1. A user-friendly dialogue web interface on the web page of the *Ключові слова* [Klyuchovi slova] (Keywords) menu with the following sections (Fig. 2):

- *Вибрати мову контенту* [Vybraty movu kontentu] (Select the content language) – one/several languages of the analyzed text.
- *Мін. вага слова*, % [Min. vaha slova, %] (Min. word weight, %) – the percentage of the weight of the keyword to the total number of words of the text, after which the keywords will be selected; format - XX.XX, within [00.01 - 99.99]; mandatory field.
- *Help* – short instructions in Ukrainian on a separate web page.

- *Контент* [Kontent] (Content) – field for analysed text content.
- *Ключові слова* [Klyuchovi slova] (Keywords) – field for displaying IS of keywords set.
- *Генерувати* [Heneruvaty] (Generate) – start the keyword identification process.
- *Очистити* [Ochystyty] (Clear) – clearing the input field *Контент* [Kontent] (Content).
- *Повторюваність слів, раз* [Povtoryuvanist' sliv, raz] (Repetition of words, times) – the number of repetitions of the keyword in the text.
- *Рекомендовані рубрики* [Rekomendovani rubryky] (Recommended headings) – a list of thematic headings according to keywords.

2. The main relations of DB: the bases of words; prohibited words; rubrics; and rules of bringing to the base of the word.

3. PHP functions for processing text content:

- get_keywords() – creating a list of keywords.
- get_word() – a record of the rules for bringing the word to the base.
- explode_str_on_words() – clears the received content from blocked words, special characters, etc.
- blocked_words() – forms a list of blocked words depending on the selected language of the context.
- count_words() – calculation of key word frequencies.
- set_keywords() – writing keywords to the DB if they are not available.
- recommend_rubric() – creation of a list of recommended rubrics.
- function error() – processing errors, sending a letter to the IS administrator.

The study of the dynamics of the module for determining the collection of keywords from 100 scientific and technical articles was carried out in two stages with analysis:

- content of the thematic dictionary and a set of blocked words.
- refined based on the ML content of the thematic dictionary and set of blocked words, since with each subsequent verification of the text through the corresponding module, an additional collection of unknown words is potentially generated (absent in the list of blocked and in the thematic dictionary).

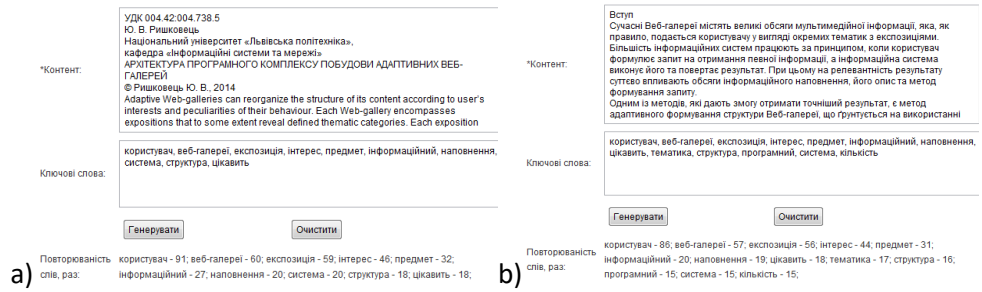


Figure 3: Results for keywords generation (http://victana.lviv.ua/index.php/kliuchovi-slova)

At each stage, the module implements the verification of the text of articles in two steps: analysis of the entire article (Fig. 3a) and without meta-data (information about authors, title, author keywords and annotations in several languages, references list, etc.) (Fig. 3b) to analyse the accuracy error of generating a collection of keywords in the presence of information noise.

3.2. An experimental study results of the Ukrainian-language content keywords identification

The statistical analysis was carried out based on a comparison of sets of keywords defined by the authors of the article and defined by the module at two different stages with different word weights within [1,5] (in the option **Мін.вага слова, %* [*Min.vaha slova, %] (*Min. word weight, %)) with full and abbreviated texts of works (Table 1) with an average arithmetic value of the author's keywords of 4.77, which approximately consist of 9-10 words. Table 2 contains the following notations: *A* (total identified keywords at a given word weight), *B* (formed significant words without pronouns and verbs), *C* (coincidence of words with the author's list), *D* (accuracy of the coincidence of identified keywords with the author's list), *E* (additional keywords defined, but not defined by the author of the publication). Known IS of keywords identification are within [100 ÷ 1000] words [28-32].

Table 1

Statistical data of volumes of analyzed texts of scientific and technical publications

Title of the article volume	Step 1		Step 2	
	In total	Arithmetic average	In total	Arithmetic average
Pages	956	9.56	828	8.28
Paragraphs	16497	164.97	15263	152.63
Rows	42553	425.53	36965	369.65
Words	345580	3455.8	291247	2912.47
Signs	2327209	23272.09	1974773	19747.73
Spaces and signs	2674889	26748.89	2265917	22659.17

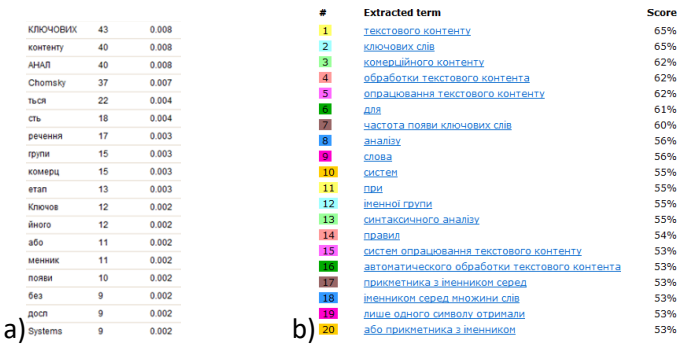


Figure 4: The result of the analysis of the article on a) [31] and b) [32]

Table 2

Statistical data of the researched content of the texts of scientific and technical publications

Name	The weight of words	Stage 1					Stage 2				
		A	*B*	*C*	*D*	*E*	*A*	*B*	*C*	*D*	*E*
Step 1	≥ 1	5.46	3.92	2.51	2.08	1.74	7.43	7.03	3.27	3	4.18

	≥ 2	1.08	0.88	0.63	0.59	0.26	2.67	2.64	1.65	1.54	1.12
	≥ 3	0.41	0.38	0.22	0.21	0.16	1.21	1.2	0.85	0.79	0.41
	≥ 4	0.15	0.13	0.09	0.09	0.04	0.46	0.45	0.33	0.31	0.15
	≥ 5	0	0	0	0	0	0	0	0	0	0
Step 2	≥ 1	6.51	5.02	2.68	2.23	2.37	8.35	7.78	3.25	2.91	4.99
	≥ 2	1.34	1.11	0.74	0.72	0.39	3.12	3.07	1.81	1.67	1.43
	≥ 3	0.51	0.45	0.29	0.27	0.17	1.42	1.4	0.93	0.85	0.54
	≥ 4	0.19	0.17	0.12	0.12	0.05	0.73	0.72	0.45	0.42	0.31
	≥ 5	0.11	0.1	0.06	0.06	0.04	0.33	0.32	0.25	0.23	0.1

The disadvantage of these IS is the inaccuracy and incorrect processing of Ukrainian-language texts in the absence of competently constructed morphological dictionaries, dictionaries of bases and blocked words. Also, the main drawback of most such IS is the limited processing of volumes of text content $[100 \div 1000]$ (Fig. 4). The best IS for processing Ukrainian-language textual content is [33] (Fig. 5), but it does not identify the set of keywords, but only the frequency of use of words, phrases and parts of words. Doesn't work with word bases at all (*ключових* [klyuchovykh] (keywords) and *ключові* [klyuchovi] (keywords) are different). The developed resource works with the basics of the word and is focused on Ukrainian/English texts (Fig. 1). For [20] in Ukrainian, the frequency of using keywords on Victana: *слово* [slovo] (word) – 120; *ключовий* [klyuchovyy] (key) – 49; *контент* [kontent] (content) – 46; *аналіз* [analiz] (analysis) – 39; *Chomsky* – 37; *система* [systema] (system) – 37. The authors identified keywords: *текст* [tekst] (text), *україномовний* [ukrayinomovnyy] (Ukrainian), *алгоритм* [alhorytm] (algorithm), *синтаксичний аналіз* [syntaksychnyy analiz] (syntactic analysis), *породжувальні граматики* [porodzhuval'ni hramatyky] (generative grammars), *лінгвістичний аналіз* [linhvistychnyy analiz] (linguistic analysis), *контент-моніторінг* [kontent-monitorinh] (content monitoring), *ключові слова* [klyuchovi slova] (keywords), *інформаційна лінгвістична система* [informatsiyna linhvistychna systema] (informational linguistic system), *структурна схема речення* [strukturna skhema rechennya] (sentence structure scheme). Authors usually define keywords more than Zipf-law patterns of word frequency distribution.

Наименование показателя	Значение
Количество символов	35927
Количество символов без пробелов	31118
Количество слов	4354
Количество уникальных слов	1589
Количество значимых слов	2873
Количество стоп-слов	1013
Вода	34.0 %
Количество грамматических ошибок	460
Классическая тошнота документа	8.12
Академическая тошнота документа	4.9 %

Слово	Количество	Частота,
слів	66	1.52
контент	54	1.24
ключових	45	1.03
chomsky	37	0.85
текст	36	0.83
система	29	0.67
текстовой	24	0.55
граматика	22	0.51
аналізу	21	0.48
крок	21	0.48
речення	18	0.41
chomsky	16	0.37
частота	16	0.37

Слово	Количество	Частота
в	85	1.95
тот	68	1.56
of	60	1.38
п	56	1.29
з	48	1.10
на	45	1.03
слово	40	0.92
the	35	0.80
для	31	0.71
р	29	0.67
і	29	0.67
and	27	0.62
у	26	0.60

Figure 5: The result of the analysis of this article on [1044]

The author of the article almost always forms at his discretion the number and content of a set of keywords in the range of 2 to 10 word combinations (usually 3-5). The developed module defines a different number of words, depending on the writing style of the corresponding author, the volume of the article, the genre, the topic, and the frequency of use of the corresponding words (from 0 to several dozen). The coincidence of the sets of found keywords with the author's without taking into account the extra words defined by the authors (repetition > 30 for a text volume of more than 4800 words) is, respectively, for [33] - 83%; [32] - 57%; [31]

- 35%; %; http://victana.lviv.ua/kliuchovi-slova - 90% (Fig. 6). Fig. 7 demonstrates the features of generating a set of probable keywords compared to an author set. The author of the article often defines a larger number of words (A_2) and a smaller number of keywords (A_1) than are present in the text. Fig. 7b shows the distribution of text density in articles, where the number of 1 – pages, 2 – paragraphs, 3 – lines, 4 – words, 5 – characters, 6 – spaces and characters, 7 – words per page, 8 – characters per page, 9 – spaces and characters on the page.

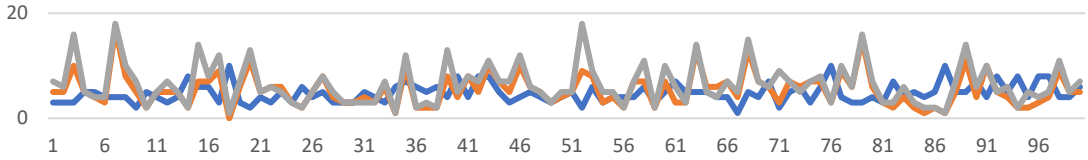


Figure 6: The results of the analysis of the set of 100 articles (blue – authors keywords, orange – stage 1, grey – stage 2)

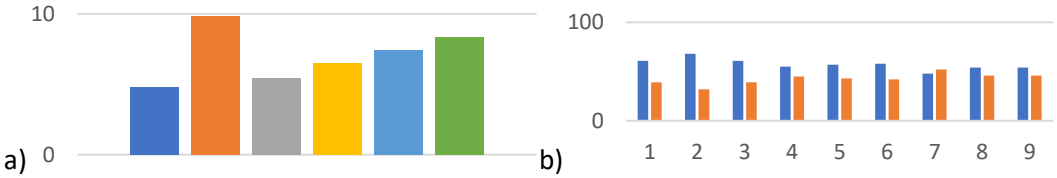


Figure 7: Analysis of verification of 100 articles (explanation in Table 3-4): a) blue - author's keywords, orange - quantity, grey - stage 1 of step 1, yellow - stage 1 of step 2, light blue - stage 2 of step 1, green - stage 2 of step 2; b) blue step - less than the average value, orange - more than the average value

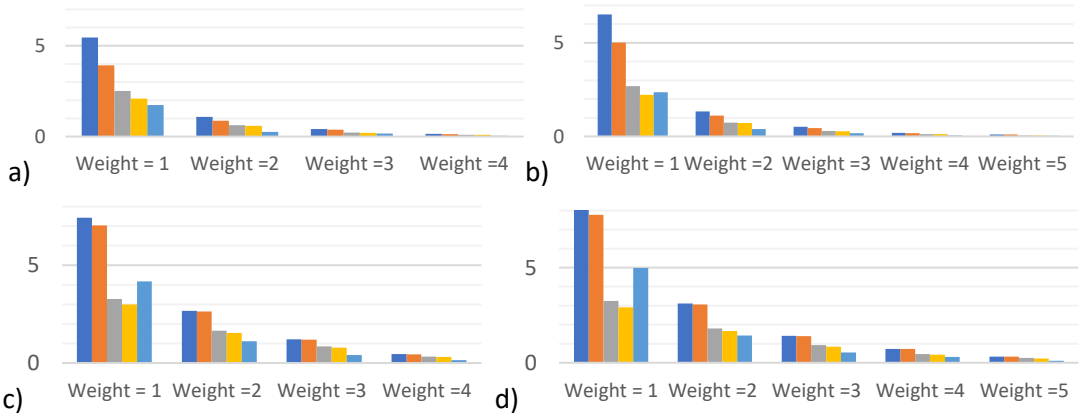


Figure 8: Obtaining meaningful words during text processing at a) stage 1, step 1, b) stage 1, step 2, c) stage 2, step 1 and d) stage 2, step 2, where blue - all words, orange - meaningful words, grey - match with author's, yellow - accuracy of the match, light blue - additional words

Table 3

Statistical data as an explanation for Fig. 7

Marking	Chart column name	Arithmetic average number of keywords	
		Explanation	Value
A_1	Author's keywords	defined by the author	4.77
A_2	Number of words	contain author's	9.82
A_3	Stage 1, Step 1		5.46
A_4	Stage 1, Step 2	probable keywords	6.51
A_5	Stage 2, Step 1	found by the module	7.43
A_6	Stage 2, Step 2	at stage X and step Y (Fig. 8-Fig. 9)	8.35

The value of A_3 differs from the value of A_1 by 0.69 (by number, but not by content); respectively, A_4 from A_1 by 1.74; A_5 from A_1 by 2.66; A_6 from A_1 by 3.58. The value of A_2 differs from the value of A_3 by 4.36; respectively, A_2 from A_4 by 3.31; A_2 from A_5 by 2.39; A_2 from A_6 by 1.47. Adaptively changing the parameters/rules of the module almost doubles the collection of identified keywords (for example, the value of A_1 is greater than A_3 by 1.144654; A_6 by 1.750524; A_5 by 1.557652; A_4 by 1.36478). The total increase in the value obtained depending on the moderation of dictionaries is, respectively, for A_3 14.46541; A_4 – 36.47799; A_5 – 55.7652; A_6 – 75.05241. When comparing A_2 more than $A_3 \div A_6$, we have a chain of such values as 1.7985; 1.5084; 1.3217; 1,176. For different stages and steps of the experiment of processing the primary text, the average coincidence of the lists of identified keywords with the author's keywords varies in the range of 52.6-68.5%. The accuracy of matching keywords with the author's keywords ranges from 43.6 to 62.9%. The average match of meaningful keywords compared to all found by the system varies between 38.9-75.8%, depending on the stages of analysis of the text of the articles. The accuracy of matching keywords compared to all found by the system ranges from 34.3-71.9%, depending on the stages of analysis of article texts.

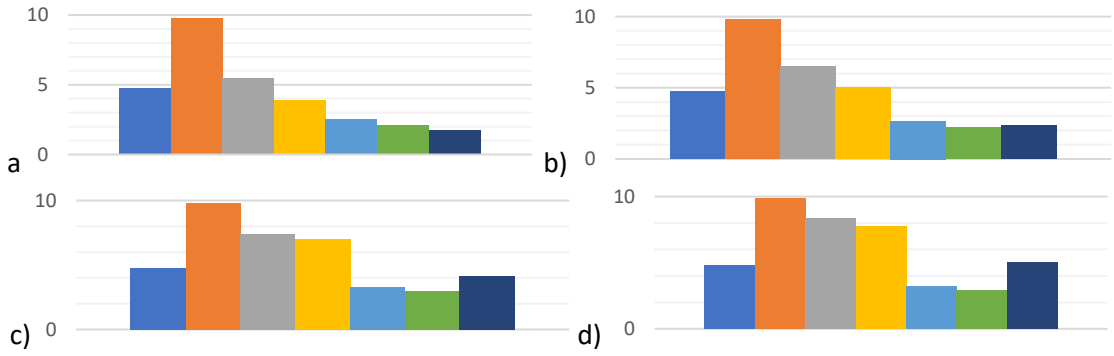


Figure 9: Arithmetic mean occurrence of significant words compared to the author's for a) stage 1, step 1, b) stage 1, step 2, c) stage 2, step 1 and d) stage 2, step 2, where blue - the author's keywords, orange - the number of words, grey - defined by the system, yellow - meaningful words, light blue - a match with the author's, green - the accuracy of the match, dark blue - additional words

For A_3, the module most often identified the number of keywords {5, 7, 3} (≥10), although the distribution of found keywords was within [1;18] words (except 17). For A_4, IS identified the number of keywords also {5, 7, 3} most often, although the distribution of found keywords is

within [1;18] (except 17), the number of identified words increased and the highest reliability index was achieved. For A_5, the module most often identified the number of keywords {7, 6, 5, 10, 8}, although the distribution of found keywords was within [2;14] (the range narrowed significantly). For A_6, the module most often identified the number of keywords {8, 5, 7, 10}, the distribution of identified keywords within [3;16] (accuracy improved). The accuracy of the definition of keywords increases in the process of the moderation of dictionaries and the ML-module. The difference between the number of keywords defined by the author and identified by the module at A_3 is 44.39919% (difference in %).

Table 4

Descriptive statistical data of keyword identification in experiments

Name	A_1	A_3	A_4	A_5	A_6
Average	4.808081	5.515152	6.565657	7.505051	8.434343
Standard error	0.180859	0.310393	0.39035	0.301297	0.324611
Median/ Mode	4/4	5/5	6/5	7/7	8/8
Standard deviation	1.799528	3.088371	3.883932	2.997869	3.229841
Sampling variance	3.238301	9.538033	15.08493	8.987219	10.43187
Excess	0.652815	1.705273	0.748643	-0.45645	-0.50438
Asymmetry	0.947939	1.125305	1.065716	0.537598	0.517047
Interval	8	16	17	12	13
Minimum/ Maximum	2/10	1/17	1/18	2/14	3/16
Sum	476	546	650	743	835
Score	99	99	99	99	99
Biggest(1)/ Smallest(1)	10/2	17/1	18/1	14/2	16/3
Reliability level (95.0%)	0.35891	0.615965	0.774637	0.597914	0.64418

Table 5

Statistical data of histogram construction for A_3 and $A_3 \div A_6$ (Fig. 10)

N	*n*	%	*N*	*n*	%	*n*	%	*N*	*n*	%	*n*	%	*N*	*n*	%
			A_1					A_3					A_4		
1	0	0.00	4	27	27.27	2	2.02	5	20	20.20	2	2.02	5	20	20.20
2	4	4.04	5	21	48.48	10	12.12	7	16	36.36	10	12.12	7	16	36.36
3	20	24.24	3	20	68.69	12	24.24	3	12	48.48	12	24.24	3	12	48.48
4	27	51.52	6	11	79.80	4	28.28	2	10	58.59	4	28.28	2	10	58.59
5	21	72.73	8	8	87.88	20	48.48	6	9	67.68	20	48.48	6	9	67.68
6	11	83.84	7	5	92.93	9	57.58	4	4	71.72	9	57.58	4	4	71.72
7	5	88.89	2	4	96.97	16	73.74	8	4	75.76	16	73.74	8	4	75.76
8	8	96.97	10	3	100.00	4	77.78	10	4	79.80	4	77.78	10	4	79.80
9	0	96.97	1	0	100.00	2	79.80	11	3	82.83	2	79.80	11	3	82.83
10	3	100.00	9	0	100.00	4	83.84	12	3	85.86	4	83.84	12	3	85.86
11	0	100.00	11	0	100.00	3	86.87	14	3	88.89	3	86.87	14	3	88.89
12	0	100.00	12	0	100.00	3	89.90	1	2	90.91	3	89.90	1	2	90.91
13	0	100.00	13	0	100.00	2	91.92	9	2	92.93	2	91.92	9	2	92.93
14	0	100.00	14	0	100.00	3	94.95	13	2	94.95	3	94.95	13	2	94.95
15	0	100.00	15	0	100.00	1	95.96	16	2	96.97	1	95.96	16	2	96,.97
16	0	100.00	16	0	100.00	2	97.98	18	2	98.99	2	97.98	18	2	98.99
17	0	100.00	17	0	100.00	0	97.98	15	1	100.00	0	97.98	15	1	100.00
18	0	100.00	18	0	100.00	2	100.00	17	0	100.00	2	100.00	17	0	100.00
More	0	100.00	More	0	100.00	0	100.00	More	0	100.00	0	100.00	More	0	100.00
			A_5					A_6							
1	0	0.00	7	15	15.15	0	0.00	8	14	14.14					
2	1	1.01	6	14	29.29	0	0.00	5	12	26.26					
3	5	6.06	5	13	42.42	1	1.01	7	11	37.37					
4	9	15.15	10	12	54.55	9	10.10	10	11	48.48					
5	13	28.28	8	11	65.66	12	22.22	4	9	57.58					
6	14	42.42	4	9	74.75	9	31.31	6	9	66.67					
7	15	57.58	12	6	80.81	11	42.42	9	9	75.76					
8	11	68.69	3	5	85.86	14	56.57	11	5	80.81					

9	4	72.73	14	5	90.91	9	65.66	14	5	85.86
10	12	84.85	9	4	94.95	11	76.77	12	4	89.90
11	1	85.86	13	3	97.98	5	81.82	13	4	93.94
12	6	91.92	2	1	98.99	4	85.86	15	3	96.97
13	3	94.95	11	1	100.00	4	89.90	16	2	98.99
14	5	100.00	1	0	100.00	5	94.95	3	1	100.00
15	0	100.00	15	0	100.00	3	97.98	1	0	100.00
16	0	100.00	16	0	100.00	2	100.00	2	0	100.00
17	0	100.00	17	0	100.00	0	100.00	17	0	100.00
18	0	100.00	18	0	100.00	0	100.00	18	0	100.00
More	0	100.00	More	0	100.00	0	100.00	More	0	100.00

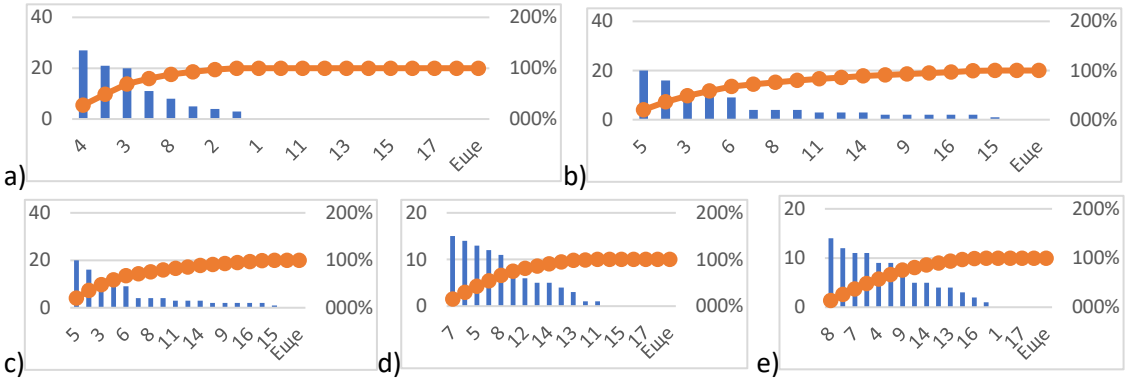


Figure 10: Histogram for sample a) A_1, b) A_3, c) A_4, d) A_5 and e) A_1

Accuracy improves with A_4 – 33.70672%, significantly improves with A_5 – 24.33809%, and with A_6 is 14.96945% (Table 4). Table 5 shows data from research articles when generating sets of keywords (Fig. 10). Analysis was performed for 100 filtered texts without metadata and unfiltered texts. The obtained average values for 100 filtered texts $\overline{Per_f} = 0{,}28$ and unfiltered $\overline{Per_0} = 0{,}19$ shows that such filtering of scientific articles improves the density of keywords by 1.48 times or by 47.83% (Fig. 11a). The obtained average values for 100 texts $\overline{Per_f^v} = 0{,}34$ and $\overline{Per_0^v} = 0{,}25$ taking into account the refinement of the thematic dictionary due to the addition of blocked words show that filtering with simultaneous moderation of the thematic dictionary improves keyword density by 1.35 times or by 35.44% (Fig. 11b).

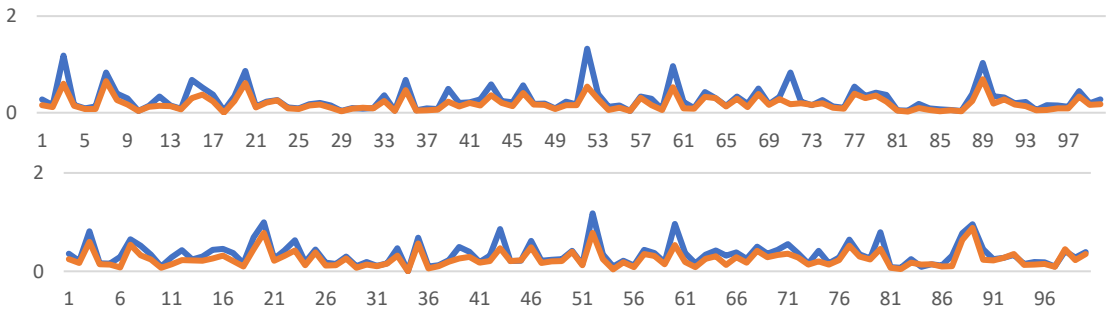


Figure 11: Percentage of keywords in the text as a result of checking articles without/with specifying the thematic dictionary (blue - filtered text, orange - general text)

A comparison of the values in the original author's text $\overline{Per_0} = 0{,}19$ and $\overline{Per_0^v} = 0{,}25$ without/with the refinement of the thematic dictionary, respectively, demonstrates the effectiveness of moderation of the thematic dictionary in the initial text - the density of keywords increases 1.34 times or by 34.33% (Fig. 12a). Values comparison in the filtered author's text $\overline{Per_f} = 0{,}28$ and $\overline{Per_f^v} = 0{,}34$ without/with the refinement of the thematic dictionary, respectively, demonstrates the effectiveness of the moderation of the thematic dictionary in the filtered text - the density of keywords increases 1.23 times or by 23.14% (Fig. 12b).

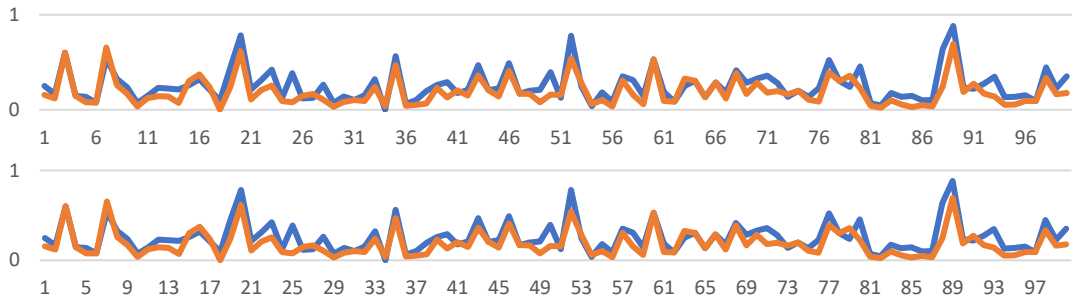


Figure 12: Percentage of keywords in the text as a result of checking primary articles with different dictionaries: a) for the general text and b) for the filtered text, where blue is specified by the dictionary and orange - without a specified dictionary

3.3. Analysis of methods for identifying stable phrases as keywords

The identification of stable phrases consists of the following stages: morphological analysis (MA), SYA, selection of key words and analysis of key phrases for stability (Fig. 13) [34-37].

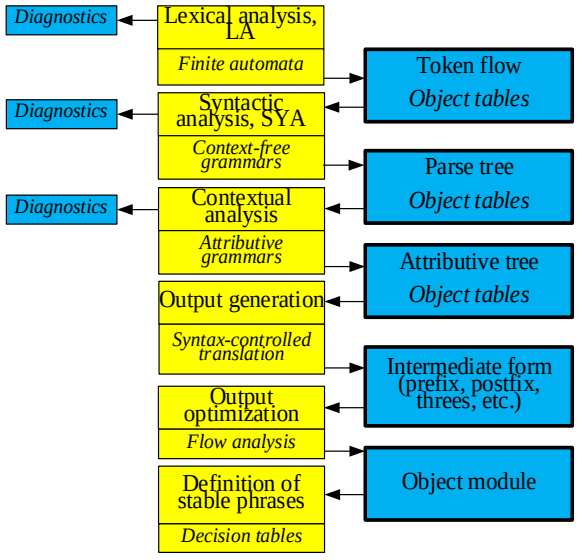
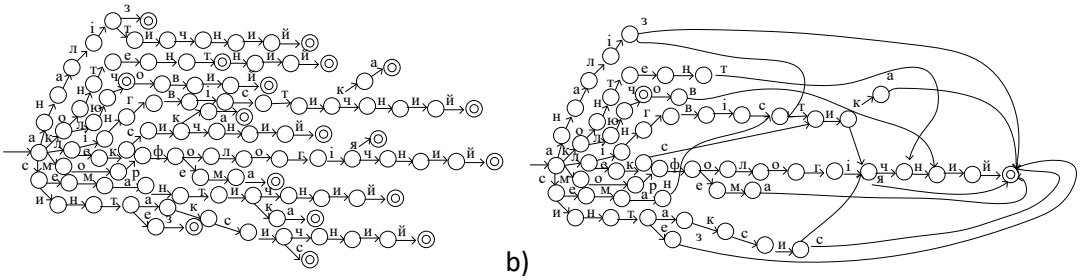


Figure 13: Identification of persistent phrases in Ukrainian-language texts

For Ukrainian-language texts, it is best to use a combination of procedural, tabular, and statistical stemming approaches. In the MA procedural approach, emphasis is placed on the use of ready-made dictionaries of bases and dictionaries of ready-made forms (DRF) in the analysis of words. Then the MA algorithm consists of the following steps: search in the SFG, base selection, and base search in the dictionary. The basis of most MAs of the Ukrainian language is a tree or Finite State Automata (FSA) (Fig. 14).



a) b)

Figure 14: MA results from storage methods: a) tree and b) FSA

The type of word is determined by the form of inflexions (Fig. 13). The algorithm works with individual words, so the content of the word is not taken into account. Parts of speech (adjective, noun, etc.) and categories of morphology (stem, suffix, etc.) are also unavailable. Variants of the rules for the stemming of Ukrainian words: short words remain unchanged, change during stemming (is an exception), do not change during stemming (is an exception), correspond to a regular expression, change the ending, has an unchanged ending, or the inflexion is cut off from the word. All this significantly complicates the keyword identification algorithm. Therefore, first of all, it is necessary to analyse widespread inflexions. Syntax - rules for combining words into correct expressions - word combinations and sentences (compare: programming language syntax). The task of the SYA (parser) is to construct the syntactic structure of the input sentence. Aspects of SYA implementation are dictionaries (information about individual language units); formal rules and interaction with neighbouring processing levels (morphological analysis, semantic analysis). Context-free grammar (CFG) rules are most often used in SYA: <*N*, *T*, *X*, *R*>, where *N* is a set of non-terminal symbols, *T* is a set of terminal symbols ($N \cap T = \emptyset$), *X* – axiom ($X \in N$), *R* is a set of transformation (substitution) rules of type $Y \to \alpha$, where $Y \in N$, α is a list of terminal and non-terminal symbols. CFG example:

$N = \{S, NP, PP, V, N, A\}$, S, T ={система, рубрикувати, україномовний, контент, за, ключовий, слово} [T ={systema, rubrykuvaty, ukrayinomovnyy, kontent, za, klyuchovyy, slovo}] (T ={system, categorize, Ukrainian-language, content, by, key, word}),

$R = \{S \to NPVP, S \to NPVPPP, NP \to AN, PP \to PNP, VP \to VNP, NP \to$ *система*$, V \to$ *рубрикувати*$, A \to$ *україномовний*$, A \to$ *ключовий*$, N \to$ *контент*$, N \to$ *слово*$, P \to$ *за*$\}$.

The disadvantage of using CFG is the periodic appearance of ambiguity with SYA, for example, "The system categorizes Ukrainian-language content by keywords" (Fig. 15). Examples of well-known SYA systems for English tests are: "Machinese Phrase Tagger" (Fig. 16) and VISL. There is

no online available information resource for SYA Ukrainian texts. "Ontology Matcher Demo" uses Machinese metadata to find ontology objects in the text (Fig. 17).

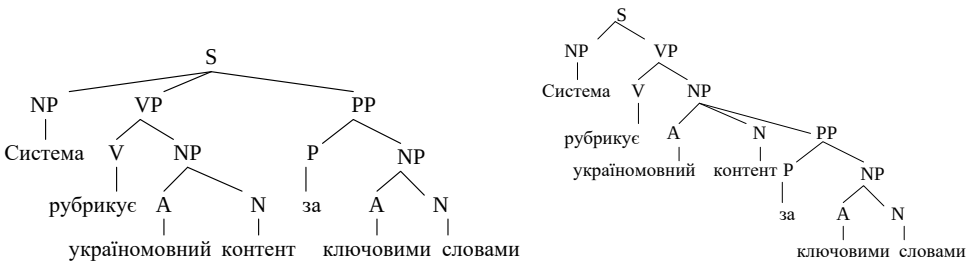


Figure 15: Examples of ambiguity in CFG

a)

Text	Baseform	Phrase syntax and part-of-speech
The	the	premodifier, determiner
train	train	nominal head, noun, single-word noun phrase
went	go	main verb, indicative past
on	on	adverbial head, adverb
up	up	preposed marker, preposition
the	the	premodifier, determiner
track	track	nominal head, noun, single-word noun phrase
out	out	adverbial head, adverb
of	of	preposed marker, preposition
sight	sight	nominal head, noun, single-word noun phrase
,	,	
around	around	preposed marker, preposition
one	one	nominal head, pro-nominal
of	of	postmodifier, preposition

b)

0	4	This	this	PRON		
5	2	is	be	V		
8	1	a	a	DET		
10	4	test	test	N	test	V

Figure 16: a) Machinese Phrase Tagger 4.9.1 analysis; b) Machinese Tokenizer

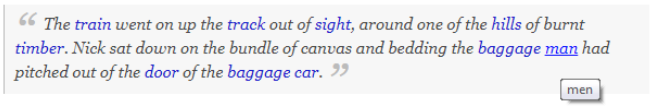


Figure 17: Ontology Matcher

Fig. 18-19 show SYA results on VISL information resource. Such informational resources do not exist for SYA Ukrainian-language texts. And the SYA process itself is quite cumbersome. For the input sentence: *Він зробив це так незручно, що зачепив образок мого ангела, який висів на дубовій спинці ліжка, і що вбита муха впала мені прямо на голову* [Vin zrobyv tse tak nezruchno, shcho zachepyv obrazok moho anhela, yakyy vysiv na duboviy spyntsi lizhka, i shcho vbyta mukha vpala meni pryamo na holovu] (He did it so awkwardly that it caught the picture of my angel that was hanging on the oak headboard and that the killed fly fell right on my head) example of SYA using pre-syntax (or Parsing by chunks - breaking the sentence into phrases, which do not intersect, (flat structure) $\neq$ a complete analysis, for example, (the boy (with the

hat)) ←→ (the boy) with (the hat)) of Ukrainian-language texts for the identification of stable word combinations when defining keywords is presented in Fig. 19.

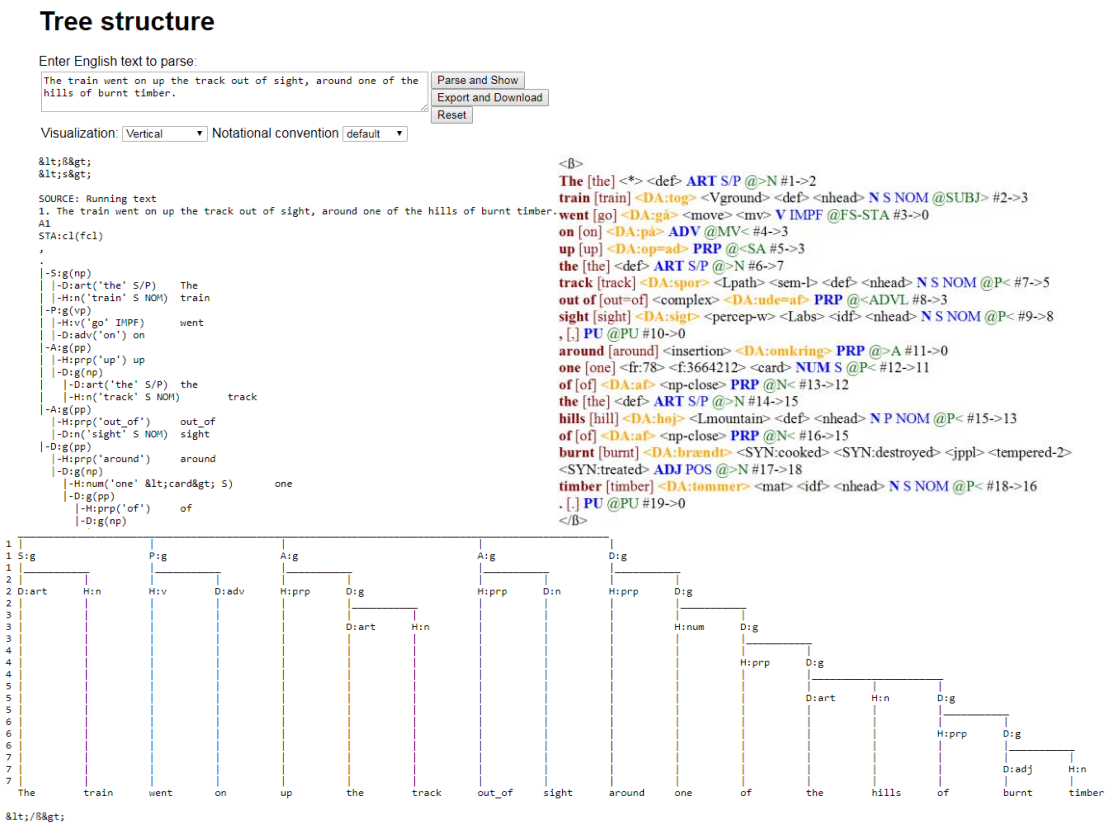


Figure 18: SYA Result on VISL

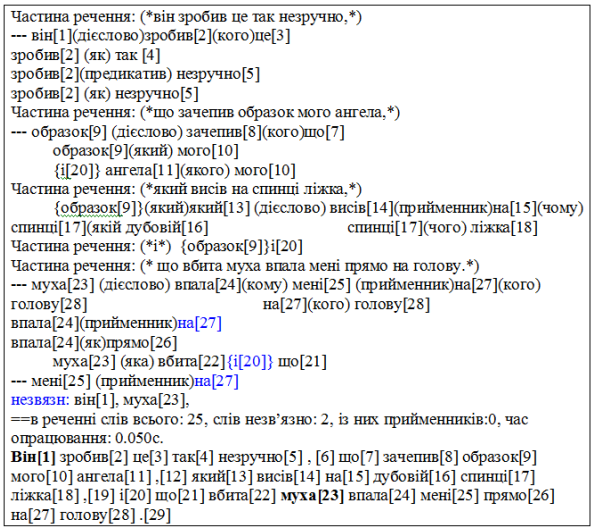


Figure 19: Result of SYA Ukrainian sentence

To select stable word combinations in the analysed texts and carry out their comparative analysis, we will use 4 different methods: FREG (frequency + morphological patterns, i.e. direct counting of the number of words); t-test; statistics χ^2; LR is the likelihood ratio.

Collocations is a word combination as a semantically and syntactically linguistic unit, where one part is chosen according to meaning, and the other depends on the first (for example, *ставити умови* [stavyty umovy] (to set conditions) – the choice of the verb *ставити* [stavyty] (to set) is determined by tradition and depends on the noun of *умови* [umovy] (the condition), with the word *пропозицію* [propozytsiyu] (offer) there will be another verb – *вносити* [vnosyty] (to enter)). This is a limited (selective) combination of words: phraseological units, idioms, proper names and trademarks. Collocations often include complex names (for example, *крейсер москва* [kreyser moskva] (moscow cruiser), *руський корабль* [rus'kyy korabl'] (russian ship), *безпілотник Байрактар* [bezpilotnyk Bayraktar] (Bayraktar drone), *від'ємний наступ* [vid"yemnyy nastup] (negative attack), *німецькі леопарди* [nimets'ki leopardy] (German leopards), *жест доброї волі* [zhest dobroyi voli] (goodwill gesture), etc.). Another name for the same phenomenon is stable phrases, N-grams. Examples of collocations –

- *Грати роль* [hraty rol'] (to play a role), *мати значення* [maty znachennya] (to have a meaning), *впливати* [vplyvaty] (to influence), *справляти враження* [spravlyaty vrazhennya] (to make an impression);
- *Засоби масової…* [zasoby masovoyi…] (means of mass...), *зброя масової…* [zbroya masovoyi…] (weapons of mass...), *вищий навчальний ….* [vyshchyy navchal'nyy ….] (higher education);
- *глибокий старець* [hlybokyy starets'] (deep old man) ↔ *поверхневий/мілкий невеликий юнак* [poverkhnevyy/milkyy nevelykyy yunak] (superficial/shallow little young man);
- *міцний чай* [mitsnyy chay] (strong tea) ↔ *сильний чай* [syl'nyy chay] (strong tea);
- *Кока-кола* [Koka-kola] (Coca-Cola), *Microsoft Windows*;
- *Гола Пристань* [Hola Prystan'] (Hola Prystan), *Нова Каховка* [Nova Kakhovka] (Nova Kakhovka), *Володимир Волинський* [Volodymyr Volyns'kyy] (Volodymyr Volynsky), *Володимир Зеленський* [Volodymyr Zelens'kyy] (Volodymyr Zelensky), *Нью Йорк* [N'yu York] (New York), *Стив Джобс* [Styv Dzhobs] (Steve Jobs).

1. The FREG method is a direct calculation of the frequency of use of pairs (threes). For example, FREG for the sentence *В літературі описано декілька підходів до автоматичного виділення стійких словосполучень* [V literaturi opysano dekil'ka pidkhodiv do avtomatychnoho vydilennya stiykykh slovospoluchen'] (In the literature, several approaches to the automatic selection of stable word combinations are described) «.» → *в літературі* [dekil'ka pidkhodiv] (in the literature); *літературі описано* [literaturi opysano] (described in the literature); *описано декілька* [opysano dekil'ka] (several are described); *декілька підходів* [dekil'ka pidkhodiv] (several approaches); *підходів до* [pidkhodiv do] (approaches to); *до автоматичного* [do avtomatychnoho] (to automatic); *автоматичного виділення* [avtomatychnoho vydilennya] (automatic selection); *виділення стійких* [avtomatychnoho vydilennya] (allocation of persistent); *стійких словосполучень* [stiykykh slovospoluchen'] (stable phrases). Unfortunately, as a result of using this method on large volumes of text, we get

the so-called "garbage" due to the high frequency of service words. The method also requires consideration of the frequency of occurrence and patterns of word combinations.

2. The t-test method consists of statistical hypotheses testing and MA statistical model using H_0: the words met by chance; $P(w^1w^2) = P(w^1)P(w^2)$; taking into account not only pairs but also the individual words use frequency (those that make up a pair); $t = \frac{\bar{x}-\mu}{\sqrt{\frac{s^2}{N}}}$, де $\bar{x}$ is empirical average, μ is theoretical average, s^2 is empirical dispersion, N is empirical sample size. The method is not completely correct for the language, but it allows to obtain results in practice, for example, the frequency of appearance of the stable phrase *контент аналіз* [kontent analiz] (content analysis) in [37] with $P(\textit{контент}) = 85/4338$ and $P(\textit{аналіз}) = 53/4338$ is $H_0: P(\textit{аналіз}) = P(\textit{контент})P(\textit{аналіз}) \approx 2{,}39 \cdot 10^{-4}$. In the Bernoulli scheme, $s^2 = p(1-p) \approx p$ at values of $\bar{x} = 18/4338$ and $t \approx 3{,}997955$.

3. Pearson's χ^2 method is applied to 2x2 tables (Table 6). Normality is not expected in the calculations. Example, $\chi^2 = \frac{N(O_{11}O_{22}-O_{12}O_{21})^2}{(O_{11}+O_{12})(O_{11}+O_{21})(O_{12}+O_{22})(O_{21}+O_{22})} \approx 286{,}0595$.

Table 6

An example of using Pearson's χ^2 method

w_i	$w_1 =$ контент	$w_1 \neq$ контент
$w_2 =$ аналіз	18 (контент аналіз)	35 (e.g., статистичний аналіз)
$w_2 \neq$ аналіз	67 (including, контент моніторинг)	4218 (including, статистичний моніторинг)

4. The LR method consists of the calculation of hypotheses ($p_1 >> p_2$)

$$H_1: P(w^2|w^1) = p = P(w^2|\neg w^1) \text{ and } H_2: P(w^2|w^1) = p_1 \neq p_2 = P(w^2|\neg w^1)$$

where $p = \frac{c_2}{N}$; $p_1 = \frac{c_{12}}{c_1}$; $p_2 = \frac{c_2-c_{12}}{N-c_1}$. Then, using the binomial distribution $b(m,n,p) = C_m^n p^m (1-p)^{n-m}$, we get the LR likelihood ratio

$$L(H_1) = b(c_{12}, c_1, p)b(c_2 - c_{12}, N - c_1, p),$$

$$L(H_2) = b(c_{12}, c_1, p_1)b(c_2 - c_{12}, N - c_1, p_2),\ log\ \lambda = {L(H_1)}\Big/{L(H_2)},$$

where $-2\,log\,\lambda$ is asymptotically distributed as χ^2. The term extraction experiment was conducted on 3 articles from different SAs. The template for experimenting is: [Adjective + Noun], [Adjective + Noun], [Noun + Noun, Genitive Distinctive], [Noun + Noun, Instrumental Distinctive], [Noun + '-' + Noun]. During the experiment, 6 methods were used: manually determined by the authors of the articles (A); determined by the Victana.lviv.ua system, taking into account Zipf's law (B); frequency+morphological patterns FREG (C); t-test (D); likelihood ratio LR (F); statistic χ^2 (G). An analysis of 3 articles in Ukrainian and translated into English was conducted (Table A -Table B of Appendix). Key words that occur in the results of all methods are highlighted in bold, in italics only in methods B-G, and underlined in methods A and C-G. When conducting a linguistic analysis, the following features were used to form alphabetic-frequency dictionaries of two words each:

- Bigrams were formed within the boundaries of punctuation marks (if there was at least some punctuation mark between the words - these words were not considered a 2gram);

- An alphabetic-frequency dictionary of two words was formed based on their bases (bigrams) and content analysis of these bigrams;
- When analysing the inflexions of the analysed words, verbs were not taken into account when forming the bigram alphabetic-frequency dictionary (verbs were considered one of the punctuation marks);
- Before the linguistic analysis of the texts, all stop words (participles, adverbs, conjunctions) and pronouns were removed.

Statistical methods allow taking into account the use of individual words. Subtleties are associated with applying the methods to different data volumes and probability ranges (better than t-test for larger *p* where normality is violated; likelihood ratio is better approximated by χ^2 than 2x2 tables for small volumes). It is more often used not for accepting/rejecting hypotheses, but for ranking candidate phrases. For comparison with the obtained results, we will use the library from Google - Word2Vec, which has proven itself as an alternative to TF-IDF (A_1 - Table C of Appendix). We will also use the built-in methods for searching for word combinations in Python. But it didn't work very well on these datasets, because it needs huge corpora to work well. The most interesting thing is that it allows you to do this after translating each word from the corpus into a space, the size of which is set by the user, for example,

'король' + *'жінка'* - *'чоловік'* = *'королева'* ['king' + 'woman' - 'man' = 'queen'] ('king' + 'woman' - 'man' = 'queen')

After translation into a space of a certain dimension, each word becomes a vector, so you can use them to form basic operations of addition, subtraction, multiplication, etc. We will also consider the analysis through bigrams (A_2 – Table C of Appendix) and skip grams (A_3 – Table C of Appendix). The results are better than Word2Vec, namely the analysis of skipgrams with a value of 3 and also the cleaning of stop words in English were the best (A_4 – Table C of Appendix). However, these results are quite far from those obtained in Table A of the Appendix. The result is worsened by not taking into account punctuation marks and the use of stop words in the linguistic analysis as meaningful.

3.4. Parametric classification of the text in Ukrainian

When classifying the text, the definition of the grammatical meta-data of the word is implemented based on grapheme/morphological analysis (Fig. 20, algorithm 3) [38-41].

Algorithm 3. Thematic classification of Ukrainian-language content

Stage 1. Splitting the Ukrainian-language text C_r into parts (paragraphs/paragraphs, etc.).

Step 1. Loading into the C_r text tree generation module.

Step 2. Formation of a new array of tapes in the structure.

Step 3. Parsing of strings of symbols of parts of the text C_r.

Step 4. Identify the period as the end of a sentence, not part of the contraction and go to step 5, otherwise store it in an array and go to step 3.

Step 5. Identification of the end-of-text character and go to step 6, otherwise mark the end of a part of the text and go to step 2.

Step 6. Saving the tree of parts of text C_r as a structure $U_{CT}^B \in U_{CT}$.

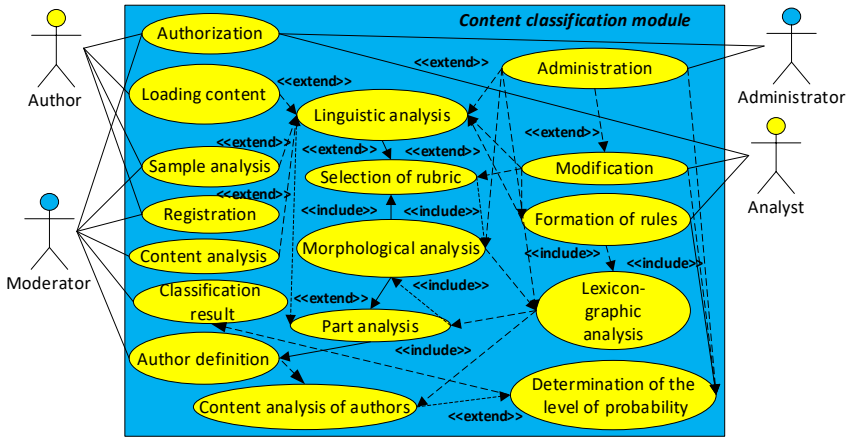


Figure 20: Diagram of text classification use cases

Stage 2. Splitting the part into expressions while preserving the structure of the text C_3.

Step 1. Analysis of the new structure of part of the text $U_{CT}^{B} \in U_{CT}$. Formation of the structure of the expression (paragraph/sentence, etc.) $U_{CT}^{R} \in U_{CT}$ with the ID_part key of type *n-to-*1 with the structure of text parts C_r.

Step 2. Formation of a new array in the structure of sentences $U_{CT}^{R} \in U_{CT}$.

Step 3. Parsing characters to the next punctuation mark.

Step 4. If the abbreviation or special entry (date, money, etc.) is according to the regular expression, then the corresponding marking of this sequence and the transition to step 5, otherwise, saving in the structure $U_{CT}^{R} \in U_{CT}$ and transition to step 2.

Step 5. If the end of the text part, then mark and go to step 6, otherwise go to step 2.

Step 6. Saving a tree of sentences in the form of a $U_{CT}^{R} \in U_{CT}$ structure.

Step 7. If the end of the text, then go to step 3, otherwise go to step 1.

Stage 3. Splitting sentences into lexemes while preserving the connection with the corresponding sentence $U_{CT}^{L} \in U_{CT}$ and, accordingly, the number of the position in the sentence.

Step 1. Formation of the lexeme structure $U_{CT}^{L} \in U_{CT}$ with the fields ID_lex, ID_sent, N_lex, T_lex as a description of the lexeme meta-data.

Step 2. Analysis of the sentence lexeme with $U_{CT}^{R} \in U_{CT}$.

Step 3. Formation of a new lexeme in the lexeme structure $U_{CT}^{L} \in U_{CT}$.

Step 4. Parsing characters up to the first character not from the Ukrainian alphabet or an apostrophe and saving tokens in the structure.

Step 5. If the end-of-sentence character, then go to step 6, otherwise go to step 3.

Step 6. Syntax analysis based on algorithms 2.

Step 7. Morphological analysis based on received lexeme chains.

Stage 4. Identification of the topic of the Ukrainian-language text $U_{CT}^{T} \in U_{CT}$.

Step 1. Identification of the hierarchical structure of features $U_{CT}^{T} \in U_{CT}$ of each semantically significant lexeme from the noun group, except for pronouns.

Step 2. Generating a dictionary with a hierarchy of token property types.

Step 3. Unification, if necessary, of similar tokens.

Step 4. Identification of a set of key words $KeyWords$ of the text $C'_r = \alpha_r(\alpha_m(C_r, U_K), U_{CT})$ with $U_{CT} = \{U_{CT1}, U_{CT2}, U_{CT3}, U_{CT4}\}$, where U_{CT} is a collection of classification conditions, U_{CT1} is a set of thematic keywords, U_{CT2} is a set of frequencies of occurrence of keywords, U_{CT3} is dependencies

of the occurrence of keywords according to different topics, U_{CT4} is frequencies of occurrence of thematic keywords.

Step 5. Formation of $U_{Ct}^{T} \in U_{Ct}$ in the set of $KeyWords$ with $TKeyWords$ (thematic keywords) for $Topic$ and $Category$.

Step 6. Calculation of $QuantitativeryTKey$ (frequency of occurrence of thematic keywords) and $FKeyWords$ (frequency of occurrence of keywords), as well as coefficients $Static$ (statistical importance of terms), $CofKeyWords$ (thematic keywords of the content), $Comparison$ (occurrence of keywords of different topics), $Addterm$ (measures of the presence of additional thermal baths).

Step 7. Calculation If there is a match between content keywords and topic keywords, then go to step 9, otherwise go to step 8.

Step 8. Generation of a new rubric with a set of key terms of the text C'_r.

Step 9. Assignment of the terms of the analyzed text C'_r to a certain class on the topic.

Step 10. Calculation of $Location$ – content weight factor C'_r in the topic.

Stage 5. Filling with meta-data of Ukrainian-language analyzed text for attributes $Topic$, $Category$, $Location, Static$, $Addterm$, $CofKeyWords$, $TKeyWords$, $FKeyWords$, $Comparison$, $QuantitativeryTKey$.

3.5. Detection of content duplication/plagiarism/rewriting

When identifying duplicate text content (for example, when identifying plagiarism/rewrites or duplicates of integrated content from different sources), the main NLP task is to analyse the degree of similarity of lines. It can also be used for spell checking or text input autocorrect as an intuitive prediction of what exactly the user wants to type. Another example is the identification of the key meaning of the text content or the determination of whether the two lines *Національний університет* «*Львівська політехніка*» [Natsional'nyy universytet «L'vivs'ka politekhnika»] (National University "Lviv Polytechnic") or *НУ* «*Львівська політехніка*» [NU «L'vivs'ka politekhnika»] (NU "Lviv Polytechnic") are the same keyword. The minimum editing distance allows us to quantify the assumption about the similarity of the analyzed strings as the calculation of the minimum number of editing operations through insertion (*i*), deletion (*d*), substitution (*r*), synonymization (*s*), permutation (*p*) necessary to transform one string into another (Fig. 21). An empty string/character alignment is a match between substrings of two sequences of strings/sentences/words.

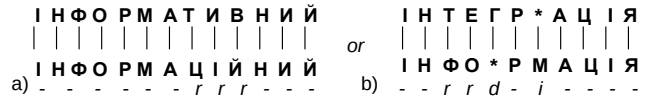


Figure 21: Scheme of the analysis of the minimum editorial distance

Each of these operations is assigned a certain value/weight. The Levenshtein distance between two lines is the simplest weighting factor in which each of the five operations has a value of 1 [42]. The Levenstein distance for the scheme Fig. 21a is equal to 3, and for the scheme, Fig. 21b equals 4. An alternative metric is where each insertion/deletion is scored as 1, and other operations are not allowed or are scored as 2 (*r*), 3 (*p*), and 4 (*s*), respectively. Then for the

schemes Fig. 21 Levenshtein distances are equal to 6 each. The process of finding the minimum editorial distance (Fig. 22a) consists of finding the shortest path - a sequence of edits from one-character line to another (Fig. 22b) based on dynamic programming [43-46].

Figure 22: Finding the editorial distance and an example of an edit path

Algorithm 4. Minimum editorial distance based on [47].

Stage 1. We define $S[0,0] = 0, n = const, m = const, i = 0, j = 0$.

Stage 2. We parse the text X and extract a string of length n for comparison. We denote the input string as A ($|A| = n$). We define $S[i,0] = S[i-1,0] + d - f_m(A[i])$.

Stage 3. We parse the text Y and select a string of length m for comparison. We denote the target string as B ($|B| = m$) . for comparison. We denote the target string as $S[0,j] = S[i,j-1] + i - f_m(B[j])$.

Stage 4. Calculation of the minimum editorial distance between two lines.

Step 4.1. We identify $S[i,j]$ as the editing distance between $A[1 \dots i]$ and $B[1 \dots j]$, that is, the distance between A and B є $S(n,m)$.

Step 4.2. We calculate $S[i,j]$ by taking the minimum of five possible paths through the matrix of reaction distances (Fig. 23):

$$S[i,j] = min \begin{cases} S[i-1,j] + d - f_{dm}(A[i]) \\ S[i,j-1] + i - f_{im}(B[j]) \\ S[i-1,j-1] + r - f_{rm}(A[i], B[j]) \\ S[i-1,j-1] + p - f_{pm}(A[i+1], A[i]) \\ S[i-1,j] + s - f_{sm}(\mathrm{A}[i], \mathrm{X}[j]) \end{cases}$$

A\B	#	І	Н	Ф	О	Р	М	А	Ц	І	Я
#	0	1	2	3	4	5	6	7	8	9	10
І	1	0	1	2	3	4	5	6	7	8	9
Н	2	1	0	1	2	3	4	5	6	7	8
Т	3	2	1	2	3	4	5	6	7	8	9
Е	4	3	2	3	4	5	6	7	8	9	10
Г	5	4	3	4	5	6	7	8	9	10	11
Р	6	5	4	5	6	5	6	7	8	9	10
А	7	6	5	6	7	6	7	6	7	8	9
Ц	8	7	6	7	8	7	8	7	6	7	8
І	9	8	7	8	9	8	9	8	7	6	7
Я	10	9	8	7	8	9	10	9	8	7	6

Figure 23: An example of a matrix for calculating the minimum editorial distance

If the values of the weights for the specified operations are known in advance, then we calculate how:

$$S[i,j] = min \begin{cases} S[i-1,j]+1 \\ S[i,j-1]+1 \\ S[i-1,j-1] + \begin{cases} 0, if\ A[i] = B[j] \\ 2, if\ A[i] \neq B[j] \end{cases} \\ S[i-1,j] + \begin{cases} 0, if\ A[i] = A[i+1] \\ 3, if\ A[i] \neq A[i+1] \end{cases} \\ S[i-1,j] + \begin{cases} 0, if\ A[i] = X[j] \\ 4, if\ A[i] \neq X[j] \end{cases} \end{cases}$$

Completing the matrix for calculating the minimum editorial distance (Fig. 24).

Stage 5. If not the end of the text Y, then $j = j + 1$, parse the text Y, select the next line of length m for comparison and go to step 4.

Stage 6. If it is not the end of the text X, then $i = i + 1$, we parse the text X, select the next line of length n for comparison and go to step 3.

Stage 7. Determination of the optimal shortest path - the sequence of edits from one character line to another (Fig. 24) in the calculation matrix of the minimum editorial distance (Fig. 23) [48].

A\B	#	І	Н	Ф	О	Р	М	А	Ц	І	Я
#	**0**	← 1	← 2	← 3	← 4	← 5	← 6	← 7	← 8	← 9	← 10
І	↑ 1	↖ **0**	← 1	← 2	← 3	← 4	← 5	← 6	← 7	← 8	← 9
Н	↑2	↑ 1	↖ **0**	← 1	← 2	← 3	← 4	← 5	← 6	← 7	← 8
Т	↑ 3	↑ 2	↑ 1	↖ **2**	← 3	← 4	← 5	← 6	← 7	← 8	← 9
Е	↑ 4	↑ 3	↑ 2	↑ 3	↖ **4**	← 5	← 6	← 7	← 8	← 9	← 10
Г	↑ 5	↑ 4	↑ 3	↑ 4	↑ **5**	← 6	← 7	← 8	← 9	← 10	← 11
Р	↑ 6	↑ 5	↑ 4	↑ 5	↑ 6	↖ **5**	← **6**	← 7	← 8	← 9	← 10
А	↑ 7	↑ 6	↑ 5	↑ 6	↑ 7	↑ 6	↑ 7	↖ **6**	← 7	← 8	← 9
Ц	↑ 8	↑ 7	↑ 6	↑ 7	↑ 8	↑ 7	↑ 8	↑ 7	↖ **6**	← 7	← 8
І	↑ 9	↑ 8	↑ 7	↑ 8	↑ 9	↑ 8	↑ 9	↑ 8	↑ 7	↖ **6**	↖ 7
Я	↑ 10	↑ 9	↑ 8	↖ 7	↑ 8	↑ 9	↑ 10	↑ 9	↑ 8	↑ 7	↖ **6**

Figure 24: An example of determining the minimum distance calculation path

Step 7.1. We define and store sequentially in each cell $S[i,j]$ the matrix for calculating the minimum editorial distance of 1-3 back pointers (from the left, above and/or diagonally) to the previous cell ($S[i-1,j]$, $S[i,j-1]$ and/or $S[i-1,j-1]$) з from which it is possible to move to the current cell (Fig. 5.28), without disturbing the change of editorial distance.

Step 7.2. Analysing from the last cell $S[n,m]$, we move through the matrix in reverse directions to $S[0,0]$, without disturbing the change in the sequence of edits and determining the shortest path of the editorial distance.

Each cell in bold represents the alignment of a pair of letters across two lines. If two adjacent cells are highlighted in one row, then the insertion operation from the source to the target is implemented, for example, the letter М after Р (5→6); two bars in a row, located in one column, indicate the deletion, for example, of the letter Г after Е (replaced before that with Ф, i.e. 4→5).

Similarly, the minimum distance algorithm can be applied to words in a sentence (plagiarism/rewrite check, speech loss calculation, machine translation) instead of symbols in a line (spell check, word error frequency calculation). For example, for spelling correction, substitutions are likely to occur between letters of the corresponding natural language located next to each other on the keyboard. The Viterbi algorithm [49] is the best option for calculating the minimum editorial distance, calculating the maximum probability of alignment of one line with another. To recognize text content as a duplicate/plagiarism or a partial rewrite, it is enough to compare the character chains of the template and analogues to find the minimum

distance. This is not enough to recognize content with a significant rewrite. Then the recognition will consist of the identification of a collection of concepts and terms of the corresponding template text based on the calculation of the degree of similarity to probable analogues of the textual content [34-49]. The collection of identified concepts and terms is supplemented from the ontology with others based on generalized relations of the IS-A type one level up and with other semantic relations whose importance weight exceeds the threshold value. Relationships between concepts and terms in the content are identified to eliminate the ambiguity of recognition to form a connected graph of the semantic image of the corresponding content. Similarity comparison results from the calculation of the semantic distance between the corresponding content (Fig. 25). The process of content comparison and similarity ranking using an ontology with text string search by pattern includes [34-49]:

1. Weighted conceptual graph G of template textual content.
2. A weighted conceptual graph G' supplemented with a content-template ontology with finding the parent of each vertex of G based on the connections between concepts.
3. Weighted conceptual graph $\hat{G} = G \cup G'$ based on SYA and SEM results.
4. Reductions of redundant elements of the weighted conceptual graph $\hat{G}$.
5. Calculation of weight centres (Fig. 28) and semantic distance between G and G'.

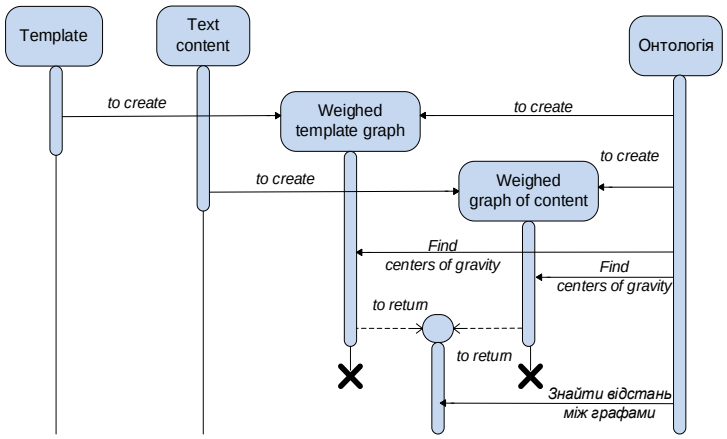


Figure 25: Sequence diagram of semantic content comparison

According to experimental testing with abstracts of scientific and technical publications, the approach based on adaptive ontology increases the accuracy of content similarity search by an average of 18% compared to the method of weighted conceptual graphs (Montez-Gómez) and 27% compared to the method based on the Dice coefficient (Fig. 26). The analysis of the effectiveness of the listed methods was carried out according to the search accuracy parameter: accuracy = (number of relevant ones found by the expert)/(number of relevant ones found by the program). The Dice coefficient method identified 63% of similar content to the template only those abstracts of scientific and technical publications, where there is the largest number of common words with the template, but did not always correlate with the content of the

prototype. At that time, the method based on adaptive ontology gave the best result considering the similarity of the template context and analogues.

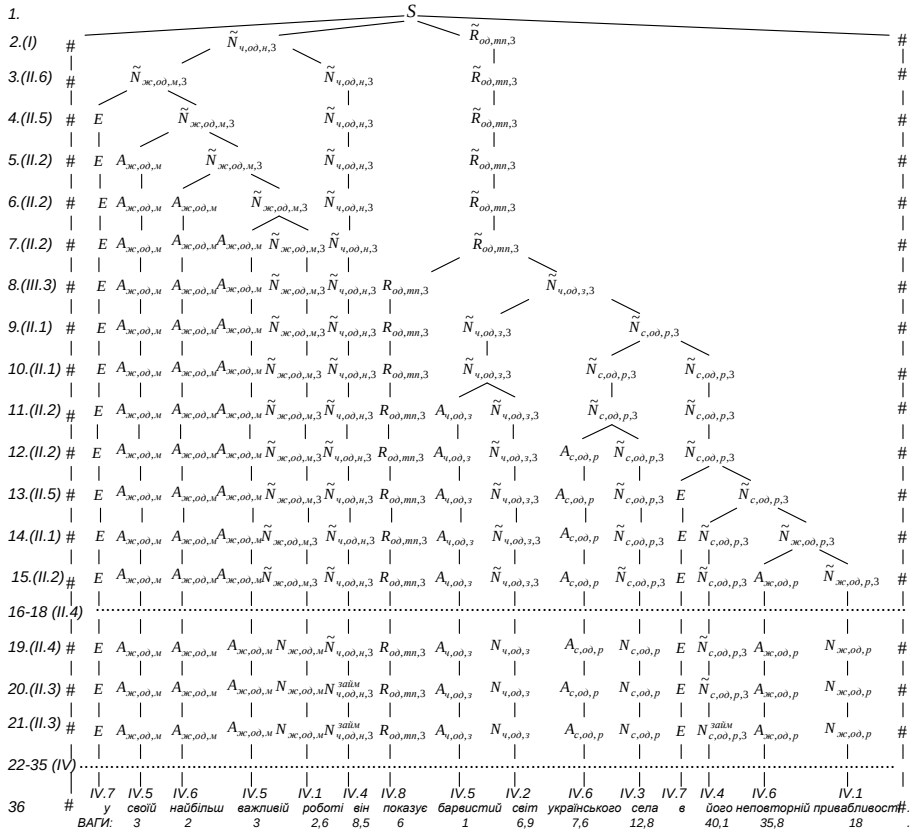


Figure 26: The result of SYA and SEM for a sentence in Ukrainian

Table 7
Comparison of methods

Method name	Accuracy χ, %
based on an adaptive ontology	90
weighted conceptual graphs (Montez-Gómez)	72
by the Dice coefficient	63

3.6. Ukrainian text processing technology for the identification of personal signs of the content author

3.6.1. Features and typical features of the author's text

Analysis of changes in the dynamics and frequency of appearance of a linguistic unit in the text is of great importance in linguistic statistics. The study of the coefficients of personal features of the author's style (Alg. 5) is based on calculations and analysis [50-51]:

- the author's text concentration degree ($I_{kt} = W_{10}/W$): the ratio of the number of words with an absolute frequency of appearance in the text ≥10 to the number of all words;
- the degree of exclusivity of the author's text ($I_{wt} = W_1/W$): the ratio of the number of words with an absolute frequency of occurrence equal to 1 to the number of all words;
- the author's speech coherence degree ($K_z = (Z + S)/(3P)$): the operative words occurrence proportion in separate sentences of Ukrainian-language textual content;
- the syntactic complexity degree of the author's speech ($K_s = 1 - P/W$): sentence number dependence in the test on the number of words (not the total number of words);
- the degree of lexical diversity of the author's speech ($K_l = W/N$): the proportion of the vocabulary of words from the text to the total volume of all words.

Algorithm 5. Study of personal features of the author's style

Stage 1. We integrate from reliable sources, use parametric filtering (eliminating information noise, such as tags, pictures, etc.), and format the Ukrainian-language text (e.g., eliminating apostrophes or replacing them with one type, eliminating them). The way the selection is organized and the size of the text sample is important: it should be at least 18 thousand words to determine the characteristics.

Stage 2. Lemmatization of Ukrainian-language text content.

Stage 3. Elimination of heterogeneity of linguistic units (for example, converting abbreviations to full text or numerical values).

Stage 4. Generating frequency dictionaries of Ukrainian-language text content based on statistical distribution in the required numerical metrics.

Stage 5. Identification/calculation of coefficients/indices of personal features of the author's style based on frequency dictionaries, for example, analysis of the share and peculiarities of the appearance of service/stop/marked words, punctuation marks, words/ sentences/ paragraphs/ chapters/ sections of different lengths, etc.

Stage 6. Analysis of coefficients/indices for accuracy and reliability.

Stage 7. Lexical and statistical modelling of the author's style features distribution.

Stage 8. Generation of author's style templates within a certain genre or topic in a certain period.

Stage 9. Experimental testing to train the system for assessing the level of belonging of Ukrainian-language texts of a certain genre/topic to a particular author's style template.

3.6.2. Determining the Ukrainian-language text's author style based on linguometry, stylometry and glottochronology technologies

Each language is characterized by a set of service words (particle, conjunction, and preposition - Table 8 - more than 70 words), and the author's style is influenced by the peculiarities of everyday speech, in particular, by the use of these words. For example, some authors prefer the word *однак* [odnak] (however), others prefer the word *отже* [otzhe] (hence), or, ignoring the rules of the Ukrainian language, they often prefer one of the conjunctions like *і* [i] (and), *та* [ta] (and), *й* [y] (and). Some prefer the preposition *тобто* [tobto] (that is), while others prefer its analogues. Analyzing and comparing the appearance and frequency of stop words as service words (there are also parasitic words characteristic of a particular author for expressing a certain topic, slang, etc.) makes it possible to model the lexical and statistical pattern of a particular author's style. Fig. 27 presents a graphical representation of the relative frequency of the

appearance of stop words in four different texts (Excerpts 1-4) and the template (Etalon) based on the statistical data of the appearance of the official word (Table B of Appendix). The results of the analysis of the four texts (Table 9) show that it is more likely that Excerpt 4 belongs to the author of the template (although there is not a significant difference between the results of the study of texts 4 and 2, if they were written in the same period, they do not belong to the author of the template, if in different periods with the template, the probability of belonging to this author increases).

Table 8

Service parts of the Ukrainian language (stop words)

Part of speech	List of stop words
Prepositions	без, біля, близько, в, вглиб, від, для, до, з, за, з-за, з-під, крізь, на, над, під, по, поза, при, про, проміж, у, через [bez, bilya, blyz'ko, v, vhlyb, vid, dlya, do, z, za, z-za, z-pid, kriz', na, nad, pid, po, poza, pry, pro, promizh, u, cherez] (without)
Connectors	а, або, але, й, і, коли, немов, одначе, проте, та, та й, так, також, тобто, через те що, хоча, чи, що, щоб, якщо [a, abo, ale, y, i, koly, nemov, odnache, prote, ta, ta y, tak, takozh, tobto, cherez te shcho, khocha, chy, shcho, shchob, yakshcho] (and, or, but, and, and, when, as if, however, but, and, and, so, also, that is, because of that, although, or, that, so that, if)
Particles	або, адже, аякже, би, вже, ж, же, ледве чи, лише, мов, немов, навіть, не, ні, он, ось, так, тільки, то, тобто, уже, це, чи [abo, adzhe, ayakzhe, by, vzhe, zh, zhe, ledve chy, lyshe, mov, nemov, navit', ne, ni, on, os', tak, til'ky, to, tobto, uzhe, tse, chy] (or, because)

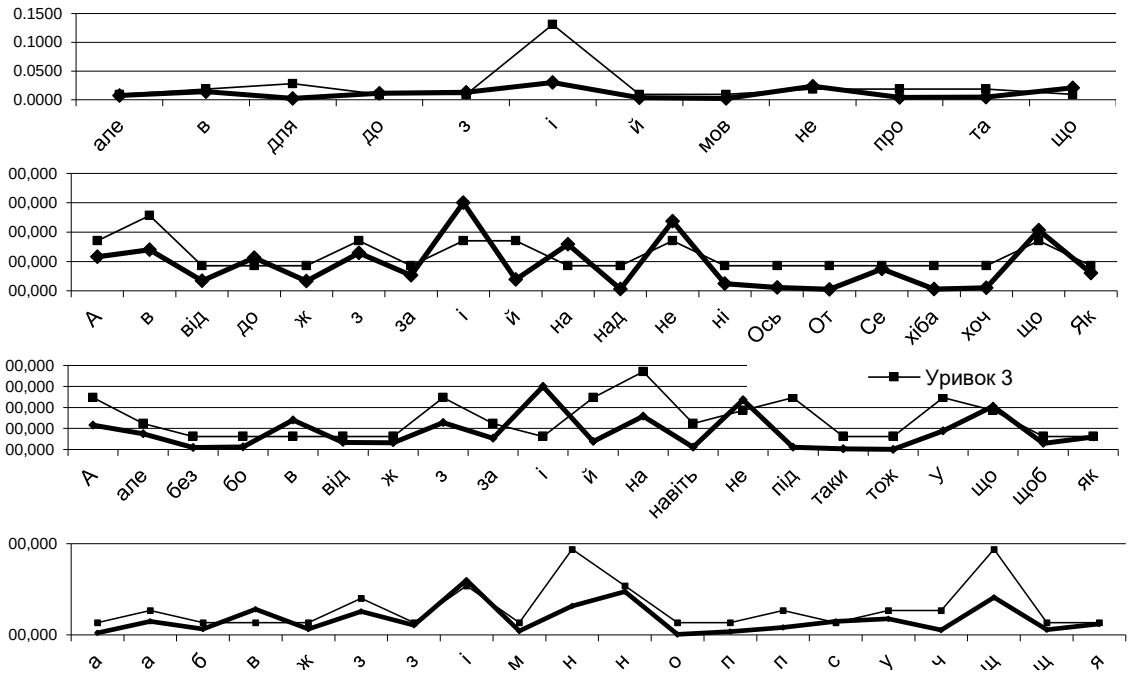


Figure 27: Probability of stop words (correlation coefficient – $R_{e\text{-}Y1}$=0,6076; $R_{e\text{-}Y2}$=0,7066; $R_{e\text{-}Y3}$=0,2810; $R_{e\text{-}Y4}$=0,7326), where the thick line is the standard, and the thin line is the excerpt, respectively 1-4 for each diagram separately

Table 9

Correlation coefficients for stop words

N	$R_{e\sim U}$	Particle	Conjunction	Preposition	$R'_{e\sim U}$
4	**0.7326**	**0.9594**	**0.9544**	**0.5639**	**0.6905**

2	0.7066	0.9580	0.5714	0.4928	0.4913
1	0.6076	1	0.79	0.72	0.6900
3	0.2810	0.8800	0.1624	0.1517	0.2254

Thus, the application of the anchor word method yielded the following results: among the studied passages, the passage most likely to belong to the standard was indeed the one authored by the author of the standard. Other results also confirm the effectiveness of the method of reference words in the attribution of texts. Thus, in the first study, the next highest probability of belonging to the standard is a passage from another work by the same author. Excerpt 1, which also belongs to the standard, "lost" to Excerpt 4 by only one-tenth in the correlation coefficient. The result for Excerpt 3, which is separated from the standard by about a hundred years, is also adequate. The assumption that the influence of the proportion as a method parameter on the results is insignificant led to a decrease in the correlation coefficients. Moreover, the difference between the correlation coefficients for Excerpt 1 and Excerpt 4 decreased significantly and amounted to 0.0005. Nevertheless, more research is needed to confirm or refute the fact that particles are not a determining factor in authorial style. To achieve the research goal, a module with the ability to select the language/languages of the analyzed content was developed and implemented on a Web resource (Fig. 28). Experimental testing of the functioning of the module for identifying and analyzing a collection of service words from 100 scientific and technical publications was carried out in 3 stages (Alg. 6).

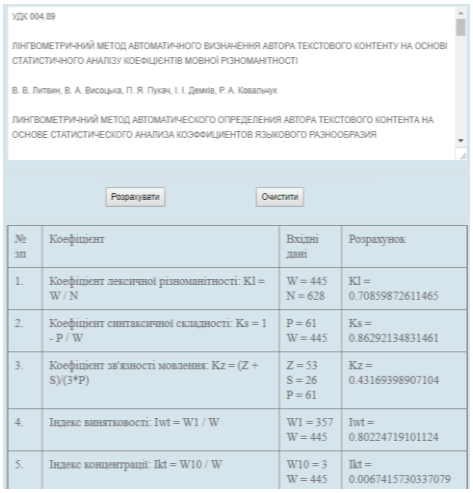


Figure 28: An example of text analysis on http://victana.lviv.ua/nlp/linhvometriia

Algorithm 6. Analysis and interpretation of linguistic and statistical studies of the author's style of speech identification

Stage I. Lexical analysis of the text to identify stop words and calculate the coefficients of lexical author's speech (text diversity).

Step 1. Filtering Ukrainian-language text content from information noise (special characters, pictures, tags, numbers, formulas, etc.).

Step 2. Sizing text content - excess is cut off.

Step 3. Identification of sentence length *P* in Ukrainian-language text content.

Step 4. Identification of the number of words *N* in Ukrainian-language text content.

Step 5. Volume identification by frequency dictionary of word bases *W*.
Step 6. Identification of the volume of W_1 words used exactly once in the text.
Step 7. Identification of the volume of W_{10} words used ≥10 times in the text.
Step 8. Identification of the volume of *Z* prepositions in text content.
Step 9. Identification of the volume of conjunctions *S* in text content.
Step 10. Calculating the degree of exclusivity of text content: $I_{wt}=W_1/W$.
Step 11. Calculating the degree of concentration of text content: $I_{kt}=W_{10}/W$.
Step 12. Calculating the degree of coherence of a text: $K_z=(Z+S)/(3*P)$.
Step 13. Calculating the degree of syntactic complexity of a text: $K_s=1–P/W$.
Step 14. Calculating the degree of lexical diversity of a text: $K_l=W/N$.
Step 15. Table presentation of the results at http://victana.lviv.ua/nlp/linhvometriia.
Stage II. Determining the author's style using stylometry methods.
Step 1. Checking the lengths of the reference text and the selected passages and adjusting the length of the reference text to the minimum length checked.
Step 2. Cleaning the reference text from special characters, etc.
Step 3. Determining the number of words in the reference text.
Step 3. Determination of the number of stop words (prepositions + conjunctions + particles) in the reference text.
Stage III. Analysing the text by the method of glottochronology according to the Swodesh list.

An example of the result of lexical analysis of one Ukrainian-language textual content to identify stop words and calculate the coefficients of lexical authorial speech (text diversity) is presented in Table 10. For stage III, the main task is to determine the number of words from Swodesh's 200-word list that appear in the works of different periods and to determine the percentage of such words in the passages. We will also investigate the number of common words from Swodesh's list for the selected passages. We will select passages written several years apart. Let the passages consist of, for example, 250 words, not including the title and proper names. A comparison of the 200-word list of Swadesh and Passage 1 is presented in Table 11 (common words are highlighted in colour).

Table 10

An example of analyzing the author's style of speech

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

Table 11

Words from Swodesh's list

N	Word	AF	RF	Word	AF	RF	Word	AF	RF
	Excerpt 1			Excerpt 2			Excerpt 3		
1	*і* [i] (and)	19	0.2500	*і*[i] (and)	6	0.1224	*і*[i] (and)	10	0.2174
2	*що* [shcho] (what)	6	0.0789	*що*[shcho] (what)	3	0.0612	*що*[shcho] (what)	4	0.087
3	*з*[z] (with)	5	0.0658	*з*[z] (with)	2	0.0408	*з*[z] (with)	2	0.0435
4	*все*[vse] (all)	4	0.0526	*все*[vse] (all)	4	0.0816	*все*[vse] (all)	3	0.0652
5	*в*[v] (in)	4	0.0526	*в*[v] (in)	7	0.1429	*в*[v] (in)	4	0.087
6	*на*[na] (on)	3	0.0395	*на*[na] (on)	1	0.0204	*на*[na] (on)	1	0.0217
7	*там*[tam] (there)	3	0.0395	*там*[tam] (there)	2	0.0408	*там*[tam] (there)	1	0.0217

8	*ні*[ni] (no)	3	0.0395	*ні*[ni] (no)	1	0.0204	*ні*[ni] (no)	7	0.1522
9	*знати*[znaty] (know)	2	0.0263	*знати*[znaty] (know)	2	0.0408	*знати*[znaty] (know)	2	0.0435
10	*який*[yakyy] (which)	2	0.0263	*який*[yakyy] (which)	4	0.0816	*який*[yakyy] (which)	1	0.0217
11	*ви*[vy] (you)	1	0.0132	*ви*[vy] (you)	1	0.0204	*вони*[vy] ()	2	0.0435
12	*what*	1	0.0132	*хто* [khto] (who)	1	0.0204	*хто*[khto] (who)	2	0.0435
13	*як*[yak] (as)	2	0.0263	*якщо*[yakshcho] (if)	1	0.0204	*якщо*[yakshcho] (if)	1	0.0217
14	*он*[on] (he)	5	0.0658	*тут*[tut] (here)	2	0.0408	*тут*[tut] (here)	1	0.0217
15	*довго*[dovho] (long)	2	0.0263	*далеко*[daleko] (long)	1	0.0204	*довго*[daleko] (long)	1	0.0217
16	*я*[ya] (I)	6	0.0789	*це*[tse] (this)	2	0.0408	*це*[tse] (this)	1	0.0217
17	*старий* [staryy] (old)	2	0.0263	*товстий* [tovstyy] (thick)	1	0.0204	*інший*[inshyy] (other)	1	0.0217
18	*слухати* [slukhaty] (old)	1	0.0132	*кидати*[kydaty] (throw)	1	0.0204	*казати* [kazaty] (say)	1	0.0217
19	*чоловік* [cholovik] husband)	1	0.0132	*потік* [potik] (stream)	1	0.0204	*приходити* [prykhodyty] (come)	1	0.0217
20	*багато*[bahato] (many)	1	0.0132	*один*[odyn] (stream)	2	0.0408			
21	*рік*[rik] (year)	1	0.0132	*назад*[nazad] (back[)	1	0.0204			
22	ім'я[im"ya] (year)	1	0.0132	*інший*[inshyy] (other)	1	0.0204			
23	*сонце*[sontse] (sun)	1	0.0132	*білий*[bilyy] (white)	1	0.0204			
24				*дещо*[deshcho] (something)	1	0.0204			
	Total	76		49			46		

In Excerpt 1, 253 words long, there are 23 words from the 200-word list of Svodesh. These words make up 30.04% of the entire passage. In Excerpt 2, 262 words long, there are 24 words from the 200-word list of Svodesh. These words make up 18.7% of the entire passage. In Excerpt 3, 246 words long, there are 19 words from the 200-word list of Svodesh. These words make up 18.7% of the entire passage. Analysing the obtained data, we note that words from Svodesh's list in Excerpt 1 make up 30% of the excerpt, which is much more than 18.7%, as in Excerpts 2 and 3 (Fig. 29a). Such results are natural and transparent: over time, a person's vocabulary also enriches. Also, these passages in Fig. 29b show such results graphically:

- nodes indicate a passage and the number of words in it from the Svodesh list;
- arcs indicate the number of common words from the Svodesh list for these passages and the correlation coefficient for these passages;
- in the centre, the total number of words common to the excerpts and the Svodesh list is indicated (Table 11 - common words are highlighted in colour).

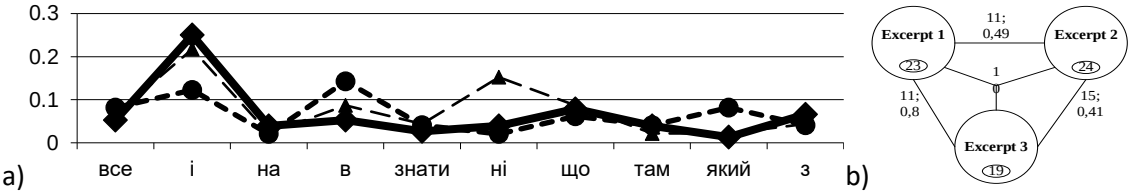


Figure 29: Numerical results of the study of passages, where squares - text excerpt 1, circles - 2, triangles - 3

During the experimental testing, an analysis of more than 300 Ukrainian-language excerpts of texts (the first 10,000 characters) of one-person (more than 100 authors) scientific and technical publications of the Bulletin of Lviv Polytechnic University of the "Information Systems and Networks" series for the period 2001–2021 was carried out (alg. 7).

Algorithm 7. Identification and analysis of a collection of service words in texts

Stage 1. Research of publications to identify the range of the optimal volume of the analysed Ukrainian textual content.

Step 1. Analysis of Ukrainian-language textual content in its entirety (alg. 6).

Step 2. Analysis of excerpts of Ukrainian-language textual content in ranges [10;1000000] characters from the beginning of the scientific and technical publication.

Step 3. Analysis of the obtained results. The optimal analysis of Ukrainian textual content is in the range [100;10000] characters. If ≤100 characters, the values of the stylistic parameters of different authors are similar, and the values of the same author in different passages in different publications are sometimes significantly different. If ≥10000 characters - the parameters almost do not change, moreover, various publications have ≤10000 characters and quite a few publications have ≥10000 characters.

Step 3. Analysis of excerpts of Ukrainian textual content in the range [100;10000] characters of more than 100 different authors to form general stylistic patterns of the author.

Stage 2. The study of the results of changes in the degree of diversity of the author's speech depends on the time interval in the range [2001; 2021] for the author's periodic stylistic patterns formation.

Stage 3. Identification of parameters that change over time and the range of change, and parameters that do not change or do not change significantly.

Stage 4. A study of publications to identify the author's speech styles according to general and periodic patterns in different periods [2001; 2021] years.

Stage 5. A study of computed speech parameters to generate a subset of potential authors with a similar style to other reference collective works from the period [2001; 2021], among whose authors are the authors of individual scientific and technical publication templates.

Stage 6. Analysis of results. If in the generated subsets of potential authors, there are real authors of the collective work, then determine the parameters that can more accurately identify it. Conduct experiments on several algorithms. Choose the best one to identify the style of a potential author in texts from different periods.

3.6.3. Linguometric analysis of determining the content author based on statistical parameters of speech diversity

Every author improves both his vocabulary and his style of writing publications over time. Therefore, it is necessary to investigate whether the parameters of the stylistic diversity of the authors' speech change over time, which exactly change and in what range (Fig. 30-31). Over time, authors use shorter words more often (Fig. 30) and the degrees of lexical diversity K_l and syntactic complexity K_s do not change significantly (Fig. 30b–d). The degree of speech connectivity K_z does not decrease significantly. In 2001, it changed within [0.5; 1.2], and in 2021 – within [0.4; 0.9] (Fig. 30e). The distribution does not change significantly over time for the exclusivity parameter I_{wt}, and there are significant changes for the concentration parameter I_{kt} (Fig. 31). For example, over time, authors for a certain topic of research use some service words more often in their publications (Fig. 32). According to the results presented in Fig. 33 over time, authors use shorter sentences to describe their research in Ukrainian-language scientific and technical texts. Also, the amount of occurrence of prepositions in Ukrainian-language scientific and technical texts is decreasing, but the distribution of occurrence of conjunctions almost does not decrease over time (Fig. 34).

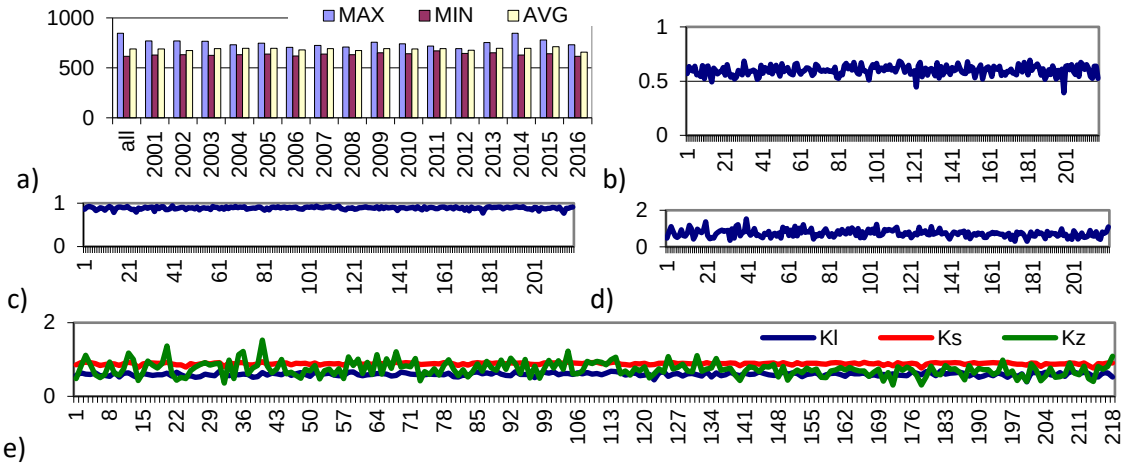


Figure 30: Distribution: a – words and speech parameters for texts of the same volume in the range [2001; 2021] year: b – K_l; c – K_s; d – K_z; e – parameters K_l, K_s and K_z

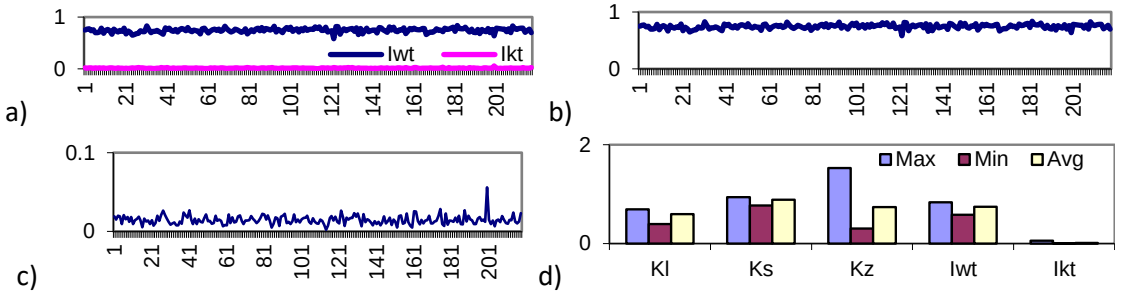


Figure 31: Distribution of the degree of speech for a – both parameters; b – I_{wt}; c – I_{kt}; d is the minimum, maximum and average value for all parameters

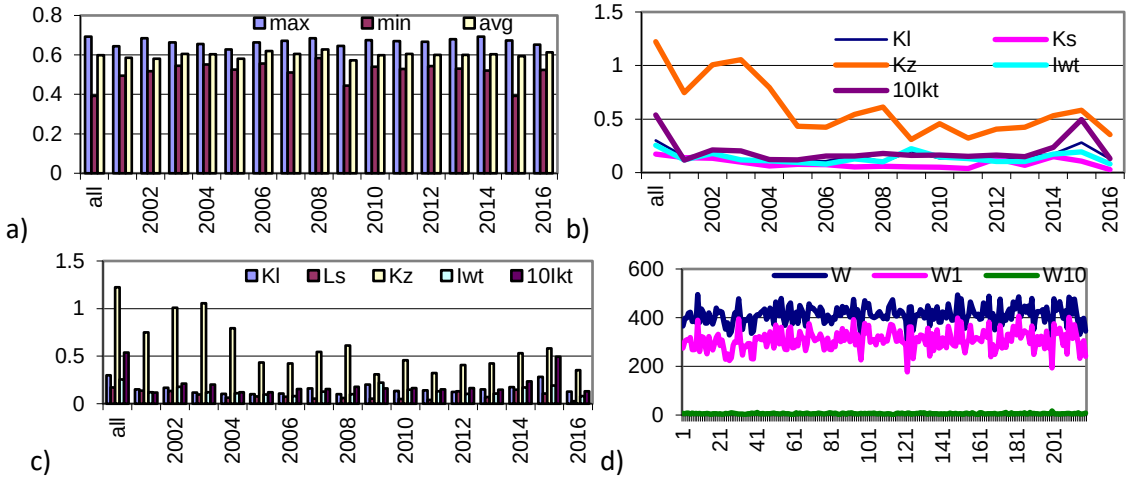


Figure 32: Distribution of speech parameters for texts of equal volume in the range of 2001–2017 years: a – maximum, minimum and average value K_l; b,c – change of parameter values; d – the appearance of word forms (all, only 1 time and ≥10)

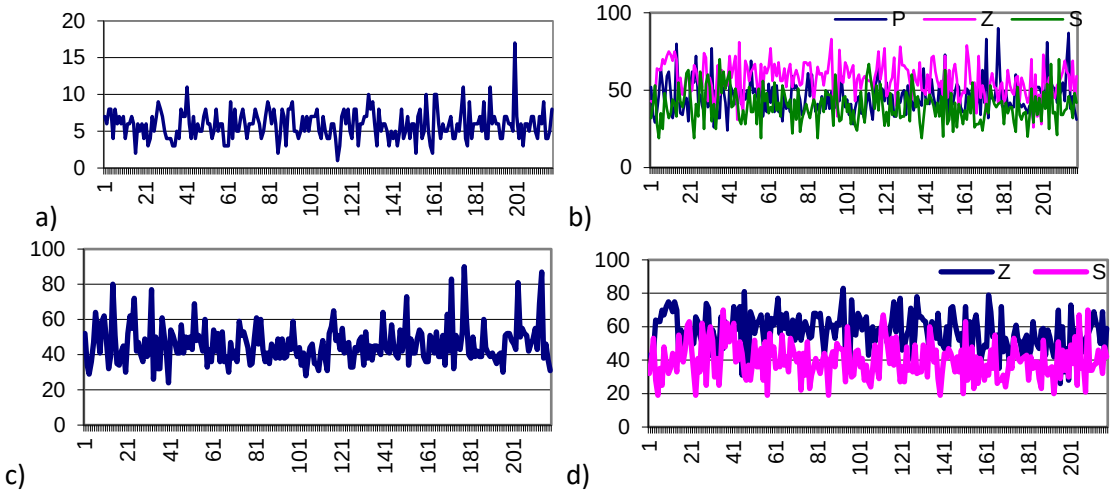


Figure 33: Distribution of occurrence of words: a – ≥10 (W_{10}); b – the degree of speech connectivity; c – a sentence; d – prepositions and conjunctions

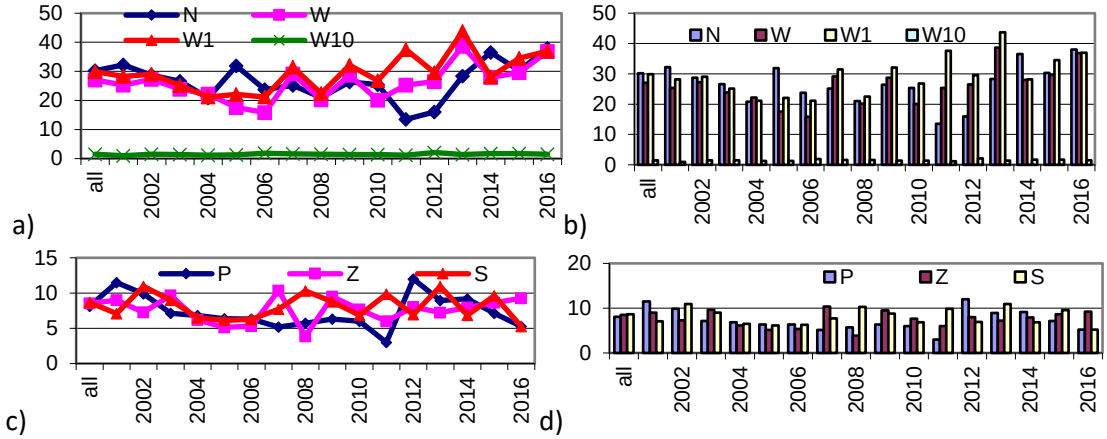


Figure 34: Changes in the distribution of features of the author's speech style over time

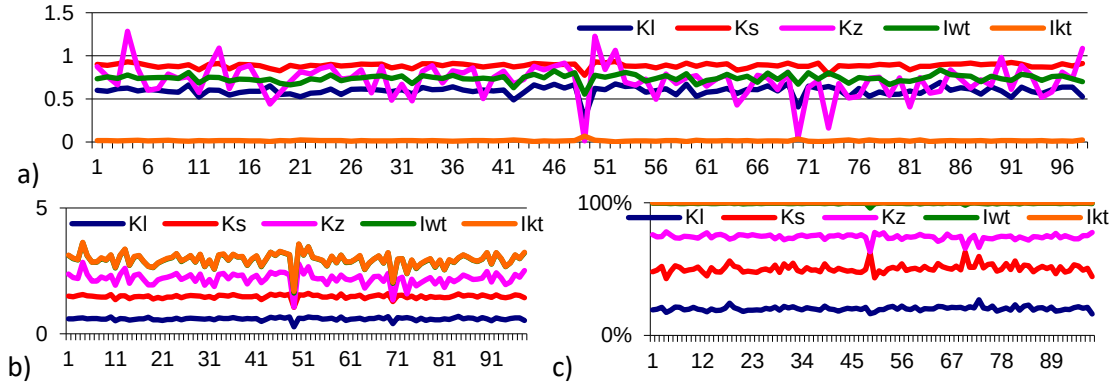


Figure 35: Study of change over time according to the features of speech: a – identification of the author's style; b – total amount; c – value embedding based on normalization

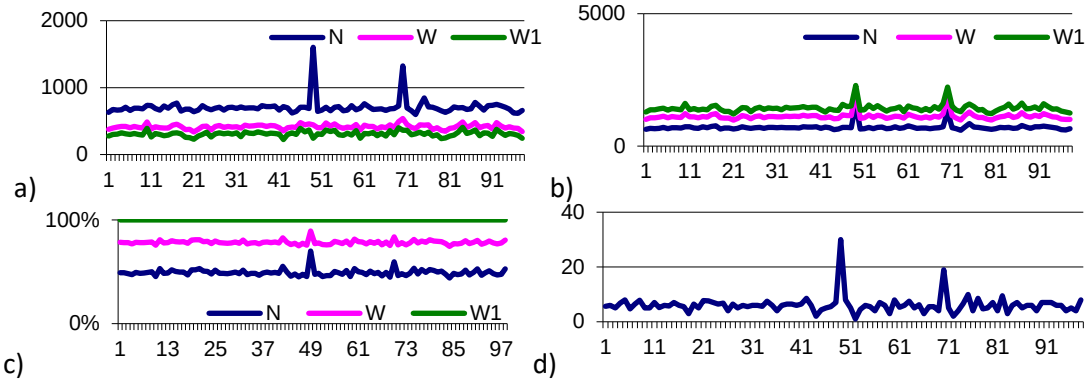


Figure 36: Study of changes in time according to the features of speech: a – identification of the author's style; b – total amount; c – value embedding; d – W_{10} changes

It is necessary to find the range of growth of each of the studied parameters (Fig. 35) since there is a dynamic change not only in the features of the author's speech style during a certain period of scientific activity, but also in individual parameters (the volume of the appearance of sentences, conjunctions and prepositions, word forms on the total volume of words, word forms that are used exactly 1 time and ≥10). The sign of the author's speech, apart from K_z, does not change significantly. Then we will examine the publications by additional parameters (Fig. 36). Introducing additional parameters will reduce the set of potential authors with similar speech styles (Fig. 37 and Table E of Appendix D).

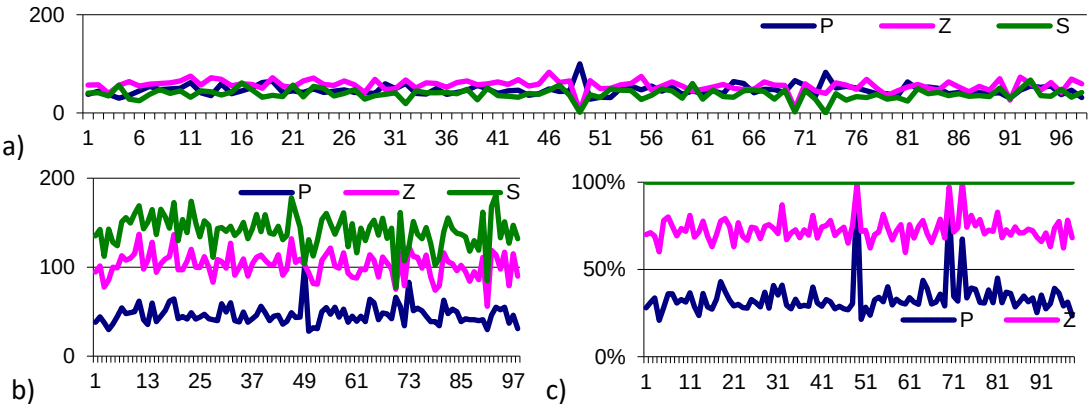


Figure 37: Changes research in time according to the features of speech: a – identification of the author's style; b – total amount; c –the nesting of each value

3.6.4. The quantitative method of determining the authorship of text content based on a statistical analysis of the distribution of N-grams

Each language has its statistical parameters. For example, for Ukrainian texts, it was found that the statistical parameters of styles can be considered the frequencies of vowels, consonants, gaps between words, as well as soft and sonorous groups of consonants (Table F of Appendix D). To achieve the goal of the research, a system was developed with the possibility of choosing

the language/languages of the analysed content, which is implemented on the Victana Web resource. For high-quality and effective content analysis, when determining the degree of authorship of a specific person, we suggest analysing the reference text and the researched one in several stages.

- Algorithm 1. Linguometric analysis of author's speech diversity coefficients (alg. 8);
- Algorithm 2. Stylometric analysis (alg. 9);
- Algorithm 3. Analysis of stable word combinations (algorithm 10);
- Алгоритм 4. Linguistic statistical analysis through N-grams (alg. 11).

The Web resource for linguometric analysis has the following fields (Fig. 38a):

- *Знаків* [Znakiv] (Signs) (he entered text must contain at least 100 and no more than 10,000 characters.) – the maximum content size is displayed.
- *Контент* [Kontent] (Content) – the field where the studied text is copied from buffer.
- *Розрахувати* [Rozrakhuvaty] (Calculate) – start the calculation.
- *Очистити* [Ochystyty] (Clear) – clearing the entered data.

Algorithm 8. Linguometric text analysis to determine authorship

Stage 1. Filtering of Ukrainian-language text content from information noise (special symbols, pictures, tags, numbers, formulas, etc.).
Stage 2. Determining the size of text content – excess is cut off.
Stage 3. Identification of the volume of sentences in Ukrainian textual content.
Stage 4. Identification of the total volume of words in text N.
Stage 5. Identification of the volume of unique W words in textual content.
Stage 6. Identification of the volume of prepositions Z in textual content.
Stage 7. Identification of the volume of conjunctions S in textual content.
Stage 8. Calculation of the coefficients of the author's speech.
Stage 9. Output of results to the end user (Table 12, Fig. 38a).

Table 12

An example of calculating the coefficients of the author's speech

Coefficient	Input data	Calculation
Coefficient of lexical diversity: $K_l=W/N$	W=184, N=295	K_l=0.62372881355932
Coefficient of syntactic complexity: $K_s=1-P/W$	P=18, W=184	K_s=0.90217391304348
Coefficient of speech connectivity: $K_z=(Z+S)/(3*P)$	Z=20, S=28, P=18	K_z=0.88888888888889
Exclusivity index: $I_{wt}=W_1/W$	W_1=141, W=184	I_{wt}=0.76630434782609
Concentration index: $I_{kt}=W_{10}/W$	W_{10}=2, W=184	I_{kt}=0.010869565217391

a) b)

Figure 38: An example of a) the result of the application of linguometric analysis and b) entering data for stylometric analysis

The Web resource for stylometric analysis has the following fields (Fig. 38b):

- *Еталонний текст* [Etalonnyy tekst] (Reference text) – the field where Reference text is copied from the buffer.
- *Вибрати Уривок* 1 (2, 3) [Vybraty Uryvok 1 (2, 3)] (Select Excerpt 1 (2, 3)) – open access to excerpts. Access to the next passage only after activating access to the previous one. Access is opened sequentially from the smallest number to the largest.
- *Уривок* 1 (2, 3) [Uryvok 1 (2, 3)] (Passage 1 (2, 3)) – the field where the text of the passage is copied from the buffer. The entered text must contain at least 100 characters. (Currently 0) – After starting the calculation, the actual number of marks of each passage will be calculated and displayed separately.
- *Розрахувати* [Vybraty Uryvok 1 (2, 3)] (Calculate) – start the calculation.
- *Очистити* [Ochystyty] (Clear) – clearing the entered data.

Algorithm 9. Stylometric analysis of the text to determine authorship

Stage 1. Checking the lengths of the reference text and selected passages and reducing the length of the reference text to the minimum of the checked ones.

Stage 2. Cleaning the reference text from special characters, etc.

Stage 3. Determination of the number of words in the standard text.

Stage 4. Determination of the number of stop words (prepositions + conjunctions + particles) in the standard text (Fig. 39, Table 13).

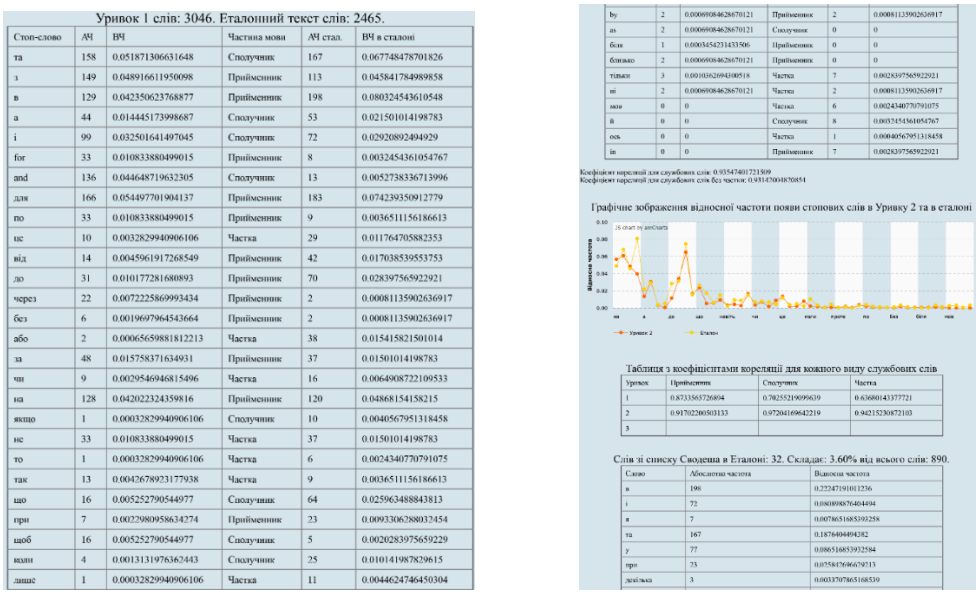
Stage 5. The length of Passage 1 is no more than the minimum text.

Stage 6. Cleaning of Passage 1 from special characters and others.

Stage 7. Determination of the number of words W1 for Passage 1.

Stage 8. Determination of the number of stop words (prepositions + conjunctions + particles) in the text.

Stage 9. Preparation of individual arrays (excerpt and standard) for calculating the correlation coefficient



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

Стоп-слово	АЧ	ВЧ	Частина мови	АЧ стал.	ВЧ в еталоні
та	158	0.051871306631648	Сполучник	167	0.067748478701826
з	149	0.048916611950098	Прийменник	113	0.045841784989858
в	129	0.042350623768877	Прийменник	198	0.080324543610548
а	44	0.014445173998687	Сполучник	53	0.021501014198783
і	99	0.032501641497045	Сполучник	72	0.02920892494929
for	33	0.010833880499015	Прийменник	8	0.003245436105476 7
and	136	0.044648719632305	Сполучник	13	0.005273833671399 6
для	166	0.054497701904137	Прийменник	183	0.074239350912779
по	33	0.010833880499015	Прийменник	9	0.003651115618661 3
це	10	0.003282994090610 6	Частка	29	0.011764705882353
від	14	0.004596191726854 9	Прийменник	42	0.017038539553753
до	31	0.010177281680893	Прийменник	70	0.028397565922921
через	22	0.007222586999343 4	Прийменник	2	0.000811359026369 17
без	6	0.001969796454366 4	Прийменник	2	0.000811359026369 17
або	2	0.000656598818122 13	Частка	38	0.015415821501014
за	48	0.015758371634931	Прийменник	37	0.015010141987 83
чи	9	0.002954694681549 6	Частка	16	0.006490872210953 3
на	128	0.042022324359816	Прийменник	120	0.048681541582 15
якщо	1	0.000328299409061 06	Сполучник	10	0.004056795131845 8
не	33	0.010833880499015	Частка	37	0.015010141987 83
то	1	0.000328299409061 06	Частка	6	0.002434077079107 5
так	13	0.004267892317793 8	Частка	9	0.003651115618661 3
що	16	0.005252790544977	Сполучник	64	0.025963488843813
при	7	0.002298095863427 4	Прийменник	23	0.009330628803245 4
щоб	16	0.005252790544977	Сполучник	5	0.002028397565922 9
коли	4	0.001313197636244 3	Сполучник	25	0.010141987829615
лише	1	0.000328299409061 06	Частка	11	0.004462474645030 4

by	2	0.000690846286701 21	Прийменник	2	0.000811359026369 17
as	2	0.000690846286701 21	Сполучник	0	0
бля	1	0.000345423143350 6	Прийменник	0	0
близько	2	0.000690846286701 21	Прийменник	0	0
тільки	3	0.001036269430051 8	Частка	7	0.002839756592292 1
ні	2	0.000690846286701 21	Частка	2	0.000811359026369 17
мов	0	0	Частка	6	0.002434077079107 5
й	0	0	Сполучник	8	0.003545436105476 7
ось	0	0	Частка	1	0.000405679513184 58
in	0	0	Прийменник	7	0.002839756592292 1

Коефіцієнт кореляції для службових слів: 0.9354701721509
Коефіцієнт кореляції для службових слів без часток: 0.9514700487085 4

Графічне зображення відносної частоти появи стопових слів в Уривку 2 та в еталоні

Таблиця з коефіцієнтами кореляції для кожного виду службових слів

Уривок	Прийменник	Сполучник	Частка
1	0.8733565726864	0.7825521909963 9	0.6368014337772 1
2	0.9170220050313 3	0.9720416964221 9	0.9421523087210 3
3			

Слів зі списку Сводеша в Еталоні: 32. Складає: 3.60% від всього слів: 890.

Слово	Абсолютна частота	Відносна частота
в	198	0.2224719101236
і	72	0.0808988764044 94
я	7	0.0078651685393258
та	167	0.1876404494382
у	77	0.0865168539325 84
при	73	0.0758426966679213
дельфін	3	0.0033707865168539

Figure 39: An example of the result of the application of stylometric analysis and the result of the application of stylometric analysis for Excerpt 2

Stage 10. Calling the function to calculate the correlation coefficient.

Stage 11. Formation of the array to form a graphic image of the relative frequency of occurrence of stop words in Excerpt 1 and the standard.

Stage 12. Calling the function for calculating the HF graph (Fig. 40a).

Stage 13. Calling the function to calculate the correlation coefficient of Excerpts 2(3) for each of the service words.

Stage 14. Form the words of the Svodesha list from the directory, and determine the number of words from the Svodesha list in the passage (for the reference text and selected passages - Table 13).

Stage 15. We form common standards for the Standard, Excerpts 1–3 and the Svodesh list.

Stage 16. The research results are displayed on the screen (Table 14).

Table 13

Passage 1 words: 153. Reference text words: 153

Word	AF	RF	Part of speech	AF in Etalon	RF in the Etalon
в[v] (in)	5	0.032679738562	Preposition	5	0.032679738562
а[a] (and)	2	0.0130718954248	Conjunction	2	0.0130718954248
це[tse] (it)	1	0.0065359477124	Particle	1	0.0065359477124
та[ta] (and)	16	0.1045751633987	Conjunction	16	0.1045751633987
для[dlya] (for)	7	0.0457516339869	Preposition	7	0.0457516339869
з[z] (with)	2	0.0130718954248	Preposition	2	0.0130718954248
ж[zh] (same)	1	0.0065359477124	Particle	1	0.0065359477124
і[i] (and)	3	0.019607843137	Conjunction	3	0.019607843137
також[takozh] (also)	2	0.0130718954248	Conjunction	2	0.0130718954248
мов[mov] (as)	2	0.0130718954248	Particle	2	0.0130718954248
у[u] (in)	1	0.0065359477124	Preposition	1	0.0065359477124
що[shcho] (what)	1	0.0065359477124	Conjunction	1	0.0065359477124
за[za] (by)	1	0.0065359477124	Preposition	1	0.0065359477124

Table 14

Common to the Standard, Excerpts 1–3 and the Svodesh list: 8 (26.67 %) of the total: 30

N	Common	AF	Etalon	Excerpt 1	Excerpt 2	Excerpt 3
1	*в*[v] (in)	5	0.167	0.167	0.167	0.167
2	*це*[tse] (it)	1	0.033	0.033	0.033	0.033
3	*та*[ta] (and)	16	0.533	0.533	0.533	0.533
4	*з*[z] (with)	2	0.167	0.167	0.167	0.167
5	*коло*[kolo](near)	1	0.033	0.033	0.033	0.033
6	*і*[i] (and)	3	0.1	0.1	0.1	0.1
7	*у*[u] (in)	1	0.033	0.033	0.033	0.033
8	*що*[shcho] (what)	1	0.033	0.033	0.033	0.033

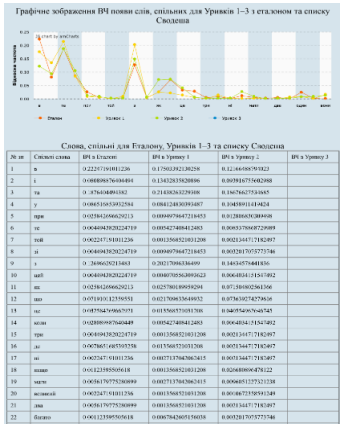
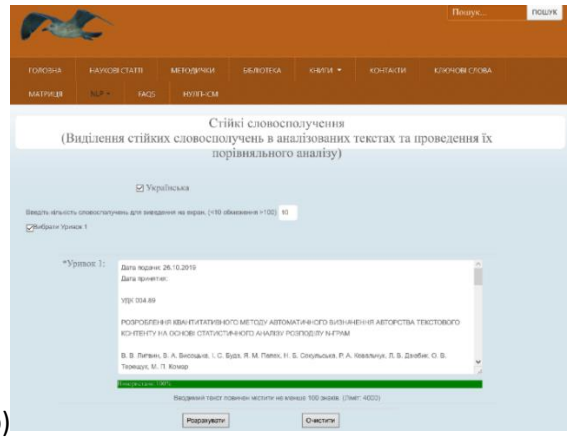
a)  b) 

Figure 40: An example of the result of a) stylometric analysis for Passages 1–3 and b) using the analysis of persistent phrases

For the automated processing of the text, it is of great importance not only what the frequency of appearance of this or that category, but in general its presence in the studied text. Summing up, it should be noted that the use of content analysis for the creation of information systems allows you to capture the distribution of various features of the analysed text content.

For example, the frequency characteristics of the text (average sentence size) may indicate a certain specificity of a person's intellectual abilities in terms of verbal presentation of thoughts. By determining the average length of sentences, it is possible to characterize the change in the individual's emotional state. One of the most significant features in the psycholinguistic analysis of textual content is the choice of analysing a dictionary variant in context dependence. Thanks to the establishment of the coefficient of vocabulary diversity of speech (Table 15), it is possible to identify, for example, the degree of the author's possible presence of schizophrenia.

Table 15

Coefficients of frequency characteristics of the text

Coefficient	Formula
Verbs (aggressiveness)	$K_{\text{Different words}}$= different words /2$N_{\text{all words}}$
Verbs (aggressiveness)	$K_{\text{дієсл.}}$= verbs/$N_{\text{all words}}$·100 %
Emotionality of the text	$K_{\text{прикм.}}$= adjectives/2$N_{\text{all words}}$
Logical connectivity	$K_{\text{лог.зв'язн.}}$= stop of words/3$N_{\text{sentences}}$

Embolism (clogging)	$К_{емб.}$= embols/$N_{all\ words}$·100 %

Another criterion of language competence is the coefficient of verbosity (aggressiveness). The essence of this coefficient is the ratio of the number of verbs and verb forms (adverbs and adjectives) to the total amount of words. A high indicator of aggressiveness indicates the presence of a high degree of negative emotionality of the author, which is reflected in the text itself by manifestations of changes in the dynamics of events and other characteristic features. The parameter of logical coherence based on the analysis of operative words indicates a sufficiently harmonious level of logical construction of the text. The coefficient of speech clutter is the ratio of the total volume without semantic load of words to the total volume of words. The composition of words without semantic load includes exclamations as *а-а-а*, *е-е-е*, *м-м-м*, *ха-ха*, *ну-ну*, *еге*, *ж*, *ой* (a-a-a, e-e-e, mm-m-m, ha-ha, nu-nu, ege, zh, oi, etc.), vulgarisms (profanity), unnecessary repetition. The coefficient of speech clogging states either the degree of a person's negative emotional state (nervousness, fear, discomfort in the environment, etc.) or a low level of speech culture and intelligence. Even taking into account the fact that the artistic text is considered androgynous in principle and is an interweaving of subordinate functions - the qualities of the author's "I" are in a certain way graded depending on the characterological profile of one or another author. In other words, the original text and the translated text depend on their authors. The following fields are available on the Web resource for the analysis of persistent phrases (Fig. 40 b):

- Enter the number of phrases to display on the screen (10;100) – how many phrases will be displayed on the screen after the calculation.
- *Вибрати Уривок* 1 (2, 3) [Vybraty Uryvok 1 (2, 3)] (Select Excerpt 1 (2, 3)) – open access to excerpts. Access to the next passage only after activating access to the previous one. Access is opened sequentially from the smallest number to the largest. (Not implemented - only one passage is analysed).
- *Уривок* 1 [Uryvok 1] (Passage 1) – the field where the text of the corresponding passage is copied from the buffer.
- *Використано*:57 % [Vykorystano:57 %] (Used: 57%) – The entered text must contain at least 100 characters. (Limit: 4000) – text size analysis.
- *Розрахувати* [Rozrakhuvaty] (Calculate) – start the calculation.
- *Очистити* [Ochystyty] (Clear) – clearing the entered data.

Algorithm 10. Linguistic statistical analysis of stable word combinations

Stage 1. Cleaning the received content from special characters, etc.

Stage 2. Forming blocked words list from the database depending on the selected language of the context.

Stage 3. Preparation for the formation of arrays of double word combinations and all words. The input is an array: the key is numbers, and the value is text, divided by sentences (dot separator). The words are checked against the database of keywords and, according to the rule described in the database, lead the given word to the base of the word, if it is not itself the base of the word.

Stage 4. Determination of persistent word combinations using the FREG method: obtain the absolute frequency of word combinations (Fig. 41a).

Stage 5. Determination of stable word combinations using the t-test method: P(W1)*P(W2) taking into account not only pairs but also the frequency of use of individual words (those that make up the pair).

Stage 6. Determination of stable phrases using the LR method.

Stage 7. Determination of stable phrases according to the X2 method (Table 16).

Stage 8. The results of the study are displayed on the screen.

a) 

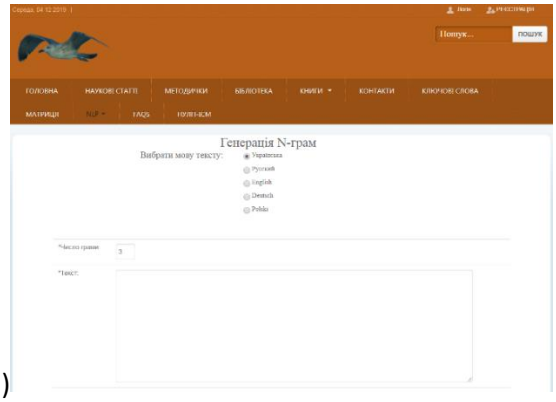
b) 

Figure 41: A persistent phrases analysis result and N-gram text analysis application

Table 16

List by frequency rating of persistent phrases for article 1, phrases: 45. Total words: 108

N	FREG			t-test		LR		X2	
	Keyword	AF	RF	Keyword	t	Keyword	logL	Keyword	X2
1	*система електронний* [systema elektronnyy] (electronic system)	4	0.088889	*система електронний* [systema elektronnyy] (electronic system)	1.822222	*інформаційний технологія* [informatsiynyy tekhnolohiya] (information technology)	5.03e–1	*прийняття рішення* [pryynyattya rishennya] (making a decision)	45.000000
2	*інформаційний система* [informatsiynyy systema] (information system)	4	0.088889	*електронний контент-комерція* [elektronnyy kontent-komertsiya] (electronic content commerce)	1.578091	*інтелектуальний система* [intelektual'nyy systema] (intelligent system)	2.13e–1	*система електронний* [systema elektronnyy] (electronic system)	45.000000
3	*електронний контент-комерція* [elektronnyy kontent-komertsiya] (electronic content commerce)	3	0.066667	*розділ науковий* [elektronnyy kontent-komertsiya] (scientific section)	1.319933	*інформаційний система* [informatsiynyy systema] (information system)	8.36e–2	*електронний контент-комерція* [elektronnyy kontent-komertsiya] (electronic content commerce)	32.946429
4	*розділ науковий* [elektronnyy kontent-komertsiya] (scientific section)	2	0.044444	*інформаційний система* [informatsiynyy systema] (information system)	1.222222	*портал науковий* [portal naukovyy] (scientific portal)	5.58e–2	*розділ науковий* [elektronnyy kontent-komertsiya] (scientific section)	29.302326
5	*портал науковий* [portal naukovyy] (scientific portal)	1	0.022222	*прийняття рішення* [pryynyattya rishennya] (making a decision)	0.977778	*курс технологія* [kurs tekhnolohiya] (technology course)	3.31e–2	*курс технологія* [kurs tekhnolohiya] (technology course)	21.988636

6	*інтелектуальний система* [intelektual'nyy systema] (intelligent system)	1	0.022222	*курс технологія* [kurs tekhnolohiya] (technology course)	0.955556	*сховище дані*[skhovyshche dani] (data storage)	3.31e–2	*сховище дані*[skhovyshche dani] (data storage)	21.988636
7	*прийняття рішення* [pryynyattya rishennya] (making a decision)	1	0.022222	*сховище дані*[skhovyshche dani] (data storage)	0.955556	*прийняття рішення* [pryynyattya rishennya] (making a decision)	8.27e–3	*портал науковий* [portal naukovyy] (scientific portal)	14.318182
8	*курс технологія* [kurs tekhnolohiya] (technology course)	1	0.022222	*портал науковий* [portal naukovyy] (scientific portal)	0.933333	*розділ науковий* [elektronnyy kontent-komertsiya] (scientific section)	1.89e–3	*інформаційний система* [informatsiynyy systema] (information system)	5.848550
9	*сховище дані*[skhovyshche dani] (data storage)	1	0.022222	*інтелектуальний система* [intelektual'nyy systema] (intelligent system)	0.777778	*електронний контент-комерція* [elektronnyy kontent-komertsiya] (electronic content commerce)	1.55e–4	*інтелектуальний система* [intelektual'nyy systema] (intelligent system)	3.579545
10	*інформаційний технологія* [informatsiynyy tekhnolohiya] (information technology)	1	0.022222	*інформаційний технологія* [informatsiynyy tekhnolohiya] (information technology)	0.688889	*система електронний* [systema elektronnyy] (electronic system)	1.37e–6	*інформаційний технологія* [informatsiynyy tekhnolohiya] (information technology)	1.890409

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

When identifying the author of a text, it is assumed that the text reflects the author's style of writing, which makes it possible to distinguish him from others. To compare texts with each other, it is necessary to compare the text with some numerical characteristic, which would be approximate for the texts of the same author and would differ significantly for the works of different authors. Such a characteristic can be the density of the distribution of letter combinations of three consecutive symbols (3-grams). It is defined as a set of empirical frequencies of the use of letters or their combinations. When analysing text based on the density of the N-gram distribution, punctuation marks, spaces, and numbers are not taken into account. The task of identifying the author of an unknown text in terms of the density of the N-gram distribution is defined as follows. A certain set of texts is given, which contains the works of Y famous authors. Let L_y be the amount of content by the y-th author. $N_{i,y}$ is the number of characters in the i-th content of the y-th participant, i=1, ..., L_y. The distribution density of N-grams of content, the volume of which is equal to $N_{i,y}$, is given as a set of values $f_{i,y}(j)=k_j/N_{i,y}$, k_j is the number of uses of N-gram under number j. The argument j=1,...,$y(n, M)$ corresponds to the number of the letter combination (N-grams) in alphabetical order, where M is the strength of the alphabet of the language of the written text, n is the order of the N-gram, that is, the number of symbols in the letter combination. $y(n, M)=M^n$ is the number of N-grams in this alphabet. Each author is identified with his weighted average density of N-gram distribution according to the formula $p_y(j) = \frac{1}{N_y}\sum_{i=1}^{L_y} p_{i,y}N_{i,y}$. They are the author's standards. To compare two texts, or a text and an author's standard, it is necessary to set the distance between the corresponding distribution functions. As distance metrics, we apply the norm in the space of functions as terms. So, for example, the distance $p_{x,y}$ between the N-gram distribution density of an unknown text p_x and any author's N-gram distribution density $p_{x,y}$ is calculated as:

$$p_{x,y} = \left|\left|p_x - p_y\right|\right| = \sum_{j=1}^{y(n,M)} \left|p_x(j) - p_y(j)\right|. \; P_{i,y} = \frac{\left|\left|p_{i,y} - p_y\right|\right|}{1 - \frac{N_{i,y}}{N_y}}.$$

The text "*x*" belongs to the author whose distance to the density of the N-gram distribution will be the smallest. When solving the classification problem, the data set was not divided into test and training sets. Weighted average distribution densities of N-grams were constructed over the entire set of content of one author. The distance from content *i* to a specific author y was calculated as $P_{i,y}$. The formula makes it possible to exclude the participation of the density of the distribution of N-grams of content *i* in the average density of the distribution of N-grams of a specific author. The Web resource for analyzing N-grams has the following fields (Fig. 41b):

- *Вибрати мову тексту* [Vybraty movu tekstu] (Select the language of the text) – the language of the text for analysis (research). The default is "Ukrainian".
- *Число грами* [Chyslo hramy] (Number of grams) – кількість знаків у грамі. Можна міняти на 1, 2, 3, 4. За замовчуванням 3.
- Limitation of text in characters.
- *Текст* [Tekst] (Text) – the field where the researched text is copied from the buffer.
- *Генерувати* [Heneruvaty] (Generate) – to start the generation of N-grams.
- *Очистити* [Ochystyty] (Clear) – clearing the entered data.

Algorithm 11. Linguistic statistical analysis of N-grams of text [52]

Stage 1. Cleaning of the researched text (numbers, special characters).
Stage 2. We calculate the number of words in the text.
Stage 3. All words of the text are translated into lowercase.
Stage 4. We remove spaces.
Stage 5. Depending on the selected language, the appropriate alphabet is substituted.
Stage 6. Depending on the set number of grams, the corresponding function is launched, which calculates all possible variants of grams and stores them in an array.
Stage 7. Next, the function of counting the number of occurrences of words is launched. Here we calculate the relative frequency of occurrence and store in the array: the serial number of the gram, the gram itself, the number of occurrences of this gram, and the relative frequency of occurrence of this gram.
Stage 8. The next function forms the array obtained in the previous function for export to a CSV file. This file is stored on the server. It can be downloaded to the user's (researcher's) computer using the link, which will be accessed after creating a form with the research results.
Stage 9. The results of the research are displayed on the screen (only those grams found in the text).
Stage 10. Access to the export file opens.
Stage 11. Summarized results are displayed:

- – the size of the alphabet;
- – the number of words in the text;
- – the number of characters in the text with spaces;
- – the number of characters in the text completely cleaned;
- – total N-grams;
- – a total of N-grams without repetitions were found;
- – a total of N-grams with repetitions were found.

We compare three scientific and technical publications [53, 54, 55] with each other based on linguistic statistical analysis of 3-grams. Articles 1, and 2 were written by the same team [53, 54], and Article 3 was written by another author [55] (Table 17). The language of the text is Ukrainian (letters in the alphabet are 33, so there are 35937 possible N-grams).

Table 17

Parameter values for analyzed articles 1–3

Parameters	Article 1	Article 2	Article 3
Total N-gram	35937	35937	35937
Total N-grams found (no duplicates)	4354	4377	3890
Total N-gram found (with repetition)	29494	29862	36383
Total words	5475	5358	6060
Total characters in raw text	39792	39663	47084
Total characters in cleared text	29967	32570	37062

But when comparing articles, we will take into account only those 3-grams that appeared in the text at the same time in three articles at least once. Therefore, for this particular example, all 3-grams are 2147. That is, for Article 1 we analyse 78.4814% of 3 grams, for Article 2 – 72.6332% and Article 3 – 84.1271%. Accordingly, the difference in the use of the corresponding 3-grams between Articles 1 and 2 is R_{12}=56,5254 %, Articles 2 and 3 – R_{23}=69,4271 %, between Articles 1 and 3 – R_{13}=62.9839 %. These indicators alone show that the characteristics of Articles 1 and 2 are more similar ($R_{23}>R_{12}$ by 12.9017%, $R_{23} > R_{13}$ by 6.4432%, $R_{13}> R_{12}$ by 6.4585%, i.e. $R_{23}>R_{13}>R_{12}$) than the characteristics of Articles 1–3 respectively and 2–3. The smaller the R_{ij}, the greater the degree to which the articles are written by the same author. In that case, Articles 1 and 2 are more likely to be written by the same author/team than Articles 2–3 and Articles 1–3 respectively. But let's analyse the use of individual 3-gram clusters in the corresponding articles and compare the obtained results (Table 18).

Table 18

The value of the parameters of the appearance of 3-grams for the analyzed articles 1–3

3-gram	The average value of 1 appearance			Range, %	Match for articles, %			Discrepancy for articles, %		
	1	2	3		1–2	2–3	1–3	1–2	1–3	2–3
а_ _	0.0393	0.0430	0.0392	6.112–6.709	4.2322	4.6322	4.197	0.0271	0.0297	0.0269
б_ _	0.0220	0.0415	0.0262	0.594–1.121	0.7046	0.7738	0.4884	0.0261	0.0287	0.0181
в_ _	0.0390	0.0367	0.0388	4.262–4.522	3.5581	4.1064	3.6523	0.0307	0.0354	0.0315
г_ _	0.0302	0.0234	0.0455	0.749–1.454	0.6551	1.3451	1.309	0.0205	0.0420	0.0409
д_ _	0.0292	0.0290	0.0354	2.263–2.764	1.5257	2.0978	1.8299	0.0196	0.0269	0.0235
е_ _	0.0438	0.0359	0.0555	3.197–4.941	3.0263	3.6893	4.0674	0.0340	0.0415	0.0457
є_ _	0.0189	0.0114	0.0321	0.252–0.707	0.2508	0.5443	0.6077	0.0114	0.0247	0.0276
ж_ _	0.0338	0.0243	0.0274	0.341–0.474	0.25	0.2302	0.2126	0.0179	0.0164	0.0152
з_ _	0.0273	0.0234	0.0352	1.311–1.973	1.1879	1.25	1.3259	0.0212	0.0223	0.0237
и_ _	0.0376	0.0338	0.0366	4.327–4.818	3.2931	4.0083	3.5984	0.0257	0.0313	0.0281
і_ _	0.0294	0.0277	0.0288	4.772–5.051	3.5963	3.9431	3.7918	0.0209	0.0229	0.0220
ї_ _	0.0114	0.0117	0.0168	0.038–0.125	0.2247	0.3031	0.2386	0.0102	0.0138	0.0108
й_ _	0.0180	0.0131	0.0188	0.301–0.432	0.3352	0.3469	0.3483	0.0146	0.0151	0.0151
к_ _	0.0383	0.0340	0.0415	2.791–3.400	2.4206	3.2381	2.4931	0.0295	0.0395	0.0304
л_ _	0.0539	0.0401	0.0364	2.073–3.070	2.4437	1.8021	2.0952	0.0429	0.0316	0.0368
м_ _	0.0238	0.0264	0.0343	2.168–3.123	1.7619	2.6603	1.8196	0.0194	0.0292	0.0200
н_ _	0.0468	0.0420	0.0474	6.421–7.257	3.8242	5.1327	4.0623	0.0250	0.0335	0.0266
о_ _	0.0473	0.0397	0.0540	6.473–8.795	5.3403	7.5276	6.3371	0.0328	0.0462	0.0389
п_ _	0.0476	0.0559	0.0720	1.858–2.809	1.6619	2.5456	2.1261	0.0426	0.0653	0.0545
р_ _	0.0384	0.0426	0.0456	3.690–4.380	3.1902	4.3566	3.4834	0.0332	0.0454	0.0363

с_ _	0.0541	0.0377	0.0381	3.169–4.541	3.3187	2.7052	3.4299	0.0395	0.0322	0.0408
т_ _	0.0445	0.0417	0.0429	5.174–5.518	3.5467	4.712	4.6607	0.0286	0.0380	0.0376
у	0.0286	0.0267	0.0332	2.193–2.726	1.7905	1.9852	1.9443	0.0218	0.0242	0.0237
ф	0.0384	0.0595	0.0401	0.276–0.495	0.3069	0.4759	0.3211	0.0345	0.0619	0.0374
х	0.0155	0.0180	0.0252	0.573–0.934	0.5083	0.7426	0.7957	0.0137	0.0201	0.0215
ц	0.0246	0.0345	0.0305	0.591–0.829	0.568	0.4416	0.4748	0.0237	0.0184	0.0198
ч	0.0425	0.0223	0.0559	0.513–1.324	1.0044	0.9368	0.6924	0.0437	0.0407	0.0301
ш	0.0145	0.0194	0.0457	0.194–0.657	0.2169	0.2917	0.6854	0.0130	0.0438	0.0378
щ	0.0200	0.0118	0.0201	0.064–0.100	0.1401	0.0828	0.1404	0.0097	0.0092	0.0142
ь	0.0317	0.0256	0.0329	0.998–1.285	0.6593	0.7983	0.7326	0.0169	0.0205	0.0188
ю	0.0173	0.0234	0.0309	0.277–0.494	0.1558	0.3005	0.2673	0.0097	0.0188	0.0167
я	0.0206	0.0216	0.0201	1.444–1.554	0.9522	1.0555	0.9361	0.0132	0.0147	0.0130

According to Table 19 and Fig. 42a some of the letters in the Ukrainian language are used most often, others are much less common. For the most frequently used letters, the frequency of appearance of 3-grams with such initial letters will have an almost identical distribution (peak values on the graph Fig. 42a), but not for other letters.

Table 19

Frequency distribution of the appearance of the 1-gram in Articles 1–3

1-gram	Article 1		Article 2		Article 3	
	Number	RF	Number	RF	Number	RF
о	2824	0.094240	2472	0.075898	3870	0.103601
н	2471	0.082460	2370	0.072766	2888	0.077312
а	2255	0.075252	2698	0.082837	2491	0.066685
т	2102	0.070146	1956	0.060055	2141	0.057315
і	1789	0.059701	1967	0.060393	2250	0.060233
и	1732	0.057799	1852	0.056862	2036	0.054504
в	1654	0.055196	1590	0.048818	1915	0.051265
с	1549	0.051692	1327	0.040743	1384	0.037050
е	1404	0.046853	1453	0.044612	2090	0.055950
р	1335	0.044550	1722	0.052871	1893	0.050676
к	1279	0.042682	1110	0.034080	1453	0.038897
л	1116	0.037242	927	0.028462	906	0.024254
у	987	0.032937	960	0.029475	1195	0.031990
д	859	0.028666	939	0.028830	1319	0.035310
м	808	0.026964	976	0.029966	1399	0.037451
п	647	0.021591	825	0.025330	1138	0.030464
я	647	0.021591	681	0.020909	864	0.023129
з	623	0.020790	644	0.019773	946	0.025325
ь	498	0.016619	418	0.012834	613	0.016410
ч	459	0.015317	289	0.008873	574	0.015366
г	408	0.013615	373	0.011452	651	0.017427
х	355	0.011847	384	0.011790	482	0.012903
б	284	0.009477	569	0.017470	428	0.011458
ж	246	0.008209	210	0.006448	176	0.004712
й	239	0.007976	260	0.007983	265	0.007094
ц	224	0.007475	334	0.010255	299	0.008004
є	188	0.006274	165	0.005066	347	0.009289
ф	179	0.005973	209	0.006417	137	0.003668
ї	174	0.005807	217	0.006663	270	0.007228
ю	156	0.005206	277	0.008505	289	0.007737
ш	117	0.003904	169	0.005189	281	0.007522
щ	95	0.003170	52	0.001597	128	0.003427

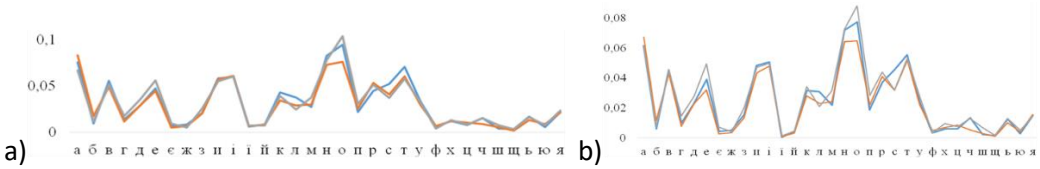


Figure 42: The graph of the frequency distribution of a) the 1-gram in Articles 1–3 and b) of 3-grams that begin with a specific letter, where blue is article 1, orange is article 2, and grey is article 3

Therefore, it is advisable to study only trigrams for initial letters that are less common in the texts of a specific language to determine the degree of belonging of the text to the corresponding author (for example, Fig. 42-Fig. 43).

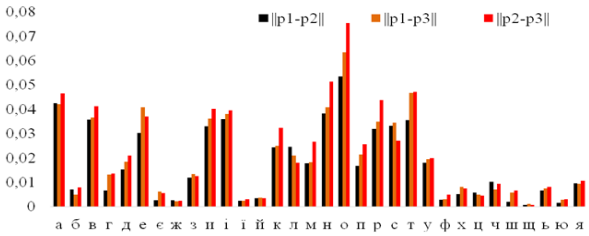


Figure 43: A graph of the difference in the use of 3-grams that begin with a specific letter

According to these graphs, it appears that Article 1 and Article 2 were most likely written by the same author, although Article 1 and Article could also have been written by the same author (but this is not true). However, articles 2–3 were written by different authors. The application of linguistic statistical analysis of 3 grams to a set of articles will allow to formation of a subset of publications similar in terms of linguistic characteristics. Imposition of additional conditions on this subset in the form of linguistic statistical analyses (set of keywords, stable phrases, stylometric, ligvometric, etc.) will allow for a significant reduce this subset, clarifying the list of more likely author's works. Thus, an analysis of the content and frequency of appearance of only official words will separate articles 1 and 3 into different subsets, leaving articles 1 and 2 in one.

3.7. Analysis of the developed method of quantitative assessment of the potential author identification of a scientific and technical publication

The method consists of six algorithms for the analysis of Ukrainian-language texts.

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

Algorithm II. Calculation and analysis of the features of the author's speech style (frequency of word usage, volume of punctuation marks, sentences, symbols, words and the ratio of the number of marks and sentences).

Algorithm III. Calculation and analysis of the parameters of the author's speech style (speech coherence, syntactic complexity, lexical diversity, degree of concentration and exclusivity of the text).

Algorithm IV. Classification by parameters and lexical features of the textual content of other publications (application of classifiers such as fuzzy, SVM and a combination of the previous two).

Algorithm V. Performance analysis based on the obtained results to determine each classifier accuracy.
Algorithm VI. Determining a subset of potential authors based on filtering from the set of all researched through the analysis of features and style parameters (algorithms VIII–XI).

A lexer-type system (tokenizer, segmenter) has been developed as part of a text analyser based on tokenization (Fig. 44a). Tokens are extracted during the operation of the parser rules and are immediately checked for compliance with the conditions in the syntax rules to avoid generating absurdity (Fig. 44b).

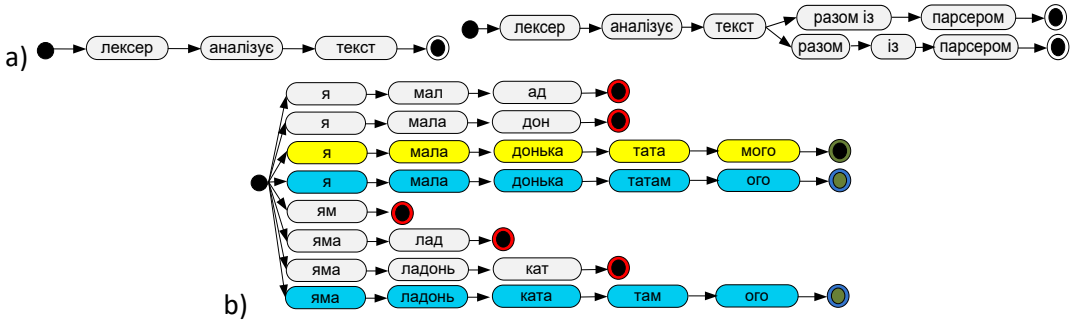


Figure 44: Illustration of a) tokenization graphs and b) tokenization graph without syntax rules

The rules help to solve several tasks, increasing the efficiency of the grammar engine, which loads the compiled rules during text parsing, without wasting time on syntax parsing. (alg. 12)

Algorithm 12 (VII). Text content segmenter

Step 1. Word recognition.
Step 2. Definition of token boundaries.
Step 3. Definition of complete word forms.
Step 4. Identification of indivisible tokens that contain dots, blanks, etc.
Step 5. Splitting the text into sentences.

In addition to defining the boundaries of tokens, the lexer also performs preliminary recognition of the morphological attributes of words, turning tokens into tokens. When constructing Ukrainian-language sentences with direct word order, a distinction is made between the noun group $\tilde{N}$ and the verb group $\check{R}$ (Fig. 45, Fig. 46).

I) $S \to \# \tilde{N}_{РД,ЧП,л,ОС}\ \tilde{R}_{ЧП,мн,ОС}\ \#.$

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

1) $\tilde{N}_{РД,ЧП,ВД,З} \to \tilde{N}_{РД,ЧП,ВД,З}\ \tilde{N}_{РД',ЧП',р,ОС}$;
2) $\tilde{N}_{РД,ЧП,ВД,З} \to A_{РД,ЧП,ВД}\ \tilde{N}_{РД,ЧП,ВД,З}$;
3) $K_1\tilde{N}_{РД,ЧП,ВД,ОС}\ K_2 \to K_1 N^{займ}_{РД,ЧП,ВД,ОС} K_2$;
4) $\tilde{N}_{РД,ЧП,ВД,З} \to N_{РД,ЧП,ВД}$;
5) $\tilde{N}_{РД,ЧП,ВД,З} \to E\tilde{N}_{РД,ЧП,ВД,З}$;
6) $\tilde{N}_{РД,ЧП,ВД,З} \to \tilde{N}_{РД,ЧП,ВД,З}\tilde{N}_{РД,ЧП,м,З}$.

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

1) $\tilde{R}_{ЧП,мн,ОС} \to R_{ЧП,мн,ОС}\ \tilde{N}_{РД',ЧП',з,ОС'}\ \tilde{N}_{РД'',ЧП'',о,ОС''}$;
2) $\tilde{R}_{ЧП,мн,ОС} \to R_{ЧП,мн,ОС}\ \tilde{N}_{РД'',ЧПС,о,ОС'}\ \tilde{N}_{РД'',ЧП'',з,ОС''}$;
3) $\tilde{R}_{ЧП,мн,ОС} \to R_{ЧП,мн,ОС}\ \tilde{N}_{РД',ЧП',з,ОС'}$;
4) $\tilde{R}_{ЧП,мн,ОС} \to R_{ЧП,мн,ОС}\ \tilde{N}_{РД',ЧП',о,ОС'}$;
5) $\tilde{R}_{ЧП,мн,ОС} \to R_{ЧП,мн,ОС} E\tilde{N}_{РД,ЧП,м,З}$;
6) $\tilde{R}_{ЧП,мн,ОС} \to E\tilde{N}_{РД,ЧП,м,З} R_{ЧП,мн,ОС}$.

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

Figure 45: Production rules for analysing a Ukrainian-language sentence, where *N* – is a noun, *A* – is an adjective, $N^{займ}$ – is a pronoun; number/число/ЧЛ (од, мн); genus/рід/РД (ч, ж, с;), person/особа/ОС (1, 2, 3); case/відмінок/ВД (н, р, д, з, о, м, к); time/час/ЧС (тп, мн, мб)

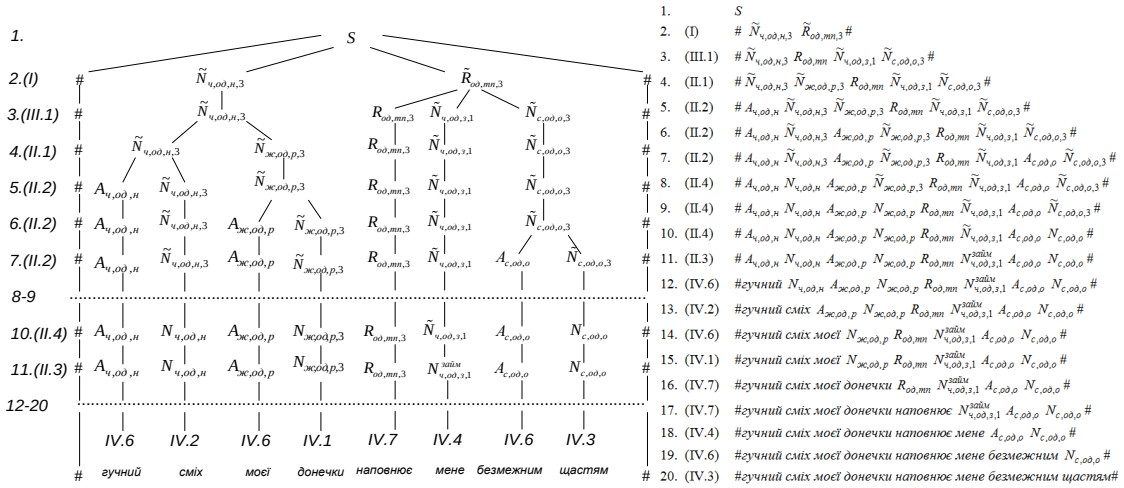


Figure 46: Illustration of the analysis of the structure of the Ukrainian sentence

We get constituents tree, or the syntactic structure of the analysed sentence (Fig. 47). For dictionary lexemes, a dictionary article whose form is the lexeme is also defined. In alphabetic-frequency dictionaries, its characteristics are determined through/for a word (Fig. 48).

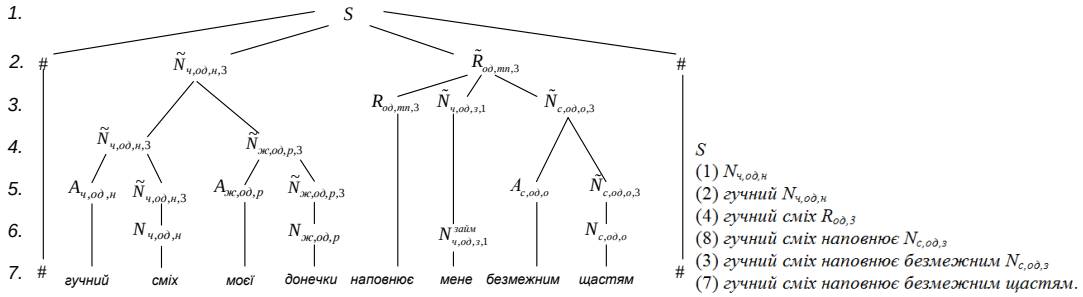


Figure 47: An illustration of the analysis of a Ukrainian sentence

a)

Уривок 1	Уривок 2	Уривок 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**
кирилічний/**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**

b)

```
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 Наявному 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.)
```

Figure 48: a) The base of rules of the alphabetic-frequency dictionary of parts of speech), where A is a verb, other capital letters are additional characteristics of a verb, V is an adjective, small

letters of the English alphabet are characteristics of a noun and b) regular expressions of morphological analysis of nouns

The database stores regular expressions for bringing the word to the base (Fig. 49a-b), where the flag is the rule for identifying the type of word (for example, noun group, singular), mask – inflexions of the word (exceptions in square brackets), find – inflexions of the word in the nominative case, repl – inflexions of the word during declension (Fig. 49c).

a)

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	ir	ir	огу
29	29	1	a	SFX	uk	ir	ir	огові
30	30	1	a	SFX	uk	ir	ir	огом
31	31	1	a	SFX	uk	ir	ir	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

b)

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

c)

```
# Іменники із закінченням на –ін з чергуванням -і -о
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.)
третій рядок описує
# Іменники із закінченням на –іг з чергуванням -і -о
SFX a іг    огу   іг      # батіг батогу       (Д.Р.)
SFX a іг    огові іг      # батіг батогові     (Д.М.)
SFX a іг    огом  іг      # батіг батогом      (О.)
SFX a іг    озі   іг      # батіг батозі       (М.)
дев'ятий рядок описує
# Іменники із закінченням на –ід з чергуванням -і -о
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.)
```

Figure 49: The base of definition rules: a – the basis of the word; b - service words and c) for determining the basis of a word

Also, in the database (Fig. 49b) there is a dictionary of service words, that is, words that are additional parameters for analysing the features of the author's speech style and taking into account during the analysis of texts significantly affect the final result.

We will determine the optimal developed algorithm out of four (VIII-XI) for identifying the style of the author of the publication based on the analysis of his collective works.

Algorithm VIII. Filtering a set of analysed author's styles

```
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 collective works (Table 20 and Table E of Appendix – highlighted in yellow), some of whose authors are numbered 6 and 30 (highlighted in blue). Array A[j][l] – style features for 94 authors. Array F[l] – average values of style features for 94 authors. The algorithm determines whether the value of the parameters and coefficients of the speech style of the *j*-th author falls within the limits $[x_i+x_{cep}; x_i-x_{cep}]$ deviation of parameter values and speech coefficients of the *i*-th collective work style. Arrays A2 (authors, the values of most parameters and coefficients are similar to the style of the team *i*) and A3 (authors, the values of most of the coefficients are similar to the style of the team *i*) are filled through the filters. Next, a new subset of authors (whose styles are more similar to the collective ones – *i*-th work) is formed from the obtained previous arrays by superimposing a new filter.

Table 20

The result of the algorithm for analyzing the style of a publication author on Victana [16] 94 authors on more than 300 individual publications for the period 2001–2021

#	N	W	W_1	W_{10}	P	Z	S	K_l	K_s	K_z	l_{wt}	l_{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

As a result, we will get the values given in Table 21 (algorithm VIII). Columns A are the results of the analysis of all the values of the coefficient vectors and speech parameters of the authors from Table 20. Column B is the result of analysing only the last 5 columns in Table 20. Unfortunately, this algorithm produced such results that the listed authors of these works are unlikely to have written them themselves (the best results are highlighted in red - and it is not enough to claim that they are the authors of more than 50% of these collective works). Although, on the other hand, this algorithm gives good results - reducing the number of authors at the first stage of authorship determination (up to 34.04% of the total number of project participants). This is necessary for further filtering through the analysis of root words (prepositions and conjunctions) and keywords, features of semantics and vocabulary when constructing sentences, etc.

Table 21

Experimental testing of algorithms I–IV on the Victana Web resource [16]

Algorithm	Collective	Average 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
	$\bar{x}_i$	6,15	2,74	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
	$\bar{x}_i$	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
	$\bar{x}_i$	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
	$\bar{x}_i$	5,64	2,34	6.3	3.0	7.5	2.8	50.3	41.5	35.8	38.0

As a result, we will get the values given in Table 21 (algorithm IX). Then we will analyse algorithm IX. It does not differ significantly from the previous one, only by the 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 an array of average absolute values of deviations of data points from the average value. The obtained results are slightly improved, but not enough to claim that authors numbered 6 and 30 are the real authors of collective works 1–4, although they wrote them. On the other hand, the number of authors (up to 38.56% of the total number of project participants) with a similar style of speech increased slightly. Now let's analyse algorithm X. In algorithm 1, we will also replace the condition in the third cycle with the following:

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

As a result, we will get the values given in Table 6.14 (algorithm X). As we can see, the obtained values make it clear that the style of authors numbered 6 and 30 is quite close (more than 75–100%) to the style of collective works 1-4, respectively (positive results are highlighted in red). Although the number of authors (up to 42.02% of the total number of project participants) with similarities in speech style has increased significantly. On the other hand, many of those who were not included in the previous stages of the study were included in that list, and those who were also included in the previous two stages of the study fell out of the crowd. Now let's try to reduce the total number by applying the XI algorithm to the obtained initial data - parameters and speech coefficients of 94 project participants. In algorithm X, we improve the condition in the third cycle:

```
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, we get the values given in Table 21 (Algorithm XI). The obtained values also confirm that the style of authors numbered 6 and 30 is quite close (more than 75–100%) to the style of collective works 1–4, respectively (positive results are highlighted in red). Also significantly reduced the number of authors (to 38.03% of the total number of project participants) with similarities in speech style. Fig. 50 provides detailed graphs of the results obtained when applying algorithms VIII–XI (numbered 1–4, respectively) for the analysis of the method of determining the author's style developed by us.

Further, to determine the author's style, an analysis of root words (prepositions and conjunctions) and keywords of the authors' works was used, as 38.03% got to those. Each individual has his special vocabulary for conveying his opinion, including the so-called "parasitic" (*тобто*, *отже*, *хоча* [tobto, otzhe, khocha] (that is, therefore, although) etc.) та службових слів (*і*, *та*, *й*, *але*, *хоч би* [i, ta, y, ale, khoch by] (and, and, and, but, although) etc.).

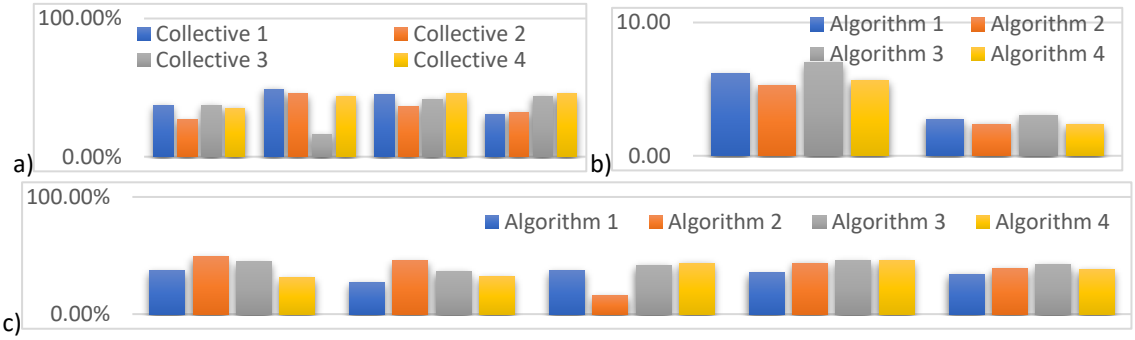


Figure 50: Style identification research: a – according to the developed algorithms; b – taking into account the signs of speech; c – for analysed collective works

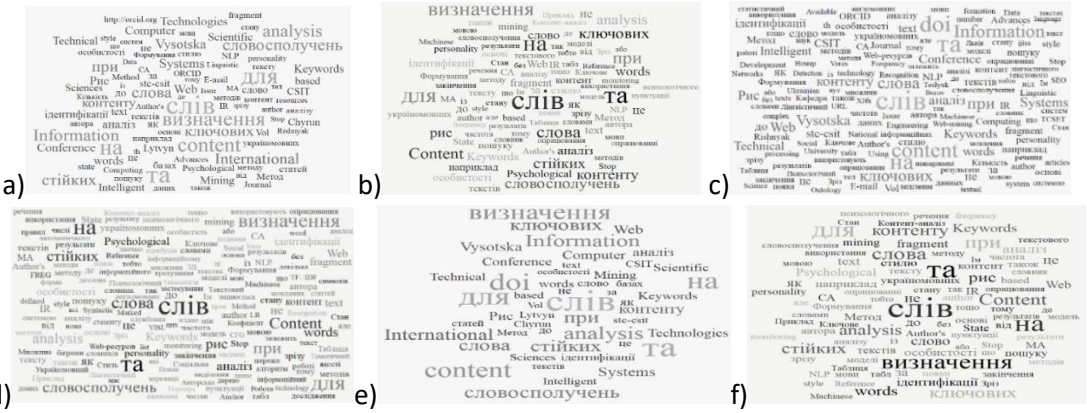


Figure 51: Study of style at stage 2 for the text with the construction of a frequency dictionary: a – complete with 100 words; b – the main one of 100 words; c – complete with 200 words; d – the main one of 200 words; e – complete with 50 words; f – the main one of 50 words

Fig. 51 presents an example of the analysis of the author's style at the second stage - through the analysis of the frequency of appearance of service and keywords taking into account various filters, the analysis of full texts with a list of references and annotations in different languages, and the analysis of only the informative part of the publication, i.e., the main text with the construction of a frequency dictionary accordingly 200, 100 and 50 words).

4. Conclusions

A method of determining stable word combinations was developed based on the identification of keywords of the Ukrainian-language text and analysis of the lexical speech coefficients of the

author of the text in reference excerpts of the content, which made it possible to improve the accuracy of the method of determining the style of the author of the text by 9% based on statistical linguistics. The method consists of the use of Zipf's law in the formation of stable word combinations as key, taking into account the following rules of preliminary linguistic processing of the text: removal of all sentence words; form bigrams only within the limits of punctuation marks; the verb and pronoun are considered punctuation marks; determine verbs by their inflexions; form bigrams based on their bases without taking into account their inflexions; definition of adjectives by their inflexions and to believe that adjectives should only be in the first place in the bigram from Ukrainian-language texts. A set of programs has been developed to identify persistent phrases as key. An approach to the development of linguistic content analysis software for the determination of stable word combinations in the identification of keywords of Ukrainian-language and English-language textual content is proposed. The peculiarity of the approach is the adaptation of the linguistic statistical analysis of lexical units to the peculiarities of the constructions of Ukrainian and English words/texts. The results of the experimental approbation of the proposed method of content analysis of English- and Ukrainian-language texts for the determination of stable word combinations in the identification of keywords of technical texts were studied.

A method of determining the author in Ukrainian-language texts has been developed based on the analysis of the coefficients of the author's lexical speech in the referenced passage of the author's text, which is based on the analysis of a collection of keywords, persistent phrases, indicators of lingumetry, stylometry, as well as the results of the analysis of N-grams based on comparisons of differences in the use of 2- gram and 3-gram for publications similar in style within [6;7]%, and for those not exactly similar – >12%), which made it possible to identify a set of potential authors of publications from more than one author (up to [9;34] % of the total number of project participants) and develop a method for identifying the author's style.

A method of identifying the style of the author of the text based on the analysis of the features of the author's speech style in a template passage of the author's text has been developed. The method consists of a comparative analysis of the author's attribution in a statistically processed work of the author (standard) with an arbitrary analysed passage. The method evaluates the degree of text belonging to the template of the author's style with the analysis of the corresponding coefficients of the lexical author's speech. Moreover, the method works under the condition that the template of the author's style is generated on reliable data. An analysis of reference words was used for attribution, the obtained results are presented in the form of correlation coefficients. Separately, we will mention the evolution of the significance of one of the parameters of the text - in the author's attribution of the texts.

An algorithm for identifying service words based on linguistic analysis of text content has been developed. For each of the passages, the absolute and relative frequencies of stop words were analysed and compared with the reference values. Therefore, the application of the method of reference words gives the following results: finding among the studied passages what most likely belongs to the standard. Other results confirm the effectiveness of the reference words method in the authorial attribution of texts. The proposed assumption about the insignificance of the influence of the share as a parameter of the method on the results led to a

decrease in the correlation coefficients. However, to confirm or refute the fact that fractions are not a determining factor in the author's style, more thorough research must be performed. An algorithm for the lexical analysis of Ukrainian-language texts and an algorithm for syntactic analysis of text content has been developed. The peculiarities of the algorithms are the adaptation of the morphological and syntactic analysis of word forms to the peculiarities of the construction of Ukrainian words/texts. Belonging to a part of speech and declension within this part of speech were taken into account based on the analysis of inflexions and word bases according to regular expressions.

A comparison of the results of content monitoring on a set of 300 one-man works of a technical direction by 100 different authors for the period 2001–2021 was carried out to determine whether and how the coefficients of text diversity of these authors change in different periods. The best results according to the density criterion are achieved by the article analysis method without initial mandatory information such as abstracts and keywords in different languages, as well as a list of references. The method of identifying a potential author is decomposed based on the analysis of speech style parameters such as speech coherence, degree of syntactic complexity, lexical diversity, degree of concentration and exclusivity. Characteristics of the author's style were also analysed, such as the total amount of words in the text, the number of unique words, the number of conjunctions/prepositions, the number of sentences, and the number of words with a frequency of 1 and ≥10. For example, 3-grams of 3 articles were analysed. 78.4814% of 3-grams were analysed for Article 1, 72.6332% for Article 2, and 84.1271% for Article 3. Accordingly, the difference in the use of the corresponding 3-grams between Articles 1–2 is R_{12}=56,5254 %, between 2 and 3 – %, between 1 and 3 – R_{13}=62.9839 %. These indicators themselves show that the characteristics of Articles 1 and 2 are more similar (R_{23}>R_{12} by 12.9017%, R_{23} > R_{13} by 6.4432%, R_{13}> R_{12} by 6.4585%, i.e., R_{23}>R_{13}>R_{12}) than the characteristics of Articles 1–3, respectively and 2–3. The smaller the R_{ij}, the greater the degree to which the articles are written by the same author. Then in this case Articles 1–2 are more likely to be written by the same author than Articles 2–3 and 1–3 respectively.

This work solved an important scientific and applied problem of CLS analysis and synthesis for solving various problems of processing Ukrainian-language textual content based on the development of new and improvement of known NLP models, methods and tools.

During the execution of the work, the following results were obtained:

1. An analysis of the current state and prospects for IT development of natural language processing was carried out, which made it possible to define the problem and research objectives, as well as to form general research directions in the absence of non-commercial CLS with open source for processing Ukrainian-language textual content and a standardized design approach.

2. The relevance of solving the problem of analysis and synthesis of CLS based on the development of the general structure of the system for processing Ukrainian-language textual content is substantiated, due to the interaction of the main processes/components of IS and methods of linguistic processing of textual content adapted to the Ukrainian language based on grapheme, morphological, lexical, syntactic, semantic, structural, ontological and pragmatic analysis allowed to improve the IT intellectual analysis of the text flow for solving a specific NLP

problem. This ensured the adaptation of NLP processes for the analysis of Ukrainian-language textual content and, based on them, increased the accuracy of the obtained results by 6-48%, depending on the specific NLP task. For example, for the NLP task of determining the keywords of the Ukrainian-language text, the density of keywords increases in the range [1.23; 1.48] times or by [23.14; 47.83]% depending on the quality/accuracy of filling the thematic dictionary through machine learning.

3. The methods of processing information resources such as integration, management and support of Ukrainian-language content have been improved, which made it possible to adapt the process of intellectual analysis of the text flow and develop metrics for the effectiveness of CLS functioning for the solution of various NLP tasks. The developed methods and tools make it possible to build CLS processing of Ukrainian-language text content according to the needs of the permanent/potential target audience based on the analysis of the history of actions of website users.

4. NLP methods based on pattern-matching regular expressions were improved, which made it possible to adapt the methods of tokenization and normalization of text by cascades of simple substitutions of regular expressions and finite state machines.

5. The MA method of the Ukrainian-language text based on word segmentation and normalization, sentence segmentation and modified Porter's stemming algorithm was improved as an effective means of identifying lem affixes for the possibility of marking the analysed word, which made it possible to increase the accuracy of keyword searches by 9%.

6. The IT of intellectual analysis of the text flow was improved based on the processing of information resources, which made it possible to adapt the generally typical structure of modules for integration, management and support of content to solve various NLP problems and increase the efficiency of CLS functioning by 6-9%. This became possible thanks to the combination of linguistic analysis methods adapted to the Ukrainian language, improved IT processing of information resources, ML and a set of metrics for evaluating the effectiveness of CLS functioning. The main principle of building such CLS is modularity, which facilitates their construction according to the requirements for the presence of appropriate processes for solving a specific NLP problem.

7. A method of determining the author in Ukrainian-language texts has been developed based on the analysis of the coefficients of the author’s lexical speech in the referenced passage of the author’s text, which is based on the analysis of a collection of keywords, persistent phrases, indicators of lingumetry, stylometry, as well as the results of the analysis of N-grams based on comparisons of differences in the use of 2- gram and 3-gram for publications similar in style within [6;7]%, and for those not exactly similar – >12%), which made it possible to identify a set of potential authors of publications from more than one author (up to [9;34] % of the total number of project participants) and develop a method for identifying the author's style.

8. A method of determining stable word combinations was developed based on the identification of keywords of the Ukrainian-language text and analysis of the lexical speech coefficients of the author of the text in reference excerpts of the content, which made it possible to improve the accuracy of the method of determining the style of the author of the text by 9% based on statistical linguistics.

9. The reliability of scientific and practical results is confirmed by relevant materials on the implementation of dissertation research, as well as by comparing the obtained practical results on different samples of reliable input data. CLS was developed on the information resource http://victana.lviv.ua using CMS Joomla! (for developing the e-framework of articles), PHP (for implementing text content processing methods), HTML (for implementing page markup), CSS (for describing page styles), and MySQL (for storing data and dictionaries). The experimental study confirmed the reliability of the method of determining keywords - for different algorithms for processing the primary text, the average coincidence of the lists of identified keywords with the authors varies in the range of 52.6-68.5%. The accuracy of matching keywords with the author's keywords ranges from 43.6 to 62.9%. The average match of meaningful keywords compared to all found by the system ranges from 38.9-75.8%, depending on the stages of analysis of article texts. The accuracy of matching keywords compared to all found by the system varies between 34.3-71.9%, depending on the stages of analysis of the texts of the articles.

References

[1] Y. H. Hu, C. T. Tai, K. E. Liu, C. F. Cai, Identification of highly-cited papers using topic-model-based and bibliometric features: The consideration of keyword popularity, Journal of Informetrics 14(1) (2020) 101004.

[2] A. Cheikhrouhou, et. al., Multi-task learning for simultaneous script identification and keyword spotting in document images, Pattern Recognition 113 (2021) 107832.

[3] T. Kumar, M. Mahrishi, G. Meena, A comprehensive review of recent automatic speech summarization and keyword identification techniques, AI in Industrial Applications: Approaches to Solve the Intrinsic Industrial Optimization Problems, 2022, pp. 111-126.

[4] P. Kenekayoro, Author and keyword bursts as indicators for the identification of emerging or dying research trends, J. Sci. Res. 9(2) (2020) 120-126.

[5] A. Berko, Y. Matseliukh, Y. Ivaniv, L. Chyrun, V. Schuchmann, The text classification based on Big Data analysis for keyword definition using stemming, in: Proceedings of IEEE 16th International conference on computer science and information technologies, Lviv, Ukraine, 22–25 September, 2021, pp. 184–188.

[6] A. Taran, The Role of Keyword Language in the Database of World Slavic linguistics "iSybislaw", CEUR Workshop Proceedings 3171 (2022) 266-276.

[7] N. Bondarchuk, et. al., Keyword-based Study of Thematic Vocabulary in British Weather News, CEUR Workshop Proceedings 3171 (2022) 451-460.

[8] O.V. Bisikalo, W. Wójcik, O.V. Yahimovich, S. Smailova, Method of determining of keywords in English texts based on the DKPro Core, in: Proceedings of SPIE - The International Society for Optical Engineering, 2016, 10031.

[9] R. Campos, et. al., YAKE! Keyword extraction from single documents using multiple local features, Information Sciences 509 (2020) 257-289.

[10] P. S. Sharma, D. Yadav, R. N. Thakur, Web page ranking using web mining techniques: a comprehensive survey, Mobile Information Systems 2022(1) (2022) 7519573.

[11] A. Rejeb, K. Rejeb, A. Appolloni, H. Treiblmaier, M. Iranmanesh, Exploring the impact of ChatGPT on education: A web mining and machine learning approach, The International Journal of Management Education 22(1) (2024) 100932.
[12] V. Kayser, E. Shala, Scenario development using web mining for outlining technology futures, Technological forecasting and social change 156 (2020) 120086.
[13] M. Karp, N. Kunanets, Y. Kucher, Meiosis and litotes in The Catcher in the Rye by Jerome David Salinger: Text Mining, CEUR Workshop Proceedings 2870 (2021) 166-178.
[14] S. Kumar, A. K. Kar, P. V. Ilavarasan, Applications of text mining in services management: A systematic literature review, International Journal of Information Management Data Insights 1(1) (2021) 100008.
[15] L. Hickman, et. al., Text preprocessing for text mining in organizational research: Review and recommendations, Organizational Research Methods 25(1) (2022) 114-146.
[16] Z. Yang, Z. Xiangyi, The Applicability of Zipf's Law in Report Text, Lecture Notes on Language and Literature 6(10) (2023) 57-64.
[17] Z. Wang, M. Ren, D. Gao, Z. Li, A Zipf's law-based text generation approach for addressing imbalance in entity extraction, Journal of Informetrics 17(4) (2023) 101453.
[18] A. Koshevoy, H. Miton, O. Morin, Zipf's law of abbreviation holds for individual characters across a broad range of writing systems, Cognition 238 (2023) 105527.
[19] C. Boyer, L. Dolamic, N. Grabar, Automated Detection of Health Websites' HONcode Conformity: Can N-gram Tokenization Replace Stemming?, Studies in Health Technology and Informatics 216 (2015) 1064.
[20] O. Bisikalo, V. Vysotska, Linguistic analysis method of Ukrainian commercial textual content for data mining, CEUR Workshop Proceedings 2608 (2020). 224-244.
[21] V. Vysotska, P. Pukach, V. Lytvyn, D. Uhryn, Y. Ushenko, Z. Hu, Intelligent Analysis of Ukrainian-language Tweets for Public Opinion Research based on NLP Methods and Machine Learning Technology, International Journal of Modern Education and Computer Science (IJMECS) 15(3) (2023) 70-93.
[22] V. Starko, A. Rysin, VESUM: A Large Morphological Dictionary of Ukrainian As a Dynamic Tool, CEUR Workshop Proceedings 3171 (2022) 61-70.
[23] V. Lytvyn, P. Pukach, V. Vysotska, M. Vovk, N. Kholodna, Identification and Correction of Grammatical Errors in Ukrainian Texts Based on Machine Learning Technology, Mathematics 11(4) (2023) 904.
[24] V. Starko, O. Synchak, Feminine Personal Nouns in Ukrainian: Dynamics in a Corpus, CEUR Workshop Proceedings 3396 (2023) 407-425.
[25] O. Synchak, V. Starko, Ukrainian Feminine Personal Nouns in Online Dictionaries and Corpora, CEUR Workshop Proceedings 3171 (2022) 775-790.
[26] V. Starko, Implementing Semantic Annotation in a Ukrainian Corpus, CEUR Workshop Proceedings 2870 (2021) 435-447.
[27] Starko, V.: Semantic Annotation for Ukrainian: Categorization Scheme, Principles, and Tools. In: CEUR workshop proceedings, Vol-2604, 239-248. (2020).
[28] Keygeneratortext. URL: http://msurf.ru/tools/keygeneratortext/.
[29] Keygeneratorurl. URL: http://webmasta.org/tools/keygeneratorurl/.

[30] Keywordstext. URL: http://www.keywordstext.therealist.ru/.
[31] Keygeneratortext. URL: http://syn1.ru/tools/keygeneratortext/.
[32] Terminology extraction. URL: http://labs.translated.net/terminology-extraction/.
[33] Advego. URL: http://advego.ru/text/seo/.
[34] V. Vysotska, S. Holoshchuk, R. Holoshchuk, A comparative analysis for English and Ukrainian texts processing based on semantics and syntax approach, CEUR Workshop Proceedings 2870 (2021) 311-356.
[35] V. Vysotska, O. Markiv, S. Teslia, Y. Romanova, I. Pihulechko, Correlation Analysis of Text Author Identification Results Based on N-Grams Frequency Distribution in Ukrainian Scientific and Technical Articles, CEUR Workshop Proceedings 3171 (2022) 277-314.
[36] V. Vysotska, S. Mazepa, L. Chyrun, O. Brodyak, I. Shakleina, V. Schuchmann, NLP tool for extracting relevant information from criminal reports or fakes/propaganda content, in Proceedings of the 2022 IEEE 17th International Conference on Computer Sciences and Information Technologies (CSIT), 2022, November, pp. 93-98.
[37] V. Lytvyn, et. al., Analysis of statistical methods for stable combinations determination of keywords identification, Eastern-European Journal of Enterprise Technologies 2/2(92) (2018) 23-37. doi: 10.15587/1729-4061.2018.126009.
[38] N. Kholodna, V. Vysotska, O. Markiv, S. Chyrun, Machine Learning Model for Paraphrases Detection Based on Text Content Pair Binary Classification, CEUR Workshop Proceedings 3312 (2022) 283-306.
[39] Y. Stepaniak, V. Vysotska, O. Markiv, L. Chyrun, S. Chyrun, L. Pohreliuk, Technology of Text Content Topic Classification Based on Machine Learning Methods, in Proceedings of the IEEE 5th International Conference on Advanced Information and Communication Technologies (AICT), 2023, pp. 121-126.
[40] Y. Hlavcheva, O. Kanishcheva, M. Vovk, M. Glavchev, Using Topic Modeling for Automation Search to Reviewer, CEUR Workshop Proceedings 3171 (2022) 81-90.
[41] N. Khairova, A. Kolesnyk, O. Mamyrbayev, G. Ybytayeva, Y. Lytvynenko, Automatic Multilingual Ontology Generation Based on Texts Focused on Criminal Topic, CEUR Workshop Proceedings 2870 (2021) 108-117.
[42] V. I. Levenshtein, Binary codes capable of correcting deletions, insertions, and reversals, Soviet physics doklady 10(8) (1966) 707-710.
[43] R. Bellman, R. Kalaba, Dynamic programming and statistical communication theory, Proceedings of the National Academy of Sciences of the United States of America 43(8) (1957) 749.
[44] R. Bellman, R. Kalaba, On the role of dynamic programming in statistical communication theory, IRE Transactions on Information Theory 3(3) (1957) 197-203.
[45] R. Bellmam, Dynamic programming. Princeton univ. press. Princeton. New Jersey, 1957.
[46] R. Bellman, On the approximation of curves by line segments using dynamic programming, Communications of the ACM 4(6) (1961) 284.
[47] R. A. Wagner, M. J. Fischer, The string-to-string correction problem, Journal of the ACM (JACM) 21(1) (1974) 168-173.

[48] D. Gusfield, Algorithms on stings, trees, and sequences: Computer science and computational biology, Acm Sigact News 28(4) (1997) 41-60.
[49] G. D. Forney, The viterbi algorithm, Proceedings of the IEEE 61(3) (1973) 268-278.
[50] V. Motyka, Y. Stepaniak, M. Nasalska, V. Vysotska, Lexical Diversity Parameters Analysis for Author's Styles in Scientific and Technical Publications, CEUR Workshop Proceedings 3403 (2023) 595–617.
[51] R. Romanchuk, V. Vysotska, V. Andrunyk, L. Chyrun, S. Chyrun, O. Brodyak, Intellectual Analysis System Project for Ukrainian-language Artistic Works to Determine the Text Authorship Attribution Probability, in Proceedings of the 2023 IEEE 18th International Conference on Computer Sciences and Information Technologies, CSIT-2023, Lviv, 19-21 October 2023 p.
[52] V. Lytvyn, et. al., Development of the quantitative method for automated text content authorship attribution based on the statistical analysis of N-grams distribution, Eastern-European Journal of Enterprise Technologies 6(2-102) (2019) 28-51. doi: 10.15587/1729-4061.2019.186834.
[53] V. Lytvyn, et. al., Development of the linguometric method for automatic identification of the author of text content based on statistical analysis of language diversity coefficients, Eastern-European Journal of Enterprise Technologies 5(2(95)) (2018) 16–28. doi: 10.15587/1729-4061.2018.142451.
[54] V. Lytvyn, et. al., Development of the system to integrate and generate content considering the cryptocurrent needs of users, Eastern-European Journal of Enterprise Technologies 1(2(97)) (2019) 18–39. doi: 10.15587/1729-4061.2019.154709.
[55] P. Kravets, The Game Method for Orthonormal Systems Construction, in Proceedings of the 9th International Conference - The Experience of Designing and Applications of CAD Systems in Microelectronics, 2007. doi: doi.org/10.1109/cadsm.2007.4297555.

Appendices

Table A

List by frequency rating of stable word combinations for 3 random articles

№	Author's	Victana.lviv.ua (according to Zipf's law)	FREG, t-test	LR	χ^2
Q	A	B	C, D	F	G
			In work[1] in Ukrainian		
1	**Стиль автора**	Стоп-слово	Відносна частота	*Коефіцієнт кореляції*	*Коефіцієнт кореляції*
2	Статистичний аналіз	Метод визначення	*Коефіцієнт кореляції*	Відносна частота	Відносна частота
3	Лінгвістичний аналіз	**Визначення стилю**	**Стиль автора**	*Частота появи*	*Частота появи*
4	Квантитативна лінгвістика	**Стиль автора**	**Визначення стилю**	Стопове слово	Авторська атрибуція
5	Авторська атрибуція	Аналіз уривку	Стопове слово	Україномовний текст	**Стиль автора**
6	**Визначення стилю**	*Частота появи*	Україномовний текст	**Стиль автора**	Україномовний текст
7	Україномовні тексти	*Автор тексту*	*Частота появи*	Поява слова	Стопове слово
8	Технологія лінгвометрії	Уривок тексту	Авторська атрибуція	Авторська атрибуція	**Визначення стилю**
9	Технологія стилеметрії	*Коефіцієнт кореляції*	Поява слова	**Визначення стилю**	Поява слова
10	Технологія глоттохронології	Дослідження тексту	*Автор тексту*	Слова уривку	Слова уривку
			In work[2] in Ukrainian		
1	**Web Mining**	**Ключове слово**	**Ключове слово**	*Текстовий контент*	*Текстовий контент*
2	Контент-моніторінг	**Контент-аналіз**	*Текстовий контент*	**Ключове слово**	Тематичний словник
3	**Ключові слова**	Визначена системою	**Web Mining**	Тематичний словник	**Ключове слово**

4	**Контент-аналіз**	*Формування системою*	Тематичний словник	*Слова контенту*	*Слова контенту*
5	Стеммер Портера	**Web Mining**	Визначення слів	Ключове словосполучення	*Множина слів*
6	Лінгвістичний аналіз	*Слова контенту*	Ключове словосполучення	Визначення слів	*Формування системою*
7	Метод визначення	*Текстовий контент*	*Слова контенту*	*Формування системою*	**Web Mining**
8	Визначення слів	Аналіз статистики	*Множина слів*	**Web Mining**	Визначення слів
9	Слов'янськомовні тексти	Ключове словосполучення	*Формування системою*	*Слова контенту*	*Слова контенту*
10	Технологія NLP	*Множина слів*	**Контент-аналіз**	Контент-моніторінг	Контент-моніторінг
			In work[3] in Ukrainian		
1	Інформаційний ресурс	**Контент-аналіз**	Психологічний стан	Психологічна особистість	Психологічна особистість
2	**Контент-аналіз**	Стоп- слово	Психологічна особистість	Психологічний стан	Психологічний стан
3	Лінгвістичний аналіз	Тематичний словник	**Контент-аналіз**	Формування зрізу	Формування зрізу
4	Морфологічний аналіз	Пости користувача	Марковане слово	*Стан особистості*	Зріз стану
5	**Соціальна мережа**	Повідомлення користувача	Психологічний зріз	Марковане слово	Марковане слово
6	Формування зрізу	Користувач мережі	*Стан особистості*	Психологічний зріз	**Контент-аналіз**
7	Зріз розуміння	*Стан особистості*	Формування зрізу	**Контент-аналіз**	Психологічний зріз
8	Розуміння особистості	Аналізована особистість	Зріз стану	Зріз стану	*Стан особистості*
9	Україномовні тексти	**Соціальна мережа**	Зріз особистості	Аналізована особистість	**Соціальна мережа**
10	Big-Five	Диспозиції особистості	**Соціальна мережа**	**Соціальна мережа**	Аналізована особистість
			In the work[1] in English		
1	**Style of the author**	*Reference fragment*	*Reference fragment*	Words fragment	Words fragment
2	Statistical analysis	**Author's style**	Words fragment	*Reference fragment*	*Reference fragment*
3	Linguistic analysis	*Author's text*	*Syntactic words*	*Stop words*	Recognition author
4	Quantitative linguistics	*Syntactic words*	Frequency fragment	Swadesh list	*Stop words*
5	Author's attribution	*Stop words*	Swadesh list	Recognition author	Swadesh list
6	Recognition of style	Formatted fragments	*Stop words*	*Syntactic words*	*Syntactic words*
7	Ukrainian texts	*Anchor words*	**Author style**	Frequency fragment	Frequency fragment
8	Linguometry technology	Author's language	Recognition author	*Author's text*	*Author's text*
9	Stylemetry technology	Method of anchor	*Author's text*	*Anchor words*	**Author style**
10	Glottochronology technology	Frequency dictionary	*Anchor words*	**Author style**	*Anchor words*
			In the work[2] in English		
1	Web Mining	*Text content*	*Text content*	Web mining	Web mining
2	**Content monitoring**	**Content analysis**	Web mining	*Text content*	*Text content*
3	**Content analysis**	Analysis of statistics	Keywords text	*Keywords content*	*Keywords content*
4	Porter stemmer	Defined systematically	Keywords defined	Keywords text	Analysis text
5	Linguistic analysis	*Stop word*	Analysis text	Keywords defined	Keywords text
6	Determining the keywords	Potential keywords	*Keywords content*	*Stop word*	Keywords defined
7	Slavic language	**Content monitoring**	**Content monitoring**	Analysis text	*Stop word*
8	Slavic texts	*Author's keywords*	**Content analysis**	*Author's keywords*	**Content monitoring**
9	Method for determining	*Keywords content*	*Stop word*	**Content monitoring**	**Content analysis**
10	Web technology	Direct word	*Author's keywords*	**Content analysis**	*Author's keywords*
			In the work[3] in English		
1	Information resource	**Content analysis**	**Content analysis**	Psychological personality	**Content analysis**
2	**Content analysis**	*Psychological state*	Psychological personality	*Psychological state*	Psychological personality
3	Linguistic analysis	Personality analysis	*Psychological state*	**Content analysis**	*Psychological state*
4	Morphological analysis	Personality disposition	Social networks	Based analysis	Based analysis
5	Social network	Psychological analysis	Marked words	State personality	Psychological base
6	Status of personality	Personality model	State personality	Psychological base	State personality
7	Personality understanding	Stop words	Based analysis	Social networks	Social networks
8	Formation of the status	Psychological disposition	Psychological base	Marked words	Psychological base
9	Stop words	Content monitoring	State based	State based	Marked words
10	Method of formation	Social network	Based content	Psychological base	State based

Table B

Differences in methods according to the rating list of 100 stable word combinations

Q	A	B	C	D	F	G	A	B	C	D	F	G	A	B	C	D	F	G
	For Ukrainian-language articles [1-3]																	
A	1	0.23	0.47	0.35	0.27	0.21	1	0.27	0.51	0.39	0.31	0.25	1	0.25	0.49	0.36	0.29	0.23
B	0.23	1	0.63	0.61	0.52	0.43	0.27	1	0.65	0.63	0.57	0.47	0.25	1	0.64	0.62	0.55	0.45
C	0.47	0.63	1	0.93	0.17	0.71	0.51	0.65	1	0.94	0.25	0.73	0.49	0.64	1	0.93	0.21	0.72
D	0.35	0.61	0.93	1	0.19	0.75	0.39	0.63	0.94	1	0.26	0.77	0.36	0.62	0.93	1	0.22	0.76

F	0.27	0.52	0.17	0.19	1	0.26	0.31	0.57	0.25	0.26	1	0.39	0.29	0.55	0.21	0.22	1	0.33
G	0.21	0.43	0.71	0.75	0.26	1	0.25	0.47	0.73	0.77	0.39	1	0.23	0.45	0.72	0.76	0.33	1
For English-language articles [1-3]																		
A	1	0.27	0.51	0.47	0.31	0.27	1	0.31	0.55	0.51	0.35	0.31	1	0.29	0.53	0.49	0.33	0.29
B	0.27	1	0.66	0.64	0.55	0.47	0.31	1	0.69	0.67	0.59	0.49	0.29	1	0.68	0.65	0.57	0.48
C	0.51	0.66	1	0.95	0.23	0.76	0.55	0.69	1	0.96	0.27	0.77	0.53	0.68	1	0.95	0.24	0.75
D	0.47	0.64	0.95	1	0.21	0.79	0.51	0.67	0.96	1	0.29	0.81	0.49	0.65	0.95	1	0.25	0.78
F	0.31	0.55	0.23	0.21	1	0.31	0.35	0.59	0.27	0.29	1	0.41	0.33	0.57	0.24	0.25	1	0.37
G	0.27	0.47	0.76	0.79	0.31	1	0.31	0.49	0.77	0.81	0.41	1	0.29	0.48	0.75	0.78	0.37	1

Table C

Differences of other methods according to the rating of the frequency of occurrence of stable word combinations

Method	Language	Work [1]	Work [2]	Work [3]
A_1	UA	('контент_моніторінгу', 13)	('тематичного_словника', 11) ('слов_янськомовних', 10)	('психологічного_стану', 16) ('формування_зрізу', 12) ('sfx_a', 12) ('структурну_схему', 7) ('відкритість_досвіду', 6) ('зрізу_психологічного', 2)
	ENG	('swadesh_list', 18) ('based_on', 15)	('based_on', 20) ('slavic_language', 15) ('author_s', 13)	('based_on', 35) ('psychological_state', 26) ('social_networks', 22) ('his_her', 11) ('following_structural', 8) ('big_five', 7) ('let_us', 7) ('structural_scheme', 4)
A_2	UA	(('службових', 'слів'), 32) (('стопових', 'слів'), 24) (('визначення', 'визначення'), 23) (('стилю', 'стилю'), 22) (('слів', 'слів'), 22) (('списку', 'сводеша'), 20) (('в', 'уривку'), 19) (('опорних', 'слів'), 18) (('стилю', 'автора'), 17) (('автора', 'автора'), 17)	(('ключових', 'слів'), 72) (('текстового', 'контенту'), 21) (('на', 'етапі'), 17) (('визначення', 'ключових'), 16) (('крок', '1'), 16) (('крок', '2'), 16) (('web', 'mining'), 15) (('слів', 'в'), 14) (('тематичного', 'словника'), 11) (('для', 'визначення'), 10)	(('на', 'основі'), 21) (('психологічного', 'стану'), 18) (('контент', 'аналізу'), 16) (('маркованих', 'слів'), 15) (('зрізу', 'психологічного'), 14) (('стану', 'особистості'), 14) (('формування', 'зрізу'), 12) (('особистості', 'на'), 12) (('sfx', 'a'), 12) (('основі', 'контент'), 11)
	ENG	(('of', 'the'), 107) (('author', 's'), 52) (('of', 'a'), 51) (('in', 'the'), 46) (('the', 'author'), 45) (('reference', 'fragment'), 31) (('analysis', 'of'), 24) (('words', 'in'), 22) (('to', 'the'), 21) (('the', 'method'), 21)	(('of', 'the'), 134) (('in', 'the'), 61) (('by', 'the'), 45) (('analysis', 'of'), 39) (('of', 'a'), 31) (('the', 'text'), 30) (('the', 'system'), 30) (('to', 'the'), 29) (('of', 'keywords'), 28) (('text', 'content'), 27)	(('of', 'the'), 134) (('is', 'the'), 117) (('the', 'content'), 45) (('of', 'a'), 43) (('analysis', 'of'), 37) (('based', 'on'), 35) (('on', 'the'), 34) (('in', 'the'), 33) (('content', 'analysis'), 30) (('the', 'process'), 27)
A_3	UA	(('слів', 'слів'), 88) (('стилю', 'автора'), 68) (('службових', 'слів'), 63) (('визначення', 'стилю'), 61) (('списку', 'сводеша'), 56) (('стопових', 'слів'), 48) (('визначення', 'автора'), 45) (('авторського', 'мовлення'), 33) (('опорних', 'слів'), 31) (('стилю', 'стилю'), 30)	(('ключових', 'слів'), 74) (('слів', 'в'), 24) (('web', 'mining'), 22) (('текстового', 'контенту'), 21) (('на', '2'), 20) (('визначення', 'ключових'), 19) (('ключових', 'в'), 19) (('визначення', 'слів'), 18) (('слів', 'для'), 18) (('на', 'крок'), 18)	(('на', 'основі'), 21) (('психологічного', 'стану'), 18) (('психологічного', 'особистості'), 17) (('контент', 'аналізу'), 16) (('стану', 'особистості'), 15) (('маркованих', 'слів'), 15) (('зрізу', 'психологічного'), 14) (('зрізу', 'стану'), 14) (('зрізу', 'особистості'), 14) (('особистості', 'на'), 14)
	ENG	(('of', 'the'), 186) (('the', 'of'), 169) (('of', 'of'), 152) (('of', 'a'), 81) (('the', 'the'), 75) (('the', 'author'), 66) (('and', 'of'), 63) (('in', 'the'), 57) (('of', 'author'), 57) (('of', 'words'), 55)	(('of', 'the'), 258) (('the', 'of'), 235) (('of', 'of'), 137) (('the', 'the'), 122) (('of', 'keywords'), 72) (('in', 'the'), 71) (('a', 'of'), 70) (('and', 'of'), 69) (('by', 'the'), 64) (('of', 'content'), 63)	(('the', 'of'), 304) (('of', 'the'), 243) (('the', 'the'), 168) (('of', 'of'), 162) (('is', 'the'), 154) (('of', 'a'), 91) (('the', 'is'), 76) (('the', 'content'), 71) (('is', 'of'), 61) (('and', 'the'), 57)
A_4	UA	(('слів', 'слів'), 88) (('стилю', 'автора'), 68) (('службових', 'слів'), 63) (('визначення', 'стилю'), 61) (('списку', 'сводеша'), 56) (('стопових', 'слів'), 48) (('визначення', 'автора'), 45) (('авторського', 'мовлення'), 33) (('опорних', 'слів'), 31) (('стилю', 'стилю'), 30)	(('text', 'content'), 30) (('web', 'mining'), 24) (('keywords', 'text'), 23) (('keywords', 'defined'), 22) (('stage', '1'), 20) (('analysis', 'text'), 18) (('step', '2'), 18) (('keywords', 'content'), 17) (('content', 'monitoring'), 17) (('step', '1'), 17)	(('на', 'основі'), 21) (('психологічного', 'стану'), 18) (('психологічного', 'особистості'), 17) (('контент', 'аналізу'), 16) (('стану', 'особистості'), 15) (('маркованих', 'слів'), 15) (('зрізу', 'психологічного'), 14) (('зрізу', 'стану'), 14) (('зрізу', 'особистості'), 14) (('особистості', 'на'), 14)

	ENG	(('fragment', 'fragment'), 37) (('reference', 'fragment'), 35) (('words', 'fragment'), 25) (('syntactic', 'words'), 21) (('frequency', 'fragment'), 19) (('swadesh', 'list'), 19) (('stop', 'words'), 18) (('author', 'style'), 17) (('fragment', '3'), 17) (('recognition', 'author'), 16)	(('ключових', 'слів'), 74) (('слів', 'в'), 24) (('web', 'mining'), 22) (('текстового', 'контенту'), 21) (('на', '2'), 20) (('визначення', 'ключових'), 19) (('ключових', 'в'), 19) (('визначення', 'слів'), 18) (('слів', 'для'), 18) (('на', 'крок'), 18)	(('content', 'analysis'), 40) (('psychological', 'personality'), 27) (('psychological', 'state'), 26) (('social', 'networks'), 22) (('marked', 'words'), 21) (('state', 'personality'), 20) (('based', 'analysis'), 19) (('psychological', 'based'), 18) (('state', 'based'), 18) (('based', 'content'), 18)

Table D

Absolute and relative frequencies of stopwords in the Excerpt and the standard

Fragment	Stop word	AF	RF	Part of speech	RF in fragment	Fragment	Stop word	AF	RF	Part of speech	RF in fragment
1 (107 words)	але	1	0.0093	Conjunction	0.0074	3 (162 words)	а	4	0.0247	Conjunction	0.0116
	в	2	0.0187	Preposition	0.0140		але	2	0.0123	Conjunction	0.0074
	для	3	0.0280	Preposition	0.0024		без	1	0.0062	Preposition	0.0008
	до	1	0.0093	Preposition	0.0113		бо	1	0.0062	Conjunction	0.0012
	з	1	0.0093	Preposition	0.0129		в	1	0.0062	Preposition	0.0140
	і	14	0.1308	Conjunction	0.0300		від	1	0.0062	Preposition	0.0034
	й	1	0.0093	Conjunction	0.0038		ж	1	0.0062	Conjunction	0.0033
	мов	1	0.0093	Participle	0.0022		з	4	0.0247	Preposition	0.0129
	не	2	0.0187	Participle	0.0237		за	2	0.0123	Preposition	0.0053
	про	2	0.0187	Preposition	0.0040		і	1	0.0062	Conjunction	0.0300
	та	2	0.0187	Conjunction	0.0047		й	4	0.0247	Conjunction	0.0038
	що	1	0.0093	Conjunction	0.0206		на	6	0.0370	Conjunction	0.0159
2 (117 words)	а	2	0.0171	Conjunction	0.0116		навіть	2	0.0123	Participle	0.0011
	в	3	0.0256	Preposition	0.0140		не	3	0.0185	Participle	0.0237
	від	1	0.0085	Preposition	0.0034		під	4	0.0247	Preposition	0.0011
	до	1	0,0085	Preposition	0.0113		таки	1	0.0062	Participle	0.0004
	ж	1	0.0085	Conjunction	0.0033		тож	1	0.0062	Conjunction	0.0001
	з	2	0.0171	Preposition	0.0129		у	4	0.0247	Preposition	0.0088
	за	1	0.0085	Preposition	0.0053		що	3	0.0185	Conjunction	0.0206
	і	2	0.0171	Conjunction	0.0300		щоб	1	0.0062	Conjunction	0.0028
	й	2	0.0171	Conjunction	0.0038		як	1	0.0062	Conjunction	0.0060
	на	1	0.0085	Preposition	0.0159	4 (149 words)	адже	1	0.00671	Participle	0.0011
	над	1	0.0085	Preposition	0.0005		але	2	0.01342	Conjunction	0.0074
	не	2	0.0171	Participle	0.0237		би	1	0.00671	Participle	0.0033
	ні	1	0.0085	Participle	0.0024		в	1	0.00671	Preposition	0.0140
	ось	1	0.0085	Participle	0.0012		ж	1	0.00671	Conjunction	0.0033
	от	1	0.0085	Participle	0.0005		з	3	0.02013	Preposition	0.0129
	се	1	0.0085	Participle	0.0074		за	1	0.00671	Preposition	0.0053
	хіба	1	0.0085	Participle	0.0006		і	4	0.02685	Preposition	0.0300
	хоч	1	0.0085	Participle	0.0010		мов	1	0.00671	Participle	0.0022
	що	2	0.0171	Conjunction	0.0206		на	7	0.04698	Preposition	0.0159
	як	1	0.0085	Conjunction	0.0060		не	4	0.02685	Participle	0.0237
							отсе	1	0.00671	Participle	0.0003
							при	1	0.00671	Preposition	0.0018
							про	2	0.01342	Preposition	0.0040
							се	1	0.00671	Participle	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

Table E

The result of the algorithm of analysis of the author's style of the publication

№	*N*	*W*	W_1	W_{10}	*P*	*Z*	*S*	K_l	K_s	K_z	I_{wt}	I_{kt}
1	671.3	395.6	299	6	44.2	57.1	41.1	0.59	0.89	0.76	0.76	0.015
2	662.5	410.3	303	5	37.8	39.8	34.8	0.61	0.9	0.67	0.74	0.012
3	668.8	418.3	325.8	6.8	29.8	56	57	0.63	0.93	1.28	0.78	0.016
4	708	419	309	8	36	64	28	0.59	0.91	0.85	0.74	0.019
5	661.1	402.7	299.7	4.7	44.7	54.7	24.8	0.61	0.89	0.6	0.74	0.012
6	694.5	417.4	313.1	6.4	54.3	58.5	38.1	0.6	0.87	0.62	0.75	0.015
7	691.8	403.4	301.6	7.8	47.8	60	47.8	0.58	0.88	0.79	0.75	0.019
8	682.5	394.2	291	5	49	61	39.7	0.58	0.88	0.74	0.74	0.013

9	733.5	486.5	392	5	50	65	45	0.66	0.9	0.76	0.8	0.01
10	729	380	261	7	62	75	32	0.52	0.84	0.58	0.69	0.018
11	686.5	414.5	312.6	5.9	41.1	56.9	45	0.6	0.9	0.86	0.75	0.012
12	665.5	399	299	6	35.5	72	43	0.6	0.91	1.09	0.75	0.015
13	724.2	394.2	278.8	5.8	59.6	68.4	36.8	0.55	0.85	0.61	0.71	0.015
14	691	396.7	289	7	39	55.3	42.3	0.57	0.9	0.85	0.73	0.018
15	745	439	319	6	45	59	61	0.59	0.9	0.89	0.73	0.014
16	768	452.5	323	5.5	51.5	58	47	0.59	0.89	0.68	0.71	0.012
17	647	422	308	3	62	50	32	0.65	0.85	0.44	0.73	0.007
18	677.5	373.5	255	6.5	64.5	72	36	0.55	0.86	0.57	0.68	0.018
19	680	379	251	5	42	55	33	0.56	0.89	0.7	0.66	0.013
20	642	337.5	230.3	7.8	44.8	52.3	56.8	0.52	0.87	0.81	0.68	0.023
21	665	376	275.7	7.7	41.7	65	32.3	0.57	0.89	0.79	0.73	0.02
22	731	420	301	7	49	71	54	0.57	0.88	0.85	0.72	0.017
23	691.7	425.7	331.3	6.5	41.8	58.2	50	0.62	0.9	0.88	0.78	0.015
24	668.8	368.3	262.5	6.8	44	55.8	34.5	0.55	0.88	0.73	0.71	0.018
25	691	421	311	4	47	65	40	0.6	0.89	0.74	0.74	0.01
26	708.5	434	323.5	6.5	42	57.5	47.5	0.61	0.9	0.84	0.75	0.015
27	665	406	309	5	41	42	28	0.61	0.9	0.57	0.76	0.012
28	700	418.5	320.5	6	40	68.5	35	0.6	0.9	0.88	0.77	0.014
29	704.5	412	303.5	5.5	59	47.5	38	0.58	0.86	0.49	0.74	0.013
30	688,8	416.8	321.9	6	49.7	49.3	41.3	0.6	0.88	0.67	0.77	0.016
31	711	396	268	6	60	67	19	0.56	0.85	0.48	0.68	0.015
32	691	436.7	336.7	5.7	40	51	44.7	0.63	0.91	0.82	0.77	0.013
33	695	422.5	318.3	7.5	38.5	61.3	41	0.6	0.91	0.89	0.75	0.018
34	699	427	314	6	49.5	60	41	0.61	0.88	0.69	0.74	0.014
35	683	438	339	4	38	52	42	0.64	0.91	0.82	0.77	0.009
36	730	440	323	6	42	62	39	0.6	0.9	0.8	0.73	0.014
37	714.5	418.5	304.5	6.5	46	65	48.5	0.59	0.89	0.86	0.73	0.016
38	717.5	433.5	321.5	6.5	56	57.5	26.5	0.6	0.87	0.5	0.74	0.015
39	728	430	313	6	49	59	51	0.59	0.89	0.75	0.73	0.014
40	667	401.5	305	6.5	40	63	35.5	0.6	0.9	0.82	0.76	0.016
41	715.5	352	223.5	8.5	45	58	34	0.49	0.87	0.68	0.63	0.024
42	699	401	302	6	46	68	32	0.57	0.89	0.72	0.75	0.015
43	620	411	323	2	36	55	40	0.66	0.91	0.88	0.79	0.005
44	645	403	302.3	4.3	39.3	58.7	37.7	0.62	0.9	0.84	0.74	0.011
45	708	475	392	5	49	83	46	0.67	0.9	0.88	0.83	0.011
46	708	442.5	336.5	5.5	43.5	62	56.5	0.63	0.9	0.91	0.76	0.012
47	689	458	369	7	44	65	36	0.66	0.9	0.77	0.81	0.015
48	1602	442	245	30	100	3	1	0.28	0.77	0.01	0.55	0.068
49	644	400	310	8	28	66	37	0.62	0.93	1.23	0.78	0.02
50	661.5	402.5	302	5	32	49.5	31	0.6	0.92	0.84	0.75	0.012
51	705	474	369	1	31	50	49	0.67	0.93	1.06	0.78	0.002
52	656	422.5	341.5	4.5	50	57.5	46	0.64	0.88	0.69	0.81	0.011
53	704.8	458.8	360	6	54.8	60	45.8	0.65	0.88	0.66	0.78	0.013
54	716	413.5	293	5.5	47	74.5	27.5	0.58	0.89	0.73	0.71	0.013
55	652	389	287	4	55	46	36	0.6	0.86	0.5	0.74	0.01
56	666	412	318	7	44	55	49	0.62	0.89	0.79	0.77	0.017
57	732	402	290	6	53	63	45	0.55	0.87	0.68	0.72	0.015
58	670	449	356	3	38	55	30	0.67	0.92	0.75	0.79	0.007
59	693	366	242	8	45	44	60	0.53	0.88	0.77	0.66	0.022
60	761	440	315.8	5.3	39.3	48.5	28.3	0.58	0.91	0.65	0.71	0.012
61	717	422	310	6	45	53	46	0.59	0.89	0.73	0.73	0.014
62	673.5	419	329	7.5	39	58	33	0.62	0.91	0.78	0.79	0.018
63	679	381	280	5	64	50	32	0.56	0.83	0.43	0.73	0.013
64	682.6	416.2	318	6.2	60	47.8	45	0.6	0.86	0.59	0.76	0.015
65	658	399	277	3	41	48	47	0.6	0.9	0.78	0.69	0.008
66	683	446	357	5.5	48.5	63	43.5	0.65	0.89	0.74	0.8	0.012
67	689.5	407.5	296	5.5	47.5	57	28	0.59	0.88	0.61	0.73	0.014
68	726	493	399	4	42	56	46	0.68	0.91	0.81	0.81	0.008
69	1325	538	360	19	66	9	2	0.4	0.88	0.06	0.67	0.035
70	697	450	361.5	5	56	59.5	46	0.65	0.88	0.63	0.8	0.011
71	652	405	296	2	34	45	28	0.62	0.92	0.72	0.73	0.005
72	598	386	309	4	83	40	0	0.65	0.78	0.16	0.8	0.01
73	726.3	441.3	332.3	6.7	51	61.3	39	0.6	0.88	0.68	0.75	0.015
74	846	440	299	10	54	57	26	0.52	0.88	0.51	0.68	0.023
75	712.5	442.5	331.5	4	51	48	33	0.62	0.88	0.53	0.75	0.009
76	706	374	275	8.5	45	68.5	31	0.53	0.88	0.74	0.73	0.023
77	682.3	398.7	296.3	4.7	39	50.3	37.7	0.58	0.9	0.75	0.74	0.012
78	654	361	240	5	39	35	28	0.55	0.89	0.54	0.66	0.014
79	631	350	249	7	34	45	31	0.55	0.9	0.75	0.71	0.02
80	661	391	275	4	63	53	24	0.59	0.84	0.41	0.7	0.01
81	709.5	399	292.5	9.5	48	58	49.5	0.56	0.88	0.75	0.73	0.024

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

Table F

Frequencies of appearance of letters in the standard and the studied passages

Letter	Ukrainian language (etalon) letters use Frequency	Fragment 1		Fragment 2		Letter	Ukrainian language (etalon) letters use Frequency	Fragment 1		Fragment 2	
		AF	RF	AF	RF			AF	RF	AF	RF
« »	0.133	80	0.14	82	0.15	я	0,024	15	0.03	6	0.01
о	0.082	37	0.07	41	0.08	з	0,018	9	0.02	8	0.01
а	0.074	43	0.08	31	0.06	б	0,016	7	0.01	5	0.01
н	0.068	33	0.06	30	0.06	ч	0,015	5	0.01	11	0.02
и	0.054	27	0.05	27	0.05	г	0,012	4	0.01	6	0.01
в	0.047	29	0.05	19	0.04	ю	0,012	2	0.00	2	0.00
т	0.046	25	0.04	20	0.04	б	0,011	7	0.01	5	0.01
е	0.038	26	0.05	45	0.08	х	0,01	4	0.01	7	0.01
р	0.036	15	0.03	16	0.03	ц	0,009	7	0.01	1	0.00
с	0.033	22	0.04	27	0.05	ж	0,007	3	0.01	7	0.01
м	0.031	10	0.02	13	0.02	й	0,007	4	0.01	6	0.01
к	0.031	22	0.04	20	0.04	ш	0,005	3	0.01	2	0.00
л	0.028	17	0.03	30	0.06	щ	0,004	3	0.01	1	0.00
д	0.028	16	0.03	4	0.01	ф	0,003	1	0.00	0	0.00
у	0.025	19	0.03	14	0.03	Others	0,0605	51	0.09	34	0.06
п	0.025	11	0.02	21	0.04						