Natural Language Texts Authorship Establishing Based on the Sentences Structure

Viktor Shynkarenko and Inna Demidovich

Ukrainian State University of Science and Technologies, 2, academician Lazaryan str., Dnipro, 49010, Ukraine

Abstract

Natural Language Texts Authorship Establishing was carried out on the basis of the hypothesis that each author has a peculiar the sentences structure forming style with different parts of speech. A natural language text was translated into a formal language generated by a formal stochastic grammar. For each product of the training sample, the corresponding stochastic formal grammar was restored. This method made it possible to reflect the author characteristic style in sentences building. On the basis of works statistical sample, inference rules and a probabilistic measure of their application were formed. The effectiveness of the proposed method was evaluated experimentally. In authorship establishing a probabilistic measure of the text belonging to a formal stochastic grammar was determined. To assess the reliability of the obtained results, the confidence interval of the probability measure was calculated. In the studies with the control sample, the possibility of the correct text authorship establishment is 75-80%.

Keywords

natural language texts, statistic analysis, text structure, text authorship, classification, parsing, confidence interval, formal stochastic grammar

1. Introduction

This article solves the problem of the text authorship determining by analyzing the sentences structure. It is important to note that the task of the texts authorship establishing, as well as the task of its attribution, is still relevant for today and covers a wide range of goals in various fields and is interesting to a number of specialists in various fields.

To determine the true author of a text, it is often necessary to turn to experts who can identify the author of an unknown text or determine whether a work belongs to another author using characteristic linguistic features and various stylistic devices. Expert text analysis takes a lot of time and is very laborious. In this regard, formal methods of different texts attribution have great prospects for automating the analysis process.

Currently, various approaches such as the theory of pattern recognition, mathematical statistics and probability theory, algorithms of neural networks and cluster analysis, and many others are used for text attribution [1, 2]. However, all methods used are not sufficiently effective. The particular difficulty is working with features that are characteristic of a particular language, which significantly complicates the task.

Working with the Ukrainian language, like other Slavic languages, has particular difficulties due to their structural complexity, as well as the variability of word forms and the possibilities of constructing sentences. Also, the complexity of the task is added by various styles of speech that are characteristic of a certain sphere of human activity, place of residence, age, education and subject of the text [3].

In this work, only literary works are used to determine authorship. The analysis of sentences in the text is carried out in order to form and formalize their structure.

COLINS-2022: 6th International Conference on Computational Linguistics and Intelligent Systems, May 12–13, 2022, Gliwice, Poland
EMAIL: shinkarenko_vi@ua.fm (V. Shynkarenko); 2019demidovichinn@gmail.com (I. Demidovich)
ORCID: 0000-0001-8738-7225 (V. Shynkarenko); 0000-0002-3644-184X (I. Demidovich)


2. Related works

Syntactic analysis of various text parts is a popular method for analyzing the author's work, its semantics, focus and main idea of the work. However, this type of various directions texts analysis is faced with the complexity of the syntactic model's automatic formation [4, 5, 6]. This is largely due to the complexity of the language structure itself, the variability of the word forms used and the very sentences structure. Despite this, method of text research like this carries the greatest amount of information about the author's style: regardless of the text subject, the syntactic structure of the author's language will clearly display his syllable.

Various studies of natural language formalization are known [7, 8]. One of the methods to work with natural language is using of grammars [9]. For example, similar studies were carried out for Italian [10] and Ukrainian [3].

Unlike the problem of text categorization, the goal of which is to determine a topic or list of topics for a text based on its content, text parsing abstracts from a specific area and tries to understand content-independent features of a text that are "linguistic expressions" of individual authors [11]. Such content-independent text properties are usually called stylometric features. For these purposes, various methods have been proposed and applied in problems of authorship attribution, including: frequency of words [12, 13], symbolic n-grams [14, 15, 16], auxiliary words, syllables [17] and parts of speech definition [18, 19].

The idea of using information about parts of speech is not new and has been successfully applied in a number of style classification problems, where, in particular, texts in English were processed [19, 2]. As a rule, their repeating sequences were extracted on the parts of speech basis.

The described approach is feasible for English texts, since the structure of the English language is quite strict and the word order in a sentence is clearly assigned to a certain part of speech. In addition, the words themselves do not have many different word forms, and when a prefix or suffix is added or removed, they go into another part of speech category. However, with using such method into Ukrainian, difficulties may arise. Unlike English, Ukrainian is a more variable language, and the number of forms for one word due to case, gender, and number changing significantly complicates the task. Moreover, the assignment of a word to one or another part of speech is ambiguous and difficult to perform within the automatic process.

The problem associated with establishing authorship need an individual approach in each case [20]. As a general rule, problems where the number of potential authors is small and the data samples are large are considered easy and high accuracy is expected. Complexity increases with an increase in the number of authors and a decrease in data volumes [21], which leads to a decrease in recognition accuracy.

3. Methods

3.1. Work structure rules formation

This paper explores a method for the texts authorship determining based on the sentence structure of the author's individual language.

Stochastic grammar is used to create rules that describe the structure of sentences in a text. For each rule, the probability of its application in a particular work is determined. The probability of inferring the whole sentence is defined as the product of the parts of speech sequences probabilities used in it. The resulting rules will generate a language that is specific for the explored and structurally similar works of a certain author.

To describe the main text structure, the parts of speech as a word characteristic were used. Thus, each word in the sentence is replaced by the part of speech that it is.

Each of the words in the text was analyzed for similarity with the parts of speech existing in the Ukrainian language. For service parts of speech: prepositions, pronouns, conjunctions and interjections, their list was used in all possible forms, and verbs, nouns, adverbs, adjectives, participles and participles were determined by comparison with the list of word endings.

When a corresponding word or its ending was found in one of the lists, the word was automatically replaced with the corresponding part of speech. If it was not possible to automatically determine the answer, the user was asked to then include the entire word or its ending (the data was entered manually by the user) in an already existing list.

The following tags were used to tag words in the text in Ukrainian: verb (v), noun (n), pronoun (prn), adjective (adj), conjunction (cnj), adverb (adv), preposition (prp), participle (prtcpl), interjection (intrj), gerund (ger).

For each part of speech, the probability of its occurrence in a certain place of the sentence in the given text is calculated. The probability of the certain part of speech appearance in the studied sequence will allow us to more accurately capture the individual writing style specific of each of the authors under study. After receiving the text in the form of parts of speech sequences set in sentences with the probability of their occurrence in a particular place, rules are formed.

To do this, all sentences starting with the same part of speech are grouped, the first word is discarded, and the procedure for calculating the probability is repeated for the next word.

After the sentences are again grouped according to the parts of speech at the beginning, the first word is again discarded and the probability for the next element is calculated, and so on. Probability is calculated as the number of cases in the text divided by their total number.

Table 1

The rules of the restored stochastic grammar according to "Etude" by I. Bahrianyi

Left part	Probability	Terminal	Nonterminal	Left part	Probability	Terminal	Nonterminal
σ	0,31	v	$A_{1,1}$	$A_{3,4}$	1,00	prp	$A_{4,3}$
$A_{1,1}$	0,17	ε		$A_{3,5}$	1,00	adv	$A_{3,3}$
$A_{1,1}$	0,13	n	$A_{2,1}$	$A_{3,6}$	1,00	v	$A_{4,4}$
$A_{1,1}$	0,04	prp	$A_{2,2}$	$A_{3,7}$	0,80	adj	$A_{3,7}$
$A_{1,1}$	0,04	cnj	$A_{2,3}$	$A_{3,7}$	0,20	ε	
$A_{1,1}$	0,21	v	$A_{2,4}$	$A_{4,1}$	1,00	intrj	$A_{5,1}$
$A_{1,1}$	0,13	prn	$A_{2,5}$	$A_{4,2}$	1,00	adj+n	G3
$A_{2,1}$	0,40	ε		$A_{4,3}$	1,00	n	$A_{5,2}$
$A_{2,1}$	0,40	n	$A_{3,1}$	$A_{4,4}$	1,00	adv	$A_{5,3}$
$A_{2,1}$	0,20	prp	$A_{3,2}$	$A_{4,5}$	1,00	v	$A_{5,4}$
$A_{2,2}$	1,00	n	$A_{3,3}$	$A_{5,1}$	1,00	n	$A_{6,1}$
$A_{2,3}$	1,00	v	$A_{3,1}$	$A_{5,2}$	0,33	ε	
$A_{2,4}$	0,20	adj+n	$A_{4,5}$	$A_{5,2}$	0,33	v	$A_{6,2}$
$A_{2,4}$	0,20	ε		$A_{5,2}$	0,33	adv	$A_{6,1}$
$A_{2,4}$	0,20	prp	$A_{2,1}$	$A_{5,3}$	1,00	n	$A_{3,3}$
$A_{2,4}$	0,20	v	$A_{3,4}$	$A_{5,4}$	1,00	prp	$A_{6,3}$
$A_{2,4}$	0,20	prn	$A_{3,5}$	$A_{6,1}$	1,00	v	$A_{7,1}$
$A_{2,5}$	0,33	n	$A_{3,6}$	$A_{6,2}$	1,00	v	$A_{7,3}$
$A_{2,5}$	0,33	adj	$A_{3,7}$	$A_{6,3}$	1,00	prtcpl	$A_{7,2}$
$A_{2,5}$	0,33	v	$A_{3,4}$	$A_{7,1}$	1,00	prp	$A_{5,3}$
$A_{3,1}$	0,50	ε		$A_{7,2}$	0,33	adj+n	$A_{7,2}$
$A_{3,1}$	0,20	cnj	$A_{4,1}$	$A_{7,2}$	0,67	ε	
$A_{3,2}$	1,00	prn	$A_{4,2}$	$A_{7,3}$	1,00	n	$A_{3,3}$
$A_{3,3}$	1,00	ε					

Thus, the substitution rules for some product T have an initial non-terminal, then terminals corresponding to each word in the sentence and the probability of applying the corresponding rule when parsing the text and have the form:

$$\sigma \xrightarrow{p_{1j}} b_{1j} A_{1,j}, \quad A_{i,j} \xrightarrow{p_{i+1,k}} b_{i+1,k} A_{i+1,k}, \quad j = 1 \ldots J_i, \; k = 1 \ldots K_i$$

where σ – initial non-terminal, b_{ij} – terminals corresponding to the *i*-th word in the sentence (and corresponding to the *i*-th rule applied when parsing the sentence or the *i*-th level of the rule), $A_{i,j}$ – *j*-th non-terminal in the *i*-th level rule, $p_{i,k}$ – the probability of applying the corresponding rule when parsing this work, J_i, K_t – is the number of different non-terminals in the right part of the rules of the *i-1*-th level and *i*-th level, respectively.

The level corresponds to the ordinal number of the word in the sentence. Several alternative rules are allowed with a non-terminal on the left side of the rule, but the terminals on the right side of such rules are different, which ensures deterministic parsing. Thus, the text is presented as a set of rules that describe its structural features using the rules described above. The symbol ε stands for empty (end of rule).

An example of the one automatically restored set of rules is presented below in Table 1. The rule describes all 24 sentences in the text "Etude" by I. Bahrianyi, with a verb in the beginning. As can be seen from the presented probabilities, in the studied work, 31% of sentences will begin with a verb. And the percentage of sentences consisting of only one word, a verb, is 17%.

Examples of the first few rules according to the table are: $\sigma \xrightarrow{0.31} vA_{1,1}$; $A_{1,1} \xrightarrow{0,17} \varepsilon$; $A_{1,1} \xrightarrow{0,13} nA_{2,1}$. On the left side of the rule is a non-terminal, then the probability of its application is indicated, and on the right side of the rule is a terminal with a non-terminal to go to the next rule.

When using this text method, a sentence from the work "Etude" by I. Bahrianyi presented as a sequence of parts of speech included in it will have the following form Table 2.

Table 2

Sentence tagging and corresponding probabilities of stochastic grammar rules

Word in a sentence	Tag	Probability
Чорні	adj	0,06
ґрати	n	0,6
розпанахали	v	0,6
небо	n	0,125

3.2. Comparison of two works

To compare two works, they must be presented in the form of a restored formal stochastic grammar with the rules, the formation of which is described above. Each sequence of rules in one text is compared with each sequence of rules in another text. Let the rules for some text T_i be formed like:

$$\sigma' \xrightarrow{p'_{1j}} b'_{1j} A'_{1,j}, \quad A'_{i,j} \xrightarrow{p'_{i+1,k}} b'_{i+1,k} A'_{i+1,k}, \quad j = 1 \ldots J'_i, \; k = 1 \ldots K'_i .$$

Let's say in texts T_i and T_k there are sentences of similar structure, such as

$$S_i : \sigma \Rightarrow^{p_{1j_1}} b_{1j_1} A_{1j_1} \Rightarrow^{p_{2j_2}} b_{2j_2} A_{2j_2} \ldots \Rightarrow^{p_{lj_l}} b_{lj_l}$$

$$S_k : \sigma \Rightarrow^{p'_{1j_1}} b'_{1j_1} A'_{1j_1} \Rightarrow^{p'_{2j_2}} b'_{2j_2} A'_{2j_2} \ldots \Rightarrow^{p'_{ij_l}} b'_{lj_l} \text{ и } b_{ij_i} = b'_{ij_i}$$

Assume that two texts under study contain sentences of a similar structure (S_i and S'_k), then the degree of their statistical structural similarity will be determined as the product of the minimum difference between the probabilities of applying the corresponding rule:

$$\rho(S_i, S'_k) = \prod_{m=1}^{l} \min(p_{mj_m} - p^i_{mj_m})$$

And the degree of two works texts statistical structural similarity as the sum of the all its sentences similarity degrees.

The degree of texts statistical structural similarity T_i and T_k :

$$\rho(T_1, T_2) = \sum_{i=1}^{N} \rho(S_i, S'_i), \quad (1)$$

where S'_i – structurally similar sentence to sentence S_i according (1), N – the number of sentences in any of these works, if the text T_2 structurally similar sentence to sentence S_i is absent, then $\rho(S_i, S'_i) = 0$.

notice, that $\rho(T_i, T_k) = \rho(T_k, T_i)$ $\rho(T_i, T_i) = 1$ complete match, $\rho(T_i, T_k) = 0$ – if in texts T_i and T_k no sentences have the same structure.

Formation of formal stochastic grammars was carried out for all the works of each author in the training sample, generating the language specific for a particular author. For determining the similarity of a work according to (1), the formal stochastic grammar corresponding to the work from the control sample was used as T_1, and the stochastic grammar for all works of the potential author altogether was used as T_2.

3.3. Calculation of the confidence interval boundary values using Student's t-test

To obtain more reliable results, we calculated confidence intervals for each of the authors in the sample. Student's t-test was applied [22].

To calculate the confidence interval in the training sample for each of the authors, their presented texts similarities to each other were calculated. Data on the similarity of texts within the training sample for each of the authors was divided into three parts with the same number of components.

The following formula was used to calculate the confidence interval:

$$t_{2,\beta} \sqrt{\frac{1}{6} \sum_{k=1}^{3} (\zeta_k - \theta_s)^2},$$

where $t_{2,\beta}$ – Student's t-test, β – confidence level, ζ_k – the average value of k-th sample part, θ_S – the average value over the entire sample.

4. Results

4.1. Training and control samples formation

During the experiment, the authorship of natural language texts was determined by two samples.

For the first experiment, 20 works of literary texts by 10 Ukrainian authors were selected in the training sample. The control sample consists of 3 works by each author.

The works of the following authors are presented: IB – I. Bahrianyi, AV – A. Vyshnia, MV – M. Vovchok, AD – A. Dovzhenko, HK – H. Kvitka-Osnovianenko, PM – P. Myrnyi, VN – V. Nestaiko, VP – V. Pidmohylnyi, IF – I. Franko, MK – M. Khvylovyi.

For the second experiment, both samples were doubled, respectively, the training sample included 40 works by the same authors, and 60 texts made up a new control sample - 6 works by each author.

The choice of literary texts is due to the availability of reliable information about the works authorship and the presence of each author specific style.

4.2. Experiment results

At Figure 1 and Figure 2 the results of the experiments are presented. Each bar in the chart represents the works of a particular author from the control sample. The column is divided into two zones, where the blue part displays the number of texts with correctly identified authorship, and the orange part shows the number of texts with erroneous ones.

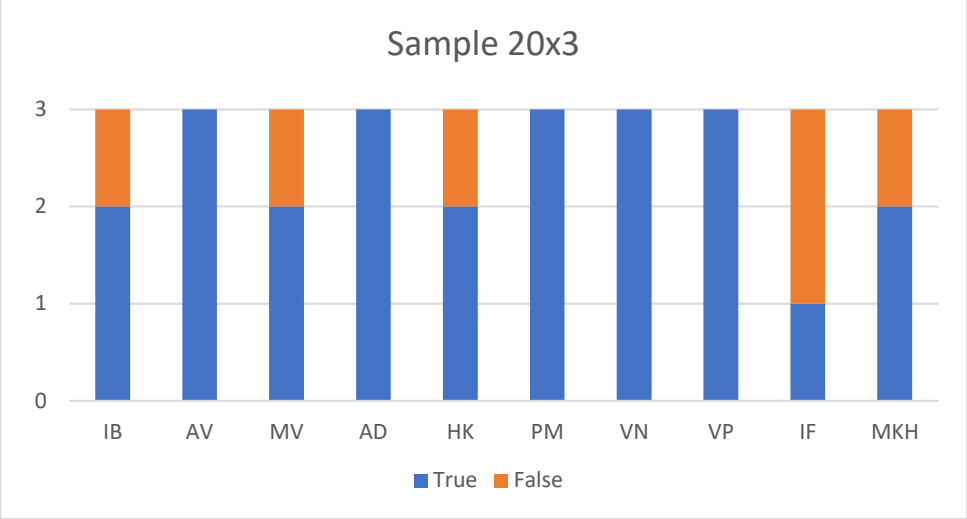

Figure 1: Authorship establishing results with a sample 20x3

According to the results obtained, working with the smaller of the two samples, containing a total of 200 works by 10 authors in the training sample (20 for each author), and 30 works in the control sample, cases of authorship correct attribution is 24, which is 80%.

The best result was obtained working with the works of A. Dovzhenko, P. Myrnyi, V. Nestaiko and V. Pidmohylnyi. I. Franko turned out to be the author with the most difficult to define style.

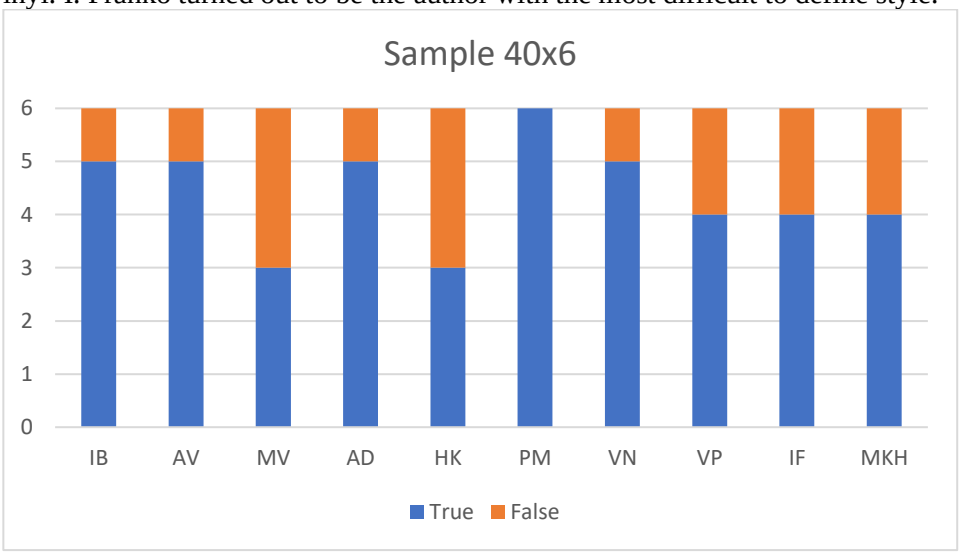

Figure 2: Authorship establishing results with a sample 40x6

According to theobtained results, working with a larger doubled sample (400 texts of 10 authors in the training sample and 60 works in the control sample), the number of correctly determined cases is 45, which is 75%.

The best result was obtained working with the works of P. Myrny. The authors of the most difficult cases to detect are H. Kvitka-Osnovianenko and M. Vovchok - the athurship of only a half of their works in the sample was correctly determined.

To obtain better result, a confidence interval was introduced Table 3.

Confidence intervals of different authors differ significantly from each other. So, on average, the interval ranges from 0.04 to 0.08, however, for A. Vyshnia and M. Vovchok it is much larger, and amounted to 0.37 and 0.12, respectively. The minimum interval was 0.02 for V. Pidmohylnyi.

For some of the authors, such as H. Kvitka-Osnovianenko and M. Vovchok, a special style of narration is characteristic, which is difficult to classify and structure, due to which they are characterized by a low recognition result.

Whereas the authors I. Bahrianyi, A. Vyshnia, A. Dovzhenko, P. Myrnbi and V. Nestaiko have a more individual style of writing, which is displayed in the sentence structure and allows establishing their authorship with high accuracy.

Table 3
Confidence interval by authors

	IB	AV	MV	AD	HK	PM	VN	VP	IF	MKH
Average	0,83	0,83	0,61	0,83	0,58	0,97	0,87	0,69	0,67	0,66
Max	0,86	1,00	0,67	0,86	0,60	0,99	0,89	0,70	0,71	0,69
Min	0,80	0,64	0,55	0,80	0,56	0,95	0,85	0,68	0,63	0,63
Range	0,05	0,37	0,12	0,06	0,05	0,04	0,03	0,02	0,07	0,07

Taking into account the confidence intervals, the following results were obtained Table 4.

Table 4
Results of authorship determination working with the confidence interval

In the interval	1 actual author	1 not actual author	more than 1 author, actual is among them	more than 1 author, actual isn't among them
Number	14	2	36	8
Per cent	23,3%	3,3%	60%	13,3%

According to the results obtained, for 14 cases, as a result of authorship determination, one correctly recognized candidate was obtained, which amounted to 23.3%. In 36 cases, the program was able to narrow the number of applicants to 3, while in 31 cases the applicant with the greatest structural similarity was the correctly recognized author (which accounted for 51.7% of the total sample) and in another 5 cases the correctly recognized author was included in the list of candidates (8.3% of the total sample). Also, in only 8 cases, the authorship of the text was not determined - none of the submitted candidates was correctly recognized, which amounted to 13.3%.

As a result of the obtained data analysis, it can be argued that the size of the confidence interval can also be considered a specific feature of the author's personal style. So, for some of them, the size of the confidence interval differs significantly from other authors. It is much larger in A. Vyshnia and M. Vovchok. It is noteworthy that A. Vyshnia is characterized by his own style of writing, which made it possible to determine with sufficient accuracy the texts of his authorship, while the style of M. Vovchok is rather difficult to classify. The minimum interval was obtained for the works of the author V. Pidmohylnyi, which, however, leads to rather high results of his classification.

Thus, the percentage of correct authorship identification, taking into account the confidence interval, was improved to 83.33% (in 50 cases out of 60, the author of the work was correctly recognized).

5. Discussion

According to our previous studies [3, 23, 24, 25] on establishing the texts authorship, the initial result of reliability ranged from 18% of the correct texts attribution to 82% [23] using character-by-character analysis. Later, the result was improved with a range from 80% to 91% using N-grams [24], genetic algorithm [25] and also working with stems and dictionaries [3].

In this paper, the proposed method was used for the first time and results of 75-80% were obtained, which, taking into account the peculiarities of the Ukrainian language and the complexity of its formalization for solving the task, correspond to the spread in the percentage of determining the texts authorship in works devoted to this topic. So, for example, the result of determining authorship is in the range from 74% to 92% correctly identified cases [26-32]. These results varied depending on the used method, the language and style of the analyzed text.

Working with different foreign languages and taking into account their distinctive features, the authors used various methods.

In a study of English poets works authorship [30] using the architectures of a convolutional neural network, a multilayer perceptron, and LSTM neural network, the results ranged from 74-83%. In another work, using various stylometric functions and algorithms, also in English authors works, the success rate was 82% [20]. Analyzing text corpora in English and Spanish, working with syllables, the achieved result was 78.8%. For the Russian language, which is similar in complexity and structure to Ukrainian, working with a combination of support vector machine and a genetic algorithm, such results as 82.3% were achieved [27]. And using N-grams in processing the Portuguese language in [31], the result reached 72%.

Working with Ukrainian texts, predominantly journalistic style, in view of the structural variability and complexity of the language, using methods such as neural networks [28] and the Quantitative Method for Automated Text Authorship Attribution Based on the Statistical Analysis of N-grams Distribution [29] and working with scientific articles, the authors obtained results of 92% and 79%, respectively.

The case of using confidence intervals in determining the authorship of texts has not been found in recent works, which allows us to assert the novelty of this approach.

Among the works related to text tagging, one can find works devoted to the N-grams [6, 33, 34].

For example, a study [6] of the Coptic language, which is the last phase of the Egyptian language family and a descendant of the ancient Egyptian script, was conducted to assess the success of tagging the study of genre, style and authorship in the Coptic language. The results of the study show a relatively high accuracy of 94-95% correct automatic tagging for literary texts.

In [33] the authors focused on the attribution of the Polish texts authorship using stylometric features based on part-of-speech tags. The Polish language is characterized by a high level of inflection, so the authors managed to distinguish more than 1000 tags, which made it possible to build a fairly large feature space by processing texts and performing their classification using machine learning methods. The use of this method in highly inflected languages, including Polish, is considered by the authors to be a promising direction in authorship attribution.

In [34] for the authorship attribution problem, the use of part-of-speech skip-grams and an in-house top-k sequential pattern mining algorithm is considered. The authors of the study come to an accuracy of 86-97% for various authors in training sample.

Given the differences in the analyzed languages in the presented studies, we can conclude that the method proposed here is a promising direction for working with Ukrainian literary texts and will significantly improve the results obtained.

6. Conclusions

The paper proposes a new method of text attribution using stochastic formal grammars. Since all known methods do not give high accuracy and do not take into account the sentence constructing rules, there is a need to search for new additional methods and new attributes. The characteristic of the author’s style in the aspect of the sentence constructing is a previously unexplored sphere. Its use, in combination with already known methods, can increase the efficiency of determining the natural language texts authorship.

Analyzing texts with a sample 20x3, the number of matches for the author was 24 out of 30 works. And working with doubled samples, the result is also positive, but to a lesser extent - 45 out of 60 matches. The results obtained were 80% and 75%, respectively.

Taking into account the confidence interval, the results were improved to 83.3%. As a result of the analysis, it can be argued that the size of the confidence interval can also be considered a characteristic feature of the author's personal style. Thus, a large confidence interval may indicate a low level of differentiation of the author's style and, as a result, a poor result in determining the authorship of his works. And vice versa - with a small confidence interval, the probability of confident the author's style differentiation increases significantly.

The average value of the works similarity in the training sample is also significant - the higher the value, the more clearly the style of the author is determined and, accordingly, the result of determining his works authorship is higher.

In the future, it is planned to improve the presented method to obtain a better result, and for the same purpose, it can be combined with previously studied methods. The possibility of more detailed tagging of sentences parts is considered - working not only parts of speech, but also forms, numbers, gender, etc. for the word under study. This approach will provide more information about the structure of sentences and the rules for their construction, specific to a particular author.

7. References

[1] R. A. Hardcastle, CUSUM: a credible method for the determination of authorship?, Science & Justice: Journal of the Forensic Science Society 37(2) (1997) 129-138. doi:10.1016/s1355-0306(97)72158-0.

[2] P Juola, GK Mikros, S Vinsick, A comparative assessment of the difficulty of authorship attribution in Greek and in English, Journal of the Association for Information Science and Technology 70 (1) (2019) 61-70. doi:10.1002/asi.24073.

[3] V. Shynkarenko, O. Kuropiatnyk, Constructive Model of the Natural Language, Acta Cybernetica. 23, 4 (2018) 995–1015. doi:10.14232/actacyb.23.4.2018.2.

[4] Y. Zhang, H. Kamigaito, M. Okumura, A Language Model-based Generative Classifier for Sentence-level Discourse Parsing, in: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, Online and Punta Cana, Dominican Republic, Association for Computational Linguistics, 2021, pp. 2432–2446. doi:10.18653/v1/2021.emnlp-main.188.

[5] J. Li, M. Liu, B. Qin, et al. A survey of discourse parsing, Front. Comput. Sci. 16 (2022). doi:10.1007/s11704-021-0500-z.

[6] A. Zeldes, C. T. Schroeder, Computational Methods for Coptic: Developing and Using Part-of-Speech Tagging for Digital Scholarship in the Humanities, Digital Scholarship in the Humanities 30 (2015) i164–i176. doi:10.1093/llc/fqv043.

[7] I. Buzhinsky, Formalization of natural language requirements into temporal logics: a survey, 2019 IEEE 17th INDIN, (2019), 400-406. doi:10.1109/INDIN41052.2019.8972130.

[8] P. Linz, SH. Rodger, An introduction to formal languages and automata, Jones & Bartlett Learning, 2022.

[9] M. Silberztein, A new linguistic engine for nooj: Parsing context-sensitive grammars with finite-state machines, in: Proceedings of the 11th International NooJ Conference Formalizing Natural Languages with NooJ and Its Natural Language Processing Applications, Kenitra, Morocco, 2017, pp. 240– 250. doi:10.1007/978-3-319-73420-0_20.

[10] A. Mazzei, V. Lombardo, Building a large grammar for Italian, in: LREC, 2004.

[11] M. Gamon, Linguistic correlates of style: authorship classification with deep linguistic analysis features, in: Proceedings of the 20th International Conference on Computational Linguistics. COLING 2004, Stroudsburg. Association for Computational Linguistics, 2004, doi:10.3115/1220355.1220443.

[12] C. Akimushkin, DR. Amancio, ON. Oliveira Jr., On the role of words in the network structure of texts: Application to authorship attribution, Physica A: Statistical Mechanics and its Applications 495 (2018) 49-58. doi:10.1016/j.physa.2017.12.054.

[13] DL. Hoover, The microanalysis of style variation, Digital Scholarship in the Humanities, Issue suppl_2 32 (2017) ii17-30. doi: 10.1093/llc/fqx022.

[14] Y. Sari, A. Vlachos, M. Stevenson, Continuous N-gram Representations for Authorship Attribution, in: Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers, Valencia, Spain, 2017, pp. 267-273. doi:10.18653/v1/E17-2043.

[15] I. Markov, J. Baptista, O. Pichardo-Lagunas, Authorship Attribution in Portuguese Using Character N-grams, Acta Polytechnica Hungarica 14(3) (2017) 59–78. doi:10.12700/APH.14.3.2017.3.4.

[16] H. Gómez-Adorno, JP. Posadas-Durán, G. Sidorov, Document embeddings learned on various types of n-grams for cross-topic authorship attribution, Computing 100 (2018) 741–756. doi:10.1007/s00607-018-0587-8.

[17] G. O. Sidorov, Automatic Authorship Attribution Using Syllables as Classification Features, Rhema journal 1 (2018) 62-81.
[18] E. Stamatatos, A survey of modern authorship attribution methods, J. Am. Soc. Inf. Sci. Technol 60 (3) (2009) 538–556.
[19] M. Koppel, J. Schler, S. Argamon, Computational methods in authorship attribution, J. Am. Soc. Inf. Sci. Technol 60 (1) (2009) 9–26. doi:10.1002/asi.20961.
[20] M. Koppel, J. Schler, S. Argamon, Authorship attribution: what's easy and what's hard?, Lecture Notes in Computer Science 7499 (2012) 282-289. doi:10.1007/978-3-642-32790-2_34.
[21] K. Luyckx, W. Daelemans, The effect of author set size and data size in authorship attribution, Lit. Linguist. Comput. 26 (1) (2011) 35–55.
[22] P. Mishra, U. Singh, C. M. Pandey, P. Mishra, G. Pandey, Application of student's t-test, analysis of variance, and covariance, Annals of cardiac anaesthesia 22(4) (2019) 407–411. doi:10.4103/aca.ACA_94_19.
[23] V. I. Shynkarenko, I. M. Demidovich, Determination of the attributes of authorship of natural texts, Artificial Intelligence 3 (2018) 27-35.
[24] V. I. Shynkarenko, I. M. Demidovich Authorship Determination of Natural Language Texts by Several Classes of Indicators with Customizable Weights, in: Proceedings of the 5th International Conference on Computational Linguistics and Intelligent Systems (COLINS 2021), Volume I: Main Conference, Lviv, Ukraine, April 22-23, 2021, pp. 832-844.
[25] I. Demidovich, V. Shynkarenko, O. Kuropiatnyk, O. Kirichenko, Processing Words Effectiveness Analysis in Solving the Natural Language Texts Authorship Determination Task, in: Proceedings of the XVI International Scientific and Technical Conference (CSIT'2021), Lviv, Ukraine, 2021. doi: 10.1109/CSIT52700.2021.9648829.
[26] R. Iyer, C. Rosé, A Machine Learning Framework for Authorship Identification FromTexts. ArXiv abs/1912.10204 (2019).
[27] J. Rygl, A. Horák, Authorship Attribution: Comparison of Single-Layer and Double-Layer Machine Learning, Lecture Notes in Computer Science 7499 (2012) 282-289. doi:10.1007/978-3-642-32790-2_34.
[28] M. Lupei, A. Mitsa, V. Repariuk, V. Sharkan, Identification of authorship of Ukrainian-language texts of journalistic style using neural networks, Eastern-European Journal of Enterprise Technologies 1(2) (2020) 30–36. doi:10.15587/1729-4061.2020.195041.
[29] 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) (2019) 28-51. doi:10.15587/1729-4061.2019.186834.
[30] V. Moshkina, I. Andreeva, N. Yarushkinaa, Solving the problem of determining the author of text data using a combined assessment, CEUR Workshop 2782 (2020) 112-118.
[31] I. Markov, J. Baptista, O. Pichardo-Lagunas, Authorship Attribution in Portuguese Using Character N-grams, Acta Polytechnica Hungarica 14(3) (2017) 59–78. doi:10.12700/APH.14.3.2017.3.4.
[32] G. O. Sidorov, Automatic Authorship Attribution Using Syllables as Classification Features, Rhema journal 1 (2018) 62-81.
[33] P. Szwed Authorship Attribution for Polish Texts Based on Part of Speech Tagging. In: Kozielski S., Mrozek D., Kasprowski P., Małysiak-Mrozek B., Kostrzewa D. (eds) Beyond Databases, Architectures and Structures. Towards Efficient Solutions for Data Analysis and Knowledge Representation. BDAS 2017. Communications in Computer and Information Science, vol 716. Springer, Cham. (2017) doi:10.1007/978-3-319-58274-0_26.
[34] J. Pokou, F. Fournier-Viger, Ch. Moghrabi, Using Frequent Fixed or Variable-Length POS Ngrams or Skip-Grams for Blog Authorship Attribution, CPCI-S 收录 (2019) 63-74.