DOI https://doi.org/10.28925/2412-2491.2025.2512
UDC 81`33

SYMBOL FREQUENCY AS A COMPONENT OF THE STATISTICAL PROFILE OF M.YATSKIV'S SHORT STORIES IDIOLECT

Tsiokh L. Y.
Lviv Polytechnic National University,
larysa.y.tsokh@lpnu.ua
ORCID ID: https://orcid.org/0000-0003-2695-4411

Shyika Y. I.
Lviv Polytechnic National University,
yuliia.i.shyika@lpnu.ua
ORCID ID: https://orcid.org/0000-0003-2474-0479

This article presents a quantitative analysis of literary text at the graphological and phonetic levels. The study is based on the experimental research corpus of short stories by M. Yatskiv. A text array has been created for statistical research at the symbol (grapheme) level. The absolute (number) and relative frequency of each symbol of the extended Ukrainian alphabet has been calculated in the entire text array. Based on these frequencies, the rank of each symbol has been determined, and entropy has been calculated using the standard formula for the corpus as a whole, as well as for sequentially and randomly selected text segments of 108 characters each. For the entire text array and separately for sequentially and randomly selected segments, the distribution of characters by type and the euphony of the text have been calculated. Euphony has been defined as the proportion of vowels, sonorants, and voiced consonants in the text. The degree of correspondence between the frequencies of characters in the entire corpus and in the segments has been assessed using Pearson's chi-squared test. The frequency distribution of characters in the research text array has been taken as the hypothetical theoretical distribution function, and the chi-squared statistics have been calculated for each segment. The null hypothesis stated that "the frequency distribution of characters in a given segment does not differ from the corresponding distribution in the full text." Simultaneously, the rank of each character in the frequency distribution has been determined for every segment. All calculations have been made using programs in Python. The results have been compared with similar analysis of short stories by Vasyl Stefanyk. The findings demonstrate that even a randomly selected segment as small as one hundredth of the corpus can approximate the overall frequency distribution of *characters with high probability. Moreover, the results indicate that for novellas as a genre with stable structural elements within a specific period of a national literature (e.g., Yatskiv and Stefanyk), linguistic statistics show minimal variation.*

Key words: *text corpus, text array, writer's idiolect, sequential samples, random samples, entropy, melodiousness of the text.*

Цьох Л.Й., Шийка Ю.І. Частота використання символів як компонент статистичного профілю ідіолекту новелістики М. Яцкова

Стаття присвячена квантитативному аналізу художнього тексту на графологічному та фонетичному рівні. Дослідження проводиться на матеріалі експериментального дослідницького корпусу новел М. Яцкова. Для проведення статистичних досліджень на рівні символів (графем) з новел М. Яцкова було утворено текстовий масив. Обраховано абсолютну (кількість) та відносну частоту кожного символу розширеної української абетки у цілому текстовому масиві. За отриманими частотами для кожного символу було визначено його ранг та пораховано ентропію за відповідною формулою у текстовому масиві в цілому, для випадково послідовних та вибраних сегментів тексту довжиною 108 символів. Для цілого текстового масиву та окремо для послідовно і випадково вибраних сегментів було обраховано розподіл символів за типами та милозвучність тексту. Милозвучність визначали як відсоток сукупності голосних, сонорних та дзвінких букв. За результатами проведених обчислень було проаналізовано відповідність частот символів у цілому тексті та вибраних сегментах. За теоретичне підґрунтя було взято критерій згоди К. Пірсона . За гіпотетичну теоретичну функцію розподілу було прийнято частотний розподіл символів у дослідницькому масиві тексту. Для кожного вибраного сегменту обчислювали статистику критерію Пірсона. За нульову гіпотезу прийняли твердження: "у тексті-вибірці розподіл частот символів розширеної української абетки не відрізняється від відповідного розподілу в тексті-репрезентанті". Паралельно для тексту-вибірки визначено ранг кожного символу в частотному розподілі. Всі обчислення виконано за допомогою власних програм на мові Python. Отримані результати порівняно з аналогічними для новел Василя Стефаника. Доведено, що при випадковому виборі сегменту, розмір якого становить навіть одну соту обсягу тексту, можна отримати з високою вірогідністю частоту символів у всьому тексті. Показано, що для новели як жанру зі стійкими жанротворчими елементами в певних часових рамках конкретної національної літератури (Яцків – Стефаник) лінгвостатистичні показники будуть мати мінімальне розходження.

Ключові слова: *текстовий корпус, текстовий масив, ідіолект автора, послідовні вибірки, випадкові вибірки, ентропія, мелодійність тексту.*

Introduction. Modern linguistic research is characterized by the integration of diverse methods and techniques, comprehensive approaches to the study of linguistic phenomena, the inclusion of perspectives from various scientific disciplines, and the use of advanced technical and informational resources.

Statistical methods are used in linguistics because language has measurable properties, the internal interdependence between the qualitative and quantitative aspects of linguistic structure, the relationship between the frequency of linguistic units in speech and certain statistical patterns, and the ability to obtain objective data independent of the researcher's subjective perception.

Quantitative methods enable the understanding of the quantitative characteristics of the research object, the exploration of the qualitative characteristics underlying these quantitative aspects, and the identification of the processes through which quantitative features transit into a new quality. In contemporary research, quantitative calculations are greatly simplified and accelerated by specialized computer software. Such software processes large datasets quickly. It provides reliable estimates of frequencies, tests the validity of selected features, and verifies conclusions reached by other methods.

Theoretical Background. In modern linguistic stylistics, terms such as "style," "author's style," "idiostyle," "idiolect," "individual style," "style of the author," "author's idiolect" are often used interchangeably to denote the characteristic features of a writer's individual speech, yet they lack clear and transparent definitions and criteria for differentiation. Researchers primarily focus on the functioning of linguistic means in literary texts. B. Stelmakh identifies the dominant features of an author's idiolect as peculiarities of use of particular vocabulary layers, sound, visual and olfactory imagery, the condensation or non-condensation of lexemes, punctuation, stylistic variation of syntactic constructions, rhythm, euphony, and narrative composition (Stelmakh, 2004, p. 231). O. Pavlyshenko identifies lexical units that are either unique to a specific author's works or appear more frequently in them than in the works of others as markers of the author's idiolect (Pavlyshenko, 2004, p. 314). A. Naumenko interprets idiolect as the style of an individual, representing a level of the speech system below functional style (Naumenko, 2003, p. 203), while O. Selivanova defines idiolect as an individual version of language, reflected in a speaker's unique set of speech characteristics. An idiolect includes both elements of linguistic norms and usage, and demonstrates the level of individual linguistic activity. In written texts, an idiolect reveals traits of idiostyle (Selivanova, 2008, p. 173).

The structure of a writer's idiostyle manifests itself at all linguistic levels, from phonetics to complex syntactic constructions. The linguistic analysis of style as a system is most effectively conducted at the level of language structure, with the lexical and syntactic levels considered the most productive in style creation, leading to numerous studies focused on these aspects (Karasov & Levchenko, 2022; Lototska & Saban, 2023; Seminck et al., 2022).

However, the graphological and phonetic levels remain the least studied, particularly in terms of quantitative analysis through corpus-based methods. In our opinion, a writer's idiolect can also be identified through statistical indicators, such as the frequency of symbols, which could serve as a basis for analysing the phonetic level – a level whose potential for distinguishing style should not be underestimated.

The works of the talented Western Ukrainian writer Mykhailo Yatskiv occupy a prominent place in the history of Ukrainian literature as a distinctive and complex phenomenon. His works have been the subject of extensive literary and scientific debates; his short stories have been widely translated into Czech, Polish, German, French, and even Japanese, and included in anthologies of the world's best short stories. Despite the considerable scholarly attention that Yatskiv's work has received (Tkachuk, 2013; Kryvuliak, 2007; Melnyk, 2011), further study remains relevant, particularly through the application of contemporary theoretical and methodological approaches, including corpus technologies and linguistics.

Methodological notes. For this study, 42 short stories by M. Yatskiv from the collection "Black Wings" (Yatskiv, 2016) have been selected. General quantitative characteristics of the research corpus are as follows: characters – 232184, word uses – 53978, word forms – 30424. For comparative purposes, the corpus of Vasyl Stefanyk's short stories has been used, with the following quantitative characteristics: characters – 232184, word uses – 53978, word forms – 30424.

To conduct a statistical analysis at the character (graphemes) level, a text array has been created from M. Yatskiv's short stories, including only the characters of the extended Ukrainian alphabet (letters, spaces, hyphens, and apostrophes). This array then has been divided into paragraphs of equal length, each consisting of 108 characters; if the last paragraph was less than 108 characters long, it was excluded from the final dataset. The study has been conducted according to the methodology proposed by I. Kulchytskyi (Kulchytskyi, 2019a; Kulchytskyi, 2019b), which involved the following steps:

Calculating the absolute (total count) and relative frequency of each character in the extended Ukrainian alphabet across the entire text array.

Determining the rank of each character based on its frequency and calculating its entropy using the formula (Seminck et al., 2022):

$$H = -\sum_{i=1}^{36} p_i \log_2 p_i \quad (1)$$

where $p_i$is relative frequency of the i-th character.

Determining the size of the text segment in paragraphs of 108 characters. The sample size has been determined by dividing the number of paragraphs in the text by 100. The remainder has been used to organize the sequential selection of text segments.

Sequentially selecting segments of the specified length. Segments of 108 characters have been selected sequentially, starting from the first paragraph of the text array. The absolute and relative frequency, rank, and entropy has been calculated for each character of the extended Ukrainian alphabet within the selected segment. After reaching the end of the text array, the selection and subsequent calculations of consecutive segments began with the second paragraph of the text array, then with the third, and so on. The number of offsets from the beginning of the text has been regulated by the remainder of dividing the number of paragraphs by 100. Once the selection of consecutive segments has been completed, the average frequency and entropy of each character have been calculated.

Making similar calculations for randomly selected text segments.

Calculating the distribution of characters by type and the euphony of the text. These calculations have been made for the entire text array and separately for sequentially and randomly selected segments. Euphony has been defined as the percentage of the total number of vowels, sonorants, and voiced letters relative to the total number of characters.

Determining the average frequency rank of each character. Based on the results, the average rank has been calculated, representing the position each character occupied both in the entire text and in each segment.

Analysing the correspondence between the frequencies of characters in the whole text and selected segments: This analysis was based on the results of the calculations. Pearson's chi-squared test served as the theoretical basis (Oakes & Farrow, 2007). The frequency distribution of characters in the text under study has been treated as a hypothetical theoretical distribution function. For each selected segment, the χ^2_{exp} statistic has been calculated. The following statement has been accepted as the null hypothesis H_0: "In the sample text, the frequency distribution of the characters of the expanded Ukrainian alphabet does not differ from the corresponding distribution in the text array".

The following steps have been taken for testing:

The absolute frequency of the characters of the extended alphabet has been calculated for the sample text;

Then, these frequencies have been used to calculate the statistics of the χ^2_{exp} criterion;

$t_{cr} = \chi^2_{1\text{-}\alpha,k\text{-}1}$ has been determined using the appropriate table (National Institute of Standards and Technology, n.d.) at the significance level α=0,05 and degrees of freedom for k=36 (corresponding to the number of characters in the extended Ukrainian alphabet);

If $\chi^2_{exp} \geq t_{cr}$, the hypothesis has been rejected; otherwise it has been accepted. Simultaneously, the rank of each character in the frequency distribution has been determined for the sample text.

All calculations have been made using our own Python programs. The results have been compared to those obtained for Vasyl Stefanyk's stories.

Results and Discussion. The frequency of characters in M. Yatskiv's short stories, in comparison with the frequency of characters in V. Stefanyk's works, is illustrated in *Fig. 1*.

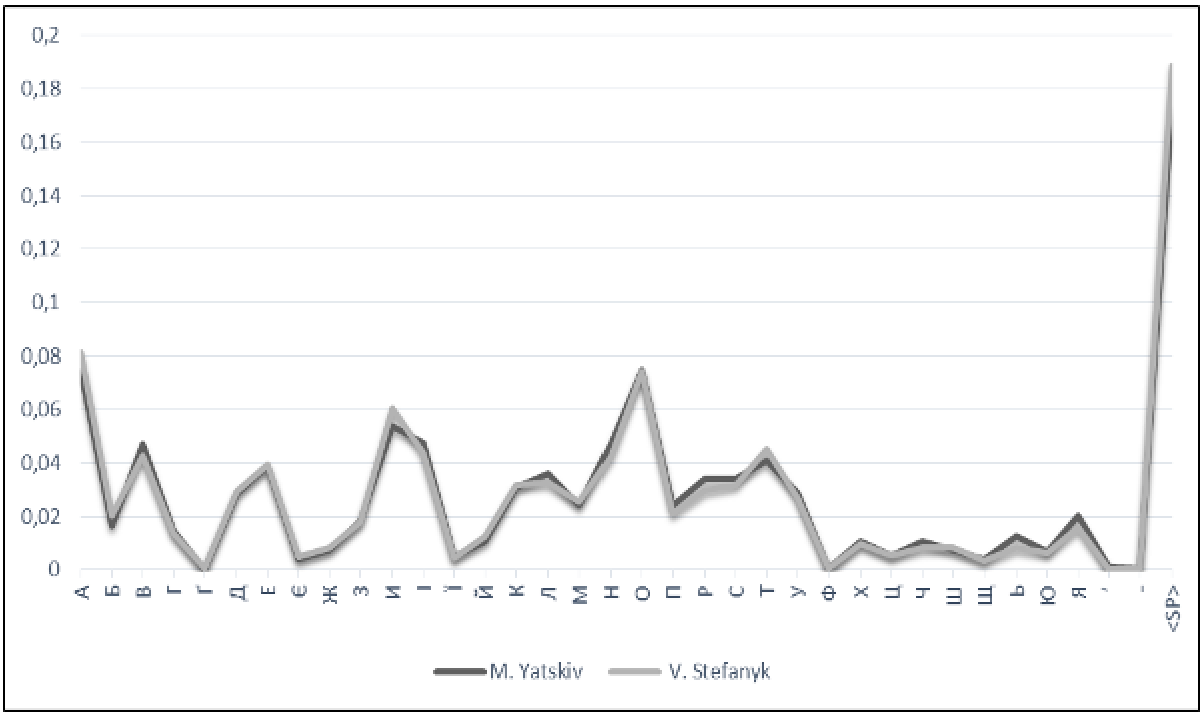

Figure 1. Frequency of characters in text arrays

As shown in *Figure 1*, the overall frequency of characters in the texts by both writers corresponds to the typical frequency patterns observed in fiction. This finding supports the thesis regarding the stability of character frequency within a given language, functional style, and genre. Excluding the space symbol, which consistently shows the highest frequency, the remaining characters can be categorized into high-frequency (О, И, А, І,), medium-frequency (Е, Р, В, С, Т, Н, К, М, Д, У, Л, П, Я, З), and low-frequency (Ь, Г, Б, Х, Ч, Ц, Й, Ю, Ж, Ї, Є, Ф, Ш, Щ, - , ‘, Ґ — 0%) groups.

Subsequently, consecutive and random samples from the text array of M. Yatskiv’s short stories have been processed and compared with the frequency of characters in the entire text (8700 samples have been analysed). The results are presented in *Table 1.*

Table 1. Comparison of the frequency of characters in the whole text and samples

Character	**Frequency**		
	Entire text	**Samples (mean)**	
		Consecutive	**Random**
А	0.07453	0.07469	0.07451
Б	0.01636	0.01633	0.01640
В	0.04742	0.04771	0.04748
Г	0.01470	0.01474	0.01468
Ґ	0.00015	0.00016	0.00015
Д	0.02857	0.02863	0.02857
Е	0.03895	0.03886	0.03892
Є	0.00405	0.00409	0.00403
Ж	0.00722	0.00716	0.00722
З	0.01833	0.01831	0.01836
И	0.05397	0.05398	0.05407
І	0.04748	0.04744	0.04749
Ї	0.00493	0.00489	0.00491
Й	0.01079	0.01083	0.01082
К	0.03085	0.03099	0.03088
Л	0.03613	0.03612	0.03609
М	0.02415	0.02404	0.02416

Character	**Frequency**		
	Entire text	**Samples (mean)**	
		Consecutive	**Random**
Н	0.04735	0.04714	0.04741
О	0.07520	0.07531	0.07518
П	0.02429	0.02432	0.02424
Р	0.03447	0.03435	0.03447
С	0.03427	0.03426	0.03424
Т	0.04118	0.04115	0.04115
У	0.02891	0.02895	0.02891
Ф	0.00090	0.00088	0.00091
Х	0.01070	0.01069	0.01068
Ц	0.00531	0.00532	0.00531
Ч	0.01089	0.01084	0.01089
Ш	0.00763	0.00765	0.00762
Щ	0.00424	0.00423	0.00423
Ь	0.01255	0.01248	0.01252
Ю	0.00710	0.00696	0.00706
Я	0.02034	0.02025	0.02036
‘	0.00095	0.00095	0.00096
-	0.00091	0.00094	0.00091
<SP>	0.17423	0.17437	0.17423

The mean value has been calculated using the formula:

$$\bar{X} = \frac{\sum x_i n_i}{\sum n_i} \qquad (2)$$

where x_i is the variant, n_i is the number of occurrences of the variant in the experiments, and *i* is the variant number.

The standard error has been calculated using the formula:

$$\sigma = \sqrt{\frac{\sum (x_i - \bar{x})^2 n_i}{\sum n_i}} \qquad (3)$$

where x_i is the variant, n_i is the number of occurrences of the variant in the experiments, *i* is the variant number, and $\bar{x}$ is the mean value.

As indicated by the results, there are no significant differences in the ratios between sequentially and randomly selected segments, nor between these segments and the entire set of works.

Thus, the set of characters consists of letters of the Ukrainian alphabet, spaces, apostrophes, and hyphens. This set has been further divided into three subsets: vowels, consonants, and special characters. Special characters include spaces, apostrophes, and hyphens. The frequency of each character has been calculated separately for each subset within the sequentially and randomly selected segments, as well as for the entire set of works. The results are presented in *Table 2*.

Table 2. Distribution of characters by type

Character type	**Entire text**	**Samples (mean)**	
		Consecutive	**Random**
Vowels	78768	758	758
Consonants	111510	1073	1073
Special characters	56610	544	545

As shown in *Table 2*, there are no significant differences in the ratios within each subset when comparing sequentially and randomly selected segments to the entire set of works.

To analyse the euphony of M. Yatskiv's texts, the set of Ukrainian language phonemes has been divided into vowels and consonants. The consonants were further categorized into sonorants, voiced obstruents, and voiceless obstruents, without distinguishing between palatalized and non-palatalized (hard vs. soft) consonants (see *Fig. 2*).

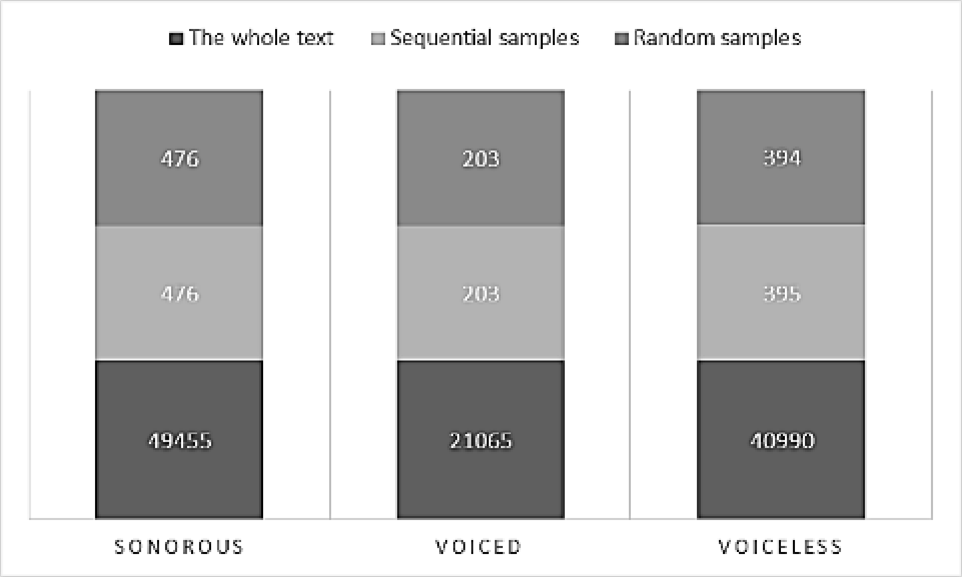

Figure 2. Distribution of consonants by type

The degree of euphony has been measured as the percentage of vowels, sonorants, and voiced obstruents taken from the text. *Table 3* presents a comparison of the statistical characteristics of euphony between the texts of M. Yatskiv and V. Stefanyk.

Table 3. Melodiousness of the Texts

Sample Type	**M. Yatskiv**	**V. Stefanyk**
Entire text	0.78	0.79
Sequential samples	0.78	0.79
Random samples	0.78	0.79

With minor differences, these characteristics are nearly identical and align closely with the average for works of fiction.

Based on the absolute (number) and relative frequency of each character of the expanded Ukrainian alphabet, calculated across sequentially and randomly selected segments as well as the entire text array, entropy has been calculated using the appropriate formula (Loukas & Chung, 2022). The results are presented in *Table 4*.

Table 4. Entropy of Characters in the Text and Samples

Coefficient	**Entire text**	**Samples**		
		General	**Consecutive**	**Random**
Number of experiments	1	17400	8700	8700
Maximum value	4.467	4.54388	4.54388	4.524274
Minimum value	4.467	4.35364	4.353641	4.374502
Average value	4.467	4.45116	4.446842	4.45547
Standard error	0	0.02694	0.03203	0.019718
Spectrum of fluctuation of average frequency	0	0.0002	0.000343	0.000211
Standard deviation error	0	0.0002	0.000343	0.000211
Relative error	0	0.00009	0.000151	0.000093

Based on the results of the calculations, the correspondence between the frequencies of symbols in the entire text and the selected segments has been analysed using Pearson's chi-squared test (χ^2). The data are presented in *Table 5*.

Table 5. Frequency of Characters According to Pearson's Chi-Squared Test

Text	**Number of samples**	**Type of samples**	**Number of matches**	**Percentage of matches**
M. Yatskiv, short stories	8700	Sequential	2508	28,83%
	8700	Random	7855	90,29%

Simultaneously, the rank of each character in the frequency distribution has been determined for the sample text, as shown in *Table 6*.

Table 6. Rank of Characters in the Research corpus

Character	**Rank averaged**	**Rank calculated**
А	3	3
Б	20	20
В	5	6
Г	21	21
Ґ	36	36
Д	15	15
Е	9	9
Є	32	32
Ж	28	27
З	19	19
И	4	4
І	6	5
Ї	32	30
Й	24	24
К	13	13
Л	11	10
М	17	17

Character	**Rank averaged**	**Rank calculated**
Н	6	7
О	2	2
П	17	16
Р	11	11
С	11	12
Т	8	8
У	14	14
Ф	34	35
Х	23	25
Ц	29	29
Ч	24	23
Ш	26	26
Щ	31	31
Ь	22	22
Ю	27	28
Я	18	18
’	33	33
-	36	34
<SP>	1	1

Analyzing the results, it can be concluded that by randomly selecting a segment as small as one hundredth of the text volume, it is possible to obtain, with high probability, a frequency distribution of characters that closely matches that of the entire text.

Conclusions and perspectives. Quantitative analysis of texts at different linguistic levels, together with automatic corpus processing and statistical calculations, allows for the identification of important characteristics that can clarify aspects of a writer’s idiolect and help to draw conclusions about the aesthetic significance of the texts in the research corpus.

The statistical characteristics of M. Yatskiv’s and V. Stefanyk’s short stories closely align to the literary style and are almost indistinguishable from each other. This finding has been proved through linguistic and statistical methods. We hypothesize that for genres with stable genre-forming elements (such as short stories, dramas, poems, etc.) within specific periods of a particular national literature, linguistic statistical indicators will show minimal variation. Proving this hypothesis will require expanding the scope of the study and incorporating a broader range of methods for analysing linguistic material.

REFERENCES

1. Karasov, V., & Levchenko, O. (2022). *Statistical characteristics of O. Zabuzhko's idiolect*. In *2022 IEEE 17th International Conference on Computer Sciences and Information Technologies (CSIT)* (pp. 138–141). IEEE. https://doi.org/10.1109/CSIT56902.2022.10000546
2. Kryvuliak, O. V. (2007). *Transformatsiia estetyky symvolizmu v novelakh M. Yatskova, O. Pliushcha, I. Lypy* [Transformation of the aesthetics of symbolism in the novellas of M. Yatskov, O. Pliushch, and I. Lypy] (Author's abstract of PhD dissertation). Taras Shevchenko National University of Kyiv. [in Ukrainian]
3. Kulchytskyi, I. (2019a). Okremi aspekty kvantytatyvnykh doslidzhen ukrayinskoyi movy [Certain aspects of quantitative studies of the Ukrainian language]. *Ukraina Moderna, 27*, 73–96. https://doi.org/10.3138/ukrainamoderna.27.073 [in Ukrainian]
4. Kulchytskyi, I. (2019b). *Statistical analysis of the short stories by Roman Ivanychuk*. In *Proceedings of the 3rd International Conference on Computational Linguistics and Intelligent Systems (COLINS-2019)* (Vol. 1, pp. 312–321).
5. Lototska, N., & Saban, O. (2023). *R. Ivanychuk's idiolect: Quantitative parameterization of the language used in the text*. In *2023 IEEE 18th International Conference on Computer Science and Information Technologies (CSIT)* (pp. 1–4). IEEE. https://doi.org/10.1109/CSIT61576.2023.10324093
6. Loukas, O., & Chung, H. R. (2022). *Entropy-based characterization of modeling constraints*. arXiv. https://doi.org/10.48550/arXiv.2206.14105
7. Melnyk, O. O. (2011). *Modernistskyi fenomen Mykhaila Yatskova: Kanon ta interpretatsiia* [The modernist phenomenon of Mykhailo Yatskov: Canon and interpretation]. Naukova dumka. [in Ukrainian]
8. Naumenko, A. M. (2003). Blukanina suchasnoho perekladu: vid hlukhoho kuta semiotyky do hlukhoho kuta kohnityvnoi linhvistyky [The wandering of modern translation: From the dead end of semiotics to the dead end of cognitive linguistics]. *Nova filolohiia, 3(18)*, 203. [in Ukrainian]
9. National Institute of Standards and Technology. (n.d.). *Critical values of the chi-square distribution*. Retrieved September 28, 2025, from https://www.itl.nist.gov/div898/handbook/eda/section3/eda3674.htm
10. Oakes, M., & Farrow, M. (2007). Use of the chi-squared test to examine vocabulary differences in English-language corpora representing seven different countries. *Literary and Linguistic Computing, 22*(1), 85–99. https://doi.org/10.1093/llc/fql044
11. Pavlyshenko, O. (2004). Markery avtors'koho idiolekta v leksyko-semantychnykh poliakh diiesliv anhlomovnoi khudozhnoi prozy [Markers of the author's idiolect in the lexico-semantic fields of verbs in English-language fiction]. *Mova i kul'tura, 7*(4, Pt. 2), 314–315. [in Ukrainian]
12. Selivanova, O. O. (2008). *Suchasna linhvistyka: napriamy ta problemy* [Modern linguistics: Directions and problems]. Dovkillia. [in Ukrainian]
13. Seminck, O., Gambette, P., Legallois, D., & Poibeau, T. (2022). The evolution of the idiolect over the lifetime: A quantitative and qualitative study of French 19th century literature. *Journal of Cultural Analytics, 7*(3), 1–24. https://doi.org/10.22148/001c.37588

14. Stelmakh, B. (2004). Indyvidual′nyi styl′ yak ob″iekt linhvostylistychnykh doslidzhen′ [Individual style as an object of linguistic-stylistic research]. *Visnyk Umans'koho Peduniversytetu. Seriia: Filolohiia (Movoznavstvo)*, 228–233. [in Ukrainian]

15. Tkachuk, O. (2013). *Naratyvni pryntsypy prozy Mykhaila Yatskova* [Narrative principles of Mykhailo Yatskov's prose]. Medobory. [in Ukrainian]

16. Yatskiv, M. (2016). *Chorni kryla*. Piramida. [in Ukrainian]

Дата надходження статті до редакції: 28.09.2025
Прийнято до друку: 27.10.2025