# Quantitative characteristics of the author's idiostyle

Tetiana Shestakevych[1,†], Yuliia Shyika[1, *,†] and Larysa Tsiokh[1,†]

[1] *Lviv Polytechnic National University, 28a,Stepana Bandery St., Building 5, Room 407, Lviv, 79013, Ukraine*

**Abstract**

The field of corpus linguistics and the significance of text corpora in linguistic research has been explored in the article. The article examines the various classifications of linguistic corpora based on factors such as linguistic data type, parallelism, literature, and purpose of creation. Furthermore, it highlights the parameters and criteria for creating high-quality linguistic corpora, including sufficiency, consistency, reproducibility, correctness, and technologic ability. The article presents a case study on the corpus of M. Yatskiv's works, discussing its typological and applicative characteristics. Finally, it provides quantitative characteristics and linguistic statistical analysis of the research corpus, offering insights into vocabulary volume, word forms, vocabulary richness, and word repetition. Overall, the value of text corpora in linguistic research has been highlighted and practical examples for analysis of the author's idiostyle has been provided.

**Keywords**

Corpus linguistics, text corpora, software tools, statistical analysis, idiostyle

## 1. Introduction

Linguistics is one of the first humanitarian sciences that uses mathematical modelling and computer-informational approaches to analyse data, draw insightful conclusions, and conduct research. To objectify and optimize linguistic research, creating text corpora is an effective method that can also provide new perspectives on traditional concepts.

Initially, any collection of texts used for linguistic studies was considered a corpus. However, with the introduction of the first electronic text corpus in the 1960s, a more precise definition appeared. A language corpus is now defined as "a collection of oral and written language data in electronic form." As corpus linguistics has progressed, the definition of a corpus has been refined. In the 1980s, it was argued that a text corpus should include several mandatory features to be considered a linguistic object. First of all, these are electronic form, standardisation, coding, representativeness, balance, research orientation and others [1].

*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.

Tetiana.v.shestakevych@lpnu.ua (T. Shestakevych); yuliia.i.shyika@lpnu.ua (Y. Shyika); larysa.y.tsokh@lpnu.ua (L. Tsiokh)

0000-0002-4898-6927 (T. Shestakevych); 0000-0003-2474-0479 (Y. Shyika); 0000-0003-2695-4411 (L. Tsiokh);

The fundamental features of a text corpus are its machine readability, which requires electronic forms and specific data coding systems, and its representativeness. There are various definitions of a corpus that emphasize the importance of these two features, such as "a collection of machine-readable texts that fully represents the language and its diversity," "a large number of natural language texts in digital form used for linguistic research," where "natural" means everything that has been expressed in oral or written form"; "written and spoken texts which are in one way or another representative for a language and are presented as an electronic database". To these criteria, N. Dash and B. Chaudhuri [2] add the parameter of corpus applicability in linguistic research, defining a corpus as "a collection of linguistic data composed of either written texts or transcribed spoken texts, the main purpose of which is to test hypotheses about language".

Within the field of corpus linguistics, various theoretical definitions exist regarding the nature of a corpus. However, J. Sinclair provides a brief and functional definition of an electronic text corpus. According to J. Sinclair, a corpus refers to a collection of carefully selected and appropriately ordered text passages or fragments that serve as a representative sample of language [3].

In brief, a text corpus refers to an electronic compilation of written and spoken texts in any natural language that is methodically structured to meet certain mandatory requirements and intended to facilitate scientific research on language. The nature and scope of text corpora may vary significantly, depending on factors such as their intended use, structure, selection principles, volume, and presentation format, among others. As such, text corpora are distinguishable from databases and other similar resources, and may differ widely from one another.

## 2. Related works

In contemporary linguistics, corpus-based studies have become a significant area of research, with numerous monographs, scientific articles, and textbooks on corpus technologies in foreign and domestic linguistics. Corpus-based studies hold an important place in world linguistics. Leading figures in corpus linguistics, such as G. Leech, D. Biber, J. M. Sinclair, S. Th. Gries, S. Granger, T. McEnery, P. Baker, P. W. Hanks, among others, have made groundbreaking contributions to the development and application of corpus-based methodologies. In the Ukrainian context, several researchers have significantly contributed to corpus-based studies, enriching the field with insights specific to the Ukrainian language and its usage. Notable Ukrainian scholars in corpus linguistics include S. Buk, N. Darchuk, O. Demska, V. Zhukovska, A. Zahnitko, I. Danyliuk, H. Sytar, V. Shyrokov, I. Kulchytskyi, and others.

The key topics of interest in this field can be broadly categorized into several areas, including an analytical review of discussions and foreign publications regarding the place of corpus technologies and corpus linguistics in modern linguistics, an overview of corpus linguistics and the history of its formation, a discussion of what a corpus of texts involves, its defining features, approaches to the classification of corpora as well as the branches and methods of their use. Additionally, the concept of the national text corpus, its prerequisites and principles of planning and compilation, the Ukrainian National

Linguistic Corpus with a volume of more than 100 million word uses, which was created in the Ukrainian National Linguistic Fund of the National Academy of Sciences, the basic principles and perspectives of the research corpus of the Ukrainian language, some aspects of the creation and use of specific research corpora, and the technical aspects of preparing texts for further corpus research has been reviewed.

Various above-mentioned scholars have proposed classification of linguistic corpora based on several factors. These include the type of linguistic data (written, oral, or mixed), parallelism (monolingual, bilingual, or multilingual), literature (literary, dialectal, colloquial, terminological, or mixed), the purpose of creation (multipurpose and specialized), genre (fictional, folklore, dramatic, or journalistic), availability (free, commercial, or closed), purpose (research and illustrative), dynamism (dynamic and static), tagging (tagged and not tagged), type of tagging (morphological, semantic, syntactic, or others), and volume of text (full-text or fragmented). However, some researchers suggest simplifying the classification system and recognizing the following categories: specialized, reference, multilingual, parallel, educational, diachronic, and mentoring [4]. It is important to note that these classifications aid in organizing and categorizing linguistic corpora for various purposes, such as research and analysis.

O. Demska-Kulchytska argues that corpora could be identified as:

- Full-text (full texts are included in the corpus)
- Fragmentary (only text fragments are included)
- Exploratory (used in linguistic research to formulate new theories and concepts)
- Illustrative (used to confirm existing theories or hypotheses about language)
- Monitoring/dynamic (provide the possibility of observing changes in the language, taking into account the aspect of diachrony)
- Statistical (show the state of the language at a particular time period)
- Diachronic (represent the language in several time lapses)
- Synchronic (represent the language or text of a certain defined period of time)
- General (represent the national language)
- Specialized (aimed at solving specific research tasks) [5].

Within the field of corpus linguistics, there exists a distinct category of multilingual linguistic corpora, parallel corpora, and comparative linguistic corpora, which hold significant value for scholars engaged in translation studies. These resources allow the effective analysis and comparison of language across diverse contexts and languages, enabling researchers to gain a deeper understanding of linguistics in practice [6, 7].

The representativeness of a corpus is a significant feature, “which means the ability of the corpus to reflect all the properties of the subject field, by which we understand the linguistic system's implementation level, which comprises linguistic phenomena subject to description; authenticity involves the selection of written or spoken text(s), excerpt(s) of text(s) created by the native speaker(s) in the process of real communication. This criterion is an essential component of empiricization, ensuring the material's authenticity. Additionally, selectivity is necessary to limit the material by selecting specific speech

fragments, while balance involves introducing a proportional number of textual resources into the corpus.

Linguistic corpora are characterized by four main parameters. Firstly, the corpus size should be significant enough to be representative of the subject field. Secondly, it should be structured and tagged for efficient use. Thirdly, the texts included in the corpus must be digitized for ease of access and analysis. Fourthly, the concept of "electronic corpus" includes special software for working with this corpus. These parameters are essential to ensure the corpus' quality and effectiveness in linguistic analysis [8].

V. Shyrokov outlines the selection criteria applied during the creation of the Ukrainian National Corpus. These included the diachronic aspect, which determined the selection of texts across time periods, as well as the stylistic aspect, which aimed to represent the substyles of the national language. Additionally, the territorial aspect was considered, taking into account the specificity of the literary language in different regions of Ukraine and the fact that the Ukrainian language can be used to create literary oral or written texts outside of Ukraine. Finally, the quantitative aspect was also taken into account, clearly defining the number of words in each text or passage included in the corpus, as well as the number of texts and/or passages [9].

## 3. Methods and materials

In corpus linguistics, users interact with the corpus through specialized software tools or corpus managers, which offer diverse means to extract the necessary information from the corpus. These tools enable users to conduct various types of searches, such as searching for specific word forms, discontinuous or continuous syntagms, or word forms based on morphological features. Additionally, users can access information on the origin or type of text, as well as obtain lexical and grammatical statistical data. Users may also save selected concordance lines in a separate file on their computer.

However, the corpus alone is insufficient for accomplishing many of the tasks aforementioned. It is also necessary for the text to contain diverse linguistic information. This led to the development of a tagged corpus, which facilitates the acquisition of more interesting results at the statistical level. Tagging makes it possible to count not only the frequency of words, but also the frequency of different parts of speech [10, 11].

The task of corpus annotation centers around the markup format. A linguistic corpus that possesses at least one linguistic parameter markup is distinguishable from other linguistic information and instrumental systems or databases. As such, specific requirements are placed upon the technique and technology of tagging. Ideally, the marking of corpora should occur in a unified and coordinated manner with previously established systems of tagging electronic arrays of information, allowing for a linguistically meaningful interpretation of introduced markers [12].

Structural annotation involves selecting structural elements of the text using a particular markup language and set of markers that indicate the external elements of text structure. To implement structural annotation, several procedures must occur:

1. Text segmentation
2. Formalization of annotation parameters of target units
3. Creation of a tagset or set of formal codes
4. Determination of the annotation scheme and its principles

Linguistic experts have identified several key criteria for standard corpora, including sufficiency, consistency, reproducibility, correctness, possibility of data collection, technologic ability, scalability, compactness, and clarity. Sufficiency refers to the need for a wide range of structural elements that can meet most requirements. Consistency is crucial, as the markup scheme must be based on consistent rules that enable precise identification of tags and attributes. Reproducibility is also essential, with the coding scheme based on clearly defined rules that enable the original text to be reproduced using simple algorithms. Correctness is maintained through software that checks the conformity of markups with structural specifications. Data collection is also an important criterion, encompassing direct data collection through manual input or automatic text recognition, as well as data coding. Technologic effectiveness is necessary to meet the needs associated with automatic processing of texts, including the selection of text according to established criteria, use of particular mechanisms, and type of intertext indexes. The possibility of ranging is also critical, ensuring that any created scheme has the ability to expand. Compactness is also a key consideration, with markup potentially affecting the file size and the speed of text data processing. Methods of achieving compactness include tag minimization, for example, omitting or shortening the final tag, use of specific end tags or XML markup schemes. Finally, clarity is crucial when direct user work with the text is required without special software support, with transparent markup essential to facilitate this process [13-15].

Therefore, the following marking system has been used:  – the beginning of a paragraph;  – the end of the paragraph; <s> – the beginning of a sentence; </s> – the end of the sentence; <head>…</head> – heading.

In analyzing an idiostyle, it is important to consider the structural elements of the text, such as the title, paragraph, and sentence. The annotated corpus of M. Yatskiv's prose offers a valuable tool for study the author's idiostyle, allowing for both qualitative and quantitative analysis of his language. This can provide invaluable insights into the characteristics of his works, making it a valuable source for those seeking to study and understand the nuances of his writing.

In terms of typological and applicative characteristics, the corpus of M. Yatskiv's prose can be classified as: – illustrative: it has been compiled for the purpose of linguistic-statistical analysis of the writer's idiolect; – full-text: contains the complete text of the story "In a Clutch (Shadow Dance)" as well as 42 short stories; – static: does not allow for the ongoing addition of texts; – author’s language: only texts by M. Yatskiv; – monolingual: includes texts only in Ukrainian; – written: the corpus is a collection of written texts; – annotated: textual data are tagged at the syntactic level.

The following software has been used to establish the quantitative and qualitative characteristics of the corpus:

- Textanz, particularly its Wordforms option, which enables the determination of not only the frequency of each word form and its length but also the variance (deviation of the values of a random variable from the center of distribution). Additionally, Summary option allows for determining the text's word count, number of sentences, average sentence length, average paragraph length, average word length, number of unique word forms, lexical diversity, lexical density, longest/shortest sentence, longest/shortest word form, and coefficient of readability
- AntConc toolkit and its Words List option, which counts all the words in the corpus and presents them in an ordered list
- Programs coded in Python by authors to work with the corpus, for example, to convert a list of word forms with structural marks into a list of word forms in txt format and Excel tables, for preprocessing text arrays before corpus analysis, to calculate the distribution of lengths of different linguistic units (words in letters, sentences in words, etc.), and to compute statistics on the distribution of word forms and words of the text by parts of speech, among other tasks

The formal model of the process of information technologies application in this research is presented as a Petri net (Fig. 1).

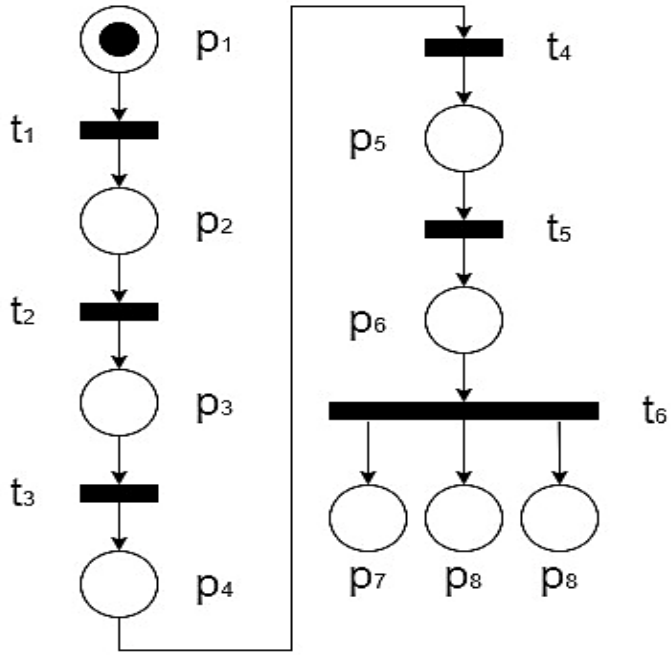


**Figure 1**: The formal model of the process of information technologies application

Petri net is a mathematical abstraction, widely used for processes modelling, as it is convenient in visualization of simultaneous and sequential tasks within a modelled process [16, 17]. Petri net *N=(I, O, P, T, W),* as a position has results of a processes, presented with transitions. Every transition, defined in Table 1, means an application of a

relevant information technology, for example, Optical Character Recognition application, Excel, Python software, text analysis software AntConc, etc. Petri net`s marking *M=(1, 0, 0, 0, 0, 0, 0, 0, 0, 0)*. To reach the goal of the research goal, the transitions should fire consequentially, *t1→ t2→ t3→ t4→ t5→ t6→ t7→ t8→ t9* (see Table 1 and 2).

**Table 1**
The transitions of Petri net

| Transition | Function |
|---|---|
| $t_1$ | Apply Optical Character Recognition |
| $t_2$ | Check the correctness of recognition |
| $t_3$ | Mark up the text and check the correctness of the markup |
| $t_4$ | Analyze the text |
| $t_5$ | Lemmatize the text |
| $t_6$ | Statistically analyze the text |

The purpose of creating a corpus of short stories based on the language of M. Yatskiv is to offer empirical data for scientific research on the author's idiostyle. The transition from traditional research methods to corpus-based ones is an evolutionary step that necessitates the existence of electronic textual data. This data enables the creation of a foundation for producing a dictionary of M. Yatskiv's language.

**Table 2**
The positions of Petri net

| Position | Explanation |
|---|---|
| $p_1$ | Text in PDF format |
| $p_2$ | Recognized text |
| $p_3$ | Correctly recognized text |
| $p_4$ | Markuped text |
| $p_5$ | Text analysis results |
| $p_6$ | Lemmas of the text |
| $p_7$ | Division by parts of speech |
| $p_8$ | Quantitative indicators of the lexical level |
| $p_9$ | Representativeness of the sample asessment |

## 4. Experiment

By meeting all the requirements for creating a linguistic corpus, we have obtained a convenient tool that can be used for work of any complexity, especially for solving the specific objectives of our research.

The story "In a Clutch" and 42 short stories by M. Yatskiv has been selected for the study. General quantitative characteristics of the corpus are presented in Table 3.

**Table 3**
Quantitative characteristics of the research corpus of M. Yatskiv's works

| No | Title | Quantity | | | |
|---|---|---|---|---|---|
| | | Symbols | Word uses | Word forms | Words |
| A | "In a Clutch" | 388201 | 72599 | 17658 | 10058 |
| C | Novels | 219590 | 42963 | 10535 | 6202 |

Texts for the research corpus were first digitalized, then normalized and tagged. The AntConc program was then used to generate a list of word forms from the story "In a Clutch." This list was subsequently transferred to MS Excel for lemmatization, reducing word forms to their dictionary form. The total number of unique words in the story was then calculated by sorting word forms using MS Excel's "Sorting and filtering" function based on the "Part of speech + Lemma" criterion. The "Interim results" function in MS Excel was then used to generate a list of subcorpus words and their frequency of use. The same function was also employed to calculate the number of word uses, word forms, and words by parts of speech in the subcorpus of M. Yatskiv's works. Thus, as a result of the analysis of both subcorpora, we obtained the following results presented in Table 4.

**Table 4**
General characteristics of the research corpus

| Index | "In a Clutch" | Short stories |
|---|---|---|
| Length of preliminary paragraphs | 108 | 108 |
| Number of preliminary paragraphs | 2286 | 2649 |
| Control number of symbols | 388201 | 219590 |
| Word uses | 72599 | 42963 |
| Word forms | 17568 | 10535 |
| Words | 10058 | 6202 |

A linguistic statistical analysis of the research corpus of M. Yatskiv's works has been conducted, following the established methodology developed by S. N. Buk and Kulchytskyi&Tsiokh and other[18, 19, 20, 21, 22]. The main characteristics of the text were identified by the researcher, and are as follows:

The volume of the text, denoted as the total number of words used (N), equals 72599 in the corpus under research.

The number of word forms in the text (Vf), or the unique words used, equals 17658.

The vocabulary volume (V), referring to the number of words used in the text, equals 10058 in our research corpus.

The vocabulary richness or diversity index (Id) is the ratio of the vocabulary volume (V) to the overall text volume (N), and is calculated using the formula:

$$\mathrm{Id} = \frac{V}{N} = \frac{10058}{10058} = 0.14. \quad (1)$$

A higher index indicates a greater variety of words used. In this case, the index of 0.14 is considered high, as the average index of fiction, as calculated by S.N. Buk [23], is 0.067.

The average repetition of a word in the text (Iwr) is inverted from the diversity index and calculated as:

$$\text{Iwr} = \frac{N}{V} = \frac{72599}{10058} = 7.22, \quad (2)$$

where N is an overall text volume, V is the vocabulary volume.

On average, each word is used approximately seven times in the text.

Hapax legomena (V1) pertains to words that appear only once in a given sample, with a frequency of 1. Our corpus of research contains a total of 5495 such words.

The exclusivity index, on the other hand, measures the variability of the vocabulary, specifically the portion of the text that comprises words that appear only once. The exclusivity index for the vocabulary (Ien) is calculated as the ratio of the number of lexemes with a frequency of 1 (V1) to the volume of the text (N), resulting in:

$$\text{Ien} = \frac{V1}{N} = \frac{5495}{72599} = 0.08. \quad (3)$$

According to the functional styles of the Ukrainian language [24], the exclusivity index for fiction is 0.029.

The vocabulary concentration index represents the portion of the text that consists of words that appear ten or more times. The vocabulary concentration index (Ivc) is determined as the ratio of the number of words in the text with an absolute frequency of 10 or more (V10) to the total number of words in the text, giving us:

$$\text{Ivc} = \frac{V10}{V} = \frac{949}{10058} = 0.09. \quad (4)$$

The average vocabulary concentration index for fiction is 0.14/

A low concentration index and a high number of words with a frequency of 1 (and, consequently, a high exclusivity index) are indicative of a significant diversity in the author's vocabulary.

The lexical density index, which is closely associated with the vocabulary concentration index, is a measure that expresses the ratio of content words (Ncw) in the text to the total number of words (N). Texts that have fewer function words tend to be more lexically dense. It is possible to calculate coefficients of lexical density for content words and separately for nouns, adjectives, verbs, and adverbs:

$$\text{Iden} = \frac{\text{Ncw}}{N} = \frac{17143}{42963} = 0.4. \quad (5)$$

The Automated Readability Index (ARI) was first developed in 1967 with the intention of evaluating the readability of technical manuals and various documents. Over time, its application has expanded to other areas. Unlike other well-known readability indices, such as the Flesch-Kincaid, Gunning Fog Index, SMOG Index, and the Fry Readability Formula, the ARI possesses a unique advantage, along with the Coleman-Liau index, in that it does not rely on a specific natural language of the printed text. This is due to the fact that it does

not take into account syllables, but rather the ratio of signs in a word and the number of sentences. The ARI formula is:

$$\mathrm{ARI} = 4.71 \times \frac{C}{W} + 0.5 \times \frac{W}{S} - 21.43 = 4.71 \times \frac{219590}{42963} + 0.5 \times \frac{42963}{3954} - 21.43 \quad (6)$$
$$= 9.08,$$

where C represents the number of characters in the text, W represents the number of words in the text, and S represents the number of sentences in the text. It is important to note that the higher the ARI index, the more challenging it is to comprehend the text.

Additionally, the ratio of parts of speech in a given text can serve as a statistical parameter of an individual author's style and a characteristic feature of a particular work [25]. The research corpus has been subjected to morphological tagging, using the classical division of words into parts of speech, and the frequency of each part of speech in the text has been automatically obtained, as shown in Table 5.

**Table 5**
Frequency of parts of speech

| Part of speech | Word use | % | Word forms | % | Words | % |
|---|---|---|---|---|---|---|
| Interjection | 176 | 0.24 | 46 | 0.26 | 41 | 0.41 |
| Verb | 13603 | 18.74 | 6112 | 34.61 | 2887 | 28.7 |
| Pronoun | 8186 | 11.28 | 450 | 2.55 | 179 | 1.78 |
| Noun | 20218 | 27.85 | 6518 | 36.91 | 4140 | 41.16 |
| Preposition | 8066 | 11.11 | 72 | 0.41 | 50 | 0.5 |
| Adjective | 6073 | 8.37 | 3327 | 18.84 | 1877 | 18.66 |
| Adverb | 4863 | 6.7 | 810 | 4.59 | 696 | 6.92 |
| Conjunction | 5355 | 7.38 | 48 | 0.27 | 37 | 0.37 |
| Particle | 4839 | 6.67 | 67 | 0.38 | 52 | 0.53 |
| Numeral | 1220 | 1.68 | 208 | 1.18 | 97 | 0.97 |
| Total | 72599 | 100 | 17658 | 100 | 10058 | 100 |

## 5. Results

The analysis of M. Yatskiv's works reveals that verbs and nouns are the most frequently used parts of speech, accounting for 18.74% and 21.48% respectively in both the novel and the writer's short stories. Function words, on the other hand, show the highest level of activity (25.39% and 34.14% respectively). Pronouns are also used in significant amounts, amounting to 11.28% and 11.5%. Adjectives and adverbs are almost equal in M. Yatskiv's short stories, with 8.44% and 8% respectively, while adjectives are more prevalent in the novel at 11.11% and 6.70%. Numerals, on the other hand, are the least used, having only 1.68% and 0.75% in both texts. Numerals are the least numerous in both texts (1.68% and 0.75%).

In linguistic statistics, it is common to calculate the quantitative relations between parts of speech, considering them as one of the components of the statistical

characteristics of the text. These relations include the index of nominal modifiers (Inat), which measures the ratio of the sum of noun uses (Vn) to the sum of adjective uses (Vadj), and the index of verbal modifiers (Ivat), which measures the ratio of the sum of adverb uses (Vadv) to the sum of verbs uses (Vv) [26]. Additionally, the degree of nominality (Inom) is also considered, measuring the ratio of the sum of noun uses (Vn) to the sum of verb uses (Vv).

The indices of epithetization, nominalization, and verbal modifiers serve as an important supplement to the qualitative analysis although they are not the defining characteristics of the stylistic interpretation of the text. The quantitative relations of parts of speech in M. Yatskiv's novel "In a Clutch" demonstrates the author's frequent use of nominal and verbal modifiers, as well as a high degree of nominality when compared to the average figures of fiction. As a result, several coefficients have been calculated to quantitatively characterize the lexical level of M. Yatskiv's works of fiction in various ways (see Table 6).

**Table 6**
Quantitative characteristics of the lexical level of M. Yatskiv works

| Coefficient | In a Clutch | Short stories |
|---|---|---|
| Richness of the vocabulary | 0.14 | 0.61 |
| Average word repetition in text | 7.22 | 1.65 |
| Text exclusiveness | 0.08 | 0.24 |
| Vocabulary concentration | 0.09 | 0.49 |
| Lexical density | 0.75 | 0.4 |
| Automated Readability Index | 10.27 | 8.08 |
| Nominal modifiers index | 0.3 | 0.51 |
| Verbal modifier index | 0.36 | 0.41 |
| Nominality degree | 1.49 | 0.73 |

To determine the significance or insignificance of the statistical difference between the coefficients of the author's long prose and short stories, $\chi^2$ has been calculated, which is also known as the homogeneity criterion in quantitative linguistics. To calculate the criterion of homogeneity, it is necessary to have a certain number of indicators for each sample. This involves constructing a table with a number of rows equal to the number of samples and a number of columns equal to the number of indicators to be compared. Based on the results of our research, the resulting table 7 is as follows:

**Table 7**
Homogeneity criterion calculation

| | k1 | k2 | k3 | k4 | k5 | k6 | k7 | k8 | k9 | ∑T |
|---|---|---|---|---|---|---|---|---|---|---|
| T1 | 0.14 | 7.22 | 0.08 | 0.09 | 0.75 | 10.27 | 0.3 | 0.36 | 1.49 | 20.7 |
| T2 | 0.61 | 1.65 | 0.24 | 0.49 | 0.4 | 8.08 | 0.51 | 0.41 | 0.73 | 13.12 |
| ∑k | 0.75 | 8.87 | 0.32 | 0.58 | 1.15 | 18.35 | 0.81 | 0.77 | 2.22 | 33.82 |
| T1 | 0.14 | 7.22 | 0.08 | 0.09 | 0.75 | 10.27 | 0.3 | 0.36 | 1.49 | 20.7 |

The $\chi^2$ calculation methodology presented by V. Perbyinis [27] has been adopted. This involves the application of a specific formula to analyze the data.

$$x^2 = s \times \sum \frac{(knTn)^2}{\Sigma kn \Sigma Tn} - 1.$$

Following the completion of the calculations for our table, the following results have been obtained:

$$\chi^2 = 0.45.$$

In order to determine whether $\chi^2$ indicates a significant difference, it is necessary to refer to the table of critical values of $\chi^2$. This involves assessing the number of degrees of freedom, which in this particular case is f = 8. If the calculated value of $\chi^2$ is greater than the table value for the given significance level, the difference is considered significant. In our case, 0.45 is significantly less than the smallest number in the series. This indicates that the difference in statistical indicators characterizing the lexical level of M. Yatskiv's short stories and novel is statistically insignificant and therefore allowable. It can be concluded that a common idiostyle is present, uniting works under research.

## 6. Conclusions

Linguistic corpora play a vital role in linguistic research, providing a systematic and structured collection of written and spoken texts in electronic form. These corpora allow for the application of mathematical modeling and computer-informational approaches to analyze language data and draw insightful conclusions. The definition and criteria for a corpus have evolved over time, emphasizing the importance of machine readability, representativeness, standardization, and other features. Various classifications of linguistic corpora have been proposed based on factors such as the type of linguistic data, parallelism, literature, purpose of creation, genre, availability, and more.

The creation of a linguistic corpus requires careful selection of texts based on diachronic, stylistic, territorial, and quantitative aspects. The corpus of M. Yatskiv's prose serves as a valuable resource for studying the author's idiostyle, offering both qualitative and quantitative analysis of language. The corpus is classified as illustrative, full-text, static, author's language, monolingual, written, and annotated.

Quantitative characteristics of the research corpus, such as word uses, word forms, and parts of speech, have been analyzed for the story "In a Clutch" and 42 short stories by M. Yatskiv. Digitalization, normalization, tagging, and analysis using software tools like AntConc and MS Excel have facilitated data processing and statistical analysis. The linguistic statistical analysis of the research corpus has provided insights into the volume of the text, the number of word forms, vocabulary volume, and vocabulary richness.

Overall, linguistic corpora and their analysis offer valuable resources and methodologies for studying language, enabling researchers to gain a deeper understanding of linguistic phenomena, language variation, and idiolects. These corpora provide a solid foundation for empirical research and facilitate the development of linguistic theories and concepts.

## References


[1] O. Demska, Tekstovyi korpus: Ideia inshoi formy. VPTs NaUKMA, Kyiv, 2011.

[2] N. Dash, B. Chaudhuri, Using Text Corpora for Understanding Polysemy in Bangla, in: Proceedings of the Language Engineering Conference, IEEE, Hyderabad, India, 13 December 2002, pp. 99-109. doi: 10.1109/LEC.2002.1182297.

[3] J. Sinclair, Developing Linguistic Corpora: a Guide to Good Practice, 2004. URL: https://users.ox.ac.uk/~martinw/dlc/chapter1.htm

[4] P. Baker, A. Hardie, T. McEnery, A Glossary of Corpus Linguistics. Edinburgh University Press, Edinburgh 2006. doi: 10.1515/9780748626908.

[5] O. Demska-Kulchytska, Deshcho pro klasyfikatsiiu tekstovykh korpusiv, Naukovi zapysky Ternopilskoho derzhavnoho pedahohichnoho universytetu im. V. Hnatiuka, 1 (2004) 153–57. URL: http://ekmair.ukma.edu.ua/handle/123456789/1704.

[6] M. Baker, Corpora in Translation Studies, Target. International Journal of Translation Studies 7 (1995) 223–43. doi: 10.1075/target.7.2.03bak.

[7] D. Barth, S. Stefan, Understanding Corpus Linguistics, Routledge, 2022.

[8] Y. Demianchuk, Riznovydy korpusu tekstiv u protsesi perekladu dokumentiv ofitsiino-dilovoho styliu. Naukovyi visnyk DDPU imeni I. Franka. Seriia «Filolohichni nauky»: Movoznavstvo 1 (2016) 104–07. URL: http://ddpu-filolvisnyk.com.ua/uploads/arkhiv-nomerov/2016/NV_2016_5-1/27.pdf.

[9] V. Shyrokov (Ed.), Korpusna linhvistyka. Dovira, Kyiv, 2005.

[10] I. Kulchytskyi, Tekhnolohichni aspekty ukladannia korpusiv tekstiv, in: O. Levchenko (Ed.), Dani tekstovykh korpusiv u linhvistychnykh doslidzhenniakh, Vydavnytstvo Lvivskoi politekhniky, Lviv, 2015, pp. 29–45.

[11] I. Kulchytskyi, Unormuvannia tekstu pid chas dokorpusnoho opratsiuvannia: Dosvid zastosuvannia, Visnyk Natsionalnoho universytetu «Lvivska politekhnika» 7 (2020) 51–58. doi: 10.23939/sisn2020.07.051.

[12] M.-L. Merten, M. Wever, M. Geierhos, D. Tophinke, Eyke Hüllermeier, Annotation uncertainty in the context of grammatical change, International Journal of Corpus Linguistics, 28:3 (2023) 430-459. doi: 10.1075/ijcl.20113.mer

[13] J. Mesch, Creating a multifaceted corpus of Swedish Sign Language, in: E. Wehrmeyer (Ed.), Advances in Sign Language Corpus Linguistics, John Benjamins, Amsterdam, 2023. doi: 10.1075/scl.108.09mes.

[14] S. Buk, Velyka proza Ivana Franka: elektronnyy korpus, chastotni slovnyky ta inshi mizhdystsyplinarni konteksty, Lviv Ivan Franko University, Lviv, 2021.

[15] I. Khomytska, V. Teslyuk, N. Kryvinska, I. Bazylevych, Software-Based Approach Towards Automated Authorship Acknowledgement – Chi-Square Test on One Consonant Group, in: Electronics, Vol. 7:1138, July 2020, pp. doi: 10.3390/electronics9071138. URL: https://www.mdpi.com/2079-9292/9/7/1138

[16] T. Shestakevych, O. Volkov, The criteria for choosing the optimal solution under the uncertainty in project management (2021) CEUR Workshop Proceedings, 2851, pp. 95 – 105.

[17] A. Gozhyj, I. Kalinina, S. Shiyan, V. Nechakhin, Building a Fuel Measurement System Model based on Colored Petri Nets (2023) International Scientific and Technical

Conference on Computer Sciences and Information Technologies. doi: 10.1109/CSIT61576.2023.10324266.
[18] I. Kulchytskyi, L. Tsiokh, M. Malaniuk, Quantitative Equivalence Level in Poetry Translation, in: Proceedings of the 2018 IEEE 13th International Scientific and Technical Conference on Computer Sciences and Information Technologies (CSIT), IEEE, 2018, pp. 51-54. doi: 10.1109/stc-csit.2018.8526715.
[19] M. Paquot, L. Plonsky, Quantitative research methods and study quality in learner corpus research, International Journal of Learner Corpus Research 1:3 (2017) 61-94. URL: http:// hdl.handle.net/2078.1/185993
[20] F. Seifart, R. Mundry, Quantitative Comparative Linguistics based on Tiny Corpora: *N*-gram Language Identification of Wordlists of Known and Unknown Languages from Amazonia and Beyond, Journal of Quantitative Linguistics 22 (2015) 202–214. doi: 10.1080/09296174.2015.1037161
[21] V. Karaban, A. Karaban, AI-translated poetry: Ivan Franko's poems in GPT-3.5-driven machine and human-produced translations, Forum for Linguistic Studies 6 (2024) doi: 10.59400/fls.v6i1.1994
[22] Y. Shi, L. Lei, Lexical Richness and Text Length: An Entropy-based Perspective, Journal of Quantitative Linguistics, 29:1 (2022) 62–79. doi: 10.1080/09296174.2020.1766346
[23] S. Buk, A. Rovenchak, Rank-Frequency Analysis for Functional Style Corpora of Ukrainian, Journal of Quantitative Linguistics 11:3 (2004) 161–171. doi: 10.1080/0929617042000314912
[24] S. Buk, Statystychni kharakterystyky leksyky osnovnykh funktsionalnykh styliv ukrainskoi movy: Sproba porivniannia, Leksykohrafichnyi biuleten 13 (2006), 166–70. URL: http://dspace.nbuv.gov.ua/handle/123456789/72846.
[25] A. Divjak, S. Sharoff, T. Erjavec, Slavic Corpus and Computational Linguistics, Journal of the Slavic Linguistics Society 25 (2017)171–199. doi: 10.1353/jsl.2017.0008.
[26] V. Vincze, The Relationship of Dependency Relations and Parts of Speech in Hungarian, Journal of Quantitative Linguistics 22:1 (2015) 44-54. doi: 10.1080/09296174.2014.974458
[27] V. Perebyinis, Statystychni metody dlia linhvistiv. Nova knyha, Vinnytsia, 2002.