Statistical Characteristics of Roman Ivanychuk's Idiolect (Based on Writer's Text Corpus)

Nataliia Lototska[1]

Lviv State University of Life Safety, Kleparivska str. 35, Lviv, 79013, Ukraine

Abstract

The paper presents the statistical study of Roman Ivanychuk's historical novels. It is pointed out that his idiolect hasn't been the subject of linguistic and statistical research yet. The analysis of author's text and its lexicon reflects the individuality of linguistic preference.

It is known that statistical methods make possible to identify quantitative markers of text and vocabulary, and, in turn, give them a qualitative interpretation. Text corpus serves as a useful tool for discovering many aspects of language use that might be otherwise left undetected.

For an integrated study of the writer's idiolect the corpora of Roman Ivanychuk's texts and Ukrainian literary prose were created based on GRAC. Its functionality enables to obtain the data on frequency of parts of speech and present various morphological statistical characteristics (index of epithetization, index of verb + adverbial complement, level of nominalization), statistical parameters of vocabulary and text (text size, size of vocabulary of lexemes, diversity index, average word frequency, hapax legomena, exclusivity index of text and vocabulary, concentration index of text and vocabulary), frequency zones of vocabulary.

The results of this study may be interpreted as individual manner of author's writing and applied for text identification and further research of individual language.

Keywords

Frequency, frequency rank, idiolect, statistical characteristics, parts of speech, text corpus.

1. Introduction

The figure of Roman Ivanychuk (1929–2016) is significant in Ukrainian literature of the second half of the 20th century and the beginning of the 21st century. Firstly, the writer is known for his historical novels and short stories. In addition, he wrote numerous novels, memoirs, interviews, and journalistic texts. Roman Ivanychuk's texts have constantly been subjects of interest to literary critics and linguists. However, the writer's idiolect hasn't been the subject of thorough linguistic and statistical analysis yet.

The research of the writer's style in linguistics is mainly carried out on his literary texts. The text is the basis for the idiolect study as well. The individuality of the author's language, his manner is reflected in the preference for certain lexical, morphological, syntactic, phonetic means in the text [41].

Representatives of text theory consider that an individual text is a system united by communicative integrity, logical, grammatical, and stylistic relations [67]. Specificities of the use of a particular unit in a certain text determine its functional properties: frequency, position, compatibility, which depends on text nature, functional or author's style and varies from text to text [47].

The author's lexicon reflects the idiolect most specifically. "The linguistic personality of the writer is revealed through the individually used word, his artistic and individual picture of the world is reflected in linguistic expression" [48, p. 11].

The writer's speech as a marker of linguistic personality makes it possible to follow his / her manner of choice and use of words, whereas the language picture of the author's world as a representative of a particular linguistic and cultural community is displayed in his literary texts [26].

COLINS-2022: 6th International Conference on Computational Linguistics and Intelligent Systems, May 12–13, 2022, Gliwice, Poland
EMAIL: nata07lototska@gmail.com (N. Lototska)
ORCID: 0000-0001-6692-196X (N. Lototska)

The author's text reveals the concept of linguistic personality through which the writer creates a certain fragment according to the world picture presented in his / her cognition [25].

2. Related works

The notion of idiolect is well-known in linguistics, although an exact definition and the very existence of the phenomenon is a subject of studies and debates. Norbert Dittmar offers a definition of idiolect as the language of the individual, which because of the acquired habits and the stylistic features of the personality differs from that of other individuals [14, p. 111].

Numerous researchers have analyzed the idiolect of certain authors synthesizing various disciplines: linguostylistics [26, 48, 63, 64], cognitive linguistics [30], quantitative linguistics [5, 6, 7, 13, 17, 19, 27, 32, 33, 41, 44, 51, 60, 61] and literature studies.

An idiolect is a set of language characteristics of a native speaker; a person's specific, unique way of speaking or writing. The past lack of interest in idiolects derives from the difficulty in obtaining appropriate data and, on a theoretical level, it arises in some cases from a general dismissal of usage as being uninteresting and in others from an understandable focus on the general rather than the particular [3]. However, the notion of idiolect remains an understudied topic, especially in quantitative linguistics, due to the insufficiency of relevant large corpora [3, 4, 39].

Corpus linguistic research offers strong support for the idea that language variation is systematic and can be described using empirical, quantitative methods [18]. Anatol Stefanowitsch defines a corpus as a large collection of authentic text (i.e., samples of language produced in genuine communicative situations) [58]. A corpus is "a collection of pieces of language text in electronic form" [56, p. 19], moreover text corpus presents useful statistical information such as number of word types, frequency, co-occurrences [4].

Writer's text corpus belongs to an independent corpus and can be a part of a general language corpus of a certain language or exists as a separate entity, and can provide a detailed study of a writer's lexicon and open prospects for further research. Text corpus possesses a huge potential in the study of writer's language both in qualitative and quantitative aspects, moreover some facts of writer's style can be revealed only with the help of a text corpus, such as vocabulary richness, indexes of variety and exclusiveness, etc. [5, 6, 7, 66]

Nowadays the texts of Taras Shevchenko [63], Hryhoriy Skovoroda [49], Yuri Shevelyov [11; 61], Mykhailo Kotsyubynsky [60], Bohdan Lepky [57], Ivan Franko [6, 7], Vasyl Stefanyk [19] and others are studied by means of corpus linguistics.

Ways of individual style expression in a natural can be automatically detected with the use of computational linguistics methods [29], as fiction works tend to be long and provide large quantities of data [53]. The use of corpus-based approach and the application of statistical method allow to solve the problems of author attribution that helps in identifying different types of texts and can be used in plagiarism detection, author's identification and resolving disputed authorship [12, 15, 21, 22, 40, 35, 36, 37, 50, 62].

Statistical studies of Ukrainian literary texts have been used to study their lexical and stylistic peculiarities. The prospects and relevance of the use of quantitative methods (and not only interpretive) in literary texts are traced in the following researches [1, 2, 3, 11, 17, 19, 27, 28, 29, 32, 33, 34, 42, 44, 45, 51, 53, 59, 68].

Ihor Kulchytskiy's study [28] is on the individual style of writing of Roman Ivanychuk researched by the means of statistics in order to find some distinctive features of the Ivanychuk's writing to reveal his special manner comparing to other Ukrainian authors. According to Solomiya Buk [5, 6] statistical parameters of Ivan Franko's dictionary such as ratio of hapax legomena and frequently used lemmas help to find out and to describe more precisely the important features of author's style.

Research of parts of speech in the writer's text and dictionary is considered an important stage to establish individual features. The peculiarities of the behavior of parts of speech are revealed in different writer's texts [8, 9, 46, 51, 52], in particular in Ukrainian fiction texts of Ivan Franko, Oleksandr Dovzhenko, Hryhir Tyutyunnyk, Yuriy Smolych, Myhailo Stelmakh, and others.

An analysis of the research studies of the writer's idiolect presents tendencies to its lexicographic parameterization, quantitative analysis and the application of computer technologies in particular on the basis of a text corpus [5, 6, 7, 11, 19, 27, 28, 32, 33, 34].

3. Methods and materials

3.1. Statistical approach in the idiolect study

The author's text reflects his / her world view and demonstrates the richness of the lexicon of his / her linguistic personality, while the statistical approach allows to identify quantitative markers of the idiolect and, in turn, gives them qualitative interpretation.

Quantitative analysis used in idiolect study allows to avoid methodological mistakes frequently caused by researcher's subjectivity [32]. Statistical approaches applied to different writers' texts may reveal characteristics which differ them one from the others and therefore present individual creative manner of a writer [45].

Yuriy Pavlov and Elizaveta Tikhomirova consider low-frequency vocabulary as an idiolect marker [44, p. 9]. Tatiana Demidova adds that the author's linguistic taste, his literary inclinations and the richness of his linguistic personality are reflected in low-frequency vocabulary [13]. Mihail Muhin offers the analysis of the writer's frequency vocabulary to identify the idiostyle features [41].

Statistics is an important tool for linguistic data analysis in modern linguistics. In addition, quantitative methods ensure reliability of results, allow to reveal language units and text structure properties, any research is impossible without statistical studies [32]. The fact that the language itself is a complex system subordinated to the laws of statistics proves the necessity of using statistical methods in linguistics [46].

Text quantitative characteristics allow to determine the qualitative characteristics of the writer's idiolect objectively [33]. It is generally acknowledged there is an internal interdependence between the qualitative and quantitative features of language structure, which determines the subordination of frequency of language units in speech to certain statistical patterns [31, p. 5].

Statistical methods are widely used in linguistics and have become one of the most efficient and time-saving tools of processing different sets of texts [27]. These methods allow to obtain accurate data of lexical units in context, to obtain data on frequency of occurrences, words, lemmas, grammatical categories. In addition, search results can be ranked by different parameters, and we are able to set threshold values thus making it possible to obtain meaningful information [20, p. 66].

Statistical studies provide the opportunity to compare the proportion of parts of speech in the writer's texts, reflect the quantitative characteristics of the writer's lexicon, which represents information about the stylistic features of the writer at the lexical level objectively, and, vice versa, identify the words that do not function in society during his creative activity [5, 6, 7].

Statistical analysis of the historical prose fiction of Roman Ivanychuk enables to demonstrate individual manner of author's writing. The topicality of the research lies in the lack of thorough idiolect research of Roman Ivanychuk's literary legacy, a need for an integrated study of his lexical system based on the text corpus and by means of modern research methods.

3.2. Writer's text corpus as a tool to reveal statistical characteristics of idiolect

The use of text corpus provides reliable criteria for determining the acceptability and the evaluation of certain linguistic phenomena use, allows to obtain accurate data on the lexical structure of language, and the relative frequency of some lexical items (words) use [58].

The creation of text corpus involves an integrated processing of writer's lexicon that represents an opportunity to carry out more advanced and perspective studies of his / her literary texts [5, 6, 7].

The subject of our research is the statistical characteristics of Roman Ivanychuk's historical novels. To accomplish similar study for Roman Ivanychuk the corpus of his prose fiction texts has been created. This corpus comprises 16 historical novels and 1 historical trilogy written throughout 1962-2016 (total corpus size is 1,295 million words): *At The Edge Of The Paven Way (Krai bytoho shliakhu), Mallows (Mal'vy), Red Wine (Cherlene vyno), Manuscript From Ruska Street (Manuskrypt z vulytsi Rus'koyi), Water From The Stone (Voda z kameniu), The Fourth Dimension (Chetvertyi vymir), Scars On The Rock (Shramy na skali), Crane's Cry (Zhuravlynyi kryk), Because War Is War (Bo viyna viynoyu), Horde (Orda), The Gospel Of Thomas (Yevanheliye vid Tomy), Pillars Of Fire (Vohnenni stovpy), Saxaul In The Sands (Saksaul u piskakh), Across The Pass (Cherez pereval), Pilgrimage (Khresna proshcha), Voices From Above The Waters Of Kinneret (Holosy z-nad vod Henisareta), I Have Not Written About Donbass Yet (Ya shche ne pysav pro Donbas).*

The texts of the novels were converted into an electronic form, the next step was the normalization of the texts in the MS Word editor [27]. "Text normalization process contains the following stages: normalization of coding, normalization of graphics, text proofreading, technical normalization of punctuation" [27, p. 58].

The next step was to upload these texts into GRAC [55] and create the Roman Ivanychuk's subcorpus (RITC). The GRAC makes it possible to search any linguistic phenomenon using NoSketchEngine interface that in turn enables search by lemma, word form and grammatical tags, visualization of frequencies as a concordance, customization of text filters (texts of a given period, style, original language, etc.) [54]. GRAC's functionality also provides automatically retrieved full information about word-forms, lemmas, parts of speech and their frequency etc. "GRAC is intended to be a universal tool for a wide range of research questions" [55].

4. Experiment and results

4.1. Statistical characteristics of Roman Ivanytchuk's lexicon as an idiolect marker

For multifaceted idiolect study linguists make a quantitative description of writers' texts, which provides accurate information about the peculiarities of vocabulary functioning in these texts [7, p. 86].

To carry out integrated research of the author's idiolect, the subcorpus of Ukrainian prose fiction (UPFTC) was created in the GRAC by applying filters like style Fiction (DOC.STYLE — FIC), original language Ukrainian (DOC.ORIGINAL — UK), time span (DOC.DATE — 1960–2016).

Roman Ivanychuk's subcorpus data are compared with the data of Ukrainian prose fiction text corpus for the period of 1960-2016 to reveal the peculiarities of Roman Ivanychuk's idiolect.

Frequency of the parts of speech analysis, vocabulary ranking, calculation of morphological and statistical indicators for vocabulary, statistical characteristics of the vocabulary and text display frequency patterns of the idiolect, which allows the authorization of the text and its further automatic processing.

The study of the parts of speech frequency in the text is essential for revealing individual writer's characteristics [46, p. 186]. To present the statistical characteristics of nouns, adjectives, verbs, adverbs, prepositions in Roman Ivanytchuk's texts CQL queries are used (other parts of speech aren't studied due to problems with removing homonymy). The relative frequency of the mentioned parts of speech in Roman Ivanychuk's subcorpus have been calculated and are manifested in the table 1.

Table 1

The frequency of parts of speech in RITC

№	Text title	Noun	Adjective	Verb	Adverb	Preposition
1	At The Edge of the Paven Way (Krai bytoho shliakhu)	38,7	14,7	19,1	10,4	10,4
2	Mallows (Mal'vy)	40,8	16,4	18,1	8,8	11,1
3	Red Wine (Cherlene vyno)	40,5	15,4	17,4	8,1	11,5
4	Manuscript from Ruska Street (Manuskrypt z vulytsi Rus'koyi)	39,8	15,9	17,7	8,9	10,7
5	Water from the Stone (Voda z kameniu)	39	15,9	17,8	9,1	10,9
6	The Fourth Dimension (Chetvertyi vymir)	41,5	16,8	16,9	9,4	10,8
7	Scars on the Rock (Shramy na skali)	41	16,9	17,2	8,8	10,9
8	Crane's Cry (Zhuravlynyi kryk)	40,9	16,3	17,7	9	10,8
9	Because War Is War (Bo viyna viynoyu)	40,9	16,2	17,4	9,1	11,7
10	Horde (Orda)	42,5	16,7	17,4	8,5	11,3
11	The Gospel of Thomas (Yevanheliye vid Tomy)	42,3	18,3	16,6	8,1	11,7
12	Saxaul in The Sands (Saksaul u piskakh)	40,2	16,5	17,1	9,3	11,5
13	Pillars of Fire (Vohnenni stovpy)	39,5	15,9	17,8	9,5	12,1
14	Across The Pass (Cherez pereval)	40,2	17,8	17,1	9,3	11,4

№	Text title	Noun	Adjective	Verb	Adverb	Preposition
15	Pilgrimage (Khresna proshcha)	41,8	18,5	16,7	8,8	12,2
16	Voices from above The Waters of Kinneret (Holosy z-nad vod Henisareta)	42,6	19,1	15,9	7,3	12,6
17	I Have Not Written About Donbass Yet (Ya shche ne pysav pro Donbas)	41,4	18,2	15,8	8,2	11,9

The obtained data in table 1 demonstrate that the relative frequency of nouns in RITC fluctuates within the range 38,7 (*Krai bytoho shliakhu*) and 42,6 (*Holosy z-nad vod Henisareta*), adjectives — 14,7 (*Krai bytoho shliakhu*) and 19,1 (*Holosy z-nad vod Henisareta*), verbs — 15,8 (*Ya shche ne pysav pro Donbas*) and 19,1 (*Krai bytoho shliakhu*), advebs — 7,3 (*Holosy z-nad vod Henisareta*) and 10,4 (*Krai bytoho shliakhu*), prepositions — 10,4 (*Krai bytoho shliakhu*) and 11,9 (*Holosy z-nad vod Henisareta*). In his early works the frequency of nouns and adjectives is higher than the frequency of verbs and adverbs, as compared to the texts of later period.

The abovementioned data in Roman Ivanychuk's subcorpus and those from the subcorpus of Ukrainian prose fiction are compared and presented in the table 2.

Table 2

The frequency of parts of speech in RITC and UPFTC

Part of speech	RITC	UPFTC
Noun	40,1	38,5
Adjective	16,6	16
Verb	17,5	17,7
Adverb	9	11,3
Preposition	11,4	10,1

The frequency of the parts of speech in RITC correlates with data retrieved in UPFTC, however the frequency of nouns in RITC is higher by ≈ 5%, prepositions is by ≈ 11% than in UPFTC, in contrast to the frequency of adverbs, which is higher by ≈ 20 %.

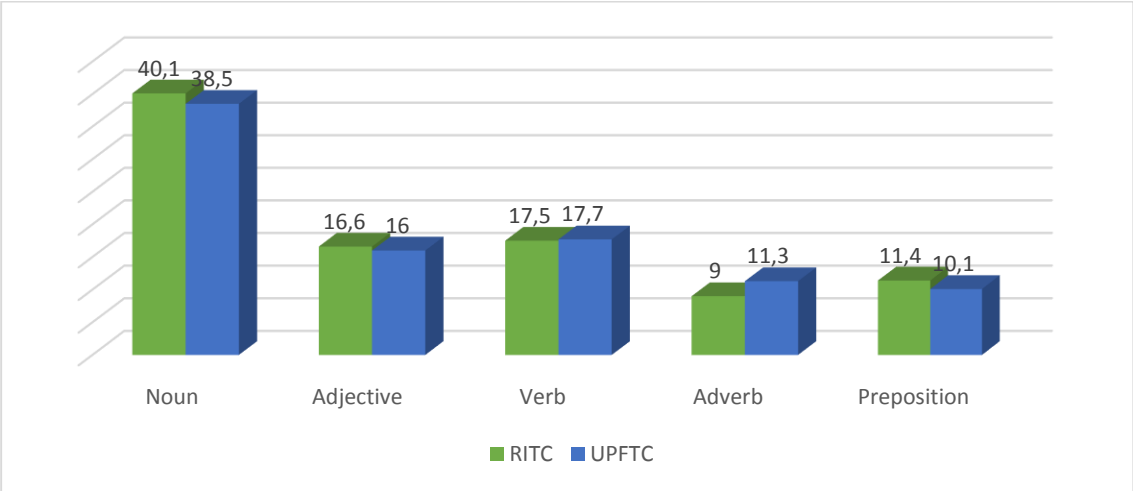

Figure 1: The frequency of parts of speech in RITC and UPFTC

The study of Roman Ivanychuk's idiolect at the morphological level represents morphological statistical characteristics, such as the index of epithetization (correlation between total noun occurrences and total adjective occurrences), the index of verbal definitions (correlation between total adverb occurrences and total verb occurrences), the level of nominalization (correlation between total noun occurrences and total verb occurrences) [5, 6]. The data received are presented in the table 3.

Table 3
Morphological statistical characteristics in RITC and UPFTC

Morphological statistical characteristics	RITC	UPFTC
index of epithetization	2,44645	2,41239
index of verbal definitions	0,517969	0,642332
level of nominalization	2,319452	2,181664

The data in the table 3 indicate that the number of adjectives per noun in RITC is higher than in UPFTC. The ratio of adverbs to verbs demonstrates higher number of verb collocations in RITC than in UPFTC. These data serve as an idiolect peculiarity.

Yuhan Tuldava considers the vocabulary in text as a system and supposes to study it by means of "quantitative" mathematics that, in turn, permits to identify and comprehend its system properties [38]. Galina Napreenko [42] suggests the word frequency in the text is an identification parameter to determine its authorship. The frequency of a lexical unit is an important characteristic of a word, as it indicates the activity of its functioning in the text, its value in statistical structure of the text.

In this study the statistical parameters of author's vocabulary and text are the following: *text size* (N) — total number of words in the text; *size of vocabulary of lexemes* (V) — a number of lemmas in the text; *diversity index* (V/N) — a ratio of size of *vocabulary of lexemes* (V) to text size (N); *average word frequency* in the text (N/V) — the ratio of text size (N) to size of *vocabulary of lexemes* (V); *hapax legomena* (V1) — a number of lexemes with frequency 1; *exclusivity index of vocabulary* (V1/V) — the ratio of number of lexemes with frequency 1 (V1) to total number of lemmas (V); *exclusivity index of text* (V1/N) — the ratio of number of lexemes with frequency 1 (V1) to text size (N); *concentration index of vocabulary* (V10/V) —the ratio of lexemes with frequency 10 (V10) and more to size of vocabulary of lexemes (V); *concentration index of text* (V10t/N) — the ratio of words with frequency 10 (V10) and more to text size (N) [5, 6].

As a result of RITC and UPFTC analysis we received the data regarding the vocabulary in the texts under study which are presented in the table 3.

Table 3
Statistical parameters in RITC

	Text title	Text size (N)	Size of vocabulary of lexemes (V)	Diversity index (V/N)	Average word frequency (N/V)	Hapax legomena (V1)	Exclusivity index of vocabulary (V1/V)	Exclusivity index of text (V1/N)	Concentration index of vocabulary (V10/V)	Concentration index of text (V10t/N)
1	At The Edge of the Paven Way (Krai bytoho shliakhu)	119231	13231	0,111	9	6133	0,051	0,464	0,757	0,108
2	Mallows (Mal'vy)	69386	10411	0,15	6,7	5247	0,076	0,504	0,715	0,090
3	Red Wine (Cherlene vyno)	46960	8723	0,186	5,4	4784	0,102	0,548	0,641	0,075
4	Manuscript from Ruska Street (Manuskrypt z vulytsi Rus'koyi)	61715	9864	0,159	6,3	5104	0,083	0,517	0,672	0,086
5	Water from the Stone (Voda z kameniu)	69148	11569	0,167	5,9	5983	0,087	0,517	0,677	0,079
6	The Fourth Dimension (Chetvertyi vymir)	60693	10745	0,177	5,6	5748	0,095	0,535	0,687	0,075
7	Scars on the Rock (Shramy na skali)	69456	12120	0,174	5,7	6432	0,093	0,531	0,679	0,072
8	Crane's Cry (Zhuravlynyi kryk)	125383	16278	0,129	7,7	7645	0,061	0,470	0,743	0,103

	Text title	Text size (N)	Size of vocabulary of lexemes (V)	Diversity index (V/N)	Average word frequency (N/V)	Hapax legomena (V1)	Exclusivity index of vocabulary (V1/V)	Exclusivity index of text (V1/N)	Concentration index of vocabulary (V10/V)	Concentration index of text (V10t/N)
9	Because War Is War (Bo viyna viynoyu)	71317	12128	0,17	5,9	6385	0,090	0,526	0,682	0,078
10	Horde (Orda)	59715	10326	0,173	5,8	5465	0,092	0,529	0,872	0,075
11	The Gospel of Thomas (Yevanheliye vid Tomy)	92015	13118	0,142	7	6416	0,070	0,489	0,773	0,095
12	Saxaul in The Sands (Saksaul u piskakh)	62087	11207	0,181	5,5	6048	0,097	0,540	0,671	0,073
13	Pillars of Fire (Vohnenni stovpy)	143849	16899	0,117	8,5	7744	0,054	0,458	0,781	0,110
14	Across The Pass (Cherez pereval)	50943	10278	0,201	4,9	5772	0,113	0,562	0,638	0,064
15	Pilgrimage (Khresna proshcha)	89272	13995	0,156	6,4	6977	0,078	0,499	0,708	0,087
16	Voices from above The Waters of Kinneret (Holosy z-nad vod Henisareta)	34223	8505	0,248	4	4868	0,142	0,572	0,624	0,053
17	I Have Not Written About Donbass Yet (Ya shche ne pysav pro Donbas)	9612	3306	0,343	2,9	2188	0,228	0,662	0,511	0,033

The novel *Vohnenni stovpy* comprises the highest rate of *text size* (143849), the largest *vocabulary of lexemes* (16899), the highest rate of *hapax legomena* (7744), the highest rate of *concentration index* of vocabulary (0,110). Meanwhile the novel *Ya shche ne pysav pro Donbas* holds the lowest indicator of *text size* (9612), the smallest *vocabulary of lexemes* (3306), the lowest rate of *hapax legomena* (2188), the lowest rate of *concentration index* of vocabulary (0,033). Although in the novel *Ya shche ne pysav pro Donbas* there is the highest rate of *diversity index* (0,343), the highest rate *of exclusivity index* of vocabulary (0,662), as compared with the novel *Vohnenni stovpy* where there are the lowest indexes of *diversity* (0,117) and *vocabulary exclusivity* (0,458).

The index of diversity is inversely proportional to text length, the longer text is the less unique words it potentially possesses [46, p. 143]. *Hapax legomena* usually cover 40-60% of text [24, p. 72]. In Roman Ivanychuk's texts *hapax legomena* index varies between 46-66%. Thus, *concentation index* of vocabulary is the opposite of *exclusivity index* of vocabulary, that is confirmed by RITC.

To study the peculiarities of statistical structure in Roman Ivanychuk's text the data of RITC and UPFTC were taken into consideration and compared (see the table 4).

Table 4

Statistical parameters in RITC and UPFTC

Statistical characteristics	RITC	UPFTC
Text size (N)	1235014	76744330
Size of vocabulary of lexemes (V)	49828	288755
Diversity index (V/N)	0,040	0,004
Average word frequency (N/V)	24,8	265,8
Hapax legomena (V1)	16540	102725
Exclusivity index of text (V1/N)	0,013	0,001

Statistical characteristics	RITC	UPFTC
Exclusivity index of vocabulary (V1/V)	0,332	0,356
Concentration index of text (V10t/N)	0,812	0,912
Concentration index of vocabulary (V10/V)	0,284	0,346

Due to the GRAC functionality the data of words with frequency 10 and more (V_{10t} = 1002836) and lexemes with frequency 10 and more (V_{10} = 14178) are presented. It is found that in Roman Ivanychuk's texts *hapax legomena* (V_1) involve 16540 words, exclusive vocabulary (33%) predominates high-frequency vocabulary (28%) in the author's lexicon. These data indicate the diversity and the richness of Roman Ivanychuk's vocabulary.

Meanwhile in UPFTC words with frequency 10 and more cover 70051172 words, lexemes with frequency 10 and more — 100050 lemmas, *hapax legomena* (V_1) — 102725 words. These data mean that exclusive vocabulary (36%) predominates high-frequency vocabulary (34%) too. The part of high-frequency vocabulary is much higher in Ukrainian prose fiction text corpus because its size is 62 times bigger and diversity index is 10 times lower that in Roman Ivanychuk's text corpus, which explains the outweigh law.

It is known that in speech speakers give preference to a small number of units, which are of high frequency [43]. They form the core of any speech subsystem, while most units are low frequent [37]. This regularity was noticed by Dewey and called the *outweigh law*, later on, it was further researched by the German linguist J. Zipf, who formulated the *Zipf's law*, which sets the dependences [18]

It should be noted that the larger the text corpus is the more informative it is. In RITC the indicator of average word frequency is 25, that is each word is used, on average, about 25 times. The relatively small number of high-frequency vocabulary (low concentration index accordingly) and relatively large number of words with frequency 1 (therefore, high index of exclusivity) indicate a great diversity of vocabulary in Roman Ivanychuk's texts.

Frequency ranking of vocabulary provides information about the core and periphery of writer's dictionary. Considering this, one of the stages of our study is the creation of a frequency dictionary based on RITC and UPFTC to detect the frequency zones. Yuhan Tuldava pointed out [62, p. 65] that stratification of vocabulary by frequency (that is identifying frequency zones of words), is important to determine text complexity, to create minimum dictionary, to process text automatically.

Frequency-rank patterns represent the text structure and are interpreted as manifestations of individual preferences of linguistic personality in the choice of use of certain lexical units [34]. The vocabulary of Roman Ivanychuk's texts and Ukrainian prose fiction texts is divided into four zones according to the interval of ranks (see the table 5).

Table 5

Frequency zones of vocabulary in RITC and UPFTC

	RITC			UPFTC		
	Frequency zone	Number of words	Coverage, %	Frequency zone	Number of words	Coverage, %
1	more than 1000	136	0,4%	more than 1000	7 331	2,4%
2	100 — 999	1 301	2,6%	100 — 999	28 402	9,6%
3	10 — 99	10 439	21%	10 — 99	64 329	22%
4	1 — 9	37 963	76%	1 — 9	188 705	65%

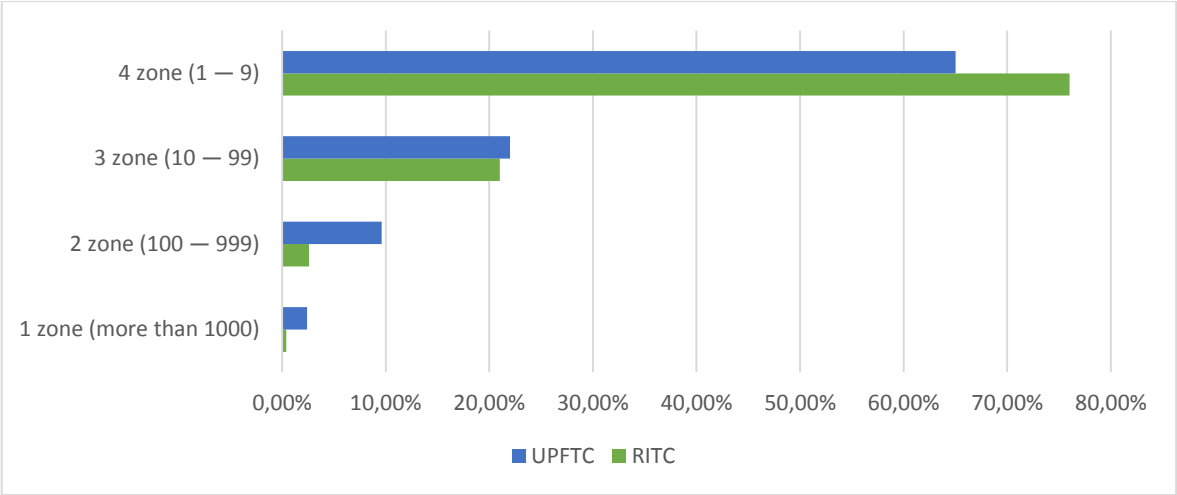

Figure 2: Frequency zones of vocabulary and their coverage in RITC and UPFTC

As the table 5 and the figure 2 show the fourth zone in RITC (76%) is the most consistent in terms of all contentious words, hapax legomena cover 33% the author's vocabulary, this data manifests the uniqueness of the author's idiolect. The second zone includes the largest number of common / general words, the third — words specific to literary style, the first zone consists of official and uninformative common / general words, which serve as formal markers for text attribution [34].

Much less coverage with high-frequency vocabulary is detected in Roman Ivanychuk's texts as compared to Ukrainian prose fiction texts, which makes his texts unique. A specific feature of frequency zone in RITC is higher coverage with low-frequency vocabulary and lower coverage with high-frequency vocabulary than in UPFTC.

5. Discussion and conclusion

In this study Roman Ivanychuk's and Ukrainian prose fiction subcorpora based on GRAC utility enables to manifest and compare quantitative characteristics and qualitative indicators of the author's lexicon. The author's idiolect is studied from the point of view of lexical arsenal by means of statistical parameters that make it possible to extract lexical markers of idiolect and identify his texts among others.

GRAC's functionality enables to get information on the number of word usages, word forms, lemmas automatically, and to compile a frequency dictionary of RITC and UPFTC as well.

Corpus-based dictionary offers an entirely new and much richer type of information, opens new possibilities enabling comparison of one single man vocabulary with that of another and allows to solve different problems. Methodologically, it is obvious that having more dictionaries of the type from various time periods offers a chance to study idiolects in a principled and objective way and follow their developments through time [9].

As the result of the research the distribution and the frequency of parts of speech, concordance, morphological statistical characteristics of vocabulary (index of epithetization, index of verbal definitions, level of nominalization) and statistical parameters of vocabulary and text (text size, size of lexemes vocabulary, diversity index, average word frequency, hapax legomena, exclusivity index of text and vocabulary, concentration index of text and vocabulary), frequency zones of vocabulary are presented.

The statistical characteristics of text and vocabulary allow to determine the qualitative features of the writer's idiolect objectively. The quantitative relations between parts of speech are an important element of statistical text characteristics. Frequency-rank regularities represent text structure and can be interpreted as manifestations of individual preferences of linguistic personality in the choice of certain lexical units.

Obtained data on parts of speech, text structure, vocabulary, frequency zones in Roman Ivanychuk's text corpus are different from those in Ukrainian fiction prose text corpus, which, in its turn, demonstrate the specificity of the author's idiolect. The practical results of the study can be applied for text identification and further research of writer's individual language. This type of study can be used not only for idiolect investigations, but can also serve as data in other contexts, like authorship attribution, stylometric studies, and for literature researches.

6. References

[1] P. Baker, Using Corpora in Discourse Analysis, A&C Black, 2006, 197 p.

[2] M. Barlow, Individual usage: a corpus-based study of idiolects, in: Proceedings of LAUD Conference. 2010.

[3] M. Barlow, Individual usage: a corpus-based study of idiolects, University of Auckland. International Journal of Corpus Linguistics 18(4), 2013. doi:10.1075/ijcl.18.4.01bar

[4] D. Biber, S. Conrad, Register, genre, and style, Cambridge University Press (2009) 344 p.

[5] S. Buk, Distinguishing Quantitative Parameters of Author's Language and Style (A Case of Ivan Franko Long Prose Fiction), Visnyk Lvivskoho universytetu, Seriia filolohichna, Vyp. 70 (2019) 299–308. [S. Buk, Distinguishing Quantitative Parameters of Author's Language and Style (A Case of Ivan Franko Long Prose Fiction), Bulletin of Lviv University, Philological Series, Vol. 70 (2019) 299–308]

[6] S. Buk, Quantitative analysis of the novel Ne Spytavńy Brodu by Ivan Franko in the Light of Statistical and Quantitative Linguistics, Speech and context, International Journal of Linguistics, Semiotics, and Literary Science, Vol. 1(VI) (2014) 100–112.

[7] S. Buk, Suchasni metody doslidzhennia movy pysmennyka u slovianoznavstvi, Problemy slovianoznavstva, Vyp. 61 (2012) 86–95. [S. Buk, Modern methods of studying the writer's language in Slavic studies, Problems of Slavic studies, Vol. 61 (2012) 86–95]

[8] F. Čermák, Slovník Karla Čapka, Praha, Nakladatelství Lidové noviny, Ústav Českého národního korpusu, 2007, 715 s.

[9] F. Čermák, V. Cvrček. Author Dictionaries Revisited: Dictionary of Bohumil Hrabal, Institute of the Czech National Corpus, Charles University Prague (2010) 592–598.

[10] Yu. O Danchevska., I. M. Kulchytskyi Deiaki aspekty stvorennia korpusu khudozhnikh tvoriv V. S. Stefanyka, MegaLing–2012 «Prykladna linhvistyka ta linhvistychni tekhnolohii», Kyiv, 2013, ss. 143–149. [Yu. O Danchevska., I. M. Kulchytskyi, Some aspects in creation of the corpus of V.S. Stefanyk's literary texts, in: Proceedings of MegaLing-2012 "Applied Linguistics and Linguistic Technologies", Kyiv, 2013, pp. 143–149]

[11] I. Danyliuk, A Zahnitko., H. Sytar, Korpus tekstiv Yuriia Shevelova: struktura, funktsii, navihatsiia, Mova: klasychne – moderne – postmoderne, Vyp. 5 (2019) 158–169. URL: http://nbuv.gov.ua/UJRN/Langcmp_2019_5_14 [I. Danyliuk, A Zahnitko., H. Sytar, Yuri Shevelyov's Corpus Texts: structure, functions, navigation, Language: classical - modern - postmodern,. Vol. 5 (2019) 158–169. URL: http://nbuv.gov.ua/UJRN/Langcmp_2019_5_14]

[12] M. Darwich, S. A. Mohd, N. Omar, N. A.Osman, Corpus-Based Techniques for Sentiment Lexicon Generation: A Review, J. Digit. Inf. Manag., 17(5), 2019, 296 p.

[13] T. D. Demidova, Periferiynaya chast leksikona kak pokazatel literaturnoy maneryi pisatelya (na materiale liricheskih miniatyur V.P. Detkova), Vestnik LGU im. A.S. Pushkina, № 2, 2011. URL: http://cyberleninka.ru/article/n/periferiynaya-chast-leksikona-kak-pokazatel-literaturnoy-manerypisatelya-na-materiale-liricheskih-miniatyur-v-p-detkova [T. D. Demidova, Peripheral part of the lexicon as an indicator of the literary manner of writer (based on lyrical miniatures of V. P. Detkov), Bulletin of the Leningrad State University named after A.S. Pushkin, No. 2, 2011. (2019) 158–169. URL: http://nbuv.gov.ua/UJRN/Langcmp_2019_5_14]

[14] N. Dittmar, Explorations in 'Idiolects', Amsterdam Studies in the Theory and History of Linguistic Science Series 4 (1996) 109–128.

[15] O. Halvani, Ch. Winter, L. Graner, Assessing the Applicability of Authorship Verification Methods, in: Proceedings of the 14th International Conference on Availability, Reliability and Security, No. 38, 2019, pp. 1–10. URL: https://doi.org/10.1145/3339252.3340508.

[16] N. Hrytsiv, I. Kulchytskyy, O. Rohach, Quantitative Comparative Analysis in Parallel Translation Corpus: building author's and translator's statistical profiles: (a case study of Lucy Maud Montgomery), in: Proceedings 2020 IEEE 15th International Conference on Computer Sciences and Information Technologies (CSIT), 2020, pp. 255–258. doi: 10.1109/CSIT49958.2020.9321893

[17] N. Hrytsiv, T. Shestakevych, J. Shyyka, Quantitative Parameters of Lucy Montgomery's Literary Style, in CEUR Workshop Proceedings, Vol. 2870, 2021, pp. 670–684.

[18] A. G. Jivani, A Comparative Study of Stemming Algorithms, Int. J. Comp. Tech. Appl., Vol. 2, Issue 6 (2011) 1930–1938.

[19] Yu. O. Kalymon, Strukturno-informatsiina model slovnyka movy novel Vasylia Stefanyka, dys. … kand. filol. nauk, Lviv, 2020, 312 s. [Yu. O. Kalymon, Structural-informational dictionary model of Vasyl Stefanyk's short stories language, Thesis for a Candidate Degree in Philology, Lviv, 2020, 312 p.].
[20] M. Khokhlova, Yssledovanye leksyko-syntaksycheskoy sochetaemosti v russkom yazyike s pomoshchiyu statystycheskykh metodov (na baze korpusov tekstov), avtoref. dys. na soysk. uch. step. kand. fylol. nauk, "Prykladnaya i matematycheskaya lynhvystyka", Sankt-Peterburg (2010) 218 s. [M. Khokhlova, The study of lexical and syntactic collocatibility in the Russian language using statistical methods (based on Text Corpus), Sankt-Peterburg, (2010) 218 p.]
[21] I. Khomytska, V. Teslyuk, Authorship and Style Attribution by Statistical Methods of Style Differentiation on the Phonological Level, volume 871 of Advances in Intelligent Systems and Computing III, AISC, Springer, 2019, pp. 105-118.
[22] I. Khomytska, V. Teslyuk, A. Holovatyy, O. Morushko, Development of methods, models, and means for the author attribution of a text, volume 3(2-93) of Eastern-European Journal of Enterprise Technologies, 2018, pp. 41-46.
[23] R. Köhler, G. Altmann, Aims and Methods of Quantitative Linguistics, Problems of Quantitative Linguistics, Chernivci (2005) 12–42.
[24] A. Kornai, Mathematical Linguistics, London, Springer, XIII, 2008, 289 p.
[25] T. Kosmeda, A. Zahnitko, Zh. Krasnobaieva-Chorna. Delineation of Linguopersonology and Linguoaxiology, Uniwersytet im. Adama Mickiewicza w Poznaniu, Wydawnictwo Naukowe UAM, Poznań, 2019.
[26] O. P. Kostetska, Indyvidualne movlennia avtora yak obiekt linhvistyky ta pidkhody do yoho doslidzhennia, Naukovi zapysky, Natsionalnyi universytet Ostrozka akademiia, Seriia: Filolohichna, № 49 (2014) 196–199. [O. P. Kostetska, The author's individual speech as an object of linguistics and approaches to its research, Scientific Notes, Ostroh Academy National University, Philological Series, No. 49 (2014) 196–199]
[27] I. M. Kulchytskyi, Unormuvannia tekstu pid chas dokorpusnoho opratsiuvannia: dosvid zastosuvannia. Visnyk Natsionalnoho universytetu "Lvivska politekhnika", Seriia: Informatsiini systemy ta merezhi, Vyp. 7 (2020) ss. 51–58. [I. M. Kulchytskyy, Text normalization during pre-corpus preparation: experience of application, Bulletin of the National University "Lviv Polytechnic", Series: Information systems and networks, Vol. 7 (2020) pp. 51–58]
[28] I. Kulchytskyi, U. Shandruk, The quantitative research of scientific texts at the symbolic level. In: Computational linguistics and intelligent systems, vol 2 (2018) 71–80.
[29] K. Lagutina et al., *A* Survey on Stylometric Text Features, 25th Conference of Open Innovations Association (FRUCT), 2019, pp. 184-195. doi: 10.23919/FRUCT48121.2019.8981504.
[30] G. Lakoff, The Contemporary Theory of Metaphor. In Metaphor and Thought, Cambridge, Cambridge University Press (1998) 202–249.
[31] V. V. Levitskiy, Kvantitativnyie metodyi v lingvistike. Nova Kniga, Vinnitsa (2007) 264 s. [Levitsky V.V. Quantitative methods in linguistics. In: Nova Kniga, Vinnitsa, 264 p. (2007)
[32] N. Lototska, Statistical analysis of collocations of the concept joy in R. Ivanychuk's text corpus, Scientific Journal of Polonia University, Vol. 37 No 6. (2019) 92–98.]
[33] N. Lototska, Statistical Research of the Colour Component ЧОРНИЙ (BLACK) in R. Ivanychuk's Text Corpus, in: Proceedings of the 5th International Conference on Computational Linguistics and Intelligent Systems (COLINS 2021),Vol. I, Lviv, Ukraine, 2021, pp. 486–497.
[34] N. Ya. Lototska, Idiolekt Romana Ivanychuka: korpusnobazovanyi ta linhvokohnityvnyi pidkhody, Dysertatsiia na zdobuttia naukovoho stupenia doktora filosofii za spetsialnistiu 035 — Filolohiia, Natsionalnyi universytet «Lvivska politekhnika», Lviv, 2021. [N. Ya. Lototska, The idiolect of Roman Ivanychuk: corpus-based and linguo-cognitive approaches, Ph.D. thesis, specialty 035 Philology, Lviv Polytechnic National University, Lviv, 2021]
[35] V.Lytvyn, V. Vysotska, I.Budz, Ya. Pelekh, N. Sokulska, Development of the Quantitative Method for Automated Text Content Authorship Attribution Based on the Statistical Analysis of N-grams Distribution, 2019 DOI: 10.15587/1729-4061.2019.186834
[36] V. Lytvyn, V. Vysotska, P. Pukach, Z. Nytrebych, I. Demkiv, A. Senyk, O. Malanchuk, S. Sachenko, R. Kovalchuk, N. Huzyk, Analysis of the developed quantitative method for automatic attribution of

scientific and technical text content written in Ukrainian, volume 6(2-96) of Eastern-European Journal of Enterprise Technologies, 2018, pp. 19-31. DOI: 10.15587/1729-4061.2018.149596
[37] V. Lytvyn, V. Vysotska, Y. Burov, O. Veres, I. Rishnyak, The Contextual Search Method Based on Domain Thesaurus, Advances in Intelligent Systems and Computing. (2017) 310–319. doi: https://doi.org/10.1007/978-3-319-70581-1_22
[38] M. Mahlberg, P. Stockwell, J. Joode, C. Smith, M. O'Donnell, CLiC Dickens: novel uses of concordances for the integration of corpus stylistics and cognitive poetics. URL: https://research.birmingham.ac.uk/portal/files/38225413/ cor_2E2016_2E0102.pdf
[39] S. Mollin,"I entirely understand" is a Blairism: The methodology of identifying idiolectal collocations. *International Journal of Corpus Linguistics*, *14*(3) (2009) 367–392. DOI: *https://doi.org/10.1075/ijcl.14.3.04mol*
[40] S. T. Mubin, S. P. Rajesh, Authorship Identification with Multi Sequence Word Selection Method, in: Thermal Stresses—Advanced Theory and Applications, 2019, pp. 653–661.
[41] M. Yu. Muhin, Kontseptualnyie profili proizvedeniy M. Bulgakova, V. Nabokova, A. Platonova i M. Sholohova (po dannyim sopostavitelnogo analiza chastotnoy leksiki), Vestnik BFU im. I. Kanta, № 8, 2010. URL: http://cyberleninka.ru/article/n/kontseptualnye-profili-proizvedeniy-m-bulgakova-v-nabokova-aplatonova-i-m-sholohova-po-dannym-sopostavitelnogo-analiza-chastotnoy [M. Yu. Mukhin, Conceptual profiles of texts by M. Bulgakov, V. Nabokov, A. Platonov and M. Sholokhov (according to the comparative analysis of frequency vocabulary), Bulletin of the BFU named after I. Kanta, No. 8, 2010. URL: http://cyberleninka.ru/article/n/kontseptualnye-profili-proizvedeniy-m-bulgakova-v-nabokova-aplatonova-i-m-sholohova-po-dannym-sopostavitelnogo-analiza-chastotnoy]
[42] G. V. Napreenko, Internet-dnevniki i problema identifikatsii lichnosti, Yurislingvistika 11, Pravo kak diskurs, tekst i slovo, pod red. N. D. Goleva, K. I. Brineva, Kemerovo, Izd-vo Kemerovskogo gosudarstvennogo universiteta (2011) 480–492. [G. V. Napreenko, Internet diaries and the problem of personal identification, Jurislinguistics 11, Law as discourse, text and word, ed. N. D. Goleva, K. I. Brinev, Kemerovo, Publishing House of Kemerovo State University (2011) 480–492]
[43] O. Naum, L. Chyrun, V. Vysotska, O. Kanishcheva, Intellectual system design for content formation, in: 12th International Scientific and Technical Conference on Computer Sciences and Information Technologies (CSIT), 2017. doi: https://doi.org/10.1109/stc-csit.2017.8098753
[44] Yu. N. Pavlov, E. A. Tihomirova, Otsenka ustoychivosti vo vremeni chastotnyih slovarey avtorov v zadachah identifikatsii tekstov, Nauka i obrazovanie, № 12, 2011. URL: http://cyberleninka.ru/article/n/77-30569-274006-otsenka-ustoychivosti-vo-vremenichastotnyh-slovarey-avtorov-v-zadachah-identifikatsii-tekstov [Yu. N. Pavlov, E. A. Tikhomirova, Time stability estimation of authors' frequency dictionaries in text identification problems, Science and Education, No. 12, 2011 URL: http://cyberleninka.ru/article/n/77-30569-274006-otsenka-ustoychivosti-vo-vremenichastotnyh-slovarey-avtorov-v-zadachah-identifikatsii-tekstov]
[45] O. O. Pavlychko, Shchodo statystychnykh parametriv avtorskoho styliu (na materiali tvoriv E.M. Remarka), Movni i kontseptualni kartyny svitu, VPTs «Kyivskyi un-t», Kyiv, Vyp. 29 (2010) 186–191. [O. O. Pavlychko, Regarding the statistical parameters of the author's style (based on the texts of E.M.Remark), Linguistic and conceptual worldview, PPC Kyiv University, Kyiv, Vol. 29 (2010) 186–191]
[46] V. S. Perebyinis., M.P. Muravytska., N. P. Darchuk Chastotni slovnyky ta yikh vykorystannia, Kyiv, 1985, 204 s. [V.S. Perebyinis., M. P. Muravytska., N. P. Darchuk, Frequency dictionaries and their use, Kyiv, 1985, 204 p.]
[47] V. I. Perebyinis, Shcho daie statystyka movoznavtsiam?, Visnyk Kyivskoho linhvistychnoho universytetu, Seriia Filolohiia, Kyiv, Vyd. tsentr KNLU, T. 6, № 2. (2003) 27–32. [V. I. Perebyynis, What does statistics give to linguists?, Bulletin of Kyiv Linguistic University, Philology Series, Kyiv, Publishing House KNLU, Vol. 6, No. 2 (2003) 27–32]
[48] O. Perelomova, Idiostyl Valeriia Shevchuka, dys. … kand. filol. nauk, 10.02.01, Sumy, 2002. 177 s. [O. Perelomova, Valery Shevchuk's Idiostyle, Ph.D. thesis, 10.02.01, Sumy, 2002. 177 p.]
[49] N. Pylypiuk, O Ilnytzkyj., S. Kozakov, Online Concordance to the Complete Works of Hryhorii Skovoroda, 2013. URL: http://www.arts.ualberta.ca/~ukr/skovoroda/NEW/index.php?glang

[50] S. Raj, B. Kannan, and V. P. Jagathy Raj, Significance of Network Properties of Function Words in Author Attribution, Intelligent Data Engineering and Analytics, Springer, Singapore, 202, pp. 171-181. https://doi.org/10.1007/978-981-15-5679-1_17
[51] A. Rovenchak, S. Buk, Part-of-speech sequences in literary text: Evidence from Ukrainian, Journal of Quantitative Linguistics, Vol. 25, No. 1, (2018) 1–21. doi: https://doi.org/10.1080/09296174.2017.1324601
[52] M. Ruszkowski, Statystyka w badaniach stylistyczno-składniowych, Kielce, Wydawnictwo Świętorszyskiej, 2004, 144 s.
[53] O. Seminck, Ph. Gambette, D. Legallois, T. Poibeau, The Corpus for Id iolectal Research (CIDRE), Journal of Open Humanities Data, Ubiquity Press, 7, 2021, pp. 15.
[54] M. Shvedova, The General Regionally Annotated Corpus of Ukrainian (GRAC, uacorpus.org): Architecture and Functionality, in: Proceedings of the 4th International Conference on Computational Linguistics and Intelligent Systems, COLINS 2020, Vol. I, Lviv, Ukraine (2020) pp. 489–506.
[55] M. Shvedova, R. von Waldenfels, S. Yarygin, A. Rysin, V. Starko, M. Woźniak, M. Kruk et al. GRAC: General Regionally Annotated Corpus of Ukrainian, 2017–2021. URL: http://uacorpus.org/
[56] J. Sinclair Corpus, Concordance, Collocation, Oxford, Oxford University Press, 1991, 200 p.
[57] H. Sytar, Osoblyvosti realizatsii frazeolohizovanykh rechen u tvorakh Bohdana Lepkoho, Linhvistychni studii, Vyp. 40 (1) (2020) 64–80. URL: http://nbuv.gov.ua/UJRN/lingst_2020_40(1)__7 [H. Sytar, Peculiarities of realization of phraseologized sentences in Bohdan Lepky's novels, Linguistic Studies, Vol. 40 (1) (2020) 64–80. URL: http://nbuv.gov.ua/UJRN/lingst_2020_40(1)__7]
[58] A. Stefanowitsch, Corpus linguistics: A guide to the methodology, Textbooks in Language Sciences 7, Berlin, Language Science Press, 2020.
[59] M. Stubbs, Quantitative Methods in Literary Linguistics, Cambridge, Cambridge University Press (2014) 46-62.
[60] H. Sytar, Syntaksychni frazeolohizmy v linhvopersonolohiinomu portreti Mykhaila Kotsiubynskoho, Teoriia linhvistychnykh paradyhm: kolektyvna monohrafiia na poshanu profesora, chlen-korespondenta NAN Ukrainy Anatoliia Zahnitka, za red. Zh. Krasnobaievoi-Chornoi, Vinnytsia: TOV «Nilan-LTD», 2019, ss. 172–195. [H. Sytar, Syntactic Phraseologisms in the Linguo-Personological Portrait of Mykhailo Kotsyubynsky, Theory of Linguistic Paradigms: A Collective Monograph in Honor of Professor Anatoliy Zagnitko, Corresponding Member of the National Academy of Sciences of Ukraine, ed. J. Krasnobayeva-Chorna, Vinnytsia, Nilan-LTD LLC, 2019, pp. 172–195]
[61] H. V. Sytar, Syntaksychni frazeolohizmy v linhvopersonolohiinomu portreti Yuriia Shevelova (na materiali korpusu tekstiv Yuriia Shevelova), Linhvistychni studii, Vyp. 37 (2019) 130–134. [H. V. Sytar, Syntactic phraseology in Yuri Shevelyov's linguo-personological portrait (on the material of Yuri Shevelyov's corpus texts), Linguistic Studies, Vol. 37 (2019) 130–134]
[62] Yu. A. Tuldava, Problemyi i metodyi kvantitativno-sistemnogo issledovaniya leksiki. Tallin, Valgus, 1987, 204 s. [Yu. A. Tuldava, Problems and methods of quantitative-systemic research of vocabulary, Tallinn, Valgus, 1987, 204 p.]
[63] A. I. Vehesh, Tradytsii ta novatorstvo ukrainskoi literaturno-khudozhnoi antroponimii posttotalitarnoi doby, dys. … kand. filol. nauk, 10.02.01, Ivano-Frankivsk, 2010. 273 s. [A.I. Vehesh, Traditions and innovations of Ukrainian literary and artistic anthroponymy of the post-totalitarian era, Ph.D. thesis, 10.02.01, Ivano-Frankivsk, 2010. 273 p.]
[64] V. I. Voloshuk, Linhvostylistychni osoblyvosti idiolektu Z. Lentsa v malykh epichnykh zhanrakh: avtoref. dys. … kand. filol. nauk, 10.02.04, Lviv, 2004. 20 s. [V.I. Voloshuk, Linguistic and stylistic features of Lenz's idiolect in small epic genres, Ph.D. thesis, 10.02.04, Lviv, 2004. 20 p.]
[65] V. Vysotska, V. Lytvyn, V. Kovalchuk, S. Kubinska, M. Dilai, B. Rusyn, L. Pohreliuk., L. Chyrun, S. Chyrun, O. Brodyak, Method of similar textual content selection based on thematic information retrieval, in: CSIT, Proceedings of the XIVth Scientific and Technical Conference, Lviv, 2019 pp. 1–6.
[66] W. Wimmer, G. Altmann, Review Article: On vocabulary richness, Journal of Quantitative Linguistics, Vol. 6, No. 1.(1999) 1–9.
[67] A. P. Zahnitko, Teoriia hramatyky i tekstu, Donetsk, 2014, 480 s. [A. P. Zahnitko, Theory and Grammar of the Text, Donetsk, 2014, 480 p.]

[68] O. Zuban, Lexicographical Database of Frequency Dictionaries of Morphemes Developed on the Basis of the Corpus of Ukrainian Language, in: Advances in Intelligent Systems and Computing IV, CSIT 2019, vol. 1080, Springer, Cham, 2020, doi: 10.1007/978-3-030-33695-0_37.