A corpus-based approach to author's idiolect study: lexicological aspect

Yuliia Kalymon – Olha Romanchuk – Nadiya Fedchyshyn –
Ulyana Protsenko – Nadiia Yurko

DOI: 10.18355/XL.2022.15.03.03

Abstract
The aim of this paper is to present the application of a multimodal approach to the analysis of the writer's lexicon. The versatility of the anthropocentric paradigm of modern linguistics and the approaches applied in our research are mindful of the choice of relevant research methods. The features of a writer's lexicon (or idiolect in a broader sense) have been of great interest among scholars though they have been mostly explored by literary critics. Nowadays, more linguists are also interested in thorough research on the subject by applying more rigorous scientific methods like quantitative and statistical ones. The paper herewith focuses on a few of them as a preliminary attempt to showcase some results. The primary concern is to make this study corpus-based. Hence it is crucial to create a corpus of works based on the generally acknowledged rules and principles. Therefore, such a corpus may contribute to the development of corpus linguistics and be applicable to other research regarding studying the writer's idiolect. Another point of interest is that such research may either support or somehow disprove established views regarding the writing style of a certain author (that is very useful when considering author attribution issues or even forensic linguistics, though that is not the case of the given research). Finally, lexical semantics plays one of the leading roles in defining dominant concepts of an author's writing style corresponding to the cognitive linguistics domain. As a result of such an approach, it may help to construct the lexicon of the author into an ideographic entity. Furthermore, similar findings or comparative analysis of corresponding objects of study may contribute to displaying the author's outlook and subsequently that of the whole nation.
Key words: corpus-based research, lexicon, author's idiolect, comparative study

Introduction

An individual's language can be like his / her business card, a kind of "fingerprint" that may distinguish him/her from other people (Coulthard, 2004: 432). In literature, this phenomenon is particularly prominent due to the unique style of a particular author, which is difficult to confuse with anyone else's one. N. Sovtys rightly notes that "through the language of literary works, the selection of language means from the national language fund and their artistic understanding is revealed through the *linguistic personality* of the writer …" (Sovtys, 2013: 477). The complexity of the approach to the problem of studying the language of fiction is explained by the specificity of the subject of study, its intermediate place at the intersection of two sciences: linguistics and literary criticism. The subject of text in linguistics is studied as a set of linguistic units, while in literary studies, the text is considered the expression of artistic, expressive, figurative language. In general, many scholars have discussed the issue of studying linguistic personality with the use of different approaches, ranging from psychological (I. Ogiyenko, O. Potebnia), language teaching (V. Sukhomlynskyi) and cultural relativism (S. Yermolenko, L. Matsko) (Romanchenko, 2015: 117-119). The concept of "language consciousness" was introduced in the study of the writer's linguistic personality and is also used in the study of the certain author writing style (Marchuk, 2012: 58 61; Selihey, 2009: 13 27; Struhanets, 2012: 128-133).

Supplementary to cognitive linguistics, the linguo-cultural approach is singled out. Firstly, this approach contributes to the understanding of cultural orientations of the ethnos through the prism of the linguistic outline of the literary work or national dictionaries, which encrypt nationally oriented and cultural information (Kostetska, 2014: 198; Mishenina, 2017). O. Ivanischeva states that by "calling the division of culture into material and spiritual a scientific abstraction, the researchers point out the unity of culture, where each material object had to evolve into an "idea" in the human brain before it had been created" (Ivanischeva, 2014). I. Berkeshchuk rightly notes that "the subjective image of the world has a basic, invariant part, common to all its bearers, and variable, which reflects the unique life experience of a human being. The invariant part is formed in the context of culture and reflects its system of meanings. Its variability is determined by the socio-cultural reality in which a person is "immersed"" (Berkeschuk, 2018: 65). According to V. Kononenko, "the idea of the unity of linguistic and cultural paradigm, which determines the common principles of national worldview, the existence of linguistic personality, determines the system of modern anthropomorphic views on language and culture" (Kononenko, 2008: 15-17). Therefore, anthropocentric aspects of linguistic phenomena within cognitive linguistics have increased the number of studies devoted to the study of the idiolect of a particular writer in Ukraine (e.g., S. Vorobkevych in the study by O. Kulbabska and N. Shatilova (Kulbabska, 2016)) as they actualize the anthropocentric paradigm of linguistics, which allows revealing the expression of the author's linguistic personality, his / her consciousness because literary texts materialize the author's linguistic picture of the world as a representative of the ethnic community. Nevertheless, none of this research was based on corpus data.

At present, corpus-based studies are quite multi-dimensional and are represented in numerous publications abroad and in Ukraine. They comprise the general theory of corpus linguistics (Demska, 2005; Demska, 2011), A. Pawlowsky (Pawlowski, 2006), V. Shyrokov and others (Shyrokov, 2005), J. Sinclair (Sinclair, 1991), corpus typology and methods of corpus data interpretation (Andrushenko, 2021; McEnery, 2012), correlation of corpus linguistics and other linguistic disciplines, principles of creating text corpora of natural languages, terminology standardization, etc., which are discussed in the works of I. Kulchytskyi (Kulchytskyi, 2015), T. McEnery and others (McEnery, 2006), A. O'Keefe and others (O'Keefe, 2007), M. Stubbs (Stubbs, 1996), J. Svartvik (Svartvik, 2007), W. Teubert (Teubert, 2007) and many others.

With the advent of corpora, their size and the range of tasks they are supposed to solve have expanded tremendously. Scientists prioritize creating balanced and representative corpora and developing a unified system of text elements marking (structural notation), lists of tags to distinguish words according to word classes, and then syntactic, semantic, and discourse connections marking. All these advances have been improved over the decades, and today some of them have already reached the level of full automation, and some are still the subject of debate (Kulchytskyi, 2020; Garside, 2016: 8–101). Ukrainian corpus linguistics is currently developing thanks to scientists working in major educational and state institutions (O. Levchenko at Lviv Polytechnic National Unieversity; S. Buk at Ivan Franko National University of Lviv, O. Demska at Kyiv-Mohyla National Academy, N. Darchuk at Taras Shevchenko National University; Ye. Karpilovska at Institute of Ukrainian language of the National Academy of Sciences of Ukraine, V. Starko at Ukrainian Catholic University, V. Shyrokov at Ukrainian Lingua-Information Fund of the National Academy of Sciences of Ukraine) and in cooperation with higher educational institutions abroad (M. Shvedova, R. von Waldenfels, V. Starko at Friedrich Schiller University of Jena, N. Kotsyba at Warsaw University, Kharkiv Pedagogical University and the University of Alberta). The vast majority of their work concern the

modern Ukrainian language and its representation in various genres (e.g., Rabus, 2021; Levchenko, 2020; Rovenchak, 2018).

Regarding literary text analysis through the prism of corpus data, the leading opinion is that such means are auxiliary and contribute to the process of triangulation of linguistic phenomena, i.e., studying using several approaches and techniques to help clarify or reconsider the object of study (Baker, 2006; Biber, 2011; Jones, 2012). Accordingly, we mention here three basic approaches: identification of linguistic facts through categorical analysis (procedural approach); correlation of facts using statistical methods (quantitative approach); and interpretation of results (cognitive approach) (Teubert, 128–129). If the first two can be fulfilled partially automated, the latter requires the intervention of the human mind because it concerns the interpretation of data.

Studying the language of literary works allows one to learn more about distinctive features of the author's stylistics and, in a broader sense, to understand the foundations of the mentality of the whole ethnic group (Baker, 2015: 5–22). M. Stubbs emphasizes that corpus linguistics allows a broader view of the literary text by "finding patterns in the use of language on the example of many different texts, texts of one genre, author or period" (Stubbs, 2014: 47). As B. Louw notes (Louw, 2013: 240–252), the leading role of the corpus lies in the perception of literature based on corpus data (reinforcement of intuitive sensations by factual data) and a new interpretation of works based on the corpus.

Ukrainian corpus linguistics may boast of separate corpora of Taras Shevchenko's poetry, Hryhoriy Skovoroda's works, Ivan Franko's selected novels. Literary works of prominent Ukrainian authors (like T. Shevchenko, I. Franko, L. Ukrayinka, Yu. Andrukhovych, S. Zhadan etc.) are parts of Ukrainian language corpora available online (e.g. http://www.mova.info/, http://uacorpus.org/Kyiv/en). There is still a need for representation of other Ukrainian authors' works in electronic form as part of the national language corpus or as a separate unit for specific educational or scientific research needs. For as far as the Ukrainian language is concerned, at the turn of the XXth century, it was treated as having two variants: so-called western Ukrainian and Naddniprianska Ukraine language (due to the fact of Ukrainian territories being separated and governed by two empires – Austro-Hungarian and Russian). At that time, there was also an explicit discussion between the greatest minds of both parts of Ukraine as to which variant was to be the core of the unified literary language. When studying the diachronic development of the language, it is of vital necessity to have examples of texts which represent various periods and types of language. Among the authors who were representatives of the western-Ukrainian variant Vasyl Stefanyk (1871-1937) was one of the most prominent authors of literary modernism at the time as the author of short stories famous for their strong expressiveness and deepened perception of human existence, though laconic nature.

Literature review

Vasyl Stefanyk's work has been analyzed, translated, and researched by many scholars. This was done by his contemporaries (I. Franko, I. Ogiienko, B. Lepkyi, B. Grinchenko, I. Trush, S. Kryzhanivskyi, etc.) and subsequent generations of scientists and researchers, not only in Ukraine (I. Kovalyk, V. Greshchuk, V. Gnatiuk, etc.), but also abroad, particularly in Canada (T. Kobzey "The great carver of Ukrainian peasant souls"; L. Lutsiv "Vasyl Stefanyk – singer of the Ukrainian land", D. Struk "A Study of Vasyl Stefanyk: The Pain at the Heart of Existence", Y. Vassiyan "Attempt of critical characterization"; K. Kysylevskyi "Nadprutskyi dialect of Stefanyk's characters") (Danchevska, 2013). French-speaking literary criticism and translation, dedicated to V. Stefanyk, have already more than centenary history. In 1912 and 1915 the, French translation of separate short stories were available. Up till now the most substantial publications in French are the book "Croix de pierre" that contained more

than 40 writer's short stories, separate chapters about V. Stefanyk in a 12-volume Belgian anthology "Patrimoine littéraire européen" (1993-2000) and Sarcelles' anthology of the Ukrainian literature of XIXth-XXth centuries (as a Scientific Society named after T. Shevchenko (NTSh) publication in Western Europe (2004)) (Kravets, 2014).

The author's lexicon has been the subject of numerous studies, and as the result, there are "Index to V. Stefanyk's works" (Kovalyk, 1972), dictionary of obscure words by I. Buksa (Buksa, 1996), dictionary of phraseology (1000 phraseological units) as part of T. Yevtushyna's thesis (Yevtushyna, 2005), as well as K. Kysylevskyi's article "Nadprutsky dialect of Stefanyk's characters: Materials for the dictionary" published in Rome in 1973 (Danchevska, 2013). As we can see, Vasyl Stefanyk's literature heritage has its own lexicographical history, but the point is that most of these studies were conducted based on the works published at different times after his death, they were hugely edited and transformed to make them comprehensible for readers unfamiliar with the specific language he used. A lot of these studies were concerned with the peculiarities of the language of the author. V. Stefanyk himself confessed in his letters to publishers and friends that he would like to make the language of his stories understandable to all readers, but then his literary characters would be deprived of their individuality and authenticity (Greschuk, 2010; Greschuk, 2010-2011). Nowadays, advances in modern linguistics allow us to apply new approaches to an integral description of his literary lexicon.

As one of such promising approaches, we consider the ideographic description of the writer's lexicon. The peculiarity of this approach is that it presents the lexical structure of the language by semantic categories of various degrees of generalization and cross-referencing. Based on hierarchical grouping, the lexical-semantic system is consistently divided into thematic groups from broader to narrower categorical meanings (e.g., Universe – Nature – Flora, Fauna, Human, etc.). Correlative groupings within such thematic groups (synonymous, antonymic, and various associative relations) are distinguished. Vocabulary units are grouped according to concepts, so to find the words themselves, their alphabetical indexes are presented. As an example of this approach, we cannot but mention the work conducted by E. Wynalek on narrative poetry "Pan Tadeusz" (http://nevmenandr.net/tadeusz/index.php). The corpus consists not only of the text but also of its three translations into Russian. The aim is the lexical and ideographic explanation of about 20,000 words. Only major word classes are described (nouns, adjectives, verbs, adverbs). This does not include words in languages other than Russian and Polish, as well as barbarisms. E. Wynalek developed her work on the classification of concepts by R. Hallig and W. von Wartburg, slightly revised and adopted to her project needs (see "Search by thematic classification": http://nevmenandr.net/tadeusz/hwru.php). At the top of the classification, there are three main sections: "Universe", "Human", "Human and Universe". At the lowest level, there are tokens that fill this or that classification. The meaning of the word and examples of its use in the text, the depth of its ideographic description are presented. This information appears in a new window. According to the developer, the current version of the dictionary is operational; further improvements lie in developing its database and creation of parallel articles in Russian and Polish for each token. As for Ukrainian authors there is no such work conducted so far. Therefore, we find such research challenging and up to date.

Research methods

The methodological basis of our research is the anthropocentric paradigm, as it reveals the author's picture of the world as a reflection of a nation's picture of the world. Modern approaches and methods of corpus and quantitative linguistics allow a

deeper interpretation of the organization of literary text at all levels. We interpret lexical units of short stories through the prism of a functional approach, the leading features of which include empiricism (corpus data), the use of quantitative methods and interdisciplinarity. Thanks to this approach, we also identify the specifics of the functioning and organization of lexical units in the contextual environment. The nature of the anthropocentric paradigm of modern linguistics and the approaches used in our study determine the choice of relevant research methods.

A corpus-based method is to create the corpus of short stories by Vasyl Stefanyk (the VSC). The task is to choose the so-called canonical texts in Slavic tradition, which means the texts that were either the text of the academic publication of the literary works or the last published edition of the texts during the lifetime of the author. In the case of Vasyl Stefanyk there is an academic edition, but the texts there are either hugely corrected or taken from publications that were out after the death of the author (Pikhmanets, 2016). Therefore, the decision was made to take the anniversary edition of his stories published in 1933 (Stefanyk, 1933) and some that were not included into it, though published during his lifetime ("Mezha" / "The Borderline" (Stefanyk, 1927), "Portret" / "The Portrait" (Stefanyk, 1929), "Shkilnyk" / "The Schoolboy" (Stefanyk, 1932).

The structural method is used to analyze idiolect features at lexical and morphological levels of the language. The method of dictionary definitions is to define the structure of the meaning of lexical units (here Ukrainian Language Dictionary in 11 volumes – ULD-11 (Slovnyk, 1970-1980) and its updated version online – ULD-20 (https://services.ulif.org.ua/expl/Entry/index?wordid=1&page=0) are used). A contextual method is to describe the meanings that are actualized in the text itself (concordance lines from the VSC are retrieved). Statistical method is applied to find out the frequency of selected units; quantitative analysis is to interpret statistical data and to establish repetition, dominance, and correlation of the lexical units; contrastive analysis is to compare the lexical units of Vasyl Stefanyk works with the ones provided in Frequency Dictionary of Modern Ukrainian Fiction (FDMUF) (Chastotnyi, 1981).

Results and Discussion

One of the reasons for the creation of the corpus is the possibility of its further use by other scientists and the exchange of data. Therefore, the primary task is to create a corpus by following generally acknowledged rules and stages. Firstly, the texts of the short stories from the accepted literature sources were transformed into electronic form and processed accordingly. Such normalization of texts specifically for the Ukrainian language is thoroughly explained in I. Kulchytskyi's article "Technical aspects of computer processing of natural texts" (Kulchytskyi, 2015). As the result, the corpus consists of 57 novels, that is 53978 words (12834 word forms).

Secondly, according to the research needs and type of texts, a special type of annotation is required for further text analysis by special computer software for natural language processing. This is especially relevant when working on texts that contain a significant proportion of dialect elements. Such are the texts of Vasyl Stefanyk's short stories.

The result of many years of work of the team of scientists and developers was the creation of a unified standard for annotation of texts called TEI (https://tei-c.org/guidelines/p5/), which was used both to create large corpora and fulfill small research projects (https://tei-c.org/activities/projects/). There are the following main types of annotation (or tagging): external (information about the author and the text), structural (section, paragraph, sentence, direct/indirect speech) and linguistic (morphological, grammatical, semantic, prosodic, syntactic). Tag elements of TEI are applied to the texts of our corpus.

Thus the short stories are structurally defined at the sentence level. The tags <hd> and </hd> are used to identify headings, to indicate the beginning of the paragraph and to indicate the end of the paragraph, dedication is marked by <pr> and </pr>, text parts (or chapters) are distinguished by and respectively.

Another characteristic structural type of literature is the presence of a character's speech and author's narrative. Character's speech in Ukrainian literature can be represented in the text as direct speech, indirect speech, indirect-direct language, etc. For the most part, direct speech is a means of characterization and therefore is of considerable interest to researchers (Bekhta, 2002: 23–30). S. Buk notes that the definition of direct and author's speech "allows to study them in quantitative and qualitative dimensions", and "the quantitative ratio of these layers of vocabulary will differ in each writer, it can be considered a statistical parameter of idiolect" (Buk, 2011a: 205–206). In the example of the corpus of novels by I. Franko, the researcher makes a statistical analysis of these types of narratives and compares its result with the study by V. Perebyinis (Buk, 2011a: 204). Distinguishing the language of the author and the characters (direct speech) is provided in author's dictionaries of K. Chapek (Cermak, 2008) and F. Dostoyevskiy (Slovar, 2002), the distribution of words by author's and character's speech is also found in the FDMUF. While developing our corpus, we pay attention to distinguishing these types of narratives as well.

Direct speech is a means of literal transmission of another's speech with full preservation of lexical, syntactic and intonational features. Direct speech is reproduced and transmitted by means of direct speech, dialogue or quotation, as well as indirect speech. Therefore, we consider the hyphens at the beginning of the speech and before the words of the author, as well as quotation marks, as signs of direct speech. Accordingly, the direct speech is marked with tags <q> and </q> within the structural tagging of the sentences. Dialogues are also marked with tags <q> and </q>. In addition to dialogues, V. Stefanyk's short stories contain examples of the use of monologues and polylogues. In the monologue, only one person speaks (first-person narrative), and there are no words of the author, so we do not use the mark <q>… </q>, unlike in the polylogue.

Indirect speech is another way of transmitting a character's speech. It helps the speaker transmit the statements of another person on his / her behalf by retelling its general meaning. We do not use <q>… </q> in sentences with indirect speech.

In addition to the above-mentioned types of speech, the short stories contain folklore in the form of songs ("She, the Earth", "Evening Hour", "Leaving the City"). They are tagged with <fl> and </fl>.

The distribution of word classes in the VSC is based on the morphological tagging of the corpus. A. Serednytska points out the tendency in modern linguistics to consider word classes as "cognitive-linguistic categories closely related to the process of human cognition of the surrounding reality", and among the common categories for all languages, we distinguish objectivity, action, and features, which are partly represented by different word classes (Serednytska, 2019: 60-63). Such analysis applied to one's literary works, in the cognitive linguistic aspect, is a promising one for building a linguistic picture of the writer's outlook and ideographic description of the language of his / her works and was not applied to literary texts of V. Stefanyk.

The absolute and relative frequencies of word forms and word use by word classes for each of the short stories are calculated (see Table 1).

Part of speech	Word form		Word form use	
	Absolute	Relative	Absolute	Relative
Exclamation	54	0,42 %	338	0,63 %
Participle	244	1,90 %	273	0,51 %
Adverbial participle	32	0,25 %	34	0,06 %
Verb	5557	43,30 %	11043	20,46 %
Pronoun	406	3,16 %	6902	12,79 %
Noun	4397	34,26 %	12597	23,34 %
Preposition	66	0,51 %	5259	9,74 %
Adjective	1316	10,25 %	2633	4,88 %
Adverb	488	3,80 %	3283	6,08 %
Conjunction	77	0,60 %	8237	15,26 %
Particle	65	0,51 %	2843	5,27 %
Numeral	132	1,03 %	536	0,99 %
Total	**12834**	**100,00%**	**53978**	**100,00%**

Table 1. Absolute and relative frequency of word classes in the VSC (by word forms and word forms use)

The frequency of words by word classes in short stories makes it possible to compare these indicators with prose works of other writers or with indicators inherent in Ukrainian fiction (as provided in work "Frequency dictionaries and methods of their use" (Perebyinis, 1985: 157) and based on FDMUF) (see Table 2).

Word classes	**The VSC**	**Ukrainian fiction**
other word classes	31,46 %	26,8 %
noun	23,34 %	25,8 %
verb	20,46 %	18,3 %
adjective	4,88 %	8,6 %
pronoun	12,79 %	11,0 %
adverb	6,08 %	8,4 %
numeral	0,99 %	1,0 %
Total	100 %	100 %

Table 2. Relative frequency of word forms use (by word classes) in the VSC and Ukrainian fiction

It can be concluded that the distribution of words by word classes is quite uniform (except for adjectives), given different time slots and genre specifics of the compared corpora.

Another promising area of literary stylistics is the analysis of the correlation between dialectal vocabulary used in characters and the author's speech. Such studies of Ukrainian literature were made while compiling FDMUF (Ukrainian prose of the mid-twentieth century) and of I. Franko's novels (Buk, 2011a: 202-203). This is the first study of the kind of Vasyl Stefanyk's short stories. Thus, the combination of characters/author's speech and general and dialectal vocabulary units tagging provide quantitative results and the percentage of these types of vocabulary and narrative. Below there are generalized indicators for the VSC (Figure 1).

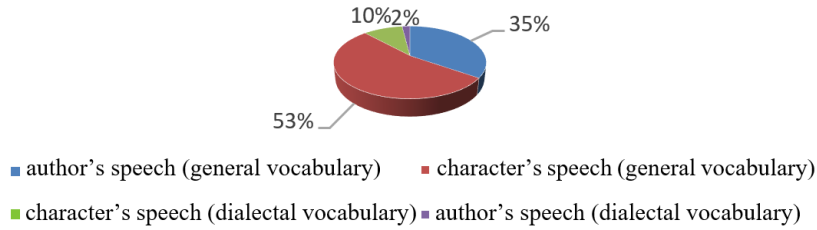

Figure 1. Correlation of author's / character's speech, general and dialectal vocabulary units of the VSC

As we may see, dialectal vocabulary units comprise 2% of the author's speech and 10% of the characters' speech. The ratio of general vocabulary in these types of the narrative is 35% and 53%, respectively. Characters' speech in Vasyl Stefanyk's short stories contains more dialectal vocabulary units, which confirms that it was used by the author to show them as authentic and genuine characters. The diagram below shows the correlation between direct speech and author's speech in general (see Figure 2). These figures indicate a high level of dialogues in the works by V. Stefanyk that is typical for the literary genre he was using – short stories.

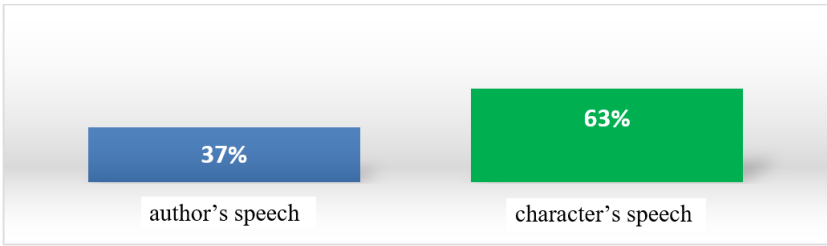

Figure. 2. The ratio of direct character's / author's speech in short stories by Vasyl Stefanyk

The information obtained from the corpus also makes it possible to identify the most frequent words among general language and dialectal vocabulary units and to carry out their comparative analysis.

The most frequent noun among general vocabulary is the word *khata / house* (249), and among the dialect vocabulary it is *gazda / farmer* (68).

According to the FDMUF, the absolute frequency of the word *khata* is almost twice as high – 482. However, in terms of the number of uses of Stefanyk's word *khata* for 500,000 words of the text processed for FDMUF, which is obtained by multiplying its frequency by 9.3, then it is 2315, i.e. the word *khata* in short stories by Vasyl Stefanyk is 4.8 times more frequent than in modern Ukrainian fiction. The dialect word *gazda* in terms of recalculation is 632 against 5, which is 126 times higher. The analysis of the first 40 most frequent words of general language vocabulary in the VSC in comparison with the FDMUF is given in Figure 3.

The striking difference is in the frequency of use of the words *дитина / child* (2167 and 115), *баба / old woman or grandmother* (1386 and 94), *Бог / God* (1516 and 230), *жінка / woman* (1358 and 450), *чоловік / man* (1274 and 354), *мама / mother* (1209 and 227), *піч / stove* (530 and 57), *мужик / bloke* (474 and 34), *церква / church* (456 and 54), *гріх / sin* (456 and 54) and *мамо / father* (428 and 86). These

findings might be useful for comparative genre analysis and the prevalence of topics among authors of different epochs and genres.

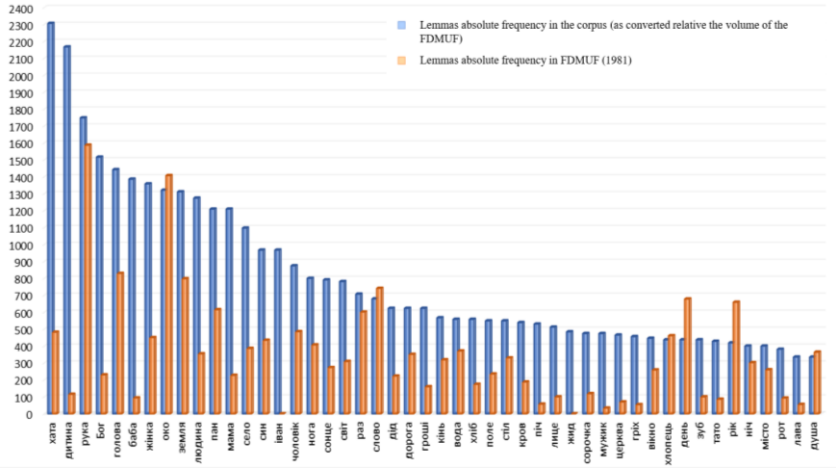

Figure 3. Correlation of 40 most frequent vocabulary units from the VSC and FDMUF

Lexical semantics has attracted much attention in literary studies as well. Defining the leading concepts of author's writing style is possible due to a thorough analysis of his / her lexicon. The concept, a global and universal unit of structured knowledge, highlights its connection with language, thinking, memory and psyche, abstraction, ethnocultural color (Zahnitko, 2010), therefore "the concept is a syncretic phenomenon that has a mental essence and material origin – linguistic expression, is a specific part of ethnocultural information that reflects the world of national perception of objects and concepts denoted by language" (Garbera, 2018: 33). The leading theme of all Vasyl Stefanyk's short stories is household and way of life of peasants, and therefore, the use of vocabulary units denoting *earth* seems natural.

In general, *earth* as a concept in its scope is verbalized in various tokens and contexts and is interesting in the study of the language of short stories by Vasyl Stefanyk, as it has not yet been the subject of a separate study. Firstly, we consider the lemma *earth*.

V. Zhaivoronok in (Zhaivoronok, 2006: 243-244) singles out the following ethnolinguistic features of the *earth*: 1) "the upper layer of the earth's crust, as well as the whole globe as a place of human life and all living things", with the identification of the earth as a mother, fertile and rich, as a sacred and personified entity, a symbol of oath and death; 2) "soil that is cultivated for growing plants; the eternal dream of a peasant is to have his own piece of soil, his own field"; 3) "country, region". All of them are reflected in many epithets, fixed phrases and sayings.

ULD-20 (https://sum20ua.com/Entry/index?wordid=36010&page=1138) provides seven meanings: "1) The third largest planet in the solar system from the Sun, which revolves around its axis and around the Sun // Place of life and human activities; 2) The upper layer of the earth's crust // Earth's surface, on which we walk; 3) Substance of dark brown color, which is part of the earth's crust; 4) Land (as opposed to water); 5) Soil that is cultivated and used for growing plants; 6) Country, region, state; 7) In Germany, Austria – the main administrative unit".

In the VSC the lemma *earth* belongs to the high-frequency vocabulary (143 occurrences) and is present in 38 short stories. Contextual analysis of word usage makes it possible to single out the actualization of the following meanings:

a) place of life and activity of people: *"He (God)* ***will give life*** *(literally: bring to earth), he will not give talent in your hands, nothing in life is a godsend, and still the*

whole world shouts: "Countrymen are thieves, robbers, murderers!" ("Klenovi lystky"/"Maple leaves");
b) the earth's surface, where we walk: *"And all that he threw from that roof* ***to the ground*** *and dragged in front of the house on the sward"* ("Davnyna"/"Antiquity");
c) substance of dark brown color, which is part of the earth's crust: *"If only they threw some* ***soil*** *into the coffin and do not get me togged up"* ("Mezha" / "Boundary");
d) the soil that is cultivated and used for growing plants: *"And I lie down on the field and thank the winds for what they are and* ***the land*** *that it gives birth to everything"* ("Mezha" / "Boundary").
e) country, region, state: *"It's none of my business, but why did you take your carts, raven horses, little children and leave your* ***land****?"* (She the Earth: 170).
We also find examples of personification and antropomorphising of the earth: *"It seemed that* ***the earth complained*** *about its wounds"* ("Mariya" / "Mary"); *"… As rye* ***begged*** *for a sickle so did* ***the earth****: "Come, Fedor, and take bread from me""* ("Paliy" / "The Firestarter").
Through the analogy of "earth is grave" the adjective *damp* acquires a sacred meaning, i.e. reflects the folk poetic tradition of the earth as the last refuge of man (which may indicate its belonging to the semantic field of the word *death*): *"… That war put many of them* ***beneath the sod****"* ("Did Hryts" / "Grandfather Hryts"); *"Put your head on the threshold, he said, and I'll cut it down: you will* ***lie beneath the sod****, I will be sent to the gallows, and the children will carry water for the Jews!"* ("Klenovi lystky" / "Maple leaves").
Other ways of verbalizing the semantics of the word *death* in Ukrainian are expressions like *pity v zemlyu, zaporpaty v zemlyu, hnyty v zemli / to go down to the ground, to dig into the ground, to rot in the ground*: *"You die and don't care, hence you are* ***rotting in the ground****, do you?'* ("Katrusia" / "Kathy")
Another component of the concept of land is the lemma *field* (from the corpus we get 64 usages) as a "forestless plain, flat large space… // Plot of land used for crops" [ULD 11]: *"The old man took care around the house and the oxen, the old woman took care of the household, and the sons sowed* ***the field****"* ("Davnyna" / "Antiquity"); *"Either when dragging sheaves from* ***the field*** *or bringing manure to* ***the field****, the veins of the horses and Ivan's bulged…"* ("Kaminnyi Khrest" / "A Stone Cross"); *"Behind the forest they stood in* ***the field****"* ("Vyvodyly z sela" / "Taken out of the village").
The lemmas *pole* and *nyva* are connected with the concept earth by a synonymous relation as "a plot or strip of land on which grain crops grow or which is intended for cultivation; field" [ULD 11]) (34 usages retrieved): *"My* ***fields*** *are like well-fed sheep, black and curly"*("Vona zemlia" / "She, the Earth"); *"He plows* ***the field*** *and can't hold the stick with his hands, because thirst burns in his throat"* ("Skin" / "Dying").
According to the same semantic feature, the lemma *pole* / *field* enters into synonymous relations with the token *lan* as "a forestless and spacious plain; a plot of arable land that has certain boundaries" [ULD 11]) (11 word forms found in the VSC). Also, in ULD-11 we find *lan* defined as obsolete: "a plot of arable land within 10-30 acres as a measure of the land": *"I mow your* ***fields*** *and forget not only about the children, but don't remember myself"* ("Klenovi lystky" / "Maple leaves") *"A lot of the sun in the lot of limitless* ***fields****"* ("Firestarter" / "Paliy").
In the meaning of "a yard around the house with all the adjacent buildings and lands "[UDL 11] we find the use of the lemma *grunt* / *ground* (12 usages): *"Do not kiss, do not lick the hands of gentlemen, because you are a lady, you are a better lady than a landlady, because you have your own* ***ground****"* ("Takyi panok" / "Such a master").
We also record in the VSC the usage of the lemma *grunt* as part of a phrase meaning "to perform a funeral rite" – *klasty na grunt* (which also indicates its belonging to the

semantic field of the word *death*): *"Grandfather Dmytro **buried** (poklav na grunt) his four sons"* ("Davnyna" / "Antiquity").
The lemma of *rillia* / *cultivated land* (6 usages) as "plowed field" and "plowed layer of soil" [ULD 11] is also referred to the concept of land: *"Harrows flew on the ground like feathers. Maxim threw his hat on **the plowed field**…"*("Syny" / "Sons").
The lemma *lis / forest* (28 word usages in the VSC) according to its first meaning in ULD 11 as "a large area of land overgrown with trees and bushes" belongs to the concept of land and is used in the works of Vasyl Stefanyk when describing the surrounding nature: *" - Man, look at the fellows and the village and **the forest** and come to your senses"* ("Sud" / "Court"); *"Vasylko, take Nastia and lead her to her uncle; over there, through the path along **the forest**, you know "*("Dytiacha pryhoda" / "Children's Adventure").
The word *doroha / road* can also be attributed to the concept of land as "the stripe of land on which one rides and walks" [UDL 11]). It should be noted that the lemma *doroha / road* in this sense is used in short stories in about half of the occurrences (22 out of 58 usages): *"Fences along **the road** cracked and fell – all the people were seeing Ivan off"*("Kaminnyi Khrest" / "A Stone Cross").
Thus, the analysis of the components of the concept of land in the context of reflecting everyday life of peasants showed that most of them are nominations of arable land, which contain a synonymous series of words (*pole, lan, rillia*) and obsolete vocabulary (*grunt*). A small part are nominations of natural objects denoting *forest / lis, road* / *doroha*, which correlate with the concept of land partially and are rarely used in short stories in this sense. In general, we can conclude that such study is promising for the compilation of an ideographic dictionary of the language of short stories and can be applied for other lexical units in the corpus.

Conclusion

The multidisciplinary approach to studying any writer's lexicon results in works on author's lexicography, text linguistics, corpus linguistics, research of the writer's idiolect, the conceptual and linguistic outlook of an individual, dialectal vocabulary units in literary works, and statistical research.
In accordance with world practice and the basic principles of creating text corpora, methodologically, the algorithm for constructing a corpus of short stories by Vasyl Stefanyk was carried out in several stages: the creation of an electronic database of short stories published during the writer's lifetime; standardization and marking the texts both structurally and morphologically. According to its characteristics, the created corpus is a reference, synchronous, static, specific, full text, written, monolingual and partially marked one. Since corpus technology allows any text to be displayed in a concordance format, which has the completeness of the thesaurus type by considering absolutely all word usages, this feature is leading in the study of literary texts.
The VSC-based studies aimed to compare the obtained statistical facts with the relevant indicators of modern Ukrainian fiction. It is established that the vocabulary of short stories contains differences in terms of dialectal word usage, i.e. the usage of some tokens are absent in modern fiction, or their frequency of usage is much lower.
A few other statistical studies have been carried out, and their results have been analyzed. In particular, the frequency of usage of word classes for all short stories was studied. It is determined that among the words that belong to major word classes, verbs, nouns, and adjectives have the largest numbers of usage in both groups.
A comparison of quantitative information by word classes with the same of Ukrainian fiction showed an even distribution.
The percentage of author's and character's speech indicates a high percentage of dialogues in short stories that is characteristic of the short stories genre.

Determining separately, the share of dialectal vocabulary in the direct speech of the author and the characters testifies to more frequent use of the dialectal language units in the direct speech of the characters. This showcases the author's strived to depict his characters as vividly and genuine as possible by providing them with the speech that was characteristic of the inhabitants of the described region.

According to research, the methods of corpus-based and quantitative analysis allow us to confirm or clarify existing ideas about the style of the author. Further development of corpus research (including the creation of new corpora of works by Ukrainian writers) will allow more diverse research and comparative studies.

We consider further research in a thorough analysis of Vasyl Stefanyk's short stories lexicon at the grammatical and semantic levels, as well as expanding the source base with corpora of the epistolary heritage of the novelist, parallel corpora of various editions and translations of his works as well.

Bibliographic references

Andrushenko, O., (2021). Corpus-based studies of Middle English adverb largely: syntax and information-structure. XLinguae, 14(2). DOI: 10.18355/XL.2021.14.02.05 (in English)

Baker, P., (2006). Using Corpora in Discourse Analysis. A&C Black.

Baker, P., & Mcenery, A., (2015). Corpora and Discourse Studies: Integrating Discourse and Corpora. Springer.

Bekhta, I., (2002). Personazhnyi dyskurs u naratyvnii strukturi khudozhnoho tekstu [Character's discourse in literary text structure]. Mova i kultura. Seriya Filolohiya. 5 (4). Kyiv, 23-34.

Berkeshchuk, I., (2018). Kartyna svitu – osnova svitohliadu [Picture of the world as the basis of outlook]. Naukovi pratsi Kamianets-Podilskoho natsionalnoho universytetu imeni Ivana Ohiienka. Filolohichni nauky, 47, 64-67.

Biber, D., (2011). Corpus linguistics and the scientific study of literature: back to the future? Scientific Study of Literature, 1 (1), 15-23. https://doi.org/10.1075/ssol.1.1.02bib

Buk, S., (2011a). Priama y avtorska mova velykoyi prozy Ivana Franka: linhvostatystychne doslidzhennia u konteksti korpusnoyi linhvistyky [Direct and indirect speech in Ivan Franko's prose: statistical corpus-based research]. Visnyk Lvivskoho universytetu. Seriia filolohichna, 52 . Movoznavstvo. 199-209.

Buk, S., (2011b). Slovianskyi dosvid ukladannia chastotnykh slovnykiv movy pysmennyka [Slavic experience of author's dictionary compilation]. Problemy slovianoznavstva. 60, 217-224.

Buksa, I., (1996). Tvorchist ta slovnyk malozrozumilykh sliv V. Stefanyka [Literature and unfamiliar words dictionary of V. Stefanyk]. Kyiv: Smoloskyp.

Cermak, F., (2008). An Author's Dictionary: The Case of Karel Čapek. In: E. Bernal, J. DeCesaris (Ed.), Proceedings of the XIII EURALEX International Congress. Barcelona: Institut Universitari de Lingüística Aplicada. Universitat Pompeu Fabra; Documenta Universitaria.

Chastotnyi slovnyk suchasnoyi ukrayinskoyi khudozhnioyi prozy : U 2 T. [Frequency Dictionary of Modern Ukrainian Fiction - FDMUF] za red. Perebyinis V. S. Kyiv, 1981.

Coulthard, M., (2004). Author Identification, idiolect, and linguistic uniqueness. Applied Linguistics, 24(4). 431-447.

Danchevska, Yu., (2013). Mova tvoriv V. S. Stefanyka: suchasnyi stan ta perspektyvy doslidzhennia [Language of V. Stefanyk's works: current status and promising research directions]. In: Liudyna. Kompiuter. Komunikatsiia: Zbirnyk naukovykh prats za red. O. P. Levchenko. Lviv: Vydavnytstvo Lvivskoi politekhniky, 44-46.

Demska, O., (2011). Tekstovyi korpus: ideia inshoi formy [Text corpus: an idea of the other shape]. Kyiv: VPTs NaUKMA.
Demska–Kulchytska, O., (2005). Osnovy natsionalnoho korpusu ukrainskoi movy : monohrafiia. [Foundations of National Corpus of the Ukrainian Language. A monograph] Kyiv.
Garbera, I. (2018). Movnoarealne pole kontseptu liudyna: frazeokodovyi riven i linhvokompiuterne modeliuvannia: monohrafiia [Areal features of the concept HUMAN BEING: phraseme-coding and computer linguistics modelling: a monograph]. Vinnytsia: TOV „Nilan-LTD".
Garside, R., Leech, G., & Mcenery, T., (2016). Corpus Annotation: Linguistic Information from Computer Text Corpora. London: Routledge.
Greshchuk, V., (2010). Pivdenno-zakhidni dialekty v ukrainskii khudozhnii movi. Narys [South-western dialects in Ukrainian language. A sketch]. Ivano-Frankivsk: Vyd-vo Prykarpat. nats. un-tu imeni Vasylia Stefanyka.
Greshchuk, V., (2010-2011). Dialektne slovo v khudozhnii movi [A dialectal word in literature]. Ukrainoznavchi studiyi, 11-12, 3-11.
Ivanishcheva, O., (2016). Culture phenomena: lexicographical description issues. XLinguae Journal, 9(2). doi: 10.18355/XL.2016.09.02.73-89
Jones, S., (2012). When Computers Read: Literary Analysis And Digital Technology. Bulletin of the American Society for Information Science and Technology, 38(4), 27-30. https://doi.org/10.1002/bult.2012.1720380408
Kononenko, V., (2008). Mova u konteksti kultury : monohrafiia [Language in the context of culture: a monograph]. Kyiv, Ivano-Frankivsk.
Kostetska, O., (2014). Indyvidualne movlennia avtora yak obiekt linhvistyky ta pidkhody do yoho doslidzhennia [Author's speech as linguistic object and approaches to its study]. Naukovi zapysky Natsionalnoho universytetu „Ostrozka akademiya", 49,. Ostroh, 196-199.
Kovalyk, I., & Oshchypko, I., (1972). Khudozhnie slovo V. Stefanyka. Materialy dlia slovopokazhchyka do novel V. Stefanyka. Metodychnyi posibnyk [A word of literature by V. Stefanyk. Materials for the word index. Methodology.]. Lviv: Vyd-vo Lvivskoho universytetu.
Kravets, YA., (2014). Vasyl Stefanyk u frantsuzkomovnomu prochytanni [V. Stefanyk in French-speaking world]. Sultanivski chytannia. Aktualni problemy literaturoznavstva v komparatyvnykh vymirakh: Zbirnyk statei. Ivano-Frankivsk: Symfoniia forte. III, 128-140.
Kulbabska, O., & Shatilova, V., (2016). „Pyshu, yak sertse dyktuye..." (Idiostyl Sydora Vorobkevycha) : monohrafiya [„I am writing as my heart is dictating…" (Sydir Vorobkevych's idiolect): a monograph]. Chernivtsi: Chernivetskyi nats. un-t.
Kulchytskyi, I., (2015). Tekhnolohichni aspekty ukladannia korpusiv tekstiv[Technologiical aspects of text corpus compilation]. Dani tekstovykh korpusiv u linhvistychnykh doslidzhenniakh: monohrafiia za red. O. Levchenko. Lviv: Vydavnytstvo Lvivskoi politekhniky, 29-45.
Kulchytskyi, I., (2020). Unormovuvannia tekstu pry dokorpusnomu opratsiuvanni: dosvid zastosuvannia [Text normalization during pre-corpus preparation: experience of applicatio] Visnyk Natsionalnoho universytetu Lvivska politekhnika. Informatsiini systemy ta merezhi. Vydavnytstvo Lvivskoi politekhniky. Lviv. 7, 51-58. https://doi.org/10.23939/sisn2020.07.051
Levchenko, O., Tyshchenko, O., & Dilai M., (2020). Associative Verbal Network of the Conceptual Domain БІДА (MISERY) in Ukrainian. CEUR Workshop Proceedings. 2604, 106-120.
Louw, B., (2013). The Role of Corpora in Critical Literary Appreciation. Teaching and Language Corpora. Routledge. https://doi.org/10.4324/9781315842677

Marchuk, L., (2012). Aktsentna leksyka tekstiv Hryhoriia Bilousa [Obscene language of texts of H. Bilous]. Naukovi pratsi Kamianets-Podilskoho natsionalnoho universytetu imeni Ivana Ohiienka. Filolohichni nauky, 29(1), 57-62.
Mcenery, T., & Hardie, A. (2012). Corpus Linguistics: Method, Theory and Practice. Cambridge University Press.
Mcenery, T., Xiao, R., & Tono Y. (2006). Corpus-based Language Studies: an Advanced Resource Book. London: Routledge.
Mishenina, T., & Dsevickaya, L. (2017). Correlation linguocultureme human within east Russian, Ukrainian and Belarusian languages: comparative analysis. XLinguae, 10(4). doi: 10.18355/XL.2017.10.04.25
O'Keeffe, A., MCCarthy, M., & Carter, R., (2007). From Corpus to Classroom: language use and language teaching. Cambridge University Press.
Pawlowski, A., (2006). Zwiazki lingwistyki i matematyki z perspektywy filologicznej na przykladzie prac Jana Czekanowskiego i Jerzego Woronczaka [Relations between mathematics and linguistics on example of works by Yana Chekhanovskogo i Yerzhego Voronchaka]. Rozprawy Komisji Jezykowej, 33, 297-304.
Perebyinis, V., Muravytska, M., & Darchuk N., (1985). Chastotni slovnyky ta yikh vykorystannia [Frequency dictionaries and methods of their use]. Kyiv: Naukova dumka, 1985.
Pikhmanets, R., (2016). Problemy naukovoho vydannia khudozhnoi spadshchyny Vasylia Stefanyka [Problems of scientific compilation of literary works of V. Stefanyk]. Prykarpatskyi visnyk NTSh. Slovo, 122-133.
Rabus, A., & Shvedova, M. (2021). Morphological variation in ukrainian regional varieties: A corpus study. Slavia. 90(1), 1–24.
Romanchenko, A., (2015). Movna osobystist yak interdystsyplinarnyi obiekt doslidzhen [Language personality as an interdisciplinary object of study]. Naukovi zapysky Natsionalnoho universytetu „Ostrozka akademiia", 58. Seriia: Filolohichna, 117-119.
Rovenchak, A., & Buk, S., (2018). Part-of-Speech Sequences in Literary Text: Evidence From Ukrainian. Journal of Quantitative Linguistics, 25(1), 1-21, doi: 10.1080/09296174.2017.1324601 (in English)
Selihei, P., (2009). Struktura y typolohiya movnoyi svidomosti [Structure and typology of language conscience]. Movoznavstvo, 5. 12-29. doi 10.33190/0027-2833
Serednytska, A. (2019). Rol chastyn movy u movnii kartyni svitu [Word classes role in language picture of the world]. Naukovyi visnyk Mizhnarodnoho humanitarnoho universytetu. Seriia: „Filolohiia". Odesa, 60-63.
Shyrokov, V., Buhakov, O., Hriaznukhina, T., & Et al., (2005). Korpusna linhvistyka: monohrafiia [Corpus Linguistics: a monograph]. Ukrainskyi movno-informatsiinyi fond NAN Ukrainy. Kyiv. Dovira.
Sinclair, J., (1991). Corpus, Concordance, Collocation. Oxford: Oxford University Press.
Slovar Yazyka Dostoyevskoho. Leksicheskij Stroi Idiolekta. Vyp. 2. (2002). [Dictionary of Dostoyevskyi's language] Glavnyi redaktor chlen-korrespondent RAN Yu. N. Karaulov. Rossiyskaya akademiya nauk, Institut russkoho yazyka im. V. V. Vinogradova. Moskva: Azbukovnik.
Sovtys, N. (2013). Khudozhnii tekst yak vidobrazhennia movnoi kartyny svitu [Literary text as a reflection of language picture of the world]. Kyivski polonistychni studii, 22, 476-479.
Struhanets, L., (2012). Poniattia „movna osobystist" v ukrainistytsi [The notion of "language personality" in Ukrainian linguistics]. Kultura slova, 77, 127-133
Stubbs, M., (2014). Quantitative Methods in Literary Linguistics. In: P. Stockwell & S. Whiteley. (Ed.). The Cambridge Handbook of Stylistics. Cambridge: Cambridge University Press, 46-62.

Stubbs, M., (1996). Text and corpus analysis: computer-assisted studies of language and culture. Oxford: Blackwell.
Slovnyk ukrayinskoyi movy: V 11 TT. [Ukrainian Language Dictionary: 11 volumes – ULD-11] Akademiya Nauk URSR. Instytut movoznavstva; za red. I. K. Bilodida. K.: Naukova dumka. 1970-1980.
Stefanyk, V., (1927). Mezha [Boundary]. „Literaturno-naukovyi visnyk", t. 92, kn. 2. Lviv, 97-98.
Stefanyk, V., (1929). Tvory [Literary works] (peredmova V. Koriaka. Do druku vyhotuvav Iv. Lyzanivskyi. 3-ye vyd.) DVU, 94-95.
Stefanyk, V., (1933). Tvory [Literary works]. Vasyl Stefanyk; z derevorytamy V. Kasiiana i M. Butovycha. Lviv: Z drukarni Vydavnychoi Spilky „Dilo".
Stefanyk, V., (1932). Shkilnyk [Schoolboy]. „Ridna shkola" Lviv, 1(2-4).
Svartvik, J., (2007). Corpus linguistics 25+ years on. In: Language and computers: studies in practical linguistics. Amsterdam, New York, 11-27.
Teubert, W., (2007). Corpus linguistics and lexicography. In: W. Teubert (Ed.). Text Corpora and Multilingual Lexicography. John Benjamins Publishing Company. Amsterdam / Philadelphia, 109-134. https://doi.org/10.1075/bct.8
Yevtushyna, T., (2005). Linhvostylistychnyi potentsial frazeolohii u tvorakh V. Stefanyka [Linguistic and stylistic potential of phraseology in V. Stefanyk's works]. Dys. … kand. fil. nauk. Kyiv.
Zahnitko, A., (2010). Klasyfikatsiini typolohii kontseptiv [Classificational typology of concepts]. Linhvistychni studii: zb. naukovykh prats. Donetsk: Don NU. 21, 12-21.
Zhaivoronok, V., (2006). Znaky ukrainskoi etnokultury. Slovnyk dovidnyk. [Signs of Ukrainian ethnoculture. Reference dictionary] Kyiv: Dovira.

Words: 7743
Characters: 51 636 (28,71 standard pages)

Yuliia Kalymon, PhD
Department of Ukrainian and Foreign languages
Lviv State University of Physical Culture named after Ivan Boberskyi
11, Kostiushko Str.
79000 Lviv,
Ukraine
kalymon.yulia@gmail.com

Prof. Olha Romanchuk, Doctor of Science in Pedagogy
Head of Department of Ukrainian and Foreign languages
Lviv State University of Physical Culture named after Ivan Boberskyi
11, Kostiushko Str.
79000 Lviv,
Ukraine
slang@ldufk.edu.ua

Prof. Nadiya Fedchyshyn, Doctor of Science in Pedagogy
Head of The Department of Foreign Languages
Ternopil State Medical University named after Ivan Horbachevskyi
1, Voli Square
46000 Ternopil
Ukraine
fedushunno@tdmu.edu.ua

Assoc. Prof. Ulyana Protsenko, PhD
Department of Ukrainian and Foreign languages
Lviv State University of Physical Culture named after Ivan Boberskyi
11, Kostiushko Str.
79000 Lviv,
Ukraine
new@ldufk.edu.ua

Nadiia Yurko
Department of Ukrainian and Foreign languages
Lviv State University of Physical Culture named after Ivan Boberskyi
11, Kostiushko Str.
79000 Lviv,
Ukraine
nau40279@gmail.com