UDC [811.161.2+811.162.1+811.111]:81'33
DOI https://doi.org/10.32447/2663-340X-2022-11.23

CORPUS-BASED APPROACH IN THE STUDY OF VERBAL PREDICATES

Nazarchuk Roksolana Zinoviivna
Candidate of Philological Sciences,
Senior Lecturer at the Department of Applied Linguistics
Lviv Polytechnic National University
12, Bandery Str., Lviv, Ukraine

Karamysheva Iryna Damirivna
Candidate of Philological Sciences,
Associate Professor at the Department of Applied Linguistics
Lviv Polytechnic National University
12, Bandery Str., Lviv, Ukraine

The article highlights the features of the corpus approach to the analysis of verbal predicates of Ukrainian, Polish and English languages. The problems of corpus linguistics are outlined, and the corpora of texts are reviewed, namely the General Regionally Annotated Corpus of Ukrainian (GRAC), the National Polish Corpus (NKJP), and the British National Corpus (BNC). These resources, meeting the requirements of authenticity, representativeness (balance), and sufficiency in volume (S. Buk, O. Demska, Ch. Fellbaum, S. Gries, A. Stefanowitsch M. Shvedova, V. Shyrokov and others), allow to optimize and to objectify the interpretation of language material as well as to obtain high reliability of results. The scientific novelty of the study is a comprehensive comparative analysis of the functional capabilities of the corpora of texts in Ukrainian, Polish, and English to determine their features, as well as outlining the prospects of applying a corpus-based approach to identify object connections of verbal predicates. The basic concepts of corpus linguistics are introduced: lemma, word form, token, label (tag), co-usage (collocation), concordance, frequency, and CQL query; they are illustrated by the implementations of the corresponding referentially specialized verbal predicates and their objects based on GRAC, NKJP, and BNC. In particular, it is shown that the simplest use of text corpora is to certify the use of a unit (concordance) and its frequency (i.e. the number of repetitions in the corpus). For languages rich in word change, it is important to be able to identify different forms of one word with its basic form − the lemma. The presence of a tag (POS, part of speech) allows acquiring the co-usage (collocations) of two words. It has been illustrated how the studied corpora allow further grouping, frequency ordering, and isolation of specific word combinations, and the markup of parts of speech facilitates the search for all collocations of verbal predicates with potential object-nouns at a distance of 4 tokens. The function of searching for co-usage (collocations) of verbs, grouped by lemmas or grammatical characteristics and ordered by frequency, is described. The advantages of semantic markup in the GRAC corpus for the automatic search of language units and their combinations are pointed out. All the described features of the text corpora are accompanied by illustrations from GRAC, NKJP, and BNC.

Key words: *corpus linguistics, corpus, GRAC, NKJP, BNC, verb, object, lemma, collocation*

Introduction. Linguistic studies using corpora, i.e. large collections of original texts, are gaining popularity along with the development of information technology that allows automatic search and analysis of language units and their combinations.

Characterizing the peculiarities of relations between a verbal predicate and the object, we consider it appropriate to use a corpus-based approach, which, according to researchers, ensures the objectivity of results and prevents introspective interpretation: "the standard procedures for accessing corpora *(concordances, collocate lists, frequency lists)* are a natural step towards identifying the relevant distributions in the first place" (Stefanowitsch, 2020, p. 59).

Problem setting. Following the works (Buk, 2001, p. 62–65; Demska, 2011, p. 83–89; Fellbaum, 2015, 2019; Gries, 2006; Stefanowitsch, 2020; Shvedova, 2020; Shyrokov, 2005, p. 12–13), text corpus means a set of language or speech data described linguistically competently, presented in the electronic form and fitted with appropriate specialized software, intended for a variety of studies, that meets the requirements of authenticity, representativeness (balance), sufficiency in volume. As O. Demska rightly points out, corpus "is formed from real fragments of written or spoken speech, without providing for the modification of speech reality, which turns it into an empirical category and allows considering the actual corpus

material as an empirical basis for linguistic study" (Demska, 2011, p. 41).

The authenticity provides for the involvement of texts of artistic, scientific, popular scientific works, periodicals, transcription of radio and television programs, etc. The Representativeness (or balance) of the collection means the reproduction of quantitative and qualitative diversity of areas of the real use of a particular language; for example, preserving the relative share of artistic and scientific texts. How big the corpus should depend on the language; the existing corpora of the main European languages range from hundreds of millions to over a billion tokens (along with a lemma as an initial form of a word, corpus linguistics uses the broader concept of a token as the smallest unit into which the corpus is divided, with any quantitative characteristic in a text; tokens are any sequence of characters between spaces or other separators: word form, number, punctuation mark, symbol (smiley, mathematical symbol, etc.). It should be noted that the mechanical increase in the size of corpora today does not require extraordinary efforts; a much bigger challenge is to maintain representativeness.

Analysis Procedure. The General Regionally Annotated Corpus of Ukrainian (GRAC) is publicly available and contains over 600 million words. The National Corpus of Polish (NKJP) comprises about 1.5 billion words, and the British National Corpus (BNC) contains about 100 million words. These corpora were created by research institutions of the respective countries, and they meet the already mentioned requirements.

An important point for the efficient use of the corpus of any language in linguistic studies is the presence of marking. Usually, corpus compilers keep relevant bibliographic data (e.g. author's name, title, edition, year, page, etc.) for each fragment of a text, and add some grammatical information to each word (e.g., belonging to a part of speech, gender, case, etc.). The branching of the marking affects the qualitative opportunities of studies based on the corpus data. Thus, the semantic marking in the GRAC corpus of Ukrainian makes it possible to distinguish between abstract and common nouns denoting living beings or non-living beings, etc. The simplest use of language corpora is to certify the use

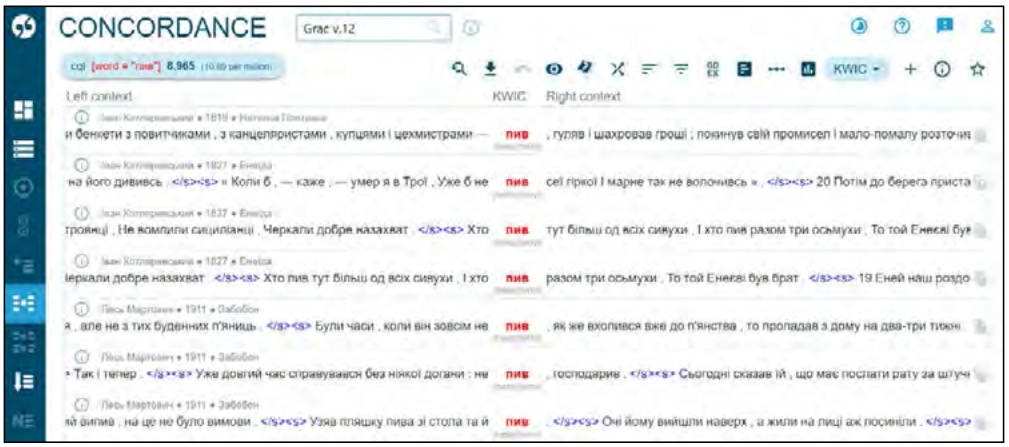

Figure 1. Examples of the use of the word form *пив* in the GRAC corpus of Ukrainian

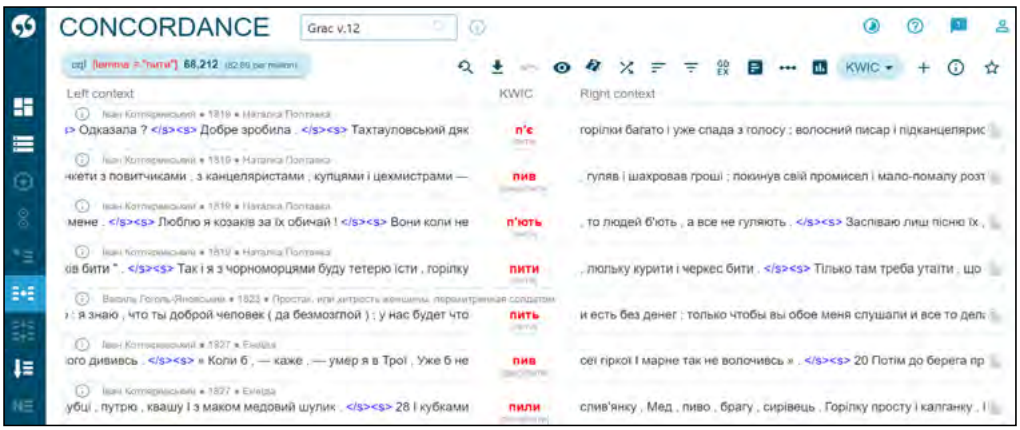

Figure 2. Examples of the use of the verb *пити* in all its forms in the GRAC corpus

of a unit (*concordance*) and its frequency (i.e., the number of repetitions in the corpus). For a balanced corpus, the relative frequency $v_w = n_w/N$ of the appearance of a particular word w in it (i.e., the ratio of the frequency n_w to the total number of words N) corresponds to its relative frequency in real language use; thus, a higher frequency of a unit w_1 than w_2 is an indicator that the unit w_1 is more commonly used in real language than the unit w_2. Thus, the query [word = "пив"] finds in the GRAC corpus of Ukrainian 8,965 cases of the use of the word form *пив* (see Fig. 1). Given

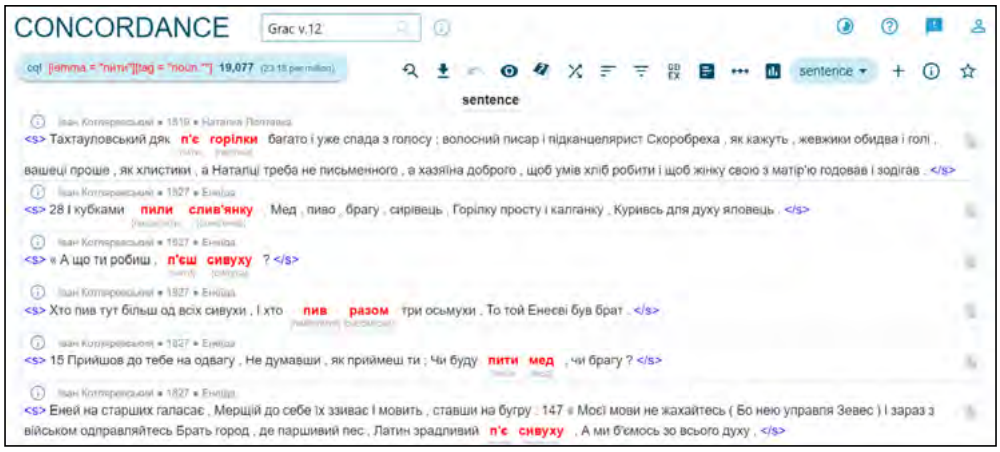

Figure 3. Search for combinations of nouns with the verb *пити* in the GRAC corpus

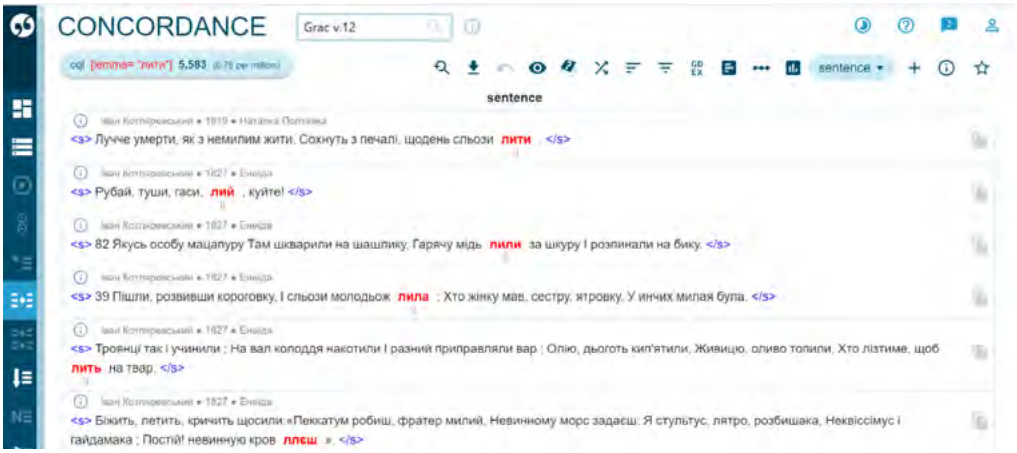

Figure 4. Examples of the use of the lemma *лити* in the GRAC corpus using the SketchEngine search server

Figure 5. Frequency of the use of different forms of the lemma *лити*

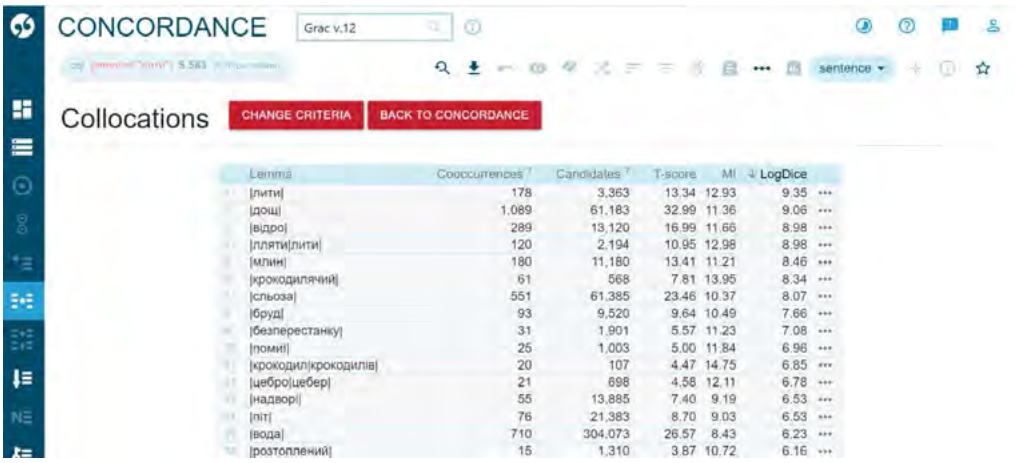

Figure 6. Collocations of the lemma *лити* at a distance of at most 4 tokens

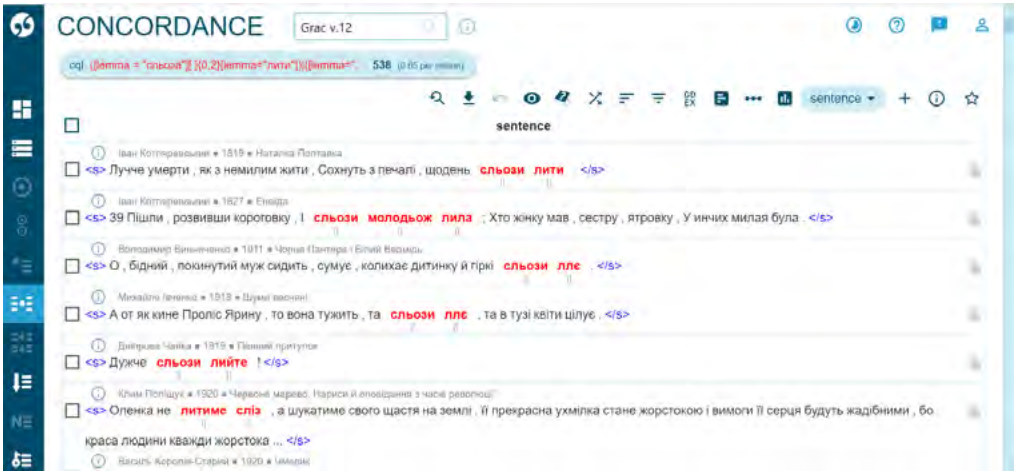

Figure 7. Collocations of the lemmas *лити* and *сльози* at a distance of at most 2 tokens

Figure 8. The search result for the lemma *lać* in the NKJP corpus

the volume of the GRAC corpus, we obtain a frequency of 10.89 per million of tokens.

For languages rich in inflection, it is important to be able to identify different forms of a word with its basic form – *lemma* (for example, the word forms *пив*, *п'ють*, etc. with the verb *numu* in the infinitive form). The availability of appropriate marking makes it possible to take into account in one search query with a given lemma also all the uses of its various derived forms. Thus, the result of the query [lemma = "пити"] in the GRAC will be 68,212 examples with the verb *numu* in all its forms (see Fig. 2).

A tag (POS, part of speech) mark makes it possible to obtain collocations of two words. For example, to identify all nouns that occur after the verb пити, you can use the search [lemma = "пити"] [tag = "noun.*"] (the result is 19,077 examples of the use in the GRAC (see Fig. 3). The corpus allows further grouping, frequency ordering and allocation of specific phrases.

Figure 9. The search results for collocations of the lemma *lać* with nouns

Figure 10. The search for collocations of the lemma *drink* with nouns at a distance of up to 4 tokens

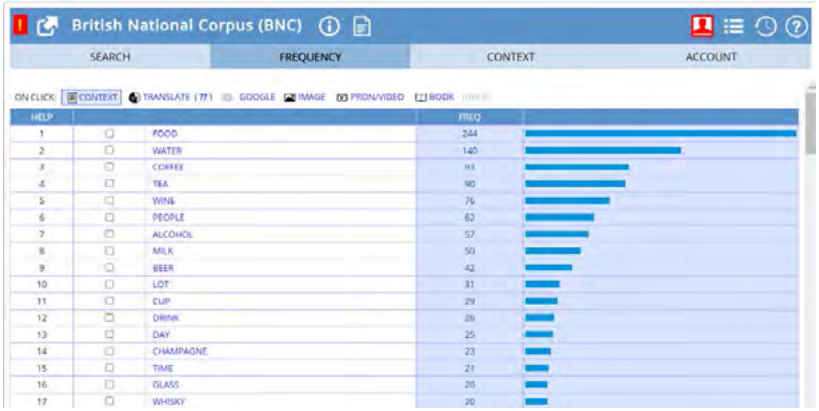

Figure 11. The list of noun collocations of the lemma *drink* sorted by frequency

Figure 12. The list of collocations of the lemmas *drink* and *coffee*

A. Stefanowitsch analyses the case (Stefanowitsch, 2020, p. 47–48) when the study of the use of the adjective *implacable* in the 450 million corpus of contemporary American COCA proves its appearance mainly in a specific context, which implies rivalry between people or even their enmity. This fact is not reflected in the illustrative materials of the Merriam-Webster dictionary, which leads to misinterpretation of the connotations of this adjective.

The greatest opportunities for researchers are offered by corpora with flexible queries using the CQL (Corpus Query Language). For example, you can search for all potential objects of the verb *ламати* in the GRAC corpus using the query ([tag = «noun.*»][]{0,3}[lemma= «ламати»])|([lemma = «ламати»][]{0,3}[tag = «noun.*»]) within <s/>. The result is a list of uses of the verb *ламати* in all its forms (lemma= «ламати») together with a noun ([tag = «noun.*»]) in one sentence (within <s/>), between which there may be up to three other words ([]{0,3}). For example: В світлому отворі дверей миготіли постаті, і деякі, переступаючи в захваті поріг, ламали тут священний затишок різким човганням взуття.

The detailed characteristics of the analysed text corpora are presented below.

The GRAC is the largest corpus of the Ukrainian language, publicly available and convenient for conducting both general and specialized linguistic studies. The GRAC contains rich grammatical and semantic marking, is integrated into the SketchEngine [https://parasol.vmguest.uni-jena.de/grac_crystal/#dashboard?corpname=grac12] search system, which allows advanced searching with full use of marking using the CQL.

The total volume of version 12 of the corpus published in 2021 is 822,959,896 tokens, 640,932,211 words (7,591,461 unique words total), among which there are 2,693,775 lemmas; the GRAC contains 50,803,520 sentences collected from 97,245 documents and is the largest digitized base of the Ukrainian language.

The SketchEngine search server allows searching for the number of uses of a word form or lemma in the corpus, as well as finding their collocations (common uses) with other units. For example, the search for the lemma *лити* gives 5,583 occurrences (see Fig. 4). If necessary, you can view the frequency of different forms of the basic lemma (see Fig. 5).

For further analysis, the server has the function of finding collocations, grouped by lemmas or other grammatical characteristics and ordered by frequency (see Fig. 6).

To identify special common uses, the SketchEngine allows complex searches using the CQL; for example, the query ([lemma = «сльоза»][ ]{0,2}[lemma=«лити»])|([lemma=«лити»][ ]{0,2}[lemma=«сьоза»]) within <s/> starts a search for collocations of the verb *лити* and the noun *сльоза* in their various forms; both units are present in the same sentence (within <s/>), and between them there may be up to two other words (see Fig. 7).

The NKJP is the corpus of Polish created in cooperation with the Institute of Fundamentals of Informatics of the PAS (the Polish Academy of Sciences), the Institute of the Polish Language of the PAS, the PWN Scientific Publishing House of the PAS, and the University of Lodz with the support of the Ministry of Science and Education of Poland.

To work with the corpus, the PELCRA search engine was used, which shows the frequency of uses of a certain word form or lemma with subsequent analysis of collocations. The NKJP contains a balanced part (about 300 million words) as well as a complete part (about 1.5 billion words). For example, the search result for the lemma *lać* is presented in Fig. 8.

The PELCRA system makes it possible to search for collocations (for example, Fig. 9 shows the search results for all the collocations of the lemma *lać* with nouns immediately before or after it).

The BNC is the result of the work of a specially formed consortium (which united the Universities of Oxford and Lancaster, several commercial publishing houses, and the British Library); contains almost 100 million words of British English of the late twentieth century. The current version of the corpus was completed in 1994 (the corpus is non-dynamic, like the NKJP); 90% of all textual information are written sources of various genres and 10% are conversational sources. The BNC was created specifically for linguistic studies and, despite its relatively small size, the corpus is well balanced, and therefore the results of searches in it are representative, i.e. give an undistorted picture of the functioning of contemporary British English (according to C. Fellbaum, experts consider the BNC to be "a reliable source for English researchers" (Fellbaum, 2019, p. 749).

The publicly available program allows making basic searches for exact matches, all forms of a given lemma, collocations, etc. The corpus has a partial morphological marking, but shows a lack of semantic one.

For example, a basic search for the lemma *drink* as a verb can be done by entering the query drink_v* in the field; the search result will be general information about the frequency of this lemma; in particular, it was found 2,981 times. The marking of parts of speech makes it possible to search for all collocations of specialized reference verbs with potential noun objects at a distance of at most 4 tokens (see Fig. 10). The search result will be the list of all relevant nouns (see Fig. 11); the examples of the use can be viewed by highlighting the desired one (see Fig. 12).

Conclusion. Linguistic studies using text corpora allow automatic search and analysis of language units and their combinations. The text corpora of Ukrainian, Polish, and English (GRAC, NKJP, BNC) meet the requirements of authenticity, representativeness (balance), and sufficiency in volume. The introduction of research text corpora into linguistic use makes it possible to optimize and objectify the analysis of language material and provides for highly reliable results.

ABBREVIATIONS

BNC – British National Corpus
GRAC – The General Regionally Annotated Corpus of Ukrainian
NKJP – Narodowy Korpus Języka Polskiego

BIBLIOGRAPHY

1. Бук С. Н. Велика проза Івана Франка: електронний корпус, частотні словники та інші міждисциплінарні контексти. Львів: ЛНУ імені Івана Франка, 2021. 424 с.
2. Демська О. Текстовий корпус: ідея іншої форми. Київ: ВПЦ НаУКМА, 2011. 282 с.
3. Широков В., Бугаков О., Грязнухіна Т. та ін. Корпусна лінгвістика. Київ: Довіра, 2005. 471 с.
4. Fellbaum Ch. How flexible are idioms? A corpus-based study// Linguistics. 2019. Vol. 57. № 4. P. 735–767.
5. Fellbaum Ch. The treatment of multi-word units// Oxford Handbook of Lexicography/ ed. by P. Durkin. Oxford: Oxford U-ty Press, 2015. P. 411–425.
6. Gries S.Th. Corpus-based methods and cognitive semantics: The many senses of to run // Corpora in Cognitive Linguistics: Corpus-based Approaches to Syntax & Lexis/ ed. by S.Th. Gries & A. Stefanowitsch. Berlin; New York: Mouton de Gruyter, 2006. P. 57–100.
7. Shvedova M. The General Regionally Annotated Corpus of Ukrainian (GRAC, uacorpus.org): Architecture and Functionality // Proceedings of the 4th International Conference on Computational Linguistics and Intelligent Systems (COLINS 2020). 2020. Vol. I. P. 489–506.
8. Stefanowitsch A. Corpus Linguistics: A Guide to the Methodology. Berlin: Language Science Press, 2020. 508 p.

REFERENCES

1. Buk S. N. (2021). Velyka proza Ivana Franka: elektronnyi korpus, chastotni slovnyky ta inshi mizhdystsyplinarni konteksty [Ivan Franko's great prose: electronic corpus, frequency dictionaries, and other interdisciplinary contexts]. Lviv: LNU imeni Ivana Franka [in Ukrainian].
2. Demska O. (2011). Tekstovyi korpus: ideia inshoi formy [Text corpus: the idea of another form]. Kyiv: VPTs NaUKMA [in Ukrainian].
3. Fellbaum Ch. (2019). How flexible are idioms? A corpus-based study. Linguistics, 57 (4), 735–767.
4. Fellbaum Ch. (2015). The treatment of multi-word units. In P. Durkin (Ed.), Oxford Handbook of Lexicography (pp. 411–425). Oxford: Oxford University Press.

5. Gries S.Th. (2006). Corpus-based methods and cognitive semantics: The many senses of to run // In S.Th. Gries & A. Stefanowitsch (Eds.), Corpora in Cognitive Linguistics: Corpus-based Approaches to Syntax & Lexis (pp. 57–100). Berlin; New York: Mouton de Gruyter.

6. Shvedova M. (2020). The General Regionally Annotated Corpus of Ukrainian (GRAC, uacorpus.org): Architecture and Functionality. Proceedings of the 4th International Conference on Computational Linguistics and Intelligent Systems (COLINS 2020), I, 489–506.

7. Shyrokov V., Buhakov O., Hriaznukhina T. a.o. (2005). Korpusna linhvistyka [Corpus linguistics]. Kyiv: Dovira [in Ukrainian].

8. Stefanowitsch A. (2020). Corpus Linguistics: A Guide to the Methodology. Berlin: Language Science Press.

КОРПУСНИЙ ПІДХІД У ДОСЛІДЖЕННІ ДІЄСЛІВНИХ ПРЕДИКАТІВ

Назарчук Роксолана Зіновіївна
кандидат філологічних наук,
старший викладач кафедри прикладної лінгвістики
Національного університету «Львівська політехніка»
вул. С. Бандери, 12, Львів, Україна

Карамишева Ірина Дамірівна
кандидат філологічних наук,
доцент кафедри прикладної лінгвістики
Національного університету «Львівська політехніка»
вул. С. Бандери, 12, Львів, Україна

У статті висвітлено особливості корпусного підходу до аналізу дієслівних предикатів української, польської, англійської мов. Окреслено проблеми корпусної лінгвістики, здійснено огляд корпусів текстів, а саме Генерального регіонально анотованого корпусу української мови (GRAC), Національного корпусу польської мови (NKJP) і Британського національного корпусу (BNC). Згадані ресурси, відповідаючи вимогам автентичності, репрезентативності (збалансованості), достатності за обсягом (С. Бук, О. Демська, К. Фельбавм, А. Стефановіч, В. Широков, М. Шведова та ін.), дають змогу оптимізувати й об'єктивізувати тлумачення мовного матеріалу, одержати високу достовірність результатів. Наукова новизна дослідження полягає у комплексному зіставному аналізі функційних можливостей корпусів текстів української, польської, англійської мов з метою визначення їхніх особливостей, а також окреслення перспектив застосування корпуснобазованого підходу для вияву об'єктних зв'язків дієслівних предикатів. Уведено основні поняття корпусної лінгвістики: лема, словоформа, токен, мітка (тег), співуживання (колокація), конкорданс, частотність, запит CQL; їх проілюстровано реалізаціями відповідних референтно спеціалізованих дієслівних предикатів і їхніх об'єктів на базі GRAC, NKJP і BNC. Зокрема, показано, що найпростіше використання корпусів текстів полягає в засвідченні вживання тої чи іншої одиниці (concordance) та її частотності (тобто кількості повторень у корпусі). Для мов, багатих на словозміну, важлива можливість ідентифікації різних форм одного слова з його базовою формою – лемою. Проілюстровано, як досліджені корпуси уможливлюють подальше групування, упорядкування за частотністю та виокремлення конкретних словосполук, а розмітка частин мови полегшує пошук усіх колокацій дієслівних предикатів з потенційними об'єктами-іменниками на відстані 4 токенів. Описано функцію пошуку співуживань дієслів, погрупованих за лемами чи граматичними характеристиками та впорядкованих за частотністю. Вказано на переваги семантичної розмітки у корпусі GRAC для автоматичного пошуку одиниць мови та їхніх поєднань. Усі описані особливості корпусів текстів супроводжено ілюстраціями з GRAC, NKJP і BNC.

Ключові слова: *корпусна лінгвістика, корпус, GRAC, NKJP, BNC, дієслово, об'єкт, лема, колокація*