UDC 811.111’322:004.9
DOI https://doi.org/10.32782/tps2663-4880/2026.45.1.16

SKETCH ENGINE AS AN INSTRUMENT OF CORPUS-BASED LINGUISTIC ANALYSIS

SKETCH ENGINE ЯК ІНСТРУМЕНТ КОРПУСНО-ОРІЄНТОВАНОГО ЛІНГВІСТИЧНОГО АНАЛІЗУ

Bober N.M.,
orcid.org/0000-0002-9639-0562
Candidate of Philological Sciences, Associate Professor,
Associate Professor at the Department of Germanic Philology
Borys Grinchenko Kyiv Metropolitan University

The relevance of the study is determined by the rapid growth of the role of digital corpus tools in contemporary linguistic analytics and the need for critical reflection on their methodological potential. The aim of the article is to provide a comprehensive investigation of the functional capabilities and limitations of the Sketch Engine platform in the context of corpus-driven and digital approaches to the analysis of linguistic data. The methodological framework is based on methods of corpus analysis, contrastive interpretation, and automated explication of lexico-grammatical characteristics, which are used to assess the performance of the key modules of Sketch Engine and to verify their relevance for working with different types of corpora.

The study identifies specific features of the tool’s operation with grammatical and semantic profiles of word usage, automatic collocational models, and frequency characteristics of lexical units. The effectiveness of Word Sketch, Sketch Diff, and Concordance is analysed in the investigation of lexico-semantic, syntactic, and pragmatic phenomena, in particular in the study of emotionally coloured vocabulary in contemporary English.

A number of technical and methodological limitations affecting the interpretation of results are identified, including the need for corpus data filtering, the complexity of processing polysemy and homonymy, and the dependence of the quality of automatically generated data on corpus specificity. Examples of the practical application of the tool for analysing linguistic categories that are difficult to formalise, in particular metaphorical models and emotional states, are generalised.

An approach is developed that combines automated procedures with critical linguistic interpretation in order to increase the reliability of large-scale textual data analysis. It is concluded that Sketch Engine is a powerful tool for analysing large volumes of textual data (Big Data); however, its use requires a high level of professional competence, methodological caution, and a clear awareness of the linguistic context and epistemological goals of scientific inquiry.

The practical significance of the study lies in the possibility of using its results by specialists in corpus linguistics, digital humanities, and applied linguistics to optimise research procedures and enhance analytical accuracy when working with different types of corpora.

Key words: digital language data, emotional vocabulary, collocations, corpus tools, tagging, Word Sketch, Concordance.

Актуальність дослідження зумовлена стрімким зростанням ролі цифрових корпусних інструментів у сучасній лінгвістичній аналітиці та необхідністю критичного осмислення їхнього методологічного потенціалу. Метою статті є комплексне дослідження функціональних можливостей і обмежень платформи Sketch Engine у контексті корпусно-цифрових підходів до аналізу мовних даних.

Методологічну основу становлять методи корпусного аналізу, зіставної інтерпретації та автоматизованого експлікування лексико-граматичних характеристик, за допомогою яких здійснено оцінювання продуктивності ключових модулів Sketch Engine та перевірено їхню релевантність для роботи з різними типами корпусів.

Встановлено особливості роботи інструменту з граматико-семантичними профілями слововживання, автоматичними колокаційними моделями та частотними характеристиками лексичних одиниць. Проаналізовано ефективність застосування Word Sketch, Sketch Diff та Concordance у вивченні лексико-семантичних, синтаксичних і прагматичних явищ, зокрема у вивченні емоційно забарвленої лексики сучасної англійської мови.

Виявлено низку технічних та методологічних обмежень, що впливають на інтерпретацію результатів: необхідність фільтрації корпусного матеріалу, складність обробки полісемії та омонімії, а також залежність якості автоматично згенерованих даних від специфіки корпусу. Узагальнено приклади практичного використання інструменту для аналізу мовних категорій, що погано піддаються формалізації, зокрема метафоричних моделей та емоційних станів.

Розроблено підхід до поєднання автоматизованих процедур із критичною лінгвістичною інтерпретацією для підвищення достовірності аналізу великих масивів текстових даних. Робиться висновок, що Sketch Engine є потужним інструментом для аналізу великих масивів текстових даних (Big Data), проте його використання вимагає високої фахової компетентності, методологічної обережності та чіткого усвідомлення лінгвістичного контексту й епістемологічних цілей наукового пошуку.

Практична цінність полягає у можливості використання результатів фахівцями з корпусної лінгвістики, цифрової гуманітаристики та прикладної лінгвістики для оптимізації дослідницьких процедур і підвищення аналітичної точності під час роботи з корпусами різних типів.

Ключові слова: цифрові мовні дані, емоційна лексика, колокації, корпусні інструменти, тегування, Word Sketch, Concordance.

Problem statement. In contemporary linguistics, digital corpora and automated text-processing tools have ceased to serve merely as auxiliary resources; instead, they increasingly shape the methodological architecture of research, establish new standards of empirical evidence, and enable the analysis of linguistic phenomena on a scale previously unattainable. Modern linguistic studies are progressively oriented toward the analysis of large datasets, as the digitalisation of communication has led to a rapid expansion of textual information and a transformation of traditional conceptions of language patterns. In this context, corpus-based methods become central to the investigation of the lexical-semantic, syntactic, and pragmatic characteristics of linguistic units, enabling the identification of usage regularities that cannot be established intuitively or through conventional linguistic interpretation. Of particular significance are tools that integrate automated data processing with sophisticated linguistic modelling, among which Sketch Engine holds a leading position. Therefore, exploring its capabilities, limitations, and potential for analysing emotionally marked vocabulary constitutes a relevant and timely research objective.

Owing to the flexibility of its search settings and a wide range of corpus tools (such as Concordance, Word Sketch, Collocations, Thesaurus, and others), this platform provides opportunities for multidimensional investigation of lexical-semantic relations, syntactic patterns, and pragmatic features of linguistic units. Particularly important is the fact that Sketch Engine allows researchers to build custom corpora tailored to specific research goals – from the analysis of media discourse to the study of professional terminology or emotional lexis. This makes the tool universally applicable within both academic and applied linguistic research.

Analysis of recent research and publications. Over the past decades, corpus linguistics has established itself as one of the key methodological paradigms in linguistic research, providing an objective, representative, and large-scale empirical basis for the study of diverse linguistic phenomena [1; 2; 3; 6; 8; 24; 31]. In the context of the intensification of digital technologies [15; 22; 32], the use of specialised software platforms that automate data collection, systematisation, and analysis processes has become particularly relevant.

One of the most functionally rich and, at the same time, accessible tools in this domain is Sketch Engine – an online platform [10] for corpus analysis that allows researchers to work with both ready-made and independently created corpora. Owing to automated analysis algorithms [26], including the generation of grammatical and semantic profiles of word usage (word sketches) [16; 18], the construction of collocational profiles [9; 12], frequency lists, and contextual exploration of key units, this tool is widely employed across various areas of linguistic studies–from terminology and phraseology to cognitive semantics and emotional lexicography.

A considerable number of scholarly works emphasise the functional advantages of Sketch Engine, particularly its ability to generate grammatical and semantic word sketches–synthetic representations of the grammatical and lexical relations of a given unit based on the frequency analysis of its contextual environments. As noted by A. Kilgarriff [19], this function provides a new level of lexicographic processing and enables highly accurate modelling of dictionary entries grounded in real language use. In addition, the tool offers researchers access to Keyword Analysis, Collocation Extraction, Thesaurus Building, and Concordance Search, which significantly broadens the range of linguistic tasks [11] that can be implemented within the corpus-based approach.

A substantial body of research is devoted to the application of Sketch Engine in lexicography. For instance, the studies by I. Kosem [20] demonstrate the effectiveness of this tool in compiling academic and learner dictionaries, where the automatic extraction of typical contexts and grammatical constructions is crucial. Within lexical-semantic research, the platform is employed to analyse synonymy, antonymy, semantic fields, and conceptual domains, as illustrated in studies of emotional vocabulary [5; 27; 30].

Sketch Engine is also integrated into discourse analysis. In the works of Baker [4] and Mautner [23], it is shown how corpus-based approaches can reveal latent patterns of ideological and sociocultural representation in media discourse. In this context, Sketch Engine functions not only as a tool for examining lexical structures but also as a powerful means of identifying discursive practices.

Ukrainian scholars likewise actively contribute to the development of corpus-based approaches. In the study by Holoshchuk [17], the current state of corpus linguistics is analysed as an interdisciplinary research paradigm grounded in the computerised processing of large textual datasets. The author concludes that corpus linguistics represents a priority and direction for contemporary philological research, as it provides an empirical basis for analysing authentic language data, enhances the reliability of linguistic findings, and expands opportunities to investigate

language change at different levels (lexical, semantic, grammatical).

According to R. Makhachashvili and A. Bakhtina [21], the use of digital methods is particularly promising for the analysis of emotional content, since such data are characterised by high contextual variability and pragmatic density. T. Diak and Yu. Hrytsiuk [13] argues that corpus tools are effective for identifying keywords and stable thematic structures in folklore genres, demonstrating their capacity to handle texts of diverse stylistic and cultural origins. In turn, I. Dilay and M. Dilay [14] emphasise that cognitively oriented corpus linguistics requires integrating instrumental precision with theoretical models of interpretation, especially for the analysis of emotional state vocabulary.

Special attention should also be paid to technical and methodological issues. A number of studies [16; 18] highlight the need for a critical approach to interpreting results from automated tools, particularly in cases of polysemy, syntactic ambiguity, or corpus non-representativeness. Since Sketch Engine relies on pre-automated annotation, the analytical results largely depend on the quality of morphological and syntactic tagging, which is not always sufficiently accurate, especially for languages with highly developed inflectional systems. No algorithm is universally optimal: complex models (including large language models) achieve higher accuracy but require substantial computational resources, whereas simpler methods are more computationally efficient.

Thus, the review of scholarly sources indicates that Sketch Engine is not only a technically advanced platform but also a powerful analytical instrument that opens up new perspectives for linguistic research. At the same time, there is an urgent need for further reflection on its limitations, the development of methodological principles for its effective use, and the integration of automated corpus analysis with qualitative interpretative approaches.

The implementation of these tasks enables a comprehensive understanding of the status, effectiveness, and perspectives of applying Sketch Engine as a corpus linguistics tool, as well as to identify directions for the further development of methodologies for its use in the humanities.

The **methodological** framework of the present study is based on the principles of corpus linguistics, which involve the analysis of linguistic units using large, representative text corpora [28; 29] processed with digital tools. The central analytical instrument of the study is the Sketch Engine platform, which enables both quantitative and qualitative analysis of linguistic data across different languages.

The study employs an integrated approach that combines the following **methods**:

1. Corpus-linguistic method, involving the selection, processing, and analysis of language data using corpora. In particular, various types of corpora (including the British National Corpus [7], TenTen corpora, OpenSubtitles, enTenTen21, etc.) available in Sketch Engine were reviewed and tested.
2. Quantitative analysis**,** which makes it possible to establish frequency characteristics of lexemes, grammatical structures, collocations, and keywords.
3. Qualitative analysis**,** aimed at interpreting usage contexts through Concordance Search, Contextual Filtering, and Corpus Query Language (CQL).
4. Comparative analysis, allowing corpus-based results to be compared with other sources (dictionaries, theoretical models, previous studies) in order to verify data validity.
5. Critical analysis of platform functionality, intended to evaluate the limitations of Sketch Engine, particularly those related to automated annotation, result interpretation, corpus specificity, and linguistic ambiguity.

The application of this methodology has enabled a multi-level analysis of Sketch Engine as a research tool, revealing its functional capacities and limitations, outlining its potential domains of application, and outlining directions for further corpus-based linguistic research.

The literature review demonstrates that, despite substantial development in corpus approaches, insufficient attention has been paid to a systematic analysis of Sketch Engine's potential for the study of emotional vocabulary, particularly in terms of its lexical-semantic, syntactic, and pragmatic behaviour. Moreover, previous studies often focus on isolated technical aspects of the platform without offering a comprehensive assessment of its analytical potential and methodological constraints. This creates a clear need for research that synthesises the tool's functional parameters and demonstrates the specific features of its application to the analysis of emotional lexis in contemporary English.

Statement of the Task. The aim of the present study is to provide a comprehensive examination of the functional capabilities and limitations of the digital platform **Sketch Engine** in its application to corpus-based analysis of linguistic data in contemporary linguistics. The study focuses on evaluating the tool as a means of automated investigation of lexical-semantic, grammatical, and discursive features of language, as well as on critically reflecting on the factors that affect the validity and interpretation of corpus-based results.

In accordance with this aim, the study seeks to address the following **objectives:** to analyse the theoretical foundations of the corpus-based approach in linguistics and to determine the role of digital tools in modern linguistic research; to characterise the functional potential of Sketch Engine as a platform for corpus analysis, including the types of corpora, available operations, and output formats; to examine the main methods of linguistic data processing in Sketch Engine, in particular word sketches, collocation analysis, concordance search, frequency lists, keyword extraction, and related techniques; to identify the technical, methodological, and interpretative limitations of using Sketch Engine in linguistic studies; and to substantiate the relevance of applying Sketch Engine to the analysis of emotional vocabulary as an example of a linguistic category characterised by a high degree of semantic variability and contextual dependence.

Presentation of the Main Material. In the present study, a systematic analysis of the functional capabilities of the **Sketch Engine** platform was conducted as a corpus analysis tool that integrates automated processing of linguistic data with flexible analytical mechanisms. The primary focus was on tools most relevant to contemporary linguistic research, particularly for investigating the lexical, semantic, and discursive characteristics of linguistic units.

The **Sketch Engine** platform provides access to more than 500 corpora across over 90 languages, enabling both cross-linguistic and intra-linguistic comparisons. The corpora may be either general or specialized (e.g., scientific texts, mass media, social networks, spoken language), allowing researchers to select an appropriate linguistic environment in accordance with specific research objectives (Fig. 1). Within the framework of the present study, several English-language corpora–namely the British National Corpus (BNC) and enTenTen21–were selected for the analysis of emotional vocabulary as an illustrative case of the functional application of Sketch Engine.

One of the most powerful and at the same time methodologically significant functions of the **Sketch Engine** platform is **Word Sketch**–a tool that enables the automated generation of a generalised lexico-grammatical profile of a word. This function is based on the statistical analysis of co-occurrence patterns of lexical items within a given corpus and provides the researcher with structured information about the typical syntactic, collocational, and semantic relations of a particular unit. Essentially, Word Sketch can be viewed as a "lexicographic portrait" of a word, revealing its linguistic behaviour through frequency-driven patterns of actual usage.

Word Sketch constructs a set of collocations grouped according to types of grammatical relations (e.g., *modifier of noun*, *object of verb*, *subject of verb*, *adjective modified by*, etc.). The algorithm automatically identifies which words most frequently co-occur with the target unit and determines the strength of these associations using statistical measures such as LogDice, MI-score, T-score, and others. As a result, Word Sketch makes it possible to describe a lexeme not merely as an isolated unit, but as a node within a system of semantic and grammatical relations that reflect its actual functioning in language.

In the study of emotive vocabulary, that is, words denoting or expressing emotional states, the Word Sketch function is particularly valuable. It enables the reconstruction of emotional scenarios and contextual

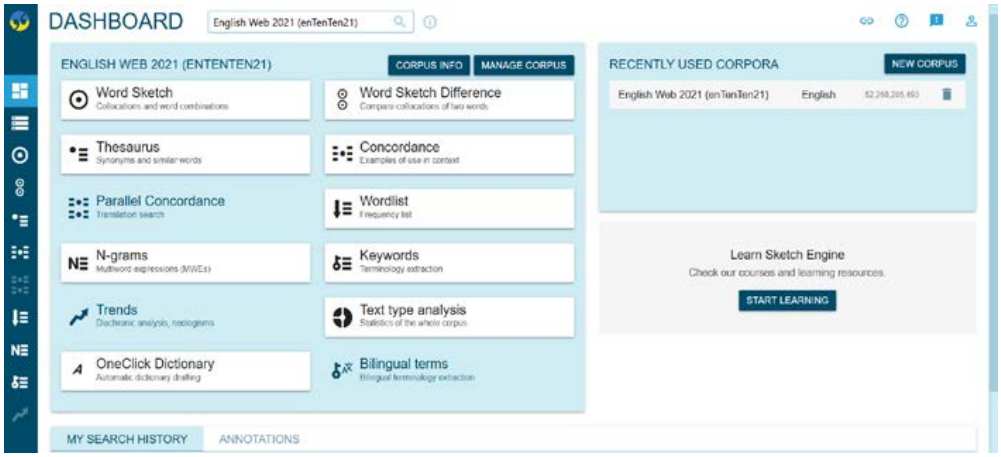

Fig. 1. Sketch Engine Dashboard

patterns in which specific lexemes reveal their semantic behaviour. For instance, an analysis of the lexeme *sad* in the British National Corpus using Word Sketch reveals typical collocations in the position of a nominal modifier (*sad story, sad eyes, sad truth*), as a predicative adjective (*feel sad, become sad, look sad*), as well as with intensifying adverbs (*very sad, deeply sad, unbearably sad*). These combinations not only capture syntagmatic relations but also reproduce the emotional continuum of Anglophone culture, in which sadness is represented through bodily, cognitive, and aesthetic images.

A similar analysis of the lexeme *happy* demonstrates an opposite semantic profile in terms of emotional valence. Word Sketch registers dominant links with the verbs *feel, look, make, and become (as verbal contexts) and with nouns such as life, people, moment, and child* functioning as subjects or objects of the emotional state. Collocations such as *truly happy, extremely happy, and perfectly happy* indicate typical patterns of intensification of positive emotion, whereas combinations like *happy to help* or *happy with results* illustrate the pragmatic function of approval or satisfaction. Thus, Word Sketch allows researchers to identify quantitatively and interpret qualitatively the typical lexico-semantic patterns that encode ways of verbalising emotional states.

For the analysis of more complex emotions–such as the lexemes *angry, anxious*, and *excited*–the Word Sketch function helps determine which verbs most frequently co-occur with a given emotion in subject or object position. In the case of *angry,* dominant constructions include *get angry, feel angry, and make someone angry, which point to the emotion's dynamic character, conceptualised* as arising in response to stimuli. By contrast, the analysis of *anxious* reveals collocations such as *anxious about the future* and *anxious to please*, which illustrate the cognitive-behavioural dimension of anxiety, while *excited* predominantly co-occurs with the verbs *feel, get, and be, and with nouns like news, idea, and audience*, indicating the affective-evaluative nature of this emotional state. For example, the analysis of the lexeme *anger* in the enTenTen21 corpus revealed typical grammatical relations (notably with *cause, feel, suppress, and explode with*), which enabled the reconstruction of recurrent usage patterns of the lexeme in authentic linguistic contexts (Fig. 2).

Word Sketch also makes it possible to compare the lexico-grammatical profiles of two or more words through the **Sketch Diff** tool. For example, a comparison of *angry* and *annoyed* reveals differences in the degree of intensity and social markedness: while *angry* more frequently occurs in constructions with verbs such as *get, make,* and *sound*, *annoyed* shows a higher collocational frequency with modifiers like *slightly, mildly,* and *clearly*, which points to a gradation in emotional intensity. This comparative approach is particularly useful for lexicographic description, the construction of emotional thesauri, and the refinement of emotional scales in corpus-based digital research.

Another important aspect is the possibility of analysing contextual variation across corpus types.

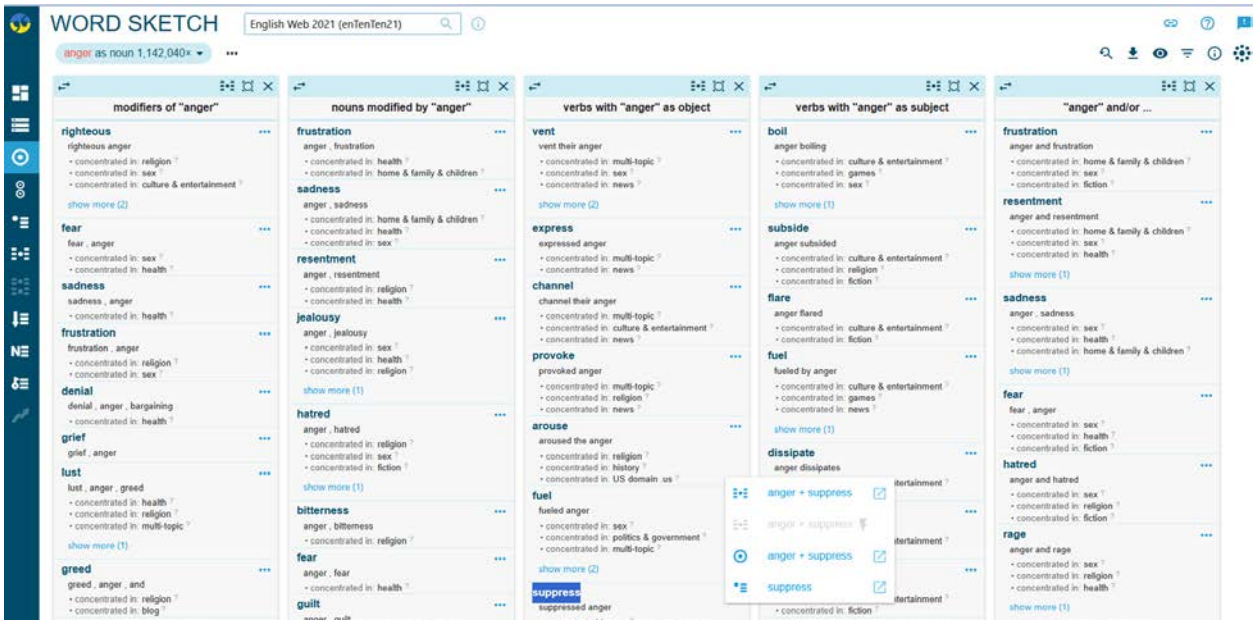

Fig. 2. Typical grammatical relations of the lexeme "anger" in the enTenTen21 corpus

For instance, in social media corpora, Word Sketch for the lexeme *sad* may reveal combinations such as *sad af, so sad rn, that's sad lol*, which reflect emerging informal patterns of emotional expression and a shift from describing an internal state to evaluating an event. In contrast, in literary corpora, more metaphorical or aestheticised combinations are typically observed–such as *sad melody, sad beauty,* or *sad silence*–which indicates genre-based differentiation in emotional expression.

Thus, the Word Sketch function in Sketch Engine is not merely a technical means of identifying collocations, but an analytical instrument for investigating the linguistic conceptualisation of emotions. It enables researchers to uncover regularities in the expression of emotional states, trace semantic nuances, compare stylistic variants, and formalise patterns of emotional discourse. In combination with the *Concordance* and *Word List* tools, *Word Sketch* forms the basis for in-depth corpus-based digital analysis, allowing scholars to integrate the statistical reliability of data with linguistic interpretation and thereby bringing corpus linguistics closer to a more cognitively oriented understanding of emotional language.

One of the key functions of the **Sketch Engine** platform that ensures the analytical depth of corpus research is **Collocation Analysis**–a tool for identifying statistically significant co-occurrences of lexical units within a specified contextual window. From the perspective of contemporary linguistics, collocational analysis constitutes an essential component of the semantic-cognitive approach to language study, as it enables empirical investigation of how associative links between words are realised in actual language use, thereby shaping conceptual and emotional profiles. In Sketch Engine, this function is implemented with a high degree of automation, combining quantitative methods with linguistically relevant parameters, such as part-of-speech relations, positional patterns, directionality of collocational dependency, and statistical measures.

The Collocation Analysis function determines which units most frequently co-occur with a target word within a predefined context window (typically ±5 or ±10 positions). The results are presented in a table containing the collocate, its part of speech, co-occurrence frequency, and a statistical index of association strength–such as LogDice, MI-score (Mutual Information), T-score, or Log-likelihood. These parameters make it possible to identify not only the frequency of combinations, but also the degree of associative stability, which is critically important for the description of emotionally marked lexical units.

Collocational analysis is particularly effective in the study of positively expressive vocabulary, as such units tend to form stable emotional-semantic fields. As an illustrative example, let us consider the key lexeme happiness in the enTenTen (English Web Corpus) (Fig. 3), which contains over 19 billion words and represents a broad range of contemporary English-language discourse.

The collocational profile of happiness reveals that its most frequent associative partners can be grouped into several semantic categories:

1. Abstract concepts reflecting internal states or their sources: joy, peace, contentment, satisfaction, love, success, freedom, fulfilment. These collocations reflect a cognitive schema of harmony, in which happiness is conceptualised as a state of balance between emotional, spiritual, and social well-being.

2. Verbs denoting the cause, acquisition, or expression of happiness: bring happiness, find happiness, experience happiness, seek happiness, share happiness, spread happiness, achieve happiness. These combinations illustrate a dynamic model of emotion, in which happiness is the outcome of action. For instance, the collocations bring happiness and spread happiness, emphasising the altruistic dimension of positive emotions, characteristic of humanistic and ethical discourses.

3. Adjectival modifiers indicating the degree or source of the emotional state: true happiness, pure happiness, lasting happiness, real happiness, eternal happiness, great happiness. Such expressions reflect the intensification of emotional meaning and reveal a tendency toward the idealisation of the concept of happiness. Particularly indicative is the use of the adjectives true and real, which foreground the emotion's authenticity and moral-evaluative dimension.

4. Contextual collocations with nouns from the social or cultural domain: Happiness Index, happiness level, happiness study, The Happiness Project.

These expressions demonstrate the sociolinguistic institutionalisation of the concept of happiness in contemporary public discourse, where emotion is redefined as a domain of measurable indicators of well-being.

The collocational analysis of happiness also enables the identification of evaluative frameworks. For example, combinations such as happiness and success or happiness and love indicate the alignment of happiness with socially approved values. By contrast, pairs like *happiness* and *sorrow*, or *happiness* and *misery*, construct antithetical conceptual oppositions that structure the emotional scale of human experience.

The use of statistical measures in **Sketch Engine** enables researchers to distinguish between high-frequency but weakly associated collocations (e.g., find happiness) and relatively infrequent but strongly associated semantic pairs (e.g., eternal happiness). In this respect, the LogDice measure is more balanced, as it takes into account both frequency and proportional strength of association, unlike the MI-score, which tends to overestimate rare combinations. Thus, the Collocation Analysis function in Sketch Engine does not merely register co-occurrence, but also allows collocations to be ranked according to their degree of lexico-semantic relevance, which is particularly valuable in the study of emotionally marked vocabulary.

The application of collocational analysis to the study of positively **marked emotional concepts**, such as *happiness*, reveals not only frequency-based patterns but also the underlying cultural and axiological structures that shape the use of this lexeme. For instance, in Western discourse, happiness is often correlated with categories such as achievement, freedom, and love, whereas in Asian corpora (e.g., the Japanese Web Corpus), contexts emerge in which happiness is associated with family harmony, gratitude, and inner peace, reflecting the cultural specificity of the emotional concept.

It can therefore be assumed that Collocation Analysis in Sketch Engine functions not merely as a tool for statistical computation, but as a methodological instrument for reconstructing the conceptual structure of emotional notions. Its application to the study of positively marked lexis, and happiness in particular, makes it possible to describe systemic patterns of emotional expression, trace the dynamics of their cultural representations, and uncover latent patterns of associative thinking.

The **Concordance Search** tool proved particularly useful for qualitative analysis of lexical usage contexts, enabling identification of ironic, metaphorical, or pragmatically conditioned uses of linguistic units. Unlike statistically oriented functions (such as *Collocation Analysis* or *Word Sketch*), concordance provides an in-depth contextual approach that allows the researcher to observe a lexeme "in action," that is, in its authentic textual realisations. This feature makes Concordance Search an indispensable tool for the interpretative analysis of emotionally, evaluatively, and pragmatically marked units.

The Concordance Search tool allows researchers to generate usage examples (concordance lines) based on queries formulated as lexemes, word forms, parts of speech, or CQL (Corpus Query Language) patterns. Each result is displayed in the KWIC (Key Word In Context) format, with the target word in the centre and its left and right context. This format ensures balanced contextual visibility, enabling the researcher to trace regularities in syntactic environments, stylistic features, and connotative shifts in meaning.

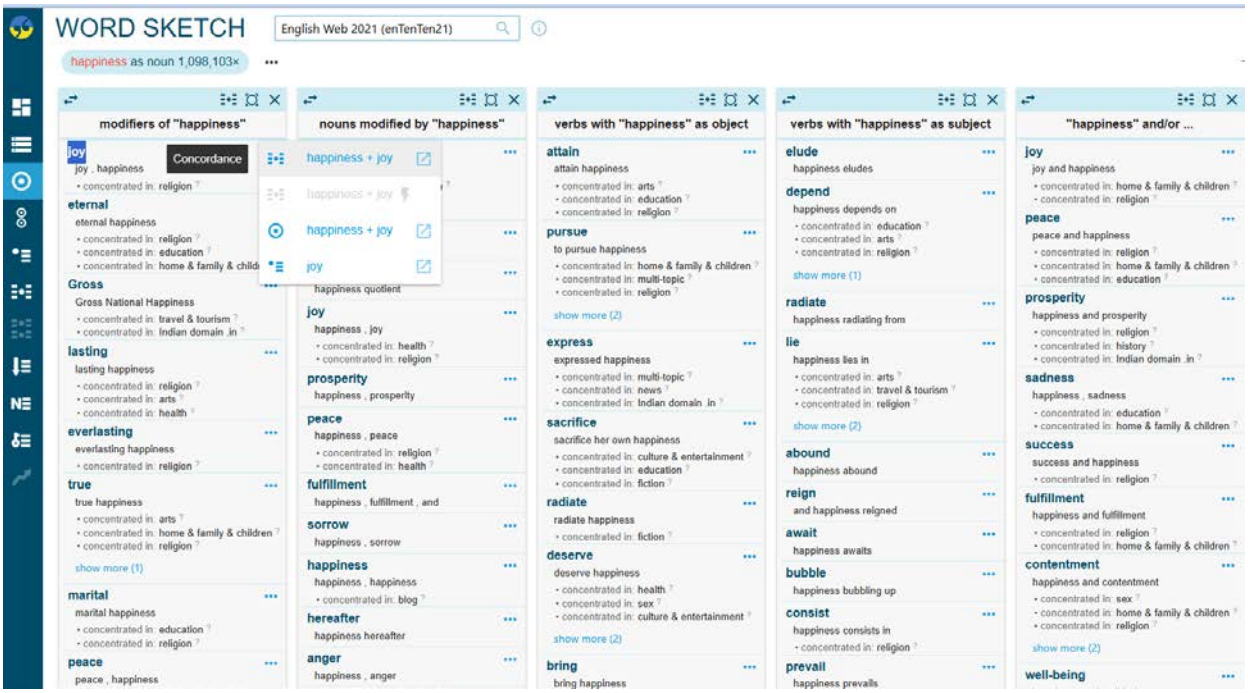

Fig. 3. Collocation Analysis: typical collocational patterns of positively marked lexis

In corpus-based studies of emotional vocabulary, Concordance Search functions as a kind of microscope, enabling the capture not only of frequency but also of the semantic flexibility of emotional concepts. For example, the analysis of concordance lines for the lexeme *fear* in the enTenTen corpus (Fig. 4) shows that this word frequently occurs in political or media discourse in ideologically loaded contexts. In this respect, Concordance Search within Sketch Engine provides a powerful qualitative complement to quantitative corpus methods, allowing researchers to integrate statistical tendencies with fine-grained semantic and pragmatic interpretation.

"Politicians exploit the public's fear to gain support for stricter policies."

"The campaign was driven by fear rather than facts."

"Media outlets amplified fear of migration to influence public opinion."

Such examples demonstrate that *fear* extends beyond its basic emotional meaning ("the feeling of fear") and acquires a discursive function of manipulation, becoming an instrument of emotional influence in the rhetoric of power or the media. In this way, qualitative concordance analysis enables the identification of pragmatic strategies by which emotional vocabulary shapes evaluations, beliefs, or social attitudes.

Similarly, the analysis of the positively marked lexeme *happiness* in the same corpus reveals a range of semantic nuances depending on the communicative genre:

"Happiness is not something ready-made. It comes from your own actions."

"Buying a new phone gave her a brief moment of happiness."

"Governments should measure citizens' happiness alongside economic growth."

In the first example, *happiness* carries an existential-philosophical nuance; in the second, a consumerist-emotional one; and in the third, a social-institutional one. Through concordance analysis, the researcher can classify the types of contexts in which particular meanings of an emotional lexeme are actualised and trace how the concept changes across different discursive domains (media, politics, advertising, personal blogs, etc.).

Particular attention should be paid to the identification of ironic and metaphorical uses of emotional lexemes. Thus, in concordance lines for *love*, secondary and humorous uses are frequently observed, where emotional expression is accompanied by ironic distancing:

"Oh, I just love waiting in traffic for hours."

"Don't you just love when your computer freezes right before a deadline?"

In these examples, the lexeme *love* operates within an inverted semantic field, producing a sense of sarcasm. Such cases are almost impossible to detect using purely statistical methods; however, they are readily identifiable through Concordance Search, which underscores the importance of the qualitative stage in corpus analysis.

Another example concerns the metaphorical use of emotional lexemes, in which a word transfers its meaning from the psychological domain to the social or political one:

"The nation was gripped by fear after the announcement."

"The city breathed happiness during the festival."

Here, *fear* and *happiness* function as anthropomorphised categories that describe not individual but collective states. Through concordance analysis, the researcher can trace a shift in focus from individual emotion to collective metaphor, which is of considerable significance for cognitive-discursive interpretation.

An important function of **Concordance Search** is the ability to sort results by various parameters: part-of-speech context, left or right context, lexical pattern, or text source. This allows the material to be structured by semantic types and enables the identification of conventional usage patterns–for example, verbal constructions such as *feel fear*, *overcome fear*, *spread happiness*, and *find happiness*. Further analysis of such patterns opens up the possibility of semantic profiling of emotional concepts, in which each lexeme emerges as the centre of a network of contextual associations.

It follows that **Concordance Search** in Sketch Engine is not merely a basic technical search tool, but a methodological platform for the interpretative analysis of emotional concepts. It facilitates the transition from the quantitative to the qualitative level of corpus research, enabling the description not only of usage frequency, but also of pragmatic strategies, semantic variability, and cultural connotations of emotional lexemes. In this way, concordance analysis becomes a key stage in constructing a comprehensive corpus-based model of emotional discourse, in which emotion is viewed as a multidimensional phenomenon–simultaneously linguistic, cognitive, and social.

Despite its evident functional power and broad analytical potential, using Sketch Engine as a corpus analysis tool entails several limitations that require the researcher to reflect critically. These concern

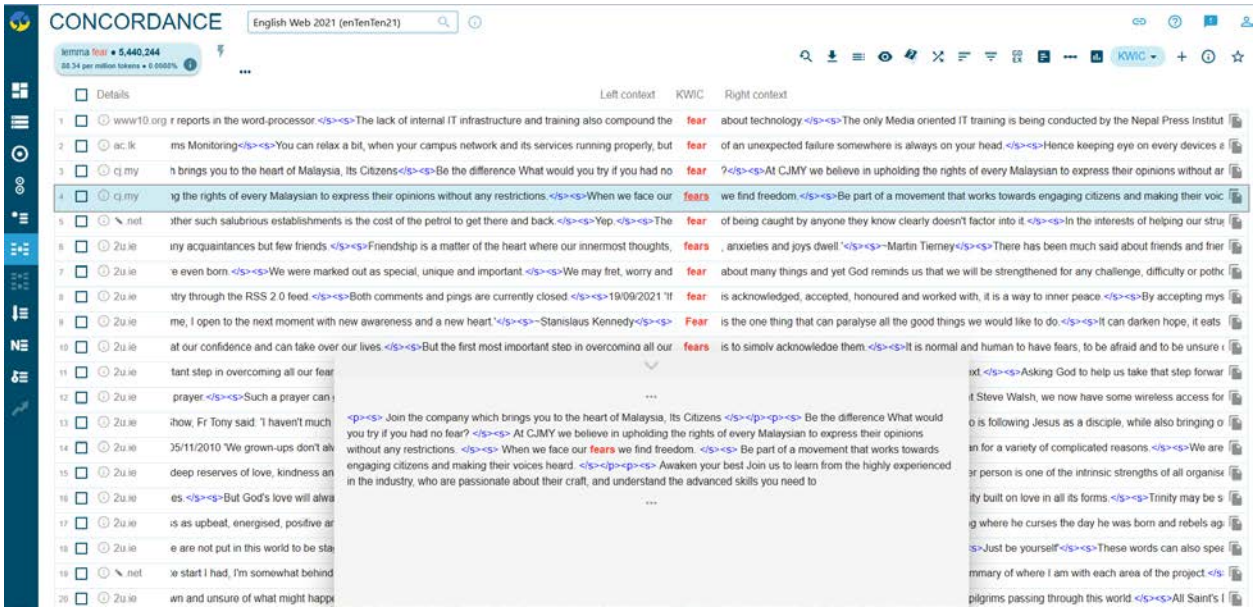

Fig. 4. Contextual analysis of the lexeme fear in corpus data using the Concordance Search tool

not only the technical aspects of working with the platform but also methodological risks associated with interpreting automatically generated results. It is important to emphasise that the effectiveness of corpus-based research largely depends not on the tool itself, but on the researcher's competence and ability to combine quantitative indicators with linguistically grounded conclusions.

One of the first and most significant limitations concerns errors in lemmatisation and part-of-speech tagging (POS-tagging). In Sketch Engine, these processes are performed automatically using built-in machine-learning algorithms that do not always adequately account for polysemy, contextual shifts, or idiomatic usage. For instance, in the case of words such as *love* or *fear*, the algorithm may incorrectly identify the part of speech in expressions like *'to love deeply' or 'a deep fear'*, or fail to distinguish emotional meanings from metaphorical or grammaticalised uses. Such errors affect frequency counts, collocational profiles, and even the results of automatic comparisons of lexical relations (Word Sketch). For emotionally marked vocabulary, where meaning is highly context-dependent, this problem is particularly acute.

A second problematic dimension concerns the structure of the corpora themselves, especially their size, genre representativeness, and balance. Although the platform offers hundreds of ready-made corpora in different languages, not all of them are of the same quality. In some corpora, certain genres (e.g., news or internet discourse) may dominate others, leading to statistical distortions in the data. In this context, analysing the lexeme *happiness* in a corpus dominated by social media texts may create the impression of its predominantly everyday-emotional or advertising-evaluative use, while academic or literary contexts remain underrepresented. Such genre disproportionality calls into question the representativeness of the sample and requires cautious interpretation of quantitative results.

A third limitation lies in the technical complexity of certain platform functions. Tools such as CQL (Corpus Query Language), custom corpus building, and advanced frequency comparison require specialised technical training. Without a basic understanding of corpus programming, query logic, and XML markup, the researcher may obtain incomplete or incorrect results. In this respect, the entry threshold for Sketch Engine remains relatively high for linguists who lack experience with formal query languages.

Another important issue is the interpretative ambiguity of automatically generated data. Even when the system functions correctly, Sketch Engine cannot guarantee that the identified collocations are linguistically relevant for emotionally marked vocabulary. For example, the frequent combination *fear + of* does not necessarily indicate a stable collocation–it may simply reflect structural regularities of English. Thus, automatic frequency does not equal semantic significance. Without additional qualitative analysis of concordances, such results may lead to misleading generalisations.

It is also necessary to take into account the difficulties of working with polycode texts (for

example, media discourse containing emojis, hyperlinks, tagged elements, or graphic symbols). Current recognition algorithms in Sketch Engine do not always accurately process these elements, which complicates the study of new forms of verbal expression, such as verbal emoticons or iconic orthography. For contemporary linguistics, which increasingly focuses on digital communication genres, this constitutes a substantial methodological limitation.

Furthermore, an important factor remains the dependence on corpus updates and algorithm versions. Sketch Engine continuously improves its tagging models, but this may lead to incompatibility in results from a diachronic perspective. Data obtained in different years may be difficult to compare due to changes in lemmatisation principles or statistical calculations. In long-term studies (e.g., diachronic or linguocultural research), such variability may distort trends or complicate the replication of results.

Among other limitations, one should also note the lack of full transparency of the algorithms underlying the calculation of collocations or the assessment of associative strength (for example, LogDice score or MI score). Although these measures are standardised, it is not always clear to the researcher how the system processes ambiguous cases or calculates frequencies in large corpora with duplicated data.

Finally, the limitations of Sketch Engine also concern the incomplete universality of its approaches across different languages. Algorithms developed for English often demonstrate lower accuracy in corpora of other languages with more complex morphological systems and greater lexical context dependence. For Ukrainian or Polish, typical errors in lemmatisation or tagging may reach 10–15%, which significantly affects the validity of quantitative conclusions.

Despite these limitations, Sketch Engine remains a leading analytical environment in corpus linguistics, combining large-scale data with flexible settings and the potential for deep interpretation. Its effectiveness depends directly on the researcher's methodological awareness, who must not only use the tool but also critically evaluate the results. Thus, Sketch Engine should be viewed not as an “automatic researcher”, but as an intellectual partner in scientific inquiry–powerful, yet demanding in terms of accuracy, representativeness, and interpretative rigour.

Conclusion. Summarising the results of the conducted study, it can be argued that the Sketch Engine platform has established itself as one of the most powerful tools in contemporary corpus linguistics, providing a multi-level analysis of linguistic phenomena. Its functional capabilities enable the combination of the empirical precision of quantitative data with the depth of qualitative interpretation, rendering this tool indispensable for research in lexis, semantics, and pragmatics, as well as for the study of language dynamics in the digital environment.

The results of the analysis have demonstrated that Sketch Engine effectively performs the tasks of identifying collocational structures, constructing lexico-grammatical profiles, and analysing the contextual use of lexemes, particularly in the domain of emotional vocabulary, where contextual variability and pragmatic load are key. The combination of tools such as Word Sketch, Collocation Analysis, and Concordance Search enables a comprehensive analysis of linguistic units–from their structural relations to their discursive realisations.

At the same time, the study emphasises that even the most advanced digital tools available today cannot replace the interpretative competence of the researcher. The accuracy of linguistic conclusions largely depends on the quality of the corpus data, the level of automatic annotation, the representativeness of the sample, and the scholar's methodological awareness. The identified limitations–such as lemmatisation errors, corpus heterogeneity, or the complexity of working with CQL queries–do not diminish the value of the platform, but rather highlight the need for a critical and reflective approach to corpus analysis.

Thus, Sketch Engine should be regarded not merely as a technical tool, but as a methodological digital system that integrates automation with analytical thinking. Its use opens new perspectives for cognitive, sociolinguistic, and emotion-semantic analysis, and also contributes to the development of new interdisciplinary approaches within the field of digital humanities.

Further research should be directed towards the development of specialised corpora for the analysis of specific thematic, genre-based, or sociolinguistic domains, as well as towards the integration of corpus tools with machine learning methods to enable in-depth analysis of dynamic language changes. In this context, Sketch Engine emerges as a platform capable not only of recording linguistic facts but also of facilitating the understanding of language as a living, variable, and cognitively rich system.

REFERENCES:

1. Anokhina T. O. Multilingual corpus as resource for working with political speeches by European public figures. *MESSENGER of Kyiv National Linguistic University. Series Philology*. 2023. Vol. 26, №. 2. P. 9–19. DOI: https://doi.org/10.32589/2311-0821.2.2023.297658

2. Andrushchenko O. Iu. Lancsbox software options for the prospective investigation of the multilingual corpus for European studies. *MESSENGER of Kyiv National Linguistic University. Series Philology*. 2023. Vol. 26, №. 1. P. 9–18. DOI: https://doi.org/10.32589/2311-0821.1.2023.286180

3. Asenciόn-Delaney Y., Collentine, J. G., Colmenares J. J., Urzúa A. Training teachers to use corpus tools in the Spanish language classroom. Journal of Spanish Language Teaching. 2022. Vol. 9, №. 2. P. 134–147. DOI: https://doi.org/10.1080/23247797.2022.2157082

4. Baker P. A Glossary of Corpus Linguistics. Edinburgh : Edinburgh University Press, 2006. 192 p.

5. Bednarek M. Semantic preference and semantic prosody re-examined. Corpus Linguistics and Linguistic Theory. 2008. Vol. 4, № 2. P. 119–139. DOI: https://doi.org/10.1515/CLLT.2008.006

6. Biber D., Conrad S., Reppen R. Corpus Linguistics: Investigating Language Structure and Use. Cambridge : Cambridge University Press, 1998. 300 p.

7. British National Corpus. *The British National Corpus*. URL: https://www.english-corpora.org/bnc/ (date of access : 26.01.2026).

8. Бобер Н. М. Матричне профілювання семантики фразово-дієслівних емотивів у Британському національному корпусі : дис. … канд. філол. наук : 10.02.04. / Національний педагогічний університет імені М.П. Драгоманова. Київ, 2020. 228 с. URL: https://npu.edu.ua/images/file/vidil_aspirant/avtoref/D_26.053.26/Bober.pdf (дата звернення : 28.11.2025)

9. Breslaw R., Laufer B. Learning new collocations: The effects of grouping and language of instruction. System. 2026. Vol. 136. Article. 103873. DOI: https://doi.org/10.1016/j.system.2025.103873

10. Буровицька Ю. М., Крива С. Е. Цифрова лінгвістика та лінгвістичний аналіз: доступні інструменти і сервіси. *Могилянські читання – 2025: досвід та тенденції розвитку суспільства в Україні: глобальний, національний та регіональний аспекти*. *Філологія* : матеріали XXVIII Всеукр. наук.-практ. конф., Миколаїв, Україна, 9–14 листоп. 2025 р. Миколаїв : ЧНУ ім. Петра Могили, 2025. С. 70–75. URL: https://dspace.chmnu.edu.ua/jspui/bitstream/123456789/3010/1/%D0%9C%D0%A7-2025.%20%D0%A4%D1%96%D0%BB%D0%BE%D0%BB%D0%BE%D0%B3%D1%96%D1%8F.pdf#page=70

11. Çyfeku J. Corpus linguistics and technology integration in SLA: Generating language tasks through Sketch Engine. Journal of Positive School Psychology. 2022. Vol. 6, №. 9. P. 1445–1458. URL: https://journalppw.com/index.php/jpsp/article/view/12428

12. Denysova N. B. Дискурс онлайн-навчання через леми й колокації: корпусний методологічний підхід. *Наукові записки. Сер. Філологічні науки*. 2025. №. 1 (214). С. 46–53. DOI: https://doi.org/10.32782/2522-4077-2025-214.1-5

13. Дяк Т. П., Грицюк Ю. І. Використання корпусних інструментів для виявлення ключових слів у стрілецьких і повстанських піснях як жанру фольклорного дискурсу. *Науковий вісник НЛТУ України*. 2024. № 7, т. 34. С. 60–71. DOI : https://doi.org/10.36930/40340708

14. Ділай І. П., Ділай М. П. Когнітивна корпусна лінгвістика: сучасний стан і перспективи. *Нова філологія*. 2021. № 83. С. 71–78. DOI : https://doi.org/10.26661/2414-1135-2021-83-10.

15. Франчук Н. П. Інноваційні методи навчання прикладній лінгвістиці з використанням інформаційних технологій. *Світові освітні тренди: навчання впродовж життя в інформаційному суспільстві* : зб. матеріалів Міжнар. наук.-практ. конф., присвяченої 190-річчю Університету та 50-річчю Інституту, Київ, 20–21 черв. 2024 р. Київ : УДУ імені Михайла Драгоманова, 2024. С. 224–227. URL: http://enpuir.npu.edu.ua/handle/123456789/46472

16. Gries S. T. Behavioral profiles: A fine-grained and quantitative approach in corpus-based lexical semantics. The Mental Lexicon. 2010. Vol. 5, №. 3. P. 323–346. DOI: https://doi.org/10.1075/ml.5.3.04gri

17. Holoshchuk S. Corpus linguistics: Modern approach and research perspective. Transcarpathian Philological Studies. 2022. Vol. 21, №. 1. P. 249–252. DOI: https://doi.org/10.32782/tps2663-4880/2022.21.1.47

18. Kilgarriff A., Rychly P., Smrz P., Tugwell D. The Sketch Engine. Proceedings of the International Conference on Language Resources and Evaluation. 2004. P. 105–115 URL: https://www.researchgate.net/publication/267223981

19. Kilgarriff A., Baisa V., Bušta J. The Sketch Engine: Ten years on. Lexicography. 2014. Vol. 1, №. 1. P. 7–36. DOI: https://doi.org/10.1007/s40607-014-0009-9

20. Kosem I., Koppel K., Zingano Kuhn T. Identification and automatic extraction of good dictionary examples: The case(s) of GDEX. International Journal of Lexicography. 2019. Vol. 32, №. 2. P. 119–137. DOI: https://doi.org/10.1093/ijl/ecy014

21. Махачашвілі Р. К., Бахтіна А. О. Hate Emoji в лінгвокриміналістичних експертизах: проблеми декодифікації смислу. *Вісник Житомирського державного університету імені Івана Франка. Сер. Філологічні науки.* 2021. № 94. С. 79–96. DOI: https://doi.org/10.35433/philology.1(94).2021.79-96

22. Makhachashvili R., Semenist I., Klochkov V. AI-enhanced multilingual lexicography for digital communication. Proceedings of IMCIC. 2025. Vol. 1, P. 247–253. DOI: https://doi.org/10.54808/IMCIC2025.01.247

23. Mautner G. Mining large corpora for social information: The case of elderly. Language in Society. 2007. Vol. 36, №. 1. P. 51–72. DOI: https://doi.org/10.1017/S0047404507070030

24. O'Keeffe A., McCarthy M. The Routledge Handbook of Corpus Linguistics. New York: Routledge, 2022. 754 p. DOI: https://doi.org/10.4324/9780367076399

25. O'Keeffe A., McCarthy M., Carter R. From Corpus to Classroom: Language Use and Language Teaching. Cambridge: Cambridge University Press, 2007. 315 p. DOI: https://doi.org/10.1017/CBO9780511497650

26. Parnus K., Svysiuk O. The algorithm of analything English texts with the help of parsers. *SWorldJournal*. 2022. Vol. 1, № 14-01. P. 108–113. DOI: https://doi.org/10.30888/2663-5712.2022-14-01-020

27. Partington A., Taylor C., Duguid A. Patterns and Meanings in Discourse. Amsterdam: John Benjamins, 2013. 372 p. DOI: https://doi.org/10.1075/ijcl.19.2.07bro

28. Stefanowitsch A. Corpus Linguistics: A Guide to the Methodology. Berlin: Language Science Press, 2020. 481 p. DOI: https://doi.org/10.5281/zenodo.3735822

29. Степанов В., Могильник А. Потенціал корпусних технологій у вивченні фразеологізмів та сленгізмів. *Філологічні трактати*. 2025. № 1, т. 17. С. 63–77. DOI : https://www.doi.org/10.21272/Ftrk.2025.17(1)-6

30. Стєкольщикова В. А., Бабич В. І., Сікорська В. Ю. Аналіз емоційного контенту у масових комунікаціях: застосування методів машинного навчання. *Вчені записки Таврійського національного університету імені В. І. Вернальского. Сер. Філологія. Журналістика.* 2025. № 75, т. 36. С. 191–196. DOI: https://doi.org/10.32782/2710-4656/2025.2.2/29

31. Zhukovska V., Mosiiuk O., Buk S. Register distribution of English detached nonfinite/nonverbal constructions with explicit subject. CEUR Workshop Proceedings. 2023. Vol. 3396. P. 63–76. URL: http://eprints.zu.edu.ua/37287/

32. Зозуля Н., Стєкольщикова В., Бабич В. Використання цифрових технологій для аналізу комунікативних процесів у соціальних мережах. *Наукові праці Міжрегіональної Академії управління персоналом. Філологія.* 2025. № 1 (15). С. 16–20. DOI: https://doi.org/10.32689/maup.philol.2025.1.3

Дата першого надходження статті до видання: 25.02.2026
Дата прийняття статті до друку після рецензування: 30.03.2026
Дата публікації (оприлюднення) статті: 07.05.2026