Corpus-Based Data Processing of Terminological Systems in Linguistic Theory

Tetiana Hromko[1], Tetiana Kovalevska[2], Hanna Truba[2] and Aryna Frumkina[2]
[1]*Department of Germanic and Oriental Languages and Translation, International University of Odesa, Fontanska Doroha Str. 33, 65062 Odesa, Ukraine*
[2]*Department of Ukrainian Languages and Language Training for Foreign Citizens, Odesa I. I. Mechnikov National University, French Blvd., 65082 Odesa, Ukraine*
hromkotv@gmail.com, tetiana.kovalevskaya@gmail.com, 3182009060@ukr.net, arynafrumkina@gmail.com

Keywords: Data Analysis, Corpus-based Modelling, Quantitative Linguistic Analysis Institutional Discourse, Terminological System, Suggestive Linguistics, Odesa Linguistic School.

Abstract: The article presents a corpus-oriented analysis of the terminological dominants of linguistic theory based on publications of the Odesa Linguistic School, which makes it possible to interpret its scholarly discourse as an institutionally coherent and quantitatively verified body of knowledge. The study is grounded in a combination of corpus-linguistic methods and linguistic data analysis, within which scholarly texts are treated as a structured informational resource. The corpus analysis encompasses the frequency-based parametrization of terms, the identification of their keyness, concordance analysis of usage contexts, modelling of collocational relations, and a diachronic assessment of the dynamics of the terminological system. The totality of these procedures provides a formalized description of the School's metalanguage and its internal structural organization. The results demonstrate the presence of a stable invariant core of the terminological system, formed by basic and derivative terms of linguistic theory, as well as a dynamic periphery associated with the development of specialized subfields. Of particular analytical value is the subcorpus of suggestive linguistics, which emerges as an autonomous, statistically verified, and methodologically mature segment of the institutional discourse. The applied approach ensures the institutional validation of the Odesa Linguuistic School as an integral scholarly formation with its own metalanguage and a controlled innovative dynamics, and demonstrates the effectiveness of corpus-based and quantitative methods for the analysis of collective scholarly discourses in contemporary humanities research.

1 INTRODUCTION

In the contemporary scholarly landscape, corpus research is increasingly viewed within an interdisciplinary framework in which the text emerges as an object of analysis and data processing, and scholarly discourse as a structured informational corpus. From the perspective of corpus analysis, this opens up new possibilities for the verification of humanities research, particularly for the reconstruction of specialized scholarly texts in which linguistic units function as carriers of conceptual and institutional knowledge [1], [2].

Of particular scholarly value in this regard are corpora compiled from the publications of individual scientific schools. Such corpora are characterized by thematic coherence, a stable terminological system, and internal methodological consistency, which allows them to be regarded as representative models of institutional knowledge [3], [4]. The analysis of terminological dominants within such corpora makes it possible to identify not only linguistic regularities but also the deeper mechanisms underlying the formation of a scientific tradition [6] - [7].

This study focuses on a corpus of publications of the Odesa Linguistic School from 2014-2024 [8], which is treated as an integral institutional body suitable for analysis within the paradigm of quantitative linguistic data processing. Terms of linguistic theory in such a corpus function as structural markers that register the School's theoretical orientations, its conceptual continuity, and the directions of its internal development.

The study builds upon the author's previous research, in particular the article devoted to the analysis of multimodality in the works of Sebastian

Unger [9], which demonstrated the possibilities of combining corpus-oriented analysis with the interpretation of complex semiotic structures. In the present work, the focus is shifted to the level of collective scholarly discourse, making it possible to broaden the scope of application of the corpus approach and to demonstrate its relevance for the analysis of institutional terminological systems.

The aim of the article is to identify the terminological dominants of linguistic theory in the corpus of publications of the Odesa Linguistic School and to substantiate the School's corpus-based identity as an institutional phenomenon of contemporary linguistics within a corpus-oriented analytical framework.

2 RELEVANCE AND ANALYSIS OF THE TOPIC AREA

The Odesa Linguistic School, well known among Ukrainian and international linguists for its significant research in cognitive, psycho- and neurolinguistics and discourse theory in their projection onto the foundational ideas of suggestive linguistics, represents a stable and methodologically coherent scholarly tradition reflected in the corpus of its publications. These are presented in a series of collective monographs titled The Odesa Linguistic School [6] as well as in the fundamental dissertation research of O. Hohorenko, M. Druzhynets, A. Kovalevska, T.Hromko, N. Kutuza, A. Romanchenko, H. Truba, and others.

Within this corpus, several groups of terms can be distinguished. The first comprises general theoretical terms of linguistic theory, which form the terminological core of the corpus and are characterized by stability throughout the analyzed period. The second group consists of terms related to the internal differentiation of linguistic theory, reflecting the development of particular theoretical directions and the expansion of the School's domain of inquiry. A separate group is formed by terms of suggestive linguistics, which constitute a specialized segment of the corpus and represent the emergence of an independent direction within the overall terminological system of linguistic theory

3 METHODOLOGY

The research methodology is based on a corpus approach and on methods of text data analysis and processing, corresponding to the thematic field of Data Analysis and Processing. Scholarly texts are treated as structured informational corpora, and the terms of linguistic theory as key units for the representation of institutional knowledge, suitable for formalized analysis and interpretation.

The corpus of publications of the Odesa Linguistic School is used as an institutionally coherent data set within which the systematization of terms is carried out and their functioning in scholarly discourse is traced. Such an approach makes it possible to interpret the corpus as an analytical model in which linguistic units function as variables, and their frequency and combinability as parameters of analysis.

The analysis of the suggestive linguistics subcorpus relies on the methodological principles developed by T. Yu. Kovalevska, in particular on the understanding of suggestiveness as a multilevel linguistic-discursive phenomenon and on the principles of its linguistic description. Within this approach, suggestive units are examined in the system of cognitive, pragmatic, and discursive interactions, which enables their analysis as an integral theoretical paradigm.

The methodological foundation of the study is also constituted by the author's conceptual framework presented in T. Hromko's monograph «Methodology and Experience of Subdialect Description» (2021), in which the linguarium is conceptualized as a corpus-grounded, stratified, and cognitively verified system. The application of this concept to the analysis of an institutional scholarly corpus makes it possible to interpret the School's terminological system as a metalanguage structure formed in the process of collective scholarly activity.

Within a data-driven analytical framework, the integration of corpus analysis with linguistic interpretation enables the processing of textual data as a formalized resource and provides a basis for identifying terminological dominants and the internal differentiation of scholarly discourse.

4 RESULTS

The corpus-based representation of the terminological dominants of linguistic theory in the publications of representatives of the Odesa Linguistic School from 2014-2024 demonstrates the presence of a clearly structured and methodologically consistent terminological system. Its core is formed by basic and derivative terms marked by the components linguistics, linguistic science, linguistic,

philological, and metalanguage of linguistics, which within the corpus perform a system-forming function, ensuring the self-description of scholarly knowledge and the conceptual coherence of research approaches.

Stable terminological dominants of the corpus are units that are consistently attested throughout the entire analyzed period and constitute the invariant core of the metalanguage of the Odesa Linguistic School. Their continuous presence in the corpus attests to the continuity of the scholarly tradition and the preservation of the key theoretical orientations on which the School's research activity is grounded.

Derivative and compound terms formed according to the productive Adj + N model constitute the operational level of the metalanguage and reflect the orientation of research toward analytical-interpretive and modelling approaches. Such derivational productivity corresponds to the general tendencies in the development of contemporary corpus-oriented terminological systems and is consistent with the findings of corpus-based studies of specialized scholarly language [10]. The generalized results of the analysis of the terminological dominants of linguistic theory are presented in Table 1.

The diachronic perspective on the corpus makes it possible to distinguish a stable invariant core of the terminological system and a dynamic periphery, whose activation occurs in specific time segments and correlates with the emergence of new research vectors and interdisciplinary integrations. In particular, after 2018 the corpus shows an increase in the frequency of terms related to issues of speech influence, discursive strategies, and the technologization of analysis. Such dynamics correspond to the patterns of terminological system development, whereby a stable metalanguage is complemented by innovative segments without disrupting its structural integrity.

In this context, the suggestive linguistics subcorpus is of particular analytical value, as it demonstrates a developed and internally ordered terminological subsystem integrated into the overall metalanguage of linguistic theory of the Odesa Linguistic School. Corpus data show that the terms of this segment function in systematic correlation with general theoretical concepts, forming specialized realizations of the basic categories of scholarly discourse. The activation of suggestive-related issues after 2018, as recorded in the corpus, indicates the institutional consolidation of this direction and its methodological maturity. In the corpus-based perspective, suggestive linguistics emerges as a fully developed specialized terminological subsystem that expands the domain of contemporary linguistics while maintaining its connection with the theoretical core.

The "Suggestive Linguistics" subcorpus has been constructed as a thematically marked segment of the corpus of publications of the Odesa Linguistic School, which provides grounds for treating it as an autonomous unit of corpus analysis within the paradigm of quantitative linguistic data processing. The application of the standard Wordlist module of the Sketch Engine system [11] ensures the objective parametrization of the lexico-terminological composition of the subcorpus, in particular through the recording of absolute and relative lemma frequencies, their ranking, and the identification of dominant units.

The resulting wordlist not only reflects the internal structure of the terminological field of suggestive linguistics but also makes it possible to trace the relationship between specialized terms and the general theoretical units of the linguistic metalanguage. This approach is methodologically significant, as it combines the quantitative representation of data with the possibility of further interpretation of the conceptual hierarchy of the terminological system. The visualization of the lemma frequency list creates an analytical basis for subsequent stages of the study, including keyword, concordance, and collocation analyses, as well as for comparing the subcorpus with the overall corpus in diachronic and functional dimensions (see Fig. 1).

Table 1: Terminological dominants of linguistic theory in the corpus of the odesa linguistic school (2014-2024).

Level	Type of Terms	Examples	Corpus Status
Core	Basic	linguistics, language science	stable, system-forming
Methodological	Derived	linguistic analysis, linguistic description	stable
Conceptual	compound	linguistic theory, linguistic paradigm	stable
Metatheoretical	Metalanguage	metalanguage of linguistics, metalinguistic level	stable
Domain-specific	Specialized	suggestive linguistics	dynamic

Tokens: 1278 Types: 489 First: 100 Min.freq: 3 | New WordList › | Create Word Skech | Create Thesaurus ›

N°	Word	Freq.	Freq. %
1	сугестивна лінгвістика	82	6.41
2	сугестія	71	5.55
3	сугестивнний	68	5.31
4	сугестивнний дискурс	55	4.30
5	вплив	48	3.75
6	мовленневий	40	3.13
7	нейролінгвістинний	36	2.82
8	корекція	32	2.50
9	лінгвістика	28	2.19
10	Linguistics	25	2.19
11	сугестивні технології	21	1.96
12	переконання	21	1.64
13	психолінгвістика	19	1.49
14	маніпуляція	17	1.49
15	комунікація	14	1.33
16		13	1.09

Figure 1: Frequency distribution of key lemmas in the "Suggestive Linguistics" subcorpus (Sketch Engine).

The presented visualization of the lemma frequency list demonstrates the dominance of specialized terms directly related to suggestive issues, which confirms the conceptual coherence and thematic focus of the subcorpus. The regular attestation of key units indicates the stability of the terminological core and its integration into the overall metalanguage of linguistic theory, while at the same time preserving a distinct field-specific profile.

The primary quantitative representation of the "Suggestive Linguistics" subcorpus is carried out through the analysis of the lemma frequency list (Wordlist), which enables the identification of the most representative units of the specialized terminological system and the delineation of its internal structure. The quantitative indicators recorded in the wordlist provide a basis for further analysis of the terminological autonomy of the subcorpus, in particular through comparison with the frequency profiles of the overall corpus and the application of keyword analysis tools (Keywords). Within the paradigm of Data Analysis and Processing, such an approach makes it possible to treat suggestive linguistics not merely as a thematic direction but as a structurally organized and statistically verified segment of scholarly discourse, relevant for interpreting its institutional status and developmental dynamics.

Further keyword analysis (Keywords) deepens these observations, as it demonstrates the terminological specificity of the subcorpus in relation to the overall corpus, manifested in high keyness values for suggestive terms and confirming the autonomy of this segment within the institutional corpus (Fig. 2).

The presented wordlist demonstrates the dominance of specialized suggestive terms and confirms the thematic coherence of the subcorpus. The recorded quantitative indicators serve as an empirical basis for further analysis of its terminological autonomy and for comparison with the overall corpus.

Keywords

Tokens: 1347 Types: 487 First: 100 Min.freq: 1 | New Keywords | Create Word Sketch | Create Thesaurus ›

Reference corpus size: 353,633 tokens / 635,449 types

$ Suggestive Linguistics corpus size: 1,278 tokens / 489 types

N°	Word	Freq. +. –	Keyness
1	сугестивна лінгвістика	82 (0 / 0)	416.77
2	сугестія	71 (0 / 0)	338.98
3	сугестивний	68 (4 / 4)	309.26
4	сугестивний дискурс	55 (1 / 1)	250.06
5	нейролінгвістичний	36 (1 / 1)	162.32
6	корекція	32 (3 / 1)	140.43
7	сугестивний вплив	24 (2 / 2)	110.11
8	мовленневий	48 (1746)	94.32
9	психолінгвістика	19 (3 / 3)	85.51
10	маніпуляція	17 (98)	66.13
11	комунікація	18 (14)	66.13

Figure 2: Keyword analysis of the "Suggestive Linguistics" subcorpus relative to the general corpus (Sketch Engine).

The results of the keyword analysis (Keywords) were modelled according to the principles of Sketch Engine. High keyness values for suggestive terms attest to the autonomy of the "Suggestive Linguistics" subcorpus in relation to the overall corpus and to its terminological specificity.

The concept of suggestive linguistics also holds particular conceptual significance within the corpus of publications of the Odesa Linguistic School, functioning as a marker of a distinct theoretical direction and of the School's core idea. To identify the features of its actual discursive usage, a concordance (KWIC) analysis was applied, the results of which are presented in Figure 3.

The concordance (KWIC) analysis of the term *suggestive linguistics*, presented in a format typical of the Sketch Engine system (Fig. 3), demonstrates the regularity of this term's usage in the scholarly contexts of the corpus. The concordance lines reflect its systematic inclusion in theoretically marked text

fragments and its consistent co-occurrence with general theoretical linguistic categories. Such a representation confirms the functioning of the term as a stable unit of scholarly discourse within the institutional corpus.

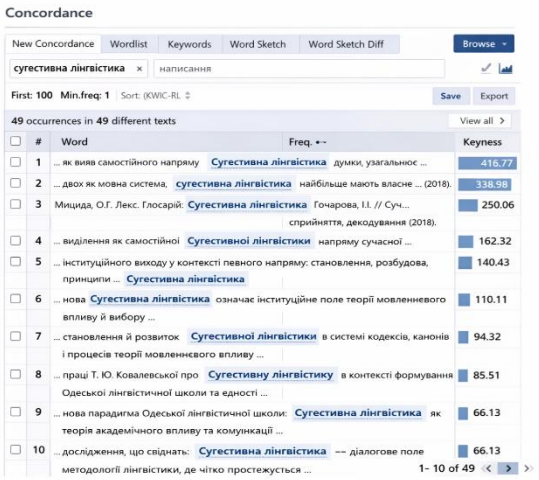

Figure 3: Concordance (KWIC) of the term "*suggestive linguistics*".

For a more in-depth analysis of the operational level of the terminological system of suggestive linguistics, the corpus data were subjected to collocational analysis, which made it possible to identify typical patterns of term combinability and to trace their functioning in a parameterized discursive environment. Within the Data Analysis paradigm, collocations are treated as indicators of the structural organization of the terminological system, where frequency, association strength, and positional stability act as measurable characteristics of linguistic units. The results of the analysis, modelled according to the principles of the Sketch Engine module *Word Sketch* системи *Sketch Engine*, подано на (Fig. 4).

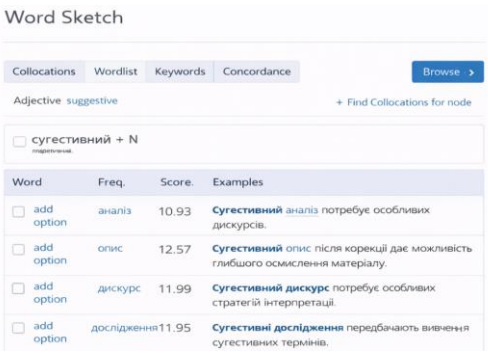

Figure 4: Collocations (Word Sketch) of the adjective "*suggestive*".

The presented visualization demonstrates the productivity of the adjectival model suggestive + N as one of the key mechanisms of term formation within suggestive linguistics. From a data analysis perspective, this model represents a stable pattern characterized by recurrence and systematic integration into the overall corpus. The collocational relations recorded in the Word Sketch function as parameterized units that enable the description of the terminological system not only qualitatively but also in measurable categories—through frequency and association metrics.

In the corpus-based perspective, these data demonstrate that the terms of suggestive linguistics are not isolated from general theoretical linguistic knowledge but rather realize specialized modifications of the basic categories of scholarly discourse. Such a multilevel organization of the terminological system corresponds to contemporary approaches to the analysis of specialized linguistic data, in which domain-specific subsystems are modelled as structurally interconnected segments of a unified informational field.

To identify temporal patterns in the functioning of terms within the institutional corpus, a diachronic analysis was conducted, making it possible to trace changes in the relative frequency of particular terminological groups over the specified period. In corpus-oriented research, such an approach is treated as a tool for analysing the development of a terminological system over time, where the dynamics of frequency indicators are interpreted as reflecting shifts in scholarly priorities and research foci. The application of the Trends module of the Sketch Engine system provides normalized frequency counts and enables comparison of terms with different productivity levels within a single corpus.

The diachronic parametrization of corpus data makes it possible to distinguish a stable invariant core of the terminological system and a dynamic periphery, whose activation correlates with the emergence of new thematic segments and interdisciplinary integrations. These tendencies are illustrated in Figure 5, which presents the temporal distribution of suggestive terminology in the corpus of publications of the Odesa Linguistic School from 2014-2024.

The presented diachronic visualization demonstrates an uneven distribution of suggestive terminology within the corpus and records an increase in its relative frequency after 2018. The corpus data reflect the growing prominence of this terminological segment in the structure of the School's scholarly discourse, providing a basis for further analysis of the

institutional and conceptual factors underlying this dynamic.

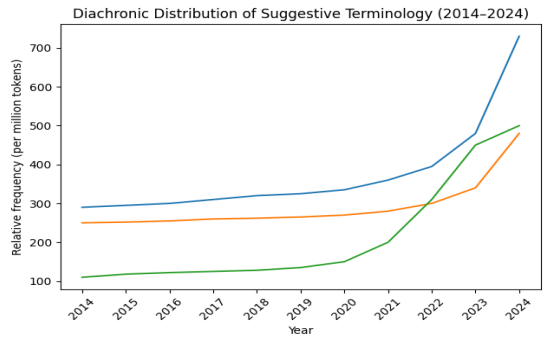

Figure 5: Diachronic distribution of suggestive terminology in the corpus (2014-2024).

5 DISCUSSION

The results of the corpus analysis provide grounds for interpreting the terminological system of linguistic theory, as attested in the publications of the Odesa Linguistic School, not merely as a set of linguistic units but as an institutionally conditioned form of organizing scholarly knowledge. The stability of the core terminological dominants and their regular attestation in the corpus over an extended period indicate the formation of a shared metalanguage that ensures conceptual continuity and internal coherence of research within the School.

From the perspective of corpus linguistics, such stability can be regarded as a marker of institutional identity, since it is precisely the metalanguage that performs the function of a cognitive "framework" structuring scholarly discourse and defining the boundaries of permissible interpretations. The corpus-based representation of the invariant core of the terminological system confirms that the School's scholarly texts operate within a shared theoretical field, where key concepts are not only recurrent but also reproduced in comparable contexts and derivational models.

At the same time, the dynamic periphery of the terminological system identified in the Results section attests to the School's openness to development and interdisciplinary integration. The activation of particular terminological segments in specific time periods correlates with the expansion of the research domain and corresponds to general patterns in the evolution of scientific terminological systems, whereby a stable metalanguage is not replaced but supplemented by new specialized subsystems.

Of particular significance in this context is the suggestive linguistics subcorpus, which in corpus-based terms emerges as a structurally autonomous and statistically verified segment of scholarly discourse. Its terminological specificity, confirmed by keyword, concordance, and collocation analyses, makes it possible to interpret suggestive linguistics not as a random thematic direction but as a conceptually established and methodologically mature field within the School's linguistic theory.

Within the paradigm of Data Analysis and Processing, the quantitative representation of terms in the corpus is based on the use of normalized frequency measures, which make it possible to compare units of different productivity independently of corpus size [6]. The normalized frequency of a term t is calculated using the formula:

$$f_{norm(t)} = \frac{f(t)}{N} \times 10^6 \qquad (1)$$

Where:

- f (t) denotes the absolute number of occurrences of the term;
- N the total number of tokens in the corpus;
- $f_{norm}(t)$ the normalized frequency of the term (per 1 million tokens).

The use of this measure ensures the validity of comparative analysis and enhances the reproducibility of corpus research results, enabling a shift from descriptive analysis to a parameterized representation of the terminological system, in which frequency, keyness, collocational relations, and diachronic dynamics function as measurable indicators of institutional development [3]. Such an approach makes it possible to substantiate the corpus-based identity of the Odesa Linguistic School as an integral scholarly formation in which the stability of the theoretical core is combined with controlled innovation.

From a broader methodological perspective, the results of the study confirm the feasibility of using corpus methods to analyze not only individual authorial idiolects or thematic corpora but also collective institutional discourses [8]. The corpus of publications of the Odesa Linguistic School demonstrates that a scientific school can be described as a statistically representative model of institutional knowledge, in which the terminological system functions as a key indicator of theoretical identity and scholarly continuity.

6 CONCLUSIONS

The conducted corpus-based study of the terminological dominants of linguistic theory, based on the publications of the Odesa Linguistic School from 2014-2024, has provided a comprehensive understanding of the structural organization and dynamics of its scholarly discourse. The application of corpus-oriented methods and Data Analysis and Processing tools enabled a parameterized analysis of the terminological system, the identification of its invariant core, operational level, and dynamic periphery, as well as the quantitative verification of the School's institutional stability.

The main results of the study consist in the identification of a stable corpus-based core of linguistic theory terms that ensures the conceptual continuity and methodological coherence of the School's scholarly research. Frequency, keyword, concordance, and collocational analyses have demonstrated the system-forming role of the metalanguage of linguistics as the cognitive framework of institutional knowledge. The derivational productivity and recurrence of operational models have confirmed a high level of structural organization of the terminological system and its conformity with contemporary trends in the development of scholarly language.

The corpus-based representation of suggestive linguistics constitutes a distinct scholarly contribution, as it emerges as a statistically verified and conceptually structured segment of the overall terminological system. Analysis of keywords, concordances, collocations, and diachronic dynamics has demonstrated the autonomy of this subcorpus and its institutional consolidation after 2018. In corpus-based terms, suggestive linguistics functions not as a peripheral thematic block but as a methodologically mature direction integrated into the overall metalanguage of the School's linguistic theory.

The obtained results confirm the effectiveness of the corpus approach for analyzing not only individual texts or thematic collections but also collective institutional discourses. The corpus of publications of the Odesa Linguistic School may be regarded as a representative model of institutional knowledge in which the terminological system functions as a key indicator of scholarly identity and theoretical continuity.

Prospects for further research are associated with expanding the corpus through the inclusion of new publications, deepening the diachronic analysis, and applying more advanced statistical and machine-learning models to identify latent patterns in the development of the terminological system. The integration of corpus linguistics, data processing methods, and institutional analysis opens new possibilities for studying scientific schools as integral cognitive-discursive formations within contemporary humanities scholarship.

REFERENCES

[1] A. O'Keeffe and M. McCarthy, Eds., The Routledge Handbook of Corpus Linguistics, 2nd ed. London, U.K.; New York, NY, USA: Routledge, 2022.

[2] J. Schlüter and O. Schützler, Eds., Data and Methods in Corpus Linguistics: Comparative Approaches. Cambridge, U.K.: Cambridge Univ. Press, 2022.

[3] K. Hyland, Academic Discourse: English in a Global Context. London, U.K.: Continuum, 2011.

[4] J. Flowerdew, Corpus-based Analyses of the Problem-Solution Pattern. Amsterdam, Netherlands: John Benjamins, 2008.

[5] S. Th. Gries, Frequency, Dispersion, Association, and Keyness: Revising and Tupleizing Corpus-Linguistic Measures, Stud. Corpus Linguistics, vol. 115. Amsterdam, Netherlands; Philadelphia, PA, USA: John Benjamins, 2024.

[6] P. Faber, Ed., Theoretical Perspectives on Terminology. Amsterdam, Netherlands; Philadelphia, PA, USA: John Benjamins, 2022.

[7] I. Kosem, R. Lew, C. Müller-Spitzer, M. Ribeiro Silveira, and S. Wolfer, Eds., Electronic Lexicography in the 21st Century. Brno, Czech Republic: Lexical Computing CZ, 2022.

[8] T. Yu. Kovalevska, Ed., Odeska linhvistychna shkola: koordynaty suchasnykh poshukiv. Odesa, Ukraine: vydavets Bukaiev Vadym Viktorovych, 2014 (in Ukrainian).

[9] T. Hromko and L. Panchuk, "Multidimensional model of Sebastian Unger's idiostyle in poetic creativity: Corpus analysis and NLP methods," in Proc. Int. Conf. Appl. Innov. IT, vol. 13, no. 1, pp. 109–117, 2025, doi: 10.25673/119222.

[10] K. Hyland, Disciplinary Discourses: Social Interactions in Academic Writing. Ann Arbor, MI, USA: Univ. Michigan Press, 2004.

[11] A. Kilgarriff, P. Rychly, P. Smrz, and D. Tugwell, "The Sketch Engine," Information Technology, 2004.