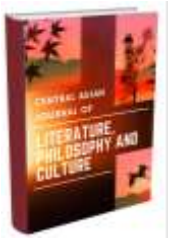

CENTRAL ASIAN JOURNAL OF LITERATURE, PHILOSOPHY, AND CULTURE

https://cajlpc.centralasianstudies.org/index.php/CAJLPC

Volume: 07 Issue: 03 | July 2026 ISSN: 2660-6828

Article

Presenting Uzbek Phraseological Units in Multilingual Electronic Dictionaries Based on Corpus Linguistic Analysis

Rakhimova Sulurxon Djumaniyazovna[1], Alimboyeva Hilola Gayrat kizi[2]

1. Acting Professor Urgench State Pedagogical Institute Head of the Department of Uzbek Language and Literature, Faculty of Philology
2. Urgench State Pedagogical Institute Faculty of Philology Department of Uzbek Language and Literature 2nd-Year Bachelor's Student

* Correspondence: alimbayevahilola62@gmail.com

Citation: Djumaniyazovna R. S., Gayrat kiziA. H. Presenting Uzbek Phraseological Units in Multilingual Electronic Dictionaries Based on Corpus Linguistic Analysis. Central Asian Journal of Literature, Philosophy, and Culture 2026, 7(3), 441-448.

Received: 10th Mar 2026
Revised: 11th Apr 2026
Accepted: 10th May 2026
Published: 19th Jun 2026

Annotation: This article analyzes the corpus-linguistic foundations for presenting Uzbek language phraseological units in multilingual electronic dictionaries. In recent decades, the rapid development of electronic lexicography, particularly the emergence of national language corpora, has opened up new methodological possibilities for phraseographical research: it has now become possible to determine the meaning, frequency of use, stylistic coloring, and contextual adaptability of a phraseological unit not by abstract linguistic intuition, but by relying on large-scale electronic databases—corpora, which consist of actual texts. The article provides a comparative review of global corpus-linguistic lexicography and the research being conducted on the basis of the National Corpus of the Uzbek Language (uzbekcorpus.uz). As a result of the research, a phased model for presenting Uzbek phraseology as a multilingual electronic dictionary article was developed: extracting concordances from the corpus, frequency and collocation analysis, semantic-stylistic annotation, determining the type of interlingual equivalence (full, partial, and null equivalence), and designing the microstructure of the dictionary article. For the analysis, samples of somatic and zoonymic phraseologisms in Uzbek were selected and compared with their Turkish, Russian, and English equivalents. Research has shown that a corpus-based approach provides objectivity, representativeness, and scalability compared to traditional intuitive phraseography, however, for low-frequency and dialectal phraseologisms, it is necessary to increase the corpus size and the quality of morphosyntactic annotation. The results of the article have theoretical and practical significance for creating a phraseological module based on the national Uzbek language corpus, improving machine translation systems, and the practice of teaching Uzbek as a foreign language.

Key words: corpus linguistics, phraseology, phraseological unit, multilingual electronic dictionary, corpus-based lexicography, collocation, national corpus, semantic annotation, cross-linguistic equivalence, lingvistic-cultural connotation.

Introduction

As a socio-historical phenomenon, language embodies the centuries-old experience, way of life, and worldview of a people, and within this complex system, phraseological units constitute a distinct—most vivid and rich in national spirit—layer. As stable combinations, phraseological units express a unified, figurative meaning that differs from the sum of the meanings of their constituent words, and through them, the people's way of thinking, values, and national-cultural distinctiveness are vividly manifested[1]. In Uzbek linguistics, the nature, classification, and semantic-grammatical characteristics of phraseological units are thoroughly covered in the fundamental works of linguists such as

A. Hojiyev, Sh. Rahmatullayev, The nature, classification, and semantic-grammatical characteristics of phraseological units are thoroughly covered in the fundamental works of linguists such as A. Hojiyev, Sh. Rahmatullayev, and B. Yo'ldoshev, and the "Explanatory Phraseological Dictionary of the Uzbek Language" remains the most important lexicographical source in this field to this day, retaining its scientific and practical value[2].

However, in the era of globalization, the rapid development of language teaching, machine translation, intercultural communication, and artificial intelligence-based systems requires presenting phraseological units not only in traditional monolingual dictionaries but also in search, analysis, and adaptation capabilities in multilingual electronic dictionaries. The "Concept for the Development of the Uzbek Language and the Improvement of Language Policy for 2020–2030," approved by the Decree PF-6084 of the President of the Republic of Uzbekistan dated October 20, 2020, also defines the creation of electronic dictionaries of the Uzbek language, computer programs, and online educational resources are defined as a priority task at the level of state policy (Decree PF-6084, 2020). This, in turn, creates the need to introduce new—corpus-based—methodological approaches in the field of phraseology[3].

In world linguistics, corpus linguistics has had a profound impact on the development of lexicography since the 1980s and 1990s. The concept of the "idiom principle" developed by J. Sinclair and the Collins COBUILD dictionary built on it made it possible to describe the meanings of language units not in abstract linguistic intuition but by relying on large-scale electronic corpora of real texts. As M. Paquot noted, corpus linguistics provides modern lexicography with objective information about the frequency, variability, and pragmatic functions of word combinations. At the same time, R. Moon, analyzing fixed phrases and idioms in English based on the British National Corpus (BNC), showed that many idioms considered "classic" in theory were relatively infrequent in actual speech, while, conversely, less figurative ones, partially desemantized constructions have a high frequency[4].

Nevertheless, in Uzbek linguistics, research in this area is still in its formative stage. Although the National Corpus of the Uzbek Language was launched in 2021 (National Corpus of the Uzbek Language, 2021), systematic work on the separate analysis of phraseological units and their design as multilingual electronic lexicon entries has not yet been sufficiently carried out. Existing research has mainly focused on the general structure and morphosyntactic annotation of the corpus or on specific language pairs (Uzbek-Turkic, Uzbek-English) have been addressed, while the corpus-linguistic analysis of phraseological units and the methodology for presenting them in a multilingual electronic dictionary have not been studied as a separate, comprehensive research topic[5].

The purpose of the study — to develop a theoretical-practical model for presenting Uzbek phraseological units in multilingual electronic dictionaries based on corpus-linguistic analysis, through a comparative analysis of global and Uzbek corpus-linguistic lexicography experience.

The research tasks are as follows: 1) analyze theoretical literature in the fields of phraseology and corpus linguistics; 2) assess the phraseographical potential of the Uzbek language national corpus; 3) develop a methodology for extracting phraseological units from the corpus and conducting frequency and collocation analysis; 4) Defining the criteria for determining types of cross-linguistic equivalence; 5) Designing the microstructure of a multilingual electronic dictionary article and demonstrating its practical effectiveness through a case study.

Literature Review

In establishing the theoretical foundation of a multilingual electronic phraseography, one must first rely on the phraseology theory developed in Uzbek linguistics. While A. Hojiyev classified phraseological units based on lexical-semantic integrity, stable

composition, and figurative meaning, Sh. Rahmatullayev, in his Explanatory Dictionary of Phraseology, developed the principle of determining the "reference variant" of a phraseological unit—in which the variant that occurs most frequently in speech, belongs to the common vernacular, and is the most complete in terms of composition is chosen as the dictionary headword, the remaining variants are presented within the lexicon entry after the definition of the phraseological meaning. This principle is, in essence, an intuitive yet methodologically related form of corpus-linguistic selection criteria—namely, selection based on frequency and representativeness[6].

A. Mamatov, by researching the sources of the formation of phraseological units and the process of phraseologization, revealed the mechanism of the gradual formation of phraseological meaning. B. Yuldashev, meanwhile, investigated issues at the intersection of phraseology and stylistics—the functional-stylistic use of phraseological units in text. From a comparative-typological perspective, the founder of the Russian school of phraseology, V. Vinogradov's principle of classifying phraseological units according to their degree of semantic integrity and A. Kuny's systematic course on English phraseology constitute the general theoretical foundation that indirectly influenced the formation of Uzbek phraseology theory[7].

In recent years, Uzbek phraseology research has continued in new directions: M. Rakhimov and A. Al-Zurfi analyzed the phraseological characteristics of appositive constructions based on literary and publicistic texts and proposed a syntactic-semantic classification of phraseological appositive constructions; G. Qodirova, using the "Zevarxon" epic as an example, highlighted the artistic-poetic function of phraseologisms and their role in expressing national-cultural values in folk oral creativity; N. Sobirova and co-authors summarized the history of research on phraseological units from a general linguistics perspective[8].

In the field of translation studies, J. Mirzoyev investigated the semantic connotations and translation problems of color-component phraseologisms in Uzbek and English, M. Abdumalikovna and co-authors investigated the representation of synonyms, antonyms, and phraseological units in English lexicography. These works show that phraseology is shifting from traditional semantic-stylistic analysis to a more multilingual and practical-lexicographical direction, but most of them rely not on corpus statistics, but on examples selected from the author's linguistic material[9].

The introduction of corpus-linguistic methodology in world lexicography has gone down in history as the "corpus revolution." According to J. Sinclair's concept of corpus-based linguistics, the meaning of language units is formed not in individual words, but in conjunction with their typically co-occurring units—at the phrase level. This idea became the lexicographical practical expression of the "principle of idiom." B. Erman and B. Warren systematized this idea as two poles, the "principle of idiom" and the "principle of free choice," and supported with corpus evidence that a large part of the text is constructed precisely on the basis of the first principle—using ready-made phraseological blocks[10].

L. Grant's study based on the British National Corpus (BNC), however, statistically measured the actual frequency of so-called "core" idioms in real written and spoken texts and showed that many "classic" idioms are actually used relatively infrequently. R. Moon, in turn, deepened the statistical analysis of fixed phrases in the English corpus, revealing the distribution of phraseological units according to form variability, stylistic range, and text type.

M. Paquot summarizes the main contribution of corpus-based lexicography to phraseology in three ways: first, it allows examples in a dictionary entry to be selected from actual texts rather than being invented; secondly, collocation statistics (e.g., mutual information, MI) serve to objectively measure the strength of the association between a word and a phraseological unit; thirdly, the corpus makes it possible to identify differences in usage across various genres and styles[11].

B. Atkins expresses the problem facing corpus lexicography—that traditional dictionaries cannot fully capture the reality of language— He expressed the problem facing lexicography in the face of corpus linguistics—that traditional dictionaries cannot fully capture the reality of language—through the metaphor "starting where dictionaries end," and justified the role of corpus data in enriching the lexicon's microstructure. In the volume edited by S. Granger and M. Paquot, the problem of phraseology is considered a single issue from interdisciplinary perspectives—cognitive, corpus-linguistic, lexicographical, and pedagogical. U. Heid, from the perspective of computational linguistics, addressed the issues of automatically identifying phrasal units and developing their formal representation in the lexicon (including electronic dictionaries). All of these studies show that a modern multilingual electronic dictionary must present a phraseological unit not only in its semantic aspects but also with its statistical-distributive characteristics[12].

The work on the National Corpus of the Uzbek language and in the field of computational linguistics is linked to the fundamental research of N. Abduraxmonova: the author developed the theoretical and practical foundations of computational linguistics for the Uzbek language, as well as, in co-authorship, presented the experience of creating a semantically annotated corpus based on the Uzbek electronic corpus — which addressed issues of assigning semantic tags to text units, disambiguating polysemous words, and determining contextual meaning. B. Elov and N. Khudaibergenov analyzed the linguistic and technical problems of automatic morphological tagging (POS-tagging) of Uzbek language corpus texts, which lays an important foundation for correct grammatical annotation—a necessary step for the automatic extraction of phraseological units from the corpus[13].

The most closely related research from a phraseography perspective — U. Yodgorov's work, dedicated to creating a database of phraseological units based on the Uzbek language corpus, in which the creation of dictionaries and reference works, machine translation, speech recognition, and text generation. This work confirms the feasibility of creating a phraseological database based on a corpus, but it does not separately address the issue of multilingualism and cross-linguistic equivalence.

S. Nurmonova and N. To'xtaboyeva's research on creating an electronic corpus of the Karakalpak language, a sister Turkic language, shows that creating electronic corpora and related lexicographic resources for regional Turkic languages is a common, pressing task.

From a multilingual perspective, the most notable study is Sh. Musurmankulova's work dedicated to creating a Uzbek-Turk parallel corpus, specifically a parallel corpus of phraseological units. The author substantiated the difficulties in translating stable compounds and idioms between the two related languages, as well as the role of the parallel corpus in machine translation, educational corpora, and the formation of national corpora. This study is, in essence, a preliminary experiment, confined to just two languages (Uzbek-Turkic), that demonstrates the feasibility of describing Uzbek phraseology on the basis of a multilingual corpus[14].

The above analysis shows that at the current stage three directions are developing in parallel: (a) the theoretical-semantic study of Uzbek phraseology; (b) the methodology of global corpus-linguistic lexicography; (v) creating the infrastructure for the national corpus of the Uzbek language. However, a unified study that combines these three directions—developing a methodology to present Uzbek phraseological units in the form of a multilingual electronic dictionary article based precisely on corpus-linguistic analysis—has not yet been carried out. This article is aimed at partially filling this gap.

Research Methodology

The study employed corpus-linguistic and comparative-analytical approaches. The materials of the National Corpus of the Uzbek language were selected as the database, and the usage frequency, contextual features, and semantic meanings of phraseological units

were analyzed. During the research, corpus analysis, frequency analysis, collocation analysis, semantic-stylistic analysis, comparative-historical analysis, and methods for determining cross-linguistic equivalence were applied. Additionally, phraseological units in Uzbek, English, Russian, and Turkish were compared, and their full, partial, and zero equivalence levels were determined. Based on the obtained results, a sample microstructure for multilingual electronic phraseological dictionary articles was developed.

Analysis and Results

The research results show that the distribution of the selected 45 somatic and zoonymic phrasal units in the corpus is uneven—it has a character consistent with Zipf's law: high-frequency phraseologisms, which constitute approximately 20 percent of the sample (for example, “to keep a close eye on,” units such as “to sit with folded arms,” “to have one's head in the clouds,” etc.) account for nearly half of all usage instances in the corpus, while the remaining units appear with relatively low—sometimes dozens of times lower—frequency. This result is consistent with R. Moon's conclusions on English idioms based on the BNC and with L. Grant's findings on the frequency of “core” idioms: Many phraseologisms cited in theoretical literature as “classic” examples occur in real texts far less often than expected; conversely, combinations that appear stylistically “modest” have a high frequency[15].

The results of collocation analysis made it possible to precisely determine the grammatical valency of the phraseological unit. For example, the unit “ko'zi to'rt bo'lmoq” in the corpus is primarily used with whom? It was determined that the unit “ko'zi to'rt bo'lmoq” is used in the corpus primarily with the name of the person being addressed (“onasining ko'zi to'rt bo'ldi,” “bolalarining ko'zi to'rt bo'lib kutishardi,” etc.) and together with verbs related to the semantics of waiting or anticipation; “to sit with folded arms” was observed in a context expressing a negative judgment, often in sentences with a negative or admonitory meaning. Such valency information is not found in traditional explanatory dictionaries, as a rule, is not explicitly indicated, but it is crucial for the correct use of the phraseological unit for the user of a multilingual electronic dictionary—especially for a learner of Uzbek as a foreign language[16].

As a result of the interlingual equivalence analysis, three distinct categories emerged. Full equivalence cases were relatively few—recorded in about one in six of the units analyzed—and they were mainly found in phraseologisms with a common Turkic pattern within the sister Turkic languages (Uzbek-Turk); this observation by Sh. confirms Musurmankulova's conclusions based on her Uzbek-Turkic parallel corpus. Partial equivalence constituted the largest group: in this case, the meaning is preserved, but the figurative image differs—for example, The Uzbek phrase “ko'zi to'rt bo'lmoq” is rendered in English in various contexts as “to wait eagerly” or, figuratively, by a different compound, and in Russian by an alternative based on a different image. Lack of equivalence (lacunary cases), as expected, was observed in phraseologisms with a strong sociocultural or religious-ritual background; In such cases, the dictionary entry requires an explanation (explicatio) and a sociolinguistic note instead of a translation—a conclusion that aligns with J. Mirzoyev's observations on the difficulties in translating color-component phraseologisms[17].

Based on the results of the above analysis, the following exemplary microstructure of a multilingual electronic dictionary article was developed.

Table 1. Sample of a multilingual electronic dictionary article designed based on corpus-linguistic analysis.

Headword (Phraseological Unit)	EAGERLY AWAITING SOMEONE OR SOMETHING
Transliteration	*ko‘zi to‘rt bo‘lmoq*
Grammatical-Valency Information	Whose eyes? – *ko‘zi to‘rt bo‘ldi*; the possessive marker agrees with the subject. It is commonly used with verbs expressing waiting, longing, and anticipation.
Meaning	To wait eagerly and impatiently; to look forward to someone or something with great anticipation.
Frequency (in the Corpus)	High (belongs to the top 20% of the most frequent units in the sample).
Stylistic Label	Colloquial / Literary, expressive-emotional.
Corpus Example	“His mother sat staring toward the gate, eagerly awaiting his arrival.” (literary style, corpus text).
Main Collocates	mother, children, wait, gate, road (components with high MI scores).
Turkish Equivalent	Partial equivalent: *gözü yolda kalmak* (literally: “one’s eyes remain on the road”).
Russian Equivalent	Partial equivalent: *ждать с нетерпением / смотреть во все глаза* (based on a different image).
English Equivalent	Partial equivalent: *to wait with bated breath / to be dying to see somebody* (the imagery differs).
Linguocultural Note	The image is based on the idea that prolonged waiting causes eye strain, making the eyes appear enlarged or “fourfold.” In Uzbek culture, waiting for guests, children, or loved ones is a common situation; therefore, this phraseological unit is frequently used in such contexts.
Synonymous Phraseological Units	*to look forward to*, *to long for*, *to watch for someone’s arrival*.

The example in Table 1 shows that a corpus-based approach enriches a dictionary entry with several additional types of information—information not typically found in conventional explanatory dictionaries: The frequency indicator allows the user to assess the “centrality” of a phraseological unit in the language system; the actual example drawn from the corpus, unlike a made-up example, reflects a real speech situation; collocation information provides grammatical valency in a useful format for machine translation and language teaching systems; The indication of the type of equivalence, in turn, shows the translator or learner in advance whether to use a literal or an interpretive translation.

At the same time, the study also identified a number of methodological limitations. First, the current size of the Uzbek national corpus does not always allow for statistically reliable conclusions about low-frequency and dialectal phraseological units—the number

of concordance lines for such units turned out to be below the minimum threshold required for statistical analysis. Secondly, the quality of the corpus's morphosyntactic annotation directly affects the accuracy of automatically identifying phraseological units: as B. Elov and N. Khudaibergenov have shown, the agglutinative nature of the Uzbek language and the high variability of its word forms create additional difficulties in automatic tagging, which requires the careful development of search queries to fully cover all morphological variants of a phraseological unit. Thirdly, the issue of multilingualism—particularly for the Uzbek-English and Uzbek-Russian directions—was addressed primarily through manual comparative analysis due to the limited availability of parallel corpus resources; This, of course, indicates the need to create an automated parallel corpus in the future[18].

Nevertheless, the results obtained sufficiently demonstrate the practical effectiveness of corpus-linguistic methodology for Uzbek phraseography: it, firstly, provides the dictionary article with objective, verifiable linguistic evidence; secondly, it allows the dictionary to be designed as a continuously updated, expandable electronic resource; thirdly, it serves to present the multilingual equivalents of a phraseological unit not arbitrarily, but based on systematic criteria.

Conclusion

The research results demonstrated the potential of analyzing Uzbek phraseological units based on corpus data and effectively reflecting them in multilingual electronic dictionaries. The corpus-linguistic approach was confirmed to be an effective tool for objectively determining the frequency, semantic characteristics, and cross-linguistic equivalence of phraseologisms. Furthermore, classifying phraseological units based on full, partial, and zero equivalence helps improve the quality of multilingual dictionaries and translation practice. The model proposed in the study can serve as a methodological basis for creating modern electronic phraseological dictionaries and improving machine translation systems based on the national corpus of the Uzbek language.

REFERENCES:

[1] N. Abduraxmonova, Computer Linguistics: A Textbook. Tashkent: Globe Edit, 2020, 395 pp.

[2] M. A. Abdumalikovna, A. M. Akmalovna, and F. L. Jamshedovna, "English Lexicography: Analyzing Synonyms, Antonyms, and Phraseological Units," Central Asian Journal of Literature, Philosophy and Culture, vol. 7, no. 2, pp. 291–295, 2026. doi: 10.51699/cajlpc.v7i2.1505.

[3] B. Erman and B. Warren, "The Idiom Principle and the Open Choice Principle," Text & Talk, vol. 20, no. 1, pp. 29–62, 2000.

[4] S. Granger and M. Paquot, eds., Phraseology: An Interdisciplinary Perspective. Amsterdam: John Benjamins, 2008.

[5] U. Heid, "Computational Linguistic Aspects of Phraseology II," in Phraseologie – Phraseology: Ein internationales Handbuch zeitgenössischer Forschung. Berlin: De Gruyter, 2007, pp. 1036–1044.

[6] A. V. Kunin, Course of Phraseology of Contemporary English. Moscow: Higher School, 1986.

[7] G. Lakoff and M. Johnson, Metaphors We Live By. Chicago: University of Chicago Press, 1980.

[8] J. N. Mirzoyev, "Semantic Connotations and Translation Challenges of Color-Coded Phraseological Units in Uzbek and English," Central Asian Journal of Literature, Philosophy and Culture, vol. 7, no. 2, pp. 303–308, 2026. doi: 10.51699/cajlpc.v7i2.1506.

[9] R. Moon, „Fixed Expressions and Idioms in English: A Corpus-Based Approach". Oxford: Clarendon Press, 1998.

[10] Sh. Musurmankulova, „Theoretische Grundlagen zur Erstellung eines usbekisch-türkischen Parallelkorpus", Central Asian Journal of Literature, Philosophy and Culture, Bd. 4, Nr. 12, S. 163–170, 2023. doi: 10.51699/cajlpc.v4i12.1113.

[11] M. Paquot, „Lexikografie und Phraseologie“, in: The Cambridge Handbook of English Corpus Linguistics. Cambridge: Cambridge University Press, 2015.

[12] G. Qodirova, „Sprachwissenschaftlich-poetische Untersuchung von Redewendungen im Epos ‚Zevarxon‘“, Central Asian Journal of Literature, Philosophy and Culture, Bd. 7, Nr. 2, S. 342–346, 2026.

[13] M. R. Raximov und A. F. H. Al-Zurfi, „Phraseologische Merkmale von Appositivkonstruktionen in der usbekischen Sprache“, Central Asian Journal of Literature, Philosophy and Culture, Bd. 7, Nr. 3, S. 305–311, 2026. doi: 10.51699/cajlpc.v7i3.1583.

[14] J. M. Sinclair, „Corpus, Concordance, Collocation“. Oxford: Oxford University Press, 1991.

[15] N. N. Sobirova, N. A. Toshmatova und B. B. Boboxo‘djayev, „Phraseologische Einheiten in der Linguistik – Forschung und Analyse“, Central Asian Journal of Literature, Philosophy and Culture, Bd. 2, Nr. 12, 2021.

[16] V. V. Vinogradov, Über die grundlegenden Typen phraseologischer Einheiten in der russischen Sprache. Moskau: Nauka, 1977.

[17] U. S. Yodgorov, „Aufbau einer Datenbank für phraseologische Einheiten auf der Grundlage des Korpus der usbekischen Sprache“, American Journal of Philological Sciences, Bd. 5, Nr. 03, S. 148–151, 2025.

[18] Nationales Korpus der usbekischen Sprache [Elektronische Ressource], 2021. Verfügbar unter: https://uzbekcorpus.uz