The OSCAR PUBLISHING Services
American Journal of Philological Sciences

VOLUME Vol.06 Issue02 2026
PAGE NO. 14-21
DOI 10.37547/ajps/Volume06Issue02-05

Corpus Linguistics And A Mixed Monolingual–Parallel Methodology For Investigating Uzbek – English Action Verbs Synonyms

Gulhida Hasanova
Doctoral student (PhD) in Comparative Linguistics, Uzbekistan State World Languages University, Uzbekistan

Received: 12 December 2025; **Accepted:** 04 January 2026; **Published:** 07 February 2026

Abstract: Corpus linguistics is a branch of linguistics concerned with compiling naturally occurring language data in electronic form and analyzing it systematically. A corpus is understood as a collection of authentic texts that is sufficiently representative to support reliable generalizations about a linguistic unit or variety. The methodological value of corpora lies in their scale and authenticity, which provide extensive empirical evidence that can validate claims that might otherwise rely on subjective intuition. This empirical orientation is especially important for synonym research because near-synonyms often differ in collocation, phraseological behavior, stylistic distribution, pragmatic constraints, and evaluative meaning, distinctions that are not always visible in dictionary definitions. This paper presents a corpus-based research design for analyzing Uzbek and English action-verb synonym sets through a combined methodology: monolingual corpora are used for intralingual semantic profiling, while a custom parallel corpus is compiled for translation-equivalent analysis. The study operationalizes lexical–semantic distinctions via collocation measures (Mutual Information and t-score), multiword pattern extraction aligned with the idiom principle, semantic prosody analysis using a two-stage procedure, and reliability control through Cohen's kappa. In addition, validity safeguards based on corpus representativeness, normalization, and dispersion are integrated into the analysis. The translation component follows corpus-based translation studies and descriptive translation studies frameworks to quantify translation equivalents, identify asymmetries, and interpret translation strategy tendencies.

Keywords: Corpus linguistics; synonymy; near-synonyms; Uzbek action verbs; English action verbs; parallel corpus; collocation analysis; Mutual Information (MI); t-score; idiom principle; semantic prosody; lexical semantics; translation equivalence; asymmetry in translation; corpus-based translation studies; frequency normalization; dispersion (DP); concordance analysis.

Introduction: Corpus linguistics investigates language by compiling naturally occurring linguistic units in electronic form and analyzing them systematically. A corpus is a collection of naturally occurring texts that is sufficiently large and representative to enable a researcher to form an evidence-based understanding of a given linguistic unit or variety. The importance of corpora for linguistic research lies in the authenticity and high volume of the texts that constitute them, because these features provide the researcher with large-scale empirical evidence. Empirical evidence should serve as the primary basis of linguistic analysis, since a researcher's intuition about language is subjective and may not reliably correspond to actual language use. For this reason, theoretical claims and analytical conclusions should be supported by extensive empirical data rather than by introspection alone [Sinclair, 1991; Biber et al., 1998; McEnery & Hardie, 2012].

This requirement is especially relevant for synonym research. The lexical–semantic properties of synonyms—particularly the distinctions that differentiate members of a synonym set—are not always evident through dictionary definitions alone.

Near-synonyms often exhibit differences in combinability with other words, typical contextual environments, stylistic distribution, and pragmatic licensing conditions [Cruse, 1986; Apresyan, 1995]. Uzbek linguistics has also emphasized that synonymy should be approached as a systematic phenomenon characterized by a semantic dominant and by differentiating features that emerge in usage [Hojiev, 1985; Shoabdurahmonov, 1999]. Consequently, corpus methods are necessary for revealing the lexical–semantic differences among synonyms through repeated authentic contexts, distributional evidence, and phraseological patterns.

Contemporary corpus linguistics recognizes multiple corpus types, differentiated by design criteria. One key criterion is the number of languages represented. Corpora may be monolingual, parallel, or multilingual. Monolingual corpora are the most common type, consisting of texts in a single language. Parallel corpora consist of authentic source-language texts aligned with their translations into another language; for example, English-language works aligned with Uzbek translations. The distinctive feature of parallel corpora is that the same content and its context can be observed simultaneously in two languages, enabling the researcher to track lexical choice and equivalence patterns across languages [Johansson, 2007; Baker, 1995]. Parallel corpora may also include more than two languages, in which case they are often referred to as multilingual corpora [Johansson, 2007]. Another relevant type is the comparable (contrastive) corpus, which contains texts in two or more languages that share key properties (e.g., topic, register, genre) but are not translations of each other. For instance, a large collection of English, Russian, and Uzbek articles on COVID-19 could form a comparable corpus. The key difference is that parallel corpora are built from originals and their translations, whereas comparable corpora are built from texts that share a linking feature or a set of features but are independently produced [Biber et al., 1998; McEnery & Hardie, 2012].

Corpora also differ by temporal coverage. Diachronic corpora include texts from different historical periods, enabling researchers to identify linguistic transformations over time. Synchronic corpora, in contrast, represent language within a restricted time span and support the description of contemporary usage [McEnery & Hardie, 2012]. Finally, corpora differ by update policy. Static corpora do not change over time, which supports replicability: results obtained from a fixed dataset remain stable. Dynamic corpora are continuously updated and are therefore useful for tracking recent linguistic trends and innovations [Kennedy, 1998; Biber et al., 1998].

The objectives of the present research require the use of both monolingual and parallel corpora. Monolingual corpora are essential for analyzing the semantic and usage-based properties of each synonym in each language, because monolingual datasets typically provide broader coverage and larger size than other corpus types. Corpus size, in turn, supports the credibility and stability of findings by minimizing the role of chance distribution. For English intralingual analysis, the study uses the Corpus of Contemporary American English (COCA), a large corpus that is regularly expanded with new texts, making it suitable for profiling current usage patterns [Davies, 2008]. For Uzbek intralingual analysis, the study uses the O‘zbek tili korpusi, which is currently among the largest available Uzbek corpora and provides the best structured foundation for Uzbek usage analysis [To‘ychiyeva, 2020]. However, because Uzbek corpus resources remain smaller than COCA, Uzbek intralingual analysis may require supplementation with additional Uzbek literary materials not yet included in the Uzbek corpus. This supplementation is treated as methodologically justified when coverage is insufficient, and it must be transparently documented.

Cross-linguistic comparison of Uzbek and English action-verb synonym sets requires a contrastive perspective. Because a unified, balanced Uzbek–English comparable corpus is not currently available, the study operationalizes a comparable design through structured comparisons between the two monolingual corpora, while paying careful attention to differences in size and composition and using normalized frequency measures where appropriate.

The translation-focused part of the study is conducted using a parallel corpus. Specifically, the research employs a custom parallel corpus built from Uzbek literary works translated directly into English and English literary works translated directly into Uzbek. This parallel corpus is compiled in Sketch Engine and is aligned only at the level of paragraphs and sentences that contain the target synonym nodes. This focused alignment strategy reduces annotation burden while preserving the analytic core of the translation dataset [Kilgarriff et al., 2014].

The contemporary status of corpus linguistics in linguistic research has been achieved through the work of many scholars, but the foundations of corpus methodology are particularly strongly associated with John Sinclair. Sinclair’s Corpus, Concordance, Collocation is regarded as a foundational work in corpus linguistics and presents key concepts that remain central to corpus-based lexical–semantic analysis, including collocation, semantic prosody, and the idiom principle [Sinclair, 1991]. Sinclair argues that

meaning cannot be adequately explained by dictionary definitions or grammar alone; rather, meaning emerges in patterned usage and must be studied through authentic contexts and recurrent combinations. By analyzing large numbers of authentic examples, Sinclair shows that corpus evidence can reveal lexical properties and meaning components that cannot reliably be discovered through intuition alone. This logic aligns with Apresyan's lexical-semantic framework, which emphasizes that synonym differences may derive from hidden semantic components (semes) that are not always explicitly represented in dictionary entries. Apresyan argues that synonym distinctions may manifest through stylistic specialization, combinability restrictions, and pragmatic factors; therefore, these semes can only be identified through systematic analysis of authentic usage patterns [Apresyan, 1995]. Because the present study aims to demonstrate lexical–semantic differences among synonyms explicitly, it treats combinability and phraseological patterns as essential dimensions of synonym differentiation.

METHODS

The research design integrates two major components: (1) intralingual analysis using monolingual corpora, and (2) translation-oriented analysis using a parallel corpus. Monolingual corpora are used to analyze the semantic and usage-based properties of each synonym within a language, while the parallel corpus is used to analyze translation equivalents, distributional asymmetries, and translation strategy tendencies [Baker, 1995; Laviosa, 2002; Toury, 1995].

For English, COCA is selected as the primary monolingual corpus because it is large, systematically designed, and continuously updated, which makes it suitable for profiling contemporary American English usage [Davies, 2008]. For Uzbek, the O'zbek tili korpusi is selected because it provides the largest available structured dataset for Uzbek usage analysis [To'ychiyeva, 2020]. Because the Uzbek corpus is smaller than COCA, Uzbek intralingual analysis may be supplemented with additional Uzbek literary texts when the corpus does not provide sufficient coverage for particular target items. Such supplementation is treated as necessary for ensuring adequate empirical support, and all added materials are documented transparently.

To compare Uzbek and English action-verb synonym sets, the study employs a contrastive corpus design. Ideally, this would be supported by a unified Uzbek–English comparable corpus. Because such a resource is not currently available, the study approximates the comparable design through careful comparisons between the two monolingual corpora. In practice, this means comparing intralingual profiles across languages while controlling for genre/register and using normalized frequencies to mitigate differences in corpus size [Biber et al., 1998].

The translation component of the study is conducted through a custom bidirectional parallel corpus compiled in Sketch Engine from (i) Uzbek originals translated directly into English and (ii) English originals translated directly into Uzbek, primarily within the fiction domain. The corpus is aligned only at the level of paragraphs and sentences containing the target synonyms. This method is designed to focus alignment effort on analytically relevant portions of the dataset while maintaining reliable extraction of translation equivalents [Kilgarriff et al., 2014].

To identify intralingual collocational properties of each node (target word), the study applies two methods associated with Sinclair's corpus approach: Pointwise Mutual Information (MI) and t-score [Sinclair, 1991]. MI is used to identify strong associations between the node and its collocates. For each occurrence of the node, all words occurring within a window of two words to the left and two words to the right are treated as potential collocates. The mean expected co-occurrence is calculated using the frequency of the node and the frequency of the collocate within the corpus and dividing their product by the total number of corpus tokens.

The example provided in this study is as follows:

- Total corpus tokens = 10,000
- Frequency of chopmoq = 450
- Frequency of halloslab = 180

Mean expected co-occurrence:

$MI = (450 \times 180) / 10{,}000 = 8.1$

Thus, the expected (mean) co-occurrence of chopmoq and halloslab is approximately 8. The next step is to calculate the observed number of cases where the two occur together and compare it with the expected value. In the example:

Observed co-occurrences of chopmoq + halloslab = 103

MI (log form) is computed as:

$MI = \log_2(103 / 8) = \log_2(12.875) \approx 3.68$

This indicates that the observed co-occurrence is 12.875 times greater than expected by chance. Therefore, chopmoq and halloslab are not a random adjacency but form a meaningful collocational pair. The use of the logarithmic form facilitates comparison across different word pairs.

Approximate interpretation of MI values on a

logarithmic scale:

- 2 times above expected → MI ≈ 1
- 4 times above expected → MI ≈ 2
- 8 times above expected → MI ≈ 3
- 16 times above expected → MI ≈ 4
- 32 times above expected → MI ≈ 5

Interpretation of MI values:

- MI = 0 → no association
- MI > 3 → stable and noticeable collocational tendency
- MI > 5 → strong association, potentially characteristic of idiomatic combinations
- MI < 0 → extremely low probability of co-occurrence in the same context

While MI helps identify semantically strong associations, Sinclair's second method—t-score—captures how frequently and consistently the pair is used in the language, favoring stable high-frequency combinations [Sinclair, 1991; Evert, 2005]. The calculation is given as:

t-score = (Observed frequency − Expected frequency) / √(Observed frequency)

For the chopmoq + halloslab pair:

t-score = (103 − 8) / √103 ≈ 9.36

Interpretation of t-score values:

- 2 → low probability of a meaningful collocation
- 3 → noticeable collocation
- 5 → frequently used, clearly established collocation
- 6–7+ → stable, widely distributed typical collocation in the corpus

Based on these results, the study draws two key conclusions regarding halloslab chopmoq:

1. The two words have a strong collocational association; they are not merely two adjacent words in a sentence (MI).

2. The combination is frequently and systematically used in Uzbek as an established unit (t-score).

These conclusions provide evidence that the action verb chopmoq carries a semantic component associated with physical exertion. Using the same procedure, the study identifies additional collocates of chopmoq and interprets the semantic features that motivate those combinations. All synonyms in the same action-verb set are analyzed via the same steps, and comparison across the results makes it possible to define the shared and differentiating lexical–semantic properties of the synonym set.

The procedure for identifying collocations in corpus material is implemented through the following steps:

1. For each node, select a window of two words on the left and two on the right within each concordance line (four total positions), and treat these as potential collocates.

2. Filter collocates: retain only those that co-occur with the node at least 5 times (for large corpora) or at least 10 times (for smaller corpora).

3. Compute MI and t-score for each candidate collocate pair.

4. Select the top 40 collocates by MI and the top 40 collocates by t-score, then compare the two lists. This comparison identifies (a) rare but strongly associated collocates and (b) frequent and stable collocates typical of the node's conventional usage.

After all synonyms in a set are analyzed through the same workflow, the shared and differentiating properties are identified. For cross-linguistic comparison, the Uzbek and English synonym sets are compared based on the same analytic outputs.

One of Sinclair's most influential proposals is the idiom principle, which argues that speakers produce language not primarily by assembling single words in isolation but by drawing on semi-preconstructed multiword units, patterns, and constructions that recur frequently in use [Sinclair, 1991]. Such constructions may consist of two words or more. Conventional compatibility and incompatibility between words and constructions can reveal semantic boundaries and hidden meaning components. Therefore, the present study also analyzes the phraseological frames in which action verbs occur.

Sinclair's procedure for extracting such multiword units is operationalized as follows:

1. Extract frequent 3-, 4-, and 5-word sequences (n-grams) containing the node.

2. Retain only those n-grams that occur more than 10 times and have MI values above 3.

However, because Uzbek is an agglutinative language and exhibits greater word order flexibility than English, extracting contiguous 3–5 word sequences may not be sufficient for capturing all verb-centered multiword constructions in Uzbek. Therefore, in addition to n-grams, the study identifies potential constructions using corpus query tools with patterns such as complement + verb, subject + verb, and adverbial modifier + verb. These construction-based searches are intended to improve coverage of Uzbek verb-centered phraseological behavior and to ensure that

syntactically relevant realizations are included in the analysis [Shoabdurahmonov, 1999].

Another key concept emphasized by Sinclair is semantic prosody, the idea that a word's meaning becomes especially visible when it is studied through its most typical contexts rather than in isolation [Sinclair, 1991; Louw, 1993]. For synonym research, semantic prosody is important because words with similar denotational meaning often differ in connotation and emotional–evaluative coloring, and these differences can emerge through recurrent contextual environments [Stubbs, 2001].

In this study, semantic prosody is identified through a two-stage procedure:

First, a list of the top 40 collocates is compiled based on highest MI and t-score values. These collocates are then categorized as positive, negative, or neutral using an evaluative lexicon (a list of words labeled as positive or negative by specialists). The proportions of positive, negative, and neutral collocates are calculated as percentages of the list. This provides an initial indication of the kinds of contexts in which the node is typically used.

Second, 200 concordance lines containing the node are selected, and each full context is coded holistically as positive, negative, or neutral. In this stage, evaluation is based on the overall contextual meaning rather than on collocates alone. Because this coding is performed by human researchers and therefore may involve subjectivity, reliability control is integrated into the procedure.

To reduce subjectivity in contextual coding, the study uses Cohen's kappa coefficient as an inter-rater reliability measure [Cohen, 1960]. Two researchers independently code the same 200 concordance lines as positive, negative, or neutral. The number of cases in which both raters assign the same category and the number of mismatches across all category pairs are recorded, and kappa is computed using an online calculator capable of handling three-category coding. For this research, the GraphPad online tool is used as the calculator platform.

Interpretation of Cohen's kappa values:

- ≤ 0 → no agreement
- 0.01–0.20 → very low agreement
- 0.21–0.40 → low agreement
- 0.41–0.60 → moderate agreement
- 0.61–0.80 → satisfactory agreement
- 0.81–1.00 → high agreement

For research data to be treated as sufficiently reliable, a kappa value above 0.60 is required. If the value is lower, the coding procedure is repeated with a new sample of concordance lines.

Once reliability is established, the proportions of positive, negative, and neutral contexts in the 200-line sample are calculated. If one category reaches 60% or more, the prosody is treated as dominant, and the Stage 2 result is compared to Stage 1. If both stages converge (for example, both indicate negative contexts), the verb is interpreted as having a strong negative prosody. This information supports more precise boundaries for synonym usage and differentiation.

The reliability of corpus-based analysis depends not only on the analytic procedures but also on the properties of the corpus itself. Issues of validity and representativeness have been widely discussed in corpus methodology, particularly in work associated with Douglas Biber and large-scale corpus grammar projects [Biber et al., 1999]. Biber's experience in building and analyzing large corpora for descriptive grammar research emphasizes that corpus results can be misleading if corpora are not carefully documented, balanced, and interpreted according to their represented registers and distributions.

Following corpus validity principles, the study integrates the following safeguards:

1. The corpus compilers must explicitly document which layer of language the corpus represents; otherwise, results may appear general but in fact reflect only the most represented variety.

2. The corpus should avoid dominance by a single author, text, or genre, because quantitative results may otherwise reflect idiosyncratic style rather than general language patterns.

3. In interpreting findings, researchers must account for which registers/styles are represented; conclusions based on one style may not hold for another.

4. Frequency values should be normalized relative to corpus size to prevent misleading comparisons; the study reports frequencies as pmw (per million words).

5. Frequency should be interpreted alongside distribution across texts; otherwise, patterns may be driven by a small number of documents. Dispersion can be measured using indices such as Gries's DP [Gries, 2008].

To implement these safeguards, the study adopts the following procedure:

- All selected texts are documented with metadata including genre, author, year of creation, and size (tokens).

- Because the translation component is based on literary texts, translation-based conclusions are interpreted as specific to the domain of literary translation.
- All frequency measures are presented in normalized form (pmw).
- For each target synonym verb, dispersion across texts in the corpus is calculated using DP.
- In monolingual corpora used for intralingual analysis, key analyses are also carried out by genre/register whenever possible.

The application of corpus methods to translation studies is closely associated with Mona Baker, who argued for a shift in translation studies from prescriptive to descriptive approaches. Baker emphasizes that translation equivalence should be analyzed empirically in parallel corpora, because such analysis reveals tendencies in how particular linguistic units are translated and can uncover patterns of translator behavior [Baker, 1995; Baker, 1996]. The importance of this approach includes identifying translators' implicit norms, comparing semantic properties of source and target languages, and detecting lexical gaps.

Based on Baker's framework, the translation component of this study is implemented through the following steps:

1. Build a working Uzbek–English parallel corpus in Sketch Engine from original texts and their translations.
2. Align sentences containing the selected Uzbek and English action-verb synonyms.
3. For each synonym, calculate the frequency of its translation equivalents.
4. For comparability, compute each equivalent's share as a percentage of all occurrences of the source synonym in the aligned dataset. This distribution is expected to reveal why translators choose particular equivalents and may also expose semantic properties that emerge specifically in translation contexts.
5. Based on an initial sample, identify which translation strategies were used in translating sentences containing each synonym. To reduce subjectivity, two researchers independently code translation strategy categories, and reliability is evaluated using Cohen's kappa.
6. Use the resulting patterns to identify translation universals or systematic tendencies relevant to synonym behavior in translation.

RESULTS

Because the provided text primarily specifies a methodological framework, the results section is presented as the concrete analytic outputs that the proposed pipeline yields. The integrated monolingual–parallel design produces several reportable outcomes:

1. Intralingual collocation profiles for each synonym node in Uzbek and English, including ranked collocates by MI and by t-score. These profiles make it possible to compare synonym members within a set and identify systematic distributional differences.
2. Interpretive semantic features inferred from collocational evidence, such as exertion, intensity, manner, speed, or typical contextual constraints, supported by repeated corpus evidence.
3. Multiword pattern inventories, including frequent n-grams and construction-based patterns (e.g., complement + verb, subject + verb, adverbial modifier + verb), which reveal phraseological frames associated with each synonym.
4. Semantic prosody profiles for each verb node, derived from Stage 1 collocate polarity distributions and validated through Stage 2 concordance-based contextual coding.
5. Reliability documentation through Cohen's kappa values for contextual polarity coding and (where applied) translation strategy coding.
6. Validity statistics, including normalized frequency values (pmw) and dispersion (DP) measures to ensure that conclusions are not driven by isolated texts or corpus imbalance.
7. Translation equivalent distributions for each synonym in each translation direction, including raw frequencies and proportional shares of equivalents, enabling identification of asymmetries and stable translator preferences.

DISCUSSION

The methodology presented in this study treats synonymy as a usage-based lexical–semantic phenomenon whose distinctions emerge most clearly through systematic patterns in authentic contexts. Collocational analysis provides distributional evidence for meaning components by demonstrating that verbs which appear similar in dictionary glosses may differ substantially in their typical co-occurrence behavior, which in turn reflects selectional tendencies and semantic specialization [Sinclair, 1991; Cruse, 1986]. Using MI and t-score together enables a more nuanced interpretation: MI highlights strong associations that may be rare but semantically diagnostic, while t-score highlights frequent, stable combinations that characterize conventional usage. The combined use of these measures supports an empirically grounded differentiation among action-verb synonyms.

Multiword pattern analysis extends this evidence to phraseology and constructional behavior. Under the idiom principle, the repeated use of semi-preconstructed patterns is not peripheral but central to meaning construction. Therefore, identifying typical frames and recurrent multiword patterns provides additional evidence for semantic boundaries and pragmatic constraints [Sinclair, 1991]. The methodological adaptation for Uzbek is also critical: because Uzbek is agglutinative and allows greater word order flexibility, contiguous n-grams alone may not capture key verb-centered constructions. Construction-based queries therefore function as an essential complement, improving the descriptive adequacy of the Uzbek analysis.

Semantic prosody adds an evaluative dimension that is often decisive for lexical choice among near-synonyms. The two-stage approach reduces methodological risk: Stage 1 provides a distributional estimate, while Stage 2 anchors claims in direct contextual reading. Because annotation can be subjective, reliability control through Cohen's kappa makes subjectivity measurable and manageable and strengthens the credibility of interpretive claims [Cohen, 1960].

Validity safeguards ensure that findings remain interpretable and replicable. Corpus results must be understood relative to corpus composition, register representation, and distribution across texts. Normalization (pmw) prevents misleading frequency comparisons across corpora of different sizes, and dispersion measures reduce the risk that apparent frequency effects are produced by a small number of texts. These controls are essential when comparing a large English corpus (COCA) with smaller Uzbek resources.

Finally, the parallel corpus component enables the study to extend synonym analysis into translation behavior. Corpus-based translation studies emphasizes describing translator behavior empirically rather than prescribing equivalence. By quantifying translation equivalents and interpreting asymmetries, the methodology can reveal where translation neutralizes distinctions present in the source language or introduces distinctions not salient in the original. At the same time, such findings must be interpreted as outcomes of the intersection between language systems and translation norms, rather than as pure reflections of either one alone [Baker, 1995; Toury, 1995].

CONCLUSION

This paper presents a comprehensive corpus-based methodology for investigating Uzbek and English action-verb synonymy through a combined monolingual–parallel design. The approach is grounded in the principle that linguistic claims should be supported by large-scale empirical evidence rather than subjective intuition. Monolingual corpora provide the basis for intralingual semantic profiling, while a custom Sketch Engine parallel corpus supports translation-equivalent analysis. Lexical–semantic differences among synonyms are operationalized through MI and t-score collocation profiling, supported by multiword pattern extraction aligned with the idiom principle and extended through construction-based querying appropriate for Uzbek morphology and syntax. Semantic prosody is identified through a two-stage procedure and validated through inter-rater reliability using Cohen's kappa. The study integrates corpus validity safeguards, including transparent metadata documentation, normalization (pmw), and dispersion measurement (DP). Finally, the translation component follows a corpus-based descriptive framework to quantify equivalent distributions, identify asymmetries, and model translation strategy tendencies with reliability control. Overall, the proposed pipeline offers a replicable and empirically grounded framework for explaining why near-synonyms are not fully interchangeable and how their semantic and pragmatic profiles are reflected across languages and in translation.

REFERENCES

1. Apresyan, Y. D. (1995). Selected works. Volume I: Lexical semantics. Moscow: Shkola "Yazyki russkoy kultury".
2. Baker, M. (1995). Corpora in translation studies: An overview and some suggestions for future research. Target, 7(2), 223–243.
3. Baker, M. (1996). Corpus-based translation studies: The challenges that lie ahead. Amsterdam: John Benjamins.
4. Biber, D., Conrad, S., & Reppen, R. (1998). Corpus linguistics: Investigating language structure and use. Cambridge: Cambridge University Press.
5. Biber, D., Johansson, S., Leech, G., Conrad, S., & Finegan, E. (1999). Longman grammar of spoken and written English. London: Longman.
6. Cohen, J. (1960). A coefficient of agreement for nominal scales. Educational and Psychological Measurement, 20(1), 37–46.
7. Cruse, D. A. (1986). Lexical semantics. Cambridge: Cambridge University Press.
8. Davies, M. (2008). The Corpus of Contemporary American English (COCA): 560 million words, 1990–present. Available online.
9. Evert, S. (2005). The statistics of word

cooccurrences: Word pairs and collocations (Doctoral dissertation). University of Stuttgart.

10. Gries, S. T. (2008). Dispersions and adjusted frequencies in corpora. International Journal of Corpus Linguistics, 13(4), 403–437.

11. Hojiev, A. (1985). O‘zbek tilida sinonimlar. Toshkent: Fan.

12. Johansson, S. (2007). Seeing through multilingual corpora: On the use of corpora in contrastive studies. Amsterdam: John Benjamins.

13. Kennedy, G. (1998). An introduction to corpus linguistics. London: Longman.

14. Kilgarriff, A., Baisa, V., Bušta, J., Jakubíček, M., Kovář, V., Michelfeit, J., Rychlý, P., & Suchomel, V. (2014). The Sketch Engine: Ten years on. Lexicography, 1(1), 7–36.

15. Laviosa, S. (2002). Corpus-based translation studies: Theory, findings, applications. Amsterdam: Rodopi.

16. Louw, B. (1993). Irony in the text or insincerity in the writer? The diagnostic potential of semantic prosodies. In M. Baker, G. Francis, & E. Tognini-Bonelli (Eds.), Text and technology: In honour of John Sinclair (pp. 157–176). Amsterdam: John Benjamins.

17. McEnery, T., & Hardie, A. (2012). Corpus linguistics: Method, theory and practice. Cambridge: Cambridge University Press.

18. Shoabdurahmonov, Sh. (1999). O‘zbek leksikologiyasi. Toshkent: O‘qituvchi.

19. Sinclair, J. (1991). Corpus, concordance, collocation. Oxford: Oxford University Press.

20. Stubbs, M. (2001). Words and phrases: Corpus studies of lexical semantics. Oxford: Blackwell.

21. Toury, G. (1995). Descriptive translation studies and beyond. Amsterdam: John Benjamins.

22. To‘ychiyeva, D. (2020). Korpus lingvistikasi asoslari. Toshkent.

23. Church, K. W., & Hanks, P. (1990). Word association norms, mutual information, and lexicography. Computational Linguistics, 16(1), 22–29.

24. Tognini-Bonelli, E. (2001). Corpus linguistics at work. Amsterdam: John Benjamins.

25. Hunston, S. (2002). Corpora in applied linguistics. Cambridge: Cambridge University Press.