METHODS OF USING CORPUS LINGUISTICS IN THE ANALYSIS OF TERMINOLOGY IN ENGLISH AND UZBEK

Nigora Rahmatillayevna Hayitova

Lecturer at Renaissance Education University

ABSTRACT: This article examines methodological approaches to using corpus linguistics in the analysis of terminology in English and Uzbek. The study identifies the role of corpora in collecting, processing, and interpreting terminological units within both languages, with particular attention to domain-specific vocabularies such as those in science, technology, and artificial intelligence. Comparative analysis shows that corpus-based methods provide objective data on term frequency, collocations, semantic shifts, and usage contexts. The article outlines monolingual and parallel corpus applications, corpus annotation and tagging techniques, as well as tools for quantitative and qualitative analysis. The results emphasize that integrating corpus linguistics into terminological studies not only increases accuracy in identifying and defining terms but also supports the standardization and modernization of Uzbek technical lexicon in alignment with international practices.

Keywords: corpus linguistics, terminology analysis, English language, Uzbek language, frequency analysis, concordance, term extraction, parallel corpora, collocations, standardization.

INTRODUCTION

In the modern era of globalization and digitalization, the demand for accurate, up-to-date, and standardized terminology in both English and Uzbek has grown significantly. Terminological precision plays a central role in scientific, technical, and academic communication. With the rapid development of specialized fields such as artificial intelligence, medicine, and engineering, the volume of new terms has increased substantially.

Traditional terminological research, based primarily on manual collection and analysis, has been gradually replaced or supplemented by corpus-based methods. **Corpus linguistics** offers systematic tools for gathering, storing, and analyzing authentic language data, enabling researchers to identify patterns in term usage, frequency trends, and cross-linguistic variations with greater accuracy.

The aim of this article is to present an overview of corpus-based approaches to terminological analysis, focusing on their application to English and Uzbek, and to discuss practical methods for their implementation in comparative linguistic studies. Monolingual corpora, such as the **British National Corpus (BNC)** for English or the **Uzbek National Corpus**, serve as primary sources for identifying the usage frequency and contextual variation of terms within a single language. They are particularly useful for:

Determining the most common forms and variants of a term.

Observing diachronic changes in terminology.

Identifying specialized domain registers.

Parallel corpora contain aligned texts in two or more languages. For English–Uzbek studies, such corpora enable direct comparison of term translations, shifts in meaning, and semantic equivalence. They are widely used in:

Translational equivalence analysis.

Identifying calques, borrowings, and adapted terms.

Quality assessment of translation in technical documentation.

These corpora focus on a specific field, for example, **artificial intelligence terminology**. Specialized corpora allow for in-depth frequency and collocation analysis of domain-specific terms, supporting the development of glossaries and specialized dictionaries. One of the most fundamental corpus methods is frequency counting. By ranking terms according to their occurrence in the corpus, researchers can distinguish between core vocabulary and peripheral terminology. Frequency lists also help in identifying emerging terms in fast-evolving domains. Concordance tools display instances of a term in its immediate context (Key Word in Context — KWIC). Collocation patterns reveal which words frequently occur together with a given term. In English, for example, *neural* frequently collocates with *network*, *processing*, and *computation*. In Uzbek, *sun'iy* most often collocates with *intellekt* or *idrok*.

Automatic term extraction (ATE) uses computational tools to identify potential terminology candidates from corpora. Criteria may include statistical measures (TF-IDF, Mutual Information) and linguistic filters (POS tagging, noun phrase detection).

Annotating corpora with metadata such as part-of-speech tags, semantic domains, and translation equivalents enhances the depth of terminological research. For bilingual English–Uzbek corpora, alignment tagging helps track translation consistency. In English, corpus-based terminological research benefits from large, well-maintained, and open-access corpora (e.g., COCA, BNC, NOW Corpus). This allows for high-precision frequency analysis, semantic mapping, and domain-specific studies.

In contrast, Uzbek corpus resources are still in the development stage, with limited publicly available specialized corpora. However, recent efforts by national linguistic research centers and universities have introduced corpora for legal, medical, and educational terminology.

Challenges in Uzbek include:

Lack of large-scale specialized corpora.

Inconsistent term standardization.

Variability in orthography for borrowed terms.

Despite these challenges, combining English corpus resources with Uzbek-language datasets in parallel or comparable corpus projects offers a powerful methodology for cross-linguistic terminological research.

Several corpus analysis tools are applicable for English–Uzbek terminology research:

AntConc — concordance and collocation analysis.

Sketch Engine — multilingual corpus query and word sketches.

ParaConc — parallel corpus alignment and analysis.

TermoStat — automatic term extraction.

These tools support both qualitative analysis (semantic nuance, usage context) and quantitative analysis (frequency counts, co-occurrence measures).

CONCLUSION

Corpus-based methods have transformed the study of terminology by providing objective, data-driven insights into usage patterns, frequency, and contextual variation. For English, the availability of extensive, high-quality corpora facilitates advanced terminological analysis. For Uzbek, further corpus development and digitization efforts are essential for effective and standardized terminology work. The integration of corpus linguistics into English–Uzbek comparative studies ensures greater accuracy in identifying equivalent terms, supports translation quality, and contributes to the modernization of Uzbek technical lexicon in line with global standards.

REFERENCES

1. Biber, D., Conrad, S., & Reppen, R. (2021). Corpus Linguistics: Investigating Language Structure and Use. Cambridge University Press.
2. Bowker, L., & Pearson, J. (2002). Working with Specialized Language: A Practical Guide to Using Corpora. Routledge.
3. Kilgarriff, A., et al. (2014). The Sketch Engine: Ten Years On. Lexicography, 1(1), 7–36.
4. McEnery, T., & Hardie, A. (2012). Corpus Linguistics: Method, Theory and Practice. Cambridge University Press.
5. Nazarov, A. (2023). Corpus-based approaches to Uzbek technical terminology. Philology Issues, 4(2), 55–68.