Translation Units as A Basis for Constructing Bilingual Lexicons on The Paratranslator.UZ Platform

Vasliddinova Kamola Qodirjon qizi

PhD student, Uzbekistan State World Language University, Tashkent, Uzbekistan

ABSTRACT

This article analyzes translation units as the fundamental elements in building effective bilingual lexicons using the Paratranslator.uz platform. In this study traditional word-level framework approach is explored by addressing specific challenges in machine translation involving Uzbek and English languages. About 3000 translation samples in literary, official, scientific, spoken styles are analyzed by demonstrating translation units incorporating contextual, cultural and grammatical features. According to the findings, machine translation quality is enhanced significantly by revealing translation unit-based lexicons while providing a scalable foundation for expanding linguistic capabilities within the platform.

Keywords: Bilingual lexicon, computational linguistics, corpus linguistics, translation units, framework, lexical unit, syntactical unit, discursive unit, Paratranslator.uz.

INTRODUCTION

The development of machina translation encourages lexicographs to carry on researches the new approaches in building bilingual lexicons. Therefore, analyzing translation units in contexts is being systematically crucial as ever particularly for languages with limited digital resources like Uzbek. Translation units can generally be grouped into three main categories depending on their contextual characteristics: lexical, syntactic, and discursive.

Lexical units consist of individual words and fixed expressions, such as idioms and collocations, which carry distinct meanings within a text.

Syntactic units include complete sentences as well as proverbs or traditional sayings, which reflect established grammatical structures and linguistic patterns.

Discursive units are larger text segments, like paragraphs and dialogues, that share thematic coherence and a unified meaning, effectively presenting a specific idea or concept. Such a classification provides a clearer framework for analyzing translation units, enhancing our understanding of how meaning is expressed and preserved across various linguistic levels. Conventional machine translation approaches have traditionally relied on direct word-for-word correspondences, which frequently fail to capture language complexities including idiomatic expressions, cultural references, and fundamental structural differences [1]. According to Jean Paul Vinay and Jean Darbelnets's theory translation unit is the smallest segment of the utterance which couldn not be translated individually [2]. In the field of lexicography, the development of bilingual lexicons is a critical task that supports multilingual communication and translation. The emergence of digital platforms, such as Paratranslator.uz, has facilitated the creation and analysis of bilingual corpora, providing new opportunities for the systematic study of translation units. This study explores the concept of translation units as the foundational basis for constructing bilingual lexicons on the Paratranslator.uz platform.

The main objectives of this study are:

1. To identify and define translation units within the Uzbek-English parallel corpus on Paratranslator.uz.

2. To analyze the contextual consistency of these units across various text types.

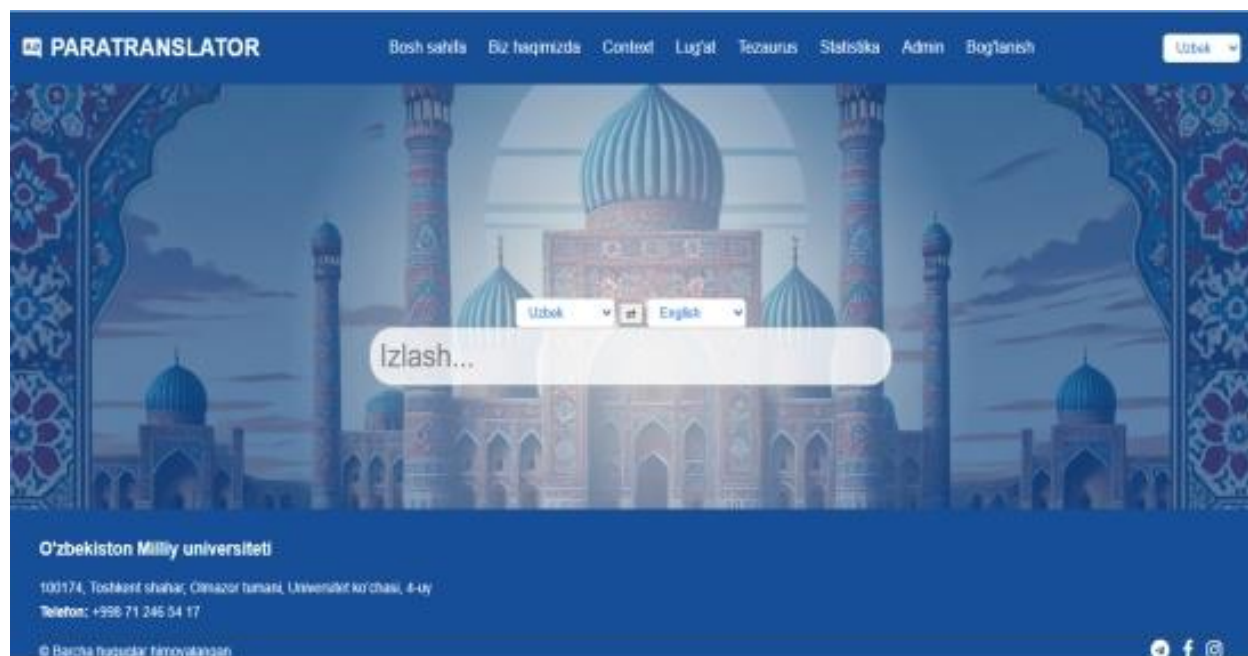3. To construct a bilingual lexicon based on the identified translation units.

METHODS

This study employs comparative analytical methods and data collection and analysis approach, systematically scholars' theory and contribution to build bilingual lexicons and translation units were discussed. We take journal articles, monographs and theoretical books as the primary sources to carry on research. In this research, we aimed to investigate translation units for the Uzbek-English bilingual lexicon by employing stylistic criteria as lexicographic sources. To achieve this, we utilized a web crawler to download texts in both Uzbek and English that possess exact translation equivalents, representing four distinct styles: literary, scientific (via Ziyonet) [6], formal (via Lex.uz) [3], and popular (via President.uz) [5]. These texts were then integrated into the "PARATRANSLATOR: a context-based electronic translation dictionary platform based on a parallel corpus." During the selection process, we adhered to corpus criteria rooted in the observational method of empirical knowledge acquisition, ensuring that the texts within each style conformed to standards of reliability, legality, and alignment with national cultural and ethical norms.

RESULTS

Over 10,000 legal-normative texts in Uzbek were sourced from Lex.uz [3], while more than 100,000 popular-style texts were obtained from Kun.uz [4]. Additionally, literary texts from various works and scientific texts, including abstracts and article annotations, were paired with their English counterparts. These aligned bilingual texts were systematically uploaded to the database of the "PARATRANSLATOR: a context-based electronic translation dictionary platform based on a parallel corpus". We developed the theoretical and methodological framework for constructing a bilingual lexicon for the Uzbek language.

The interface of "PARATRANSLATOR: a context-based electronic translation dictionary platform based on a parallel corpus".

DISCUSSION

The results demonstrate that translation units offer a robust basis for bilingual lexicon construction, particularly in contexts where traditional dictionaries may be insufficient.

The use of the Paratranslator.uz platform proved effective for data collection and analysis, highlighting the potential of digital platforms for lexicographic research.

The process of analyzing translation units on the "PARATRANSLATOR: a context-based electronic translation dictionary platform".

The identification of translation units combined automated extraction techniques with manual annotation processes. A team consisting of two computational linguists with expertise in Uzbek linguistics collaborated with three professional translators who possessed extensive experience in Uzbek-English translation. Together, they annotated 2,500 sentence pairs, identifying translation units according to established criteria:

1. Semantic unity (representing a complete conceptual meaning)

2. Structural cohesion (functioning as an integrated syntactic unit)

3. Translation integrity (requiring holistic rather than component-level translation)

CONCLUSION

This study has demonstrated significant advantages of using translation units as foundational elements for constructing bilingual lexicons on the Paratranslator.uz platform. The translation unit-based approach yielded substantial improvements in translation quality across multiple metrics, with particular benefits for handling idiomatic expressions, specialised terminology, and complex grammatical structures.

The findings carry both theoretical and practical implications. From a theoretical perspective, they support the understanding that translation operates above the word level, involving complex units of meaning that incorporate contextual, cultural, and grammatical dimensions. From a practical standpoint, the implementation of translation units in the Paratranslator.uz platform provides a model for enhancing machine translation quality, particularly for languages with limited digital resources.

Future development of the Paratranslator.uz platform will focus on expanding translation unit coverage, improving automatic extraction methods, and integrating translation unit-based lexicons with neural machine translation architectures. This research represents an important

advancement toward more nuanced and accurate machine translation systems capable of better capturing human language complexities.

REFERENCES

Baker, M., & Saldanha, G. (Eds.). (2020). Routledge encyclopedia of translation studies (3rd ed.). Routledge.

Vinay, J.-P., & Darbelnet, J. (1995). Comparative stylistics of French and English: A methodology for translation (J. C. Sager & M.-J. Hamel, Trans.

Lex.uz

Kun.uz

President.uz

Ziyonet