IMPROVING UZBEK MACHINE TRANSLATION THROUGH PARALLEL CORPORA: CHALLENGES AND SOLUTIONS

Botirova Nilufar Salimjon kizi
PhD student
Uzbekistan State World Languages University

***Annotation**: The thesis explores the significance of parallel corpora in modern translation studies, focusing on their crucial role in improving machine translation systems, specifically in the context of the Uzbek language. Parallel corpora, which consist of texts in multiple languages aligned at the sentence or paragraph level, are essential for training neural network-based translation systems. The paper outlines the main challenges in creating high-quality parallel corpora, particularly for underrepresented languages like Uzbek. These challenges include limited available resources, contextual mismatching, errors in segmentation and alignment, and copyright issues. The thesis discusses several solutions to these problems, such as building open-access databases, leveraging machine translation systems, using modern alignment tools, and engaging in crowdsourcing efforts. Additionally, it emphasizes the future potential of parallel corpora in advancing translation quality, supporting linguistic research, and promoting the global recognition of the Uzbek language. Ultimately, the paper argues that parallel corpora are not just a scientific resource but a technological tool, bridging the gap between human translators and machine translation systems.*

***Key Words**: Corpus, corpus linguistics, parallel corpus, translation corpus, comparable corpus, segmentation, machine translation*

In modern translation studies, parallel corpora are recognized as one of the key methodological foundations. A parallel corpus is a collection of texts written in two or more languages, with corresponding segments (sentences or paragraphs) aligned to each other. These corpora not only enhance the quality and accuracy of translations but also serve as a crucial tool in automating the translation process. Specifically, they form the core dataset for training neural network-based machine translation systems. For instance, international parallel corpora such as Europarl, the UN Corpus, or Tatoeba now contain millions of segments across hundreds of language pairs and contribute significantly to improving the quality of machine translation algorithms.[1] Using parallel corpora, linguistic features, translation equivalents, cultural differences, and grammatical structure variations are analyzed in depth. These corpora are invaluable for identifying subtle aspects like translation style, synonymous variants, and contextual appropriateness. Moreover, they provide a scientific basis

[1] Koehn, P. *Europarl: A Parallel Corpus for Statistical Machine Translation*. MT Summit X.2005

for demonstrating the variability of translation styles. However, the process of creating such corpora comes with several challenges. First and foremost, gathering high-quality translation resources—especially for underrepresented languages—remains a critical issue. Dividing translation texts into appropriate segments, aligning them accurately into pairs, and performing automatic alignment can often lead to errors. To address these challenges, statistical and neural aligners, the use of metadata, and intellectual filters that take into account both the translators and the context of the text are recommended solutions. For this reason, parallel corpora serve not just as an aid, but also as a mediator in the translation field—acting as a bridge between translators and translation tools (such as machine translation and CAT tools). This role makes them a central element in modern translation technologies.[2]

The initial stage of working with parallel corpora is gathering appropriate translation sources and organizing them in a clear structure. This process is a key factor that directly influences the quality of the translation. One of the major challenges, especially for Uzbek, is the limited availability or complete lack of high-quality parallel texts. There are very few parallel corpora for language pairs involving Uzbek—particularly for English-Uzbek, Russian-Uzbek, or Tajik-Uzbek—making them insufficient for scientific and practical research purposes.

Another issue encountered when collecting translation texts is contextual mismatching, where a text in one language does not fully align with the text in another language. In such cases, automatic segmentation and alignment tools (such as hunAlign, Bleualign, or more recent GPT-based matching models) may sometimes generate incorrect segments. Additionally, copyright concerns, text accessibility, and usage restrictions also pose barriers when gathering translation sources.[3]

Several approaches exist to address these challenges. One common solution is creating parallel corpora based on open-source materials such as government documents, UN texts, European Union reports, or bilingual archives from media outlets. Another method gaining traction is expanding parallel text databases through crowdsourcing, with voluntary translators contributing. In some cases, semi-automatic parallel corpora are created by comparing the initial outputs of machine translation tools with human translations.[4]

Parallel corpora offer unmatched potential as the foundation for machine translation. Models trained on these corpora can more accurately learn the grammatical, lexical, and stylistic features of a language, improving their

[2] Tiedemann, J. Parallel Data, Tools and Interfaces in OPUS. In LREC.2012.

[3] Resnik, P., & Smith, N. A. *The web as a parallel corpus*. Computational Linguistics.2003.
[4] Bojar, O., et al. *Findings of the 2014 Workshop on Statistical Machine Translation*. ACL.2014.

understanding of context in translation and reducing the issue of "ambiguous translations." In particular, neural machine translation systems (like Google Translate or DeepL) achieve high accuracy by learning from large volumes of parallel texts.[5] Thus, parallel corpora are not just a scientific resource for improving translation quality, but also a technological tool for automating translation. They serve as a bridge in the translation world, advancing collaboration between humans and machines to a new level.

Solutions to the Challenges

Several solutions are proposed to address the main challenges in creating parallel corpora. First, it's essential to build an open-access and high-quality text database for the Uzbek language. This could involve collaborating with international organizations to use official documents, scholarly articles, and other linguistic resources. For example, popular parallel corpora like Europarl or the UN Corpus should also be adapted to include Uzbek.

A second solution is to use machine translation tools to automatically segment and align parallel texts. Today, machine translation systems—especially those based on neural networks—are showing excellent results. These systems can automatically analyze Uzbek texts and help produce high-quality translations.

Additionally, using modern alignment tools (such as hunAlign) to correctly organize and align the texts is highly effective. These tools can help ensure the texts are properly segmented and aligned. Furthermore, engaging more Uzbek translators and users through a crowdsourcing approach will help expand the parallel text database. This collaborative effort can significantly contribute to creating a more robust and diverse parallel corpus for Uzbek.

In conclusion, parallel corpora play a crucial role in the development of translation systems, scientific research, and machine translation systems in Uzbek linguistics. The main challenges in creating parallel corpora for the Uzbek language include the lack of high-quality texts, the failure to account for contextual and cultural differences, and the shortage of metadata. However, with the help of modern technologies and approaches, these issues can be addressed. The development of parallel corpora will enable greater global recognition of the Uzbek language and provide opportunities for its effective use in translation systems. The future of parallel corpora in Uzbek linguistics is very promising. They not only enhance the effectiveness of translation systems but are also crucial for studying the grammatical, semantic, and stylistic features of the language. With parallel corpora, new scientific

[5] Och, F. J., & Ney, H. *A systematic comparison of various statistical alignment models*. Computational Linguistics.2003.

research, lexicographical works, and approaches to language preservation can be developed for Uzbek. Moreover, to further strengthen the role of parallel corpora in machine translation systems, it is essential to expand the existing database of Uzbek texts. This will not only improve the quality of translations but is also vital for increasing the global recognition of the Uzbek language.

References:

1. Koehn, P. *Europarl: A Parallel Corpus for Statistical Machine Translation*. MT Summit X.2005.

2. Tiedemann, J. *Parallel Data, Tools and Interfaces in OPUS*. In *LREC*.2012.
3. Bojar, O., et al. *Findings of the 2014 Workshop on Statistical Machine Translation*. ACL.2014.
4. Och, F. J., & Ney, H. *A systematic comparison of various statistical alignment models*. Computational Linguistics.2004.
5. Resnik, P., & Smith, N. A. *The web as a parallel corpus*. Computational Linguistics.2003.
6. Artetxe, M., & Schwenk, H. *Massively Multilingual Sentence Embeddings for Zero-Shot Cross-Lingual Transfer and Beyond*. Transactions of the ACL.2019.
7. Sharoff, S. *Constructing Comparable Corpora for Low-Resource Languages*. Language Resources and Evaluation.2020.
8. Translators Without Borders – https://translatorswithoutborders.org
9. OPUS corpus – http://opus.nlpl.eu
10. LaBSE (Google Research) – https://github.com/google-research/bert