

Djumaniyazova Intizor Davlatnazarovna,
teacher of the Department of Language
and Literature, Mamun University

PARALLEL CORPUS AS AN INTEGRAL PART OF THE LANGUAGE CORPUS

INTRODUCTION

The language corpus is a large electronic collection of texts used to study various aspects of language, including syntax, vocabulary, stylistics, and semantics. One of the key types of language corpus is parallel corpus. It consists of texts translated from one language to another, which allows for comparative linguistic analysis and the use of this data for machine translation tasks and the development of natural language processing (NLP) technologies. Parallel corpora occupy an important place in modern linguistics and are an integral part of the language corpus. The purpose of this work is to study the parallel corpus as an important component of the language corpus, as well as its role in various fields of linguistics and technology.

MATERIALS AND METHODS.

The research uses data from scientific publications and examples of existing parallel buildings, such as Europarl, OPUS, TED Talks. The methods of comparative analysis and systematization of information are used as a methodological approach. The main focus is on the structure of the parallel corpus, its functions in linguistic research and machine translation, as well as existing problems and challenges in the creation and application of such corpora.

RESULTS.

1. What is a parallel housing?

A parallel corpus is a type of language corpus containing texts in two or more languages that are translations of the same source text. Each passage in one language corresponds to its exact translation in another language. Parallel corpora are necessary for studying language equivalents, as well as for analyzing various aspects of translation, such as the lexical and grammatical means used to convey meanings.

2. The structure of the parallel housing The parallel housing consists of several main components:

- 1) Texts and translations: The parallel corpus includes text pairs in different languages, synchronized in fragments (for example, sentences or paragraphs), which allows you to accurately correlate parts of the text in one language with their counterparts in another language.
- 2) Metadata: Each text in a parallel corpus is accompanied by metadata such as information about genre, author, year of publication, and subject area. This allows researchers to take into account the context of the translation when analyzing the data.
- 3) Annotated data: Some parallel corpora contain syntactic and semantic annotations that can be useful for in-depth linguistic analysis, for example, when teaching machine translation systems.

3. The role of parallel corpora in linguistics and machine translation

Parallel corpora play an important role in various fields of linguistics and technology, such as machine translation and natural language processing.:

- 1) Translation Analysis: Parallel corpora are used to study how words and phrases are translated between languages. This allows researchers to study the equivalence between languages, as well as identify translation strategies and possible errors.
- 2) Machine Translation: Parallel corpora are the basis for training machine translation systems. Based on these corpora, algorithms are trained that allow you to translate text from one language to another, taking into account the grammar and stylistics of both languages.
- 3) Lexicography and Semantics: Parallel corpora serve as a source for creating accurate dictionaries and researching the lexical meaning of words in context. It also helps lexicographers to find the most appropriate translations for different meanings of words depending on their context.
- 4) Natural Language Processing (NLP): Parallel corpora are used in text classification, sentiment analysis, information retrieval, and other NLP tasks where it is important to consider multiple languages simultaneously.

4. Examples of parallel enclosures

- 1) Europarl: One of the most famous parallel corpus containing translations of the debates of the European Parliament into 21 languages. This corpus is actively used for teaching and testing machine translation systems.
- 2) OPUS: A large project containing translations of texts in various genres in several languages. OPUS is used for machine translation training and translation analysis in various fields such as engineering, science, and literature.

Discussion Parallel enclosures, despite their importance, have a number of problems associated with their creation and use.:

- 1) Translation quality: One of the main difficulties is to ensure high-quality translation. Automatic translations used in parallel corpora often contain errors, which can affect the accuracy of the study.
- 2) Limited volume: For some languages, parallel corpora may be limited in volume, which makes it difficult to train models for less common languages.
- 3) Identification of equivalence: Different languages may have unique grammatical and lexical structures, which makes it difficult to find exact equivalents when forming parallel corpora.

DISCUSSION

Parallel enclosures, despite their importance, have a number of problems associated with their creation and use.:

- 1) Translation quality: One of the main difficulties is to ensure high-quality translation. Automatic translations used in parallel corpora often contain errors, which can affect the accuracy of the study.
- 2) Limited volume: For some languages, parallel corpora may be limited in volume, which makes it difficult to train models for less common languages.

3) Identification of equivalence: Different languages may have unique grammatical and lexical structures, which makes it difficult to find exact equivalents when forming parallel corpora.

CONCLUSION

The parallel corpus is an important component of the language corpus, providing valuable data for linguistic analysis and solving problems in the field of machine translation and natural language processing. Despite existing problems such as translation quality and limited data for some languages, parallel corpora continue to play a key role in modern research and technology. The development of parallel corpora will open up new opportunities for creating more accurate and efficient machine translation systems, as well as improving text processing technologies in different languages.

LIST OF USED LITERATURE:

1. Barsova, I. V. (2014). The language corpus as a tool for studying translation: problems and prospects. Moscow: Nauka Publ.
2. Kunts, V. (2017). Parallel corpora in linguistics and machine translation. St. Petersburg: St. Petersburg State University Publishing House.
3. Rakhmatullina, A. R. (2018). Using parallel corpora for translation analysis in the context of machine learning. Kazan: Kazan University.
4. Tiedemann, J. (2012). OPUS – A Collection of Multilingual Parallel Corpora for Computational Linguistics. Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC 2012), 2371–2376.
5. Koehn, P. (2005). Europarl: A Parallel Corpus for Statistical Machine Translation. Machine Translation, 20(2), 1–20.
6. Lavergne, T., Marcu, D., & Koehn, P. (2010). Predictive Models for Statistical Machine Translation. Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (ACL 2010), 317–326.
7. Vasilescu, L., & Barbu, I. (2020). Parallel Corpora for Language Technologies and Computational Linguistics. Journal of Linguistic Research, 27(4), 501–513.
8. Zhang, M., & Zhao, W. (2016). The Role of Parallel Corpora in Machine Translation Systems. Computational Linguistics Journal, 42(3), 75–89.
9. Liu, Q., & Xu, W. (2019). Applications of Parallel Corpora in Cross-Linguistic Semantics and Lexicography. Lexicographic Studies, 35, 145–159.
10. TED Conferences. (2020). TED Talks: A Multilingual Parallel Corpus. [online] Available at: <https://www.ted.com/translate>.