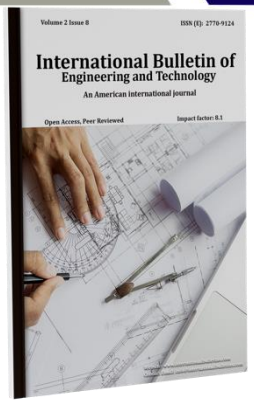

THE DEVELOPMENT OF CORPUS LINGUISTICS AS AN INDEPENDENT FIELD IN UZBEK LINGUISTICS

Umirova S.M.
Associate Professor of Samarkand State University
https://doi.org/10.5281/zenodo.17096641

Abstract. After computational linguistics was established as an independent discipline, its internal branches emerged. One of these is corpus linguistics, whose main task is to collect data in a specific language in one place, process it, and transmit it to future generations while preserving the purity of the language. This article provides information on how corpus linguistics has developed as an independent field in Uzbek linguistics.

Keywords: language, natural language, natural language processing, computational linguistics, corpus linguistics, database, text.

Corpus linguistics is a collection of written data, coordinated within one language or across multiple languages. A corpus includes a compilation of texts written in various languages, translated texts, written and spoken language data, articles, books, and other textual materials. It is used for organizing, distributing, and analyzing data. This information is valuable for research on language grammar, lexicon, phonetics, pragmatics, sociolinguistics, and other linguistic concepts. The corpus can also encompass various languages, their dialects, and unique features, which creates numerous opportunities for linguistic research. This data is used for linguistic studies, facilitating language learning, acquisition, and correction, studying the original language, and compiling language-related information.

The first information about corpus linguistics in the world appeared in the 1940s [8]. The practical implementation of ideas for automatic text translation, along with the emergence of computational linguistics and mathematical linguistics, created the need to develop databases. For automatic translation to be realized, databases in the languages intended for translation were first necessary. Issues of corpus linguistics were initially proposed in accordance with the idea of implementing automatic translation [5].

Despite the achievements in Uzbek corpus linguistics, numerous challenges still exist in this field. One significant problem is the lack of a corpus or collection of text data specifically designed for linguistic research. This scarcity of resources makes it difficult for researchers to create accurate models, as these require large volumes of data. Another crucial issue is the lack of standardization in the Uzbek language and the presence of various dialects. This complicates the development of a universal algorithm capable of correctly analyzing the language. Developing an Uzbek language corpus is essential both for research purposes and for machine learning, natural language processing, and machine translation. An additional noteworthy challenge is the need to develop a standardized orthography and writing system for the Uzbek language, which would enable the standardized processing of data input into machine learning systems.

Despite the challenges, the future of Uzbek computational linguistics appears promising. With the development of machine learning algorithms, researchers can now harness the

power of big data to create more accurate models. Researchers may also focus on developing algorithms capable of learning from unlabeled data, which reduces the need for labeled datasets. Furthermore, the advancement of deep learning algorithms now enables the creation of models that can analyze word contexts, facilitating the development of more precise models. Interesting research is being conducted in areas such as automatic speech recognition, named entity recognition, and machine translation. Additionally, the growth of the internet and social networks has led to a surge in online communication in the Uzbek language, opening up new opportunities for research and applications. Key areas requiring further research attention include the development of more comprehensive corpora, sentiment analysis tools, and natural language generation. In Uzbek corpus linguistics, there are several applications, such as language teaching, machine translation, and language technology development. For example, language teachers can use corpora to design textbooks and learning materials that reflect the authentic usage of the language. Machine translation developers can utilize corpora to improve the accuracy of translation systems. Moreover, corpora can serve as a source of linguistic data for natural language processing applications, such as speech recognition and text-to-speech synthesis.

Another study conducted in Uzbek computational linguistics pertains to the field of machine translation. In 2015, Akmaljon Ibragimov and his team developed a machine translation system for Uzbek-English called YMT (Yandex-Microsoft Transconnector). The system utilizes a statistical translation model trained on parallel Uzbek-English texts. The system was tested and evaluated using a bilingual assessment metric, with results showing that the system achieved an accuracy score of 29.63 points.

One of the most significant achievements in the field of Uzbek language corpus development was the creation of computational tools for the Uzbek language. These tools enabled the processing and analysis of large volumes of Uzbek texts, which proved crucial for language modeling, machine translation, and speech recognition. Furthermore, research was conducted in the field of lexicography, leading to the development of electronic dictionaries for the Uzbek language. These dictionaries include searchable word combinations with various features, inflection classes, and other forms of information essential for natural language processing.

In the early years of Uzbekistan's independence, the study of computational processing of Uzbek texts was limited. However, with the development of tools such as morphological analyzers, morphological generators, and part-of-speech taggers, this field began to expand. Among those who contributed to the development of Uzbek computational linguistics are scientists from the National University of Uzbekistan, Tashkent University of Information Technologies, and Tashkent State University of Uzbek Language and Literature. In Sh. Khamroeva's research, the goals and objectives of author corpora, their unique characteristics, structure, composition, and derivatives are highlighted, along with the similarities and differences between the author corpora of A.S. Pushkin, F.M. Dostoevsky, A.S. Griboyedov, and W. Shakespeare [6]. As a result of the research, the researcher created a dictionary of terms related to corpus linguistics [7].

In various studies, the corpus has been defined and described differently. V.V. Rikov examines the corpus as logical thought and logical analysis. E. Finegan states: "A corpus is a collection of texts, typically in a computer-readable format, that contains information about the context in which the text was produced, the information provider, the author, the

addressee, or the audience" [2]. T. McEnery and A. Wilson acknowledge that "A corpus is a linguistic domain that can be used as a language sample, selected according to specific language criteria" [4]. V.P. Zakharov defines the "linguistic corpus of texts as a large, electronically presented, unified, structured, tagged, philologically representative series of linguistic data intended for solving specific linguistic problems" [3.3.]. S. Adilova, summarizing the definitions of corpora, outlines the following main characteristics in the formation and use of a corpus:

- To create a corpus, it is necessary to provide large volumes of text (from the internet or stored on disk);
- The material must be accurate, not require editing, demonstrate the variability of the language, and include elements of natural speech;
- To conduct linguistic analysis, it is necessary to identify and mark language units;
- As a result of the analysis, it should be possible to distribute the language material in a well-founded manner [1. 390-394].

The corpus contains a wealth of linguistic data from various sources, enabling the study of different aspects of the language. The field of Uzbek computational linguistics has come a long way since its inception, achieving significant progress in creating computational tools for the Uzbek language. Uzbek computational linguistics is a developing field that holds great promise. Researchers have made notable advancements in developing language resources and creating models capable of language analysis. Researchers, language teachers, and language technology developers can utilize the insights gained from analyzing the Uzbek language corpus. However, despite the progress, the field still faces numerous challenges, including the lack of resources and the non-standardization of the language, as well as the need to develop a robust corpus and standardized orthography. Uzbek corpus linguistics serves as a valuable tool for language learning and the development of language technology applications. The future of the field looks promising with the development of machine learning algorithms and deep learning models that can learn from unlabeled data.

References:

1. Адилова C. Corpus linguistics and its role in teaching uzbek as a foreign language. O‘zbek tili taraqqiyoti va xalqaro hamkorlik masalalari. – Toshkent, 2019.
2. Finegan E. LANGUAGE: its structure and use. – N.Y.: Harcourt Brace College Publishers, 2004.
3. Захаров В.П. Корпусная лингвистика. Учебно-методическое пособие. – Санкт-Петербург, 2005.
4. McEnery T, Wilson A. Corpus Linguistics. Edinburgh: Edinburgh University Press, 2nd edition, 2001.
5. Rahmonova. A. O‘zbek tili milliy korpusini yaratishda kompyuter usullari. PhD dessertatsiyasi. –Toshkent, 2021.
6. Хамроева Ш. Ўзбек тили муаллифлик корпусини тузишнинг лингвистик асослари: Филол. ф. бўйича ф. д. (PhD) дис. автореф. – Қарши, 2018.
7. Хамроева Ш.М. Корпус лингвистикаси атамаларининг қисқача изоҳли луғати. – Тошкент: Камалак нашриёти, 2018.

8. Курс “Корпусная лингвистика” (А.Б.Кутузов) Лицензия Creative commons Attribution Share-Alike 3.0 Unported (Электрон ресурс) - //lab314.brsu.by/kmp-lite/kmp-video/CL/CorporeLingva.pdf.