SOME REFLECTIONS ON NATIONAL LANGUAGE CORPORA

Umirova S.M.
Associate Professor of Samarkand State University
https://doi.org/10.5281/zenodo.17096654

Abstract. This article discusses one of the achievements of modern linguistics - the emergence of language corpora, their capabilities, the demand and need for language corpora, challenges in this field, as well as the work carried out in Uzbek linguistics, its analysis, and the perspectives of scholars on this matter.

Keywords: language, national language, linguistics, corpus linguistics, language corpora, text, linguistic analysis, natural language, electronic data.

Although corpus linguistics is a branch of computational linguistics that has developed in recent years, considerable work has already been accomplished in this field within global linguistics. In the current era of globalization, languages worldwide have begun to compile comprehensive collections of their vocabulary, along with works from all styles and genres created in these languages, with the aim of enhancing their national status and preserving their pure form. Simultaneously, they have set out to demonstrate the internal capabilities of languages by performing various linguistic operations on the collected texts. These efforts encompass electronic text data and the operations performed on them in corpora created in national languages.

While national language corpora are considered electronic databases containing all written and oral information produced in a given language, they differ from electronic libraries in terms of the linguistic processing applied to the collected information.

In world linguistics, national language corpora are being effectively utilized for the integration of language with information technologies, its processing, digitization, electronic storage, and conducting various research. Indeed, language corpora are not only electronic collections, but also serve as guides and lexicographic sources for language learners and those conducting various linguistic studies. The earliest corpora appeared in the form of card files, later evolving into written dictionaries and electronic dictionaries. Today, true language corpora have been created, encompassing millions of words and word forms. National language corpora have now become an integral part of linguistics.

Language corpora can be created for any language and are considered collections of texts covering various materials - works of fiction, scientific and journalistic articles, as well as oral language - including samples of colloquial speech.

Language corpora can be utilized for various purposes by representatives of all fields. Corpora are particularly important tools for linguistic research, as they provide objective results and clear conclusions by summarizing various scientific hypotheses based on original texts. In this process, researchers perform various linguistic operations using large volumes of texts and study collected examples from authentic materials. With the help of corpora, it is possible to examine the lexical, morphological, and syntactic features of a particular language, analyze the usage, frequency, semantic development, and pragmatics of words and phrases in

speech, as well as compare similarities or differences between various dialects and languages, and identify changes in the language over time. Furthermore, corpora are valuable in the fields of natural language processing, creating various computational models of language, and machine translation.

Language corpora play a crucial role in fields such as computational linguistics, corpus linguistics, and applied linguistics. Specifically, they make a practical contribution to enhancing our understanding of language, its structure, and its usage in speech processes, as well as to developing new methods and approaches for linguistic research.

In today's era of globalization, the need for national language corpora is growing, and we can cite several reasons for this:

- Linguistic corpora, created in various languages and encompassing numerous genres, provide researchers with access to large volumes of authentic texts, enabling them to study different dialects and styles of the language, utilize the language more extensively and comprehensively, and analyze it based on original texts;

- Studying areas such as grammar, vocabulary, and stylistics of the language through corpora allows for the identification of changes across time and space, as well as the comparison of different languages;

- Since corpora clearly display statistics and frequency of language units, the scientific results and conclusions obtained are reliable;

- Corpora play a crucial role not only in linguistic analysis but also in improving natural language processing, machine translation, development of other language-related computer programs, and refining modeling systems.

In short, language corpora are an integral part of linguistic research, serving as a means to test various scientific assumptions and hypotheses, obtain accurate data, and study language through authentic texts.

However, creating a corpus has its own challenges and difficulties. The primary requirement for corpus creation is a large volume of textual data in the target language. The texts included in the corpus must be of high quality and fully meet research requirements; otherwise, the analysis results will be inaccurate. It is desirable for the texts to be diverse in style and genre. It is crucial to ensure that there are rights to use the selected texts for the corpus, as some texts may be protected by copyright and have restricted usage.

One of the main features of the corpus is that the entered texts are processed, meaning linguistic characteristics are assigned to the used words, word forms, phrases, and syntactic units. This process primarily requires manual labor, thus demanding significant time and effort.

The software development for the corpus is also a complex process, requiring equal participation from both programmers and linguists. In particular, the programmer bears more responsibility for language modeling, automating various linguistic operations to the extent possible, managing the program, updating it, adding data, and ensuring it functions dynamically.

Generally, it is necessary to carefully plan all these challenges, utilize appropriate tools and methods, work in teams as much as possible, and collaborate with field specialists.

Another crucial aspect of working with a corpus is that it should be accessible to researchers. While some corpora are available on the Internet, they may not be accessible; access to a corpus might be possible, but conducting research, using it, or downloading data

may not be allowed. Some corpora are free, while others require payment for use. If the use of a corpus is unrestricted, it facilitates collaboration among researchers.

The field of corpus linguistics has been garnering interest among Uzbek linguists in recent years, with numerous studies being conducted in this area. It is known that before corpus linguistics developed as a separate field, it was an integral part of computational linguistics. The achievements of linguostatistical methods or mathematical linguistics should be recognized as the pioneering works in computational linguistics.

In recent years, research work conducted under the leadership of M. Hakimov for the "Translator L-MX" system has provided a lexical-syntactic analysis of the Uzbek and Russian languages [13], while the studies of Sh. Khamroeva, N. Abdurakhmanova, and D. Urinboeva [14;2;10] have employed a corpus-based approach to text analysis. This aspect has highlighted the crucial importance of developing linguistic modules in the creation of linguistic software programs.

B. Mengliyev, S. Bobojonov, and Sh. Khamroyeva, in their initial published articles about the national corpus, view the corpus as a tool that operates in either an online or offline system and express the following opinion: "A corpus is a collection of texts subjected to a search program to determine the characteristics of language units, a compilation of written or oral texts stored in electronic form in natural language, functioning within a computerized search system" [9].

B. Mengliyev's article published in the "Ma'rifat" newspaper about the national corpus of the Uzbek language [9], Sh. Hamroyeva's work on the linguistic foundations of compiling an author's corpus [14], M. Abjalova's linguistic modules for a program editing and analyzing Uzbek texts, which represent the initial processing stage of corpus texts (for a program editing official and scientific style texts) [4], A. Eshmuminov's research on creating a synonym database for the national corpus of the Uzbek language [8], and G. Toirova's monographic study on the theoretical and practical issues of creating a national corpus of the Uzbek language [11] are significant for their methodological recommendations in developing the language corpus.

As a result of N. Abdurakhmanova's research [2], the conceptual framework for the theoretical and practical foundations of developing electronic corpora of the Uzbek language was studied. For the linguistic analysis stages of lemmatization and tokenization processes in the Uzbek language corpus, a morphological database of the Uzbek language and a system of morphotactic rules have been created. N. Abdurakhmanova also briefly discusses the use of various modern computer technologies in creating translation memory and segmenting the equivalence of translation units when developing parallel texts as one of the internal corpora (subcorpus) of the Uzbek language electronic corpus. Based on these theoretical conclusions, a concordance search system was developed for the Uzbek language corpus manager (search system), utilizing lemma, token, and n-gram model-based searches.

In her monographic study on the theoretical and practical aspects of creating a national corpus of the Uzbek language [11], G. Toirova theoretically substantiated the technology for developing the Uzbek language national corpus based on the grammatical description of word forms and lexical units in Uzbek, as well as the principles of presenting information in the national corpus. The scholar proposed a method for automatic analysis of templates related to plain text, headlines, and poetic excerpts, which regulates the modeling approaches of existing standards for the National Corpus markup based on SGML/XML language.

O. Khidirov, directly utilizing the achievements of Uzbek theoretical linguistics, developed syntactic tag categories and tag models for the Uzbek language. He elucidated the principles of syntactic tagging for simple concise and simple extended sentences in Uzbek based on linguistic-syntactic patterns. Furthermore, he substantiated the system of syntactic tags and search parameters for compound sentences in Uzbek, including coordinated, asyndetic, subordinate, and complex compound sentences, as well as compound sentence constructions with direct speech [15].

O. Abdullayeva's research on the theoretical and practical foundations of creating a corpus of Uzbek language internet information texts [1] examined the similarities and differences between the internet as a corpus and traditional corpora. The study substantiated the internet's potential as a corpus through observations and developed software and linguistic support for the Uzbek language internet information text corpus. It is worth noting that O. Abdullayeva's views on developing linguistic annotations for the Uzbek language internet information text corpus align with Sh. Hamroyeva's ideas on linguistic tagging. The linguist succeeded in implementing morphological and semantic annotation of speech units in the Uzbek language internet information text corpus and demonstrating the results of linguistic analysis within the corpus.

G. Begmatova's research on creating an idiom database for the Uzbek national corpus [7] systematizes the content and chronology of research on idiomatic combinations in world linguistics, Turkology, and Uzbek linguistics.

N. Ataboev, using the example of the functional features of the COCA English corpus, expressed his views on the criteria for creating a corpus, such as representativeness, purposefulness, and limited coverage. These views were based on the results of a statistical study of the history of corpus linguistics, its developmental stages, and classification problems.

Based on N. Atabaev's research, the possibility of comparing linguocultural units has been demonstrated through statistical analysis of inter-corpus metaphors, collocations (semantic combinations), and phraseological units. This comparison was made possible by examining the representative nature of corpora, specifically using the British National Corpus (BNC) and the Corpus of Contemporary American English (COCA) as examples. The scholar proved that the COCA corpus, which was first studied as a dictionary corpus, provided phonetic, orthographic, grammatical, and lexical-semantic features of lexical units in the field of lexicography. Additionally, he substantiated its importance in improving English dictionaries [6].

In A. Eshmuminov's research on the synonym database of the Uzbek language national corpus, the linguistic foundations for developing a synonym database were established. The study identified types of linguistic markup, determined methods and tools of semantic markup used in creating a synonym database, and revealed the characteristics of semantic markup [8]. Based on these theoretical generalizations, an electronic database of Uzbek language synonyms was created.

In recent years, efforts have been made to create a national corpus of the Uzbek language, resulting in the launch of several corpora. These include the "Educational Corpus of the Uzbek Language" - http://uzschoolcorpara.uz/, developed within the framework of a project implemented at the Alisher Navoi Tashkent State University of Uzbek Language and Literature, as well as the National Corpus of the Uzbek Language (based on materials from the

epic "Alpomish") - uzbekcorpora.uz, developed as part of a project carried out by the Samarkand branch of Tashkent University of Information Technologies and a team from Samarkand State University.

The national language corpus is also one of the primary means of achieving global recognition for our native language, as well as improving and elevating national spirituality. Enriching it and establishing its use are among the urgent tasks of today.

References:

1. Абдуллаева О. Ўзбек тилининг интернет ахборот матнлари корпусини шакллантиришнинг назарий ва амалий асослари. Филол. фан. бўйича фалсафа доктори (PhD)… дис. афтореф. – Тошкент, 2022. – 25 б.
2. Абдураҳмонова Н. Инглизча матнларни ўзбек тилига таржима қилиш дастурининг лингвистик таъминоти: Филол. фан. бўйича фалсафа доктори (PhD)… дис. афтореф – Тошкент, 2018. – 47 б.
3. Абдураҳмонова Н. Ўзбек тили электрон корпусининг компютер моделлари: Филол. фан. доктори (DSc) диссер. – Тошкент, 2022. – 168 б.
4. Абжалова М. Ўзбек тилидаги матнларни таҳрир ва таҳлил қилувчи дастурнинг лингвистик модуллари (Расмий ва илмий услубдаги матнлар таҳрири дастури учун). Филол. фан. бўйича фалсафа доктори (PhD)… дис. автореф. – Фарғона, 2019.
5. Арзикулов Х.А., Пиотровская К.Р. Информатика и переработка текста средствами вичислителной техники (учебное пособие). – Самарканд, 1986.
6. Атабоев Н. Инглиз тили корпусининг функсионалц хусусиятлари (COCA мисолида). Филол. фан. бўйича фалсафа доктори (PhD) дис. автореф. – Тошкент, 2020. – 58 б.
7. Бегматова Г. Ўзбек миллий корпусида идиомалар базасини яратиш. Филол. фан. бўйича фалсафа доктори (PhD)…дис. – Термиз, 2021. – 145 б.
8. Эшмўминов А. Ўзбек тили миллий корпусининг синоним сўзлар базаси: Филол. фан. бўйича фалсафа доктори (PhD) дис. афтореф. – Қарши, 2019.
9. Менглиев Б., Бобожонов С., Хамроева Ш. Ўзбек тили миллий корпуси. 2018-йил, 26-апрел, http://marifat.uz/marifat/ruknlar/fan/1241.htm
10. Ўринбоева Д. Халқ оғзаки ижоди: жанрий-лисоний ва лингвостатистик тадқиқ муаммолари: Филол. фан. бўйича докт. (DSc) дисс. автореф. – Самарқанд, 2019. – 74 б.
11. Тоирова Г. Ўзбек тили миллий корпусини яратишнинг назарий ва амалий масалалари. – Globeedit, 2020. – 168 б.
12. Тоирова Г. Ўзбек тили миллий корпусини яратишнинг назарий ва амалий масалалари: Филол. фан. доктори (DSc) диссер. – Бухоро, 2021. – 165 б.
13. Хакимов М.Х. Расширяемий входной язик математического моделирования эстественного язика для много-язичной ситуасии машинного перевода // ЎзМУ хабарлари, 2009. № 1. – С.75-80.
14. Хамроева Ш. Ўзбек тили муаллифлик корпусини тузишнинг лингвистик асослари: Филол. фан. бўйича фалсафа доктори (PhD) дис. афтореф. – Қарши, 2018. – 53 б.
15. Хидиров О. Миллий корпус учун парсинг дастури яратишнинг лингвистик асослари. Филол. фан. бўйича фалсафа доктори (PhD) дис. афтореф. – Жиззах, 2021. – 26 б.
16. http://uzschoolcorpara.uz/
17. uzbekcorpora.uz