O 'zbekiston Respublikasi
oliy ta'lim, fan va
innovatsiyalar vazirligi

Alisher Navoiy nomidagi
Toshkent davlat o'zbek tili
va adabiyoti universiteti

O'ZBEKISTON RESPUBLIKASI OLIY TA'LIM, FAN VA INNOVATSIYALAR VAZIRLIGI

ALISHER NAVOIY NOMIDAGI TOSHKENT DAVLAT O'ZBEK TILI VA ADABIYOTI UNIVERSITETI

"O'ZBEK TILI MILLIY VA TA'LIMIY KORPUSINING NAZARIY VA AMALIY MASALALARI"

mavzusidagi
Respublika ilmiy-amaliy konferensiyasi materiallari

(2023-yil 5-may)

Toshkent – 2023

O‘ZBEKISTON RESPUBLIKASI OLIY TA’LIM, FAN VA INNOVATSIYALAR VAZIRLIGI

ALISHER NAVOIY NOMIDAGI TOSHKENT DAVLAT O‘ZBEK TILI VA ADABIYOTI UNIVERSITETI

“O‘ZBEK TILI MILLIY VA TA’LIMIY KORPUSINING NAZARIY VA AMALIY MASALALARI”

mavzusidagi

Respublika ilmiy-amaliy konferensiyasi

materiallari

(2023-yil 5-may)

Toshkent – 2023

O‘zbek tili milliy va ta’limiy korpusining nazariy va amaliy masalalari / Respublika ilmiy-amaliy konferensiya to‘plami. Elektron nashr / ebook. – Toshkent: ToshDO‘TAU, 05.05.2023. – 238 b.

Mazkur to‘plamdan o‘zbek tilining milliy korpusi hozirgi holati natijalari, muammolari, vazifalari, o‘zbek tilining ta’limiy korpusini yaratish masalalari, amaliy filologiya yo‘nalishlari istiqbollari, o‘zbek tili ta’limida korpuslardan foydalanish, tabiiy tilga avtomatik ishlov berish muammolariga oid materiallar joy olgan.

Ushbu ilmiy to‘plamdan korpuslingvistikasi, pedagoglar, amaliy filologiya sohasi mutaxassislari, tadqiqotchilar, magistrantlar va sohaga qiziquvchi talabalar foydalanishlari mumkin.

Mas’ul muharrir:

B.Mengliyev – Alisher Navoiy nomidagi Toshkent davlat o‘zbek tili va adabiyoti universiteti Amaliy filologiya kafedrasi mudiri, filologiya fanlari doktori, professor.

Taqrizchilar:

S.Muhamedova – Alisher Navoiy nomidagi Toshkent davlat o‘zbek tili va adabiyoti universiteti O‘zbek tili ta’limi fakulteti dekani, filologiya fanlari doktori, professor.

B.Elov – Alisher Navoiy nomidagi Toshkent davlat o‘zbek tili va adabiyoti universiteti Kompyuter lingvistikasi va raqamli texnologiyalar kafedrasi mudiri, texnika fanlari bo‘yicha falsafa doktori (PhD).

O‘ZBEKISTON RESPUBLIKASI OLIY TA’LIM, FAN VA INNOVATSIYALAR VAZIRLIGI

ALISHER NAVOIY NOMIDAGI TOSHKENT DAVLAT O‘ZBEK TILI VA ADABIYOTI UNIVERSITETI

O‘ZBEK TILI MILLIY VA TA’LIMIY KORPUSINING NAZARIY VA AMALIY MASALALARI

mavzusidagi Respublika ilmiy-amaliy konferensiyasi materiallari

(2023-yil 5-may)

Toshkent – 2023

TASHKILIY QO‘MITA

Sirojiddinov Shuhrat Samariddinovich (Toshkent davlat o‘zbek tili va adabiyoti universiteti)

Jamoliddinova Odinaxon Rustamovna (Toshkent davlat o‘zbek tili va adabiyoti universiteti)

Mengliyev Baxtiyor Rajabovich (Toshkent davlat o‘zbek tili va adabiyoti universiteti)

Muhamedova Saodat Xudoyberdiyevna (Toshkent davlat o‘zbek tili va adabiyoti universiteti)

Hamroyeva Shahlo Mirjonovna (Toshkent davlat o‘zbek tili va adabiyoti universiteti)

Zaripov Rafiqjon Ergashboy o‘g‘li (Toshkent davlat o‘zbek tili va adabiyoti universiteti)

TAHRIR HAY’ATI

Mengliyev Baxtiyor Rajabovich

Islomov Ikrom Xushboqovich

Hamroyeva Shahlo Mirjonovna

Zaripov Rafiqjon Ergashboy o‘g‘li

Elova Dilrabo Qudratillayevna

Murtazayev Abror Odilovich

Musulmonova Kamola Husniddin qizi

CORRESPONDENCES, ANALOGIES AND DIFFERENCES OF CORPORA

KHURSANOV Nurislom

Doctor of philosophy (PhD) in Philological Sciences, Senior Teacher

Department of English Philology,

Alisher Navo’i Tashkent State University of Uzbek Language and Literature

E-mail: nurislomkhursanov92@gmail.com

https://orcid.org/0000-0001-5714-2745

Abstract: It is difficult to imagine today's linguistics without artificial intelligence. The development of time, new technologies in science and society, requirements and principles of science teaching and learning are changing. Linguistics is no exception. In this article, the role of language corpora in the way of creating technologies and tools that ensure the storage, search, analysis and transfer of information and knowledge, improving systems for recognizing and creating written and spoken texts in natural language, and developing artificial intelligence systems is talked about.

Keywords: corpora, sub-corpus, method, extralinguistic factors, time, genre, retsenziya.

Annotatsiya: Bugungi kun tilshunosligini sun’iy intellektsiz tasavvur qilish qiyin, albatta. Zamon rivojlanganligi, fan va jamiyatda yangi texnologiyalar, fanni o‘qitish va o‘rganish talab va tamoyillari o‘zgarmoqda. Tilshunoslik fanlari ham bundan mustasno emas. Mazkur maqolada axborot va bilimlarni saqlash, qidirish, tahlil qilish va uzatishni ta'minlaydigan texnologiyalar va vositalarni yaratish, tabiiy tilda yozma va og'zaki matnlarni tanib olish va yaratish tizimlarini takomillashtirish, sun'iy intellekt tizimlarini rivojlantirish yo‘lida til korpuslarining o‘rni haqida so‘z boradi.

Kalit so‘zlar: korpus, sub-korpus, metod, ekstralingvistik omillar, janr.

Why do we need historical corpora? It would seem that the answer to this question is simple and obvious: of course, for the same purposes as the corpora of modern texts: to search for material for the linguistic analysis of medieval texts, both traditional and, first of all, corpus - quantitative, statistical, distributive - methods for searching for illustrations, for acquainting students of medieval languages with textual practice to a greater extent than grammars, textbooks, and anthologies allow.

But this answer needs significant clarifications, which are determined by the features of the composition and structure of historical documents, the chosen forms of storage and marking of their electronic versions, the ultimate goals and tasks of historical and linguistic research, some extralinguistic factors, in particular, the number of handwritten monuments that have survived to our time.

“The expediency of creating and the meaning of using corpora,” writes V.P. Zakharov about modern corpora, is determined by the following prerequisites:

1) a sufficiently large (representative) volume of the corpus guarantees the typicality of the data and ensures the completeness of the presentation of the entire spectrum of linguistic phenomena;

2) data of different types are in the corpus in their natural contextual form, which creates the possibility of their comprehensive and objective study;

3) once created and prepared data array can be used repeatedly, by many researchers and for various purposes” [Zakharov, 2005: 3].

The second and third provisions undoubtedly characterize historical corpora, but the first is not applicable to them due to objective reasons. The first feature of the historical corpus is the impossibility of achieving a volume comparable to modern ones, which is necessary to obtain a statistically significant amount of studied facts for solving some problems related primarily to the study of lexical units and syntactic constructions, and the unattainability of equal representation of texts of different types and genres, different time periods and book schools (“the volume and completeness of the historical corpus is limited by the number of documents that have survived from one time or another, and this also limits the possibility of achieving a balanced corpus” [Baranov, 2010: 224]. But both are compensated by the possibility of using a subcorpus randomly formed by history and known by its parameters (time, genre, recension, etc.), which also represents unsurvived sources of a certain period, and examples extracted from the subcorpus allow not only to observe textual usage, but also when comparing them with the corresponding facts of the previous and subsequent periods, identify the directions of changes and explain them. Taking into account these two characteristics of the historical corpus - the inclusion of surviving sources in a wider range of documents of a particular era and the comparability of the data extracted from the subcorpus with the materials of other subcorpuses – “allows us to remove the restrictions associated with the underrepresentation of linguistic phenomena in the texts, and even single, statistically indicative cases of variability considered as sufficiently reliable on a wider time background” [Baranov, 2010: 225].

The third feature of the corpus is the inclusion of a handwritten document as a unity of a physical copy of the codex, a letter, fragments that have come down to us and the text (work) on them [Khursanov, 2021: 311-318]. Therefore, the object of a meta description is not only a work (text), but also a physical medium - a book, a letter or parts thereof, which have parameters that should be classified as purely reference in terms of corpus technologies - a place and organization of storage, a fund, a collection, a collection , number, size, material, watermarks, number of columns and lines per page, etc., and such characteristics that can be used in the formation of a research subcorpus and in the preparation of a linguistic sample. For example, such properties of the manuscript, describing the execution of the document, include the time and place of the production of the list, the number and boundaries of handwriting, and some. Therefore, when developing the principles for storing and marking up a historical document, it is necessary to take into account the presence of two equally important units of metadescription in the corpus - the physical carrier of the text and the text of the work.

Text edition or manuscript edition?

The uniqueness of each manuscript, each list, each surviving fragment implies the presentation in electronic form not only of the text of the work (and not in fragments, but in the volume that has come down to our time), as is often done in the corpora of modern texts, but also the preservation of page and linking it line by line to a specific physical medium - the manuscript, which, together with the meta descriptions of both the manuscript and the text, in fact, is the preparation of two electronic editions - the publication of a specific list and the publication of the text.

This situation requires answers to questions that do not arise when creating a modern corpus, for example: should a machine-readable copy, and if so, how accurately convey structural, graphic-orthographic, paleographic (that is, those that are given to us directly for observation) features list? what ways to store and demonstrate these features: with the help of descriptions, comments in a free format, or additional information structured in accordance with a certain format, or storing structural features in the transcription itself and calculating their presence and parameters automatically?

The inseparable connection between the manuscript and the text requires taking into account their relationship with each other: the codex may contain several works, each of which needs a separate meta-description, including information about the author(s), translator(s), time of writing or translation of the text, genre, and some. etc., the text of a work may not be listed in the order of parts, may contain mechanical inserts from other texts - these and similar features should be taken into account when marking the corpus.

Features of a medieval written monument impose additional requirements on the demonstration of texts, concordances, lists: units in them must be identified using the names of the list and text, page number and address on the page, which allows us to speak of a complex resource that combines the demonstration of full texts - fragments - concordances, as well as special types of lists of linguistic units - alphabetical, frequency, lists of n-grams.

Original and transcription: degree of correspondence

Information about the code, the location of the text on it and its structure can be given both with the help of descriptive characteristics (for example, the number of sheets in the code, the place of additions), and by transferring the maximum number of such features in an electronic copy, followed by automatic calculation [Khursanov, 2021: 247-53].

In the second case, such markup tools are used that allow, during the demonstration, to transfer the manuscript, as is customary in historical-linguistic publications of sources, sheet to sheet, page to page, line to line, sign to sign. The problem of the accuracy of the transmission of graphic, paleographic and spelling features of the writing of medieval Slavic texts in electronic publications has been repeatedly discussed in special works devoted to the standards and formats for encoding characters of the Old Slavonic alphabet, as well as descriptions of those solutions that are adopted in specific systems for input and storage of transcriptions of medieval documents.

In any case, adherents of the exact transfer of the structure and composition of the manuscript, graphical orthographic and paleographic variability face two differently directed tasks: to find ways to reflect the features of the original as accurately as possible in a machine-readable copy, on the one hand, and to create algorithms, procedures and programs that would necessity (search, ordering, demonstration) eliminated variability, leveled formal differences.

The fourth feature of the historical corpora of medieval Slavic monuments is the presence of multiple non-standard correlations between the manuscript and the text located on it and its non-standard graphic and orthographic record, which must simultaneously be stored in an electronic copy and eliminated during data processing (for example, during lemmatization, search and display of data) [Raupova & Sultonova, 2021: 116-124].

The main problem for automating the linguistic markup of the Slavic historical corpus is the lack of electronic morphological (grammatical) dictionaries. Therefore, their creation is today one of the main tasks in the field of applied Paleo-Slavistics. From publications, it is known that at present, several teams are creating on the basis of different written sources a) databases of textual precedents with manually assigned morphological marks, b) dictionaries based on dictionaries, including contexts, and, accordingly, word forms) electronic grammar dictionaries containing information about fixed and unfixed, but potentially possible word forms, d) procedures and programs based on probabilistic methods and other algorithms. Which of these ways will be the most effective for creating systems for automated morphological analysis of texts, the main characteristic of which is an unnormalized, varying graphic and spelling form of a word, time will tell.

Conclusion

The correspondence between the original, on the one hand, and the translation(s) and editions, on the other hand, is much different than in the modern corpus: the original from which the translation was made may not be known, and its role is played by one of the most lists close to it in the original language. In the second and third types of corpora, lists of one work should be considered as a whole equal in terms of corpus construction. Therefore, it should provide for the possibility of showing aligned units in the direction from any of the lists, and not just, for example, from ancient to later.

References

1. Baranov, V.A. Software Tools and User Interfaces Designed for HistoricalLinguistic Purposes of Project ‘Manuscript’. Newsletter of the Association "History and Computer" // Proceedings of the International Conference "Computer Technologies and Mathematical Methods in Historical research”, Petrozavodsk, July 11–16, 2011.

2. Khursanov NI. (2021) Creating a corpus of texts. Journal of Education and Innovative Research;4(1):247-53.

3. Khursanov, N. I. (2021). On the theoretical and practical foundations of language corpora. *Asian Journal of Multidimensional Research*, *10*(9), 311-318.

4. Raupova, L., & Sultonova, Sh. (2021). Scientific Basis of Compiling a Dictionary of Grammatical Terms. *Current Research Journal of Philological Sciences*, *2*(10), 116-124.

5. Zakharov, V.P. (2005) Corpus Linguistics: Educational and Methodological Guide. – St. Petersburg.

MUNDARIJA

1. SHO‘BA. O‘ZBEK TILINING MILLIY KORPUSI: NATIJALAR, MUAMMOLAR, VAZIFALAR

2. SHO‘BA. O‘ZBEK TILINING TA’LIMIY KORPUSINI YARATISH MASALALARI

3. SHO‘BA. AMALIY FILOLOGIYA YO‘NALISHLARI ISTIQBOLLARI