ISSN NO: 2769-996X
Date of Publication: 21-06-2022

General Principles of Creating a Marked Language Body

Abbosjon Mirzayev
Samarkand branch of Tashkent University of Information
Technologies Named after Muhammad Al-Khwarizmi Master

Abstract: The article discusses the basic principles of corpus linguistics, a new field in applied marked Uzbek linguistics, as well as the process of designing and constructing corpus. Examples of achievements of world linguistics in the creation of corpus resources in the Uzbek language are given. Practical linguistic experience proves how to set up the first stage of corpus linguistics.

Keywords: corpus, marked corpus linguistics, genre-thematic structure, VP Zakharov, SY Bogdanova, LN Zasorina, frequency dictionary, computerization.

The hull project is the stage of its creation, the main stage of creating a perfect hull, which will include ways to improve it later. The concept of corpus, a new form of traditional card indexing for linguists; they were computerized by the twentieth century, making them widely available. Of course, the Internet has played an important role in the transformation of card indexes. As a result, there is a common version of large texts that allow for a variety of linguistic research. In this regard, the issue of the volume and balance of language material, which is the basis for dictionaries and grammars, was on the agenda, especially in the process of creating national corps. The question of the representativeness of the corpus was solved by the sufficiency and diversity of the texts. According to VP Zakharov and SY Bogdanova, when considering the genre-thematic structure of the corpus, it is necessary to pay special attention to the problem of what unit to take as a corpus text [1.36]. For example, should small advertising text in newspapers be considered as separate text or should they be combined into one text? Is a newspaper article a text or should a single issue of a newspaper be evaluated as a single text? Should each poem be a single text or should a collection of poems be included as a whole? Should published letters written in response to each other, essentially discussing a single topic, be approached as a single text or as a separate corpus unit? These questions are answered based on the type of corpus that composes, and the function it performs later. Depending on whether it is a national corps or a special corps, a corps unit is determined. VP Zakharov and SY Bogdanova also cite the issue of chronology as an important aspect of the process of designing the building. Another important question is what part of the body is taken from the original form of the text and what is excluded. Since the images contained in the text do not belong to the language material, it is also important to remove them from the body of the text, and to adapt the tables to the body. They are important for expressing the content of the text, but if left in the body, they can be difficult to mark. Quotations, quotations, units of measurement, units of measurement also require special attention. If the listed issues are solved on the basis of certain principles at the design stage, some of them will be solved in the process of building the building, in the use of the building. However, feedback from the user should also be considered before starting the case. Experts distinguish the following stages in the technological process of building the case [2]:

1. Ensure that the text enters the body in accordance with the specified source.

2. Automatic text processing. Electronic text input to the case can be obtained in a variety of ways: handwritten, scanned, copyrighted, gift, exchange, Internet, original models provided by the publisher to the case.

3. Analysis, initial processing of the text. At this stage, the texts received from various sources are subject to philological examination and editing.

4. Conversion, graphical analysis. Some texts go through the first machine process where the re-encoding process takes place, and the non-text parts (pictures, tables) are deleted or changed. Copying links in text, deleting borders (in MSDOS texts), hyphens, and other characters are the same.

Graphic analysis consists of dividing the text into parts (words, links) and deleting the noun element.

5. Defining, formalizing a non-standard (non-lexical) element, a special text element (abbreviated name (name, surname), assimilation lexeme written in another alphabet, name given to the picture, comment, title,

ISSN NO: 2769-996X
Date of Publication: 21-06-2022

bibliography, etc.) revision based on different criteria. Of course, these actions are performed automatically by the text editor.

The next step in the design of the case is the selection of the source. The importance of the corpus is not in the fact that it compiles and arranges a wide range of texts in one language, so it is based on several criteria. What is the basic unit of the corpus when sorting the corpus material, what is its size (preferably how many words it contains), what source is the written text based on, how much is it, what area of the language does the text belong to? issues such as The first answer to this question was given by RG Piatrovsky and his students in 1965-80. They developed a dictionary of frequencies and the principles of text selection for linguistic and statistical research. This problem is also raised in the preface to the frequency dictionary [3,936] edited by L.N. Zasorina. Then for the first time a statistical method of text selection Factors such as i, size, quantity are listed. The basic units of the body are the word, the stem (base, lemma) and the sentence. The size of the building to be built is determined by the purpose of the building. If it aims to study letters, letter combinations, sounds, diphthongs, it doesn't have to be that big. If the lexical units of the text are formed in order to study the morphological phenomenon, syntactic, methodological specificity, a large volume is required. According to SA Sharov [4], in the selection process, the choice of the genre of the text (prose, drama, poetry, scientific text, newspaper, magazine material, etc.), the period of the text (modern, 10-year, 50-year and classical text) issues such as whether the text should be in literary language only or whether other sources should be included. In this process, the corpus compiler will, of course, consult a linguist, a linguist, or a questionnaire method. In the process of creating a case, based on his experience, the author considers the total size of the case, the time of publication of the text, the number of texts, the size of the elementary selection, the type of genre to be selected. The survey was used by the authors of The American Heritage Intermediate Corpus. There are 5 million words and 22 types of texts in English for children and teenagers. A survey was sent to 221 schools in the United States to determine which text was appropriate. After studying the results of the survey, a list of 19,000 books was compiled. Based on this, 1,045 texts of 500 words each were selected. In conclusion, it should be noted that the most important stage in the design of the case is the selection of material (text), sorting, its technical adaptation to the case.

References

1. Shaykhislamov, N. (2020). Cognitive Linguistics, the Symbolic and Interactive Functions of Language. Education and Science in the XXI Century, 1 (6), 390-393.
2. Shaykhislamov, N. (2020). Linguistic and Cultural Aspects of Bread Baking Terms in English and Uzbek Languages. Integration of Science and Education: Opportunities and Trends, 62-64.
3. Shaykhislamov, N. (2020). Use of modern pedagogical technologies in teaching mother tongue. Problems of science and education in Uzbekistan: problems and solutions, (3), 71-73.
4. Shayxislamov, N. (2020). Paralinguistic and extralinguistic means of speech. Scientific and practical research in Uzbekistan, 36-37.
5. Shayxislamov, N. (2020). Fundamental theoretical foundations of gender linguistics and gender issues. Current issues of linguistics and translation studies in the XXI century: theory, practice, innovation, 144-146.
6. Shayxislamov, N. (2020). Use of interactive methods in mother tongue lessons. Prospects of Development of Science and Education, 1 (2), 247-249.
7. Shayxislamov, N. (2020). Comparative typological analysis of investment and credit terms used in Uzbek and Russian languages. General Problems of Philology, 204-206.
8. Shayxislamov, N. (2020). Pedagogical bases of mastering students' native language and literature. Education and Science in the XXI Century, 1 (6), 421426.
9. Shayxislamov, N. (2020). Gender aspects in the anthropocentric study of zoonym component metaphors. Education and Science in the XXI Century, 1 (6), 304-309.
10. Shayxislamov, N. (2020). History and theoretical foundations of linguocultural studies. Problems of science and education in Uzbekistan: problems and solutions, (2), 219-221.
11. Shayxislamov, N. (2020). The philosophical yellow in Abay's poetry. Issues in the Study of Abay's Creativity and Scientific Heritage, 154-157.

ISSN NO: 2769-996X
Date of Publication: 21-06-2022

12. Shayxislamov, N. (2020). Fundamentals of technology, basic language and gender theory. Continuity of foreign language teaching in preschools, secondary schools and higher education institutions, 734-735.
13. Shofqorov, A. M. (2018). Phonetic repetitions and their methodological features. Social Sciences and Humanities in the Education System, (4), 79-86.
14. Shofqorov, A. M. (2019). Use interactive teaching methods in the study of contract types. Language and Literature Issues, (5), 18-19.
15. Shofqorov, A. M. (2020). Phonetic repetitions in Hamid Olimjon's poetry. General Problems of Philology, 60-62.