International Journal of Academic Pedagogical Research (IJAPR)
ISSN: 2643-9123
Vol. 5 Issue 3, March - 2021, Pages: 155-156

Linguistic Bases Of Creating A Corpus Of Grammatical Terms

Sultanova Shahnoza Akmalovna

*Lecturer at the Department of Foreign Languages
of the Faculty of Translation Theory and Practice
at the Tashkent State University of Uzbek Language and
Literature named after Alisher Navoi*

Abstract: *The article deals with the theoretical aspects of corpus linguistics. The author presents an analysis of the declared problem and a systematization of existing theories, terms and concepts. It includes examples of English, Uzbek and Russian linguistic corpora.*

Keywords: *corpus, corpus linguistics, markup, representativeness, balance.*

INTRODUCTION

Definition of the term “corpus”:

Any research carried out by a linguist should be focused at least on the following stages of activity:

1) Selection of principles and bases (standards) for the classification of objects under study.

2) The process of distributing objects into classes in accordance with these bases (standards).

3) Understanding, interpretation, interpretation of the results of the distribution of objects into classes, explanation of the reasons for such distribution [1, P.29]. At the same time, the first stage of this activity implies the presence of objects to studied, i.e. collection of empirical material for constructing a theory at the final stage of research. Currently, corpus linguistics is becoming increasingly popular in the collection and analysis of practical material. In addition, this is a natural step in linguistics following the rapid development of information technologies.

DISCUSSION

Corpus linguistics appeared in the 60s of the twentieth century, mainly on the material of the English language, but very quickly began to arise Corpus because of other languages. At Brown University in the USA in 1963, scientists W. N. Francis and G. Kucera created the first corpus of texts on an electronic medium (Brown Corpus, free access from the website of the University of Leeds: http://corpus.leeds.ac.uk/ protected/). It contained 500 texts of the 15 most popular genres of English-language prose in the United States, with 2,000 words each. The case accompanied by a frequency index and an alphabetical frequency index, as well as some statistical distributions.

A corpus is a collection of texts in one or more languages that linked by certain parameters.

MATERIALS AND METHODS

The corpus is a collection of written and oral statements. The corpus data is usually digitized, i.e. stored on computers and available electronically. In this case, the constituent parts of the corpus, the texts, consist of data, as well as, possibly, of metadata describing this data, and of linguistic annotations that organize this data.

Corpus linguistics as a separate branch of linguistics finally formed in the first half of the 90s of the twentieth century. At the same time, the conceptual apparatus began to take shape. Thus, J. Sinclair describes the corpus as «a collection of naturally-occurring language text, chosen to characterize a state of variety of a language» [3, P. 171]. This definition emphasizes one of the fundamental principles when choosing texts for building a corpus – we are talking about unedited texts, that is, unedited texts. The language presented in the form in which it manifests itself in speech (whether oral or written). In addition, the corpus contains not the existing “samples” and “prescriptions” for the correct construction of the message, but as many “variants” of the language as possible, even if some of them are located on the periphery of the language system.

In the following years, the concept of “Corpus” increasingly concretize «A corpus is a collection of texts, designed for some purpose, usually teaching or research. [...] A corpus is not something that a speaker does or knows, but something constructed by a researcher. It is a record of performance, usually of many different users, and designed to be studied, so that we can make inferences about typical language use. Because it provides methods of observing patterns of a type which have long been sensed by literary critics, but which have not been identified empirically, the computer-assisted study of large corpora can perhaps suggest a way out of the paradoxes of dualism» [2, P. 239-240].

Thus, each of the presented definitions of the concept of “corpus” emphasizes the following:

1) Many texts must be submitted in electronic form (on the Internet or on disk);

2) Language data should be marked up for analysis for linguistic purposes;

3) As a result of the analysis, there should be a possibility of different distribution of the resulting language material (by genre, year of creation of the text, topic, etc.).

If we consider the first point, then the essential criterion here is the availability of the corpus of texts in electronic form. The entire existing set of text corpora can be divided into three broad categories: 1) freely available; 2) partially available; and 3) commercial.

The first category includes a fairly limited number of currently existing text corpus. The most extensive (with a total volume of more than 500 million words) is the National Corpus of the Russian Language (www.ruscorpora.ru). Most of the existing corpora belong to the second category, but for solving specific linguistic problems, such partial access is often sufficient. So, in the British National Corps (http://www.natcorp.ox.ac. uk/) the output of the result is limited to 50 random examples, in addition, many features of the search interface that comes with the full (paid) version of the corpus are missing. Along with this, there is a non-commercial version of this corpus (http://corpus.byu.edu/bnc/), available after a simple registration procedure, in which 100 million words in 1980-1993 texts are presented for search. A fairly representative selection from the Turkic corpus the Uzbek language is available in a special program either on the basis of the Academy of Sciences or on the platform of the National library of Uzbekistan.

The third group includes, for example, the Bank of English (Bank of English) with a free trial subscription for one month to get access to Collins Wordbanks Online (553 million words) (http://www.collinslanguage. com/ content-solutions/wordbanks), after which you need to purchase the paid version of the corpus.

RESULT

Thus, the corpus is a representative array of unedited texts presented in electronic form, usually marked up for analysis for linguistic purposes, provided by a relatively easy-to-use search engine, representing as many “variants” of the language as possible.

In the period of the origin of corpus linguistics, the questions of computerization of this direction were not raised, and “researchers pointed out the possibility of neglecting language variability, i.e. territorial, social, professional, age, gender, individual and similar differentiation of the language” [3, P.76-77]. Today, by neglecting it, we consciously limit ourselves to various limits when studying texts of a particular language, which calls into question the objectivity of this kind of research. With the advent of electronic corpora, the variety of forms of language existence has become more visible, and the possibilities for studying language data have expanded. The modern linguistic corpus contains hundreds of millions of word usages, and the fact that with the help of an electronic corpus, the results of examples of word usages can be obtained in a few fractions of seconds, greatly simplifies the task for linguists. The presented typology of corpora, without claiming to be all-inclusive, shows us the existing variety of text corpora and allows us to orient ourselves in it for further scientific research.

REFERENCES:

1. Kryvnova O. F. Areas of application of speech corpora and experience of their development // Tr. XVIII Session of the Russian Acoustic Society of RAO. Taganrog, 2006.
2. Melnikov G. P. System typology of languages: Principles, methods, models / Ed. by L. G. Zubkov. Moscow: Nauka, 2003.
3. Sinclair J. Corpus, Concordance, Collocation. Oxford, 1991.
4. Stubbs M. Words and phrases: corpus studies of lexical semantics. Oxford, 2001.