EUROPEAN INTERNATIONAL JOURNAL OF MULTIDISCIPLINARY RESEARCH AND MANAGEMENT STUDIES

VOLUME04 ISSUE05
DOI: https://doi.org/10.55640/eijmrms-04-05-40 Pages: 255-257

ANALYZING CORPUS LINGUISTICS

Shayusupova Nargiza Bakhtiyorovna
Assisitant teacher, The Branch of Astrakhan State Technical University, Uzbekistan

ABOUT ARTICLE

Key words: Computer linguistics, computer technologies, machine- readable format, unified, structured.

Received: 21.05.2024
Accepted: 26.05.2024
Published: 31.05.2024

Abstract: Corpus linguistics is the section of computer linguistics which engaged in development of general principles for the construction and use of linguistic corpora using computer technologies. Under the linguistic or language corpus of texts is understood large, machine- readable format, unified, structured, marked, philologically competent array of linguistic data designed to solve specific linguistic problems.

INTRODUCTION

Corpus linguistics is the section of computer linguistics which engaged in development of general principles for the construction and use of linguistic corpora using computer technologies. Under the linguistic or language corpus of texts is understood large, machine- readable format, unified, structured, marked, philologically competent array of linguistic data designed to solve specific linguistic problems. Presently, there are many definitions of corpus. For instance, E. Finegan states that corpus – representative collection of texts in machine- readable format including information about the author, reader or the audience and situations in which the text was produced. [1]

Corpus contains a large collection of representative samples of texts covering different varieties of language used in various domains of linguistic interactions. Theoretically, corpus is (C)apable (O)f (R)epresenting (P)otentially (U)nlimited (S)elections of texts. It is compatible to computer, operational in research and application, representative of the source language, processable by man and machine, unlimited in data, and systematic in formation and representation [2]

Corpus linguistics appeared in 60s XX century, mainly on the material of English, but very quickly Corpus began to appear on the base of others

languages. At Brown University USA in 1963 by scientists W. N. Francis and

G. Kuchera created the first electronic media (Braunovsky building, free access from the university website Leeds City: http://corpus.leeds.ac.uk/pro-

tected/). It contained 500 texts 15 the most popular genres in English

US prose, 2,000 words each. K core enclosed was a frequency index and

alphabetical frequency index, as well as some statistical distributions.

Corpus linguistics as a separate the section of linguistics is finally formed

took place in the first half of the 90s. XX century the same time, concepts began to take shape of apparatus. Thus, J. Sinclair describes corpus like «a collection of naturally-occurring language text, chosen to characterize a state of variety of a language» [3] This definition emphasizes one from the fundamental principles when choosing texts for building a corpus – speech.

We are talking about unedited texts, i.e. language is presented as it is manifested himself in speech (whether it was oral speech or written). In addition, in the corpus non-existent “samples” are presented and “prescriptions” for proper construction message, and as much as possible quality of language “variants”, even if some of them are located on the periphery of the linguistic systems. Thus, in each of the representations of definitions of the concept “corpus” under the following is highlighted:

1) a lot of texts must be before delivered in electronic form (on the Internet

or on disk);

2) language data must be sized for analysis for linguistic purposes;

3) as a result of the analysis there must be a possibility of differences

distribution of the received language material (by genre, year of creation of the text, subject, etc.).

While scientific researcher scholars came across with a number of problems. First of all due to note the problem of representativeness these corpuses, i.e. ability to reflect everything properties of the

problem area. Representatives are determined by phonetic, morphological, syntactic, stylistic parameters. Corpus creator first raises the question, what kind of corpus and for whom is it creates. It is impossible to imagine a computer all texts or all conversations of this language, so the creators of the corpus are guided by focus on researchers who this building intended.

Lexicographers have enough of a large corpus with examples of rarely used words and/or their forms, but for specialized research by grammarians, stylists, etc.

It is necessary to cover different styles and genres of national language. And here the researchers began to face with the problem of selecting texts for

housings. For example, W. Francis and G. Kuchera Our goal was to present a corpus of texts that looking for clear and precise selection criteria:

1. Origin and composition of the text (Thor had to be native born American English, dialogue into semantic nests, i.e. belongs more to the field of semantics than logic, and is difficult to formalize.

Thirdly, buildings are built with the aim of giving objective information, therefore, before The creators of the corpus are tasked with reducing

factor of subjectivity when selecting texts for the body and develop strict and clear selection criteria

REFERENCES

1. Finegan E. LANGUAGE: its structure and use. – N.Y.: Harcourt Brace College Publishers, 2004
2. Dash, N.S. (2005) Corpus Linguistics and Language Technology: With Reference to Indian Languages. New Delhi: Mittal Publications.
3. Sinclair J. Corpus, Concordance, Collocation. Oxford, 1991
4. Волоснова Ю.А. Корпусная лингвистика: Проблемы и перспективы //ЛЕСНОЙ ВЕСТНИК/ 7/2006