Academicia Globe:
Inderscience Research

ISSN: 2776-1010 Volume 3, Issue 2, Feb, 2022

BASIC THEORETICAL PRINCIPLES OF CORPUS LINGUISTICS

Azizbek Vosiljonov
Magistrant, Fergana State University
azizbekvosiljonov@gmail.com

Abstract
The article discusses the basic principles of corpus linguistics, a new field in applied 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 linguistics, corpus, science, computer translation; authentic and semi authentic texts.

Introduction

Important features of global development are determined by the penetration of technology into the industry, the formation of computer programs, the process of integration. In the process of global integration, the creation of a natural language-based information style has become a vital necessity. Today, one of the most important tasks is to raise the status of the Uzbek language and make it one of the most influential languages. Computer linguistics is one of the convenient opportunities created to fulfill this need. The science of computer linguistics plays an important role in bringing the Uzbek language to the world stage, making it one of the By the beginning of the 21st century, with the development of information technology, the process of globalization has reached a new level. The rational use of the achievements of mankind in the development of science, technology, culture, industrial relations, democracy, the rule of law and justice will pave the way for Uzbekistan to become one of the most developed countries in the world. The process of globalization requires rapid development in all areas. The computer system, which is a product of technical progress, provides convenience in all areas, the rapid delivery of information, translation, editing processes in a short time, an artificial language that serves as a tool between communicators of different nationalities, ie information -provides the formation of computer style.

Corpus (corpus) means "body" in Latin. "A body is a collection of electronic texts that means words, phrases, grammatical forms, and the meaning of a word to be found through a specific search engine." world languages, and learning and teaching languages.[1]

The purpose of corpus linguistics is to introduce the basics of corpus technology, as well as corpus linguistics, based on an empirical approach to language learning. The tasks are to develop skills in corpus linguistics, to show the theoretical and practical importance of corps in conducting scientific linguistic research, to determine the role of computer technology in the system of sciences related to linguistics.[2] Corpus linguistics is the creation of corpus and the implementation of linguistic research based on them, the objective and linguistic directions of linguistic systems: lexicographic research, the

ISSN: 2776-1010 Volume 3, Issue 2, Feb, 2022

description of the lexical layer of the language, the ratio of words in the language vocabulary, lexical-semantic, structural changes, the study of the grammar of natural languages, the essence of the language system and the description of its use

The first records of the corpus in world linguistics date back to the 1940s. When we talk about the history of the corps, we first mention the Brown corps, which was built in 1961-1964. Created at Brown University, it contains 500 text fragments of 2,000 words each.

The main directions of modern corpus linguistics are: first, the creation of these dictionaries and lexicographic research, all dictionaries of modern English are corpus-based (Collins, Webster, MacMillan, etc.); second, to obtain accurate information about the lexical structure of languages through the study of corpuscles, to establish the frequencies of use of words. [3] As a result of the search given to determine the frequency of a particular word on the basis of the corpus, using diagrams and graphs, the ordinal number of the word is inversely proportional to its frequency, because the word in the second ordinal number is less than the first digit word. it is clear that it is used less than the third. No frequency dictionary can provide accurate body information, because language is constantly changing and the frequency of words is also relative. Based on this practice of the corpus known as the Sipfa law, the chances of identifying frequently used words in any language are now high.

In order to provide accurate and effective areas of language processing, conversion to computer language, and language modeling for artificial intelligence, first of all, linguistic research must be performed with high accuracy that fully meets the technical requirements.[4] Thus, the effectiveness of language corporations and their wide range of possibilities are closely linked to the results of linguistic research, the level of perfection of lexicographic interpretations, and the semantic differentiation of lexemes and terms. Approaches to solving practical problems of computer linguistics can be divided into the following classes:

- Rule-based approaches;
- Approaches based on machine learning;
- Hybrid approaches.

Language corporations serve as an important linguistic source in all approaches. In particular, in machine learning and hybrid approaches, the computer understands natural language using corpses. Quality corps is one of the main tools in this process. In rule-based approaches, an algorithm created from language corpora is used to evaluate the performance of a program.[5] Language corpora are widely used in the fields of artificial intelligence technology development, machine learning, in-depth teaching, as well as in the validation of linguistic theories, language teaching, and other fields of linguistics.

Experts distinguish the following stages in the technological process of building the case [6]:

1. Ensure that the text enters the body in accordance with the specified source.
2. Automatic text processing. The electronic text included in the case can be obtained in various ways: handwritten, scanned, copyrighted, gift, exchange, Internet, original models provided by the publisher to the case developer.

ISSN: 2776-1010 Volume 3, Issue 2, Feb, 2022

3. Analysis, initial processing of the text. At this stage, texts received from various sources are subject to philological examination and editing.
4. Conversion, graphematic 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 syllables, removing borders (in MS DOC texts), hyphens, and other characters are the same. Graphic analysis involves performing actions such as dividing the text into parts (words, links) and deleting the noun element.[7]
5. Defining, formalizing a non-standard (non-lexical) element, a special text element (abbreviated name (name, surname), a lexeme in another alphabet, a name for a picture, a comment, a title, a list of references, etc.

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. http://rusorpora.ru
2. https://uzbekcorpus.uz
3. Dash N. S. Corpus Linguistics: A General Introduction, CIVIL, Mysore, 2010,
4. Захаров В.П. Корпусная лингвистика: Учебно-метод. пособие. –СПб., 2005.
5. Language. Literature, Education. 2018(3):68.
6. Abdurakhmonova N, Urdishev K. Corpus based teaching Uzbek as a foreign language. Journal of Foreign Language Teaching and Applied Linguistics (J-FLTAL). 2019;6(1-2019):131-7.
7. Abdurakhmonov N. Modeling Analytic Forms of Verb in Uzbek as Stage of Morphological Analysis in Machine Translation. Journal of Social Sciences and Humanities Research. 2017;5(03):89-100.
8. Kubedinova L. Khusainov A., Suleymanov D., Gilmullin R., Abdurakhmonova N. First Results of the TurkLang-7 Project: Creating Russian-Turkic Parallel Corpora and MT Systems. Proceedings of the Computational Models in Language and Speech Workshop (CMLS 2020) co-located with 16th International Conference on Computational and Cognitive Linguistics (TEL 2020) .2020/11: 90-101.