Impact Factor:					
	ISRA (India) = 4.971	SIS (USA) = 0.912	ICV (Poland) = 6.630
	ISI (Dubai, UAE) = 0.829	РИНЦ (Russia) = 0.126	PIF (India) = 1.940
	GIF (Australia) = 0.564	ESJI (KZ) = 8.997	IBI (India) = 4.260
	JIF = 1.500	SJIF (Morocco) = 5.667	OAJI (USA) = 0.350

SOI: 1.1/TAS DOI: 10.15863/TAS

International Scientific Journal

Theoretical & Applied Science

p-ISSN: 2308-4944 (print) **e-ISSN:** 2409-0085 (online)

Year: 2020 **Issue:** 11 **Volume:** 91

Published: 09.11.2020 **http://T-Science.org**

QR – Issue QR – Article

Azizakhan Abdugafurovna Rakhmanova
National University of Uzbekistan
PhD researcher of the department
of "Uzbek filology"

THE ROLE OF PARALLEL TEXT IN CORPUS LINGUISTICS

Abstract: *This article describes the source of corpus linguistics, the general description of corpus, its incentive as a database, its role in linguistic research and its importance in solving linguistic problems. The role of the national corpus, the author's corpus, the corpus of parallel texts in solving linguistic problems are analyzed. The investigation of phonetic, lexical, grammatical features of the general stages of language development, the definition of the size of the vocabulary, the role of language in determining the principles of language development are shown. The linguistic importance of the corpus in the study of lexicography, lexicology, syntax, methodology, the role of research in the study of linguodidactics, mother tongue, its place in foreign language education, the role of corpus linguistics methods in the analysis of language issues are discussed and were shown in the article. The importance of grammatical analysis acquisition of computer analysis and demonstrating are played a huge role in the article, to be more concrete, explicit words or concepts that semantically combined with information in a corpus of text, lemmatization, skimming, tokenization processes in computer analysis play an important role in the execution and authority of morpheme analysis.*

Key words: *corpus, the corpus of texts, the corpus of virtual texts, the corpus of parallel texts, concordance, the corpus of authorship, methodology, creative style, phonetic, lexical, grammatical features, lexicology, lexicography, syntax, kalka, semantic kalka, lemmatization, lemma, skimming, skimmer, tokenization, token.*

Language: *English*

Citation: Rakhmanova, A. A. (2020). The role of parallel text in corpus linguistics. *ISJ Theoretical & Applied Science, 11 (91)*, 66-70.

Soi: http://s-o-i.org/1.1/TAS-11-91-15 ***Doi***: https://dx.doi.org/10.15863/TAS.2020.11.91.15

Scopus ASCC: *1203.*

Introduction

As the language experiences a particular procedure of development, there was a need to collect and generalize language units, to summarize all their structures according to the lexical layer, to reflect historical units and to sum up information about a limited lexical layer. One of the urgent issues of worldwide development is the creation of a common database of national languages using current technical capabilities and with the help of this basis to determine the semantic capabilities of the language, the scope of content expression.

In world linguistics, the opportunity of research on applied linguistics, computer linguistics, corpus linguistics is extending. The development of corpus linguistics and the formation of a database in the national language is one of the key factors in increasing and expanding the vocabulary of the language. It is important to characterize the standards of the advancement of corpus linguistics, to feature the importance of computer dictionaries as a database, to clarify the linguistic support of thesauruses, concordances, to investigate corpus types, and their role in the development of national language.

Corpus linguistics was formed as a direction of computer linguistics. Because of the issues which ought to be tackled and the wide range of tasks, corpus linguistics is developing as a separate field. Computer linguistics and corpus linguistics are the directions of applied linguistics.

The situation of every language is controlled by its place in information and communication. Computer linguistics and corpus linguistics play a significant role in ensuring the capacity of the information exchange of language. Corpus plays a practical role in illuminating the features of a particular language, reflecting its capabilities, improving the field of linguistics, specifically,

developing computer lexicography and the concepts of the social sphere.

In recent years, the field of applied linguistics has received attention as one of the determining components of social development. The socio-economic need is growing for applied linguistics, because, first of all, it is based on improving teaching methods, and by generalizing grammatical features it teaches foreign languages and the national language as a mother tongue.

The main linguistic factor that promotes the development of computational linguistics is the corpus of electronic texts or the corpus of parallel texts. The word *Corpus* (corpus) is taken from Latin word which has the meaning "body". "Corpus is a collection of electronic text which serves to find words, phrases, grammatical forms and the meaning of a word with the help of a particular search engine [22]". In computer linguistics, the word "corpus" is widely used as a "corpus of texts". "The body of texts is a set of specific language units that are stored electronically to solve different problems for linguists. These can range from phonemes, graphics, morphemes to larger units - lexeme, sentences, and texts (fiction or scientific works, texts of newspapers and magazines). A special program which depends on how they are stored, it may find examples of each word or phrase, spelling variants, and synonyms. As a result of increased research of the corpus of texts, corpus linguistics was formed in linguistic.

Corpus linguistics studies various issues such as the main terminological apparatus of this direction, framework and fundamental qualities of the corpus, typology of the corpus, the aim of the corpus, factors which cause by its formality, history of linguistic corpus formation, modern status, the role of corpus technology, corpus technology, factors of forming corpus technology, the first generation of the corpus, the second generation of corps and schools of corpus linguistics. The fundamental task of it is to give information about Corpus compilation methodology, representative issues of corps, linguistic studies of the basis of the corpus, granular concept, corpuscular interfaces between an internet search engine and linguist, symbol, types of symbol, contemplated corps, the design process, and the basic standard case definition, linguistic plan, extra-linguistic definition, methods of creating extra linguistic corps, automatic morpheme and syntactic analysis, linguistic means of presenting texts, standardization of corpus linguistics; to find information by corpus and print it out, type of information, the usage of found information and others. The scientific literature on corpus linguistics also provides information about the usage of concordance, programs for working with concordant corpses, concordance, and parallel corpus, and the use of corps in the social sphere.

First of all, linguists collected computer-generated corpora in 1960. "The first corpus of computer-generated texts was the Brown Corps (VK, in English Brown Corpus, VS), which was created at Brown University in 1961, and it consisted of 500 text fragments of 2000 words each. Texts from The Brown Corpus are taken from magazines, American books, and newspapers which were published in the United States. Corpus authors U. Francis and G. Kucher formulated it as a large volume of material that was initially statistically processed: a frequency and alphabetical-frequency dictionary which was based on different statistical distributions.

The purpose of creating the Brown Corpus is to study and compare the written genres in English. Scientists, who developed the idea, paid attention to solve the problem from problem-solving and followed the principles of making and sorting the text. As an alternative, the corpus was built based on a statistical procedure, on the other hand, the statistics were determined by the free decisions of the corpus authors based on a professional awareness base. To achieve maximum objectivity in this complex process, there was a requirement for maximum formalized, procedure verification and control. Later, European researchers created a corpus which was based on this principle and first published in the UK in 1961, it consisted of 15 different genres (registers), 2000 words (word-forms) of 5000 texts. It covered 1 million British words of the English language, and they gave the name "Lancaster-Oslo-Bergen Corpus" named after British and two Norwegian Universities or The LOB Corpus for short.

Other inventions which were created in the Brown type were also very important for researchers. In 1963, The Brown Standard Corpus of American English was established at Brown University, in the United States. This corpus was created in the field of linguistics and served for linguistic description and analysis.

The first two large corps were created in the written American and British versions of the English language, and these corps have not lost their relevance even today, and still serves as a basic unit for many studies on the English language.

In the decade since these corpses were created, computers have become cheaper, more powerful computer classes, keyboard methods of typing text and scanner options have emerged. These abilities helped to develop and to result in billions of word-size types of corpus.

Although the first information about the corpus in world linguistics appeared in the 40s of the XX century [9], the aim, purpose, theoretical issues of the corpus linguistics, the principles of corpus formation were mentioned in the 60s of the XX century. Brown corpus (1961-1964) is the first source that gives information about the theoretical and practical foundations of corpus linguistics [4]. Since then, several Brown Corpus-type databases have been formed. In the 1970s, a frequency dictionary of the

Russian language was created based on a corpus and it contained 1 million words. In the 1980s, a corpus of texts in Russian was also created at the University of Uppsala, Sweden. Later, as a result of the development of computer lexicography, there was a need for a large text corpus. To be more concrete, 1 million words were not enough for an electronic dictionary database. On this basis, a large corpus of texts began creating. In many countries, such kind of corpus began forming in the 80s of the twentieth century. Different kinds of projects have been developed, among them were the Bank of English in Great Britain, as well as the British National Corps (BN), the Russian Machine Fund in Russia, and the Russian National Corpus .

“By 1990, more than 600 computer corpus had been registered”.

Studies have noted four main periods of corpus formation where corpus was created in the 1960s, 1970s, 1980s, and 2000s. It is based on the period of formation of the British and Russian corpus.

In the world of information, the corpus is emerging as a search system. In English, Russian, and several European languages among the search system, the corpus-based search was widespread.

The corps serves as an important resource for research in all spheres. The important role of corpus linguistics in communication, information exchange, research requires the researcher to have the skills to analyze, compare and evaluate information in the recommended literature, the ability to work with corpus linguistics software and information resources.

We tried to emphasize the practical significance of the corpus for linguistic research in the article.

The formation, development, and theoretical foundations of corpus linguistics are mentioned in the research. Corpus linguistics and it's subject which is explained and defined in the scientific literature, the language corpus is generalized as a set of special software-based texts in a certain period, different genres, different styles, regional and social variants**.**

The corpus of texts reflects the vocabulary of a particular language. “Text corpus is a collection of data corpus with units of text or integrity”. The vocabulary of the dictionary is not synchronous but also includes lexical units based on diachronic development. This allows developing general stages of the language by phonetic, lexical and grammatical features. At present, at the lexical level is far more difficult to understand and to study the semantics of historical, archaic, dialectal, argo, jargon. This is because lexemes of the given circle are not regularly used. While historical and archaic words are used in the classical sources, when dialectal words, argo and jargon are actively used in oral speech. Dictionaries are tools for mastering the meaning of lexical units with the help of a semantic framework. The problem is that the Uzbek language dictionary does not have a complete dictionary that covers the historical, archaic, dialectical units, argo, and jargon. Besides, the integration of language and literature in philological education is not formed, moreover, there is no attention to the skills of working with the dictionary. Formation of text corpus allows to know and to learn all layers of the dictionary level.

The corpus has a unique value as a database in the development of the national language. The corpus plays a special role in the coverage of linguistic issues, in the translation of lexical units, in the analysis of the semantic value of reality units, grammatical forms, and grammatical tools in the education system.

“There are several types of corps: single author's corpus, single book corpus, and a national corpus. The features of the National Corps are to develop language in a certain period, as well as regional and social variants, which encompasses all aspects. Below we consider the importance of corpora in linguistic research in terms of corpus types.

The role of the corpus of virtual texts in the arrangement of philological creativity. In recent years, the development of the Internet affected the emergence of a corpus of virtual texts. Internet search sites, electronic libraries, virtual encyclopedias serve as a corpus. The genre and thematic diversity of the corpus depend on the interests of the internet user. For example, in the scientific sphere “Wikipedia” is used as a corpus of large volumes of text. The corpus of virtual texts serves to develop creativity, increase it and improve imagination.

In corpus linguistics, the corpus of parallel texts is very important. The corpus of parallel texts is electronic versions of fiction, manuals, mass media, and various documents in two or more languages. With the help of a parallel corpus, it is conceivable to know the variants of a single word, sentence, paragraph, super syntactic integrity in different languages. A parallel corps is an important event for today’s intercultural dialogue. There is a possibility to identify the universals in different language environments, cultures and mental features of languages, realities and lacunar units through the parallel corpus. The parallel text corpus will also help the development of automatic translation and computer lexicography. With the help of parallel texts special concordance programs will be developed, moreover, it helps to create different special dictionaries.

“The corpus of parallel texts is being used for scientific and practical purposes (including to teach foreign languages). Source language and target language of the text is the structures of parallel texts. For example, the English text “Alice in Wonderland” and its translation in German, French, and Russian are the basis for the creation of parallel texts”.

Formation of parallel texts will enhance the prestige of the Uzbek language and strengthen its position. First of all, the parallel text corpus acts as a communicative database. Translation of one language

text into another language is aimed to identify intercultural relationships and highlighting differences. It provides an opportunity for semantic analysis of lexical units through translation options in other languages.

First of all, the corpus of parallel texts on work allows analyzing the similarities and differences in the grammar of two or more languages. The corpus of parallel texts is important in comparing the features of artistic style in both languages. In the artistic style, the means of image, movements, figurative expressions are described with the tradition of each language. For example, the attribution of “musk” (black fragrant substance) describes as eyebrows and hair, “shahd” (honey) describes lips or words, “tulip”, “ruby” to the lips, the sloping body to the “bow”, the usage of camels, caravans, horses, and dogs as artistic symbols reflects the style of depiction typical of Uzbek classical texts. The grammatical features of languages are expressed in the author's speech, the harmonious use of literary language and dialect in the speech of the characters serves to illuminate the figurative imagery.

The corpus-based on literary texts also includes units specific to certain languages, to be more concrete, to the units of reality. The corpus of parallel texts gives the possibility to define the principles of real units in translation. It will be possible to obtain information about the translations of reality units in such methods as calculus, semantic calculus, and equivalent word selection.

Understandably, there will be problems with phrases in the corpus of parallel texts which is based on works of art. Phrases in the Uzbek language consist of two or more words, which serve to form a new lexical meaning of the word based on the semantics of it. Therefore, if we use machine translation in the formation of the corpus of parallel texts, it causes problems in the correct illumination of semantics. To solve problems, phrases should be distinguished from simple and compound lexical units, word combinations and of course should be marked with special tags. The translation of phrases, of course, requires expert supervision.

The corpus of parallel texts allows comparing different cultures based on different languages, to master the content of lexical units representing cultural relations. Through the corpus of parallel texts, it is possible to compare and contrast the phonetic, lexical, morphological, syntactic features of languages. Such corpora are also important in that it can provide information about the categories specific to word groups, the expression of grammatical meaning, and the system of word-formation.

The manual and literature manuals are designed to provide the features of the scientific method of the corpus and give information about theoretical information in a particular field. The corpus of parallel texts created within the framework of mass media provides complete information about the type and content of mass publications. It helps to control the content of issues covered in the press. Placing an oral version of TV and radio texts (audio texts) in the corpus of parallel texts increases the illustration of the database.

Parallel texts in official, office documents are important in determining the style of official office, normative legal documents in different languages. This view of parallel texts serves as a source for research aimed at comparing the features of formal style in different languages, highlighting their universal and different aspects.

The corpus of parallel texts can be used as a linguistic database in language education, language teaching system. The role of parallel text corpus is very important in learning the content of works of art, analyzing conceptually, studying the basics of text linguistics and specific text features in different languages.

There are several aspects of the corpus and its feature fully covers the problems of a particular language.

The information corpus contains materials in the field of large-scale problems based on a particular selection method. The information corpus focuses on solving specific aspects of the problem part. The information corpus is structured based on speech norms and also takes into account the potential capabilities of language speakers. The corpus also can reflect historical forms of language. Therefore, the data corpus serves as the main source for linguistic research. The information corpus provides factual information in enlightening the properties of language units from sound to text. The corpus is used today as the most reliable source. Based on the information corpus, the linguist can draw certain conclusions about the development of language as a system in the field of activity.

The corpus of research is a particular type of case which is designed for a separate study. The function of language units which studies and devoted to different aspects called a corpus of research. This corpus is not built on the facts (post-factum) on which any research is based, but on the evidence that preceded in the research. This type of information is used in the sphere of linguistic issues. This type of corpus is important for providing a wide range of information on a variety of topics. Designed corpora served as a material for research in various fields.

Initially, the corpus of texts was created as statistical information reflecting a particular state of the language system. A typical example of this type of corpus is the author's corpus. Linguistic and non-linguistic issues require the identification of language phenomena on a time scale, such as changes in word meanings, the frequency of use of this or that syntactic construction. New technologies were made in the

corpus of dynamic texts and it developed the procedural aspect of the problem area.

References:

1. Navoiyning, A. (2012). *"Hayrat ul-abror" dostoni konkordansi*. Toshkent: TDShI.
2. Aripov, M., & Norov, A.M. (2019). *Razrabotka ontologicheskoy modeli sintaksicheskix pravil uzbekskogo yazyka. "Amaliy matematika va informatsion texnologiyalarning dolzarb muammolari".* Xalqaro anjuman tezislari to'plami. Of the international scientific conference "Actual problems of applied mathematics and information technologies". (pp.279-280). Toshkent, 14-15 noyabr.
3. Baranov, A.N. (2001). *Vvedenie v prikladnuyu lingvistiku*. (p.61). Moscow: Editorial URSS.
4. Frensis, N., & Kuchera, G. (1967). *Vychislitelnyy analiz sovremennogo amerikanskogo varianta angliyskogo yazika*. Moscow.
5. Hamroeva, Sh. M. (2018). *Ŭzbek tili mualliflik korpusini tuzishning lingvistik asoslari*: Filol. fan. f. d. (PhD) dis. avtoref, (p.9). Qarshi.
6. Karimov, S., Qarshiev, A., & Isroilova, G. (2007). *Abdulla Qahhor asarlari tilining lug'ati*. Alfavitli lug'at. Chastotali lug'at, Toshkent.
7. (2011). *Kompyuter lingvistikasi asoslari*. Toshkent: Akademnashr.
8. Kurbonova, F. (2014). *Kompyuter lug'atlari: tezaurus*. Toshkent.
9. (n.d.). *Kurs "Korpusnaya lingvistika"* (A.B. Kutuzov) Litsenziya Creative commons Attribution Share-Alike 3.0 Unported (Elektron resurs). lab314.brsu.by/kmp-lite/kmp-video/CL/CorporeLingva.pdf.
10. Kozlova, N.V. (2013). Lingvisticheskie korpusa: opredelenie osnovnyx ponyatiy i tipologiya. *Vestnik NGU. Seriya: Lingvistichekaya i mejkulturnaya kommunikatsiya*.T.11.Vypusk 1.