Journal of Critical Reviews

ISSN- 2394-5125 Vol 7, Issue 12, 2020

REPRESENTATION OF LINGUISTIC ISSUES IN CORPUS

[1]Rakhmanova Azizakhan Abdugafurovna, [2]Abdurakhmanova Mukaddas Tursunalievna, [3]Xolmanova Zulkhumor Turdievna

[1]Doctorate student, National University of Uzbekistan named after Mirzo Ulugbek, Tashkent, Uzbekistan.
[2]Candidate of Philological Sciences, Docent, National University of Uzbekistan named after Mirzo Ulugbek, Tashkent, Uzbekistan.
[3]Doctor of Philological Sciences, Professor, National University of Uzbekistan named after Mirzo Ulugbek, Tashkent, Uzbekistan.

Received: 02.03.2020 ***Revised: 04.04.2020*** ***Accepted: 06.05.2020***

Abstract
Corpus linguistics is a branch of applied linguistics that assumes a significant role in the communication of the information age. Corpora play a unique role in linguistic research as a database. 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.

Keywords: 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.

DOI: http://dx.doi.org/10.31838/jcr.07.12.20

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 [3,61].

MATERIALS AND METHODS

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 extralinguistic 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.

FORMATION OF CORPUS LINGUISTICS.

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 [21,16-17].

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 [23].

"By 1990, more than 600 computer corpus had been registered [21, 16-17]".

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 [5,9].

ANALYSIS AND RESULTS

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 [5,8].

The role of the corpus in linguistic research. 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" [3,62]. 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.

The significance of the corpus of parallel texts as a basis for comparative and relative research. 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 [3,62]”.

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 [3,63]. 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. In the existing literature, such corpora are also called monitor corpus.

The assembly of monitor corpora is characterized by a one-time and continuous collection of text. In pre-defined intermediate forms of time, updating of the text or the correction of the text corpus is observed [3,63]".

The peculiarity of the dynamic corpus is that the user who conducts research can easily distinguish the working corpus from the general corpus. "Birmingham corpus" of the English language was made as a dynamic corpus [3,62]".

The role of authorial corpora in the coverage of the creative style. The role of authorial corpora in linguistic research is very important. In Uzbek linguistics, research on computational linguistics and the formation of the basics of corpus linguistics helped to the creation of the author's corpora. The author's corpus serves as a source of information on the artistic style of the Uzbek literary language, as well as a source of insight into the author's ability to use words and language skills.

Authorship corpus is important in the study of linguopoetics, style, textual linguistics, artistic skill and examples of linguistic imagery. The author's vocabulary, the usage of this vocabulary, the ratio of own and mastered words in the works of the artist, the participation of dialects, archaisms, historians, barbarism, vulgarisms are reflected in the text of the author's corpus.

such kind of division of the corpus indicates its development. Because corpus is not just a database, but a set of information about the vocabulary, historical development, phonetic, lexical, grammatical features of a particular language.

Nowadays, computer technologies aimed at storing and processing texts are of great interest [3,62]. The corpus of texts is being formed as a widely used database on the Internet. Metalanguage, which serves to provide international communication in the world, further expands the usage of these languages. From this, it is clear that the formation of the national corpus is one of the main factors in the survival of national languages as a mother tongue, their widespread use and as a natural language.

Many of the world's languages have their national corporations of excellence, text processing, language corporations have become an indisputable tool of modern linguistics for linguistic research and practical tasks, a corpus which differs from an ordinary electronic library, corpus annotation, concordance (a relatively simple view of the search engine) or the corpus manager, its search capability, the general requirements for the corpus manager are discussed in it. The role of the corpus in the study of lexicography, lexicology, syntax, methodology, linguodidactics, mother tongue, foreign language education is analyzed [5,9]".

In research, the corpus is seen as a separate sphere from computer linguistics. We support the views that interpret it as a direction of computer linguistics which includes the creation of the corpus and the problems of the corpus, the purpose, and the tasks. While computer linguistics deals with the problems of natural language processing, solving language problems through computer systems, language teaching, text editing large-scale integrity reflects the language richness and a database for linguistic research. Therefore, it is appropriate to interpret the corpus as a direction of computer linguistics, and corpus linguistics as a separate field dealing with corpus types, corpus creation principles, computer methods, and problems. Areas of computer linguistics, automatic translation, automatic editing programs, information retrieval system, computer lexicography, information methods based on natural language processing, ie artificial language, language teaching programs are areas that improve corpus linguistics, which in turn is used as a source. Corpus linguistics is interrelated with computer linguistics, taking advantage of its achievements and developing it in turn [21,16-17].

The publication of the case with comments (posted on the Internet) is useful in conducting various experiments and research through the corpus. In terms of statistics, the corpus is seen not only as a branch of linguistics but also as the most reliable source of knowledge. So far the question of whether corpus linguistics is practical, general methodology, or whether it is an independent scientific discipline has not yet been answered clearly. The fact that many branches of linguistics, from theoretical linguistics to criminalistic linguistics, use an empirical method for analysis confirms the view that corpus linguistics is also a method. However, this was not accepted as a real and original object of corpus linguistics. The strict usage of it as an object of knowledge and information from the elements of language encourages the recognition of this field as an independent science and distinguishes it from other fields of linguistics by its uniqueness. Nowadays, Humboldt University in Berlin and the University of Birmingham have corpus linguistics departments. The fact that several prestigious journals (International Journal of Corpus Linguistics) are published specifically proves that corpus linguistics is gaining its place as an independent science.

It can be said that corpus linguistics views language as a social phenomenon that can be empirically verified based on original texts. Hence, it relies on precise calculations in the examination of linguistic knowledge and does not describe phenomena that cannot be determined by empirical knowledge or experience.

While the corpus covers the basic aspects of literary language and lively speech, according to the history of the words can be divided into three such as archaism, historian, neologism, etc.; it also allows us to distinguish the appearance of words according to their opportunity: dialectal words, terms, slang, and jargon, as well as features of speech style. Corpora is a source that reflects the vocabulary, as well as provides information about lexical units, lexemes, grammatical forms, and grammatical means.
The national corpus includes all types of corpus which is created in the language of this nation. As a result a database is created that reflects all its capabilities. Each corpus that forms the national corpus is important in terms of its characteristics.

One of the important aspects of global development are determined by the penetration of information and communication technologies in the social sphere, the formation of automated programs, the process of integration and innovation. One of the important tasks in the period of rapid exchange of information is to raise the status of the Uzbek language, to make it one of the most influential languages. Computer linguistics and corpus linguistics are good opportunity to fulfill this need. Computer linguistics and corpus linguistics play an important role in the development of the national language, its inclusion in the list of secular languages, language learning and teaching. The process of globalization requires rapid development in all areas. The computer system, which is a product of technical progress, has created conveniences in all areas, provides fast data transmission, translation, editing processes in a short time using a machine, the formation of artificial language, to be more concrete, information-computer style as a means of intercultural communication.

In the information age, along with all other areas, national linguistics is developing rapidly. The essence of the field of science, its place in society, its place in social life is determined by the nature of creativity, the driving forces in the field of production. The development of social spheres is giving an opportunity for the creation of new spheres.

Until today, ideas on language theory in national linguistics have been extensively analyzed. Theoretical issues of linguistics have been studied in various aspects. The deep analysis of scientific and theoretical issues has given rise to various views and interpretations. This situation affects that the scientific debate is still ongoing based on scientific research on certain issues. Scientific debates also entered manuals in schools, academic lyceums, and higher education. As a result, different approaches have emerged in the interpretation of theoretical issues. This situation began to negatively affect the study of the Uzbek language. Even without this, students who were far from using their mother tongue had difficulty learning grammatical rules. As a result, there were problems in learning both theoretical rules and practical use of language. To solve such kind of problems there was a need for practical linguistics.

The role of national corpora in the analysis of linguistic issues. The development of the field of applied linguistics serves to expand the opportunities of the national language and to create skills for its practical use. The development of the national language, the expansion of its capabilities, the increase in its functional level is determined by the level of development of computer capabilities. The first factor that provides the computer capabilities of the national language is the formation of the national language corpus. National corpora are developing as a determining factor in the development of national languages. Large-scale national corpora have been created in English and Russian which is considered as the world's languages.

The National Corpus of the Russian Language has been operating on the Internet since 2003 and currently contains various Russian texts with the usage of 149 million words. In the future, the national corpus of the Russian language is expected to consist of texts containing 200 million words [22].

The national corpus of the Russian language is valuable not only as a database in that language but also as a source that reflects the capabilities of that language. “Another function of the corpus is to demonstrate all the capabilities of the fields of lexis, grammar, accentology and language history. Previously, experts copied the necessary examples from the text by handwriting: this activity was labor-intensive and did not allow the processing of large volumes of material. Now the volume of studied material and the speed of information searching are not limited, and it allows the scientist or teacher to work with a range of different types of texts. The main users of the National Corpus are researcher-linguists in various fields. Corpus users are not limited. Reliable statistics on a particular period or style of writing are of interest to literary critics, historians, and other scientists who work in the humanitarian sphere [22]”.

The development of corpus linguistics in the Russian language demonstrates the development of computational linguistics in this language. In Russian linguistics, there has been researched on computer linguistics such as machine translation, automatic editing, computer lexicography. [24]. The National Corps plays an important role in determining the size of the national language, providing a literary language and vocabulary of certain areas.

While the National Corpus is the most needed helper for professionals, it makes research easier and it is the most reliable source of time savings. It can be widely used by philologists, lexicographers, programmers, editors, translators, journalists, teachers, students, and any other specialists, that means “….. corpora are created to provide a rich language resource to linguists, as well as to a variety of professionals [24]”. Currently, in world linguistics, effective work is being done to develop various national and language-related corpora of many languages. For example, the syntactic corpus, the corpus of poetic texts, the corpus of oral speech, the corpus of famous writers (for instance, the corpus of Alisher Navoi's collection of perfect works) and other various levels of research are being conducted on software development [5,43].

National corps is an important factor in learning the basics of the language in the information century. In the national corpus, the glossary of language reflects relatively. Especially, the national corpus provides a wide range of opportunities for teaching the national language as a mother tongue, for communicative, emotional-expressive, and accumulative functions of the language.

“National corpus is also important in teaching language as their mother tongue or as a foreign language. Many textbooks and syllabuses are currently served as a corpus-based. With the help of a corpus, it is possible to quickly and efficiently examine unknown words or grammatical forms of foreign authors, schoolchildren, teachers, journalists and writers [22]”.

There are several factors to the development of the National Corps. First of all, creating a national language corpus is a very difficult process. For this purpose, it is important to select all texts of the language on a computer system, to integrate them into certain basic concepts, and to adapt them to the requirements of the criteria of the searching system. Subsequently, the volume of periodic information of the corpus texts also varies and expands because of a long time. For example, “The National Corpus of the Russian Language includes the period from the beginning of the 19th to the beginning of the 21st century: in this period different kind of sociolinguistic versions such as literal, in conversation, simple conversational style and dialects are displayed in it. The corpus includes 2 things such as original works (not translated) which is the cultural value and the second one is fiction (prose, dramaturgy, poetry) which is interesting from the point of language. However, the National Corps is not just a corpus of literary language. The corpus also contains a large number of examples of written speech (including present-day oral speech), that are: memoirs, essays, journalism, popular and scientific literature, public performances, private correspondence, diaries, and documents [22].

The role of corpus linguistics methods in the analysis of language issues. The role of the corpus as a scientific material is analyzed in the researches covering the directions and content of corpus linguistics [10]. During the creation of Corpus Linguistics, some research methods were established. S. Wallis and G. Nelsons [11] distinguishes the following methods: the explanatory method, the generalization method, and the analytical method.

1. In the explanatory method, the texts are placed in a certain order. Explanations include structural and grammatical terms, syntactic analysis, and other instructions. The assignment of grammatical terms in this method ensures

the value of the corpus as a scientific source. Syntactic analysis is important in illuminating the grammatical structure of a language, distinguishing language units and speech units, and identifying speech phenomena.

2. In the generalization method, the terms of the system are mapped with their content and translations. This method focuses on specific terminology and includes linguistic searches, such as a system of rules that a student needs to learn.
3. The analytical method includes checking statistical data and periodicity in finding information [16]. While statistics serve for accuracy, the periodicity of information helps to form an idea of certain stages of language development.

The issues of computer systems and the role of computer science in statistical analysis in Uzbek linguistics were discussed [16]. The importance of thesaurus dictionaries as a database was analyzed [8]. Dictionaries have developed that serves as a database for the formation of the authorship corpus. A frequency dictionary of Abdullah Qahhor's works has been created [6].

Concordance dictionaries which are created in Uzbek linguistics are an important stage in the formation of the national [1]. In particular, the creation of concordance of classical sources, along with the improvement of computer lexicography, plays an important role in transmitting the spirituality, enlightenment and moral views of our ancestors to today's generation, the formation of the historical foundations of the Uzbek language, historical words and archaisms [19,4].

Many scientific articles on the formation of the Uzbek language information style have been published. A system for modeling the syntax of the Uzbek language has been developed [2,279-280]. The corpus is interpreted as the main database that makes up the discourse [17].

In Uzbek linguistics, research on computer linguistics, processing of natural language, the statistical analysis also focuses on corpus linguistics [14]. Areas of computer linguistics began to be studied as the object of monographic research [15,117]. Issues of corpus linguistics began to be studied in monographic terms in recent years [5]. Sh. Hamroeva researched “The linguistic basis of the Uzbek language in the author's corpus”. The study reveals the corpus in Uzbek linguistics, its peculiarities, theoretical foundations, the linguistic, practical and educational significance of the language corpus. Formation and development of corpus linguistics, peculiarities of the first and next-generation corpus, history of Russian and English linguistic corpus, the current state of corpus linguistics, features of modern Russian, English, Turkish, Tajik corpus, their commonalities, and differences are compared and corpus types are classified. General principles of building the case are studied and described, they are the technological process of the design and construction, the specifics of the case manager and its types.

The importance of computer analysis and modeling given in corpora in understanding grammatical analysis. In a text corpus, the data is semantically combined under certain words or concepts. Such words and expressions are represented by the term “storage unit”. Database Storage Unit. The data corpus is a unit of the problem area which is formed by certain principles, and the storage unit is defined by how the compositions are sorted. The storage unit is a concept defined by the corpus formation procedure, which represents one description in several languages. Discussing the base units of the corpus, U. Francis notes that, they may be separate words, short phrases, sentences and phrases (syntax). If the corpus is taken for syntactic analysis, whole texts or large fragments are taken as a basic unit. [3,62].

Storage units serve as the storage of data in the corpus of the text, their properties, and functions as an expression of the semantic description of the text. If the text serves as an advertisement, the storage unit will indicate which factory or business company it belongs to. The storage unit also includes information about the type of product to be advertised, to whom or to what is it belong to and the aggregate status of the product. Storage units allow text users to select products or objects that are intended for the user's purpose. The linguistic value of the text corpus increases because of the textual descriptions as a storage unit, what language is used, whether original or translated. Providing a source of text will ensure corpus reliability. To get a clear idea of the storage unit, let's look through the following example:

Linguistic corpora are a collection of a variety of styles of a particular language in a given period, and also a unique of texts of social content which is based on special software. The corpus consists of an array of texts, which is enriched with special additional information and serves as the basis for linguistic research. Language corpora are divided into simple and descriptive language corpora. Linguistically described or marked texts are descriptive language corpora.

In the 1980s, standard characters like SGML (Standard Generalized Markup Language) were introduced into electronic texts. Initially, this symbol was designed for the printing industry and later began to be widely used in all fields. SGML, which are language symbols, are considered to be language constructors. Because of its complexity in appearance and usage in its original state, HTML and XML were created in its database. Nowadays, all over the world uses this program in the internet sphere. Based on this program, it is possible to mark texts according to all parameters: underline, citations, identify words from another language, make a list and perform similar research tasks.

SGML is based on the idea of tags. Tags are a system of conditional symbols adopted to represent a specific piece of text. In the S + O + V model, S is a subject, O is a object, the signs which is accepted for word groups are: **N**=noun, **Ns**= singular form of nouns, **Nprop**= proper noun, **NP**=noun phrase, **Adj** or **A**= adjective, **Q**=question form of the words, **V**= verb, **Vt**=transitive form of the verb, **V**=intransitive form of the verb, **Vp**=predicative form of the verb (finite form), Vnp= the non-predicative form of the verb (non finite form), **V**=gerund, **Vcn**= past participle or passive verb, **VP**= verb conjunction, **ayx**=auxiliary verb, **mod**= modal verb, **Adv** or **D**= adverb, **Pron**=pronoun, **Art**=article, **Prep**= preposition, **Conj**=conjunction, **Part**= participle, **Interj**= exclamation, **Mim**= imitation word: parts of speech: **S**= subject, **P** or **V** = predicate, **O**= object, **M**=case phrase can be an example to the tags. Only the use of the symbols V and P can be as a model more than one concept which is prevented from being designated as tags: **V** = predicate; **V**= verb, **P** = predicate, **P**= preposition.

Corpus linguistics studies the issues such as the basic terminological apparatus which is related to corpus, system and a basic description of corpus, corpus typology, technology, the function of the corpus, types according to formal factors, history of the linguistic corpus, modern state, factors of the formation of corpus technology, corpus development stages, and corpus linguistics schools.

There is a theoretical and practical basis for the development of corpus linguistics in the era of global development. Besides, information technology and communication capabilities also play an important role in research in the field of corpus linguistics.

Thesauruses are one of the computer dictionaries that have a special place in the development of corpus linguistics, information retrieval system, semantic analysis, and have value as an electronic database.

Thesaurus (this word is taken from Latin and the meaning is “treasure”) is a computer dictionary based on a database of keywords, terms, basic concepts that reflect the main content of the text. Thesauruses, unlike encyclopedic and annotated dictionaries, are structured according to the frequency of use of language units in the text and the degree to which they reflect the subject content of the text. Thesaurus is a collection of data that provides a content-based information retrieval system. The terms are added into the thesaurus based on strict semantic principles, taking into account the hypo-heperonymic (species-gender), holo-meronic (whole-part), hierarchical (hierarchical) relations and associative semantic connections of units. It should be noted that in recent years, thesauruses in the database of search engines are also provided with hyperlinks, which creates some convenience for the user. That is, it also makes it easier to find information that is related to other close concepts of the terms are being searched [15,17].

It is beneficial to analyze the structure, principles of operation, capabilities as a computer database of the existing thesauruses. One of the most common thesaurus databases today is the WordNet system. “WordNet-type thesauruses which are developed by several countries have been successfully used in the processing of natural languages. The detailed construction and organization of lexical information in the thesaurus type are described in the research of V.N. Lukashevich” [12,61-97]. Given information of WordNet is available on the website of the United Nations WordNet Association [25]. This information serves as a program for the planned Uzbek language thesaurus.

Creating a thesaurus, the development of a thesaurus-based search system contributes to the development of corpus linguistics. Firstly, thesauruses are important as a large-scale, information-search-based database. This information has a special value in the development of the parallel text corpus and the research corpus.

In the corpus of parallel texts, words in a particular language are given in parallel with the translation of the texts into another language. It is known that in the process of translation there are logical, methodological contradictions in the translation of some words, combinations, and texts. The main reason for this is the semantic structure of lexemes and the lack of information about their sema. When a lexeme is translated from one language to another, in many cases it is not enough to simply express the lexical meaning. To make its correct translation, it is necessary to choose a logical, methodologically appropriate alternative which is based on the semantics of the lexemes. Information about the semantics of lexemes is expressed in thesauruses. Thesauruses reproduce synonyms, textual meanings, hypo-hyperonymic series, antonyms, paronyms, and contextual meanings of a lexeme.

“Development of the WordNet thesaurus of the Uzbek language requires a combination of traditional Uzbek lexicography and modern information technologies. The use of corpus technologies allows creating a resource that reflects the separated expression of Uzbek words and their lexical-semantic variants in a real contextual situation [13,286-287]”.

Special morphological features in the corpus system services for the analysis process. “Morphological analysis of algorithms. As a result of morphological analysis of the word, the elements of its morphological structure: core, base, affixes, suffixes are determined. The most commonly used algorithms for morphological analysis are stemming and lemmatization algorithms. Stemming is the process of finding the basis of a given word. The word base is a tangible part of the word and represents its lexical meaning. The base of the word may also be unpredictable with the morphological core of the word. A morpheme is the lexical meaning of a word. The problem of finding the basis of the word is one of the pre-existing problems of computer science. The purpose of stemming is to determine the compatibility of similar word forms in semantics”[18,290-291]. “Stemmers began to be created in the late 50s of the twentieth century. Stemmers are divided into two types such as algorithmic and lexical stemmers. The algorithmic stemmer works based on files consisting of a list of word-formative affixes and inflections. The lexical stemmer, on the other hand, works based on a dictionary of word bases, where the program begins morphological analysis with the first letter of the word. The base of the words in the text is compared with those in the dictionaries [18,290-291]”.

Lemmatization is also a process of determining the basis of a word, only in which the given word form is prearranged in advance to which word group it belongs. For example, the stemmer takes the garden as a basis for gardening, gardener and gardener’s. Lemmatayzer, on the other hand, defines the verb forms as the basis for gardening, the gardener in the noun category, the gardener’s lexeme as the basis for the word gardener’s. The concept of lemma represents a lexeme. The problem of lemmatization, on the other hand, is to identify the word forms that match to a lexeme. When analyzing the word my windows’ based on lemmatizer, the word base is formed as morphological information of the window, number suffixes is -s, possessive suffix –my and conjunction suffix: apostrophe ’ [18,290-291].

When it comes to factors influencing text classification, linguistic resources play an important role. One such resource is the linguistic (national) corpus of the Uzbek language. Designed corpora do not come across the basic requirements. We need to create WordNet-type linguistic resources here.

“Lexical analysis of algorithms. One of the fundamental algorithms of lexical analysis is lexical decomposition, which involves dividing a given text into tokens. The program that implements this algorithm is called a tokenizer. A “token” is used to represent lexical units. The tokenizer first divides the text which is based on the spaces between words, then removes the punctuation marks from the words [18,290-291]”

DISCUSSIONS

In addition to highlighting the importance of corpus in linguistic research, it should note the following statements:

1. The corpus of texts which is created in each language primarily serves to reflect the lexical richness of that language. It depicts the grammatical and methodological features. It is also important in the study of text linguistics.
2. The lexical-semantic possibilities of language are expressed in corpora. Thesauruses, which are an integral part of the corpus, are valuable in that they provide extensive information about the lexemes' private and transferrable meanings, communicative features, and emotional-expressive functions.
3. National corpora ensure that the national language has its place in the world communication system. It is the basis for active participation in the international information and communication system.
4. The corpus of parallel texts in translation allows for the full expression of the lacunar and real units between the languages.
5. The creation of a national language corpus will also ensure the development of all social spheres that use this language as a means of communication.

CONCLUSION

1. 1As a matter of first importance, corpora are a database for linguistic research. Research is being conducted to raise the status of language, extend its social capacities, characterize its place in communication, and improve it following the requirements of the information age. One of the urgent issues of worldwide improvement in the production of a typical database of national languages using current specialized capacities and based on this to determine the semantic capabilities of the language.
2. Firstly, the database which is created in each language serves to reflect the lexical richness, grammatical features, lexical-semantic possibilities of this language. National corpora guarantee that the national language has its place in the world communication system. Besides, it will be the basis for active participation in the international information and communication system. The creation of the national corpus of the Uzbek language will increase its prestige and status as the state language, allow Uzbeks in our country and abroad to learn the basics of the past and modern information on a worldwide scale.
3. The creation of a national language corpus will not only increase the capability of the national language but also ensure the development of all social spheres that use this language as a means of communication.
4. Research on corpus linguistics in national languages, specifically in Uzbek, is now being formed. It is important to expand the scope of research on corpus linguistics, to form a wide database which is based on semantic and syntactic notations, to examine the Uzbek lexicon and historical foundations such as historian, archaism, slang and social structures.
5. It is important to take into account the typological and genetic features of the national language in the development of computer methods, marking in the creation of the national corpus of the Uzbek language.

REFERENCES

1. Alisher Navoi. Concord of the epic "Hayrat ul-abror". - Tashkent, TDShI, 2012.
2. Aripov M., Norov A.M. Development of ontological models of syntactic rules of the Uzbek language. "Current problems of applied mathematics and information technology." Proceedings of the International Conference. Of the international scientific conference "Actual problems of applied mathematics and information technologies" .- Tashkent, November 14-15, 2019. V.279-280.
3. Baranov A.N. Introduction to applied linguistics. - M .: Editorial URSS, 2001.-S.61.
4. Francis N. Kuchera G. Computational analysis of the modern American English. - M., 1967
5. Hamroeva Sh. M. Linguistic bases of creation of the author's corpus of the Uzbek language: Philol. fan. f. d. (PhD) dis. avtoref. - Karshi, 2018. –p.9
6. Karimov S., Karshiev A., Isroilova G. Dictionary of the language of Abdulla Qahhor's works. Alphabetical dictionary. Frequency Dictionary. - Tashkent, 2007.
7. Fundamentals of computer linguistics.-Tashkent: Akademnashr, 2011
8. Kurbanova F. Computer dictionaries: Thesaurus. -Tashkent, 2014.
9. Course “Corpus Linguistics” (A.B. Kutuzov) License Creative commons Attribution Share-Alike 3.0 Unported (Electronic resource) - //lab314.brsu.by/kmp-lite/kmp-video/CL/CorporeLingva.pdf.
10. Kozlova N.V. Linguistic corpus: definition of basic concepts and typology. Vestnik NGU. Series: Linguistic and intercultural communication.2013.T.11.Vypusk 1.
11. Wallis, S. and Nelson G. Knowledge discovery in grammatically analyzed corpora'. Data Mining and Knowledge Discovery, 5: 307–340. 2001.
12. Lukashevich V.N. Thesauruses in the tasks of information retrieval.-M.: Publishing house Mosk. on.-that, 2011.61-97.
13. Matlatipov GR, Madatov H.A.Methodology of automation of some ontological types of WordNet for the Uzbek language through the creation of WordNet. "Current problems of applied mathematics and information technology." Proceedings of the International Conference. Of the international scientific conference “Actual problems of applied mathematics and information technologies”. -Tashkent, November 14-15, 2019. –V.286-287.
14. Polatov A. Computer linguistics. –T., 2011;
15. Rakhimov A. Basics of computer linguistics. –Tashkent: Akademnashr, 2011. –P.117
16. Rizayev S. Problems of linguostatistics in Uzbek linguistics. -Tashkent: Fan, 2006
17. Sadullaeva N.A., Abdurashitova E.T.Relevance of corpus to discourse. "Current problems of applied mathematics and information technology." Proceedings of the International Conference. Of the international scientific conference "Actual problems of applied mathematics and information technologies" .- Tashkent, November 14-15, 2019.
18. Tursunov M., Qarshiev A. Algorithms and programs of morphological and lexical analysis of Uzbek texts. Actual problems of applied mathematics and information technologies ”Proceedings of the International Conference. Of the international scientific conference “Actual problems of applied mathematics and information technologies” .-Tashkent, November 14-15, 2019.- V.290-291
19. Umarov E. Concordances of Alisher Navoi's "Khamsa" epics. –Tashkent, 2011.-p.4.
20. 20.Thye Global WordNet Association htpp: // global wordnet.org/ wordnets-in-thye –world.
21. 21.Zaxarov V.P., Bogdanova S.Yu. Corpus lingvistika.-Irkutsk: IGLU, 2011. –P.16-17.
22. 22.http://rusorpora.ru
23. http:www.corpus
24. http://www.wikipedia
25. 25.Thye Global WordNet Assoiation htpp://global wordnet.org/ wordnets-in-thye –world.