Creation and importance of language corps in Uzbekistan

Guli Toirova[1]*

[1]Bukhara State University, M. Iqbol, 11, 200100 Bukhara City, Uzbekistan

Abstract. The article discusses the transformation of language into the language of the Internet , computer technology, mathematical linguistics, its continuation and the formation and development of computer linguistics, in particular the question of modeling natural languages for artificial intelligence. The Uzbek National Corps plays an important role in enhancing the international status of the Uzbek language. The work carried out in the field of computer linguistics plays an important role in resolving existing problems in the Uzbek language. The question of the linguistic and extralinguistic separation of special tags for marking texts and their components is studied in particular.The coding requirements for important text information are defined. The state analyzes the linguistic module and the algorithm and its types from independent components of the linguistic program code. The need for algorithms for phonological, morphological and spelling rules for the formation of the lexical and grammatical code is scientifically substantiated. The importance of such linguistic modules as phonology, morphology and spelling in the formation of the linguistic base of the national corpus of the Uzbek language is emphasized. The article examines the corpus's primary purpose as a complex linguistic source, as well as the fact that it primarily contains two sorts of information and its types. The key effective capabilities of the corpus, according to the paper, are reducing time spent on the text analysis process and being able to explain the properties of language units in speech with thousands of instances. The national corpus, the educational corpus, and the parallel corpus are all discussed in the subject of computer linguistics. It was stressed that linguistic and extralinguistic tagging of them, the development of corpus formation algorithms, and the establishment of corpus linguistic support are all societal need. It recognizes the urgency of developing the basis for the creation of the Uzbek language corpus, conducting research in the field of computer linguistics as a scientific and theoretical source.

1 Introduction

Artificial intelligence has enabled a wide range of benefits in the use of language thanks to modern information technology. He is capable of doing a variety of things that the human intellect is capable of. Electronic sources, which are the result of artificial intelligence, are designed to keep humans safe and reduce their weight. Among the most pressing problems

* Corresponding author: r.a.quldoshev@buxdu.uz

are the conversion of the Uzbek language to the Internet and electronic language, as well as the enhancement of national language electronic resources (Uzbek language corpus, electronic dictionaries, and website contents).

2 Research Question(s)

We've previously mentioned that languages that have attained world linguistic civilisation have already done work on information processing using computer technology, machine translation, electronic lexicography, the establishment of thesauruses, and the creation of the language corpus. English, Russian, Arabic, French, German, Spanish, and Tajik are just a few of them. The scientific and theoretical aspects of creating a language corpus in the Internet system in these languages have also been established, emphasizing the necessity to speed up efforts to turn the Uzbek language into one that is "understood" by the Internet.

In world linguistics, the generation of language corpora on the Internet is the primary means of maintaining a particular language by the second decade of the twenty-first century, broadening the scope of its research, and demonstrating language skills. Computer technology, in particular, which is a great invention of the twentieth century, opens the door to a wide range of opportunities for linguistics as well as other fields, and imposes enormous tasks on computer language, the emergence of computer linguistics is crucial for the success of natural languages.

In global language studies, the study of linguistic modeling of language, the development of algorithms for word lemming and tags, as well as the electronic use of oral and written monuments, samples of spiritual heritage created in a specific language, in order to increase the use of national and cultural heritage. Particular emphasis is placed on information processing via computer technology, the development of necessary software and methodological software for the introduction of information resources, the development of the language corpus on the Internet, and, on this basis, scientific and theoretical aspects of the national language corpus.

A variety of studies on automatic translation, development of linguistic bases of the author's corpus, processing of lexicographic texts, and linguostatistical analysis have been conducted in Uzbek linguistics. Special emphasis was placed on "enhancing the education system and increasing the capacity of quality educational services." Given that raising the international status of the Uzbek language, elevating it to the level of a world language of communication, learning and teaching Uzbek abroad, expanding opportunities, and polishing our national language can all be accomplished directly through the national corpus, "theoretical and practical issues of Uzbek national corpus." Solution is relevant. In this sense, there is a need to further deepen research on the linguistic basis of the text corpus and the national corpus, the technology of creating its software.

3 Literature Review

The corpus is the subject of corpus linguistics. This term is variously defined in the scientific literature. For example, it is used in English with terms such as linguistic corpus or text corpus. Recognition of the scientific research of A.N. Khomsky, G.N.Luch, Ch.F.Meer, J.Sinkler, M.Z.Kurdi in solving such problems as creation of the national corpus of a certain language, its analytical technology, development of the field of corpus linguistics should (Mohamed Zakaria Kurdi, 2016; Toirova, 2020, p.57; Charlez, 2004, p. 7; Shomsku, 1962).

John Sinclair defines the term “corpus” as follows: “The corpus consists of a fragment of texts in electronic form selected according to visible criteria for the study of language or linguistic diversity, to be presented as a source of information” (Sinclair, 2004).

Large set of massive texts in Russian corpus linguistics, principles of corpus formation, linguistic database VG Britvin, VP Zakharov, IA Melchuk, AB Kutuzov, RG Kotov, LI Belyaeva , Reflected in the targeted research of E.V.Nedoshivina, V.V.Rykov, V.Plungyan (Britvin, 1983; Bloomfield, 1968; Belyaeva & Chizhakovsky, 1983; Zakharov, 2011; Nedoshivina, 2006; Rykov, 2005; Plungyan, 2005; Kutuzov, 2017; Kotov, 1977).

Russian scientist V.P. Zakharov explains the term "corpus" as follows: "corpus - a set of linguistic data units of language, compiled on the basis of oral and written texts" (Zakharov, 2011).

H.Iskhakova, S.Muhammedov, S.Riza on the linguistic-statistical analysis of the text in Uzbek linguistics, lexicographic processing, linguistic support of the automatic editing program, linguistic modules of the editing and analytical program, synonymous vocabulary of the national corpus, linguistic bases of the author's corpus. S.Muhammedova, B.Mengliev, D.Urinbaeva, A.Pulatov, U.Dysimova, G.Valieva, G.Jumanazarova, N.Abdurahmonova, Sh.Hamroeva, M.Abjalova, A.Eshmominov, O.Kholiyorov, R. Karimov's work is noteworthy. Our scientists, such as S.Karimov, S.Muhammedova, Sh.Hamroeva, conducted research on the specialty 10.00.01 of corpus linguistics.

Uzbek linguists define the term "corpus" as follows: Uzbek linguists interpret the term "corpus" as follows: "corpus is a set of linguistic units that make up a set of texts collected for a specific purpose" (Eshmuminov, 2019), a set of written or oral texts stored in electronic form in a language, placed in a computerized search engine” (Bongers, 1947). Research in Uzbek linguistics describes the essence of the corpus as follows: “The corpus is the ability to present existing information in the form of text; the ability to provide as much information as possible depending on the size of the case; it is an opportunity to use the data of a once-created corpus repeatedly to solve various problems” (Pulatov, 2011).

"A corpus is a set of texts that are subject to a search engine in order to determine the characteristics of language units, written or oral, stored in electronic form in a natural language, placed on a computer-based search engine software-based on-line or off-line system" (Mengliev, Bobojonov & Hamroeva, 2018) source.

O.Khaliyorov, who conducted research on the "educational corpus", in his work states the following: The educational corpus of the Uzbek language is a corpus designed to teach the possibilities of the Uzbek language, has a linguodidactical character, contains electronic texts, acts as a special site” (Hamroeva, 2018).

Regarding the parallel corpus, R. Karimov says: “parallel electronic analogue of translated texts; consists of several “original texts and one / several translations of them” (Karimov, 2021).

Language corpora can be divided into different forms in terms of structure, purpose, stability, variability. For example, V.P. Zakharov lists the following forms: “according to the form of data storage: audio, written, mixed; according to the language of the text: monolingual and multilingual; by genre: literary, dialectal, oral, journalistic, mixed; according to the access to the building: free, commercial buildings, closed; by purpose: research, illustrative; according to variability: dynamic and stable; marked and unmarked according to the possession of additional information (annotated)” (Zakharov, 2011).

V.V. Rykov, on the other hand, focuses on the following aspects in the classification of corpus types: “According to the level and structure of the data, according to the chronological sign (position) of the language, according to the language of use, according to the purpose of use” (Rykov, 2005).

In her prohibition, Sh. Hamroeva divides the corpus into the following types: “According to a certain period of language or a certain type of its occurrence (genre, style, a social or age group, the language of a writer or scientist); according to the type of linguistic mark; by type of speech: written, oral, mixed; they look like a multimodal corpus, a corpus of special texts” (Hamroeva, 2018).

"Specialized corpus: a group of texts of a specific type: newspaper text, scientific articles; common building; comparative corpus; parallel corpus; educational building; didactic corpus," says U.Kholiyorov (Kholiyorov, 2021).

4 Methodology

Each academic defined linguistic corpora from his or her own perspective and categorised them in various ways. What features of the Uzbek linguistic corpus are mirrored in it, and what corpuses are now being created?

The creation of the Uzbek language national corpus is a relatively new direction in both Uzbek linguistics and modern information technology. The language corpus is a major source and powerful information resource for compiling large-scale dictionaries. The language corpus allows for the rapid creation and processing of dictionaries using a computer. The importance of the corpus in the field of lexicography is that no tool can match the corpus in determining the period and frequency of use of a word. In the near future, the need for a dictionary today for a student learning a language or a researcher exploring any aspect of it will undoubtedly shift to the corpus.

5 Results and Discussion

Linguists at Tashkent State University of Uzbek Language and Literature named after Alisher Navoi are now working on a project dubbed "Educational Corpus" that is both scientific and practical. The creation of the Uzbek language educational corpus aims to gradually form data based on foreign experience, and includes an electronic textbook containing modern vocabulary of the Uzbek literary language, multilingual speakers, and non-translated lexical units of the Uzbek language, as well as a set of multimedia products, including audio and video materials, as well as a mobile application, aimed at the formation of correct pronunciation skills in Uzbek.

The educational system permits students to study Uzbek as a state language, a second language, and a foreign language in depth. Users may study the Uzbek language freely thanks to the educational building's electronic material, which includes audio, video, multimedia apps, pronunciation and spelling programs, and e-learning dictionaries. Unlike other curriculum, this complex focuses on developing the capacity to utilize the Uzbek language in unusual contexts. Beginners, students, parents, instructors, and students of the Uzbek language may all benefit from it. This will contribute to the formation of scientific and technological resources that will ensure the economic growth and social development of the republic.

As a result of the focus on scientific research in the field of Uzbek computer linguistics on the processing of the Uzbek language using modern information technologies, the first appearances of the national corpus are emerging as practical work.

Participation in the international scientific-practical conference "Theoretical and practical issues of creating Uzbek national and educational corpus" in May 2021, initiated by scientists of the Tashkent State University of Uzbek Language and Literature named after Alisher Navoi, with practical proposals not only domestic but also international did. At the end of the conference, specific tasks were identified to create excellent national and educational corpus, drawing on international experience.

Scientists of Samarkand State University are also working on a project called "Design and development of the national corpus of the Uzbek language" and scientists of Bukhara State University are conducting research on theoretical and practical issues of creating a national corpus of the Uzbek language. The results of the research are of great importance in raising the international status of the Uzbek language, raising it to the level of a world language of communication, learning and teaching the Uzbek language abroad, expanding the capabilities of our national language.

The creation of a national corpus of Uzbek language will make it possible to "digitize" the Uzbek language and turn it into an Internet language.

It opens the door to new opportunities to increase learning effectiveness. It is very easy to find a word, phrase or phrase that is rarely used through the corpus, or the problem with its use and spelling (spelling) is solved in a very short time. Today, more than a grammar scholar, the average researcher needs to know the status, level of application of a particular word, phrase, or construction, who used it when, when, and for what style. The corpus is focused on solving similar problems. The National Corpus is necessary to study the lexicon and grammar of the existing language. Another function of the corpus is to provide relevant information in the specified areas (lexicon, grammar, accentology, history of language). The National Corpus is a comprehensive universal information retrieval system that can be used not only by linguists, but also by all those who use the Uzbek language: experts in various fields, scientists, politicians, dictionary designers, researchers and others.

The formation and study of language corpus began some time before the development of the field of corpus linguistics. Examples include eighteenth-century biblical studies (e.g., Cruden), dictionaries (Johnson, Oxford English Dictionary, Webster Dictionary), language teaching (frequency corpus to Thorndike, 1921), and the Quirk Corpus (Survey of English Usage).

With the advent of computer technology, corpus linguistics began to develop rapidly. The Brown Corpus includes texts published in 1961 in the United States. Its capacity is from 2,000 to 500 plates per word. According to the special hierarchy of genres, 5 plates from the daily edition, 2 samples from the weekly editions, 4 plates of detective stories and 20 samples of novels were taken. The original version was presented as a plain text format with no characters47. In the selection of the texts, the authors Nelson Francis and Henry Kusera first developed the criteria for corpus formation:

- The origin and content of the text (the author, of course, the English version of the American version, not less than half the volume of the dialogue text);
- synchronization (it is known that it includes texts first published in 1961);
-Selection of individual texts on the basis of the presentation of different genres, the ratio of their numbers and special probability operations;
- Convenience of texts for computer analysis (inserting special characters to convey the originality of the text, etc) (Zakharov, 2011).

The full name of the Brown University is the Brown University Standard Corpus of Present-Day American English. The corpus consists of written versions of the American version of the English language, with a use of 1 million words. The Brown Corpus set the standard with 1 million words and became a benchmark for creating such corpus in other states. It later emerged that such a standard was unsatisfactory. Appropriate representation of the entity during the application of statistical methods requires only representative selection and large volumes of texts.

The Upsal Corpus (University of Uppsala, Sweden), based on Brown's principles, is also 1 million words in size, which is quite limited in terms of the number of genres it contains (Zakharov, 2011).

As the power of computers increased, it became possible to create relatively large and powerful enclosures. In the UK, the Bank of England Project (BANC) and the British National Corpus (BNC) emerged, which changed the corpus representation standard to 100 million words. It should include full texts, sample words that are common in oral speech patterns, and be easy to access via the Internet.

National corpus of many European languages (Spanish, Italian, Croatian) were created on the basis of British principles. In the Czech Republic, for example, the Czech National Corpus, which includes 100 million-word forms, has been open since 2000.

The representativeness of the corpus is very important. The corpus will not only have to study the variety of events being studied, but it will also have to correctly determine the place of this phenomenon in the lives of the speakers of that language.

The following criteria can be distinguished in the selection of the text for the corpus and the correct assessment of its representativeness:
- Text corpora, which seeks to fully reflect the diversity of the objective existence of speech;
-Cases designed for specific purposes of interest to the researcher (Zakharov, 2011)..

Corpus linguists select the representative corpus based on specific conditions. This is mainly due to the fact that the corpus consists of "10-20 million word usages." At the same time, they point out that a “much more alternative by genre” has been created to ensure corpus completeness. In particular, it should cover a wide range of artistic, dramatic, poetic and other texts.

The existing buildings in the global network are characterized by large volume (abundance of materials), as well as deep sockets. For example, Pushkin's and Chekhov's corpuses are morphologically and semantically annotated corpuses. One of the most perfect corpuses. Only modern software does not meet the design requirements. Because they were 8-10 years old. The next is the perfection of the search system of Shakespeare’s authorial corpus, distinguished by the brilliance of the whole design, but not morphologically and semantically marked. The Pushkin and Chekhov corpuses do not have syntactic markings. Many of the world’s recognized languages have their own national corpus, which differs in its level of excellence and ability to scientifically process the text. There are about 70 language corporations currently operating on the Internet, including English, Spanish, Chinese, Arabic, French, Russian, German, Polish, Polish-Ukrainian, Czech, Slovak, Serbian, Croatian, Bosnian, Bulgarian, Bulgarian-Russian, Macedonian, Scottish. , Netherlands, Dutch-French, Swedish, Dutch, Norwegian, Icelandic, Faroese, Medieval French, Spanish, Italian, Portuguese, Romanian, Lithuanian, Latvian, Greek, Eastern Armenian, Ossetian, Albanian, Indian, Hittite, Finnish, Uralic languages, Estonian, Veps, Hungarian, Udmurt, Georgian, Anglo-Georgian, Lezgin, Turkish, Tatar, Tajik, Bashkir, Crimean Tatar, Kalmyk, Buryat, Mongolian, Arabic, Hebrew, Amharic, Japanese, ancient Japanese, Baman, Esperanto corpora of languages. Each of the corpus named above has its own advantages and disadvantages. For example, the Tajik language corpus can only act as an electronic library and a photo and video gallery. It is not marked at all, the search engine is imperfect and inconvenient. He can only point to a whole work.

Existing corporations are used for purposes such as statistical analysis of language use, natural language processing (NLP) software, lexical resource creation, language teaching or learning. The texts presented in the corpus are important in the study of the dynamic state of language or in the analysis of the subject of various branches of linguistics. For example, computer analysis and database creation of linguistic resources is one of the tasks of corpus linguistics. It therefore serves as an important electronic resource for the creation of software such as data classification, data processing, machine translation, sentiment analysis.

“Uzbek computer linguistics is formed on the basis of features of the Uzbek language that are completely different from the English language. This shows that before the creation of Uzbek computer linguistics, there is a need to perfectly systematize and formalize the Uzbek language. Bringing rich, broad and deeply developed language issues, such as Uzbek, to the level of computer-based solutions requires a greater amount of work than English” said Pulatov. Agreeing with the scientist, it is possible to rely on his main ideas, although it is not possible to use English computer linguistics directly in the creation of Uzbek computer linguistics. In the preparation of the linguistic base and the bank of national texts for the formation of the language corpus of the Uzbek language, reference is made to the research work on the national corpus of the Russian language. The national corpus of the Uzbek language should function as both an electronic library and a linguistic corpus, have morphological, semantic, syntactic markings and meet the latest requirements of design, with the advantage of improved search system.

Below we present the plan for the formation of the Central Bank to create a national corpus of the Uzbek language:

MB in Microsoft Access There are two ways to create a table in MB MS Access MBBT. The simplest way to create an MB is to create all the necessary tables, forms, and reports using the MB Wizard. However, you can create a blank MB and then add tables, forms, reports, and other objects to it - this is the most convenient method, but it requires a separate defined object of the MB. In both cases there is a possibility to modify and expand the created MB. To create a new MB, choose Create, New Database, and then Create from the File menu.

MB stores millions of records, from which it is possible to find the necessary information at any time. The data in the MB tables should have simple tools in searching for the required information. Search and sorting is done in tabular mode and by special queries. A matching query is created, resulting in the required records.

The search for information is done through queries, and as a result of the query we have a new table that satisfies the given conditions.

In MB, information can be sorted, words can be arranged alphabetically or numbered. Sorting is done for ease of data search. Typically, the table is sorted by key field value. Sorting can be done on one or more fields. To do this, select the required fields and select the sort condition. Database modeling is done step by step (Fries, 1969).

The process of examination includes the collection of materials, their design in the form of technical specifications. They justify the expediency of creating a bank and a database. The following factors have been identified and cited as key factors:

- commonly used information;
- providing users with interactive data access;
- the existence of complex connections between data;
- the need to update the system.

Materials containing conclusions and proposals for the creation of a bank and database based on certain conditions and capabilities are included in the feasibility study of the project, as well as they form the basis for the formation of technical conditions for the development of the database system.

The system. It defines the goals and scope of problems to be solved, the scale and scope of the system, the global constraints.

At the technical design stage, development results and design decisions are formalized in the form of a technical project. It covers common issues: defining the configuration of computing tools, creating a logical database model, updating and configuring it in the form of other level models, selecting the operating system and database, and physical design. Then special programs are developed for the user database, the submodels available for each user are identified.

A technical design is a basic design document that provides development and descriptions for all components of a created database. Database modeling uses a variety of methods and tools to select a particular database. This includes the initial basic changes of data preparation and working with it, the identification of technological features for all processes associated with the creation and implementation of the database. The technical project reflects the organizational changes associated with the operation of the hardware and software with the organization of new information (Leech, 1991).

Technical design solutions are presented in more detail at the design stage and are described in detail. The working draft has a technically similar structure, but is clarified by in-depth study and verification. At this stage, the collection and pre-preparation of normative references, the development of official and technological guidelines for working with new information technologies will be carried out.

The purpose of this paper is to review the research in the field of linguistic database and to see the possibilities of using this technology in lexicographic projects. It is also possible to present a variant of such a project in the form of a lexicographic database that reflects the vocabulary of the Uzbek language with sound semantics. Database technology is used in the process of creating traditional and electronic dictionaries. Dictionary bases of special and terminological dictionaries are being actively developed.

5.1 What is the linguistic base?

The linguistic base is the basic morphological building blocks that prepare your text for further analysis and allow for effective search or processing in tokens, lemmas, parts of speech, and much more. The basic tools of the language enrich the original text in the native language for the best processing, speed and clarity of the natural language.

The world has developed smart, successful search semantics that work on the Internet. A person sends a request to a computer search engine and wants an answer. Does the computer understand human language? Is it possible to get answers to questions entered in multiple languages? Currently, open source platforms provide a basic system for reverse full-text search engines, but explicit search problems become more difficult because you add more language (the language you are looking for) to queries and results. Today, the internet offers tools that you can search in 40 languages. The words entered into the linguistic database perform a search in the requested language based on the search engine.

For example, Arabic linguistics has already incorporated Arabic text into a linguistic database designed to connect to major search engines and data search programs in order to facilitate the analysis of documents written in Arabic. However, it is not suitable for standard automated analysis methods that take into account Arabic writing, which is traditionally spoken. In such a linguistic base, Arabic words often contain grammatical elements that signify features such as the direction of the verb, object, person, number, gender, and so on. This system automatically performs the following actions: 1) forms the linguistic form of the word; 2) identifies parts of speech; 3) normalizes spelling norms, including the removal of vowels and nouns, the unification of hamza forms, and the permanent broken irregular "broken" plural forms; 4) can work in Persian (Persian and Dari), Pashto and Urdu.

5.2 Linguistic base: what are the problems and solutions of the problem?

The development of this technology in linguistics and the creation of similar sources solve the following problems:

1) The problem of the structure of the empirical material and its primary analysis allows to establish complete texts, starting from the fixation of units of language levels (grammars, dictionaries, phonetic databases). On the one hand, the completion and definition of the structural model of the language system, on the other hand, the creation of national models of discursive regions and a model of the general language system;

2) the task of finding new ways to install and store language information, as well as to organize access to these materials;

3) the task of finding new ways of processing the material to optimize research and obtain new results;

4) such as the task of reviewing learning outcomes by referring to large material (Nedoshivina**,** 2006; Sinclair, 2004).

5.2.1 Contribution to the science

In summary, the creation of a national corpus is done in two stages: identifying the list of sources and digitizing the texts (converting them to computer form). Its technological process consists of: creating a dictionary of repetition of lexemes and word forms on the basis of selected texts; review the text for any unit of the obtained repetition dictionary; to divide a graphic word into syllables and to compile a dictionary of repetitions of syllables; word resource sorting; processing unlimited files at the same time; create text corpora with external characters; calculation of statistical data for the corpus of created texts and separate texts included in the corpus; working with original texts in txt, doc i rtf format, automatic setting of encoding, etc. Considerations have been made that specific linguistic model forms should be developed for the marking of each word group. The text marking format, the coding requirements for important text information are studied and the existing body marking standards are taken into account. In view of the fact that there is currently no system for automatic text processing and searching on the basis of different characters from the text, it was noted that the layout is the main task of creating a corpus.

5.3 The national corpus of the Uzbek language

The national corpus of the Uzbek language is the hierarchy of lexical units, including synonyms, antonyms, homonyms, words that exist in the Uzbek language; must be able to automatically analyze the morphological structure of a word, the construction of a word, the meaning of words, its morphological features. That is, in the process of composing, lemmaging, marking the corpus, it is necessary to find such words that are part of the corpus in the texts on the basis of individual searches and interpret them specifically. For this, it is necessary to carry out algorithmic, linguistic modeling work.

The word “interface” is derived from English and is used to mean “appearance”. The word is often used in computer technology. A computer is the only communication system that provides a variety of information exchange between a person and a machine. An interface is two elements of a single system and a connecting link that works using that system. The interface is a system of communication between various nodes and complex hardware blocks, as well as technology and the user. It is expressed in the form of logical (information representation system) and formal (information properties). It is used to issue commands for specific tasks. Such an interface is called a user interface. The interface of any device is divided into external and internal views depending on the functions it performs. The user will not have direct access to the internal interface, it has a private option. With the external interface, the user can communicate directly and use it to control the device. These two types of interfaces always fit into a single device and ensure its operation, they cannot exist separately. The user interface can be divided into 2 parts. For example, this is the part that is responsible for entering data on the device and the user is responsible for its output. If we are talking about a simple work computer, then in the first category we have everything that works on a computer. Accordingly, everything belongs to the second category, through which the computer transmits information to the user in response to commands given by the same keyboard, mouse and other input devices, ie monitors, speakers, headsets, printers, plutors, etc. The interfaces used in computer technology come in the following types:

Visual. A standard computer interface that transmits data using visual images displayed on a monitor.

Gesture. As a rule, it serves as an interface for phones or tablets. In most cases, it is a touch panel that responds to the movements of the fingers of the person controlling the system and responds to a certain degree to each specific movement. It can be called a simplified version of the simple visual interface.

Sound. This type of interface has emerged relatively recently. Allows you to control the system using voice commands. The system, in turn, responds through user communication. Interestingly, modern technology allows us to control not only the sound of phones or computers, but also the sound of home appliances and even on-board computers.

The interface of the national corpus has a different design, structure, the perfection of which is entrusted to the author who created the corpus. Because the interface is an attractive overall look that makes a first impression on the body. Decorations that reflect the national colorite, as well as symbols that reflect the classic or modern, should be taken into account when creating the interface.

Hence, it is important that the interface is perfectly and systematically developed in order to demonstrate its convenient and most efficient capabilities in the use of the national corpus. Therefore, the interface is created in a user-friendly, easy-to-use form that meets the requirements of modern software design.

6 Conclusion

In short, Corpus linguistics is the most advanced branch of linguistics, and the corpus is a necessary tool for linguists; oral, written monuments are a source of information reflecting the national-cultural heritage. The corpus is a collection of texts subject to a search program, and a well-defined corpus serves as a stable linguistic base in ensuring the effectiveness of linguistic research. As a product of artificial intelligence, the linguistic corpus includes an electronic dictionary, a translation portal, a terminological database, a virtual (electronic) library, e-government, e-publishing, e-textbooks and manuals. The general view of the Uzbek national corpus is divided into several windows and right and left columns. It will have the following windows: "Lexical search", "Morphological search", "Syntactic search". Words and phrases from it are automatically analyzed in a matter of seconds. Linguistic and extralinguistic markings are created in a single format of data expression in the Uzbek national corpus, as well as in the world language corporations. Reconsideration of the theoretical foundations of morphological and syntactic markup based on academic grammar, practical work related to the reduction of the system of semantic markup tags will be carried out. The importance of the socket in the case is incomparable, because the width or narrowness of access to the case depends on the socket of the case. Perfect layout is a guarantee of a wide range of options, universal housing.

References

1. L. I. Belyaeva, V. A. Chizhakovsky, Thesaurus in automatic text processing systems., Chisinau (1983)
2. L. Bloomfield, Language. Moscow, "Progress" (1968)
3. H. Bongers, The history and principles of Vocabulary control, Woerden: WOCOPI (1947)
4. V. G. Britvin, Applied modeling of syntagmatic semantics of scientific and technical text (by the example of automatic indexing), Moscow State University (1983)
5. M. Charlez, English corpus linguistics: An introduction. Cambridge University Press, UK, 168 (2004)
6. A. Eshmuminov, Synonymous database of the Uzbek language national corpus. Dissertation of PhD in Philology, Tashkent (2019)
7. Ch. C. Fries, The structure of English. An introduction to the construction of English sentences, London (1969)

8. Sh. Hamroeva, Linguistic bases of creation of the author's corpus of the Uzbek language: Author’s Abstract of the Dissertation of PhD in Philology, Tashkent (2018)
9. R. Karimov, Linguistics and programming issues of creating a parallel corpus of Uzbek and English, Author’s Abstract of dissertation of PhD, Bukhoro, 151 (2021)
10. R. Kuldoshev, et al., *Mathematical statistical analysis of attainment levels of primary left handed students based on pearson's conformity criteria*, E3S Web of Conferences EDP Sciences, 371 (2023)
11. R. Kuldoshev, et al., *Mathematical statistical analysis of attainment levels of primary left handed students based on pearson's conformity criteria*, E3S Web of Conferences **371**, 05069 (2023)
12. R. G. Kotov, Linguistic aspects of automated control systems. Moscow, Nauka (1977)
13. A. B. Kutuzov, Corpus linguistics (2017) http://lab314.brsu.by/kmp-lite/kmp-video/CL/CorporeLingva.pdf (Last accessed 14.05.2023)
14. G. Leech, The State of Art in Corpus Linguistics, English Corpus Linguistics, London (1991)
15. I. A. Melchuk, Word order in the automatic synthesis of the Russian word (preliminary messages), Scientific and technical information, **12**, 12-36 (1985)
16. B. Mengliev, Is the Uzbek language corpus being created? Ma'rifat newspaper. April 3, (2018) marifat.uz/marifat/v_pomosh_uchitelu-marifat/savol/1142.htm. (Last accessed 14.05.2023)
17. B. Mengliev, S. Bobojonov, Sh. Hamroeva, Uzbek National Corpus. April 26, (2018) http://marifat.uz/marifat/ruknlar/fan/1241.html (Last accessed 14.05.2023)
18. M. Z. Kurdi, Natural Language Processing and Computational Linguistics: Speech, Morphology and Syntax, Great Britain, USA: Wiley-ISTE, 300 (2016)
19. E. V. Nedoshivina, Programs for working with text corpuses: an overview of the main corpus managers. Study guide, St. Petersburg, 26 (2006)
20. V. Plungyan, Why are we making the National Corpus of the Russian language? (2005) http://magazines.russ.ru/oz/2005/2/2005_2_20-pr.html. (Last accessed 15.07.2023)
21. A. Q. Pulatov, Computer Linguistics. Tashkent, Akademnashr, 520 (2011)
22. V. V. Rykov, A course of lectures on corpus linguistics (2005) http://rykov-cl.narod.ru/c.html
23. N. Shomsku, The logical basis for linguistic theory, Proceedings of the IX International Congress of Linguists (1962)
24. D. Sinclair, How to use corpuses in teaching a foreign language, Preface to the book, Studies in Corpus Linguistics, 12, **VIII**, 308 (2004) http://www/ruscorpora.ru/corpora-info.html (Last accessed 15.07.2023)
25. G. Toirova, The Role of Setting in Linguistic Modeling. International Multilingual Journal of Science and Technology, **4(9)**, 722-723 (2019) http://imjst.org/index.php/vol-4-issue-9-september-2019/ (Last accessed 15.07.2023)
26. G. Toirova, About the technological process of creating a national corpus. Foreign languages in Uzbekistan, 2(31):57-64, (2020) https://journal.fledu.uz/uz/2-31-2020.
27. G. Toirova, The importance of the interface in the creation of the corpus. Ínternauka, **7**, (2020) doi: /10.25313/2520-2057-2020-7-5944.
28. V. P. Zakharov, Corpus linguistics: a textbook for students of humanitarian universities, Irkutsk, 161 (2011)

29. G. Toirova, G. Astanova, N. Rahimova, Artistic Expressions of a Situational Pragmatic System, International Journal of Recent Technology and Engineering (IJRTE) ISSN: 2277-3878, **8(3)**, 4591-4593 (2019)
30. G. Toirova, M. Yuldasheva, l. Elibaeva, Importance of Interface in Creating Corpus. International Journal of Recent Technology and Engineering (IJRTE) ISSN: 2277-3878, **8(2S10)**, 352-355 (2019)
31. G. Toirova, N. Abdurahmonova, A. Ismoilov, *Applying Web Crawler Technologies for Compiling Parallel Corpora as one Stage of Natural Language Processing*. 2022 7th International Conference on Computer Science and Engineering (UBMK) Sep. 14 - 16, Diyarbakir /Turkey, 73–75. (2022)