AJMR

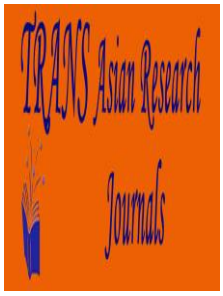
ISSN: 2278-4853 Vol 10, Issue 9, September, 2021 Impact Factor: SJIF 2021 = 7.699

Published by: TRANS Asian Research Journals

AJMR:

Asian Journal of Multidimensional Research

(A Double Blind Refereed & Peer Reviewed International Journal)



DOI: 10.5958/2278-4853.2021.00690.X

ESTABLISHMENT OF A NATIONAL CORPUS THE UZBEK LANGUAGE IS A REQUIREMENT OF A NEW ERA

Guli Toirova Ibragimovna*

*Associate Professor,
Doctor of Philosophy,
Bukhara State University, UZBEKISTAN
Email id: tugulijon@mail.ru

ABSTRACT

The article is scientifically substantiated by the need to create a national corpus of the Uzbek language. Suggestions are given on the structure of the corpus, the program interface, the algorithm of the program, the technology for obtaining the results. Based on the experience of world scientists, the requirements for the data encoding format for the national corpus of the Uzbek language are described. The article says that the interface of the national corpus and the author's corpus has a different design, structure, its improvement is the responsibility of the author of the corpus, and the interface should evoke the first impression of the corpus, an attractive appearance. The interface should take into account decorations that reflect the national color, as well as symbols reflecting the classics or modernity, the interface should reflect the life and work of the artist, the works created by him should open in separate windows, partly in photo galleries. The types of internal and external interface are also discussed. The article analyzes the linguistic module and algorithm and its types from independent components of linguistic programs. The need for an algorithm for phonological, morphological and spelling rules for the formation of the lexical and grammatical code is scientifically substantiated. The importance of language modules, such as phonology, morphology and spelling, in the formation of the linguistic base of the national corpus of the Uzbek language is emphasized.

KEYWORDS: *Corpus, Coding, Transformation And Graphic Analysis, Technological Process, Automatic Marking, Metadata, Text Information Of A Large Array, Morphological Marking, Spelling Module, Morphological Module, Linguistic Module, Phrase Modules, Word Algorithm, Formula Algorithm, Tabular Algorithm, Graphic Algorithm.*

INTRODUCTION

In world linguistics, by the second decade of the twenty-first century, the creation of language corpora on the Internet is the main means of maintaining a particular language, expanding its field of research and demonstrating language skills. In particular, computer technology, which is the great invention of the twentieth century, opens the door to a wide range of possibilities for linguistics as well as other fields and poses important challenges for the computer language, the emergence of computational linguistics is critical to the success of natural languages.

In global language studies, the study of linguistic language modeling, the development of algorithms for lemming words 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 attention is paid to the processing of information using computer technology, the development of the necessary methodological and software for the introduction of information resources, the development of the language corpus on the Internet and, based on this, the scientific and theoretical aspects of the national language.

In Uzbek linguistics, various studies have been carried out on automatic translation, the development of the linguistic foundations of the author's corpus, the processing of lexicographic texts and linguistic-statistical analysis. Particular emphasis was placed on "improving the education system and increasing the ability to provide quality educational services." Considering that raising the international status of the Uzbek language to the level of the world language of communication, studying and teaching the Uzbek language abroad, expanding the possibilities and polishing our national language can be achieved directly through the national corpus "Theoretical and Practical Issues of the Uzbek National Corpus". In this sense, it is necessary to further deepen the research of the linguistic foundations of the text corpus and the national corpus, the technology for creating its software.

Modern information technology has opened the door to a wide range of usability of language through artificial intelligence that it can perform many of the functions that the human mind can perform. It has been suggested that it was created to make people's time easier. The development of computer technology has led to the creation of electronic resources such as electronic dictionaries, translation portals, a terminology database, virtual (electronic) library, electronic text corpus, electronic government, electronic publications, electronic textbooks and manuals. Artificial intelligence consists of algorithms and software systems designed to perform a variety of tasks, and it can perform a range of tasks that the human mind can perform.

If we compare the human thinking system and artificial intelligence, we can conclude that human thinking has the advantages of being creative, flexible, capable of using emotional perception using a comprehensive, all-encompassing knowledge. It has the following disadvantages: complex conductive (expressive), unable to quickly document, unstable human thinking. An information retrieval system has the advantages of consistency, ease of presentation and uniformity, and easy documentation. It also has a number of disadvantages, such as the fact that it is artificial, limited, preprogrammed, of course, using symbolic perception and special knowledge. By analyzing the advantages and disadvantages of both systems, it has been proven that the main advantages of human thinking, including in many areas such as creativity, ingenuity, information transfer and content in general, are superior to artificial intelligence.

The linguistic corpus is not only the presentation of available information in the form of text, but also the analysis of the text, which, in terms of its analytical ability, is considered better than an electronic library.

Differences between the electronic library and the linguistic corpus: the search unit of the electronic library is the text of the entire work. You can search for a specific work in it. The source providing automatic search for information on the Internet is an electronic or virtual library. The electronic library has a different name. For example: virtual library, e-library, e-library, e-library. In such a library, books, magazines and newspapers will be located in the computer's memory, and not on the bookshelves. It comes in the form of a set of data stored digitally on a computer or a special device on a computer. Such data can include print, audio, video and multimedia data. There is no need for a special place to store books, since the site contains various information from the electronic library.This page is regularly visited, collected and filled in by specialists from a special center in libraries. The corpus search unit can be in the form of a language unit and a speech unit. Such a text can be used not only for reading, but also due to the presence of various grammatical interpretations of these texts, linguistic operations can be performed on them. It differs from the thesaurus in that the thesaurus searches for a concept, while the corpus looks for a word and its use. The corpus is important as a lexicographic unit containing various dictionaries (frequency, toponyms, grammatical words, phrases, etc.). Corpus is of great importance in modern lexicography. Therefore, it serves as a resource for compiling large vocabularies. Over time, the corpus becomes a large (extensive) information resource, becoming important for many linguistic directions. Corporate dictionaries are created and processed faster than ever before. The texts available in the operating system of the corpus have a sorting function. The researcher will be able to distinguish the example he needs from all the texts, and not only from the one that is important for the study. The electronic library does not have the listed functions.

Corpus is a set of texts in electronic form that find the meaning of words, phrases, grammatical forms through a specific search engine. There are different types of enclosures. For example, a corpus of authors, a corpus of books (including the first corpus of the Bible). The national corpus of a given language includes all aspects, genres, and methods, territorial and social variants of this language.

As a linguodidactician, the language corpus is equally important in the study of native and foreign languages. This opens the door to new opportunities for improving learning efficiency. It is very easy to find a word, phrase or phrase that is rarely used in the corpus, or the problem with their use and spelling (orthography) is solved in a very short time. It should be noted that the information in the linguistic corpus is not the same as described in the grammar or textbook, but the same as in the society. This is the most productive tool in the study of the folk and literary language. Today, not only grammar, the average researcher needs to know the status, level of application of a particular word, phrase or construction, who used it, when and for what style. The corpus is focused on solving similar problems.

The national corpus is required to study the vocabulary and grammar of an existing language. Another task of the corpus is to provide up-to-date information on the levels and areas of linguistics (lexicology, accentology, language history). The electronic corpus of the language is useful not only for linguists, but also for all people using the Uzbek language: specialists in various fields, scientists, politicians, lexicographers, researchers. It is a complex universal information retrieval system that can be used for various purposes [20].

The creation of a national corpus - a method of statistical research, computer translation, speech synthesis and recognition, the implementation of linguistic activities such as spell checking will help to realize the next stage in the development of corpus linguistics.

The criteria for creating corporations created in the world are: creation and filling of text, synchronization, presentation of different genres, sorting of individual texts by the ratio of numbers and special probabilistic operations, simplicity of computer analysis (placement of special characters to convey intertextuality).

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. It should be noted that L. Abyalova created linguistic modules for editing and analyzing natural language processing programs, studied the processes of graphic, morphological and syntactic analysis of texts [3]. The texts presented in the corpus are important in the study of the dynamic state of the language or in the analysis of the subject of various branches of linguistics.

The distribution of world corporations and corporations created over the years, the main periods of the creation of the corpus of texts, the corpus of the English and Russian languages, their various classifications are reflected in the research on corpus linguistics.

Research in the field of Uzbek linguistics and computational linguistics has provided information on some of the created world corpora, but the classifications are not fully covered in the study.

Theoretically, it was studied that the specific symbols and representativity that allow electronic search (at the morphological, syntactic level) are an important factor in the corpus (a complete reflection of the originality of many genres in the language). Suggestions are given on the structure of the corpus, the program interface, the algorithm of the program, the technology for obtaining the results.

On the technological process of building a body that provides the stages of the technological process VV Rykov, Yu.N. Marchuk, I. Melchuk, Sh. Khamroeva [7,8,9,18]:

1. The stage of preliminary processing of the text. At this stage, all texts from different sources are corrected and edited. The text is prepared for bibliographic and extralinguistic description.

a) the stage of transformation and graphical analysis. Most of the texts are considered initially. In particular, it removes elements (figures, tables) that are not needed for coding and automatic language analysis for a computer format, as well as underscores in the text.

b) the stage of automatic marking. This is done by automatically correcting the marking results, i.e. correcting and separating errors (manual or semi-automatic).

2. The stage of text marking. At this stage, the required corpus data (metadata) is entered. Meta-descriptions of corpus texts include: bibliographic information, symbols describing genre and stylistic features of the text, information about the author, and much more. This information is usually entered manually. Text components (paragraphs, sentences, word selection) and purely linguistic writing are often done automatically.

3. The stage of providing access to the case. The case display looks like this: it can be distributed on CD-ROM and is available in WAN mode. Different categories of users will have different rights and different capabilities.

4. The final stage is making changes (corpus manager) to the structure of a specialized linguistic information system that provides fast multi-parameter search and statistical processing of marked-up texts.

Of course, the composition and number of stages in each case may differ from those listed above, and the actual technology may be more complex.

The main requirements for the search engine of the National Corpus of the Uzbek language are as follows:

1) Search for words and phrases by their characteristics (grammatical, semantic, etc.);

2)Take into account the distance between the text (a whole passage of speech or work) and words;

3)Search for metatext information;

4)Extended language requirements, including boolean references, parentheses, and text operators;

5)The efficiency of indexing;

6)Quickly find the answer to the most difficult question;

7) Wide range, use of words up to the largest size (use of hundreds of millions of words).

Corpus data coding is based on the most authoritative standards. For example, TEI (Text Encoding Initiative), XCES (XML Corpus Encoding Standard), EAGLES (European Advisory Group on Language Engineering Standards). When presenting data in the National Corpus, the formatting of the text that carries linguistic information is based on the SGML / XML language.

There are two main types of textual information in the corpus:

A. Text information of a large array. Includes characters that fully represent the text: author name, gender, date of birth, text title, text creation time, word size, subject, text type, style, scope, etc.

V. Lexical information. Lexical information includes the following symbols: represents individual words, i.e. can use a word form in a specific place in the body of the text. This includes:

V.1. Morphological features:

• lexeme (word form);

• grammatical features of a lexeme (a group of words, living beings, passing events);

• grammatical features of the word form (number, contract, slope, time, person).

V.2. Semantic symbols:

Semantic classification, taxonomic class, mereology, assessment, causation, word-formation relations, etc. [1,10,11].

In the body, text is made up of a sequence of paragraphs, paragraphs are made up of sentences, and sentences are made up of words. In this case, the basic unit of analysis is the word, and the unit of text is the sentence. With the help of a search engine in the corpus, you can find words and phrases related to a specific character, related only to this sentence. The search result is a list

of sentences in which the found words are highlighted in a separate font. If necessary, the search text can be extended to the border of the paragraph, but no more.

Thus, it is possible to identify the main structural units in the body: word, sentence, paragraph, text. It does not use units that represent the structural division of the text (parts, chapters, sections), units that are outside the paragraph, and units that represent the syntactic structure of a sentence (sentences, groups).

"Uzbek computational linguistics is based on the features of the Uzbek language, which are completely different from English. This shows that before the creation of Uzbek computational linguistics, it was necessary to perfectly systematize and formalize the Uzbek language. To bring rich, extensive and deeply developed language issues, such as Uzbek, to the level of a computer solution, requires much more work than English, "A. Pulatov said [11].

Agreeing with the scientist, one can rely on his main ideas, although it is impossible to directly use English computational linguistics when creating Uzbek computational linguistics. When preparing the linguistic base and the bank of national texts for the creation of the linguistic corpus of the Uzbek language, a reference was made to the research work on the national corpus of the Russian language. In a study based on the observations of V.P. Zakharova [5], A.E. Polyakov [11], the process of preparing texts for the corpus is divided into the following parts:

1) the first layout of the text in minimal HTML format;

2) determination of morphological marks and homonymy (in a part of the body);

3) metatext markup;

4) Change the output format for the Yandex server.

The encoding of lexical information in the electronic body is adapted to the HTML / XML rules. This opens up a wide range of possibilities for fast processing of text in programs of various types, search index, morphological parser, converters, editing stages and automation of markup in the body. The texts for the National Corpus are imported from different sources and are presented in different formats such as plain text, HTML, RTF, PDF.

In the process of preparing the text, the following elements are removed from the text that do not belong to the author or are not important for learning the language: page numbers, column headings, title pages, table of contents, output data, systematic spelling, annotations, editor comments (comments written by the author are saved), drawings , diagrams, formulas (but captions are stored under them);

Linguistic and extralinguistic markings are the only data expression formats that facilitate the exchange of information in a corpus.

The technological process of the national corpus consists of: creating a dictionary of repetitions of lexemes and word forms based on the selected texts; view the text for any unit of the received dictionary of repetitions; divide a graphic word into syllables and compose a dictionary of repetitions of syllables; sorting word resources; simultaneous processing of an unlimited number of files; create text corpora with external symbols; the text being created is a corpus and the calculation of statistical data for individual texts included in the corpus.

Database - information, software that provides storage, updating, search and delivery of data [6]. It is an automated system that is a combination of equipment and personnel. The development of this technology and the creation of similar sources in linguistics solve the following tasks:

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

2) the task of finding new ways of installing and storing language information, as well as organizing access to these materials;

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

4) solves the problem of checking the research results, referring to a large amount of material.

Linguistic support is a set of language tools that ensure the adequate functioning of a language in a specific area. When creating an electronic corpus of the Uzbek language, its software is created only with impeccable linguistic support.

Today it is popular in the context of "virtual dictionaries, technologies for working with them and their creation." At the same time, you need to understand the industry on the Internet. In this sense, the term cyberlexicography refers to the theoretical basis for creating electronic dictionaries on the Internet - general and special types of academic, encyclopedic and linguistic dictionaries [23]. The creation of a corpus, which is a prime example of cyber dictionaries, has taken lexicography to a new level. It is known that the corpus is an electronic collection of written texts, created on the basis of the introduction of a set of dictionaries collected using specially developed computer programs [13, 14, 15]. The corpus covers all aspects of the language in the form of a computer program. Thus, corpus linguistics leads to a reevaluation of the language in terms of its characteristics.

The corpus, which is a collection of dictionaries built into an information retrieval system, is the main source of the cyberlexicographic corpus. Cyberlexicographic corporations are implemented through a specially designed computer program that displays words as needed. Most importantly, the corpus program examines the given target word, determines the number of samples in the corpus and calculates the binding frequency, displays examples specific to the target part, from which the user can continue further research. The creation of national cyber dictionaries of the Uzbek language is an urgent task that allows you to connect and feed the world virtual world with Uzbek lexicography, national Internet dictionaries as its product. The development of cyberlexicography requires the creation of a complete and complete database of cyber dictionaries in the Uzbek language, automatic editing, the creation of excellent programs that translate from Uzbek into another language or vice versa.

In computational linguistics, the term "linguistic module" plays an important role. For example, translation of a natural language into a computer language, that is, the creation of methods for processing text through a computer system. To do this, use extended translations of programs created in other languages. The linguistic module is an independent component of such programs. For example, if a lexical unit is surrounded by a vocabulary layer (words), the grammar unit edits symbols, punctuation marks, letters and other symbols, spelling rules of the

spelling unit, morphological unit of word analysis (parsing of a word-token) and synthesis (lexeme formation), a supersyntactic unit in syntactic module - the phenomenon of the relationship of sentences or words is analyzed. When creating the national corpus of the Uzbek language, its algorithm is based on the specifics of the language.

The national corpus of the Uzbek language should be able to automatically analyze the lexical units available in the Uzbek language, including synonyms, antonyms, homonyms, assimilation words, word ranking, morphological structure of a word, word formation, word meaning, its morphological features. That is, in the process of compiling, lemmaging, marking up the corpus, it is necessary, on the basis of individual searches, to find such words included in the corpus in the texts and interpret them specifically. To do this, it is necessary to perform the above linguistic modeling algorithm. M. Abdzhalova's research "Linguistic modules of the program for editing and analyzing texts in the Uzbek language" [4], A. Eshmuminov's research on lexical units "Synonymous base of words of the Uzbek national corpus" [16], automatic analysis of the morphological features of the words of Sh. parts of the study "Linguistic foundations of the author's corpus" [18], research by N. Abdurakhmanova "Linguistic support of the program for translating English texts into Uzbek" [2] on issues related to the translation of lexical units from the Uzbek language. "Dictionary of synonyms of the Uzbek language", "Explanatory dictionary of Uzbek words", "Dictionary of obsolete words of the Uzbek language", "Dictionary of synonyms of the Uzbek language", "Dictionary of words of the Uzbek language", which are available in Uzbek linguistics to designate lexical units. Linguistic support can be the "Dictionary of contradictory words of the Uzbek language", "Dictionary of classification of words of the Uzbek language", "Educational etymological dictionary of the Uzbek language", "Educational toponymic dictionary of the Uzbek language". Only such dictionaries should be revised, lemmas of words, to distinguish their number depending on the nature of words and to connect the members of a number of lemmas with each other.

Only then can the revised dictionary become the basis of the programmer's software. Linguistic modeling of marking is advisable, since in the linguistic model the morphological tag takes the form of a conditional abbreviation. To designate each group of words, special linguistic model forms have been developed. It is necessary to develop an algorithm for morphological marking of the language base. It is necessary to define ways of supplying the linguistic base with semantic markup. Linguistic labeling is of great importance in the creation of a national corpus and the formation of its linguistic base.

When creating the linguistic base of the "National Corpus of the Uzbek Language", it is important to create models of artificial words. In this case, using the linguistic modules proposed by M. Abdzhalova, we can offer the following model of word formation in 3 different forms by the affixation method:

Hence: B = stem, DW = derived word

1. DW = base + "whether"; DW = acos + "la"; DW = base + "size"; DW = base + "lik";

DW = base + qi; DW = base + "xon"; DW = base + "don" .....

2. In word formation with the affixation method, affixes are usually added after the base. Accordingly, artificial words formed by this method have the form "base + suffix" (for example, "taste + less", "oppress + red").

3. Artificial words also have the form "prefix + base". This phenomenon manifests itself mainly in word formation with the help of affixes borrowed from the Tajik language: dishonest, inconvenient, unworthy.

4. DW = be + base + lik; DW = none + base + lik; DW = ham + base + lik; DW = bad + base + lik .....

The suffix may be followed by a suffix and a prefix: be-sabr-lik, be-parvo-lik, be-saranjom-lik, be-sarishta-lik, ham-yaral-lik, ham-nafas-lik, no-insof-lik, no-mard-lik, no-makul-chilik, no-mahram-lik, no-anik-lik, bad-bakht-lik.

We can say that the use of the modeling method in the field of linguistics, especially in the process of grammatical analysis, at first glance seems to be a movement from simplicity to complexity, but serves to increase the level of comprehensibility and accuracy of the studied subject.

The corpus of managers (the corpus browser or corpus query system) is a tool for multilingual corpus analysis and is usually a complex system used to find the linguistic forms and sequences of the corpus of managers. It can provide information about the text or information provided by the caller in terms of the location of the data property (such as lemma and tag). This is called "concordance". Other actions include co-location searches, frequency statistics, and metadata that are processed in the text. A relatively short description of the corpus of managers refers to the server or query engine of the corpus. In this case, the client-side aspects are called user interfaces. The manager's body can be presented as a program on a personal computer or as a web service [24].

An integral part of the "text corpus" concept is a text or linguistic data management system. Recently, he is more commonly referred to as the corps manager. Corpus Manager is a specialized search engine. It includes proprietary data retrieval software, statistics collection and user-friendly delivery of results.

For example: search in the Russian National Corpus is carried out on the basis of Yandex.Server Professional. Yandex.Server, on the other hand, searches for hidden functions and separate tabular information in grammar and metatext. Search data is generated through the Yandex.Server. Provides full-text information search, taking into account the morphological features of the Russian language, on a web server in the corporate network. The search is carried out taking into account the morphology of the Russian, English, Ukrainian languages. He also works for Yandex on the Internet. If you enter the word "go", you will see documents containing the words "go", "going", "shell", "walked". The search result will be documents sorted by relevance. They take into account not only the number of documents, but also the contrast of words, the frequency of their use and the distance between words.

Queries are analyzed in terms of their subject matter and formal content, and interpreted in a glossary of scientific terms that work with the corpus. The search consists of comparing the individual elements of the corpus in order and determining their compatibility. In this case, the corpus texts are considered relevant and are recommended for sending [25].

The query language model of the Uzbek National Corpus usually includes the following elements:

1) Direct search elements (terms and information requests);

2) Means of morphological standardization of text query elements;

3) Operators (conjunction, disjunction, negation);

4) Linear grammar tools (distance and position operators);

5) Additional search terms:

-Search in designated places of the body (for example, in tags);

- Limiting the search area (for the works of some authors, some documents and their types);

6) Qualification (rating) requirements for the results obtained;

7) Requirements for the form and type of results [22].

At the first stage of development, it is important to choose the data backup and database management system (DBMS) required for the search engine. The ability to use a DBMS and data backup allows you to quickly and reliably access large volumes in real time and must meet the following criteria:

-Responsiveness (1 request per second, including the speed of the database, including a table with 100 million rows);

-Scalability (application of requirements for system functionality in accordance with processes distributed on several machines);

- The cost of the corpus (the analysis includes free commercialization and data storage);

- Interaction with software (support for the ability to work with systems such as PHP and Unix);

-Availability of documents (full availability of documents in Russian, English and Tatar languages);

- Development prospects (dynamics of project development, user community, developers' plans);

The database and system architecture is designed to answer the following types of questions:

- For direct search by word form or lemma;

-Re-examine the morphological features of the phenomena presented in the form of conjunctions, disjunctions, forms of negation, such as and, or, yo;

- For the type of hybrid search for word forms and morphological features in the lemma.

Using the architecture created for the Manager Corps allows solving many problems. In the future, this architecture can be easily applied to integrate linguistic data analysis, including morphological analyzer, multivalued morphological module solution, and various other services.

This approach to solving the problem of operating systems for the linguistic corpus. This specially developed system can be used not only for the operation of the electronic corpus of the Tatar language, but also when making changes to the Uzbek corpus [17].

In the second chapter of the volume, entitled "Syntactic analysis of texts in the corpus (syntactic analysis)", syntactic analysis, its functions, morphological analyzer, etc. are theoretically analyzed.

Parsing is the computer definition of parsing. For this, a mathematical model is created to compare tokens described by one of the programming languages with the official grammar. For

example PHP, Perl, Ruby, Python. When a person reads, from the point of view of philological science, he syntactically parses the words (tokens) that he sees on paper, comparing them in his dictionary (official grammar). A machine-readable "script" is a program (script) that allows you to compare suggested words with words on the Internet. The scope of application of such programs is very wide, but they all work according to practically the same algorithm. The algorithm for working with parsing is as follows: no matter what official programming language it is written in, the algorithm for processing it remains unchanged. Internet access, access to the code of the web resource (access) and its download; reading, receiving and processing data; present the received data in processed form - files in .txt, .sql, .xml, .html and other formats.

Parsing solves the following tasks:

1. Lexical analysis is the division of the text into sentences and phrases.

2. Morphological analysis of words (tokenization and lemmatization) - to determine the part of speech, conjugation, type (homeland) and other grammatical features of words, taking into account the text (meaning priority).

3. Syntactic syntactic analysis - to determine the relationship of words in sentences, to search for possessive and participle, to divide sentences into groups according to the possessive, complementary and case types.

4. Simplified parsing (chunking) - division of complex text into subclauses.

They are performed for all of the above, including words that are not in the dictionary. It will also be possible to connect the spelling correction mode of the morphological analyzer.

Parsing is designed to quickly parse very large amounts of text (tens of kilobytes or hundreds of megabytes). To achieve high parsing efficiency, the entire dictionary is entered into its RAM at startup. The analysis does not limit the size of the analyzed text or the time it takes to use it.

Tokens are morphological analyzer objects whose function is to perform stemming, lematization, and morphological analysis. The Uzbek language is agglutinative in its structure, which affects the algorithm of the morphological analyzer. The analysis was based on a dictionary approach. It identifies each stem of a word, and determines which paradigm a particular word belongs to. The grammar dictionary is based on the "Explanatory Dictionary of the Uzbek Language", which is the only lexicographic academic research.

Hence, parsing is the process of establishing syntactic links and attaching certain syntactic symbols to words or phrases.

The word "interface" comes from the English language and means "appearance". This word is often used in computer technology. The computer is the only communication system that provides a variety of information exchange between man and machine. An interface is two elements of a single system and a link that works with this system. An interface is a communication system between various nodes and complex hardware units, as well as technology and the user. It is expressed in the form of a logical (information presentation 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, he has a private option. With the help of the external interface, the user can communicate directly and use it to control the device. These two types of interfaces always fit

into one device and ensure its operation, they cannot exist separately. The user interface can be divided into 2 parts [21]. For example, this is the part that is responsible for entering data on the device, and the user is responsible for outputting them. If we are talking about a simple working 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 issued by the same keyboard, mouse and other input devices, that is, monitors, speakers, headsets, printers, plutors, etc. technique are of the following types:

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

Gestures: Typically, it serves as an interface for phones or tablets. In most cases, this is a touch panel that responds to the movements of the fingers of the person operating the system and reacts to a certain extent to each specific movement. It can be called a simplified version of a simple visual interface.

Sound: This type of interface has appeared relatively recently. Allows you to control the system using voice commands. The system, in turn, responds through user interaction. Interestingly, modern technologies make it possible to control not only the sound of phones or computers, but also the sound of household appliances and even on-board computers.

One of the newest trends in this area is the touch interface. The principle of its operation is based on the physical interaction of the user and the machine, which is carried out through certain objects.

The interface of the national corpus differs in a different design, structure, the revision of which is entrusted to the author who created the corpus. Because the interface has an attractive overall look that makes the first impression on the body. The interface should take into account decorations reflecting the national flavor, as well as symbols reflecting the classics or modernity.

Therefore, it is important that the interface is ideally and systematically designed to showcase the user-friendly and most effective experience when using the national corpus. Therefore, the interface should be created in a user-friendly format that meets the requirements of modern software.

CONCLUSION

1. Corpus linguistics is the most advanced branch of linguistics, and corpus is a necessary tool for linguists; oral, written monuments are a source of information reflecting the national and cultural heritage. A corpus is a set of texts to be searched for, and a well-defined corpus serves as a stable linguistic base to ensure the effectiveness of linguistic research. As an artificial intelligence product, the linguistic corpus includes an electronic dictionary, translation portal, terminology database, virtual (electronic) library, electronic government, electronic publications, electronic textbooks and manuals. Linguistic electronic sources, which are a product of artificial intelligence, are considered raw materials for creating a certain linguistic corpus.

2. The creation of the National Corpus is carried out in two stages: determination of the list of sources and digitization of texts (transformation into computer form). Its technological process consists of: creating a dictionary of repetitions of lexemes and word forms based on selected texts; view the text for any unit of the received dictionary of repetitions; divide a graphic word into syllables and compose a dictionary of repetitions of syllables; sorting word resources; simultaneous processing of an unlimited number of files; create text corpora with external

symbols; calculation of statistical data for the corpus of created texts and individual texts included in the corpus; work with source texts in txt, doc i rtf format, automatic encoding setting, etc.

3. The most effective standards for encoding corpus data have been selected. The presentation of its data is based on an SGML / XML text layout. When encoding, lexical information is adapted to HTML / XML rules. The texts selected for the National Corpus are taken from various sources and presented in different formats: plain text, HTML, RTF, PDF.

REFERENCES:

1. Abroskin A. A. Corpus search: problems and methods of their solution // National corpus of the Russian language: 2006–2008. New results and prospects. SPb .: Nestor-History, 2009. – 277–282 p.

2. Abdurakhmonova N.Z. Inglizcha matnlarni ўzbek tiliga tarzhima қilish dastouring linguistic taminoti (Sodda haplar misolid). Filol.fan.byich falsafa doctor (PhD) ... dis. author. - Tashkent, 2018.

3. Abzhalova M. Taxrir va taxlil dasturlaring linguistic modulari: Monograph - T., 2020. - 176 b.

4. Abzhalova M. Ўzbek tilidagi matnlarni taxrir va taxlil қiluvchi dasturning linguistic modulari (Rasmiy va ilmiy uslubdagi matnlar taxriri dasturi uchun). Philol.fan.byich falsafa doctorate (PhD) ... dis. author. –Farғona, 2019.

5. Zakharov V.P. Corpus linguistics. Study guide. - St. Petersburg, 2005 .-- 48 p.

6. DateK. J. Introduction to database systems. 8th edition .: Per. from English - M .: Publishing house "Williams", 2005. –1328s.

7. 7. Rykov V.V. A course of lectures on corpus linguistics. URL: http://rykov-cl.narod.ru/c.html.

8. Marchuk Yu.N. Fundamentals of Computational Linguistics. - M .: Publishing house of MPU, 2000.

9. Melchuk I.A. Word order in the automatic synthesis of the Russian word (preliminary messages) / Scientific and technical information. 1985, No. 12. - P.12-36.

10. Kustova GI, Lyashevskaya ON, Paducheva EV, Rakhilina EV Semantic markup of vocabulary in the National corpus of the Russian language: principles, problems, prospects // National corpus of the Russian language: 2003-2005. Results and prospects. –M., 2005.– pp. 155–174.

11. Polyakov AE Technology of information preparation in the National Corpus of the Russian language Text. / A.E. Polyakov // National Corpus of the Russian Language: 2003-2005. Results and prospects. - M., 2005. -S. 192.

12. Platov A. Қ. Computer linguistics / A.Қ.Platov; masul muharrir: A.A. Abduazizov, M.M. Oripov. - T .: Akademnashr, 2011 .-- 520 b. (-B. 7.)

13. ChGU named after I. N. Ulyanov. [Electronic resource]. http://www.myshared.ru/slide/10492/

14. Sivakova N.A. Lexicographic description of English and Russian phytoonyms in the electronic glossary. diss. . Dr. filol. sciences in the form of scientific. report Tyumen., 2004. - 72 p ..;

15. Sazhenin, II Vocabulary corpus: problems of definition and structural organization / II Sazhenin; otv. ed. I.P. Matkhanov. // Problems of interpretive linguistics: types of perception and their linguistic embodiment: interuniversity collection of scientific papers. - Novosibirsk: Publishing house of NGPU, 2013. - P. 294 - 298

16. Eshmuminov A. Ozbek tili milliy korpusining synonym sozlar bazasi. Filol.fan.byich falsafa doctor (PhD) ... dis. author. - arshi, 2019.

17. Khakimov B.E., Gilmullin R.A., Gataullin R.R. Resolution of grammatical polysemy in the corpus of the Tatar language // Scientific notes of Kazan University: Humanities. 2014.Vol. 156. No. 5. P. 236–244.

18. Hamroeva Sh. Ozbek tili mualliflik korpusini tuzishing linguistic asoslari. Filol.fan.byicha falsafa doctor (PhD) ... dissertation - arshi, 2018. -B.45.

19. Toirova G. About the technological process of creating a national corps. // Foreign languages in Uzbekistan. Electronic scientific-methodical journal. - Tashkent. 2020, № 2 (31), –B.57–64. https://journal.fledu.uz/uz/millij-korpus-yaratishini-tehnologiya-zharayoni-hususida/

20. Toirova G. The importance of linguistic module forms in the national corpus // Current problems of modern science, education and upbringing (Current problems of modern science, education and upbringing in the region) (Electronic scientific journal), - Urgench. 2020, № 5, –B.155-166. http://khorezmscience.uz/public/archive/2020_5.pdf

21. Toirova G. The importance of the interface in the creation of the corpus. Índternatíonal Scíentífíc Journal «Índternauka», // Mejdunarodnyy nauchnyy zhurnal «Internauka». - 2020. - №7. Online magazine. https://doi.org/10.25313/2520-2057-2020-7-5944

22. Toirova G. The Role Of Setting In Linguistic Modeling. // International Multilingual Journal of Science and Technology. ISSN: 2528-9810 Vol. 4 Issue 9, September - 2019, -P.722-723 http://imjst.org/index.php/vol-4-issue-9-september-2019/

23. http://://studfile.net/preview/1619320/page:3/Кибернетическая лексикография Захаров В. 1 А -10 ФИЯ

24. Suleymanov D., Nevzorova O., Gatiatullin A., Gilmullin R., Khakimov B. National corpus of the Tatar language "Tugan Tel": grammatical annotation and implementation. Procedia-Social and Behavioral Sciences, 2013, vol. 95, pp. 68–74. DOI: 10.1016/j. sbspro.2013.10.623.

25. Kilgarriff A., Baisa V., Bušta J., Jakubíček M., Kovář V., Michelfeit J., Suchomel V. The Sketch Engine: ten years on. Lexicography, 2014, no. 1, pp. 7–36. DOI: 10.1093/ijl/ecw029.