Using As A Corpus Search System

Urokova Dilnoza Odil qizi

2nd-year graduate student at the National University of Uzbekistan named after Mirzo Ulugbek
adilovnadilnoza09@gmail.com

ABSTRACT

This article covers the concept of a corpus, its role in language, types of corpus searches, the principles of operation, and various approaches to searching and studying it.

Keywords: corpus, context, dictionary, intertextuality, latent models, language comparison, metadata, text annotation, and linguistic annotation

In world linguistics, the necessity of creating corpora in languages to solve issues such as broader exploration of linguistic possibilities, identifying problematic aspects of language grammar in context, defining grammatical patterns in language, facilitating the creation of multi-domain electronic dictionaries, improving the efficiency of using modern information technologies in language learning, implementing automatic translation, search, and computer analysis in language, and preparing electronic textbooks and dictionaries, determines the relevance of our article. The development of theoretical and practical foundations for creating corpora in languages and the need to build specialized corpora for specific fields of language is highlighted. A corpus is not only a large collection of texts but also the foundation for written or spoken materials based on linguistic analysis. In a corpus, the following types of texts can be used:

- Texts by specific writers or authors;
- Texts related to a specific decade or century;
- Contemporary texts on specific topics;
- Contemporary texts that are sufficiently prevalent in the language or society.

Based on the types of texts defined above, the following can be searched when forming a corpus:

- All forms of a phrase in context;
- Changes and consistency in the dictionary;
- Words most commonly associated with certain phrases;
- Major differences between two texts;
- Specific peculiarities in the use of words and phrases by a particular author;
- Intertextuality: the meaning of a word as the sum of its usage occurrences;
- Latent models using word units;
- Language comparison.

The search system of the corpus and its operating principle involve three main components that assist in checking the database within the corpus. These are: metadata, text annotation, and linguistic annotation (explanation). Metadata provides some information about the author of the text,

when it was published, and in which language it was written. Metadata can be encoded in the corpus text or stored separately as a document in the database. Text annotation or marking is used to represent text formatting, for instance, marking the beginning and end of a sentence. Metadata identifies the speakers in the text and provides useful information about their age and gender. Text annotation is then used to show when each speaker starts and finishes speaking. The combination of metadata and text annotation together answers a range of research questions. Linguistic information can be encoded within the text corpus, which will later be analyzed systematically and precisely. In this case, the corpus is explained from an analytical or linguistic point of view. Annotation uses coding rules such as text annotation, for example, XML (extensible markup language) tags, where a phrase starting point (<np>) and endpoint (</np>) are used. For example, np The student </np> <np> went to the board </np> (<np> The children </np> sat in <np> the room </np>).

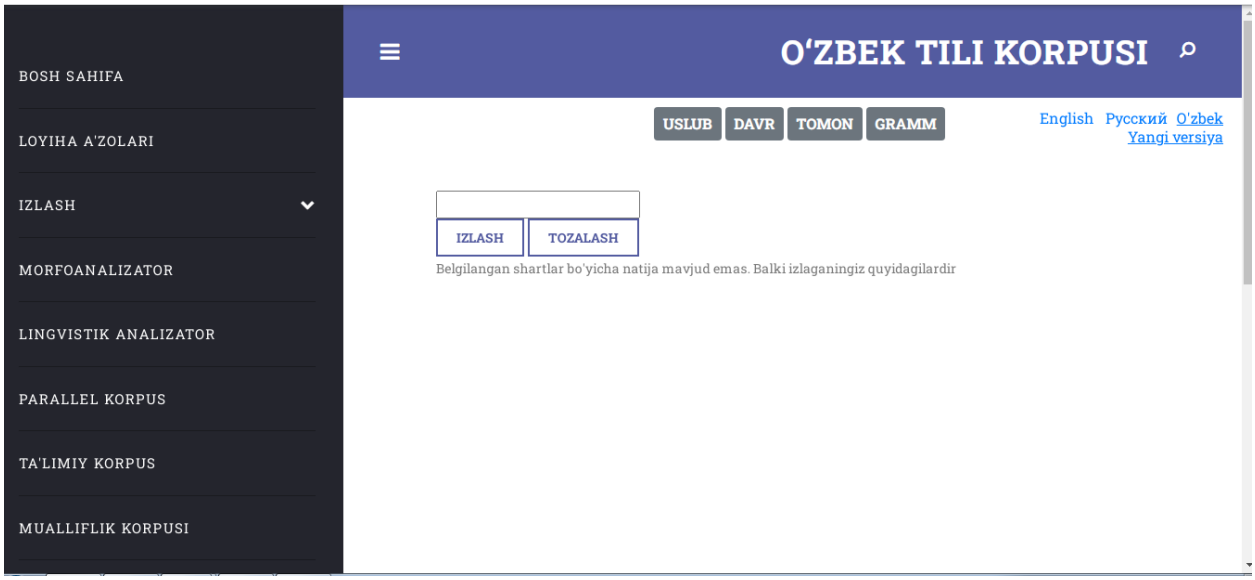
In corpus linguistics, the use of software that allows users to search quickly and reliably is indispensable. Some of these programs, such as concordancers, allow users to search for words within a text. Many such programs enable the generation of frequency data from the texts. For example, they can create a list of all the words appearing in the corpus, along with a frequency list that shows how often each word occurs in that specific corpus. Concordancers and frequency facts are considered two important aspects of analysis, serving as significant parts of corpus linguistics in both qualitative and quantitative terms.

One of the most important sites in the corpus search system is uzbekcorpus.uz. For the Uzbek language corpus, lemma and token are the primary units for searching. However, the corpus search manager is also designed for searching tokens, word combinations, and sentences. The linguistic database forms the morphological basis of the Uzbek language and serves as the foundation for all search algorithms. It enables the extraction of statistics on the words available in the corpus. Additionally, through the search system, it is possible to generate concordances based on the n-gram occurrences of any given word in the corpus. Lemma refers to the dictionary form of a word, and for most languages, it can serve as the material for derivation and grammatical formation. One of the distinctive features of the Uzbek language in relation to Turkic languages is the combinatory formation of morphemic units.

The functions of the Uzbek electronic corpus are as follows:

- Homepage
- Search (by lemma, token, or concordance)
- Morphological analyzer
- Subcorpora (educational, parallel corpora)
- User guide
- Electronic dictionaries
- Thesaurus

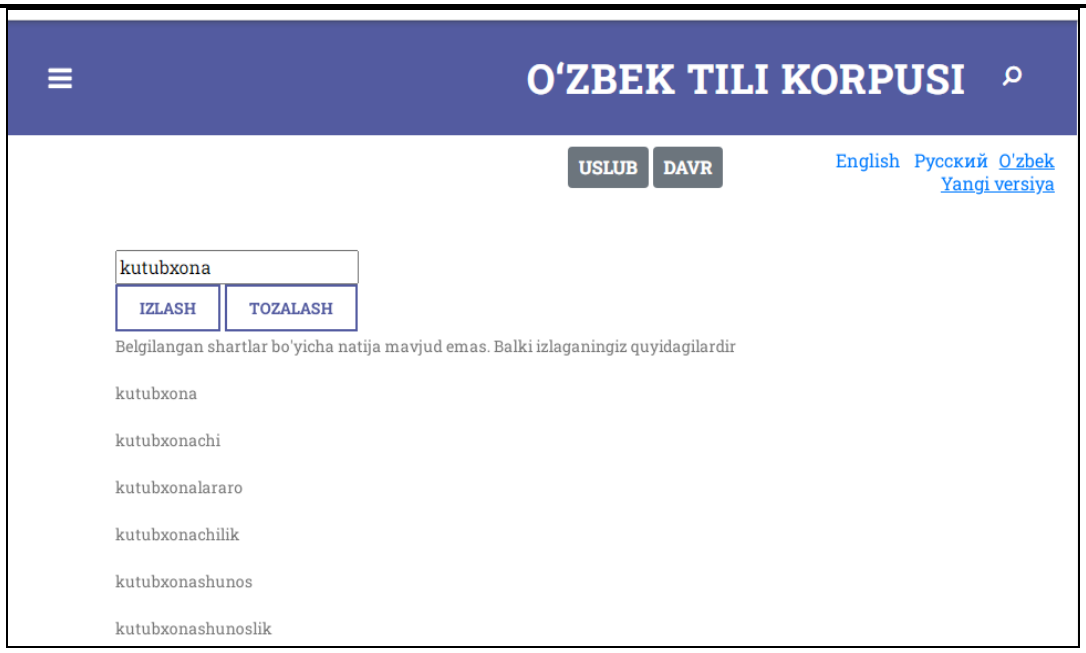
The corpus search system performs functional tasks based on three main units.



1) Searching texts by style
2) Searching texts chronologically, i.e., by period
3) Searching for concordances using the n-gram model, from both the left and right sides

This electronic corpus has the capability to generate statistical data based on the following aspects:
1) Searching by lemma – for example, when the word "kitob" (book) is searched, the root form of this word is identified within the text.

When searching by lemma, words are queried in their dictionary form, either with zero grammatical features or with specific grammatical affixes, and the result is returned according to the form requested in the query.
The lemma is considered the main unit for information retrieval in the corpus. As a result of morphological analysis, statistical data on independent word categories are obtained. Morphologically formed words will have various forms in terms of combination.
2) Searching by token – creating a query based on the [root + grammatical category] model. In this case, derivational affixes or the morphological analysis of the unit involving the word are examined.



3. Searching by concordance – n-gram left [W1 + W2 + W3 + Lt] / n-gram right [Lt + W1 + W2 + W3 + Wn]. In this case, W represents a word, W1 refers to the n-th word at a specified distance from the searched word, and Lt refers to [lemma + grammatical category]. This method allows for the identification of words within a specific context by examining the left and right proximity of the target word and its grammatical features.

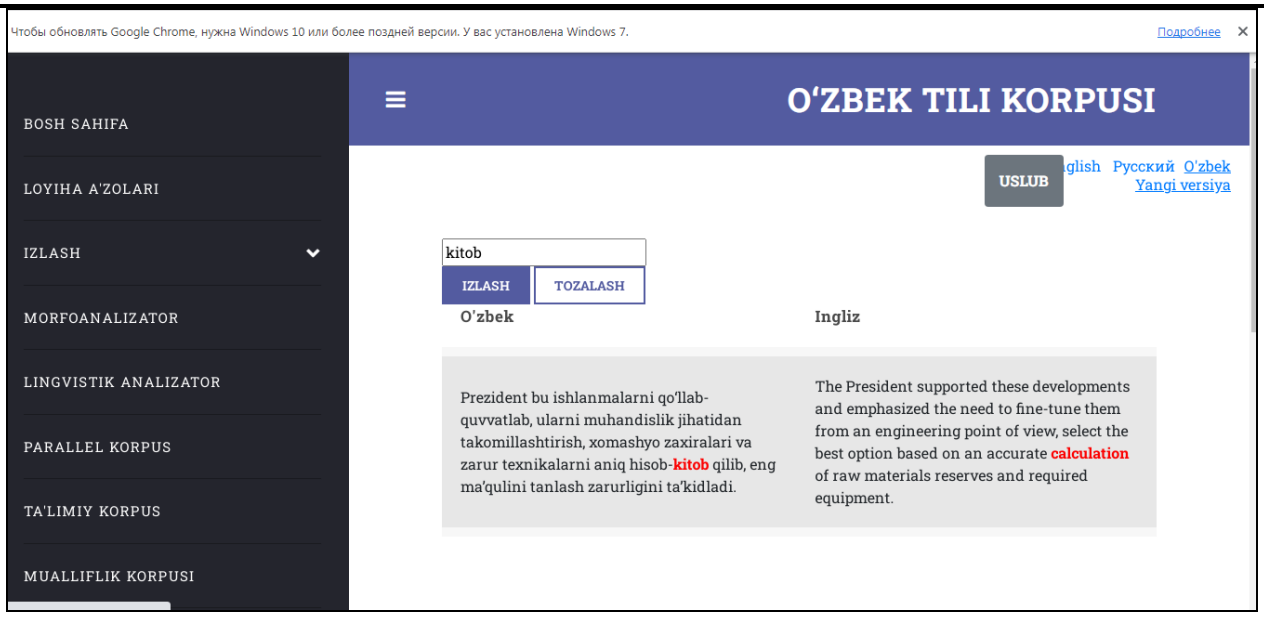
In concordance searching, the relationship of a word with the preceding or following units in the search system is indicated using the n-gram model. This approach allows for the examination of the word's context by looking at the sequence of words that appear before or after it, thereby providing a clearer understanding of its usage in different linguistic contexts.

In the system, the lemma of any word in the text, its part of speech, and the morphological annotation derived from the FST method will be displayed in the window.

4) Phrase-based search – [W + W + Wn + grammatical category]. This type of search is useful for identifying fixed expressions, such as "forest king" or "field queen," where grammatical markers in phraseology are considered. The system takes into account the possibility of grammatical suffixes serving as components in phrases like these.

This system supports searches in both Latin and Cyrillic scripts.

The internal corpus of the Uzbek language electronic corpus is a parallel corpus, which allows searching for segmented units from Uzbek-English and English-Uzbek parallel texts stored in translation memories.



The confirmation of a word's accuracy or the identification of its error by searching for similar words in the corpus database is considered a crucial aspect of corpus technology.

The evolution of corpus linguistics development shows its close connection with computational linguistics. Scientific achievements created in one field complement each other and contribute to the optimal use of available possibilities. Definitions of corpora describe them as collections of words used for linguistic analysis, and they are described as linguistic datasets composed of written texts or transcriptions of spoken language. The primary goal is to validate hypotheses about language.

Although definitions and assumptions about corpora may vary slightly, they all align in defining corpora as a collection of texts, stored in electronic form, which are governed by a search program, designed to explore the characteristics of linguistic units. These collections, stored in a computerized search system, aim to clarify language structure and facilitate research in natural language processing.

List of References:

1. Мирзиёев Ш. Миллий тараққиёт йўлимизни қатъият билан давом эттириб, янги босқичга кўтарамиз. – Тошкент: "Ўзбекистон" НМИУ, 2017. – Б.168.
2. Abdurahmonova N.O'zbek tili electron korpusining kompyuter modellari. Monografiya.Toshkent-2021.
3. https://www.merriam-webster.com/dictionary/corpus
4. Crystal D. An Encyclopedic Dictionary of Language and Languages. - Oxford, 1992.