American Journal of Philological Sciences

VOLUME Vol.05 Issue 03 2025
PAGE NO. 148-151
DOI 10.37547/ajps/Volume05Issue03-37

Formation of a database of phraseological units based on the corpus of the Uzbek language

Yodgorov Umidjon Saydilla oglu

Teacher, Department of Computer Linguistics and Digital Technologies, Tashkent State, University of Uzbek Language and Literature named after Alisher Navoi, Uzbekistan

Received: 28 January 2025; **Accepted:** 28 February 2025; **Published:** 31 March 2025

Abstract: This article discusses the issues of forming a database of phraseological units based on the Uzbek language corpus. A linguistic database is a structured collection of information that stores words, phrases, grammatical forms, idioms and other linguistic elements, and is used in lexicography to create dictionaries and reference books, in natural language processing to support technologies related to automatic translation, speech recognition and text generation, in linguistic research to analyze the structure of the language, the change and use of language units, and in language education to create educational materials and tools for language learning.

Keywords: Linguistic corpus, phraseological unit, search, database.

Introduction: A linguistic database is an organized collection of information that stores words, phrases, grammatical forms, idioms, and other linguistic elements. Such databases may be used for a variety of purposes, including:

1. In lexicography: creation of dictionaries and reference books.

2. Natural language processing: support for technologies related to automatic translation, speech recognition and text generation.

3. Linguistic studies: analysis of language structure, changes and use of language units.

4. In education: creating educational materials and tools for language learning.

A database can be composed of tables with specific fields, where each language unit's definition, part of speech, illustrative examples, and other metadata are covered.

The creation of a database of linguistic data (language units) is carried out in several stages. Here are the step-by-step steps to create such a database:

1. The goal and requirement are determined. Before creating a database, it is determined what the database will be used for:

- It is determined which language units (words, phrases, grammatical devices) the database will contain;

- What information will be stored about each unit (definition, word class, examples, synonyms, antonyms, etc.);

- The criteria for searching and filtering data are determined.

2. Platform selection. The platform for storing the database is selected and determined:

- Relational databases (e.g., MySQL, PostgreSQL, SQLite).

- NoSQL databases (e.g., MongoDB).

- Spreadsheet editors for small projects (e.g., Microsoft Excel, Google Sheets).

3. Design the database structure. The table structure for the database is created. The main tables can include:

- the table of language units: contains the main information about each unit;

- table of definitions (meanings): definition, explanation and examples of cases of use are described;

- table of grammatical signs: information about word group, person, number, etc. grammatical signs.

4. Data entry. The process of data collection and tabulation can be done manually or automated using scripts (for example, using Python to analyze textual

data).

5. Development of the input interface. If the database is used by other users, it is appropriate to create an interface for easy use. Such an interface can be created in the following form:

- web application with data entry and search forms;

- a mobile application.

6. Testing and iteration. Testing is required to ensure that the database functions as intended. Bugs will be fixed based on feedback from the user(s).

7. Documentation and database maintenance. The database is kept up-to-date by adding documentation, new language units, and corrections.

8. Renewals and Extensions. The database is regularly updated by adding new units, fixing errors and improving the structure according to the user's needs.

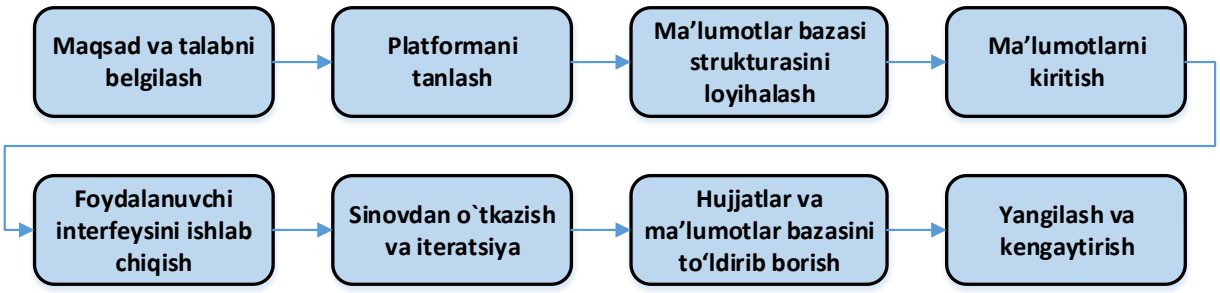


Fig. 1. These are the steps to create a database like this

A functional and efficient linguistic database can be achieved by following the above steps.

Above is a general description of the creation of linguistic data. Below we will consider the phraseological database.

The phraseological database serves for the enrichment of lexicographic resources, language teaching and linguistic research. There are some specific aspects to consider when compiling this view of the linguistic database.

1. Requirements for the content of the database:

1. To determine the type of phraseological units (phrase, article, fixed combination).

2. Determination of attributes:

- phraseology (phrase);

- description;

- translation (if the database also includes data in other languages);

- examples;

- scope of the topic, category: for example nature, food, etc.;

- level of use;

- synonyms and antonyms (if there are any).

3. The design of the database content is required:

Data model: formation of database structure. For example:

- "phrases" table (ID, phraseology, explanation, category)

- "examples" table (ID, phraseology ID, example)

- "translations" table (ID, phraseology ID, translation, language)

Creating a database of phraseological units using a linguistic corpus is an effective way to collect, analyze and create a database of phraseological units based on real examples from the language. This process is carried out according to the following algorithm:

1. Selection of linguistic corpus. In this process, a corpus is selected to be the source for the database. It is appropriate to choose a linguistic corpus that includes various text materials (for example, news, fiction, scientific articles). A special corpus software (eg Sketch Engine, AntConc) can be used for easy search and extraction of phraseological units (idioms).

2. Extract phraseological units. Searching for phraseological units (finding fixed phrases) requires the use of search and filtering functions in the selected corpus (for example, by searching for combinations of the "verb + noun" pattern).

3. Automation of the search.

- scripts and algorithms: if the object is a large amount of text, it is possible to perform a search using scripts in Python or another programming language to automatically extract phrases using regular expressions or other word processing methods;

- machine learning: the use of machine learning methods to classify and extract idioms from large texts gives good results.

These are the general algorithm and description of creating a database of phraseological units, and in this section we will focus on the specific algorithm of developing a database of phraseological units based on

the corpus of the Uzbek language.

As mentioned in the previous sections, the object of research is the corpus of the Uzbek language, and the phraseological database was formed on the basis of this corpus. For this, as a result of querying in the following sequence, a list of each phraseological unit was found, a grammatical paradigm, and then a database was formed on the basis of these.

To determine phraseological units from the database of the Uzbek language corpus, we use the following ER scheme:

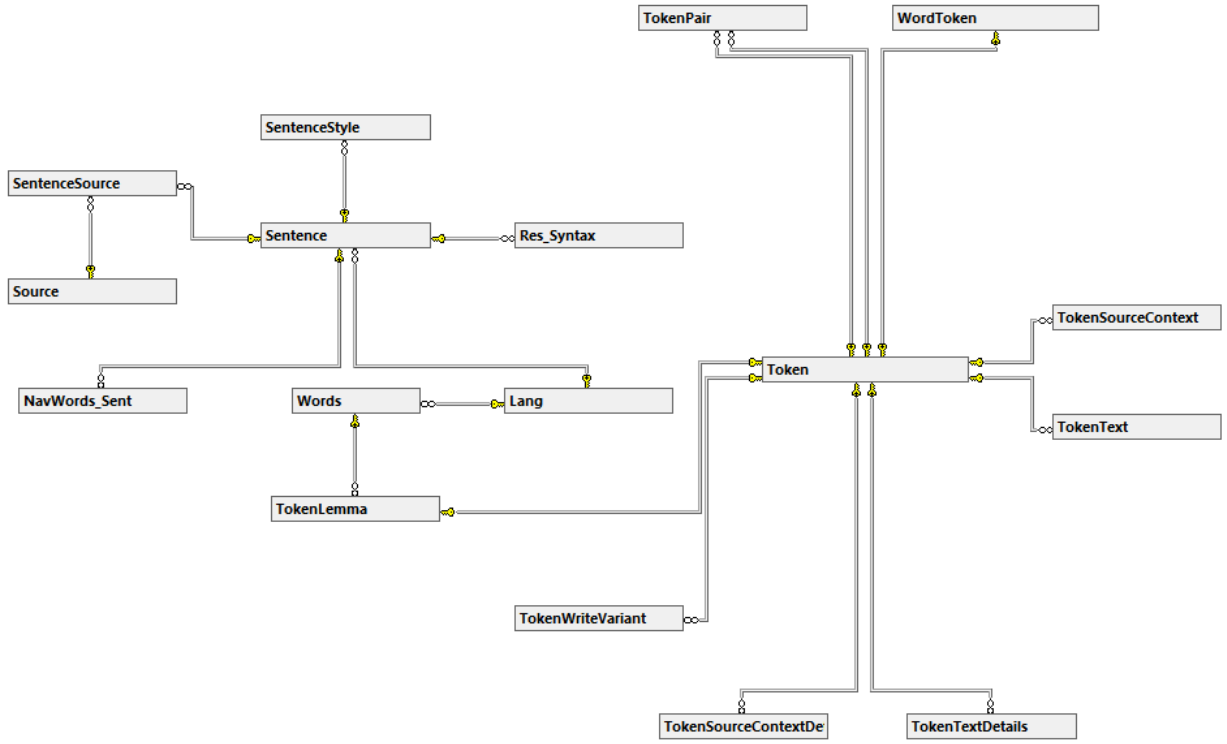


Figure 2. the following ER scheme in determining phraseological units

In this section, we will analyze how phraseological units of different frequencies occupy a place in the database, the structure and content of the database, statistics of phraseological units of the Uzbek language.

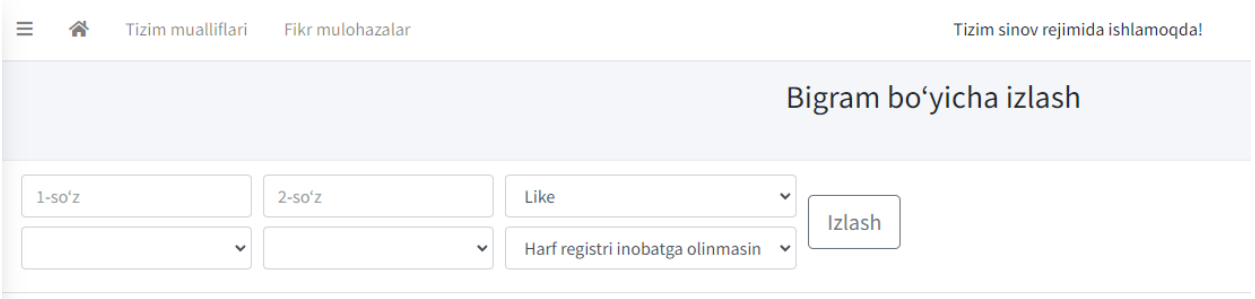
SELECT dbo.Token.Id, dbo.Token.Value, dbo.Token.CorrectVersion,
dbo.WordToken.LemmaValue,
dbo.WordToken.LemmaSuffixAffsPart,
dbo.WordToken.StemValue,
dbo.WordToken.StemPrefixAffsPart,
dbo.WordToken.StemSuffixAffsPart,
dbo.WordToken.RootValue

FROM dbo.Token INNER JOIN
dbo.WordToken ON dbo.Token.Id = dbo.WordToken.TokenId

WHERE dbo.WordToken like % “ajali yet(moq)”

For example, to find the phrase "ajali yetmoq", the following algorithm is used to query:

1. You need to open the following link: https://uznatcorpara.uz/uz/Bigram/Search



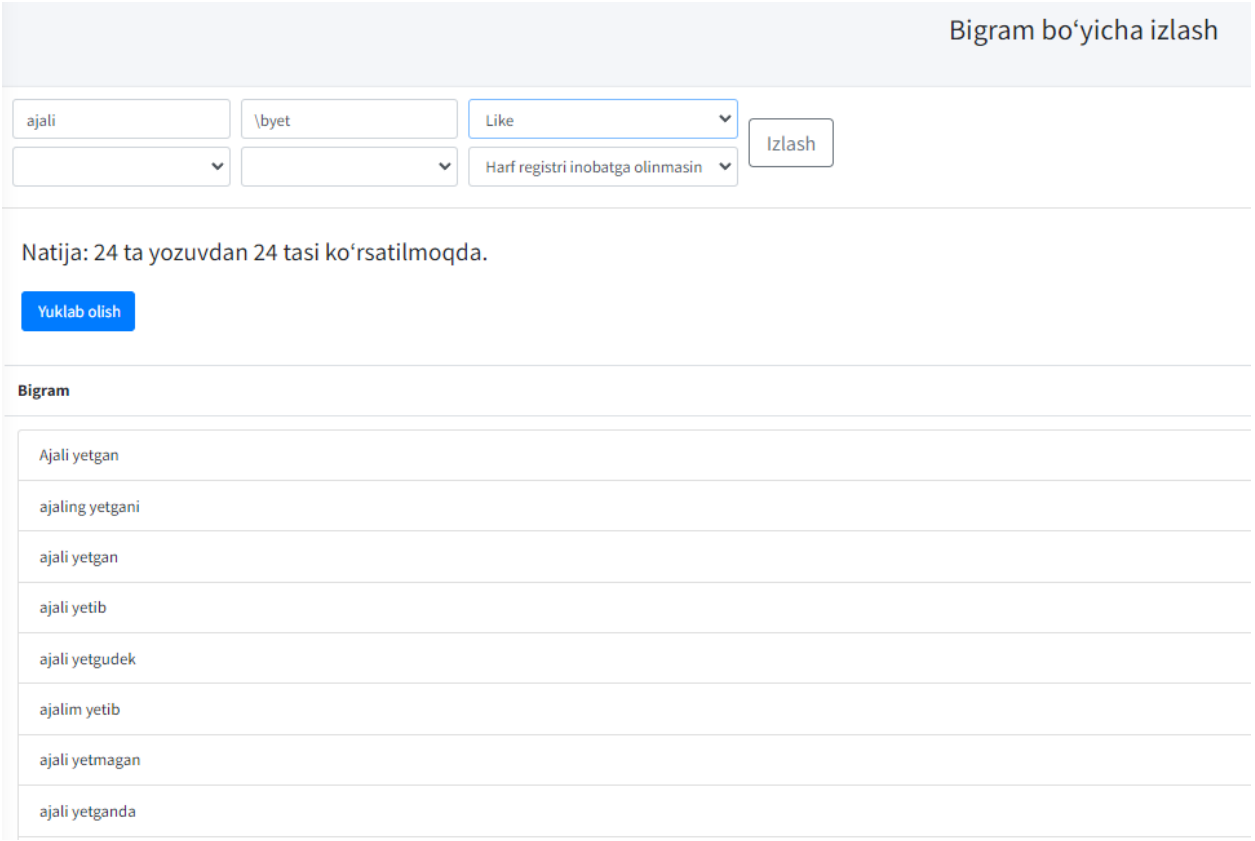
In this interface, you need to enter 2 tokens ("ajali" and "yet(moq)") in the corresponding cells.

1. You need to select the appropriate item from the section “Harf registri inobatga olinsin/olinmasin”.

2. From the corpus database, you can search for the above-mentioned phrase using "Like" (%) or " Regulyar ifoda ".

3. After setting the required search parameters, click the "Izlash" button.

4. The result will be generated in the following form:



As a result, the grammatical paradigm of the phrase “ajali yetmoq” was formed, as shown below:

1. ajaling yetgani
2. ajali yetgan
3. ajali yetib
4. ajali yetgudek
5. ajalim yetib
6. ajali yetmagan
7. ajali yetganda
8. ajalimiz yetganida
9. ajali yetgani
10. ajali yetar
11. ajalim yetguncha
12. ajaling yetmay
13. ajali yetganini

As the list shows, these morphological forms are unique and do not repeat each other. The resulting paradigm does not contain an erroneous morphological form or variant. It can be seen that 13 morphological forms of the expression “ajali yetmoq” are included in the morphological database.

REFERENCES

https://uzschoolcorpara.uz/

Раҳматуллаев Ш. Ўзбек тилининг изоҳли фразеологик луғати. – Тошкент: Ўқитувчи, 1978. – Б. 29-30.

Erman B., & Warren, B. (2000). The idiom principle and the open choice principle. Text & Talk 20(1), 29–62.

Grant, L. (2005). Frequency of core idioms in the British National Corpus (BNC). International Journal of Corpus Linguistics, 10(4), 429-451.

Heid U. Computational linguistic aspects of phraseology II // Phraseologie - Phraseology: Ein Internationales Handbuch Zeitgen?ssischer Forschung - An International Handbook of Contemporary Research / hrsg. von H. Burger, G. Ungeheuer, H. E. Wiegand. Berlin: De Gruyter, 2007. S. 1036-1044.