The OSCAR PUBLISHING Services
International Journal Of Literature And Languages

VOLUME Vol.05 Issue10 2025
PAGE NO. 163-170
DOI 10.37547/ijll/Volume05Issue10-36

Linguistic Issues Of Search Formation Of Units Of The Uzbek Language Based On Linguistic Tags

Yuldashev Aziz Uyg‘un o‘g‘li
Teacher at TashDO'TAU, Uzbekistan

Received: 28 August 2025; **Accepted:** 24 September 2025; **Published:** 27 October 2025

Abstract: This article covers the creation of the corpus of the Uzbek language and the creation of a search system based on linguistic tags for its effective use. First of all, various linguistic units in the scope of search - word form, lemma, syntactic unit, collocation, phrase and grammatical constructions were analyzed. Also, the interface elements, filters, methods of displaying results and statistical indicators used in the corpus search are described. The article shows the main types of corpus tagging - lexical (POS), morphological, syntactic and semantic tags, and their effect on search capabilities. At the same time, linguistic search methods through n-gram analysis, collocation detection, and regular expression (regex) are widely covered in the corpus. These approaches provide an opportunity to effectively use the corpus of the Uzbek language in the fields of scientific research, language teaching and automatic language processing.

Keywords: Uzbek language corpus, linguistic tags, search engine, lemmatization, syntactic analysis, morphological tagging, semantic tagging, collocation, n-gram analysis, regular expression.

Introduction: Linguistic units within the scope of the search. It is desirable that the corpus search system allows you to search for different language units. Searching of the Uzbek language corpus involves searching for the following units: word form, lemma (basic form), syntactic unit, collocation, phrase (phraseological combination) and grammatical construction. For each type of unit, a special search method and mechanism must be developed.

1. Search by word form. Searches text for the exact word or form entered by the user. This is a plain text search that finds the exact sequence of letters entered without any linguistic normalization. For example, when a user searches for the word " kitoblar ", only examples of exactly the form " kitoblar " will be extracted from the corpus. This method often comes in handy when looking for a specific form or spelling of a particular word. Word form search can often be extended using regular expressions and wildcards.

2. Search by lemma. This is a lemmatization-based search that covers all grammatical forms of the word entered by the user: the search engine first determines the lexical base (lexeme) of the word and finds all the word forms corresponding to this lemma. For example, if the user enters the word " kitob " in the search engine in the lemma position, all forms such as kitob, kitobni, kitoblar, kitobimizning are found in the results. Lemma search can also be used in the context of word groups: like searching only verb lemmas or only noun lemmas.

3. Search by syntactic unit. This allows searching for specific syntactic constructions if the corpus contains syntactic tags and structure. A syntactic unit means sentence fragments (possession, participle, complement, determiner, etc.) or syntactic connections. For example, if the user searches for the structure "possession + participle", the system will extract all possessive-participle sentences based on the results of syntactic analysis of the examples of the participle with the owner in the sentence. For this purpose, the corpus should be syntactically tagged. The syntactically annotated sub-corpus of the Russian National Corpus has a similar syntactic search: it provides a connection graph for each sentence, and the user can query this graph. As an example of a syntactic search in Uzbek, you can search for templates such as "possessive adjective construction" or " baquvvat [NOUN] [VERB]". The advantage of this type of search is that the user searches for an entire grammatical structure or sentence and retrieves relevant examples

from the corpus.

4. Search by collocations. A collocation is a word (or group of words) that occurs together in a text. Collocative search can be of two types:

1) the user searches for cases where two or more words appear close to each other (for example, within a few words);

2) automatic output of the most typical collocations for a single word by the user.

The first case is to search with the parameter of distance between words. For example, there are examples where the word " kitob " comes with the verb " olmoq " among the 3 words. In this case, the search engine sets a condition such as " kitob... olmoq " (dots - indicates an intermediate position), and as a result, contexts such as "bring a book " and "wanted to take a book " will be displayed. In the second case, the corpus itself generates a list of the most frequently occurring words for the word “kitob” (for example, with MI – mutual information measure or LL – log-likelihood indicator). If the user uses the Collocates function in the COCA corpus, he can see the most typical collocations for the entered word (high collocations such as words and books for “library”). Collocations are very important in the study of phraseological combinations or free combinations in language learning, and the search system should also provide such an opportunity.

5. Search by phrase and phraseology. This means searching for a specific stable combination, idiomatic expression or word combinations. Some phrases have a fixed form and can be found by searching for a simple word (if the phrase " koʻpchilikning ogʻziga tushmoq " is in this sequence of words). However, in some cases, the words in the phrase can change or replace each other. Therefore, it is necessary to introduce pattern search in the search engine.

For example, the construction "not only X but also Y" is a template in English, and the user can find expressions with different values of X and Y with an expression like not_only * but_also *. In the Uzbek language, the construction " nafaqat X, balki Y ham " can be found in the form of a template. In this case, the variable parts in the phrase (*) indicate that it can be any word. Also, the user can search for a similar phrase or options. For example, all variants of the expression " koʻz ochib yumguncha " can be searched in the corpus.

6. Search by grammatical constructions. This is the process of finding examples in the corpus that correspond to a specific grammatical rule or device. For example, if a user wants to search for a passive voice, he can search for verbs with the suffix “BE + Participle II” (in English) or “-ilgan/-lgan” (in Uzbek), or search for the possessive construction in Uzbek – [NOUN+ning] + [NOUN+i]. In this case, the search system will extract all examples that correspond to the pattern “nominal case noun + 3rd person possessive affix noun” using morphological characters. For example, the corpus will produce results such as “talabaning daftari”, “bogʻning gullari”, because these examples correspond to the grammatical device “X ning Y si”. Searching for grammatical devices often requires morphological and syntactic tagging information: the system “reads” the desired device from the characters. For example, if the tags in the corpus include OT (noun) and QAR (nominal case), EGA3 (3rd person possessive), the above pattern can be searched for as OT+QAR ... OT+EGA3. Such complex searches are very important for linguistic research: they help to collect all examples of a particular grammatical phenomenon from the corpus.

The ability to search by all the above-mentioned linguistic units makes the corpus search system versatile and convenient. In particular, the corpus manager should have a broad search scope so that it can meet the needs of users of different levels - language researcher, language teacher or ordinary user. As can be seen from the example of the British National Corpus (BNC), modern corpora offer a wide range of search types: word form, lemma, syntagmatic groups (compounds), search by morphological characters, context display, statistics output, etc. Therefore, in the design of search functions for the Uzbek language corpus, maximum rich and flexible search options are provided, relying on international experience.

Search interface and user capabilities. The search interface is a graphical window that provides interaction between the user and the corpus system. The user formulates a query through the search form, and the system displays the results according to this query. Below are examples of the main elements that should be in the user interface and their functions:

1. Search field (query input box): the place where the user enters the word or phrase that he is looking for. It is in the form of a simple text line, with a search button next to it. The user enters, for example, a kitob in this field and presses the Enter key or the search icon (magnifying glass). The search field has an auto-complete function (suggestions for continuing the text), and when the user types “kit”, suggestions such as “kitob”, “kitoblari” appear, which is connected to the dictionary database.

2. Search Type Options: Allows the user to choose which type of search to perform. For example, there are radio buttons or drop-down lists such as "By word", "By lemma", "By phrase", "By morphological

characters", "By syntactic pattern". If the user selects “By lemma”, the system searches for all lemma forms of the entered word. If "Morphological" is selected, the user will be expected to put a morphological filter along with the word. There is also an “Advanced search” mode, in which a template consisting of several words, distance, etc. can be specified (search in the form of X ... Y).

3. Language Unit Filter: This allows the user to filter word groups or linguistic units. For example, the user can limit the search to a specific word class by using a checkbox or list such as "Nouns Only", "Verbs Only". A similar syntactic role filter can be applied: search only for words in the possessive position. In order for such filters to work, the words in the corpus are required to be tagged with POS tags and syntactic tags.

4. Morphological properties filter: the user can restrict the grammatical categories of the word to make the search more precise. For example, searching only for nouns in the plural and accusative case, or only for verbs in the past tense, passive participle. The interface will have the appropriate menus to do this: number (singular/plural), case (genitive, accusative, dative, etc.), tense (present, past, future), aspect, etc.

5. Spacing and order parameters: if the user is searching for a combination of several words, he may be required to specify the spacing between them. The interface will have fields such as “Distance between words: min __, max __”. For example, if the user searches for the words “xalq” and “maqol” with the condition $0 \le \text{distance} \le 3$, the result will be cases where these two words appear next to each other or with 1, 2 or 3 other words between them (“xalq maqollari”, “xalq orasida mashhur maqol” kabi).

6. Corpus composition and subcorpus selection: in most cases, the corpus is divided into several subcorpora. For example, by genres (fiction, scientific texts, mass media), by periods (for example, 20th century texts vs. 21st century), by regional dialects, etc. The interface should have a subcorpus selection menu so that the user can choose which section to search from. For example, a user may want to search only a corpus of fiction or only a corpus of dialect texts. The Russian National Corps has such an opportunity. It is possible to search only in them by selecting specific authors or a set of texts through the "My Corp" menu on the website. For the Uzbek language corpus, it can be “Press Corpus”, “Fictional Literature Corpus”, “Scientific and Methodological Texts”, and the user can specify the appropriate one.

7. Results display method: the user can choose the format in which they want to view the search results. The most common format is KWIC (Key Word in Context) . In the KWIC format, each found word (or phrase) is displayed on a separate line, along with the surrounding context (a certain number of words on the left and right). The searched key is usually centered and given a different color or font to make it stand out. For example, if the user searches for the word "ilm", the result may look like this:

... bilim va **ilm** orqali ... (Manba: "X" gazetasi, 2020)

... zamon talabi – **ilm** egallash ... (Manba: "Y" kitobi, 2018)

Each concordance line also displays the source or identifier of the text (text name or ID number). The interface should allow the user to expand the context or view the entire sentence. For example, when the user clicks on a line, they can view the entire paragraph of this verse or the entire sentence. There is also a mode for displaying results by sentence. This is especially useful for complex queries: when searching by syntactic structure, it is necessary to see the entire sentence.

1. Sorting and filtering the results: it is necessary to offer sorting options so that the user can see the results found in the order convenient for him. For example, sorting according to the alphabet (according to the left or right neighbor of the keyword), according to the source (according to the name of the text or the author), according to the date (according to the date of creation of the texts), according to relevance (accuracy), etc. In the RNC example, the results can be viewed by author or year, or even in random order. In our system, the user can also choose to sort, for example, "in ascending order by text year" or "in alphabetical order by word after keyword". Filtering the results is also important: when there are too many results, the user can limit them using filters such as, for example, author = Alisher Navoi or genre = Scientific (for this, the corpus should have the corresponding metadata defined).

2. Results statistics and indicators: It would be good if the search interface presented the number of results and some statistical information at a glance. For example, after clicking the search button, the following information could appear at the top of the screen: “Found: 256 matching examples, in 89 documents.” This will help the user to understand how broad the query is. Also, the interface can offer additional tools for analyzing the results found: for example, a statistical table by word frequency, a graph (the dynamics of use by years), a table of collocations, distribution by location, etc. Such functions are considered an integral part of the corpus manager and

are present in many modern corpuses. For example, when a user searches for a word in the BNC web interface, its overall frequency and genre distribution are also displayed.

3. Help function and additional facilities: The interface for users should have a "Help" section or mention the language of the request. In the case of complex search syntax (regex, CQL), the user can see the main rules and examples here. In addition, options such as saving the search history (list of last searched words), exporting results (downloading in CSV or Excel format) , marking results and adding them to favorites make the user's work easier. Multilingualism can also be considered in interface design. For example, the menu and buttons should be Uzbek, Russian, English (according to the language chosen by the user).

Types of linguistic tagging and their impact on search. Linguistic tags are formal labels assigned to word and sentence units in the corpus; they "glue" linguistic information to each unit in the text. The corpus of the Uzbek language is expected to have four main tagging layers: lexical/POS tagging, morphological tagging, syntactic tagging, and semantic tagging. Each layer makes its own contribution to the search of the corpus, expands the possibilities of the search.

1. Word class (POS) tagging. Each word in the corpus is designated as a word class: noun, verb, adjective, adverb, pronoun, auxiliary, preposition, modal, etc. POS-tagging allows the search engine to filter by keyword and search by template. For example, a user can search for the combination "noun + adjective" with the pattern [OT SIF]. As a result, adjective + noun combinations such as “yaxshi odam”, “katta uy” come out of the corpus. The user can also limit the type of words that come around a word: queries such as OT examples after the "tarixiy" quality cannot be executed without POS-tagging. POS-tags are usually a limited number of categories (about 12 categories in Uzbek language), and machine-trained POS-taggers can be used to automatically tag them. In the Uzbek language, POS-tagging research has been conducted in recent years, in particular, BERT-based models have been tested.

2. Morphological tagging. This layer determines the morphological categories (grammatical categories) of each word. In Uzbek, nouns include features such as number (singular/plural), case (genitive, accusative, possessive, etc.), possessive form (I/II/III person possessive, possessive suffix with/without), tense (present-future, past, narrative, probable), degree/form (imperfect, passive, subjunctive, imperative), person and number (1st person singular, 3rd person plural, etc.), mood (command, condition, wish) for verbs. Morphological tagging is usually done by concatenation of tags: a set of tags is written together in one word. For example, combinations such as OTBir (noun, singular), OTBirQar (noun, singular, demonstrative), OTKo’pO'r (noun, plural, case) can be used for NOUN.

In order for morphological tagging to be complete and systematic, a list of tags (tagset) is carefully developed first. Given the rich inflectional structure and affixation system of the Uzbek language, the tagset is much more detailed (for example, the RNC for the Russian language has more than 600 morphological character combinations). We also need probably hundreds of different tag combinations. In this regard, Universal Dependencies or other international standards can be used as a basis and adapted to the Uzbek language.

1. Syntactic tagging. Syntactic tagging means specifying the syntactic structure of the sentences in the corpus. There are two main ways to do this: dependency tagging and constituent tagging. In the dependency method, each word plays a different role in grammatical relationships with other words. For example, Subject, Predicate, Object, Adverbial, etc.

From a search perspective, syntactic tagging capabilities are as follows: the user has the ability to search by syntactic relations or phrase structures. For example, if sentences are tagged with UD standard linking tags, the user can search for X words in the possessive case with the query “nsubj: X” (nsubj is possessive in UD). Or, by searching for “combination of obj and amod” he can find sentences with a complement and a qualifying adjective attached to it. Although these are very complex searches, the result will be very valuable for scientific research. In cases where RNC syntactic search is used to study word order in Russian, users make queries specifically on syntactic relations.

Semantic tagging. Semantic tagging can be understood in two ways: lexical-semantic tags - tags related to the meaning category or lexical semantics of a word (for example, a hypernym of a word, a lexical field code, a WordNet synset number, etc.); text semantic tags - tags related to the topic, pragmatic functions or speech acts.

In our project, the first type - word-level semantic tags - is more likely to be considered. For example, the BNC Sampler of the British National Corpus contains a corpus tagged with semantic categories using the UCREL Semantic Tagger, where each word is divided into 21 large semantic fields (eg People, Animals, Food, Actions, Frequency, etc.) and 232 smaller subcategories. As a result, the user can perform unusual searches, for example, searching for all words on the topic "Food and Drink". Or if it is tagged with

"Place Names" (toponyms), it is possible to extract from the corpus the contexts in which all the place names occur.

Semantic tagging is a developing area in the Uzbek language. Several studies have been conducted to create an Uzbek WordNet network, and the issue of automatic disambiguation of word meanings has been raised. In terms of impact on search, if semantic tags are available, the user will be able to perform searches based on semantics and topics. For example: search for synonyms of the adjective " katta ": if a list of synonyms for the word " katta " is obtained through a semantic network, all occurrences of them in the corpus can be extracted.

N-gram based corpus search

Types of N-grams. In corpus linguistics, n-gram is understood as a sequence consisting of n consecutive units of words. 2nd gram (bigram) - a combination of two words, for example, the combination " yaxshi odam " is considered one bigram. Likewise, the 3rd gram (trigram) consists of a sequence of three words, and so on. N-grams are used in linguistic research for collocation detection, automatic translation, speech modeling, and many other tasks. Such types as bigram, trigram, 4-gram are often included in the analysis and show the frequency of connections between units in the text. In the framework of this work, the types of bigram, trigram, quadruple and, if necessary, larger n-grams are searched and analyzed in the corpus. With the help of N-gram analysis, it is possible to identify pairs of words and combinations that often occur together, which are called collocations. Collocations are combinations of words that regularly occur together in a text, and they can form phrases or terms that have a specific meaning. For example, in the Uzbek language, phrases such as "ilmiy ish" or "katta rahmat" can be common collocations. By searching the corpus for n-grams, the most frequently repeated combinations are found, their frequency is calculated, and the strength of the connections between lexical units is analyzed. The results of such collocation analysis are useful in studying the phraseological wealth of the language, in automatic translation systems, in text indexing, and even in determining meaningless words. For example, it is known that researches have identified the most common auxiliary words and their combinations in the Uzbek language and used n-gram and collocation methods to list them as insignificant words. In order to find all bigrams, trigrams, etc. in the corpus, it is first necessary to break the text into tokens (words). Then the process of generating n-grams is carried out: at each position shift of the given text, n tokens are sequentially taken and recorded as n-grams. This process goes through the entire corpus, creating a dictionary of n-grams and their frequency of occurrence. Due to the large size of the corpus, it is important to perform such calculations efficiently; below, the issue of providing this when programming the search function is considered.

The frequency (total number of repetitions) of pairs or triplets obtained as a result of an N-gram search is one of the main indicators. At the same time, it is possible to determine the strength of the collocational bond using statistical indicators. For example, the Mutual Information (MI) indicator measures the probability of two words occurring together compared to their probability of occurring separately. If we explain the meaning of the MI formula in a simple way, two words usually occur rarely independently, but when they occur together a lot, then the MI value is high. However, MI can give a falsely high estimate in small corpora and for rare words, so other indicators such as T-score and log-likelihood are also used. In addition, the dispersion indicator represents the distribution of a word or combination throughout the corpus: it is determined whether it is evenly distributed in all texts or is found more often in texts collected in a certain genre, author or time. Dispersion analysis helps to understand whether collocations are specific to certain fields or are a universal feature of the language.

A variety of visual tools are used to effectively explain the most significant n-grams found in the corpus and their statistical performance. For example, a list of n-grams can be presented in the form of a table in descending order of frequency. It is also useful to display collocations graphically: for example, by plotting the highest-frequency bigrams as a horizontal bar chart, their relative speed is visible at a glance. One such example is given below.

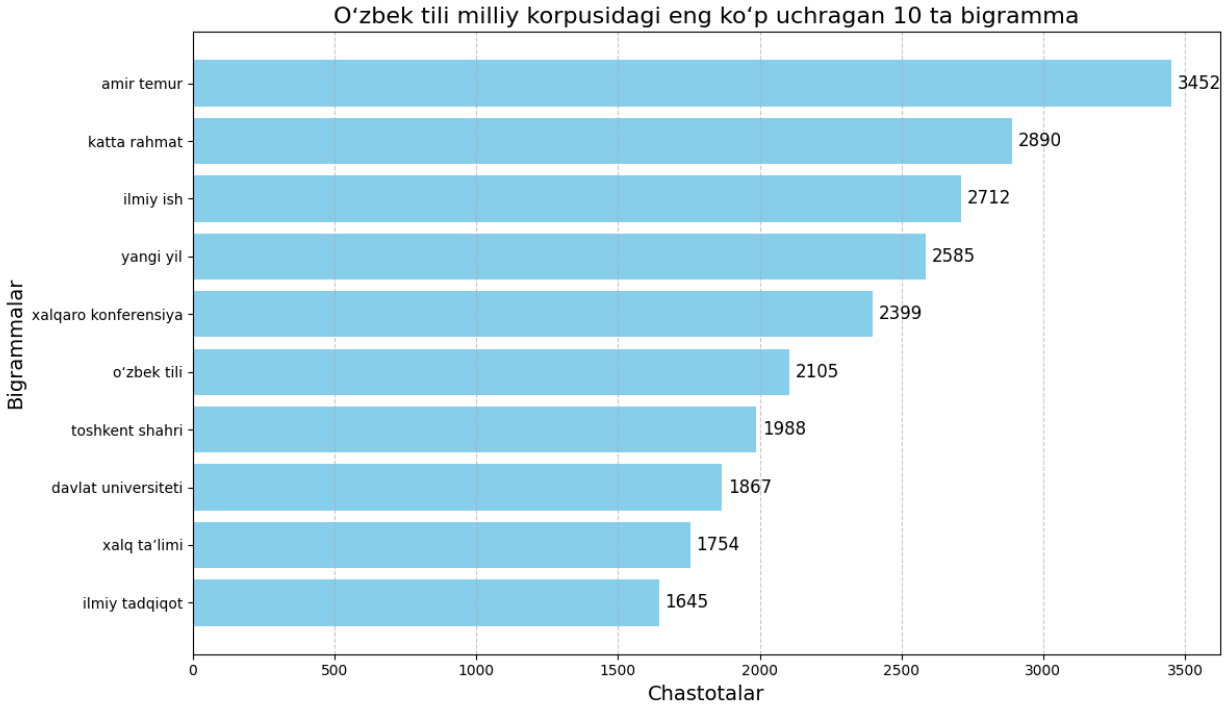

Figure 3.3.1. 10 most frequent bigrams in sample text from the corpus

In addition, the most active words or phrases from a large corpus can be represented in the form of a word cloud. In this case, the frequency of occurrence of a word in the text is proportional to its font size in the cloud, which quickly highlights the most frequently used elements.

Figure 3.3.2. Representation of the 10 most frequent bigrams in the sample text taken from the corpus in the form of a word cloud

Network graphs are also used to analyze collocations. In such a graph, words are represented as nodes, and the lines between them indicate their occurrence; the thicker the line or the closer the distance between nodes, the stronger the collocation is understood to be. Such visual representations are more understandable to the user than dry numbers, and provide a clearer picture of the linguistic analysis of the results of the corpus.

Linguistic search using regular expressions. A regular expression (regex) is a flexible search pattern for a sequence of characters in text, which is used to find any part of the text that exactly matches the specified pattern. Simply put, using a regular expression, you can send a request saying “find words or phrases that match such and such conditions”. For example, the regex ^[A-Za-z]+$ will find all words consisting only of Latin letters, or the regular expression lar$ will find elements ending in the sequence "lar". Searching the corpus using regular expressions is a powerful tool that offers more possibilities than simple text search, because with the help of regex, an infinite number of forms of words can be covered with a single expression: the user can search for all potential matches using a single general pattern that cannot be pre-listed.

In agglutinative languages, including Uzbek, multiple meanings and grammatical forms are formed by adding various affixes to the word structure. Regular expressions are very useful for searching for these affixes. For example, in Uzbek, the plural suffix is “-lar”, and a regex like \w+lar\b will find all words ending with the suffix “-lar” in the text. Here \w+ is a word body of any length, lar is a sequence of these letters, and \b is

a word boundary. Thus, using regex, it is possible to automatically extract all plural forms from the corpus and study their statistical distribution. It is also possible to check the rate of occurrence of a specific grammatical event by searching for certain tense or relative affixes of the verb. For example, in the Uzbek language, the verb of the present continuous tense takes the suffix "-yapti". The regex query \w+yapti\b extracts all verbs in the corpus that are used in the present continuous.

If the words in the corpus are tagged by word categories, regular expressions can be used to search for words in a given category or sequence of categories. For example, if each token in the corpus has its class next to it (in the form of book/N or good/Adj), then the expression \b\w+/N\b will find all nouns, or the expression Adj\s+\w+/N can find adjective + noun compounds (like (“qizil gul”, “yangi loyiha”). This method makes it possible to automatically separate syntagmatic structures (adjective + noun combinations, adverb + verb combinations) from the text and analyze them. Even if there are no tags in the corpus, regex can be used to roughly determine a word class based on a given morphological pattern. For example, in the Uzbek language, most nouns end in a vowel (such as -a, -o), and adjectives can end with suffixes such as -li, -siz. Based on such a rule, the expression \w+li\b can help in the search for adjectives, although the accuracy of such a search is not as high as in an annotated corpus, but it can reveal general trends.

In Uzbek, the sequence of word-forming and form-forming suffixes creates certain morphological patterns. Regular expressions are also used to detect these patterns. For example, when a possessive suffix and a declension suffix are combined in nouns, the forms “-ning” (singular possessive + declension) or “-lari” (plural possessive) appear at the end of the word. The expression \w+ning\b finds all words in the possessive declension, while \w+lari\b separates the plural possessive forms. Thus, it is possible to search for a specific morphological form in the corpus, find out how often it is used, and compare it with other forms. Patterns in the inflection of verbs: for example, imperative verbs often end with the suffix -(V)ng (“boring”, “oling”). Using regex, the expression \w+ng\b finds all such imperative verbs and determines the proportion of imperative mood in the corpus.

Regular expression search results are usually displayed in the form of a list. This list shows which words or phrases match the expression and in which part of the text they occur. In corpus linguistics, it is common to present such results in the form of a concordance. In a concordance, each found example is displayed on a separate line, along with the text context to the left and right of it. This method allows you to analyze the found linguistic phenomenon in context. For example, if the past tense verbs of the form “gan edi” found using regex are displayed in a concordance, the researcher can see what meaning they express in combination with the words around them (for example, narrative tone or irony). Such a description of the results is important for drawing linguistic conclusions: rather than just listing them, the context in the concordance helps to understand how they are used.

Below is a systematic analytical table of linguistic search types using Regular Expressions (RegEx) in the Uzbek language corpus (see Table 3.3.1).

Regular expressions provide the ability to quickly and accurately identify linguistic phenomena from the language corpus. In particular, in the Uzbek language, which is an agglutinative language, it is very convenient to automatically identify affixes, suffixes, and morphological forms. By extracting and analyzing the results in the form of a concordance, it is possible to study the contexts of use of certain linguistic cases. This method can also be used to conduct statistical analysis of complex grammatical and morphological structures in the corpus.

CONCLUSION

In this article, the possibilities of linguistic tagging and search of the corpus of the Uzbek language were thoroughly analyzed. Search methods for word form, lemma, syntactic unit, collocation, phrase and grammatical constructions were considered in detail. The corpus interface, options for filtering and displaying results for the user, and statistical analysis methods were also described. It was noted that the use of morphological, syntactic and semantic tagging will increase the accuracy and coverage of the search. The possibilities of quickly finding complex language phenomena using N-gram analysis, collocation detection, and regular expressions are demonstrated. These approaches, developed on the basis of international experience, serve the effective use of the Uzbek language corpus not only in linguistic research, but also in the fields of language teaching, translation systems and natural language processing.

REFERENCES

1. Bober, N., Kapranov, Y., Kukarina, A., & Tron, T. (2021). British National Corpus in English language teaching of university students.
2. Sharipov, M., Mattiev, J., Sobirov, J., & Baltayev, R. (2022). Creating a morphological and syntactic tagged corpus for the Uzbek language. arXiv preprint arXiv:2210.15234.
3. Bobojonova, L., Akhundjanova, A., Ostheimer, P., &

Fellenz, S. (2025). BBPOS: BERT-based Part-of-Speech Tagging for Uzbek. arXiv preprint arXiv:2501.10107.

4. Xudayberganov, N. (2024). O‘zbek tili korpusiga morfologik ishlov berish. Computer linguistics: problems, solutions, prospects, 1(1).
5. Sharipov, M., Mattiev, J., Sobirov, J., & Baltayev, R. (2022). Creating a morphological and syntactic tagged corpus for the Uzbek language. arXiv preprint arXiv:2210.15234.
6. Elov, B., & Ahmedova, M. (2024). N-gramlar asosida imloni tuzatish tizimini ishlab chiqish. Uzbekistan: Language and Culture, 3(3).
7. Rasulov Z.I. Tilshunoslikning zamonaviy yo‘nalishlari. moduli bo‘yicha o‘quv-uslubiy majmua. – Buxoro, 2025.
8. Madatov, K., Bekchanov, S., & Vičič, J. (2022). Dataset of stopwords extracted from Uzbek texts. Data in Brief, 43, 108351.
9. Smith, G. (2003). Searching for morphological structure with regular expressions. Tiger Projektbericht, Univ. Potsdam.
10. Avgustinova, T., & Zhang, Y. (2009, September). Exploiting the Russian national corpus in the development of a Russian Resource Grammar. In Proceedings of the workshop on adaptation of language resources and technology to new domains (pp. 1-11).
11. Xolmo‘minovna, A.O. (2022, September). Morphological Annotation System in The Corpus of Internet Information Texts in The Uzbek Language. In 2022 7th International Conference on Computer Science and Engineering (UBMK) (pp. 154-158). IEEE.
12. https://kunansy.github.io/RNC/
13. http://web-corpora.net/
14. https://www.sketchengine.eu/glossary/mi-score/