IEEE TÜRKİYE SECTION

UBMK'24

Bildiriler Kitabı
Proceedings

Editor Eşref ADALI

9. Uluslararası Bilgisayar Bilimleri ve Mühendisliği Konferansı

9th International Conference on Computer Science and Engineering

26-27-28 Ekim (October) 2024 Antalya - Türkiye

The Problem of the Archaic Words' Semantic Description in the Alisher Navoi Authorship Corpus

Shuhrat Sirojiddinov
Tashkent State University Uzbek Language and Literature named after Alisher Navoi
Tashkent, Uzbekistan
shsirojiddinov1961@gmail.com
ORCID: 0000-0001-5032-6648

Manzura Abjalova
Tashkent State University Uzbek Language and Literature named after Alisher Navoi
Tashkent, Uzbekistan
abjalova.manzura@gmail.com
ORCID: 0000-0002-1927-2669

Sulton Normamatov
Tashkent State University Uzbek Language and Literature named after Alisher Navoi
Tashkent, Uzbekistan
normamatovsulton1@gmail.com

Nargiza Gulomova
Navoi innovation university, department of philology and language teaching.
Navoi, Uzbekistan
gulomovamoi@mail.ru
ORCID: 0000-0002-7716-1799

***Abstract* — Creation and use of authorship corpora is becoming a social demand in the world. Authorship corpora are important as they are considered to be one of the modern innovative educational tools in the educational process, and the most convenient sources of information, research object and subject for researchers. This article is based on the fact that author corpora are interactive tools for learning the language in a diachronic aspect, understanding the language of the Middle Ages. Additionally, the linguistic database of Alisher Navoi's authorship corpus is statistically analyzed. The possibilities of Alisher Navoi's authorship corpus in the new version are also revealed.**

Keywords* — *statistical analysis, semantic description, semantic tag, author's corpus of Alisher Navoi, linguistic support, dictionary, ghazal.

I INTRODUCTION

In recent years, the issues of modelling lexical units, morphological and semantic analysis of corpus texts, identification of homonyms, grammatical and semantic tagging of lexical units have been studied in Uzbek corpus linguistics. Today, along with the creation of morphoanalyzer, parser, and semantic analyzer capabilities for the Uzbek language corpora, creation of the author corpus (AC), a private/structural corpus of the National Corpus of the Uzbek language, is one of the most urgent tasks facing corpus scientists. Maintaining the purity of the state language, enriching it and improving the speech culture of the population, ensuring the active integration of the state language into modern information technologies and communications is becoming a priority. Achieving widespread use of our national heritage in the modern information and communication system, promoting the works of our ancestors among young people, creating philological corpora that present examples of classic literature in a readable and understandable manner, including improving Alisher Navoi AC and for this Semantic tagging of Navoi's creative heritage is an important task [1]. The main purpose of improving the corpus is as follows:

- Increasing the importance of Alisher Navoi's authorship corpus as an innovative base and educational resource;
- Enriching the base of Alisher Navoi AC with the semantic base of ghazals from the "Khazayin ul-maoni" collection of the writer in order to read and understand classic literary materials using the corpus, to analyze the artistic work of the students, as well as to form their linguistic, literary and speech competences;
- Providing a semantic description of lexical units in ghazals, including historical words, and forming their database;
- Semantic tagging of the names of historical and cultural famous places and famous people in order to expand the spiritual and mental thinking of users, to increase their knowledge of history;
- Compilation of the classification of content-semantic categories of the lexicon of Alisher Navoi's ghazals [2] forms the basis of our research work.

The principles of authorship corpora improvement technology, the advantages of using a new type of Alisher Navoi's authorship corpus and the requirements for its lexical-semantic tag base have been determined [3], and the principles of conceptualization of providing information to the corpus have been theoretically justified.

II. METHODOLOGY

It was mainly relied on the rule-based method and statistical method to improve the author's corpus of a new type. It should be noted that Alisher Navoi's collection "Khazayin ul-maoni" was sorted based on logical criteria. Kulliyat consists of 4 divans, each divan contains 650 ghazals [4]. Currently, in Uzbek literature, Navoi's works are used a lot to understand and study the language of the 15 th century, therefore, the ghazals in this collection, which contains the best poems of the creator, have been semantically tagged, and the historical words have been manually translated into the current language. Additionally, contextual semantics was described in Uzbek language.

A. Rule-based method

Alisher Navoi lived and worked in the 15th century. There is a big difference between the Uzbek language of the Middle Ages and the modern Uzbek language. It contains many lexical units that are out of use for today's readers or whose formal expression has changed. We call them explanatory words [5]. It is unreasonable to show the

semantic group or semantic field of such words without clarifying them to the user. Currently, 60% of the language of the Navoi period is incomprehensible to today's readers, that is, many words from that period have become archaic and have fallen out of use. As a result, special dictionaries are required to properly read and understand the works of Alisher Navoi. In order to interactively present the semantic explanation of words that are difficult to understand in the process of reading Navoi's ghazals in the corpus, the historical and obsolete words in the corpus materials were analyzed by Navoi experts according to their meaning in the context using 12 dictionaries. It was identified in the work and included in the database. For example, in the ghazals of the "Khazayin ul-maoni" collection, the word "ganj" is used in 65 places, expressing different semantic meanings. In the "Annotated Dictionary of the Language of Alisher Navoi's Works" published under the editorship of E. Fazilov, 6 different meanings of the word "ganj" are explained on the example of stanzas [9]. In most ghazals, the word ganj means treasure, wealth, it is used in ghazals, as well as in other meanings (Fig. 1).

If the user of the corpus moves the cursor over the word ganj, which is considered as a word, it will be known which lexical meaning it means in the verse (Fig. 2). Through this innovative approach, the user immediately understands the exact meaning of the word (form) by interpreting it. [8]. The lexeme ganj is multi-meaningful, and it is necessary to take into account all its meanings found in verses in semantic tagging. Another example is given in table 1 based on the word *subh* (Table I).

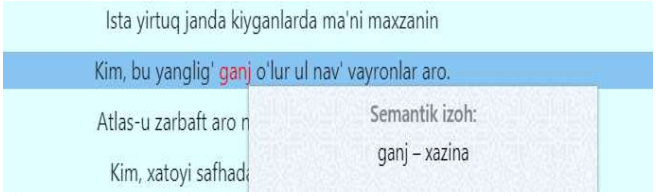

Fig. 2. Semantic description of the word ganj in the 20th ghazal of "Badoe' ul-wasat" divan on the page [13] http://v1.alishernavoicorpus.uz/gazalin_gazal/20/?csrfmiddlewaretoken=17ppht7iQCLs7LfqxXGw2uvtjcJADs JLlQx2615fi8aaWg11oE2CcsCHOcxAw9oD&search_word=ganj in the prototype of the Alisher Navai authorship corpus

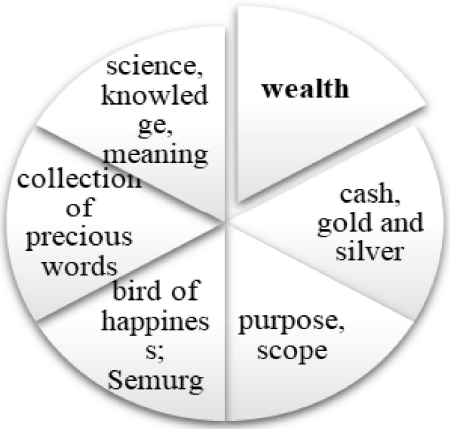

Fig. 1. Semantic meanings of the word "ganj"

TABLE I. THE MEANING OF THE *SUBH* LEXEME EXPRESSED IN GHAZALS

Ghazal number	word (form)	The basis-explanation / meaning	Verses
1	subhig'a	subh – beauty	Ey navbahori orazing subhig'a jonparvar havo
16	subh	subh – morning	G'am tuni subh tilarsen, ko'ngul uz, zaxming ko'p
39	subh	subh – morning	Subh chun esti sabo, ichkil qadahkim, bo'lmog'ung
76	subhi	subh – morning	Jamoling subhi ochib erdi yuz gul
101	subhg'a	subh – morning	Ne subhg'a bu sabohat, ne mehr muncha sabih
148	subhin	subh – morning	Vasl subhin to abad ko'z tutmasun ushshoq aro
161	subhdin	subh – morning	Qamarg'a subhdin ravnaq, quyoshqa shomdin zevar
170	subhidek	subh – morning	Shomi hijron subhidek bag'rimni chok etti sipehr
334	subh	subh - facial light	Garchi ruxsori erur subhi dilafro'z, valek
448	subhning	subh – morning	Subhning bir nafase asra damin

The main focus of this contextuality principle is to determine the meaning of a specific obsolete word in conjunction with other words in the context based on the rule. This method serves for automatic semantic description of corpus materials using N-gram method.

B. Method of statistical analysis

It was statistically determined that 40% of the words in Navoi's ghazals are related to the modern Uzbek language, and their contextual semantics was studied. As a result of the research, it was found that 15% of such words (for example, kitob, olam, dunyo, yer; yosh, quyosh, ko'z; men, sen, u, biz, siz, ular; sari, uzra)) do not require a special semantic description. For example, the word "Sun" is used not only as a cosmic body, but also in meanings such as love and affection. Therefore, such meanings require a contextual semantic explanation in their verse. The remaining 15% of words that do not have a figurative meaning were

automatically given the opportunity to express their semantic meaning in the corpus.

In 2600 ghazals of "Khazain ul-maoni" collection, Navoi's style of creation, skill of using words, range of words used in ghazals, unique writing style were determined using statistical analysis method. For Alisher Navoi's corpus of authorship, the total words in the lexicon were divided into two groups: 1) historical, obsolete - words that are difficult to understand today; 2) words used in modern Uzbek language. The scope and volume of the work are shown statistically in Table II: the table shows 4 divans in the kulliyat, each of which has 650 ghazals, the number of verses in the ghazals, the total number of words in the kulliyat and 2600 ghazals the total number of word forms in the past.

TABLE II. THE AMOUNT OF WORD FORMS IN "KHAZAYIN UL-MAONI" COLLECTION

Name of divans in "Khazain ul-maoni" collection	**Statistics of ghazals, stanzas and verses**	**The total number of words in a divan and the number of words in 650 ghazals**
"G'aroyib us-sig'ar"	650 ghazals: 4975 stanzas – 9950 verses	75,350,000 words in total, 66,954 words in ghazals
"Navodir ush-shabob"	650 ghazals: 4998 stanzas – 9996 verses	71,527,000 words in total, 66,596 words in ghazals
"Badoye' ul-vasat"	650 ghazals: 5001 stanzas – 10002 verses	71,580,000 words in total, 66,539 words in ghazals
"Favoyid ul-kibar"	650 ghazals: 5029 stanzas – 10058 verses	75,911,000 words in total, 66,722 words in ghazals
Total:	**2600 ghazals: 20003 stanzas – 40006 verses**	**There are 294,368,000 words in total, 266,811 words in ghazals.**

III RESEARCH RESULTS

Ghazals received a semantic classification in the database based on the principle of conceptuality, that is, a specific word is surrounded by other words that appear in the same stanza, in harmony with the main content of the stanza [6]. As can be seen from Table III, in our research, 266,811 word forms in 2,600 ghazals available in Alisher Navoi's "Khazayin ul-maoni" collection were semantically processed based on method of statistical analysis. Table III gives the statistics of literary-genre units specific to classic literature with semantic description and tags in the corpus base.

TABLE III. STATISTICS OF SEMANTICALLY DESCRIPTION UNITS IN THE DATABASE OF ALISHER NAVOI'S AUTHOR CORPUS

Explanatory lexical units in Alisher Navoi's ghazals	**Words**	**Proverb**	**Idiomatic expressions**	**Literary poetic arts**					
	Tagged words	***Proverb***	***Expressions***	***Talmeh***	***Tanosib***	***Tazod***	***Tashbih***	***Tasdir***	***Tasbe'***
statistics	266 811	124	92	883	More than 3000	253	124	145	32

As a result of the research, the following information was added to the gazelles in the SQL database:

1) text type of ghazals;
2) audience age;
3) serial number of the ghazal;
4) the historical, obsolete word in the verse and its current semantic explanation was given in separate fields;
5) each comment was coded by ID (Table IV).

In the process of tagging, it was taken into account that a certain word in one verse of Alisher Navoi's ghazals can give a contextual meaning in the case of collocation, or more precisely, that the totality of meaning intended by Alisher Navoi occurs, in which case they are coded with a single character. Thus, word combinations with a certain meaning were included in one semantic group. For example, when the explanatory words in the verse of ghazal 590 "Dayr aro tushsa falak engiga xirqam yuruni" were semantically tagged, a special code number 13536 was attached to all the word forms in the verse [7]. This is, firstly, a way of introducing the poem to the machine, and secondly, the machine understands that the semantic description given to the words in this line is exactly the semantic description of ghazal number 590 with code 13536 [10].

In the verse shown in Table 4:1) dayr – dunyo; 2) falak – osmon; 3) eng – yuz; 4) xirqa – eski juldur kiyim; 5) yurun – yamoq word forms were identified with the number 13536.

TABLE IV. ENCODING OF VERSES

Sequence number in the annotated word base	**ghazals number**	**Word (form)**	**Explanation / meaning**	**ID number**
20720	590	dayr	dayr – dunyo (world)	13536
20721	590	engiga	eng – yuz (face)	13536
20722	590	falak	falak – osmon (sky)	13536
20723	590	yuruni	yurun – yamoq (patch)	13536
20724	590	engiga	eng – yuz (face)	13536
20725	590	xirqam	xirqam – eski juldur kiyim (old clothes)	13536
20726	590	hajr	hajr – ayriliq (loss)	13537
20727	590	yog'ini	yog'in – yomg'ir (rain)	13537
20728	590	ashk	ashk – ko'z yosh (tears)	13537
20729	590	juvuni	juvun – daryo (rever)	13537

A search filter was created in Alisher Navoi's author corpus [http://alishernavoicorpus.uz/uz] using the above fields in the SQL database. The search engine [11] shows the place of use and frequency of the keyword (form).

In the project, six Navoi scientists worked on writing a semantic description of archaic words in 1950 gazelles, in the first version of the corpus – 650, a total of 2,600 gazelles (Fig. 3). This process lasted a year. After completing the semantic description process, each archaic word ID was coded (Table V). It also makes repetitive archaisms recognizable in software and links each word form to its contextual meaning.

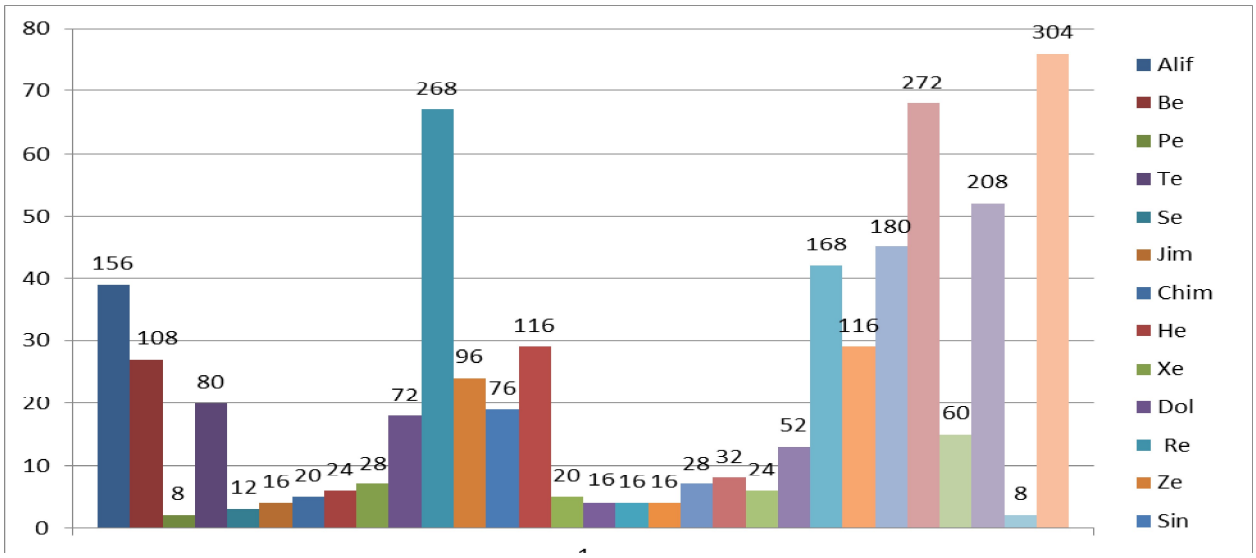

Fig. 3. Arrangement and statistics of 2,600 ghazals in the "Khazayin ul-maoni" kulliyat based on the arabic alphabet

TABLE V. ID ENCODING OF THE WORD "DIMOG"

Ghazal number	Explanation / meaning *of* the word "dimog'"	ID code	Ghazal lines
167	dimog' – kayfiyat, ruhiy holat ***(moodt)***	14898	Birovki mushkdin osoyishi *dimog'* istar
300	dimog' – aql, miya ***(mindt)***	28550	Xabt o'lg'an elga mash'ala dudi bila *dimog'*
318	dimog' – burun ***(nose)***	31565	Binafshaning nega bo'lmish *dimog'i* muncha uluq
395	dimog'– istak, xohish ***(dream)***	37612	Sen dog'i bir gul ishi birla *dimog'imni* qizit
507	dimog' – ko'ngil ***(heart)***	44946	Kim erur andin *dimog'im* ichra o't, ko'zimda – su

The word "dimog'" is used 41 times in the ghazals of collection "Khazayin ul-Maoni". The lexeme "dimog" is considered a polysemantic word (in a special explanatory dictionary it has 6 meanings) (Fig. 4), the context is necessary to determine its meaning in the poems. At this stage, the semantic valences of the archaic word on the left and right should be taken into account based on the N-gram method. Only then will the original semantics appear. As a result of providing a semantic description that does not depend on the contextual meaning of an individual word, the main idea expressed in the entire ghazal may change. This is, in fact, a subtle and complex aspect of the semantic analysis of ghazals. The subtle spiritual differences in the word "dimag'" in stanzas in poems were revealed only with the help of explanatory dictionaries and Navoists. Dictionaries are the main source, but contextual semantics not reflected in dictionaries is quickly promoted by Navoists. Thus, as a result of the line-by-line semantic description of archaic words in Alisher Navoi's ghazals, they have a contextual explanation, and the exact dictionary meaning of a certain archaism in a line is presented to the user interactively when clicking on it in the corpus [14] (Fig. 5).

Based on the performed works and prepared databases, the workflow in the corpus was algorithmized based on the architecture shown in Fig. 6.

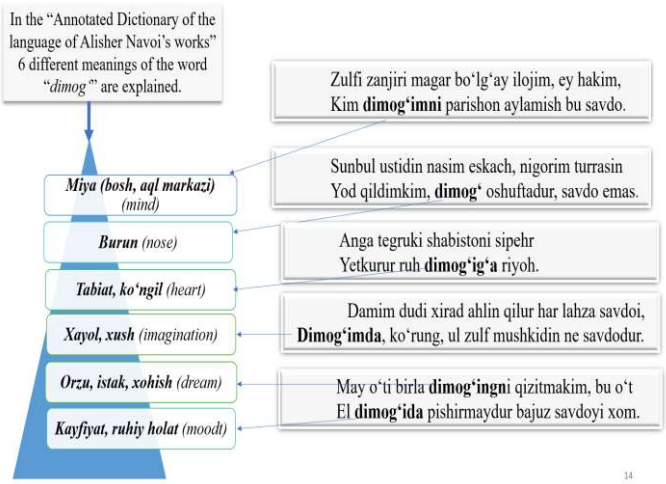

Fig. 4. "Dimog" is given 6 comments in the "Dictionary of Alisher Navoi's works"

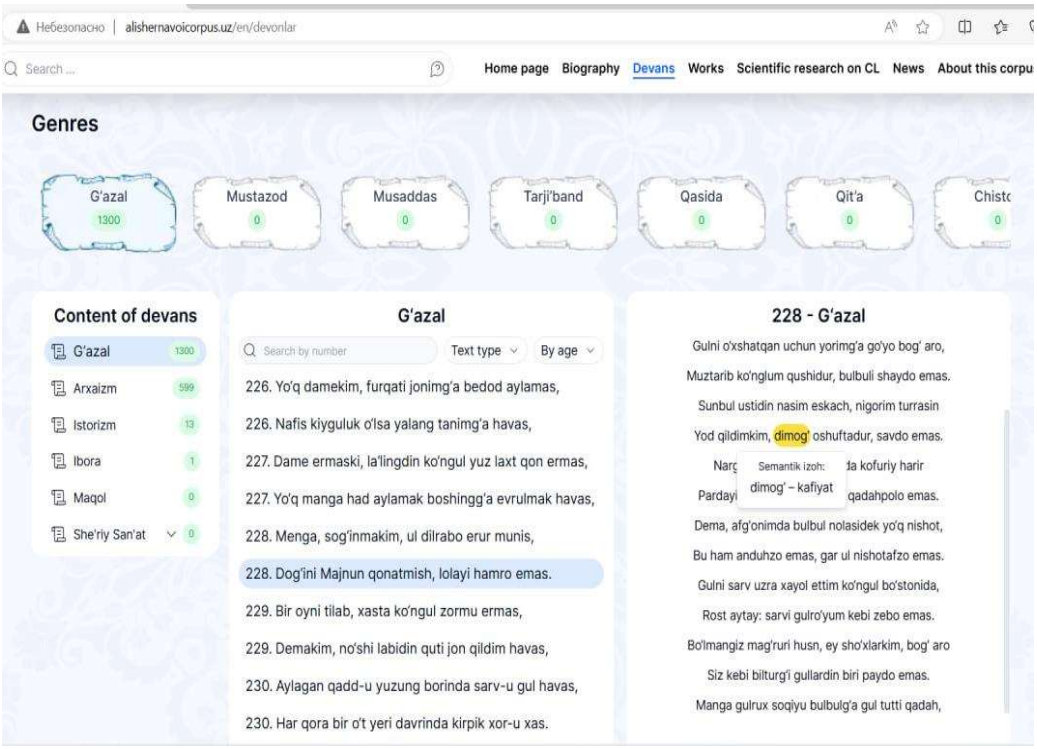

Fig. 5 In the corpus, the semantic description of the word “dimog” in verse 4 of ghazal 228 is presented interactively (by clicking the mouse on this word) [14] (http://alishernavoicorpus.uz/uz/devonlar?search=dimog%27)

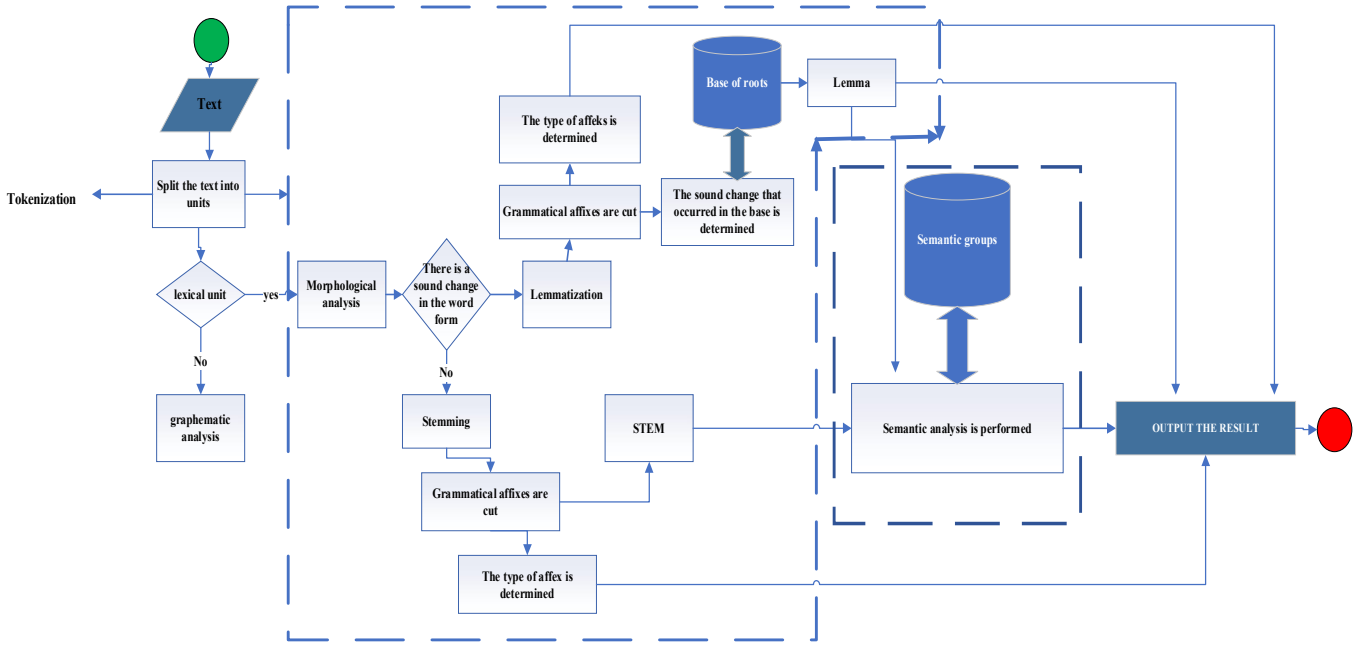

Fig. 6. Alisher Navoi authorship corpus workflow has the following architecture (e.i. BPMN)

IV. CONCLUSION

In this study, the semantic description of archaic words in the ghazals of the “Khazayin ul-maoni” collection for Alisher Navoi author’s corpus was highlighted.

The lexicology of the time Navoi lived and the current lexicon differ in many aspects, in particular, in order to ensure that the archaism and historicisms found in ghazals and explanatory lexemes with narrowed meaning are understandable and readable, semantic tagging and lexical, morphological should be analyzed [12]. Currently, in the corpus database, each semantic relationship of polysemous words expressed in verses is being semantically tagged on the basis of faceted classification. Such an approach has a semantic group of all the words in the corpus materials, and helps to correctly derive their contextual semantics. By dividing the words in the ghazal into lexical-semantic groups, it becomes possible to know the poet’s style, the skill of using words, in a clear and complete way, in a short time [13].

REFERENCES

[1] M. Abjalova, Korpus lingvistikasi: uslubiy qo‘llanma / M.A. Abjalova. – Toshkent: Nodirabegim, 2022. – 110 b.

[2] M.Abjalova, “Tagging and Annotation of Corpus Units”, International Journal of Language Learning and Applied Linguistics, ISSN: 2835-

1924 Volume 2 | No 12 |– pp. 103-107. December 2023. https://inter-publishing.com/index.php/IJLLAL/article/view/3228/2733
[3] M. Abjalova, E. Adalı, and O. Iskandarov, “Educational Corpus of the Uzbek Language and its Opportunities,” 8th International Conference on Computer Science and Engineering (UBMK), Burdur, Turkiye, pp. 590-594, September 2023. doi: 10.1109/UBMK59864.2023.10286682.
[4] Sh. Sirojiddinov, Alisher Navoiy. Manbalarning qiyosiy-tipologik, tekstologik tahlili. Toshkent: Akademnashr. 2011. – 329. b.
[5] M. Abjalova and N. Gulomova, “Author’s Corpus of Alisher Navoi and its Semantic Database,” IEEE – UBMK – 2022: 7th International Conference on Computer Science and Engineering.– Diyarbakir, Turkey. – pp. 182-187. 24-26 September 2022. Impakt Factor 5.5. DOI: 10.1109/UBMK55850.2022.9919546
[6] M. Abjalova and N. Gulomova, “Alisher Navoi and the third renaissance period,” Procedia of Theoretical and Applied Sciences. Vol. 4 (2023).– pp. 111-115. February 2023.
[7] M. Abjalova and O. Iskandarov, “Methods of Tagging Part of Speech of Uzbek Language,” *IEEE – UBMK – 2021:* 6th International Conference on Computer Science and Engineering. Ankara – Turkey.– pp. 82-85. 15-16-17 September 2021. Impakt Factor 5.5 DOI: 10.1109/UBMK52708.2021.9558900.
[8] M.A. Abjalova, O‘zbek tili ontologiyasi: yaratish texnologiyasi va konsepsiyasi, monografiya / M.A. Abjalova. –Toshkent: Nodirabegim, 2021. – 215 b. ISBN 978-9943-7804-5-3
[9] E. Fozilov and others, Alisher Navoiy asarlari tilining izohli lug‘ati, 1 jild – Toshkent: Fan, 1983-1985. – p. 32.
[10] M. Abjalova and N. G‘ulomova, “Alisher Navoiy mualliflik korpusi va uning imkoniyatlari,” Kompyuter lingvistikasi: muammolar, yechim va istiqbollar. Xalqaro ilmiy-amaliy konferensiya to‘plami. Elektron nashr / ebook. – Toshkent: ToshDO‘TAU,– pp. 89-93. April 2022.
[11] A. Rahimov, Kompyuter lingvistikasi asoslari. – Toshkent: Akademnashr, 2011. – p. 103.
[12] M. Abjalova and N. G‘ulomova, “Alisher Navoiy korpusini yaratish texnologiyasi,” Alisher Navoiy nomidagi Toshkent davlat o‘zbek tili va adabiyoti universiteti. O‘zbekiston: til va ma’naviyat. Lingvistika. ISSN 2181-922X. Vol. 4 (1). – pp. 14-32. December 2023.
[13] http://v1.alishernavoicorpus.uz/ – Version 1 of the author's corpus of Alisher Navoi (Application time: 31.07.2024)
[14] http://alishernavoicorpus.uz/en – New version of the author's corpus of Alisher Navoi (Application time: 18.08.2024)