JOURNAL OF APPLIED SCIENCE AND SOCIAL SCIENCE
eISSN 2229-3113 pISSN 2229-3205

Volume 15 Issue 12, December 2025
Impact factor: 2019: 4.679 2020: 5.015 2021: 5.436, 2022: 5.242, 2023: 6.995, 2024 7.75

JADID LEGACY INTO THE CORPUS: ELBEK-FITRAT LEXICOGRAPHY, HOMONYMY CRITERIA, AND A MODEL OF SEMANTIC TAGGING OF DRAMA VOCABULARY

Rustamov Doniyorbek Avazbekovich
Tashkent University of Applied Sciences
Doctor of Philosophy (PhD) in Philology
E-mail: doniyorbekrustamov@utas.uz
ORCID: 0009-0000-0038-5324

Abstract. The article reinterprets the Jadid heritage using lexicographic and corpus linguistic approaches. First of all, based on the lexicographical works of Elbek, Abdurauf Fitrat, and other enlighteners (EOL, etymological commentaries of Fitrat, etc.), the phonetic-graphic, lexical, and grammatical manifestations of the language of the period are characterized. Based on the analysis by Sh.Bobojonova, the differences between EOL (416 units) and ROL (497 units), the need to revise the criteria for determining homonyms (similarity of pronunciation/writing, difference in meaning) due to the spelling norms of 1929, based on Latinization and synharmonism, are substantiated. For the author's corpus, lemmatization, token-lemma-compound-concordance search, operator/constant tags, synonym-antonym layers, and a markup model are provided.

Keywords: EOL, ROL, homonym, synharmonism, etymology, semantic tagging, lemmatization.

A number of studies have been conducted on the study of the heritage of the Jadids and the creation of scientific research on them. In particular, dictionaries compiled by linguists such as Elbek and Abdurauf Fitrat during the Jadid period allow us to describe the language situation of that time, collect and summarize the necessary information, and conduct a comparative analysis with today. In his work devoted to the formation of Uzbek lexicography, S. E. Normatov extensively researched the problems of lexicography of the Jadid period, highlighting the lexicographical views of enlighteners such as Abdulla Qodiriy, Elbek, Abdurauf Fitrat, Ashurali Zohiriy, Ishoqxon Ibrat, and provides valuable evidence regarding the linguistic features of the period. Also, M. M. Latipov specifically analyzes the lexical and grammatical aspects of Ishoqxon Ibrat's works; in this study, a number of lexical units are divided into thematic groups and explained from an etymological point of view. F. Bobozhov, in his dissertation devoted to the linguistic features of Uzbek Jadid dramas, analyzes the dramas of Abdulla Avloni and Mahmudkhoja Behbudi. Based on the results of these studies, it is shown that the lexicon of Jadid dramas can be semantically tagged.

Sh. Bobojonova who studied Elbek's lexicographical activity, analyzes the forms of linguistic terms, homonyms, and polysemous units in the interpretation of Elbek. According to it, 416 homonymous units are recorded in "EOL (Elbek's Dictionary of Homonyms)," and 497 in "ROL (Shavkat Rakhmatullayev's Dictionary of Homonyms)." 178 homonyms in EOL are also included in ROL, while the remaining 238 are not included due to the following factors: ROL takes into account three criteria at once, such as pronunciation uniformity, spelling compatibility,

JOURNAL OF APPLIED SCIENCE AND SOCIAL SCIENCE
eISSN 2229-3113 pISSN 2229-3205
Volume 15 Issue 12, December 2025
Impact factor: 2019: 4.679 2020: 5.015 2021: 5.436, 2022: 5.242, 2023: 6.995, 2024 7.75

and sharp differences in meanings when defining the homonym; In EOL, formal equality is taken as the main criterion. In 1929, during the transition to the Latin script, 9 vowels, characteristic only of synharmonic languages, were adopted as the basis, and the rule of spelling according to thick-thin oppositions was followed: in thick words, letters characteristic of thin vowels (k, g, a) were not written, and in thin words, graphemes characteristic of thick vowels (q, g', o) were not written. Therefore, lexemes that differ slightly in pronunciation are presented in writing in the same form. For example, in our modern literary language, pairs such as olmos-olmas, boyroq-bayroq, qozi-qozi, chalgi-cholgu, chal-chol, chay-choy, barmoq-bormoq, olaman-olomon, which differ in pronunciation and writing, were written in the same form in the spelling of that period, and this situation was considered the norm for the spelling norms of that time. Thus, Elbek's dictionary of homonyms differs from the modern Uzbek literary language in phonetic, lexical, and grammatical indicators; these differences show that they caused certain differences from the literary language of the period as a result of attempts to update the spelling concepts created by the Jadids.

Abdurauf Fitrat has special research in the field of lexicography, which relies on the etymological approach in the interpretation of Turkic lexemes. For example: Yatika - yitika, seven stars; Arabic banāt al-na'sh ("banotunnash"), Tajik haft dodaron - constellation (45). Mung'ilamak (yaman'ilamak) means "to eat the brain": since ancient Turks served the brain of a sheep to their most honored guest during a feast, later this word was used to mean "to eat a delicious meal." Also, such units as sav - sur; qazirmoq - qayirmoq, qaytarmoq; qub - emphatic particle ("juda ham"); bulmoq - topmoq; kuvazlik - kattalik, takabburlik, shodlik are also explained. These lexemes are units in Mahmud Kashgari's "Dīwān Lughāt al-Turk," and Fitrat clarifies them with various sources and examples of living folk language. Fitrat's linguistic views are described in detail in the research of M. Kurbanova, which extensively illuminate Fitrat's activities as a linguist. At the same time, Fitrat was also a prolific playwright; the language of his historical dramas and the semantic aspects of the lexemes in them require deep analysis.

The linguistic and stylistic features of the works of Mahmudhoja Behbudi, one of the Jadid enlighteners, were studied by D.Rakhmatova. The study mainly analyzes the language of Behbudi's journalistic texts. However, Behbudi's dramas also have their own linguistic and stylistic features as a prelude to Uzbek dramaturgy. The significance of dividing words in dramaturgy into thematic groups and presenting them in corpus format lies in the fact that during the period when the Jadids lived and created, serious reforms began in language and spelling issues; protecting the language from foreign elements, forming the literary language, the syncretic nature of the living folk language is clearly manifested in the plays. Therefore, it is advisable to systematize the explanation of lexical units used in Jadid dramas in a separate dictionary order and semantically tag them as "Jadid drama terms."

The main requirements for the Uzbek language corpus search engine are as follows:

- search for words and phrases according to their grammatical, semantic, and other features;

- taking into account the text (a fragment expressing a complete thought) and the distance

JOURNAL OF APPLIED SCIENCE AND SOCIAL SCIENCE
eISSN 2229-3113 pISSN 2229-3205
Volume 15 Issue 12, December 2025
Impact factor: 2019: 4.679 2020: 5.015 2021: 5.436, 2022: 5.242, 2023: 6.995, 2024 7.75

between units (inter-word distance);

- Application of metamatin information search;

- coverage of advanced logical connectors, parentheses, and text operators;

- ensuring the effectiveness of indexation;

- answering complex questions with sufficient speed;

- scalability: ability to work with resources consisting of hundreds of millions of words.

Corpus search is carried out in four ways: by token, lemma, combination, and concordance. In the search process, along with the chosen method, it is necessary to specify the method and period parameter. Searching by token finds word forms, therefore it is necessary to include the exact form of the lexeme used in the search. Because the grammatical suffixes in the works of Abdurauf Fitrat (in general, Jadid literature) differ significantly from modern literary norms. For example, if the user enters the form of otg'a, which was encountered during the Jadid period, in the desire to search for the word noun by token, the result will be narrowed; this unit was also used in the form of otg'a at that time (quote from a Cyrillic source): "When a look sign is added to a horse, a third type of 'pronoun traces' are added to the horse that comes after it" [Abdurauf Fitrat, 2006:171; 1].

In addition to these considerations, it should be noted that in Jadid texts, especially in dramas, there are different morphological allomorphs of the same word: for example, the form borurg'a in the plays of Abdulla Avloni can be found in the dramas of Niyazi in the form of borarg'a; there are also parallel formations such as ishqig'a-ishqig'a. Analyses show that the lexemes of Jadid dramas, although belonging to the same literary period, differ from the point of view of literary language and dialect; morphologically, literary norm, dialect, phonetic variation, and context have a combined effect. This circumstance complicates the search for the token in the corpus. Therefore, when including the lexical fund of Jadid dramas in the corpus, it is important to separate synonymous lines. When searching by lemma, the lexical (original) form of the word is found; searching by compound helps to identify colloquial forms of linguistic terms found in the texts of Abdurauf Fitrat. Concordance, on the other hand, clearly indicates the environment in which the lexeme is used - the collocation surrounding units. The study of Jadid dramas within the framework of the author's corpus is of great importance from the point of view of periodicity, linguistics, and content, but it creates some difficulties in the search engine: the language of the Jadid period does not fully correspond with the language of the modern period, as a result of which the user may not receive the expected result.

The process of semantic tagging is crucial in creating separate authorial corpora for Jadid dramas. Sh. Hamrayeva, Sh. G'ulomova, D. Ahmedova, and A. Eshmo'minov. In the research of D. Akhmedova, devoted to the lexico-semantic tagging of nominative units for Uzbek language corpora, operator and constant tags are distinguished: corpus units are tagged according to their semantic field, belonging to the group and group; the group/group is designated by the operator tag, the field - by the constant tag. I. Islamov, who considered the issue of including

JOURNAL OF APPLIED SCIENCE AND SOCIAL SCIENCE
eISSN 2229-3113 pISSN 2229-3205
Volume 15 Issue 12, December 2025
Impact factor: 2019: 4.679 2020: 5.015 2021: 5.436, 2022: 5.242, 2023: 6.995, 2024 7.75

geographical terms in the corpus, emphasizes that the division of the language into several thematic groups in Uzbek language sources can serve as a natural basis for operator tags. Consequently, common thematic groups represent operator tags, and relatively stable semantic fields represent constant tags.

The following are indicated as linguistic support for the semantic markup of the corpus:

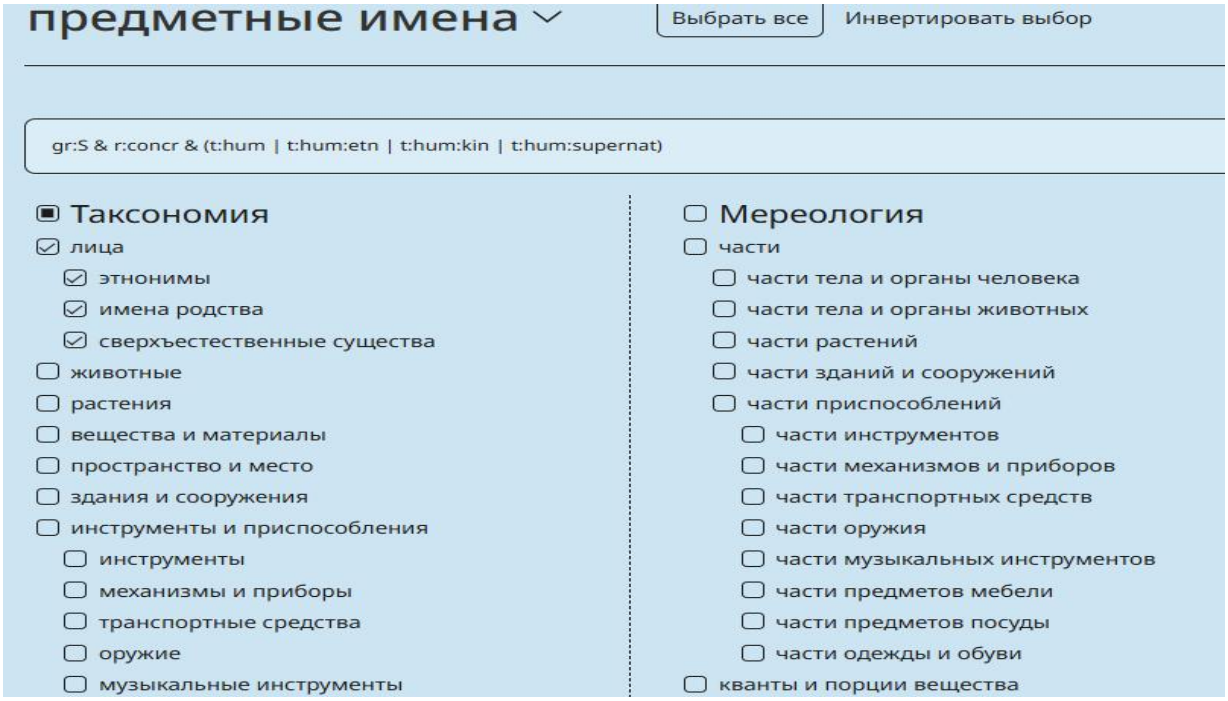
- Dictionary (lemma list)

- Semantic dictionary or semantic database

- Linguistic models of semantic markup

-Criteria for distinguishing polysemy and homonymy.

In corpus linguistics, the placement of lexical units at semantic levels is determined by the term taxonomy. In Jadid dramas, as indicated above, there are various semantic fields; when conveying homonyms, synonyms, antonyms, and polysemous units, it is advisable to introduce separate semes similar to the experience of O'TIL. A. Eshmuminov conducted research on the tagging of lexemes with taxonomic and meronymic relations in the Uzbek language in the corpus and developed a tagging mechanism through JavaScript codes based on words and their statistics in O'TIL; a linguistic base and statistics for the selection of taxonomic/meronymic units based on search parameters are proposed. The corps of Jadid dramas can also be divided into thematic groups. The Ruscorpora.ru database includes taxonomic and meronymic search engines that support searching by desired tags within selected topics.



JOURNAL OF APPLIED SCIENCE AND SOCIAL SCIENCE
eISSN 2229-3113 pISSN 2229-3205
Volume 15 Issue 12, December 2025
Impact factor: 2019: 4.679 2020: 5.015 2021: 5.436, 2022: 5.242, 2023: 6.995, 2024 7.75

We have previously divided Jadid dramas into separate thematic groups; after compiling a dictionary of Jadid drama vocabulary, semantic groups of words are also distinguished in the dictionary. The sequence can be as follows: Semantic group → word → synonym → antonym → explanation.

Example:

№	Semantic group	Word	Synonym	Antonym	Series
1	Education and science	Enlightenment	–	–	noun
2	Education and science	Jadid method	–	–	izafet
3	Religious	Superstition	–	–	noun
4	Social	Poor	destitute	rich	adjectives

Another way to create a linguistic base for Jadid dramas is to compile author's dictionaries: in which the word, word form, semantic group and field, synonymous series, antonyms, and contextual differences of polysemy are systematically recorded. Special author corpora covering the work of a particular writer are widely used in the world. The author's lexicography usually consists of two parts: the base (corpus or concordance) and the product (dictionary). In recent years, especially in corpus linguistics, the concept of "concordance" has become widespread: it refers to a list of all contextual uses of a particular language tool with a reference to the source; it is also used in the sense of an uncontextual/alphabetical index for the keywords of a work. In the corpus, which includes all completed literary texts of P. Chekhov, the texts are compiled based on the author's complete collection of works.

When creating author corpora on the works of the Jadids, information about dramatic texts can be used as linguistic support. For example, for the Behbudi corpus, the following stages of marking are proposed:

1) word form; 2) the initial (lemma) form of the word; 3) special applications; 4) extralinguistic designation: (1) the title of the work - for example, the drama "Padarkush"; (2) title/dedication; (3) year of writing; (4) thematic group; (5) main topic; (6) semantic group (science-education, religious, socio-political, socio-domestic, scientific, artistic).

In the grammatical markup of the dictionary, each case of use is determined by: lexeme (vocabulary form), grammatical sign of the lexeme (part of speech, concrete-abstract), grammatical indicators of the word form (number, case, conjugation, tense). The places "understood" by the computer are manually re-marked, and the peculiarities of the language of the period are clarified. On the basis of the specified corpus, it is possible to create various types of dictionaries - concordance, frequency, reverse, grammatical, special dictionaries by parts of speech, text and characters.

JOURNAL OF APPLIED SCIENCE AND SOCIAL SCIENCE
eISSN 2229-3113 pISSN 2229-3205

Volume 15 Issue 12, December 2025

Impact factor: 2019: 4.679 2020: 5.015 2021: 5.436, 2022: 5.242, 2023: 6.995, 2024 7.75

REFERENCES.

1. Nuritdinov A.S. Jadid davri adabiy muhitiga doir asarlardan korpusda foydalanish //Kompyuter lingvistikasi: muammolar, yechim, istiqbollar. Vol. 1 №. 01 (2024).

2. Mahmudxo‘ja Behbudiy. Tanlangan asarlar. To‘plovchi: B.Qosimov. Ma’naviyat, 1999.

3. Бобомуродова Ш. Ўзбек тилшунослиги ривожида Элбекнинг роли: Филол. фан. номз. ... дисс. автореф. Тошкент, 2002.

4. Сайидов Ё. Фитрат бадиий асарлари лексикаси: Филол. фан. номз... дисс. автореф. – Тошкент: 2001.

5. Ҳожиев А. Ўзбек тилида сўз ясалиши. – Тошкент: Ўқитувчи, 1989.