HSR (LONDON),
HOUGHTON STREET REVIEW JOURNAL
ISSN: 2058-5519|Impact factor-15.9
http://houghtonstreetreview.com/index.php/hsr

METHODOLOGICAL FOUNDATIONS FOR CREATING A DIGITAL CORPUS AND CONDUCTING POETIC ANALYSIS OF ALISHER NAVOI'S LYRIC POETRY

Pardayeva Baxtiniso
Senior Lecturer, Termez City 2-Technical School
doi:https://doi.org/10.5281/zenodo.22974103

Abstract: This article examines the digitalization and computational analysis of the lyrical heritage of the great Uzbek poet and thinker, Alisher Navoi, specifically focusing on the ghazals included in his Khazayin ul-Ma'oniy ("Treasury of Meanings") collection. The study analyzes the principles of creating an authorial corpus, a multi-layered structural, morphological, and semantic tagging (annotation) system for classical texts, and computational algorithms for the automatic detection of aruz meters, rhyme schemes, and poetic devices. Furthermore, the role and practical significance of Digital Poetics and Stylometry methods in addressing contemporary issues in Navoi studies are demonstrated through specific computational models.
Keywords: Alisher Navoi, authorial corpus, digital poetics, semantic tagging, computational linguistics, Khazayin ul-Ma'oniy, poetic devices, aruz algorithm, NLP.

Introduction

In contemporary global science, the reformatting of humanities research through digital technologies and artificial intelligence tools has led to the emergence of the field known as Digital Humanities. Conducting computational analysis on classical poetic texts opens up new horizons not only for linguistics but also for literary criticism, textual studies, and cultural history. The lyrical heritage of Alisher Navoi, considered the pinnacle of Uzbek classical literature, stands out due to its vast volume, complex genre characteristics, and intricate poetic structure.

To date, traditional Navoi studies have predominantly investigated the artistic-aesthetic and philosophical-mystical aspects of his text using descriptive and comparative-historical methods. However, without precise statistical data, it remains difficult to completely grasp the mathematical regularities of aruz meters, rhyme systems, the cross-connections of poetic figures, and the distributive spread of lexical units across a lyrical heritage encompassing thousands of ghazals. Consequently, constructing a digital corpus of Alisher Navoi's lyric poetry and implementing automatic poetic analysis on this foundation stands out as one of the most urgent tasks of modern computational linguistics.

Literature Review

The digitalization of classical poetic texts, the modeling of authorial corpora, and the study of their stylistic-linguistic nature using computational linguistics tools are deeply established in international academic traditions, particularly within the Western school of Digital Humanities through the

HSR (LONDON),
HOUGHTON STREET REVIEW JOURNAL
ISSN: 2058-5519|Impact factor-15.9
http://houghtonstreetreview.com/index.php/hsr

fundamental theories of scholars such as John Sinclair, Douglas Biber, and David L. Hoover. Specifically, J. Sinclair interprets authorial corpus analysis not merely as a collection of electronic texts, but as an objective critical and stylometric benchmark (Corpus Criticism) that reveals the aesthetic-linguistic invariants of a creator's individual style, free from subjective evaluations. In turn, D. Biber analyzed the multi-dimensional stylistic features of texts (Multi-Dimensional Analysis) via computational corpora, proving that the distribution of linguostatistic frequencies in authorial texts is a primary linguistic operation that determines genre and aesthetic architecture while facilitating the cognitive processing of the text.

Similarly, within Russian and international computational linguistics, the linguopoetic and statistical features of designing corpora for artistic and classical texts have been extensively researched. Fundamental works by scholars like A.A. Zaliznyak and M.N. Mihaylov hold a prominent place regarding the principles of building poetry corpora within the Russian National Corpus (RNC) and establishing electronic databases for Eastern classical texts. In Uzbek linguistics, the theoretical and practical foundations of creating a national corpus, along with the problems of automatic text processing and lemmatization, have been thoroughly explained in the research of M. Abjalova and other local representatives of computational linguistics. Meanwhile, the system of classical poetic structures, meters, and imagery devices has been perfectly described within traditional Navoi studies by prominent scholars such as Yo. Is'hoqov and A. Hojiahmedov.

Despite these advancements, the multi-layered annotation (tagging) of poetic units and artistic devices within a digital corpus framework, as well as the stylometric modeling of authorial style dynamics specifically using Alisher Navoi's ghazals as a case study, have not yet been investigated as a dedicated research object. Filling this specific gap defines the theoretical-methodological foundation of our current work. Synthesizing the conceptual stylometric perspectives of Western schools with the discursive-cognitive approaches in national Navoi studies and Uzbek applied linguistics integrates the international methodological foundation of this research. It enables an objective evaluation of the analyzed linguistic-poetic mechanisms based on universal digital standards.

In the direct course of analysis, this study relies on the methodological principles of the logical matrix developed by linguist D.A. Kopil, which serves to structure argumentation in scientific communication. In our approach, the components of this logical matrix (informing, differentiating, evaluating, citing authority, and appealing to background knowledge) do not remain confined within a purely pragmatic and communicative shell. Instead, they are connected to the linguostatistic formulas extensively used to mathematically measure authorial style—specifically the Type-Token Ratio (TTR) and the text-size-independent Brunet's Index—during the creation of the digital corpus of Alisher Navoi's lyric poetry. This enables a complete reinterpretation of these components as functional-stylistic and linguocognitive phenomena. The rationale behind this approach is that the distribution of lexical frequencies and symbolic images within ghazal discourse is not merely a statistical index; it

HSR (LONDON),
HOUGHTON STREET REVIEW JOURNAL
ISSN: 2058-5519|Impact factor-15.9
http://houghtonstreetreview.com/index.php/hsr

acts as the primary rhetorical mechanism for cognitively embedding the author's emotional-intellectual concepts into the reader's mind and activating stable aesthetic frames within the audience.

The primary objective of this research is to model a multi-layered digital authorial corpus for Alisher Navoi's ghazals and to execute a comparative-evolutionary analysis of the poet's style based on corpus data, focusing on its linguostatistic, stylometric, and poetic features. In doing so, the manifestation of this platform through lexical-morphological, syntactic, and poetic tagging (tagset) tools is investigated. Furthermore, the study aims to uncover the linguocultural and methodological foundations of digitalizing classical poetic texts according to international TEI (Text Encoding Initiative) standards.

Methodology

This study utilizes the principles of forming an authorial corpus from computational linguistics, object-oriented programming methods, and the international standard metadata framework, TEI (Text Encoding Initiative). The ghazals included in Alisher Navoi's Khazayin ul-Ma'oniy collection were selected as the object of research.

The creation of the corpus model was executed using a four-stage filtering pipeline:

1. **Grapheme-Phoneme Analyzer:** Converting texts from Cyrillic/Latin scripts according to the phonetic regularities of the Old Uzbek (Chagatai) language.
2. **Morphological Analyzer:** Extracting word forms (lexicon, affixes) from the transcribed text and reducing them to their lemmas.
3. **Semantic Tagset:** Assigning specific digital identifiers (IDs) to each part of speech and poetic imagery device.
4. **Algorithmization:** Integrating scripts that automatically detect poetic meters (aruz) and artistic figures using Python programming libraries such as NLTK, Regex, and Stanza.

```
xml
<!-- Example of Tagging Ghazal Couplets in XML Format -->
<bayt id="0023_01">
  <misra tartib="1">Ko‘zi yoshidin ul oy bag‘rini qon etti nihon,</misra>
  <misra tartib="2">Navoiy, dardi aro chorasiz qon etti nihon.</misra>
  <poetik_tahlil>
    <sanat turi="tashbih" vosita="ul oy" obyekt="yor"/>
    <sanat turi="tazod" komponentlar="yosh/qon"/>
    <aruz bahr="hazaj" vazn="mafoiylun mafoiylun mafoilun"/>
  </poetik_tahlil>
</bayt>
```

Use code with caution.

Scientific Novelty

The scientific novelty of this research is determined by the following factors:

HSR (LONDON),
HOUGHTON STREET REVIEW JOURNAL
ISSN: 2058-5519|Impact factor-15.9
http://houghtonstreetreview.com/index.php/hsr

- For the first time, Alisher Navoi's ghazal poetry, which has historically been studied using descriptive and comparative-historical approaches, is systematic, mapped, and transferred into computer memory based on the standards of Digital Humanities and Authorial Corpus technologies.
- A conceptual template for migrating classical poetic texts into digital formats (XML/JSON) was engineered based on TEI (Text Encoding Initiative) standards, accurately reflecting the grammatical and semantic characteristics of the Old Uzbek (Chagatai) language at an international level.
- The study established principles for multi-dimensional tagging across morphological, stylistic, and poetic layers (Poetic Tagset), automatically identifying the system of symbolic images and artistic devices within the ghazals.
- Logical-pragmatic matrices from computational linguistics (specifically D.A. Kopil's approach) were adapted to the environment of an authorial corpus, proving that the distribution of lexical frequencies and the evolution of authorial style (Stylometry) are not mere mathematical data, but metadiskursiv mechanisms that activate stable aesthetic and linguocognitive frames in the mind of the reader.
- A digital methodology for the mathematical detection and unbiased evaluation of the stages of Alisher Navoi's creative development (specifically analyzing the vocabulary richness of his youth and late-age divans) was introduced via the integration of TTR and Brunet's indices within a software algorithm.

Results

In the process of developing the digital corpus of Alisher Navoi's poetry, statistical criteria for the multi-layered tagging of texts were established. To map the artistic devices, imagery systems, and lexical composition of the ghazals into a digital format, a structural model consisting of 4 primary layers was designed:

Table 1. Structural Layers of the Navoi Digital Corpus

Analysis Layer	Applied Parameters and Tags	Detection Method / Algorithm
1. Metatextual Layer	Divan title, year of composition, genre, ghazal number, number of couplets.	Relational Database (SQL).
2. Linguistic-Morphological	Lemma, part of speech, archaisms, Arabic-Persian loanwords.	Morphological Dictionary and Lemmatizer.
3. Phonetic-Rhythmic	Syllable structure (long/short), aruz meter (bahr), rhyming words, radif.	Regular Expressions.
4. Poetic-Semantic	Artistic devices (tazod, talmeh, tanosub), lyrical and mystical imagery.	Contextual Filter of Semantic Fields.

HSR (LONDON),
HOUGHTON STREET REVIEW JOURNAL
ISSN: 2058-5519|Impact factor-15.9
http://houghtonstreetreview.com/index.php/hsr

Based on this architecture, the software module processes the words within the ghazal couplets not just as a flat vocabulary list, but by carefully factoring in their poetic functions. For instance, to automatically identify the device of tanosub (harmony/affinity), words such as "gul" (rose), "bulbul" (nightingale), "bog‘" (garden), "bahor" (spring), and "xazon" (autumn) are grouped into a single contextual cluster within the system. When these words appear together within a single couplet, the program automatically appends the <sanat turi="tanosub"> tag to that segment.

Discussion

The methodology of Digital Poetics allows researchers to verify subjective assumptions in traditional literary criticism through objective statistical data. Yo. Is’hoqov, in his fundamental research, emphasized that the descriptive scope of Navoi's poetics is exceptionally vast. Through corpus analysis, we have gained the ability to express this descriptive scope in exact numbers.

For example, when investigating color semantics in Navoi's ghazals, traditional approaches typically assume that "qora" (black—as a symbol for hair, beauty marks, or night) or "qizil" (red—as a symbol for roses, blood, or wine) are the most frequently used colors. However, automated frequency dictionary generation can reveal unexpected statistical ratios. Furthermore, employing Stylometry algorithms allows researchers to mathematically analyze the dynamics of vocabulary richness—measured via the Type-Token Ratio (TTR) coefficient—between the divan of his youth (G‘aroyib us-sig‘ar) and the divan compiled at the end of his life (Favoid ul-kibar). This stands out as a critical factor in definitively resolving problems surrounding the periodization of Alisher Navoi's creative evolution.

The continuous refinement of this authorial corpus will serve as a baseline for training future Natural Language Processing (NLP) models to deeply comprehend the semantics of the Old Uzbek language.

Conclusion

Building a digital corpus of Alisher Navoi's lyric poetry and evaluating its poetic features using computer programs represents a strategic step toward adapting Uzbek philology to the requirements of the digital era.

The research yields the following key conclusions:

1. First, the developed multi-layered tagging template enables the seamless integration of classical literature into international TEI metadata standards.
2. Second, digital algorithms designed to detect aruz meters and artistic figures accelerate the speed and precision of text analysis in Navoi studies by several orders of magnitude.
3. Third, this corpus model will serve as a universal framework for the systematic digitalization of works by other classical poets of our nation, such as Babur, Fuzuli, Ogahi, and Mashrab

References

HSR (LONDON),
HOUGHTON STREET REVIEW JOURNAL
ISSN: 2058-5519|Impact factor-15.9
http://houghtonstreetreview.com/index.php/hsr

1. Abjalova M.A. O‘zbek tili korpusi muammolari va istiqbollari (Problems and Prospects of the Uzbek Language Corpus). – Toshkent: Fan, 2021. – 180 b.
2. Hojiahmedov A. Mumtoz she’riyat qoidalari (Rules of Classical Poetry). – Toshkent: O‘qituvchi, 1998. – 140 b.
3. Is’hoqov Yo. Navoiy poetikasi (Navoi's Poetics). – Toshkent: Fan, 1983. – 260 b.
4. Alisher Navoiy. Xazoyin ul-maoniy. 20 jildlik mukammal asarlar to‘plami (20-Volume Complete Collection of Works). T. 3-6. – Toshkent: Fan, 1987-1990.
5. Biber D., Conrad S., Reppen R. Corpus Linguistics: Investigating Language Structure and Use. – Cambridge University Press, 1998. – 300 p.
6. Sinclair J. Corpus Criticism: Looking at Authorial Style. – Oxford University Press, 2004. – 245 p.
7. Safarov Sh. Pragmalingvistika (Pragmalinguistics). – Toshkent: O'zbekiston milliy ensiklopediyasi, 2008. – 318 b.
8. Mahmudov N. Tilning mukammal tadqiqi yo'llarini izlab... (In Search of Perfect Ways of Language Research...) // O'zbek tili va adabiyoti. – Toshkent, 2012. – № 5. – B. 3-16.
9. G‘ulomova N., Eshmo‘minov A. Mualliflik korpuslarida semantik teglash prinsiplari (Principles of Semantic Tagging in Authorial Corpora) // Kompyuter lingvistikasi va raqamli texnologiyalar. – 2023. – №2. – B. 45–52.
10. Копыл Д.А. Современный англоязычный научный дискурс: коммуникативно-прагматический потенциал убеждения: Автореф. ... канд. филол. наук. – Ростов-на-Дону, 2013. – 26 с.