

Computational Agrarian Poetics: A Digital-Humanities Framework for Studying the Peasant Image in Contemporary Uzbek Prose

SHOYIMOVA GULSHOD SHOYIM QIZI
Asia International University, Bukhara, Uzbekistan

ABSTRACT

This article develops a reproducible Digital Humanities framework for studying the peasant image in contemporary Uzbek prose. Literary research on agrarian characters has traditionally relied on close reading, historical interpretation, typology and thematic analysis. These methods remain indispensable, but they make it difficult to test the distribution of character-associated lexical, ecological and technological patterns across larger corpora. The study addresses this methodological gap by proposing Computational Agrarian Poetics, a seven-stage workflow that combines corpus design, script and orthographic normalisation, morphological processing, seed-lexicon construction, keyword and collocation analysis, character-centred concordance, semantic-domain modelling and humanistic validation. Togʻay Murod's Otamdan qolgan dalalar is used as a pilot case because its representation of Dehqonqul links labour, land, machinery, institutional power and ecological knowledge. The article does not fabricate large-scale statistical results; instead, it establishes an auditable protocol for future empirical research and demonstrates how textual hypotheses can be translated into measurable corpus questions. Particular attention is paid to the agglutinative morphology of Uzbek, Cyrillic-Latin variation, authorial syntax and the risk of treating generative AI output as textual evidence. The originality of the study lies in combining Uzbek literary criticism with corpus stylistics and computational literary studies while defining a human-in-the-loop model in which

quantitative signals guide, but do not replace, interpretation. The framework offers a methodological foundation for comparative research across Soviet, post-independence and twenty-first-century Uzbek prose.

Keywords: Computational literary studies, corpus stylistics, characterization, Uzbek NLP, semantic domains, human-in-the-loop analysis, agrarian humanities

1. INTRODUCTION

The expansion of digital humanities has changed the scale at which literary texts can be investigated. Computational literary studies now routinely employ corpus building, annotation, lexical analysis, stylometry, topic modelling, network methods and other procedures to identify patterns that are difficult to observe through unaided reading [14; 10; 16]. The methodological survey produced by CLS INFRA describes corpus building, preprocessing, analysis and evaluation as interconnected stages of a reproducible research process [4]. For literary scholars, the key challenge is not whether computation should replace interpretation – it should not – but how computational procedures can generate transparent evidence for questions that remain fundamentally humanistic.

Characterisation is one of the most productive areas for such integration. Stockwell & Mahlberg [5] demonstrate that corpus stylistics can identify textual traces surrounding a fictional character and use those patterns to strengthen cognitive and stylistic interpretation. Their work is important because literary characters are not constructed only through explicit descriptions. They emerge cumulatively through verbs, body references, recurring objects, spatial relations, speech contexts and thematic associations. A corpus approach can retrieve these distributed cues systematically.

The peasant image in Uzbek literature is particularly suitable for relational analysis. A literary farmer may be named only intermittently, while the character's identity is continuously produced through lexical fields associated with soil, field, water,

cotton, labour, family, weather, machines, officials and bodily effort. In other words, the *dehqon* image is not a single word but a network. Traditional close reading can interpret individual instances of that network with great sensitivity; however, it cannot easily establish whether a pattern is recurrent across hundreds of pages, whether it differs between authors, or how it changes across historical periods.

Existing studies of Togʻay Murod's *Otamdan qolgan dalalar* have examined historical oppression, semiotic structure, anthropocentric semantics and stylistic repetition [6; 7; 3]. Some recent studies mention corpus-linguistic methods, yet there remains a methodological difference between naming corpus linguistics as one method and publishing a fully reproducible workflow that another researcher can re-run. The missing elements typically include corpus-selection criteria, preprocessing rules, lemmatisation procedures, query design, window size, statistical thresholds, annotation categories and validation strategy.

This article addresses that gap by proposing computational agrarian poetics: A digital humanities framework for studying how farmer characters are constructed through recurring relations among language, environment, labour, technology and institutions. The framework is designed for Uzbek prose and therefore gives particular attention to agglutinative morphology, Cyrillic-Latin variation, orthographic normalisation, literary syntax and limited NLP resources. It also addresses a contemporary methodological problem: the use of generative artificial intelligence in literary research. AI can assist with coding and exploration, but its outputs must never be treated as primary textual evidence without verification.

The article asks three research questions:

1. RQ1: Which computational procedures can make the linguistic construction of the peasant image measurable without reducing literary meaning to frequency?
2. RQ2: What adaptations are necessary when corpus-stylistic methods are applied to Uzbek literary texts?

3. RQ3: How should AI-assisted operations be documented and validated so that computational results remain reproducible? The scientific novelty lies in developing a domain-specific, human-in-the-loop methodology that connects Uzbek literary criticism with corpus stylistics and computational literary studies.

Corpus stylistics emerged from the integration of corpus-linguistic techniques with literary interpretation [12; 13]. Its value lies in making recurrent textual patterns visible while preserving the need for contextual reading. Instead of relying only on memorable quotations selected by the critic, the researcher can retrieve every occurrence of a word, phrase or character reference and examine the regularities surrounding them. This produces a more auditable evidential base for stylistic claims.

Stockwell & Mahlberg's [5] study of David Copperfield is especially relevant because it uses corpus techniques to investigate character information outside direct speech, extending earlier stylistic models of characterisation [9; 11]. Their approach demonstrates that characterisation is cumulative and distributed. For the study of the peasant image, this means that a researcher should not search only for the lexeme *dehqon* or a character's proper name. The analysis must capture the semantic environment in which the character is repeatedly located: work, land, pain, tools, authority, family and ecological perception.

Computational stylistics has also shown that pattern discovery can proceed from bottom-up extraction to human interpretation. Frontini, Boukhaled & Ganascia [1] demonstrate how computational procedures can identify stylistic patterns in literary corpora, after which scholars determine their significance. This sequence is methodologically useful for Uzbek agrarian prose because it prevents the research from being entirely constrained by a pre-selected list of themes.

The maturation of computational literary studies has increased attention to reproducibility. Schöch, Dudar & Fileva [4] distinguish corpus building, preprocessing and annotation, analysis, and evaluation as separate research steps. Each step

creates methodological choices that can alter results. Which editions are included? How are OCR errors corrected? Is punctuation preserved? Are lemmas automatically generated or manually checked? Which statistical measure defines a collocation? A publication that reports only final visualisations without documenting these decisions is difficult to evaluate and impossible to reproduce.

For literary studies, reproducibility does not mean that interpretation must become identical across researchers. Quantitative literary scholarship has repeatedly stressed that measurement is most valuable when its assumptions and scale remain visible to interpretation [15]. Humanistic arguments can remain plural. What should be reproducible is the computational path that produced the evidence: the corpus, processing rules, queries, thresholds and outputs. A second scholar should be able to retrieve the same concordance or frequency table even if they interpret its literary significance differently.

Uzbek introduces several complications. First, it is an agglutinative language in which roots combine with multiple derivational and inflectional suffixes. Token counts can therefore fragment semantically related forms. Second, twentieth- and twenty-first-century texts may exist in Cyrillic, Latin or mixed orthography. Third, literary prose frequently includes dialect, phonetic stylisation, non-standard punctuation and author-specific spelling. Fourth, OCR can confuse apostrophes, quotation marks and visually similar letters, producing false frequency patterns.

These issues make preprocessing an interpretive as well as technical stage. Over-normalisation can erase stylistic features; under-normalisation can make comparable forms appear unrelated. The methodology must therefore preserve two versions of each text: a diplomatic version that remains faithful to the source and an analytical version prepared for computational processing. Every transformation should be logged.

The research gap is thus methodological rather than merely thematic. Uzbek literary scholarship contains valuable studies of individual texts and characters, and international corpus stylistics offers robust analytical tools, but the two have not yet been

systematically integrated around the *dehqon* image through a transparent, reusable protocol. The present study contributes a seven-stage workflow that transforms a literary question into a sequence of computational operations and then returns the results to close reading.

A second contribution is conceptual. Computational Agrarian Poetics treats the farmer not as a fixed label but as a relational textual node. This allows the analysis to capture the co-occurrence of character, ecology, labour and technology and to compare those relationships across periods. A third contribution concerns AI ethics: the framework specifies where generative AI may assist and where it must not replace source-based evidence.

2. METHODS

The study is methodological and uses *Otamdan qolgan dalalar* as a pilot case. The purpose is not to claim final large-corpus statistical findings but to establish a protocol that can be implemented on a future corpus of modern and contemporary Uzbek prose. A methodological article is appropriate at this stage because the reliability of computational results depends on decisions made before analysis begins. The framework therefore makes those decisions explicit.

The proposed research corpus should contain prose texts published across at least three historical strata: late Soviet literature, early independence literature and twenty-first-century prose. Texts should be selected purposively if they contain sustained representations of rural life, agriculture, land, village economies, farmer characters or agrarian transformation. Each text should be accompanied by metadata including author, title, year, genre, edition, script, source type, geographic setting where relevant, and copyright status.

The corpus should be stored in UTF-8 plain text or a transparent XML format. Page boundaries and chapter divisions should be preserved so that computational findings can be returned to the printed source. The pilot text should never be treated as a substitute for the full corpus; it is used only to

demonstrate how hypotheses about Dehqonqul can be operationalised.

Stage 1: Corpus construction and documentation

Corpus construction begins with an inclusion protocol. The researcher defines the temporal range, genre boundaries and minimum relevance of agrarian content before collecting texts. This prevents cherry-picking after patterns are already known. For every file, a metadata record documents bibliographic information and the method of digitisation. If OCR is used, the original page image should remain available for verification.

A version-control principle is recommended. The project preserves (a) the source or page-image version, (b) an exact diplomatic transcription and (c) a computational copy. Only the computational copy is normalised. This separation allows future researchers to audit whether a frequency difference is literary or merely the result of preprocessing.

Stage 2: Script and orthographic normalization

The second stage addresses Cyrillic–Latin variation and orthographic inconsistency. Conversion should be rule-based and reversible where possible. Apostrophes deserve special attention because Uzbek electronic texts may contain several Unicode characters that appear visually similar but are computationally distinct. A normalisation map should standardise these for analysis while retaining the original spelling in the diplomatic layer.

Historical and stylistic forms should not be automatically modernised. If a form carries dialectal or authorial significance, it should be tagged rather than erased. The guiding principle is minimal normalisation: change only what is necessary for comparability and document every change.

Stage 3: Morphological processing and lemmatization

Because Uzbek is agglutinative, word-form frequency is often a poor approximation of concept frequency. A lemma such as *dala* may appear with possessive, case and plural suffixes; verbs associated with labour may appear in many tense, person and

derivational forms. The analytical pipeline should therefore include lemmatisation or carefully tested stemming. Automatic output must be evaluated against a manually annotated sample.

For publication-quality research, the paper should report the size of the validation sample, the categories of errors and the correction procedure. Literary texts are likely to produce more errors than standard news or administrative corpora because of dialogue, dialect, expressive punctuation and neologism. A human-in-the-loop approach is therefore preferable to fully automatic preprocessing.

Stage 4: Seed lexicon and corpus-driven expansion

The analysis begins with a seed lexicon organised into domains rather than a single list. The LAND domain can include *dala*, *yer*, *tuproq*, *suv*, *ariq* and related forms; LABOUR can include *ishlamoq*, *termoq*, *chopmoq*, *ekmoq* and bodily effort; TECHNOLOGY can include *ketmon*, *traktor*, *mashina* and technical vocabulary; INSTITUTION can include *rais*, *kolxoz*, *reja*, *mutaxassis*, *olim* and administrative terms; AFFECT can include pain, fear, pride, attachment and exhaustion.

The seed lexicon is only a starting point. Corpus-driven procedures should expand it by examining keywords, collocates and concordance lines. Unexpected recurrent items may reveal a dimension that the researcher did not initially anticipate. This is where computation can genuinely challenge rather than merely confirm a prior reading.

Stage 5: Keywords, collocation and character-centered concordance

Keyword analysis compares the agrarian corpus with a reference corpus to identify unusually frequent items, following established corpus-linguistic principles in which frequency is interpreted relative to an explicit comparison set [8]. The reference corpus should be similar in period and genre wherever possible; otherwise differences may reflect chronology rather than agrarian content. Statistical measures and minimum-frequency thresholds must be reported. Keyword lists are then interpreted, not treated as results in themselves.

Collocation analysis asks which words systematically occur near target terms. For example, the researcher can examine the lexical environment of *dala*, *yer*, *paxta*, *traktor* and *Dehqonqul*. Window size should be justified and sensitivity-tested. A narrow window may capture phrase-level associations, while a sentence or paragraph window may better represent narrative relations.

Character-centered concordance retrieves all textual segments around a character's name or unambiguous reference. Stockwell and Mahlberg's characterisation model suggests that such contexts can reveal recurring actions, objects, body states and thematic roles. For *Dehqonqul*, the retrieved passages can be coded into LAND, BODY, LABOUR, TECHNOLOGY, AUTHORITY, FAMILY, AFFECT and KNOWLEDGE categories.

Stage 6: Semantic-domain modelling and relational analysis

The sixth stage moves from individual words to conceptual relations. Semantic-domain modelling groups lemmas into broader fields and compares their distribution across characters, chapters or historical strata. This allows the research to ask whether the farmer is predominantly surrounded by bodily labour vocabulary, ecological vocabulary, institutional vocabulary or technological vocabulary.

A relational matrix can then model character-domain associations. The purpose is not to convert a literary character into a number but to identify where interpretation should concentrate. If TECHNOLOGY terms cluster strongly around one farmer but not another, the scholar returns to those passages and asks what kind of agency the pattern constructs.

Stage 7: Humanistic validation and AI-assisted research

Every computational signal must return to the text. The validation cycle is: quantitative signal -> concordance -> source verification -> close reading -> theoretical interpretation. This sequence protects the study from common errors such as assuming that frequency equals importance or that co-occurrence automatically proves a semantic relation.

Generative AI may assist with exploratory coding, code generation, translation of metadata or the clustering of already retrieved concordance lines. However, it should not be used to invent quotations, infer unobserved frequencies or replace the corpus. Any AI-assisted stage should record the model name and version, date of use, prompt or classification protocol, input size, validation sample and human corrections. The final claim must remain traceable to the literary text and documented computational output.

3. RESULTS

3.1. *Pilot operationalisation*: Otamdan qolgan dalalar

The pilot case demonstrates how familiar literary observations can be translated into testable corpus questions. A close reading of *Otamdan qolgan dalalar* suggests that Dehqonqul is repeatedly constructed through five relational fields: land, bodily labour, technology, institutional power and ecological judgement. A computational study would not begin by declaring these fields true; it would formulate queries capable of confirming, complicating or rejecting them.

Hypothesis H1 states that lexical items from the LAND and LABOUR domains occur disproportionately in contexts surrounding Dehqonqul compared with non-farmer characters. H2 states that TECHNOLOGY terms are more likely to co-occur with evaluative verbs and consequence vocabulary in Dehqonqul-centred contexts than in general narration. H3 states that INSTITUTION terms cluster with obligation, command and productivity vocabulary, while farmer-centred contexts more frequently contain sensory and ecological terms. These hypotheses are measurable, yet their literary significance still requires qualitative interpretation.

A practical query sequence would begin with lemmatised searches for the character name and major agrarian roots. Concordance lines would then be exported with stable source identifiers. Collocation measures would identify recurrent associations, after which manual coding would classify the relation as action, perception, evaluation, coercion or affect. The

coding guide should include examples and exclusion rules so that a second annotator can reproduce the decisions.

The pilot also demonstrates why Uzbek-specific preprocessing matters. If possessive and case endings on *dala* or *yer* are counted as unrelated forms, the relational importance of land will be underestimated. If Cyrillic and Latin versions of a term are separated, diachronic comparison may be distorted. If expressive repetition is normalised away, one of Murod's central stylistic devices may disappear from the computational representation.

3.2. *Results of the methodological synthesis*

The methodological synthesis yields seven principles for computational research on the peasant image. These principles function as the main results of the article because the study is designed to produce a research protocol rather than to report a fabricated numerical corpus analysis.

Principle	Risk addressed	Required procedure	Expected research value
Dual-text preservation	Loss of stylistic evidence	Keep diplomatic and computational versions	Auditability of preprocessing
Morphology-aware counting	Fragmented Uzbek word forms	Lemmatise/stem and validate	More reliable concept frequencies
Relational character model	Reducing the farmer to a name	Analyse domains around character contexts	Captures ecology, labour and technology
Reference-corpus control	False keyword salience	Compare with period/genre-matched prose	Better thematic specificity
Human validation	Frequency interpreted as meaning	Return every key result to concordance and source	Preserves literary interpretation
Transparent AI use	Hallucinated or opaque evidence	Log model, prompt, inputs and corrections	Reproducible AI assistance
Diachronic comparability	Historical change confused with preprocessing	Stable metadata and normalisation rules	Tracks transformation of the <i>dehqon</i> image

The first result is that digital literary analysis of Uzbek prose requires a layered text model. The source, diplomatic and computational versions serve different scholarly functions and should not be collapsed. The second result is that morphology must be treated as a core methodological issue rather than a minor technical detail. Without lemmatisation or comparable processing, claims about frequency may be systematically misleading.

The third result is conceptual: the peasant image should be modelled relationally. In computational terms, this means combining character-centred context windows with semantic domains. In literary terms, it means recognising that a farmer is constructed through recurrent relations with land, labour, tools, institutions and emotions. The fourth result is that reproducibility is compatible with interpretive plurality. Scholars may disagree about the meaning of a collocation pattern while still agreeing on the procedure that retrieved it.

The fifth result concerns AI. Generative systems can accelerate routine tasks but create a new layer of methodological opacity if their role is not documented. A human-in-the-loop model therefore becomes not merely a safety preference but a scholarly requirement. AI should be used to assist observable procedures, never to substitute for the primary text.

4. DISCUSSION

The proposed framework changes the relationship between close and distant reading. Distant or corpus-based analysis does not replace the critic's attention to a passage; it changes how passages are selected. Instead of choosing only memorable scenes, the researcher can identify recurrent patterns across a corpus and then perform close reading where the pattern is strongest, weakest or unexpectedly absent. This produces a more disciplined dialogue between scale and detail.

For the study of the *dehqon* image, the approach opens new comparative questions. Does the late Soviet farmer appear primarily through bodily hardship and institutional vocabulary? Does early independence prose increase the frequency of

ownership, memory and national identity terms? Does twenty-first-century prose introduce entrepreneurship, migration, ecology, smartphones, platforms or agritech? These questions cannot be answered reliably through one novel, but a documented corpus can transform them into a research programme.

The method also helps connect literary studies with contemporary science and technology without turning literature into a decorative supplement. A computational analysis of historical agrarian prose can reveal how technological vocabulary, scientific authority and ecological language have changed over time. Such evidence can inform environmental humanities and science-and-technology studies by showing how societies have imagined the farmer before and during digital transformation.

At the same time, computational methods introduce their own biases. A corpus contains what has been digitised, accessible and selected. NLP tools perform better on some registers than others. Frequency privileges repetition, while literature often places great weight on rare images. For this reason, the framework insists on evaluation at every stage rather than celebrating automation as inherently objective.

Theoretical and practical implications. The theoretical implication is a redefinition of literary characterisation as a multi-layered textual network. Computational Agrarian Poetics treats character not as an isolated semantic label but as a distribution of recurring relations. This provides a bridge among stylistics, environmental humanities and agrarian literary studies.

The methodological implication is the possibility of building a reusable Uzbek literary research infrastructure. A carefully curated corpus with documented normalisation, morphological annotation and metadata could support not only studies of the peasant image but also research on gender, landscape, historical memory, dialect, emotion and technological change.

The practical implication concerns research training. Graduate researchers should learn to report corpus decisions with the same precision expected in empirical social science: data source, sample criteria, preprocessing, software versions,

parameter settings and validation. Such transparency strengthens the international readability of Uzbek literary scholarship and reduces the risk that 'digital humanities' becomes only a fashionable label.

Limitations and future research. The main limitation is that the present article proposes and demonstrates a framework rather than executing a full multi-author corpus experiment. This is deliberate: reporting invented quantitative results would undermine the reproducibility the article seeks to promote. The next stage should implement the workflow on a copyright-compliant corpus and publish descriptive statistics, concordance samples, code and preprocessing logs.

A second limitation is tool availability for Uzbek. Morphological analysers and lemmatisers may vary in performance across literary registers. Future work should benchmark available tools against manually annotated literary samples. A third limitation concerns historical orthography and dialect, which require specialist philological judgement that cannot be fully automated.

Future studies could build a balanced corpus of 20–40 prose works across three periods, compare farmer and non-farmer characters, and use mixed methods to test diachronic hypotheses. Inter-annotator agreement could strengthen qualitative coding, while network analysis could model relations among characters, agricultural objects and institutions.

5. CONCLUSION

This article has proposed computational agrarian poetics as a reproducible digital humanities framework for studying the peasant image in contemporary Uzbek prose. The framework integrates corpus construction, dual-text preservation, orthographic normalisation, morphology-aware processing, keyword and collocation analysis, character-centred concordance, semantic-domain modelling and humanistic validation.

Its core methodological claim is that computation should make evidence more traceable, not interpretation more

mechanical. For Uzbek literary studies, the agglutinative structure of the language, script variation and authorial style require a human-in-the-loop approach at every stage. For AI-assisted research, transparency about models, prompts, inputs and corrections is essential.

Applied to a future multi-period corpus, the method can reveal how the *dehqon* image shifts from bodily labour and collective production toward new configurations of ownership, ecological awareness, migration, entrepreneurship and digital technology. It therefore provides both a practical research protocol and a conceptual bridge between national literary studies and international computational humanities.

REFERENCES

1. Frontini, F., Boukhaled, M. A. & Ganascia, J.-G. 2017. Mining for characterising patterns in literature using correspondence analysis: An experiment on French novels. *Digital Humanities Quarterly*, 11/2.
2. Murod, T. 1994. *Otamdan qolgan dalalar*. Tashkent: Sharq.
3. Niyozmetov, B. 2025. *Tog'ay Murod asarlarida qo'llanilgan leksik birliklarning ma'no tahlilini antroposentrik yondashuv asosida o'rganish*. Tadqiqotlar.
4. Schöch, C., Dudar, J. & Fileva, E. (Eds.) 2023. *Survey of Methods in Computational Literary Studies*. CLS INFRA. <https://doi.org/10.5281/zenodo.7892112>.
5. Stockwell, P. & Mahlberg, M. 2015. Mind-modelling with corpus stylistics in David Copperfield. *Language and Literature*, 24/2, 129-147. <https://doi.org/10.1177/0963947015576168>
6. Tolkun, S. 2015. Soviet reality and Soviet criticism in Tog'ay Murod's novel *Otamdan Qolgan Dalalar*. *International Journal of Humanities and Education*, 1/2, 323-348.
7. Tuğlacı, G. 2024. Semiotic analysis of Togay Murad's novel *Otamdan Qolgan Dalalar*. *Journal of Modern Turkish Studies*, 21/1, 62-84. <https://doi.org/10.58785/MTAD.21.2024.1.5>
8. Biber, D., Conrad, S. & Reppen, R. 1998. *Corpus Linguistics: Investigating Language Structure and Use*. Cambridge University Press.
9. Culpeper, J. 2001. *Language and Characterisation: People in Plays and Other Texts*. Longman.

10. Jockers, M. L. 2013. *Macroanalysis: Digital Methods and Literary History*. University of Illinois Press.
11. Leech, G. & Short, M. 2007. *Style in Fiction: A Linguistic Introduction to English Fictional Prose*. 2nd ed. Pearson Longman.
12. Mahlberg, M. 2013. *Corpus Stylistics and Dickens's Fiction*. Routledge.
13. McEnery, T. & Hardie, A. 2012. *Corpus Linguistics: Method, Theory and Practice*. Cambridge University Press.
14. Moretti, F. 2005. *Graphs, Maps, Trees: Abstract Models for Literary History*. Verso.
15. Piper, A. 2018. *Enumerations: Data and Literary Study*. University of Chicago Press.
16. Underwood, T. 2019. *Distant Horizons: Digital Evidence and Literary Change*. University of Chicago Press.

SHOYIMOVA GULSHOD SHOYIM QIZI

DEPT. OF UZBEK AND RUSSIAN PHILOLOGY,

ASIA INTERNATIONAL UNIVERSITY,

BUKHARA, UZBEKISTAN.

ORCID: 0009-0001-9924-8764

E-MAIL: <GULSHODA199409@GMAIL.COM>