

A Hierarchically Integrated Annotation Model for Literary Texts: Theoretical Foundations and Empirical Validation on Uzbek Prose

 Aliqulova Mohinur O'ktam qizi

Teacher, Department of Foreign Languages and Literature, Faculty of History and Philology, Tashkent University of Applied Sciences, Tashkent, Republic of Uzbekistan, Uzbekistan

Received: 16 April 2026; **Accepted:** 04 May 2026; **Published:** 31 May 2026

Abstract: Digital literary study relies on linguistic annotation, but standard models fail to integrate textual layers. This paper introduces the Hierarchically Integrated Annotation Model for literary texts. Synthesizing seven analytical levels on a token reference system, HIAM unifies major international systems with Uzbek corpus-linguistic standards. Tested on a 1000-token 20th-century Uzbek prose sample, the model achieved 90.9% accuracy and strong consensus ($\kappa = 0.85$), proving its viability for computational philology in agglutinative languages.

Keywords: Hierarchically integrated annotation, literary text, multilayer annotation, corpus linguistics, cognitive poetics, Uzbek prose, agglutinative languages, semantic tagging, stylistic markup, computational philology.

Introduction: The contemporary convergence of literary scholarship and computational text processing has revolutionized philological methodologies, reframing creative works as stratified sign systems that necessitate rigorous empirical exploration [1; 2]. Within this scientific paradigm, linguistic annotation - defined as the systematic attachment of metalinguistic information to textual segments - mediates between the aesthetic density of creative discourse and the formal constraints of algorithmic processing. This methodological shift is propelled by three prevailing paradigms: corpus linguistics, which converts raw text into queryable empirical evidence [4; 5; 27]; cognitive linguistics, which interprets meaning as a broad conceptual architecture requiring markup models that extend beyond the lexical surface [6; 7; 19; 26]; and digital humanities, which uses TEI standards to validate annotation as a core practice of modern textual scholarship [8; 20].

Nevertheless, traditional markup frameworks often fail to synthesize the structural, semantic, stylistic, and

conceptual dimensions of literature within a single operational matrix. Standard analytical tools generally function in strict disciplinary isolation. Conversely, literary writing operates simultaneously across multiple interconnected planes; isolating these levels strips away the dense internal dynamics that fundamentally define literariness.

This structural integration task is uniquely challenging for typologically distinct languages. The majority of dominant annotation frameworks target inflecting or isolating language families, meaning that their direct application to agglutinative configurations like Uzbek introduces serious morphological complexities [33; 34; 35]. National corpus-linguistic scholarship in Uzbekistan has addressed these fundamental challenges through several milestone achievements: Mengliyev's structural triad for systematic corpus architecture [33], Abdurahmonova's mathematical models for agglutinative morphology [34], Hamroyeva's lexicographic and statistical criteria for authorial corpora [35], Gulyamova's educational

corpus frameworks [38], and Eshmo'minov's automated morphological tagging algorithms [39]. While these foundational works establish a reliable computational baseline, they leave literary-specific annotation parameters - including stylistic, narrative, and conceptual layers - largely unexamined.

The current investigation addresses this scientific gap by formulating and validating the Hierarchically Integrated Annotation Model on 20th-century Uzbek prose. This study aims to build a theoretically grounded, operationally viable framework that explicitly maps cross-layer interrelations within agglutinative literary texts. To accomplish this, the investigation: defines the theoretical boundaries of literary annotation using text linguistics and semiotics; identifies key annotation strata along with their categorical inventories; establishes the operating principles and analytical affordances of the system; and empirically validates the framework via a pilot corpus experiment.

Consequently, this paper investigates three primary questions: How can the multilayered parameters of creative texts be formally unified. Which annotation strata are structurally non-negotiable. What level of empirical reliability does multilayer tagging guarantee when applied to agglutinative prose? We hypothesize that an explicit, hierarchically integrated model significantly increases both the descriptive granularity and interpretive scope of literary text analysis while remaining computationally viable.

THEORETICAL FRAMEWORK

This investigation treats the literary text not as a flat surface to be decoded, but as a polystratal sign formation whose holistic meaning is generated via the concurrent operation of multiple linguistic and semiotic layers. Drawing on the stratificational paradigm of systemic-functional linguistics [1] and Lotman's cultural semiotic theory [11], we argue that surface-level lexical units are inherently tied to complex semantic, syntactic, pragmatic, stylistic, narrative, and conceptual matrices. Under this view, annotation shifts from a neutral technical step into a theoretically inflected act of analytical reconstruction, externalizing interpretive logic into highly structured, empirical objects.

Three theoretical pillars anchor this position:

The Stratificational Principle: The literary text consists of distinct, interpenetrating linguistic layers that maintain independent categorical inventories while remaining empirically interdependent [1].

The Semiotic Principle: Every linguistic token carries denotational data along with rich connotative, intertextual, and cultural resonances [11; 21].

The Cognitive-Conceptual Principle: Lexical and grammatical selections activate underlying cognitive structures - such as frames, image-schemas, and conceptual metaphors - that demand a specialized annotation layer [6; 12; 13; 19].

These core pillars are crucial for agglutinative languages. In Uzbek literature, morphological structural richness directly generates stylistic and semantic implications on the lexical surface. Furthermore, the deep cultural networks shaped by Turkic oral genres, classical Persian-Chagatai borrowings, and 20th-century socio-cultural shifts require a cognitive layer calibrated to local, rather than imported, metaphorical systems.

The HIAM architecture merges international research frameworks with national corpus-linguistic methodologies. Mengliyev's tripartite engineering paradigm - which asserts that an operationally reliable Uzbek corpus must integrate a descriptive linguistic model, a mathematically formal representation, and a programmatically realized implementation [33] - provides the core logic for the HIAM scheme. These three elements correspond to our defined categorical inventories, token-anchored relational structures, and digital environment specifications.

Furthermore, Abdurahmonova's mathematical-linguistic modeling of Uzbek morphology [34] provides an important methodological precedent for the detailed subcategorization within HIAM's morphological layer. Hamroyeva's authorial corpus criteria [35] inform our stylistic parameters, which are applied here to compare four canonical authors. Gulyamova's educational corpus frameworks [38] contribute crucial quantitative parameters, while Eshmo'minov's automated tagging algorithms [39] guarantee the structural scalability of the HIAM model beyond the pilot stage.

Beyond computational linguistics, the framework incorporates traditional Uzbek grammatical and

stylistic traditions. The morphological foundations established by G'ulomov, Tixonov, and Qo'ng'urov [40], Rahmatullayev's etymological reconstructions [41], Hojiyev's terminological inventory [42], and Mahmudov's pragmatic studies [43] support the lexical, semantic, and pragmatic strata. Finally, Sarimsoqov's analysis of literary devices [44], Quronov's literary theory [45], Karimov's research on Cho'lpon [46], and Boboyev's theoretical insights [47] guide the stylistic and conceptual-cognitive layers.

The HIAM model differentiates seven primary strata, operationalized through defined categorical inventories and unified via a token-based reference system:

Lexical-Morphological Stratum: Captures tokens, lemmas, POS tags, and morphological markers. Features extended tagsets for archaisms, dialectal variations, and individual authorial idiolects [9; 22]. Encodes the complex agglutinative morpheme ordering unique to Uzbek text [33; 34; 39].

Syntactic Stratum: Maps clause structural hierarchies. Explicitly tags creative syntactic deviations (fragmentation, hypotaxis, inversion) [1; 23] while maintaining strict compatibility with Universal Dependencies.

Semantic Stratum: Oversees word-sense disambiguation and semantic roles via frame semantics [14]. Resolves classical and modern Uzbek lexical fluctuations utilizing Rahmatullayev [41] and Hojiyev [42].

Pragmatic Stratum: Tracks speech acts, deictic indices, and presuppositional structures. Captures narrative pragmatics, including irony, free indirect speech, and shifting focalization [2; 24; 43].

Stylistic Stratum: Registers aesthetic deviations, tropes, figures, and rhythmic features [15; 16]. Integrates Sarimsoqov's extensive taxonomy of native figural devices as its primary inventory [44].

Discourse Stratum: Evaluates supra-sentential cohesion, coherence, and rhetorical structures. Adapts Rhetorical Structure Theory and Penn-Discourse-Treebank systems for creative conventions [17; 25].

Conceptual-Cognitive Stratum: Marks conceptual metaphors, mental spaces, and cultural archetypes [6; 12; 19; 26]. Incorporates specific cultural frames

extracted from Quronov [45], Karimov [46], and Boboyev [47].

Five core structural rules regulate the HIAM system:

Explicitness: All markers must be traceable to explicitly documented categories.

Layer Independence: Each individual stratum remains internally consistent and self-contained.

Inter-stratal Mappability: Cross-references different layers through shared token anchors.

Interpretive Transparency: Exposes, rather than conceals, the underlying human interpretive decisions.

Scalability: Simultaneously supports micro-level close analysis and macro-corpus inquiry [5; 27].

METHODOLOGY

This research utilizes a mixed-methodological framework. Comparative analysis positions HIAM relative to TEI P5, Universal Dependencies, FrameNet, PropBank, and GrAF. Semantic, discourse, and conceptual modeling organize these diverse inputs into a unified formal architecture adhering to Mengliyev's design logic [33].

To validate operational robustness, an empirical pilot investigation was conducted on a 1000-token prose sample extracted from four canonical 20th-century Uzbek writers: Abdulla Qodiriy, Cho'lpon, G'afur G'ulom, and Odil Yoqubov. Selection criteria prioritized canonical standing, historical representativeness, and stylistic heterogeneity. Source files were gathered from critical academic editions and pre-processed to fix OCR-induced distortions [33; 39].

Two independent specialists processed the full dataset across all seven strata using a unified token reference system. Quality control involved parallel annotation, conflict resolution, and statistical validation. The morphological layer referenced data from Abdurahmonova [34] and Eshmo'minov [39]; the stylistic layer applied Sarimsoqov's taxonomy [44]; the conceptual layer integrated native literary frameworks [45; 46; 47]. Inter-annotator agreement was calculated using Cohen's kappa (κ) [36], and overall accuracy was verified against a post-resolution consensus reference.

Consider the constructed Uzbek sentence "Eski daraxt uyqudan uyg'ondi va shamol bilan suhbatlashardi" ('The old tree awakened from sleep and was conversing

with the wind’), modeled on personification patterns. A basic monolayer lexical-morphological scheme supplies only token boundaries, lemmas, and POS tags. The HIAM model, conversely, registers multiple data points simultaneously: syntactically, the coordinated predicate hierarchy with *daraxt* as the shared subject; semantically, the anomalous assignment of agentive predicates to a non-volitional entity; conceptually, the metaphor *NATURE IS A SENTIENT INTERLOCUTOR*, stylistically, the extended figure of personification amplified by the imperfective aspect; pragmatically, the implicature of pastoral isolation; and discursively, the scene-setting function. Shared token anchors connect these distinct levels, generating an integrated analytical record.

Empirical Validation

The 1000-token evaluation corpus was distributed equally among the four selected authors based on textual integrity, prioritizing stylistically dense literary passages over expository prose. The final sample comprised 87 sentences, with an average length of 11.5 tokens and a maximum length of 27 tokens. Lexical diversity, evaluated via the type-token ratio, reached 0.62 - directly matching figures reported for stylistically marked passages in larger Uzbek corpora established by Mengliyev [33] and Hamroyeva [35].

The aggregate Cohen’s kappa coefficient computed across all seven strata reached $\kappa = 0.85$, confirming strong operational reliability. Agreement metrics followed an expected downward gradient from formal surface layers to deeply interpretive ones: lexical-morphological ($\kappa = 0.95$), syntactic ($\kappa = 0.91$), semantic ($\kappa = 0.86$), pragmatic ($\kappa = 0.82$), stylistic ($\kappa = 0.80$), discourse ($\kappa = 0.78$), and conceptual-cognitive ($\kappa = 0.74$). This distribution matches international annotation research trends [3; 17]. Overall marking accuracy reached 90.9% against the consensus reference, with remaining discrepancies mostly clustered within the top two interpretive levels.

The pilot study yielded valuable layer-specific insights. The lexical-morphological level displayed an extremely high density of complex agglutinative suffix chains (with some lexemes carrying up to seven affixal slots), validating the extended subcategorization models designed by Abdurahmonova [34] and Eshmo’minov [39]. Syntactically, marked configurations accounted

for 14.3% of all clauses, with hypotactic embedding and structural inversion dominating in Cho’lpon’s prose, and shorter coordinated patterns prevailing in Qodiriy. The semantic stratum revealed significant polysemy within kinship and emotional vocabularies in G’ulom and Yoqubov’s texts, where Rahmatullayev’s etymological dictionary [41] served as an indispensable disambiguation tool.

Stylistically, personification, antithesis, and ethno-poetic similes constituted 60% of all marked passages, demonstrating that Sarimsoqov’s taxonomy [44] is highly effective without further expansion. Mechanically, the recurring conceptual metaphors *NATURE IS A SENTIENT INTERLOCUTOR*, *TIME IS A FLOWING SUBSTANCE*, and *HOMELAND IS A MOTHER* captured the majority of conceptual mappings, aligning with cognitive-poetic findings on agrarian-cultural literary traditions [19; 26] and local critical frameworks [46; 47].

RESULTS AND DISCUSSION

Multilayer annotation exposes critical cross-layer relations - such as syntactic anomalies driving unique stylistic effects - that traditional flat structures conceal. Within the Uzbek evaluation sample, the semantic and conceptual strata systematically separated conventionalized cultural metaphors from novel poetic innovations [12; 28]. Additionally, the model successfully isolated “emergent literariness”, where standard lexical tokens gain aesthetic value through interaction with syntactic shifts or conceptual reframing. This cross-layer dynamic was present in 17.2% of all marked structures, confirming that aesthetic effects emerge from inter-stratal dynamics rather than isolated levels.

The core principles of explicitness and inter-stratal mappability ensure that higher-level interpretive claims remain firmly anchored in lower-level linguistic evidence. This directly addresses a long-standing critique of literary stylistics: the historical reliance on impressionistic, un-warranted readings [16; 29]. Tying conceptual claims to exact syntactic and lexical anchors generates a highly replicable, cumulative evidential framework. This directly mirrors the development of Uzbek corpus linguistics, which emphasizes mathematically formalised, programmatically realized, and linguistically grounded models [33; 34; 35; 38; 39].

The HIAM architecture successfully unifies micro-level close analysis with large-scale corpus inquiry:

Micro-scale: Annotated text segments provide granular linguistic evidence for deep qualitative readings.

Macro-scale: Annotated corpora enable macro-level distributional analyses of stylistic and conceptual trends.

The model effectively blends computational distant reading [30; 31; 32] with the qualitative precision of corpus stylistics [15], enabling seamless movement between different analytical scales.

While TEI remains primarily structural [8], Universal Dependencies handles syntax, and PropBank/FrameNet address general semantics [14], they all lack specialized literary-specific layers. Similarly, discourse models like the Penn Discourse Treebank do not align with complex creative registers [17; 25]. HIAM integrates these external systems where applicable while adding strata vital for analyzing literariness. In the Uzbek context, it extends the foundational work of Mengliyev [33], Abdurahmonova [34], Hamroyeva [35], Gulyamova [38], and Eshmo'minov [39], moving beyond basic structural or morphological tagging into true multilayered literary annotation.

The model's performance on Uzbek prose confirms that multilayer annotation functions effectively beyond flecational or isolating languages. Capturing the complex affixal profiles of an agglutinative language required extending the morphological subcomponent [33; 34; 39]. It also highlighted culturally specific conceptual profiles that resist universal taxonomies, offering a valuable template for other Turkic literatures. However, certain limitations must be noted: the conceptual-cognitive stratum requires human judgment, resulting in a lower agreement rate ($\kappa = 0.74$) [3]; tropes and rhetorical frameworks demand careful local calibration [11; 21; 44]; and the 1000-token pilot is modest, meaning that expanding to larger corpora represents the next research phase [5; 27].

By aligning structural levels with TEI-compatible markup while exploring cognitive-stylistic dimensions, the HIAM framework ensures seamless interoperability with digital scholarly editions. It supports creating richly annotated literary corpora that serve as empirical substrates for stylistic typology, authorial

fingerprinting, and translation studies [18; 32]. In an era where opaque neural language models increasingly dominate text analysis, this explicit, theoretically warranted architecture preserves vital analytical visibility and reproducibility.

CONCLUSION

This study developed and validated the Hierarchically Integrated Annotation Model for literary texts, grounded in systemic and cognitive-conceptual principles. The model unifies seven mutually constituting strata - lexical-morphological, syntactic, semantic, pragmatic, stylistic, discourse, and conceptual-cognitive - under five core operational principles.

Theoretical Value: Integrates international corpus linguistics, cognitive poetics, and digital humanities with the rich legacy of Uzbek philological scholarship into a single relational system.

Practical Value: Enables the construction of multi-layered literary corpora for advanced computational philology.

Empirical Robustness: A 1000-token pilot on 20th-century Uzbek prose (A. Qodiriy, Cho'lpon, G'. G'ulom, O. Yoqubov) yielded 90.9% accuracy and strong inter-annotator agreement ($\kappa = 0.85$), proving its cross-typological viability.

Future research will focus on standardizing conceptual metaphor tagsets, testing cross-linguistic extensibility, and integrating neural language technologies for semi-automated, scalable annotation. Reframed in this light, the linguistic annotation of literary text emerges not as a peripheral technical operation but as a substantive theoretical practice mediating linguistic description, computational processing and literary interpretation.

REFERENCES

1. Halliday M.A.K., Matthiessen C.M.I.M. Halliday's Introduction to Functional Grammar. 4th ed. – London: Routledge, 2014. – 808 p.
2. Van Dijk T.A. Discourse and Context: A Sociocognitive Approach. – Cambridge: Cambridge University Press, 2008. – 267 p.
3. Ide N., Pustejovsky J. (Eds.). Handbook of Linguistic Annotation. – Dordrecht: Springer, 2017. – 1295 p.
4. McEnery T., Hardie A. Corpus Linguistics: Method, Theory and Practice. – Cambridge: Cambridge

- University Press, 2012. – 294 p.
5. Sinclair J. Trust the Text: Language, Corpus and Discourse. – London: Routledge, 2004. – 224 p.
6. Lakoff G., Johnson M. Philosophy in the Flesh: The Embodied Mind and Its Challenge to Western Thought. – New York: Basic Books, 1999. – 624 p.
7. Langacker R.W. Cognitive Grammar: A Basic Introduction. – Oxford: Oxford University Press, 2008. – 562 p.
8. Burnard L. What is the Text Encoding Initiative? How to Add Intelligent Markup to Digital Resources. – Marseille: OpenEdition Press, 2014. – 142 p.
9. Leech G. Adding Linguistic Annotation. // Wynne M. (Ed.). Developing Linguistic Corpora: A Guide to Good Practice. – Oxford: Oxbow Books, 2005. – P. 17–29.
10. Jakobson R. Language in Literature. – Cambridge, MA: Harvard University Press, 1987. – 548 p.
11. Lotman Y.M. Universe of the Mind: A Semiotic Theory of Culture. – Bloomington: Indiana University Press, 1990. – 288 p.
12. Stockwell P. Cognitive Poetics: An Introduction. 2nd ed. – London: Routledge, 2020. – 222 p.
13. Fauconnier G., Turner M. The Way We Think: Conceptual Blending and the Mind's Hidden Complexities. – New York: Basic Books, 2002. – 464 p.
14. Pustejovsky J. The Generative Lexicon. – Cambridge, MA: MIT Press, 1995. – 312 p.
15. Semino E., Short M. Corpus Stylistics: Speech, Writing and Thought Presentation in a Corpus of English Writing. – London: Routledge, 2004. – 248 p.
16. Mahlberg M. Corpus Stylistics and Dickens's Fiction. – London: Routledge, 2013. – 232 p.
17. Mann W.C., Thompson S.A. Rhetorical Structure Theory: Toward a Functional Theory of Text Organization. // Text. 1988. Vol. 8. No. 3. P. 243–281.
18. Hoover D.L., Culpeper J., O'Halloran K. Digital Literary Studies: Corpus Approaches to Poetry, Prose, and Drama. – London: Routledge, 2014. – 272 p.
19. Kövecses Z. Metaphor: A Practical Introduction. 2nd ed. – Oxford: Oxford University Press, 2010. – 400 p.
20. McCarty W. Humanities Computing. – Basingstoke: Palgrave Macmillan, 2005. – 297 p.
21. Eco U. The Role of the Reader: Explorations in the Semiotics of Texts. – Bloomington: Indiana University Press, 1979. – 273 p.
22. Biber D., Conrad S. Register, Genre, and Style. 2nd ed. – Cambridge: Cambridge University Press, 2019. – 360 p.
23. Toolan M. Narrative Progression in the Short Story: A Corpus Stylistic Approach. – Amsterdam: John Benjamins, 2009. – 220 p.
24. Sperber D., Wilson D. Relevance: Communication and Cognition. 2nd ed. – Oxford: Blackwell, 1995. – 326 p.
25. Prasad R., Webber B., Lee A. Discourse Annotation in the PDTB: The Next Generation. // Proceedings of the 14th Joint ACL-ISO Workshop on Interoperable Semantic Annotation. – Santa Fe: ACL, 2018. – P. 87–97.
26. Gibbs R.W. Metaphor Wars: Conceptual Metaphors in Human Life. – Cambridge: Cambridge University Press, 2017. – 343 p.
27. Wynne M. (Ed.). Developing Linguistic Corpora: A Guide to Good Practice. – Oxford: Oxbow Books, 2005. – 92 p.
28. Steen G.J., Dorst A.G., Herrmann J.B., Kaal A.A., Krennmayr T., Pasma T. A Method for Linguistic Metaphor Identification: From MIP to MIPVU. – Amsterdam: John Benjamins, 2010. – 238 p.
29. Wales K. A Dictionary of Stylistics. 3rd ed. – London: Routledge, 2014. – 496 p.
30. Moretti F. Distant Reading. – London: Verso, 2013. – 244 p.
31. Jockers M.L. Macroanalysis: Digital Methods and Literary History. – Urbana: University of Illinois Press, 2013. – 224 p.
32. Underwood T. Distant Horizons: Digital Evidence and Literary Change. – Chicago: University of Chicago Press, 2019. – 240 p.
33. Mengliyev B.R. Computer Linguistics of the Uzbek Language: Theoretical Foundations and Applied

Issues. – Tashkent: Fan, 2020. – 224 p.

- 34.** Abdurahmonova N. Mathematical-Linguistic Model of the Morphological Analyser of the Uzbek Language: Doctor of Philological Sciences (DSc) Dissertation Abstract. – Tashkent, 2020. – 60 p.
- 35.** Hamroyeva Sh. Linguistic Foundations of the Authorial Corpus of the Uzbek Language: Doctor of Philosophy (PhD) Dissertation Abstract. – Tashkent, 2021. – 54 p.
- 36.** Cohen J. A Coefficient of Agreement for Nominal Scales. // Educational and Psychological Measurement. 1960. Vol. 20. No. 1. P. 37–46.
- 37.** Plungian V.A., Sichinava D.V. The Russian National Corpus: A New Stage of Development. // Voprosy yazykoznaniya. 2018. No. 6. P. 7–23.
- 38.** Gulyamova Sh. Linguomethodological Foundations of the Educational Corpus of the Uzbek Language: Doctor of Philosophy (PhD) in Pedagogical Sciences Dissertation Abstract. – Tashkent, 2022. – 52 p.
- 39.** Eshmo'minov A. Algorithms for the Automatic Morphological Analysis of Uzbek Language Texts: Doctor of Philosophy (PhD) Dissertation Abstract. – Tashkent, 2019. – 48 p.
- 40.** G'ulomov A., Tixonov A.N., Qo'ng'urov R. Morphemic Dictionary of the Uzbek Language. – Tashkent: O'qituvchi, 1977. – 464 p.
- 41.** Rahmatullayev Sh. Etymological Dictionary of the Uzbek Language. – Tashkent: Universitet, 2000. – 600 p.
- 42.** Hojiyev A. Explanatory Dictionary of Linguistic Terms. – Tashkent: O'zbekiston Milliy Ensiklopediyasi, 2002. – 168 p.
- 43.** Mahmudov N. Communicative-Pragmatic Investigation of Language. – Tashkent: Mumtoz so'z, 2017. – 320 p.
- 44.** Sarimsoqov B. Literary Devices. – Tashkent: O'qituvchi, 1988. – 224 p.
- 45.** Quronov D. Introduction to Literary Theory. – Tashkent: O'zbekiston, 2007. – 224 p.
- 46.** Karimov N. Cho'lpon: His Life and Creative Heritage. – Tashkent: Fan, 1991. – 240 p.
- 47.** Boboyev T. Foundations of Literary Theory. – Tashkent: O'zbekiston, 2002. – 560 p.