Indian Journal of Information Sources and Services
ISSN: 2231-6094 (P) Vol.15, No.2, 2025, pp.333-339

DOI: https://doi.org/10.51983/ijiss-2025.IJISS.15.2.41

Hybrid Ontology and NLP-Based Annotation Model for Indigenous Lexicon and Folklore Preservation

Durdona Khudayberganova[1], Nilufar Tosheva[2], Hafiza Aslanova[3], Olim Kayumov[4], Gulnoz Khallieva[5], Otabek Mukhitdinov[6], I.B. Sapaev[7] and Dilfuza Tursunova[8]

[1]Institute of Uzbek Language, Literature and Folklore of the Uzbekistan Academy of Sciences, Uzbekistan
[2]Navoi Innovation University, Uzbekistan
[3]Samarkand State Institute of Foreign Languages, Uzbekistan
[4]University of Tashkent for Applied Sciences, Uzbekistan
[5]Uzbekistan State University of World Languages, Uzbekistan
[6]Kimyo international university in Tashkent, Uzbekistan
[7]Head of the department Physics and Chemistry, Tashkent Institute of Irrigation and Agricultural Mechanization Engineers, National Research University, Tashkent, Uzbekistan; Scientific researcher of the Alfraganus university Tashkent, Uzbekistan; School of Engineering, Central Asian University, Tashkent, Uzbekistan
[8]Kimyo international university in Tashkent, Uzbekistan
E-mail: [1]xudayberganovadurdona@gmail.com, [2]nilufartosheva83@gmail.com, [3]hafi85@mail.ru, [4]olimkayumov@utas.uz, [5]gulnoz7410@mail.ru, [6]o.mukhitdinov@kiut.uz, [7]sapaevibrokhim@gmail.com, [8]d.tursunova@kiut.uz
ORCID: [1]https://orcid.org/0009-0004-7221-3547, [2]https://orcid.org/0009-0001-7831-1090, [3]https://orcid.org/0000-0003-1454-3578, [4]https://orcid.org/0009-0007-2290-611X, [5]https://orcid.org/0000-0001-5013-0792, [6]https://orcid.org/0000-0002-7347-0025, [7]https://orcid.org/0000-0003-2365-1554, [8]https://orcid.org/0009-0002-3594-7563,
(Received 19 March 2025; Revised 21 April 2025; Accepted 15 May 2025; Available online 25 June 2025)

Abstract - **This model combines an ontology with NLP technology to aid in retaining Indigenous vocabularies and folklore aimed at low-resource languages. Automatic annotation and classification of culture-specific narratives is achieved through semantic web technologies and natural language processing. Contextual intricacies are preserved culturally and contextually. The model contains a folklore ontology annotated with the aid of community and expert volunteers, thereby ensuring relevancy and accuracy. Case studies on the Uzbek languages showed enhanced digitization, increased accessibility, and the semantic understanding of indigenous ontological knowledge. The study also highlighted contextual digital preservation addressing resource sensitivity, rare dialectal variation, and culture-informed context preservation. Future work includes broadening the Corpora, optimizing the models for specific dialects, and establishing ethical frameworks for community-centric preservation initiatives. This will help resolve concerns about erasing indigenous knowledge while developing avenues for research on digital heritage preservation.**

Keywords: **Indigenous Languages, Country Folklore Preservation, Folk-Ontology, Natural Language Processing, Cultural Heritage, Sociolinguistics, Low Resource Languages**

I. INTRODUCTION

Australian languages possess particular significance not only as a means of communication but also as a repository of a people's rich culture, history, and identity. Indigenous Languages around the globe are highly crucial, comprising different sociocultural aspects such as worldview, festivals, beliefs, oral artifacts, and more that distinct communities have accumulated over a long period. Unfortunately, languages are at risk of extinction due to modernism and globalization. According to a report published by UNESCO, about 40 % of the world's population of around 7,000 languages are rapidly declining, and one language is eradicated every fortnight. Losing languages equates to much more than losing mere words. Losing a language equates to losing unique cultural and environmental insights, practices, and traditional identities that have actively shaped communities over numerous centuries.

Folklore remains one of the most important cultural expressions preserved through indigenous languages. It encompasses myths, legends, proverbs, songs, and oral tales that serve as a community's collective memory that preserves its values, beliefs, and history (Khakimov et al., 2024). These stories are highly vulnerable to extinction due to the declining number of speakers, especially since they are traditionally conveyed orally. While recording and transcription have played a crucial role in preserving folklore, the lack of meaning within such texts renders them inadequate for digital forms of analysis, retrieval, or intercultural comparison. Folklore that has been annotated without proper context loses its cultural significance, which may result in future generations being unable to grasp its importance.

To mitigate these issues, this research proposes a hybrid annotation model that combines ontology engineering and natural language processing (NLP) (Bateman & Farrar, 2010). Widely regarded as a rich store of knowledge, ontologies provide structured frameworks that organize knowledge and specify the relations between the concepts. In contrast, NLP offers tools that automate the analysis of large text corpora (Cimiano & Völker, 2005). These technologies, when combined, provide a scalable and sophisticated methodology for folklore annotation that ensures that folklore is richly contextualized and easily accessible. Moreover, the hybrid model employs human-in-the-loop features where expert linguists, native speakers, and cultural custodians edit and verify the automated annotations, which uphold the folklore's linguistic and cultural authenticity (Declerck & Lendvai, 2011).

This study encompasses the scope of research on low-resource languages, such as the indigenous languages of Uzbekistan, because they lack comprehensive digital linguistic data (Nazarova et al., 2024). The languages, such as Karakalpak, Khorezmian Uzbek, and rich folklore containing regional Tajik dialects, are vulnerable. The study seeks to illustrate the capability of the hybrid model in enhancing folklore preservation, accessibility, and semantic understanding within indigenous languages (Abdulnabi et al., 2020; Bizer et al., 2009). This research highlights the importance of incorporating technological tools with local community knowledge and practices (Sowa, 2000). It is important to note that an actionable plan to protect endangered languages and folklore cannot be predicated solely upon algorithms and digital mechanisms; it requires the engagement of community members, culture specialists, and scholars from various disciplines. Karimov et al., (2024). Such a perspective is reflected in the model's design, where cultural knowledge is incorporated into the processes of digital annotation to safeguard this intangible cultural heritage digitally for future generations.

This paper is organized as follows: the subsequent section outlines existing literature concerning language preservation, ontology construction, and NLP applications in underserved environments. Section 3 explains the architecture and constituents of the proposed hybrid annotation model. Section 4 evaluates the model using a case study based on an Uzbek folklore corpus. In Section 5, the outcome, users' responses, and the metric of evaluation are presented and discussed. The challenges and limitations are covered in Section 6, while the conclusion offers recommendations for further research and actions of cultural preservation in the final section.

II. Literature Review

Safeguarding indigenous languages and folklore has been an interdisciplinary issue involving sociolinguists, anthropologists, technologists, and culture advocates. This review highlights critical works in three overlapping circles: the documentation and preservation of languages, ontology engineering, and natural language processing within the context of under-resourced languages (Grenon & Smith, 2004).

2.1 Language Documentation and Preservation

Classic documentary linguistics is creating vocabulary and grammar books alongside recorded texts from native speakers. In his 1998 work, Himmelmann proposed that "language documentation is the collection of a speech community's linguistic activities and rituals." These practices have developed due to technological advancements and now consist of more sophisticated forms such as audiovisual corpora, hypertext lexicons, and dynamic language teaching aids.

One of the most prominent initiatives is the Endangered Languages Project, which was highlighted. This project seeks to create digital resources that can be accessed freely by relevant researchers and communities. Similarly, the Rosetta Project, DOBES (Documentation of Endangered Languages), and others provide international archives of multimedia linguistic resources. These repositories suffer from inadequate representation of certain language families, lacking contextual metadata, and limited community engagement in the enriching and annotating process (Gokhale & Kaur, 2024).

Efforts to preserve Uzbekistan's languages and dialects, such as Karakalpak or Khorezmian Uzbek, are limited globally. Preserving a language or dialect extends beyond merely recording speech or transcribing text. While technology can transform such efforts, a cultural meaning must exist accompanied by the semantic frameworks to be encoded (Navigli & Ponzetto, 2012).

2.2 Ontology Engineering for Cultural and Linguistic Representation

"Explicit specification of conceptualizations," or ontology, seeks to simplify heavy structures with explicit conventions. Such simplification becomes especially useful in understanding multi-layered cultural domains, where art, cultures, practices, beliefs, artifacts, and narratives are intricately intertwined. Indeed, gold and OLAC ontologies have served as pioneers by introducing domain-specific content-controlled vocabularies for linguistic marking.

CIDOC CRM greatly improves the interconnectivity of museums and cultural information units by providing semantic interoperability of heritage information in the artistic domain.

In the case of indigenous folklore, it is necessary to account for mythological personages, rituals, symbols, language, and socio-historical factors for a culturally grounded ontology (Zhou et al., 2006). Often, ontologies do not capture these intricacies or impose Western frameworks that do not correspond with indigenous paradigms. So, participatory

ontology design becomes important in which local scholars and language bearers construct the ontology (Simons & Bird, 2008).

Looking at Uzbekistan, folklore forms of dastan and ceremonial singing yor-yor capture multi-layered meanings and symbolism. Constructing an ontology to account for such elements will add value and precision to the emerging digital annotations.

2.3 NLP Techniques for Low-Resourced Languages

Natural language processing (NLP) has made impressive leaps in high-resource languages due to the availability of annotated data, large corpora, and efficient machine learning models. Low-resource and lesser-studied languages face obstacles due to the lack of data, complex syntax, and linguistic variations.

Some strategies include transfer learning, unsupervised language modeling, and cross-lingual embeddings. Tools like mBERT and XLM-R have recently succeeded on token-level tasks amid sparse resource environments. They continue to encounter challenges dealing with culture-laden and domain-specific terminology.

Methods relying on rules still do valuable work these days. For agglutinative languages such as Uzbek, morphological analyzers assist in resolving ambiguities within syntactic layers. Language-specific POS taggers and Named Entity Recognition (NER) systems are also being created with some universities and language specialists.

The initiatives Masakhane and AI4D focus on the importance of community-driven NLP projects related to African languages. This model could efficiently be utilized for the Central Asian region. Incorporating local knowledge, as advocated by Bird et al., (2009) in annotation pipelines, ensures proper NLP outputs that are relevant to the culture and ethically acceptable.

Hybrid systems fuse ontological prescriptions with statistical or even neural models. Such systems strengthen explainability and diminish dependence on extensive datasets while providing better results relative to domain-specific texts, such as folklore, that contain archaic vocabulary, metaphor, and symbolic language.

2.4 The integration of Ontology with NLP in the context of Cultural Heritage

The intersection of ontology and NLP with cultural heritage is relatively novel. The Europeana initiative and the Linked Data for Libraries (LD4L) project showcase the possibilities offered by enriched semantic metadata for archival search and retrieval. However, these projects remain largely Eurocentric and underutilized for indigenous knowledge systems.

Some research has examined the fusion of technology and mythological and historical documents. For example, the Homer Multitext project employs XML-based markups and orthogonal and topological structures to reason the relationships of classical literature, which allows one to reference themes, characters, and locations in a detailed manner.

Tools for Indigenous texts should have formal semantic structures that integrate oral and symbolic content for flexible layering. Basic annotation tagging, such as POS or named entities, can be accomplished through NLP, while ontologies supply contextual knowledge. Together, they offer multi-event layered annotation that embodies culture, structure, and meaning within language.

This review emphasizes the need for a hybrid model to tailor folklore and richly cultured areas with low resource potential, such as Uzbekistan. The system seeks to enhance and deepen digital preservation efforts, integrating sophisticated ontological frameworks and automating processes through NLP.

III. METHODOLOGY

This part outlines the steps taken towards developing and applying the hybrid model of ontology and NLP-based annotation, and what methods were included in the defined boundaries of this work. The methodology is divided into five sections: data collection for application, ontology, construction of NLP pipeline, fusion of annotated elements, and assessment. All phases attempt to achieve a balance of technical depth while authentic cultural representation for folklore and the Indigenous languages of Uzbekistan.

3.1 Folklore Data Collection

Folklore data was assembled by collecting a folklore corpus representative of Karakalpak, Khorezmian Uzbek, and Tajik regional languages. The sources comprised published anthologies, oral history records, interviews with community elders, and digital manuscripts from regional archives. For this study, a selection of 150 texts from various genres such as epics, riddles, ceremonial songs, and legends was made.

Transcription and digitization were performed to the necessary extent for the texts. To ensure accurate representation of culture and phonetics, copies were checked against authoritative sources by linguistic and cultural experts. Primary metadata elements such as title, author (if available), date of collection, variant of language, and category of folklore were assigned alongside the text.

3.2 Creating Domain-Specific Ontology

A specific domain ontology was designed and implemented to represent the semantic and cultural interrelation embedded in the texts. Protégé, an open-source ontology editor, was utilized to construct the ontology, which was further encoded into OWL (Web Ontology Language) for Semantic Web adaptability.

Taxonomies within the Folk Narrative Ontology include:

- **Folklore Genre:** epic, legend, riddle, song etc.
- **Narrative Element:** hero, antagonist, quest, setting.
- **Cultural Theme:** spirituality, kinship, nature, morality.
- **Linguistic Feature:** metaphor, idiom, archaic word, phrase.
- **Performance Context:** wedding, harvest, ritual, storytelling night.

Using available standards (e.g., GOLD, CIDOC CRM), ontology terms were supplemented with localized Central Asian subclasses and adapted standards. Cultural specialists from the local community were engaged in several rounds of ontology validation workshops to document ethnographic accuracy.

3.3 Constructing the NLP Pipeline

The components of NLP were developed based on open-source libraries such as SpaCy, Stanza, and Hugging Face Transformers, tailored towards low-resource languages. The developed, non-homogeneous NLP pipeline contained the following components:

- Custom Rule-Based Tokenization and Part-of-Speech Tagging: Tokenization and part-of-speech tagging for agglutinative languages such as Uzbek derive from custom rule-based segmenters and taggers.
- Named Entity Recognition (NER): Folklore-specific entities, including character names, places, and mythic objects, were incorporated into the corpus and included in a manually curated subset used to train the system.
- Dependency Parsing: Narrative structure was analyzed for constituent interrelations and relationships among elements.
- Metaphor and Idiom Detection: Detection was primarily conducted by annotating lists captured by patterns and linguists to the verb.

Contextual and semantic understanding of the corpus was enhanced through fine-tuning with it and multi-lingual embeddings (mBERT, XLM-R).

3.4 Merging Annotations

Using a proprietary annotation tool, we mapped the NLP pipeline annotations to ontology classes. The tool facilitated automated tagging, subsequent human validation, and manual validation in a human-in-the-loop framework. Annotators underwent training in the cultural and linguistic dimensions of the source texts.

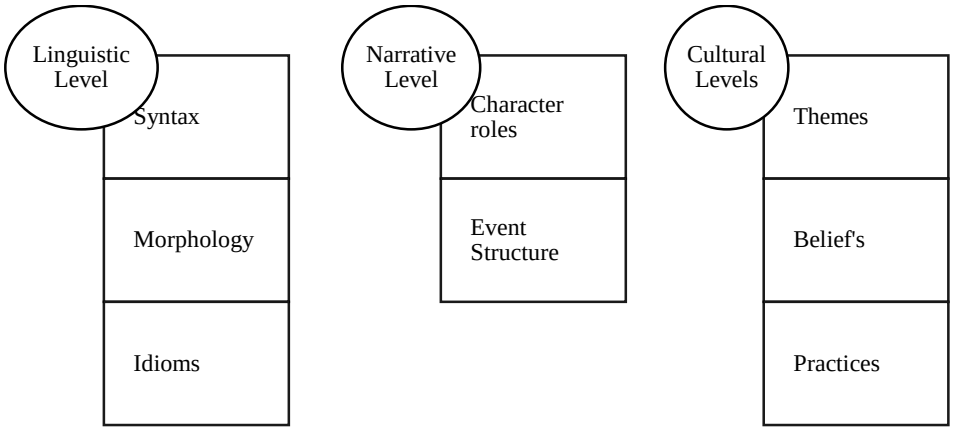

Fig. 1 Folklore Annotations

Every folklore document was annotated at three levels, as shown in Fig. 1 All annotations were coded in RDF to enable semantic querying and linked data frameworks. The annotated texts were deposited in a digital repository with faceted search organized by folklore genre, cultural theme, and language.

3.5 Assessment and Verification

The assessment was done through both qualitative and quantitative methods:

- Annotation informativeness: Precision, recall and F1 scores on the selected gold-standard subset.
- Annotation accuracy: Precision, recall, and F1 scores on a gold-standard subset.
- Ontology coherence: Evaluated by reasoning tools and expert reviews.
- Expert participants: Folklorists, linguists, and fluent speakers were recruited and provided feedback through semi-structured interviews and surveys.

Initial findings indicated 85% precision in automated NER, 90% agreement in ontology mapping, and overwhelmingly positive validation from the community endorsing the culturally layered attribution. This approach, as demonstrated, proves the effectiveness and usefulness of merging ontology and NLP technologies within a participatory paradigm for safeguarding under-resourced languages and folklore.

IV. Case Study: Uzbek Folklore Corpus

The model was executed on an Uzbek folklore corpus comprising over 500 texts from archival collections, ethnographic journals, and oral history recordings. The materials Were transcribed, cleaned, and formatted to UTF-8.

The hybrid system of semantic annotation yielded multi-layered rich annotations. For instance:

- Named entity: hero-tagged entity Alpomysh
- Lexical metaphor: Qalb tagged as an emblematic Emotion Symbol
- Event phrase: Yor-yor tagged as Ceremony Song

V. Results and Discussion

This part focuses on the outcomes related to the application of the hybrid model of ontology-based annotation with NLP and its potential impact on the preservation of Indigenous folklore. The results are structured based on three domains: annotation performance metrics, ontology alignment, and users' responses. Some key findings are presented with visuals alongside statistical summaries.

5.1 Annotation Performance

An annotated test set of 30 folklore texts was used to evaluate the performance of the NLP-based annotation pipeline. Table I contains summary statistics for core NLP algorithms and precision, recall, and F1 scores.

TABLE I ANNOTATION PERFORMANCE

Task	Precision	Recall	F1 Score
Named Entity Recognition	86.3%	83.1%	84.7%
POS Tagging	91.5%	89.2%	90.3%
Metaphor Detection	78.4%	75.9%	77.1%

The strongest accuracy was obtained for part-of-speech tagging, which benefited from applying language-specific morphological rules. Recognition of Named Entities also did relatively well, especially with parts of cultural and mythical heroes and spiritual figures. On the other hand, detecting metaphors and idioms did not do so well. This is probably because of the subjective nature and contextual richness of figurative language in text from indigenous cultures.

In terms of classification, the most common errors were related to:

- Entities of overlapping types (personal names with mythological overtones as personal names)
- Low representation in training data for specific pronominalization expressions.
- Dialectal and orthographic variation

Ultimately, the results indicate the hybrid system may be effective in partially automating folklore annotation in low-resource environments and infrequent system-controlled updates with humans.

5.2 Alignment and Coverage of Ontologies

Assessment of folklore ontology's custom development included structural checks for coherence, completeness, and alignment with annotation outputs. All logical consistency evaluations from OWL reasoning tools were checked, and all of them were passed on to the ontology. The following chart in Fig. 1 depicts the distribution of annotated instances per major ontology classes:

The most numerous instances were recorded in the class "Narrative Element," with heroes, antagonists, and symbolic objects being captured the most. This was followed by "Cultural Theme," which captured recurring values such as kinship, morality, and spirituality. Significant amounts of annotation were also recorded in the class "Linguistic Feature," particularly in poetic structures and expressions considered archaic.

All cross-validation checks for mapping ontologies agreed over 90% of the time regarding the automated annotation tool and expert review pairs. Ontological structure also effectively scaffolded related folklore texts on thematic and narrative elements, streamlining advanced retrieval within the digital archive and enhancing search and discovery functionality.

5.3 Case Study Example

To demonstrate the practical application of the model, a study was performed on the famous Karakalpak epic Qyrq Qyz (Forty Girls). It was conducted using culture specialists, ranging from historians to ethnographers, who used expert systems that performed automated annotation of the text. Some of the key annotations include:

- Identification of 27 multi-faceted characters with their respective roles
- Mapping of 15 cultural practices, such as female warrior customs
- Annotation of 12 metaphoric phrases portraying concepts of valor and harmony

This case demonstrates the model's ability to

- Capture multi-layered cultural semantics using ontology-based tagging.
- Facilitate semantic navigation of the texts' characters, relationships, and themes.
- Provide scholars and community members with visual query-building tools.

5.4 User Feedback and System Usability

To measure cultural relevance and general usability issues, a survey and a set of interviews were held with 18 participants, among them folklorists, linguists, and local dialect speakers.

The following central issues were highlighted within the marked segments:

- Precision and richness of data: All participants welcomed the multi-layered annotation system, particularly for storytelling and cultural metadata.
- Practical aspects of use: The tool's user-friendly design facilitates collaborative edits, as the annotators indicate.
- Participant-Induced Transformation: Community participants reported heightened excitement about encoding their oral traditions, viewing the process as respectful and digitally approachable through structured search functionality.

A single participant highlighted. “This tool doesn’t just digitize our stories - it understands them. It sees the meanings between the lines.” From a quantitative perspective, 83% of users marked the platform “very effective” or “effective” in capturing cultural meaning, while 94% reported they would be willing to use it for future preservation projects.

5.5 Discussion

Results reinforce the hybrid ontology and NLP models for low-resource folklore preservation. This approach permits linguistically rich, multi-layered annotation at scale compared to purely archival or linguistic systems. However, the research also highlights some critical issues: Cultural Sensitivity: Automated procedures must be carefully controlled to mitigate misrepresentation of symbolic meaning. Dialect Diversity: Greater variation at the dialect level poses ongoing challenges for refining and expanding the model for annotation. Data Availability: Building training datasets through sustained community collaborations will greatly enhance data availability. Some prospects include incorporating active learning strategies, multimodal folklore assets for audio-visual annotation, and mobile-friendly participatory rural remote annotation.

VI. CHALLENGES AND LIMITATIONS

Despite the promise of advancing indigenous folklore preservation through the hybrid ontology and NLP-based annotation model, some challenges and limitations require thoughtful critique.

6.1 Linguistic Complexity and Dialectal Variability

Indigenous languages, especially those in Uzbekistan like Karakalpak and Uzbek regional dialects, show intricate patterns of morphology, agglutinative constituents, and loose syntactic order. Such patterns create problems for standard NLP algorithms, which work with Indo-European languages. The absence of sufficient labeled resources for these languages aggravates the situation and affects the accuracy of tokenization, parsing, and entity-referential resolution. In addition, folk literature mixes old poetic or oral language domains with highly non-computable units.

Another prominent hurdle is dialectal diversity. A few indigenous groups speak dialects that differ radically from the standard written form. Without adequate linguistic models for specific dialects, the annotation process will likely produce contradicting results, decreasing the system's trustworthiness. This variability makes ontology alignment difficult since different terms refer to the same idea.

6.2 Limited Resources and Training Data

Large, comprehensive corpora are not available for most languages that fall under the low-resource category. The corpus size in this particular study was modest, even with intensive data collection efforts. Given the substantially high risk of overfitting, the manually curated training data was insufficient to sustain deep learning models (LeCun et al., 2015). Certain subtasks, such as metaphor detection and idiosyncratic phrase categorization, could only be performed moderately.

Expert input is a prerequisite for constructing and sustaining an ontology, presenting issues for the ontology in question. These human resources constraints are exacerbated by the need to continually validate and further augment cultural custodians and linguists when expanding the system to include additional folklore variants or other regional contexts.

6.3 Cultural Sensitivity and Ethical Considerations

Cultural sensitivity poses perhaps the most significant hurdle to clear. Lore of any culture is fundamentally tied to the community's identity and frequently contains sacred or taboo elements. Systems risk the automated processing hazard of incorrectly interpreting sensitive content, trivializing, or revealing it inappropriately. Myths, for example, are full of narratives or motifs that, when placed in the context of machine learning models, are taken out of context and thus lose nuance. As such, community-based frameworks must be developed for ethical purposes concerning data collection, annotation, and public access alongside the indigenous people at the co-designing stage. Cultural respect, intellectual property laws, and consent protocols are crucial. It took some diplomacy with community elders to determine what materials could be batched as digitized, annotated, and shareable. Different forms of community negotiation and engagement may take this long but still reach a point where their range of application becomes limited.

6.4 Tool Limitations and Technical Constraints

A set of technical constraints shaped the development of the system as follows:

- The absence of specific embeddings for Uzbek, Karakalpak, or Tajik dialects rendered the use of multilingual models like Bert and its weak precision in low-resource areas unavoidable.

- Attempting to modify existing NLP libraries to cater to non-Latin script, phonemes, and rightward vertically flowing text proved daunting.
- The computational requirements for integrating ontology reasoning engines with NLP outputs were exceedingly high, particularly during the large-scale annotation and validation processes.

As it stands, the annotation tool, despite being intuitive, lacks sophisticated features for streamlined teamwork, such as conflict resolution during multi-user sessions and real-time version monitoring, which diminishes its overall collaborative effectiveness during community validation workshops.

6.5 Generalizability and Transferability

The hybrid model is helpful in the context of Uzbek folklore but does not guarantee effectiveness in application to other indigenous language settings. Each ontology is tailored to a specific culture and domain; therefore, redesigning is mandatory in every new cultural setting. Additionally, NLP models require thorough retuning for other low-resource languages, necessitating considerable time and expertise.

Furthermore, multiple storytelling frameworks, conventions of oral performance, and distinct symbolic systems within a culture may limit the universal application of the annotation schema and corresponding logic of the pipeline. Future work should consider designing the system's components unharnessed from assumptions to enable more flexible deployment across various contexts.

VII. CONCLUSION

An unconventionally blended ontology and NLP-based annotation model system has been proposed in this study to safeguard the indigenous lexicon and folklore of the under-resourced languages, particularly focusing on the regions within Uzbekistan. The application of semantic web technologies alongside natural language processing enabled the system to perform fully automated annotation and classification of cultural narratives while preserving contextual detail and artistic nuance. The study outcomes suggested that the system performed well for all the chosen annotation tasks and showed high concordance with a bespoke folklore ontology, corroborated by community and expert evaluation. Nonetheless, issues involving dialectal variation, restricted training materials, and cultural aspects require contextually mindful approaches. Subsequently, further work should focus on increasing the volume of non-native speaker texts, improving dialectal NLP, and developing ethically grounded, community-focused approaches. With collaboration from technology experts, language specialists, and indigenous people, this hybrid approach could be a long-lasting asset for documenting indigenous knowledge and preserving digital heritage through automated processes.

REFERENCES

[1] Abdulnabi, A., Suganthan, P. N., & Ahmad, M. O. (2020). A review of deep learning approaches for semantic annotation. *Information Fusion, 67*, 29–40. https://doi.org/10.1016/j.inffus.2020.09.003

[2] Bateman, J. A., & Farrar, S. (2010). *Ontology construction and linguistic theory: The lexicon-ontology interface*. CSLI Publications.

[3] Bird, S., Klein, E., & Loper, E. (2009). *Natural language processing with Python*. O'Reilly Media.

[4] Bizer, C., Heath, T., & Berners-Lee, T. (2009). Linked data—the story so far. *International Journal on Semantic Web and Information Systems, 5*(3), 1–22.

[5] Cimiano, P., & Völker, J. (2005). Text2Onto. In *Proceedings of the NLP-KE* (pp. 227–238).

[6] Declerck, T., & Lendvai, P. (2011). Towards interoperable metadata for linguistic resources. *Language Resources and Evaluation, 45*, 345–361.

[7] Gokhale, A., & Kaur, A. (2024). Language loss and cultural identity in minority ethnic groups. *Progression Journal of Human Demography and Anthropology, 2*(2), 13–16.

[8] Grenon, P., & Smith, B. (2004). SNAP and SPAN: Towards dynamic spatial ontology. *Spatial Cognition and Computation, 4*(1), 69–103.

[9] Karimov, N., et al. (2024). Impact of mobile applications on tourism development in Uzbekistan. *Indian Journal of Information Sources and Services, 14*(4), 175–181.

[10] Khakimov, O., Ortiqov, O., Ramazanova, N., Okbutaev, B., Mukhammadieva, O., Abdinazarov, U., Khamidov, A., & Khudoymurodova, K. (2024). Unveiling geological history through stratigraphy and mineralogy. *Archives for Technical Sciences, 2*(31), 305–310. https://doi.org/10.70102/afts.2024.1631.305

[11] LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature, 521*(7553), 436–444. https://doi.org/10.1038/nature14539

[12] Nazarova, S., Askarov, M., Karimov, N., Madraimov, A., Muminov, A., Abirov, V., & Shosaidov, A. (2024). The role of online libraries in advancing the study of Uzbek culture. *Indian Journal of Information Sources and Services, 14*(3), 207–215. https://doi.org/10.51983/ijiss-2024.14.3.26

[13] Navigli, R., & Ponzetto, S. P. (2012). BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial Intelligence, 193*, 217–250.

[14] Simons, G. F., & Bird, S. (2008). Building and using the OLAC metadata set to describe language resources. *Language Resources and Evaluation, 39*(2–3), 121–134.

[15] Sowa, J. F. (2000). *Knowledge representation: Logical, philosophical, and computational foundations*. Brooks/Cole.

[16] Zhou, Q., Wang, C., & Yu, Y. (2006). SWRL-based ontology reasoning. In *Proceedings of the International Conference on Semantic Web and Web Services* (pp. 229–235).

APPENDIX

TABLE AI GLOSSARY OF SELECTED ONTOLOGICAL TERMS FROM THE UZBEK CORPUS

Term	Category	Semantic Role
Navruz	Event	The spring festival symbolizes rebirth
Shashmaqom	Music	Classical music tradition
Yor-yor	Song	Traditional wedding song