Enhancing Data Authenticity: Leveraging Humanities Annotation Practices for NLP

Urmishree Bedamatta
Department of English
Ravenshaw University

ABSTRACT

This paper explores the potential of applying textual criticism practices, traditionally a core aspect of humanities research, to enhance the authenticity and interpretability of linguistic data for Natural Language Processing (NLP) applications. By proposing a multi-layered annotation model, this work argues that annotations extending beyond syntactic and semantic labels, encompassing historical, cultural, and rhetorical contexts, can provide NLP systems with a deeper, context-aware understanding of language. Drawing on examples from the digital edition of the Odia Mahabharata, the paper illustrates how annotations that capture word evolution, cultural nuances, and stylistic choices can mitigate challenges in transcription, while preserving the authenticity of texts. The paper further demonstrates how such annotation practices enable NLP systems to address linguistic subtleties such as ambiguity, irony, and sentiment, making them more effective for complex tasks like machine translation, sentiment analysis, and content generation. Ultimately, this study argues that integrating humanities-driven annotation practices into NLP can not only improve the quality of computational models but also ensure the preservation and accessibility of culturally and historically significant language forms.

General Terms

Natural Language Processing

Keywords

Textual criticism, Multi-layered annotation, Odia Mahabharata, Natural language processing, Digital humanities

1. INTRODUCTION

The concept of ‘authentic data’ is multifaceted, particularly in linguistics and computational fields, where it refers to data that is contextually accurate, true to real-life use, and often gathered directly from human interactions or natural environments. In contrast to synthetic or constructed data, which may be artificially generated or lack real-world grounding, authentic data retains its integrity and credibility by reflecting the linguistic, cultural, and social contexts from which it originates.

In linguistic theory, authentic data is closely tied to the notion of real-life language use. According to Labov (1972), authentic data in sociolinguistics refers to the language produced in natural, unscripted environments, such as casual conversation or oral traditions, which provides insight into the variations and dynamics of language in its everyday context. For Labov, the value of authentic data lies in its ability to reveal the subtleties of dialects, sociolects, and community-specific linguistic forms that might be overlooked in more formal or controlled settings. Kress & van Leeuwen (2001) extend this understanding of authentic data into multimodal analysis, where authentic data includes not just spoken language but also visual and cultural expressions in texts, advertisements, and media. Authentic data, for them, is not limited to the spoken or written word, but encompasses all semiotic resources that contribute to meaning-making in a given context.

In computational linguistics and Natural Language Processing (NLP), authentic data is essential for building accurate models of language understanding, translation, and generation. According to Bender (2019), authentic data refers to text or speech corpora that are collected from real-world sources, such as news articles, social media, books, or recorded speech, rather than being artificially constructed or simplified for algorithmic use. Authentic data preserves the complexity, diversity, and unpredictability of natural language, which is crucial for training machine learning models that can generalize well in real-world applications. In a similar vein, Bird, Klein, & Loper (2009) emphasise that authentic data in NLP is indispensable for tasks like part-of-speech tagging, named entity recognition, and syntactic parsing. Real-world linguistic data provides the varied contexts needed for models to learn the intricacies of language use, including rare words, idiomatic expressions, and cross-cultural variations that may not be represented in artificially curated datasets. In the context of machine learning, the definition of authentic data extends to include data that accurately represents the phenomenon being modeled. As Charniak (1993) suggests, authentic data for machine learning tasks, such as speech recognition or machine translation, must be sourced from actual human interactions, ensuring that it reflects the language’s idiosyncrasies, including informal language, cultural references, and situational variations.

Blodgett et al. (2020) discuss the importance of authentic data in ensuring that AI models do not propagate harmful stereotypes or biases. The use of real-world data ensures that the AI system can better understand and replicate human-like reasoning and interactions without relying on biased or skewed datasets that might distort the reality of human language. The notion of authenticity in data derived from the internet has gained prominence in the digital age, particularly as vast amounts of information become increasingly accessible for research and applications such as NLP. While the internet provides a broad and diverse pool of data, the authenticity of such data must be critically examined, as its provenance, quality, and contextual relevance often vary. Tufekci (2014) argues that the reliance on internet-sourced data, a primary source of authentic data for many NLP tasks, can often lead to issues like bias, misinformation, and ethical concerns. Data scraped from the web may reflect the dominant voices in society, marginalizing minority dialects, older speakers, and less widely spoken languages. This is particularly problematic for languages which are low-resource in the digital domain, like Odia, where authentic data may be scarce, leading to underrepresentation in computational models.

2. AUTHENTICITY BEYOND SOURCE: THE ROLE OF ANNOTATION

Authenticity in data is often assessed by its source, whether it originates from credible, verifiable, or authoritative origins. However, this perspective overlooks a crucial dimension: the quality and depth of the annotation accompanying the data. Data's authenticity should not only be determined by its provenance but also by the extent to which its context, structure, and interpretative layers are systematically and rigorously annotated.

2.1 Annotation as a Measure of Authenticity

While traditional linguistic corpora, such as digitized texts, are an essential resource for preserving authentic data, Müller (2016) advocates for detailed contextual metadata, while calling for comprehensive data collection strategies that incorporate both written and spoken forms. Ide and Pustejovsky (2017) argue that annotations are not mere metadata; they embody a secondary layer of meaning-making that is essential for high-quality data. They emphasize that annotations imbue raw data with linguistic, contextual, and domain-specific knowledge, transforming it into a resource that is ready for computational analysis. Similarly, Bird and Liberman (2001) argue that the depth and rigor of annotation directly influence the data's reusability across various NLP tasks. In the context of historical linguistics or cultural studies, annotations addressing the grammatical, lexical, stylistic, and discourse-level aspects of data ensure that the nuances of language use and cultural expression are preserved and made accessible.

2.1.1 The Need for Enhanced Annotation in NLP

Traditional NLP annotation often focuses on labeling linguistic features such as part-of-speech tags, named entities, syntactic dependencies, and sentiment. While these annotations are critical for machine learning tasks, they often lack the depth of meaning and context seen in humanities-oriented annotations. Many NLP models rely on shallow annotations that miss out on complex layers such as historical context, cultural nuances, literary devices, or interpretative layers that are integral in understanding text deeply, as done in the humanities.

In the humanities, annotation is not just about labeling surface-level linguistic features, but also about adding layers of interpretation that encompass historical, cultural, philosophical, and socio-political contexts. For example, annotating historical word forms and their evolution can help NLP systems understand diachronic changes in language use. Literary analysis often highlights stylistic features such as tone, metaphor, irony, or symbolism, which can be essential for tasks like sentiment analysis or literary text classification. Humanities annotations also explore discourse-level features, such as narrative perspective, character arcs, or rhetorical devices. These can provide NLP systems with deeper understanding of text structure and meaning.

2.1.2 Enriching Data Annotation for NLP: Lessons from the Humanities

Projects such as *The William Blake Archive* or the *Perseus Digital Library* integrate humanities-style annotations for deep textual analysis. These projects highlight the potential of combining rich, multi-layered annotations with digital technologies to enhance NLP models. In the Bodleian Libraries' Digital Manuscripts project, scholars annotate medieval texts with insights about historical spelling, word usage, and meaning shifts, which can inform NLP models for historical language processing.

In this paper, I argue that the principles and practices of textual criticism can serve as a valuable framework for improving annotation quality in NLP applications. To support this claim, I draw attention to the recently released digital edition of the Odia Sarala Mahabharata and compare it to similar critical editions to highlight its unique contributions as an ideal annotation model.

The digital edition exemplifies a robust and multi-layered approach to text representation, organizing each chapter into the following sub-parts: a) the Odia critical text, presenting the edited version based on rigorous scholarly evaluation; b) a collection of textual variants for each line or verse, capturing alternative readings from different manuscripts; c) notes that provide justification for the selection of specific variants, supported by historical and linguistic evidence; and d) an English translation of the critical text, making the edition accessible to a broader audience, and e) cultural index.

The critical notes are particularly noteworthy, as they document not only the historical forms of words but also their etymological development, contextual usage, and modern equivalents. This level of detailed annotation parallels the layered analysis required in NLP, where understanding diachronic language change, regional variations, and contextual nuances is crucial for accurate computational models.

Similar editions, such as the critical edition of the Mahabharata by the Bhandarkar Oriental Research Institute or the *Chaucer Variorum Edition*, adopt comparable multi-layered methodologies. These editions annotate their texts with apparatuses that include variant readings, philological commentary, and linguistic analysis, ensuring that each interpretation is grounded in textual evidence. By capturing such granular details, these critical editions provide exemplary models for annotating texts in NLP. They bridge linguistic forms across time and space, offering rich, structured data that can be directly used to train systems in historical linguistics, lexical semantics, and translation.

The digital edition of the Odia Sarala Mahabharata demonstrates how the praxis of textual criticism aligns closely with the needs of NLP annotation. It emphasizes transparency in editorial decisions, contextual richness in variant readings, and accessibility for interdisciplinary use, making it a compelling annotation model for computational linguistics. This approach can guide the creation of annotated corpora that meet both the depth required in the humanities and the precision demanded by NLP applications.

2.1.3 A Sample of Multi-Layered Annotation in the Digital Sarala Mahabharata

Preserving Linguistic Nuance: The example of the Odia word ଭୟେଣୀ (bhayeṇī) illustrates the importance of a multi-layered annotation model in NLP to address issues arising from linguistic and cultural assumptions. In one version of the text, the word appeared as ଭୟେଣ (bhayeṇa), which seemed consistent with the morpheme patterns prevalent in the text. Based on this observation, the editors revised the text to align with the assumed linguistic logic. However, this editorial choice inadvertently overlooked the cultural and contextual relevance of the original word. Upon further examination, it was revealed that ଭୟେଣୀ (bhayeṇī) means "sister" and fits seamlessly within the context of the verse, aligning with its intended meaning.

This case highlights the risks of erasing culturally vibrant forms of expression through assumptions that prioritize modern linguistic norms over historical and colloquial usage. Such editorial decisions not only disrupt the fidelity of the text but also diminish the rich cultural and historical value embedded in its language. A multi-layered annotation model addresses this by incorporating the following layers: a) Historical context: Annotating words like ଭୟେଣୀ (bhayeṇī) with their historical and etymological significance ensures that these forms are not misinterpreted or lost. This includes documentation of their evolution, usage patterns, and meanings in varying contexts; b) Cultural nuance: Words such as ଭୟେଣୀ (bhayeṇī) often carry cultural meanings that extend beyond their dictionary definitions. Annotation capturing such cultural significance helps NLP systems to interpret texts within their sociocultural frameworks; c) Contextual coherence: By annotating the intended meaning and situational context of words, the model ensures alignment with the text's overall narrative or discourse, preventing misinterpretations like the editorial assumption that altered the verse; d) Colloquial and Stylistic Features: Colloquial forms like ଭୟେଣୀ (bhayeṇī) often convey stylistic richness. Annotation of these forms preserves their expressive depth and highlights their relevance compared to their modern equivalents.

Annotating Emotional States: In another instance, an editorial choice was to be made among terms like "କଠୋର ବ୍ରତଭାବ" (rigid ascetic disposition), "ନିଷ୍ଠାପର ଭାବ" (devotional resolve), "ନିଷ୍ଠୁର ଭାବ" (harsh disposition), and "ବୈଷ୍ଣବ ଭାବ" (Vaishnava disposition) to describe Shantanu's emotional state in distancing himself from his wife, Ganga. While each term captures an aspect of Shantanu's emotional state, "ବୈଷ୍ଣବ ଭାବ" was selected for its specificity within the Vaishnava tradition. The editorial decision was rooted in its specific connotations within the Vaishnava tradition, particularly its association with the Udasina sect. This term not only reflects Shantanu's renunciation of worldly ties but also resonates with the characteristics of the Udasina sect of Vaishnavism, known for their melancholy, detachment, and spiritual rigor. Contextual annotations underscore how Shantanu's observance of the Ekadashi ritual, a fasting day dedicated to Vishnu, further aligns with the themes of self-restraint and spiritual dedication encapsulated in "ବୈଷ୍ଣବ ଭାବ."

Annotating Semantic Nuances: Misinterpretations in manuscripts often arise from similarities in letter shapes, leading to transcription errors and incorrect word boundaries. Such challenges necessitate a careful, layered approach to annotation, which combines paleographic analysis and contextual interpretation to preserve textual fidelity. One example of this issue is the misreading of "ପୁରୁର ବାମନ୍ଦନ" instead of "ପୁରୁରବା ନନ୍ଦନ." This error likely resulted from the visual resemblance between the letters ମ (ma) and ନ (na). By visualizing similar letter shapes, the editors reconstructed the probable cause of the transcription error. The editors also explain cases of incorrect segmentation while correcting the words.

Spelling errors further complicate manuscript interpretations. For instance, the distinction between ଭାଗିରଥି (Bhāgirathi), typically a male name, and ଭାଗିରଥୀ (Bhāgirathī), the goddess Ganga, is crucial for maintaining semantic accuracy. Similarly, nuances between terms such as କେଳି (Keḷi) and କେଳୀ (Keḷī) or ବାଉନି (Bāuni, to disapprove) and ବାହୁନି (Bāhuni, to bewail) may go unnoticed without sufficient familiarity with regional idioms and linguistic contexts. To address these ambiguities, we carefully annotated terms that might otherwise introduce confusion. For example, Ganga's father's name was standardized as ନୀର୍ଘାତ (Nīrghāta) to distinguish it from ନିର୍ଘାତ (Nirghāta), which could mean "force," "ferocious," or "roaring sound." While most manuscript versions used the second spelling, the adjustment ensures clarity by avoiding unintended dual meanings.

When the principal witness diverged significantly from other witnesses, we retained these unique forms, providing relevant synonyms in our notes to support interpretive consistency. For words absent from Odia dictionaries, the editors cite phonetically or contextually resonant terms in the inline notes, recognizing that such lexemes may have faded from contemporary usage.

Interpretative Annotation: In the process of editing critical texts, it is often necessary to choose between variations of a phrase that appear to convey a similar meaning at first glance. However, closer examination may reveal significant differences in interpretation, making the editor's annotation crucial in justifying the chosen reading.

For example, one version (ଯାହାର ପ୍ରସନ୍ନେ or jāhāra prasanne) suggests that the blessing is contingent upon the deity's happiness: "if you are happy you bless them," while the other version (ଯାହାର ଦର୍ଶନେ or jāhāra darśane) indicates that the mere sight of the deity brings blessings: "when people see you they are blessed." At first glance, this may seem like a minor variation, as both versions emphasise the bestowal of blessings. However, upon closer examination, the difference in meaning could be significant. In the first version, blessings are conditional, depending on the deity's state of happiness or contentment. This introduces a relationship where divine grace must be earned through devotion, ritual, or the deity's pleasure. It suggests the importance of maintaining the deity's favour, presenting a more transactional or merit-based dynamic between the divine and the worshipper. In the second version, blessings are unconditional, conferred simply by witnessing the deity without any need to first please them. This interpretation shifts the focus to the power of the deity's presence itself, indicating that the deity's grace is abundant and accessible to all who behold them. The act of seeing the divine is enough to bring blessings, creating a more spontaneous and generous relationship. Choosing the second meaning highlights the immediacy and universality of the deity's grace, available to all who approach them with reverence. In contrast, the first version emphasises a more transactional nature of worship. While both interpretations share the central idea of divine blessing, the shift from a conditional to an unconditional framework significantly alters the understanding of divine grace and the worshipper's relationship with the deity.

This annotation not only justifies the editorial choice but also highlights how seemingly minor textual variations can lead to distinct interpretative frameworks. By carefully considering these differences, the annotation highlights the nuanced implications of each version and their broader significance in understanding the text.

Contextual Annotation: Sometimes, an annotation is made to highlight how the choice between phrases impacts the contextual interpretation, for example, in understanding the relationship between royal responsibilities and spiritual practices. The choice is between the phrases *ରାଜପଦେ ବସି*

(rājapade basi) and *ରାଜ୍ୟପଦ ଛାଡ଼ି* (rājyapada chhāḍi). The former conveys the meaning “while seated on the throne, why do you observe Ekadashi?” whereas the latter suggests “leaving the affairs of your kingdom, why do you engage in Ekadashi?” These variations reflect differing nuances in the interpretation of royal duties in relation to spiritual practices.

Such annotations illuminate the interconnected layers of linguistic, cultural, and religious meaning in the text. By situating Shantanu’s emotional state within these frameworks, annotations offer a richer interpretive lens for both readers and NLP systems, ensuring the preservation and accessibility of culturally embedded nuances. Similarly, by identifying misinterpretations due to letter shape similarities and contextualizing spelling variations, these annotations help resolve ambiguities that may arise in transcription. This approach, integrating paleographic analysis, semantic precision, and cultural awareness, mitigates transcription challenges and enhances the interpretability of texts, making them more accessible for both scholars and NLP systems.

3. CONCLUSION

Annotation practices rooted in textual criticism, a cornerstone of humanities research, provide an essential framework for preserving the authenticity and depth of linguistic data. These practices involve detailed, multi-dimensional annotations that account for historical, cultural, linguistic, and semantic nuances. By carefully documenting word usage, etymology, and contextual meanings, and variations as well as addressing transcription errors and interpretative ambiguities, they safeguard the subtleties often overlooked in automated or reductive approaches. Such annotations serve as a bridge between ancient or culturally specific texts and modern computational methods, ensuring the preservation and accessibility of linguistic diversity in classical literature, historical manuscripts, and regional dialects. The integration of these annotations into NLP systems enhances the ability to navigate complex linguistic features like ambiguity, metaphor, irony, and sentiment, with greater cultural and contextual sensitivity. As this discussion highlights, annotations are not merely tools for textual preservation; they are active contributors to the evolving intersection of humanities and computational linguistics. By leveraging the practices of textual criticism, one can enrich the authenticity of linguistic data and expand the horizons of NLP technologies, promoting a deeper and more inclusive understanding of language.

4. ACKNOWLEDGMENTS

I thank the Odisha Higher Education Programme for Equity and Excellence, Government of Odisha, India, for funding support for the digital edition of the Adi Parba of the Odia Sarala Mahabharata (ravenshawmahasarala.nic.in).

5. REFERENCES

[1] Bender, E. M. (2019). The #Bender Rule: On Naming the Languages We Study and the Languages We Use. ACL 2019.

[2] Bird, S., Klein, E. & Loper, M. 2009. Natural Language Processing with Python. O’Reilly Media.

[3] Bird, S., & Liberman, M. 2001. A Formal Framework for Linguistic Annotation. Speech Communication, 33(1-2), 23-60.

[4] Blodgett, S. L., Barocas, S., Dastin, J., & Wallach, H. 2020. Language (technology) is Power: A Critical Survey of “Bias” in NLP. ACL 2020.

[5] Charniak, E. 1993. Statistical Language Learning. MIT Press.

[6] Ide, N., & Pustejovsky, J. 2017. Handbook of Linguistic Annotation. Springer.

[7] Kress, G., van Leeuwen, T. 2001. Multimodal Discourse: The Modes and Media of Contemporary Communication. Edward Arnold.

[8] Labov, W. 1972. Sociolinguistic Patterns. University of Pennsylvania Press.

[9] Muller, T. 2016. Digital Humanities and Computational Linguistics: Exploring the Potential of Annotated Corpora. Language Resources and Evaluation.

[10] Tufekci, Z. 2014. Big Questions for Social Media Big Data: Representations and Biases in the Big Data Paradigm. Proceedings of the 2014 ACM Conference on Web Science.