Engaging with an Indian Epic: A Digital Approach

Urmishree Bedamatta
School of Languages,
Ravenshaw University,
Cuttack, Odisha

Bhabani Shankar Prasad Mishra
School of Computer Engineering,
KIIT Deemed to be University,
Bhubaneswar, Odisha

Santwana Sagnika
School of Computer Engineering,
KIIT Deemed to be University,
Bhubaneswar, Odisha

Anshuman Pattanaik
School of Computer Engineering,
KIIT Deemed to be University,
Bhubaneswar, Odisha

ABSTRACT
India's heritage texts have had a long history of being mined for knowledge of language and culture by Christian missionaries to India, colonial officers of the East Indian Company and the British Raj, German, European and American Indologists and later by native scholars driven by nationalist sentiments. It was during their investigative exercises that a vast body of India's heritage texts was recovered and made the subject of rigorous study. A large number of editions in English translation as well as in modern Indian vernacular languages started appearing on the scene. The focus then was primarily on patthoddhar [retrieval of the 'ur'-text] or making shuddhasanskarana [correct edition]. The exercise was purely manual and time-consuming and concentrated on a limited number of texts. But there still lies a vast treasure of ancient knowledge in India's palm leaf manuscripts, waiting to be discovered, deciphered and interpreted for contemporary readers and scholars. It is impossible to ignore the ubiquitousness of Information Technology based tools and the scope that they offer for large-scale data mining. Of late, a large body of historical texts is being made available digitally by repositories and institutions worldwide. The time is ripe for digitally inspired editions, beginning with studies in corpus linguistics. This paper throws light on the challenges to be addressed for the preparation of a digital historical corpus edition of Sarala Mahabharata, a local version of the famous Sanskrit Mahabharata by Vyasa, from Odisha in the eastern part of India.

General Terms
Natural Language Processing

Keywords
Sarala Mahabharata, Text Mining, Odia, Digitization, Palm leaf manuscript

1. INTRODUCTION
This article is inspired by a research project, led by the main author, on a critical edition of Sarala Mahabharata (SM, henceforth), an Odia epic purportedly written in the fifteenth century. The project team includes humanities scholars, linguists, computer science professionals and transcribers.

All about this epic, all epics for that matter, is between history and literature. However, when historical "evidences" are discussed, no evidence provided so far about Sarala's time and place is conclusive or wholly positive. Scholarship on the epic has remained largely referral or ambivalent in its opinion about whether the epic may be taken to represent historical time and circumstances. Every 'evidence' excavated from the messy corpus proves self-attenuating, which may be taken to elucidate, partly, "the Mahabharata problem" (Aurobindo's phrase). Before going to the nature of Sarala corpus, it is necessary to bring forth Aurobindo's statement on the possible accrual of cultural capital and scholarly acumen through finding a solution to the Mahabharata problem:

> For the solution of the Mahabharata problem is essential to many things, to any history worth having of Aryan civilisation and literature, to a proper appreciation of Vyasa's poetical genius and, far more important than either, to a definite understanding of the great ethical gospel which Sri Krishna camedown on earth to teach as a guide to mankind in the dark Kali Yuga then approaching. But I fear that if the inquiry is to be pursued on the lines the writer of this article seemed to hint, if the Society is to rake out 8000 lines […] and dub the result the Mahabharata of Vyasa, then the last state of the problem will be worse than its first. It is only by a patient scrutiny and weighing of the whole poem, disinterestedly, candidly and without preconceived notions, a consideration canto by canto, paragraph by paragraph, couplet by couplet, that we can arrive at anything solid or permanent. But this implies a vast and heart-breaking labour. (Aurobindo, p. 180; emphasis mine)

The article Aurobindo refers to is Velandai Gopala Aiyer's "The Date of the Mahabharata War" published in Indian Review (vol. II, January-December 1901), a monthly journal edited by G. A. Natesan. The subject of Aiyer's article is beyond the scope of this paper, which shall instead focus on the method problem of doing the Mahabharata, as hinted at by Aurobindo (see emphasis in the quote above).

2. LITERATURE REVIEW
Sarala Mahabharata in Odisha: The Mahabharata which was in wide circulation in early twentieth century Odisha is a mixture of the compositions of Vyasa, Sarola and Kashiram Das of Bengal. The composers are Phakir Charan Mohanty and Mohan Charan Das (1927, Manmohan Pustakalay Press). The introduction says thus: "As we found different versions in the compositions of the Odia poet Sarola Das and the Bengali poet Kashiram Das, we have sincerely rendered the matter of the eighteen parvas of the Mahabharata in simple, easy verses and published this compilation of the grantha for the benefit of the people". It is, therefore, not possible to say who actually is the composer of the Mahabharata published by Manmohan Pustakalaya Press.

Among the important Odia reworkings of Vyasa's Mahabharat are *Biramitrodaya Mahabharata* composed by HariharRath of Puri and published with a cash grant from the Maharaja of Sonepur, Biramitrodaya Singhdeo. (Gopal Chandra Praharaj, in the introduction to this book, writes, "This book is a translation of a redaction of Vyasa's Sanskrit

Mahabharat"); Bal Mahabharata: Prose Version of Mahabharata for Children" (1923) by Madhusudan Das; Nitiratnanidhi Mahabharata (1967) by Bishnu Prasad Mishra and A compilation of some special stories of Vyasa's Mahabharata, Gopinath Das's *Tika Mahabharata*, 18th century; Krushna Singh's translation of Vyasa's Mahabaharata, 1859. Even Phakir Mohan had translated Vyasa's Mahabharata, the copy of which is yet to be traced. Vyasa circulated in Odisha not only in the form of poetry but also through the prose composition of Gobinda Chandra Mohapatra. It started in 1892 through the medium of *UtkalPrabha* magazine. Thereafter, it took almost 15 years for Sarola to rise in his own land. He appears in 1913 in *chaudaakshari* and *dwipada* style through a publication of Purana Prakashan Company. Subsequently, many more readings were brought forth by other publishers. Sarola began to be interpreted, based on these mis(?)readings. At last, in 1964-66, Arta Ballabh Mohanty tries to reconstruct a picture of Sarola, which is today acknowledged as the standard edition of SM.

But none ventured to edit Sarola until the great litterateur Arta Ballav Mohanty, who, as Gauri Brahma, the author of the introduction to the text says, requested the government to support the exercise, and Arta Ballabh was appointed the "chief compiler" for the exercise.

In the early stages of the project, the purpose was primarily to reconstruct Sarala. ArtaBallav and his team came out with an edition of Sarala Mahabharata which was published by the Directorate of Culture, 1964-66. A preliminary review of Mohanty's approach (as is evident from the footnotes) yields no systematic account of the emendatory process adopted by the editor. The professor's study, apparently, was based on 22 palm leaf manuscripts available to him at the time. The preface to this edition, which was written by someone other than the editor called it *shuddhapatha*[corrected version]. And in the meagre footnotes one finds mention of କ, ଖ, ଗ, ଘ, etc. which are codes for around 21 different manuscripts (called 'witnesses' in critical edition projects) used for the production of the edition of 18 cantos. So uncritical is the apparatus that one might as well pick up any witness today and say here it is: this is the କ, ଖ, ଗ, or ଘ witness that Prof. Arta Mohanty used. But what queers the pitch is the finding of Gopinath Mohanty, an equally eminent researcher, that the author of the Mahabharata was not "Sarala" but "Sarola" and that the text was written in the tenth century and not in the fifteenth century (*Gabeshaka Drushtire Sarola Mahabharata*, [Sarola Mahabharata from a researcher's perspective; published in 2017, the book is a compilation of a series of articles which appeared in the Odia magazine *Jhankaar* in the 1950s]).

However, the cultural importance of ArtaBallav's edition was so great that Sarala Sahitya Sansad, a literary organisation, with financial assistance from the Odisha State Department of Culture, published a second edition. This edition not only deleted the meagre footnotes of the original edition but also inserted fresh mistakes, hundreds of them.

The present state of SM corpus: With so many copies of Sarala's text in Odisha State Museum, more than 70 copies in the Utkal University library and around 30 copies in the National Archives in Bhubaneswar, it is also not possible to say with any degree of certainty if the variant readings in the print editions would be found in any of these manuscript copies. There are hundreds and thousands of manuscript copies in private collections across the Odisha region, unlisted and undeciphered. Any edition of SM thus would be just an exercise in 'editorial best guess'. Institutional cataloguing of these palm leaf manuscript copies is largely incomplete and unreliable. For example, the handwritten catalogue of the Odisha State Museum cites 277 palm leaf manuscript copies of SM. These manuscripts are catalogued by arrival — some are purchased or collected while others have been donated. Cataloguing is mostly incomplete: For example, some manuscripts have the colophon listed against them, others do not. The reason for this is because not all manuscripts have the colophon at the end; one comes across colophonic information somewhere in the middle of the text. Identification of all the manuscripts according to time and provenance is a work in progress. No descriptive catalogue of the manuscript copies of SM exists, as of now.

3. PROPOSED WORK

The ideal situation for any attempt at a reconstruction of Sarala would be to study all the manuscript copies, taking into account all peculiarities of a copy as regards canto, paragraph and couplet. However it is through morphosyntactic forms that the particulars of variation become explicit. The variety and difference in scribal approach from word-formation to story construction, once made explicit, would probably make it clear that what really matters is not a 'correct' edition, for one doubts if there can be one, but a complete Sarala variorum for all to decipher. Among the 277 palm leaf manuscript copies in the Odisha State Museum listed by the project team, 50 copies, based on their readability, are being identified by time and provenance. These copies, most of which belong to the nineteenth and the first half of the twentieth century and are mostly from the coastal districts of Odisha, constitute the text corpus for our project. Just about eight per cent of the copies belong to the western and southern regions of the state. Hence, it is proposed to create a digital historical corpus edition beginning with linguistically annotated dataset, involving part-of-speech (POS) tagging and chunking, to enable the preparation of a critical edition, based on these manuscript copies, in future.

State of research in Odia corpora and the current project: Odia is an Indo-Aryan language. Corpus annotation in Odia has been done for modern Odia prose only. The Indian Languages Corpora Initiative (ICLI), has been, by far, the most ambitious attempt by the Government of India to develop parallel annotated corpora consisting of 25,000 sentences each from the domains of health, tourism and agriculture for eleven modern Indian languages, including Odia. POS tagging in ICLI was based on the POS tag set developed by the Bureau of Indian Standards (BIS) to ensure uniformity in POS tagging. However, as Vaz et al. (2012) report on tagging for Konkani language under ICLI, the BIS POS tag set, being lexically driven, proved to be inadequate for higher levels of natural language processing (NLP). Hence there came about a need for extensive manual post-processing to ensure the correctness of POS tagging. For Odia specifically, Das et al. (2015) developed a Support Vector Machine (SVM)-based tagger using the BIS guideline. However, no experiment on POS tagging an Odia historical variorum corpus has been done.

This project by the authors is the first-of-its-kind attempt in any Indian language to prepare a digital historical edition which also constitutes corpora. The texts of palm leaf manuscript copies are being digitally transcribed in Unicode, which is a labour-intensive and time-consuming process. Given the budgetary limitations and deadline restrictions, as well as the awkward and complex situation in which the digitized surrogates of the manuscript copies are needed to be

obtained from the museum, the experiment is necessarily to be conducted in a low-resource scenario with little or no training data. For pre-modern Odia historical corpora as identified for the project, there is no annotated training data. To begin with, it is proposed to use the ICLI tagger trained on modern Odia to tag the corpus. However, it is anticipated that there would be need for specialized taggers. Given the nature of the text corpus, as mentioned above, there are considerable variations in orthography, word forms and meaning giving rise to extensive variations at the level of the couplet. The text corpus is being prepared keeping the following in mind:

i) Faithful transcription: it constitutes authentic data as against the text of the printed editions which are heavily emended or constructed by the editor; it preserves dialectal variations

ii) Minimal editorial intervention except where readability is affected

iii) Multiple-level user interface to allow the user to suggest changes / corrections to the transcription and the editorial interventions as well as to evaluate the annotated language corpora.

iv) The format for presentation such that textual variations and the attendant linguistic annotation can be collated simultaneously

4. COMPUTING IN ODIA LANGUAGE: CHALLENGES AND OPPORTUNITIES

In this section the challenges to be addressed for preparing a digital historical corpus edition of Sarala Mahabharata are discussed.

4.1Issues of font: While processing non-English scripts, one has to grapple with inconsistent letters and keystrokes. In Odia, for example, desktop publishing commonly uses several fonts such as Kalinga, Saral, Ashoka, Aprant, Mahanadi and Tara Tarini. While these fonts have been in use traditionally, Unicode type is rarely used. The processing and porting of SM corpus need to be done in Unicode-based fonts, for generalized usage. Alternatively, other encodings might be used, but it needs to have wide browser support, in order to be accessible globally and to facilitate communication within the system. Moreover, the SM variorum constitutes texts written in variable letters. Some letters are no more in use in modern Odia and, hence, are not encoded in the Unicode / Unicode-based fonts. Digital representation of such letters is a major challenge.

4.2Issues of OCR: Optical Character Recognition refers to electronic identification of printed or handwritten text and making it available in a computer-readable, preferably ASCII format. OCR makes documents fully searchable. OCR technology permits reading of documents containing a mixture of fonts of various sizes and styles. [2] In this process, there might arise some problems, as mentioned [3]:

(i) Source documents such as palm leaf manuscripts have a rough surface and may be extremely degraded, i.e. blurred or faded, owing to which the OCR device cannot identify the text. Parts of these manuscripts are worm-eaten or broken, and hence the text may not be read wholly.

(ii) Even if one finds a good readable manuscript, the OCR input still needs to go through error and grammatical correction, for further comprehension and processing like summarization or translation.

(iii) Odia palm leaf manuscripts are written in *karaninabaja* [running script and a language 'incorrect' by modern standards; *karani*, however, may simply be taken to mean the Odia script before it was standardized]. Punctuation marks are very rare and there is no spacing in between words, which leads to a laborious and time-consuming post-processing phase.

4.3 Tagging a variorum corpus: With variable spellings, words and word forms, one may opt for normalisation to plan a POS tag set, but this would not help in preserving the dialectal and other variations needed for diachronic study of language. In such a scenario, information extraction and mining by search engines or other tools can be severely affected.

4.4Issues of collation: Collation refers to character-by-character comparison of texts, traditionally by editors to process transcripts and identify errors and reliability of copies [4]. In most algorithms, collation is done based on a collation sequence, specific to the application in hand. Algorithms like these are quite complex and require multiple passes over the text. When the algorithm works on different languages, sequencing becomes difficult and ambiguous. Collation also faces trouble with processing numbers, especially decimals. IBM knowledge center provides collation standards for various languages, including Odia [5]. Due to variations in SM, the digital system needs to project different versions and highlight the changes, which include morphological changes, semantic variations and their respective POS tags simultaneously collated.

4.5Issues in topic modeling: Topic modeling deals with identifying abstract topics that occur in a document, normally using statistical methods. Current machine learning models perform topic modeling with considerable accuracy, but they do face a few issues. Some of them are mentioned [6].

(i) As the models work on purely unsupervised learning, their performance is not up to the standard so as to be able to eliminate human intervention.

(ii) Unsupervised models can perform better after some generalization of supervised learning, which is difficult to implement, due to lack of sufficient annotated resources.

(iii) For languages like Odia having limited corpus, it is difficult to perform an effective unsupervised learning technique.

(iv) Sarala Mahabharata contains verses, which are not divided into chapters or segments. Proper segmentation is an ambiguous and subjective matter, which can affect the process of topic modeling.

(v) Effective topic modeling is dependent on proper part-of-speech tagging and sentiment analysis of the text.

(vi) Proper word representations or embeddings are still not available for Odia language, which restricts the capabilities of any machine leaning model.

4.6 Issues in summarization: Text Summarization[7] is one of the biggest issues in natural language processing(NLP). Automatic generation of summary out of text data suffers challenges like coverage of context [8], data redundancy, correlation between the sentences of the summaries and many more. In the case of historic data, the summarization method needs special attention because it may suffer event, data loss which hamper the original context. We discuss regional language as well as the historical data focusing on Sarala's timeline, and more specifically SM [10]. In any language for natural language processing, we need a rich corpus, tools, web dictionaries and word vectors.

5. CONCLUSION

A few recent attempts in computational analysis in Odia language include lemmatization and developing a tag set for sentence stops and pauses. However, these attempts are project-driven and the output of the project is not shared by the lead investigators. Hence, researchers are forced to develop modules from scratch, leading to delayed attempts at enhancing the features of software which are already developed. Moreover, owing to the extremely varied nature of the corpus data which is not generally shared by researchers, it is not known if any given software could be replicated for use. In such cases, Python, a programming language, is most preferred for natural language processing as it contains various libraries and packages such as NLTK, Stanford NLP, Spacy and many more to handle highly unstructured text data. These packages can be used for simple operations such as performing tokenization and identifying regular expressions. Algorithms understand only numeric representations, ascribing numbers to words such that each word is represented by different sets of numbers. Unlike in the English language in which Word2Vec is available, vector format is yet to be developed for Odia language. Semantic similarities between sentences are a concern. Odia language is enriched with ornamental words and many words have different meanings depending on the context. Named Entity Recognition (NER) is also one of the vital issues in Odia data. Preparation of the digital historical corpus edition of Sarala Mahabharata, thus, is a ground zero scenario. It is not only the manuscripts which need to be digitally transcribed for the first time but also every unit of the Odia language system as represented in the Sarala Mahabharata corpus which needs to be manually arranged to constitute separate and elaborate datasets to be exploited for natural language processing.

6. ACKNOWLEDGEMENT

This paper is the result of preliminary work done as part of a project funded by Odisha Higher Education Programme for Equity and Excellence (OHEPEE) assisted by World Bank, at Ravenshaw University, India.

7. REFERENCES

[1] SriAurobindo. "Valmiki and Vyasa: Notes on the Mahabharata". *The Harmony of Virtue: Early Cultural Writings 1890-1910*. pp. 179-196. Sri Aurobindo Ashram, Pondicherry (1972).

[2] Ahmed, Fazluddin. *Digitization as a Means of Preservation of Manuscripts: Case study of Osmania University Library*. (2009).

[3] Arlitsch, Kenning, and John Herbert. "Microfilm, paper, and OCR: Issues in newspaper digitization. the Utah digital newspapers program." *Microform & Imaging Review* 33.2 (2004): 59-67.

[4] collation | Lexicon of Scholarly editing. Available: http://uahost.uantwerpen.be/lse/index.php/lexicon/collation/

[5] Unicode Collation Algorithm based collations. Available: https://www.ibm.com/support/knowledgecenter/en/SSEPGG_11.5.0/com.ibm.db2.luw.admin.nls.doc/doc/r0050489.html

[6] Blei, D. "Probabilistic Topic Models: Origins and Challenges." 2013 Topic Modeling Workshop at NIPS. 2013.

[7] Luhn, Hans Peter. "The automatic creation of literature abstracts."*IBM Journal of Research and Development* 2.2 (1958): 159-165

[8] Pattanaik, Anshuman, et al. "Extractive Summary: An Optimization Approach Using Bat Algorithm."*Ambient Communications and Computer Systems*.Springer, Singapore, 2019.175-186.

[9] Mishra, Mahendra Kumar. "A Hero of the Mahabharata in the Folklore of Central India."*The Mahabharata in the Tribal and Folk Traditions of India* (1993): 157-170.

[10] Bhoi, Debendra Nath, and PriyadarshiniBakshi. "SaralaDasa, The Originator of the Oriya Literature."*Orissa Review* (2004): 57.

[11] Loper, Edward, and Steven Bird. "NLTK: the natural language toolkit." arXiv preprint cs/0205028 (2002).

[12] Nadeau, David, and Satoshi Sekine. "A survey of named entity recognition and classification." *LingvisticaeInvestigationes* 30.1 (2007): 3-26.